Stochasticity and Skip Connection Improve Knowledge Transfer

Luong Trung Nguyen
Dept. of Electrical and Computer Engineering
Seoul National University, Seoul, Korea
ltnguyen@islab.snu.ac.kr

Kwangjin Lee
Dept. of Electrical and Computer Engineering
Seoul National University
Seoul, Korea
kjlee@islab.snu.ac.kr

Byonghyo Shim
Dept. of Electrical and Computer Engineering
Seoul National University
Seoul, Korea
bshim@islab.snu.ac.kr

***Abstract*—Deep neural networks have achieved state-of-the-art performance in various fields. However, DNNs need to be scaled down to fit real-word applications where memory and computation resources are limited. As a means to compress the network yet still maintain the performance of the network, knowledge distillation has brought a lot of attention. This technique is based on the idea to train a student network using the provided output of a teacher network. Deploying multiple teacher networks facilitates learning of the student network, however, it causes to some extent waste of resources. In the proposed approach, we generate multiple teacher networks from a teacher network by exploiting stochastic block and skip connection. Thus, they can play the role of multiple teacher networks and provide sufficient knowledge to the student network without additional resources. We observe the improved performance of student network with the proposed approach using several dataset.**

***Index Terms*—convolutional neural network, Knowledge transfer, image classification, multiple teacher networks**

I. Introduction

In recent years, we have witnessed great success in deep neural networks (DNN) in various applications such as driverless vehicles and drone-based deliveries, smart cities and factories, remote medical diagnosis and surgery, and artificial intelligencebased personalized assistants [1]–[6]. Despite the superior performance, it is not easy to use the DNN-based models for the embedded systems having limited memory and computation requirement. As a means to handle this issue, knowledge distillation (KD), an approach to make the DNN-based models smaller but efficient, has received much attention in recent years [7].

The basic idea of KD is to train a smaller network (a.k.a. *student network*) with the help of distilled knowledge (a.k.a., *soft targets*) provided by a larger network (a.k.a., *teacher network*) (see Fig 1). Weights of the student network are trained in a way to minimize the sum of two cross-entropy losses: 1) the cross-entropy between the outputs of the student network and soft targets, and 2) the cross-entropy between the outputs of the student network and labeled data. Recently, it has been shown that the learning of student networks can be much improved with the help of multiple teacher networks [7]–[9]. Multiple teacher networks have different knowledge and views each other so that they generate different outputs on the same input. Hence, the student network can fuse distinct knowledge from multiple teacher networks to establish its own comprehensive and in-depth understanding of the dataset [9]. However, DNN-based multiple teacher networks might require more resources as compared with a single teacher network.

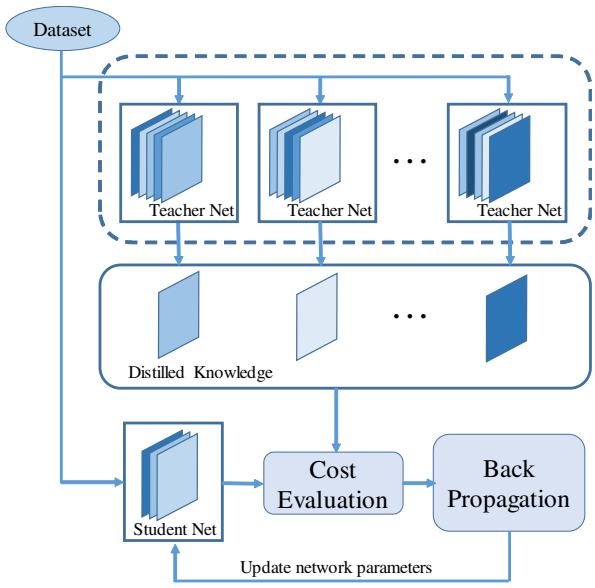


Fig. 1. Multiple teacher networks providing distilled knowledge to train the student network.

The primary goal of this paper is to improve the performance of a student network with the help of multiple teacher networks but clearly distinct from previous studies in the sense that there is no additional resource constraints. Our approach

builds on a single teacher network using stochastic block and skip connection so that it can provide multiple outputs to the student network. Recall that stochastic blocks are conceptually similar to the dropout operation, which sets part of the output of individual units to be zero at random during training [10]. Stochastic blocks consist of one or more layers, which are skipped at random during training. Recall also that skip connections are identity mappings designed to bypass one or more layers [1]. For example, consider a network consisting of 3 blocks S_1, S_2, and S_3 (see Fig 2). We can easily generate multiple networks from the given network using stochastic blocks and skip connections. Here, skip connections are represented using the identity mapping *Id*. When the blocks are dropped at random, the network generate different outputs (see Fig 2). This example is obviously simple, but the fundamental principle to extend a large network is not much different from this. Stochastic blocks and skip connections play an important role to generate multiple networks using multiple valid paths in the network diagram (see Fig 2).

In this paper, we propose a DNN-based scheme to generate multiple teacher networks from a teacher network using stochastic blocks and skip connections. By dropping DNN blocks at random, each generated teacher network consists of valid paths of batch data. In doing so, we can generate multiple teacher networks from one teacher network without additional resources. We demonstrate that the generated teacher networks are proper for transferring knowledge and also show that student networks improve further with the help of teacher networks.

II. Related Works

KD has been exploited to compress the knowledge in an ensemble of models into a single model [7]. An attention-based distillation approach has also been proposed [11]. In this approach, attention maps made from intermediate feature maps of networks are used to transfer knowledge. Also, mutual learning-based KD has been suggested as a new paradigm of bidirectional knowledge transfer, which exchanges knowledge in a mutually beneficial way [8].

Recently, knowledge transfer using multiple teacher networks have been proposed [9]. Multiple teacher networks provide comprehensive guidance to help the learning of student networks [8], [12]. A simple way is to build and train multiple teacher networks in a separate manner on the same data and then to transfer their distilled knowledge to the student network. However, a whole ensemble of the teacher networks is cumbersome and might be too computationally expensive, especially if the individual models are large neural networks. To overcome this, an approach to use a single teacher network and then to generate multiple outputs using noise perturbation has been proposed [13]. This might be problematic since the random noise can corrupt the knowledge in the outputs [14]. In contrast, our proposed approach generates multiple networks (see Figure 2) so that reliable and various knowledge is ensured. Also, in our approach, the generated networks have higher entropy outputs compare to the original network. Note that encouraging high entropy output [15] and smoothing label [16] are proved to help the training of DNNs since regularizing the high confident outputs prevents the network from overfitting and increases the adaptivity of the network.

III. Proposed Approach

In the proposed KD scheme, we use multiple teacher networks to provide different distilled knowledge to the student network (see Fig. 1). The main concept of our proposed approach is to generate multiple teacher networks of various outputs from a single network. In essence, the key point of this approach is to incorporate a single teacher network with stochastic blocks and skip connections (see Fig. 2). We explain how to generate teacher networks and demonstrate that these networks are appropriate for helping the learning of student networks.

A. Generating Multiple Networks

We exploit skip connections and stochastic blocks to generate multiple networks from a single teacher network. First, we add skip connections from the input of each DNN block to the corresponding output of the block. Recall that skip connections have been popularly used in DNN as shortcuts to jump over some layer [17]. Skip connections benefit from two folds: 1) to avoid the problem of vanishing gradients in DNN by bypassing blocks causing this problem and 2) to generate multiple paths of data in the network so that the network is still working well even when there are some dropped blocks (see Fig. 2). To express the output of a DNN block using skip connections, let f_i be the function representation of the i-th block in DNN[1]. Then, the output (o_i) of the i-th block can be expressed as

$$o_i = f_i(o_{i-1}) + o_{i-1}. \tag{1}$$

When the i-th block is dropped, $f_i(o_{i-1}) = 0$ and thus $o_i = o_{i-1}$. We can use a binary tree to intuitively represent the DNN with skip connections (see Fig. 2).

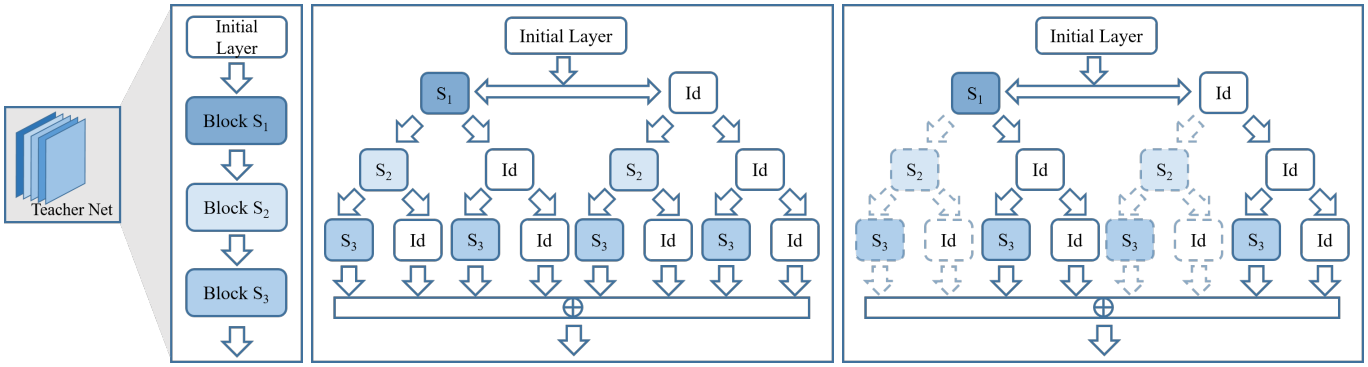
Second, in the proposed scheme, we drop the blocks at random using the mechanism of stochastic blocks. Recall that a stochastic block is assigned with a survival probability p, i.e., the block is dropped with probability $1-p$. To be specific, let p_i be the survival probability of the i-th block ($1 \leq i \leq n$). We use linear decay mode in which p_i satisfies

$$p_i = 1 - \frac{i-1}{n-1}(1-p_n). \tag{2}$$

Obviously, we have $p_1 = 1$ and $p_1 > p_2 > \cdots > p_n$. That is, in the binary tree, the blocks near the root have higher chance to survive. As a result, dropping a few blocks in the network does not harm the performance much[2]. The proposed scheme is a natural combination of the DNN and stochastic blocks and skip connections. From this, multiple teacher networks can be generated with reliable performance.

[1] f_i is the function composition of a nonlinear activation function (e.g., rectified linear unit (ReLU) or leaky ReLU) and a linear function of weights and biases in the i-th DNN block.

[2] If k blocks are dropped from n blocks, 2^{n-k} valid paths still exist.



(a) (b) (c)

Fig. 2. Diagram of a teacher network: (a) conventional network, (b) network with skip connections represented by identity mapping Id, and (c) network with stochastic blocks in which the blocks are dropped at random.

Note that in (2), all the survival probabilities p_i are expressed as functions of p_n. The initialization of p_n implies a trade-off between reliability and variety of teacher networks. When p_n is large, there are just a few of generated teacher networks whose the size is as large as the original DNN. In contrast, when p_n is small, there are multiple generated teacher networks whose the size is much smaller then the original DNN. In the proposed scheme, we set $0.5 \leq p_n \leq 0.9$ with 0.1 interval and choose the best p_n for each pair of teacher network and student network. We also note that since the computational complexity of each block is the same as in the single network, the overall complexity of the proposed scheme is comparable with that of the single network.

B. High-Entropy Output

Teacher networks provide student network with high entropy output distribution which helps the learning process of student networks (see Fig 3). It is known that regularizing a neural network to be less confident improves the performance [8], [15]. This is because training a network to have high confident output let it overfit to the train dataset and reduce the adaptivity by bounding the gradient. Also, the high entropy distribution contains important information like relations between classes which are the salient cues how the teacher network generalizes. In [8], it has been shown that using an ensemble of n networks as a teacher is less helpful than using n individual networks as n teachers. This is because the ensemble makes the secondary values of outputs low entropy distribution. In our proposed approach, generating teacher networks from the teacher network is analogous to using individual networks instead of the ensemble of them in [8]. As a result, the knowledge that teacher networks provide contains important information and is learned easily by student networks.

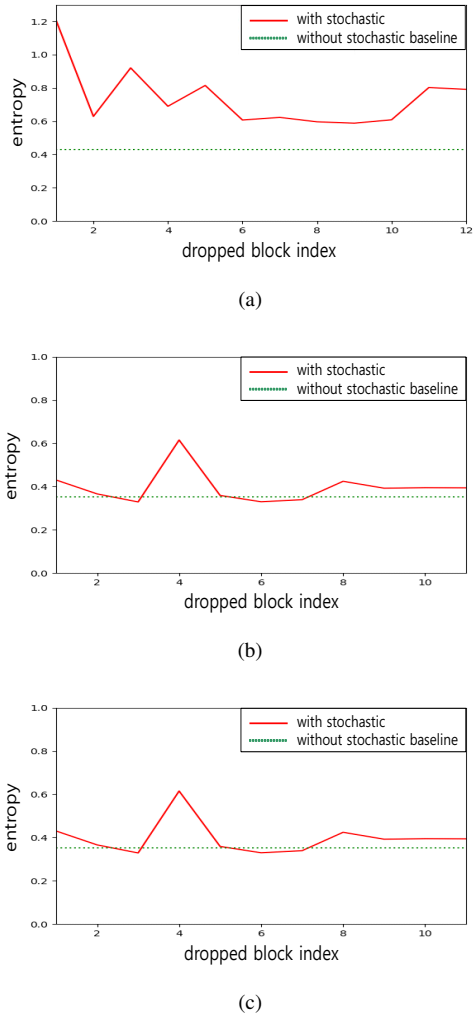


Fig. 3. Entropy when each block is dropped from (a) residual network 32, (b) mobilenet, and (c) wide residual network 28-10.

IV. Experiment

We test the performance of the proposed scheme using two datasets - CIFAR-100 [18] and tiny imagenet [19]. CIFAR-100 dataset consists of 32×32 RGB color images drawn from 100 classes, which are split into $50,000$ train and $10,000$ test images. Tiny imagenet dataset is a down-sampled version of ImageNet dataset. It consists of 64×64 RGB color images

TABLE I
ACCURACY PERFORMANCE OF MUTUAL LEARNING (ML) ON CIFAR 100

Net 1	**Net 2**	**Independent Learning**		**Conventional ML**		**Proposed Technique**	
ResNet 32	ResNet 32	69.86	69.86	71.14	71.21	73.68	73.58
MobileNet	ResNet 32	74.08	69.86	75.62	71.1	76.2	72.76
WRN 28-10	ResNet 32	78.98	69.86	78.53	72.18	80.65	73.08
MobileNet	MobileNet	74.08	74.08	75	75.16	75.5	76.1
WRN 28-10	MobileNet	78.98	74.08	78.34	76.41	81.03	76.82
WRN 28-10	WRN 28-10	78.98	78.98	78.83	78.95	81	80.66

TABLE II
ACCURACY PERFORMANCE OF MUTUAL LEARNING (ML) ON CIFAR 100

Net 1	**Net 2**	**Independent Learning**		**KD**	**Proposed Technique**
Res 32	VGG 13	69.86	67.74	71.5	72.2
Res 110	Res 20	71.69	68.32	68.72	70.99
WRN 28-10	Res 32	78.98	69.86	69.85	74.87
MobileNet	Res 32	74.08	69.86	69.88	71.77
Res 110	Res 32	71.69	69.86	70.12	73.36
MobileNet	VGG 13	74.08	67.74	68.83	71.12

TABLE III
ACCURACY PERFORMANCE OF ATTENTION TRANSFER (AT) ON TINY IMAGENET

Net 1	**Net 2**	**Independent Learning**		**AT**	**Proposed Technique**
Res 110	Res 20	52.32	46.85	51.49	51.9
WRN 28-10	Res 32	58.91	49.01	53.56	54.15
Res 110	Res 32	52.32	49.01	54.52	54.91
WRN 40-4	Res 32	55.19	49.01	54.33	54
WRN28-10	WRN40-4	58.91	55.19	60.98	61.36

drawn from 200 classes, which are split into $100,000$ train and $10,000$ test images.

For CIFAR-100, we normalize each image and augment the train images as in [1]. Each network is trained for 200 epochs with batch size of 128. The initial learning rate is divided by 10 for every 60 epochs. For tiny imagenet, we use its pure dataset version without augmentation. Each network is trained for 100 epochs with batch size of 128. The learning rate is divided by 10 for every 40 epochs. We use stochastic gradient descent optimizer with momentum of 0.9. The initial learning rate is 0.01 for ML and 0.1 for the others.

We present simulation results of knowledge transfer on CIFAR-100. Here, we use a bidirectional knowledge transfer, which exchanges knowledge between two networks *Net 1* and *Net 2* in a mutually beneficial way [8]. From the results, we observe that the proposed technique improve the accuracy of the conventional technique [8], resulting in up to 5% improvement of the accuracy (see Table I and II).

We also present simulation results of knowledge transfer on tiny imagenet. From the results, we observe that the proposed technique improve the accuracy performance of KD, resulting in up to 61.36% (see Table III).

V. CONCLUSION

In this work, we propose to change the structure of a teacher network to get the effect of multiple teacher networks. In our proposed approach, we obtain multiple teacher networks without additional resources so that compact networks improve further with the help of more extensive knowledge. The proposed structure can be easily applied to other transfer methods and tasks, e.g object detection and segmentation.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, June 2016.

[2] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," arXiv preprint arXiv:1602.02410, 2016.

[3] Y. Wu, "Google's neural machine translation system: Bridging the gap between human and machine translation," CoRR, Sep. 2016.

[4] W. Kim, Y. Ahn, and B. Shim, "Deep Neural Network Based Active User Detection for Grant-free NOMA Systems," IEEE Trans. Commun., vol. 68, no. 4, pp. 2143-2155, Apr. 2020.

[5] L. Nguyen, J. Kim, and B. Shim, "Low-Rank Matrix Completion: A Contemporary Survey," IEEE Access, vol. 7, no. 1, pp. 94215-94237, Jul. 2019.

[6] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," IEEE Wireless Commun., vol. 25, no. 3, pp. 124–130, Jun. 2018.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," Comput. Sci., vol. 14, no. 7, pp. 38–39, 2015.

[8] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4320-4328.

[9] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in Proc. SIGKDD, 2017, pp. 1285–1294.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[11] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, 2016.

[12] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in Interspeech, 2016, pp. 3439–3443.

[13] B. B. Sau and V. N. Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," arXiv preprint arXiv:1610.09650, 2016.
[14] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "Disturblabel: Regularizing cnn on the loss layer," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4753–4762
[15] G. Pereyra *et al.*, "Regularizing neural networks by penalizing confident output distributions," arXiv preprint arXiv:1701.06548, 2017.
[16] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in Proc. CVPR, 2016, pp. 2818–2826.
[17] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 550–558.
[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Rep. TR-2009, 2009.
[19] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.