Stochasticity and Skip Connection Improve Knowledge Transfer

Kwangjin Lee, Junhan Kim and Byonghyo Shim
Department of Electrical and Computer Engineering, Seoul National University
Email: {kjlee, junhankim, bshim}@islab.snu.ac.kr

Abstract

Deep neural networks have achieved state-of-the-art performance in various fields, but they have to be scaled down to be used for real-world applications. As a means to reduce the size of a neural network while preserving its performance, knowledge transfer has brought a lot of attention. One popular method of knowledge transfer is knowledge distillation (KD), where softened outputs of a pre-trained teacher network help train student networks. Since KD, other transfer methods have been proposed, and they mainly focus on loss functions, activations of hidden layers, or additional modules to transfer knowledge well from teacher networks to student networks. In this work, we focus on the structure of a teacher network to get the effect of multiple teacher networks without additional resources. We propose changing the structure of a teacher network to have stochastic blocks and skip connections. In doing so, a teacher network becomes the aggregate of a huge number of paths. In the training phase, each sub-network is generated by dropping stochastic blocks randomly and used as a teacher network. This allows training the student network with multiple teacher networks and further enhances the student network on the same resources in a single teacher network. We verify that the proposed structure brings further improvement to student networks on benchmark datasets.

Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performances on complex tasks like computer vision (He et al. 2016), language modeling (Jozefowicz et al. 2016), and machine translation (Wu et al. 2016). Moreover, they surpass human ability in several fields including image classification (He et al. 2016), the go game (Silver et al. 2016), voice generation (Oord et al. 2016), and so on. Despite their superior performance, it is difficult to use DNN-based models because of limited memory and computational resources in the embedded systems. To deal with this problem, many studies have been done to make DNNs smaller but efficient to be applicable in resource limited cases. One of them is knowledge transfer (KT), which train a smaller network with the information of large model's information. Knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), which is an initial work of KT, has brought much attention. The basic idea of KD is to train a small network (called *student network*) with the help of softened outputs of a pre-trained large network (called *teacher network*). The teacher network is usually more suitable for extracting the structure of the data than the student network. Thus, if the student network mimics the structure taught by the teacher network, the student network can be further optimized to the dataset. This is achieved by minimizing cross-entropy loss not only with labels but also with the outputs of the teacher network. Over the years, various studies on KT have been done. Some works propose to get additional knowledge from intermediate layers of teacher networks (Romero et al. 2014; Zagoruyko and Komodakis 2016a; Yim et al. 2017). In another work, a peer-teaching paradigm where networks exchange knowledge each other is used instead of a teacher-student paradigm (Zhang et al. 2018).

The primary goal of this paper is to make a single teacher network to behave as multiple teacher networks. Since multiple teacher networks provide various outputs on a given input, they can provide more extensive knowledge than a single teacher network does. It has been shown that student networks improve further with multiple teacher networks which are used as an ensemble or separately (Hinton, Vinyals, and Dean 2015; You et al. 2017; Zhang et al. 2018). However, using multiple teacher networks is a resource burden and delays the training process. In this work, we propose to add stochastic blocks and skip connections to a teacher network. In doing so, we can get the effect of multiple teacher networks in the same resource of single teacher network. A stochastic block is a block that falls with a fixed probability in the training phase and weighted by its survival probability in the inference phase (Huang et al. 2016). Skip connections make huge number of paths in the network and function as memory which link the information of previous parts and later parts even if stochastic blocks drop. In the training phase, different sub-networks are generated resulting from stochastic drop in the teacher network for each batch. The sub-networks still have reliable performances since there still exist valid paths. Each sub-network becomes a teacher network for each batch, so the student

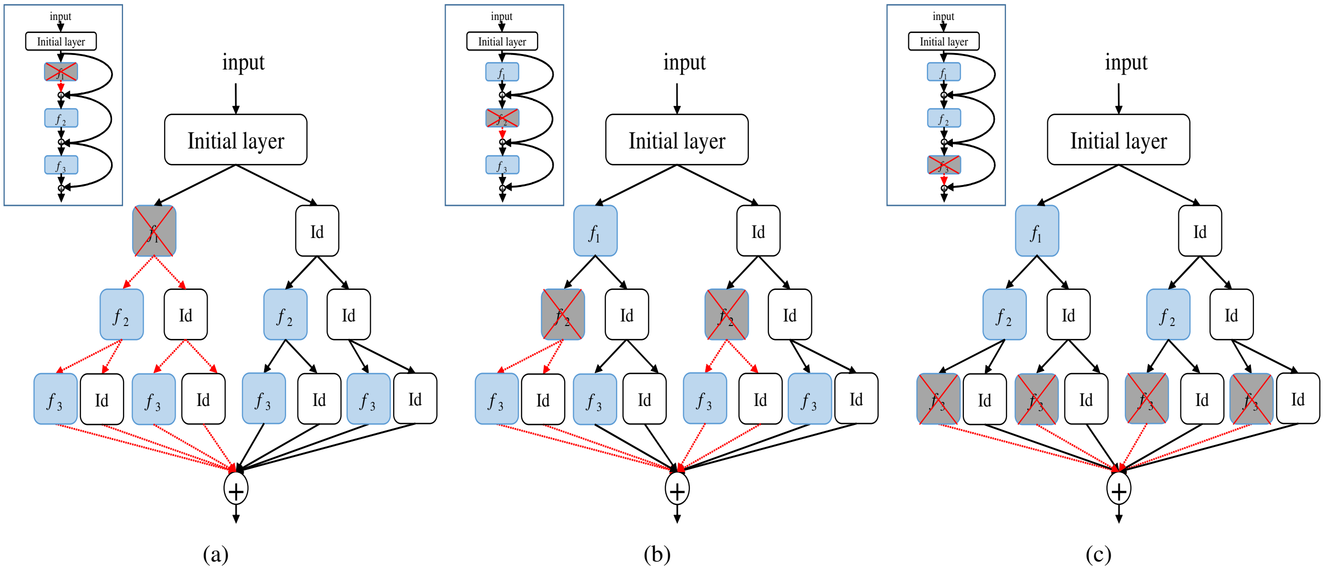


Figure 1: (a), (b), (c) Residual network with stochastic blocks when one block is dropped.

network is trained with multiple teacher networks in the entire training phase. Figure 1 is example of sub-networks generated by dropping one block each from a network with the proposed structure. The networks consists of 3 blocks and f_i, *Id* represents the ith block of the network $(i \in 1, 2, 3)$ and an identity block generated by a skip connection respectively. Red arrows in the figure mean that the outputs of the blocks are 0. In Figure 1, even if one block drops, each sub-network still has 4 valid paths of 8 total paths. We observe that : (i) multiple teacher networks are generated from a single teacher network with no more resources; (ii) generated networks provide different knowledge to a student network; (iii) the performances of student networks improve with the help of a teacher network of the proposed structure. We succeeded in training the student network to perform better than the ones with the same architecture trained by the knowledge transfer methods (KD) (Hinton, Vinyals, and Dean 2015), attention transfer (AT) (Zagoruyko and Komodakis 2016a), and mutual learning (ML) (Zhang et al. 2018)) over CIFAR-100 (Krizhevsky, Hinton, and others 2009) and tiny-imageNet (Russakovsky et al. 2015) datasets.

The rest of this paper is organized as follows. First, we review recent studies related to our work. Then, we demonstrate the proposed scheme with details. After this, we present experiments and discuss the results. Finally, summary and concluding remarks are given in the conclusion.

Related Works

Knowledge Transfer

Knowledge transfer of neural networks has been proposed over a decade ago (Buciluǎ, Caruana, and Niculescu-Mizil 2006) but has recently received much attention with some intuitions and a generalized approach (Hinton, Vinyals, and Dean 2015). There, softened outputs of a teacher network are used to transfer knowledge to a student network. They demonstrate that the softened outputs of a teacher network provide a student network with additional supervision and prevent the student network from overfitting. Later, distillation has been applied in transferring knowledge from powerful and easy-to-train networks to small but hard-to-train networks (Romero et al. 2014). Romero et al. suggest intermediate outputs of teacher networks to be used as *hints* for student networks. An attention-based distillation method makes use of attention maps of teacher networks which are made from feature maps (Zagoruyko and Komodakis 2016a). To transfer knowledge while avoiding direct mimicry, (Yim et al. 2017) exploits flows calculated by Gram matrix of feature maps from two layers of a teacher network, then a student network is trained to mimic the flows of the teacher network. Recently, mutual learning (Zhang et al. 2018) suggests a new paradigm of bidirectional knowledge transfer. All networks in mutual learning are not fixed and exchange knowledge unlike conventional teacher-student paradigm where teacher networks are fixed and student networks only get knowledge.

Multiple Teacher Networks

Student networks can be improved further with the help of multiple teacher networks (You et al. 2017). The dissimilarity between teacher networks provide extensive knowledge to a student network and help to further enhance the student network. Similarly, (Zhang et al. 2018) shows that a neural network can be further improved with the help of multiple neural networks for vision tasks such as image classification and person re-identification. Also, (Chebotar and Waters 2016) shows that multiple teacher networks are more helpful than a single teacher in speech recognition. Most of the distillation methods improve the performance of student networks with multiple teacher networks, but deploying them is demanding due to additional resources. Instead of directly using multiple teacher networks, (Sau and Balasubramanian 2016) suggest perturbing the outputs of a teacher network to get the effect of multiple teacher networks. However, perturbing outputs with noise can be problematic as it changes the values of the outputs so that corrupted knowledge of the teacher network is transferred. In our proposed structure, multiple networks of valid paths are generated (see Figure 1) so that reliable and various outputs are transferred to the student network and provide flexible knowledge.

Regularizing Output

In reinforcement learning, encouraging the policy to have an output distribution with high entropy has been used to improve exploration. This prevents the policy from converging early and leads to improved performance (Williams and Peng 1991; Mnih et al. 2016). Also, penalizing confident outputs (Pereyra et al. 2017) and smoothing label (Szegedy et al. 2016) are proved to help the training of a deep neural network. Regularizing the high confident outputs helps training the deep neural network since it prevents over-fitting of the network and a big difference between values of outputs so that the adaptivity of the network increases.

In the same vein, high confident outputs of a teacher network are challenging for student networks to learn. In ML (Zhang et al. 2018), it has been shown that the ensemble of multiple networks is a worse teacher than the individual networks. Individual networks provide higher entropy outputs than the ensemble, so that the salient secondary values in the outputs can be more helpful generalizing student networks.

Proposed Structure

To get the effect of multiple teacher networks from a single teacher network, we propose to add stochastic blocks and skip connections to the teacher network. In this section, first, we explain in detail how to change the structure of the teacher network to make multiple sub-networks. Then, we demonstrate that multiple sub-networks can be used as multiple teacher networks.

Generating Multiple Networks

For ResNet (He et al. 2016) and Wide ResNet (Zagoruyko and Komodakis 2016b), they consist of blocks and contain skip connections. For MobileNet (Howard et al. 2017), we grouped a depth-wise convolution and a point-wise convolution as one block and add skip connections from each input of the block to the corresponding output.

Skip connections in residual networks prevent vanishing gradient problem, so that deeper networks can be trained well. In other respect, skip connections let a residual network to be viewed as an ensemble of multiple paths of different lengths (Veit, Wilber, and Belongie 2016). When we set ith block of a residual network as f_i, then the output (o_{i+1}) of the $(i+1)$ th block is expressed as follows.

$$o_{i+1} = f_{i+1}(o_i) + o_i \tag{1}$$

Since there are two paths from a previous output to the next output, if there are n blocks in the network, 2^n paths exist from the input layer to the output layer.

It has been shown that, to some extent, changing the structure of a residual network do not harm the performance much (Veit, Wilber, and Belongie 2016). Especially, deleting blocks of residual networks does not harm the performance much. This is because there still exist valid paths even if some blocks of a residual network drop (if k blocks are dropped from n blocks, valid 2^{n-k} paths still exist). Therefore, when a neural network consists of blocks and contains skip connections, multiple neural networks with adequate performances are generated by dropping blocks randomly. To implement this idea in training phase, we set blocks of neural networks to be stochastic as in (Huang et al. 2016). Since initial blocks extract low-level features that will be used by later blocks, we choose *linear decay* mode to set the survival probability of each block. P_{end} denotes the survival probability of the last block and $p^i_{survival}$ denotes that of the ith block expressed as

$$p^i_{survival} = 1 - (1 - p_{end}) \times \frac{i}{N-1}, \tag{2}$$

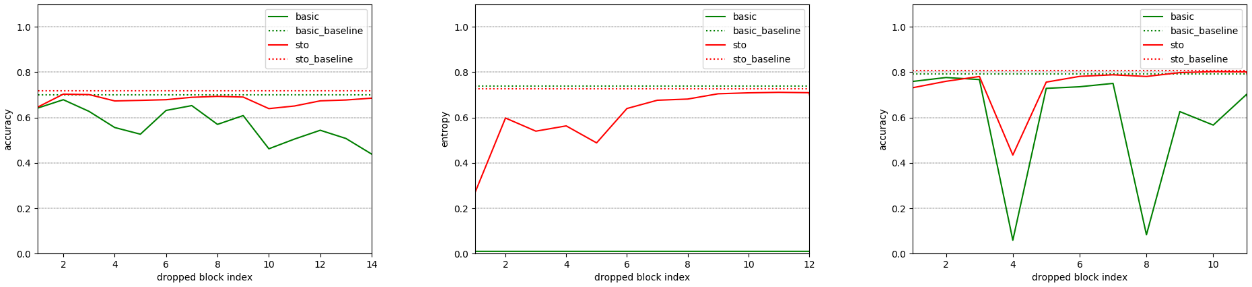
where N is the number of total blocks and $i = 0 \in \{0, 1, ..., N-1\}$.

P_{end} implies a trade-off between quantity and quality of sub-networks. If p_{end} is high, each generated sub-network will have longer length, so the performance will be better than sub-networks with shorter lengths. However, high p_{end} generates less sub-networks. In the opposite case, more sub-networks are generated but each performance can be a bit lower. Optimal p_{end} seems different for teacher and student pairs. We tried p_{end} ranging [0.5, 0.9] with interval 0.1 and choose p_{end} that improves student networks most for each teacher and student pair.

Attributes of the Proposed Structure

One might wonder if sub-networks can play the role of teacher networks and provide independent knowledge so that student networks get sufficient knowledge. For a residual network of 110 layers, it has been shown that sub-networks generated by dropping some blocks show competent performance and are independent each other (Veit, Wilber, and Belongie 2016).

Convolutional neural networks (CNNs) based on residual networks will probably have the same characteristic, however, other networks like mobile network are not guaranteed to generate reliable sub-networks with the proposed structure. To verify the efficacy, we show the accuracy when each block is dropped from pre-trained networks for three kinds of networks. Figure 2 is the accuracy result when each block drops from residual network of 32 layers, mobile network, and wide residual network 28-10. In Figure 2, *sto* and *basic* represents networks with the proposed structure and original networks respectively. *Sto* networks are stronger against dropping blocks than basic networks, so we pre-train teacher networks with the proposed structure. It seems that dropping initial blocks of mobile network and 4th block of wide resnet 28-10 degrades the performance significantly. To observe the impact of such blocks that are fatal to drop, we compare cases where the blocks drop or does not drop like other blocks in ablation study.

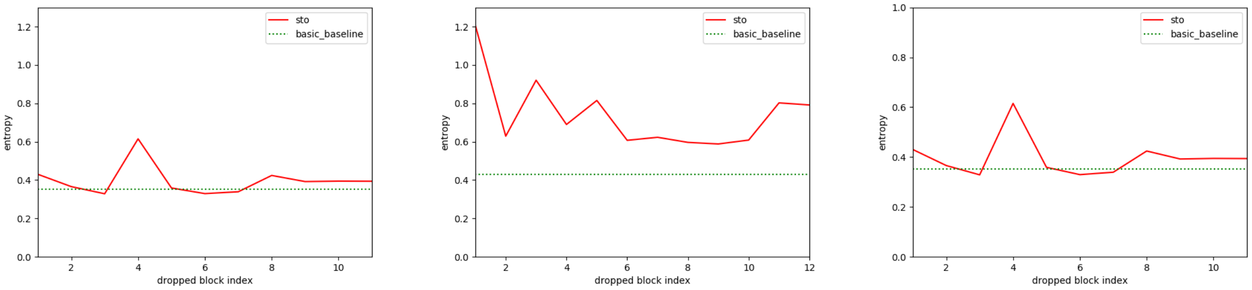


(a)

(b)

(c)

Figure 2: Entropy when each block is dropped from (a) residual network 32, (b) mobilenet, and (c) wide residual network 28-10.



(a)

(b)

(c)

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

The performances of sub-networks lag behind the original network. However, they sometimes predict correctly while the original network does not. Also, they generate outputs with high entropy which are easier for student networks to learn (see Figure 3). It is known that regularizing a neural network to be less confident improves performance (Pereyra et al. 2017). Similar results are also observed in deep mutual learning. In (Zhang et al. 2018), they show 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 outputs have low entropy, which means that the secondary values of outputs becomes small. The secondary values are salient cues in transferring knowledge as it provides important information like relations between classes. Dropping blocks of the teacher network is analogous to using individual networks instead of the ensemble of them. Thus, sub-networks can provide student networks with rich knowledge. Also, knowledge of the original network is fully utilized as all the blocks of the network are used in generating sub-networks in the end.

Generated sub-networks share considerable parts of the original network but provide different knowledge to a student network. The degree of the difference is similar to that of individual neural networks. We confirm the similarity with resnet 32 and attach the related table in the appendix.

Application to Other Distillation Techniques

We apply the proposed method to other distillation techniques, KD, AT, and ML. To apply to KD and AT, the teacher network is changed to have skip connections and stochastic blocks and other settings are not changed. In mutual learning, the notions of teacher and student vanish since both networks give and take knowledge each other. But for convenience, we denote a network with large capacity as a teacher network and the other as a student network. The teacher network is changed into proposed structure as in KD and AT. To apply the proposed structure to mutual learning, both networks should be pre-trained since teacher networks are not fixed. If the networks are not pre-trained, they cannot be improved because of the stochastic property of the teacher network. Let's assume a situation when both networks are not pre-trained. At the beginning of training process, sub-networks of a teacher network are randomized. In mutual training, each sub-network and a student network exchange knowledge. However, since a different sub-network is used each time, for many epochs, the student gets random knowledge from randomized sub-networks so that it does not improve. Also, sub-networks are not optimized due to the disturbing knowledge from the student network.

In our simulation, multiple sub-networks are not used at the same time but one sub-network generated by stochastic drop is used as a teacher network for each batch.

Table 1: Improvement of knowledge distillation (KD) with the proposed structure on CIFAR 100

Net 1	**Net 2**	**independent**		**KD**	**ours**	p_{end}
ResNet 32	VGG 13	74.08	67.74	68.83	71.12	0.8
ResNet 110	ResNet 20	71.69	69.86	70.12	73.36	0.6
WRN 28-10	ResNet 32	78.98	69.86	69.85	74.87	0.5
MobileNet	ResNet 32	74.08	69.89	69.88	71.77	0.7
ResNet 110	ResNet 32	71.69	69.86	70.12	73.36	0.7
MobileNet	VGG 13	74.08	67.74	68.83	71.12	0.7

Table 2: Improvement of attention transfer (AT) with the proposed method on CIFAR 100

Net 1	**Net 2**	**independent**		**AT**	**ours**	p_{end}
ResNet 110	ResNet 20	71.69	68.32	68.34	68.67	0.6
WRN 28-10	ResNet 32	78.98	69.86	69.87	70.64	0.7
ResNet 110	ResNet 32	71.69	69.86	70.22	71.23	0.7
WRN 40-4	ResNet 32	75.67	69.86	70.03	70.59	0.7
WRN 28-10	WRN 40-4	78.98	75.67	75.36	76.09	0.7

Table 3: Improvement of mutual learning (ML) with the proposed structure on CIFAR 100

Net 1	**Net 2**	**independent**		**ML**		**ours**		p_{end}
ResNet 32	ResNet 32	69.86	69.86	71.14	71.21	73.68	73.58	0.9
MobileNet	ResNet 32	74.08	69.86	75.62	71.1	76.2	72.76	0.8
WRN 28-10	ResNet 32	78.98	69.86	78.53	72.18	80.65	73.08	0.5
MobileNet	MobileNet	74.08	74.08	75	75.16	75.5	76.1	0.9
WRN 28-10	MobileNet	78.98	74.08	78.83	76.41	81.03	76.82	0.5
WRN 28-10	WRN 28-10	78.98	78.98	78.83	78.95	81	80.66	0.5

Table 4: Improvement of knowledge distillation (KD) with the proposed structure on tiny imagenet

Net 1	**Net 2**	**independent**		**KD**	**ours**	p_{end}
ResNet 32	VGG 13	49.01	44.61	55.76	57.56	0.9
ResNet 32	ResNet 20	49.01	46.85	49.57	50.6	0.9
MobileNet	ResNet 20	55.38	46.85	51.8	52.15	0.7
MobileNet	ResNet 32	55.38	49.01	54.48	54.85	0.8
MobileNet	ResNet 110	55.38	52.32	58.15	58.2	0.9
WRN 28-10	ResNet 32	58.91	49.01	55.7	55.34	0.6

Experiment

Dataset and Simulation Setting

We evaluate the proposed method with two datasets - CIFAR-100 (Krizhevsky, Hinton, and others 2009) and tiny imagenet (Russakovsky et al. 2015). CIFAR-100 dataset

Table 5: Improvement of attention transfer (AT) with the proposed method on tiny imagenet

Net 1	Net 2	independent		AT	ours	p_{end}
ResNet 110	ResNet 20	52.32	46.85	51.49	51.9	0.6
WRN 28-10	ResNet 32	58.91	49.01	53.56	54.15	0.7
ResNet 110	ResNet 32	52.32	49.01	54.52	54.91	0.8
WRN 40-4	ResNet 32	55.19	49.01	54.33	54	0.7
WRN 28-10	WRN 40-4	58.91	55.19	60.98	61.36	0.8

Table 6: Comparison

Net 1	Net 2	independent		partial	full	p_{end}	
WRN 28-10	ResNet 32	78.98	69.86	74.81	74.87	0.5	0.5
mob	ResNet 32	74.08	69.86	70.3	71.77	0.9	0.7

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 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. The data augmentation includes horizontal flips and random crops from image padded by 4 pixels on each side, filling missing pixels with reflections of original image. Each network is trained for 200 epochs with batch size of 128 and learning rate which is decreased at every 60 epochs. For tiny imagenet, we simulate with the pure dataset without augmentation. Each network is trained for 100 epochs with batch size of 128 and learning rate which is decreased at every 40 epochs. Stochastic gradient descent optimizer with momentum of 0.9 is used for the whole simulation. The initial learning rate is 0.01 for ML case and 0.1 for the other cases. 4 CNNs are used - wrn, resnet, mobilenet, and vgg net (Simonyan and Zisserman 2014). CNNs are modified to the proposed structure when they are used as teacher networks. All the results in the simulation are averaged over 3 times.

CIFAR-100

Here, we present simulation results of knowledge transfer methods on CIFAR-100. Table 1 is the simulation results of KD and KD with the proposed structure. As you can see in Table 1, we confirm that the proposed structure further improves performances of student networks on KD. In case of (WRN 28-10, ResNet 32) pair, the accuracy of ResNet 32 trained with the proposed structure improves more than 5% compared to when ResNet 32 is trained the pure WRN 28-10.

Table 2 is the simulation results of AT and AT with the proposed structure. In AT, attention maps of teacher and student networks should have same spatial size. So, we used residual networks and wide residual networks to fit the spatial size conveniently. Attention maps are made by square sum via channel axis and l2 normalization. The proposed structure show further improvement over the pure AT method. We confirm that the proposed structure improve student networks further with AT method in all the pairs.

Table 3 is the simulation results of ML and ML with the proposed structure. Only teacher networks are change to the proposed structure, as mentioned in the previous section. And a teacher networks and a student network exchange knowledge each other. The network pairs in Table 3 are same with those of the paper (Zhang et al. 2018). The proposed structure still show further improvement in peer learning paradigm, so both networks are improved further.

Tiny Imagenet

Here, we present simulation results of knowledge transfer methods on tiny imagenet. Table 4, 5 are the simulation results of KD, AT, and proposed structure. As the simulation results on CIFAR-100, the proposed structure improves student networks generally, but there exist one pair each for KD, AT that the student network is not improved.

Ablation Study

In Figure 2, when some blocks drop, then, performances of neural networks drop significantly. The blocks are the 4th block of wrn 28-10 and 1st to 6th blocks of mobilenet. We name these blocks significant blocks. Sub-networks generated by dropping significant blocks have low performance so that the networks might not be adequate teacher networks. Hence, we observe if student networks improve further, when sub-networks generated by dropping the other blocks except the significant blocks are used as teacher networks. We use KD and CIFAR-100 dataset for (wrn 28-10, resnet 32) and (mobilenet, resnet 32) pairs.

In Table 6, partial means that significant blocks do not drop and full means that all the blocks drop stochastically in training phase. The results show that using more teacher networks is more helpful improving a student network even if some of them do not perform well. This is in line with the result of ML (Zhang et al. 2018) where larger network still benefits from being trained together with a smaller network.

Conclusion

In this work, we propose to change the structure of a teacher network to get the effect of multiple teacher networks in the same resource of one teacher network. In our proposed structure, we obtain multiple teacher networks without additional resource so that compact networks improve further than those trained from conventional transfer methods. The proposed structure can be easily applied to other transfer methods and tasks, e.g. segmentation or object detection.

Acknowledgments

References

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541. ACM.

Chebotar, Y., and Waters, A. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, 3439–3443.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *European conference on computer vision*, 646–661. Springer.

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

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.

Sau, B. B., and Balasubramanian, V. N. 2016. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In *Advances in neural information processing systems*, 550–558.

Williams, R. J., and Peng, J. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science* 3(3):241–268.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4133–4141.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294. ACM.

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

Zagoruyko, S., and Komodakis, N. 2016b. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE Confer-*

ence on Computer Vision and Pattern Recognition, 4320–4328.