Learning Student Networks in the Wild

Hanting Chen[1,2], Tianyu Guo[1,2], Chang Xu[3], Wenshuo Li[2], Chunjing Xu[2], Chao Xu[1], Yunhe Wang[2*]

[1] Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University.

[2] Noah's Ark Lab, Huawei Technologies. [3] School of Computer Science, Faculty of Engineering, The University of Sydney.

htchen@pku.edu.cn, yunhe.wang@huawei.com

Abstract

Data-free learning for student networks is a new paradigm for solving users' anxiety caused by the privacy problem of using original training data. Since the architectures of modern convolutional neural networks (CNNs) are compact and sophisticated, the alternative images or meta-data generated from the teacher network are often broken. Thus, the student network cannot achieve the comparable performance to that of the pre-trained teacher network especially on the large-scale image dataset. Different to previous works, we present to maximally utilize the massive available unlabeled data in the wild. Specifically, we first thoroughly analyze the output differences between teacher and student network on the original data and develop a data collection method. Then, a noisy knowledge distillation algorithm is proposed for achieving the performance of the student network. In practice, an adaptation matrix is learned with the student network for correcting the label noise produced by the teacher network on the collected unlabeled images. The effectiveness of our DFND (Data-Free Noisy Distillation) method is then verified on several benchmarks to demonstrate its superiority over state-of-the-art data-free distillation methods. Experiments on various datasets demonstrate that the student networks learned by the proposed method can achieve comparable performance with those using the original dataset. Code is available at https://github.com/huawei-noah/Data-Efficient-Model-Compression

1. Introduction

Deep convolutional neural networks have been widely used in various computer vision tasks such as image recognition [14, 18], object detection [29, 11, 42, 41] and image segmentation [23]. However, these networks usually consist of enormous number of parameters and requires heavy computation cost, which prevent their usage in edge devices such as mobile phones and autonomous cars. For example, VGG-net [31] requires 548*MB* memory for saving parameters and 20*G* floating point operations for processing a single image. To this end, a great number of techniques including quantization [13], pruning [20] and distillation [15] have been proposed to accelerate and compress convolutional neural networks.

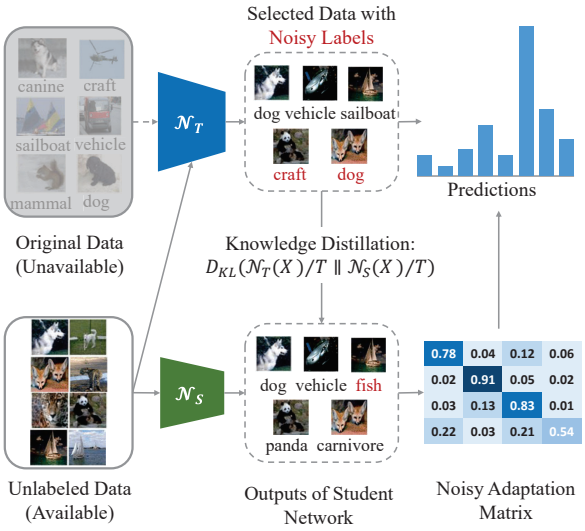


Figure 1. The diagram of the proposed method for learning student networks in the wild. Useful data will be first selected from the external unlabeled data and then utilized for training the desired student network. The noise adaption matrix is exploited for correcting labels of unlabeled data estimated by the teacher network.

Admittedly, we can obtain considerable compression ratios on benchmark datasets and models using these method when we can access the original training data of the pre-trained network. However, the training data is often unavailable in some practice constrains such as privacy or transmission. For example, we want to compress a deep model trained on millions of images, while the dataset is difficult to transfer and restore. Furthermore, people are willing to share their trained models to public, while they are very anxious about the training data especially some private data, *e.g.*, face, voice and fingerprint. Thus, a recent trend for

*Corresponding author

model compression algorithms is to develop data-free techniques that can reduce the computational complexities of pre-trained networks without original training data.

To this end, Lopes *et al*. [24] first formulated the data-free learning problem and use the "meta-data" to reconstruct the original images. However, its performance is limited since the useful knowledge information in the teacher network has not been fully investigated. Chen *et al*. [2] developed a GAN (Generate Adversarial Network [10]) based method, which used a generator network to approximate the training samples from the given teacher network. Besides using generators to obtain training data, other methods [27, 1] synthesized training data by directly optimizing the input random images on the pre-trained network.

However, it is hard to generate images which have enough information for training the compressed network, since the size of training data is usually much larger than the given network. For example, ImageNet dataset consists of over 10 million numbers of images with 224×224 size and requires over 138*GB* storage, while a ResNet-50 model contains only $\sim$100*MB* parameters. Therefore, the quality of these generated images cannot be ensured. Moreover, the running time and overhead for generating enough images for large scale dataset (*e.g*., ImageNet) are expensive. Thus, an efficient and effective method for learning portable student network without training data is urgently required.

In this paper, we present to utilize the large amount of unlabeled data in the wild to address the data-free knowledge distillation problem. Instead of generating images from the teacher network with a series of priori, images most relevant to the given pre-trained network and tasks will be identified from a large unlabeled dataset (*e.g*., Flickr [17]) to conduct the knowledge distillation task. We first analyze the bound of distance between the outputs of the teacher and the student networks, and then explore a data selection method for searching useful unlabeled data. Then, these data with the noisy labels derived from the teacher network is collected. To further improve the performance of the student network, a noise adaptation matrix is exploited for refining the labels provided by the teacher network. The portable student network is supervised by the conventional knowledge distillation approach on the collected data and the proposed noisy distillation using the adaptation matrix, as shown in Figure 1. Experiments conducted on several benchmarks demonstrate that the proposed DFND (Data-Free Noisy Distillation) algorithm can surpass all data-free distillation methods and achieve the state-of-the-art performance, the accuracy of the resulting student is comparable to that of the student network trained using original data.

2. Related Works

Here we briefly review the related works of model compression and acceleration, which consists of data-driven and data-free methods.

2.1. Data-Driven Model Compreesion

In order to compress and speed-up pre-trained heavy deep models, various effective approaches have been proposed recently, including quantization [39, 3], pruning [33, 22, 6], distillation [37] and neural architecture search [32, 38, 21]. Han *et al*. [13] combined pruning, quantization and Huffman coding together and then obtained a compressed deep model with extremely lower computation and storage cost. Hinton *et al*. [15] proposed to distill the knowledge from a heavy teacher network to a portable student network. Li *et al*. [20] proposed a filter pruning method to remove the filters in the convolutional neural network with small ℓ_2 norm. Luo *et al*. [25] proposed ThiNet, which formulate filter pruning as an optimization problem and prune filters based on the information from the next layer. Courbariaux *et al*. [5] proposed binary neural network, which utilize binary weights and activations to largely reduce the computation complexity and storage consumption of networks. Howard *et al*. [16] introduced depthwise separable convolution, which decompose the traditional convolution into 1×1 convolution and depthwise convolution, to accelerate the inference of deep neural networks. Han *et al*. [12] proposed GhostNet to utilize cheap operations to generate more features from existing features, which achieve the state-of-the-art performance in mobile settings.

Although these methods can achieve promising compression and speed-up ratios, they cannot be directly used when the original training dataset is unavailable. For instance, pruning and quantization methods requires training data to fine-tune the compressed networks. Therefore, data-free network compression methods have become a research hotspot.

2.2. Data-Free Model Compreesion

Only few works focus on compressing networks without original training data. Lopes *et al*. [24] used the "meta data", which is the activation statistics of original data in the teacher network, to reconstruct the dataset. Chen *et al*. [2] introduced a generator which is trained by regarding the teacher network as a fixed discriminator to generate the images which have similar distribution with the original dataset. Nayak *et al*. [27] synthesized the Data Impressions from the teacher netwwork as training samples. Fang *et al*. [7] proposed a data-free adversarial distillation scheme to generate "hard samples" for the student networks. Yin *et al*. [40] proposed DeepInversion to invert the trained network to synthesize input images from random noise. Beside knowledge distillation, some works focus on data-free quantization and pruning. Nage *et al*. [26] introduced a data-free compression method, which can quantize network to 8bit without fine-tuning by correcting the quantization bias. Choi *et al*. [4] proposed a adversarial training method to generate samples

for network quantization and pruning. Gong *et al*. [9] utilize vector quantization and k-means clustering to weights and lead to 16-24 times compression with little loss of accuracy.

Although the above data-free methods can compress networks without obtaining the original dataset, their performance is limited since generating training samples using only the teacher network is a hard attempt, especially on large scale dataset (*e.g.*, ImageNet). To this end, we propose to leverage the unlabeled data for data-free knowledge distillation with better performance.

3. Preliminary

Here we first review the conventional knowledge distillation [15] method for learning a portable student network from a heavy pre-trained teacher network. Denote the training data as X, and Y is the ground-truth label, the student network $\mathcal{N}_S$ and the teacher network $\mathcal{N}_T$, respectively. The loss function for distilling a student network is formulated as:

$$\mathcal{L}_{KD}(\mathcal{N}_S) = \mathcal{H}_{CE}(\mathcal{N}_S(x), y) + \lambda \mathrm{D}_{KL}(\mathcal{N}_S(x), \mathcal{N}_T(x)), \quad (1)$$

where $x \in X, y \in Y$, $\mathcal{H}_{CE}$ indicates the cross entropy loss, D_{KL} is the Kullback–Leibler (KL) divergence, and λ is the trade-off parameter. The second term in Eq. 1 minimizes the distance between output distributions of teacher and student networks, which can be considered as a strong regularizer for helping the training of the student network $\mathcal{N}_S$.

The traditional knowledge distillation requires the original training data , *i.e.*, the dataset used for training the teacher network $\mathcal{N}_T$. However, as mentioned above, this dataset is usually unavailable due to the privacy or transmission problems. Instead of using these data, there are massive data available on the Internet (*e.g.*, ImageNet and Flickr1M). Although there are a seies of data-free compression methods [2, 7] generating images from the given teacher network, the visual quality and computational costs limit their performance. Thus, we are motivated to explore an effective method for learning student networks with these unlabeled but available data.

A straightforward idea for using public unlabeled data $x^U \in X^U$ to perform the knowledge distillation is:

$$\mathcal{L}^U_{KD}(\mathcal{N}_S) = \mathrm{D}_{KL}(\mathcal{N}_S(x^U), \mathcal{N}_T(x^U)). \quad (2)$$

However, compared with Eq. 1, the limitations of Eq. 2 mainly lies in two parts. First, Eq. 2 aims to minimize the distance between the outputs of $\mathcal{N}_T$ and $\mathcal{N}_S$ over the distribution of the unlabeled data X^U instead of the original data X^O as claimed in Eq. 1. Therefore, the performance of the student network learned by Eq. 2 on the original dataset is not guaranteed. On the other hand, the classification loss is absent as the ground-truth labels are not accessible for the unlabeled data. These two parts prevent Eq. 2 from a suitable objective function for learning a student network with acceptable performance on the original data.

To overcome the shortage mentioned above and leverage the advantage provided by the unlabeled data, we propose a Data-Free Noisy Distillation (DFND) algorithm that help the student network to learn useful and correct information from the teacher. In brief, we select the most valuable samples from unlabeled data which helps the student to achieve good performance over the distribution of original data, and equip the unlabeled data with a noisy pseudo label to implement the classification loss which is missing in Eq. 2.

4. Data Collection in the Wild

As mentioned above, performance of the student network trained on the unlabeled dataset cannot be guaranteed. To this end, we aim to collect useful data from the huge unlabeled dataset X^U, which can ensure the student network trained by the selected data achieve good performance in the original data. Thus, our goal can be formulated as follows,

$$\hat{X}^U = \arg\min_{X^U} \mathrm{D}_{KL}(\mathcal{N}_S(x^O), \mathcal{N}_T(x^O)) \quad (3)$$

where x^O denotes the original training data, and $\mathcal{N}_S$ is the student network trained by the selected data $x^U \in X^U$ according to Eq. 2. However, sample selection principle described in Eq. 3 is intractable as the original data x^O is not accessible. In the following, we provide a surrogate selection principle to collect useful data from the huge unlabeled dataset X^U.

Hinton *et al*. [15] has proved that minimizing the KL divergence of the soft targets between the teacher and student can be regarded as minimizing the MSE (mean squared error) loss of their outputs when the temperature for knowledge distillation is relative high compared with the magnitude of the logits. Thus, instead of using KL divergence in Eq. 3, we will analyze the MSE loss (ℓ_2^2 distance) between the output of teacher and student to collect useful data, *i.e.*,

$$\hat{X}^U = \arg\min_{X_U} \mathcal{L}_{MSE}(\mathcal{N}_S(x^O), \mathcal{N}_T(x^O)) \quad (4)$$

With the help of Eq. 4, we propose the surrogate principle to select samples in Proposition 1.

Proposition 1: Given a pre-trained teacher network $\mathcal{N}_T$, and a huge unlabeled dataset X^U, the noisy value of an unlabeled sample x^U can be expressed as,

$$V(x^U) = \mathcal{D}_{KL}(\mathcal{N}_T(x^U), \hat{y}^U), \quad (5)$$

where $\hat{y}^U = \arg\max_i \mathcal{N}_T(y = y_i | x^U)$ is the pseudo label of x^U which is predicted by the teacher network $\mathcal{N}_T$. A useful sample is expected to have a small noisy value $V(x^U)$.

Proof. For the convenience of proof, we calculate the distance between domain instead of specific images. Denote

the domain of original data as D^O and the domain of the unlabeled training data for distillation as D^U. Using the triangle inequality of distance metric, the ℓ_2^2 distance between the teacher output and student output taking data in D^O as inputs is bounded as:

$$d_{\ell_2}(D_S^O, D_T^O) \leq d_{\ell_2}(D_S^O, D_S^U) + d_{\ell_2}(D_S^U, D_T^U) + d_{\ell_2}(D_T^U, D_T^O), \tag{6}$$

where D_S^O, D_S^U denote the student outputs in D^O, D^U and D_T^O, D_T^U are the teacher outputs in D^O, D^U, respectively.

Since we cannot directly minimize $d_{\ell^2}(D_S^O, D_T^O)$ as the original data is unknown, we then turns to minimize its upper bound. The first term in inequation 6, *i.e.*, $d_{\ell_2}(D_S^U, D_S^O)$, measures the distance between the outputs of original data and unlabeled data, which can be viewed as the generalization ability of the student network and it is determined by the network itself. The second term of inequality 6 on the right is exactly the goal of distillation on the unlabeled domain D^U, *i.e.* loss function 2. The third term $d_{\ell_2}(D_T^U, D_T^O)$ measures the distance between the teacher outputs in the original domain and unlabeled domain. As the pre-trained teacher network $\mathcal{N}_T$ and the original domain D^O is fixed, we can only select the unlabeled data for the domain D^U to minimize this term.

Since the true distribution of D_T^O is unknown, we cannot directly minimize the $d_{\ell_2^2}(D_T^U, D_T^O)$. Taking the distribution of the label in the original domain as D^Y, we have the inequality:

$$\begin{aligned} d_{\ell_2}(D_T^U, D_T^O) &\leq d_{\ell_2}(D_T^U, D^Y) + d_{\ell_2}(D_T^O, D^Y) \\ &\leq d_{\ell_2}(D_T^U, D^Y) + d_{\ell_1}(D_T^O, D^Y) \\ &\leq \sqrt{\mathrm{D}_{KL}(D_T^U, D^Y)} + \sqrt{\mathrm{D}_{KL}(D_T^O, D^Y)}. \end{aligned} \tag{7}$$

The second inequality is hold since ℓ_2 distance is bounded by ℓ_1 distance while the third inequality is hold since the square of ℓ_1 distance is bound by the KL divergence (theorem 1.3 in [28]). The teacher network have been well trained in the original domain D_T^O, which means the second term of inquation 7 on the right can be very small. Then we have $d_{\ell_2}(D_T^U, D_T^O) \leq \sqrt{\mathrm{D}_{KL}(D_T^U, D^Y)}$.

Therefore, the inequation 6 can be rewritten as:

$$d_{\ell_2}(D_S^O, D_T^O) \leq d_{\ell_2}(D_S^O, D_S^U) + d_{\ell_2}(D_S^U, D_T^U) + \sqrt{\mathrm{D}_{KL}(D_T^U, D^Y)}. \tag{8}$$

Utilizing the approximation of MSE loss and KL divergence, the optimization problem 4 can be reformulated as:

$$\hat{X}^U = \arg\min_{X_U} \big[\sqrt{\mathrm{D}_{KL}(\mathcal{N}_T(x^U), y)} + \sqrt{\mathrm{D}_{KL}(\mathcal{N}_T(x^U), \mathcal{N}_S(x^U))}\big], \tag{9}$$

where y are the labels predicted by the teacher network. As the second term in Eq. 9 is exactly the goal for training the student network and will be minimized to a very small value. The objective for selecting useful data from the optimal unlabeled dataset $\hat{X}^U$ can be formulated as:

$$\hat{X}^U = \arg\min_{X_U} \mathrm{D}_{KL}(\mathcal{N}_T(x^U), y) \tag{10}$$

Denote x^U and y as the data and corresponding labels of the dataset X^U. we will calculate the noisy value $V(x^U) = \mathcal{D}_{KL}(\mathcal{N}_T(x^U), y)$ of each image $x^U \in X^U$ and select a certain number of samples with smallest values to construct the optimal unlabeled dataset $\hat{X}^U$ for knowledge distillation. Note that the data is unlabeled, we use pseudo label y to calculate the noisy value, where $y = \arg\max_j(\mathcal{N}_T(x))_j$. □

Following the Proposition 1, samples with higher confidence score provided by the teacher are more likely to be selected as training data. The intuitions behind the data collection method are straightforward. First, with a teacher trained with the original data, samples assigned with a lower confidence score have a lower probability of being from the original distribution. Thus the proposition 1 can prevent most out-of-distribution samples from being selected. What's more, the information provided by the teacher on samples with a higher confidence score is less likely to be incorrect. As a result, we select samples with high value defined in Proposition 1.

It should be noted that, although selecting unlabeled data with low entropy pseudo labels is sometime used in semi-supervised learning [19, 35], we are the first to apply this technique in the data-free knowledge distillation setting. Moreover, we provide thorough analysis to guarantee the effectiveness of this data collection method theoretically.

5. Noisy Distillation from the unlabeled Data

Besides the KL-divergence on the outputs, the student network is also supervised by the cross entropy loss with ground-truth labels in the conventional knowledge distillation (Eq. 1). However, in the data-free distillation setting, the training data is unlabeled thus we do not have enough supervised information for learning student networks with better performance. Although the teacher network can be utilized for generating labels for the unlabeled data, the teacher could also make mistakes and provide incorrect label. To address this problem, we propose a novel noisy distillation method for distilling the student network with unlabeled data and noisy labels produced by the teacher network.

A straightforward method to generate labels for the first term in Eq. 1 is to use the pseudo labels predicted by the teacher network as their one-hot labels. The loss function

Algorithm 1 Learning Student Network with Unlabled Data and Noise Labels.

Input: A given teacher network $\mathcal{N}_T$, unlabeled dataset D^U, number of selected data K.

Module 1: Unlabeled data selection.
for each sample x_i in unlabeled dataset D^U **do**
  Employ the teacher network to obtain $N_T(x_i)$;
  Predict the pseudo label: $y_i = \arg\max_j(\mathcal{N}_T(x_i))_j$;
  Calculate and restore the noisy value $V(x_i) = \mathcal{D}_{KL}(\mathcal{N}_T(x_i), y_i)$ for the sample x_i;
end for
Select K samples with smallest noisy values and establish an alternative dataset $\hat{X}^U$;
Module 2: Noisy distillation.
Initialize the student network $\mathcal{N}_S$ with fewer parameters and lower computation cost;
Construct the noisy adaptation layer with the matrix Q according to Eq. 14;
repeat
  Randomly select a batch $\{x_i^s\}_{i=1}^n$ from $\hat{X}^U$;
  Employ the teacher network and the student network simultaneously: $\mathcal{N}_T(x), \mathcal{N}_S(x)$;
  Calculate the noisy distillation loss $\mathcal{L}_{ND}(\mathcal{N}_S, \mathcal{N}_T, Q, x)$ according to Eq. 13;
  Update $\mathcal{N}_S$ and Q according to their gradients;
until convergence

Output: The resulting student network $\mathcal{N}_S$ with acceptable performance.

for knowledge distillation can be reformulated as:

$$\mathcal{L}_{ND}(\mathcal{N}_S) = \mathcal{H}_{CE}(\mathcal{N}_S(x), \hat{y}) + \lambda \mathrm{D}_{KL}(\mathcal{N}_S(x) \| \mathcal{N}_T(x)), \quad (11)$$

where x, y denotes the training data and $\hat{y} = \arg\max_i(\mathcal{N}_T(x))_i$. For the knowledge distillation term D_{KL}, the student can directly learn from the teacher with the unlabeled data. For the cross entropy term, a more accurate label y will definitely help the training of the student network. However, teacher predicted labels $\hat{y}$ is noisy labels, since $\mathcal{N}_T$ is not learned for capturing information in the unlabeled data. To address this problem, we utilize the following function to discover the probability of noisy labels and true labels, *i.e.*,

$$p(\hat{y} = i|x) = \sum_{j=1}^{k} p(\hat{y} = i|y = j)p(y = j|x), \quad (12)$$

where k is the number of categories of y, $p(\hat{y}|y)$ is a noisy adaption matrix Q where $Q_{ij} = p(\hat{y} = i|y = j)$. It transforms the probability of true label $p(y|x)$ to the noisy probability $p(\hat{y}|x)$. Therefore, we can learn the true labels by adding a noisy adaptation matrix Q after the softmax layer of the student network. Then, the cross entropy is calculated between the transformed outputs and the pseudo labels. The noisy distillation in Eq. 3 can be reformulated as:

$$\begin{aligned}\mathcal{L}_{ND}(\mathcal{N}_S) =& \mathcal{H}_{CE}(Q(\mathcal{N}_S(x)), \hat{y}) \\ &+ \lambda \mathrm{D}_{KL}(\mathcal{N}_S(x), \mathcal{N}_T(x)).\end{aligned} \quad (13)$$

In practice, the actual values in the matrix Q are unknown. We thus initialize values in Q with the prior information of the teacher network. Specifically, denote the accuracy of the teacher network for the i-th class in the original dataset as a_i, the diagonal value of Q can be set as $Q_{ii} = a_i$. Since the row of Q can be regarded as a probability distribution, we have $\sum_{j=1}^{k} Q_{ij} = \sum_{j=1}^{k} p(\hat{Y} = i|Y = j) = 1$. Therefore, we initialize the Q as:

$$Q = \begin{bmatrix} a_1 & \frac{1-a_2}{k-1} & \cdots & \frac{1-a_k}{k-1} \\ \frac{1-a_1}{k-1} & a_2 & \cdots & \frac{1-a_k}{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-a_1}{k-1} & \frac{1-a_2}{k-1} & \cdots & a_k \end{bmatrix}. \quad (14)$$

During optimizing the student network $\mathcal{N}_S$, the noisy adaptation matrix Q is optimized simultaneously, keeping the constraints $\sum_{j=1}^{k} Q_{ij} = 1$. Note that Goldberger and Ben-Reuven [8] also proposed to utilize a noisy adaptation layer for learning noisy labels. Our work apply the noisy adaptation technique to the knowledge distillation scheme and simultaneously learn the soft label sand noisy predicted labels from the teacher network, which have different setting and loss function with this work.

The proposed method is summarized in Algorithm 1. First, we select data from the unlabeled dataset with the smallest noisy values, which implicitly minimizes the divergence between the teacher and student outputs in the original dataset. Second, the student is also trained by a classification loss with noisy labels predicted by the teacher, and an adaptation matrix is explored for correcting label noise. Note that there are few works [30, 34] also utilize noise to improve the knowledge distillation, their original training set is available and they focus on adding noise to labels and images and improve the generalization ability of the student networks. In contrast, we focus on eliminate the noise of the labels predicted by teacher networks for the unlabeled images.

6. Experiments

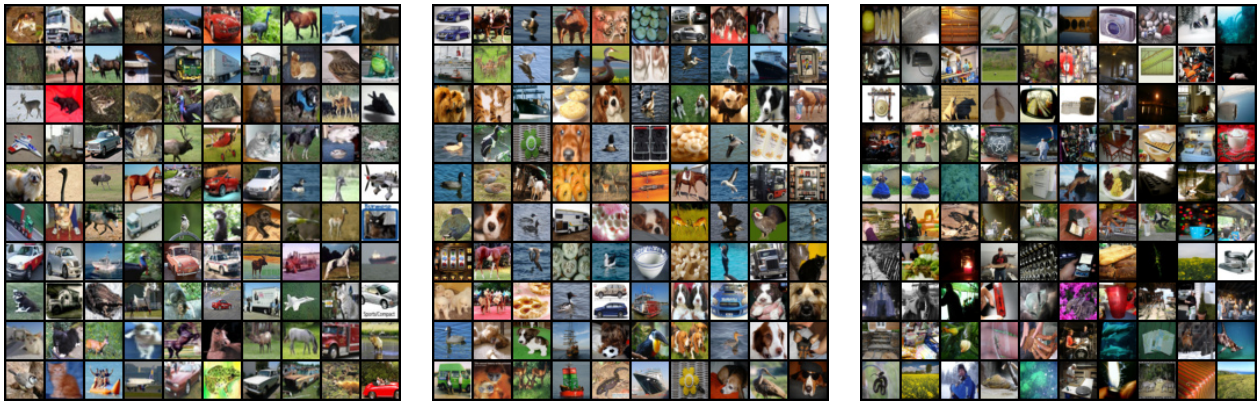
In this section, We conduct extensive experiments to verify the effectiveness of the proposed method on the image classification task and semantic segmentation task.

6.1. Classification Results on CIFAR

We first test the proposed method on the CIFAR-10 and CIFAR-100 dataset. Both of them contain 60,000 RGB 32×32 size images, including 50,000 training images and 10,000 test images of 10 and 100 categories, respectively. To

Table 1. Classification result on the CIFAR-10 and CIFAR-100 datasets.

Algorithm	Required data	FLOPS	#params	CIFAR-10	CIFAR-100
Teacher	Original data	∼1.16G	∼21M	94.85%	77.34%
Student	Original data	∼557M	∼11M	93.92%	76.53%
Knowledge Distillation [15]	Original data	∼557M	∼11M	94.34%	76.87%
DAFL [2]	No data	∼557M	∼11M	92.22%	74.47%
DFAD [7]	No data	∼557M	∼11M	93.30%	67.70%
DeepInversion [40]	No data	∼557M	∼11M	93.26%	-
PU-compression [36]	PU data	∼557M	∼11M	93.75%	-
Randomly selected	unlabeled data	∼557M	∼11M	91.25%	73.25%
DFND (ours)	unlabeled data	∼557M	∼11M	**94.02%**	**76.35%**



(a) original data (b) selected data (c) unselected data

Figure 2. Visualization of selected images, unselected images from ImageNet and original images on CIFAR-10 using ResNet-34. Each part consists of 100 images.

make a fair comparison, we select the ResNet-34 model as the teacher network and ResNet-18 as the student network following [2]. The teacher network and student network are optimized using Nesterov Accelerated Gradient (NAG). Weight decay and momentum are set as 5×10^{-4} and 0.9, respectively. We train the teacher networks for 200 epochs, where the initial learning rate is set as 0.1 and divided by 10 at 80 and 120 epochs, respectively. Random flipping, random crop and zero padding are used for data augmentation as suggested in [2]. The student networks of the proposed method are trained for 40000 iteration. For the proposed method, we select 600,000 images from the ImageNet dataset and the hyper-parameters T in knowledge distillation is set as 2, which is tuned by grid search. λ in Eq. 11 is set as 4 following [15]. The ImageNet dataset are resized to $32 \times 32 \times 3$ so that the unlabeled images can be put into the teacher network. When training the student network using knowledge distillation, the training data is normalized and cropped into 32 size with a 4 pixel zero padding.

Table 1 reports the distillation results on the CIFAR-10 and CIFAR-100 datasets. The teacher network achieves a 95.58% accuracy on the CIFAR-10 dataset and the student network using knowledge distillation on the original dataset achieves a 94.34% accuracy.

We then observe the performance of the same student network without the original training data. Chen *et al*. [2] proposed DAFL to use a generator for approximating the original training data from the pre-trained teacher network. Since it is difficult to generate images with only a teacher network, this method only achieves an accuracy of 92.22%. Data-Free Adversarial Learning [7] achieves a 93.30% accuracy without any training data, which is slightly higher than that of DAFL. Yin *et al*. [40] utilize DeepInversion to achieve a 93.26% accuracy. PU Compression [36] utilizes few original training data (100 images from CIFAR-10) and massive unlabeled data (all images from ImageNet) and select useful data using a PU classifier, which achieves a 93.75% accuracy. However, the positive data (few original training data) is not always available due to privacy and transmission reasons. Therefore, we propose to directly select useful from the massive unlabeled data (ImageNet data) using the teacher network using Eq. 10. As a result, the resulting student network learned using the proposed method achieves a 94.02% accuracy, which is very closed to that of the baseline knowledge distillation using the original data. In contrast, using the randomly selected data from the unlabeled dataset can only achieve a 91.25% accuracy. Note that the number of selected data used here is same as that in these two methods.

Besides CIFAR-10, we further verify the capability of the

Table 2. Classification result on the ImageNet dataset.

Algorithm	Required data	FLOPS	#params	Top-1 acc	Top-5 acc
Teacher	Original data	∼3.67G	∼22M	73.27%	91.26%
Student	Original data	∼1.82G	∼12M	67.00%	87.60%
Knowledge Distillation [15]	Original data	∼1.82G	∼12M	68.67%	88.76%
PU-compression [36]	PU data	∼1.82G	∼12M	61.92%	86.00%
DFND (ours)	unlabeled data	∼1.82G	∼12M	61.75%	85.93%

proposed method on the CIFAR-100 dataset. The accuracy of the teacher network is 77.84% and the performance of the student network is only 76.53%. Data-Free Learning (DAFL) can obtain a 74.47% accuracy without any real-world training data while DFAD only achieves a 67.70% accuracy. Using the randomly selected data achieves a 73.25% accuracy and the proposed noisy distillation method achieves a 76.35% accuracy, which is much higher than other approaches.

6.2. Ablation Study

To further demonstrate the effectiveness of the proposed method, we conduct visualization experiments and ablation study on the CIFAR dataset.

Visualization of selected data. We visualize the selected data from the ImageNet dataset using the ResNet-34 network pre-trained on CIFAR-10 dataset. Figure 2 shows the images in the CIFAR-10 dataset, the selected and unselected images of the ImageNet dataset, respectively. In practice, the select images are similar with the original data. For example, there are dogs and cats in these images, which are also included on the categories of the CIFAR-10 dataset. In contrast, the unselected images consists of classes of humans and scenes which are not exist the original CIFAR-10 dataset. This visualization result demonstrates that the proposed data selection scheme in Eq. 10 can successfully select the useful training data from the massive unlabeled data. By using these selected data, the student network can successfully learn useful information from the pre-trained teacher network.

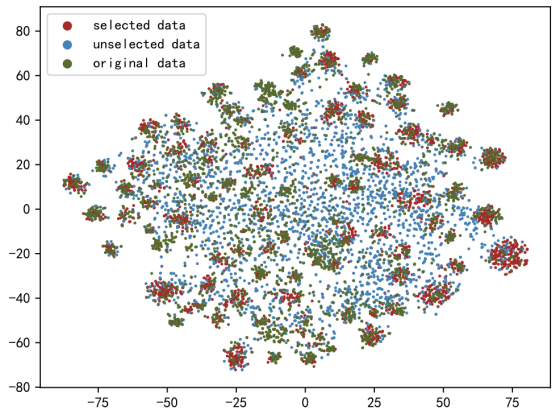


Figure 3. T-sne visualization of features generated by selected images, unselected images from ImageNet and original images on CIFAR-10 using ResNet-34. Each part consists of 500 images.

Visualization of features. To further investigate the effectiveness of our data collection method, we visualize the features before the fully connect layer generated by different data from the teacher networks. The pre-trained ResNet-34 network in the CIFAR-100 dataset is used as the teacher network. We take 5000 images from the CIFAR-100 dataset, the selected images using the proposed data collected method and unselected images from the ImageNet dataset for visualization. As shown in Figure 3, the features of selected data has similar distribution with those of original data. The visualization results demonstrate that the selected data can inherit similar information with the original data from the teacher network.

Table 3. Effectiveness of different learning strategies used in our method.

Initialization of Q	$Q = I$	Eq. 14
Fixed Q	75.60%	75.85%
Learnable Q	76.07%	**76.35%**
Without Q	75.73%	75.73%

Ablation study. To have an explicit understanding of the proposed method, we then evaluate the effectiveness of the noisy adaptation matrix in our method on the CIFAR-100 dataset. Table 3 shows the impacts of using different initializations for the noisy adaptation matrix Q. The impact of using a fixed or learnable Q is also investigated. The line "Without Q" means that we only use Eq. 2 to train the student network instead of Eq. 11, which can only achieve a 75.73% accuracy. By introducing the noisy adaptation matrix Q taking Eq. 14 as initialization, we can achieve a 75.85% accuracy. If Q is learnable, the proposed method can achieve the best performance, *i.e.*, 76.35% accuracy in the CIFAR-100 dataset, which is closed to the accuracy of the student network trained with original data (76.57%).

6.3. Classification Results on ImageNet

We then conduct experiments on the ImageNet dataset, which consists of ∼1.2 million images. Images in this dataset are resized into $3 \times 256 \times 256$ RGB images and are randomly cropped into 224×224 size for training the teacher and student networks. It is not easy for the existing data-free methods [2, 7] to generate so many high-resolution images only using the information from a pre-trained teacher network. Thus, these methods can hardly be applied on the ImageNet dataset. To this end, we propose the data collection method to utilize the massive unlabeled data in the wild. We use the

Table 4. Semantic segmentation results on the NYUv2 dataset.

Algorithm	Required data	FLOPS	#params	mIOU
Teacher	Original data	∼41.0G	∼24M	0.517
Student	Original data	∼5.54G	∼3.4M	0.375
Knowledge Distillation [15]	Original data	∼5.54G	∼3.4M	0.380
DAFL [2]	No data	∼5.54G	∼3.4M	0.105
DFAD [7]	No data	∼5.54G	∼3.4M	0.364
DFND (ours)	unlabeled data	∼5.54G	∼3.4M	0.378

Flicker1M dataset as the unlabeled dataset, which contains 1 million images. The pre-trained ResNet-34 network on the ImageNet dataset is used as the teacher model and a randomly initialized ResNet-18 is used as the student model. For training the student network, we use 0.1 initial learning rate with 5×10^{-4} weight decay and 0.9 momentum. We train 110 epochs in both steps and divide the learning rate by 10 every 30 epochs as suggested in [36] for maintaining the performance of the student networks.

Table 2 reports the classification results on the ImageNet dataset using different methods. The teacher ResNet-34 model achieves 73.27% top-1 accuracy and 91.26% top-5 accuracy, while requires $41.0G$ FLOPs and $24M$ parameters, which is not affordable for mobile devices. The student ResNet-18 model trained using the original ImageNet dataset achieves a 67.00% top-1 accuracy and a 87.60% top-5 accuracy, and have a lower computational complexity. By applying knowledge distillation, the student network can achieve a 68.67% top-1 accuracy and a 88.76% top-5 accuracy. The PU-compression method [36] utilizes the few original data, *i.e.*, 1000 images from the ImageNet data and the unlabeled dataset to conduct the experiment. This method achieves a 61.92% top-1. However, few data from the original dataset is usually unavailable. Therefore, we apply our data-free noisy distillation method by using the Flicker1M dataset as the unlabeled dataset. As a result, the proposed method achieves a 61.75% top-1 accuracy and a 85.93% top-5 accuracy without any original data and with only unlabeled data, which is very closed to those of PU compression method. These results demonstrate the effectiveness of the proposed method for selecting suitable training images from massive unlabeled data and learning student networks on the large-scale dataset.

6.4. Segmentation Results

Besides the image classification, the proposed method can also be applied on segmentation tasks, since semantic segmentation can be regarded as a pixel-level visual recognition problem. We adopt the calculation noisy value on each pixel of images and average the per-pixel noisy values for each image.

We conduct experiment on the NYUv2 dataset as that in [7]. This dataset contains 1449 images from 13 different classes. The images in the experiments are resized and cropped to 128×128. We use FCN with ResNet-50 backbone as the teacher model while that with MobileNetV2 backbone as the student model. The teacher model is trained on the NYUv2 dataset and the student networks are randomly initialized. For the proposed method, we use the ImageNet dataset as the unlabeled dataset.

Table 4 reports the segmentation results using different algorithms. The teacher network achieves a 0.517 mIOU while requires 41.0G FLOPs. The student networks trained using the original data achieves a 0.375 mIOU and requires only 5.54G FLOPs. Applying the knowledge distillation technique brings up the performance of the student network to a 0.380 mIOU. When the original training data is unavailable, traditional compression methods cannot be directly applied. Chen *et al*. [2] and Fang *et al*. [7] use the generated data to train the student network and achieve 0.105 and 0.364 mIOU values, respectively, which are lower than that of the baseline student network using the original dataset. In contrast, the proposed noisy distillation method achieves a 0.378 mIOU, which surpasses all the existing data-free approaches and demonstrates that the proposed method can be successfully applied on semantic segmentation task.

7. Conclusion

Since original data is often unavailable when compressing the pre-trained networks, a lot of data-free model compression methods are developed for generating the training data. However, the performance of compressed networks using these methods is limited due to the difficulty of image generation, especially on large-scale datasets. In this paper, we propose a two-step framework to compress the given network using massive unlabeled data effectively. First, we develop a data selection method by analyzing the bound of the distillation loss of the original data. Second, since the selected data is unlabeled, we propose a noisy distillation scheme by introducing a noisy adaptation layer to eliminate the noisy of the labels generated by teacher network. As a result, the proposed method achieves the state-of-the-art performance among all the data-free compression methods.

Acknowledgment This work is supported by National Natural Science Foundation of China under Grant No. 61876007, and Australian Research Council under Project DE180101438 and DP210101859.

References

[1] Kartikeya Bhardwaj, Naveen Suda, and Radu Marculescu. Dream distillation: A data-independent model compression framework. *arXiv preprint arXiv:1905.07072*, 2019. 2

[2] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *ICCV*, 2019. 2, 3, 6, 7, 8

[3] Hanlin Chen, Baochang Zhang, Xiawu Zheng, Jianzhuang Liu, David Doermann, Rongrong Ji, et al. Binarized neural architecture search. In *AAAI*, 2020. 2

[4] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *CVPR*, 2020. 2

[5] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016. 2

[6] Minjing Dong, Hanting Chen, Yunhe Wang, and Chang Xu. Crafting efficient neural graph of large entropy. In *IJCAI*, 2019. 2

[7] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019. 2, 3, 6, 7, 8

[8] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 20167. 5

[9] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014. 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[11] Jianyuan Guo, Kai Han, Yunhe Wang, Chao Zhang, Zhaohui Yang, Han Wu, Xinghao Chen, and Chang Xu. Hit-detector: Hierarchical trinity architecture search for object detection. In *CVPR*, 2020. 1

[12] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *CVPR*, 2020. 2

[13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 6, 7, 8

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[17] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008. 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[19] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML workshop*, 2013. 4

[20] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. 2016. 1, 2

[21] Mingbao Lin, Rongrong Ji, Yuxin Zhang, Baochang Zhang, Yongjian Wu, and Yonghong Tian. Channel pruning via automatic structure search. *arXiv preprint arXiv:2001.08565*, 2020. 2

[22] Shaohui Lin, Rongrong Ji, Chenqian Yan, Baochang Zhang, Liujuan Cao, Qixiang Ye, Feiyue Huang, and David Doermann. Towards optimal structured cnn pruning via generative adversarial learning. In *CVPR*, 2019. 2

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[24] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 2

[25] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *CVPR*, 2017. 2

[26] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, 2019. 2

[27] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, R Venkatesh Babu, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. *arXiv preprint arXiv:1905.08114*, 2019. 2

[28] Pantelimon G Popescu, Sever S Dragomir, EMIL I Sluşanschi, and OCTAVIAN N Stănăşilă. Bounds for kullback-leibler divergence. *Electronic Journal of Differential Equations*, 2016, 2016. 4

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[30] Bharat Bhusan Sau and Vineeth N Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016. 5

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[32] Yehui Tang, Yunhe Wang, Yixing Xu, Hanting Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. A semi-supervised assessor of neural architectures. In *CVPR*, 2020. 2

[33] Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing Xu, Chao Xu, and Chang Xu. Scop: Scientific control for reliable neural network pruning. *NeurIPS*, 2020. 2

[34] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019. 5

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 4

[36] Yixing Xu, Yunhe Wang, Hanting Chen, Kai Han, XU Chunjing, Dacheng Tao, and Chang Xu. Positive-unlabeled compression on the cloud. In *NeurIPS*, 2019. 6, 7, 8

[37] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *CVPR*, 2020. 2

[38] Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. In *CVPR*, 2020. 2

[39] Zhaohui Yang, Yunhe Wang, Kai Han, Chunjing Xu, Chao Xu, Dacheng Tao, and Chang Xu. Searching for low-bit weights in quantized neural networks. *NeurIPS*, 2020. 2

[40] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020. 2, 6

[41] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *CVPR*, 2019. 1

[42] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, 2019. 1