Deep Hypersphere Feature embedding-based Knowledge Distillation

Eunsom Jeon
Arizona State Univ.
ejeon6@asu.edu

Yongbaek Cho
Arizona State Univ.
ycho56@asu.edu

Sanggu Park
Arizona State Univ.
spark230@asu.edu

Abstract

Most previous studies based on knowledge distillation aim to develop lightweight and efficient models by increasing their model size and computation scale. However, this traditional approach has a limitation on deploying the cumbersome deep models to devices with restricted resources (e.g., memory limitation). Neither does it provide low computational complexity. This paper investigates the following question: How to capture best the knowledge of a teacher model and how to improve the performance for training the small student model. This paper proposes a new knowledge distillation loss that is motivated by attention transfer and angular margin based training. Deep hypersphere feature embedding-based knowledge distillation guides the student to learn angularly attentive features from the teacher and get benefits for classification performance by a discriminative angular distance metric on a hypersphere manifold. The student can produce similar activations of the teacher as well as improve the performance. Our method is validated on two public datasets.

1. Introduction

In the recent decade, an area of 'Convolutional Neural Network' (CNN) has been tremendously and deeply studied so that industries commercialize the method to predict the market volatility with accumulated meta-data. However, a challenge still needs to overcome: a CNN's inherent challenge that inevitably demands high computational complexity and large storage requirements [6]. For this reason, the application is still limited to the environments which provide massive computational abilities. In reality, the demand for applying CNN on mobile devices is rapidly increasing according to the revolutionary development of IoT environments [10, 21]. To counter this challenge, many of the studies have developed a lightweight form of CNN models which assure the performance while lightening the network scale [7, 9, 21].

Knowledge distillation (KD) is one of the improvements on reducing the network size and developing an efficient network model to which recent studies have paid attention [2, 6, 23]. The concept of knowledge distillation is that the network consists of two networks, a larger one (teacher) and a smaller one (student) [9]. While the student is trained, the teacher transfers its knowledge to the student. So, the student, the smaller network model, can retain the teacher model's classification performance. However, contrary to a universal knowledge that the model performs better when its size is larger and trained time is longer, an experiment showed that KD performs better when a specific size of teacher model and the number of epochs are set [2]. For this reason, a proper scale of the teacher model and the number of epochs should be set in advance when implementing KD.

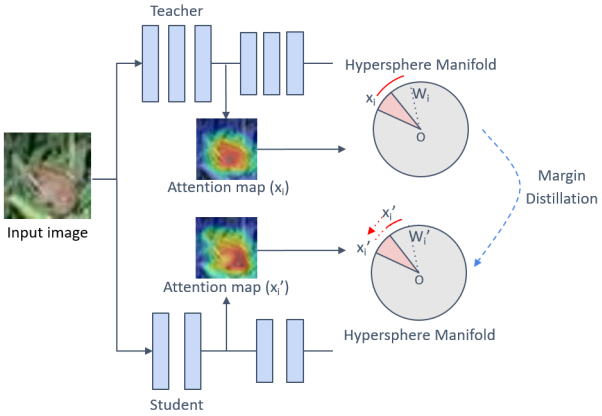


Figure 1. An overview of deep hypersphere feature embedding-based knowledge distillation. The distillation method encourages student to produce similar activations of the teacher and get more discriminative distance on a hypersphere manifold.

Many studies have mainly focused on the structure of teacher and student network [12] and repetitive distillation using various teacher-student models [5, 22], but not much on the knowledge itself and approach to complement traditional KD.

Recently, feature-based distillation methods [6] have been studied for better-mimicking teacher and separating the classes. Romero *et al.* [15] firstly introduced to use intermediate representations in FitNets. The distillation loss enables the student to mimic the teacher's matrices from

feature maps in intermediate layers. In attention transfer [25], spatial attention maps are used for distillation, computed by summation of the squared activations. The distillation loss also encourages the student to generate similar normalized maps as the teacher. However, these studies have focused on mimicking features of the teacher only. How to boost the student's classification ability for the student while knowledge is transferred from the teacher is still required to be explored.

In this paper, we propose a robust loss function for knowledge distillation that is motivated by transferring attentive features and training discriminative angular distance metric on a hypersphere manifold. Deep hypersphere feature embedding-based knowledge distillation encourages the student network to produce similar activations of teacher network and get better classification ability. Figure 1 shows the overall procedure. To match the dimension size between teacher and student and get more activated features, spatial attention maps are computed by using output activation maps. An angular distance metric, geodesic distance on a hypersphere manifold, is constructed by projecting the attention maps for the teacher and student, respectively. By angular margin based distillation, the student tries to mimic more separated decision regions of the teacher, simultaneously training to enlarge the inter-class and compress the intra-class angular distribution. Therefore, the proposed method improves the performance of the student more effectively.

The contributions of this paper are:

- We introduce deep hypersphere feature embedding-based knowledge distillation, which uses attentive features on angular margin based distillation.
- We implement the proposed distillation method with traditional KD to complement the previous method.
- We perform KD with different sizes of teacher and student networks. We corroborate results from previous studies which suggest that the performance of a higher capacity teacher model is not necessarily better.
- We experimentally show that our approach provides significant improvements with various deep network architectures. Also, our study is evaluated across two public datasets.

The rest of the paper is organized as follows. In Section 2, we describe related work. In Section 3, we provide a overview of the proposed method. In Section 4, we describe our experimental results and analysis. In Section 5, we discuss our findings and conclusions.

2. Related Work

2.1. Knowledge Distillation

Knowledge distillation, a transfer learning method, trains a smaller model called a student model by shifting knowledge from a larger model called a teacher model. KD is firstly introduced by Buciluǎ *et al.* [1] and more explored by Hinton *et al.* [9]. The main concept of KD is using the soft labels by a trained teacher network. That is, mimicking soft probabilities helps students get knowledge of teachers, which improves beyond using hard labels (training labels) alone. Cho *et al.* [2] explore which combination of student-teacher is good to obtain the better performance. They find that using a teacher trained by early stopping the training and stopping KD close to convergence improve the efficacy of KD. KD can be categorized into two approaches which use the outputs of the teacher. One is response-based KD, traditional KD, which uses the posterior probabilities with softmax loss [6]. The other is feature-based KD using the intermediate features with normalization [6]. Feature-based methods perform with the response-based method to complement traditional KD [6].

2.2. Attention Transfer

To allow a student to mimic the aspect of a teacher network, how to capture the knowledge of the teacher and distill the knowledge has been explored. Zagoruyko *et al.* [25] suggest activation-based attention transfer (AT), which uses a sum of squared attention mapping function computing statistics across the channel dimension. Although the depth of teacher and student is different, knowledge can be transferred by the attention mapping function, which matches the depth size as one. The activation-based spatial attention maps are created by the sum of absolute values raised to the power of p: $\mathrm{F}^{p}_{sum}(\mathrm{A}) = \sum_{j=1}^{c} |A_j|^p$, where F is a created attention map, A is an output of a layer, c is the number of channels for the output, j is the number for the channel, and $p > 1$. Putting more weight (using power value p of the equation) corresponds to put weight to the most discriminative parts defined by activation level. AT can perform as a feature-based distillation method and be combined with traditional KD.

2.3. SphereFace

Most of the existing methods were dependent upon Euclidean distance for feature distinction. These approaches could not solve the problem that classification under open-set protocol only shows a meaningful result when successfully narrowing maximal intra-class distances. To solve this problem, an angular-softmax (A-softmax) function is proposed to distinguish the features by increasing angular margins between features [13]. According to its geometric interpretation, using A-softmax function equivalents to the projection of features onto the hypersphere manifold, which intrinsically matches preliminary condition that features also lie on a manifold. Applying angular margin penalty corresponds to the geodesic distance margin penalty in the hypersphere [13]. For this reason, the A-

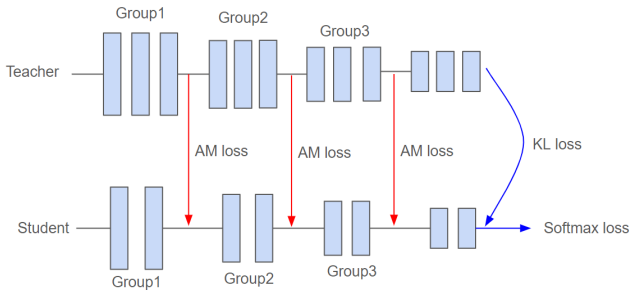


Figure 2. Schematics of teacher-student knowledge transfer with the proposed method

softmax function shows superior performance than the previous works when tested on several classification problems.

3. Proposed Method

In this work, we adopt the four components as the deep hypersphere feature embedding-based knowledge distillation. Our overall learning scheme is illustrated in Figure 2. We learn an efficient and lightweight deep hypersphere feature embedding knowledge distillation model for all the four components.

Our method is composed of four modules:

1. Training teacher and student models based on early stopped KD
2. Generating normalized attention maps to transfer knowledge
3. Constructing discriminative angular distance metric
4. Transferring attentive features by our proposed distillation method

In order to achieve highly accurate and better results of student models, it is critical to train strong student models for all the four components.

3.1. Traditional knowledge distillation

Based on the traditional knowledge distillation [9], the loss function for training student is:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathcal{C}} + \lambda\mathcal{L}_{\mathcal{K}} \tag{1}$$

where $\mathcal{L}_{\mathcal{C}}$ denotes the standard cross entropy loss, $\mathcal{L}_{\mathcal{K}}$ is KD loss, and λ is hyper-parameter; $0 < \lambda < 1$. The error between the output of the softmax layer of a student network and ground-truth label is penalized by the cross-entropy loss:

$$\mathcal{L}_{\mathcal{C}} = \mathcal{H}(softmax(a_S), y_G) \tag{2}$$

where $\mathcal{H}(\cdot)$ is a cross entropy loss function, a_S is logits of a student (inputs to the final softmax), and y_G is a ground truth label. The output of student and teacher are matched by KL-divergence loss:

$$\mathcal{L}_{\mathcal{K}} = \tau^2 KL(z_T, z_S) \tag{3}$$

where $z_T = softmax(a_T/\tau)$ is a softened output of a teacher network, $z_S = softmax(a_S/\tau)$ is softened output of a student, and τ is a hyperparameter; $\tau > 1$.

We adopt early-stopped KD (ESKD) [2] for training teacher and student models, referring to its effects across the board at improving the efficacy of knowledge distillation. The early stopped model of a teacher tends to train student models better than the standard KD strategy, which uses a fully trained teacher. The traditional KD is to be combined with distillation using features obtained intermediate layers.

3.2. Generating attention maps

To transfer more activated features from teacher to student, output of intermediate layers can also be used. The distillation methods using outputs of intermediate layers can be combined with the traditional KD. It is beneficial to guide the student network towards to induce more similar patterns of teacher and get the better ability for classification.

In order to match the dimension size between teacher and student models, we create the normalized attention maps [25], which has benefits in generating maps discriminatively between positive and negative features. Also, another additional training procedure is not required for only matching the channel dimension size between teacher and student. We use the power value p of 2 for generating the attention maps, which shows the best results for the previous method [25]. Let an output map is $A \in \boldsymbol{R}^{c\times h\times w}$, where, c is the number of output channels, h is height for the size of output and w is width for the size of output. The attention map for teacher is $\mathrm{F}_T^l = \sum_{j=1}^{C} |A_{T,j}^l|_T^2$, where l is a specific layer, c is the number of channels, j is the number for the output channel, and T denotes a teacher network. The attention map for student is $\mathrm{F}_S^{l'} = \sum_{j'=1}^{C'} |A_{S,j'}^{l'}|_S^2$, where l' is the corresponding layer of l, c' is the number of channels for the output, j' is the number for the output channel, and S denotes a student network. If the student and teacher use the same depth for transfer, l' can be the layer at the same depth as l; if not, l' can be the end of the same block for the teacher.

3.3. Discriminative angular distance metric

In our proposed loss function, the key is to stand out the difference between positive and negative information (enlarge the gap between inter-class to classification) from using the attention map through the feature of the attention angle probability function.

Based on the attention maps, probability related to positive and negative features can be written in:

$$\tilde{G}_{ang1} = log(\frac{e^{Q_p(x_i W_i)}}{e^{Q_p(x_i W_i)} + e^{Q_n(x_i W_i)}}) \tag{4}$$

where x is input vectors of a layer, W is weights for the layer, xW represents the output feature vector from the layer, Q_p is a function for obtaining the attention map as a positive map for a class i, and Q_n represents to get adverse features as a negative map.

To apply general equation for constructing angular distance metric [19], normalized attention maps for teacher and student maps are computed, respectively. Let output of the function Q as Q' corresponding to an attetion map. For simplicity, we transform Q' to $\|Q'\|cos(\theta)$ and rescale it to s, where θ is the angle between the feature and weight for Q'. The normalized feature and weight make the predictions depend on the angle between the feature and the weight only. With the normalized $\|Q'_p\|$, the negative feature $\|Q'_n\|$ is simply obtained; $\|Q'_n\|$ = 1 - $\|Q'_p\|$. The embedding features can be mapped on the hypersphere:

$$\tilde{G}_{ang2} = log(\frac{e^{s\cdot(cos(\theta_{p_i}))}}{e^{s\cdot(cos(\theta_{p_i}))} + e^{s\cdot(cos(\theta_{n_i}))}}) \tag{5}$$

where s is the scaling factor to determine the sphere size, θ_{p_i} is the angle between feature and weight for positive vectors for an attention map, θ_{n_i} is the angle between feature and weight for negative vectors for the attention map.

To enlarge the gap between positive and negative vectors, we apply angular margin penalty m to the angle. Using margin to the angle enhances the inter-class discrepancy, and intra-class compactness [13]. The feature encoding attentive angle from projecting the attention map on a hypersphere manifold can be written as follow:

$$\tilde{G}_{ATang} = log(\frac{e^{s\cdot(cos(m\theta_{p_i}))}}{e^{s\cdot(cos(m\theta_{p_i}))} + e^{s\cdot(cos(\theta_{n_i}))}}) \tag{6}$$

where m is a constant for the margin. $\tilde{G}_{ATang}$ contains the information of inter-class distance between each features, which is explained by positive or negative. For transferring knowledge, teacher and student generate the features ($\tilde{G}_T$ and $\tilde{G}_S$) from intermediate layers, respectively.

3.4. Angular margin based distillation

To transfer attentive feature from teacher to student, we derive a loss function that mitigates the gap between the teacher and student model. The loss function for angular margin based distillation is:

$$\mathcal{L}_{\mathcal{A}} = \alpha \sum_{(l,l')\in I} \Big\| \frac{\tilde{G}^l_T}{\|\tilde{G}^l_T\|_2} - \frac{\tilde{G}^{l\prime}_S}{\|\tilde{G}^{l'}_S\|_2} \Big\|^2_F \tag{7}$$

where I collects the layer pairs (l and l'), $\|\cdot\|_F$ is the Frobenius norm [18], and α is a constant determined by the number of layer pairs for distillation; $\alpha = \frac{1}{\#pairs}$.

This criterion encourages the student to train for equalizing the distance between features from teacher network and student network, simultaneously mimic more separated decision regions of the teacher. It corresponds to the function guides student to get the of the teacher's discriminative features and enlarge the inter-class and compress the intra-class.

To complement the traditional KD, the loss function is combined with the previous ones. Our overall learning objective can be written as a loss function:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathcal{C}} + \lambda\mathcal{L}_{\mathcal{K}} + \gamma\mathcal{L}_{\mathcal{A}} \tag{8}$$

where $\mathcal{L}_{\mathcal{C}}$ is a entropy loss, $\mathcal{L}_{\mathcal{K}}$ is a knowledge distillation loss, $\mathcal{L}_{\mathcal{A}}$ denotes the angular margin based loss function, and λ and γ are hyper-parameter to control the balance between different losses.

In this section, we show the experimental validation of our approach. We evaluate our methods with different combinations of teacher and student which have different architectures. We take experiments with two public datasets which are CIFAR-10 and CINIC-10. We report results with several different hyperparameter (γ and m) for the proposed distillation function to show the sensitivity and which one is the best.

4. Experiments

4.1. Dataset Description

4.1.1 CIFAR-10

CIFAR-10 is a commonly adopted dataset for comparing distillation methods [25, 2, 18]. CIFAR-10 dataset consists of 60,000 colored images in 10 classes, which is 6,000 images per class. Since the images are set to be low-resolution (32×32 pixels), it is able to apply the dataset to multiple trials of training the networks in a relatively short time. Due to its convenience and time efficiency, it is utilized as one of the benchmark datasets in the area of image learning. The application of CIFAR-10 helps validate the efficacy of our models with less time consumption[1]. For the following experiments on CIFAR-10, 200 epochs using SGD with momentum 0.9 and setting the initial learning rate lr = 0.1 are applied. The lr is decayed by a factor of 0.2 at epochs 40, 80, 120, and 160. The CIFAR-10 dataset includes 10 classes with 5000 training images per class and 1000 testing images per class. Each image is an RGB image of size 32×32. We use the 50,000 images of the training set and 10,000 as the testing set. For the hyperparameters regarding our proposed method, we stayed consistent with the popular choice (τ = 4 and λ = 0.9) empirically figured out throughout the experiments [2].

[1] considering the given time for the project (EEE515 Spring 2021) is only about three weeks

Group Name	Output Size	WRN16-k	WRN28-k
conv1	32×32	3×3, 16	3×3, 16
conv2	32×32	$\begin{bmatrix} 3\times3, 16k \\ 3\times3, 16k \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 16k \\ 3x3, 16k \end{bmatrix} \times 4$
conv3	16×16	$\begin{bmatrix} 3\times3, 32k \\ 3\times3, 32k \end{bmatrix} \times 2$	$\begin{bmatrix} 3x3, 32k \\ 3\times3, 32k \end{bmatrix} \times 4$
conv4	8×8	$\begin{bmatrix} 3\times3, 64k \\ 3\times3, 64k \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64k \\ 3\times3, 64k \end{bmatrix} \times 4$
	1×1	average pool, 10-d fc, softmax	

Table 1. Structure of WRN networks used in CIFAR-10 experiments. Downsampling is performed in the first layers of conv3 and conv4. 16 and 28 means depth and k is width of the network.

4.1.2 CINIC-10

We extend our experiments on CINIC-10 [3] which is designed to be an option relative to CIFAR-10 and ImageNet. CINIC-10 consists of augmented extension in the style of CIFAR-10, but the dataset contains 270,000 images whose scale is closer to that of ImageNet. The images are equally split into the each set named 'train', 'test', and 'validate' for each. The size of the images is 32×32. There are ten classes in each 90,000 subsets of images and 9,000 images per class. For experiments with CINIC-10 datasets, we followed the similar parameter setting closely with CIFAR-10 to the best of our knowledge. We used SGD with momentum 0.9, initial learning rate lr is 0.1, and weight decay 1×10^{-4}. We set τ = 16 and λ = 0.6 [18] for KD, and 180 epochs. The lr drops down by 0.2 at epochs 40, 80, 120, and 160. We adopted CINIC-10 to consider rapid experimentation and computational resources for training and validation for our method and all baselines.

4.2. Various depth and width architectures

Settings for experiments: We performed the proposed method using WideResNet (WRN) [24] models for teacher and student models to evaluate the classification accuracy, which is popularly used for KD [2, 25, 23, 18]. WRN16-3 was used as a medium teacher model in the previous study [2]. To validate with different capacity of models, WRN16-3 and WRN28-3 are chosen as teachers. Their network architectures are described in Table 1. After training WRN16-3 and WRN28-3 with scratch learning, the models are frozen while student models are trained. We perform baseline comparisons with traditional KD [9] and attention transfer [25]. We set the weight for the distillation loss function for attention transfer (β = 1000 for CIFAR-10 and β = 50 for CINIC-10). The constant parameter α and margin parameter m for the proposed method are $\frac{1}{3}$ and 1.35, respectively. The loss weight γ of the proposed method for CIFAR-10 is 5000 when the student is WRN16-3 and 1000 when a student is WRN28-3. The distillation loss weight γ of the proposed method for CINIC-10 is 500 for CINIC-10. We determined the parameters by empirical experiments considering the distillation effects by the capacity of models. The settings are applied to the following experiments as well. All following experiments were repeated five times.

4.2.1 Evaluation with various depth and width

We implemented with the same depth or width of WRN architectures for teacher and student. The results with classification accuracy $(\%)$ for the student models are shown in Table 2 for CIFAR-10 and Table 3 for CINIC-10.

In both of the Tables, our method, ATang+KD, shows the highest accuracy among the target models followed by AT+KD, KD, and scratch learned model in order. When comparing KD and AT+KD, in most of the cases, AT+KD shows better performance. That is, we proved again that the attention map helps the teacher to transfer its knowledge better than its traditional one. For CIFAR-10, the result of the proposed method with WRN16-3 student and WRN28-3 teacher is even better than the teacher. Therefore, referring to the fact that our method outperforms AT+KD in all cases, we conclude that applying discriminative angular distance metric for KD maximizes the attention map's efficacy of transferring the knowledge and performs to complement the traditional KD.

Also, when applying the larger teacher model (WRN28-3) and the smaller student model (WRN16-1), implying the capacity difference between teacher and student is larger than the other combinations, the performances were the worst for every student model. This consequence repeatedly verifies the previous study [2] that the larger teacher does not always guarantee to produce the better student.

4.2.2 Analysis with activation maps

In order to analyze results with intermediate layers, we adopt Grad-CAM [17] which uses the class-specific gradient information to visualize the coarse localization map of the important regions in the image. The activation maps from intermediate layers across various methods are shown in Figure 3. Our method, ATang+KD, shows intuitively similar activated regions with the traditional KD [9] in the low-level. However, in mid-level and high-level, our method has higher activations around the region of a target object, which is different from the previous methods [9, 25]. Thus, the proposed method can classify positive and negative areas discriminatively, compared to the previous methods [9, 25]. The high-level activation maps with various input images are described in Figure 4. The activated area of the proposed method is presented in the center of a target in an image. The result shows that our method

Student	Teacher	Student	KD [9]	AT+KD [25]	ATang+KD (ours)	Teacher
WRN16-1 (0.2M)	WRN16-3 (1.5M)	84.11±0.21	85.29±0.15	85.33±0.10	**86.27**±0.09	87.76±0.12
WRN16-1 (0.2M)	WRN28-3 (3.3M)	84.11±0.21	85.15±0.59	85.26±0.03	**85.70**±0.09	88.65±0.10
WRN16-3 (1.5M)	WRN28-3 (3.3M)	87.76±0.12	89.07±0.10	89.16±0.07	**89.30**±0.10	88.65±0.10

Table 2. Experiments on CIFAR-10 with various combinations of teacher and student which have the same depth or width of WRN. Knowledge distillation losses are softened class scores (traditional KD), attention transfer (AT), and angular margin based attention distillation (ATang). Brackets indicate the number of trainable parameters for the model (model size).

Student	Teacher	Student	KD [9]	AT+KD [25]	ATang+KD (ours)	Teacher
WRN16-1 (0.2M)	WRN16-3 (1.5M)	79.49±0.08	80.87±0.04	80.83±0.06	**80.98**±0.02	84.15±0.04
WRN16-1 (0.2M)	WRN28-3 (3.3M)	79.49±0.08	80.15±0.03	80.25±0.06	**80.28**±0.04	85.58±0.03
WRN16-3 (1.5M)	WRN28-3 (3.3M)	84.15±0.04	85.36±0.14	85.50±0.03	**85.53**±0.02	85.58±0.03

Table 3. Experiments on CINIC-10 with various combinations of teacher and student which have the same depth or width of WRN. Knowledge distillation losses are softened class scores (traditional KD), attention transfer (AT), and angular margin based attention distillation (ATang). Brackets indicate the number of trainable parameters for the model (model size).

performs better in focusing on the foreground object distinctly with high weights. On the other hand, less focusing on the background compared to the other methods [9, 25] which are focusing on both or the background. By more weighting on the regions, the student based on the proposed method has stronger discrimination ability. Therefore, our method, angular margin based attention distillation (ATang+KD), guides student models to not only enlarge the inter-class but also compress the intra-class.

4.3. Different student and teacher architectures

There are wide varieties of model combinations applicable to KD. For the additional experiments, We implemented with students and teachers from different architecture networks including different depth and width of WRN. We used ResNet [8] and MobileNetV2 [16] which are well-known benchemarks for image classification. We applied the same settings with the experiments of the previous section.

The experimental results with multiple combinations of the models are described in Table 4 and 5. Our method, ATang+KD, shows the best accuracy when compared to AT+KD, KD, and scratch learnt model. In most of the cases, AT+KD shows the better performance than the traditional KD. It is verified that applying discriminative angular distance metric with attention improves the efficacy of transferring the knowledge from teacher to student and complements the traditional KD.

4.4. Analysis of hyperparameters for distillation

In this Section, we investigate sensitivity for hyperparameters (γ and m) used for the angular margin based attention distillation. And we present errors with $\mathcal{L}_{\mathcal{A}}$ to explain the relationship between accuracy and distillation.

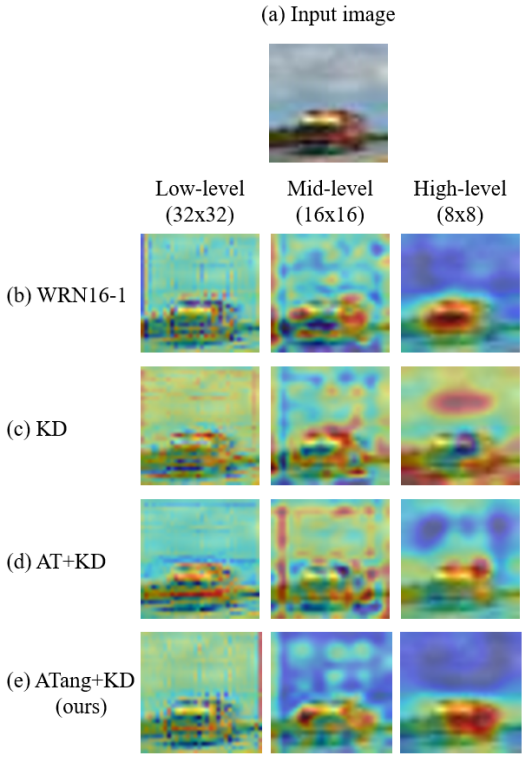


Figure 3. Activation maps over different levels of a student (WRN16-1) trained with a teacher (WRN16-3) for CIFAR-10

4.4.1 Analysis of angular distillation hyperparameter γ

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) by using various γ on CIFAR-10 are illustrated in Figure 5 (m = 1.35). When γ is 5000, two results show the best accuracy. For WRN16-3 as a teacher, the accuracy of γ = 7000 is higher than the one of γ = 3000. However, for WRN28-3 as a teacher, the accu-

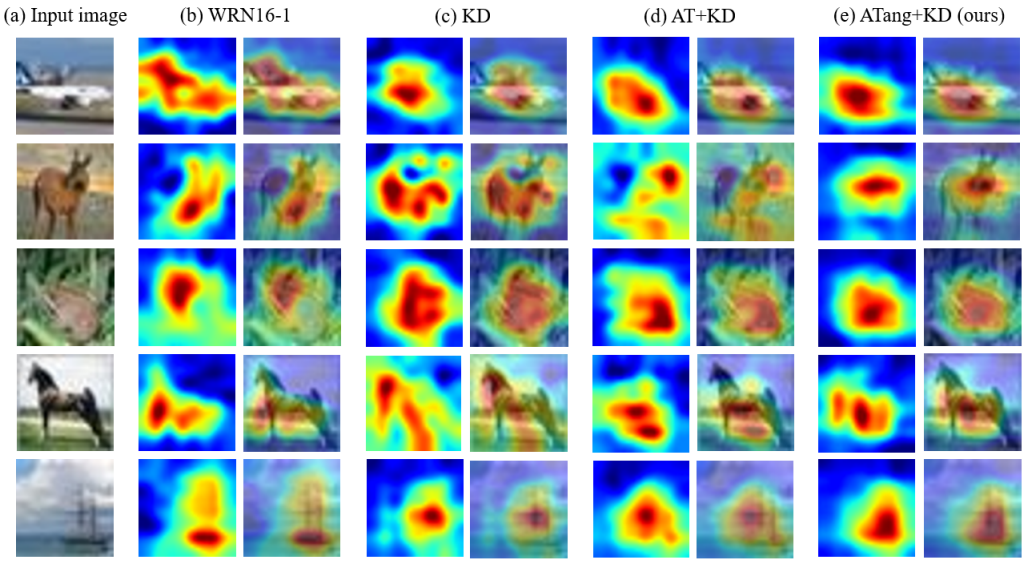


Figure 4. Activation maps of high-level from a student network (WRN16-1) trained with a teacher (WRN16-3) over different input images for CIFAR-10.

Student	Teacher	Student	KD [9]	AT+KD [25]	ATang+KD (ours)	Teacher
WRN16-1 (0.2M)	ResNet44 (0.7M)	84.11±0.21	85.44±0.06	85.15±0.16	**85.97**±0.17	86.41±0.20
ResNet20 (0.3M)	MobileNetV2 (0.6M)	85.20±0.17	87.28±0.15	87.44±0.10	**87.88**±0.07	89.61±0.11

Table 4. Experiments on CIFAR-10 with different combinations of teacher and student which have different architectures. Knowledge distillation losses are softened class scores (traditional KD), attention transfer (AT), and angular margin based attention distillation (ATang). Brackets indicate the number of trainable parameters for the model (model size).

Student	Teacher	Student	KD [9]	AT+KD [25]	ATang+KD (ours)	Teacher
ResNet20 (0.3M)	WRN28-3 (3.3M)	80.72±0.05	81.31±0.20	81.70±0.04	**81.73**±0.03	85.58±0.03
ResNet20 (0.3M)	WRN16-3 (1.5M)	80.72±0.05	82.22±0.03	82.23±0.03	**82.25**±0.01	84.15±0.04
WRN28-1 (0.4M)	WRN16-8 (11M)	81.32±0.07	81.56±0.03	82.40±0.04	**82.58**±0.08	86.30±0.04
WRN16-8 (11M)	WRN28-6 (13M)	86.30±0.04	86.94±0.03	87.26±0.02	**87.42**±0.03	86.67±0.03

Table 5. Experiments on CINIC-10 with different combinations of teacher and student which have different architectures. Knowledge distillation losses are softened class scores (traditional KD), attention transfer (AT), and angular margin based attention distillation (ATang). Brackets indicate the number of trainable parameters for the model (model size).

racy of γ = 3000 is higher than the one of γ = 7000. Therefore, when a teacher is the high capacity network, compared to the student, using γ = 3000 may produce better results. When a student has high capacity network as well, using the smaller value of γ can produce the better results, compared to the larger value. When a student is WRN16-3 and teacher is WRN28-3, the accuracy of γ = 1000 was 89.3% and one of γ = 5000 was 88.83%.

4.4.2 Analysis of angular margin m

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) by various angular margin m on CIFAR-10 are illustrated in Figure 6 (γ = 5000). m = 1.0 means there is no additional margin dis-

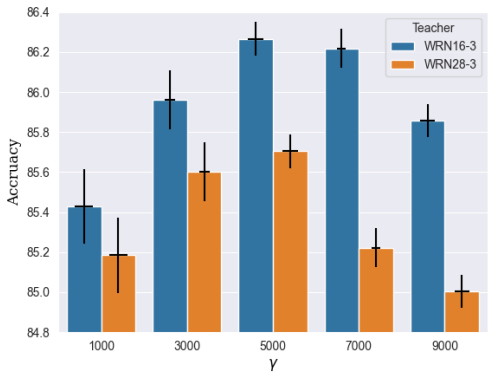


Figure 5. Accuracy (%) of a student (WRN16-1) across various γ, trained with teachers (WRN16-3 and WRN28-3) for CIFAR-10.

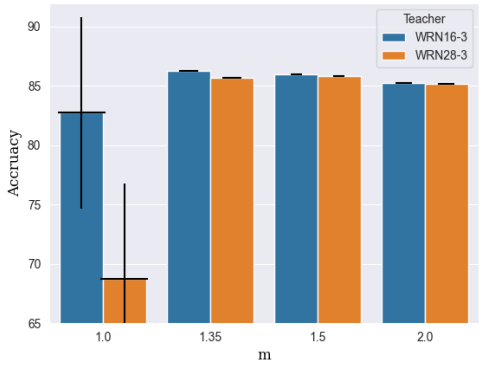


Figure 6. Accuracy (%) of a student (WRN16-1) across various angular margin m, trained with teachers (WRN16-3 and WRN28-3) for CIFAR-10.

tance between positive and negative features on the hypershpere for distillation. Using the large value of m corresponds to make a large gap between positive and negative features for distillation. When m is 1.35 for the teacher WRN16-3, the student WRN16-1 shows the best performance 86.27%. When the teacher is WRN28-3, the student produces the best accuracy with m = 1.5. The both results are better than the results with m is 1.0 or 2.0. Therefore, margin parameter m affects to the distillation and applying the larger margin does not guarantee to generate the better result. When teacher is larger capacity model, using margin m with larger than 1.35 can produce the good performance. Therefore, determining proper m with considering the capacity of teacher and student is important for distillation. We recommend to use near 1.35 for margin m to obtain the best results.

4.4.3 Analysis of $\mathcal{L}_{\mathcal{A}}$ with error

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) with γ = 3000, 5000, and 7000 on CIFAR-10 are shown in Figure 7. The results of teacher WRN16-3 show that the higher accuracy tends to be presented when distillation loss is the smaller. For teacher WRN28-3, the high accuracy tends to be presented when distillation loss is around 0.087×10^{-3}, but, some of the lower accuracy can be shown as well.

The results from angular margin based distillation corroborate the previous study [2] that the larger teacher does not always guarantee to produce the better student. Also, it is verified that $\mathcal{L}_{\mathcal{A}}$ is correlated with the accuracy.

5. Discussion and future work

In this article, we presented "Deep Hypersphere Feature Embedding based Knowledge Distillation", a newly suggested angular margin based attention distillation. According to the experiment, the results have shown our method has a novelty in maximizing the student model's ability for classification to enlarge the inter-class and compress the intra-class. We have verified the consistent efficiency of our methods by implementing multiple combinations of the models. We have shown nearly all of the improvements that are achieved by training with an angular margin based attention distillation.

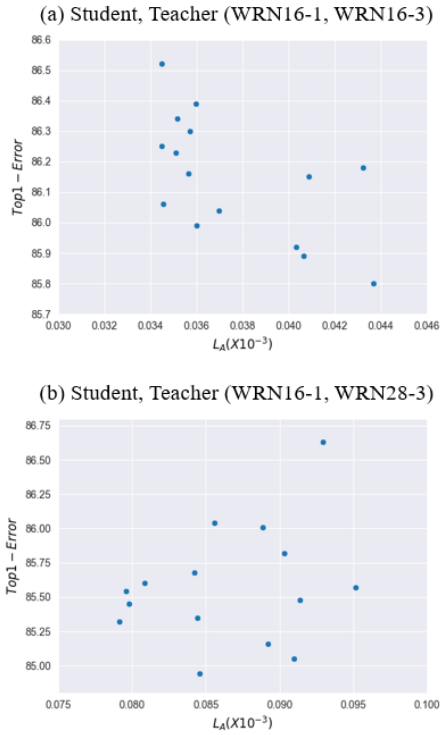


Figure 7. Accuracy (%) of a student (WRN16-1) across various m, trained with teachers (WRN16-3 and WRN28-3) for CIFAR-10.

In future work, we would like to extend the idea presented in this paper in many different applications. Even though we have shown great performances on the more challenging dataset, the accuracy is not much improved compared to the easier one. For the further exploring the distillation effects with the angular margin based method, we would like to apply the idea with various hypersphere feature embedding based methods [20, 4]. Also, our method can be applied to other classification problems with higher resolution images and more diverse classes (e.g., CelebA [14] and CIFAR-100 [11]). Furthermore, our novel method can be extended to object detection since we demonstrate consistent improvements over various experimental setups, which could provide good insights to further advancement in other fields. Larger object detector models can be used as a teacher to guide the learning process of the student models, which can obtain benefits from our method.

References

[1] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, 2006.

[2] J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4802, 2019.

[3] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[5] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[6] J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

[7] 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.

[8] 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*, pages 770–778, 2016.

[9] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *Proceedings of the International Conference on Neural Information Processing Systems Deep Learning and Representation Learning Workshop*, 2015.

[10] I. Jang, S. Kim, H. Kim, C.-W. Park, and J. H. Park. An experimental study on reinforcement learning on iot devices with distilled knowledge. In *International Conference on Information and Communication Technology Convergence*, pages 869–871, 2020.

[11] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] X. Lan, X. Zhu, and S. Gong. Knowledge distillation by on-the-fly native ensemble. *arXiv preprint arXiv:1806.04606*, 2018.

[13] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[15] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning and Representation*, 2015.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[18] F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[19] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[20] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[21] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016.

[22] C. Yang, L. Xie, S. Qiao, and A. Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv preprint arXiv:1805.05551*, 2018.

[23] 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*, pages 4133–4141, 2017.

[24] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

[25] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning and Representation*, 2017.

Supplementary Material

6. Additional experimental results and details

6.1. Empirical experiments for λ of KD

In order to find an optimal parameter λ and to corroborate the previous study [2], we tested with different λ for training based on KD on CIFAR-10 dataset. As shown in Figure S1, when λ is 0.9 ($\tau = 0.4$) with KD, the accuracy of a student (WRN16-1) trained with WRN16-3 as a teacher is the best. If λ is large, the distillation effect of KD is increased. Since the accuracy depends on λ, we referred to previous studies [2, 18] to choose the popular parameters for experiments.

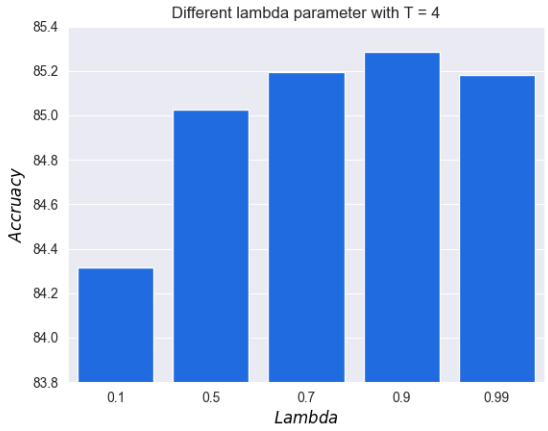


Figure S1. Accuracy (%) of a student (WRN16-1) trained with teachers (WRN16-3) for various λ on CIFAR-10.

6.2. Analysis of γ on CINIC-10

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) by using various γ on CINIC-10 are illustrated in Figure S2 ($m = 1.35$). When γ is 500, two results show the best accuracy. When the teacher is WRN16-3, the accuracy of $\gamma = 1000$ is much higher than the one of $\gamma = 100$. However, when the teacher is WRN28-3, the accuracy of $\gamma = 100$ and $\gamma = 1000$ are not that much different; difference is only 0.02%. Therefore, as described in Section 4, when the capacity between teacher and student is similar, using larger γ can produce the better performance. If the difference between their capacities is high, using smaller one can show the better performance. We recommend to use $\gamma = 500$ which shows good performances in general cases on complicated dataset such as CINIC-10.

6.3. Analysis of margin m on CINIC-10

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) by using various m on CINIC-10 are shown in Figure S3 ($\gamma = 500$). When m is

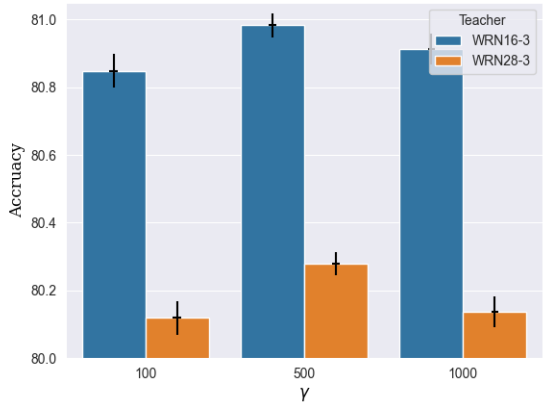


Figure S2. Accuracy (%) of a student (WRN16-1) across various γ, trained with teachers (WRN16-3 and WRN28-3) for CINIC-10.

1.35, the student trained with teacher of WRN16-3 shows the best accuracy, 80.98%. The accuracy of $m = 1.5$ and 2.0 are 80.92% and 80.90%, respectively. However, when the teacher is WRN28-3, the best accuracy, 80.35%, is produced with $m = 2.0$. The accuracy of $m = 1.35$ and 1.5 are 80.30% and 80.32%, respectively. These results show that adding margin m affects the distillation procedures. As described in Section 4, when teacher has much larger capacity, compared to student, using m with larger than 1.35 can generate the better performance. Therefore, using an optimal m is important to get the better results. We recommend to use approximate $m = 1.35$ in general cases.

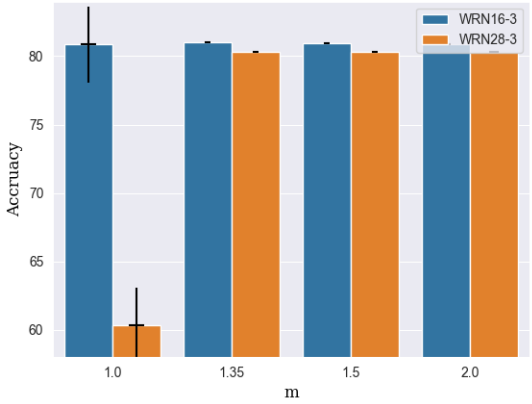
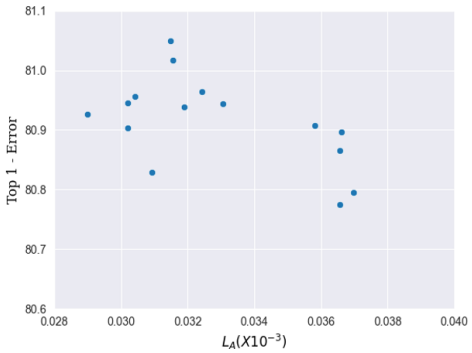


Figure S3. Accuracy (%) of a student (WRN16-1) across various m, trained with teachers (WRN16-3 and WRN28-3) for CINIC-10.

6.4. Analysis of $\mathcal{L}_{\mathcal{A}}$ with error on CINIC-10

The results of a student model (WRN16-1) trained with teachers (WRN16-3 and WRN28-3) on CINIC-10 are shown in Figure S4 with γ = 100, 500, and 1000 (m = 1.35). For both of the results, the higher accuracy tends to be produced when distillation loss is the smaller. The results verify that angular margin based attentive feature distillation loss $\mathcal{L}_{\mathcal{A}}$ is correlated with the accuracy.

(a) Student, Teacher (WRN16-1, WRN16-3)



(b) Student, Teacher (WRN16-1, WRN28-3)

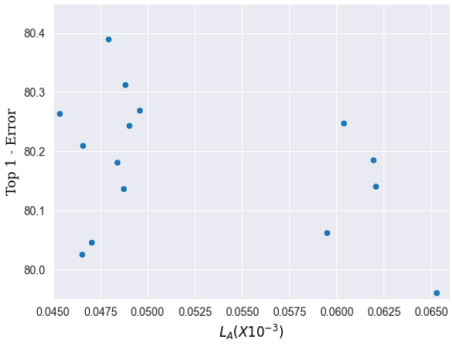


Figure S4. Accuracy (%) of a student (WRN16-1) across various m, trained with teachers (WRN16-3 and WRN28-3) for CINIC-10.