AGKD-BML: Defense Against Adversarial Attack by Attention Guided Knowledge Distillation and Bi-directional Metric Learning

Hong Wang¹, Yuefan Deng¹, Shinjae Yoo², Haibin Ling¹, Yuewei Lin²†
¹Stony Brook University, Stony Brook, NY, USA
²Brookhaven National Laboratory, Upton, NY, USA
{hong.wang.2,yuefan.deng,haibin.ling}@stonybrook.edu, {sjyoo,ywlin}@bnl.gov

Abstract

While deep neural networks have shown impressive performance in many tasks, they are fragile to carefully designed adversarial attacks. We propose a novel adversarial training-based model by Attention Guided Knowledge Distillation and Bi-directional Metric Learning (AGKD-BML). The attention knowledge is obtained from a weight-fixed model trained on a clean dataset, referred to as a teacher model, and transferred to a model that is under training on adversarial examples (AEs), referred to as a student model. In this way, the student model is able to focus on the correct region, as well as correcting the intermediate features corrupted by AEs to eventually improve the model accuracy. Moreover, to efficiently regularize the representation in feature space, we propose a bidirectional metric learning. Specifically, given a clean image, it is first attacked to its most confusing class to get the forward AE. A clean image in the most confusing class is then randomly picked and attacked back to the original class to get the backward AE. A triplet loss is then used to shorten the representation distance between original image and its AE, while enlarge that between the forward and backward AEs. We conduct extensive adversarial robustness experiments on two widely used datasets with different attacks. Our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches. The code of AGKD-BML will be available at: https://github.com/hongw579/AGKD-BML.

1 Introduction

Deep neural networks (DNNs) have achieved great breakthrough on a variety of fields, such as computer vision [22], speech recognition [17], and natural language processing [8]. However, their vulnerability against the so-called adversarial examples (AEs), which are the data with carefully designed but imperceptible perturbations added, has drawn significant attention [38]. The existing of AEs is a potential threat for the safety and security of DNNs in real-world applications. Thus, many efforts have been made to defend against adversarial attacks as well as improve the adversarial robustness of the machine learning model. In particular, adversarial training [16, 27]-based models are among the most effective and popular defending methods. Adversarial training solves a min-max optimization problem, in which the inner problem is to find the strongest AE within an $\epsilon$-ball by maximizing the loss function, while the outer problem is to minimize the classification loss of the AE. Madry et al. [27] provided a multi-step projected gradient descent (PGD) model, which has become the standard model of the adversarial training. Following PGD, a number of recent works have been proposed to improve adversarial training from different aspects, e.g., [6, 11, 28, 32, 35, 42, 49, 51].

However, the adversarial training-based models still suffer from relatively poor generalization on both clean and adversarial examples. Most of the existing adversarial train-
ing based models focus only on the on-training model that utilizes adversarial examples, which may be corrupted, but have not well explored the information from the model trained on clean images. In this work, we aim to improve the model adversarial robustness by distilling the attention knowledge and utilizing bi-directional metric learning.

The attention mechanism plays a critical role in human visual system and is widely used in a variety of application tasks [34, 53]. Unfortunately, one of our observations shows that the perturbations in the adversarial example (AE) will be augmented through the network, and thus significantly corrupts the intermediate features and attention maps. It is shown in the Figure 1, the AE confuses the model by letting it focus on different regions from the clean image. Intuitively, if we can transfer the knowledge of clean images from the teacher model to the student model to 1) obtain right attention information, and 2) correct the intermediate features corrupted by AE, we should be able to improve the model’s adversarial robustness.

With this motivation, we propose an Attention Guided Knowledge Distillation (AGKD) module, which applies knowledge distillation (KD) [18] to efficiently transfer attention knowledge of the corresponding clean image from the teacher model to the on-training student model. Specifically, the teacher model is pre-trained on the original clean images and will be fixed during training, while the student model is the on-training model. The attention map of a clean image obtained from the teacher model is used to guide the student model to generate the attention map of the corresponding AE against the perturbations.

We further use t-distributed Stochastic Neighbor Embedding (t-SNE) to study the behavior of the AE in the latent feature space (see Figure 3), and observe that the representations of the AE are usually far away from their original class, similar as shown in [28]. While AGKD transfers information of clean image to the student model from the teacher model and thus provides the constraints on the similarity between the AE and its corresponding clean image, there is no constraint of samples from different classes taken into account. Previous works [24, 28, 52] proposed using metric learning to regularize the latent representations of different classes. Specifically, a triplet loss is utilized, in which latent representations of the clean image, its corresponding AE and an image from another class are considered as the positive, anchor, and negative example, respectively. However, this strategy only considers the one-directional adversarial attack, i.e., from the clean image to its adversarial example, making it less efficient.

To address the above issue, we propose a Bi-directional attack Metric Learning (BML) to provide a more efficient and strong constraint. Specifically, the original clean image (positive) is first attacked to its most confusing class, which is the class that has the smallest loss other than the correct label, to get the forward adversarial example (anchor). Then, a clean image is randomly picked from the most confusing class and is attacked to the original image to get the backward adversarial example as the negative.

By integrating AGKD and BML, our AGKD-BML model outperforms the state-of-the-art models on two widely used datasets, CIFAR-10 and SVHN, under different attacks. In summary, our contribution is three-fold:

- An attention guided knowledge distillation module is proposed to transfer attention information of clean image to the student model, such that the intermediate features corrupted by adversarial examples can be corrected.
- A bidirectional metric learning is proposed to efficiently constrain the representations of the different classes in feature space, by explicitly shortening the distance between original image and its forward adversarial example, while enlarging the distance between the forward adversarial example and the backward adversarial example from another class.
- We conduct extensive adversarial robustness experiments on the widely used datasets under different attacks, the proposed AGKD-BML model outperforms the state-of-the-art approaches with both the qualitative (visualization) and quantitative evidence.

2 Related Works

Adversarial Attacks. Generally, there are two types of adversarial attacks: white-box attack where the adversary has full access to the target model, including the model parameters, and the black-box attack, where the adversary has almost no knowledge of the target model. For white-box attack, Szegedy et al. [38] discovered the vulnerability of deep networks against adversarial attacks. They used a box-constrained L-BFGS method to generate effective adversarial attacks. After that, several algorithms were developed to generate adversarial examples. As a one-step attack, the fast gradient sign method (FGSM) proposed in [16] uses the sign of the gradient to generate attacks, with $\ell_{\infty}$-norm bound. In [23], Kurakin et al. extended FGSM by applying it iteratively and designed basic iterative method (BIM). A variant of BIM was proposed in [12] by integrating momentum into it. DeepFool [29] tried to find the minimal perturbations based on the distance to a hyperplane and quantify the robustness of classifiers. In [31], the authors introduced a Jacobian-based Saliency Map Attack. The projected gradient descent (PGD) was proposed in [27] as a multi-step attack method. The CW attack, a margin-based attack, was proposed in [4]. Recently, Croce et al. introduced a parameter-free attack named AutoAttack [10],
which is an ensemble of two proposed parameter-free versions of PGD attacks and the other two complementary attacks, i.e., FAB [9] and Square Attack [1]. It evaluates each sample based on its worst case over these four diverse attacks which includes both white-box and black-box ones. Besides the additive attacks, [14, 15, 20] show that even small geometric transformations, such as affine or projective transformation can fool a classifier. In addition to those attacks on the input image to the model, attempts are made to design adversarial patches that can fool the model in the physical world [13, 19, 23]. On the other side of the coin, adversarial attacks may also be used to improve the model performance [44, 25, 30].

**Adversarial defense.** Adversarial training-based models, which aim to minimize the classification loss to the strongest adversarial examples (maximal loss within a $\epsilon-$ball), are believed as one of the most effective and widely used defense methods. In practice, they iteratively generate adversarial examples for training. In [16], Goodfellow et al. generated the adversarial examples by FGSM, while Madry et al. [27] used the Projected Gradient Descent (PGD) attacks during adversarial training. Many variants based on adversarial training were proposed in recent years. For example, [35] computed the gradient for attacks and the gradient of model parameters at the same time, and significantly reduced the computation time. Adversarial logit paring [21] constrains distance between the logits from a clean image and its adversarial example, while [28] and [52] built a triplet loss between a clean image, its corresponding adversarial example and a negative sample. TRADES [51] computed the gradient for attacks and the gradient of model parameters at the same time, and significantly reduced the computation time.

**3 Proposed Method**

In this section, we present the framework of our proposed AGKD-BML model in detail. As illustrated in Figure 2, AGKD-BML framework consists of two modules, i.e., the attention guided knowledge distillation (AGKD) module and the bidirectional attack metric learning (BML) module. The AGKD module is used for distilling attention knowledge of the clean image to the student model, to obtain a better attention map for adversarial example, as well as correcting the corrupted intermediate features. The BML module efficiently regularizes the representation in feature space by using bidirectional metric learning. In the rest of this section, we first briefly introduce the standard adversarial training (AT) and (non-)targeted adversarial attack, and then describe the two modules of our proposed model and the integration of them.

**3.1 Preliminaries**

We first briefly describe the standard adversarial training (AT) [27]. Suppose we have a labeled $C$-class classification dataset $D = \{(x, y)\}_{i=1}^{N}$ of $N$ samples, where the label $y \in \{1, 2, \ldots, C\}$. There are two types of adversarial attacks, i.e., the non-targeted attacks and the targeted attacks, which can be formulated as eq. (1) and (2), respectively:

$$\max_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y)$$ (1)

$$\min_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y_t)$$ (2)

where $\delta$ is the perturbation added to the image $x$, $\Delta$ provides an $\ell_\infty$-norm bound of the perturbation, and $f_\theta(\cdot)$ and $\mathcal{L}(\cdot)$ to denote the network with model parameters $\theta$ and the loss function, respectively. Non-targeted attacks maximize the loss function given the correct label $y$, while targeted attacks minimize the loss function given the target label $y_t$.

Standard AT uses non-targeted PGD (Projected Gradient Descent) attack [27] during training, which can be formulated as a min-max optimization problem:

$$\min_{\theta} \mathbb{E}_{x \in \mathcal{D}} \left[ \max_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y) \right]$$ (3)

In the objective function, the outer minimization is the update of the model parameters while the inner maximization is for generating adversarial attacks. Specifically, PGD is used to generate attacks, which is an iterative non-targeted attack with random start at the beginning. In this paper, following [41], we use targeted attacks during training with the most confusing class as target class.

**3.2 Attention guided knowledge distillation**

To distill attention information of clean images to the student model, we propose an attention guided knowledge
3.2.1 Class irrelevant attention map

We generate the class irrelevant attention map at the last convolutional layer. Specifically, we treat the backbone neural network until the last convolutional layer as a feature extractor, denoted by $\mathcal{F}(x)$ for a given image $x$, where $\mathcal{F}(x) \in \mathbb{R}^{C \times H \times W}$. We then produce an operator, denoted by $\mathcal{A}(\cdot)$, to map the feature map to the two-dimensional attention map, $\mathcal{A}(\mathcal{F}(x)) \in \mathbb{R}^{1 \times H \times W}$. In this paper, we simply pick the average pooling through the channel dimension (or identical weights $1 \times 1$ convolution) as $\mathcal{A}$.

3.2.2 Knowledge distillation

The knowledge distillation (KD) [18] utilizes a student-teacher (S-T) learning framework to transfer information learned from the teacher model to the student model. In this paper, we treat the model trained on the natural clean images by standard training as the teacher model and the one under adversarial training as the student model. The attention information is what we expect to transfer from the teacher model to the student model. As the teacher model is trained on the clean images with high testing accuracy, it is able to provide correct regions that model should focus on. Therefore, the attention map of the clean image extracted by the teacher model will transfer to the student model. The loss function of this attention guided knowledge distillation is written as:

$$
\mathcal{L}_{KD}(x_t, x_s) = D(\mathcal{A}(\mathcal{F}_t(x_t)), \mathcal{A}(\mathcal{F}_s(x_s)))
$$

where $x_t$ and $x_s$ are input images of the teacher and student model, respectively, and $\mathcal{F}_t$ and $\mathcal{F}_s$ are feature extractors of the teacher and student models, respectively. $D(\cdot)$ is the distance function (e.g., $\ell_1$) to measure the similarity between these two attention maps. Given an adversarial example,
the AGKD guides the student model to focus on the same regions as its clean image.

### 3.3 Bidirectional attack metric learning

In our work, we use the targeted attack to obtain the adversarial examples. Let \( x^s \) refers to a samples with the label \( y = s \), and \( x^t \) refers to an adversarial example of \( x^s \) with the target label \( y_t = t \). In this paper, the forward adversarial example is targeted towards the most confusing class, which is defined as follow:

\[
y_{mc} = \arg \min_{y \neq s} L (f(x^s), y).
\]  

(5)

Given an original clean image \( x^s \), we first generate the targeted adversarial example \( x_{mc}^s \) towards its most confusing class. Then, we randomly select a sample \( x_{mc}^y \) from the most confusing class, and generate its adversarial example \( x_{mc}^y \) that targeted back to the original label \( s \). We utilize \( x^s, x_{mc}^y, x_{mc}^y \) as positive, anchor and negative sample, respectively. The triplet loss is defined as:

\[
L_{tr} (x_a, x_p, x_n) = [d(E(x_a), E(x_p)) - d(E(x_a), E(x_n)) + m]^+,
\]  

(6)

where \( x_a, x_p, x_n \) denote to positive, anchor and negative samples, respectively. \( E(\cdot) \) is the representation from the penultimate layer of the model. \( d(a, b) \) denotes the distance between two embeddings \( a \) and \( b \), which is defined as the angular distance \( d(a, b) = 1 - \frac{|a \cdot b|}{||a||_2 ||b||_2} \), following [28]. \( m \) is the margin. Comparing to the previous metric learning based adversarial training, e.g., [28] and [52] which only consider forward adversarial example, we consider both the forward and backward adversarial examples. Therefore, we name it the bidirectional metric learning.

By adding an \( \ell_2 \)-norm regularization on the embedding, the final BML loss function is written as:

\[
L_{BML} = \lambda_1 L_{tr} (x_{ymc}^s, x^s, x_{mc}^y) + \lambda_2 L_{norm},
\]  

(7)

where \( L_{norm} = ||E(x_{ymc}^s)||_2 + ||E(x^s)||_2 + ||E(x_{mc}^y)||_2 \) is the normalization term, and \( \lambda_1 \) and \( \lambda_2 \) are the trade-off weights for the two losses.

### 3.4 Integration of two modules

We integrate the attention guided knowledge distillation and bidirectional metric learning together to take the benefits from both modules. As we consider bidirectional adversarial attack, we have two clean/adversarial image pairs, \( x^s_{ymc} \) and \( x_{mc}^{ymc} \). For both pairs, we apply the AGKD from the attention map of the clean image obtained by teacher model to the student model, which can be formulated as:

\[
L_{AGKD} = L_{KD}(x^s, x_{ymc}^s) + L_{KD}(x^{ymc}, x_{mc}^{ymc})
\]  

(8)

where the first term denotes the AGKD loss for the forward attack pair, i.e., \( x^s \) and \( x_{ymc}^s \), while the second term denotes the backward attack pair, i.e., \( x_{mc}^{ymc} \) and \( x^{ymc} \).

By combining the standard cross entropy loss used in the traditional adversarial training, the BML loss, and the AGKD loss, the final total loss is:

\[
L_{total} = L_{ce} + L_{AGKD} + L_{BML}
\]  

(9)

The overall procedure of AGKD-BML model is shown in Algorithm 1.

### 4 Experiments

#### 4.1 Experimental settings

**Dataset** We evaluate our method on two popular datasets: CIFAR-10 and SVHN. CIFAR-10 consists of 60k 3-channel color images with size of \( 32 \times 32 \) in 10 classes, in which 50k images for training and 10k images for testing. SVHN is the street view house number dataset, which has 73257 images for training and 26032 images for testing. We evaluate model on a larger datasets: Tiny ImageNet, and the results are shown in supplementary materials.

**Comparison methods** We use comparison methods include: (1) undefended model (UM), where the model is trained by standard training; (2) adversarial training (AT) [27], which uses non-targeted PGD adversarial examples (AEs) for training; (3) single-directional metric learning (SML) [28]; (4) Bilateral [41], which generates AEs on both images and labels; (5) feature scattering
where adversarial attacks for training are generated with feature scattering in the latent space; (6) and (7) utilize the channel-wise activation suppressing (CAS) [3] on TRADES [50] and MART [43], respectively, which showed the superior compared to the original version. Note that Bilateral, FS generate AEs by using single-step attacks in training, while AGKD-BML uses 2-steps attacks, “AT” and “SML” use 7-step attacks, and “TRADES+CAS” and “MART+CAS” use 10-step attacks. To fairly compare to these multi-step attack models, we also train a 7-step attack variant of AGKD-BML, referred as to “AGKD-BML-7”. We test the models with various attacks including FGSM [16], BIM [23], PGD [27], CW [4], MIM [12] with different attack iterations. We also evaluate the models in a per-sample manner using AutoAttack (AA) [10], which is an ensemble of four diverse attacks. Finally, we also test the black-box adversarial robustness of the model.

**Implementation details** Following [27] and [28], we use Wide-ResNet (WRN-28-10) [48], and set the initial learning rate $\gamma$ as 0.1 for CIFAR-10 and 0.01 for SVHN. We use the same learning rate decay points as [41] and [49], where decay schedule $[100, 150]$ for CIFAR-10 and $[60, 90]$ for SVHN, with 200 epochs in total. “AGKD-BML-7” has the learning rate that decays at 150 epochs and the training stops at 155 epochs, following the suggestions in [43, 33]. In training phase, the perturbation budget $\epsilon = 8$ and label smoothing equals to 0.5 following [49]. In the AGKD module, we adopt $\ell_1$ norm to measure the similarity between attention maps. For the BML module, parameters are the same as [28], i.e., margin $m = 0.03$, $\lambda_1 = 2$ and $\lambda_2 = 0.001$.

### 4.2 Evaluation of adversarial robustness

We evaluate our model’s adversarial robustness and report the comparisons in Table 1. The results on “clean” images are used as a baseline for evaluating how much the accuracy of the defenders will drop as increasing the adversarial robustness. It is shown in Table 1, AGKD-BML overall outperforms the comparison methods on CIFAR-10. AGKD-BML also shows better adversarial robustness on SVHN dataset with a large margin.

Interestingly, in Table 1, we observed that AGKD-BML showed different superiorities to different attacks, i.e., AGKD-BML trained on 7-step attack has higher performance than that trained on 2-step attack against AA, but much lower performance against the regular attacks, e.g., PGD and CW. The reason for this phenomenon is, in our opinion, that compared to the regular attacks, AA is an ensemble of four different types of attack, including white-box and black-box ones, which requires the generalization capability of defense against different types of attack. The generation of the 7-step attack significantly increases the diversity of AEs used for training and thus, it improves the robustness against AA with some sacrifice on accuracy against regular attacks. On the other hand, the generation of the 2-step attack focuses more on the regular attacks but less diverse, which makes it has lower performance against AA. As an empirical defense method, we argue the model trained by small-number-step attack is still useful in some scenarios that the adversarial attacks are known. We provide more results of AGKD-BML model trained on large-number-step attacks against AA in supplementary materials.

### 4.3 Ablation study

We analyze the ablation effect of each component of AGDK-BML on CIFAR-10 dataset. The quantitative and qualitative results are shown in Table 2 and Figure 3, respectively. “UM”, “AT” and “SML” are the same models described above. “BML” denotes the bidirectional metric learning without using any knowledge distillation. “AGKD” denotes the model applied attention map guided knowledge distillation without any metric learning. In Figure 3, we provide the t-SNE plots to show the sample representations in feature space. The triangle points with different colors represent the clean images in different classes, while the red circle points are the AEs under PGD-20 attack. We show AEs from two classes (i.e., deer and frog).

“UM” shows how the adversarial attacks behave if a model dose not have any defense. A simple one step attack FGSM drops UM’s accuracy to $\sim 30\%$, while the multi-step attacks, e.g., PGD-20 and CW-20, drop its accuracy to $0\%$. It is also visualized in the first column of Figure 3, where all the AEs locate far from their original class, and fit into the distributions of other classes. As a standard benchmark defense model, “AT” provides a baseline for improvements on both single-step and multi-step attacks.

**The Effect of Bidirectional Metric Learning** “SML” and “BML” both apply metric learning to constrain the clean image and its AE to keep a short distance, while push away the images in different classes. The difference between them is that SML only considers forward attacks and BML considers both forward and backward attacks. In the second and third columns of Figure 3, we can see that the SML does pull many of the AEs back to their original class, i.e., purple in the first row and cyan in the second row. However, one of the side effects of SML is that it makes classes confusing for clean images and thus may make a significant accuracy drop on clean images. In contrast, BML keeps better separations between different classes, and has much less amount of AEs located far away compared to SML. It demonstrates the benefit of the bidirectional strategy.

**The Effect of Attention Guided Knowledge Distillation** Utilizing “AGKD” alone is able to obtain a good accuracy, which is better than BML. By integrating both AGKD
Table 1. Evaluation results on CIFAR-10 and SVHN, under different widely used attacks. For CIFAR-10 dataset, we grouped the models by small-number or larger-number steps attack in training. “Bilateral” and “FS” use one-step attack, AGKD-BML uses two-step attack, and others use large-number steps with the step numbers show followed by the model names. “AGKD-BML-7” is a variant of AGKD-BML that uses 7-step attack for training. The best accuracy for each group is illustrated as bold, and the overall best accuracy is highlighted.

<table>
<thead>
<tr>
<th>Attacks(steps)</th>
<th>clean</th>
<th>FGSM</th>
<th>BIM(7)</th>
<th>PGD (20)</th>
<th>PGD (100)</th>
<th>CW (20)</th>
<th>CW (100)</th>
<th>AA [10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UM</td>
<td>95.99%</td>
<td>31.39%</td>
<td>0.38%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Bilateral [41]</td>
<td>91.2%</td>
<td>70.7%</td>
<td>-</td>
<td>57.5%</td>
<td>55.2%</td>
<td>56.2%</td>
<td>53.8%</td>
<td>29.35%</td>
</tr>
<tr>
<td>FS [49]</td>
<td>90.0%</td>
<td>78.4%</td>
<td>-</td>
<td>70.5%</td>
<td>68.6%</td>
<td>62.4%</td>
<td>60.6%</td>
<td>36.64%</td>
</tr>
<tr>
<td>AGKD-BML</td>
<td>91.99%</td>
<td>76.69%</td>
<td>73.81%</td>
<td>71.02%</td>
<td>70.72%</td>
<td>63.67%</td>
<td>62.55%</td>
<td>37.07%</td>
</tr>
<tr>
<td>AT-7 [27]</td>
<td>86.19%</td>
<td>62.42%</td>
<td>54.99%</td>
<td>45.51%</td>
<td>45.22%</td>
<td>46.26%</td>
<td>46.05%</td>
<td>44.04%</td>
</tr>
<tr>
<td>SML-7 [28]</td>
<td>86.21%</td>
<td>58.88%</td>
<td>52.60%</td>
<td>51.59%</td>
<td>46.62%</td>
<td>48.05%</td>
<td>47.39%</td>
<td>47.41%</td>
</tr>
<tr>
<td>TRADES+CAS-10 [3]</td>
<td>85.83%</td>
<td>65.21%</td>
<td>-</td>
<td>55.99%</td>
<td>-</td>
<td>67.17%</td>
<td>-</td>
<td>48.40%</td>
</tr>
<tr>
<td>MART+CAS-10 [3]</td>
<td>86.95%</td>
<td>63.64%</td>
<td>-</td>
<td>54.37%</td>
<td>63.16%</td>
<td>-</td>
<td>48.45%</td>
<td>-</td>
</tr>
<tr>
<td>AGKD-BML-7</td>
<td>86.25%</td>
<td>70.06%</td>
<td>64.97%</td>
<td>57.30%</td>
<td>56.88%</td>
<td>53.36%</td>
<td>52.95%</td>
<td>50.59%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attacks(steps)</th>
<th>clean</th>
<th>FGSM</th>
<th>BIM(10)</th>
<th>PGD (20)</th>
<th>PGD (100)</th>
<th>CW (20)</th>
<th>CW (100)</th>
<th>MIM (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UM</td>
<td>96.36%</td>
<td>46.33%</td>
<td>1.54%</td>
<td>0.33%</td>
<td>0.22%</td>
<td>0.37%</td>
<td>0.24%</td>
<td>5.39%</td>
</tr>
<tr>
<td>Bilateral [41]</td>
<td>94.1%</td>
<td>69.8%</td>
<td>-</td>
<td>53.9%</td>
<td>50.3%</td>
<td>-</td>
<td>48.9%</td>
<td>-</td>
</tr>
<tr>
<td>FS [49]</td>
<td>96.2%</td>
<td>83.5%</td>
<td>-</td>
<td>62.9%</td>
<td>52.0%</td>
<td>61.3%</td>
<td>50.8%</td>
<td>-</td>
</tr>
<tr>
<td>AT-7 [27]</td>
<td>91.55%</td>
<td>67.13%</td>
<td>54.03%</td>
<td>45.64%</td>
<td>44.02%</td>
<td>47.14%</td>
<td>45.66%</td>
<td>52.13%</td>
</tr>
<tr>
<td>SML-7 [28]</td>
<td>83.95%</td>
<td>70.28%</td>
<td>57.58%</td>
<td>51.91%</td>
<td>49.81%</td>
<td>51.25%</td>
<td>49.31%</td>
<td>43.80%</td>
</tr>
<tr>
<td>TRADES+CAS-10 [3]</td>
<td>91.69%</td>
<td>70.79%</td>
<td>-</td>
<td>55.26%</td>
<td>-</td>
<td>60.10%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MART+CAS-10 [3]</td>
<td>93.05%</td>
<td>70.30%</td>
<td>-</td>
<td>51.57%</td>
<td>-</td>
<td>53.38%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AGKD-BML</td>
<td>95.04%</td>
<td>89.32%</td>
<td>75.06%</td>
<td>74.94%</td>
<td>69.23%</td>
<td>69.85%</td>
<td>62.22%</td>
<td>76.86%</td>
</tr>
</tbody>
</table>

Figure 3. t-SNE plots for illustrating the sample representations in feature space. The triangle points with different colors represent the clean images in different classes, and the red circle points are the adversarial examples under PGD-20 attack. Best viewed in color.

and BML, the proposed AGKD-BML obtain the best performance in terms of both quantitative and qualitative results. In the fourth column of Figure 3, AGKD-BML pulls most of the AEs back to their original class, while keeps better separation between classes than BML does. We also provide the attention maps of the AEs obtained by the trained models in Figure 4. Compared to AT, AGDK-BML obtains better attention maps which are more identical to the ones
of clean images. This suggests that the AGKD does help on correcting the representation of AEs in feature space.

**Table 2. Ablation study on CIFAR-10 dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>FGSM</th>
<th>PGD (20)</th>
<th>CW (20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UM</td>
<td>31.39%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>AT [27]</td>
<td>62.42%</td>
<td>45.57%</td>
<td>46.26%</td>
</tr>
<tr>
<td>SML [28]</td>
<td>58.88%</td>
<td>51.59%</td>
<td>48.05%</td>
</tr>
<tr>
<td>BML</td>
<td>71.08%</td>
<td>60.51%</td>
<td>56.53%</td>
</tr>
<tr>
<td>AGKD</td>
<td>75.57%</td>
<td>65.93%</td>
<td>60.71%</td>
</tr>
<tr>
<td>AGKD-BML</td>
<td>76.69%</td>
<td>71.02%</td>
<td>63.67%</td>
</tr>
</tbody>
</table>

**4.4 Black box adversarial robustness**

To evaluate the black-box adversarial robustness, i.e., the adversary has no knowledge about the model, we generate an AE for each clean image in CIFAR-10 testing set by using natural models under PGD-20 attack with $\epsilon = 8$. Then the AGKD-BML model, as well as the comparison models, are tested on the generated adversarial example data. As demonstrated in Table 3, AGKD-BML model achieves the best accuracy among the models suggesting that AGKD-BML is robust to the black-box attacks as well.

**Table 3. Black-box adversarial robustness.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>85.4%</td>
<td>88.9%</td>
<td>89.9%</td>
<td>86.4%</td>
<td>90.75%</td>
</tr>
</tbody>
</table>

**5 Conclusion**

We proposed a novel adversarial training based model, named as AGKD-BML, that integrates two modules, i.e., the attention guided knowledge distillation module and the bi-directional metric learning module. The first module transfers attention knowledge of the clean image from a teacher model to a student model, so as to guide student model for obtaining better attention map, as well as correcting the intermediate features corrupted by adversarial examples. The second module efficiently regularizes the representation in the feature space, by shortening the representation distance between original image and its forward adversarial example, while enlarging the distance between the forward and backward adversarial examples. Extensive adversarial robustness experiments on two popular datasets with various attacks show that our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches.

**Acknowledgement.** This work is supported by the U.S. Department of Energy, Office of Science, High Energy Physics under Award Number DE-SC-0012704 and the Brookhaven National Laboratory LDRD #19-014, and in part by National Science Foundation Award IIS-2006665.
References


