Deep Learning Final Project Proposal
: Improved structure for Knowledge Distillation

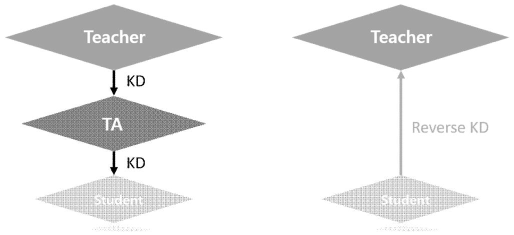
Chanhwa Lee 2014-18323, Seung Hoon Paik 2015-12868

Ⅰ. Previous work

Knowledge Distillation (KD) is one of the popular ways that compress model. The idea of KD was first introduced by Bucilua et al. [1], and popularized by Hinton et al. [2]. The pre-trained large model is called "Teacher", and the objective small model is called "Student". In Hinton's model, teacher transfer the "soft label" of each given data to student, which takes an important role in student's loss function. Soft label is the class score, the result of softmax at the last layer. Hinton's approach is called "Baseline Knowledge Distillation (BLKD)". Interestingly, for fixed objective model size, student with BLKD performs better than the model using a classical learning method.

In [3], Mirzadeh et al. found the size gap between teacher and student is important in training. The main point is, for fixed student size, the larger teacher does not always guarantee a better student. To reduce the negative effect of the size gap, they proposed "Teacher Assistant Knowledge Distillation (TAKD)" with adding a mid-size model called "TA" between teacher and student.

Yuan et al. [4] suggested two other variations of KD. They are named as "Reversed Knowledge Distillation (Re-KD)" that student teaches teacher, and "Defective Knowledge Distillation (De-KD)" that poorly trained teacher teaches student. One of the interesting results is poorly trained teacher also enhances the accuracy of student.



KD via Teaching Assistance

Reverse KD

Figure 1. Variations of KD model

Ⅱ. Goal

Our goal is to find better KD methods. We suggest two possible candidates, WSKD and GSKD. We will implement both structures, and compare the computation cost and performance of candidates with the other existing KD variations.

Ⅲ. Significance

A large cost of time and computation is naturally required for an accurate neural network. The critical problem is that a larger model gives better accuracy, but has more limitations in deployment. Unlike developers, most users have tight restrictions (e.g. cell phone, IoT device, etc). Also, to use the model in real-time, reducing the computation is important. Therefore, model compression is a significant issue, and KD is an effective training method for appropriate model compression.

In this regard, finding a better student has been a popular topic for KD related works. We expect that our candidates, inspired by the principles of learning in the real world, will become another improved version of KD.

Ⅳ. Research Plan

We will follow the settings of previous related works: use MNIST, CIFAR-10, and CIFAR-100 data sets; use the variation of ResNet such as ResNet8, ResNet14, and ResNet26 for the structure of teacher and student.

▶ WSKD (Wise Student Knowledge Distillation)

The goal of WSKD model is to build student that has strength in which the teacher network could not. This idea comes from a Korean idiom, "청출어람", which means "the pupil has become the master." In detail, student will be trained using samples that were misclassified by teacher with a special method. There are two back propagation steps in this model.

Step 1 Update student once using normal KD loss function. We expect that student network will absorb general knowledge of teacher's.
Step 2 For a sample misclassified by teacher, modify the soft label of this sample to have the maximum value at true class and preserve correlation of other classes. Next, update the student once only using misclassified samples and new loss function defined as a weighted sum of cross entropy loss and KL-divergence between student's soft label and teacher's modified soft label.

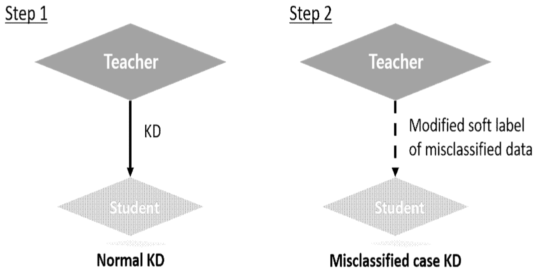


Figure 2. Structure of Wise Student KD model

By repeating Step 1 and Step 2, student network gets general knowledge to classify samples imitating teacher's performance and special knowledge to perform better than teacher in samples which teacher has failed to classify well.

▶ GSKD (Growing Student Knowledge Distillation)

The idea of GSKD comes from the human life cycle. It reflects the academic level and brain growth of corresponding stages in the cycle. Teacher will train the smallest student. Then larger student for the next stage is affected by the previous stage student. A detailed description of the structure is as follows:

Step 1 Teacher trains the smallest size student, with normal KD loss function. This is the student for the first stage.
Step 2 Train the student for the next stage. This is larger than the previous one. In this current stage, the update requires the soft labels of both teacher and the previous stage student. We will use the loss function similar to the normal KD, with the only difference is adding KL-divergence term for the current and previous stage student. Repeat Step 2 until we get the objective size of student we want.

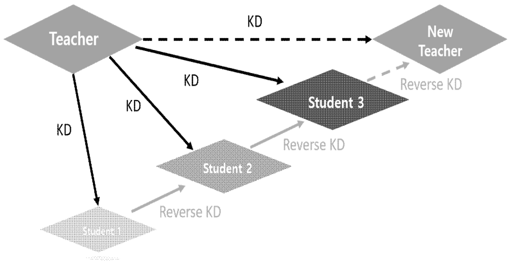


Figure 3. Structure of Growing Student KD model

If we repeat Step 2 until we reached the student with the same size as teacher, let this student be a new teacher. Then, we may compare the accuracy of the new teacher and the original teacher. Since this iteration can be interpreted as a stage-wise reverse KD, we expect to observe the interesting result.

V. Roles

Chanhwa Lee : Focus on WSKD
Seung Hoon Paik : Focus on GSKD
The other part will be handled in a co-operative manner :)

VI. References

[1] Buciluǎ, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2006).
[2] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
[3] Mirzadeh, Seyed-Iman, et al. "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher." arXiv preprint arXiv:1902.03393 (2019).
[4] Yuan, Li, et al. "Revisit Knowledge Distillation: a Teacher-free Framework." arXiv preprint arXiv:1909.11723 (2019).