Hierarchical Knowledge Squeezed Adversarial Network Compression

Peng Li,[1,2*] **Changyong Shu,**[1,2*] **Yuan Xie,**[1†] **Yanyun Qu,**[3†] **Hui Kong**[2]
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Nanjing Institute of Advanced Artificial Intelligence, Horizon Robotic, Nanjing, China
[3]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Fujian, China
{peng03.li, changyong.shu, hui.kong}@horizon.ai, yxie@cs.ecnu.edu.cn, yyqu@xmu.edu.cn

Abstract

Deep network compression has been achieved notable progress via knowledge distillation, where a teacher-student learning manner is adopted by using predetermined loss. Recently, more focuses have been transferred to employ the adversarial training to minimize the discrepancy between distributions of output from two networks. However, they always emphasize on result-oriented learning while neglecting the scheme of process-oriented learning, leading to the loss of rich information contained in the whole network pipeline. Whereas in other (non GAN-based) process-oriented methods, the knowledge have usually been transferred in a redundant manner. Observing that, the small network can not perfectly mimic a large one due to the huge gap of network scale, we propose a knowledge transfer method, involving effective intermediate supervision, under the adversarial training framework to learn the student network. Different from the other intermediate supervision methods, we design the knowledge representation in a compact form by introducing a task-driven attention mechanism. Meanwhile, to improve the representation capability of the attention-based method, a hierarchical structure is utilized so that powerful but highly squeezed knowledge is realized and the knowledge from teacher network could accommodate the size of student network. Extensive experimental results on three typical benchmark datasets, i.e., CIFAR-10, CIFAR-100, and ImageNet, demonstrate that our method achieves highly superior performances against state-of-the-art methods.

Introduction

Deep neural networks (DNNs) greatly enhance the development of artificial intelligence via dominant performance in diverse perception missions. Due to the fact that the highly computational consumption problem in modern DNNs usually restricts their direct implementation on embedded systems, there is a trend in expediting the development of network compression. The network compression can accelerate neural networks for real-time applications on edge-computing devices in the following aspects: low-rank decomposition (Denton et al. 2014; Yang et al. 2015), network pruning (Han, Mao, and Dally 2015; Belagiannis and Zisserman 2017), quantization (Courbariaux et al. 2016; Rastegari et al. 2016), knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), and compact network design (Iandola et al. 2016; Zhang et al. 2018).

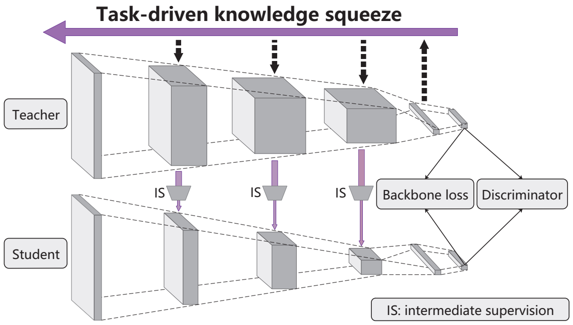

Figure 1: The overview of proposed Hierarchical Knowledge Squeezed Adversarial Network Compression (HKSANC) via intermediate supervision.

Among the above categories, KD is somewhat different due to the utilization of information from the pre-trained teacher network. (Hinton, Vinyals, and Dean 2015) forced the output of student network to match the soft targets produced by a teacher network via KL divergence. In recent years, many researchers resort to process-oriented methods, and many kinds of knowledge representation algorithms have been proposed (Zagoruyko and Komodakis 2016; Yim et al. 2017). Empirically, the loss learned by adversarial training usually has advantages over the predetermined one in the student-teacher strategy, (Belagiannis, Farshad, and Galasso 2018) and (Xu, Hsu, and Huang 2018) proposed the GAN-based distillation approaches by introducing the discriminator to match the output distribution between teacher and student.

Motivations: 1) To inherit the information from teacher network, the aforementioned GAN-based methods usually focus on result-oriented learning. While reasonably effective, rich information encoded in the intermediate layers

[*]Equal contribution

[†]Corresponding author

of teacher network might be ignored. As for the process-oriented methods, the knowledge is always represented as feature- or heat-maps.

We acknowledge that the small network cannot mimic a large one perfectly due to the large difference in the number of layers and the large gap in representation capability between large and small networks. To reduce this gap, we would better squeeze the redundancy knowledge contained in the teacher network into a compact form to accommodate the size of student network. Therefore, we propose the knowledge-squeeze method, a task-driven attention mechanism in this paper. It can convert the knowledge in the form of 3D feature map into a vector in an elegant manner. Moreover, by incorporating the intermediate supervision, the squeezed knowledge can be effectively injected into the student network.

2) In the above mentioned task-driven attention mechanism, the squeezed knowledge can be achieved with a global descriptor vector as input. However, due to the semantic gap, the feature maps of shallow layer cannot directly match the global descriptor on semantic level, resulting in a low effectiveness of the squeezed knowledge of shallow layer for knowledge transfer. To overcome this, we need to gradually transfer high-level semantic information from deep to shallow. Therefore, we introduce hierarchical connections between high-level squeezed knowledge and low-level attention blocks in an attention mechanism.

The major contribution of this paper is a Hierarchical Knowledge Squeezed Adversarial Network Compression (HKSANC) via intermediate supervision, shown in Figure 1. Specifically, (i) A novel knowledge transfer method, which involves an effective intermediate supervision, is proposed based on the adversarial training framework. A task-driven attention mechanism is introduced to achieve the highly compact knowledge representation, which can accommodate the small size of student network. (ii) To improve the representation capability of low-level attention block, a hierarchical connection structure is introduced in the attention-based method so that the highly squeezed knowledge could be extracted. (iii) We conduct an extensive evaluation of our method on several benchmark datasets, where the experimental results demonstrate that our method achieves highly competitive performance compared with some other knowledge transfer approaches, while maintaining smaller accuracy degradation.

Related Work

Network Compression: We briefly review the following five kinds of approach for deep network compression. (1) Low rank decomposition: In this case, the main idea is to construct low rank basis of filters to effectively reduce the weight tensor. Related approaches have also been explored by the principle of finding a low-rank approximation for the convolutional layers (Rigamonti et al. 2013; Lebedev et al. 2014; Yang et al. 2015). (2) Network pruning: Removing network connections not only reduces the model size but also prevents over-fitting. Parameter sharing has also contributed to reduce the network parameters with repetitive patterns (Schmidhuber 1992; Belagiannis and Zisserman 2017). (3) Quantization: The main goal of quantization are reducing the size of memory requirement and accelerates the inference by using weights with lower precision representations (Soudry, Hubara, and Meir 2014; Courbariaux, Bengio, and David 2015; Rastegari et al. 2016). (4) Compact network design: Many researchers resort to derive more efficient network architectures, such as ResNets (He et al. 2016a), SqueezeNet (Iandola et al. 2016) and ShuffleNet (Zhang et al. 2018), to shrink the number of the parameters while maintaining the performance. (5) Distillation: The proposed method belongs to this category, which will be discussed in detail in the next paragraph.

Knowledge Distillation: Knowledge distillation (Ba and Caruana 2014) is used to transfer knowledge from teacher network to student network by the output before the softmax function (logits) or after it (soft targets), which has been popularized by (Hinton, Vinyals, and Dean 2015). As it is hard for student network with small capacity to mimic the outputs of teacher network, several researches (Belagiannis, Farshad, and Galasso 2018; Xu, Hsu, and Huang 2018) focused on using adversarial networks to replace the manually designed metric such as $L1/L2$ loss or KL divergence.

Attention-based Knowledge Transfer: Early work on attention based tracking was motivated by human attention mechanism theories (Rensink 2000), and was accomplished via Restricted Boltzmann Machines. It was exploited in many computer-vision-related tasks. In transfer learning, attention transfer facilitates the fast optimization and improves the performance of a small student network via the attention map (Zagoruyko and Komodakis 2016) or the flow of solution procedure (FSP) matrix (Yim et al. 2017). Attention transfer is also introduced in machine reading comprehension (Hu et al. 2018).

The above transfer methods emphasize on how to represent the intermediate information more effectively, while neglecting to construct a compact form of knowledge for transfer, which is more important in compression since we have acknowledged that the capability of the small network is far below the deep one. Different from them, we propose to squeeze the intermediate knowledge by using a task-driven attention (Jetley et al. 2018), such that the highly compact knowledge from teacher network could accommodate the size of student network.

Method

The Architecture of HKSANC

As illustrated in Figure 2, our method consists of the teacher, student and discriminator networks. We denote the teacher network (require pre-train) as T, and the student network as S. Both the teacher and student network are built by a backbone-subnetwork (e.g., VGG, ResNet) Net_b and an attention-subnetwork Net_a. The Net_b is a standard CNN pipeline with N ($N = 3$ in Figure 2) replicated blocks, where we can obtain N corresponding intermediate features $L_{T/S} = \{L^1_{T/S}, L^2_{T/S}, \ldots, L^N_{T/S}\}$.

In attention-subnetwork, the attention block produce the squeezed knowledge vector ($\widetilde{g}^{a,i}_{T/S}$, $i = 1, \ldots, N$) corre-

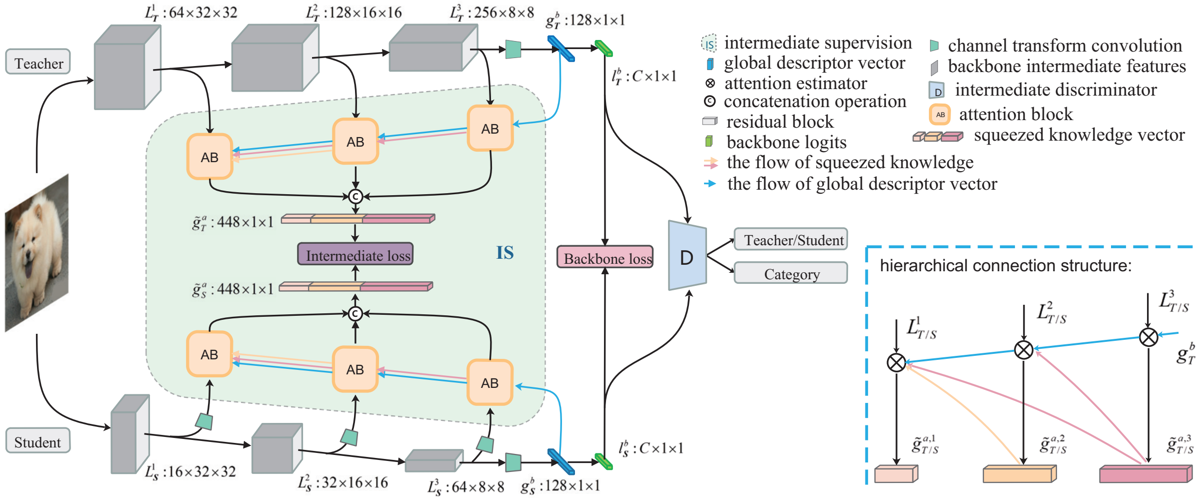
Teacher
Student
$L_T^1:64\times32\times32$
$L_T^2:128\times16\times16$
$L_T^3:256\times8\times8$
$g_T^b:128\times1\times1$
$l_T^b:C\times1\times1$
AB
$\tilde{g}_T^a:448\times1\times1$
Intermediate loss
IS
$\tilde{g}_S^a:448\times1\times1$
Backbone loss
D
Teacher/Student
Category
$L_S^1:16\times32\times32$
$L_S^2:32\times16\times16$
$L_S^3:64\times8\times8$
$g_S^b:128\times1\times1$
$l_S^b:C\times1\times1$
IS intermediate supervision
global descriptor vector
⊗ attention estimator
© concatenation operation
residual block
backbone logits
the flow of squeezed knowledge
the flow of global descriptor vector
channel transform convolution
backbone intermediate features
D intermediate discriminator
AB attention block
squeezed knowledge vector
hierarchical connection structure:
$L_{T/S}^1$
$L_{T/S}^2$
$L_{T/S}^3$
g_T^b
$\tilde{g}_{T/S}^{a,1}$
$\tilde{g}_{T/S}^{a,2}$
$\tilde{g}_{T/S}^{a,3}$

Figure 2: The architecture of our proposed HKSANC. The sub-figure in the left is the paradigm of our teacher-student strategy. The sub-figure in the bottom right corner of the dotted line is the description of hierarchical connection structure of attention block. This is an example where the teacher is ResNet-164 and the student is ResNet-20, and C is the category number (best viewed in color).

sponding to the i-th block. Another output of the attention block is the vectors that send to the low-level attention block. As shown in the bottom right corner of Figure 2, for a certain attention estimator, the squeezed knowledge vectors obtained from higher-level are connected with it in a hierarchical manner, which is beneficial to produce effective representation. As a result, the low-level attention estimator not only takes the global descriptor vector, but also the squeezed knowledge vectors from high-level as inputs. The detailed structure of the task-driven attention estimator is presented in Figure 3. Firstly, the input vectors are concatenated. Then the channel alignment convolution is introduced to achieve the feature vector $\hat{g}_{T/S}^{b,i}$ as the channel number of input might differ from that of $L_{T/S}^i$. Next, the element-wise sum between $\hat{g}_{T/S}^{b,i}$ and $L_{T/S}^i$ is conducted along the dimensionality of channel, so that the $\hat{L}_{T/S}^i$ is obtained. Finally, the squeezed knowledge descriptor $\widetilde{g}_{T/S}^{a,i}$ for the i-th block can be gained by the following equation:

$$M = \text{softmax}(\widehat{W} * \hat{L}_{T/S}^i) \tag{1}$$

$$\widetilde{g}_{T/S}^{a,i} = \text{average\_pooling}(M \odot \hat{L}_{T/S}^i) \tag{2}$$

where the $\widehat{W}$, a $C \times 1 \times 1$ convolution kernel, is used to compute the attention score M, and $*$ denotes the convolutional operator, and $\odot$ is the element-wise product.

With the help of attention estimator, the knowledge contained in each backbone block of the teacher network, which in the form of 3D feature map could be converted into the compact form of a vector by integrating the task-specific information and reducing the redundancy, such that the so-called knowledge squeeze can be realized.

As for the discriminator, it aims to distinguish where the input vectors come from (teacher or student). The discriminator is composed of three sequentially stacked fully-connected layers. The number of nodes in all hidden layers is the dimension of the input. We use backbone logits (denoted by $l_{T/S}^b$) as the input of the discriminator in our experiments.

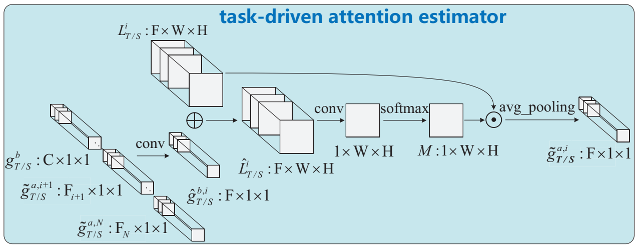
task-driven attention estimator
$L_{T/S}^i:\text{F}\times\text{W}\times\text{H}$
$g_{T/S}^b:\text{C}\times1\times1$
$\tilde{g}_{T/S}^{a,i+1}:\text{F}_{i+1}\times1\times1$
$\tilde{g}_{T/S}^{a,N}:\text{F}_N\times1\times1$
conv
$\hat{g}_{T/S}^{b,i}:\text{F}\times1\times1$
$\hat{L}_{T/S}^i:\text{F}\times\text{W}\times\text{H}$
conv
$1\times\text{W}\times\text{H}$
softmax
$M:1\times\text{W}\times\text{H}$
avg_pooling
$\tilde{g}_{T/S}^{a,i}:\text{F}\times1\times1$

Figure 3: The structure of designed task-driven attention estimator for i-th block, where the backbone intermediate feature $L_{T/S}^i$, the global descriptor $g_{T/S}^b$ and the squeezed knowledge vectors $\widetilde{g}_{T/S}^{a,k}(k = i+1, ..., N)$ are taken as input to produce the squeezed knowledge descriptor $\widetilde{g}_{T/S}^{a,i}$.

Overall Loss Function

To train our network, we define a loss function in Eqn. (3) including three components, i.e., the adversarial loss L_{adv}, the backbone loss L_b, and the intermediate loss L_{is}:

$$L = \lambda_1 L_b + \lambda_2 L_{adv} + \lambda_3 L_{is} \tag{3}$$

where λ_1, λ_2, λ_3 are trade-off factors. During the process of knowledge transfer, the backbone loss L_b is utilized to directly match the output of two networks, while the adversarial loss L_{adv} is employed to minimize the discrepancy between distributions of logits from the two networks.

Both of them are served for result-oriented learning. On the contrary, the intermediate loss L_{is} could facilitate the process-oriented learning via intermediate supervision. Consequently, both process-oriented and result-oriented learning can complement each other. We found that the result is insensitive to the scope (from 0.1 to 10.0) of trade-off factors in the experiments, thus the weight of each loss is set to be equal for simplify, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

Backbone Loss: Backbone loss is a result-oriented constraint which makes the student mimic the teacher by minimizing the $L2$ loss between backbone logits from the teacher and student networks:

$$L_b = \|l_S^b(x) - l_T^b(x)\|_2^2 \tag{4}$$

Adversarial Loss: In the proposed model, a GAN based approach is introduced to transfer knowledge from teacher to student. The teacher and student networks convert an image sample x to the logits $l_T^b(x)$ and $l_S^b(x)$ respectively, where the student is considered as a generator in vanilla GAN. While the discriminator aims to distinguish whether the input comes from teacher or student. As our goal is to fool the discriminator in predicting the same output for teacher and student networks, the objective can be written as:

$$\begin{aligned} L_{adv}^o = \min_{l_S^b} \max_D \quad & E_{l_T^b(x)\sim p_T}[\log(D(l_T^b(x)))] + \\ & E_{l_S^b(x)\sim p_S}[\log(1 - D(l_S^b(x)))] \end{aligned} \tag{5}$$

where p_T and p_S correspond to logits distribution of the teacher and student network, respectively.

In order to get more valuable gradient for student, the regularization and category-level supervision are introduced to further improve the discriminator. We utilize three regularizers to prolong the minimax game between the student and discriminator as follows:

$$L_{reg} = -\mu\bigg(|\omega_D| + \|\omega_D\|_2^2 - E_{l_S^b(x)\sim p_S}[\log(D(l_S^b(x)))]\bigg) \tag{6}$$

where ω_D is the parameters of discriminator, and μ controls the contribution of regularizer in optimization, the negative sign denotes that the loss term is updated in the maximization step. The first two terms force the weights of discriminator to grow slowly, the last term is referred to as the adversarial sample regularization, and the above loss terms are originally designed in (Xu, Hsu, and Huang 2018). It utilizes the additional student samples (labeled as teacher) to confuse the discriminator such that the capability of discriminator can be restricted to some extent.

Note that the adversarial loss defined in Eqn. (5) only focus on matching the logits on distribution-level, while missing the category information might result in the incorrect association between logits and labels. Consequently, the discriminator is further modified to simultaneously predict the "teacher/student" and the class labels. On this occasion, the output of discriminator is a $1{+}C$ dimensional vector (the first element represents "teacher/student", while the remaining denote the category by using one-hot encoding), and the category regularizer for the discriminator can be written as:

$$\begin{aligned} L_{adv}^C = \min_{l_S^b} \max_D \quad & E_{l_T^b(x)\sim p_T}[\log(P(l(x)|C_T(x)))] \\ & + E_{l_S^b(x)\sim p_S}[\log(P(l(x)|C_S(x)))] \end{aligned} \tag{7}$$

where $l(x)$ means the label of the sample x, $C_{T/S}(x)$ corresponds to $D(l_{T/S}^b(x))$[1:], where the [:] denotes the vector slice in python. As the discriminator has to learn with the extended outputs to jointly predict "teacher/student“ and category, the adversarial learning becomes more stable.

To sum up, the final loss for adversarial training can be formulated as:

$$L_{adv} = L_{adv}^o + L_{reg} + L_{adv}^C \tag{8}$$

Intermediate Loss: The loss of intermediate supervision is an effective term to inject the squeezed knowledge, i.e., $\widetilde{g}_T^{a,i}, (i = 1, \ldots, N)$, into the student network. Due to the highly compact form of squeezed knowledge, it can be given by the $L2$ distance:

$$L_{is} = \|\widetilde{g}_S^a(x) - \widetilde{g}_T^a(x)\|_2^2 \tag{9}$$

where $\widetilde{g}_{T/S}^a(x)$ is the concatenation of $\widetilde{g}_{T/S}^{a,i}(i = 1, ..., N)$. It is noteworthy that other loss function, such as $L1$ loss and cross-entropy loss, can also be applied. Besides, we find that $L2$ loss outperforms others empirically in our experiments.

Optimization

The optimization procedure of the proposed HKSANC contains two stages. First, the teacher is trained from scratch by using labeled data. The attention-subnetwork with additional auxiliary layers (i.e., one fully-connected layer and a softmax output layer) and backbone-subnetwork are trained simultaneously by optimizing the two cross-entropy losses, while the auxiliary layers are removed during the process of transfer learning. Second, fixing the teacher network, the student and discriminator are updated under the framework of adversarial training, where the number of steps inside each component is simply set to 1 in our experiments. Both student and discriminator are randomly initialized. We use Stochastic Gradient Descent (SGD) with momentum as the optimizer, and set the momentum as 0.9, weight decay as $1\mathrm{e}-4$. The learning rate, initialized as $1\mathrm{e}-1$ and $1\mathrm{e}-3$ for student and discriminator separately, is multiplied by 0.1 at three specific epochs during the training process. As for the regularization, having been examined the different values for weight factor μ in our experiments, we conclude that setting μ to 1 is a good compromise for all evaluations empirically. For all experiments, we train on the standard training set and test on the validation set. Besides, data augmentation (random cropping and horizontal flipping) and normalization (subtracted and divided sequentially by mean and standard deviation of the training images) is applied to all the training images.

Experiments

Experimental Setting

Datasets: We consider three image classification datasets: CIFAR-10, CIFAR-100, and ImageNet ILSVRC 2012. Both CIFAR-10 and CIFAR-100 contain 50K training images and 10K validation images, respectively. The ImageNet

ILSVRC 2012 contains more than 1 million training images from 1000 object categories and 20K validation images with each category including 20 images. The image size of CIFARs and ImageNet is 32×32 and 224×224, separately.

Evaluation Measures: We evaluate different models from the following two aspects: 1) the testing error of student network; 2) the convergence stability (S) for training procedure. As for the former, the Top-1 error is calculated for all datasets, while the Top-5 error is additionally adopted for ImageNet. The testing error in ablation study is the average of twenty runs. The convergence stability is computed by the concussion range of the testing error:

$$S = \mathrm{Var}(\mathit{Err}) \tag{10}$$

where Var denotes the variance calculation, *Err* is $[\mathit{err}_1^{\max} - \mathit{err}_1^{\min}, ..., \mathit{err}_E^{\max} - \mathit{err}_E^{\min}]$, $\mathit{err}_e^{\max}$ and $\mathit{err}_e^{\min}$ denote the maximum and the minimum error rate over twenty runs on e-th epoch ($e = 1, \ldots, E$), respectively.

Competitors: Since our proposed method is closely related with the knowledge transfer based on attention mechanism and the adversarial training, the following works should be included in our experiments. As for knowledge transfer, four representative knowledge transfer methods need to be analyzed: the attention map computed by the statistics of feature values across the channel dimension (Zagoruyko and Komodakis 2016), the FSP matrix generated by the inner product of two features (Yim et al. 2017), minimizing the maximum mean discrepancy to match the distributions of neuron selectivity patterns from two networks (Huang and Wang 2017), and the method that formulates knowledge transfer as maximizing the mutual information (Ahn et al. 2019). These four approaches produce the transferred knowledge in a heuristic manner, while our model achieves a more compact one via the way of task-driven learning. For fair comparison, we adopt their representation of transferred knowledge into our framework.

As for adversarial training, two recently representative methods, i.e., adversarial network compression (ANC, Belagiannis, Farshad, and Galasso 2018) and training student network with conditional adversarial networks (TSCAN, Xu, Hsu, and Huang 2018), are included. We implement the above two GAN based approaches on our own. Note that the backbone network in TSCAN, i.e., wide residual networks (WRN), is replaced by the ResNet, and the cross-entropy loss as well as the KD loss for student update are removed for fair comparison.

Finally, for the purpose of comprehensive comparison, we introduce four additional knowledge distillation methods: mimic learning with *L*2 loss (L2-Ba, Ba and Caruana 2014), distillation with soft targets via KL divergence (KD, Hinton, Vinyals, and Dean 2015), knowledge transfer with FSP matrix (FSP, Yim et al. 2017), and Fitnets (Romero et al. 2014). Four quantization methods: weights binarization during training process except parameters update (BinaryConnect, Courbariaux et al. 2016), reducing the precision of the network weights to ternary values (Quantization, Zhu et al. 2016), binaryzation of the filters (BWN) and the additional input (XNOR) (Rastegari et al. 2016).

Implementation Details: For CIFAR-10 and CIFAR-100, we set the pre-trained teacher as ResNet-164, the student as ResNet-20*, where preact block (He et al. 2016b) is employed since it is currently the standard architecture for recognition. We select the minibatch size as 64 and total train epoch as 600 with the learning rate multiplied by 0.1 at epoch 240 and epoch 480. We adjust the teacher network to ResNet-152 like ANC and TSCAN since the pre-trained model is available, the student network is changed to ResNet-50/18, the mini-batch size is set to 128, and the total epoch is 120 for ImageNet dataset, where the learning rate is divided by 10 at epoch 30, 60 and 90. Our implementation is based on Pytorch, with 1 and 4 NVIDIA GTX 1080ti GPU for CIFAR-10/100 and ImageNet, separately.

Ablation Study

Comparison of Connection Methods of Attention Block: In our proposed model, the squeezed knowledge is achieved by task-driven attention mechanism with hierarchical connection structure, which is different from the original attention method (Jetley et al. 2018), whose attention estimators merely take the global descriptor vector as input (the method would be presented if the pink and orange lines are removed in the bottom right corner of Figure 2). To proof that the squeezed knowledge achieved with the hierarchical structure is more effective, we apply the two methods in knowledge distillation and observe the classification error of student on the two benchmark datasets. The result is shown in Table 1. L_b and L_{is} denote the backbone loss and intermediate loss respectively. The only difference between original and our method is the connection structure of attention block. Obviously, the method with hierarchical structure performs better than the one without the architecture, which demonstrates that the squeezed knowledge achieved by using the hierarchical connection structure is more valuable for knowledge transfer.

Loss composition	Error [%]	
	CIFAR-10	CIFAR-100
$L_b + L_{is}$[original]	7.55	31.97
$L_b + L_{is}$ [our]	**7.51**	**31.86**

Table 1: The comparison of connection methods of attention block. The same backbone loss L_b is employed.

Comparison of Knowledge Transfer Methods: We aim to demonstrate that the task-driven attention mechanism is a more effective way to squeeze the knowledge transferred from teacher to student than the methods of the attention map (AT, Zagoruyko and Komodakis 2016), FSP matrices (FSP, Yim et al. 2017), matching the distribution of neuron selectivity patterns (NST, Huang and Wang 2017) and the mutual information maximization (VID, Ahn et al. 2019). To do so, we compare these five types of intermediate knowledge transfer methods.

*Note that the teacher and student networks do not restrict to one certain type, any other network, such as WRN, can be used in the same way.

As illustrated in Table 2, it demonstrates that the backbone loss L_b, combining with intermediate loss L_{is} with any form of transferred knowledge (appointed in bracket in the first column in Table 2), could facilitate the training of student network, especially our squeezed transferred knowledge, which outperforms other representations by a significant margin. It indicates that, by integrating the task-driven attention scheme, our squeezed knowledge is a more suitable representation to be adapted into the small scale network.

Loss composition	Error [%]	
	CIFAR-10	CIFAR-100
L_b	8.19	32.60
$L_b + L_{is}$ [AT]	7.97	32.31
$L_b + L_{is}$ [VID]	8.02	32.51
$L_b + L_{is}$ [FSP]	8.03	32.45
$L_b + L_{is}$ [NST]	8.07	32.36
$L_b + L_{is}$[ours]	**7.51**	**31.86**

Table 2: The evaluation of different knowledge transfer methods. The same backbone loss L_b is applied. The effect of different intermediate losses L_{is} and ours is studied.

Loss Functions for Discriminator: Since the proposed HKSANC is built upon the adversarial training framework, it is necessary for us to find the reasonable combination of loss functions for the discriminator. As it is shown in Table 3, generally speaking, any adversarial loss can improve the performance of student network. Specifically, by comparing the difference between the line 1 and 2, and the difference between the line 1 and 3, either category regularizer L^C_{adv} or discriminator regularizer L_{reg} could boost the performance slightly. However, their joint constraint (see the last row in Table 3) will lead to a remarkable improvement, which indicates that both L^C_{adv} and L_{reg} play critical roles in our adversarial training model.

Loss composition	Error [%]	
	CIFAR-10	CIFAR-100
$L_b + L_{is}$	7.51	31.86
$L_b + L_{is} + L^o_{adv} + L_{reg}$	7.44	31.55
$L_b + L_{is} + L^o_{adv} + L^C_{adv}$	7.36	31.65
$L_b + L_{is} + L_{adv}$	**7.29**	**31.35**

Table 3: The evaluation of different components of adversarial loss. The same backbone loss L_b and intermediate loss L_{is} are employed.

Benefits of Intermediate Supervision: We look into the effect of enabling and disabling different loss components of HKSANC model, as shown in Table 4. We can see that even merely using backbone loss L_b could be able to obtain a better effect than student network without any knowledge transfer (directly supervised learning by using sample-label pair). Moreover, both L_{adv} and L_{is} could improve the performance of student network.

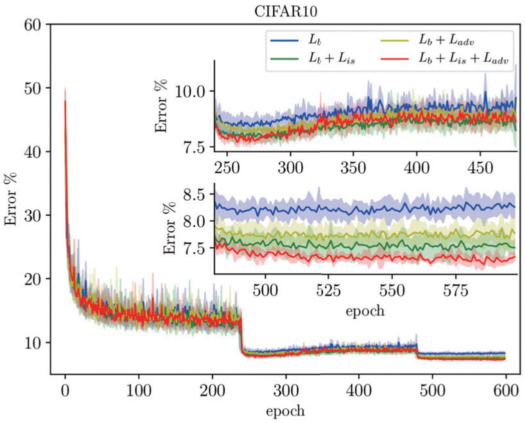

Figure 4: The training procedure of four different models on CIFAR-10, where the X-axis denotes training epochs and the Y-axis denotes testing error.

Interestingly, utilizing L_{is} get better result than L_{adv}. We give the explanation as follows: Recall the subsection of loss function, both L_b and L_{adv} give service to the result-oriented learning, which is naturally without the guidance from intermediate supervision (see line 4 in Table 4). As a result, significant improvement can not be acquired by incorporating another result-oriented loss function. On the contrary, further improvement can be realized by adding the intermediate loss L_{is} to L_b (see line 3 in Table 4), which is an evidence that both of them can complement each other. Furthermore, the final approach combining all of the loss components preforms the best, this is attributed to the advantage of the adversarial training.

Loss composition	Error [%]	
	CIFAR-10	CIFAR-100
supervised learning	8.58	33.36
L_b	8.19	32.60
$L_b + L_{is}$	7.51	31.86
$L_b + L_{adv}$	7.72	32.17
$L_b + L_{adv} + L_{is}$	**7.29**	**31.35**

Table 4: The effect of loss components in HKSANC.

Intuitively, the benefit of combining different losses will be presented in the training procedure, e.g., the training curve. Figure 4 represents the test error of different models over time on CIFAR-10 (the result on CIFAR-100 is presented in supplemental material). Our model ($L_b + L_{adv} + L_{is}$, the red line in Figure 4) has a relatively lower testing error during the training process, especially after epoch 240, see the zoom-in windows in the upper right.

Moreover, we select the last 100 epochs to calculate the convergence stability of different models, which is illustrated in Table 5. By comparing the first two lines and last two lines respectively, the convergence stability becomes more apparent after integrating the loss function of intermediate supervision with squeezed knowledge, which further indicates that the intermediate loss could improve the stability of transfer learning.

Model	Stability	
	CIFAR-10	CIFAR-100
L_b	$2.46\text{e}-3$	$5.12\text{e}-3$
$L_b + L_{is}$	$2.18\text{e}-3$	$4.51\text{e}-3$
$L_b + L_{adv}$	$3.02\text{e}-3$	$6.41\text{e}-3$
$L_b + L_{adv} + L_{is}$	$2.36\text{e}-3$	$5.80\text{e}-3$

Table 5: Convergence Stability. The variance of testing error concussion range on CIFAR-10 and CIFAR-100 through last 100 epochs are shown.

Comparison with State-of-the-art

We first compare our model with several cutting edge compression approaches, including six distillation algorithms and four quantization ones. Eight of them, i.e., six distillations and two quantization approaches, are available for CIFAR-10 and CIFAR-100 datasets, other unavailable methods, whose results are not provided by their authors, are omitted in the reported table. Similar way is adopted in the experiment for ImageNet dataset.

Model	Param	Error [%]	
		CIFAR-10	CIFAR-100
Teacher RN-164	2.6M	6.57	27.76
Student RN-20	0.27M	8.58	33.36
FSP	0.27M	11.30	36.67
L2-Ba	0.27M	9.07	32.79
KD	0.27M	8.88	33.34
FitNets	2.5M	8.39	35.04
Quantization	0.27M	8.87	–
Binary-Connect	15.20M	8.27	–
ANC	0.27M	8.08	32.45
TSCAN	0.27M	7.93	32.57
HKSANC	0.27M	**7.29**	**31.35**

Table 6: Comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100. The teacher and student networks are RN-164 and RN-20 respectively. (RN denotes ResNet.)

As shown in the Table 6, the deep teacher preforms much better than the shallow student network with supervised learning (line 2 in Table 6), and the error rate of small network learned by using distillation models is bounded by the teacher's performance, as expected. Both distillation and quantization approaches obtain relatively good performance with a small model size. Specifically, two GAN based competitors (ANC and TSCAN) achieve desirable results which outperform the supervised learning for student with 0.5% and 0.65%, respectively. Obviously, the proposed HKSANC further boost the capability of student network. More noticeable improvements can be seen on CIFAR-100 dataset. To sum up, we can see that our method acquires the lowest error with the same or less number of parameters, which demonstrates that our model benefits from effective representation of transferred knowledge and intermediate supervision.

Model	Param	Error [%]	
		Top-1	Top-5
Teacher RN-152	58.21M	27.63	5.90
Student RN-50	37.49M	30.30	10.61
Student RN-18	13.95M	43.33	20.11
XNOR (RN-18)	13.95M	48.80	26.80
BWN (RN-18)	13.95M	39.20	17.00
L2-Ba (RN-18)	13.95M	33.28	11.86
ANC (RN-18)	13.95M	32.89	11.72
TSCAN (RN-18)	13.95M	32.72	11.49
HKSANC(RN-18)	13.95M	**31.34**	**10.85**
L2-Ba (RN-50)	37.49M	27.99	9.46
ANC (RN-50)	37.49M	27.48	8.75
TSCAN (RN-50)	37.49M	27.39	8.53
HKSANC (RN-50)	37.49M	**26.72**	**7.97**

Table 7: Comparison with state-of-the-art methods on ImageNet. The teacher network is RN-152, and the student networks are RN-50 and RN-18. (RN denotes ResNet.)

More adequate evidence is provided by the comparison on large-scale dataset, i.e., ImageNet. Two widely used networks, ResNet-50 and ResNet-18, are employed as the student network, and the comparison results are presented in Table 7. By analyzing the results in the first group (line 1 to 3), we can deduce that ResNet-152 might contain redundancy since the difference between ResNet-152 and ResNet-50 is only 2.67%. In that case, the ANC, TSCAN, and our method could obtain desirable results which even beat the teacher network, as the ResNet-50 is a more concise architecture trained by the three distillation methods.

When the size of student network becomes smaller (ResNet-18), the error of the proposed method increases nearly 4.6% (from 26.72% to 31.34%), where our assumption that, "the small network can not perfectly mimic a large on especially when there exists significant difference in the number of layer.", can be confirmed. Nevertheless, the proposed method obtains the performance gain w.r.t the second best (TSCAN, see the last two lines in the second group of Table 7) better than that in the case of ResNet-50. This indicates that, the transferred knowledge in our model is much more suitable for injecting into the small scale network.

Conclusion

In this paper, a novel knowledge transfer method involving intermediate supervision is proposed via the framework of adversarial training for distillation. To inherit the information from teacher to student effectively, the task-driven attention mechanism is designed to squeeze the knowledge in a compact form for intermediate supervision. Moreover, the hierarchical connection structure is introduced in the attention mechanism to achieve more powerful knowledge representation. Extensive evaluation of our HKSANC is conducted on three challenging image classification datasets, where a clear outperformance over contemporary state-of-the-art methods is achieved. Additionally, the experimental results demonstrate that, the proposed attention transfer method could further facilitate the convergence stability via the intermediate supervision.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant 61876161, Grant 61772524, Grant U1065252 and partly by the Beijing Municipal Natural Science Foundation under Grant 4182067, and partly by the Fundamental Research Funds for the Central Universities associated with Shanghai Key Laboratory of Trustworthy Computing.

References

Ahn, S.; Hu, S. X.; Damianou, A.; Lawrence, N. D.; and Dai, Z. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9163–9171.

Ba, J., and Caruana, R. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, 2654–2662.

Belagiannis, V., and Zisserman, A. 2017. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 468–475. IEEE.

Belagiannis, V.; Farshad, A.; and Galasso, F. 2018. Adversarial network compression. In *The European Conference on Computer Vision (ECCV) Workshops*.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, 3123–3131.

Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; and Bengio, Y. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*.

Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, 1269–1277.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 630–645. Springer.

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

Hu, M.; Peng, Y.; Wei, F.; Huang, Z.; Li, D.; Yang, N.; and Zhou, M. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

Huang, Z., and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Jetley, S.; Lord, N. A.; Lee, N.; and Torr, P. H. 2018. Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

Lebedev, V.; Ganin, Y.; Rakhuba, M.; Oseledets, I.; and Lempitsky, V. 2014. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*.

Rastegari, M.; Ordonez, V.; Redmon, J.; and Farhadi, A. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, 525–542. Springer.

Rensink, R. A. 2000. The dynamic representation of scenes. *Visual Cognition* 7(1-3):17–42.

Rigamonti, R.; Sironi, A.; Lepetit, V.; and Fua, P. 2013. Learning separable filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2754–2761.

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

Schmidhuber, J. 1992. Learning complex, extended sequences using the principle of history compression. *Neural Computation* 4(2):234–242.

Soudry, D.; Hubara, I.; and Meir, R. 2014. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In *Advances in Neural Information Processing Systems*, 963–971.

Xu, Z.; Hsu, Y.-C.; and Huang, J. 2018. Training student networks for acceleration with conditional adversarial networks. *BMVC. British Machine Vision Association*.

Yang, Z.; Moczulski, M.; Denil, M.; de Freitas, N.; Smola, A.; Song, L.; and Wang, Z. 2015. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, 1476–1483.

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

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

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6848–6856.

Zhu, C.; Han, S.; Mao, H.; and Dally, W. J. 2016. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*.