Object Recognition and Computer Vision - Assignment 3

Lucas Versini
Institut Polytechnique de Paris
lucas.versini01@gmail.com

Abstract

This challenge aims to tackle the problem of classifying 500 *types of sketches sampled from [6], with* 40 *training images and* 5 *validation images per class, and* $5,455$ *unlabeled images for evaluation.*
This task is particularly challenging because of the limited training data per class.

1. Introduction

In this project, we explored several strategies to address the challenges posed by sketch classification: the choice of the models and optimizers, data augmentation techniques, and ensemble methods.
We also thought of several other techniques, including Dropout, self-supervised learning, test-time strategies, or even stroke removal (to make the sketches easier for the model to interpret), but no improvement was observed.

2. Model, optimizer, scheduler

We tested, among other models, EVA-02 [2], DeiT [5], ViT [1], as well as convolutional models.
We used different optimizers: for DeiT and ViT, Stochastic Gradient Descent (SGD) for a few epochs, followed by Adam [4] for the next epochs led to better results on the validation set, while for EVA-02, it was simply SGD. For all models, we used an exponential scheduler with $\gamma = 0.9$.
After training the models with different configurations, we decided to discard ViT and the convolutional models, and to keep both EVA-02 and DeiT.
Note that because of the different number of parameters ($304,567,732$ for EVA-02, $87,657,716$ for DeiT), we had to use different batch sizes (between 32 and 64 for DeiT, up to 8 for EVA-02).

3. Data augmentation

In order to reduce overfitting (especially since the number of training images per class is limited), we decided to use data augmentation techniques.
We used some basic transformations (cropping, rotations, flips...), as well as more advanced techniques such as MixUp [8], CutMix [7] and AugMix [3] (though this last one is usually used with colored images).
Using basic transformations typically led to a higher accuracy for DeiT, and to a lower accuracy for EVA-02, whereas both AugMix and CutMix improved the accuracy of both these models, and were therefore kept for the rest of the experiments.

4. Ensemble methods

We used ensemble methods to combine different models.
Specifically, we considered three methods:

- Average the probability distributions from all models, and predict the class with highest probability.
- Predict the class with the highest probability among all models.
- Predict the class which is predicted by the maximum number of models. In case of equality, the class with highest estimated probability is predicted.

The method that yielded the best results was the average, see Table 1 for some results.
In the end, we kept one EVA-02 model with no data augmentation technique, a DeiT model with CutMix, and a DeiT model with AugMix.

5. Conclusion

In conclusion, we are satisfied of the obtained results.
Through careful selection of models, optimization techniques, and data augmentation strategies, we achieved satisfying classification performance, despite material limitations.

Model	DeiT (CutMix)	EVA-02	One EVA-02 model + DeiT (CutMix) + DeiT (AugMix)
Validation set	0.902	0.9236	0.948
Kaggle public score	0.90521	0.93026	0.95260

Table 1. Scores of some models

References

[1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[3] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[4] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[6] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[8] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.