A Gift from Knowledge Distillation

Hongzhi Liu

May 25, 2018

1 Overview of Research Method

Over the past several years, various deep neural network (DNN) models have provided perfect performance in many tasks, ranging from computer visionto natural language processing. Recently, several studies on the knowledge transfer technique have been conducted. Hinton *et al.* [1] first proposed the concept of knowledge distillation (KD) in the teacher–student framework by introducing the teacher's softened output. Although the KD training achieved improved accuracy over several datasets, this method has limitations such as difficulty with optimizing very deep networks.

In the thesis of Professor Yim, the team propose a novel technique to distill knowledge [2]. As the DNN maps from the input space to the output space through many layers sequentially, they define the distilled knowledge to be transferred in terms of flow between layers, which is calculated by computing the inner product between features from two layers.

This approach is useful for fast optimization. Using the proposed distilled knowledge to find the initial weight can improve the performance of a small network. Even if the student DNN is trained at a different task from the teacher DNN, the proposed distilled knowledge improves the performance of the student DNN.

2 Learning Procedure of DNN

In the case of people, the teacher explains the solution process for a problem, and the student learns the flow of the solution procedure. In this manner, the team believe that demonstrating the solution process for the problem provides better generalization than teaching the intermediate result.

In the case of a DNN, the relationship can be mathematically considered by the direction between features of two layers. They designed the FSP matrix to represent the flow of the solution process. The FSP matrix $G \in \mathbb{R}^{m \times n}$is generated by the features from two layers. Let one of the selected layers generate the feature map $F^1 \in \mathbb{R}^{h \times w \times m}$ where h, w, and m represent the height, width, and number of channels, respectively. The other selected layer generates the feature map $F^2 \in \mathbb{R}^{h \times w \times n}$. Then, the FSP matrix $G \in \mathbb{R}^{m \times n}$is calculated as Equation (1):

$$G_{i,j}(x;W) = \sum_{s=1}^{h} \sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}, \tag{1}$$

where x and w represent the input image and the weights of the DNN, respectively. They prepared residual networks with 8, 26, 32 layers that were trained with the CIFAR-10 dataset. There are three points in the residual network for the CIFAR-10 dataset where the spatial size changes. The team selected several points to generate the FSP matrix, as shown in Figure 1.

In this study, Professor Yim considers a pair of FSP matrices between the teacher and student networks $(G_i^T, G_i^S), i = 1, \ldots, n$ with the same spatial size. We took the squared L norm as the cost function for each pair. The cost function of transferring the distilled knowledge task is defined as Equation (2):

$$\begin{aligned} &L_{FSP}(W_t, W_s) \\ &= \frac{1}{N} \sum_{x} \sum_{i=1}^{n} \lambda_i \times \parallel G_i^T(x;W_t) - G_i^S(x;W_s) \parallel_2^2, \end{aligned} \tag{2}$$

where λ_i and N represent the weight for each loss term and the number of data points, respectively. They assumed that whole loss terms are equally significant. Therefore, the team used the same λ_i for all experiments.

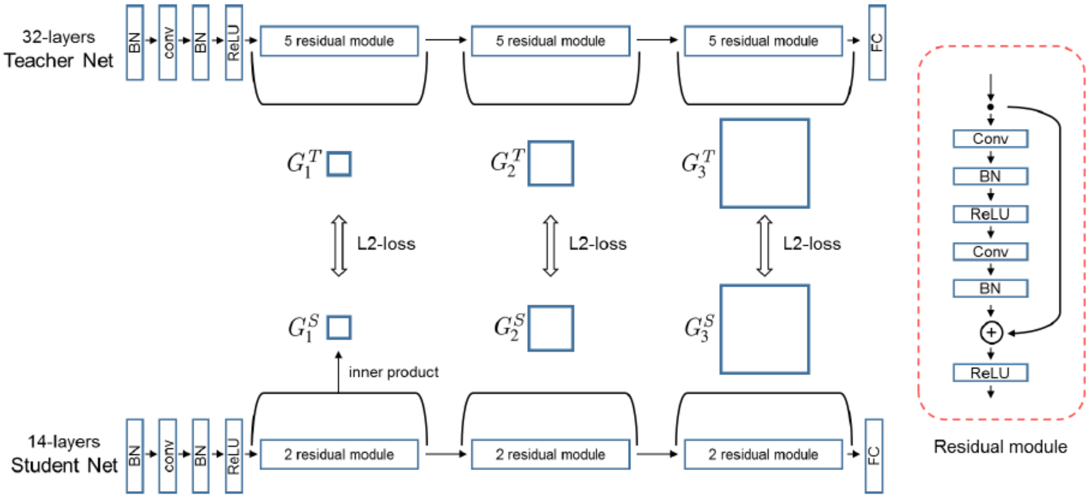

Figure 1: Complete architecture of their proposed method. The numbers of layers of the teacher and student networks can be changed. The FSP matrices are extracted at the three sections that maintain the same spatial size. There are two stages of their proposed method. In stage 1, the student network is trained to minimize the distance between the FSP matrices of the student and teacher networks. Then, the pretrained weights of the student DNN are used for the initial weight in stage 2. Stage 2 represents the normal training procedure.

The learning procedure contains two stages of training. First, the team minimize the loss function LFSP to make the FSP matrix of the student network similar to that of the teacher network. The student network that went through the first stage is now trained by the main task loss at the second stage.

3 Evaluation of the effectiveness of technique

Professor Yim compared the distilled knowledge transfer method FitNet with their proposed method. As given in Table 1, the student network with FitNet outperformed the teacher network with fewer iterations. However, when an ensemble of three networks was used, the teacher network with fewer iterations and student network with FitNet had similar accuracies. There was not that much improvement. In terms of the performance and number of iterations, the proposed method was more efficient than the existing method of FitNet, as presented in Table 1.

Table 1: Recognition rates (%) on CIFAR-100. The symbol * indicates that the network was trained with one-third of the iterations for the original case, which used 64000 iterations.

	Net 1	Net 2	Net 3	Avg	Ensemble	Iter
Teacher	64.06	64.19	64.21	64.15	69.3	192k
Teacher*	61.29	61.26	61.41	61.32	67.2	63k
FitNet [3]*	62.85	62.46	62.35	62.55	67.6	98k
Student *	64.66	64.64	64.65	64.65	68.8	95k

References

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *in Computer Science*, 2015. 1

[2] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network mini-

mization and transfer learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[3] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *in Computer Science*, 2014. 2