Machine Learning
hw3 report

Han Hao Chen, r05922021

November 8, 2016

1 Supervised Learning

Convolution neural network The hw3 is an image classification problem. We have some label image data with 32*32*3 image size. I use convolution neural network with 3 convolution layer and one hidden dense layer with 0.7 dropout rate in each layer to solve this classification problem. The training accuracy can achieve 93% and validation can achieve 60%. The testing accuracy can achieve 56.74% on the kaggle leaderboard.

2 Semi-supervised Learning 1

Self-training Since, we only have 5000 labeled data, the training result is tend to overfit these 5000 labeled data. However, we can use lot of unlabeled data to help us improving this result. By using self-training method we can gain a better model. I use three self-training cycles on this task. For each cycle, I use labeled data to train a model and use the model to predict unlabeled data, then adding top K confidence unlabeled prediction results to labeled data. The training accuracy can achieve 91% and validation can achieve 64%. The testing accuracy can achieve 59.38% on the kaggle leaderboard.

3 Semi-supervised Learning 2

Auto-encoder I also use auto-encoder to leverage the unlabeled data. First, I train an auto-encoder on unlabeled and label data until Euclidean distance error down to 2000. Then, I use each labeled data's auto-encoder features to find K nearest neighbor in unlabeled data, adding these unlabeled data (with its labeled neighbor as its label) to labeled data set. Finally, I use CNN to training on this labeled data set and get the final result. The training accuracy can achieve 91% and validation can achieve 61%. The testing accuracy can achieve 57.86% on the kaggle leaderboard.

4 Comparison

Comparison The result of self-training and auto-encoder improve the cnn's performance. However, it seems that these two semi-supervise methods don't have great progress. The cutting threshold in both self-training and auto-encoder may be a key parameter to improve the performance.

Table 1: Accuracy on three different methods

model	accuracy
CNN	0.5674
Self-training	0.5938
Auto-encoder	0.5786

note. This is testing accuracy on kaggle leaderboard

5 Training Tips

- Using mini-batch with size 100 in both methods to speed up the training.
- The learning rate is set to 5e-3 in CNN with dropout rate 0.7 for each layer
- In self-training, adding top 1000 confidence unlabeled data in each self-training cycle