ECE1512 - Project A: Knowledge Distillation

UNIVERSITY OF TORONTO

Yiyang (Fred) Shi, Hsuan-Ling (Celene) Chen

Department of Electrical and Computer Engineering

Introduction

Knowledge distillation is a form of model compression methodology used to “capture” and “distill” the knowledge in a complex machine learning model or an ensemble of models (known as teacher model) into a smaller single model (known as student model) that is much easier to deploy [1].

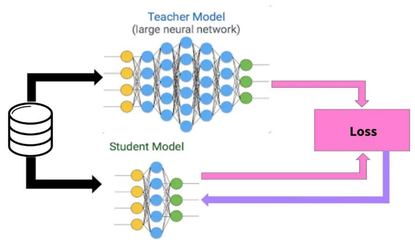


Figure 1: Knowledge distillation architecture

The project focuses on the setting of Knowledge distillation as a model compression technique to transfer knowledge from a larger model to a smaller one that can be used practically in real-world settings. There are two tasks accomplished in this project:
(1) using **conventional and state of the art knowledge distillation** frameworks as a model compression method for a popular digit classification dataset, “MNIST”
(2) using **transfer learning and knowledge distillation** to train a lightweight model for mimicking a pre-trained larger model in a clinical histopathology dataset, “MHIST”

The MNIST and MHIST datasets are listed below for direct visual comparison.

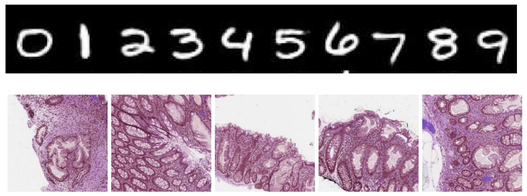


Figure 2: MNIST and MHIST dataset

MNIST dataset provides grey scale representations of numbers in different classes, while MHIST dataset includes complicated and unbalanced histological pattern of images with higher resolution at 224x224 pixels. The patterns are hard to differentiate and locate to annotate between the following 2 classes.

1. Hyperplastic Polyp (HP)

2. Sessile Serrated Adenoma (SSA)

Methods

To quantitatively measure the effectiveness of knowledge distillation, we selected a cumbersome teacher model and a lightweight student model to perform distillation on MNIST dataset first.

Table 1: Teacher and student model FLOPs comparison

Model	FLOPs (G)
Teacher model	0.022
Student model with KD	0.00248

Knowledge distillation is regulated by the two hyper-parameters (temperature **T** and scaling factor **α**), and they require manual tuning to achieve the best results on different applications. Therefore, we trained the student model and investigated the effect of each parameter separately, while maintaining the rest of configurations. The test set accuracy is used as comparison metrics between the students with and without knowledge distillation.

Subclass knowledge distillation is builds on the existing paper and proposes manual creation of subclasses to enhance the distillation process and achieve better results [2]. We performed a similar verification process and compare the effectiveness of this method to the original one.

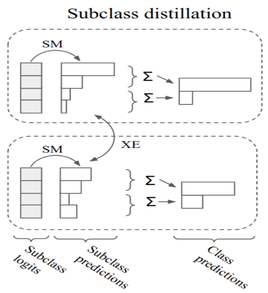


Figure 3: Subclass predictions and logits

With a good grasp of knowledge distillation and its variants, the same methods are then applied to the more complicated dataset MHIST to investigate cross dataset performance of proposed methods. Majority of the experiment setups are kept the same with addition of transfer learning process [3].

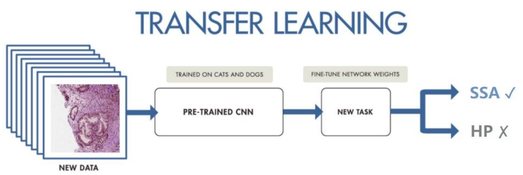


Figure 4: Transfer learning on MHIST data

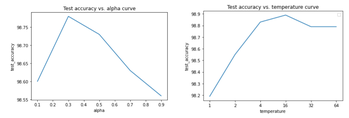
The distillation hyper-parameters are investigated, and hand tuned again for this dataset and the test accuracy results from base KD and subclass KD are compared quantitatively against the student baseline and their teacher model.

Results

Table 2: Test accuracy between models with and without Subclass KD

Model	T	α	Accuracy(%)
Teacher with subclasses	-	-	99.31
Teacher without subclasses	-	-	99.29
Student with subclasses KD	4	0.5	98.93
Student with convential KD	4	0.5	98.85
Student without KD	4	0.5	98.11

Student with KD achieves 0.74% higher accuracy than the one without KD. The introduction of subclasses increased both the teacher and student accuracy. The student with subclass KD has a higher accuracy jump of 0.08 percent from conventional KD.

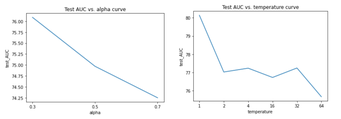


Final distillation hyperparameters values used: the task balance is 0.3 and the temperature is 16.

Table 3: Model complexity between teacher and student MNIST

Model	Number of parameters	FLOPs (G)	Accuracy(%)
Teacher model	1,011,466	0.022	99.29
Student model with KD	1,238,730	0.00248	98.89

The student model's lower FLOPs than the teacher's model indicates that it needs lower computing power and faster training and inference time compared to the teacher's model. The student model with KD achieved comparable accuracy results as the teacher with only 10 percent of FLOPs.



Final distillation hyperparameters values used: the task balance is 0.3 and the temperature is 1.

Table 4: Model complexity between teacher and student MHIST

Model	Number of parameters	FLOPs (G)	AUC(%)
Teacher model	23,568,898	8.19e-06	89.22
Student model with KD	2,260,546	5.12e-06	79.53

The ResNet50v2 model has 10 times the number of parameters as Mobilenetv2. Teacher model's FLOPs is higher than the student’s FLOPs. Model compression is achieved.

Table 5: Test AUC between models with and without Subclass KD and transfer learning

Model	T	α	AUC(%)
Teacher with subclasses	-	-	86.61
Teacher without subclasses	-	-	89.22
Student with subclasses KD	1	0.3	85.28
Student with convential KD	1	0.3	80.72
Student without KD	-	-	77.32

Model	Learning rate	Initial Epochs	Fine-tune Epochs	AUC(%)
Teacher with TL	1E-3/1E-4	10	25	89.22
Teacher without TL	1E-3/1E-4	10	25	73.17
Student KD with TL	1E-3/1E-4	10	25	80.72
Student KD without TL	1E-3/1E-4	10	25	63.18
Student TL without KD	1E-3/1E-4	10	25	77.32

In general, transfer learning and knowledge distillation both helped to improve the performance of the student model.

Conclusions

The setting of Knowledge distillation as a model compression technique can be successfully used to transfer knowledge from a larger model to a smaller one that can be used practically in real-world settings.

The work from the first task showed that the performance of the student model has been increased by 0.74% by knowledge distillation while maintaining its size on the digit classification dataset, “MNIST”

The results in the second task showed that using the guidance from a larger pre-trained network and fine-tuning, this helped the student model to achieve a better validation performance by 17.54% on the clinical histopathology dataset, “MHIST”.

To further quantify the distillation performance and test the method capabilities across different objectives and datasets, a few areas of investigation should be considered for improvements. With consideration of time and resource, we should

- Test distillation methods on a high-resolution dataset such as ImageNet subsets.

- Modify the loss function and apply concepts on a model with different learning objective, such as object detection or segmentation.

Bibliography

1. Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7), 2015. https://arxiv.org/abs/1503.02531.
2. Rafael Müller, Simon Kornblith, and Geoffrey Hinton. “Subclass distillation”. In: arXiv preprint arXiv:2002.03936 (2020)
3. D. (D. J. Sarkar, “A comprehensive hands-on guide to transfer learning with real-world applications in Deep learning,” *Medium*, 17-Nov-2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a. [Accessed: 14-Dec-2022].