# Assignment 3 - ImageNet Sketch Classification

Tess Breton

## Approach Overview

My approach consisted in fine-tuning the last classification layer of several models pretrained on ImageNet-1K, and applying data augmentation to try to prevent overfitting.

## 1. Data Exploration

**Visualizing Images** : I first used a Plotly dashboard to explore the classes. I noticed that there were many black images, which I decided to remove from the training set since thet do not carry meaningful features and very few were in the test set. I also identified mislabeled images, which I did not remove since similar mislabeled data could be in the test set. Moreover, there were even images present in two different classes, clearly limiting the achievable accuracy.

**Data Augmentation** : The ImageNet-Sketch documentation states that the dataset was built using flips and rotations. So it seemed natural to use such augmentations during training, especially horizontal flips. I also tried using other custom augmentations that seemed relevant, some of which are displayed in Figure 1.

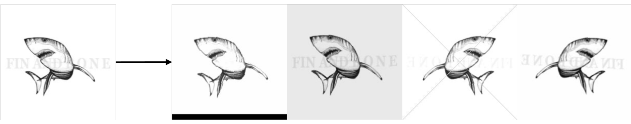

Figure 1. Reference image and some augmented versions

## 2. Models

**Vision Transformers** : First, I trained several Vision Transformers[2] of different sizes, and ended up using a 'huge' ViT from `timm`, pretrained on multiple datasets, including ImageNet-1K. I replaced the final layer with a fully connected layer of 500 outputs and froze the other layers, to leverage the pretrained model's feature extraction.

**EVA Models** : Then, when looking at `timm`'s ImageNet benchmark to see which models performed best, I realized that there were many EVA[3] and EVA-02[1] models that outperformed ViTs. I tried two of them, among which the 'large' EVA-02, which performed best on ImageNet-1K.

## 3. Training

**Set-up** : The models were trained for 20 epochs with batch size 64, SGD optimizer, learning rate 0.01 and momentum 0.9. I also tried using AdamW, dropout, different learning rates and a scheduler, but did not get any better results.

**Resizing** : I later realized that most images were not square. The default Resize transform alters their aspect ratio by stretching the shortest dimension. I have thus tried training while preserving this ratio, by adding a white padding if necessary (see Figure 2).

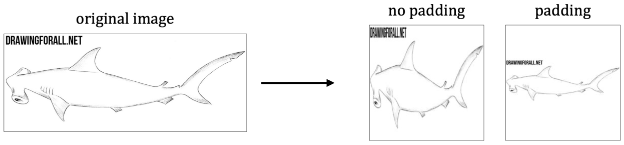

Figure 2. Default resizing and resizing with white padding

**Final Submission** : I used the validation set to track performance during training with `wandb` to select the best model. For my final Kaggle submission, I retrained the selected model on the combined training and validation sets.

## 4. Results

Tab. 1 summarizes the best results obtained. Stronger augmentations did not provide improvement, and padded resizing offered a little gain. There was always a strong overfit, with final training accuracies around 99%.

Table 1. Best validation accuracy (%) obtained during training

| Model | Baseline | Random H-flip | Padded resizing |
|---|---|---|---|
| ViT base 16-224 | 87.20% | N/A | N/A |
| ViT huge 14-224 | 90.80% | 91.36% | N/A |
| EVA giant 14-224 | 91.56% | 91.76% | N/A |
| EVA-02 large 14-448 | 92.68% | 92.60% | **92.97%** |

My final Kaggle submission (EVA-02 large trained on both training and validation data with padded resizing) scored **93.432%** on the public test set.

## References

[1] 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. 1

[2] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1

[3] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 1