Semi-Supervised Knowledge Distillation Via Teaching Assistant

Limuxuan He, Bei Hua *

School of Computer and Electronic Information Guangxi University Guangxi University Guangxi, China

* Corresponding Author Email: 549383258@qq.com, limuxuanhe@163.com

Abstract. As deep neural networks are widely used with computer vision, Model Compression Methods for knowledge distillation are being actively investigated in order to deploy them into smaller devices. However, when there are significant differences between models for students and teachers, a large amount of labeling wastes a great quantity of manpower while the performance of student learning decreases. In this paper, we propose an approach that uses semi-supervised teacher assistants to fuse knowledge distillation with teachers, effectively bridging the large teacher-student gap on a large number of unlabeled datasets and a small number of labeled datasets. We enable unsupervised teacher pre-training, fine-tuning, followed by teacher-assistant offline distillation, student-teacher-assistant oline distillation, and student-teacher offline distillation. We validate the effectiveness of the proposed method for the classification task using CIFAR10, CIFAR-100, and ImageNet.

Keywords: deep neural networks, adversarial learning, semisupervised, teacher assistant, knowledge distillation.

1. Introduction

In recent years, convolutional neural networks (CNNs) have been used in various fields of artificial intelligence [1], [11], however, when the neural network is huge, the large computational power and storage costs requirements greatly hinder its use in small devices such as smartphones, therefore, in order to address the problem of retaining both the high performance of large neural networks and the low cost of small neural networks, the knowledge distillation (KD) method [12] is proposed.

KD has been used in various fields such as speech recognition, image recognition, NLP (natural language processing), etc. The KD technique solves the problem of high computational complexity and huge storage in real-time applications. However, due to the uncertainty of knowledge, students cannot effectively learn the teacher network knowledge when the network parameters differ greatly between the teacher and the student [6], [17], therefore, Teacher Assistant Knowledge Distillation (TAKD) [18], an intermediate model, is proposed to solve the problem, and since students are extracting knowledge from the Teacher Assistant (TA), they will learn the wrong knowledge when the TA has problems, therefore, the teacherTA integration approach to improve the robustness of students.

However, most of the KDs, such as [15], [19], are based on large datasets and high computational cost, when labeling large amount of data is particularly important, so to reduce the labeling effort and train using a small amount of labeled and a large amount of unlabeled datasets, a semisupervised learning algorithm is proposed [24], where semisupervised learning consists of unsupervised or self-supervised pre-training followed by supervised fine-tuning. The method of [2] greatly improves accuracy on large neural networks by the ”unsupervised pre-training, supervised fine-tuning” paradigm. Supervised contrast learning is a variant of contrast learning [14], in which positive/negative samples are data with the same/different labels. In order to improve the generalization ability of the model with severely distorted data and to maintain the high performance of the teacher assistant while reducing the tedious process of labeling, the proposed semisupervised TA with teacher cooperation knowledge distillation(SCKD) is proposed.

In summary, several steps are included in the proposed SCKD, they are (1) unsupervised pre-training of unlabeled data by the teacher network, (2) supervised fine-tuning of labeled data by the teacher network, (3) distillation by the teaching assistant, a medium model smaller than the teacher network and larger than the student network, using unlabeled data, and finally (4) distillation by the student network using un labeled data, while using the TA and teacher networks for knowledge extraction. We evaluate the effectiveness of this approach on CIFAR-10 , CIFAR-100 and ImageNet, and our contributions are mainly in the following areas.

(1) In order to bridge the gap between teachers and students, we introduce TA, and by collaborating with teacher, students extract knowledge from both TA and teacher, expanding the scope of student's learning knowledge and effectively avoiding errors caused by learning errors of TA.

(2) For large neural network models, the required dataset will be larger. In order to reduce the dataset labeling work, semi-supervised learning is added to the distillation process, which greatly improves the accuracy of student model prediction in the case of a huge gap between teacher and student models.

(3) In the process of distillation, a combination of offline distillation and online distillation is used. Due to the huge parameters of the teacher network, the TA extracts the teacher knowledge by offline distillation, which is simple and fast, and the student extracts the teacher knowledge by offline distillation and the TA knowledge by online distillation, and the TA and the student learn from each other, which effectively improves the performance of the student network by combining offline distillation and online distillation.

2. Background

The idea of knowledge distillation i.e. the student network (y_S) is trained using not only the information provided by the true labels y but also fitting the information provided by the soft labels of the teacher network (y_S). The student model is learned using the optimized cross-entropy loss:

$$L_{CE} = H(softmax(z_S), y) \tag{1}$$

Where τ is the temperature parameter controlling the softening of the signal, $H(x,z) = -\sum_{c=1}^{C} x_c \log(z_c)$ is the crossentropy, z_T, z_S and refer to the teacher and student logits, respectively.

The Kullback-Leiblr (KL) scatter of the loss L_{KD} is derived from the softened output of the teacher and student networks as follows:

$$L_{KD} = \tau^2 KL\ (y_S, y_T) \tag{2}$$

Where $KL(x,z) = \sum_{c=1}^{C} x_c \log(x_c/z_c)$ is the KL divergence, $y_S = softmax(z_S/\tau)$, $y_T = softmax(z_T/\tau)$ are the outputs of the student and teacher networks.

Thus, the final loss function of the regular KD is written in terms of the equilibrium parameter λ as follows:

$$L = (1-\lambda)L_{CE} + \lambda L_{KD} \tag{3}$$

Where $\lambda \in [0,1]$ is a constant hyperparameter weighing the contribution of each term.

KD is an effective method to address model compression. KD is a method by which a large teacher network instructs a small student network for training by pre-training. Hinton [12] proposed that the student network extracts knowledge from the teacher network. Since the optimal ratio before soft and hard labels is difficult to determine, [25] proposed efficient regularization methods such as contrast learning [22]to prevent overfitting and improve performance. However, KD is not always effective when there are huge differences between teacher and student models, Hariharan [10] proposed to stop teacher training as early as possible to recover a more suitable student network, [18] proposed a weighted integration of multiple TAs with gradual distillation to improve student performance, and [21] proposed to use multiple TAs with dense bootstrapping to improve robustness. [7] proposed

online co-distillation by using a group of students trained simultaneously and students making predictions from their peers. One-stage online distillation [23], [27] effectively avoids the high computational cost of two-stage offline distillation [9].

In summary, in order to simultaneously balance low computational complexity and high accuracy in the case of a vastly different teacher network and student network, we propose a combination of weighted online distillation and offline distillation using teacher to collaborate with TA, while introducing adversarial learning to prevent overfitting and at the same time reducing data annotation in the case of large data sets.

3. Methodology

3.1. Semi-supervised learning

Adversarial learning gives better results compared to supervised learning when we need more long-term training and larger teacher network sizes. Inspired by semi-supervised learning [2], [13], [5] proposed the SimCLRv2 framework and found that self-training and fine-tuning using a larger network in pre-training can greatly improve accuracy, and then use unlabeled data to encourage the student network to mimic the predicted labels of the teacher network. We use the SimCLRbased framework to incorporate semi-supervised learning in the joint distillation.

Our KD is divided into four stages as in Fig. 1, i.e., teacher pre-training, teacher fine-tuning, teaching assistant learning from the teacher, and student learning from the teaching assistant and the teacher, using unlabeled data when the teacher learns in a task-independent manner, learning general image features through unsupervised pre-training, and then applying general features to specific features through fine-tuning. When teaching assistants on unlabeled data, it is a task-specific approach to further improve predictive performance, and students continue to use unlabeled data to improve model prediction and compact models. The method can be summarized into four main steps: pre-training, fine-tuning, TA distillation, and student distillation.

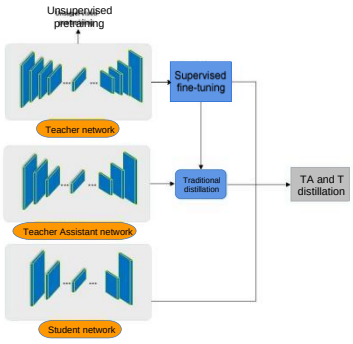


Fig. 1. Semi-Supervised Teacher Assistants and Teacher Integration Distillation Process

Contrast learning aims to help our model distinguish features that are invariant to input transformations, including data enhancement and corruption. In the first stage, self-supervised contrast loss L_C is proposed to adjust image representations:

$$L_C = -\frac{1}{N}\sum_{(h,h^+)\in\mathbb{P}} \log \frac{e^{s(h,h^+)/\tau_C}}{\sum_{h'\neq h}^{N} e^{s(h,h')/\tau_C}} \tag{4}$$

where P is composed of 2N positive examples and $s(.,.)/\tau_C$ is a cosine similarity function. In the second stage for finetuning, supervised contrast learning takes identically labeled data as positive samples, using the supervised contrast loss *Lhard*:

$$L_{hard} = -\frac{1}{N}\sum_{i}^{N}\sum_{j\neq i}^{N} \log l_{y_i=y_j} \frac{e^{s(h_i,h_j)/\tau_{SC}}}{\sum_{k\neq i}^{N} e^{s(h_i,h_k)/\tau_{SC}}} \tag{5}$$

3.2. Teacher Assistant

In the third stage we distill the fine-tuned network into a smaller network, when the gap between the teacher and student model networks is small, KD can optimize the student model to a level close to that of the teacher. However, when the difference in weight parameters or number of layers between the student model and the teacher model is too large, the student will not learn effectively from the teacher, so TAKD is proposed to bridge the large gap between the student and teacher models using the TA medium-sized network.TAKD improves the student's learning efficiency by gradual distillation. Although teacher assistants are smarter than students, however, TA still lacks the broad expressiveness of teacher. Also, since a single model is used for knowledge transfer, when TA makes mistakes in learning, it will cause an error avalanche problem if the wrong knowledge is transferred to students. Therefore, to achieve a lightweight network that can still reach the teacher level by distillation in the presence of huge model differences, we propose semi-supervised teacherTA fusion distillation, which we call SCKD. In the third stage we distill the fine-tuned teacher network into TA , the loss of TA L_{TA} as follows:

$$LTA = LT \rightarrow TA \tag{6}$$

where the $T \rightarrow TA$ indicates the teaching direction of teacher to TA.

The student's loss L_S is derived as follows:

$$L_S = L_T \rightarrow S + L_{TA} \rightarrow S \tag{7}$$

where the $TA \rightarrow S$ indicates the teaching direction of TA to student.

We used the knowledge extracted by a teacher assistant with a teacher to teach students and obtained the following distillation losses:

$$L = (1 - \alpha)LCES + \alpha LKDT \rightarrow S + \tag{8}$$

$(1 - \beta)LCES + \beta LKDTA \rightarrow S$

where α and β are the balance parameter.

3.3. Oline Distillation

Hinton [12] proposed knowledge distillation with offline distillation. Most knowledge distillation applies offline distillation, which is divided into two stages. The first stage is to train the teacher model to obtain the teacher's knowledge. The second stage is to extract knowledge from the teacher model, and then use logits or intermediate The features guide the student model. Since the first stage is pre-trained, this paper will mainly focus on the effectiveness of knowledge transfer. The main advantage of offline distillation is that it is quick and easy to use and understand. The student's performance will drop for the offline distillation has a large gap between the teacher and the student. As one possible solution, an online distillation is proposed to improve this performance. The teacher and the student update at the same time during the distillation to realize online mutual learning. The major advantage of online distillation is the enhancement of efficacy in view that no pre-training for teachers is required, and online learning outperforms offline learning.

Therefore, in the case of a huge gap in the model,as in Fig.2 we propose a joint distillation method of offline distillation and online distillation, and use different distillation methods at different stages to improve the effect of the model.

offline distillation

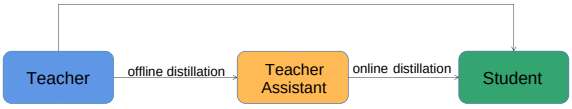


Fig. 2. Offline distillation between TA and teacher, online distillation between student and TA, and offline distillation between student and teacher

4. Experiment

4.1. Experimental Setup

Datasets: We evaluate the method on two standard datasets, CIFAR-10 [3] and CIFAR-100, as well as the large-scale ImageNet [8] dataset, which consists of 32 × 32 RGB images. These datasets consist of 32 × 32 RGB images. CIFAR10, CIFAR-100, and ImageNet contain 10, 100, and 1000 classes, respectively. CIFAR contains 50,000 training images and 10,000 test images. imageNet contains 1282167 training images and 50,000 test images.

Network architecture: we used ResNet [11] network for experiments, and when performing teacher pre-training and finetuning, we performed experiments by changing the depth and width of the network. When performing distillation, teachers used ResNet-152(2×+SE) for teachers, ResNet-101 for TAs, and ResNet-50 for students to perform the experiments.

Tab. 1 Accuracy after fine-tuning using 1% as well as 10% of the labeled examples on top-1 and top-5 of imagenet

	Model	Param	Fine-tuned on 1%10%100%			Linear eval	Supervised
Top-1	ResNet-50 ResNet50(2×) ResNet101 ResNet101(2×)	24 94 43 170	57.8 66.2 62.3 69.0	68.2 74.0 71.2 75.9	76.3 79.2 78.1 80.3	70.7 75.6 73.9 77.2	76.8 77.9 78.1 79.1
	ResNet152	58	64.1	73.2	79.2	74.6	78.5
	ResNet152(2×)	233	70.1	76.2	81.2	77.1	79.2
Top-5	ResNet50 ResNet50(2×) ResNet101 ResNet101(2×)	24 94 43 170	82.4 87.2 85.1 90.0	89.5 91.8 90.8 93.1	93.2 95.0 94.5 95.7	90.5 92.6 91.8 93.2	93.2 94.0 94.1 94.3
	ResNet152	58	86.5	92.0	95.0	92.3	94.4
	ResNet152(2×)	233	89.5	93.4	96.0	93.5	94.6

Tab. 2 The accuracy of resnet after adding se and cbam attention modules.

Dataset	ResNet-152(2×)	ResNet-152(2× +SE)	ResNet-152(2× +CBAM)	ResNet-152(2× +SE+CBAM)
Top-1	79.2	80.9(+1.7)	79.4(+0.2)	81(+1.8)
Top-5	94.6	95.4(+0.8)	95.0(+0.4)	95.7(+1.1)

Implementation details: we used the Pytorch [20] framework for implementation, starting with the same set of simple enhancement methods as SimCLR [4], i.e., random cropping, color distortion, and Gaussian blur, using ResNet-50 as the base encoder network and a 3-layer MLP projection head to project the representation into a 128-dimensional latent space. The LARS optimizer [26] was used for teacher pretraining and fine-tuning, and we trained the network with teacher assistant and student distillation, we use the Adam [16] optimizer, where we set the initial learning rate to $1e^{-5}$, the maximum learning rate to $4e^{-4}$, and the learning rate decay factor to 0.8, with the learning rate increasing linearly at 0 ∼ 10th epochs , remaining the at 11th ∼ 60th epochs and decreasing at 61st ∼ 100th epochs. The nesterov momentum of the stochastic gradient descent (SGD) optimizer is 0.9. To obtain reliable results, we performed all experiments using the hyperparameter optimization toolkit. The hyperparameters include the distillation weight λ and the temperature τ explained in Section III.

4.2. Discussions

1) Pre-training and Fine-tuning: When pre-training of the teacher model was initially performed, it is found in [4] that the gap between the supervised model and the linear classifier trainedon the semi-supervised model shrinks as the model size increases, suggesting that supervised learning benefits more from a larger model than supervised learning. As shown in the Table I, the accuracy of self-supervised and supervised learning under different ResNet model sizes and evaluation protocols are compared. Apparently, with the deepening of the width and depth of the ResNet model, the performance has been improved. Based on analysis of the minimum model and maximum model accuracy on Top-1 for traditional supervised learning,it only increased by 2.4%. However, the selfsupervised model revealed a 6.4% increase, and fine-tuning of 10% of the labeled images witnessed the increase by 6.0% as well as for 1% of the labeled images.The fine-tuning of is improved by 12.3%, which shows that the performance of the model is improved more obviously as the width and depth of the model become larger.

In order to further widen the network and improve performance, attention mechanism SE and CBAM are introduced to the attention module for comparative experiments during fine-tuning. As shown in the Table II, after adding the SE attention module, the accuracy increased by 1.1% on Top-1 and 0.8% on Top-5, while the CBAM attention module did not significantly improve the performance. Therefore, only the SE attention module is selected to improve the performance of the teacher.

Tab. 3 The acruracy on imagenet.teacher network is resnet-152 (2×) after pre-training fine-tuning, ta is resnet-101, and student network is resnet-50

Dataset	Teacher	Student	BLKD	TAKD	SCKD
Top-1	79.2	69.4	70.2	73.4	77.1
Top-5	94.6	80.9	84.3	89.1	91.3

Tab. 4 The acruracy on ciraf-10 and ciraf-100.teacher network is resnet-152 (2×) after pre-training fine-tuning, ta is resnet-101, and student network is resnet-50.

Dataset	Teacher	Student	BLKD	TAKD	SCKD
CIRAF-10	98.6	92.1	93.5	93.4	94.1
CIRAF-100	89.0	77.3	78.4	78.5	79.0

2) Teacher Assistant: We experiment the proposed method on the ResNet network and first test the classification performance of the student model before delving into our proposed method. The second stage implements the traditional knowledge distillation method, i.e., baseline knowledge distillation (BLKD) of the teacher and the student. As can be seen in Tables 1 and 2, the student performance rises by less than 1% because although the student performance is improved using the traditional knowledge distillation method, the teacher model is too large and when the student model is too small, the students still do not have sufficient ability or mechanism to imitate the teacher's behavior despite the cues they receive. In the third stage, we will use a single teacher assistant's for stepwise distillation and use this strategy to alleviate the problem of too large a gap between the teacher and student networks. Compared with the first two results, when the assistant is used between the teacher and student networks, the student performance is closest to that of the teacher on CIRAF-10 and the student is more effective in fitting the logit distribution of the TA. In the fourth stage, the teacher and the weighted average of the teaching assistant's knowledge of both outperformed on all three datasets, improving by 0.5% on CIRAF-10, 0.7% on CIRAF-100, and 3.7% on Top-1 and 0.8% on Top-5 of ImageNet relative to TAKD, effectively improving the robustness and accuracy of the students. As can be seen in Tables III and IV, the proposed method outperforms existing methods and remains effective when the teacher network differs significantly from the student network.

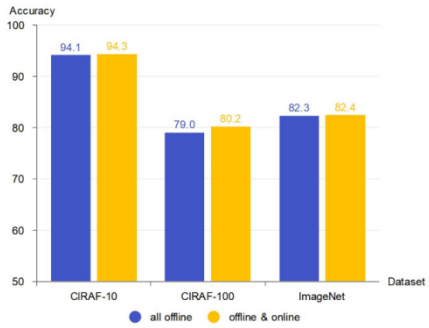


Fig. 3. We compared online distillation of students with TAs to offline distillation of all of them

3) Online Distillation: Since the two-stage KD requires more parameters and computational cost, and the introduction of TAs greatly increases the computational complexity, we improved the robustness by allowing teachers to distill offline with TAs, teacher assistants to distill online with students, and students to distill offline with teachers, as shown in Fig.3, CIRAF-10 improved by 0.2%, CIRAF-100 improved by 0.3%, and ImageNet improved by 0.1%.

5. Conclusions

In this paper, we propose a semi-supervised teacher assistant and teacher fusion distillation method that improves student learning efficiency and reduces the labor cost of label production in the presence of vastly different teacher-student models and a large number of unlabeled datasets and a small number of labeled datasets. Due to the introduction of TAs, we use offline distillation with online distillation to reduce computational cost and storage capacity, further effectively improving the performance of student.

Acknowledgement

This work was supported in part by the National Natural Science Founda-tion Project under Grant No.62062009 and the Guangxi Innovation-DrivenDevelopment Project under Grant No. AA17204058-17 and No. AA18118047-7.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[3] Rasim Caner C¸alik and M Fatih Demirci. Cifar-10 image classification with convolutional neural networks for embedded systems. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–2. IEEE, 2018.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semisupervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[6] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[7] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Featuremap-level online adversarial knowledge distillation. In *International Conference on Machine Learning*, pages 2006–2015. PMLR, 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li FeiFei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701, 2017.

[10] Parameswaran Hariharan, Bozenko F Oreb, and Tomoaki Eiju. Digital phase-shifting interferometry: a simple error-compensating phase calculation algorithm. *Applied optics*, 26(13):2504–2506, 1987.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[15] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

[18] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[19] SeongUk Park and Nojun Kwak. Feed: Feature-level ensemble for knowledge distillation. *arXiv preprint arXiv:1909.10754*, 2019.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[21] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9395–9404, 2021.

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[23] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Selftraining with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[25] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[26] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[27] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-thefly native ensemble. *Advances in neural information processing systems*, 31, 2018.