

Bringing Giant Neural Networks Down to Earth with Unlabeled Data

Yehui Tang, Shan You, Chang Xu, Boxin Shi, *Member, IEEE* and Chao Xu

Abstract—Compressing giant neural networks has gained much attention for their extensive applications on edge devices such as cellphones. During the compressing process, one of the most important procedures is to retrain the pre-trained models using the original training dataset. However, due to the consideration of security, privacy or commercial profits, in practice, only a fraction of sample training data are made available, which makes the retraining infeasible. To solve this issue, this paper proposes to resort to unlabeled data in hand that can be cheaper to acquire. Specifically, we exploit the unlabeled data to mimic the classification characteristics of giant networks, so that the original capacity can be preserved nicely. Nevertheless, there exists a dataset bias between the labeled and unlabeled data, which may disturb the training and degrade the performance. We thus fix this bias by an adversarial loss to make an alignment on the distributions of their low-level feature representations. We further provide theoretical discussions about how the unlabeled data help compressed networks to generalize better. Experimental results demonstrate that the unlabeled data can significantly improve the performance of the compressed networks.

Index Terms—Model Compression, Channel Pruning, Unlabeled Data, Dataset Bias

I. INTRODUCTION

DEEP learning has demonstrated state-of-the-art performance in many tasks, such as object classification [1], [2], speech recognition [3], [4], image generation [5], [6] and so on. The major component underlying these successes is the development of sophisticated deep neural networks (DNNs), *e.g.*, AlexNet [7], VGGNet [8], Inception [9], and ResNet [10]. However, large volume of parameters, huge run-time memory cost and heavy dependence on GPU devices hamper the deployment of these giant DNNs in real-world applications [11], [12]. For example, ResNet-50 [10] needs 95MB memory to store parameters, 97MB memory to store feature maps and 3.8×10^9 times of floating number multiplications to interface with a single image [11]. It has been well known that there is significant redundancy in a large over-parameterized network, and fewer parameters can express the same amount of information as well [13], [14]. It therefore motivates the research on neural network compression.

Yehui Tang and Chao Xu are with the Key Laboratory of Machine Perception (Ministry of Education) and Cooperative Medianet Innovation Center, School of EECS, Peking University, Beijing 100871, P.R. China. E-mail: yhtang@pku.edu.cn, xuchao@cis.pku.edu.cn.

Shan You is with SenseTime Research. E-mail: youshan@sensetime.com. Boxin Shi is with the National Engineering Laboratory for Video Technology, School of EECS, Peking University, Beijing 100871, P.R. China. E-mail: shiboxin@pku.edu.cn.

Chang Xu is with the UBTech Sydney Artificial Intelligence Centre and the School of Information Technologies in the Faculty of Engineering and Information Technologies at The University of Sydney, J12 Cleveland St, Darlington NSW 2008, Australia. Email: c.xu@sydney.edu.au.

Basically, network compression methods can be categorized into several aspects, including parameter quantization [15], [16], low-rank approximation [17], [18], knowledge distillation [19], [20] and pruning (non-structured pruning [21]–[24] & channel pruning [25]). Quantization methods represent weights or activations with low bit integers [15]; even binary weights are used [26], [27] while low-rank approximation takes advantage of tensor factorization techniques and decomposes one giant filter into multiple smaller components. Knowledge distillation focuses on the training of a compact light network and advocates the soft supervision from pre-trained powerful giant networks [19]. As for pruning methods, non-structured pruning [22] removes unimportant weights, which can get extremely high compression rate without accuracy loss; however, special hardware is needed to accelerate the computation in practice. In contrast, channel-wise pruning [25], [28] chooses to remove the whole spatial filters over channels and results in simultaneous reduction on the parameters, memory footprint and computation cost. Note that channel pruning does not destroy the structure of the giant network, so it is compatible with other compression methods and has attracted much attention recently [29].

Existing neural network compression methods have received impressive performance in experiments, but they usually need many iterations of retraining to preserve the original accuracies, especially when the compression ratio is fairly high. An immediate question therefore arises: How to retrain the compressed network if the original training dataset is incomplete? Demo models (*e.g.*, well-trained DNNs) are usually released and ready to import for users. However, due to the consideration of security, privacy or commercial profits, the model providers only supply sample data (sometimes with unknown sources) for verification purpose instead of the complete training set. For example, the website of EmotionNet¹ for emotion recognition only provides example images for display, and the original dataset can only be obtained by the approval from administrators with a rigorous agreement. Likewise, powerful CNNs are trained to predict hashtags on massive Instagram images [30], but the original training dataset is not released as mentioned in the paper except for a few of example images. This situation is more common in medical diagnosis [31], drug discovery and toxicology [32], since their used datasets are usually not completely open-source as discussed in [33]. In addition, for model compression service on the cloud, uploading the entire training dataset (*e.g.*, several GB file size) to the cloud is unpractical and time-consuming due to the limited accessing speed (*e.g.*, several MB/s), while

¹<https://github.com/co60ca/EmotionNet>

the trained networks (usually with MB level size) and part of the dataset can be transmitted easily. All these practical scenarios imply the necessity of compressing models with no access to the complete training dataset. However, retraining the compressed models only with the limited labeled data may incur severe over-fitting issue.

In this paper, we propose to bring giant neural networks down to earth with unlabeled data. Instead of struggling to search for original training data, we turn our attention to unlabeled data in hand that can be cheaper to acquire. The output of giant network reflects its classification characteristics and contains the necessary information for its capacity. Thus we regard unlabeled data as a portal to distill its intrinsic information, and concretely, we exploit unlabeled data to mimic the softened output of the giant network, so that its powerful classification ability can be well preserved. However, since unlabeled and labeled data are usually collected in different ways, there is a dataset bias hampering the performance. We fix this issue by make alignment on the distributions of low-level features between unlabeled and labeled data. A confidence coefficient is also introduced to weight the unlabeled data to reduce the disturbance of improper unlabeled examples. Furthermore, we provide theoretical discussions about how the unlabeled data help compressed networks to generalize better. Experimental results on benchmark datasets demonstrate the effectiveness of exploiting unlabeled data to assist the network compression.

The rest of this paper is organized as follows. Section II reviews the related work for compressing and accelerating networks. Section III makes a brief introduction of channel pruning with scaling factors as preliminaries. Then we formally elaborate in Section IV how to prune networks with unlabeled data and analyze the proposed method from a theoretical perspective. The experimental results and analysis are presented in Section V, with concluding remarks given in Section VI.

II. RELATED WORK

For the compression and acceleration of CNNs, the mainstream works are mainly divided into four categories: quantization, sparse or low-rank approximation, knowledge distillation and pruning (non-structured pruning & channel pruning).

Quantization. It aims to reduce the number of bits for representing each weight or activation in the CNNs. For example, Vanhoucke *et.al.* [34] finds that the 8-bit quantization of weights can induce significant speed-up with almost no drop of accuracy. Binary weights are even investigated to obtain extremely compressed networks, which constrains weights to only two values (*i.e.*, 1 or -1) and most time-consuming multiply-accumulate operations are replaced by simple accumulations [26], [27]. However, binarizing very large networks (*e.g.*, GoogleNet) will incur large accuracy loss. To improve the performance of quantized networks, Li *et.al.* [35] proposes Ternary Weight Network (TWN) constraining weights to ternary values (*i.e.*, -1, 0, 1). Zhu *et.al.* [36] further develops it by learning both ternary values and assignment during training. The proposed Trained Ternary Quantization (TTQ) can be trained from scratch as easy as a normal full-precision model.

Low-rank approximation. Since convolutional filters can be seen as 4D tensors, based on low-rank assumption, they can

be decomposed to multiple components with fewer parameters. Both storage and computation cost can be reduced in the way. For example, SVD method has been studied widely [17], [18] to decompose a tensor into two-layer compact convolutional filters. Those components with large sizes may be still computationally time-consuming, which can be further decomposed [18]. Thus an original redundant filter is replaced by multiple compact filters.

Knowledge distillation. By distilling the knowledge from the pre-trained giant networks, the performance of the target light and small networks can be boosted [19], [37]. Hinton *et.al.* [19] proposes to mimic the informative softened outputs of the teacher network. In addition to the output level, intermediate representations of the giant network can be transferred as hints to assist training. In [38], the attention maps via activations or gradients are also used for mimicking. Besides, You *et.al.* [39] proposes to combine multiple teacher networks. The relative dissimilarity between different examples serves as guidance and a voting strategy is used to unify dissimilarity information provided by various teacher networks. There are also some works [40], [41] extending knowledge distillation to more applications such as multi-task learning [40].

Non-structured pruning. To prune the redundant parameters, an intuitive method is to remove each weight with small magnitude and get a more sparse network. Han *et.al.* [42] proposed to apply l_1 or l_2 regularization to make weights sparse and prune tiny weights in an iterative way. The pruned network can be further compressed with quantization and Huffman encoding, resulting in $35\times$ compression rate on AlexNet without sacrifice for accuracy [43]. To avoid accuracy drop incurred by incorrect pruning weights, splicing operation [44] was introduced to recover the mistakenly removed connections. Pruning and splicing operations constitute the dynamic network surgery framework and obtain more sparse networks with fewer training epochs. However, although high compression and acceleration rates are obtained theoretically, the hardware is needed to be designed specially for realizing practical speed-up. Compared to the fine-grained pruning methods, group-wise pruning methods [45] are more common in practice. Nevertheless, structures of the original networks are destroyed as a result and real inference speed-up also depends on dedicated libraries badly [46], [47].

Channel pruning. Channel pruning methods aim to directly remove the redundant channels without destroying the structure of original networks. After pruning a whole filter of a layer, the channels of the corresponding feature maps are also pruned. Parameters, computation cost and memory footprint are reduced simultaneously. There are mainly two strategies for channel pruning. The first one [48], [49] seeks to identify the important channels layer-by-layer by minimizing the gap of feature maps between the pruned network and the original pre-trained network. Though the layer-wise reconstruction can preserve the information of each individual layer well, it ignores the global information of the entire network which obstructs deciding the width of each layers automatically and achieving high compression rate. The more prevalent strategy is to select important channels for various layers simultaneously and train the whole networks with sparsity regularization [24], [25],

[50], [51]. Slimming method [25] uses the scaling factors of batch normalization layers [52] to measure the importance for each channel. During training, the sparse constraint is imposed on the scaling factors and then channels with tiny scaling factors are pruned. The pruned networks are then fine-tuned with common cross-entropy loss to recover the performance. Recently, it was empirically found that when the training data are available, training the pruned network from scratch might achieve higher performance [53]. However, for situation that the entire training dataset is not accessible, the pre-trained weights in the giant network are vital to get a pruned network with high performance.

III. CHANNEL-WISE PRUNING UNDER SPARSE REGULARIZATION

A number of well-trained DNNs can be easily obtained from the Internet and tailored for various tasks. Most of the time these downloaded networks are too cumbersome to be applied directly in practical tasks especially for those on edge mobile devices. So some questions arise immediately: How many parameters would be sufficient for DNNs to reach decent performance? How much computational budget can be offered by our computing platforms? Answers to these questions are not unique, depending on different real-world applications. It is therefore impossible to request model providers to release well-trained models of various sizes from a few hundred KB to several hundred MB to meet all users' demand. A practical solution is to compress the released giant models to an appropriate size that can meet different requirements. In the sequel, we will revisit how the giant neural networks can be compressed to a specific size using channel pruning techniques. Moreover, we also illustrate that why channel pruning would be degraded when the labeled data are quite limited.

Suppose a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ ² containing N examples $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and corresponding labels $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^K$ are released with a well-trained DNN, where \mathcal{X} and \mathcal{Y} are raw feature space and label space, respectively. Denote the released well-trained neural network of L layers as a function $f \in \mathcal{F}$, where \mathcal{F} denotes the hypothesis space of DNNs. A (channel-wise) sparse network is usually trained by minimizing the objective function defined on these labeled data, *i.e.*,

$$\mathcal{L}_l(f) = \frac{1}{N} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \mathcal{L}_o(f(\mathbf{x}_i, W), \mathbf{y}_i) + \lambda \cdot \mathcal{R}(W) \quad (1)$$

where \mathcal{L}_o is a supervision loss (*e.g.*, cross entropy for image classification) to guarantee the network performance and $\mathcal{R}(\cdot)$ is a sparse regularizer imposed on the weights. λ balances the classification accuracy and the sparsity of weights, and a larger λ induces a sparser network. To get a network with channel-wise sparsity that can be inferred fast without specific hardware/software support, $\mathcal{R}(\cdot)$ can serve as a group sparsity over weights directly [50]. However, implementing the group sparsity on the massive weights also increases the difficulty in optimization [25], incurring slow convergence and unsatisfied results. Hence the scalar factor $\gamma = \{\gamma^{(j)}\}_{j=1}^m$ is introduced as surrogates to control the channel-wise sparsity. Denoting

²In our problem, the dataset capacity N for labeled data are usually small.

\mathbf{z}_i as the l -th layer in the feature map of the network f for example $\mathbf{x}_i \in \mathcal{X}$, the feature map $\hat{\mathbf{z}}_i$ after scaling is obtained as:

$$\hat{\mathbf{z}}_i^{(j)} = \gamma^{(j)} \mathbf{z}_i^{(j)}, \quad (2)$$

where $\hat{\mathbf{z}}_i^{(j)}$ and $\mathbf{z}_i^{(j)}$ are the j -th channel of $\hat{\mathbf{z}}_i$ and \mathbf{z}_i , respectively. The scaling factor γ control the information flow and a small $\gamma^{(j)}$ blocks the information of the corresponding channels. In practice, the trained parameters in Batch Normalization (BN) layers can work as the scaling factors [25], [54] and for those networks without BN layers, it is also easy to insert the extra scaling factors to scale feature maps in the training phase.

When training the network, sparsity regularization is imposed on scaling factors γ 's, and the objective function evolves into:

$$\mathcal{L}_l(f) = \frac{1}{N} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \mathcal{L}_o(f(\mathbf{x}_i, W, \Gamma), \mathbf{y}_i) + \lambda \cdot \|\Gamma\|_1, \quad (3)$$

where $\|\cdot\|_1$ is ℓ_1 norm to encourage only part of channels are selected to establish the network, and Γ is a vector composed of all scaling factors over the whole network. A network with spare scaling factors is obtained via Eq. (3), and then the network is pruned based on the values of scaling factors. A global threshold across layers is set according to the percentage of channels users plan to keep. For example, if users want to prune 60% channels of the network, the smaller 60% elements in Γ are removed with their corresponding channels. The structure of the network will automatically decided according to the threshold, and a pruned network is obtained as a result. After pruning, the network usually has limited classification accuracy. Fine-tuning the pruned network is therefore essential to restore the accuracy. Typically, when we fine-tune the pruned network, the objective is only the supervision loss (*i.e.*, the first term in Eq. (3)).

However, only with the extremely limited data released, Eq. (3) (with or without sparsity term $\|\Gamma\|_1$) cannot be well optimized. Serious overfitting will occur and the accuracy on the test set will drop rapidly, incurring poor performance of the pruned network.

IV. PRUNING NETWORKS WITH UNLABELED DATA

Few labeled data limit the performance of channel-pruned networks. Instead of accepting the poor performance of the pruned network with only limited sample data or struggling to search for original training data, we turn attention to cheaper unlabeled data in hand. In this section, we will present a solution to bring giant networks down to earth with channels pruned, in which we will investigate the potential benefits from unlabeled data.

A. Exploiting Unlabeled Data by Mimicking the Giant Network

Unlabeled data are much easier to collect. For example, one can easily find a large number of natural images from the Internet to assist compression of giant networks trained on large natural images set. The unlabeled data collected by users may be different from the original data used for the well-trained

giant network, but still can provide helpful information for the compression task.

Suppose the collected unlabeled dataset $\mathcal{D}^u = \{\mathbf{x}_i^u\}$ contains N^u examples $\mathbf{x}_i^u \in \mathcal{X}^u \subset \mathbb{R}^d$ ³. In the sequel, we distinguish the labeled and unlabeled examples with notations $(\mathbf{x}^l, \mathbf{y}^l)$ and \mathbf{x}^u , respectively.

Similarly, given $\mathbf{x}_i^u \in \mathcal{X}^u$, the *softened* output of the network f and the released well-trained network \tilde{f} after softmax function are represented as \mathbf{p}_i^u and $\tilde{\mathbf{p}}_i^u$, respectively,

$$\mathbf{p}_i^u = \frac{\exp(f(\mathbf{x}_i^u)/\tau)}{\|\exp(f(\mathbf{x}_i^u)/\tau)\|_1}, \quad \tilde{\mathbf{p}}_i^u = \frac{\exp(\tilde{f}(\mathbf{x}_i^u)/\tau)}{\|\exp(\tilde{f}(\mathbf{x}_i^u)/\tau)\|_1}, \quad (4)$$

where τ is a temperature parameter [19] to control the smoothness, so that a higher value of τ produces a softer probability distribution over classes. The softened output of the giant network reveals clues about its classification characteristics as well as the similarity among classes. Thus we can use unlabeled data to mimic its classification performance and distill its knowledge into the target sparse network by minimizing the cross-entropy $\mathcal{H}(\mathbf{p}_i^u, \tilde{\mathbf{p}}_i^u)$ between softened outputs of labeled and unlabeled data.

Confidence on the unlabeled data. However, the extra collected unlabeled data may have different categories compared with the labeled data for training the giant network, and many examples may not be well understood even by the giant network. Directly imposing the standard distillation loss $\mathcal{H}(\mathbf{p}_i^u, \tilde{\mathbf{p}}_i^u)$ will also enforce the sparse network to mimick the outputs of the giant network on those aberrant data, which may disturb the learning process on labeled data and incur performance degradation. Thus we propose to treat the unlabeled examples differently by weighting them with the confidence of the giant network \tilde{f} . The confidence can be reflected by their corresponding outputs, and for the network with a high confidence on example $\mathbf{x}_i^u \in \mathcal{X}^u$, one element in its output vector $\tilde{f}(\mathbf{x}_i^u)$ will be far larger than other elements. Thus the maximum of the output vector normalized by softmax function can be used as the confidence. However, the original softmax function (*i.e.*, with $\tau = 1$) is so sharp that the maximum in $\tilde{f}(\mathbf{x}_i^u)$ would be very close to 1 in most cases. Temperature τ is used to soften the output vector so that the maximum in $\tilde{f}(\mathbf{x}_i^u)$ would be more sensitive when the confidence changes. The weight C_i^u for the example $\mathbf{x}_i^u \in \mathcal{X}^u$ is defined:

$$C_i^u = \max \frac{\exp(\tilde{f}(\mathbf{x}_i^u)/\tau)}{\|\exp(\tilde{f}(\mathbf{x}_i^u)/\tau)\|_1}, \quad (5)$$

where $\max(\cdot)$ is the maximum value of a vector. In this way, we encourage the sparse network f to have similar softened outputs with those of the giant network \tilde{f} on the examples with high confidence. Then the objective of the sparse retraining on unlabeled data (*distillation loss*) can be written like Eq. (3) as

$$\mathcal{L}_m(f) = \frac{1}{N^u} \sum_{\mathbf{x}_i^u \in \mathcal{D}^u} C_i^u \cdot \mathcal{H}(\mathbf{p}_i^u, \tilde{\mathbf{p}}_i^u) + \lambda \cdot \|\Gamma\|_1, \quad (6)$$

³Here we assume the labeled data and unlabeled data have the same dimension d , which can be easily implemented by image resizing or cropping.

where $\mathcal{H}(\cdot, \cdot)$ is the cross-entropy. Considering both labeled and unlabeled data, the objective function for training network f is

$$\mathcal{L}(f) = \mathcal{L}_l(f) + \alpha \cdot \mathcal{L}_m(f), \quad (7)$$

with a constant weight $\alpha \geq 0$. In this way, the unlabeled data can help the labeled data by further supplying more information about the giant network. The confidences automatically weight different unlabeled data and discard those improper and noisy data. Note that the labeled data can also be used to mimic the output of the pre-trained giant network⁴. However, the effect of distillation on labeled data is subtle due to their limited quantity, which is further verified in the experiments.

B. Fixing the Dataset Bias between Unlabeled and Labeled Data

In Eq. (6) above, the unlabeled data are usually utilized to make a consistency of the probabilistic output between the sparse network and the original giant network. However, in practice, the collected unlabeled data by users are usually different from the original labeled data. And there exists a dataset bias (or domain shift) [55] between the unlabeled data and labeled data. As a result, the consistency in unlabeled data may not hold on labeled data. Since both the pruned network and the giant network are designed for labeled data, the classification performance of the pruned network would be degraded due to this dataset bias.

To make the massive unlabeled data act as the supplement of limited labeled data well, the representations of unlabeled data are expected to be consistent with those of labeled data, which can be fulfilled by make an alignment on the representations from two datasets. We use a few low-level layers of the network f as an aligner f_1 and denote the remaining part as f_2 . Both labeled and unlabeled data are first sent to f_1 to produce representations with same distributions and then based on these coincident representation, f_2 outputs the classification results, *i.e.*, $f(\cdot) = f_2(f_1(\cdot))$.

Generally, the gap between two distributions can be measured by multiple measures such as maximum mean discrepancy (MMD) [56] and various divergence metrics [57]. Following the recent success of adversarial training methods to reduce the gap between different distributions in many tasks, such as image style transfer [58], image super-resolution [59] and domain adaptation [60], we minimize the discrepancy between the unlabeled and labeled representation distributions by introducing a discriminator [61], which distinguishes the distributions from different datasets and is co-trained with a generator in an adversarial learning manner.

The adversarial training [61] can be regarded as a two-player minimax game. Given examples $\mathbf{x}_i^l \in \mathcal{X}^l, \mathbf{x}_i^u \in \mathcal{X}^u$, the discriminator D is to make binary predictions about whether their low-level feature representations $f_1(\mathbf{x}_i^l), f_1(\mathbf{x}_i^u)$ are from unlabeled dataset or not. In this case, the aligner f_1 plays

⁴Then the first term in Eq.6 can be replaced by $\frac{1}{N^u} \sum_{\mathbf{x}_i^u \in \mathcal{D}^u} C_i^u \mathcal{H}(\mathbf{p}_i^u, \tilde{\mathbf{p}}_i^u) + \frac{1}{N^l} \sum_{\mathbf{x}_i^l \in \mathcal{D}^l} \mathcal{H}(\mathbf{p}_i^l, \tilde{\mathbf{p}}_i^l)$, where $\mathbf{p}_i^l = \frac{\exp(f(\mathbf{x}_i^l)/\tau)}{\|\exp(f(\mathbf{x}_i^l)/\tau)\|_1}, \tilde{\mathbf{p}}_i^l = \frac{\exp(\tilde{f}(\mathbf{x}_i^l)/\tau)}{\|\exp(\tilde{f}(\mathbf{x}_i^l)/\tau)\|_1}$.

the role as the generator, which tries to confuse the two feature maps. The game can be modelled with a value function $V(f_1, D)$:

$$\begin{aligned} \min_{f_1} \max_D V(f_1, D) &= \mathbb{E}_{\mathbf{x}^l \sim p^l(\mathbf{x}^l)} [\log(D(f_1(\mathbf{x}^l)))] \\ &\quad + \mathbb{E}_{\mathbf{x}^u \sim p^u(\mathbf{x}^u)} [\log(1 - D(f_1(\mathbf{x}^u)))] . \end{aligned} \quad (8)$$

Eq. (8) is usually solved by alternatively optimizing the D and f_1 , whose loss functions are

$$\mathcal{L}_D(D) = -\hat{V}(f_1, D) \text{ with fixed } f_1, \quad (9)$$

$$\mathcal{L}_{f_1}(f_1) = \hat{V}(f_1, D) \text{ with fixed } D, \quad (10)$$

where \hat{V} is the empirical loss of the value function V , i.e.,

$$\begin{aligned} \hat{V} &= \frac{1}{N^l} \sum_{\mathbf{x}_i^l \in \mathcal{D}^l} \log(D(f_1(\mathbf{x}_i^l))) \\ &\quad + \frac{1}{N^u} \sum_{\mathbf{x}_i^u \in \mathcal{D}^u} \log(1 - D(f_1(\mathbf{x}_i^u))). \end{aligned} \quad (11)$$

The optimization of $\mathcal{L}_D(D)$ and $\mathcal{L}_{f_1}(f_1)$ is iterative and at last the distance of low-level features from labeled data and unlabeled data will be minimized. Thus the aligner f_1 can produce accordant representations of labeled and unlabeled data for the main body f_2 . Note that our scenario is different from the typical unsupervised domain adaptation task [60], [62], which adapts a network trained with massive labeled data to achieve high performance in an unlabeled domain by aligning the features in late layers, while neglecting the performance degradation on the original domain. As the unlabeled data here are only used to supply the labeled data and assist compressing networks, the representation alignment is conducted on the low-level layers to adapt the unlabeled data to be consistent to labeled data and compatible with the pre-trained network. This strategy adequately utilizes the information of the unlabeled data while reduce the disturbance to the normal training from the alignment process, and hence we can obtain a well-performed pruned network.

In practice, the aligner f_1 for feature aligning and the main body f_2 for outputting prediction results can be trained end-to-end by augmenting the original objective Eq. (7) with the *adversarial loss* $\mathcal{L}_{f_1}(f_1)$, i.e.,

$$\mathcal{L}(f) = \mathcal{L}_l(f) + \alpha \cdot \mathcal{L}_m(f) + \beta \cdot \mathcal{L}_{f_1}(f_1), \quad (12)$$

where $\beta \geq 0$ is the weight coefficient. In Eq. (12), the unlabeled data are first adapted to be compatible with the network by representation alignment, and then different examples are weighted by the example-wise confidence weight C_i^u (Eq. (5)). Hence the information in the massive unlabeled data are adequately explored and improper examples are also filtered. As a consequence, the sparse network f can well receive the help from unlabeled data for mimicking the giant network, but with subtle influence by the dataset bias.

C. Theoretical Discussions

Now we attempt to investigate how the unlabeled data help the pruned network to generalize better than that with only

a few labeled data. For simplicity of theoretical discussions, the training of entire system can be divided into two steps sequentially. In the first step, we adversarially train the aligner f_1 and the discriminator D ; then in the second step, the main body of network f_2 is trained to mimic the output of the original giant network \tilde{f} .

First, using the unlabeled and labeled data at low-level layers, we train the network via Eq. (8). Then in theory, we can make their distributions identical on feature representations, via the following Theorem 1 [61].

Theorem 1 (Feature alignment): With f_1 being fixed, the optimal discriminator D is $D^*(\mathbf{x}) = p^l(\mathbf{x})/(p^l(\mathbf{x}) + p^u(\mathbf{x}))$. Then the global optimality is achieved if and only if $p^l = p^u$.

As a result, we can align the distribution of unlabeled data's low-level features with that of labeled data. Since $p^l = p^u$, the input of the main body f_2 would have no dataset bias in theory. Then f_2 can be trained with the loss of Eq. (7), which can be cast into the framework of empirical risk minimization (ERM) with regularization. To facilitate the analysis on the unlabeled data's effect, we leave out the regularization term. Then we investigate the generalization ability of the learned f_2 by checking its generalization error bound, which is related to its population risk and empirical risk defined as

$$R(f_2) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{Q}} [\ell(f_2(\mathbf{x}), \mathbf{y})], \quad (13)$$

$$\hat{R}(f_2) = \frac{1}{N} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \ell(f_2(\mathbf{x}_i), \mathbf{y}_i), \quad (14)$$

where $\mathcal{Q} \in \mathcal{X} \times \mathcal{Y}$ is the ground-truth distribution of (\mathbf{x}, \mathbf{y}) ⁵. Usually, there exists a gap between the population risk $R(f)$ and empirical risk $\hat{R}(f)$. A desired model should have small gap. Via MaDiarmid's inequality, the gap can be bound by Theorem 2 [63].

Theorem 2 (Generalization error bound): Given a fixed $\rho > 0$, for any $\delta > 0$, with probability at least $1 - \delta$, for all $f \in F$

$$R(f_2) \leq \hat{R}(f_2) + \frac{2K^2}{\rho N} R'_N(F) + \sqrt{\frac{\ln \frac{1}{\delta}}{2N}}, \quad (15)$$

where K is the number of classes, and $R'_N(F)$ is the Rademacher complexity.

In Theorem 2, the third term shows that a large dataset capacity N induces a tight bound. In this way, with the unlabeled data involved, we boost the generalization ability of f_2 by increasing the training examples, which have identical distributions after the feature alignment. The second term refers to the Rademacher complexity as follows.

Definition 1: Given rademacher variables σ_i (independent uniform random variables in $\{-1, +1\}$) and $\mathbf{x}_i \in \mathcal{X}$, the Rademacher complexity of hypothesis $F \ni f$ is defined as

$$R'_N(F) = \mathbb{E}_{\mathbf{x}, \sigma_i} \left[\sup_{k, f^k} \sum_{i=1}^N \sigma_i f^k(\mathbf{x}_i) \right], \quad (16)$$

where $f^k(\mathbf{x}_i)$ is the k -th element in the output vector $f(\mathbf{x}_i)$.

⁵Here we do not distinguish the hard label vector and softened output vector in Eq. (7), and regard both as the target space \mathcal{Y} for simplicity.

Algorithm 1 Pruning with Unlabeled Data (PUD)

Input: Pretrained network \tilde{f} , released labeled dataset \mathcal{D}^l and collected unlabeled dataset \mathcal{D}^u

- 1: Initialize the sparse network f with \tilde{f} .
- 2: **repeat**
- 3: Randomly select $\mathbf{x}_i^l \in \mathcal{D}^l$ and $\mathbf{x}_i^u \in \mathcal{D}^u$ as a minibatch.
- 4: Forward the pre-trained giant network:
 $\{\tilde{\mathbf{p}}_i^l, \tilde{\mathbf{p}}_i^u, C_i^u\} \leftarrow \{\tilde{f}(\mathbf{x}_i^l), \tilde{f}(\mathbf{x}_i^u)\}$.
- 5: Forward the network:
 $\{\mathbf{p}_i^l, \mathbf{p}_i^u, f_1(\mathbf{x}_i^l), f_1(\mathbf{x}_i^u)\} \leftarrow \{f(\mathbf{x}_i^l), f(\mathbf{x}_i^u)\}$.
- 6: Calculate the loss of discriminator D with Eq. (9) and update the parameters of D .
- 7: Calculate the loss of network f with Eq. (18).
- 8: Update the parameters of network f .
- 9: **until** convergence
- 10: Prune channels with small scaling factors in network f .
- 11: Fine-tune the pruned network.

Output: A pruned network ready to deploy.

$R'_N(F)$ is directly related with the complexity of the hypothesis and the generalization gap as well. Thus to make the gap tighter, we can train the network f_2 by minimizing $R'_N(F)/N$. However, its computation is very hard. In practice, we usually use its upper bound or estimation [64] for each minibatch, e.g.,

$$R_c(f) = \frac{1}{N'} \max_k \sum_{i=1}^{N'} |f^k(\mathbf{x}_i)|, \quad (17)$$

where N' is the number of both labeled and unlabeled samples in a minibatch. Then term $R_c(f)$ can thus serve as a regularization term (called *Rademacher loss*) to control the generalization ability during the training of the network; the loss function is

$$\mathcal{L}_{all}(f) = \mathcal{L}_l(f) + \alpha \cdot \mathcal{L}_m(f) + \beta \cdot \mathcal{L}_{f_1}(f_1) + \eta \cdot R_c(f), \quad (18)$$

where $\eta \geq 0$ is a constant parameter, and $R_c(f)$ is calculated per minibatch. Although unlabeled data are much cheaper than labeled data, they are not free and large disks are needed to store the collected unlabeled data. Over-fitting usually occurs when tailoring or retraining the giant network with insufficient data, and thus in this case promoting the generalization with $R_c(f)$ will work. Our proposed method is summarized in Algorithm 1. Similarly, it also contains three steps, and unlabeled data play an important part in the sparse retraining and fine-tuning to assist the labeled data.

V. EXPERIMENTS

In this section, we empirically verify the effectiveness of the proposed method pruning networks with unlabeled data (PUD), which is compared with several state-of-the-art methods including SSL [50], PFEC [46], Slimming [25] and ISTA-Pruning [54]. SSL [50] and PFEC [46] identify unimportant filters by directly checking the network weights, while Slimming [25] prunes networks according to the scaling factors in BN layers. More recently, ISTA-Pruning [54] adopts



(a) CIFAR-10 dataset.

(b) STL-10 dataset.

Fig. 1. Sample images in the labeled CIFAR-10 dataset [66] and the unlabeled STL-10 dataset [68].

ISTA [65] to solve the spare-constrained optimization and achieves good performance. A Vanilla Pruning method is also conducted as a baseline, which just removes the small scaling factors of the giant networks and then fine-tunes the pruned networks with the labeled data. In addition, we train the pruned networks from scratch by randomly initializing their parameters, which is denoted as ‘Scratch’ method in our experiment. As these existing compression methods have been well optimized to achieve state-of-the-art performance on benchmark datasets including CIFAR-10 [66] and large-scale Imagenet (ILSVRC2012) [67], for fair comparison, we conduct experiments on them together with the prevalent VGGNet and ResNet following [25]. As for the assistant unlabeled data, we adopt the STL-10 dataset [68] and COCO dataset [69], respectively.

A. Experiments on CIFAR-10 Dataset

Dataset. The CIFAR-10 dataset [66] is composed of 60,000 32×32 color images from ten categories, 50,000 for training and 10,000 for testing. In our setting, only a small fraction of images are randomly selected as labeled data. The standard data augmentation [10] is adopted, including padding (with size 4), random cropping and horizontal flipping. As for the unlabeled data, we choose STL-10 dataset [68], which is also an image recognition dataset containing a large number of 96×96 (labeled and unlabeled) RGB images. Actually, STL-10 dataset has similar categories with CIFAR-10 dataset, however, their collection approaches are different. Some example images of the two datasets are shown in Figure 1. In our experiment, we randomly sample 5000 images from the unlabeled part of the STL-10 dataset to assist the compression. All unlabeled images are then rescaled into the same size 32×32 .

Networks. We experiment with VGGNet [8] and ResNet-56 [10], which are deep and powerful baseline networks broadly used in many tasks, such as image recognition, objection detection and video action analysis. The original VGGNet is designed for ImageNet dataset, thus we tailor its structure slightly to fit CIFAR-10 dataset following [25]. The features extracted by the convolution layers are pooled by a 2×2 pooling layer and then directly sent to a fully-connected layer to obtain predictions. The 56-layer ResNet is stacked by bottleneck blocks with pre-activation structure [70]. We train the VGGNet and ResNet from scratch in CIFAR-10 dataset as the giant pre-trained networks. For the adversarial loss in Section IV-B, we

TABLE I

CLASSIFICATION ACCURACY OF THE PRUNED VGGNET ON CIFAR-10 DATASET WITH THE UNLABELED STL-10 DATASET. ALL METHODS ACHIEVE APPROXIMATELY $11\times$ COMPRESSION RATE (1.8M PARAMS) AND $2.5\times$ ACCELERATION RATE (160M FLOPs).

	N^l	Scratch	Vanilla Pruning	SSL [50]	PFEC [46]	Slimming [25]	ISTA-Pruning [54]	PUD (Ours)
VGGNet	100	41.47	59.28	61.01	61.04	62.49	63.32	75.04
	500	56.97	78.31	79.42	80.21	84.52	85.93	88.51
	1K	69.86	82.56	83.23	84.69	87.23	88.05	91.14

TABLE II

CLASSIFICATION ACCURACY OF THE PRUNED RESNET-56 ON CIFAR-10 DATASET WITH THE UNLABELED STL-10 DATASET. ALL METHODS ACHIEVE APPROXIMATELY $2.0\times$ COMPRESSION RATE (0.3M PARAMS) AND $2.5\times$ ACCELERATION RATE (35M FLOPs).

	N^l	Scratch	Vanilla Pruning	SSL [50]	PFEC [46]	Slimming [25]	ISTA-Pruning [54]	PUD (Ours)
ResNet	200	42.32	51.49	52.65	54.01	55.48	55.76	62.03
	500	52.78	65.78	66.21	66.43	67.02	68.61	73.29
	1K	65.24	77.19	78.38	78.49	79.22	79.82	82.42

adopt the second pooling layer in VGGNet and the first block in ResNet as low-level feature layers, and the discriminator D is a simple 3-layer CNN. Feature maps are first delivered into two convolution layers followed by ReLU nonlinear operation, then forwarded to an average-pooling layer and a fully-connected layer to predict whether the image comes from labeled dataset or unlabeled dataset. The number of output channels of the first convolution layer is equal to the number of its input channels while the second convolution layer has double channels.

Training. For sparse retraining with Eq. (18), roughly equal iterations (15K~20K) are used. We experimentally find that this training iteration number suffices for both comparison methods (using only labeled data) and our method (using both labeled and unlabeled data). As for fine-tuning the pruned network, we use half of the iterations, *i.e.*, 7.5K~10K. For VGGNet, the initial learning rate is set to 0.003 in sparse retraining and 0.001 in fine-tuning, and for ResNet, it is set to 0.02 and 0.005, respectively. Learning rate drops by 0.1 at 1/2 and 3/4 of the maximum iterations for training with only labeled data. For training with additional unlabeled data, it drops by 0.3 at 40%, 70% and 90% of the maximum iterations. We empirically find that the two learning rate schemes fit their own setting well. For VGGNet, we select λ in the interval [0.0010, 0.0015] with step 0.0001 to control the sparsity of the network via the term $\|\gamma\|_1$, and for ResNet we select in the set {0.001, 0.002, 0.003}. When calculating the loss of discriminator, the labeled data are weighted by a coefficient equal to N^u/N^l for balance. The weight α and the temperature parameter τ are set to 0.7 and 3, respectively. The weight of adversarial loss β is set to 10^{-6} , while the weight of Rademacher loss η is select from {0.01, 0.001}. Parameters are determined with cross-validation.

Results. The classification accuracy of the pruned networks on CIFAR-10 dataset assisted by STL-10 dataset is presented in Table I and Table II for VGGNet and ResNet, respectively. The pre-trained VGGNet (ResNet) achieves 93.78% (93.96%) accuracy with 20.1M (0.59M) parameters and 398.6M (88.3M) float-point-operations (FLOPs). For fairness of comparison, all methods prune 70% channels of the pre-trained models assisted with 5K unlabeled images from STL-10 dataset, and obtain pruned networks with about 1.8M (0.3M) parameters

and 159M (35M) FLOPs.⁶

From Table I and Table II, we can see that with various numbers of labeled images N^l , the proposed PUD method significantly outperforms the comparison methods in all cases. This indicates the effectiveness and superiority of exploiting unlabeled data even when the distributions of labeled and unlabeled images are not exactly identical. When the labeled data are not sufficient, the comparison methods tend to be trapped in serious over-fitting problem. For example, with 1K labeled data, the state-of-the-art Slimming method only achieves 87.23% accuracy, with a large accuracy drop (6.55%) from the pre-trained VGGNet (93.78%). However, with the assistance of unlabeled data, our method can improve the performance by a large margin and achieves accuracy of 91.14%. Note that the pre-trained ResNet with shortcut connections and bottleneck blocks [70] is originally parameter compact, thus when pruning a similar percentage of channels, ResNet is more challenging and usually has larger accuracy drop than that of VGGNet.

Table I and Table II also show how the number of labeled images affects the performance of the pruned networks. Fewer data incur larger accuracy drop inevitably, however, the drop of our proposed method is much slower owing to the unlabeled data. For example, with only 100 labeled images, the state-of-the-art Slimming method [25] only achieves 62.49% accuracy, which is unacceptable for real applications. However, the improvement by unlabeled data are very prominent (*i.e.*, accuracy improved more than 12% comparing to Slimming [25]). The results show that unlabeled data provide a good platform to transfer the knowledge of the giant network and improve the accuracy accordingly, which is essential when labeled data are extremely limited.

The detailed structure of the pruned VGGNet and ResNet are shown in Table III and Figure 2, respectively. For VGGNet on CIFAR-10 dataset, more than 90% channels can be pruned in the later layers, implying much redundancy. For ResNet with bottleneck structure, a large number of channels in the

⁶The actual compression rate and acceleration rate are related to the percentage of channels pruned in each layer and may vary in a small range.

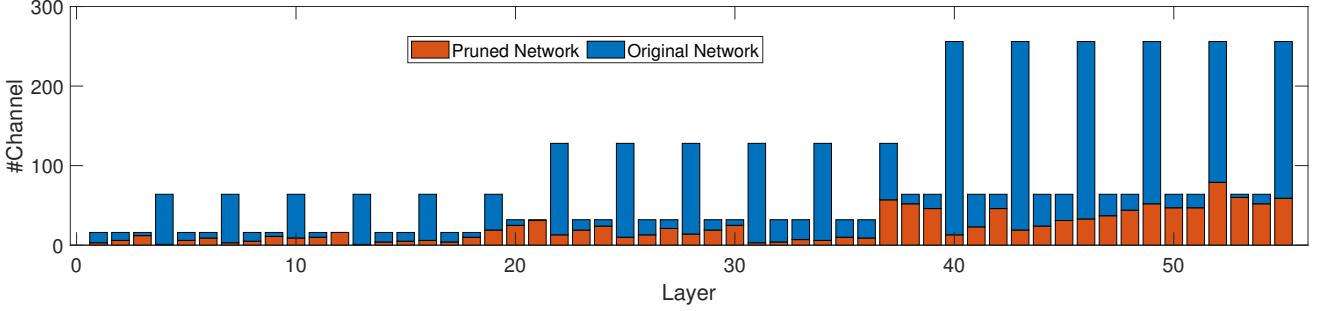


Fig. 2. Detailed structure of ResNet-56 on CIFAR-10 dataset by the proposed PUD method. The blue bar denotes the number of channels of the giant network while the red bar denotes that of the pruned network.

TABLE III

DETAILED STRUTURE OF VGGNET ON CIFAR-10 DATASET BY THE PROPOSED PUD METHOD. “# CHANNEL” AND “# CHANNEL*” DENOTE THE NUMBER OF OUTPUT CHANNELS OF CONVOLUTIONAL LAYERS IN THE GIANT NETWORK AND THE PRUNED NETWORK, RESPECTIVELY.

Layer	# Channel	# Channel*	Pruning rate (%)
conv 1-1	64	45	29.69
conv 1-2	64	60	6.25
conv 2-1	128	120	6.25
conv 2-2	128	112	12.50
conv 3-1	256	218	14.84
conv 3-2	256	211	17.58
conv 3-3	256	205	19.92
conv 3-4	256	124	51.56
conv 4-1	512	64	87.50
conv 4-2	512	59	88.48
conv 4-3	512	61	88.09
conv 4-4	512	37	92.77
conv 4-5	512	41	92.00
conv 4-6	512	39	92.38
conv 4-7	512	44	91.41
conv 4-8	512	248	51.56

“wider” layers can be pruned.

B. Experiments on ImageNet Dataset

Dataset. The ImageNet (ISLVRC2012) dataset [67] contains over 1.2M training images and 50k validation images from 1000 categories. For training, all images are cropped with size 224×224 and then randomly horizontally flipped. As for the unlabeled data, COCO dataset [69] is adopted since it is also a large-scale benchmark image dataset, which is widely used for object detection, segmentation and captioning. COCO dataset has 80 object categories, much fewer than the ISLVRC2012 dataset. We randomly sample 100k images as the unlabeled data. Using COCO dataset to assist the ISLVRC2012 dataset is a very challenging task because of the large difference of their distributions and categories. Some example images are shown in Figure 3.

Networks. Following [25], we use the “VGG-A” network model [8] with batch normalization [52] released by PyTorch⁷ as pre-trained model and evaluate performance with top-5 single-center-crop validation accuracy. The pre-trained model has 89.81% top-5 accuracy with 7.62B FLOPs. The feature

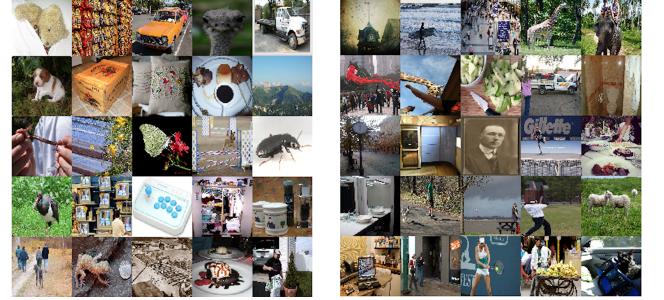


Fig. 3. Sample images in the labeled ISLVRC2012 dataset [67] and the unlabeled COCO dataset [69].

maps after the second pooling layer are sent to a 4-layer convolutional discriminator, which has a similar structure with that for CIFAR-10 dataset, but a convolution layer is added in the beginning. All the convolution layers have 3×3 kernels with stride 2.

Training. For all comparison methods, we use 100k iterations for the sparse retraining and 50k iterations for fine-tuning. The initial learning rate is set to 0.01 for the sparse retraining and determined from $\{0.001, 0.003, 0.005\}$ for fine-tuning. The learning rate drops by 0.3 at 40%, 70% and 90% of the total iterations. The weights α, β are respectively set to 0.5, 10^{-6} and η is selected from $\{0.001, 0.01\}$. The sparsity weight λ is set to 0.005 for the case $N^l = 50K$ and 0.003 for the case $N^l = 100K$. For all methods, we prune 50% channels of the pre-trained network.

Results. We randomly sample 50k and 100k labeled images from ISLVRC2012 dataset to implement the compression, assisted with 100K unlabeled samples from the COCO dataset. As Table IV shows, after 50% channels pruned, all the pruned networks have approximately $5 \times$ acceleration rate. However, the proposed PUD method achieves the best classification accuracies in all cases. It can be safely concluded that the usage of unlabeled data does enable to boost the compression performance on large-scale datasets. Comparing our results with that of ISTA-Pruning, e.g., 78.41% vs 75.51% and 82.21% vs 77.84% for top-5 accuracy, we can infer that the classification ability of the pre-trained model is well preserved by the unlabeled data via mimicking softened output and fixing the dataset bias. Considering the difference between ISLVRC2012

⁷<https://pytorch.org/docs/master/torchvision/models.html>

TABLE IV
CLASSIFICATION ACCURACY (TOP-5) OF THE PRUNED VGGNET ON ISLVRC2012 DATASET WITH THE UNLABELED COCO DATASET. ALL THE METHODS ACHIEVE APPROXIMATELY $5 \times$ ACCELERATION RATE (1.5B FLOPS).

N^l	Scratch	Vanilla Pruning	SSL [50]	PFEC [46]	Slimming [25]	ISTA-Pruning [54]	PUD (Ours)
50K	65.46	74.76	74.64	74.81	74.96	75.51	78.41
100K	70.37	76.94	77.12	77.31	77.52	77.84	82.21

TABLE V
DETAILED STRUCTURE OF THE PRUNED VGG-A MODEL ON ISLVRC2012 DATASET BY THE PROPOSED PUD METHOD. “# CHANNEL” AND “# CHANNEL*” DENOTE THE NUMBER OF OUTPUT CHANNELS OF CONVOLUTIONAL LAYERS IN THE GIANT NETWORK AND THE PRUNED NETWORK, RESPECTIVELY.

Layer	# Channel	# Channel*	Pruning rate (%)
conv 1-1	64	30	52.13
conv 2-1	128	57	55.47
conv 3-1	256	85	66.80
conv 3-2	256	123	51.95
conv 4-1	512	172	66.41
conv 4-2	512	223	56.45
conv 4-3	512	238	53.52
conv 4-4	512	499	2.54

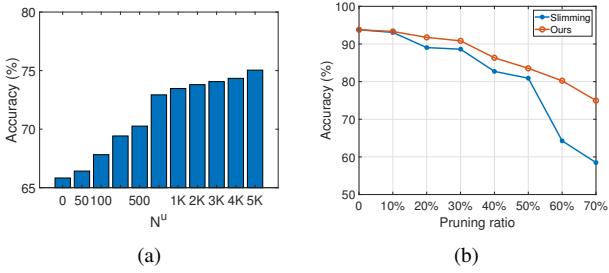


Fig. 4. Classification accuracy of the pruned networks on CIFAR-10 dataset w.r.t., (a) different number N^u of unlabeled data with 100 labeled data ($N^l = 100$) and 70% pruning ratio, and (b) various pruning ratio with $N^l = 100$, $N^u = 5K$.

dataset and COCO dataset, the significant improvement on the accuracies shows the effectiveness and superiority of our proposed method. The detailed structure of the pruned VGG-A is shown in Table V. For the VGG-A on ISLVRC2012 dataset, most of the layers has similar redundancy.

C. Ablation Studies

1) *Effect of the number of unlabeled data:* Furthermore, we investigate how the number of unlabeled data influences the classification accuracy of the pruned networks. In Figure 4(a), we report the corresponding accuracies with 100 labeled examples and various numbers of unlabeled ones. As shown in the results, when the number of unlabeled data are fairly limited, the help is also limited and the accuracy is low accordingly. But with the increase of unlabeled data, the accuracies rise steadily. When the unlabeled data are much more than labeled data, e.g., 1K vs 100, the accuracy tends to stabilize. Note that more unlabeled data also bring more training cost, thus in practice for the sake of training efficiency, users do not need to collect too many unlabeled examples.

2) *Effect of pruning ratio:* We also investigate how the accuracy of the pruned networks changes when we prune different ratios of their channels. As Figure 4(b) shows, drop of accuracy occurs as more channels are pruned since more information stored in the giant network loses and cannot be recovered totally due to limited labeled data. The accuracy of the proposed PUD method is always higher than that of pruning without unlabeled data, especially for a high pruning ratio (e.g., 60%). This might be because our method can leverage the unlabeled data to decrease the loss of information in the sparse retraining and restore information in fine-tuning as well.

3) *Effect of each individual component:* We now investigate the effect of each individual component in the proposed PUD method, i.e., the distillation loss \mathcal{L}_m , the adversarial loss \mathcal{L}_{f_1} , Rademacher loss R_c and confidence on unlabeled data C_i^u . The accuracies with (“✓”) or without (“✗”) each component for both VGGNet and ResNet on CIFAR-10 are shown in Table VI and Table VII. Distillation loss directly involves unlabeled data in the retraining process, and improves the performance by a large margin (i.e., from 87.23% to 89.84%), which verifies the prominent effect of unlabeled data as a good platform to distill knowledge from the pretrained network, as well as further alleviate over-fitting. Weighting unlabeled data with confidence C_i^u further improves performance (i.e., from 89.94% to 90.21%). However, due to the bias between labeled and unlabeled data, there is still a large room to boost performance. The adversarial loss \mathcal{L}_{f_1} alleviates the bias in low-level feature space, making unlabeled data have more positive effect and resulting in the improvement from 90.21% to 90.72%. Loss R_c derived from the theoretical generalization error bound strengthens the robustness of the proposed method as well as enhances performance slightly (i.e., from 90.21% to 90.50%). With all the components and their mutual effect, the proposed method achieves the best performance (i.e., 91.14%).

To further study how each individual loss and its corresponding weight coefficient affect the final performance, we vary weights α , β and η by fixing the others at the optimal parameter configuration with 100 labeled images and 5K unlabeled data, as shown in Figure 5.

Distillation loss \mathcal{L}_m and weight α . The main function of loss \mathcal{L}_m is to encourage the spare network to mimic the classification characteristics of the pre-trained model on unlabeled data. When varying α from 0 to 1, the degree of distillation increases accordingly. From Figure 5(a), the accuracy achieves a high level when α exceeds 0.001 then increases steadily with α . We also observe that an overlarge α (e.g., 1) would induce the accuracy to drop a bit. This might result from that in practice, the distribution of unlabeled data is different from that of labeled data, and an overemphasis on

TABLE VI
EFFECT OF EACH INDIVIDUAL COMPONENT FOR PRUNING VGGNET ON CIFAR-10 DATASET WITH THE UNLABELED STL-10 DATASET. $N^u = 5K$.

	Distillation	Confidence	Adversarial	Rademacher	$N^l = 100$	$N^l = 500$	$N^l = 1K$
VGGNet	✗	✗	✗	✗	62.49	84.52	87.23
	✓	✗	✗	✗	70.34	87.62	89.84
	✓	✓	✗	✗	71.26	88.18	90.21
	✓	✓	✓	✗	72.61	88.10	90.72
	✓	✓	✗	✓	73.82	88.19	90.50
	✓	✓	✓	✓	75.04	88.51	91.14

TABLE VII
EFFECT OF EACH INDIVIDUAL COMPONENT FOR PRUNING RESNET-56 ON CIFAR-10 DATASET WITH THE UNLABELED STL-10 DATASET. $N^u = 5K$.

	Distillation	Confidence	Adversarial	Rademacher	$N^l = 200$	$N^l = 500$	$N^l = 1K$
ResNet	✗	✗	✗	✗	55.48	67.02	79.22
	✓	✗	✗	✗	59.31	68.89	80.97
	✓	✓	✗	✗	59.92	69.37	81.28
	✓	✓	✓	✗	60.51	71.95	82.33
	✓	✓	✗	✓	61.18	72.45	82.16
	✓	✓	✓	✓	62.03	73.29	82.42

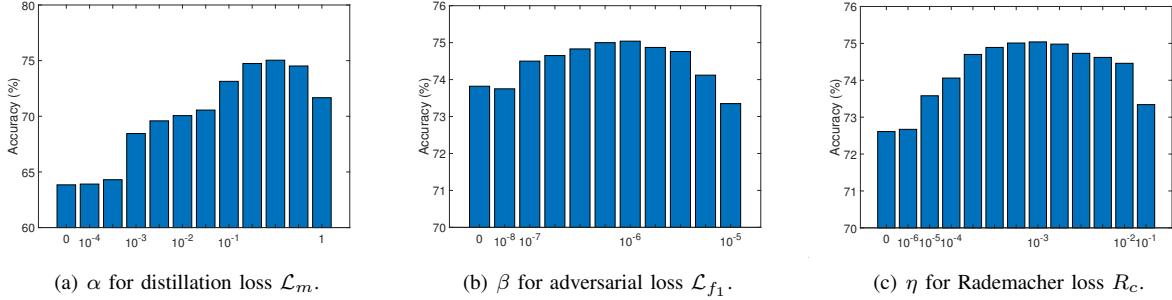


Fig. 5. Analysis on VGGNet of the three losses in Eq. (18) by varying their weights.

TABLE VIII
EFFECT OF LABELED DATA AND UNLABELED DATA FOR THE DISTILLATION LOSS \mathcal{L}_m .

L	U	$N^l = 100$	$N^l = 500$	$N^l = 1K$
✗	✗	63.57	85.26	87.86
✓	✗	65.83	85.89	88.14
✗	✓	74.71	88.44	91.05
✓	✓	75.04	88.51	91.14

TABLE IX
EFFECT OF LABELED DATA AND UNLABELED DATA FOR THE RADEMACHER LOSS \mathcal{R}_c .

L	U	$N^l = 100$	$N^l = 500$	$N^l = 1K$
✗	✗	72.61	88.10	90.72
✓	✗	72.66	88.11	90.95
✗	✓	74.53	88.42	91.10
✓	✓	75.04	88.51	91.14

the unlabeled data would disturb the network's fitting ability on the labeled data.

Adversarial loss \mathcal{L}_{f_1} and weight β . The low-level features on the pre-trained model are usually fairly different between unlabeled and labeled data, thus the adversarial loss is very large at the beginning of the retraining. We empirically find a small β can still have a more significant impact on the update of low-level features than that on the output layer since the adversarial loss is directly imposed on the low-level layers. The adversarial loss aligns the distributions of the unlabeled and labeled low-level features; however, this might cause that the feature distributions of labeled data on the pruned network drift away slightly from that on the giant network. In Figure 5(b), when the weight β is too large, much information of the original giant network will lose and the accuracy drops slightly.

Rademacher loss \mathcal{R}_c and weight η . Rademacher loss acts as a regularization term to boost the generalization ability of the pruned networks. The Rademacher loss complements with the two losses \mathcal{L}_u and \mathcal{L}_{f_1} , and work best with $\eta = 0.001$. Nevertheless, stronger regularization (e.g., 0.1 in Figure 5(c)) may also hamper the classification accuracy.

Verification of limitation of few labeled data. Note that the distillation loss \mathcal{L}_m and Rademacher loss \mathcal{R}_c can be imposed both on labeled data and unlabeled data. By fixing other components, we additionally conduct experiments whether the implementation of \mathcal{L}_m and \mathcal{R}_c cover the labeled data or unlabeled data. When implementing \mathcal{L}_m (or \mathcal{R}_c) on both labeled and unlabeled data we try different weights on them, which has subtle affect due to the limitation of labeled data, and thus equal weights are used for simplicity. Accuracies of the pruned VGGNet on CIFAR-10 are presented in Table VIII and Table

TABLE X
COMPARISON WITH STATE-OF-THE-ART KNOWLEDGE DISTILLATION METHODS.

Method	$N^l = 100$	$N^l = 500$	$N^l = 1K$
W/o unlabeled data	62.49	84.52	87.23
KD [19]	70.34	87.62	89.84
AT [38]	70.78	87.73	90.06
NST [71]	71.23	87.76	90.25
Ours	75.04	88.51	91.14

IX, and "L" represents labeled data while "U" is for unlabeled data. We can see that for both distillation loss and Rademacher loss, implementing them only with labeled data has a small effect. For example, with 100 labeled data for distillation loss (Rademacher loss), the improvement of performance is only 2.26% (0.05%). However, when introducing unlabeled data, the performance can be improved for a large margin. Even implementing distillation loss (Rademacher loss) only with unlabeled data, the accuracy is improved by 10.84% (1.92%) accordingly. We can safely conclude that the unlabeled data do play a vital part in helping the performance improvement of pruned networks.

Comparison with state-of-the-art knowledge distillation methods. For utilizing unlabeled data to assist compression, several existing knowledge distillation methods can be also directly applied, which pushes the spare network to behave similar as the giant network on those unlabeled data. We compare the proposed method with several state-of-the-art knowledge distillation methods and the accuracies of the pruned VGGNet on CIFAR-10 are shown in Table X. Unlabeled data can improve the performance of the pruned network with different distillation manners, implying the effectiveness of unlabeled data. However, the unlabeled data may be different from labeled data in many aspects such as distribution and categories, which limits the potentiality of exploring unlabeled data and the inappropriate data may disturb the training procedure. The proposed method alleviates dataset bias by accommodating the unlabeled data to the pre-trained network and then uses the instance-wise confidences to weight different unlabeled data, which further improves the performance dramatically.

VI. CONCLUSION

We solved a practical problem of compressing giant neural networks given only a few labeled examples instead of the original and complete training dataset. We exploited the unlabeled data to distill the knowledge from the giant network into the pruned network and boosted the compression performance. To alleviate the dataset bias between labeled and unlabeled data, we utilized the low-level layers of the network as an aligner to make an alignment on their representations. Experimental results validated the effectiveness of the proposed PUD method. For the future work, we plan to investigate an extreme situation even if no single example is released with the giant networks, which might demand higher generalization ability of the compressed networks.

REFERENCES

- [1] S. Qiao, Z. Zhang, W. Shen, B. Wang, and A. L. Yuille, "Gradually updated neural networks for large-scale image recognition," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 80. JMLR.org, 2018, pp. 4185–4194.
- [2] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu, "Deep predictive coding network for object recognition," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 80. JMLR.org, 2018, pp. 5263–5272.
- [3] T. Nagamine and N. Mesgarani, "Understanding the representation and computation of multilayer perceptrons: A case study in speech recognition," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 2564–2573.
- [4] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 2632–2641.
- [5] D. J. Rezende, S. Mohamed, I. Danihelka, K. Gregor, and D. Wierstra, "One-shot generalization in deep generative models," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 1521–1529.
- [6] L. Maaloe, C. K. Sonderby, S. K. Sonderby, and O. Winther, "Auxiliary deep generative models," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 1445–1453.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [11] Y. Wang, C. Xu, C. Xu, and D. Tao, "Beyond filters: Compact feature map for portable deep model," in *International Conference on Machine Learning*, 2017, pp. 3703–3711.
- [12] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.
- [13] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *CoRR*, vol. abs/1510.00149, 2015.
- [14] J. M. Alvarez and M. Salzmann, "Compression-aware training of deep networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 856–867.
- [15] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in neural information processing systems*, 2016, pp. 4107–4115.
- [16] Q. Hu, P. Wang, and J. Cheng, "From hashing to cnns: Training binaryweight networks via hashing," *arXiv preprint arXiv:1802.02733*, 2018.
- [17] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas, "Predicting parameters in deep learning," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 2013, pp. 2148–2156. [Online]. Available: <http://papers.nips.cc/paper/5025-predicting-parameters-in-deep-learning.pdf>
- [18] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1943–1955, 2016.
- [19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [20] Y. Wang, C. Xu, C. Xu, and D. Tao, "Adversarial learning of portable student networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [21] B. Reagen, U. Gupta, B. Adolf, M. Mitzenmacher, A. M. Rush, G. Wei, and D. Brooks, "Weightless: Lossy weight encoding for deep neural network compression," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 80. JMLR.org, 2018, pp. 4321–4330.
- [22] M. A. Carreira-Perpinán and Y. Idelbayev, "Learning-compression algorithms for neural net pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8532–8541.

- [23] Y. Wang, C. Xu, C. Xu, and D. Tao, "Packing convolutional neural networks in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [24] J. M. Alvarez and M. Salzmann, "Learning the number of neurons in deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 2270–2278.
- [25] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2755–2763.
- [26] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Advances in neural information processing systems*, 2015, pp. 3123–3131.
- [27] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 525–542.
- [28] R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V. I. Morariu, X. Han, M. Gao, C.-Y. Lin, and L. S. Davis, "Nisp: Pruning networks using neuron importance score propagation," *Preprint at https://arxiv.org/abs/1711.05908*, 2017.
- [29] J. Cheng, P. Wang, G. Li, Q. Hu, and H. Lu, "Recent advances in efficient computation of deep convolutional neural networks," *Frontiers of IT & EE*, vol. 19, no. 1, pp. 64–77, 2018.
- [30] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 181–196.
- [31] U. Djuric, G. Zadeh, K. Aldape, and P. Diamandis, "Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care," *NPJ precision oncology*, vol. 1, no. 1, p. 22, 2017.
- [32] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: support vector machines for pharmaceutical data analysis," *Computers & chemistry*, vol. 26, no. 1, pp. 5–14, 2001.
- [33] J. Cheng, P.-s. Wang, G. Li, Q.-h. Hu, and H.-q. Lu, "Recent advances in efficient computation of deep convolutional neural networks," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 64–77, 2018.
- [34] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on cpus," 2011.
- [35] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *arXiv preprint arXiv:1605.04711*, 2016.
- [36] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," *arXiv preprint arXiv:1612.01064*, 2016.
- [37] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.
- [38] 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.
- [39] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1285–1294.
- [40] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [41] J.-C. Su and S. Maji, "Adapting models to signal degradation using distillation," *arXiv preprint arXiv:1604.00433*, 2016.
- [42] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.
- [43] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [44] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 1379–1387. [Online]. Available: <http://papers.nips.cc/paper/6165-dynamic-network-surgery-for-efficient-dnns.pdf>
- [45] V. Lebedev and V. S. Lempitsky, "Fast convnets using group-wise brain damage," in *CVPR*. IEEE Computer Society, 2016, pp. 2554–2564.
- [46] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.
- [47] X. Gao, Y. Zhao, L. Dudziak, R. Mullins, and C.-z. Xu, "Dynamic channel pruning: Feature boosting and suppression," *arXiv preprint arXiv:1810.05331*, 2018.
- [48] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1389–1397.
- [49] J.-H. Luo, H. Zhang, H.-Y. Zhou, C.-W. Xie, J. Wu, and W. Lin, "Thinet: pruning cnn filters for a thinner net," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [50] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 2074–2082.
- [51] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance estimation for neural network pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 264–11 272.
- [52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [53] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *arXiv preprint arXiv:1810.05270*, 2018.
- [54] J. Ye, X. Lu, Z. Lin, and J. Z. Wang, "Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers," *arXiv preprint arXiv:1802.00124*, 2018.
- [55] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, "Covariate shift and local learning by distribution matching," 2008.
- [56] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 136–144.
- [57] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures," in *null*. IEEE, 2003, p. 487.
- [58] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [59] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, vol. 2, no. 3, 2017, p. 4.
- [60] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4.
- [61] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [62] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.
- [63] V. Koltchinskii, D. Panchenko *et al.*, "Empirical margin distributions and bounding the generalization error of combined classifiers," *The Annals of Statistics*, vol. 30, no. 1, pp. 1–50, 2002.
- [64] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in deep learning," *arXiv preprint arXiv:1710.05468*, 2017.
- [65] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [66] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.
- [67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [68] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *AISTATS*, ser. JMLR Proceedings, vol. 15. JMLR.org, 2011, pp. 215–223.
- [69] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV (5)*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.
- [70] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.
- [71] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.



Yehui Tang received the B.E. degree from Xidian University in 2018. Currently, he is a Ph.D candidate with the Key Laboratory of Machine Perception (Ministry of Education) at Peking University. His research interests lie primarily in machine learning and computer vision.



Shan You is now a researcher at SenseTime Research. He obtained the B.E degree from Xi'an Jiaotong University in 2014, and a Ph.D degree with the Key Laboratory of Machine Perception (Ministry of Education) at Peking University. His research interests lie primarily in machine learning and computer vision.



Chang Xu is a Lecturer in Machine Learning and Computer Vision at the School of Information Technologies, The University of Sydney. He obtained a Bachelor of Engineering from Tianjin University, China, and a Ph.D. degree from Peking University, China. While pursuing his PhD degree, Chang received fellowships from IBM and Baidu. His research interests lie in machine learning, data mining algorithms and related applications in artificial intelligence and computer vision, including multi-view learning, multi-label learning, visual search and face recognition. His research outcomes have been widely published in prestigious journals and top tier conferences.



Boxin Shi is currently a Boya Young Fellow Assistant Professor at Peking University, where he leads the Camera Intelligence Group. Before joining PKU, he did postdoctoral research at MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University from 2013 to 2016, and worked as a Researcher at the National Institute of Advanced Industrial Science and Technology from 2016 to 2017. He received the B.E. degree from Beijing University of Posts and Telecommunications in 2007, M.E. degree from Peking University in 2010, and Ph.D. degree from the University of Tokyo in 2013. He won the Best Paper Runner-up award at International Conference on Computational Photography 2015. He has served as Area Chairs for ACCV 2018, BMVC 2019, and 3DV 2019.



Chao Xu received the B.E. degree from Tsinghua University in 1988, the M.S. degree from University of Science and Technology of China in 1991 and the Ph.D degree from Institute of Electronics, Chinese Academy of Sciences in 1997. Between 1991 and 1994 he was employed as an assistant professor by University of Science and Technology of China. Since 1997 Dr. Xu has been with School of EECS at Peking University where he is currently a Professor. His research interests are in image and video coding, processing and understanding. He has authored or co-authored more than 150 publications and 5 patents in these fields.