Few-Shot Classification on ImageNet-Sketch

Mathis WAUQUIEZ
Ecole des Ponts ParisTech
mathis.wauquiez@enpc.fr

Abstract

We compare methods for few-shot classification on ImageNet-Sketch, highlighting the effectiveness of K-Nearest Neighbors with self-supervised backbones. This approach, simpler than linear classifiers, achieves competitive results, demonstrating the potential of pre-trained models for challenging datasets

1. Methodology

We compared the different approaches:

- Using models pretrained on ImageNet, and training a linear classifier on top of it.
- Using models pretrained on ImageNet, and fine-tuning the whole model on ImageNet-Sketch.
- Using models pretrained on ImageNet as backbone for the self-supervised learning method SimCLR, and training a linear classifier on top of it.
- Using the same feature encoders, but using K-Nearest Neighbors instead of a linear classifier.
- Using large transformers available in the Hugging Face library, and using K-Nearest Neighbors on top of it.

For reproducibility, we created a pipeline that uses YAML files to keep track of the models, hyperparameters, augmentations, ... used for each experiment.

1.1. SimCLR training

SimCLR is a "Simple Framework for Contrastive Learning of Visual Representations" [3]. It is a self-supervised learning method that learns representations by maximizing the similarity between augmented views of the same image, and minimizing the similarity between augmented views of different images. We reproduced this training method using Pytorch, and used it to learn representations of the images in ImageNet-Sketch, for several backbones, using the SimCLR head proposed in the original paper:

- ResNet18 [5]
- MobileNetV3-Small [6]
- EfficientNet-B0 [8]
- EfficientNetV2-S [9]
- SqueezeNet [7]
- ViT-B/16 [4]

The common point between all these models is that they are relatively small, and can be trained on a single GTX 1050 Ti GPU with 4GB of VRAM, in only a few hours each. We then trained a linear classifier on top of the learned representations, using the same hyperparameters for all the models, and evaluated them on the ImageNet-Sketch validation set. Building on top of the analysis of the transforms analysis in SimCLR, we used their recommended augmentations for the training of the models.

1.2. Large transformers

We also used CLIP ([2]) and DINO ([1]) trasnformers, as well as a transformer pretrained on ImageNet, to extract features from the images in ImageNet-Sketch for K-Nearest Neighbors classification using cosine similarity.

2. Results

2.1. SimCLR training

Our findings show that:

- Using vision transformers and EfficientNets was impractical, because SimCLR requires large batch sizes.
- MobileNetV3 and ResNet18 got the best contrastive accuracy, botch reaching about 0.83, superior to the 0.67 of SqueezeNet.
- MobileNetV3 took only 3.3 hours to train, while ResNet18 took 7.5 hours.
- MobileNetV3 proved superior accuracy-wise to ResNet18 when using a linear classifier on top of the learned representations, reaching a top-1 accuracy of 0.406.

2.2. K-Nearest Neighbors

Our findings show that:

- Using ViT models trained on datasets bigger than ImageNet granted the best results, with a top-1 accuracy always greater than 0.85.
- The best model was DINO/ViT-H/14, with a top-1 accuracy using KNN with cosine similarity of 0.90250.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1

[2] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training, 2024. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1

[4] 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, 2021. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. 1

[7] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size, 2016. 1

[8] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 1

[9] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021. 1