Exploiting CLIP-based Multi-modal Approach for Artwork Classification and Retrieval

Alberto Baldrati[1,2], Marco Bertini[1], Tiberio Uricchio[1], and Alberto Del Bimbo[1]

[1] Università degli Studi di Firenze - MICC,
[name.surname]@unifi.it
Firenze, Italy
[2] Università di Pisa
Pisa, Italy

Abstract. Given the recent advances in multimodal image pretraining where visual models trained with semantically dense textual supervision tend to have better generalization capabilities than those trained using categorical attributes or through unsupervised techniques, in this work we investigate how recent CLIP model can be applied in several tasks in artwork domain. We perform exhaustive experiments on the NoisyArt dataset which is a dataset of artwork images crawled from public resources on the web. On such dataset CLIP achieves impressive results on (zero-shot) classification and promising results in both artwork-to-artwork and description-to-artwork domain.

Keywords: image retrieval, zero-shot classification, artwork, CLIP

1 Introduction

Image Classification and Content-Based Image Retrieval (CBIR) are fundamental tasks for many domains, and have been thoroughly studied by the multimedia and computer vision communities. In the cultural heritage domain, these tasks allow us to simplify the management of large collections of images, allowing us to annotate, search and explore them more easily and with lower costs.

In the latest years neural networks have proved to outperform engineered features in both tasks. These networks are typically used in an unimodal fashion, i.e. only one media is used to train and use a network. This may limit the types of application that can be developed and may also reduce the performance of the networks. Several recent works are showing how using multi-modal approaches may improve the performance in several tasks related to visual information. In [17] it has been shown that CLIP, a model trained using an image-caption objective alignment on a giant dataset consisting of over 400 million (image, text) pairs, obtains impressive results on several downstream tasks. The authors pointed out that, using only textual supervision, the CLIP model is able to carry out a broad set of tasks including geo-localization, action recognition, OCR and

many others. This task learning can be exploited via natural language prompting enabling zero-shot transfer capabilities over many existing dataset.

In this work we try to exploit the zero-shot capabilities of CLIP in the artworks domain, in particular we focus on the NoisyArt [2] dataset which is originally designed to support research on webly-supervised recognition of artworks and Zero-Shot Learning (ZSL). Webly-supervised learning is interesting since it allows to greatly reduce annotation costs required to train deep neural networks, thus allowing cultural institutions to train and develop deep learning methods while keeping their budgets for the curation of their collections rather than the curation of training datasets. In Zero-Shot Learning approaches visual classes are learned without any training examples, leveraging the alignment of semantic and visual information learned on some training dataset. ZSL in artwork recognition is a problem of instance recognition, unlike the other common ZSL problems that address class recognition. Zero-shot recognition is particularly appealing for cultural heritage and artwork recognition, although it is an extremely challenging problem, since it can be reasonably expected that museums have a set of curated descriptions paired with artworks in their collections.

To get a better idea of how CLIP behaves in the artworks domain we started with a classification task using a shallow classifier and CLIP as the backbone. Subsequently, thanks to the descriptions of the artworks in the dataset, we performed experiments in the field of zero-shot classification where CLIP was able to demonstrate its abilities in this task. Finally, we performed experiments on the tasks of artwork-to-artwork and description-to-artwork retrieval obtaining very promising results and superior performance to a ResNet-50 pre-trained on ImageNet [19]

2 Related Works

Regarding CBIR, after the successful introduction of BOW (Bag-of-Visual Words) in [23] that use engineered visual features such as SIFT points, several works have strengthened the performance tackling various aspects such as improving local features aggregation [15, 9, 4], learning improved codebooks [14], approximating local descriptors [8]. In recent years, after the successful use of Convolutional Neural Networks (CNN) to address the image classification problem [12], CNN-based features have started to be used also for image retrieval tasks. A complete survey that compares CNN-based and SIFT-based methods for instance-based image retrieval is presented in [28]. Commonly used backbone networks are VGG [22] and ResNet [7], typically pretrained on ImageNet and then fine tuned for a specific domain. CNN features have been pooled using techniques like Regional maximum activation of convolutions (R-MAC) [25]. R-MAC considers a set of fixed squared patches at different resolutions, collecting the peak response in every channel and then sum-pooling them to generate the final R-MAC descriptor. More recent works follow an end-to-end approach: in [1] a layer called NetVLAD was proposed, it is embeddable into any CNN architecture and trainable via back-propagation, which allows to train a network end-to-end using an aggrega-

tion of VGG16 convolutional activations. Multi-scale pooling of CNN features followed by NetVLAD has been proposed in [26], obtaining state-of-the-art results using VGG16. In [16] a trainable pooling layer named Generalized-Mean (GeM), along with learning whitening, has been proposed for short representations. In this work a dual-stream Siamese network is trained using contrastive loss. The authors employ up to five image resolutions in the feature extraction stage.

In [17] is presented CLIP (Contrastive Language-Image Pre-training): it is a visual model trained by natural language supervision with the simple task of predicting which caption is associated with which image, using a contrastive loss. With such a multimodal model, the authors attempt to close the gap between weakly-supervised methods where annotations are low quality but large in scale [13, 11], and current methods that attempt to learn visual representations using natural language supervision with a fairly restricted dataset [5, 20, 27]

Fig. 1 displays how the CLIP model is trained and how it can be easily used for zero-shot prediction.

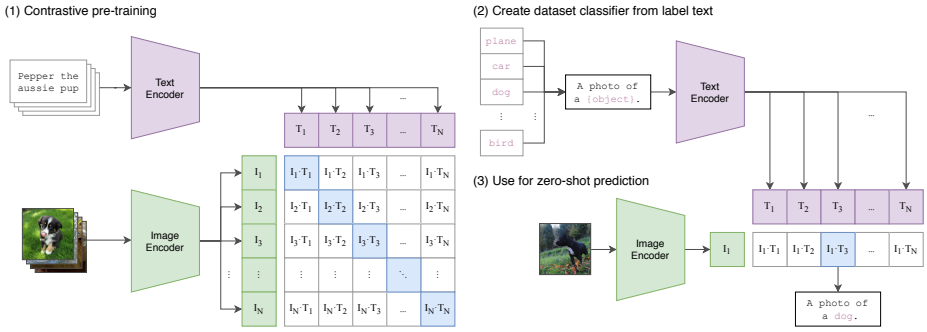

Fig. 1. Summary of CLIP approach. CLIP jointly trains an image and a text encoder to predict the correct pairings in a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. Figure and caption taken from [17]

3 Dataset

NoisyArt [2] is a dataset of artwork images gathered using structured queries to metadata repositories and web-based image search engines. According to the creators, the goal of NoisyArt is to support research on webly-supervised artwork recognition for cultural heritage applications.

In Table 1 the characteristics of the NoisyArt dataset are summarized

NoisyArt is a complex dataset which can be used for a large range of automated recognition problems. The dataset is particularly tailored to webly su-

		(webly images)		(verified images)
	classes	training	validation	test
	2,920	65,759	17,368	0
	200	4,715	1,253	1,355
totals	3,120	70,474	18,621	1,355

Table 1. Characteristics of the NoisyArt dataset

pervised instance recognition. In the dataset, for testing purposes, a subset of classes with manually verified test images is provided (*i.e.* the test set images does not contain label noise).

The NoisyArt dataset is collected from numerous public resources available on the web. These resources are DBpedia (where also the metadata are retrieved), Google Images and Flickr. Fig. 2 shows some examples of artworks with their respective sources.

From these sources the authors collected over 89K images divided into over 3K classes. Each class contains a minimum of 20 images and a maximum of 33. To be sure of having a reliable test set the authors decided to create a supervised test set using a small subset of the original classes: 200 classes containing more than 1,300 images taken from the web or from personal photos. This test set is not balanced: for some classes we have 2/3 images while for some others over ten. The different method of collecting training and test sets also raises the issue of a strong domain shift between these images and those in the training set. Finally, each artwork has a description and metadata retrieved from DBpedia, from which a single textual document was created for each class. These descriptions are included in the dataset to support research on zero-shot learning and other multi-modal approaches to learning over weakly supervised data.

3.1 GradCAM visualization

In order to have a better idea of the portions of the image that CLIP considers most important when it associates a text with an image, before moving on to the quantitative experiments, we carried out some qualitative tests using the well-know visualization technique gradCAM [21]. The technique we used is a generalization of gradCAM, where, instead of computing gradients with respect to an output class, gradients are computed with respect to textual features computed with CLIP's text encoder from the description. This approach makes each heat-map calculated by gradCAM dependent on the individual description, showing us the portions of the image that CLIP most closely associates with it. As a common practice the *saliency layer* used is the last convolutional layer of CLIP's visual encoder.

Fig. 3 shows four examples of gradCAM visualization. We can see how, using the descriptions in the dataset, CLIP places attention to the most significant portions of the image. This fact made us confident that CLIP would work very well in the domain of artwork.

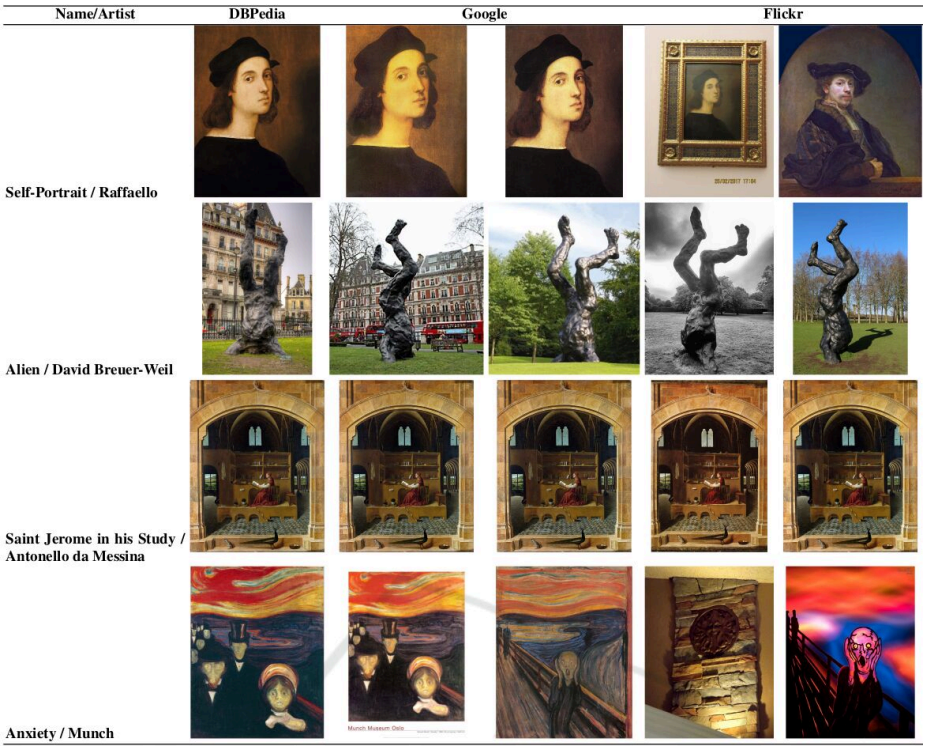

Fig. 2. Sample classes and training images from the NoisyArt dataset. For each artwork/artist pair we show the seed image obtained from DBpedia, the first two Google Image search results, and first two Flickr searches. Image taken from [2]

4 Experiments

4.1 Webly-supervised Classification

To test the performance of CLIP in the art domain, following the experimental setup followed by the authors of the dataset, we performed a webly-supervised classification on the 200 classes that are also available in the test set.

Experimental Setup Given an input image $\mathbf{x}$, we extract a feature vector using only the CLIP image encoder and then we pass it through a shallow classifier, consisting of a single hidden layer and an output layer that estimates class probabilities $p(c \mid \mathbf{x})$. The hidden layer is followed by an L^2-normalization layer which, as noted in [3], helps to create similar representations for images with different visual characteristics because the magnitude of features is ignored by the final classification layer. Such normalization is therefore useful to alleviate the effects of the domain shift between training and test set.

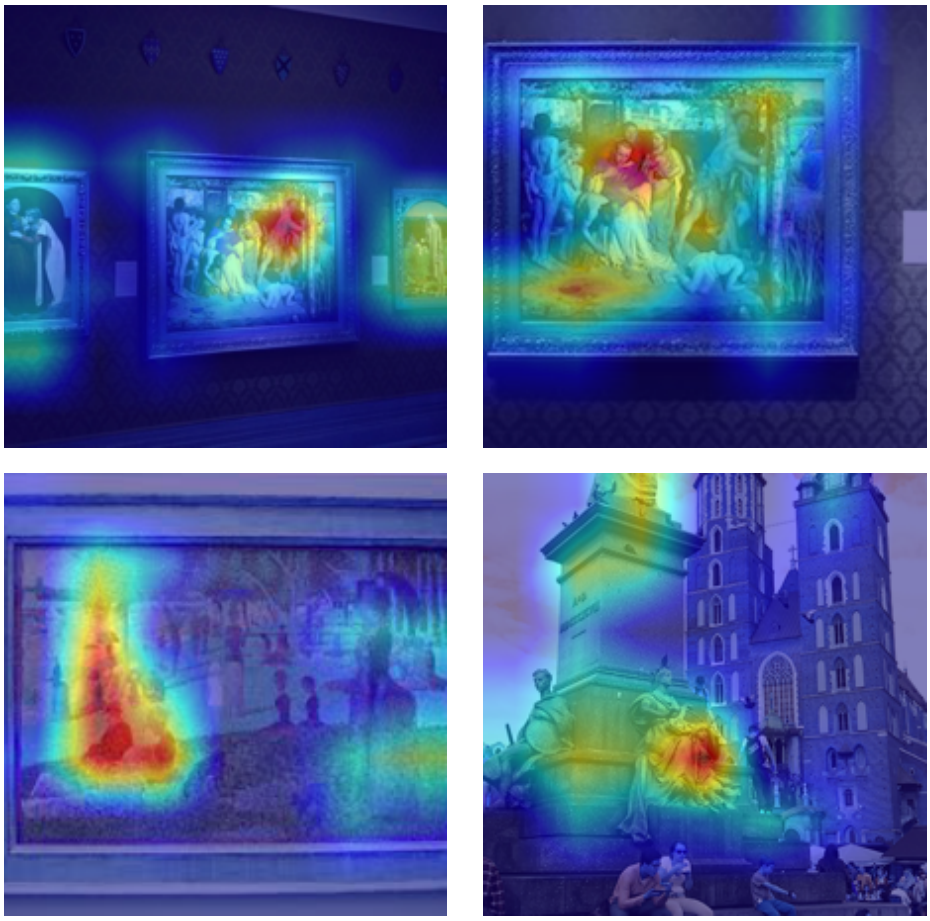

Fig. 3. Examples of gradCAM visualization on NoisyArt computing the gradients with respect to the description CLIP text features

The structure of the shallow classifier is basically the same of [2, 10]. This choice was made intentionally to analyze the effects of using the CLIP image encoder instead of a convolutional backbone trained on ImageNet. For mitigating and identifying label noise during training in [2] several techniques like Labelflip noise, entropy scaling for Outlier Mitigation and Gradual Bootstrapping are used. In our experiments however, following [3], we only use the L^2-normalization layer after the hidden layer.

We trained such a shallow classifier for 300 epochs with a batch size of 64, the learning rate used was $1e-4$. We used the CLIP model which has as a convolutional backbone a slightly modified version of the ResNet-50. The hidden layer has an input dimension of 1024 (CLIP output dimension) and an output dimension of 4096.

Model	test		validation	
	acc	**mAP**	**acc**	**mAP**
RN50 BL [2]	64.80	51.69	76.14	63.08
RN50 BS [2]	68.27	57.44	75.98	62.83
RN50 $\alpha = 0.4$ [3]	74.89	62.86	77.14	63.71
CLIP RN50	**86.63**	**77.88**	**83.56**	**72.23**

Table 2. Recognition accuracy (acc) and mean Average Precision (mAP) on NoisyArt dataset

Experimental Results Table 2 summarizes the experimental results we obtained in this classification setting. In the table *BL* refers to the baseline network [2] without any sort of label mitigating approach, *BS* refers to the noisy mitigating approach of [2] and *RN50* $\alpha = 0.4$ refers to the normalization approach of [3] where the L^2-normalization is scaled by α.

From the table it is immediately evident that with the use of CLIP as a backbone it is possible to obtain very significant improvements both on the test and the validation set. It is very interesting to see that [2, 3] have better results on validation than on the test set. In our case, however, the situation is reversed by having comparable and slightly better results on the test set. This demonstrates how CLIP is quite robust to domain shift being able to extract the semantic of an image regardless of its raw content.

4.2 Zero-shot Classification

The availability of descriptions associated with artwork made it possible to perform experiments in the area of zero-shot classification by exploiting CLIP's ability to assign a similarity score between text and images.

Experimental Results We are going to compare our results with the following baselines:

- **DEVISE [6]**: it learns a linear mapping between image and semantic space using a ranking loss.
- **EsZSL RN50 [18]**: it uses a square loss to learn the bilinear compatibility and regularizes the objective w.r.t the Frobenius norm.
- **COS+NLL+L2 RN50 [3]**: zero shot technique which integrates a non-linear extension of CMT [24] with an additional negative log-likelihood loss and an L_2 normalization step

Table 3 shows the immense potential in the zero-shot classification domain of CLIP. As a matter of fact, comparing the results with those found in the literature, we notice that by using CLIP, improvements of over 20% can be achieved. It is also worth noting that the results we have compared ours with have been achieved through a training process that uses a three-fold cross validation

Model	acc	mAP
DEVISE RN50 [6]	24.79	31.90
EsZSL RN50 [18]	25.63	29.89
COS+NLL+L2 RN50 [3]	34.93	45.53
CLIP RN50	**60.27**	**69.23**

Table 3. Zero-shot recognition accuracy (acc) and mean Average Precision (mAP) on NoisyArt dataset

where the 200 verified classes are divided into 150 for training/validation and 50 for zero-shot test classes. On our side we used CLIP out-of-the-box without any training on NoisyArt dataset.

In order to make a complete argument, it is also necessary to mention that since the data on which CLIP was actually trained is not public, we do not know if any images from this dataset were used in its training process. If so we would have some sort of leak of information that would make the comparison less fair.

4.3 Image Retrieval

Seeing the excellent behavior of CLIP in the (zero-shot) classification of artwork, we decided to perform some experiments in image retrieval.

In all the experiments that we are going to present, the images contained in the validation set (1253 images belonging to the 200 verified classes) were used as queries, while those of the test set (1379 images of the same 200 classes) were used as index images.

Experimental Setup We have conducted numerous experiments to make sure that we have a complete idea of how CLIP performs in this task on the NoisyArt dataset. As in classification experiments the CLIP model which has as visual backbone a modified version of the ResNet-50 is used.

The most natural way to use CLIP in retrieval is obviously to use the output of the visual encoder as a global descriptor comparing only the visual features, that is exactly what we did initially as a first experiment.

To take advantage of the CLIP textual encoder and of its goodness in zero-shot classification, we then reinterpreted the image-to-image retrieval task as zero-shot classification followed by text-to-image retrieval. This reinterpretation was made possible by the description and the metadata associated with each class. Thus given a query, zero-shot classification of that image was performed as the first phase, by exploiting CLIP's ability to link images and texts. We therefore used CLIP to assign a similarity score to each possible (query image, artwork description) pair using the description with the highest score in the second phase.

The second phase consists in comparing the description chosen at the end of the first one with all the images in the dataset, assigning a similarity score to each possible (query description, index image). For a complete comparison in the

results we have also reported an experiment where the first part of classification is bypassed and the correct monument is always used in the text-to-image retrieval phase.

Another setup we experimented with consists in adding to the zero-shot classification followed by text-to-image retrieval experiment as a re-ranking phase where the first 100 retrieved images are re-ordered using the similarity of the visual features.

Finally, the CLIP network was fine-tuned for adapting to this task. The fine-tuning process was done by inserting a shallow classifier composed of two linear layers at the output of the visual encoder. The learning rate was set to $1e-7$ for the CLIP encoder (keeping the normalization layers frozen) and $1e-4$ for the shallow classifier. For ease of use, a classification loss (categorical cross-entropy) was used during this fine-tuning process. We fine-tuned the model for 30 epochs using the 2,920 classes not included in the test set.

Experimental Results Before commenting on the results obtained we summarize the experimental setups:

- **RN50 image features**: We compare the image features extracted with a ResNet-50 pretrained on ImageNet
- **CLIP image features**: We compare the image features extracted with the CLIP image encoder
- **CLIP class + text-to-image**: We perform a zero-shot classification of the query followed by a text-to-image retrieval using CLIP text and visual encoder
- **CLIP class + text-to-image + visual re-ranking**: We perform a visual re-ranking of the first 100 retrieved results after CLIP zero-shot classification and text-to-image retrieval
- **Oracle + CLIP text-to-image**: We perform only the text-to-image retrieval using the ground-truth class for the description
- **CLIP fine-tuned image features**: We compare the image features extracted with the CLIP image encoder after fine-tuning

Experimental Setup	mAP
RN50 image features	36.32
CLIP image features	46.40
CLIP class + text-to-image	40.54
CLIP class + text-to-image + visual re-ranking	47.41
Oracle + CLIP text-to-image	54.21
CLIP fine-tuned image features	**69.60**

Table 4. Retrieval results on NoisyArt dataset using as queries the validation set and as index images the test set.

Table 4 summarizes the results of the experiments performed in the image retrieval setting previously described. It can be seen that CLIP visual features perform better than features extracted with a ResNet-50 pre-trained on ImageNet. It is interesting to note that the re-ranking process makes the retrieval process performed by a zero-shot classification followed by a text-to-image retrieval operation more performing than the approach which uses only the visual features pre fine-tuning. This fact is obviously made possible by CLIP's good results in zero-shot classification illustrated in previous section. It is also worth mentioning that using the ground-truth class and performing the text-to-image retrieval operation yields surprisingly good results: this confirms the goodness of CLIP in the text-to-image retrieval task. These results are even greater, by a significant margin, than those obtained using only visual features pre fine-tuning. This is probably due to the domain shift between validation and test set the visual features are more subject to. Finally, we can see that CLIP fine-tuning was very successful, bringing a very significant performance boost and achieving better results than all other approaches.

5 Conclusions

In this paper we propose to use the zero-shot capabilities of CLIP in the artworks domain, showing how this approach can greatly improve over competing state-of-the-art approaches in the challenging NoisyArt dataset. Experiments show that in addition to zero-shot classification, the proposed approach can be used for content-based image retrieval, again outperforming by a large margin other competing approaches. A benefit of using the proposed method is that it can be trained using very small datasets, thanks to the extensive pretraining of CLIP, and thus the method can be deployed also to be used on relatively small collections like those of small and medium-sized museums.

Acknowledgments This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

Bibliography

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. of CVPR*, 2016.

[2] R. Del Chiaro, A. D. Bagdanov, and A. Del Bimbo. Noisyart: A dataset for webly-supervised artwork recognition. In *VISIGRAPP (4: VISAPP)*, pages 467–475, 2019.

[3] R. Del Chiaro, A. D. Bagdanov, and A. Del Bimbo. Webly-supervised zero-shot learning for artwork instance recognition. *Pattern Recognition Letters*, 128:420–426, 2019. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2019.09.027. URL https://www.sciencedirect.com/science/article/pii/S0167865519302739.

[4] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez. Revisiting the VLAD image representation. In *Proc. of ACM MM*, 2013.

[5] K. Desai and J. Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021.

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[8] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010. doi: 10.1007/s11263-009-0285-2. URL https://doi.org/10.1007/s11263-009-0285-2.

[9] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, Sep. 2012. ISSN 1939-3539. doi: 10.1109/TPAMI.2011.235.

[10] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features, 2016.

[11] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning, 2020.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*, 2012.

[13] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining, 2018.

[14] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning vocabularies over a fine quantization. *International Journal of Computer Vision*, 103(1): 163–175, 2013.

[15] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. of ECCV*, 2010.

[16] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[18] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/romera-paredes15.html.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[20] M. B. Sariyildiz, J. Perez, and D. Larlus. Learning visual representations with caption annotations, 2020.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336359, Oct 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proc. of ICCV*, Oct 2003. doi: 10.1109/ICCV.2003.1238663.

[24] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

[25] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Proc. of ICLR*, 2016.

[26] F. Vaccaro, M. Bertini, T. Uricchio, and A. Del Bimbo. Image retrieval using multi-scale CNN features pooling, 2020.

[27] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2020.

[28] L. Zheng, Y. Yang, and Q. Tian. Sift meets cnn: A decade survey of instance retrieval, 2017.