Knowledge Distillation —— Model Compression

UNIVERSITY OF TORONTO

YuHao Wang 1007993520

Department of Mechanical & Industrial Engineering

Introduction

With the accuracy in model training become more and more precise and satisfied, we want more!!! So the next goal is to decrease the reaction time of the model. We want to deploy those high-performance model on the mobile device with less computational power but with the structural complexity of the model, it is hard to make the model respond fast.

Thus, a solution for this problem is purposed – knowledge distillation. A model with simple structural will be trained under a teacher model, which is the "high-performance model" mention early, resulting in the "compression" of the model complexity but maintain the same level of the performance with the teacher model.

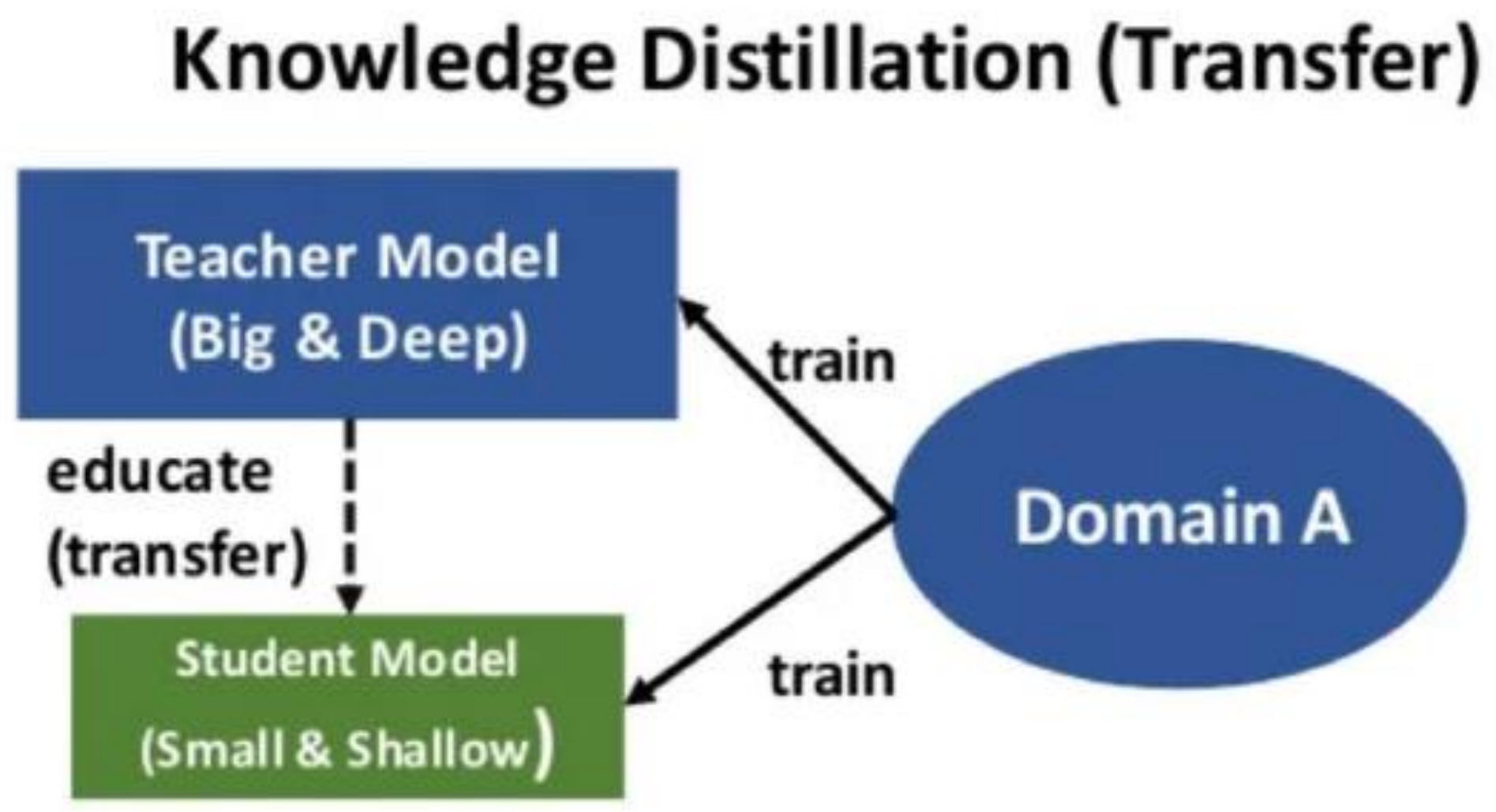


Figure 1: Simplified diagram for Knowledge Distillation

This experiment use MNIST and MHIST dataset and try to train "student" model with less parameters and FLOPS that has a faster processing time but the same level of performance with the teacher model.

For MNIST, teacher model will be a convolution network and student model will be a fully-connected network. For MHIST, teacher will be ResNet50V2 with last 5 layers unfreezed. Student model will be MobileNetV2 with last 5 layers unfreezed.

Methods

The teacher model will be trained first to act as a benchmark as well as the source of the knowledge. The input of the teacher model is the image input and its probabilities prediction for each class is the output.
The prediction from the teacher will then transferred into a "soft-target" with a temperature parameter T and fed into the student model:

$$Soft\ target = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

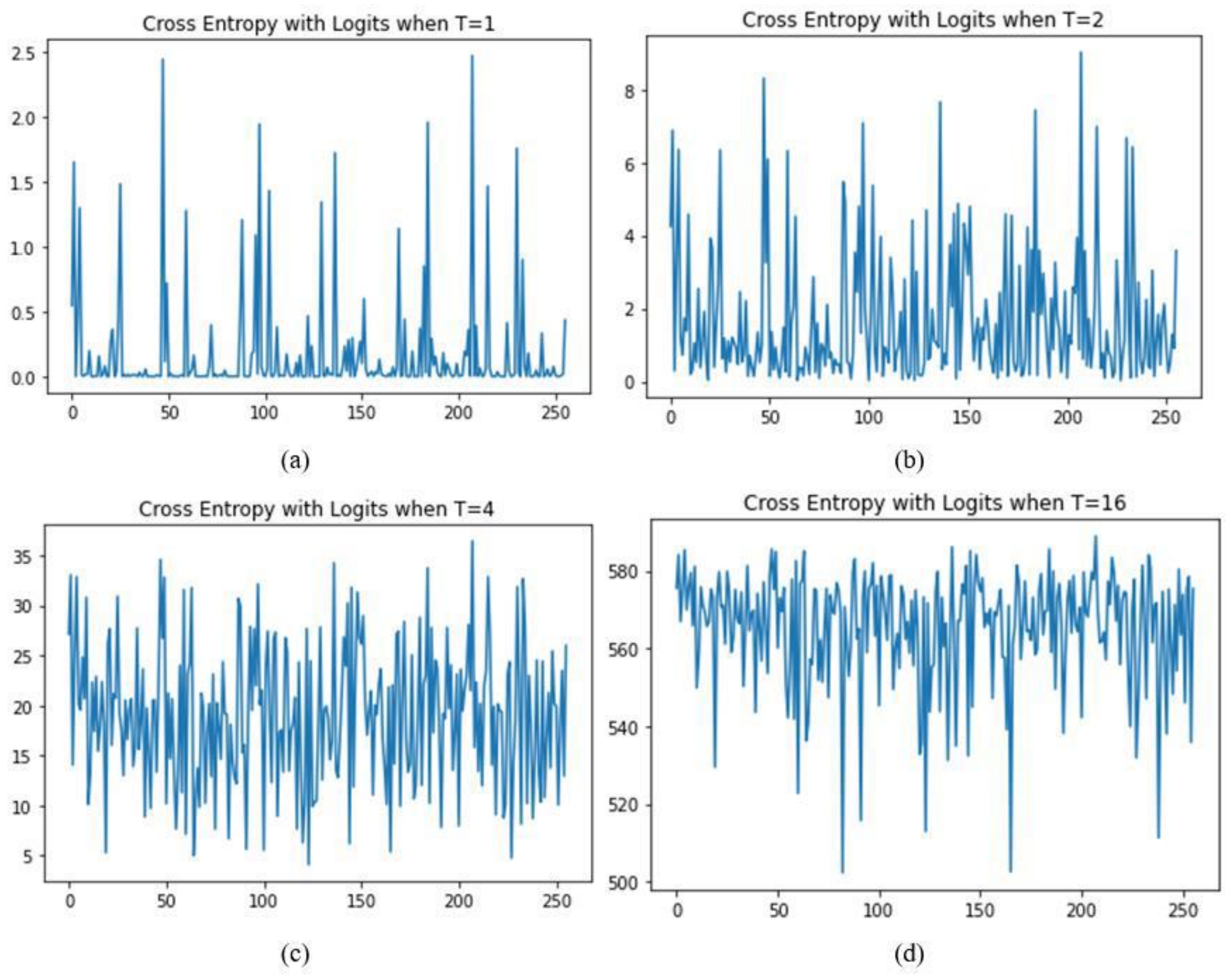
Along with the "soft-target" that fed into the student model, it also takes the ground-truth of the prediction as input to train the student model. The loss function for the student model is defined as distillation loss, which is made up by two parts as follow:

$$Loss_{Student} = \alpha * Loss_{w.r.t\ teacher} + (1 - \alpha) Loss_{w.r.t\ g.t}$$

$Loss_{w.r.t\ teacher} = KL\ divergence(Student\ pred., Soft\ targets\ from\ teacher)$ CE works fine
$Loss_{w.r.t\ g.t} = CrossEntropy(Studnet\ pred., Ground\ Truth)$

There are two hyperparameters in the loss α and T. α can change how much the student model trust on the soft-target teacher predict or the ground truth.

T will affect how much the "wrong" prediction of teacher will affect the student. With T=1, the "wrong prediction" contribute nothing to the student as the output is zero. With a higher T, the "wrong prediction" will have a non-zero output, providing information as well for the student.



(a) (b) (c) (d)

Figure 2: The effects of different temperature T

Results

Once the teacher is trained, the student model will be initialized and trained with both soft-targets and the ground truth. The students will be trained under the same parameter combination.

With the same optimizer, learning, batch and training epochs, models trained on MNIST and MHIST dataset has the following result.

For each group, the hyperparameters for distillation α and T are tunned to the best performance.

Note: MHIST dataset is unbalanced so the f1 score and AUC will be the metric for evaluation.

	Scratch	Distilled	Teacher
Accuracy	98.11%	98.65%	98.87%
FLOPS	1*10e6	1*10e6	1*10e7
# Para.	6*10e5	6*10e5	5*10e6

Table 1: Results for MNIST dataset

	Scratch	Distilled	Teacher
f1 score	0.6838	0.6885	0.7375
AUC	0.7509	0.7545	0.7921
FLOPS	6*10e8	6*10e8	7*10e9
# Para.	2*10e3	2*10e3	1*10e6

Table 2: Results for MHIST dataset

With knowledge distillation, student model in both dataset that distilled from teacher has improvement in performance compared to its own trained from scratch. Also, the drop of performance compared with the drop of FLOPS is extremely small from its teacher.

With the FLOPS for each dataset decreased by 91.4% and 90% compared to teacher model, the performance only decreased by 0.22% and 6% respectively.

One reason for the 6% drop for MHIST is the dataset have more features than the MNIST dataset. The reduce in structural complexity may result in a slight worse performance compared to the teacher model.

Conclusions

With Knowledge Distillation, a student model with less parameters and structural complexity is able to learn the knowledge from a teacher model. The performance of the student model has an acceptable drop from its teacher compared to the significantly drop of its FLOPS and the number of parameters.

The student model FLOPS is a tenth of the teacher model which means those model can be deploy on the device with less computational power. The student model can, theoretically, process the prediction ten time faster than the teacher model.

Here are some other potential task to do that can improve the results:

- Trying different network structure on MHIST dataset
- Add cross-validation between two dataset

Bibliography

1. **Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Proceedings of the 12th ACM SIGKDD international conference of Knowledge discovery and data mining, pages 535–541, 2006. https://dl.acm.org/doi/10.1145/1150402.1150464.**
2. **Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arXivpreprint arXiv:1503.02531, 2(7), 2015. https://arxiv.org/abs/1503.02531**
3. **Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arXivpreprint arXiv:1503.02531, 2(7), 2015. https://arxiv.org/abs/1503.02531**
4. **Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016. https: //arxiv.org/abs/1603.05027**