Supplementary Material - 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

1. Training convergence

We compare ViT-B/32 [2] fine-tuning for fine-grained art classification using the weights from three different pre-training strategies:

1. base contrastive pretraining CLIP [6] on 400 million images, open-sourced by OpenAI.
2. our CLIP$_{Art}$ contrastive pretraining using artwork images and their natural language descriptions,
3. ImageNet pretraining [5].

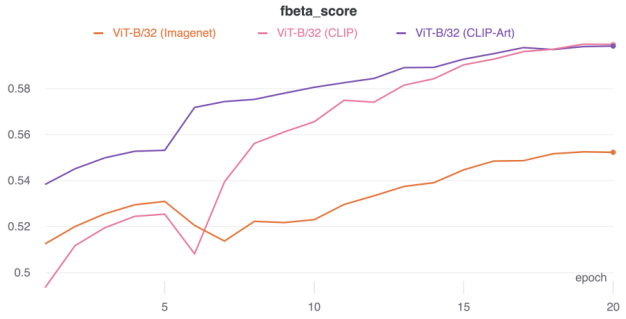


Figure 1. Convergence plot for the first stage of training. Our CLIP$_{Art}$ improves convergence and performance (F2-score). For fair comparisons, we train to convergence using the same training setup (loss, optimizer, etc.) and images in all experiments.

2. Large Scale Art Dataset

Supervised CNNs: We train more complex models for fine-grained art classification using 384 image size and exhaustive augmentations (random crops, horizontal and vertical flips, random erasing, mixup, etc.). These models represent our set of teachers, their results on iMet [9] are shown at Table 1. Each teacher model trained with labeled data will infer pseudo-labels on unlabeled artwork data, which can be scrapped from the internet. This serves as dataset for self-training / distillation of task predictions using smaller versions of these models as noisy students [8, 1].

Network	F2-score
SEResNext-50 [4]	0.701
EfficientNet-b7 [7]	**0.712**
ViT-L-16 [2]	0.707

Table 1. Supervised SOTA CNNs trained on iMet dataset [9] and evaluated on 2020 Benchmark https://www.kaggle.com/c/imet-2020-fgvc7

Scrapped images as Figure 2 include an extensive free-form description from an expert, these are involuntary transferences from human visual attention to textual attention, which implies that textual attention can help to discriminate significant parts or features for categorization [3].

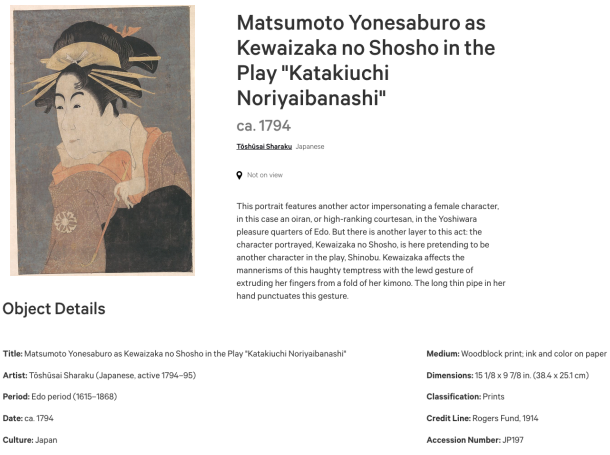


Figure 2. Scrapped image and its natural language description.

Using the teacher's predicted pseudo-labels and scrapped free-form descriptions from experts, we conform (image, text) pairs for learned visual attention from natural language supervision using CLIP [6]. In this way, we aim to build an artwork dataset consisting of more than 1 million (image-text) pairs, which, together with this work,

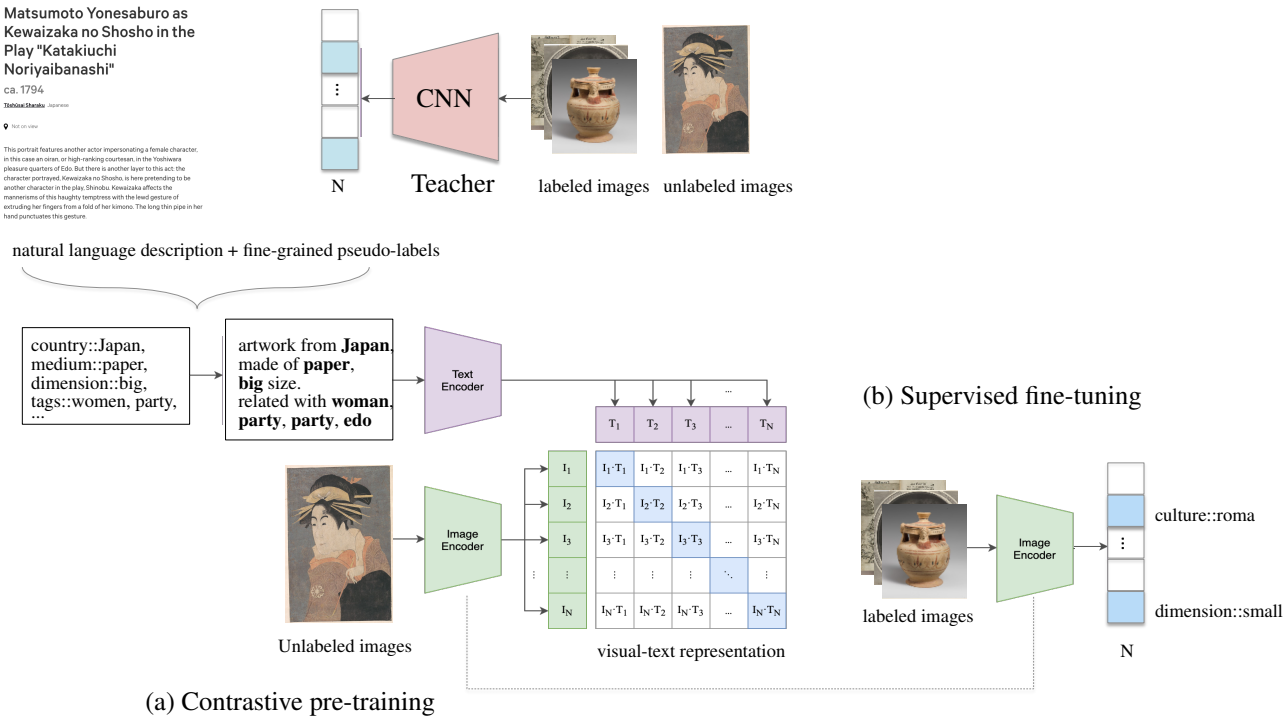


Figure 3. Summary of our semi-supervised approach based on CLIP from OpenAI [6]. We show our teacher networks trained on iMet [9] labeled data as explained in Section 2 and Table 1. Scrapped text and pseudo-labels inferred from unlabeled images are processed into free-form descriptions. We also show (a) Contrastive pre-training using unlabeled images and their noisy generated descriptions. Using a task-agnostic image encoder and text encoder, we learn a visual-textual representation, discovering discriminative visual-textual pairwise information [3]. Further supervised fine-tuning (b) can be done using labeled images.

will represent a breakthrough in artwork classification and retrieval. Figure 3 shows the explained approach.

References

[1] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 1

[2] 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] 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

[4] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. 1

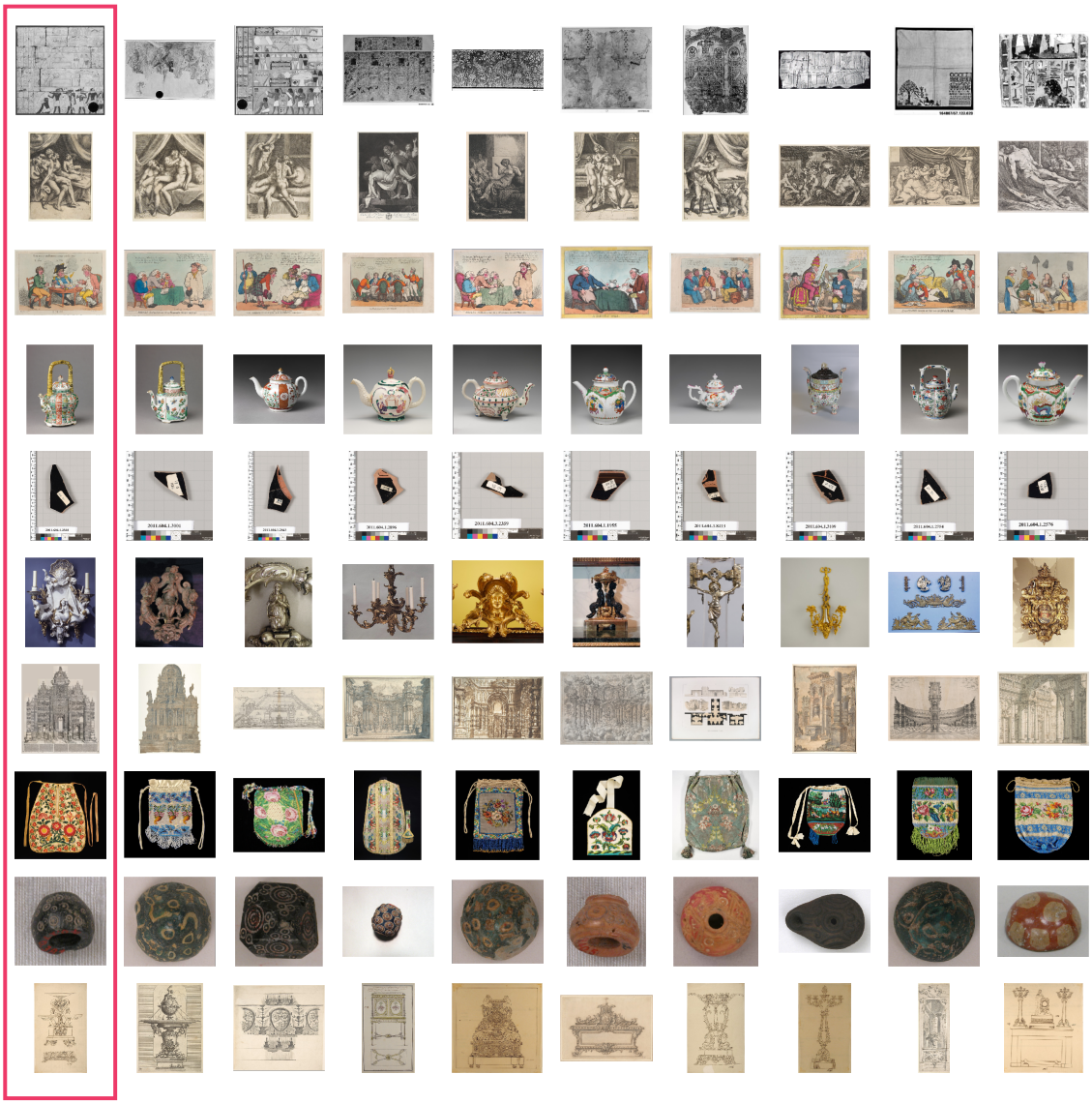
[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 1

[6] 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

[7] 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

[8] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. 1

[9] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler, and Serge Belongie. The imet collection 2019 challenge dataset, 2019. 1, 2



(a)

Figure 4. Results for artwork retrieval. We highlighted query images in column (a). For each query image we rank 20.000 validation candidates based on cosine similarity, resultant top-9 are shown in each row.