Quick Draw! Doodle Recognition Deep Learning Strategy

Rui Ning, Yumeng Yue, Zewen Zhang
ruining@stanford.edu, yuey3@Stanford.edu, zwzhang@stanford.edu

Abstract

In this project, we have explored different neural network architectures for "Quick Draw!" tasks, an experimental human-AI interactive game released by Google in 2016.

Three different CNN models are adopted here: **4-layer CNN, MobileNetV2,** and **DenseNet169**. We present qualitative result analysis of these models and conclude that DenseNet169 model with RGB images as input performs better.

Figure 1. Doodle Grayscale Images.

Data Processing

To test models, we split the data into three different folds: **80% for training, 20% for test.** To reduce computation time and storage of the data, a smaller subset of the original dataset produced by randomly sampling 2% of the totals.

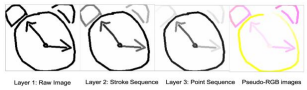

Figure 2. Schematic of data processing.

Figure 3. Pseudo-RGB Generation

In order to include **Time Sequence Information** in the input images, we generated **3-channel** pseudo-RGB images. The first channel contains **shape info**, the second channel contains **stroke sequence info**, and the last channel contains **point sequence info** in each stroke.

Figure 4. as-rendered RGB Doodle Images

Neural Network Models

For 4-layer CNN model, we applied a 3*3 filter with same padding followed by a 2*2 max pooling and 0.1 dropout for each layer, and followed by 2 fully connected layers.

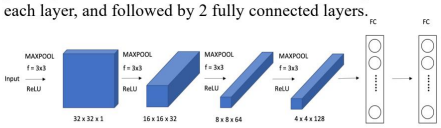

Figure 5. 4-Layer CNN

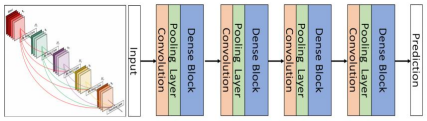

Figure 6. Densenet169

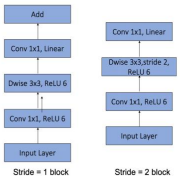

Figure 7. MobilenetV2

MobileNetV2 and DenseNet169 were directly adopted public architecture trained from scratch. All models are trained with a minibatch size of 128.

Results

All models yield an accuracy of **higher than 70%** despite different types of image input, respectively. DenseNet169 trained on pseudo-RGB images gave out the best test accuracy of 75%.

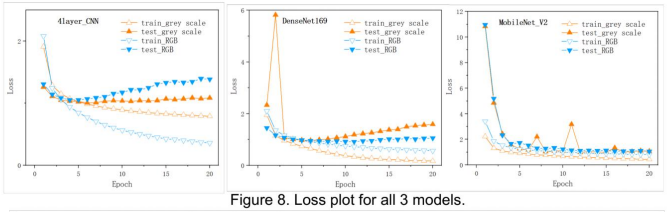

Figure 8. Loss plot for all 3 models.

Accuracy	4layer CNN	DenseNet169	MobileNetV2
1channel-greyscale image	71.8%	71.4%	74.1%
3channel-pseudo RGB image	71.4%	75.0%	70.8%

Table 1. Test Accuracy for Different Models

Surprisingly, 4layer CNN model worked quite well taking into account its shallow nature. Also it's important to notice that 4layer CNN and DenseNet169 over-fit quickly only after 5 epochs.

Discussions

- Larger training sample size should help test accuracy which is limited by computation hardware currently.
- Noisy nature of doodle images, such as bad drawing incompleteness, or oversimplicity, prevent us from achieve high accuracy with these neural networks.
- Shallower neural networks worked well on this task, deeper network not necessarily perform better.
- The inclusion of more information like time sequence in drawing only helped for DenseNet169, but not for MobileNetV2 or shallow CNN models. Deeper model better utilizes more complex information.

Future

- Examine prediction results to figure out which classes the models could not perform well on.
- Fine tune parameters to achieve higher test accuracies.
- Train models with larger sample size on computation hardware with larger memory to improve accuracy.
- Experiment with RNN models which hold promise in dealing this kind of tasks since there are a lot of sequence information.
- Implement ensemble for different models to make better prediction on doodle images
- Implement the best model on an mobile device

Reference

[1] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 4510–4520, 2018.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition , pages 4700–4708, 2017.

[4] Karen Simonyan and Andrew Zisserman. Very deep convoluvolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014