Assignment 3: Sketch image classification competition

Nayoung KWON

1. Introduction

This assignment involves classifying images from the ImageNet-Sketch dataset [7], a subset of the ImageNet [1]. Unlike the natural and high-resolution images of ImageNet, ImageNet-Sketch consists of minimalistic line drawings. By removing visual details such as color, texture and complex backgrounds, this dataset presents a challenge for classification models, making them to focus on essential structural and shape-based features. This report outlines the methods explored to achieve robust performance on this dataset.

2. Dataset

The dataset used in this competition is derived from ImageNet-Sketch, which originally contains 50,000 sketch images of 1,000 object categories. For this competition, the dataset has been modified to include images representing 500 object categories. The training set consists of 20,000 images, the validation set contains 2,500 images, and the test set includes 5,455 images. A manual inspection of the dataset revealed that the sketches have a wide range of styles, including cartoon-like and realistic representation, variations in shading intensity, and differences in line thickness. Despite this variability, these stylistic features appear to be well-balanced across the dataset.

3. Data augmentation

To enhance the performance of the models, I implemented a data augmentation process using *AugMix* [5]. This technique generates augmented versions of input images by combining multiple transformations, ensuring diversity while maintaining the semantic of the images. Specifically, I applied transformations such as rotation, horizontal and vertical flipping, adjustments to brightness, contrast, saturation, and hue,conversion to grayscale and more. These augmentations introduce variability into the training data, helping the models generalize better to unseen samples.

4. Models

I experimented with five different models for this task:

- ResNet152 [4]
- DeiT [6] pretrained on ImageNet-22k and fine-tuned on ImageNet-1k
- VisionTransformer [2] pre-trained on ImageNet-21k
- EVA02 [3] pre-trained on ImageNet-22k with masked image modeling and fine-tuned on ImageNet-22k then on ImageNet-1k

For training, I experimented with various hyperparameter configurations, including learning rate, batch size, and the number of epochs. After several trials, I determined that a learning rate of 0.001, 20 epochs, and a batch size of 16 provided optimal performance across most models. While the majority of models converged around 20 epochs, training for the EVA02 model was limited to 10 epochs due to the significant computational time required for each epoch. For optimization, I utilized the Stochastic Gradient Descent (SGD) optimizer, given its potential for better generalization on medium-sized datasets like ImageNet-Sketch.

5. Results

Model	Validation set	Test set
ResNet152	0.838	0.83615
DeiT	0.9012	0.89708
ViT	0.8912	0.91469
EVA02	0.916	0.92484

The EVA02 model demonstrated the best performance, achieving an accuracy of 0.92417. This result can be attributed to its architecture, which utilize transformers with SwiGLU activation functions, Rotary Position Embeddings (ROPE), and mean pooling mechanisms. Additionally, the use of masked image modeling during pre-training and intensive fine-tuning on both ImageNet-22k and ImageNet-1k could have contributed to its ability to generalize effectively to the ImageNet Sketch dataset.

6. Discussion

The results of this challenge are satisfying, with the EVA02 model achieving the best accuracy of 0.92417. With more time and resources, I would have explored more intensive data augmentation, collected additional data, or tested larger models. Overall, this was an interesting experience, combining technical challenges with the fun of friendly competition.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[3] 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.

[4] M HeK, Q RenS, et al. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] 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.

[6] 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.

[7] 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.