

CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification

Marcos V. Conde
Universidad de Valladolid
drmarcosv@protonmail.com

Kerem Turgutlu
Adobe
keremturgutlu@gmail.com

Abstract

Existing computer vision research in artwork struggles with artwork’s fine-grained attributes recognition and lack of curated annotated datasets due to their costly creation. In this work, we use CLIP (Contrastive Language-Image Pre-Training) [12] for training a neural network on a variety of art images and text pairs, being able to learn directly from raw descriptions about images, or if available, curated labels. Model’s zero-shot capability allows predicting the most relevant natural language description for a given image, without directly optimizing for the task. Our approach aims to solve 2 challenges: instance retrieval and fine-grained artwork attribute recognition. We use the iMet Dataset [20], which we consider the largest annotated artwork dataset. Our code and models will be available at https://github.com/KeremTurgutlu/clip_art

1. Introduction

How to tell in which culture a sculpture was made? There are hundreds of possibilities: Greek, Roman, Arabic, and more. Fine-Grained Visual Classification (FGVC) aims to classify the sub-categories under coarse-grained large categories, such as the author of a painting, material of a sculpture, country of origin of an instrument. FGVC is challenging because objects that belong to different categories might have similar characteristics, but differences between subcategories might be remarkable (small interclass variations and large intra-class variations). Because of these reasons, it is hard to obtain accurate classification results using classical Convolutional Neural Networks [8, 4, 14, 13].

Recent work [9, 2, 21] shows the key step of FGVC is identifying and extracting more informative regions and features in an image. However, labeling fine-grained categories is an expensive and time-consuming process which often requires expertise in a specialized domain, thus, FGVC datasets [7, 17, 11] have limited training data.

For this reason, research focuses on weakly-supervised learning using noisy labels, and unsupervised learning scheme to recognize informative regions [6, 18, 23].

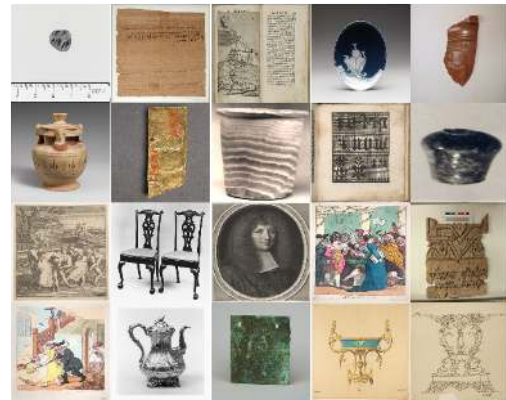


Figure 1. The artworks in iMet [20] include paintings, instruments, prints, clothing, sculpture, furniture, metalwork, etc.

Our main contributions are:

- A general Contrastive Pre-training [12] framework for fine-grained visual-textual representation learning [5] by using natural language free-form descriptions of artwork and images.
- A solution based on interpretable Visual Transformers and self-attention [3].
- Our task-agnostic model performs zero-shot fine-grained classification, and achieves better results than few-shot supervised SOTA models [4, 15].
- A multimodal representation learning for classification and image-text retrieval.

1.1. Dataset and Benchmark

The iMet Collection Dataset [20] from The Metropolitan Museum of Art in New York (The Met), presents the largest fine-grained artwork collection. Some samples are shown in Figure 1. Each image is labeled with its associated artistic attributes. The attributes can relate to what one “sees” in the work or what one infers as the object’s “utility”.



Figure 2. Image and noisy fine-grained categories.

Figure 2 shows images and their attributes description. These are grouped into 5 parent classes: country, culture, dimension, medium, tags. In total, there are 3471 unique attributes. Research-grade Museum experts curated and verified attribute labels to ensure high quality. However, each object is annotated by a single annotator without a verification step, and sometimes they added free-form text descriptions. For this reason, the authors recommend considering attributes as **noisy** labels. iMet hosts a yearly competition since 2019, providing a public benchmark based on more than 40,000 unknown test images.

2. Approach

Our approach consists of multiple stages which can be seen in Figure 3; free-form text generation, contrastive pre-training and finally fine-tuning on the downstream fine-grained art recognition task.

First, we convert noisy fine-grained categorical annotations into natural language text for a given image. We achieve this by using natural language templates and using different permutations. This process is similar to data augmentation but for text descriptions.

We use different combinations of attribute values when there might be an image with multiple attribute values for a given category, such as multiple tags which can describe different things in an art object. At the end of this data generation process, we end up having more than 15 text descriptions per image in the iMet dataset.

Second, we fine-tune ViT-B/32 CLIP model which is open-sourced by OpenAI. This model uses 2 transformer encoders for jointly embedding the text and image pairs; a ViT-B/32 for image encoding and another 12-layer transformer for text encoding. Similar to the original CLIP model [12] we minimize InfoNCE loss [16] during contrastive pre-training. In a given batch, each image-text pair or text-image pair forms a positive sample and every other image or text is considered negative. Having this symmetry we calculate pairwise cosine similarity between L_2 -normalized image-text embeddings and calculate cross-entropy loss with a learnable temperature parameter. In our synthetic dataset, a given image has multiple text descriptions, for that reason we randomly sample one text with equal probability during training. This can be viewed as data augmentation. Additionally, we apply dropout to attribute values if there are multiple values for a given category to further diversify these augmentations. For the remainder of this paper, we refer to OpenAI ViT-B/32 model as $CLIP_{base}$ and our fine-tuned version as $CLIP_{art}$.

Finally, we use the domain adapted $CLIP_{art}$ for further fine-tuning on the downstream fine-grained art recognition task.

2.1. Contrastive Pre-training

In our experiments, contrastive pre-training shows the following advantages: it can leverage free-form text to learn more generalized and robust visual features even in the pres-

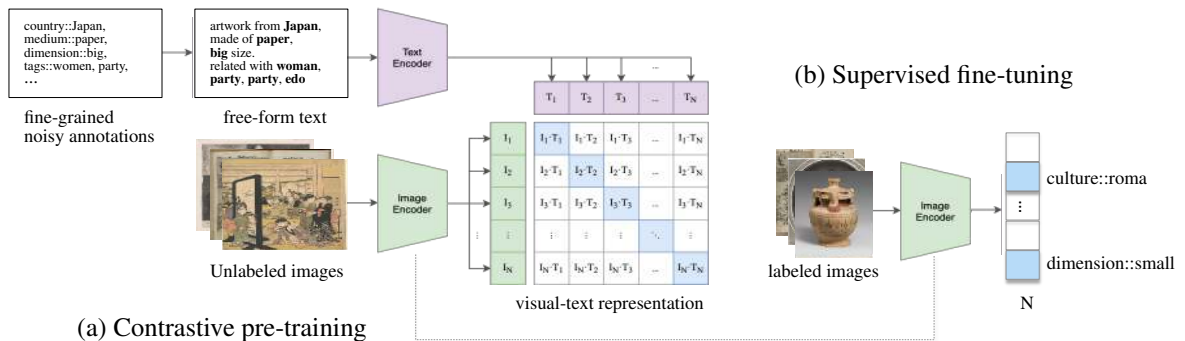


Figure 3. Summary of our approach based on CLIP from OpenAI [12]. We show (a) Contrastive pre-training using unlabeled images (or noisy annotated). We process noisy or scrapped annotations into natural language free-form descriptions as explained in Section 2. Using a task-agnostic image encoder and text encoder, we learn a visual-textual representation, discovering discriminative visual-textual pairwise information [5]. Further supervised fine-tuning (b) can be done using small labeled datasets.

ence of noise, it allows faster and better convergence for the downstream task at hand, it can be used for retrieval with any natural language query at inference time beyond the closed set of predefined labels and different loss functions from any state-of-the-art self-supervised learning method can be used [1, 19] during training.

We fine-tuned models using Ranger optimizer, a combination of Lookehead and RAdam [10, 22].

No image data augmentation is used besides random resized cropping and horizontal flip. All contrastive pre-training models are trained for 20 epochs. To test our hypothesis that free-form text descriptions help to learn good fine-grained representations, we fine-tuned CLIP_{art} model with 2 different versions of text pairs; one which is generated using all 5 categories and another which does not include **tags** category. Later retrieval performance for these 2 versions are reported in Table 2.

2.2. Fine-tuning

For all downstream fine-tuning experiments same setup; image size, data augmentation, MLP layers, and learning rate schedulers are used for a fair comparison. We treated the downstream fine-grained art attribute recognition as a multilabel classification task where each attribute is assumed to be independent and an image can be assigned multiple attributes as can be seen from Figure 3. For the first 5 epochs encoder weights are frozen and for the remaining 15 epochs all model weights are updated.

3. Experiments

In this section, we describe our experimental setup and results at fine-grained classification and artwork retrieval. All CLIP-based models have as backbone a Visual Transformer (ViT) [3]. We conducted experiments for assessing the zero-shot, few-shot and fully supervised performance of variety of models including CLIP_{base} and CLIP_{art}.

3.1. Zero-Shot Experiments

Using visual encoder, ViT-B/32, of CLIP_{base} and CLIP_{art} models we extracted image representations of 512-dimension for the full iMet 2020 training set, which consists of a total of 142,119 images. Later, we predicted on a 20K hold-out set using a query image and assigning the labels from the nearest neighbor in the training set.

3.2. Multimodal Retrieval Experiments

In order to test our hypothesis that rich free-form text helps with learning better representations we train 2 versions of CLIP_{art} using 2 different datasets with different text descriptions. We evaluate different versions of CLIP_{art} with several retrieval metrics after encoding all the 20k validation images and their corresponding text descriptions. Once embedded we calculate normalized pairwise cosine similarity between all the image and text embeddings. Using this similarity matrix we report results in Table 2 on retrieval percentage at 5, retrieval percentage at 20, mean retrieval rank, and median retrieval rank.

3.3. Few-Shot and Fully Supervised

In few-shot experiments, we trained models with a fraction of data to compare against the zero-shot performance. In fully supervised experiments we used full training data and compared a variety of models including ViT-B/32 from CLIP_{art}, ViT-B/32 from CLIP_{base}, ViT-B/32 pre-trained on ImageNet. As well as a variety of ResNets and EfficientNets for benchmarking.

3.4. Results

We report results for zero-shot, few-shot and fully supervised training using the iMet dataset [20]. We calculate F2-score metric to provide some robustness against noisy labels, favoring recall over precision. We use the validation consisting of 20K images.



Figure 4. Attention map samples. For each sample we show (left) image, (middle) attention map from CLIP-Art (right) attention map from CLIP-Base. Our contrastive learning of visual-text features helps to discriminate better the most discriminative regions in the image.

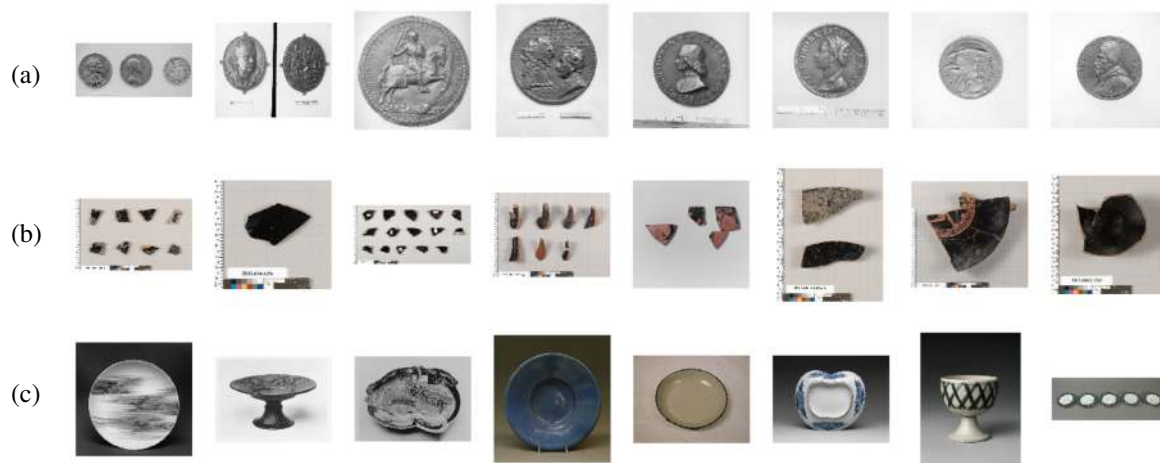


Figure 5. Multimodal Retrieval results. Each row is the result of the following text queries: (a) “italian, rome artwork with portraits, profiles, men, popes made from bronze”, (b) “art from greek, attic made of terracotta” and (c) “small japan artwork made from matsugatani type with dishes”. Note that these images are completely unknown for our model. We obtain these results as explained at Section 3.2.

In zero-shot benchmarks we used the KNN approach explained in Section 3.1. For few-shot benchmarks, only 10% and 20% random subset of the training data is used for training classification SOTA CNNs.

Table 1 shows that CLIP_{art}, a task-agnostic model, without any supervision outperforms ResNet-50 [4] and EfficientNet-B0 [15], both SOTA classification models trained with a fraction of the complete dataset and optimized for the task. Moreover, at a complicated fine-grained categorization task. We show that a simple fine-tuned vision transformer can achieve results as full-supervised CNNs. Furthermore, our CLIP_{art} ViT achieves better results than ViT pre-trained on ImageNet, and in half number of epochs. Note that the idea of this work is to explore multi-modality and image-text representations, for this reason, we do not use complex models, ensembles, aggressive augmentations, etc, as many solutions for this benchmark propose.

More information at the supplemental material.

Method	Backbone	Data (%)	F2 score
CLIP _{base} [12]	ViT [3]	0	0.5161
CLIP _{art} (ours)	ViT [3]	0	0.5507
ResNet [4]	ResNet-50	10	0.5210
ResNet [4]	ResNet-50	20	0.541
EfficientNet [15]	EffNet-B0	10	0.511
EfficientNet [15]	EffNet-B0	20	0.550
CLIP _{art} (ours)	ViT [3]	100	0.60
ResNet [4]	ResNet-50	100	0.615
ViT _{imet} [3]	ViT [3]	100	0.58

Table 1. Ablation study of the proposed methods. Data 0% corresponds to zero-shot experiments, 10-20% corresponds to few-shot and 100% corresponds to completely supervised.

Dataset	ret5	ret20	mean ret	median ret
All Categories	0.3052	0.5467	175.84	16
All (no “Tags”)	0.1658	0.3578	353.99	48

Table 2. Retrieval Results. We tested that removing a highly descriptive category such as **tags** hurts retrieval performance and supports that representations when learned conditionally on descriptive text help with fine-grained retrieval.

We show our multi-modality capability at Table 2. Our model is able to get the correct complete text pair for a given query image within the first 20 ranked predictions out of 20,000 candidates for the 54% of the time. Table 2 shows that removing such descriptive text hurt the retrieval performance significantly. Qualitative results using images as queries can be found at the supplemental material.

4. Conclusion

To solve art-related computer vision main challenges, retrieval and fine-grained attribute recognition, we have presented an approach based on Contrastive Language-Image Pre-Training (CLIP) [12] using a wide variety of artwork images and natural language supervision. By its design, the network can be instructed in natural language to perform fine grained artwork retrieval and recognition in a zero-shot manner without directly optimizing for the iMet benchmark [20]. We also proposed a way for constructing natural language text from the available closed set of attribute labels by augmenting them. Future work can focus on building an artwork dataset consisting of more than 1 million image-text pairs scrapped from iMet’s database, which, together with this work, will represent a breakthrough in artwork classification and retrieval.

References

- [1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 3
- [2] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 1, 3, 4
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1, 4
- [5] Xiangteng He and Yuxin Peng. Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):520–531, Feb 2020. 1, 2
- [6] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, 2019. 1
- [7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 1
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. NIPS’12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 1
- [9] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1
- [10] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2020. 3
- [11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 1
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 4
- [13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 1
- [14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. 1
- [15] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 1, 4
- [16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2
- [17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1
- [18] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification, 2018. 1
- [19] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. 3
- [20] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler, and Serge Belongie. The imet collection 2019 challenge dataset, 2019. 1, 3, 4
- [21] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization, 2020. 1
- [22] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019. 3
- [23] H. Zheng, J. Fu, Z. Zha, and J. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5016, 2019. 1