Assignment 3 — Sketch Image Classification
MVA Object recognition and computer vision 2024-2025

Bryan Chen
École Polytechnique, Palaiseau, France
bryan.chen@polytechnique.edu

1. Introduction

Sketch image classification challenge on ImageNet-sketch [1] dataset was tackled using various deep learning models. Starting from EfficientNet [2], we progressively explored more advanced models like OpenAI's CLIP with different ViT architectures. We experimented with fine-tuning strategies by unfreezing layers incrementally, and trying multiple schedulers. To improve training efficiency, we precomputed image features and trained only the classifier. Finally, we achieved a validation accuracy of 93% using EVA-CLIP [3] with eva_giant_patch14_336.clip_ft_in1k. Throughout the experiments, we utilized Weights and Biases (wandb) for monitoring.

2. Methodology

In order to evaluate how hard the image classification is on our problem, we began with pre-trained EfficientNet, adding layers to fine-tune the model on the ImageNet-Sketch dataset. Recognizing the potential of foundational models for zero-shot image classification, which are trained on a substantial number of data, we transitioned to OpenAI's CLIP [4]. We experimented with different ViT sizes and fine-tuning strategies, like unfreezing only the last layer of CLIP's ViT and incrementally unfreezing up to four layers to assess performance gains.

To accelerate training, we precomputed image features and labels, saving them as `.pt` / `.pth` files. This allowed us to train only the classifier on these features, significantly reducing computation time.

Moreover, realizing that models like CLIP were state-of-the-art (SoTA) in 2021-2022, we investigated newer models (eg. EVA-CLIP, CoCA, CLIPA, LLaVA, LiT) discovered through PapersWithCode. Based on availability and performance benchmarks, we selected EVA-CLIP for further experimentation.

3. Experiments and results

For each model, we monitored training and validation performance using wandb. Our key findings include:

CLIP Fine-tuning

Unfreezing layers in CLIP's ViT did not show incremental improvements. Indeed, when the last layer is unfreezed, we obtained slight improvement, and up to four layers unfreezed, we got moderate gains but with increased training time. Therefore, due to the poor gain, we decided to freeze the entire foundation model and embed the training and validation images into image features, and train only the classifier on these features, where we achieved faster training times and an improvement in accuracy

EVA-CLIP with eva_giant_patch14_336.clip_ft_in1k

Using EVA-CLIP, we achieved a test accuracy of 93%, and trying with multiple schedulers (CosineAnnealingLR and OneCycleLR), and optimizers (Adam and AdamW) surpassing previous models. For the choice of the scheduler, the choice of OneCycleLR was shown to be experimentally better and for the optimizer, AdamW was shown to be better. Moreover, we try to add some weight_decay, but it seems that by removing this hyperparameter, we obtained better test accuracy, maybe due to the ”small” training dataset that we have.

Data augmentation

By using different data augmentation techniques such as HorizontalFlip, VerticalFlip, RandomRotation, GaussianBlur, RandomErasing, etc. we observe that these techniques do not guarantee necessarily a test accuracy improvement, in fact, we have close performances with and without these techniques.

4. Conclusion

Our exploration demonstrates that modern models like EVA-CLIP significantly enhance sketch image classification (93% test accuracy) with models like EfficientNet (61% test accuracy) and CLIP (92% test accuracy). Fine-tuning strategies and efficient training methods like precomputing features contribute to performance gains while optimizing computational resources.

References

[1] H. Wang and al., “Learning robust global representations by penalizing local predictive power,” 2019. arXiv preprint arXiv:1905.13549.

[2] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” 2019. arXiv preprint arXiv:1905.11946.

[3] X. Wang and al., “Eva-clip: Improved training techniques for clip at scale,” 2020. arXiv preprint arXiv:2011.09157.

[4] A. Radford and al., “Learning transferable visual models from natural language supervision,” 2021. arXiv preprint arXiv:2103.00020.