# Modification and Experiments of Neural Network Distillation

Minke Yu：Distillation on MNIST+remove1+ self-distiilation+ Tuning work + some report work
Ziling Yuan：Distillation on cifar10 + reverse distillation +poster work +some paper work

Duke
PRATT SCHOOL of ENGINEERING

## 1 Introduction

From Caruana and Hinton's proof and practice, using distillation technology, knowledge of integrated models can be compressed into a single model, making it easier to deploy. Based on the MNIST and Cifar10 data sets, this study investigated the validity of the Distillation model, the optimization of the simpleNN and resnet distillation model (Temperature adjustment and loss function selection), and the validity of the variant of the distillation model (Reversed Distillation; Distillation using incomplete transfer dataset; Self-Distillation) was explored and discussed.

## 2 Methodology

### 2.1 Distillation on MNIST

The key of Distillation is to use the soft-labels of teacher network instead of hard label of original data as the input of student model. Two simpleNN are designed as the two networks. Teacher network is much more complex.

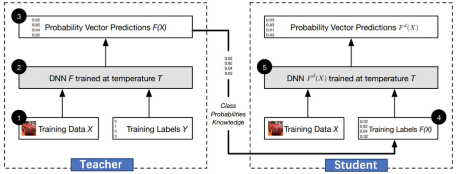
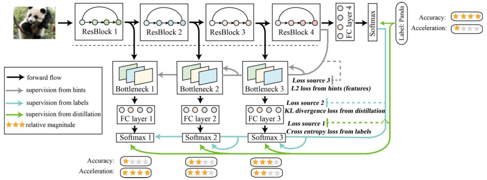

3
0.02
0.92
0.04
0.02
Probability Vector Predictions F(X)
2
DNN F trained at temperature T
1
Training Data X
0
1
0
0
Training Labels Y
0.02
0.92
0.04
0.02
Class Probabilities Knowledge
0.03
0.93
0.01
0.03
Probability Vector Predictions $F^d(X)$
5
DNN $F^d(X)$ trained at temperature T
Training Data X
4
0.02
0.92
0.04
0.02
Training Labels F(X)
Teacher
Student

Table 1: Baselines of Teacher and Student Network on MNIST

| Network | FLOPS Number | Parameters Number | Best Accuracy |
|---|---|---|---|
| Teacher | 33,716,800 | 33,716,800 | 0.9928 |
| Student | 11,206,400 | 11,206,400 | 0.9907 |

### 2.2 Self-Distillation of Resnet50 on Cifar10

Follow the framework of Linfeng's paper, we do the following work. ResNet50 was divided into 4 sections according to ResBlocks. Shallow classifiers were set respectively, combined with bottleneck layer (used to reduce the impact of shallow classifiers before) and full connection layer (used only in training and removed during predict).



The graph of Self-distillation model from the paper' Be your own teacher'

In addition, the loss function is composed of three parts: 1) the deep and shallow classifier softmax layer predicts the cross entropy loss of tags; 2) KL divergence computed by softmax outputs between students and teachers; 3) L2 loss between feature maps of deep & shallow classifier.

### 2.3 Model modification: Temperature and different loss function

Temperature: By dividing the T term when calculating the softmax loss function, it magnifies the loss value corresponding to the probability value of other classes, and magnifies the contribution of them.

Loss function: We tried 3 types of loss function: i) KL-Divergence; 2)Cosine Embedding; 3)MSE. We test the efficiency of them by integrate them into the softloss function.

## 3 Experimental Evaluations

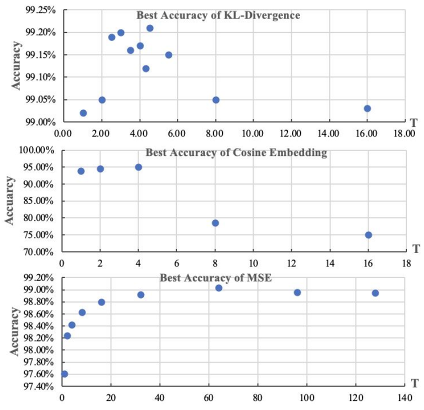
### 3.1 Distillation on MNIST

After learning the soft label, student network's accuracy increased from 0.9907 to 0.9921, much closer to the accuracy of teacher network.

Table 2: Comparison of Distilled Model and Baseline on MNIST

| Model | Best Accuracy on MNIST |
|---|---|
| Distilled Student Model | 0.9921 |
| Baseline Student Model | 0.9907 |

### 3.2 Model Modification with Temperature and Loss functions



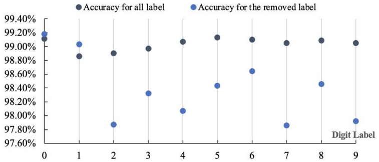
The chart of loss function selection and T parameter tuning

Code: https://github.com/ymkymkymkymx/ECE661_knowledge_distilation_project_minke-ziling

Considering T and loss function together, KL-Divergence with T = 4.5 perform the best with accuracy on MNIST is 0.9921.(We also do the T and loss function selection for resnet on cifar10, check details in the paper)

### 3.3 Distilled Model with one digit training data removed

This experiment reveals the power of Distilled model.Omit one digit from MNIST as transfer set and repeat the distillation process, the student network can still reach the accuracy around 99%. Especially, the prediction of digit which we omit for student training reaches over 95% accuracy, the knowledge is totally from the information of other digits' labels learnt by teacher.



### 3.4 Reverse the teacher and student network of distilled model

We used Resnet11 with accuracy 0.8968 as the teacher and Resnet20 as the student. The baseline for Resnet20 is 0.9188 and with the reverse distillation, it indeed reached a better result which is 0.925.

Table 6: Reversed Distilled Model on Cifar10

| Network | Best Accuracy |
|---|---|
| Teacher(Previous Student) | 0.8968 |
| Student(Previous Teacher) Distilled | 0.925 |
| Student(Previous Teacher) Baseline | 0.9188 |

### 3.5 Self distillation for ResNet50 on Cifar10

Self-Distilled Resnet50 has higher validation accuracy than baseline on Cifar10. To some extent proved the efficiency of self-Distillation''s logic, which is learn from itself for higher performance.

Table 5: Self-Distillation for ResNet50

| Network | Best Accuracy |
|---|---|
| ResNet50 Self-Distilled on Cifar10 | 0.9052 |
| ResNet50 Baselineon Cifar10 | 0.875 |

## 4 Conclusion

The significance of distillation is to transfer the knowledge of the complex teacher model to the lightweight student model by training with soft label, thus greatly improving the prediction accuracy of the student model.

In terms of 1) The optimization of the distillation model, taking KL-Divergence as the loss function and selecting T = 4.5, our distillation model had the best performance in MNIST, reaching 99.21%validation Accuracy. 2) Variation of distillation Model：Distilled Model with one digit training data removed, the prediction of digit which we omit for student training reaches over 95% accuracy, the knowledge is totally from the information of other digits' labels learnt by teacher, which proved the meaning of DIstillation, Reversed Distillation makes sense, and self-Distillation of resnet50 has a better than baseline performance on cifar10.