Sketch Image Classification

Franki NGUIMATSIA TIOFACK
MVA, ENS Paris Saclay
franki.nguimatsiatiofack@ensae.fr

Abstract

This report present our approach and results in sketch classification assignment. Aiming to develop an accurate model, we combined transfer and ensemble learning, integrating CNN and Vision Transformers. Despite the limited training set size, this approach excelled, attaining an 82.78% accuracy on a Kaggle competition's public test set. The paper summarizes our methods for data preparation, model training, and the strategic design of our ensemble model.

1. Introduction

To address this challenge, we employed an ensemble learning approach, leveraging transfer learning by adapting pretrained models from the ImageNet dataset and fine-tuning them for enhanced performance on sketch classification.

1.1. Dataset

The dataset used for this task is the TU-Berlin Sketch Dataset, a large collection of 20,000 hand-drawn sketches evenly distributed across 250 categories. These sketches, contributed by non-expert users, encapsulate a wide array of objects and concepts, characterized by their abstract and stylistic variations. This dataset, with its unique and diverse representations, provides an extensive platform for evaluating sketch classification algorithms.

1.2. Pretrained Models and Ensemble Method

We utilized a series of advanced pretrained models, including ResNet[5], EfficientNet[4], ResNext[2], Vision Transformer (ViT)[1], MobileNet[3], and DEiT[6], Swin Transformer (Swin)[7]. All these models were initially trained on the ImageNet dataset, which provided a diverse and comprehensive foundation for feature extraction. We replaced the original final linear layers with new ones, adjusted to match the 250 unique classes of our dataset.

For our ensemble method, we combined the strengths of these varied architectures by aggregating their individual predictions.

2. Pre-processing, Training and regularization

We employed transformative pre-processing methods on our dataset, applying random flips, rotations, perspective changes, and Gaussian blurring to augment the training data.

For training, we utilized SGD with learning rate 0.01, momentum 0.9 and weight decay as our optimization strategy, iterating through a maximum of 70 epochs. To combat overfitting and improve model generalization, we incorporated a Dropout layer in our models with a rate of 0.4.

We also implemented early stopping based on validation accuracy. If a model's performance on the validation set did not improve for five consecutive epochs, training was halted to prevent overfitting. This approach balanced the need for sufficient training against the risk of overtraining on the training dataset.

3. Results

The results of our experiments are summarized in table 1.

Model	validation set	Public set
Efficientnet	72.74%	x
Mobilenet	72.86%	x
Resnet50	74.22%	x
Resnext50_32x4d	73.77 %	x
DEiT	75.96%	x
Swin	78.7 %	x
ViT	78.42 %	x
Ensemble method	83.22 %	82.72%

Table 1. Accuracy on validation set and public test set

4. Conclusion

Our ensemble learning approach, combining CNNs and Vision Transformers, effectively addressed the sketch classification challenge. Despite limited training data, our method achieved a notable 82.78% accuracy on Kaggle's public test set. This success underscores the potential of integrating diverse models and strategies in complex classification tasks.

References

[1] 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.

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

[3] Howard, Andrew G and Zhu, Menglong and Chen, Bo and others. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017. arXiv:1704.04861.

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

[5] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.

[6] Touvron, Hugo and Cord, Matthieu and Douze, Matthijs and Massa, Francisco and Sablayrolles, Alexandre and Jégou, Hervé. Training data-efficient image transformers & distillation through attention, 2020. arXiv:2012.12877.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.

[8] Xingyuan Zhang, Yaping Huang, Qi Zou, Yanting Pei, Runsheng Zhang, Song Wang. A Hybrid convolutional neural network for sketch recognition.