



UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ DEGLI STUDI DI FIRENZE
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)
CORSO DI DOTTORATO IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM: INFORMATICA

VISION AND LANGUAGE TASKS:
APPLICATIONS TO REAL SCENARIOS
AND IMAGE QUALITY ASSESSMENT

Candidate

Ing. Pietro Bongini

Supervisors

Prof. Andrew D. Bagdanov

Prof. Alberto Del Bimbo

PhD Coordinator

Prof. Fabio Schoen

CICLO XXXV, 2019-2022

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Engineering. Copyright © 2023 by
Ing. Pietro Bongini.

Abstract

The human brain has always been one of the most fascinating fields of study. The first theories and research results about machine learning date back to around fifty years ago, but only in the last few years - thanks to increasing computational power - these theories have been put into practice with applications in different fields such as autonomous driving, human-computer interaction, medical imaging, and many others. Perception is perhaps the most important way humans understand the physical world, and language is how humans communicate their experiences. For this reason, the integration of vision and language has been gaining attention and language-aligned visual features have been shown effective for vision-language tasks. Recently these tasks have received significant attention from the Artificial Intelligence community, however many tasks in this field are far from solved and require further research. In this dissertation we focus on three vision and language tasks: Visual Question Answering (VQA), Image Captioning (IC), and Cross-Modal Retrieval (CMR).

Visual Question Answering systems are capable of answering visual questions (that is, questions referring to the semantic content of images), but a significant limitation is the inability to answer contextual questions (that is, those referring to image content but that require external information to be answered). For this reason, we investigate the use of external knowledge in support of answer generation. In the first part of this thesis, we propose two approaches to handle and extract external textual information and improve VQA in the Cultural Heritage domain - a domain where external information is crucial. Moreover, we propose a data collection and annotation technique, as well as a large dataset for VQA in the Cultural Heritage domain.

In the second part of this thesis, we investigate the application of Image Captioning to Image Quality Assessment (IQA). IQA is the task of evaluating the perceptual quality of images. IQA approaches are severely limited by the lack of data for training. After preliminary work on generative data augmentation, we propose a completely novel approach to exploiting visual captioning in order to infer quality scores in both No-Reference and Full-Reference scenarios.

Finally, Cross-Modal Retrieval approaches perform ranking of images based on text (and vice versa) at a merely descriptive level (focusing on what objects are in the image and their number). To address this problem, in the last part of this thesis we propose a new architecture that exploits scene

text to improve the performance of cross-modal retrieval tasks on multiple datasets that vary in the percentage of scene-text images and the type of caption (contextual, visual).

Keywords: machine learning, deep learning, multi-modal learning, visual question answering, image captioning, cross-modal retrieval, image quality assessment, representation learning, computer vision.

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.

International Journals

1. L. Galteri, L. Seidenari, **P. Bongini**, M. Bertini, AD. Bimbo. “LANBIQUE: LANguage-based Blind Image QUality Evaluation”, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18 issue 2s, 2022.

Submitted

1. F. Becattini, **P. Bongini**, L. Bulla, AD. Bimbo, L. Marinucci, M. Mongiovi, V. Presutti “VISCOUNT: A Large-Scale Visual and Contextual Question Answering Dataset for Cultural Heritage”, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

International Conferences and Workshops

1. **P. Bongini**, R. Del Chiaro, AD. Bagdanov, AD. Bimbo. “GADA: Generative adversarial data augmentation for image quality assessment”, in *International Conference on Image Analysis and Processing (ICIAP)*, 2019.
2. **P. Bongini**, Becattini, AD. Bagdanov, A. Del Bimbo “Visual question answering for cultural heritage”, in *IOP Conference Series: Materials Science and Engineering*, 2020.
3. L. Seidenari, L. Galteri, **P. Bongini**, M. Bertini, A. Del Bimbo “Language based image quality assessment”, in *ACM Multimedia Asia, 2021 (Best paper award)*.

4. **P. Bongini**, F. Becattini, A. Del Bimbo “Is GPT-3 all you need for Visual Question Answering in Cultural Heritage?”, in *European Computer Vision Conference (ECCV) Workshop on Vision and Art (VisArt)*, 2022.

Submitted

1. **P. Bongini**, AF. Biten, AM. Delgado, D. Karatzas, AD. Bagdanov, “STILT: Scene-Text Image and Language Transformer for Cross-Modal Retrieval”, *Computer Vision and Pattern Recognition*, 2023.

Demo

1. F. Vannoni, **P. Bongini**, F. Becattini, AD. Bagdanov, Alberto Del Bimbo “Data Collection for Contextual and Visual Question Answering in the Cultural Heritage Domain”, in *International Conference on Pattern Recognition (ICPR)*, 2020.

Contents

Publications	vii
Contents	ix
1 Introduction	1
1.1 Vision and Language Tasks	1
1.2 Visual Question Answering	2
1.3 Cross-Modal retrieval	7
1.4 Language and Visual Quality	9
1.5 Contributions of this thesis	13
2 Visual Question Answering for Cultural Heritage	17
2.1 Introduction	17
2.2 Method	19
2.2.1 Visual Question Answering with visual and contextual questions	19
2.2.2 Question Classifier Module	20
2.2.3 Contextual Question Answering Module	21
2.2.4 Visual Question Answering Module	21
2.3 Experimental Results	21
2.3.1 Datasets	22
2.3.2 Experimental Protocols	23
2.3.3 Experimental results	24
2.4 Conclusions	26
3 Is GPT-3 all you need for Visual Question Answering in Cultural Heritage?	27
3.1 Introduction	27

3.2	GPT-3	28
3.3	Method	29
3.4	Experiments	31
3.4.1	Dataset	31
3.4.2	Experimental Protocol	32
3.4.3	Experimental Results	34
3.5	Qualitative Analysis	36
3.6	Considerations on complexity and accessibility of GPT-3 . . .	38
3.7	Conclusions	38
4	VISCOUNTH: A Large-Scale Visual and Contextual Question Answering Dataset for Cultural Heritage	41
4.1	Introduction	42
4.2	Building VISCOUNTH: A large Visual and Contextual Question Answering Dataset for Cultural Heritage	43
4.2.1	A Semi-Automatic Approach for Generating the VQA Dataset	44
4.2.2	A Large and Detailed VQA Dataset for Cultural Heritage	48
4.3	A VQA Model for Cultural Heritage	50
4.4	Results and Discussion	52
4.4.1	Evaluation Metrics	52
4.4.2	Evaluation	52
4.4.3	Qualitative Analysis	58
4.5	Conclusions	59
5	GADA: Generative Adversarial Data Augmentation for Image Quality Assessment	61
5.1	Introduction	62
5.1.1	Auxiliary Classifier GANs.	62
5.2	Generative Adversarial Data Augmentation for NR-IQA . . .	63
5.2.1	Overview of Proposed Approach	63
5.2.2	Auxiliary Classifier GANs for NR-IQA	64
5.2.3	The GADA Architecture	65
5.3	Experimental Results	66
5.3.1	Comparison with the state-of-the-art	68
5.4	Conclusions	69

6	Language Based Image Quality Assessment	73
6.1	Introduction	74
6.2	Image Restoration	75
6.3	Evaluation Protocol	76
6.3.1	Subjective Evaluation	78
6.4	Results	79
6.4.1	Language Based IQA	79
6.4.2	Comparison with MOS	83
6.4.3	Comparison with Full-Reference Metrics	84
6.4.4	Comparison with No-Reference Metrics	85
6.5	Conclusion	86
7	LANBIQUE: LANGUAGE-based Blind Image QUality Evaluation	87
7.1	Introduction	88
7.2	Image Restoration	90
7.3	Evaluation Protocol	92
7.3.1	Evaluation with Reference Captions	93
7.3.2	Evaluation without Reference Captions	94
7.3.3	No-Reference Evaluation	95
7.3.4	Subjective Evaluation	96
7.4	Experimental Results	98
7.4.1	Results on JPEG Artefacts	98
7.4.2	Results on all distortions	105
7.5	Conclusion	108
8	STILT: Scene-Text Image and Language Transformer for Cross-Modal Retrieval	111
8.1	Introduction	112
8.2	The STILT Approach	113
8.2.1	Model Architecture	113
8.2.2	Pre-Training	114
8.2.3	Implementation Details	116
8.3	Experimental Results	116
8.3.1	Datasets	117
8.3.2	Evaluation Method	118
8.3.3	Cross-Modal Retrieval with Abstract Image-Text Alignment	119

8.3.4	Full Scene-Text Image and Text Cross-Modal Retrieval	120
8.4	Qualitative Analysis	121
8.5	Conclusions	121
9	Conclusions	123
9.1	Summary of Contributions	123
9.2	Directions for Future Work	124
A	Data Collection for Contextual and Visual Question Answering in the Cultural Heritage Domain	127
A.1	Introduction	128
A.2	Data Collection	128
A.3	Visual and Contextual Question Answering	129
A.4	Conclusion	130
B	Publications	131
	Bibliography	133

Chapter 1

Introduction

In recent years, research in Artificial Intelligence has increased significantly, giving birth to new applications in real scenarios such as facial recognition used on phones, object detection used in the medical field, object tracking, etc. AI algorithms are gradually becoming more and more accurate but still have weaknesses that make more research in this field necessary. For example, many methods cannot be used in real scenarios due to their complexity and the fact that they only work in specific domains.

In this thesis, we focus on a specific area of Computer Vision concerning Vision and Language. Research in this area concerns both applications in real scenarios and applications to other research topics such as Image Quality Assessment. In Sec. 1.1 we present an overview on vision and language tasks, in Sec. 1.2 and Sec. 1.3 we introduce, respectively, the Visual Question Answering task and the Cross-Modal Retrieval task. Sec. 1.4 gives an introduction to language and visual quality where we describe the Image Captioning and Image Quality Assessment tasks and how they can be combined together. Finally, in Sec. 1.5 we outline the contributions of this thesis.

1.1 Vision and Language Tasks

Computer vision (CV) and Natural Language Processing (NLP) are two of the most studied sub-fields of Artificial Intelligence (AI). The aim of these two tasks is to replicate the human activity for vision (CV) and language (NLP). In the last decade, research in these two fields has led to impressive

improvements in the performance of models on a wide range of tasks. This improvement is also due to the evolution of the hardware that allows faster training of large models and the collection of larger datasets.

Thanks to this hardware empowerment and the rise of AI research, a series of powerful neural networks have been developed. Traditional neural networks are typically multi-layer perceptrons (MLPs) consisting of multiple stacked linear layers and nonlinear activations [103, 104]. LeCun et al. [63] (1998) proposed Convolutional Neural Networks (CNNs) to incorporate shift-invariance as a better inductive bias for 2D visual input, which inspired a large number of deep neural networks, including AlexNet [62], VGGNet [115], GoogleNet [119], and ResNet [42]. Another seminal breakthrough were Recurrent Neural Networks (RNNs) in the field of natural language processing (NLP), which are incorporate recurrent cells for sequential data modeling [44, 105]. To mitigate the vanishing and exploding gradient problems in long sequence training, LSTM [44], a type of RNN, and GRU [23], a more efficient version of LSTM, were proposed. Another important breakthrough in NLP is the Transformer [128], which utilizes an attention mechanism to learn better language representations. Using a sequence of attention layers, Transformers can globally fuse information over language tokens, leading to rich and powerful models.

As Bisk et al. [9] explained in their work, visual perception is the most important way humans understand the physical world and language is how humans communicate their experiences. For this reason, the integration of vision and language receives much attention, and it has been amply demonstrated that language-aligned visual features are effective for vision-language tasks. Many visual-language tasks have been studied, the most relevant of them to our work are Visual Question Answering (VQA) [2, 3, 69, 120, 143, 144], Image Captioning (IC) [2, 25, 65, 150], and Cross-Modal Retrieval (CMR) [20, 66, 78].

1.2 Visual Question Answering

A Visual Question Answering system takes an image and a related question as input. The aim of the system is to generate the correct answer. In almost all state-of-the-art works this task is seen as a classification problem where the answer is chosen from a predefined dictionary. For example, the reference dataset VQA v2 [40] has a set of around 2K pre-defined answers.

Antol et al. [3], in the first work on VQA, propose an architecture in which a CNN [115] is used to encode the image, and an LSTM is used to encode the question. The output image embedding and question embedding are simply fused by point-wise multiplication and then passed through a linear layer followed by a softmax to output the probability of each candidate answer.

From this first work we observe that a typical VQA model consists of three different sub-modules: the first for image encoding, the second for question encoding, and a final sub-module for visual and textual information combination and answer prediction. Most works on VQA add novelty to one or more of these components.

Since merging image representation and question representation as in [3] is too simple and coarse, numerous works investigate the use of attention mechanisms for fusing image and question. In particular, Yang et al. [144] proposed a Stacked Attention mechanism adding multiple attention layers. At each attention layer, the question representation is used as a query for the attention mechanism. A groundbreaking approach for image representation was proposed by Anderson et al. [2] who employ a bottom-up attention mechanism based on salient objects in images. Instead of considering the entire image as a grid (as done in previous methods [3, 144]) they use object features as attention candidates. These features are extracted using a detector such as Faster R-CNN [38] trained on the Visual Genome dataset [61]. This technique was an important step forward for the VQA community and increased VQA performance considerably. After this work most of the subsequent approaches used just such visual features for the VQA task.

More recently, with the advent of Transformer models [128], multiple approaches based on this architecture were proposed. Tan et Al. [120] designed an architecture consisting of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. Fig. 1.1 gives a diagram of this architecture. The object relationship encoder (that uses the same visual features as [2]) and the language encoder have the same number of self-attention layers. The cross-modality encoder consists of multiple cross-attention and self-attention layers (N_e). As in other works [66, 69] the model is pre-trained with a large number of image-sentence pairs via diverse representative pre-training tasks: masked language modeling, masked object prediction (feature regression and label classification), cross-modality matching, and image question answering. Other works about Multi-modal Transformers showed how attention layers [29, 68, 76] and Visual-Textual

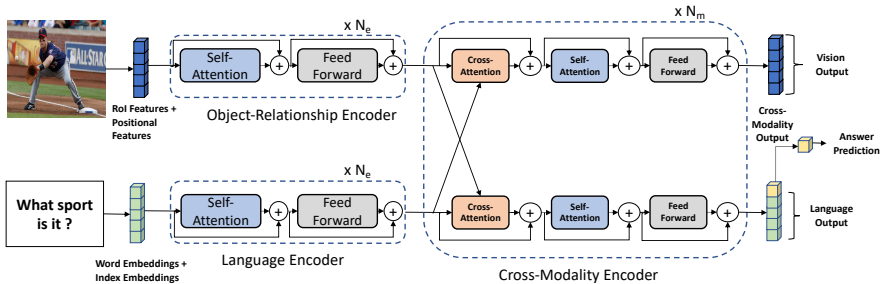


Figure 1.1: The LXMERT [120] model for learning vision-and-language cross-modality representations. Image and question are encoded separately through an object-relationship encoder (image) and a language encoder (question) then a cross-modality encoder merges together the image and text embedding to predict the answer.

features alignment [66, 69] are effective for multi-modal tasks.

Existing VQA approaches are capable of correctly answering visual questions related to image content, but are not capable of answering questions whose answer requires external knowledge. In order to answer this type of question, the VQA community recently introduced a new task known as Knowledge-Based VQA (KB-VQA). In the first work in this direction [81] the authors show the performance of standard VQA approaches OK-VQA [81]. Fig. 1.2 shows some example image-caption pairs from the OK-VQA dataset. We see that the questions belong to completely different categories that their answers require different knowledge. Garderes et al. in [37] use a knowledge-based graph based on ConceptNet [117] as contextual knowledge. The question is encoded with a BERT [27] model and then passed to two different Transformers one for image-question attention and the other for KG-embeddings-question attention. This architecture allows considering both visual information and contextual information from the knowledge graph to infer the correct answer. Following this approach, other state-of-the-art techniques use Knowledge Graphs to address this problem. The weakness of these approaches lies precisely in the use of Knowledge Graphs which leads to multiple problems:

- Standard Knowledge Graphs [79, 117] are extremely large and impossible to process entirely during training and testing.



Question: Levi is a popular brand of what item shown here?
Answer: jeans



Question: Is this a healthy meal or unhealthy meal?
Answer: healthy



Question: What nationality is this chair?
Answer: canadian



Question: Is this a modern or ancient system?
Answer: modern



Question: What city does the batter play for?
Answer: Baltimore



Question: In this country what numbers do you dial for emergencies?
Answer: 911

Figure 1.2: Samples of image-question pairs from the OK-VQA [81] dataset. These questions require external knowledge to infer the correct answer.

- As a consequence, Knowledge Graphs are usually pruned and the loss of information in the resulting graph leads to a drop in VQA performance.

Most recent approaches [41, 143] solve this problem by prompting GPT-3 [14] to generate contextual text useful to answering the question. These systems exploit the fact that GPT-3 is trained with massive amounts of data and is capable of generating realistic and correct information.

As described above VQA approaches obtain high performance but they still have important problems making it of limited use in real scenarios:

- They are capable of answering only visual questions referring to the image content and are not capable of answering contextual questions whose answers require external information.
- Visual questions are very simple (e.g. what color is...? what is on...?) and visual answers consist of a reduced number of words (varying from one to three). This weakness prohibits longer and more comprehensive answers.
- VQA does not easily incorporate structural data like ontologies, documents, knowledge graphs, etc.

In this thesis, we address this problem by proposing multiple solutions to the application of the VQA task in real scenarios. In particular, since a key problem of VQA answering contextual questions that are in many domains the most frequently asked type of questions by people, we propose two methods based on external knowledge to answer both visual and contextual captions. Our methods allow VQA to generate a comprehensive answer for contextual questions. We apply these methods to the Cultural Heritage Domain where most questions of interest involve information not deducible from the Artwork (e.g. When was the painting painted? Who is the person portrayed?). To the best of our knowledge, we are the first to apply VQA in the Cultural Heritage domain. Moreover, since no annotated dataset for VQA is available for Cultural Heritage, we propose a data collection and annotation framework for Visual Question Answering and we collect a large-scale dataset for Cultural Heritage VQA.

1.3 Cross-Modal retrieval

In general, cross-modal retrieval is the task of ranking samples expressed in one modality according to a specific query in another modality. Taking into account image and text modalities, the cross-modal retrieval task can be split into two sub-tasks:

- Image-to-text retrieval (I2T). Given an image, the aim of this task is to rank texts according to the similarity to the content of the image.
- Text-to-image retrieval (T2I). Given a text, the aim of the task is to rank images according to the similarity to the information in the text.

Cross-Modal Image-Text retrieval approaches can be divided into two main categories:

- Coarse-grained correspondence methods. The architecture consists of two branches: one for image encoding and the other for text encoding. The image features and text features are mapped in a joint embedding space to learn visual-semantic similarity.
- Fine-grained correspondence methods. The network learns to align relevant regions of the image and text words. Often these methods use some attention mechanism to learn image-text correspondence.

Regarding coarse-grained matching methods, Kiros et al. [59] use a CNN to encode the image and an LSTM to encode the text. In [141] the authors represent text with term frequency-inverse document frequency (TF-IDF) representation and pass them through a fully connected layer and the image is encoded with a deep convolutional network (DCNN). In [43] the text is represented by a matrix of continuous vectors each representing a single word and it is fed to a CNN. These approaches map image and text directly into a common latent space and learn to pair image and text through a variety of contrastive ranking losses [32, 96, 122, 131].

Subsequent approaches focus on capturing fine-grained correlations. In particular, Li et al. [67] propose a Visual Semantic Reasoning Network (VSNR) where image region features obtained as in [2] are fed to a Graph Convolutional Network (GCN) to perform visual semantic reasoning and generate features enriched with semantic relationships. Subsequently, these features are given as input to a memory mechanism that gradually generates the representation for the whole scene. This model is trained by optimizing

both matching (final representation and caption) and text generation (from final representation to caption). In [18] the authors capture correspondences between image and text progressively with multiple steps of alignment and attention.

As for VQA and Image Captioning, the advent of Transformers boosted the performances of the Cross-Modal retrieval task. Some of these models are described in the subsections related to VQA and Image Captioning [66, 69, 76]. In VILLA [35] the authors propose a two-step training approach exploitable for multiple vision and language tasks. In [20] the authors design an effective pretraining phase based on three tasks: Masked Language Modeling (MLM), Image-Text-Matching (ITM), and Masked Region Modeling (MRM). The model is then finetuned on different vision and language tasks, achieving state-of-the-art performances. Jia et al. [50] leverage a one billion image alt-text pairs dataset for vision and language pretraining. In particular, they show that pretraining a vision and language model with noisy text annotations achieves impressive performance in zero-shot vision and language tasks without finetuning.

Finally, a recent work by Redford et al. [96] the image is encoded with a Vision Transformer (ViT) [28] and the caption is encoded with a BERT Transformer; the model is trained with a large dataset of 400M images (taken from web) using InfoNCE loss and achieving state-of-the-art performance.

Despite the intense research in this field, cross-modal image-text retrieval approaches still have many limitations:

- Their performance is acceptable only when image-text pairs are well-aligned and the overlap between images and text is significant (both image and text mention or show the same objects).
- They do not work in real scenarios where image and text contain complementary information, and for this reason their alignment is abstract and symbolic. Fig. 1.3 shows examples of images and compares the visual description with a real complementary description that contains more contextual information.
- Only a few methods exploit additional information like scene-text to improve performance, but consequently, have a reduction in performance in scene-text-free context.

In this thesis, we address these problems by investigating the use of scene text in image-text cross-modal retrieval in real scenarios. News datasets

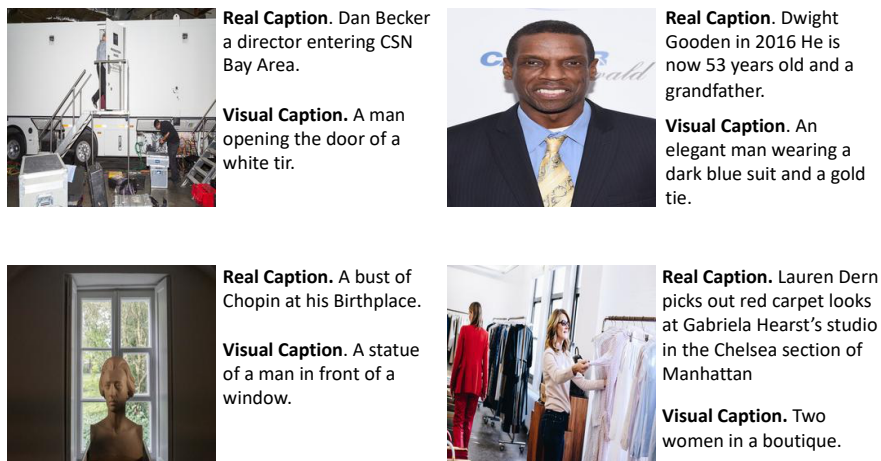


Figure 1.3: Comparison between visual description and real description for images taken from news articles from GoodNews [10] dataset. We can observe that real descriptions contain additional information not deducible from the image content.

represent a real environment where an image-text retrieval system can be a useful instrument. These datasets as described above and illustrated in Fig. 1.3 are very challenging for current cross-modal retrieval methods since the alignment between the content of the image and the caption is abstract.

We propose a transformer-based architecture that exploits scene-text to perform cross-modal retrieval. Experiments show the effectiveness of our approach and our performance surpasses state-of-the-art methods on news datasets. Finally, we show that our approach also reaches state-of-the-art performance on full scene-text datasets.

1.4 Language and Visual Quality

Image captioning is the task of describing the visual content of an image in natural language. It requires a visual understanding system and a language model capable of generating meaningful and syntactically correct sentences. Like VQA models, Image Captioning models have improved significantly in the last few years.

The standard image captioning systems consist of two sub-modules: a

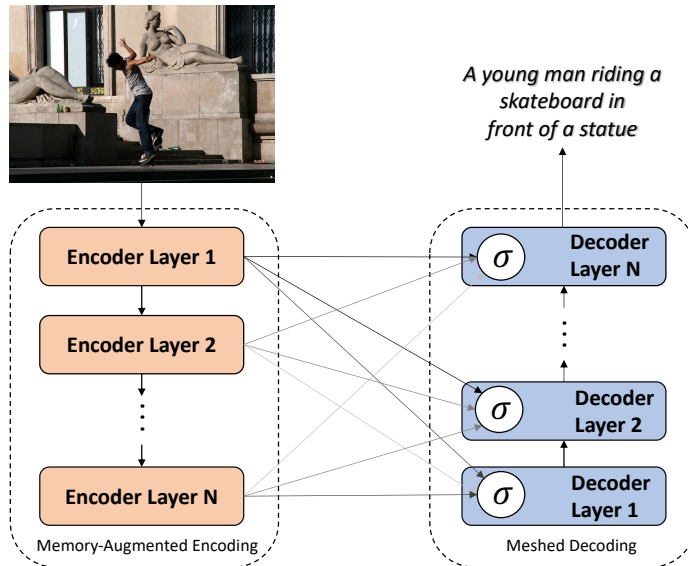


Figure 1.4: Meshed Memory Transformer [25] for image captioning architecture. The encoding phase consists of multiple levels and generates a multi-level representation of the relationships between image regions, exploiting learned a priori knowledge. These multi-level encodings of image regions are connected to a language decoder through a meshed and learnable connectivity to exploit low-level and high-level features.

visual encoder that takes image pixels as input and outputs visual features, and a language model that takes visual features (generated by the visual encoder) as input and produces a sequence of words according to a dictionary.

In the very first captioning approaches, Vinyals et al. [130] use a Convolutional Neural Network (CNN) to produce a rich representation of the image then an LSTM is used to generate the description. In [139] the authors add an attention mechanism in the LSTM in order to improve the model’s capability to focus on the correct regions during caption generation. Since the generated captions tend to not involve all the salient objects in the image, Cornia et al. in [24] propose a model capable of generating a caption on explicitly selected image regions.

As described above for VQA, taking the features of detected objects as image embeddings [2] leads to impressive improvements for the captioning

task. These visual features are also used by other captioning methods [69, 113, 150].

With the advent of Transformer models, multiple studies were carried out showing the effectiveness of this architecture. Some of these works use object features [2] as image embeddings. In particular, OSCAR [69] uses object tags detected in images as anchor points to significantly ease the learning of image-text alignments. In [25] Cornia et al. proposed a mesh-like connection between encoding and decoding layers weighted through a learnable gating mechanism. Fig. 1.4 shows a scheme of the approach. Image regions are encoded through multiple encoded layers to learn object relationships between image regions. Each encoding layer is connected to a language decoder through a meshed and learnable connectivity. In VinVL [150] the authors improve the object detector of [2] generating richer image encodings feeding them into OSCAR Transformer [69].

Since the image representation is strongly limited by the power of the object detector and the extraction of region-based features is computationally expensive, other Transformer-based architectures remove this object detector and perform direct alignment between the image and text representations. In VILT [58] the authors use a simple linear projection that operates on image patches in order to obtain visual embeddings. In SimVLM [136] the model takes row images as input and is trained with weakly labeled datasets. Transformers achieve state-of-the-art performance for many vision and language tasks, but they present two problems: self-attention for visual sequences is significantly more expensive than textual sequences; and there is an asymmetry between caption text, which is usually short and contains abstract information, and image that contains more detailed information. In order to address these problems Li et al. [65] designed an asymmetric Transformer architecture based on skip-connection achieving State-Of-The-Art in different Vision and Language tasks.

In this thesis, we use Image Captioning as an auxiliary task for Image Quality Assessment. To the best of our knowledge, we are the first to combine these two tasks. Image Quality Assessment (IQA) [135] refers to a range of techniques developed to automatically estimate the perceptual quality of images. IQA estimates should be highly correlated with quality assessments made by multiple human evaluators (commonly referred to as the Mean Opinion Score (MOS) [93, 111]). IQA has been widely applied by the computer vision community for applications like image restoration [6],

image super-resolution [127], and image retrieval [142].

IQA techniques can be divided into three different categories based on the available information on the image to be evaluated: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA). Although FR-IQA and RR-IQA methods have obtained impressive results, the fact that they must have knowledge of the undistorted version of the image (called the reference image) for quality evaluation, makes these approaches hard to use in real scenarios. On the contrary, NR-IQA only requires the knowledge of the image whose quality is to be estimated, and for this reason, is more realistic (and also more challenging).

In the last few years, Convolutional Neural Networks (CNNs) have obtained great results on many computer vision tasks, and their success is partially due to the possibility of creating very deep architectures with millions of parameters, thanks to the computational capabilities of modern GPUs. Massive amounts of data are needed for training such models, and this is a big problem for IQA since the annotation process is expensive and time-consuming. In fact, each image must be annotated by multiple human experts, and consequently, most available IQA datasets [93, 112] are too small to effectively train CNNs from scratch.

With the advent of deep neural networks, multiple works have applied them to learning for NR-IQA. These techniques require large amounts of data for training and IQA datasets are especially lacking in this regard. Therefore, to address this problem different approaches have been proposed. Kang et al. [52] use small patches of the original images to train a shallow network and thus enlarging the initial dataset. A similar approach was presented in [53] where the authors use a multi-task CNN to learn the type of distortion and the image quality at the same time. Bianco et al. [8] used a pre-trained DCNN fine-tuned with an IQA dataset to extract features, and then train a Support Vector Regression model that maps extracted features to quality scores. Liu et al. in [74] use a learning-from-rankings approach. They train a Siamese Network to rank images in term of image quality and subsequently the information represented in the Siamese network is transferred, through fine-tuning, to a CNN that predicts the quality score. Another interesting work is from Lin et al. [71] who use a GAN to generate a hallucinated reference image corresponding to a distorted version and then give both the hallucinated reference and the distorted image as input to a regressor that predicts the image quality.

In the last couple of years with the development of new larger datasets (KonIQ10K [45], PieApp [94], PIPAL ([51])) for IQA, more complex and weight architecture have been proposed for this task. In particular, Transformer-based approaches obtain impressive performances. You et al. [146] propose a shallow Transformer architecture inspired by ViT adding an adaptive positional embedding strategy to handle images with different resolutions. Cheon et al [22] inspired by [146] designed a Transformer architecture to compare distorted and reference images in a full-reference manner. Finally, in [54] Ke et al. designed a multi-scale image quality Transformer capable to process images with varying sizes and aspect ratios

In this thesis, we address the limitations of Image Quality Assessment systems with three different approaches. In a first work we design a data augmentation approach based on Auxiliary Classifier-GAN (AC-GAN). We train this architecture to generate new distorted samples according to a specific distortion type (JPEG, Blur, White Noise, etc.) and a specific distortion value. A very shallow quality evaluator is then trained with both original and augmented data obtaining performance comparable to the state-of-the-art while requiring less annotated data.

We then exploit Image Captioning for Image Quality Assessment in both No-Reference and Full-Reference scenarios. Image captioners are capable of describing the content of high-quality images, but if the image is distorted the distortion will affect the image content leading to a different caption. Thus, we can use captioning metrics to compare high-quality image-generated captions and distorted image-generated captions in order to infer a quality score. We demonstrate that there is a high correlation between the predicted score and the quality score of the distorted image. With this approach, we can avoid training on small IQA datasets since we are using a captioning system (that is trained on large amounts of data). Moreover, we demonstrate that our system is capable of correctly evaluating high-quality images and restored images.

1.5 Contributions of this thesis

This thesis focuses mainly on vision and language tasks. As described above, these include other sub-tasks such as object detection, object recognition, visual reasoning, etc. State-of-the-art approaches achieve good results where contextual information is not needed, but in real scenarios, this becomes

extremely relevant. For example, the standard Visual Question Answering task involves answering questions about the visual content of an image but in realistic situations users tend to ask questions that are not answerable by looking at the image but that instead require external contextual knowledge. The same problem can be found in the generation of captions by an image captioner or in the ranking generated by retrieval systems that are weak in handling contextual information.

To address this problem we have largely focused on the Cultural Heritage domain. In this context, the standard models of Visual Question Answering fail as most of the questions about artworks require information that cannot be inferred directly from the artwork itself (e.g. name of the artist, style, year of production). Furthermore, there are no datasets or other works in this specific area. Therefore we have designed several architectures to solve this problem and created a related dataset for VQA in the cultural heritage domain.

Retrieval systems are also affected by the presence of contextual data. In fact, when the captions no longer concern a visual description (e.g. a man speaks sitting in front of a microphone) but contain contextual information (e.g. Joe Biden speaking from the oval office), the retrieval task is much more challenging and leads to low performance by standard approaches. To address this problem, we designed a Transformer-based architecture. This model exploits scene-text in the input images as additional information for the retrieval task. We demonstrate the effectiveness of this approach in real scenarios where there is not an exact matching between image-text pairs.

We investigate image captioning and its novel application to Image Quality Assessment (IQA). Since IQA approaches are affected by the lack of data for training, we investigate the potential of generating training samples and the use of Image Captioning as an instrument to infer quality scores in both No-Reference and Full-Reference scenarios. This approach avoids training models for the IQA task relying completely on Image Captioner accuracy. To demonstrate the effectiveness of our approach we considered several types of architectures.

In summary, this dissertation makes a number of contributions to the state-of-the-art in language and vision tasks:

Visual Question Answering for Cultural Heritage. In Chapter 2 We define a novel approach for Visual Question Answering in the Cultural Heritage Domain capable of answering both visual and contextual questions.

Differently from the standard VQA approaches that cannot answer contextual questions, our architecture classifies the category of the question (visual, contextual) in order to exploit the most suitable information. To the best of our knowledge, this is the first work on VQA applied to the Cultural Heritage domain. Moreover, the proposed architecture can be used in different real environments to cope with limitations of standard VQA.

Generation of Visual and Contextual Descriptions for VQA. In Chapter 3 we introduce another approach to VQA in the Cultural Heritage domain that is capable of generating visual and contextual descriptions of artworks in order to answer both visual and contextual questions. For this purpose, we investigate the use of GPT-3 for generating contextual information. The proposed approach allows answering questions about an artwork using only its name and avoiding a costly annotation process.

A large-scale Cultural Heritage Question Answering Dataset. Since there are no datasets for Visual Question Answering in the Cultural Heritage domain, in Chapter 4 we present the VISCOUNTH dataset. This dataset contains around 6.5M image-text pairs associated with 500k images. To the best of our knowledge, this is the largest VQA dataset in the Cultural Heritage domain and the one with the widest variety of artworks.

Generative Data Augmentation for Image Quality Assessment. In Chapter 5 we present a work on classical Image Quality Assessment. In this method, we address the problem of lack of data in IQA with a generative adversarial data augmentation approach. In particular, we train an Auxiliary-Classifer GAN (AC-GAN) to generate new samples with a specific distortion type and value. We demonstrate the effectiveness of our data augmentation reaching results comparable with the state-of-the-art.

Language-based Image Quality Assessment. In Chapter 6 we present a Full-Reference Image Quality Assessment approach based on image captioning. To the best of our knowledge, this is the first method that combines IQA and image captioning. Our system generates a caption for both reference images and distorted images and then compares the two captions with standard captioning metrics. We demonstrate that the metric score is highly correlated with the quality score of the distorted image. Our framework outperforms standard approaches and avoids training on IQA datasets (that usually leads to overfitting).

Language-based Blind Image Quality Evaluation. In Chapter 7 we

extended our previous work on Full-Reference IQA based on Image Captioning to the No-Reference scenario. Since in No-Reference IQA only the distorted image is available, we train a GAN to enhance the distorted image and use the enhanced version as reference and thus transforming the No-Reference problem into a Full-Reference one.

Scene Text Image and Language Transformer for Cross-Modal Retrieval. Since most of the cross-modal retrieval approaches work only on well-aligned image-text datasets, we investigate the use of image-text cross-modal retrieval systems in real scenarios where there is an abstract and symbolic alignment between image and text pairs. In Chapter 8 we present STILT, a transformer-based architecture that exploits scene-text in images to learn better image-text representations for the cross-modal retrieval task. We demonstrate the effectiveness of our approach multiple datasets.

Chapter 2

Visual Question Answering for Cultural Heritage

*In this chapter we propose a new method for Visual Question Answering in the Cultural Heritage domain. This approach combines the capabilities of Visual Question Answering models to answer visual questions and the capabilities of Question Answering models to answer contextual questions. This approach addresses the limitation of standard VQA approaches that are not capable to answer contextual questions and therefore not really usable in real scenarios.*¹

2.1 Introduction

Museum visits have adapted throughout the years to exploit technological advances. Nowadays cultural heritage heavily relies on some form of multimedia content to deliver information to the user in ways that limit cognitive burden and engage the visitor as much as possible. This is especially true for young visitors, where gamification techniques have often proven effective [7, 46]. Technology can help bridge the gap between user interests and the message the museum wants to convey.

¹The work described in this chapter has been published as "Visual Question Answering for Cultural Heritage" in IOP Conference Series: Materials Science and Engineering 2020. Vol 949

Videos, 3D reconstructions and augmented realities, among others, have become an integral part of the visit, which has now shifted its focus not solely on artworks but also on how they are organized and presented. To offer a richer experience, smart audio guides have also been developed, gradually replacing information sheets or offering some sort of augmented visit relying on sensors available on personal smartphones. Despite the increasing diffusion of devices to help guide the visitor, the most effective way to convey most information still remains a human guide with whom the visitor may interact to ask for clarifications or deeper discussions on topics of interest. In fact, the user requires a natural way to interact with whoever is providing the information, be it an actual museum guide or a piece of software.

At the same time, the diffusion of personal assistants on smartphones is aiding an increasing number of people with everyday tasks. These assistants, though, still offer little or no help in the area of cultural heritage. This is due to the need to process complex pieces of structured information, which are often transversal to several domains.

Machine Learning is starting to reach out to the complexity of these tasks. In particular, the emerging topic of *Visual Question Answering* is able to engage a user by answering questions about visual media [3, 40]. VQA algorithms merge the capabilities of Computer Vision to understand image content and those of Natural Language Processing to reason about questions and provide relevant answers. VQA builds upon the Question Answering literature, where questions are answered related to text instead of visual content. Interest in VQA has grown quickly, but it has still not been applied to cultural heritage since the knowledge required to answer the variety of questions a user might ask about artworks is not contained within the opera itself. A full understanding requires external knowledge usually obtainable only from experts (e.g. museum guides) or information sheets. This knowledge can be processed separately since it is often available in a textual form, whether it is provided directly from the museum or retrieved from online resources. Therefore, to be able to address the dual nature of the task, i.e. answering both visual and contextual questions, VQA and QA must be combined.

In this work, we make a first step towards the development of a Visual Question Answering model for cultural heritage by combining the capabilities of a VQA model and a QA model. Our first contribution is to introduce a module that accurately discriminates between visual and contextual ques-

tions. Our second contribution is to design a model made of two branches able to answer both kinds of questions. Our experiments demonstrate the effectiveness of our technique for question classification and the performance of our general question-answering model. To evaluate our model, we annotated a subset of ArtPedia [118] with visual and contextual question-answer pairs.

In Sec. 2.2 we describe our approach to integrating Visual and Contextual Question Answering and Contextual for the cultural heritage domain, and in Sec. 2.3 we report on a number of experiments we performed to quantify the performance of our approach. We conclude in Sec. 2.4 with a discussion of our contribution.

2.2 Method

In this section, we describe our approach to open-ended visual question answering. We first show the general model that characterizes our technique then we describe the sub-modules.

2.2.1 Visual Question Answering with visual and contextual questions

The main idea of this work is to classify the type (visual or contextual) of the input question so that the question can be answered by the most suitable sub-model. We rely on a question classifier to understand if the question concerns exclusively visual traits of an image or if an external source of information is needed to provide a correct answer. The question is then fed to a VQA or a QA model, depending on the output of the classifier. In both cases, the question must be analyzed and understood, yet the usage for two separate architectures is driven by the need to process different additional sources of information. If the question is visual, then the answer is generated from the image, whereas if the question is contextual then the answer is generated using external textual descriptions.

The overall pipeline (see Fig. 2.1) used by our approach to answer a question is the following:

1. **Question Classification.** The question is given as input to the question classifier module that determines if the question is contextual or visual.

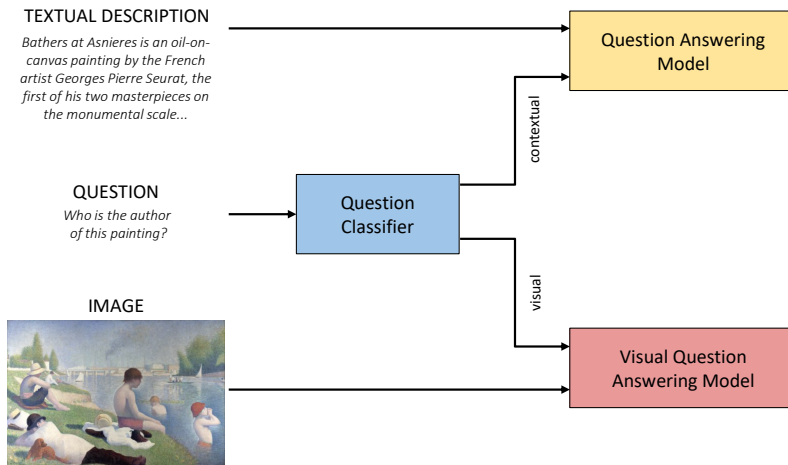


Figure 2.1: Model overview. A question classifier categorizes the question as visual or contextual. The correspondent module is used to answer the question relying either on the image or external descriptions.

2. **[Visual] Question Answering.** Depending on the predicted question type, the corresponding module is activated to generate the answer.
 - (a) If the question is contextual, it is given as input to a Question Answering module that also exploits useful external information to answer the question. This system produces an output answer only based on this external information.
 - (b) If the question is visual, the question and the image are given as input to a Visual Question Answering module. This system produces an output answer based on the content of the image.

2.2.2 Question Classifier Module

The question classifier module consists of a Bert [27] module for text classification. BERT makes use of a Transformer [128], an attention mechanism that learns contextual relations between words (or sub-words) in a text. The Transformer is trained in a bidirectional way in order to have a deeper knowledge of language context and flow. This language model is extremely versatile since it can be used for different tasks like text classification, next word in sentence prediction, question answering and entity recognition. This model

is turned into a question classification architecture by adding a classification layer on top of the Transformer output. The input question is represented as the sum of three different embeddings: the token embeddings, the segmentation embeddings and the position embeddings. Moreover, two special tokens are added at the start and in the end of the question.

2.2.3 Contextual Question Answering Module

The Model used for the Question Answering task is another Bert module that focuses on this task. In this case the module takes as input both a question and a textual description. Since this system uses the textual information to answer the question, the text must contain relevant information to generate an appropriate answer.

2.2.4 Visual Question Answering Module

The architecture of the Visual Question Answering module is similar to the one used by Anderson et al. [2] in their Bottom-up Top-Down approach. Here the salient regions of the image are extracted by a Faster R-CNN [100] pre-trained on the Visual Genome dataset [60]. The words of the question are represented with a Glove embedding [91] and then the question is encoded by a Gated Recurrent Unit (GRU) to condense each question into a fixed size descriptor. An attention mechanism between the encoded question and the salient image regions is developed to weigh the candidate regions that are useful to answer the question. Then the weighted region representations and the question representation are projected into a common space and are joined with an element-wise product. Finally the joint representation passes two fully connected layers and a softmax activation that produces the output answer.

2.3 Experimental Results

In this section we describe experiments conducted to evaluate the performance of our approach. We first introduce the datasets used for training and testing our network, then we describe the protocols adopted for the experiments and the obtained results.



QUESTION CLASSIFIER	QA MODEL	VQA MODEL
Who is the author of this painting? Contextual		
What is there in the background? Visual	Who is the author of this painting? george william joy	How many clocks are there in the figure? Two
When was this painting depicted? Contextual	When was this painting depicted? 1895	What is the color of the sand? Brown
Where is the painting now? Visual	What did Joy do while working on this painting? borrowed bus from a company	Are we at the beach? Yes
How many people are there in the image? Visual	Who mainly used this form of transport? middle classes	What is there in the background on the left? Surfboard
When was this painting depicted? Contextual		
What is she wearing on her head? Contextual		
Who is portrayed in this painting? Contextual		
What is hanging on the wall? Visual		

Figure 2.2: Sample outputs of the three components of our architecture. Correct answers are shown in black, wrong answers in red.

2.3.1 Datasets

For our experiments we used the standard VQA v2 [3] dataset, OK-VQA [81] and Artpedia [118], a dataset containing images of famous paintings.

VQA v2 This dataset contains 443,757 training questions/answers referred to 82,783 training images. The number of test examples is about the same of the training examples, instead the validation examples are about the half. Each image has more questions referred to it and these are of multiple types like relation between objects, activity recognition, counting, object detection and so on. Each question is answered by ten annotators and the given answers compose the ground truth. VQA v2 is currently the most used benchmark for Visual Question Answering tasks.

OK-VQA OK-VQA is a subset of the VQA v2 dataset and it contains 14,055 open-ended questions where each of these has five ground truth answers. In particular OK-VQA contains all the questions of VQA v2 that cannot be answered with processing only the corresponding image but require external knowledge. We use OK-VQA jointly with the original VQA dataset to obtain sets of questions related to the image (visual questions) or to external knowledge (contextual questions).

Artpedia The Artpedia dataset contains a collection of 2,930 paintings, each associated to a variable number of textual descriptions collected from Wikipedia. Each sentence is labelled either as a visual sentence or as a contextual sentence, if it does not describe the visual content of the artwork. Contextual sentences can describe the historical context of the artwork, its author, the artistic influence or the place where the painting is exhibited. The dataset contains a total of 28,212 sentences, 9,173 labelled as visual sentences and the remaining 19,039 as contextual sentences. This is not a Visual Question Answering dataset, so we manually annotated a subset of images with both visual and contextual question-answer pairs, based on the available images and descriptions.

2.3.2 Experimental Protocols

Our model is composed by three sub-modules: the question classifier that classifies if a question requires visual or contextual information, the question answering module which answers to contextual questions and the visual question answering module which answers to visual questions. The three modules generate different outputs and we evaluate each one of them independently. The Visual Question Answering module answers with short sentences of at most three words chosen from the set of answers. For this reason, as common practice in the VQA literature, we can consider the problem as a classification task and estimate the accuracy to assess its performance:

$$Accuracy = \frac{N_c}{N_a} \quad (2.1)$$

where N_c is the number of correct answers and N_a the number of total answers. The same metric can be used for the question classifier module, since it solves a binary classification task.

The question answering model instead, since it can potentially rely on structured and more complex information from the meta-data, is able to answer to questions with more words, articulating short sentences. For this reason we evaluate its performance not only with Accuracy but also with F1-measure, a metric that takes into account the global correctness of the answer:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.2)$$

where *Precision* is defined as the number of correct words divided by the

	OK-VQA/VQA v2	Artpedia
Question Classifier	0.868	0.938

Table 2.1: **Question classifier:** accuracy of our question classifier on questions from both the OK-VQA and VQA v2 datasets and from Artpedia.

QA model				VQA model		
Contextual	Visual	Accuracy	F1-score	Contextual	Visual	Accuracy
✓	✗	0.684	0.832	✓	✗	0.000
✗	✓	0.176	0.150	✗	✓	0.524
✓	✓	0.504	0.417	✓	✓	0.251

Table 2.2: Results of the two answering models on contextual questions, visual questions and both visual and contextual questions from Artpedia. Note that the VQA model does not have access to the external information required to answer the contextual questions, making it unable to answer correctly. See Sec. 2.3.3 for analysis of the performance of our full model on combined Visual/Contextual Question Answering

length of the answer and *Recall* as the number of correct words divided by the length of the ground truth.

2.3.3 Experimental results

In order to evaluate the performance of our model we make different experiments. We measure the performance of the model analyzing each component independently.

Question Classifier

We train the question classifier module with questions of both the OK-VQA and VQA v2 datasets. We take from VQA v2 a number of visual questions equal to the number of questions that require external knowledge from OK-VQA. The obtained dataset is then split into train and test sets. The question classifier is supposed to understand from the structure of the question whether the answer concerns the visual content or not. This is a generic classifier, agnostic from the domain of the task. In fact, VQA v2 and OK-VQA contain generic images, while we are interested in applications in the cultural heritage domain. We demonstrate the effectiveness of our approach

and its ability to transfer to the cultural heritage domain by evaluating it both on the VQA/OK-VQA dataset and on a new dataset comprised of a subset of Artpedia [118]. Since this dataset does not contain questions but only images and descriptions, we took 30 images from this dataset and annotated them with a variable number of both visual and contextual questions (from 3 to 5 for both categories). The accuracy of our question classifier module is shown in Tab. 2.1. We can observe that it is able to predict the type of the question correctly in most cases.

Contextual Question Answering

We test our question answering module on the subset of Artpedia containing 30 images that we annotated. In particular, we test the accuracy of our module in three different experiments: test on contextual questions, test on visual questions and test with both visual and contextual questions. Note that the outputs of the visual and contextual modules are different, since VQA is treated as a classification problem, while for QA From the results shown in Tab. 2.2 we can deduce that our question answering module works very well with contextual questions and obtains worse results with visual questions. This can be justified from the fact that visual questions refer to visible details of paintings that cannot be described in visual sentences of ArtPedia.

Visual Question Answering

Similarly to the tests conducted for the question answering module, we evaluate the visual question answering module on both visual and contextual questions. In Tab.2.2 results of our visual question answering model are shown. We can observe that conversely from the question answering module this model performs well on visual questions and is not able to answer correctly to contextual questions. This is motivated by the fact that contextual questions require external knowledge (e.g. author, year) that a purely visual question answering engine does not have access to.

Full pipeline

Finally, we combine the capabilities of all the modules together and we test on both visual and contextual questions, obtaining an accuracy of 0.570.

The full pipeline, thanks to the question classifier, is able to correctly distinguish between visual and contextual questions. The visual question answering module and the question answering module receive as input almost all questions that they are able to answer (contextual question for the question answering module and visual questions for visual question answering module). For this reason the complete model exceeds the performances of both single answering modules. Fig. 2.2 shows some qualitative result of the three components of the pipeline. The components correctly handle most of the questions but some common failure cases can be observed. For instance the Question Answering model might add details to the answer that are not present in the ground truth and the Visual Question Answering model might confuse some elements of the painting with similar objects.

2.4 Conclusions

In this first work we presented an approach for Visual Question Answering in the Cultural Heritage domain. We have addressed two important issues: the need to process both image and contextual knowledge contained and the lack of data availability. The model we presented combines the capabilities of a VQA and a QA model, relying on a question classifier to predict whether it refers to visual or contextual content. To assess the effectiveness of our model we annotated a subset of the ArtPedia dataset with visual and contextual question-answer pairs. Referring to this annotation process, in Appendix A we describe the a data collection and annotation system for Visual Question Answering.

Chapter 3

Is GPT-3 all you need for Visual Question Answering in Cultural Heritage?

*In this chapter we propose a method for Visual Question Answering that enables generation at runtime of a description sheet that can be used for answering both visual and contextual questions about the artwork, completely avoiding the image annotation process. For this purpose, we investigate on the use of GPT-3 for generating descriptions of artworks and analyze the quality of generated descriptions through captioning metrics. Finally, we evaluate the performance for Visual Question Answering and captioning tasks.*¹

3.1 Introduction

In Chapter 2, we observed that in the Cultural Heritage domain most questions posed by users concern contextual information rather than what is actually depicted in a painting. And the proposed solution is an evolution of VQA known as Contextual Question Answering (CQA). The contextual information is derived from textual meta-data, which is fed to the model along

¹The work described in this chapter has been published at the European Conference on Computer Vision (ECCV) Workshop on VISion and Art (VISArt).

with the question and the image. In this way, the VQA/CQA model has to learn to attend either relevant parts of an image or relevant sections of the text to provide an adequate answer. The need of a textual data nonetheless opens a new issue, namely where to obtain such description. Information sheets for artworks may already be available to museum curators yet extending this kind of application to new data becomes time-consuming and requires a domain expert.

In this work we explore the usage of a generative natural language processing model to automatically create contextual information to be fed to a CQA model. In fact, recently, generative text models have been finding large diffusion with groundbreaking results. Among these we find GPT-3, a generative model trained on a massive corpus of textual data regarding several domains, including art [14]. GPT-3 is capable of generating a description starting from a textual query and it has been demonstrated that the model includes knowledge of the entities described in the training data, for example paintings and artworks. We therefore investigate the possibilities and the limitations of GPT-3 in applications for cultural heritage, with a specific focus on question answering. In particular, we explore the quality of the textual description of artworks that the model is able to generate and we evaluate their applicability for visual and contextual question answering.

The main contributions of our work are the following:

- We propose an automatic approach to generate textual information sheets of artworks exploiting GPT-3. We find that the model has excellent knowledge of art concepts and event details of specific paintings.
- We propose a method to answer both visual and contextual questions which is artwork agnostic, i.e. it does not require any additional data or training to be adapted to a new set of images.
- We explore the applicability of GPT-3 in cultural heritage applications. To the best of our knowledge we are the first to apply GPT-3 to the art domain.

3.2 GPT-3

To provide to the reader a better understanding of our work, here we present a brief background context about GPT-3, the third version of Generative Pre-Trained Transformer [14]. This is an autoregressive language model

with 175 billion parameters that can be used for different tasks without any finetuning, achieving strong performances.

The architecture of the GPT-3 Transformer model is made of 96 attention layers. While language models like BERT [27] use the Encoder to generate embeddings from the raw text which can be used in other machine learning applications, GPT-3 use the Decoder half, so it takes embeddings as inputs and produces text. In particular the GPT-3 language model has the ability to generate natural language text that can be hard to distinguish from human-written text, to the point that research has been carried out to asses whether GPT-3 could pass a written Turing test [31].

Concretely, during inference, the target of the new task y is directly predicted conditioned on the given context C and the new task’s input x , as a text sequence generation task. Note that all C , x and y are text sequences. For example, $y = (y^1, \dots, y^T)$. Therefore, at each decoding step t we have

$$y^t = \arg \max_{y^t} p_W(y^t | C, x, y < t) \quad (3.1)$$

where W are the weights of the pretrained language model, which are frozen for all new tasks. The context $C = h, x_1, y_1, \dots, x_n, y_n$ consists of an optional prompt head h and n in-context examples $(\{x_i, y_i\}_{i=1}^n)$ from the new task.

3.3 Method

In a Cultural Heritage context, the information useful to answer questions about a specific artwork is contained in the artwork image and in its contextual description. Finding such a description might not be trivial, since it might require a domain expert to write it down. At the same time, it is quite costly to train a Visual Question Answering model that takes as input both the image and the description. This is also not straightforward, since the two modalities need to be blended and matched together. Consequently, the main idea of this work is to generate new descriptions for artworks based on a specific prompt or a specific question and directly use these descriptions to answer visual and contextual questions. The overall pipeline of our proposed work is as follows:

1. **GPT-3 caption generation.** We use GPT-3 to generate descriptions of artworks, leveraging its memorization capabilities that allowed the model retain relevant information about training instances. An important aspect in this phase is to feed the correct prompt as input to

GPT-3 in order to obtain realistic and correct descriptions. We consider two different types of input prompt:

- **General** - A general prompt where the expected output is a general description of the artwork. The input text follows the structure:

```
"Describe and Contextualize the painting < painting_name >"
```

- **Question-based** - A specific question based prompt. The input text follows the structure:

```
"Painting < painting_name > < question >".
```

The expected generated text by GPT-3 is a small text snippet that consists in a couple of sentences, focused on the topic of the question.

2. **Question answering.** Once the description has been generated in the previous step, we can exploit it to answer both visual and contextual questions through a Question Answering language model. For this purpose we use a pretrained version of DistilBert [108] fine-tuned on the SQUAD [97] dataset. We feed as input to the DistilBert model the generated text from the previous step together with the question. The answer given as output will be the final answer of our method.

Fig. 3.1 and Fig. 3.2 show a scheme of the two variants of our method. More precisely, in Fig. 3.1 the general input prompt for GPT-3 yields the generation of a long description of the artwork (similar to a museum information sheet). On the other hand, the question-based prompt in Fig. 3.2 yields only the generation of a brief output text, which we find suitable for answering the question. In conclusion, these two schemes follow roughly the same structure. The difference is in the input prompt that in the case of Fig. 3.1 is more general and in Fig. 3.2 is more task oriented.

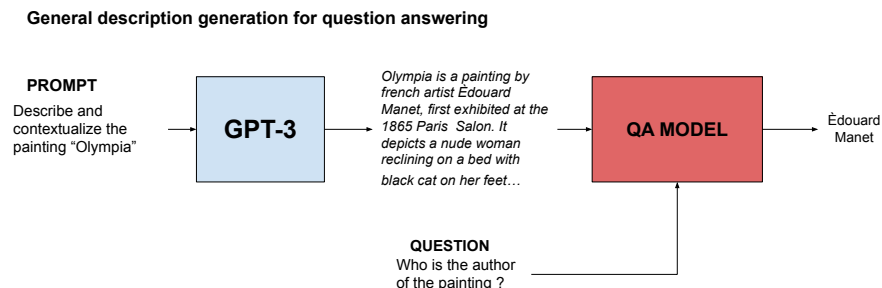


Figure 3.1: Scheme of our method for answering questions using a general generated description. A prompt with a specific structure is given as input to GPT-3. Subsequently the generated text is fed together with the question to a Question Answering model that outputs the answer.

3.4 Experiments

In this section we first outline the experimental setting for the experiments carried out in this work, presenting dataset and experimental protocol and we then move on to a discussion of the results.

3.4.1 Dataset

For our experiments, we use the Artpedia dataset [118]. Artpedia contains a collection of 2,930 artworks, associated to a variable number of textual descriptions gathered from Wikipedia. Sentences are labelled as a visual descriptions or as a contextual descriptions. Contextual descriptions regard information about the artwork that does not directly describe its visual content. For instance, contextual descriptions can describe the historical context of the artwork, its author, the artistic influence or the museum where a painting is exhibited. The dataset contains 28,212 descriptions, 9,173 of which are labelled as visual and the remaining 19,039 as contextual. The Artpedia dataset has been extended with Question-Answer annotations as in Sec. 2. In fact, a subset of the images have been associated with visual and contextual questions, derived from the corresponding captions. In this work we follow the dataset split used in Sec. 2.

Question-based description generation for question answering

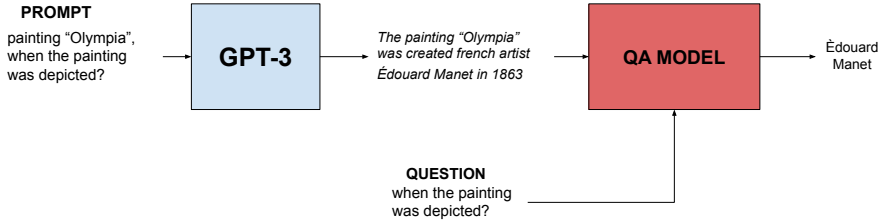


Figure 3.2: Scheme of our method for answering questions using a question-based generated description. A prompt containing the name of the painting and the question is given as input to GPT-3. Subsequently the generated text is fed together with the question to a Question Answering model that outputs the answer.

3.4.2 Experimental Protocol

Following prior work described in Sec. 2, we evaluate visual questions and contextual questions with different metrics. In fact, visual question answering and traditional text-based question answering are often treated in two different ways. Visual Question Answering is considered as a classification problem, meaning that a model has to pick an answer from a predefined dictionary of possible candidates containing a few words each. This stems from the fact that questions in most datasets are a way of guiding attention towards specific objects or attributes in the image, without requiring any complex form of language reasoning. Question Answering on the other hand is based on a set of sentences, which may contain rare or out-of-dictionary words. The task is in fact defined as identifying a subset of the textual description that contains the answer.

In light of this, to evaluate visual questions we rely on accuracy:

$$Accuracy = \frac{N_c}{N_a}, \quad (3.2)$$

where N_c is the number of correct answers and N_a the number of total answers.

For text-based question answering, instead, we use both accuracy and F1-measure, a metric that takes into account the global correctness of the

answer:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (3.3)$$

where *Precision* is defined as:

$$Precision = \frac{N_{Cw}}{|ans|}, \quad (3.4)$$

and N_{Cw} is the number of common words between the output answer and the ground truth answer and *ans* the number of words in the generated answer.

Recall is instead defined as:

$$Recall = \frac{N_{Cw}}{|gt|}, \quad (3.5)$$

where $|gt|$ is the number of words in the ground truth.

We also evaluate the quality of the descriptions generated by GPT-3, considering it as a standalone image captioning model. We use the following standard metrics for captioning:

- *BLEU1* [89]: BiLingual Evaluation Understudy (BLEU) is the most commonly used metric for machine translation and image captioning. BLEU scores are based on how similar a generated caption is to a reference caption, computing the precision of the generated words. The downside of BLEU is that it is very sensitive to small changes, such as synonyms or different word order.
- *ROUGE* [70]: differently from BLEU, which measures the precision of the caption, Recall Oriented Understudy of Gisting Evaluation (ROUGE) focuses on quantifying the amount of correct words with respect to the reference. Thus, this metric is recall-based and tends to reward long sentences.
- *CIDEr* [129]: Consensus-based Image Description Evaluation (CIDEr) is an automatic consensus metric that measures the similarity of captions against a set of ground truth sentences written by humans. This metric has been shown to yield a higher agreement with humans generated text since it captures notions of grammar, importance and precision and recall.
- *Cosine Similarity*: we compute the cosine similarity between feature vectors for the generated caption and the reference caption. Features are extracted with the algorithm TF-IDF [107].

Table 3.1: Image captioning results. We compare our method which generates captions with GPT-3 with the *General* and the *Question-based* approaches. In the *Question-based* approach we concatenate all the outputs of GPT-3 after conditioning it with different questions related to the image. We compare the results against visual captions, contextual captions or both.

Description type	Metric	OFA [132]	Ours General	Ours Question-based
Visual	BLEU1	0.048	0.181	0.137
	ROUGE	0.138	0.188	0.16
	CIDEr	0.091	0.079	0.172
	COSINE	0.113	0.157	0.110
Contextual	BLEU1	0.002	0.168	0.160
	ROUGE	0.062	0.178	0.179
	CIDEr	0.000	0.248	0.129
	COSINE	0.082	0.218	0.324
All	BLEU1	0.000	0.113	0.185
	ROUGE	0.053	0.158	0.184
	CIDEr	0.000	0.016	0.098
	COSINE	0.122	0.253	0.341

3.4.3 Experimental Results

Captioning Results

We start by assessing the quality of the captions generated by GPT-3. First of all, we ask GPT-3 to generate captions with our *General* approach. In Tab. 3.1 we compare the captions using as reference visual captions, contextual captions and both. All reference captions are ground truth captions taken from the Artpedia dataset [118].

Interestingly, the model appears to better results for visual captions using BLEU1 and ROUGE metrics, while using CIDEr and cosine similarity, the model obtains higher results for contextual captions. This may seem counter-intuitive but can be explained looking at the nature of the metrics. BLEU1 and ROUGE in fact respectively check for word-wise precision and recall, while CIDEr and cosine distance perform a sentence level scoring, which is closer to human consensus. We observe that the model is able to obtain good results, especially with the cosine metric, even when using all the captions as reference.

We then evaluate the method by taking a concatenation of the outputs generated by GPT-3 after being conditioned by different questions related to the image. This obviously introduces a strong bias, given also the fact that

Table 3.2: Experimental results for Visual Question Answering. We compare our approach against VQA-CH [11] described in Chapter 2 to understand whether GPT-3 can replace information sheets for artworks either for visual or contextual questions. We compare two versions of our model, the *General* version, which produces generic descriptions of artworks and the *Question-based* version, where prompts are conditioned with the input question to generate more specific descriptions.

	Visual	Contextual	Accuracy	F1 score
VQA-CH [11]	✗	✓	0.684	0.832
VQA-CH [11]	✓	✗	0.176	0.150
VQA-CH [11]	✓	✓	0.504	0.417
Ours - General	✗	✓	0.557	0.719
Ours - General	✓	✗	0.070	0.055
Ours - General	✓	✓	0.239	0.360
Ours - Question-based	✗	✓	0.473	0.602
Ours - Question-based	✓	✗	0.134	0.202
Ours - Question-based	✓	✓	0.256	0.330

questions have been generated from information contained in the captions, but at the same time proves the usefulness of such captions for more advanced applications such as visual question answering. As can be seen in Tab. 3.1, conditioning GPT-3 with the captions leads to better captions according to most metrics.

In Tab. 3.1 we also provide a baseline as reference, i.e. the output of the state of the art OFA captioning model [132]. We observe that captions generated by OFA do not align well with the ground truth sentences. We attribute this to a domain shift between the datasets commonly used to train captioning models and descriptions of artworks. In fact, the former are sentences written by non-experts while for applications in cultural heritage a domain knowledge is required. This further motivates the usage of GPT-3, which seems to have integrated sufficient knowledge to articulate complex sentences with a domain specific jargon.

VQA Results

To evaluate the Visual Question Answering capabilities of our proposed method, we follow the setting used in Chapter 2. However, we do not rely

on any vision-based model but rather on a fully textual question answering model based on DistilBert [108], as explained in Sec. 3.3. In Tab. 3.2, we compare our approach to the one of VQA-CH described in Chapter 2. It has to be noted that, contrary to the work in Chapter 2, we do not rely on real textual descriptions, which are known to contain the answer, but we only extract information from GPT-3. This is a strong disadvantage for our method. However, we are not interested in obtaining better results than VQA-CH, but rather our goal is to demonstrate if GPT-3 can act as a substitute of textual descriptions handcrafted by domain experts.

We test our method evaluating the accuracy for visual questions, contextual questions and both together. Quantitative results indicate that captions generated by GPT-3 can yield to high results for contextual questions, yet very low accuracy for visual questions. As for the captioning setting, we impute this behavior to the fact that GPT-3 generates generic descriptions, without including a fine-grained description of the visual content. Thus, on the one hand the question answering model is capable of extracting meaningful information from the generated captions. This means that GPT-3 is indeed capable of integrating domain knowledge during training and is capable of generating a complete information sheet of the artwork. On the other hand, captions appear to be too generic to obtain information about specific details in the image.

To overcome this limitation, we test the model using captions generated with out *Question-based* approach. By feeding the answer to GPT-3 along with the title of the artwork, the model is able to generate more specific captions. Such captions, as explained in Sec. 3.5 are usually shorter but are focused on the prompt. This is particularly interesting since it means that a purely text-based model is capable of addressing a vision-based task. In Tab. 3.2 it can be seen that for visual questions alone, our method with question-based captions performs on par or better than the vision-based VQA-CH model.

3.5 Qualitative Analysis

In this section we provide a qualitative analysis of the captions generated by GPT-3 in order to characterize which kind of information they contain in both the *General* and *Question-based* formulation.

Since the prompts that we feed to GPT-3 are different, with one being

ARTEPIDIA DESCRIPTION	QUESTION	GENERAL GENERATED DESCRIPTION	GENERAL ANSWER	QUESTION-BASED GENERATED DESCRIPTION	QUESTION-BASED ANSWER
<p><i>Bathers at Asnières (French: Une Baignade, Asnières)</i> is an oil-on-canvas painting by the French artist Georges Pierre Seurat, the first of his two masterpieces on the monumental scale. The canvas is of a suburban, placid Parisian riverside scene. A combination of complex brushstroke techniques, and a meticulous application of contemporary colour theory bring to the composition a sense of gentile...</p>	Who is the author of this painting?	<p><i>Bathers at Asnières</i> is a painting by French artist Georges Seurat. The painting depicts a group of men relaxing by the river Seine on a sunny day. The background includes industrial buildings and smokestacks, which contrast with the leisurely scene in the foreground. Seurat used small dots of color to create a sense of light and movement in the painting, which has been called an early example of Pointillism. The painting was completed in 1884, just before Seurat's death at the age of 31. It was exhibited posthumously at the Salon des Indépendants in 1886, where it received mixed reviews from critics. Some praised its technical mastery, while others found it cold.</p>	Georges Seurat ✓	The painting <i>Bathers at Asnières</i> was created by french artist Georges Seurat in 1884	Georges Seurat ✓
	When was the painting depicted?		1884 ✓	<i>Bathers at Asnières</i> was painted by George Seurat in 1884	1884 ✓
	What is the animal in the ground?		✗	The animal on the ground is a dog	dog ✓
	Is there a man lying down?		✗	There is a man lying down in the painting <i>Bathers at Asnières</i> , but it is difficult to see. He is on the left side of the painting, near the bottom.	There is a man lying down ✓
<p><i>Dogs Playing Poker</i>, by Cassius Marcellus Coolidge, refers collectively to an 1894 painting, a 1903 series of sixteen oil paintings commissioned by Brown & Bigelow to advertise cigars, and a 1910 painting. "Critic Amelie Ferrara has described <i>Dogs Playing Poker</i> as "indelibly burned into ... All eighteen paintings in the overall series feature anthropomorphized dogs, but the eleven in which dogs are...</p>	Who is the author of this painting?	<p><i>Dogs Playing Poker</i> is a painting by American artist Cassius Marcellus Coolidge. The painting depicts a group of dogs sitting around a table playing poker. The dogs are all wearing human clothes and have cigars in their mouths. The painting is set in a dimly lit room, with a fireplace in the background. The painting is often seen as an example of kitsch art, due to its subject matter and style. It is also seen as a classic example of Americana, due to its depiction of a group of dogs engaged in a typically American activity.</p>	Cassius Marcellus Coolidge ✓	The painting is anonymous.	✗
	Who is sitting on the chair?		dogs ✓	The artist who painted <i>Dogs Playing Poker</i> is named C. M. Coolidge.	✗
	What are the dogs doing?		playing poker ✓	The dogs in the painting are playing poker.	dog ✓
	What is the color of the table?		✗	table in <i>Dogs Playing Poker</i> is green	green ✓
<p><i>The Singing Butler</i> is an oil-on-canvas painting made by Scottish artist Jack Vettriano in 1992. As a contemporary cultural icon, <i>The Singing Butler</i> has been compared to Grant Wood's <i>American Gothic</i>. It depicts a couple dancing on the damp sand of a beach on the coast of Fife, with grey skies above a low horizon. To the left and right, a maid and a man hold up umbrellas against the weather...</p>	When was the painting depicted?	<p>The painting "The Singing Butler" was painted by Scottish artist Jack Vettriano in 1992. The painting depicts two people, a man and a woman, standing on a beach with a Butler who is singing and playing the guitar. The background of the painting is a blue sky with white clouds. The painting is set in the early 20th century.</p>	1992 ✓	This painting was depicted in 1992.	1992 ✓
	What are the two people in the middle doing?		dancing ✓	The two people in the middle are dancing	dancing ✓
	How many umbrellas are there?		✗	There are two umbrellas in the painting.	two ✓

Figure 3.3: Qualitative results of our method. *Green*: ground truth description from the Artpedia dataset [118] and input question. *Yellow*: general descriptions provided by GPT-3 and answer obtained based on such text. *Blue*: Question-based description and correspondent answer. General descriptions are longer and more detailed than question-based generated descriptions. However, question-based generated descriptions are customized for the specific question.

more general and the other being question-based, we expect that the corresponding generated text by GPT-3 will be different. In Fig. 3.3 we can observe these differences. Generated general descriptions are very long and have the aspect of artwork information sheets in which we can find some visual and contextual information. Question-based generated descriptions are instead shorter and contain the knowledge needed to answer to the specific questions. From Fig 3.3 we can observe that the general description is very useful to answer to contextual questions but fails on some visual questions. This is likely due to different reasons:

- The generated text does not take into account any specific question and this can lead to the generation of a description without specific information useful to answer to the question.
- Visual questions are very specific since they refer to object relationships, colors, counting, etc. and the GPT-3 model tends to be more

shallow in generating its descriptions.

On the other hand, question-based generated descriptions are helpful to answer visual questions but the small generated description useful to answer those specific questions could contain incorrect information leading to wrong answer predictions. In conclusion these two ways of generating text to answer visual and contextual questions have some pros and cons:

- General descriptions are longer and contain several pieces of information about the artwork. However this is fixed and could not contain the information needed to answer some questions.
- Question-based descriptions are generated for specific questions and contain only the information needed to answer the question on which GPT-3 has been conditioned. If the model has not memorized any specific information regarding such questions it may contain mistakes and descriptions will have to be re-computed for each question.

3.6 Considerations on complexity and accessibility of GPT-3

In the previous sections we have demonstrated that GPT-3 could indeed replace the usage of an information sheet handcrafted by a domain expert. However, we need to understand the actual applicability of GPT-3 in a real case application. GPT-3 has 175B parameters, which approximately amounts to 700GB. This means that inference on a single GPU is unfeasible due to current technological limits. The model however has been made available from OpenAI and is accessible through API that have a pricing fee per generated token. These considerations somewhat limit a large-scale usage of the model, especially if a description has to be generated for each question to be answered. On the other hand, generating fixed descriptions offline, one for each artwork, appears a viable solution at least for addressing contextual questions.

3.7 Conclusions

In this work we presented a method for Visual Question Answering in the Cultural Heritage domain. In particular we have addressed the problem of

data annotation for artworks, generating descriptions with GPT-3. The performances for the VQA task show that the generated descriptions are useful to answer the questions correctly. This technique allows to answer visual and contextual questions focusing only on the generated description and can be used for any artwork. In fact, there is no need to retrain the model to incorporate new knowledge. This is possible thanks to the memorization capabilities of GPT-3, which at training time has observed millions of tokens regarding domain-specific knowledge. Finally the generated description can be integrated as textual input (textual description) in a more complex architecture as the one described in Chapter 2 in order to address tasks like Visual Question Answering. This is of particular interest for Cultural Heritage due to the domain shift between common VQA and captioning datasets compared to the technical jargon that is needed to properly address questions about art.

Chapter 4

VISCOUNTH: A Large-Scale Visual and Contextual Question Answering Dataset for Cultural Heritage

In this work we present a large-scale heterogeneous and multi-language dataset for cultural heritage that comprises approximately 500K cultural assets and 6.5M question-answer pairs. We propose a novel formulation of the task that requires reasoning over both the visual content and an associated natural language description, and present baselines for this task. Results show that the current state of the art is reasonably effective, but still far from satisfactory, therefore further research in this area is recommended. We also present a holistic baseline to address visual and contextual questions and foster future research on the topic.

1

¹The work presented in this chapter has been submitted to ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).

4.1 Introduction

In Chapters 2 and 3 we presented two approaches for Visual Question Answering in the Cultural Heritage Domain. Moreover, we observed that in the literature there are no large datasets for VQA in Cultural Heritage and for some of our tests in the previous chapters we annotated a small dataset with the framework described in Appendix A.

To address the problem of lack of data for VQA in the cultural heritage, in this work we generate a large-scale dataset for cultural heritage in Italian and English by means of a semi-automatic approach that exploits data from an existing ontology-based knowledge graph. We first obtain a set of question templates asking expert and non-expert users to provide relevant questions for observed artworks. The question templates are then used to automatically extract answers from the knowledge graph, thus associating question-answer pairs with entities belonging to the cultural domain. We produce both short synthetic answers, useful for validating correctness of the prediction, and long colloquial answers, useful for user interaction through dialogue. A preliminary version of the dataset has been presented in [4]. We significantly extend the dataset by considering a broader variety of question verbal forms (from 282 to 427), in particular by considering verbal forms that are specific for certain cultural assets (e.g. “who is the author of this painting”, specific for paintings) and including additional details (e.g. the span of the answer for contextual question). Furthermore, we present baselines for our proposed VQA task and discuss current state-of-the-art performances, criticality and research directions. Overall the main contributions of our work are the following:

- We present the first complete large-scale multi-language visual question answering dataset for cultural heritage comprising approximately 500K images and 6.5M question-answer pairs in Italian and English. We detail our data collection process based on ArCO, the Italian cultural heritage knowledge graph.
- We rise the issue of domain shift in Visual Question Answering datasets for cultural heritage, which does not allow the exploitation of off-the-shelf VQA models without a re-training phase. We also take into account visual and contextual question answering, exploring the limitations of existing image-based and text-based question answering models for artworks.

- We propose baselines for the proposed dataset, analyzing the results according to different criteria such as question type and artwork type. We believe that this will foster the advancement and development of interactive smart assistants in museum visits enabling visual and contextual question answering capabilities.

4.2 Building VISCOUNTH: A large Visual and Contextual Question Answering Dataset for Cultural Heritage

The need for large datasets in the Cultural Heritage domain has motivated us to exploit the large and detailed amount of structured data in the ArCo Knowledge Graph [16] to produce a comprehensive VQA dataset, useful for training and evaluating VQA systems.

ArCo consists of (i) a network of seven ontologies (in RDF/OWL) modeling the cultural heritage domain (with focus on cultural assets) at a fine-grained level of detail, and (ii) a Linked Open Data dataset counting $\sim 200\text{M}$ triples, which describe $\sim 0.8\text{M}$ cultural assets and their catalog records derived from the *General Catalog of Italian Cultural Heritage* (ICCD), i.e. the institutional database of the Italian cultural heritage, published by the *Italian Ministry of Culture* (MiC). The ArCo ontology network is openly released with a CC-BY-SA 4.0 license both on *GitHub*² and on the official *MiC website*³, where data can be browsed and acceded through the SPARQL query language⁴.

Extracting information from ArCo to generate a dataset for VQA is not free of obstacles. First, ArCo does not give us a measure of which kind of questions might be interesting for average users in a real scenario. Second, ArCo data need to be suitably transformed and cleaned to produce answers in a usable form and questions need to be associated to corresponding answers. Third, the dataset we aim at generating is huge, and therefore manual validation of produced data cannot be performed.

²<https://github.com/ICCD-MiBACT/ArCo/tree/master/ArCo-release>

³<http://dati.beniculturali.it/>

⁴<https://www.w3.org/TR/rdf-sparql-query/>

4.2.1 A Semi-Automatic Approach for Generating the VQA Dataset

To create our VQA dataset, we resorted to a semi-automatic approach that involves the collaboration of expert and non-expert users and the use of text processing and natural language processing techniques to obtain an accurate list of question-answer pairs. We considered a scenario where an image is associated to available knowledge either manually (e.g., artworks in a museum can be associated with their descriptions) or by object recognition (e.g., architectural properties identified by taking pictures), and generated a dataset as a list of question-answer pairs, each one associated to an image, a description and a set of available information items. An instance of question-answer pair is: “Who is the author?” - “The author of the cultural asset is Pierre François Basan”.

Our semi-automatic approach consisted in two main steps. The first part of the process focused on generating a list of question types with associated verbal forms by considering both expert and non-expert perspectives, the latter assessed by surveys. Then, for each question type, we automatically generated a list of question-answer pairs by combining question forms and associated answer templates with information from relevant cultural assets in ArCo, and accurately cleaning the results. This process was performed by an ad-hoc tool, developed following a build-and-evaluate iterative process. At each step we evaluated a sample of the produced dataset to propose new data cleaning rules for improving results. The process ended when the desired accuracy was achieved. Eventually, question-answer pairs from different question types were combined. Next, we first detail our question types generation process, then fully describe the question-answer pairs generation by drawing from question types.

The *question types generation* process was based on the following two perspectives carried out independently: a *domain experts’ perspective*, represented by a selection of natural language competency questions (CQs) [95] previously considered to model the ArCo ontology network [16], and a *user-centered perspective*, represented by a set of questions from mostly non-expert (65 out of 104) users, collected through five questionnaires on a set of different images of cultural assets belonging to ArCo (five cultural assets per questionnaire). In the questionnaires, the users were asked to formulate a number of questions (minimum 5, maximum 10) that they considered related to each image presented (questions they would ask if they were enjoying the

cultural asset in a museum or a cultural site). In this way, we collected 2,920 questions from a very heterogeneous group of users in terms of age (from 24 to 70 years old and 42 years average age), cultural background and interests. Then, the questions were semi-automatically analyzed and annotated in order to recognize their semantics, associate them (when possible) with ArCo’s metadata, and create corresponding SPARQL queries for data extraction.

In the clustering process, we grouped user-produced questions into semantic clusters, named *question types*, with the purpose of grouping together questions that ask for the same information. Clustering was first performed automatically by text analysis and sentence similarity, then validated and corrected manually. The automatic procedure consisted in the following steps. We initially aggregated sentences that resulted to be identical after tokenization, lemmatization and stop words removal. Then, for each question, we identified the most semantically similar one in the whole set by Sentence-BERT [99] and aggregated sentences whose similarity was above 84% (we found empirically that this value resulted in a low error rate). Eventually, we performed average linkage agglomerative clustering with a similarity threshold of 60%. To prepare for manual validation, we extracted a list of question forms, each one associated to a numerical ID representing the cluster it belongs to. Questions in the same cluster (e.g., “Who is the author?” and “Who made it?”) were placed close to each other. After removing identical sentences, we obtained about 1,659 questions, grouped in 126 clusters. Each question was then manually associated to a textual (human meaningful) ID (e.g., “AUTHOR”) agreed by the annotators and a special “NODATA” ID (about 10%) was introduced for questions that refer to information that is not contained in ArCo. At the end of the process, after excluding clusters that ask for unavailable and unusable information, we obtained 29 clusters, each of them representing a question type. Obtained question types (labeled as “User”) were aggregated with the ones from the domain experts (labeled as “Expert”) obtaining 43 question types, with 20 of them in common. In addition, the experts grouped the question types into three categories based on their nature. As depicted in Fig. 4.1, most questions (31) were labeled as “contextual”, as it was not possible to find the appropriate answers in the images associated with the question type considered (e.g., DATING). Instead, eight question types were defined as “visual” since the answers can be inferred from the images associated to the cultural asset, while for four “mixed” question types the answers derive both from vi-

sual and contextual information. Eventually, the experts defined an answer template and a SPARQL query for each question type.

We employed SparqlWrapper⁵ for executing the SPARQL queries and extracting textual data and pictures from ArCo. We removed cultural assets that have zero or more than one associated pictures. For each record of the query results we generated a question-answer pair by randomly drawing a question verbal form by the set of appropriated verbal forms in the associated question cluster, with the same distribution of the results of the user questionnaires (frequently proposed questions were selected with higher probability), and building the associated answer from the answer template.

Some question verbal forms are appropriate only for specific types of cultural assets (e.g., “who was it painted by?” makes sense only for paintings). To establish the appropriated verbal forms for a cultural assets we mapped both question verbal forms and cultural assets with corresponding macro-categories (we defined nine macro-categories, i.e., SCULPTURE, OBJECT, PHOTO, FRESCO, CHURCH, FIND, PRINT, PAINTING, OTHER). Since this information is not available in ArCo, we considered the available textual description of the cultural asset category to build the mapping. Due to the multitude of categories, we performed a filtering and mapping operation to bring the wide range of types back into a small but explanatory set. As a state-of-the-art work on Italian cultural heritage, we took into account the controlled vocabularies defined by the ICCD-MiC⁶, which also provided the data for ArCo KG [16]. These controlled vocabularies ensure a standardized terminology for the description and cataloging of cultural heritage and help overcome the semantic heterogeneity that is often present in creating such catalogs. First, we filtered the vocabularies’ elements closest to the type of artworks to which users refer in their questions. We mapped each textual description of category with an entry in the controlled vocabularies. As detailed in [15], we used a string matching algorithm that takes as input a list of words from a well-defined taxonomy and a general description in free text and returns the equivalent term from the reference taxonomy.

In order to improve both the form of the answer itself and its rendering in its context, we adopted two approaches. First, we applied a set of cleaning rules, such as removing data with errors and changing patterns of verbal forms (e.g., from “Baldin, Luigi” to “Luigi Baldin”)⁷. Second, we employed

⁵<https://github.com/RDFLib/sparqlwrapper>

⁶<http://www.iccd.beniculturali.it/it/strumenti-terminologici>

⁷a complete list is available on <https://github.com/misael77/IDEHADataset>

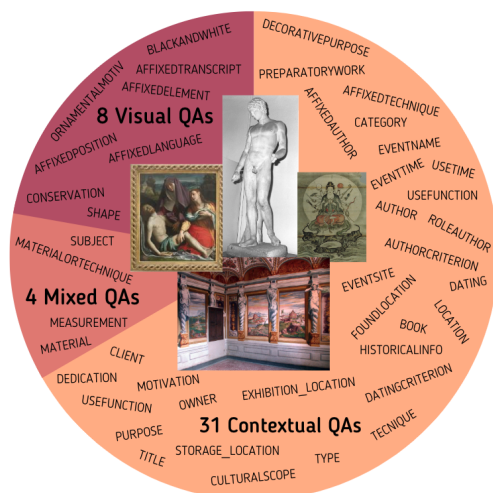


Figure 4.1: Overview of the 43 question types of QA labeled as "visual", "contextual" and "mixed". At the center some images of the types of cultural assets (e.g., PAINTING, SCULPTURE, PRINT, FRESCO) present in VISCONTATH.

pre-trained language models to improve the form of conversational answers by adapting each sentence to its associated datum (e.g., Italian prepositions and articles have to be chosen according to the gender and number of corresponding nouns or adjectives). To solve this problem we applied the cloze task of BERT [27] on the generated answers, asking to infer words whose genre and number depend on the specific datum and cannot be previously determined.⁸ Furthermore, we applied a final grammar correction task by automatic translating the sentence from Italian to English and back to Italian by means of a pre-trained language models for translation⁹.

Eventually, we automatically generated the description of each cultural asset by combining the long answers of all associated question-answer pairs, since this information is not available in ArCo.

⁸<https://huggingface.co/dbmdz/bert-base-italian-uncased>

⁹<https://huggingface.co/Helsinki-NLP/opus-mt-it-en> and [opus-mt-en-it](https://huggingface.co/Helsinki-NLP/opus-mt-en-it)

4.2.2 A Large and Detailed VQA Dataset for Cultural Heritage

The generated VQA dataset contains 6.49M question-answer pairs covering cultural assets, 43 question types and 427 verbal forms. The number of question-answer pairs per template ranges from 35 to 576K. Each question-answer pair is associated with the corresponding cultural asset and its information, including its picture, a description and its URI in ArCo. The number of question types associated to each image depends on the cultural asset’s type and ranges from a minimum of 1 to a maximum of 26

The final dataset is the largest resource available for training and validating VQA models in the cultural heritage domain. It comprises 6.493.867 question-answer pairs, with associated visual, textual and structured information. In Tab. 4.2, we report this data in comparison to the AQUA [36] dataset statistics. In contrast to AQUA, we consider a new dimension that incorporates mixed (contextual and visual) question types. Additionally, our dataset is two orders of magnitude larger than AQUA. We associate each cultural asset in our dataset with a set of question-answer pairs, with both a long conversational answer and a short synthetic answer, an image, a natural language description, its URI in ArCo, the reference ontology class and its type. In addition, we provide information on the text span of the answer in the description, when possible.

We make our dataset available on GitHub¹⁰. We also provide two samples in Italian and English of 50 question-answer pairs per question type that we manually evaluated. Results show an overall accuracy of the long answers (percent of correct entries) of 96,6% for the Italian sample, and of 93% for the English one. We also provide statistics that reports, for each question type, its usage, the number of associated question forms, the number of question-answer pairs generated, and the accuracy. Tab. 4.1 shows the breakdown of the number of question-answer pairs by cultural asset type and question type. The distribution of cultural asset types in the dataset is provided in Fig. 4.3. The most common question type are “TYPE”, “TITLE” and “MATERIALORTECHNIQUE” while “EVENTSITE”, “PURPOSE” and “BLACKANDWHITE” have fewer associated cultural assets. Excluding cultural assets not classified in a specific category (“OTHER”), the macro categories with more elements are “OBJECT” (26%) and “PAINTING” (13%) while the less populated one is “FRESCO” (1%).

¹⁰Cf. <https://github.com/misael177/IDEHADataset>

Table 4.1: Number of question-answer pairs by cultural asset typology

question type	PHOTO	FINDS	PAINTING	SCULPTURE	OBJECT	CHURCH	FRESCO	PRINT	Other	Total
TYPE	27,244	0	68,938	24,832	157,849	1,907	19	51,829	244,379	576,997
CONSERVATION	0	0	66,890	21,560	115,554	308	3	51,518	184,124	439,957
DATINGCRITERION	0	0	64,075	21,107	116,134	560	4	50,074	187,720	439,674
CULTURALSCOPE	0	0	26,744	13,765	96,606	1,828	3	9,976	140,848	289,770
DATING	25,247	0	68,589	23,343	130,031	957	4	51,598	192,023	491,792
OWNER	0	0	65,991	23,443	142,577	1,308	17	50,195	241,347	524,878
PREPARATORYWORK	0	0	14,256	4,790	33,646	15	3	18,672	37,295	108,677
CLIENT	0	0	4,310	1,170	641	0	0	1,963	4,153	11,937
TITLE	0	0	68,364	24,683	157,037	1,753	18	50,975	267,023	569,853
SUBJECT	0	0	64,307	19,904	67,791	0	3	48,102	94,701	294,898
MATERIALORTECHNIQUE	0	0	68,871	24,177	150,141	0	19	51,220	244,285	538,713
AUTHOR	21,432	0	37,994	7,523	34,128	221	0	40,507	40,105	181,910
LOCATION	0	104,210	47,797	14,580	103,088	0	0	48,426	138,830	456,931
MEASUREMENT	0	0	17,131	5,666	84,490	7	19	45,719	116,900	269,932
ROLEAUTHOR	0	0	10,207	2,949	27,387	228	0	18,014	35,828	94,613
AFFIXEDTECHNIQUE	0	0	17,987	2,721	20,012	0	0	22,817	61,846	125,383
AUTHORCRITERION	0	0	36,710	7,393	28,452	95	0	41,122	55,648	169,420
AFFIXEDPOSITION	0	0	19,864	3,235	38,381	50	0	24,442	56,950	142,922
AFFIXEDELEMENT	0	0	23,092	4,186	49,996	68	0	34,567	78,517	190,426
CATEGORY	0	0	0	1,186	29,216	12	15	0	75,102	105,531
AFFIXEDTRANSCRIPT	0	0	21,272	3,420	31,908	33	0	31,117	62,372	150,122
HISTORICALINFO	0	0	18,912	4,776	21,591	3	6	11,807	35,719	92,814
EVENTNAME	0	0	7,764	1,546	4,344	0	0	3,044	4,182	20,880
AFFIXEDLANGUAGE	0	0	6,922	1,082	15,536	0	0	5,890	26,202	55,632
USEFUNCTION	0	0	37	313	4,181	1,392	0	8	12,594	18,252
TECHNIQUE	0	0	36	315	4,016	0	0	0	13,543	17,910
USETIME	0	0	0	3	551	44	0	0	1,171	1,769
FOUNDLOCATION	0	11,173	25	1	557	0	0	16	129	11,901
EVENTTIME	0	0	7,318	1,536	4,247	0	0	3,509	3,810	20,420
MOTIVATION	0	0	2,151	960	319	0	0	1,402	2,756	7,588
MATERIAL	0	0	36	318	5,716	0	0	8	16,716	22,794
SHAPE	0	0	7,180	715	3,255	0	0	3,052	5,617	19,819
AFFIXEDAUTOR	0	0	2,439	225	3,599	0	0	4,325	1,067	11,655
USECONDITIONS	0	0	20	299	1,878	0	0	0	3,998	6,195
DECORATIVEPURPOSE	0	0	0	6	647	0	0	0	1,349	2,002
DEDICATION	0	0	0	0	914	0	0	354	1	1,269
STORAGE.LOCATION	0	0	2,412	58	411	0	0	1,185	862	4,928
EXHIBITION.LOCATION	0	0	758	24	27	0	0	4	92	905
BOOK	0	0	0	0	588	0	0	315	151	1,054
PURPOSE	0	0	0	0	8	11	0	0	104	123
ORNAMENTALMOTIV	0	0	0	0	432	0	0	0	753	1,185
BLACKANDWHITE	0	0	0	0	0	0	0	0	128	128
EVENTSITE	0	0	0	0	2	0	0	0	33	35
Total	73,923	115,383	869,399	267,810	1,687,884	10,800	133	777,472	2,691,063	6,493,867

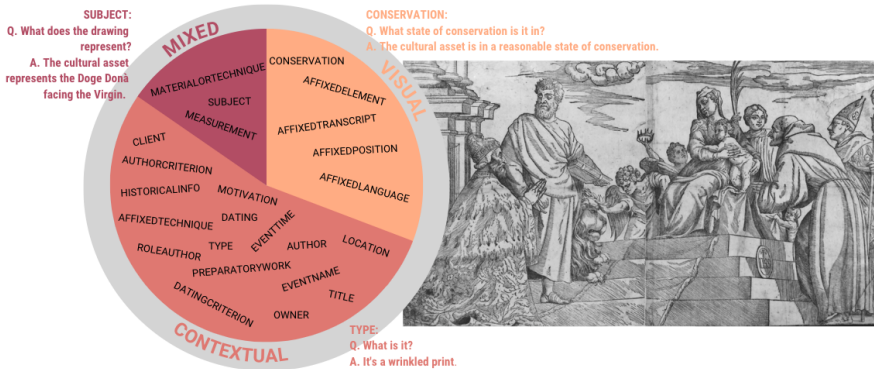


Figure 4.2: Overview of the 26 question types associated to the PRINT representing the Doge Donà facing the Virgin. 16 question types are labeled as “contextual”, five question types are “visual”, and three are “mixed”. For each group three examples of natural language question types (i.e. TYPE, CONSERVATION and SUBJECT) are given.

4.3 A VQA Model for Cultural Heritage

Visual Question Answering for Cultural Heritage requires to analyze two heterogeneous sources of information: an image depicting the artwork and a textual description providing external contextual knowledge. A model capable of effectively providing answers to both visual and contextual questions must therefore combine computer vision and natural language processing. In literature, however, most approaches deal with either one of the two modalities. To understand the challenges posed by our proposed dataset, we first propose single-modality baselines from the state of the art:

- DistilBert [108] is a very common language transformer trained by distilling the Bert base model [27]. It results to be lighter and faster with respect to Bert thanks to knowledge distillation used at training time. For this reason the size of the DistilBert model is 40% lower, while retaining 97% of its language understanding capabilities and being 60% faster. This model can then be fine-tuned with good performances on a wide range of tasks.
- RoBERTa [75] has the same architecture of Bert [27] but is trained with optimized parameters, uses a different tokenizer and a different

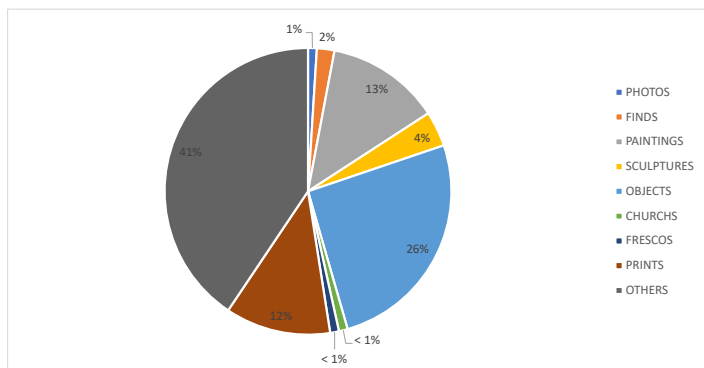


Figure 4.3: Distribution of cultural asset typologies in the VISCOUNTH dataset.

Table 4.2: Comparison of statistics from the VISCOUNTH and AQUA [36] datasets.

	AQUA			VISCOUNTH		
	Train	Val	Test	Train	Val	Test
Visual QA pairs	29,568	1,507	127	800,440	100,003	99,748
Contextual QA pairs	40,244	3,617	3,642	3,492,984	437,101	437,254
Mixed QA pairs	0	0	0	901,672	112,281	112,384
QA pairs	69,812	5,124	4,912	5,195,096	649,385	649,386

pretraining scheme.

- LXMERT [120] is a Large multimodal transformer for vision and language. It consists of three encoders: a visual encoder, a language encoder and a cross-modality encoder. This model is pretrained with large amounts of image-and-sentence pairs via diverse pretraining tasks. It has been shown that this model can achieve impressive results on different downstream multimodal tasks after an appropriate finetuning.

We then propose a multi-modality baseline model by combining DistilBert and LXMERT with a question classifier, that predicts whether the question is contextual or visual and thus if a text-based model (DistilBert) or a vision-based model (LXMERT) is required. Similar approaches have been previously adopted in VQA for cultural heritage [11, 36]. The question

classifier is based on Bert [27]. We finetuned a Bert model with a binary classifier on top. The model predicts if a given question is visual or contextual. Depending on the classifier prediction, the question is passed to the most suitable branch (vision model or text-based model) together with additional information (image or textual description).

4.4 Results and Discussion

4.4.1 Evaluation Metrics

To evaluate VQA models on the collected dataset, we follow the standard evaluation setting proposed in [97]. We rely on two metrics, Exact match and Macro-averaged F1 score:

- *Exact match* measures the percentage of predictions that exactly match the ground truth answer.
- *Macro-averaged F1 score* measures the average overlap between the prediction predicted answer and the ground truth. Both answers are considered as a set of unordered words among which the F1 score is computed. F1 scores are averaged over all questions in the dataset.

Note that for both metrics we do not consider articles and punctuations.

In addition, text-based models generate variable length sentences as a subset of the textual description, whereas vision-based models pick a candidate among a predefined dictionary of possible answers. In both cases, we take the set of words and compare it to the ground truth to compute Exact match and F1 score.

4.4.2 Evaluation

We carry out a quantitative evaluation by first testing off-the-shelf language pre-trained models. We do not expect such models to perform well on visual questions but we want to assess whether such models can exploit their language understanding to comprehend questions relative to the cultural heritage domain. As detailed in Sec. 4.3, we use as text-based models RoBERTa [75] and DistilBert [108]. Both datasets have been pre-trained on SQUAD [97], a reading comprehension dataset with more than 100.000 questions-answer pairs crowd-sourced on a set of Wikipedia articles.

Table 4.3: F1-score and Exact Match (EM) for different models on contextual questions.

Metric	Pretrained		Finetuned				Ours	
	RoBERTa [75]	Distilbert [108]	Distilbert [108]	LXMERT [120]	F1	EM	F1	EM
AFFIXEDTECHNIQUE	0.00	0.06	0.28	0.16	0.00	0.00	0.00	0.00
CULTURALSCOPE	0.00	0.10	0.84	0.40	0.00	0.00	0.84	0.40
EVENTNAME	0.00	0.03	0.97	0.86	0.00	0.00	0.97	0.86
OWNER	0.01	0.10	0.93	0.92	0.00	0.00	0.49	0.27
TECHNIQUE	0.14	0.58	0.46	0.23	0.00	0.00	0.46	0.23
ROLEAUTHOR	0.00	0.15	0.64	0.57	0.00	0.00	0.64	0.57
TYPE	0.03	0.08	0.29	0.20	0.00	0.00	0.22	0.18
LOCATION	0.03	0.15	0.96	0.91	0.00	0.00	0.96	0.91
TITLE	0.03	0.21	0.98	0.97	0.00	0.00	0.93	0.90
DATING	0.01	0.40	0.73	0.71	0.00	0.00	0.73	0.71
DATINGCRITERION	0.00	0.01	0.81	0.66	0.00	0.00	0.81	0.66
HISTORICALINFO	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00
AUTHORCRITERION	0.12	0.03	0.52	0.43	0.00	0.00	0.52	0.43
CATEGORY	0.00	0.06	0.39	0.16	0.00	0.00	0.39	0.16
AUTHOR	0.01	0.19	0.99	0.91	0.00	0.00	0.99	0.91
DEDICATION	0.24	0.38	0.98	0.96	0.00	0.00	0.98	0.96
USEFUNCTION	0.38	0.33	0.96	0.92	0.00	0.00	0.96	0.92
FOUNDLOCATION	0.01	0.29	1.00	1.00	0.00	0.00	1.00	1.00
EVENTTIME	0.03	0.03	0.32	0.03	0.00	0.00	0.32	0.03
PREPARATORYWORK	0.14	0.02	0.99	0.99	0.00	0.00	0.99	0.99
STORAGE.LOCATION	0.01	0.08	0.96	0.96	0.00	0.00	0.96	0.96
CLIENT	0.07	0.21	0.95	0.91	0.00	0.00	0.95	0.91
DECORATIVEPURPOSE	0.13	0.18	0.00	0.00	0.00	0.00	0.00	0.00
USECONDITIONS	0.04	0.07	0.96	0.47	0.00	0.00	0.96	0.47
MOTIVATION	0.01	0.13	0.89	0.49	0.00	0.00	0.98	0.49
EXHIBITION.LOCATION	0.01	0.03	0.67	0.63	0.00	0.00	0.67	0.63
AFFIXEDAUTHOR	0.01	0.46	0.86	0.67	0.00	0.00	0.89	0.67
USETIME	0.18	0.04	0.95	0.75	0.00	0.00	0.95	0.75
PURPOSE	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
BOOK	0.10	0.08	0.57	0.54	0.00	0.00	0.57	0.54
EVENTSITE	0.00	0.00	0.55	0.55	0.00	0.00	0.55	0.55
Mean Contextual	0.06	0.15	0.69	0.58	0.00	0.00	0.67	0.55

Interestingly, when evaluated on contextual questions, such models perform poorly as can be seen in Tab. 4.3. Both models are capable of answering with a certain degree of correctness to a few question categories, namely DEDICATION and USEFUNCTION, with DistilBert obtaining good F1 scores on an additional restricted number of categories such as TECHNIQUE and AFFIXEDAUTHOR. For most of the remaining question categories we report an F1 close to 0. This suggests the presence of a domain shift between standard question answering datasets (such as SQUAD) and VISCOUNTH. In fact, in art related question-answers, as well as descriptions, there is often usage of domain specific jargon that is not present in generic text corpora, making the models unable to understand the question or identify the answer within the description.

Nonetheless, although unlikely given the proven capabilities of such pre-trained models, a low F1 could be caused by intrinsic limits in the architectures. To further confirm the presence of a domain shift, rather than some form of model limitation, we fine-tuned the best of the two models, DistilBert, on the VISCOUNTH dataset. This leads to a significant improvement. The model gains on average 54 points of F1-score, obtaining close to perfect results for question types such as TITLE, AUTHOR, FOUNDLOCATION and PREPARATORYWORK. Interestingly, for other categories instead DistilBert still reports low scores, close to zero (HISTORICALINFO, DECORATIVEPURPOSE, PURPOSE). These categories however either are less represented in the data as shown in Tab. 4.1 or are intrinsically harder. For instance, the HISTORICALINFO category presents a high variability in how questions are formulated and frequently asks for generic concepts, which require a high level reasoning on the description content.

We also perform a similar evaluation with the vision-based model LXMERT [120]. However, two issues must be taken into account. First, as in most vision-based models since they cannot rely on textual descriptions, the VQA task is treated as a classification task. Answering a question corresponds to selecting the most relevant answer among a dictionary of pre-defined words or short sentences. For this reason, the domain shift is much more emphasized: if the dictionary does not contain terms suitable for cultural heritage the model will not perform well. Second, whereas a text-based model could answer visual questions if the requested information is also in the description, a vision-based model cannot answer contextual questions in any way. As a consequence, we cannot apply a pre-trained vision-model due to signifi-

cant differences in the answer dictionary. But even fine-tuning the model on VISCOUNTH leads to an F1-score of 0. In order to perform such finetuning, we create a new dictionary of answers by filtering the most frequent answers in the training set. More precisely we selected the answers that appear more than 8 times.

Moving to mixed questions (Tab. 4.4), on the one hand we can observe a similar behaviour for text-based models, although the overall F1-score is much lower since visual knowledge is required to answer correctly. On the other hand, LXMERT is able to provide correct answers to some of the questions. Notably, for the MATERIAL question type, LXMERT surpasses text-based models by a considerable margin, yet it is unable to answer to MEASUREMENT questions, contrary to DistilBert.

As expected, for visual questions we can observe an opposite trend compared to contextual questions. In Tab. 4.5 we report the results, showing that LXMERT can provide for almost all question categories a high rate of correct questions. However, after being fine-tuned on VISCOUNTH, DistilBert is capable of addressing questions related to AFFIXEDTRANSCRIPT and BLACKANDWHITE. This is due to the fact that sometimes the answers can also be found in the textual description.

For most experiments we report both the macro-averaged F1-score and the Exact Match (EM) metrics. It can be noticed that the F1 score is a relaxation of the EM metric in the sense that it allows an answer to be loosely compared to the ground truth, even when not all words are the same, thus accounting for synonyms or different phrasings.

Finally, we evaluate our combined model. We exploit the question classifier to understand which model is more suitable to address a specific question, without looking at the description nor the image. The BERT-based classifier, described in Sec. 4.3, obtains a question classification accuracy of 98.4% on the test set, indicating that it is fully capable of understanding the nature of the questions. We do not include mixed questions in training and at inference time we consider the question to be either visual or contextual based on the output of the classifier.

As can be seen from Tab. 4.3, Tab. 4.4 and Tab. 4.5, the model is able to exploit both models to accurately answer visual and contextual questions, with only a slight drop for language-based samples. For mixed questions, our model is able to improve compared to LXMERT but exhibits a drop compared to DistilBert. This confirms that mixed questions indeed pose a

Table 4.4: F1-score and Exact Match (EM) for different models on mixed questions.

Metric	Pretrained		Finetuned				Ours	
	RoBERTa [75]	Distilbert [108]	Distilbert [108]		LXMERT [120]		F1	EM
	F1	F1	F1	EM	F1	EM	F1	EM
MATERIALORTECHNIQUE	0.00	0.27	0.36	0.32	0.27	0.16	0.36	0.32
SUBJECT	0.04	0.13	0.00	0.00	0.00	0.00	0.00	0.00
MEASUREMENT	0.00	0.04	0.84	0.68	0.00	0.00	0.00	0.00
MATERIAL	0.00	0.39	0.09	0.04	0.29	0.14	0.29	0.14
Mean Mixed	0.01	0.21	0.32	0.26	0.14	0.07	0.16	0.11

challenge yet to be solved in question answering applications.

In Tab. 4.6 we report the overall average scores in terms of F1 and Exact Match. The average is computed as the mean of all category scores, i.e. contextual, mixed and visual together. Our combined model retains the best results, providing a baseline for future work in visual question answering for cultural heritage.

To better understand the challenges in the dataset, we show a breakdown of results divided by question category and type of cultural property in Tab. 4.7. We do this only for visual questions, since contextual questions do not exploit visual information. This table shows how the performance of our approach vary depending on the type of artwork. We can observe, as expected, that there is a gap between the score obtained for different types of artwork on specific question classes. As example the question category CONSERVATION (that includes questions about the conservation state of the artwork) results easier for prints than sculptures. Vice-versa, the category AFFIXEDLANGUAGE (that has questions about the language of the writing attached to the cultural asset) has better results for sculptures. Finally, we can observe that the category AFFIXEDTRANSCRIPT, that refers to the text present in the artwork, obtains very low results. This is due to the fact that these kind of questions are very challenging and require the extraction and the understanding of text in images and currently this can be done only with specific networks.

Table 4.5: F1-score and Exact Match (EM) for different models on visual questions.

Metric	Pretrained		Finetuned				Ours	
	RoBERTa [75]	Distilbert [108]	Distilbert [108]		LXMERT [120]		F1	EM
CONSERVATION	0.00	0.01	0.00	0.00	0.79	0.53	0.79	0.53
AFFIXEDLANGUAGE	0.13	0.66	0.01	0.01	0.67	0.66	0.66	0.66
AFFIXEDELEMENT	0.00	0.01	0.00	0.00	0.83	0.83	0.83	0.83
AFFIXEDTRANSCRIPT	0.02	0.08	0.80	0.69	0.05	0.04	0.04	0.04
AFFIXEDPOSITION	0.00	0.01	0.00	0.00	0.47	0.32	0.47	0.32
SHAPE	0.00	0.00	0.00	0.00	0.68	0.68	0.68	0.68
ORNAMENTALMOTIV	0.00	0.00	0.00	0.00	0.54	0.54	0.54	0.54
BLACKANDWHITE	0.00	0.00	0.70	0.70	0.96	0.96	0.96	0.96
Mean Visual	0.02	0.10	0.19	0.17	0.62	0.57	0.62	0.57

Table 4.6: F1-score and Exact Match (EM) for different models averaged over all question types.

Metric	Pretrained		Finetuned				Ours	
	RoBERTa [75]	Distilbert [108]	Distilbert [108]		LXMERT [120]		F1	EM
Mean Overall	0.05	0.14	0.57	0.47	0.13	0.11	0.61	0.51

Table 4.7: F1-score breakdown for cultural asset category and question type. We do not report the PHOTO and FIND categories since no visual question is present for such artworks.

	PRINT	OBJECT	OTHER	PAINTING	SCULPTURE	FRESCO	CHURCH
CONSERVATION	0.81	0.79	0.78	0.78	0.77	1.00	0.34
AFFIXEDLANGUAGE	0.61	0.63	0.69	0.78	0.87	-	-
AFFIXEDELEMENT	0.89	0.89	0.78	0.96	0.82	-	0.57
AFFIXEDTRANSCRIPT	0.07	0.09	0.03	0.01	0.01	-	0.00
AFFIXEDPOSITION	0.54	0.61	0.40	0.32	0.22	-	0.11
SHAPE	0.81	0.73	0.71	0.59	0.46	-	-
ORNAMENTALMOTIV	-	0.56	0.54	-	-	-	-
BLACKANDWHITE	-	-	0.96	-	-	-	-

4.4.3 Qualitative Analysis

In this section we provide a qualitative analysis of the answers given by our approach to questions in the VISCONT^H dataset.

The dataset is divided into three main question types: visual, contextual and mixed. For each type there are multiple question categories, which refer to different types of cultural assets. We thus expect the answers given by our model to be affected by all these aspects. In Fig. 4.4 we show the behaviour of our model in answering different kinds of questions for different types of cultural assets. For contextual questions we expect that the answer has to be extracted from a natural language description, therefore a language model is sufficient to answer these questions. As we can see in Tab. 4.1 and Tab. 4.3, our model is able to answer the most common contextual questions in the dataset but has lower performance for questions that appear in few examples. In Fig. 4.4 we can observe how our model is able to answer correctly to different categories of contextual questions (**LOCATION**, **AUTHOR**, **TITLE**, **DATING**, etc.) for different types of artworks. For these types of questions we do not observe different performances for different types of artworks. This is due to the fact that in these cases, our question answering language model is agnostic to visual information, being solely based on textual descriptions.

Confirming the results of Tab. 4.4, we observe that our model obtains low performances on mixed questions. This kind of questions result to be very challenging since they require both visual knowledge and contextual knowledge. For instance, for the **MATERIAL** category, the model should be able to describe the different materials the artworks are made of and learn how to recognize them visually. Our model selects either the vision-based model or the textual-based model to answer a question, hence there is not a specific way to handle this kind of questions, thus leading to a lack of performance.

Regarding visual questions, we can observe from Tab. 4.7 that we have a variation in the performances based on the type of artwork for different classes of visual questions. For example we can observe that the questions of the **SHAPE** category, that refers to the shape of the artwork, as expected, perform better for prints than for sculptures. Moreover, as shown in Fig. 4.4, several artworks contain transcripts and there is a specific question category (**AFFIXEDTRANSCRIPT**) for this detail. Our model obtains very low performance on this question class since it does not contain a specific trained model





	<p>Q: CONTEXTUAL / LOCATION: Where is the painting kept? A: Uffizi Gallery ✓</p> <p>Q: MIXED / SUBJECT: Who does it represent? A: ✗</p> <p>Q: MIXED / MATERIALORTECHNIQUE: What is the material and technique used? A: table, oil painting ✓</p> <p>Q: CONTEXTUAL / AUTHORCRITERION: On the basis of what criterion is the cultural asset attributed to the author? A: Bibliography ✓</p>		<p>Q: MIXED / MATERIALORTECHNIQUE: What are the techniques and materials? A: Paper, Etching ✓</p> <p>Q: CONTEXTUAL / ROLEAUTHOR: What role did the author play in creating the work?: A: Engraver ✓</p> <p>Q: CONTEXTUAL / DATING: When was the drawing made? A: 1559 ✓</p> <p>Q: VISUAL / AFFIXEDPOSITION: Where in the cultural asset is the element posted? A: Bottom right corner ✓</p>
	<p>Q: VISUAL / AFFIXEDTRANSCRIPT: What do the written sentences say? A: ✗</p> <p>Q: VISUAL / CONSERVATION: What is the state of preservation of the work?: A: mediocre state of conservation ✓</p> <p>Q: CONTEXTUAL / AUTHOR: Who's the author? A: De Finetti Gino ✓</p> <p>Q: CONTEXTUAL / TITLE: What's it called? A: Horse show in Trieste ✓</p>		<p>Q: CONTEXTUAL / AUTHOR: Who took the photograph? A: Anonymous ✓</p> <p>Q: MIXED / MATERIALORTECHNIQUE: What are the materials and techniques of realization?: A: Collodium, glass ✓</p> <p>Q: CONTEXTUAL / AFFIXEDTECHNIQUE: What are the technical characteristics of the element attached to the cultural asset? A: Graffito ✓</p> <p>Q: VISUAL / AFFIXEDTRANSCRIPT: What does it say in the captions? A: ✗</p>

Figure 4.4: Qualitative Results. Answers given by our approach for different question category/class on different artwork types.

for scene text extraction.

4.5 Conclusions

We presented a large scale heterogeneous multi-language dataset for visual question answering in the cultural heritage domain. Our dataset contains approximately 6.5M question-answer pairs in Italian and English, spanning 500K cultural assets of different types, including artworks, churches, historical objects and others. Each cultural asset is associated to an image, a natural language description and other information. We presented some baselines that employ and combine machine learning models for both contextual (natural language description) and visual processing. Our results show that fine-tuning on a domain-specific dataset is crucial for this task, thus confirming the utility of our dataset. Our best model achieves an overall accuracy (F1 average) of 0.61. Although these result is promising, we found out that certain question categories are hard to compute, especially the ones that require mixed (visual and contextual) reasoning. We believe that further research in this direction would be beneficial for the cultural heritage field, as well as for other fields where multi-modal (visual and natural

language) reasoning is required.

Chapter 5

GADA: Generative Adversarial Data Augmentation for Image Quality Assessment

We propose a No-reference Image Quality Assessment (NR-IQA) approach based on the use of generative adversarial networks. To address the problem of lack of adequate amounts of labeled training data for NR-IQA, we train an Auxiliary Classifier Generative Adversarial Network (AC-GAN) to generate distorted images with various distortion types and levels of image quality at training time. The trained generative model allows us to augment the size of the training dataset by introducing distorted images for which no ground truth is available. We call our approach Generative Adversarial Data Augmentation (GADA) and experimental results on the LIVE and TID2013 datasets show that our approach – using a modestly sized and very shallow network – performs comparably to state-of-the-art methods for NR-IQA which use significantly more complex models. Moreover, our network can process images in real time at 120 image per second unlike other state-of-the-art techniques.¹

¹The work described in this chapter was presented at the International Conference on Image Analysis and Processing (ICIAP), 2019.

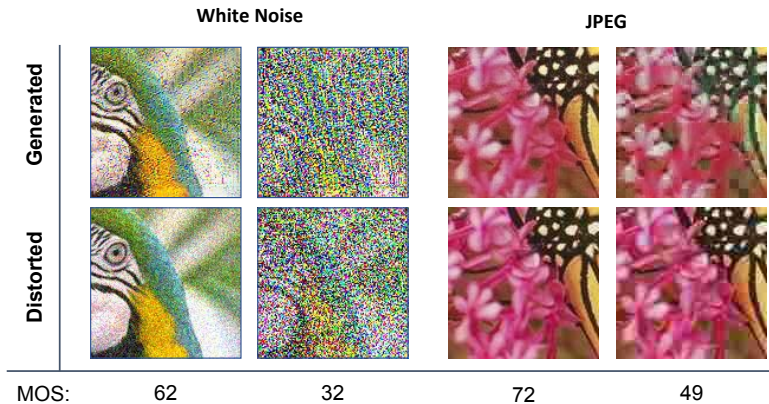


Figure 5.1: Patches extracted from images generated by the proposed method compared with the same patches from true distorted images with the same image quality and distortion type.

5.1 Introduction

In Chapters 2,3,4 we focused on the task of Visual Question Answering in the Cultural Heritage domain addressing different problems connected to real scenarios. In the next three chapters, we introduce another visual and language task: image captioning. In this thesis, the research related to this topic is strictly connected with its application to Image Quality Assessment. In this preliminary work, we propose an approach to address the lack of large labeled datasets for IQA. Since obtaining annotated data to train the network is difficult, we propose a technique to generate new images with a specific image quality and distortion type. We learn how to generate distorted images using Auxiliary Classifier Generative Adversarial Networks (AC-GANs), and then use these generated images in order to improve the accuracy of a simple CNN regressor trained for IQA. In Fig. 5.1 we show patches of images generated with our approach alongside their corresponding patches with real distortions.

5.1.1 Auxiliary Classifier GANs.

In the last few years GANs have been widely used in different areas of computer vision. The Auxiliary Classifier GAN (AC-GAN) [88] is a variant

of the Generative Adversarial Network (GAN) [39] which uses label conditioning. This kind of network produces convincing results. Our aim is to use this architecture to generate distorted images conditioned to a distortion category and image quality value. Since the main objective of the work is NR-IQA and the performance of the quality regressor is highly related to the generated image, it is crucial that the generator produce convincing distortions.

5.2 Generative Adversarial Data Augmentation for NR-IQA

In this section we describe our approach to perform data augmentation for NR-IQA datasets. We first show the general steps that characterize our technique, and then describe the use of AC-GAN in this context.

5.2.1 Overview of Proposed Approach

The main idea of this work is to generate new distorted images with a specific image quality level and distortion type to partially solve the problem of the poverty of annotated data for IQA. We use an AC-GAN to generate new distorted images. Once the generator has learned to produce distorted images convincingly we use it to generate new examples to augment the training set as we train a deep convolutional regressor to estimate IQA. The pipeline of our technique is as follows:

1. **Training the AC-GAN.** Using patches of the training images we train an AC-GAN. The generator learns to generate distorted images with a given distortion class and quality level starting from reference images. The regressor, which aims is to predict the image quality, is trained with both generated and real distorted images using the adversarial GAN loss.
2. **Generative data augmentation.** Once the training of the AC-GAN converges, the generator is able to produce convincing distortions and we can stop its training. We continue training the discriminator branch, augmenting the training data via the trained generator. The regressor is trained with both real distorted images from the training set and images artificially distorted using the generator.

3. **Fine-Tuning of the regressor.** Once convergence is reached in step 2 we perform a final phase of fine-tuning: the regressor is trained with only real distorted images from the IQA training set.

5.2.2 Auxiliary Classifier GANs for NR-IQA

An Auxiliary Classifier Generative Adversarial Network is a GAN variant in which it is possible to condition the output on some input information. In the AC-GAN every generated sample has a corresponding class label, $c \sim p_c$, in addition to the noise z . This information is given as input to the generator which produces fake images $X_{\text{fake}} = G(c, z)$. The discriminator not only distinguishes between real and generated examples but predicts also the class label of the examples. The sub-network that classifies the input is called the *classifier*. The objective function is characterized by two components: a log-likelihood on the correct discrimination L_S and a log-likelihood on the correct class L_C :

$$L_S = E[\log P(S = \text{real} \mid X_{\text{real}})] + E[\log P(S = \text{fake} \mid X_{\text{fake}})] \quad (5.1)$$

$$L_C = E[\log P(C = c \mid X_{\text{real}})] + E[\log P(C = c \mid X_{\text{fake}})] \quad (5.2)$$

The discriminator is trained to maximize $L_S + L_C$ and the generator is trained to minimize $L_C - L_S$.

Our approach is slightly different from a standard AC-GAN: the latter expects only noise and class label as input, but in our case we want to generate an output image that is a *distorted* version of a reference one, so we also need to feed the reference image and force a reconstruction with an $L1$ loss. Moreover, we want to distort the reference image so that the output matches a target *image quality*, so we feed also this value as input. Because we would like to reconstruct a distorted version of the reference image given as input, we can write the additional $L1$ loss as it follows:

$$\mathcal{L}_{L1} = E[\|y - G(z, x, c, q)\|_1]$$

where y is the distorted ground truth image, z is a random Gaussian noise vector, x is the reference image, c is the distortion class and v is the image quality.

The goal of this work is to predict the quality score of images, so we introduce a regressor network whose aim is to predict the quality score of

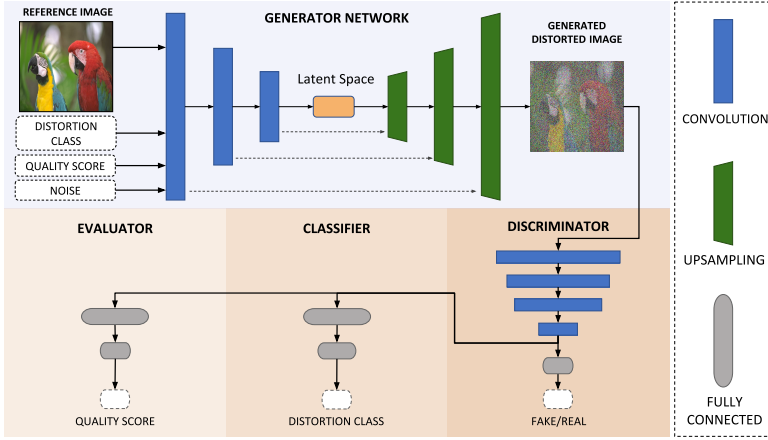


Figure 5.2: A schematic representation of the proposed network.

input images. The loss used to train this component is a mean squared error (MSE) between the predicted quality score and the ground truth:

$$L_E = E[(q - \hat{q})^2] \quad (5.3)$$

where q and \hat{q} are the ground truth and the prediction of the image quality score, respectively.

The expectations for all losses defined here are taken over minibatches of either generated or labeled training samples.

5.2.3 The GADA Architecture

In Fig. 5.2 we give a schematic representation of the proposed model. The components of the GADA network are as follows.

Generator. The Generator follows the general auto-encoder architecture. It takes as input a high quality reference image, a distortion class, and a target image quality. The input information is encoded through three convolutional layers (one with 64 feature maps and two with 128). Before up-sampling we concatenate a noise vector z to the latent representation, together with an embedding of the distortion category and image quality. We use skip connections [48, 102] in the generator, which allows the network to generate qualitatively better results.

Discriminator. The Discriminator takes as input a distorted image and through three convolutional layers (one with 64 feature maps, and two with 128 to mimic the encoder) followed by a 1×1 convolution extracts 1024 feature maps (that are also fed to the classifier and the regressor). A single fully-connected layer reduces these feature maps to a single value and a sigmoid activation outputs the prediction of the provenance of the input image (i.e. real or fake). This output is used to compute the loss defined in equation 5.1.

Classifier. The Classifier takes as input the feature maps described for the Discriminator. This network consists of two fully-connected layers. The first layer has 128 units and the second has a number of units equal to the number of distortion categories and is followed by a softmax activation function. The output of this module is used in the classifier loss for the AC-GAN as defined in equation 5.2.

Evaluator. The Evaluator takes as input the feature maps described for the Discriminator and should accurately estimate the image quality of the input image. This module consists of two fully-connected layers, the first with 128 and the second with a single unit. The MSE loss defined in equation 5.3 is computed using the output of this module.

5.3 Experimental Results

In this section we describe experiments conducted to evaluate the performance of our approach. We first introduce the datasets used for training and testing our network, then we describe the protocols adopted for the experiments.

Datasets. For our experiments we used the standard LIVE [109] and TID2013 [93] datasets for IQA. LIVE contains 982 distorted versions of 29 reference images. Original images are distorted with five different types of distortion: JPEG compression (JPEG), JP2000 compression (JP2K), white noise (WN), gaussian blur (GB) and fastfading (FF). The ground truth quality score for each image is the Difference Mean Opinion Score (DMOS) whose value is in the range $[0, 100]$. TID2013 consist of 3000 distorted images versions of 25 reference images. The original images are distorted with 24 different types of distortions. The Mean Opinion Score of distorted images varies from 0 to 9.

Experimental Protocols. We analyze the performance of our model using the standard IQA metrics. For each dataset we randomly split the reference images (and their corresponding distorted versions) in 80% used for training and 20% used for testing, as described in [52, 149]. This process is repeated ten times. For each split we train from scratch and compute the final scores on the test set.

Training Strategy. At each training epoch, we randomly crop each image in the training-set using patches of 128×128 pixels and feed it to the model. For all the three phases we train using these crops with a batch size of 64. During the first one we use Adam optimizer with a learning rate of $1e^{-4}$ for the discriminator and $5e^{-4}$ for the generator, classifier and evaluator. During the second and third phases we divide the learning rate by 10.

Testing Protocol. At test time we randomly crop 30 patches from each test image as suggested in [8]. We then pass all 30 crops through the discriminator network (with only the evaluator branch) to estimate IQA. The average of the predictions for the 30 crops gives the final estimated quality score.

Evaluation Metrics. We use two evaluation metrics commonly used in IQA context: the Linear Correlation Coefficient (LCC) and Spearman Correlation Coefficient (SROCC). LCC is a measure of the linear correlation between the ground truth and the predicted quality scores. Given N distorted images, the ground truth of i -th image is denoted by y_i , and the predicted score from the network is \hat{y}_i . The LCC is computed as:

$$LCC = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}} \quad (5.4)$$

where \bar{y} and $\bar{\hat{y}}$ are the means of the ground truth and predicted quality scores, respectively.

Given N distorted images, the SROCC is:

$$SROCC = 1 - \frac{6 \sum_{i=1}^N (v_i - p_i)^2}{N(N^2 - 1)}, \quad (5.5)$$

where v_i is the *rank* of the ground-truth IQA score y_i in the ground-truth scores, and p_i is the *rank* of \hat{y}_i in the output scores for all N images. The SROCC measures the monotonic relationship between ground-truth and estimated IQA.

Generative Data Augmentation with AC-GAN. As described in Sec. 5.2.1 our approach consists of three phases: a first one where we train the generator, a second phase where we perform data augmentation, and the final fine-tuning phase of the evaluator over the original training-set. As a first experiment, we calculated the performance obtained after each of the three different phases and compared with the performance of a direct method which consists of training *only* the evaluator and classifier branches of the discriminator directly on labeled training data (e.g. no adversarial data augmentation). We trained and tested the proposed method and the direct baseline on the LIVE dataset as described in Sec. 5.3, but for this preliminary experiment we used crops of 64×64 pixels and a shallower regression network.

In Tab 5.1 we give the LCC and SROCC values computed for the baseline and after each of the three phase of our approach. We note first that each phase of our training procedure results in improved LCC and SROCC, which indicates that generative data augmentation and fine-tuning both add to performance. At the end of phase 3 the LCC and SROCC results surpass the direct approach by $\sim 2\%$, confirming the effectiveness of GADA with respect to direct training.

5.3.1 Comparison with the state-of-the-art

Here we compare GADA with state-of-the-art results from the literature.

Results on LIVE. We trained on LIVE dataset following the protocol described in 5.3. The results are shown in Tab. 5.2. Each column of the table represents the partial scores for a specific distortion category of LIVE dataset. Our method seems to be very effective on this dataset despite the fact that many other approaches process larger patches (e.g. 224×224 , the input size of the VGG16 network) and capture more context information. We observe from the table that our model performs very well on Gaussian noise (GN) and JPEG2000 (JP2K). We obtain worse results for Fast Fading (FF), which is probably due to the fact that FF is a local distortion and we process patches of small dimension, so for each crop the probability of picking a distorted region is not 1.

TID2013 We follow the same test procedure for TID2013 and report our SROCC results in Tab. 5.3. We see that for 11 of the 24 types of distortion we obtain the best results. For local and challenging distortions like #14,

		JP2K	JPEG	WN	GBLUR	FF	ALL
Baseline	LCC	0.950	0.964	0.973	0.938	0.933	0.943
	SROCC	0.938	0.931	0.977	0.939	0.898	0.935
Phase 1	LCC	0.944	0.952	0.967	0.920	0.912	0.933
	SROCC	0.933	0.930	0.980	0.926	0.889	0.930
Phase 2	LCC	0.958	0.958	0.974	0.939	0.924	0.942
	SROCC	0.941	0.933	0.988	0.945	0.891	0.939
Phase 3	LCC	0.959	0.973	0.993	0.953	0.935	0.962
	SROCC	0.955	0.941	0.990	0.953	0.912	0.955

Table 5.1: Comparison of baseline and each phase of the GADA approach in LCC and SROCC. In the first block results for the direct baseline method (directly training the evaluator with only labeled IQA data) are shown. In the second block results for our method are shown after each of the three phases: training of the AC-GAN (Phase 1), generator data augmentation (Phase 2), and evaluator fine-tuning (Phase 3).

#15 and #16 the performance of our model is low, and again we hypothesize that the small size and uniform sampling of patches could be a limitation especially for extremely local distortions.

5.4 Conclusions

In this work we proposed a new approach called GADA to resolve the problem of lack of training data for No-reference Image Quality Assessment. Our approach uses a modified Auxiliary Classifier GAN. This technique allows us to use the generator to generate new training examples and to train a regressor which estimates the image quality score. The results obtained on LIVE and TID2013 datasets show that our performance is comparable with the best methods of the state-of-the-art. Moreover, the very shallow network used for the regressor can process images with an high frame rate (about 120 image per second). This is in stark contrast to state-of-the-art approaches which typically use very deep models like VGG16 pre-trained on ImageNet.

We feel that the GADA approach offers a promising alternative to labo-

	LCC						SROCC					
	JP2K	JPEG	GN	GB	FF	ALL	JP2K	JPEG	GN	GB	FF	ALL
DIVINE [86]	.922	.921	.988	.923	.888	.917	.913	.91	.984	.921	.863	.916
BLINDS-II [106]	.935	.968	.980	.938	.896	.930	.929	.942	.969	.923	.889	.931
BRISQUE [84]	.923	.973	.985	.951	.903	.942	.914	.965	.979	.951	.887	.940
CORNIA [145]	.951	.965	.987	.968	.917	.935	.943	.955	.976	.969	.906	.942
CNN [52]	.953	.981	.984	.953	.933	.953	.952	.977	.978	.962	.908	.956
SOM [149]	.952	.961	.991	.974	.954	.962	.947	.952	.984	.976	.937	.964
BIECON [57]	.965	.987	.970	.945	.931	.962	.952	.974	.980	.956	.923	.961
PQR [147]	-	-	-	-	-	.971	-	-	-	-	-	.965
DNN [13]	-	-	-	-	-	.972	-	-	-	-	-	.960
RankIQA+FT [74]	.975	.986	.994	.988	.960	.982	.970	.978	.991	.988	.954	.981
Hall-IQA [71]	.977	.984	.993	.990	.960	.982	.983	.961	.984	.983	.989	.982
NSSADNN [140]	-	-	-	-	-	.984	-	-	-	-	-	.986
GADA (ours)	.977	.978	.994	.968	.943	.973	.963	.948	.991	.958	.917	.964

Table 5.2: Comparison between GADA and the state-of-the-art on LIVE.

riously annotating images for IQA. Significant improvements can likely be made, especially for highly local distortions, through saliency-based sampling of image patches during training.

Method	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10	#11	#12	#13
BLIINDS-II [106]	0.714	0.728	0.825	0.358	0.852	0.664	0.780	0.852	0.754	0.808	0.862	0.251	0.755
BRISQUE [84]	0.630	0.424	0.727	0.321	0.775	0.669	0.592	0.845	0.553	0.742	0.799	0.301	0.672
CORNIA-10K [145]	0.341	-0.196	0.689	0.184	0.607	-0.014	0.673	0.896	0.787	0.875	0.911	0.310	0.625
HOSA [138]	0.853	0.625	0.782	0.368	0.905	0.775	0.810	0.892	0.870	0.893	0.932	0.747	0.701
RankIQ+FT [74]	0.667	0.620	0.821	0.365	0.760	0.736	0.783	0.809	0.767	0.866	0.878	0.704	0.810
NSSADNN [140]	-	-	-	-	-	-	-	-	-	-	-	-	-
HALLUCINATED IQA [71]	0.923	0.880	0.945	0.673	0.955	0.810	0.855	0.832	0.957	0.914	0.624	0.460	0.782
GADA (ours)	0.932	0.897	0.943	0.825	0.949	0.920	0.919	0.790	0.881	0.775	0.886	0.435	0.702
Method	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	ALL	
BLIINDS-II [106]	0.081	0.371	0.159	-0.082	0.109	0.699	0.222	0.451	0.815	0.568	0.856	0.550	
BRISQUE [84]	0.175	0.184	0.155	0.125	0.032	0.560	0.282	0.680	0.804	0.715	0.800	0.562	
CORNIA-10K [145]	0.161	0.096	0.008	0.423	-0.055	0.259	0.606	0.555	0.592	0.759	0.903	0.651	
HOSA [138]	0.199	0.327	0.233	0.294	0.119	0.782	0.532	0.835	0.855	0.801	0.905	0.728	
RankIQ+FT [74]	0.512	0.622	0.268	0.613	0.662	0.619	0.644	0.800	0.779	0.629	0.859	0.780	
NSSADNN [140]	-	-	-	-	-	-	-	-	-	-	-	0.844	
HALLUCINATED IQA [71]	0.664	0.122	0.182	0.376	0.156	0.850	0.614	0.852	0.911	0.381	0.616	0.879	
GADA (ours)	0.206	0.200	0.196	0.739	0.688	0.950	0.679	0.937	0.895	0.843	0.889	0.790	

Table 5.3: Comparison between GADA and the state-of-the-art on TID2013 (SROCC).

Chapter 6

Language Based Image Quality Assessment

Evaluation of generative models, in the visual domain, is often performed providing anecdotal results to the reader. In the case of image enhancement, reference images are usually available. Nonetheless, using signal based metrics often leads to counter-intuitive results: highly natural crisp images may obtain worse scores than blurry ones. On the other hand, blind reference image assessment may rank images reconstructed with GANs higher than the original undistorted images. To avoid time consuming human based image assessment, semantic computer vision tasks may be exploited instead. In this work we advocate the use of language generation tasks to evaluate the quality of restored images. We show experimentally that image captioning, used as a downstream task, may serve as a method to score image quality. Captioning scores are better aligned with human rankings with respect to signal based metrics or no-reference image quality metrics. We show insights on how the corruption, by artifacts, of local image structure may steer image captions in the wrong direction.¹

¹The work described in this chapter was presented at ACM Multimedia Asia, 2021, where it the Best Paper Award.

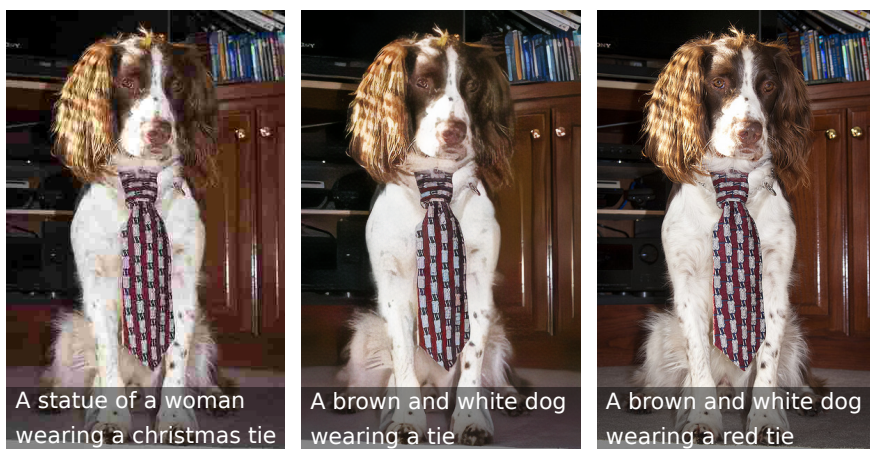


Figure 6.1: Caption generated on Compressed, Reconstructed and Original image (left to right) using [2]. Sample ground truth caption: “A brown and white dog wearing a neck tie”. Best viewed in color on computer screen.

6.1 Introduction

In Chapter 5 we introduce a first work about Image Quality Assessment that addresses the problem of lack of training data with a generative adversarial data augmentation technique. In the current chapter, we propose a completely novel approach for IQA that exploits Image Captioning. The main contribution of this work are the following:

- We propose an image quality assessment method based on language models. To the best of our knowledge, language has never been used to evaluate the quality of images.
- Our evaluation protocol shows consistency across different captioning algorithms [2, 25] and language similarity metrics. Interestingly, improving the language generation model also improves the correlation between our score and MOS.
- Experiments shows that our approach does not suffer from drawbacks of common full-reference and no-reference metrics when evaluating GAN enhanced images and keeps a high accordance with human scores for compressed and for images restored via deep learning.

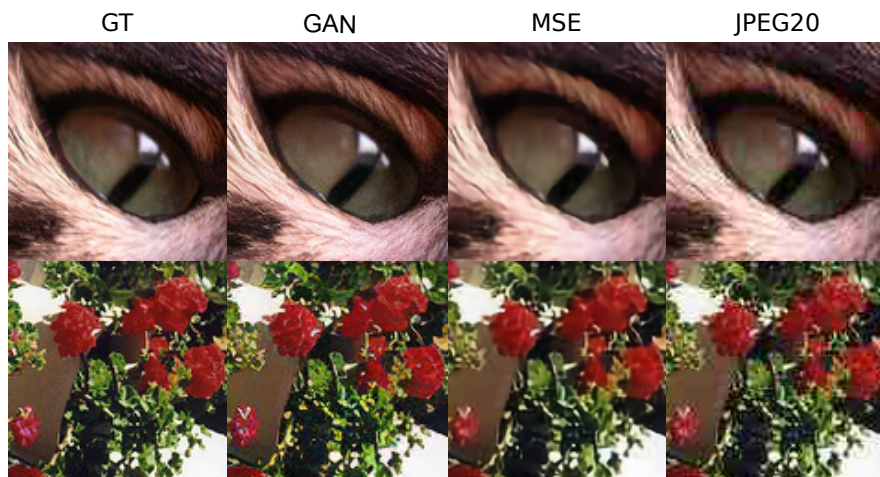


Figure 6.2: Qualitative comparison of reconstruction methods: GAN produces images more pleasant for the human eye. Best viewed in color on computer screen. GT: original image; JPEG 20: JPEG compression with quality factor 20; MSE: CNN-based restoration using MSE loss and direct training; GAN: GAN-based restoration using perceptual loss.

6.2 Image Restoration

Here we formalize the image restoration task. Given some image processing algorithm D , such as JPEG image compression, a distorted image is defined as $I_{LQ} = D(I_{HQ})$, where I_{HQ} is a high quality image undergoing the distortion process, image enhancement aims at finding a restored version of the image $I_R \approx G(I_{LQ})$.

In this work we pick a state-of-the art image enhancement method aimed at compression artifact removal, originally presented in [33]. In this work Galteri *et al.* try to learn a generative model G which, conditioned on the input distorted images, is optimized to invert the distortion process D so that $G \approx D^{-1}$. Their generator architecture is loosely inspired by [42]. They employ LeakyReLU activations and 15 residual layers in a fully convolutional network. The final image is obtained by a nearest neighbor upsampling of a

convolutional feature map and a following stride-one convolutional layer to avoid gridlike patterns possibly stemming from transposed convolutions.

The set of weights ψ of the D network are learned by minimizing:

$$\mathcal{L}_d = -\log(D_\psi(I|I^C)) - \log(1 - D_\psi(I^R|I^C)),$$

where I is the uncompressed or high-quality image, I^R is the restored image created by the generator and I^C is a compressed image.

The generator is trained combining a perceptual loss with the adversarial loss:

$$\mathcal{L}_{AR} = \mathcal{L}_P + \lambda\mathcal{L}_{adv}, \quad (6.1)$$

where \mathcal{L}_{adv} is the standard adversarial loss:

$$\mathcal{L}_{adv} = -\log(D_\psi(I^R|I^C)) \quad (6.2)$$

that rewards solutions that are able to mislead the discriminator, and \mathcal{L}_p is a perceptual loss based on the distance between images computed projecting I and I^R on a feature space by some differentiable function ϕ and taking the Euclidean distance between the two feature representations:

$$\mathcal{L}_P = \frac{1}{W_f H_f} \sum_{x=1}^{W_f} \sum_{y=1}^{H_f} \left(\phi(I)_{x,y} - \phi(I^R)_{x,y} \right)^2. \quad (6.3)$$

In [33] it has been shown that using a GAN approach instead of direct training of the network for image enhancement, results in improved subjective perceptual similarity to original images and, more importantly, in much improved object detection performance. Qualitative examples of GAN and direct training method are shown in Fig. 6.2.

6.3 Evaluation Protocol

Classic full-reference image quality evaluation methods rely on the similarity between an image which has been processed by some enhancement method and a reference undistorted image. GANs are great at filling in high frequency realistic details in image enhancement tasks. Unfortunately this often results in lower performance in full-reference assessment as can be seen in Tab. 6.4, although the restored images appear as “natural” and pleasant to human evaluators. It is clear from such results that while measuring SSIM

and PSNR, optimizing MSE or SSIM losses without adversarial learning is best. For this reason, in [33,34] semantic tasks are used to evaluate the quality of restored images. Measuring the performance of a semantic task such as detection on restored images gives us an understanding of the “correctness” of output images. Given some semantic task (e.g. object detection), a corresponding evaluation metric (e.g. mAP) and a dataset, the evaluation protocol consists in measuring the variation of such metric on different versions of the original image. Interestingly, this evaluation methodology gives hints on what details are better recovered by GANs.

In certain cases, detection is a task describing scene semantics in a very approximate fashion; usually detectors do not degrade for object classes that are clearly identifiable by their shape since even high distortions in the image are not able to hide such features. The gain in image quality provided by GANs, according to object detection based evaluation, resides in producing high quality textures for deformable objects (e.g. cats, dogs, etc).

In this work we advocate the use of a language generation task for evaluating image enhancement at a finer level. The idea is that captioning maps the semantics of images into a much finer and rich label space represented by short sentences. To be able to obtain a correct caption from an image many details must be identifiable.

We devise the following evaluation protocol for image enhancement. We pick an image captioning algorithm \mathcal{A} . Image captioning is the task of generating a sequence of words which is possibly grammatically and semantically correct, describing the image in detail. We look at performance of a captioning algorithm \mathcal{A} on different versions of a dataset (e.g. COCO): compressed, original and restored. In particular we analyze results from two highly performing captioning methods [2, 25] which combine a bottom-up model of visual entities and their attributes in the scene with a language decoding pipeline. Both methods are trained over several steps incorporating semantic knowledge at different levels of granularity. In particular the bottom-up region generator is based on Faster R-CNN [100] which is based on a feature extractor pre-trained on ImageNet [26] and then fine-tuned to predict object entities and their attributes using the Visual Genome dataset [61]. In [2], further knowledge is incorporated into the model by training the caption generation model using a first LSTM as a top-down visual attention model and a second level LSTM as a language model. Meshed memory transformers [25] share the exact same visual backbone as [2] but exploit a stack of

memory-augmented visual encoding layers and a stack of decoding layers to generate caption tokens.

No matter how captioning models are optimized, our results show that the behavior of the captioning model for image quality assessment is consistent over several metrics as shown in Tab. 6.1.

Captioning is evaluated with several specialized metrics measuring the word-by-word overlap between a generated sentence and the ground truth [89], in certain cases including the ordering of words [5], considering n-grams and not just words [70, 129] and the semantic propositional content (SPICE [1]). These metrics evaluate the similarity with respect to a set of reference captions (usually this is five references).

6.3.1 Subjective Evaluation

In this evaluation we assess how images obtained with the selected GAN based restoration method [33] are perceived by a human viewer, evaluating in particular the preservation of details and overall quality of an image. In total, 16 viewers have participated to the test, a number that is considered enough for subjective image quality evaluation tests [137]; no viewer was familiar with image quality evaluation or the approaches proposed in this work. A Single-Stimulus Absolute Category Rating (ACR) experimental setup has been developed using avrateNG², a tool designed to perform subjective image and video quality evaluations. We asked participants to evaluate images' quality using the standard 5-values ACR scale (1=bad, up to 5=excellent). A set of 20 images is chosen from the COCO dataset, selecting for each image three versions: the original image, a JPEG compressed version with QF=10 (a high compression quality factor) and the restored version of the JPEG compressed image with QF=10 compressed image; this results in a set of 60 images. Each image was shown for 5 seconds, preceded and followed by a grey image, also shown for 5 seconds. Considering our estimation of test completion time we chose this amount of images to keep each session under 30 minutes as recommended by ITU-R BT.500-13 [49].

To select this small sample of 20 images to be as representative as possible of the whole dataset for the captioning performance we operate the following procedure. Let $\mu^*(v)$ and $\sigma^{2*}(v)$ be the mean of a captioning metric score (in our case we used CIDEr) for a given version of the image v . We iteratively

²<https://github.com/Telecommunication-Telemedia-Assessment/avrateNG>

extract 20 random image ids out of the whole 5,000 testing set from the Karpathy split, without repetition. We attempt to minimize

$$e_\mu = \sum_{v \in \mathcal{V}} |\mu^*(v) - \bar{\mu}(v)| \quad (6.4)$$

and

$$e_{\sigma^2} = \sum_{v \in \mathcal{V}} |\sigma^{2*}(v) - \bar{\sigma}^2(v)| \quad (6.5)$$

by iterative resampling images until we find e_μ and e_{σ^2} such that $e_\mu \leq 10^{-3}$ and $e_{\sigma^2} \leq 10^{-4}$. Where \mathcal{V} is the set of different version of an image, namely: JPEG compressed at QF=10 (referred to as JPEG 10 in the following), its GAN reconstruction and the original uncompressed image. The selected images contain different subjects, such as persons, animals, man-made objects, nature scenes, etc. Both the order of presentation of the tests for each viewer, and the order of appearance of the images were randomized.

6.4 Results

In the following, we report results on two datasets: MS-COCO [72] and LIVE [110]. We use COCO, in particular the Karpathy split, since it is the reference benchmark for image captioning, accounting for 5000 images for training and validation each with 5 ground truth sentences per image. LIVE is a widespread benchmark for image quality assessment. LIVE consists of 29 high resolution images compressed at different JPEG qualities for a total of 204 images. For each LIVE image a set of user scores is provided indicating the perceived quality of the image.

6.4.1 Language Based IQA

In Tab. 6.1 we report results using various captioning metrics. Interestingly all metrics show that captions over reconstructed images (REC rows) are better with respect to caption computed over compressed images (JPEG rows). This shows that image details that are compromised by the strong compression induce errors in the captioning algorithm. On the other hand the GAN approach is able to recover an image which is not only pleasant to the human eye but recovers details which are also semantically relevant to an algorithm. In Fig. 6.1 we show the difference of captions generated by [2] over original, compressed and restored images. A human may likely

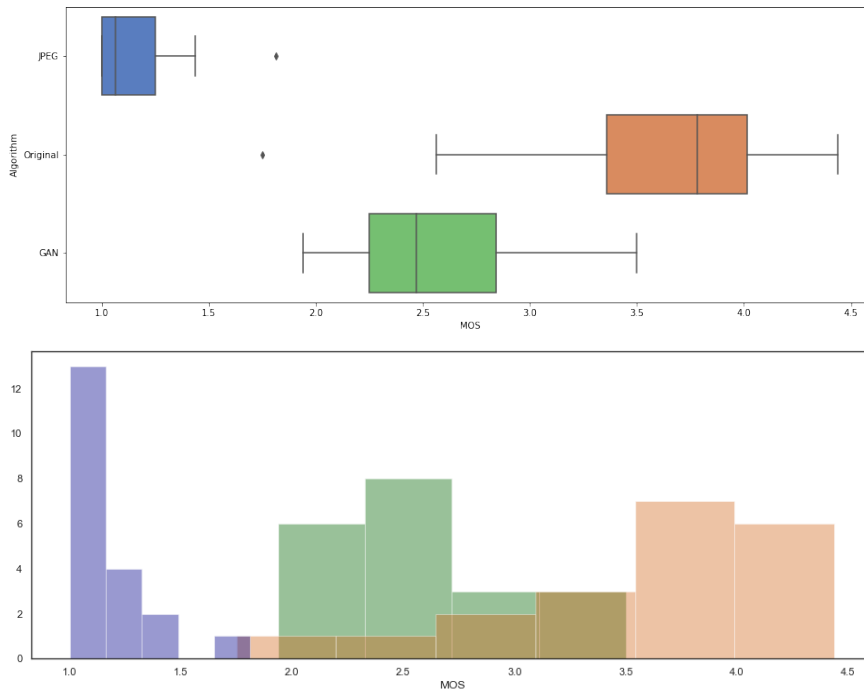


Figure 6.3: *Top*) Subjective image quality evaluation of original COCO images (orange), heavily compressed JPEG images (blue) and their restored version obtained with the GAN-based approach (green). Restored images are perceived as having a better quality than their compressed versions. *Bottom*) Histograms of MOS scores of the three types of images.

succeed in producing a almost correct caption for highly compressed images, nonetheless state-of-the art algorithms are likely to make extreme mistakes which are instead not present on reconstructed images.

In Fig. 6.5 we show the different performance of captioning algorithms in terms of CIDEr measure on the same split of test of compressed and restored images, considering different quality factors of JPEG. The captioner proposed in [25] outperforms [2] as expected, but interestingly we may observe that the range of CIDEr values of [25] is significantly higher than [2]. We argue that this could be considered a strong feature of our evaluation approach, as a wider range of value may imply that a good captioner is able

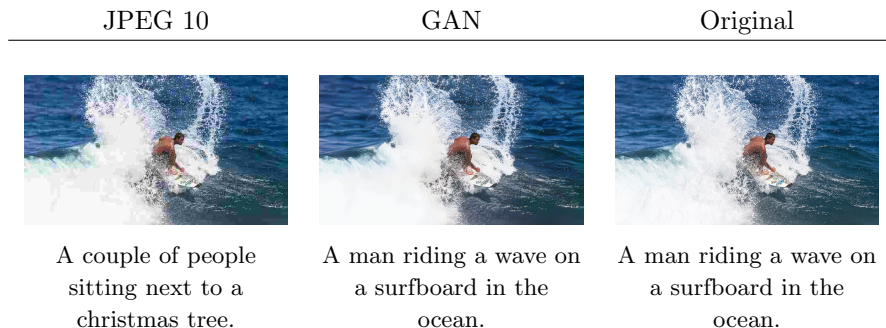


Figure 6.4: Examples of captions for COCO images used in the subjective quality evaluation. Left column) JPEG compressed with QF=10; Center column) GAN-based restoration from JPEG compressed images with QF=10; right column) original images.

Table 6.1: Evaluation of image restoration over compression artifacts using GAN and captioning as a semantic task (best results highlighted in bold). Captions created from reconstructed images obtain a better score for every metric.

QUALITY	BLEU_1 \uparrow	METEOR \uparrow	ROUGE \uparrow	CIDEr \uparrow	SPICE \uparrow
JPEG 10	0.589	0.173	0.427	0.496	0.103
REC 10	0.730	0.253	0.527	1.032	0.189
JPEG 20	0.709	0.241	0.513	0.937	0.174
REC 20	0.751	0.266	0.543	1.105	0.201
JPEG 30	0.740	0.258	0.535	1.054	0.194
REC 30	0.757	0.269	0.549	1.133	0.205
JPEG 40	0.748	0.263	0.542	1.087	0.200
REC 40	0.758	0.270	0.549	1.132	0.206
JPEG 60	0.755	0.267	0.546	1.117	0.204
REC 60	0.760	0.270	0.550	1.137	0.207
ORIGINAL	0.766	0.274	0.556	1.166	0.211

to predict the image quality in a finer manner than other weaker captioning algorithms.

Fig. 6.6 shows the bottom-up captioning process performed on an image used in the subjective evaluation. The left image shows the JPEG 10 version, while the right one shows the GAN reconstruction. The images show

the bounding boxes of the detected elements. In the first case the wrong detections of indoor elements like “floor” and “wall” are likely reasons for the wrong caption, as opposed to the correct recognition of a “white wave” and “blue water” in the GAN-reconstructed image.

In order to understand better what metric could be used instead of human evaluation we computed the correlation coefficient ρ between BRISQUE [83], NIQE [85], CIDEr and MOS for all versions of the images. As shown in Tab. 6.2, it turns out that using a fine-grained semantic task as image captioning is the best proxy (highest correlation) of real human judgment.

Fig. 6.4 show a captioning example from the COCO images used in the subjective quality evaluation experiment. On the left we show a sample compressed with JPEG with a QF=10, on the center we show the image restored with [33] and on the right we show the original one. It can be observed that the caption of the restored image is capable of describing correctly the image content, on par with the caption obtained on the original image. Instead, the caption of the highly compressed JPEG image is completely unrelated to image content, probably due to object detection errors.

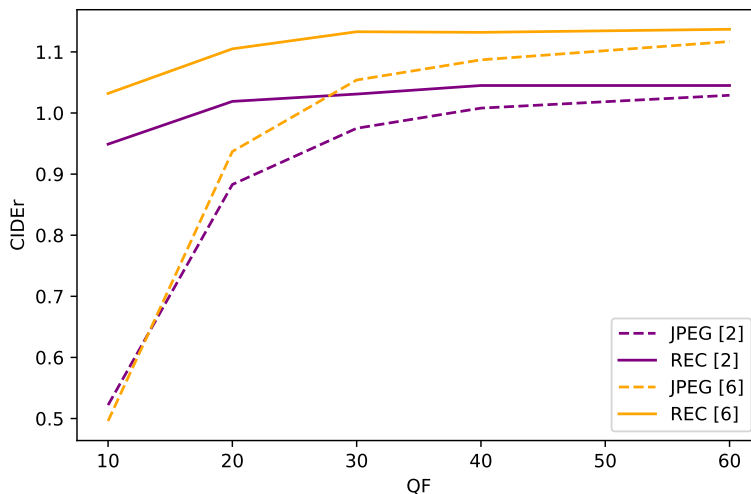


Figure 6.5: CIDEr scores using [2] and [25] on compressed and restored images for different QFs from MS-COCO.

Table 6.2: Correlation coefficient between no-reference and captioning based metrics and MOS on COCO.

Metric	ρ
NIQE	0.84
BRISQUE	0.89
CIDEr	0.96

6.4.2 Comparison with MOS

In Fig. 6.3 *top*) are reported subjective evaluation results as MOS (Mean Opinion Scores) as box plots, showing the quartiles of the scores (box), while the whiskers show the rest of the distribution. The plots are made for the original images, the images compressed with JPEG using a QF=10, and the images restored with the GAN-based approach [33] from the heavily compressed JPEG images. The figure shows that the GAN-based network is able to produce images that are perceptually of much higher quality than the images from which they are originated; the average MOS score for JPEG images is 1.15, for the GAN-based approach is 2.56 and for the original images it is 3.59. The relatively low MOS scores obtained also by the original images are related to the fact that COCO images have a visual quality that is much lower than that of dataset designed for image quality evaluation. To give better insight on the distribution of MOS scores, Fig. 6.3 *bottom*) shows the histograms of the MOS scores for the three types of images: orange histogram for the original images, green for the JPEG compressed images and blue for the restored images.

We further show that our language based approach correlates with perceived quality using a IQA benchmark test on the LIVE dataset, which contains the opinion scores for each image. However, no caption is provided in this dataset. For this reason, we consider the output sentences of captioning approaches over the undistorted image as the ground truth in order to calculate the language similarity measures. In Tab. 6.3 we show the Pearson correlation score of different captioning metrics and other common full-reference quality assessment approaches. The experiment shows an interesting behaviour of our approach in terms of correlation. In the first place, we can observe that each captioning metric has a correlation index that is higher or at least comparable with the other full-reference metrics. In particular, METEOR and CIDEr perform better than the other metrics independently of which captioning algorithm is used. Moreover, we observe

Table 6.3: Pearson score, correlating scores with users’ MOS for different captioning metrics and image based full-reference approaches on LIVE. CIDEr obtains a superior score with respect to image based methods.

Metric	Ours w/ [25]	Ours w/ [2]
BLEU 1	0.873	0.838
METEOR	0.900	0.846
SPICE	0.895	0.844
ROUGE	0.861	0.832
CIDEr	0.901	0.854
PSNR		0.857
SSIM		0.893
LPIPS		0.859

that the correlation metric significantly improves if we employ a more performing captioner. In this particular case, the visual features used by the two captioning techniques are exactly the same, the main difference lies in the overall language generation pipeline of the approaches. Hence, we argue that language is effectively useful for quality assessment, and the more a captioning algorithm is capable to provide detailed and meaningful captions the better we could use the generated sentences to formulate good predictions about the quality of images.

6.4.3 Comparison with Full-Reference Metrics

A common setting that is used to evaluate image enhancement algorithms is full reference image quality assessment, where several image similarity metrics are used to measure how much a restored version differs with respect to the uncorrupted original image. This kind of metrics, measuring pixel-wise value differences are likely to favor MSE optimized networks which are usually prone to obtain blurry and lowly detailed images. In Tab. 6.4 we report results on COCO for full-reference indexes. In this setup, we compress the original images at different quality factors and then we restore them with a QF specific artifact removal GAN. We use the uncompressed image generated caption as GT, as in Tab. 6.3. The results show that, for restored images, PSNR accounts for a slight improvement while SSIM indexes lower than the compressed counterparts. This is an expected outcome, as in [33] it is shown that state of the art results on PSNR can be obtained only when MSE is optimized and on SSIM if the metric is optimized directly.

Table 6.4: Evaluation using no-reference and full-reference metrics on MS-COCO. NIQE and BRISQUE rate better GAN images than the ORIGINAL. SSIM always rate restored images worse than compressed. PSNR shows negligible improvement.

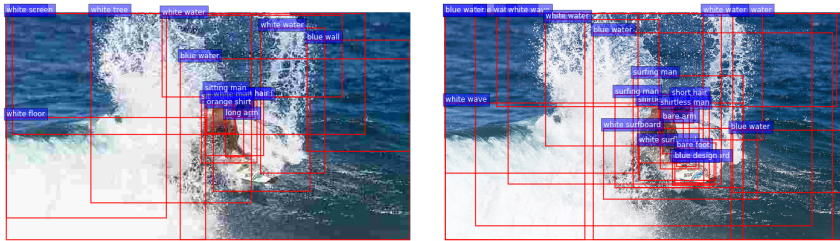
QUALITY	NIQE↓	BRISQUE↓	Ours w/ [25] ↑	PSNR ↑	SSIM↑	LPIPS↓
JPEG 10	6.689	52.67	0.542	25.45	0.721	0.305
GAN 10	3.488	17.93	1.118	25.70	0.718	0.144
JPEG 20	5.183	43.99	0.956	27.46	0.796	0.187
GAN 20	3.884	17.85	1.289	27.60	0.784	0.085
JPEG 30	4.474	37.72	1.165	28.61	0.831	0.134
GAN 30	3.601	18.32	1.370	28.81	0.819	0.060
JPEG 40	4.011	33.61	1.260	29.41	0.852	0.105
GAN 40	3.680	18.68	1.424	29.44	0.836	0.048
JPEG 60	3.588	28.15	1.366	30.71	0.880	0.067
GAN 60	3.885	19.45	1.482	30.61	0.862	0.032
ORIGINAL	3.656	21.79	-	-	-	-

Nonetheless, as can be seen in Fig. 6.2, GAN enhanced images are more pleasant to the human eye, therefore we should not rely just on PSNR and SSIM for GAN restored images. Our approach, using [25], is in line with LPIPS [151]. Unfortunately, LPIPS, as shown in Tab. 6.3 has low correlation with scores determined by human perceived quality.

6.4.4 Comparison with No-Reference Metrics

In certain cases it is not possible to use full reference metrics quality metrics, e.g. if there’s no available original image. These kind of metrics typically evaluate the “naturalness” of the image being analyzed. In the same setup we used previously, we perform experiments using NIQE and BRISQUE which are two popular no-reference metrics for images. We report in Tab. 6.4 the results.

Interestingly, these metrics tend to favor GAN restored images instead of the original uncompressed ones. Most surprisingly, NIQE and BRISQUE obtain better results when we reconstruct the most degraded version of images (QF 10-20), but these values increase as we reconstruct less degraded images. We believe that BRISQUE and NIQE favor crisper images with high frequency patterns which are distinctive of GAN based image enhancement.



A couple of people sitting next to a Christmas tree. A man riding a wave on a surfboard in the ocean.

Figure 6.6: Bottom-Up detection process of captioning on two images: left) JPEG compressed; right) GAN reconstruction. Note that several mistaken detections on the left image are avoided in the right one. In particular on the left “surfboard” is missed and “white floor” and “blue wall” are wrongly detected. This two indoor details are the one that likely mislead the captioning.

6.5 Conclusion

In this work we propose a new idea to evaluate image enhancement methods. Existing metrics based on the comparison of the restored image with an undistorted version may give counter-intuitive results. On the other hand the use of naturalness based scores may in certain cases rank restored images higher than original ones.

We have shown that instead of using signal based metrics, semantic computer vision tasks can be used to evaluate results of image enhancement methods. Our claim is that a fine grained semantic computer vision task can be a great proxy for human level image judgement.

We show that employing algorithms mapping input images to a finer output label space, such as captioning, leads to more discriminative metrics. Future work will regard the evaluation of captions provided by humans over compressed and restored images. Moreover, we will take into account the accuracy of captions as a further metric to optimize.

Chapter 7

LANBIQUE: LANguage-based Blind Image QUality Evaluation

Image quality assessment is often performed with deep networks which are fine-tuned to regress a human provided quality score of a given image. Usually, these approaches may lack generalization capabilities and, while being highly precise on similar image distribution, it may yield lower correlation on unseen distortions. In particular they show poor performances whereas images corrupted by noise, blur or compressed have been restored by generative models. As a matter of fact, evaluation of these generative models is often performed providing anecdotal results to the reader. In the case of image enhancement and restoration, reference images are usually available. Nonetheless, using signal based metrics often leads to counterintuitive results: highly natural crisp images may obtain worse scores than blurry ones. On the other hand, blind reference image assessment may rank images reconstructed with GANs higher than the original undistorted images. To avoid time consuming human based image assessment, semantic computer vision tasks may be exploited instead. In this chapter we advocate the use of language generation tasks to evaluate the quality of restored images. We refer to our assessment approach as LANguage-based Blind Image QUality Evaluation (LANBIQUE). We show experimentally that image captioning, used as a downstream task, may serve as a method to

*score image quality, independently of the distortion process that affects the data. Captioning scores are better aligned with human rankings with respect to classic signal based or No-Reference image quality metrics. We show insights on how the corruption, by artifacts, of local image structure may steer image captions in the wrong direction.*¹

7.1 Introduction

In Chapter 6 we introduce a novel approach about Image Quality Assessment based on Image Captioning. This method works in a Full-Reference Scenario where high quality reference image is available to evaluate the quality of its corresponding distorted version. In real scenarios reference image is not available. In this work, we focus on this aspect introducing a new method. We refer to the new approach as LANguage-based Blind Image QUality Evaluation (LANBIQUE). Fig. 7.2 shows the gist of the proposed approach: the effects of image compression lead to a wrong captioning of the image on the left with respect to the original high quality image on the right; captioning an image that has been obtained enhancing the compressed image with a GAN-based approach (center) leads to a caption that is very similar to the caption of the high quality image. The main contributions of our work are the following:

- LANBIQUE show consistency across different captioning algorithms [2, 25] and language similarity metrics. Interestingly, improving the language generation model also improves the correlation between our score and MOS.
- Experiments shows that LANBIQUE does not suffer from drawbacks of common Full-Reference and No-Reference metrics when evaluating GAN enhanced images and keeps a high accordance with human scores for compressed and for images restored via deep learning.

In this extended version, we propose the following improvement with respect to the work described in Chapter 6.

- We show that LANBIQUE can be used also for distortions different from JPEG compression.

¹The work described in this chapter was published in ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2022.



Figure 7.1: Qualitative comparison of reconstruction methods: GAN produces images more pleasant for the human eye. Best viewed in color and zoomed on computer screen. GAN: GAN-based restoration using perceptual loss. MSE: CNN-based restoration using MSE loss; JPEG 20: JPEG compression with quality factor 20.

- We tested LANBIQUE on the larger and more diverse PieAPP dataset, showing strong results against learning and non-learning based methods.
- Finally, the basic version of LANBIQUE is extended in order to make it possible to work also without a reference image. To get to this goal we employ a blind restoration GAN, which can restore images without the knowledge nor the intensity of the distortion, to recover a pseudo-reference image.

The rest of this chapter is organized as follows: in Section 2 we describe the related works. In Section 3 we briefly discuss about prior GAN-based image restoration approaches. In Section 4 we describe LANBIQUE in detail. In Section 5 we show experimental results of LANBIQUE on different settings and datasets. Finally, in Section 6 we draw the conclusions about our approach.

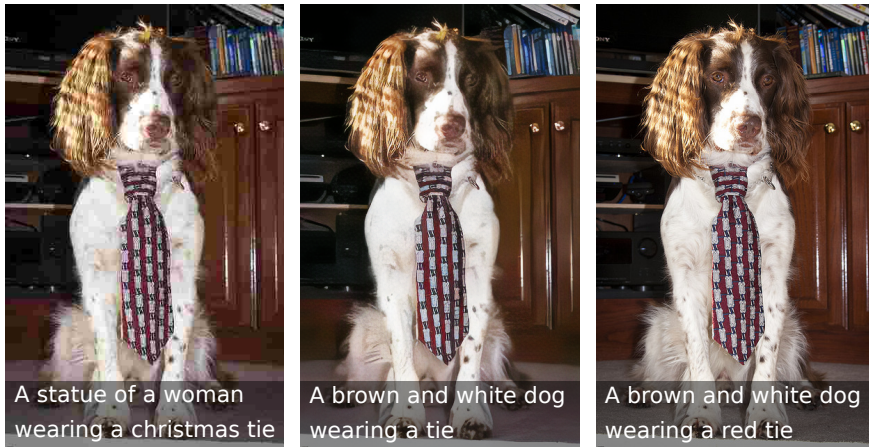


Figure 7.2: Caption generated on Compressed, Reconstructed and Original image (left to right) using [2]. Sample ground truth caption: “A brown and white dog wearing a neck tie”. Best viewed in color on computer screen.

7.2 Image Restoration

Even if this work does not propose novel image restoration approaches, to make the chapter self-contained here we formalize the image restoration or enhancement task. The main motivation that lead us to work on an alternative to image quality assessment is the poor performance of standard IQA methods on images that have been enhanced by GANs, e.g. for denoising [55,125], deblurring [124,148] or compression artefact removal [33,80,126]. Furthermore, we leverage image restoration as a tool to extend the capabilities of LANBIQUE in order to evaluate those images that lack an uncorrupted high quality counterpart, extending our approach to the No-Reference scenario, as show in Sect.7.3.3.

Problem formulation. Given some image processing algorithm D , such as JPEG image compression, a distorted image is defined as $I_{LQ} = D(I_{HQ})$, where I_{HQ} is a high quality image undergoing the distortion process, image enhancement aims at finding a restored version of the image $I_R \approx G(I_{LQ})$. In this work we use two image enhancement networks, one that is specific for JPEG artifacts [33], and a more generic approach, which can work without prior knowledge of the degradation [133].

In [33] Galteri *et al.* try to learn a generative model G which, conditioned on the input distorted images, is optimized to invert the distortion process D so that $G \approx D^{-1}$. Their generator architecture is loosely inspired by [42]. They employ LeakyReLU activations and 15 residual layers in a fully convolutional network. The final image is obtained by a nearest neighbor upsampling of a convolutional feature map and a following stride-one convolutional layer to avoid grid-like patterns possibly stemming from transposed convolutions.

The set of weights ψ of the D network are learned by minimizing:

$$\mathcal{L}_d = -\log(D_\psi(I_{HQ}|I_{LQ})) - \log(1 - D_\psi(I_R|I_{LQ})),$$

where I_{HQ} is the uncompressed or high-quality image, I_R is the restored image created by the generator and I_{LQ} is a compressed image.

The generator is trained combining a perceptual loss with the adversarial loss:

$$\mathcal{L}_{AR} = \mathcal{L}_P + \lambda\mathcal{L}_{adv}, \quad (7.1)$$

where \mathcal{L}_{adv} is the standard adversarial loss:

$$\mathcal{L}_{adv} = -\log(D_\psi(I_R|I_{LQ})) \quad (7.2)$$

that rewards solutions that are able to mislead the discriminator, and \mathcal{L}_p is a perceptual loss based on the distance between images computed projecting I_{HQ} and I_R on a feature space by some differentiable function ϕ and taking the Euclidean distance between the two feature representations:

$$\mathcal{L}_P = \frac{1}{W_f H_f} \sum_{x=1}^{W_f} \sum_{y=1}^{H_f} \left(\phi(I_{HQ})_{x,y} - \phi(I_R)_{x,y} \right)^2. \quad (7.3)$$

They employ a generator inspired by [42], with a residual architecture using LeakyReLU activations, Batch-Normalization [47] and Nearest-neighbour upsampling layer is used to recover original size [87], and a fully convolutional Discriminator. In [33] it has been shown that using a GAN approach instead of direct training of the network for image enhancement, results in improved subjective perceptual similarity to original images and, more importantly, in much improved object detection performance. Qualitative examples of GAN and direct training method are shown in Fig. 7.1.

Real-ESRGAN [133] is a more recent approach, that has the advantage of not requiring to know the type of distortion nor the intensity of it in advance

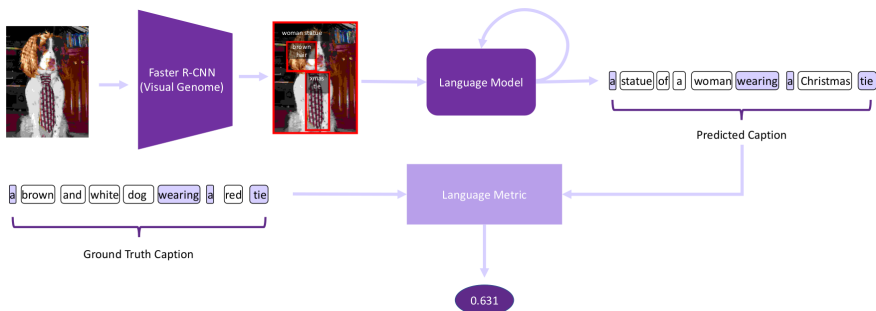


Figure 7.3: Overview of LANBIQUE. An image is first processed by an object detector, each box feature is then fed to a captioning model [2, 25]; then a metric for captioning evaluation is used to score the quality of the image. In this example a highly JPEG corrupted images yields a *low* CIDEr score of 0.631.

to restore an image. In [133] Wang *et al.* introduce a high-order degradation modeling process to better simulate complex real-world degradations. Differently from [33] they use a U-Net discriminator with spectral normalization to increase discriminator capability and stabilize the training dynamics. As in ESRGAN [134] the generator is built by several residual-in-residual dense blocks (RRDB).

7.3 Evaluation Protocol

Classic Full-Reference image quality evaluation methods rely on the similarity between an image which has been processed by some algorithm D and a reference undistorted image. Considering the use case of image enhancement of an image that was compressed, GANs are a good solution since they are great at filling in high frequency realistic details in image enhancement tasks; in this case the resulting enhanced image is compared to the reference. Unfortunately, when using classical MSE based Full-Reference metrics such as SSIM and PSNR GAN restored images yield lower performance as can be seen in Tab. 7.2, although they appear as “natural” and pleasant to human evaluators, as also seen in examples of Fig.7.1. For this reason, in [33, 34] semantic tasks are used to evaluate the quality of restored images. Measuring the performance of a semantic task such as detection on

restored images gives us an understanding of the “correctness” of output images. Given some semantic task (e.g. object detection), a corresponding evaluation metric (e.g. mAP) and a dataset, the evaluation protocol consists in measuring the variation of such metric on different versions of the original image. Interestingly, this evaluation methodology gives hints on what details are better recovered by GANs.

In certain cases, detection is a task describing scene semantics in a very approximate fashion; usually detectors do not degrade for object classes that are clearly identifiable by their shape since even high distortions in the image are not able to hide such features. The gain in image quality provided by GANs, according to object detection based evaluation, resides in producing high quality textures for deformable objects (e.g. cats, dogs, etc).

In this chapter we advocate the use of a language generation task for evaluating image enhancement. The idea is that captioning maps the semantics of images into a much finer and rich label space represented by short sentences. To be able to obtain a correct caption from an image many details must be identifiable.

7.3.1 Evaluation with Reference Captions

We devise the following evaluation protocol for image enhancement. We pick an image captioning algorithm \mathcal{A} . Image captioning is the task of generating a sequence of words, possibly grammatically and semantically correct, describing the image in detail. Given a set of reference captions S and the caption generated from an input image $\mathcal{A}(I)$, we want to measure their similarity with a language metric \mathcal{D} :

$$\text{LANBIQUE}(\mathcal{D}, \mathcal{A}; I, S) = \mathcal{D}(\mathcal{A}(I), S) \quad (7.4)$$

We look at the performance of a captioning algorithm \mathcal{A} on different versions of a dataset (e.g. COCO): compressed, original and restored. The pipeline of this evaluation approach is depicted in Fig. 7.3.

In particular, we analyze results from two highly performing captioning methods [2, 25] which combine a bottom-up model of visual entities and their attributes in the scene with a language decoding pipeline. Both methods are trained over several steps incorporating semantic knowledge at different levels of granularity. In particular, the bottom-up region generator is based on Faster R-CNN [100] which is based on a feature extractor pre-trained on ImageNet [26] and then fine-tuned to predict object entities and their

attributes using the Visual Genome dataset [61]. In [2], further knowledge is incorporated into the model by training the caption generation model using a first LSTM as a top-down visual attention model and a second level LSTM as a language model. Meshed memory transformers [25] share the exact same visual backbone as [2] but exploit a stack of memory-augmented visual encoding layers and a stack of decoding layers to generate caption tokens.

No matter how captioning models are optimized, our results show that the behavior of the captioning model for image quality assessment is consistent over several metrics as shown in Tab. 7.1.

Captioning is evaluated with several specialized metrics measuring the word-by-word overlap between a generated sentence and the ground truth [89], in certain cases including the ordering of words [5], considering n-grams and not just words [70, 129] and the semantic propositional content (SPICE [1]). These metrics evaluate the similarity with respect to a set of reference captions S , that is usually composed of five references.

7.3.2 Evaluation without Reference Captions

Unfortunately, in most of the cases reference captions are not available as they often must be collected with great expense of effort and resources; in fact, standard datasets used for image quality evaluation do not include captions. However, it is possible to evaluate any kind of test image with our language based approach by modifying the pipeline. The idea is that the reference image is enough high quality to provide a valid caption for the evaluation of LANBIQUE. We caption the reference image I_{HQ} using the same captioner \mathcal{A} we use for the test image I , then we maintain the same procedure we previously described:

$$\text{LANBIQUE-NC}(\mathcal{D}, \mathcal{A}; I, I_{HQ}) = \mathcal{D}(\mathcal{A}(I), \mathcal{A}(I_{HQ})). \quad (7.5)$$

This evaluation approach is represented in Fig. 7.4. Since we change the evaluation pipeline with respect to the previous case, we argue that there may be a drawback with respect to the original version of the approach. As a matter of fact, modern captioners provide just one description per image and this means that the computation of \mathcal{D} metric is done just between two sentences. However, this does not affect the performance of our approach significantly, provided that the \mathcal{A} generates high quality captions.

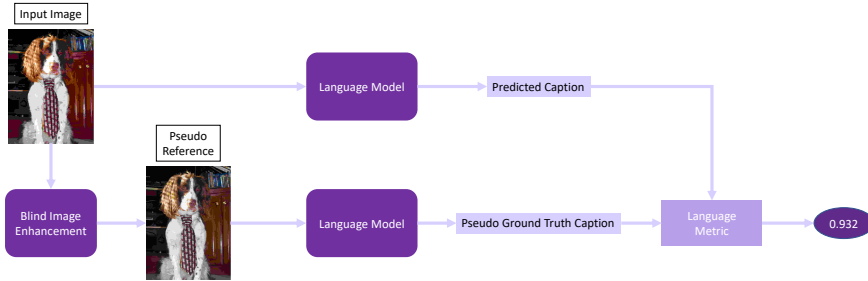


Figure 7.4: LANBIQUE without a reference caption available. The reference image is captioned as well by the same language model to generate a description of the image. This output is used as pseudo ground truth caption and compared to the predicted caption.

7.3.3 No-Reference Evaluation

In this section we show how our approach can be extended to work in a No-Reference setting. In many occasions we may not have a high quality image available to be compared with the one to be tested. For this reason, we modify our language based pipeline by adding an additional blind restoration module \mathcal{R} . We assume that the images to be tested are corrupted by one or a combination of unknown distortions that are responsible of a global reduction of the visual quality. In this extended model, our aim is to restore corrupted input image I in order to use the enhanced version as the reference image. After this operation is completed, we are able to feed both the corrupted image and the restored one to the same captioning module, hence we generate their text descriptions and finally we calculate the ultimate score based on some language metric \mathcal{D} :

$$\text{LANBIQUE-NR}(\mathcal{D}, \mathcal{A}, \mathcal{R}; I) = \mathcal{D}(\mathcal{A}(I), \mathcal{A}(\mathcal{R}(I))). \quad (7.6)$$

This No-Reference approach is depicted in Fig.7.5

Typically, image distortions are not known a priori so it may be a difficult task to train many networks capable of handling all the possible combinations of corruption processes and then select the best one for a specific restoration. For this reason, we choose to train a single network following a degradation model, so that it can restore most types of distorted images and recover their original quality as best as possible. In order to ensure a good output quality, we employed Real-ESRGAN [133] as the restoration module. We have

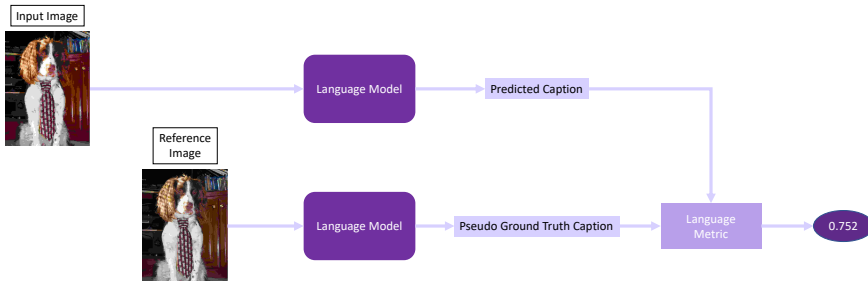


Figure 7.5: LANBIQUE in the No-Reference setting. A blind image enhancement method is used to recover a high quality version of the image, then a captioning model is applied to both images. Input image predicted caption is then compared with the pseudo ground truth caption obtained from the restored image.

modified the original model by adding JPEG2000 in the training procedure, then we have fine-tuned a pre-trained version of such network with the new introduced distortion.

In most of the cases, recovered images represent a solid reference for our evaluation model, as they are very close to real images from the point of view of human perception. In this setup, our LANBIQUE-NR assigns high scores to slightly distorted images, as their reconstruction is likely very perceptually close, and the captions generated are pretty close. On the other hand, highly distorted images are transformed into better quality data that differ significantly from input. In this case, the captions between the two versions may differ much more, thus leading to lower scores of language metrics.

7.3.4 Subjective Evaluation

In this evaluation we assess how images obtained with the selected GAN based restoration method [33] are perceived by a human viewer, evaluating in particular the preservation of details and overall quality of an image. In total, 16 viewers have participated to the test, a number that is considered enough for subjective image quality evaluation tests [137]; no viewer was familiar with image quality evaluation or the approaches proposed in this work. A Single-Stimulus Absolute Category Rating (ACR) experimental setup has

been developed using avrateNG², a tool designed to perform subjective image and video quality evaluations. We asked participants to evaluate images' quality using the standard 5-values ACR scale (1=bad, up to 5=excellent). A set of 20 images is chosen from the COCO dataset, selecting for each image three versions: the original image, a JPEG compressed version with QF=10 (high compression quality factor) and the restored version of the JPEG compressed image with QF=10 compressed image; this results in a set of 60 images. Each image was shown for 5 seconds, preceded and followed by a grey image, also shown for 5 seconds. Considering our estimation of test completion time, we chose this amount of images to keep each session under 30 minutes as recommended by ITU-R BT.500-13 [49].

To select this small sample of 20 images to be as representative as possible of the whole dataset \mathcal{D} for the captioning performance we operate the following procedure. Let $\mu^*(v)$ and $\sigma^{2*}(v)$ be the mean and variance of a captioning metric score (in this case we used CIDEr) for a given version v of the image i . We iteratively extract 20 random image ids, yielding set \mathcal{D}^* out of the whole 5,000 testing set from the Karpathy split, without repetition. We attempt to minimize:

$$e_\mu = \frac{1}{|\mathcal{D}^*|} \sum_{i \in \mathcal{D}^*} \sum_{v \in \mathcal{V}_i} |\mu^*(v) - \bar{\mu}| \quad (7.7)$$

and

$$e_{\sigma^2} = \frac{1}{|\mathcal{D}^*|} \sum_{i \in \mathcal{D}^*} \sum_{v \in \mathcal{V}_i} |\sigma^{2*}(v) - \bar{\sigma}^2| \quad (7.8)$$

by iterative resampling images until we find e_μ and e_{σ^2} such that $e_\mu \leq 10^{-3}$ and $e_{\sigma^2} \leq 10^{-4}$. \mathcal{V}_i is the set of different versions of an image i in the smaller dataset \mathcal{D}^* , namely: JPEG compressed at QF=10 (referred to as JPEG 10 in the following), its GAN reconstruction and the original uncompressed image; and $\bar{\mu}$ and $\bar{\sigma}^2$ are the mean and variance of the considered captioning metric computed on the whole set of images \mathcal{D} . The selected images contain different subjects, such as people, animals, man-made objects, nature scenes, etc. Both the order of presentation of the tests for each viewer, and the order of appearance of the images were randomized.

²<https://github.com/Telecommunication-Telemedia-Assessment/avrateNG>

7.4 Experimental Results

7.4.1 Results on JPEG Artefacts

First, we study in detail the behavior of LANBIQUE on a single distortion. This way we can easily control the amount of image corruption and evaluate the behavior of our metric on GAN restored images.

Results with reference captions. In order to use a dataset of images with a set of associated captions, we selected the 5,000 images testing set from the Karpathy split of COCO dataset [19]. The images have then been compressed at different JPEG Quality Factors (QF), and then they have been reconstructed using the GAN approach of [33]. In Tab. 7.1 we report results of LANBIQUE using various captioning metrics \mathcal{D} . Interestingly, all metrics show that captions over reconstructed images (REC rows) are better with respect to caption computed over compressed images (JPEG rows). This shows that image details that are compromised by the strong compression induce errors in the captioning algorithm. On the other hand, the GAN approach is able to recover an image which is not only pleasant to the human eye but recovers details which are also relevant to a semantic algorithm. In Fig. 7.2 we show the difference of captions generated by [2] over original, compressed and restored images. A human may likely succeed in producing an almost correct caption for highly compressed images, nonetheless state-of-the-art algorithms are likely to make extreme mistakes which are instead not present on reconstructed images.

In Fig. 7.6 we show the different performance of captioning algorithms in terms of CIDEr measure on the same split of test of compressed and restored images, considering different quality factors of JPEG. The captioner proposed in [25] outperforms [2] as expected, but interestingly we may observe that the range of CIDEr values of [25] is significantly higher than [2]. We argue that this could be considered a strong feature of our evaluation approach, as a wider range of value may imply that a good captioner is able to predict the image quality in a finer manner than other weaker captioning algorithms.

Fig. 7.7 shows the bottom-up captioning process performed on an image used in the subjective evaluation. The left image shows the JPEG 10 version, while the right one shows the GAN reconstruction. The images show the bounding boxes of the detected elements. In the first case the wrong

Table 7.1: Evaluation of image restoration over compression artifacts with GAN using LANBIQUE with different captioning metrics (best results highlighted in bold). For each metric we denote higher(\uparrow) or lower(\downarrow) is better. JPEG q indicates a JPEG compressed image with $QF = q$ (e.g. 10), while (REC q) indicates the corresponding reconstruction using [33]. Captions created from reconstructed images obtain a better score for every metric.

QUALITY	BLEU_1 \uparrow	METEOR \uparrow	ROUGE \uparrow	CIDEr \uparrow	SPICE \uparrow
JPEG 10	0.589	0.173	0.427	0.496	0.103
REC 10	0.730	0.253	0.527	1.032	0.189
JPEG 20	0.709	0.241	0.513	0.937	0.174
REC 20	0.751	0.266	0.543	1.105	0.201
JPEG 30	0.740	0.258	0.535	1.054	0.194
REC 30	0.757	0.269	0.549	1.133	0.205
JPEG 40	0.748	0.263	0.542	1.087	0.200
REC 40	0.758	0.270	0.549	1.132	0.206
JPEG 60	0.755	0.267	0.546	1.117	0.204
REC 60	0.760	0.270	0.550	1.137	0.207
ORIGINAL	0.766	0.274	0.556	1.166	0.211

detections of indoor elements like “floor” and “wall” are likely reasons for the wrong caption, as opposed to the correct recognition of a “white wave” and “blue water” in the GAN-reconstructed image.

Results without reference captions. A common setting that is used to evaluate image enhancement algorithms is Full-Reference image quality assessment, where several image similarity metrics are used to measure how much a restored version differs with respect to the uncorrupted original image. This kind of metrics, measuring pixel-wise value differences are likely to favor MSE optimized networks which are usually prone to obtain blurry and lowly detailed images.

In certain cases, it is not possible to use Full-Reference quality metrics, e.g. if there’s no available original image. These kind of metrics typically evaluate the “naturalness” of the image being analyzed. In the same setup we used previously, we perform experiments using NIQE and BRISQUE which are two popular No-Reference metrics for images. Interestingly, these metrics tend to favor GAN restored images instead of the original uncompressed ones. Most surprisingly, NIQE and BRISQUE obtain better results

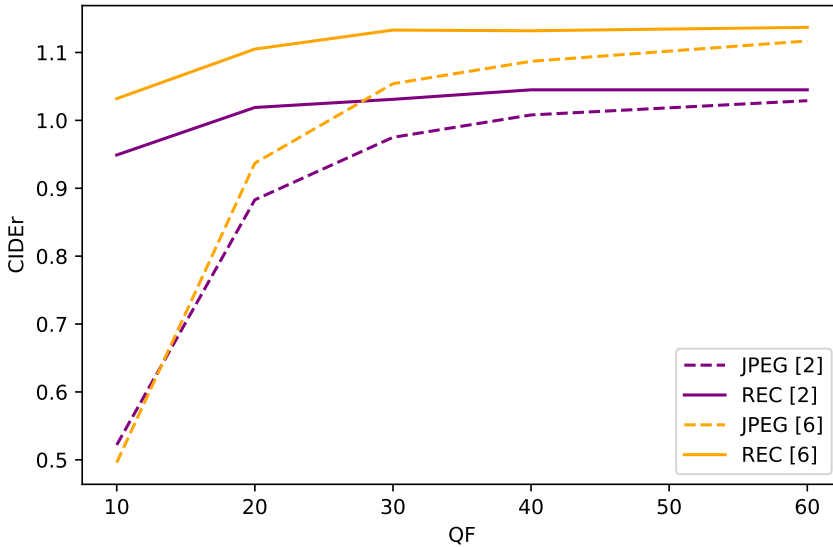
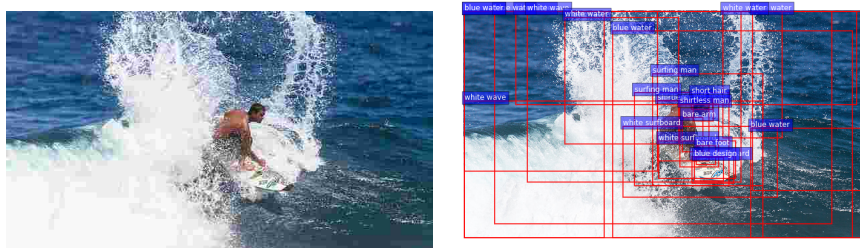


Figure 7.6: CIDEr scores using [2] (purple) and [25] (yellow) on compressed and restored images for different QFs from MS-COCO.

when we reconstruct the most degraded version of images (QF 10-20), but these values increase as we reconstruct less degraded images. We believe that BRISQUE and NIQE favor crisper images with high frequency patterns which are distinctive of GAN based image enhancement and they are typically stronger when reconstructing heavily distorted images.

In Tab. 7.2 we report results on COCO for Full-Reference and No-Reference indexes. In this setup, we compress the original images at different QFs and then we restore them with a QF specific artifact removal GAN. We use the uncompressed image generated caption as ground truth, as in Tab. 7.3. The results show that, for restored images, PSNR accounts for a slight improvement while SSIM indexes lower than the compressed counterparts. This is an expected outcome, as in [33] it is shown that state of the art results on PSNR can be obtained only when MSE is optimized and on SSIM if the metric is optimized directly. Nonetheless, as can be seen in Fig. 7.1, GAN enhanced images are more pleasant to the human eye, therefore we should not rely just on PSNR and SSIM for GAN restored images. LANBIQUE, using [25], is in



A couple of people sitting next to a Christmas tree. A man riding a wave on a surfboard in the ocean.

Figure 7.7: Bottom-Up detection process of captioning on two images: left) JPEG compressed; right) GAN reconstruction. Note that several mistaken detections on the left image are avoided in the right one. In particular on the left “surfboard” is missed and “white floor” and “blue wall” are wrongly detected. These two indoor details are the one that likely misled the captioning.

line with LPIPS [151]. Unfortunately, LPIPS, as shown in Tab. 7.3 has low correlation with scores determined by human perceived quality.

Correlation with Mean Opinion Score. In Fig. 7.8 *left*) are reported subjective evaluation results as Mean Opinion Scores (MOS) as box plots, showing the quartiles of the scores (box), while the whiskers show the rest of the distribution. The plots are made for the original images, the images compressed with JPEG using a $QF=10$, and the images restored with the GAN-based approach [33] from the heavily compressed JPEG images. The figure shows that the GAN-based network is able to produce images that are perceptually of much higher quality than the images from which they are originated; the average MOS score for JPEG images is 1.15, for the GAN-based approach is 2.56 and for the original images it is 3.59. The relatively low MOS scores obtained also by the original images are related to the fact that COCO images have a visual quality that is much lower than that of dataset designed for image quality evaluation. To give better insight on the distribution of MOS scores, Fig. 7.8 *right*) shows the histograms of the MOS scores for the three types of images: orange histogram for the original images, green for the JPEG compressed images and blue for the restored images.

We further show that our language based approach correlates with per-

Table 7.2: Evaluation using No-Reference and Full-Reference metrics on MS-COCO. For each metric we denote higher(\uparrow) or lower(\downarrow) is better. JPEG q indicates a JPEG compressed image with $QF = q$ (e.g. 10), while (REC q) indicates the corresponding reconstruction using [33]. NIQE and BRISQUE rate better GAN images than the ORIGINAL. SSIM always rate restored images worse than compressed. PSNR shows negligible improvement. [25] and CIDEr have been used by LANBIQUE-NC respectively as language model and language metric.

QUALITY	NIQE \downarrow	BRISQUE \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	LANBIQUE-NC \uparrow
JPEG 10	6.689	52.67	25.45	0.721	0.305	0.542
REC 10	3.488	17.93	25.70	0.718	0.144	1.118
JPEG 20	5.183	43.99	27.46	0.796	0.187	0.956
REC 20	3.884	17.85	27.60	0.784	0.085	1.289
JPEG 30	4.474	37.72	28.61	0.831	0.134	1.165
REC 30	3.601	18.32	28.81	0.819	0.060	1.370
JPEG 40	4.011	33.61	29.41	0.852	0.105	1.260
REC 40	3.680	18.68	29.44	0.836	0.048	1.424
JPEG 60	3.588	28.15	30.71	0.880	0.067	1.366
REC 60	3.885	19.45	30.61	0.862	0.032	1.482
ORIGINAL	3.656	21.79	-	-	-	-

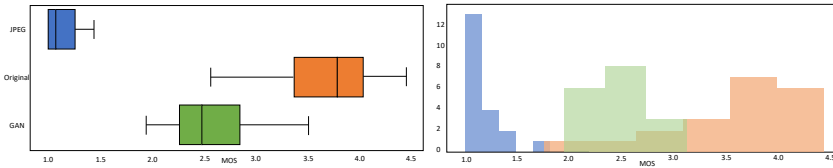


Figure 7.8: *Left*) Subjective image quality evaluation of original COCO images (orange), heavily compressed JPEG images (blue) and their restored version obtained with the GAN-based approach (green). Restored images are perceived as having a better quality than their compressed versions. *Right*) Histograms of MOS scores of the three types of images.

ceived quality using a IQA benchmark test on the LIVE dataset [111] that consists of 29 high resolution images compressed at different JPEG qualities for a total of 204 images. For each LIVE image a set of user scores is provided indicating the perceived quality of the image. However, no caption is provided in this dataset. For this reason, we consider the output sentences of captioning approaches over the undistorted image as the ground truth in order to calculate the language similarity measures, following the LANBIQUE-NC protocol presented in Sect. 7.3.2. In Tab. 7.3 we show the Pearson correlation score of different captioning metrics and other common Full-Reference quality assessment approaches. The experiment shows an interesting behaviour of our approach in terms of correlation. In the first place, we can observe that each captioning metric has a correlation index that is higher or at least comparable with the other Full-Reference metrics. In particular, METEOR and CIDEr perform better than the other metrics independently of which captioning algorithm is used. In the following experiments LANBIQUE, LANBIQUE-NC and LANBIQUE-NR have been computed using CIDEr metric. Moreover, we observe that the correlation metric significantly improves if we employ a more performing captioner. In this case, the visual features used by the two captioning techniques are exactly the same, the main difference lies in the overall language generation pipeline of the approaches. Hence, we argue that language is effectively useful for quality assessment, and the more a captioning algorithm is capable of providing detailed and meaningful captions the better we could use the generated sentences to formulate good predictions about the quality of images.

In order to better understand what metric could be used instead of human evaluation, we computed the correlation coefficient

$$\rho = \frac{\sum_{i \in \mathcal{D}} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i \in \mathcal{D}} (x_i - \bar{x})^2 \sum_{i \in \mathcal{D}} (y_i - \bar{y})^2}} \quad (7.9)$$

between BRISQUE [83], NIQE [85], the proposed LANBIQUE and MOS for all versions of the images. As shown in Tab. 7.4, it turns out that using a fine-grained semantic task as image captioning is the best proxy (highest correlation) of real human judgment.

Fig. 7.9 shows a captioning example from the COCO images used in the subjective quality evaluation experiment. On the left we show a sample compressed with JPEG with a QF=10, on the center we show the image restored with [33] and on the right we show the original one. It can be observed that



Figure 7.9: Examples of captions for COCO images used in the subjective quality evaluation. Left column) JPEG compressed with QF=10; Center column) GAN-based restoration from JPEG compressed images with QF=10; right column) original images.

Table 7.3: Pearson score, correlating scores with users’ MOS for different captioning metrics and image based Full-Reference approaches on LIVE dataset. CIDEr obtains a superior score with respect to image based methods.

Metric	LANBIQUE-NC w/ [25]	LANBIQUE-NC w/ [2]
BLEU 1	0.873	0.838
METEOR	0.900	0.846
SPICE	0.895	0.844
ROUGE	0.861	0.832
CIDEr	0.901	0.854
PSNR		0.857
SSIM		0.893
LPIPS		0.859

Table 7.4: Pearson’s Correlation coefficient, $\rho(X, Y)$ between No-Reference and captioning based metrics ($x_i \in X$) and MOS ($y_i \in Y$), as defined in Eq. 7.9 on for a sample set \mathcal{D} sampled from COCO.

Metric	NIQE	BRISQUE	LANBIQUE
$\rho \uparrow$	0.84	0.89	0.96

the caption of the restored image is capable of describing correctly the image content, on par with the caption obtained on the original image. Instead, the caption of the highly compressed JPEG image is completely unrelated to image content, probably due to object detection errors.

7.4.2 Results on all distortions

We further show the performance of our approach in full reference image quality assessment on other types of distortion. In this experiment we keep using LIVE dataset, as it contains images corrupted with other processes, such as Gaussian blur, fast-fading, JPEG2000 and white noise, but we add also a recent large scale PieAPP dataset.

Results on LIVE

We repeat the same experiment done for JPEG images on LIVE dataset, firstly considering each distortion separately and then all the distortions together. In Tab. 7.5 we show the Pearson score for LANBIQUE and several Full-Reference approaches. As we can see, our approach seems to underperform on each distortion except for JPEG, while SSIM and LPIPS are consistent despite the diversity of decaying processes. This is somehow expected, as blur and white noise tend not to harm detection significantly unless they are used with high intensity. Fast fading on the other hand, is to be considered as local distortion. For this reason, objects may not be corrupted at all, thus leading to unchanged detection performances and consequently low correlation scores for our assessment approach. As expected LANBIQUE-NR obtains a lower score than LANBIQUE-NC: in fact LANBIQUE-NC is an upper bound for the No-Reference version since this latter protocol would require a perfect blind restoration method capable of obtaining the reference images to obtain the same score.

Table 7.5: Pearson’s correlation of our approach (Full-Reference and No-Reference) on all distortions present on LIVE compared with other Full-Reference metrics. For the No-Reference approach (LANBIQUE-NR) fast fading score is not reported since actual State-Of-The-Art restoration approaches perform poorly on this distortion.

	GBLUR	FASTFADING	JP2K	JPEG	WN	TOTAL
PSNR	0.767	0.763	0.83	0.857	0.732	0.752
SSIM	0.886	0.845	0.89	0.893	0.951	0.789
LPIPS	0.951	0.836	0.885	0.859	0.910	0.785
LANBIQUE-NC	0.786	0.651	0.787	0.901	0.735	0.792
LANBIQUE-NR	0.676	-	0.679	0.796	0.667	0.701

However, we experience a totally different scenario when the distortions are evaluated all together. We can see that for each IQA approach we have tested, there is a significant drop in the correlation coefficient with respect to single distortion experiments. We argue this is due to the fact that the scores for single distortion types are well correlated but considering the scores for multiple distortion classes there is a bigger discrepancy between them that leads to a decrease of the total score. On the other hand, our approach does not suffer from this phenomenon, as the performance we measure in these

conditions is consistent, if not higher, with single distortions. Moreover, our language based approach slightly overperforms the other measures on the same data and at the same conditions.

Results on PieAPP

Finally, we use a more recent large scale dataset [94]. Prashnani *et al.* collected a very large dataset increasing the number of distortions with respect to existing IQA benchmarks. Moreover, they designed the testing procedure differently. Specifically, instead of collecting multiple subjective scores from a set of users, they rely on the fact that for humans is easier to tell which of two distorted images I_A , I_B is closer to a reference undistorted one I_R . Then images are labelled by the percentage of users that preferred an I_A with respect to I_B . If there is an even split between these two populations, it means that both images are equally different from the reference I_R . Starting from 200 reference images and combining a diverse set of 75 distortions, with a total of 44 distortions in the training set, and 31 in the test set which are distinct from the training set, the PieAPP dataset accounts for a total of 77,280 pairwise comparisons for training (67,200 inter-type and 10,080 intra-type). In Tab. 7.6 we report results in term of Kendall’s Rank Correlation Coefficient: $KRCC = 1/\binom{n}{2} \sum_{i<j} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j)$; Pearson’s Linear Correlation Coefficient (PLCC or $\rho(X, Y)$ as defined in Eq. 7.9) and Spearman’s Rank Correlation (SRCC), $\rho(R(X), R(Y))$ where $R(X)$ are the ranks of sample X .

Interestingly, both image and type of distortions do not overlap between training and testing. In Tab. 7.6, we show how our LANBIQUE-NC approach (using CIDEr and [25]) ranks with respect to non-learning (top) and learning based (bottom) approaches. We refer to non-learning methods when the algorithm is not relying in any way on any kind of supervision for the IQA task. Our approach exploits learned deep networks and features but those are not the result of training on PieAPP or on any other IQA dataset. Instead, the lower portion of the Table reports methods [12,17,56,77], that are specifically trained to score image similarity. Very interestingly our LANBIQUE-NC approach is consistently better than any non-learned image similarity metric and outperforms all both [12] and [94], with [12] being a close comparison.

Table 7.6: Evaluation on PieAPP dataset. Column FR indicates if the method is used in a Full-Reference fashion or not. For all metrics higher is better. We report Kendall’s Rank Correlation Coefficient (KRCC), Pearson’s Linear Correlation Coefficient (PLCC) and Spearman’s Correlation Coefficient (SRCC). KRCC is computed for the whole set ($pAB \in [0, 1]$) and for a set for which there is more agreement between human labels ($pAB \notin [.35, .65]$). LANBIQUE-NC has better KRCC with respect to all non-learning based methods and is also better than most of the methods that exploit some sort of supervision to perform IQA.

Method	FR	Learning	KRCC ($pAB \in [0, 1]$)	KRCC ($pAB \notin [.35, .65]$)	PLCC	SRCC
MAE	yes	no	.252	.289	.302	.302
RMSE	yes	no	.289	.339	.324	.351
SSIM	yes	no	.272	.323	.245	.316
MS-SSIM	yes	no	.275	.325	.051	.321
GMSD	yes	no	.250	.291	.242	.297
VSI	yes	no	.337	.395	.344	.393
PSNR-HMA	yes	no	.245	.274	.310	.281
FSIMc	yes	no	.322	.377	.481	.378
SFF	yes	no	.258	.295	.025	.305
SCQI	yes	no	.303	.364	.267	.360
LANBIQUE-NC	yes	no	.342	.412	.316	.310
DOG-SSIMc [90]	yes	yes	.263	.320	.417	.464
Lukin et al. [77]	yes	yes	.290	.396	.496	.386
Kim et al. [56]	yes	yes	.211	.240	.172	.252
Bosse et al. [12]	no	yes	.269	.353	.439	.352
Bosse et al. [12]	yes	yes	.414	.503	.568	.537
PieAPP [94]	yes	yes	.668	.815	.842	.831

7.5 Conclusion

In this work we propose LANBIQUE, a new approach to evaluate image quality using language models. Existing metrics based on the comparison of the restored image with an undistorted version may give counter-intuitive results. On the other hand, the use of naturalness based scores may in certain cases ranks restored images higher than original ones.

We show that instead of using signal based metrics, semantic computer vision tasks can be used to evaluate results of image enhancement methods. Our claim is that a fine grained semantic computer vision task can be a great proxy for human level image judgement. Indeed we find out that employing algorithms mapping input images to a finer output label space, such as captioning, leads to more discriminative metrics.

LANBIQUE is capable to evaluate the quality of images corrupted by

different distortions and its performance is comparable to other image quality assessment methods. Moreover, we have modified our evaluation pipeline to transform our original solution into a No-Reference method and we have demonstrated that it keeps performing fair on standard benchmarks.

Finally, we have tested LANBIQUE on a large scale dataset that contains unknown distortions. Despite the lack of learning and of knowledge on data, our approach outperforms every baseline that does not use learning for the evaluation, and it is comparable to most of the learned approaches on the same data.

As a final note, we would like to remark that our approach will continuously improve thanks to the advancement of image captioning and enhancement networks. Indeed, we have shown that without changing the visual features, switching to a better captioning algorithm we get a higher performance. Moreover, being LANBIQUE-NC an upper bound for LANBIQUE-NR, as image enhancers gain quality, the gap between the performance of these two methods will shrink.

Chapter 8

STILT: Scene-Text Image and Language Transformer for Cross-Modal Retrieval

*Image-text cross-modal retrieval tasks are capable of understanding visual semantics, leading to correct ranking of captions (and vice versa). However, their performance is limited when images are associated with complementary descriptions like scene-text and visual information is less relevant. In this work we propose a **Scene-Text Image and Language Transformer (STILT)** for cross-modal retrieval. STILT combines image and scene-text representations according to scene-text position in the image, and a fusion token representation is learned to merge visual and scene-text information. Scene-text and caption representations are aligned using a contrastive loss and are then given as input to two cross-modal encoders which improve representation learning. We demonstrate the effectiveness of our approach in two challenging contextual captioning datasets, GoodNews and Politics, and our experiments show that STILT outperforms the state-of-the-art by about 6% on both.*¹

¹Part of this work was conducted while the author was a visiting Ph.D. student at Universitat Autònoma de Barcelona, Barcelona (Spain), from October to March 2021-2022. The work has been submitted to the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

8.1 Introduction

In traditional image-text cross-modal retrieval tasks, image descriptions are descriptive of the visual content of the image (for example “a man holding a microphone”) and do not contain contextual information (for example “Bruce Springsteen singing at Madison Square Garden”). Thus, standard Transformer-based approaches only work when text is well-aligned with the visual content. This is a significant weakness since in real scenarios users tend to introduce contextual information when describing objects or looking for something in order to communicate more specific and fine-grained information. A few works address this problem, designing specific training objectives to learn the degree of abstractness of an image-text pair [121,122].

Most existing cross-modal retrieval approaches do not consider contextual information decodable from the image such as scene-text. To address this, we propose a Scene-Text Image and Language Transformer (STILT) model for image-text cross-modal retrieval. STILT works with both scene-text-free images and images containing scene-text. We first encode the image and caption with independent image and language Transformer encoders. The extracted scene-text information is encoded with the same language encoder used for the caption. Image and scene-text embeddings are then combined according to the position of scene-text in the image, and we use two multi-modal encoders to fuse scene-text image features with caption features through cross-modal attention. We use three losses for model pre-training: an image-text contrastive (ITC) loss on representations from the unimodal encoders for feature alignment, a masked-image modeling (MLM) objective to improve language encoding representation, and an image-text matching (ITM) loss on the top of the two multi-modal encoders for a more fine-grained image-text interaction.

The contributions of this work are the following:

- We propose a novel Transformer-based architecture that exploits scene-text information (OCR, and text position in the image) for robust cross-modal representation. We demonstrate the effectiveness of our pre-training approach on datasets with different percentages of scene-text images.
- We compare our approach with state-of-the-art approaches on abstractly aligned image-text datasets and full scene-text datasets, and

our experiments demonstrate that STILT significantly outperforms existing methods.

- We show through qualitative analysis the effectiveness of using scene-text for aligned and non-aligned image-text datasets, demonstrating that scene-text is useful information to learn more abstract image-text alignment.

8.2 The STILT Approach

In this section we introduce and describe the STILT model architecture and then define our pre-training strategy.

8.2.1 Model Architecture

As shown in Fig. 8.1, the STILT architecture consists of an image encoder, a text encoder, and two multi-modal encoders:

- The image encoder (I_{enc}) consists of a 6-layer Visual Transformer ViT-B/16 as initialized with weights pre-trained on Imagenet-1k.
- The textual encoder (T_{enc}) consists of a 6-layer BERT_{base} model.
- The two multi-modal encoders consist of a 6-layer ViT-B/16 (V_{mm}) and a 6-layer BERT_{base} (T_{mm}).

The input image I is given as input to the image encoder (I_{enc}) and outputs a sequence of visual tokens $\{v_{cls}, v_0, v_1, \dots, v_n\}$ corresponding to different visual regions of the image. The textual encoder T_{enc} encodes the caption and outputs a sequence of tokens $\{c_{cls}, c_0, c_1, \dots, c_m\}$. The extracted scene-text from the image I is encoded with the textual encoder T_{enc} which outputs a sequence of tokens $\{s_{cls}, s_0, s_1, \dots, s_z\}$. We consider the coordinates relative to each scene-text bounding box $\{p_0, p_1, p_2, \dots, p_z\}$ and we fuse them with the corresponding scene-text token embedding using the following strategy:

$$sp_i = (W_{ms}s_i + b_{ms}) + (W_{ps}p_i + b_{ps}) \quad (8.1)$$

where W_{ms} , W_{ps} , b_{ms} , b_{ps} are models parameters. Moreover, we filter the encoded visual regions that contain scene-text and we fuse them with the corresponding scene-text bounding box coordinates as follows:

$$vp_i = (W_{nv}v_i + b_{nv}) + (W_{np}p_i + b_{np}) \quad (8.2)$$

where W_{ms} , W_{ps} , b_{ms} , b_{ps} are again models parameters.

Finally, we fuse the visual [CLS] token v_{cls} and the scene-text [CLS] token s_{cls} as following:

$$f_{cls} = W_{fn}((W_{nv}v_{cls} + b_{nv}) \oplus (W_{ns}s_{cls} + b_{ns})) + b_{fn} \quad (8.3)$$

where \oplus is the concatenation operation and W_{nv} , W_{ns} , W_{fn} , b_{nv} , b_{ns} , b_{fn} are once again model parameters. We consider as final encoded image representation the concatenation of fused [CLS] token f_{cls} , visual embeddings, position-aware scene-text tokens (SP) and position-aware visual tokens:

$$V = [f_{cls}, v_{cls}, v_0, \dots, v_n, vp_0, \dots, vp_k, s_{cls}, sp_0, sp_1, \dots, sp_z] \quad (8.4)$$

These visual features V are fused with the caption features through cross-attention in the multi-modal encoders T_{mm} and V_{mm} .

8.2.2 Pre-Training

Inspired by [66], we pre-train STILT with three different losses: an image-text contrastive loss on the image encoder I_{enc} and the text encoder T_{enc} , Masked Language Modeling (MLM) on the multi-modal encoder T_{mm} , and Image-Text-Matching (ITM) on both the multi-modal encoders T_{mm} and V_{mm} .

Image-text Contrastive Loss. We use this loss to align image and text representations after the encoding phase. We project the [CLS] token of the visual representation v_{CLS} and the contextual representation c_{cls} into the same space after normalizing the output embeddings. We define a similarity function $s(v, c) = p_v(v_{cls})^T p_c(c_{cls})$, where p_v and p_c are linear projections that maps v_{cls} and c_{cls} into the same number of dimensions. For each image and text pair in a batch we use as predictions the softmax-normalized image-to-text p^{i2t} and text-to-image p^{t2i} similarities (as defined in [66]) and compute the loss as sum of cross-entropy (CE) losses:

$$L_{itc} = CE(y^{i2t}(I), p^{i2t}(I)) + CE(y^{t2i}(T), p^{t2i}(T)) \quad (8.5)$$

where $y^{i2t}(I)$ and $y^{t2i}(T)$ are the ground-truth one-hot vectors indicating positive matching pairs.

Masked Language Modeling. Masking tokens in the input text and training the model to predict the masked token has been demonstrated to

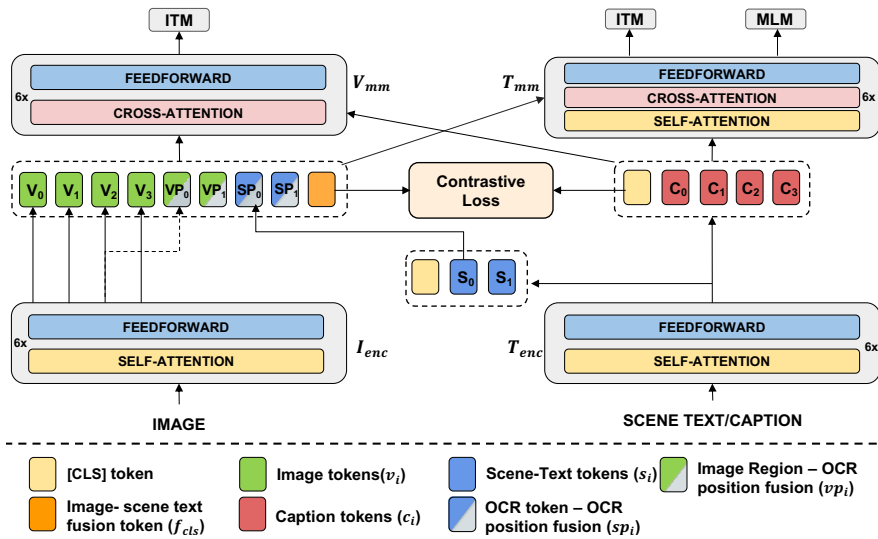


Figure 8.1: The STILT architecture. The image is encoded with a ViT encoder I_{enc} , and both scene-text and caption are encoded with the same BERT encoder T_{enc} . Image tokens corresponding to image patches that contain scene text are replicated and combined with scene-text position, while scene text embeddings are combined with corresponding positions. The two multi-modal encoders V_{mm} and T_{mm} merge visual and caption representations through cross-attention.

be good practice for pre-training language transformers [27, 66]. For the implementation of the L_{mlm} loss, we follow [27], randomly masking 15% of tokens with a [MASK] token (80% of the time) or a random token (10% of the time) or the same token (10% of the time).

Image-Text Matching. We use this loss on the top of the two multi-modal encoders T_{mm} and V_{mm} for predicting when an image-caption pair is matching or not. To compute the two-class probability p^{itm} (match/not match), we use each of the [CLS] tokens in output from the two multi-modal encoders, projecting them into a two-dimensional space and using a softmax as activation function. More specifically:

$$L_{itm}^v = CE(y^{itm}, p^{itm}(I, T)) \quad (8.6)$$

$$L_{itm}^c = CE(y^{itm}, p^{itm}(T, I)), \quad (8.7)$$

where y^{itm} is ground-truth two-dimensional one-hot vector.

The final aggregated L_{itm} loss is then:

$$L_{itm} = L_{itm}^v + L_{itm}^c, \quad (8.8)$$

and the final loss function for STILT pre-training is:

$$L = L_{itc} + L_{mlm} + L_{itm}. \quad (8.9)$$

8.2.3 Implementation Details

For STILT we use BERT_{base} [27] for caption and scene text encoding and ViT-B/16 [146] for image encoding. For our experiments, we consider a light version of our model, (STILT_{light}) where the uni-modal encoders have 9 layers and the multi-modal encoders have 3 layers. We freeze the first 8 layers of the uni-modal encoders in order to have a faster training phase. We pre-trained our model using ITC, ITM and MLM losses for 10 epochs and then fine-tuned it for another 10 epochs using ITC and ITM losses with batch size of 128 on 4 NVIDIA A100s.

8.3 Experimental Results

We performed a range of experiments in order to evaluate the effectiveness of our approach and compare STILT to the state-of-the-art. In the next

section we described the datasets used in all experiments, and in Sec. 8.3.2 we show the effectiveness of our approach with varying input information and loss objectives. In Sec. 8.3.3 we compare our approach with the state-of-the-art on complementary image-text datasets where image and text are symbolically aligned, and in and Sec. 8.3.4 we compare on full scene-text datasets.

8.3.1 Datasets

For our experiments, we use three different types of datasets: contextual datasets, scene-text images datasets, and general-purpose datasets.

Contextual Datasets. The Politics [121] dataset consists of around 246K image-text pairs where the text is part of news articles. Images and related descriptions are complementary and have a more abstract and symbolic alignment. The GoodNews [10] dataset consists of around 466K images paired with captions from the New York Times. Like Politics, captions contain additional information not inferable from the visual content.

Scene-Text Image Datasets. TextCaps [114] consists of around 28K images containing OCR taken from OpenImages and each image is associated with 5 captions. Coco Text Captions [78] (CTC) is a subset of images containing scene-text images from the COCO training set.

General Purpose Datasets. Visual Genome [60] is a dataset annotated with visual question-answer pairs and contains about 5.4M region descriptions associated with around 108K images. The Flickr30k [92] dataset contains 31,000 images collected from Flickr, Each image is paired with 5 reference sentences provided by human annotators.

Contextual datasets, scene-text images datasets, and general purpose datasets are very different. General purpose datasets are standard datasets for image-text retrieval where semantic visual information contained in the image is described in the caption. Consequently, image-text pairs are well-aligned for these datasets. In contrast, contextual datasets consist of abstractly-aligned image-text pairs since the semantic visual information is not described in the caption that tends to add complementary information to that shown in the image. Finally, full scene-text datasets consist of only images that contain scene-text with their corresponding captions and thus images and captions are well-aligned.

Input	Pre-training Losses	I → T			T → I		
		5-way	10-way	20-way	5-way	10-way	20-way
Image + Caption	ITC	64,48	51,28	40,13	63,59	50,88	40,13
Image + Caption	ITC + ITM + MLM	65,68	52,33	41,52	65,13	51,95	41,86
Image + Caption + OCR	ITC	68,38	55,73	44,86	68,29	56,13	44,87
Image + Caption + OCR	ITC + ITM + MLM	69	55,88	44,97	68,43	56,34	44,95
Image + Caption+ OCR + OCR Positions	ITC + ITM + MLM	70,61	56,12	45,34	69,98	57, 41	45,3

Table 8.1: Performance of our approach for different training losses and inputs on the Politics dataset. We report the results for image-to-text (I→T) and text-to-image (T→I) retrieval for recall@1 in 5, 10, and 20-way retrieval tasks.

8.3.2 Evaluation Method

To evaluate the effectiveness of our approach we consider different variants of our $STILT_{light}$ model varying the pre-training losses and network inputs. In Tab. 8.1 we show the performance of our method for the cross-modal retrieval task on the Politics dataset. The results reported in the table are for both image-to-text and text-to-image retrieval and are computed as described in [122] by picking random (non-paired) samples from the test set, along with the ground-truth paired sample and computing Recall@1 in 5, 10, and 20-way tasks. When using only the ITC loss we rank images according to texts (and vice versa) computing cosine similarity between the representations in output from the uni-modal encoders. Adding the ITM loss, the rank is computed as the sum of the ITM scores in output from the two multi-modal encoders.

We extracted scene text and its position from the Politics dataset using the Google vision API for optical character recognition. Since this dataset contains both scene-text images and images with no scene-text, we input into our model an empty string when there is no scene-text in the image.

We observe that adding image-text matching (ITM) and masked language modeling (MLM) to the contrastive loss (ITC) as pre-training losses always improves model performance. Moreover, adding scene-text (OCR) information as input to our model results in a significant performance boost. Finally, adding information about the scene-text positions in the image further improves the performance. STILT is able to effectively exploit scene text and its position to improve cross-modal retrieval.

method	Politics						GoodNews					
	I→T			T→I			I→T			T→I		
	5-way	10-way	20-way	5-way	10-way	20-way	5-way	10-way	20-way	5-way	10-way	20-way
PVSE [116]	59.19	-	-	60.57	-	-	85.16	-	-	85.26	-	-
HAL [73]	59.19	-	-	59.03	-	-	86.23	-	-	85.79	-	-
AMRANI [82]	61.17	-	-	61.17	-	-	86.29	-	-	86.78	-	-
THOMAS ² [123]	62.74	-	-	62.39	-	-	87.91	-	-	87.82	-	-
THOMAS [123]	64.67	-	-	64.92	-	-	88.49	-	-	88.65	-	-
STILT _{light}	70.61	56.12	45.34	69.98	57.41	45.3	93.45	87.15	79.13	93.27	87.12	78.96
STILT _{light} + Finetuning	71.83	57.47	47.55	70.95	57.96	45.91	94.51	88.07	79.98	94.36	88.04	79.80

Table 8.2: Comparison between STILT and State-Of-The-Art approaches on non-literal image-text datasets. STILT outperforms other techniques on Politics and GoodNews datasets.

Model	CTC-1K						CTC-5K					
	I→T			T→I			I→T			T→I		
	R@1	R@5	r@10	R@1	R@5	r@10	R@1	R@5	r@10	R@1	R@5	r@10
SCAN [64]	36.3	63.7	75.2	26.6	53.6	65.3	22.8	45.6	54.3	12.3	28.6	39.9
VSRN [67]	38.2	67.4	79.1	26.6	54.2	66.2	23.7	47.6	59.1	14.9	34.7	45.5
STARNet [78]	44.1	74.8	82.7	31.5	60.8	72.4	26.4	51.1	63.9	17.1	37.4	48.3
ViSTA-S [21]	52.5	77.9	87.2	36.7	66.2	77.8	31.8	56.6	67.8	20.0	42.9	54.4
STILT _{light}	53.0	78.2	88.3	37.5	66.8	78.4	33.3	58.3	68.7	21.2	43.4	55.6

Table 8.3: Comparison between STILT and state-of-the-art approaches on CTC-1K and CTC-5K test datasets.

8.3.3 Cross-Modal Retrieval with Abstract Image-Text Alignment

Tab. 8.2 shows the performances of STILT_{light} on the Politics and Goodnews datasets. These two datasets contain a different proportion of scene-text images (in Politics about 50% of images, in GoodNews only about 20%). In this experiment, we trained our model on the training set of the reference dataset and then finetuned it on the same. We report the results of recall@1 in 5-, 10-, and 20-way tasks (differently from other approaches that show only 5-way results). We see that our model outperforms state-of-the-art approaches by more than 7% on Politics and 6% on GoodNews in the 5-way test. Excluding STILT, the methods listed in the table do not use scene-text and to the best of our knowledge, we are the first to exploit OCR to perform both scene-text and scene-text aware cross-modal retrieval in abstract-aligned image-text datasets. Moreover, we demonstrate that taking advantage of scene text for this kind of dataset is extremely helpful to find additional symbolic correlations between images and captions.



Figure 8.2: Qualitative results. The first five ranked captions for image-to-text retrieval on the Politics dataset. Our model exploits scene-text effectively (left column) and still correctly ranks when no scene-text is present (right column).

8.3.4 Full Scene-Text Image and Text Cross-Modal Retrieval

For full scene-text aware retrieval, we evaluate STILT on the CTC-1K and CTC-5K test sets using the train and test splits described in [78]. We pre-trained $STILT_{light}$ on Visual Genome and fine-tuned it on Flickr30K, TextCaps, and the training split of CTC. Tab. 8.3 shows the comparison between STILT and the state-of-the-art for recall@1 recall@5 and recall@10. We see that $STILT_{light}$ outperforms other techniques on both CTC-1K and CTC-5K. VISTA-S also uses a transformer-based architecture with three separated encoders for scene-text, caption, and image. STILT uses the same BERT encoder to encode both caption and scene-text, but we replicate image embeddings containing scene-text and combine them with scene-text positions to obtain more consistent scene-text-aware image representation before passing it to the multi-modal encoders. Note that for $STILT_{light}$ we are training only the last four layers of each ViT and BERT model, unlike the other approaches that train a full model.

8.4 Qualitative Analysis

In order to qualitatively demonstrate the effectiveness of the STILT model we consider examples from the GoodNews dataset. This dataset represents a real scenario since it contains a wide variety of image-caption pairs that come from New York Times news. For this qualitative analysis, STILT_{light} model is tested in a 50-way task (ranking 50 captions according to the content of an image) with the purpose of outlining its limits. In Fig. 8.2 we give some samples for the image-to-text retrieval task. The first column contains of examples with scene-text images, and the second of images without scene-text. We see that STILT is capable of correctly ranking the captions for images that would have never been ranked correctly without exploiting scene-text. In particular, in the second example, the first five ranked captions describe a specific person, and in a real scenario if a human does not know the name of the person the only way to generate a correct ranking is by focusing on the scene-text. This is a very frequent problem in news datasets since in most cases articles are related to well-known persons and ranking algorithms are not trained to recognize them. The last two examples in the first column show examples where the scene-text is useless to generate the correct rank. Our model fails and ranks the correct caption as second but the rank 1 captions have similar meaning. The second column gives results for images with no scene-text where we see that STILT is able to associate well-aligned image-caption pairs.

8.5 Conclusions

State-of-the-art cross-modal retrieval models only work with well-aligned image-text pairs. In this work we propose a transformer-based approach to address this limitation by exploiting scene-text to perform cross-modal retrieval. We demonstrate the effectiveness of our approach on news datasets and full scene-text datasets, and our experiments show that STILT consistently outperforms the state-of-the-art. To the best of our knowledge ours is the first work exploiting scene-text for cross-modal retrieval applied to abstract image-text datasets. Considering that STILT_{light} achieves very impressive performance despite having most layers frozen during pretraining, the next interesting step would be to train the full model end-to-end.

Chapter 9

Conclusions

9.1 Summary of Contributions

In the first part of this dissertation we focused on one of the main problems of vision and language systems, Visual Question Answering, which is notably poor at incorporating contextual data. This makes many of these approaches useful in research contexts, but unusable in real-world scenarios. We consider the concrete scenario of Visual Question Answering in the cultural heritage domain and proposed two systems capable of answering both contextual and visual questions. These approaches are also suitable for use in other scenarios. Since there were no Visual Question Answering datasets for cultural heritage, we created a dataset containing about 6.5M question-answer pairs associated with about 500K images. For completeness, an annotation system for question-answer pairs in the cultural heritage domain has also been developed, which can also be used to test the accuracy of VQA models.

A key factor limiting progress in Image Quality Assessment is the lack of annotated data. We proposed a generative approach to data augmentation that partially mitigates this problem, and additionally developed a novel approach to exploiting image captioning as an image quality evaluator. This second approach does not require model training on annotated IQA datasets and is completely based on the robustness of the captioner. Furthermore, the experiments carried out also allow an evaluation from other points of view of the captioning systems.

Finally, our contributions to the field of cross-modal retrieval are very much connected to those of the Visual Question Answering described above.

The STILT model uses scene-text to address the retrieval problem of contextual data (image-text pairs taken for example from newspaper articles). Our experiments show that STILT achieves excellent performance even on standard (non-contextual) datasets.

9.2 Directions for Future Work

In the last year, research in the field of image and language tasks has mainly concerned the development of models based on transformers which, thanks to extensive pre-training, obtain impressive results in all vision and language downstream tasks. These models are extremely large (some even require 40 GPUs for training) and are hardly usable in real-world scenarios. In addition, models trained on billions of samples have been developed whose performance is truly impressive. Notable examples include CLIP [96] for cross-modal retrieval model, as well as DALL-E [98] and stable diffusion models [101] for text-to-image generation and manipulation.

A possible research direction will be to reduce the complexity of these extremely large models without reducing their performance so that they can be used in real scenarios. Moreover, since the aforementioned approaches obtain excellent results, they can be exploited to create synthetic dataset annotations for specific purposes or be applied to improve the performance of models in other areas of computer vision.

As described in various chapters of this thesis, vision and language tasks do not work in real scenarios since they are not capable to process contextual information that is not inferable from the image content. The Computer Vision community seems to be still anchored to this limitation. New datasets for contextual tasks have to be collected to improve the ability of models to abstract from a visual representation to a more general one. This branch of research needs also new metrics that involve the contextual aspect to have a better evaluation of the approaches. Another limitation is due to the way of managing external knowledge from the current approaches when addressing contextual datasets. Existing knowledge graphs, ontologies, and document libraries are too big to be processed by a model at run time. A great step forward in this direction will be a knowledge representation quick to explore that contains as much information as possible. An alternative direction to address vision and language tasks would be also to study self-supervised learning strategies avoiding long supervised training.

Transformer models have undoubtedly changed Computer Vision research improving the performance of vision and language tasks but they are not yet easily exploited in real contexts. Another important direction of research for the AI community will be to study the representation learning of transformer models in order to avoid the large pretraining phase or easily adapt an already pretrained transformer to other completely different domains. This will be extremely useful and will probably lead to new applications in multiple scenarios where the amount of training data is low as in some real contexts.

Appendix A

Data Collection for Contextual and Visual Question Answering in the Cultural Heritage Domain

In this demonstration we propose an annotation tool to collect question-answer samples for artworks, necessary to train and evaluate visual and contextual question answering models. The tool is completely web-based, and relies on an automatic question-answer generation model to aid the annotation process. Through the annotator, users can inspect and refine the generated annotations and obtain statistics on their quality. A pre-trained visual and contextual question answering model is also provided to the final user to be able to interact with the system by asking questions about artworks.¹

This appendix is related to the VQA approaches for Cultural Heritage, previously presented in Chapter 2, and Chapter 3

¹The work presented in this chapter has been presented as a Demo to the International Conference on Pattern Recognition (ICPR) 2021

A.1 Introduction

The usage of VQA for Cultural Heritage has been explored in Chapter 2, where questions have been categorized into two categories: visual if they refer to the content of the artwork and contextual if they refer to knowledge deductible only from an external source. Interacting through questions and dialogs will likely be the evolution of smart audio guides for museum visits and simple image browsing on personal smartphones. In this way, the classic audio guide turns into a smart personal instructor with which the visitor can interact by asking for explanations focused on specific interests. The advantages are twofold: on the one hand, the cognitive burden of the visitor will decrease, limiting the flow of information to what the user actually wants to hear; and on the other hand, it proposes the most natural way of interacting with a guide, favoring engagement. However, realizing such an interactive system is not straightforward. The biggest obstacle towards this goal is the lack of specialized data for the cultural heritage domain, which will require an expensive and temporally demanding annotation campaign. In particular, there is the need for question-answer pairs related to both the visual and contextual information of artworks. To address this limitation, we propose a semi supervised approach that relies on automatic question generators to adapt textual descriptions of artworks to data that can be used to train visual/contextual question answering models. The system we demonstrate is a web based annotation tool to browse, edit and validate datasets of automatically generated questions relative to images of artworks. The tool offers the advantage of lowering the annotation burden of building a dataset manually, while allowing the user to perform an analysis of question generation

A.2 Data Collection

The purpose of the proposed system is to aid users in the annotation of artwork images with visual and contextual questions/answers. Each artwork is paired with a picture and a textual description, which can be easily gathered from online sources such as Wikipedia or DBpedia. Users can assign a label to sentences to mark them as visual or contextual. These sentences are then fed to a text-based question generation model which converts them into questions and answers. The visual and contextual labels are automat-

ically transferred from sentences to questions. To obtain question-answer pairs, we first gather a collection of visual and contextual sentences relative to artworks. We use data from Artpedia [118], a dataset containing 2.930 paintings and a total of 28.212, manually labeled as visual or contextual (9.173 visual sentences and 19.039 contextual sentences). On average, an artwork is labeled with 3.1 visual sentences and 6.5 contextual sentences. The user of our system can browse all images and their textual labels and can enter new descriptions or modify existing ones. We then automatically generate question-answer pairs with *1* [30], a recently proposed end-to-end trainable sequence-to-sequence model. We have obtained more than 100.000 generated questions and answers. Each generated item can be inspected through the web interface and can be edited by the user. Once a sample has been inspected, it is marked as "revised" and is then considered to be part of the dataset under construction. If a new question is instead added directly by the user, it is automatically flagged as revised. To ease the data collection process, we developed a fast annotation web interface where multiple users can revise questions in parallel, inspecting a sequence of random question-answer pairs. In this way we have collected a dataset of 1027 manually revised question-answers out of a subset of 1500 automatically generated samples. During the revision process, if a mistake is identified in the automatically generated sample, the user can label it with a customizable error category. This provides us with statistics on the quality of the question generator, which offers interesting insights about the model. We have identified 10 error categories for questions, among which the most common are too long questions, two words questions (such as what is?) and nonsense questions.

A.3 Visual and Contextual Question Answering

We have integrated in the system the model presented in Chapter 2, which answers both visual and contextual questions, relying on a question classifier to understand whether a Question Answering or Visual Question Answering sub-module is better suited to answer. The user can therefore test the model by interacting with it, asking questions about an artwork and its contextual information through a chat. In this way it is possible to collect data with the proposed annotation tool, train a custom model and deploy it through the web interface to the final user. In addition, the collected data can be

used to test pre-trained models for visual and contextual question answering with a joint evaluation. In fact, in literature no dataset for visual question answering in the cultural heritage domain has been collected yet.

A.4 Conclusion

In this demonstration we have proposed a web based annotation tool to collect in a semi automatic way questions and answers relative to artworks. The tool relies on a text-based question/answer generator. The generated samples can then be manually inspected and revised. The system also offers to inspect the quality of the generated questions by gathering error statistics and provides an interface for the user to interact with a pre-trained question answering model that answers both visual and contextual questions.

Appendix B

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.

International Journals

1. L. Galteri, L. Seidenari, **P. Bongini**, M. Bertini, AD. Bimbo. “LANBIQUE: LANguage-based Blind Image QUality Evaluation”, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18 issue 2s, 2022.

Submitted

1. F. Becattini, **P. Bongini**, L. Bulla, AD. Bimbo, L. Marinucci, M. Mongiovi, V. Presutti “VISCOUNTH: A Large-Scale Visual and Contextual Question Answering Dataset for Cultural Heritage”, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

International Conferences and Workshops

1. **P. Bongini**, R. Del Chiaro, AD. Bagdanov, AD. Bimbo. “GADA: Generative adversarial data augmentation for image quality assessment”, in *International Conference on Image Analysis and Processing (ICIAP)*, 2019.
2. **P. Bongini**, Becattini, AD. Bagdanov, A. Del Bimbo “Visual question answering for cultural heritage”, in *IOP Conference Series: Materials Science and Engineering*, 2020.

3. L. Seidenari, L. Galteri, **P. Bongini**, M. Bertini, A. Del Bimbo “Language based image quality assessment”, in *ACM Multimedia Asia, 2021* (**Best paper award**).
4. **P. Bongini**, F. Becattini, A. Del Bimbo “Is GPT-3 all you need for Visual Question Answering in Cultural Heritage?”, in *European Computer Vision Conference (ECCV) Workshop on Vision and Art (VisArt), 2022*.

Submitted

1. **P. Bongini**, AF. Biten, AM. Delgado, D. Karatzas, AD. Bagdanov, “STILT: Scene-Text Image and Language Transformer for Cross-Modal Retrieval”, *Computer Vision and Pattern Recognition, 2023*.

Demo

1. F. Vannoni, **P. Bongini**, F. Becattini, AD. Bagdanov, Alberto Del Bimbo “Data Collection for Contextual and Visual Question Answering in the Cultural Heritage Domain”, in *International Conference on Pattern Recognition (ICPR), 2020*.

Bibliography

- [1] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *Proc. of ECCV*, 2016.
- [2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [4] L. Asprino, L. Bulla, L. Marinucci, M. Mongiovì, and V. Presutti, “A large visual question answering dataset for cultural heritage,” in *Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part II*, 2021, pp. 193–197.
- [5] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proc. of the ACL workshops*, 2005.
- [6] M. R. Banham and A. K. Katsaggelos, “Digital image restoration,” *IEEE signal processing magazine*, vol. 14, no. 2, pp. 24–41, 1997.
- [7] F. Becattini, A. Ferracani, L. Landucci, D. Pezzatini, T. Uricchio, and A. Del Bimbo, “Imaging novecento. a mobile app for automatic recognition of artworks and transfer of artistic styles,” in *Euro-Mediterranean Conference*. Springer, 2016, pp. 781–791.
- [8] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, “On the use of deep learning for blind image quality assessment,” *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.
- [9] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich *et al.*, “Experience grounds language,” *arXiv preprint arXiv:2004.10151*, 2020.

- [10] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, “Good news, everyone! context driven entity-aware captioning for news images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 466–12 475.
- [11] P. Bongini and et al., “Visual question answering for cultural heritage,” in *IOP Conference Series: Materials Science and Engineering*, vol. 949, no. 1. IOP Publishing, 2020.
- [12] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.
- [13] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, “A deep neural network for image quality assessment,” in *Proceedings of ICIP*. IEEE, 2016, pp. 3773–3777.
- [14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [15] L. Bulla, M. C. Frangipane, M. L. Mancinelli, L. Marinucci, M. Mongiovi, M. Porena, V. Presutti, and C. Veninata, “Developing and aligning a detailed controlled vocabulary for artwork,” in *European Conference on Advances in Databases and Information Systems*. Springer, 2022, pp. 529–541.
- [16] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, and C. Veninata, “Arco: The italian cultural heritage knowledge graph,” in *Proc. of ISWC, Part. II*, 2019, pp. 36–52.
- [17] H. Chang, M. K. Ng, and T. Zeng, “Reducing artifacts in JPEG decompression via a learned dictionary,” *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 718–728, Feb 2014.
- [18] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, “Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 655–12 663.
- [19] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft COCO captions: Data collection and evaluation server,” *arXiv preprint arXiv:1504.00325*, 2015.
- [20] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *European conference on computer vision*. Springer, 2020, pp. 104–120.

- [21] M. Cheng, Y. Sun, L. Wang, X. Zhu, K. Yao, J. Chen, G. Song, J. Han, J. Liu, E. Ding *et al.*, “Vista: Vision and scene text aggregation for cross-modal retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5184–5193.
- [22] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, “Perceptual image quality assessment with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 433–442.
- [23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [24] M. Cornia, L. Baraldi, and R. Cucchiara, “Show, control and tell: A framework for generating controllable and grounded captions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.
- [25] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-memory transformer for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.
- [26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proc. of CVPR. Ieee*, 2009, pp. 248–255.
- [27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of NAACL-HLT*, 2019, pp. 4171–4186.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [29] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, “An empirical study of training end-to-end vision-and-language transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.
- [30] X. Du, J. Shao, and C. Cardie, “Learning to ask: Neural question generation for reading comprehension,” *arXiv preprint arXiv:1705.00106*, 2017.
- [31] K. Elkins and J. Chun, “Can gpt-3 pass a writer s turing test?” *Journal of Cultural Analytics*, vol. 5, no. 2, p. 17212, 2020.
- [32] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” *arXiv preprint arXiv:1707.05612*, 2017.

- [33] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, “Deep universal generative adversarial compression artifact removal,” *Transactions on Multimedia*, 2019.
- [34] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, “Deep generative adversarial compression artifact removal,” in *Proc. of ICCV*, 2017. [Online]. Available: <https://arxiv.org/abs/1704.02518>
- [35] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, “Large-scale adversarial training for vision-and-language representation learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6616–6628, 2020.
- [36] N. Garcia and et al., “A dataset and baselines for visual question answering on art,” in *European Conference on Computer Vision*. Springer, 2020, pp. 92–108.
- [37] F. Gardères, M. Ziaefard, B. Abeloos, and F. Lecue, “Conceptbert: Concept-aware representation for visual question answering,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 489–498.
- [38] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [40] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.
- [41] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, “Kat: A knowledge augmented transformer for vision-and-language,” *arXiv preprint arXiv:2112.08614*, 2021.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [43] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, “Cross-modal retrieval via deep and bidirectional representation learning,” *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.
- [44] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [45] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, “Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

- [46] M. Ioannides, N. Magnenat-Thalmann, and G. Papagiannakis, *Mixed Reality and Gamification for Cultural Heritage*. Springer, 2017, vol. 2.
- [47] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.
- [48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.
- [49] *Rec. ITU-R BT.500-13 - Methodology for the subjective assessment of the quality of television pictures*, ITU, 2012.
- [50] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.
- [51] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, “Pipal: a large-scale image quality assessment dataset for perceptual image restoration,” in *European Conference on Computer Vision*. Springer, 2020, pp. 633–651.
- [52] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.
- [53] —, “Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks,” in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 2791–2795.
- [54] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, “Musiq: Multi-scale image quality transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.
- [55] D.-W. Kim, J. Ryun Chung, and S.-W. Jung, “Grdn: Grouped residual dense network for real image denoising and GAN-based real-world noise modeling,” in *Proc. of CVPR Workshops*, 2019.
- [56] J. Kim and S. Lee, “Deep learning of human visual sensitivity in image quality assessment framework,” in *Proc. of CVPR*, 2017, pp. 1676–1684.
- [57] —, “Fully deep blind image quality predictor,” *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2017.
- [58] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

- [59] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *arXiv preprint arXiv:1411.2539*, 2014.
- [60] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” in ., 2016. [Online]. Available: <https://arxiv.org/abs/1602.07332>
- [61] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [63] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [64] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.
- [65] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao *et al.*, “mplug: Effective and efficient vision-language learning by cross-modal skip-connections,” *arXiv preprint arXiv:2205.12005*, 2022.
- [66] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.
- [67] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4654–4662.
- [68] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, 2019.
- [69] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

- [70] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Text Summarization Branches Out*, 2004.
- [71] K.-Y. Lin and G. Wang, “Hallucinated-iqa: No-reference image quality assessment via adversarial learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 732–741.
- [72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proc. of ECCV*, 2014.
- [73] F. Liu, R. Ye, X. Wang, and S. Li, “Hal: Improved text-image matching by mitigating visual semantic hubs,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 563–11 571.
- [74] X. Liu, J. Van De Weijer, and A. D. Bagdanov, “Rankiqa: Learning from rankings for no-reference image quality assessment,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1040–1049.
- [75] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pre-training approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [76] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic violinguistic representations for vision-and-language tasks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [77] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, “Combining full-reference image visual quality metrics by neural network,” in *Proc. of Human Vision and Electronic Imaging XX*, vol. 9394. SPIE, 2015, pp. 172–183.
- [78] A. Mafla, R. S. Rezende, L. Gomez, D. Larlus, and D. Karatzas, “Stacmr: Scene-text aware cross-modal retrieval,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2220–2230.
- [79] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, and A. Bielefeldt, “Getting the most out of wikidata: semantic technology usage in wikipedia’s knowledge graph,” in *International Semantic Web Conference*. Springer, 2018, pp. 376–394.
- [80] F. Marni, M. Bertini, L. Galteri, and A. Del Bimbo, “A nogan approach for image and video restoration and compression artifact removal,” in *Proc. of ICPR*, 2021, pp. 9326–9332.
- [81] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, “Ok-vqa: A visual question answering benchmark requiring external knowledge,” in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

- [82] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 19–27.
- [83] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.
- [84] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [85] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [86] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.
- [87] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016, <http://distill.pub/2016/deconv-checkerboard>.
- [88] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.
- [89] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [90] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual dog model fused with random forest," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3282–3292, 2015.
- [91] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [92] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [93] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database tid2013: Peculiarities and preliminary results," in *European workshop on visual information processing (EUVIP)*. IEEE, 2013, pp. 106–111.

- [94] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “Pieapp: Perceptual image-error assessment through pairwise preference,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1808–1817.
- [95] V. Presutti and et al., “Pattern-based ontology design,” in *Ontology Engineering in a Networked World*, 2012, pp. 35–64.
- [96] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [97] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.
- [98] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.
- [99] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proc. of the EMNLP*, 2019.
- [100] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [101] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.
- [102] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [103] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.
- [104] —, “Principles of neurodynamics. perceptrons and the theory of brain mechanisms,” Cornell Aeronautical Lab Inc Buffalo NY, Tech. Rep., 1961.
- [105] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.
- [106] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the dct domain,” *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3339–3352, 2012.

- [107] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [108] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
- [109] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “Live image quality assessment database,” <http://live.ece.utexas.edu/research/quality>.
- [110] —, “LIVE Image Quality Assessment Database Release 2,” Apr. 2014.
- [111] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.
- [112] H. Sheikh, “Live image quality assessment database release 2,” <http://live.ece.utexas.edu/research/quality>, 2005.
- [113] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, “Improving image captioning with better use of captions,” *arXiv preprint arXiv:2006.11807*, 2020.
- [114] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, “Textcaps: a dataset for image captioning with reading comprehension,” in *European conference on computer vision*. Springer, 2020, pp. 742–758.
- [115] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [116] Y. Song and M. Soleymani, “Polysemous visual-semantic embedding for cross-modal retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1979–1988.
- [117] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [118] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara, “Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain,” in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 729–740.
- [119] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [120] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” *arXiv preprint arXiv:1908.07490*, 2019.

- [121] C. Thomas and A. Kovashka, “Predicting the politics of an image using webly supervised data,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [122] —, “Preserving semantic neighborhoods for robust cross-modal retrieval,” in *European Conference on Computer Vision*. Springer, 2020, pp. 317–335.
- [123] —, “Emphasizing complementary samples for non-literal cross-modal retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4632–4641.
- [124] H. Tomosada, T. Kudo, T. Fujisawa, and M. Ikehara, “Gan-based image deblurring using dct discriminator,” in *Proc. of ICPR*, 2021, pp. 3675–3681.
- [125] L. D. Tran, S. M. Nguyen, and M. Arai, “GAN-based noise model for denoising real images,” in *Proc. of ACCV*, 2020.
- [126] F. Vaccaro, M. Bertini, T. Uricchio, and A. Del Bimbo, “Fast video visual quality and resolution improvement using sr-unet,” in *Proc. of ACM MM*, 2021, pp. 1221–1229.
- [127] J. Van Ouwerkerk, “Image super-resolution survey,” *Image and vision Computing*, vol. 24, no. 10, pp. 1039–1052, 2006.
- [128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [129] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [130] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [131] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, “Deep metric learning with angular loss,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2593–2601.
- [132] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, “Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” *arXiv preprint arXiv:2202.03052*, 2022.
- [133] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in *Proc. of ICCV*, 2021, pp. 1905–1914.

- [134] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. of ECCV workshops*, 2018.
- [135] Z. Wang, A. C. Bovik, and L. Lu, “Why is image quality assessment so difficult?” in *2002 IEEE International conference on acoustics, speech, and signal processing*, vol. 4. IEEE, 2002, pp. IV–3313.
- [136] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “Simvlm: Simple visual language model pretraining with weak supervision,” *arXiv preprint arXiv:2108.10904*, 2021.
- [137] S. Winkler, “On the properties of subjective ratings in video quality experiments,” in *Proc. of QME*, 2009.
- [138] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, “Blind image quality assessment based on high order statistics aggregation,” *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.
- [139] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [140] B. Yan, B. Bare, and W. Tan, “Naturalness-aware deep no-reference image quality assessment,” *IEEE Transactions on Multimedia*, vol. PP, pp. 1–1, 03 2019.
- [141] F. Yan and K. Mikolajczyk, “Deep correlation for matching images and text,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3441–3450.
- [142] J. Yan, S. Lin, S. Bing Kang, and X. Tang, “A learning-to-rank approach for image color enhancement,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2987–2994.
- [143] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, “An empirical study of gpt-3 for few-shot knowledge-based vqa,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.
- [144] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.
- [145] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1098–1105.

-
- [146] J. You and J. Korhonen, “Transformer for image quality assessment,” in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1389–1393.
- [147] H. Zeng, L. Zhang, and A. C. Bovik, “A probabilistic quality representation approach to deep blind image quality prediction,” 2017.
- [148] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, “Deblurring by realistic blurring,” in *Proc. of CVPR*, 2020, pp. 2737–2746.
- [149] P. Zhang, W. Zhou, L. Wu, and H. Li, “Som: Semantic obviousness metric for image quality assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2394–2402.
- [150] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.
- [151] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.