Diverse Knowledge Distillation (DKD): A Solution for Improving The Robustness of Ensemble Models Against Adversarial Attacks

Ali Mirzaeian* , Jana Kosecka*, Houman Homayoun†, Tinoosh Mohsenin‡, Avesta Sasan*
*Department of ECE, George Mason University, e-mail: {amirzaei, kosecka, asasan}@gmu.edu
†Department of ECE, University of California, Davis, e-mail: hhomayoun@ucdavis.edu
‡Department of ECE, University of Maryland, Baltimore County, e-mail: tinoosh@umbc.edu

2021 22nd International Symposium on Quality Electronic Design (ISQED) | 978-1-7281-7641-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISQED51717.2021.9424353

***Abstract*—This paper proposes an ensemble learning model that is resistant to adversarial attacks. To build resilience, we introduced a training process where each member learns a radically distinct latent space. Member models are added one at a time to the ensemble. Simultaneously, the loss function is regulated by a reverse knowledge distillation, forcing the new member to learn different features and map to a latent space safely distanced from those of existing members. We assessed the security and performance of the proposed solution on image classification tasks using CIFAR10 and MNIST datasets and showed security and performance improvement compared to the state of the art defense methods.**

***Index Terms*—Adversarial Examples, Ensemble Learning, Knowledge Distillation**

I. Introduction

In the past decade, the research on Neuromorphic-inspired computing models, and the applications of Deep Neural Networks (DNN) for estimation of hard-to-compute functions, or learning of hard-to-program tasks have significantly grown, and their accuracy have considerably improved. Early research on learning models mostly focus on improving the accuracy of the models [1, 2], but as models matured, researchers explored other dimensions, such as energy efficiency of the models [3–8] and underlying hardware [9–15], as well as wider adoption and application of learning solutions in many other research fields including security applications [16–18] for solving problems that either have no closed-form solution or are too complex for developing a programmable solution. The wide adoption of these capable solutions then started raising concerns over their security.

Among many security aspects of learning solutions, the vulnerability of models to adversarial attacks [19, 20], has attracted lots of attention. Researchers have shown that subtle, yet targeted adversarial perturbation to the input (*i.e.* image, audio, or video input) of neural networks can dramatically drop their performance [21, 22].

The vulnerability of DNNs to adversarial attacks has raised serious concerns for using these models in critical applications in which an adversary can slightly perturb the input to fool the model [23]. This paper focuses on adversarial attacks on image classification models where an adversary manipulates an input image, forcing the DNN to misclassify.

Non-robust features are those features that strongly associate within a specific class, yet have small variation across categories [24]. Ilyas and et al. at [25] showed that the high sensitivity of the underlying model to the non-robust features existing at the input dataset is a significant reason for the vulnerability of the model to the adversarial examples. So an adversary crafts a perturbation that accentuates the non-robust features to achieve a successful adversarial attack.

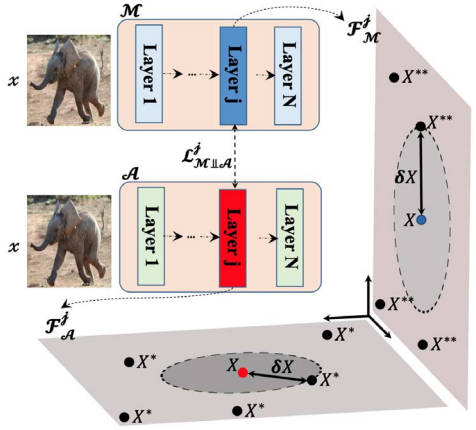

Fig. 1: In our proposed solutions, we train an auxiliary model(s) that tracks the main model's classification while learning a diverse set of features, latent representation of which maps to space far apart from the teacher's. $\mathcal{L}^j_{\mathcal{M}\perp\!\!\!\perp\mathcal{A}}$ indicates the diversity between the j^{th} feature space of the model $\mathcal{M}$ and $\mathcal{A}$ which are $\mathcal{F}^j_{\mathcal{M}}$ and $\mathcal{F}^j_{\mathcal{A}}$, respectively.

From this discussion, a means of building robust classifiers is identifying robust features and training a model using only robust features (that have a sizeable intra-class variation), making it harder for an adversary to mislead the classifier, [25]. Motivated by this discussion, we proposed a simple yet effective method for improving the resilience of DNNs, by introducing auxiliary model(s) trained in the spirit of knowledge distillation, while forcing diversity across features formed in their latent spaces.

We argue that adversarial attacks are transferable across models because they learn similar latent spaces for non-robust features. This is either the result of using the 1) same training set or 2) knowledge distillation that solely focused on improving the classification accuracy. In other words, for sharing the dataset or the knowledge of a trained network (on a dataset), the potential vulnerabilities of models coincide. Hence, an attack that works on one model is very likely to work on the other(s). This conclusion is also supported by the observations by Ilyas and et al. at [25]. From this argument, we propose to augment the task of knowledge distillation

with an additional and explicit requirement that the learned features by the student (*i.e.* auxiliary) model(s) should be distinct and independent from the teacher (*i.e.* main) model. For this reason, as showed in Fig. 1, we introduce the concept of Latent Space Separation (LSS), forcing the auxiliary model to learn features with little or no correlation to those of the teacher's. Hence, an adversarial attack on the main model will have minimal impact on the latent representation of features learned by the auxiliary model(s).

II. Background

Prior researches on adversarial learning have produced different explanations on why leaning models are easily fooled by adversarial input perturbation. The early investigations blamed the non-linearity of neural networks for their vulnerability [20, 26]. However, this perception was later challenged by Goodfellow and et. al. [20], who developed the Fast Gradient Sign Method (FGSM), explaining how neural network linearity can be exploited for rapidly building adversarial examples.

Building robust learning models that could resist adversarial examples has been a topic of interest for many researchers. Some of the most notable prior art on this topic includes 1) Adversarial Training, 2) Knowledge Distillation (KD), and 3) de-noising and refinement of the adversarial examples.

Adversarial Training: It is the process of incremental training of a model with the known adversarial examples to improve its resilience, see Fig2. The problem with this approach is that the model's resilience only improves when the model is attacked with similarly generated adversarial examples [27, 28].

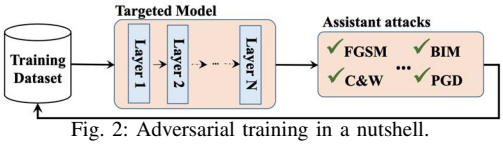

Fig. 2: Adversarial training in a nutshell.

Knowledge Distillation (KD), see Fig. 3: In this method, a compact (student) model learns to follow one or more teacher's models behavior. It was originally introduced to build compact models (students) from more accurate yet larger models (teachers). Later, it was also used to diminish the sensitivity of the student's output model concerning the input's perturbations [29]. However, the work in [30] showed that if the attacker has access to the student model, with minor changes, the student model could be as vulnerable as the teacher. Specifically, knowledge distillation can be categorized as a gradient masking defense [31] in which the magnitude of the underlying model's gradients are reduced to minimize the effect of changes in the model's input to its output. Although grading masking defenses can be an effective defense against white-box attacks, they are not resistant against black-box evasion attacks [30]. Our proposed solution is motivated by KD. However, we do not force the auxiliary (*i.e.* student) network(s) to follow the output layers of the main network (*i.e.* teacher); in contrary to the KD, the auxiliary network has to learn a different latent space while being trained for the same task and on the same dataset.

Refining the Input Image, see Fig 4: The adversarial defenses that rely on refining the input samples try to denoise the input image using some sort of autoencoder (variational, denoising, etc.) [32]. In this approach, the image is first

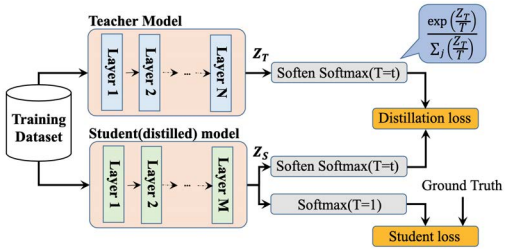

Fig. 3: Defensive Distillation in a nutshell.

encoded using a deep network to extract a latent code (a lossy, compressed, yet reconstructable representation of the input image). Then the image is reconstructed using a decoder. Next, the decoded image is fed to the classifier [32]. However, this approach suffers from two main weaknesses: (1) The reconstruction error of decoder can significantly reduce the classifier's accuracy and such reconstruction error increases as the number of input classes increases; (2) the used encoder-network is itself vulnerable to adversarial attacks which means new adversarial examples can be crafted on the model which also include the encoder-network.

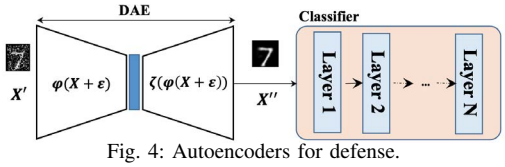

Fig. 4: Autoencoders for defense.

III. The Proposed Method

Our objective is to formulate a knowledge distillation process in which one or more auxiliary (student) models $\mathcal{A}_i$ are trained to closely follow the prediction of a main (teacher) model $\mathcal{M}$, while being forced to learn substantially different latent spaces. For example, in Fig. 5, let's assume three auxiliary models $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ have been trained alongside the main model $\mathcal{M}$ to have the maximum diversity between their latent space representations. Our desired outcome is to assure that an adversarial perturbation that could move the latent space of the input sample x out of its corresponding class boundary, $\mathcal{F}^j_{\mathcal{M}}$, has a negligible or small impact on the movement of the corresponding latent space of x in the class boundaries of the auxiliary models $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$. Hence, an adversarial input that could fool model $\mathcal{M}$, becomes less effective or ineffective on the auxiliary models. This objective is reached by the way the loss function of each model is defined. The details of $\mathcal{M}$ and $\mathcal{A}_i$ network(s), learning procedure and objective function of $\mathcal{A}_i$ are explained next:

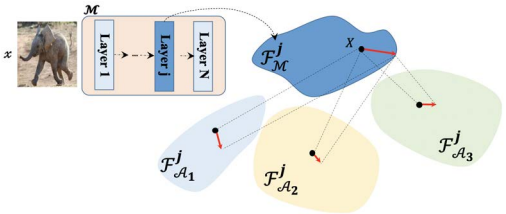

Fig. 5: $\mathcal{F}^j_{\mathcal{M}}$ is the latent space regards to jth layer of the main model, $\mathcal{F}^j_{\mathcal{A}_1}$, $\mathcal{F}^j_{\mathcal{A}_2}$, $\mathcal{F}^j_{\mathcal{A}_3}$ are the corresponding latent spaces of the auxiliary models 1, 2, and 3. X is the latent feature of an input sample X i.e., $X = \mathcal{F}^j_{\mathcal{M}}(x)$. The red arrow at the main model shows the direction of the adversarial perturbation, the red arrows for the auxiliary models show the projection of the perturbation on the latent spaces of the auxiliary model.

Main Model ($\mathcal{M}$:) In this paper, we evaluated our proposed method on two datasets MNIST [33] and CIFAR10 [34]. So depending on the underlying dataset, the structure of the main model (teacher) $\mathcal{M}$ is selected as showed in Table I. We also employed cross-entropy loss $\mathcal{C}$, see Eq. 1, as the objective function for training the model $\mathcal{M}$.

$$\mathcal{L}_{\mathcal{M}} = \mathcal{C}(Y, \mathcal{M}(X)) = -\sum_{i=1}^{i=N} Y_i \log(\mathcal{M}(X_i)) \tag{1}$$

In this equation, X_i and Y_i are i^{th} training sample and its corresponding label, respectively.

Auxiliary Model ($\mathcal{A}$): Each auxiliary model is a structural replica of the main model $\mathcal{M}$. However, model $\mathcal{A}$ is trained using a modified KD training process: let's denote the output of j^{th} layer of model $\mathcal{A}$ (*i.e.* latent space of model $\mathcal{A}$) and $\mathcal{M}$ by $\mathcal{F}^j_{\mathcal{A}}$ and $\mathcal{F}^j_{\mathcal{M}}$ respectively. Our training objective is to force model $\mathcal{A}$ to learn very different latent space compared to $\mathcal{M}$ where both do the same classification task on the same dataset. To achieve this, the term $\mathcal{L}^j_{\mathcal{M} \perp\!\!\!\perp \mathcal{A}}$ which shows the similarity of the j^{th} layer of model A to j^{th} layer of model $\mathcal{M}$ is defined as follows:

$$\mathcal{L}^j_{\mathcal{M} \perp\!\!\!\perp \mathcal{A}} = \frac{\mathcal{F}(X)^j_{\mathcal{M}} \cdot \mathcal{F}(X)^j_{\mathcal{A}}}{|\mathcal{F}(X)^j_{\mathcal{M}}||\mathcal{F}(X)^j_{\mathcal{A}}|}, \quad X \in \mathbf{T} \tag{2}$$

In this objective function, $\mathbf{T}$ is the dataset, and the $(.)$ is the inner product function. This similarity measure then is factored to define the loss function $(\mathcal{L}_{\mathcal{A}})$ for training the model $\mathcal{A}$, which increases the dissimilarity of the layer j of the model $\mathcal{A}$ with respect to the $\mathcal{M}$:

$$\mathcal{L}_{\mathcal{A}} = \underbrace{(1-\zeta)\mathcal{L}_{\mathcal{C}}(Y, \mathcal{A}(X))}_{targets\,the\,accuracy} + \underbrace{\zeta \mathcal{L}^j_{\mathcal{M} \perp\!\!\!\perp \mathcal{A}}}_{targets\,the\,diversity} \tag{3}$$

In this equation, ζ is a regularization parameter to control the contribution of each term.

Let's assume the adversarial perturbation δ, when added to the input X, forces the model $\mathcal{M}$ to misclassify X, or more precisely $\mathcal{M}(X+\delta) \neq \mathcal{M}(X)$. For this misclassification (evasion) to happen, in a layer j (close to the output) the added noise has forced some of the class-identifying features outside its related class boundary learned by model $\mathcal{M}$. However, the class boundaries learned by $\mathcal{A}$ and $\mathcal{M}$ are quite different. Therefore, as showed in Fig. 5, although noise δ can move a feature out of its learned class boundary in model $\mathcal{M}$, it has very limited power in displacing the features learned by model $\mathcal{A}$ outside of its class boundary in layer j. In other words, although the term $\mathcal{F}(X)^j_{\mathcal{M}} \cdot \mathcal{F}(X)^j_{\mathcal{A}}$ between the main and the auxiliary models has a low value, the term $\mathcal{F}(X+\delta)^j_{\mathcal{A}} \cdot \mathcal{F}(X)^j_{\mathcal{A}}$ between the auxiliary model before and after adding perturbation δ has a high value, subsequently the student model $\mathcal{A}$ has a low sensitivity to the perturbation δ, meaning $\mathcal{A}(X+\delta) = \mathcal{A}(X)$.

A. Black and White-Box defense:

The auxiliary models could defend against both white and black box attacks, description and explanation for each is given next:

Black-box Defense: In black box attack, an attacker has access to model $\mathcal{M}$, and can apply her desired input to the model and monitor the model's prediction for designing an attack and adding the adversarial perturbation to the input X. Considering no access to the model $\mathcal{A}$, and for having very different feature space, the model $\mathcal{A}$ remains resistant to black-box attacks and using a single $\mathcal{A}$ is sufficient.

White-box Defense: In white-box attack, the attacker knows everything about the model $\mathcal{M}$ and $\mathcal{A}$, including model parameters and weights, full details of each model's architecture, and the dataset used for training the network. For this reason, using a single model $\mathcal{A}$ is not enough, as that model could be used for designing the attack. However, we can make the attack significantly more difficult (and improve the classification confidence) by training and using multiple robust auxiliary models. However, each of our $\mathcal{A}$ models learns different features compared with all other auxiliary models. Then, to resist the white-box attack, we create a majority voting system from the robust auxiliary models.

TABLE I: The Architecture of the ensemble models

MNIST Architecture		CIFAR10 Architecture	
Relu Convolutional	32 filters (3×3)	Relu Convolutional	96 filters (3×3)
Relu Convolutional	32 filters (3×3)	Relu Convolutional	96 filters (3×3)
Max Pooling	2×2	Relu Convolutional	96 filters (3×3)
Relu Convolutional	64 filters (3×3)	Max Pooling	2×2
Relu Convolutional	64 filters (3×3)	Relu Convolutional	192 filters (3×3)
Max Pooling	2×2	Relu Convolutional	192 filters (3×3)
Relu Convolutional	200 units	Relu Convolutional	192 filters (3×3)
Relu Convolutional	200 units	Max Pooling	2×2
Softmax	10 units	Relu Convolutional	192 filters (3×3)
		Relu Convolutional	192 filters (1×1)
		Relu Convolutional	192 filters (1×1)
		Global Avg. Pooling	
		Softmax	10 units
System Configuration and training hyper parameters			
OS: Red Hat 7.7, Pytorch: 1.3, FoolBox: 2.3.0, GPU: Nvidia Tesla V100, EPOCH: 100, MNIST Batch Size: 64, CIFAR10 Batch Size:128, Optimizer: ADAM, learning rate: 1e-4			

Let's assume we want to train $k > 1$ auxiliary models, $\mathcal{A}_{i=1}^{i=k}$, each having a diverse latent space (*i.e.* $\mathcal{M} \perp\!\!\!\perp \mathcal{A}_1 \perp\!\!\!\perp \mathcal{A}_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp \mathcal{A}_k$). To learn these networks, firstly, based on Eq. 3, the $\mathcal{A}_1$ is learned aiming $\mathcal{M} \perp\!\!\!\perp \mathcal{A}_1$. Then, the $\mathcal{A}_2$ is learned to be diverse of both $\mathcal{A}_1$ and $\mathcal{M}$. This process continues one by one, reaching the $\mathcal{A}_i$ model, till its latent space is diverse from all previous models (*i.e.* $\mathcal{A}_i \perp\!\!\!\perp \{\mathcal{M}, \mathcal{A}_{j=1}^{j=i-1}\}$). According to this discussion for learning the i^{th} auxiliary model, the loss function is defined as:

$$\mathcal{L}_{\mathcal{A}} = (1-\zeta)\mathcal{L}_{\mathcal{C}}(Y, \mathcal{A}_i(X)) + (\zeta/i)\mathcal{L}^i_{\mathcal{M} \perp\!\!\!\perp \mathcal{A}_i} + (\zeta/i) \sum_{j=1}^{y=i-1} \mathcal{L}^j_{\mathcal{A}_y \perp\!\!\!\perp \mathcal{A}_i} \tag{4}$$

Finally, to increase the confidence of the prediction, instead of a simple majority voting system for the top-1 candidate, we consider a boosted defense in which the voting system considers the top n candidates of each model $\mathcal{A}$ for those cases that majority on top-1 fails (there is no majority between the top-1 predictions). This gives us two benefits 1) if a network misclassifies due to adversarial perturbation, there is still a high chance for the network to assign a high probability (but not the highest) to the correct class. 2) if a model is confused between a correct class and a closely related but incorrect class and assigns the top-1 confidence to the wrong class, it still helps identify the correct class in the voting system.

IV. Experimental Results

In this section, we evaluate the performance of our proposed defense against various white and black-box attacks. We trained all models using Pytorch framework [35], and all attack scenarios using Foolbox [36] (which is a toolbox for crafting adversarial examples). The details of the hyperparameters and system configuration are summarized in Table I.

A. Enforcing Latent Space Hetromorphisim

To quantify the diversity of the latent space representations of the ensemble trained on a dataset D, we first define the Latent Space Separation (LSS) measure between latent spaces of two models A_1 and A_2 as:

$$LSS_D(\mathcal{F}^j_{\mathcal{A}_1}, \mathcal{F}^j_{\mathcal{A}_2}) = \frac{2}{\|W\|}. \tag{5}$$

In which $\mathcal{F}^j_{\mathcal{A}_1}$, $\mathcal{F}^j_{\mathcal{A}_2}$ are latent space representations of the dataset D for the j^{th} layer of models A_1 and A_2, respectively. W is the normal vector of the hyperplane obtained by Support Vector Machine (SVM) classifier [38] for linearly separate the latent spaces obtained on the dataset D. More precisely, LSS between two latent spaces is obtained by following these 4 steps: 1) training both models $\mathcal{A}_1$, $\mathcal{A}_2$ on a dataset i.e., MNIST. 2) generating the latent space of each model on the evaluation set i.e., $\mathcal{F}^j_{\mathcal{A}_1}$ and $\mathcal{F}^j_{\mathcal{A}_2}$. 3) turning the latent representations into a two-class classification problem tackled by SVM classifier 4) using SVM margin as LSS

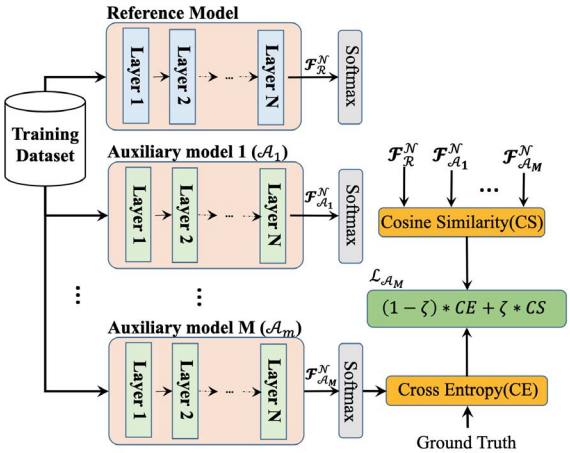

Fig. 6: training method of the DKD for training an ensemble of m auxiliary models $\mathcal{A}_1$ to $\mathcal{A}_m$ from a reference model $\mathcal{R}$. $\mathcal{F}^{\mathcal{N}}_{\mathcal{R}}$, $\mathcal{F}^{\mathcal{N}}_{\mathcal{A}_1}$, and $\mathcal{F}^{\mathcal{N}}_{\mathcal{A}_m}$ are latent space of the reference model and auxilary models $\mathcal{A}_1$ and $\mathcal{A}_m$, related to output of N^{th} layer of each one, respectively. Cross-Entropy [37] and Cosine-Similarity are the objective functions that have been used for obtaining the total loss.

distance between two latent representations of the dataset MNIST, $LSS_{MNIST}(\mathcal{F}^j_{\mathcal{A}_1}, \mathcal{F}^j_{\mathcal{A}_2})$. This process can be expanded for obtaining LSS of more than two models. For instance Fig. 7 shows the LSS between model $\mathcal{A}_1$ for two models $\mathcal{A}_2$ and $\mathcal{A}_3$ regards to their j^{th} layer, i.e., $LSS_D(\mathcal{F}^j_{\mathcal{A}_1}, (\mathcal{F}^j_{\mathcal{A}_2}, \mathcal{F}^j_{\mathcal{A}_3}))$. Note that the SVM classifier should be set in the hard margin mode, meaning no support vector can pass the margins. When the SVM classification fails, it means that the latent spaces were not linearly separable i.e., there is either an overlap between latent spaces or the decision boundary cannot be modeled linearly. So in both cases, the more the marginal distance between latent spaces are, the higher is the diversity of formed latent spaces.

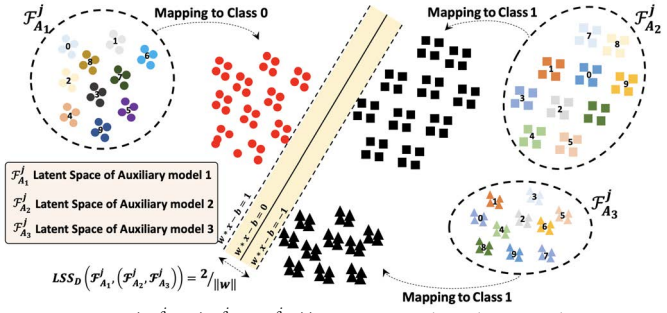

Fig. 7: $LSS_D(\mathcal{F}^j_{\mathcal{A}_1}, (\mathcal{F}^j_{\mathcal{A}_2}, \mathcal{F}^j_{\mathcal{A}_3}))$ in which $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ are three models with latent space representations $\mathcal{F}^j_{\mathcal{A}_1}$, $\mathcal{F}^j_{\mathcal{A}_2}$, and $\mathcal{F}^j_{\mathcal{A}_3}$, respectively. W is the norm vector of a hard margin support vector machine (SVM) in two class classification model.

Using Eq. 5, we define more generalized formula of LSS using an ensemble model comprised of N models, see Eq. 6. In fact the total LSS of an ensemble model is obtained by averaging the LSS of each model latent space versus all other models'. For instance, let's imagine the ensemble model comprised of three models A_1, A_2, and A_3. Then the total LSS is calculated by $1/3(LSS_D(F^j_{A_1}, (F^j_{A_2}, F^j_{A_3})) + LSS_D(F^j_{A_2}, (F^j_{A_1}, F^j_{A_3})) + LSS_D(F^j_{A_3}, (F^j_{A_1}, F^j_{A_2})))$. LSS measures the marginal distance of SVM in a two-class-classification task, so $LSS_D(A_1, (A_2, A_3))$, indicates that a SVM classification has been performed between the latent space of the model A_1 and an aggregation of latent spaces of other two models A_2 and A_3.

$$\frac{1}{N}\sum_{i=1}^{N} LSS_D(F^j_{A_i}, (F^j_{A_1}, ..., F^j_{A_{i-1}}, F^j_{A_{i+1}}, ..., F^j_{A_N})). \quad (6)$$

To investigate the effectiveness of LSS, as a metric for measuring the diversity between different latent spaces, we considered three different scenarios for training an ensemble of 3 models with the same structure: I) Random Initialization (RI) where 3 models trained independently with a random initial value. II) Knowledge Distillation (KD), where three models trained collaboratively, as shown in Fig. III) Diversity Knowledge Distillation (DKD), where three models are trained in a collaborative yet different manner of KD, see Fig. 6. Note that KD and RI methods deal with the softmax probabilities, However, DKD uses a mix of softmax and the latent spaces. Alongside each one of the designs DKD, KD, and RI, a boosted version of each one is implemented and denoted by DKD^*, KD^*, and RI^*, respectively. Both KD and DKD are trained in a one-by-one manner, meaning the i^{th} model considers the $i-1$ previously trained models at its training phase while those $i-1$ models are frozen i.e., their parameters (weights) are not updated while the i^{th} model is being trained.

TABLE II: The number of failed majority between an ensemble of the three models at the original and boosted version, indicated with *, of KD, DKD, and RI on the datasets MNIST and CIFAR10. The accuracy improvement using the boosted version is shown with Accuracy Improved (A.I). We investigated our proposed method against well-know attacks like DeepFool [39], C&W [40], JSMA [21], and FGSM [20].

Param.	MNIST						CIFAR10					
	KD	KD*	DKD	DKD*	RI	RI*	KD	KD*	DKD	DKD*	RI	RI*
Deep Fool							Deep Fool					
1	4	0	15	0	16	0	576	6	1234	51	891	33
200	7	1	93	0	61	0	628	5	1319	52	959	32
A.I.(%)	0.02		0.12		0.35		1.36		2.88		7.16	
C&W							C&W					
1	188	3	172	0	225	6	584	1	1201	2	859	13
200	560	5	794	2	921	16	630	2	1322	2	942	12
A.I.(%)	0.01		0.33		0.22		13.97		1.47		2.79	
JSMA							JSMA					
1	0	0	15	0	16	6	584	8	1201	51	859	31
200	4	1	52	0	42	8	654	8	1410	56	999	35
A.I.(%)	0.01		0.26		0.15		1.39		2.94		4.18	
FGSM							FGSM					
0.04	190	5	342	16	235	20	821	3	1801	106	1560	66
0.08	239	6	589	27	349	34	924	9	1842	71	1872	77
0.1	349	3	844	36	504	46	950	12	1829	79	2053	105
A.I.(%)	0.03		0.21		0.12		1.33		3.74		2.94	

Fig. 8-top shows the LSS for an ensemble of three models (for MNIST and CIFAR10 datasets), with the structures described in Table. I. Fig. 8-A and C show that for the MNIST dataset, increasing the value of the ζ causes 1) rapid increase at the LSS 2) slight drop at the classification accuracy. Considering the Eq. 4, increasing the ζ means putting less emphasis on the cross-entropy term, which reflects the slight drop of accuracy and increases the LSS of the latent spaces of the ensemble models. A similar pattern happens for the CIFAR10 dataset, Fig. 8-B and D, in which LSS increases to its maximum at the $\zeta = 0.5$ while the classification accuracy slightly increases. Between all the different values for acceptable accuracy, the ζ value that leads to a higher LSS is selected as the parameter. In other words, to have a diverse latent space, the LSS between them should be maximized while the accuracy is kept in the acceptable range. For example, in Fig. 8-B, when ζ is 0.5, the LSS is at the maximum level while accuracy is also at an acceptable level. Accordingly, the evaulations has been done when the parameters ζ is set to 0.5 and 0.9 for the CIFAR10 and MNIST datasets.

One instant observation in Fig. 8 is that the LSS between the latent spaces obtained by DKD approach is noticeably larger than RI, and the LSS between latent spaces obtained by RI is slightly larger than KD. This observation is aligned with our expectations because the DKD is designed to increase the diversity between the latent spaces while the KD in essence, increases the similarity between models and this is because the student model(s) imitates the behavior of the teacher(s).

B. Resistance to Black-Box Attacks

For launching a black-box attack, the adversary uses a reference model and trains it based on the available dataset. Then knowing the transferability of the adversarial example, the adversary extracts the adversarial sample on the reference model and applies it to the models under attack. Assume the adversary used LENET and VGG16 as the reference model for MNIST and CIFAR10, respectively. The underlying models under attack are an ensemble of three models with the structure shown in Table I. For investigating the performance of the proposed method (DKD) on black-box attacks, we also considered

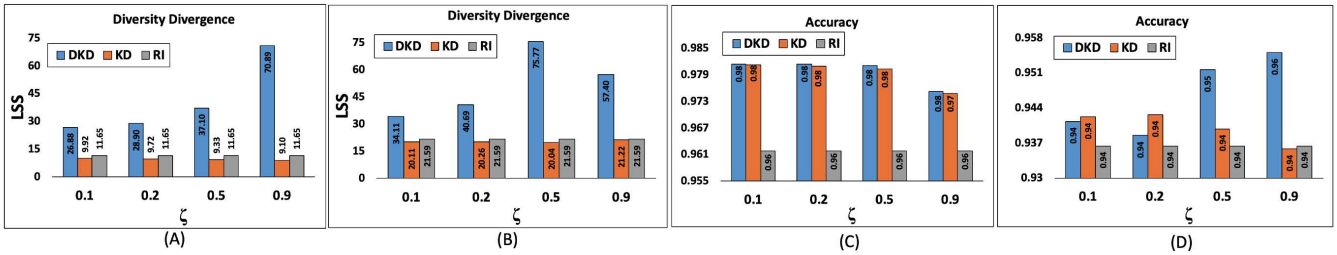

Fig. 8: (A), (B) show the LSS of an ensemble of three models on the MNIST, and CIFAR10 datasets. (C), (D) show the classification accuracy of the ensemble model on the MNIST, and CIFAR10 datasets. In this figure, three method Random Initialization(RI), Knowlege Distillation (KD), and Diversity Knowledge Distillation(DKD) are shown.

TABLE III: Black Box adversarial attack on an ensemble of three models. The bold numbers show the most resistant defense mechanism. The perturbation size, ϵ, is set to 0.1 and 0.2 for MNIST, and 0.04, and 0.08 for CIFAR10. The initial constant C is set to 10 and 0.1 for MNIST and CIFAR10 respectively. Iterative attacks are executed for 200 iterations.

MNIST								CIFAR10							
Param.	3 Ensemble Models						Ref.	Param.	3 Ensemble Models						Ref.
	KD	KD*	DKD	DKD*	RI	RI*			KD	KD*	DKD	DKD*	RI	RI*	
DeepFool [39]								DeepFool							
1	0.9510	0.9542	0.9721	**0.9726**	0.962	0.9682	0.9754	1	0.8355	0.8475	0.92	**0.9463**	0.8596	0.92	0.6505
200	0.864	0.8642	0.8904	**0.8916**	0.8433	0.8768	0.5835	200	0.7972	0.8108	0.8981	**0.9269**	0.8265	0.8981	0.1586
C&W [40]								C&W							
1	0.9612	0.9614	0.988	**0.9882**	0.9721	0.9726	0.9841	1	0.8475	**0.959**	0.9534	0.9575	0.9326	0.9571	0.9516
200	0.7829	0.783	0.8587	**0.862**	0.8127	0.8149	0.2	200	0.8014	0.9311	0.9245	**0.9321**	0.9022	0.9301	0.1543
JSMA [21]								JSMA							
1	0.9612	0.9614	0.9882	**0.9884**	0.9721	0.9726	0.9854	1	0.8475	0.8599	0.9326	**0.9571**	0.8348	0.8703	0.9516
200	0.8147	0.8148	0.9086	**0.9112**	0.8403	0.8418	0.4322	200	0.7804	0.7943	0.8851	**0.9145**	0.7644	0.8062	0.1564
FGSM [20]								FGSM							
0.04	0.9553	0.9555	0.9842	**0.9846**	0.9638	0.9644	0.952	0.04	0.8412	0.8548	0.899	**0.933**	0.8296	0.8676	0.4822
0.08	0.9247	0.9249	0.9619	**0.9629**	0.9333	0.9341	0.864	0.08	0.7476	0.7615	0.7393	**0.7763**	0.7267	0.7627	0.2534
0.1	0.8937	0.894	0.9374	**0.9395**	0.9059	0.9071	0.7847	0.1	0.6998	**0.7131**	0.6839	0.7213	0.6817	0.7111	0.0867
No Attack								No Attack							
-	0.9612	0.9614	0.9882	**0.9884**	0.9721	0.9726	0.9854	-	0.9501	0.9512	0.9561	**0.9580**	0.9385	0.9523	0.9607

TABLE IV: white box attack on the ensemble of three models. Bold numbers at each column show the most resistant method against white-box attacks. The perturbation size, ϵ, is set to 0.1 and 0.2 for MNIST, and 0.04, and 0.08 for CIFAR10. The initial constant C is set to 10 and 0.1 for MNIST and CIFAR10 respectively. Iterative attacks are executed for 200 iterations.

	MNIST						CIFAR10					
Defense	**Clean**	**FGSM 0.1**	**FGSM 0.2**	**JSMA**	**C&W**	**DeepFool**	**Clean**	**FGSM 0.04**	**FGSM 0.08**	**JSMA**	**C&W**	**DeepFool**
No Defense	0.9661	0.651	0.119	0.2421	0.4409	0.0	0.9304	0.2033	0.1846	0.2525	0.2548	0.1441
DKD*$_{Prj}$	0.9884	**0.9835**	**0.9681**	0.8849	0.9594	0.8991	**0.9604**	0.7377	0.7188	0.7788	0.8878	0.7934
DKD*$_{Agg}$	0.9884	0.9756	0.9329	**0.9725**	**0.9808**	**0.9605**	**0.9604**	**0.8923**	**0.8623**	**0.9221**	**0.9538**	**0.8751**
Yu et al. [41]	0.984	0.916	0.703	0.8014	0.791	0.6518	0.9421	0.485	0.382	0.824	0.629	0.721
Ross et al. [42]	0.992	0.916	0.604	0.9191	0.753	0.7394	0.9491	0.395	0.205	0.836	0.478	0.751
Pang et al. [43]	**0.995**	0.963	0.528	0.9465	0.781	0.7921	0.9219	0.716	0.474	0.882	0.549	0.695
AdvTrain [44]	0.991	0.73	0.527	0.645	0.392	0.627	0.905	0.446	0.314	0.781	0.501	0.73

two other methods RI, and KD for training an ensemble of three models.

We used the majority voting between the ensemble models. However, in some cases, each one of the models results in a different prediction regards to other models. We refer to these cases as failed majorities. So for each one of the possible attacks and two datasets (MNIST and CIFAR10) we counted the number of failed majorities. Table II, shows the difference between the regular and boosted version of each benchmark with regard to the number of the failed majority voting. From this table, we observe that 1) the number of majority voting failures at the DKD is always higher than the other two regular methods. This confirms that our objective function could successfully train diverse models because the majority voting fails whenever the models couldn't agree on a label. So at the presence of the adversarial example, the disagreement between the models trained with DKD is higher than the others 2) The number of majority failures drops by going from a regular to a boosted model. The effect of this drop is shown by Accuracy Improvement percentage, A.I.

Table III shows the results of applying some of the state-of-art attacks on the DKD, KD, and RI on the MNIST, and CIFAR10 datasets. Investigating these two Tables two trends reveals 1) Boosted version has better performance than the regular version 2) Almost at all attack scenarios, DKD^* outperform the other defense scenarios.

C. Resistance to White-Box Attacks

We evaluated 2 white-box attacks: 1) Projected attack, in which the adversary obtains adversarial perturbation on one model and applies it to the target model(s). 2) The Aggregated attack, in which the adversary obtains the adversarial perturbation on each model

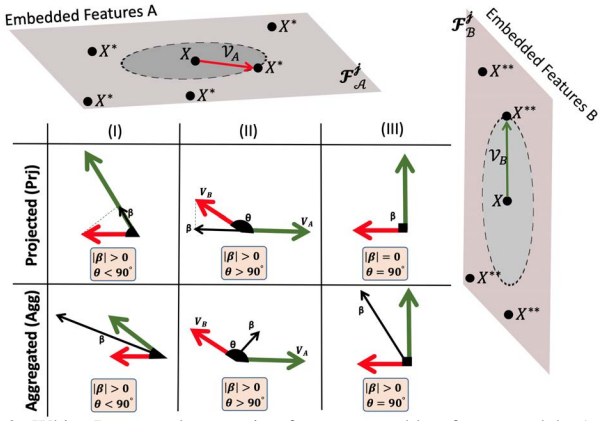

Fig. 9: White Box attack scenarios for an ensemble of two models A and B in which $\mathcal{F}^j_{\mathcal{A}}$ and $\mathcal{F}^j_{\mathcal{B}}$ are corresponding embedded features for layer jth of each model respectively. $\mathcal{V}_A$ and $\mathcal{V}_B$ are the adversarial perturbations on input sample x, i.e., $X^* = \mathcal{F}^j_{\mathcal{A}}(x) + \mathcal{V}_A$ and $X^{**} = \mathcal{F}^j_{\mathcal{B}}(x) + \mathcal{V}_B$. The angle between two vectors $\mathcal{V}_B$ and $\mathcal{V}_A$ is annotated with θ, while β is the magnitude of projected or aggregated perturbation for Projected or aggregated attack, respectively.

independently, then aggregates them to form one perturbation to apply to all models simultaneously. These two scenarios have been explained through a toy example on two models A and B in Fig. 9. For the projected attack, the adversary finds the direction τ_B as an adversarial perturbation for the input x of the model B, i.e., $\mathcal{F}^j_B(x+\tau_B) = X + \mathcal{V}_B = X^{**}$. In the second step, the adversary applies the obtained perturbation τ_B on the sample of the model A, with the hope that it may transfer to model A. Noted, projected attack is similar to black box scenario with only difference that attacker knows about the structures of auxiliary models and in this paper the auxiliary models have a same structure. Hence, based on the angle between the perturbations in the latent spaces of models A and B which are $\mathcal{V}_A$ and $\mathcal{V}_B$ three scenarios are possible, see Fig. 9. When both perturbations $\mathcal{V}_A$ and $\mathcal{V}_B$ are orthogonal (case III), a successful adversarial perturbation of one model does not transfer to the other model, and vice versa. In two other cases (I, II), the smaller the perturbations' angle, the more likely the adversarial perturbation transfer across models. Note that the projection of the perturbation τ_B on the embedded feature plane A, $\mathcal{F}^j_{\mathcal{A}}(\tau_B)$, shows the direction of the adversarial perturbation for the model A. If this projection is large enough to move the data point X out of its class boundary then input X misclassifies as X^*.

For the aggregated attack, the adversary obtains the adversarial perturbation of input sample x on the model A, τ_A, and separately obtains the adversarial perturbation B, τ_B. Then, the adversary calculates the aggregated perturbation by adding up these two perturbations i.e., $\tau_A + \tau_B$. In this scenario, let's assume $\mathcal{V}_A$, and $\mathcal{V}_B$ are the corresponding mapping of τ_A and τ_B on embedded feature planes A, and B, respectively. Based on the value of θ between $\mathcal{V}_A$, and $\mathcal{V}_B$, three different outcomes can be assumed, 1) Fig.9 Aggregated-II, $\theta > 90$, in which the aggregated perturbation i.e., β is smaller than either of perturbations $\mathcal{V}_A$, and $\mathcal{V}_B$. In this case, β as an adversarial perturbation cannot be a successful attack on either of the models. 2) Fig.9 Aggregated-I, $\theta < 90$, in which β is greater than either of perturbations $\mathcal{V}_A$, and $\mathcal{V}_B$, which means β as an adversarial perturbation leads to a misclassification in both models A, and B. However the resulted adversarial perturbation in this scenario is large and most likely perceptible to human. 3) Fig.9 Aggregated-III, $\theta = 90$, in which β is equal to either of perturbations $\mathcal{V}_A$, and $\mathcal{V}_B$, which means β as an adversarial perturbation lead to a misclassification on both models A, and B and most likely β is imperceptible to the human eyes.

Table IV captures the result of various adversarial attacks on our proposed solution. DKD^*_{Prj} shows the evaluation of DKD* when the adversary uses the projected attack and DKD^*_{Agg} shows the aggregated attack. As indicated in this table, our proposed solutions outperform prior art defense, illustrating the effectiveness of learning diverse features using our proposed solution.

V. Acknowledgement

This work was supported by Centauri Corp. and the National Science Foundation (NSF) through Computer Systems Research (CSR) program under NSF award number 1718538.

VI. Conclusion

To build robust models that could resist adversarial attacks, we proposed a solution for building an ensemble learning solution in which member models are forced to extract different features and learn radically different latent spaces. We also introduced Latent Space Separation (which is defined as the distance between the latent space representations of models in the ensemble) as a metric for measuring the ensemble's robustness to adversarial examples. The evaluation of our proposed solutions against the white and black box attacks indicates that our proposed ensemble model is resistant to adversarial solutions and outperforms prior art solutions.

References

[1] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. He et al., "Deep residual learning for image recognition," in *proc. of the IEEE conf. on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] K. Neshatpour et al., "Icnn: An iterative implementation of convolutional neural networks to enable energy and computational complexity aware dynamic approximation," in *2018 Design, Automation & Test in Europe Conf. & Exhibition (DATE)*. IEEE, 2018, pp. 551–556.

[4] ——, "Exploiting energy-accuracy trade-off through contextual awareness in multi-stage convolutional neural networks," in *20th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2019, pp. 265–270.

[5] ——, "Icnn: The iterative convolutional neural network," *ACM Transactions on Embedded Computing Systems (TECS)*, 2019.

[6] M. Hosseini *et al.*, "On the complexity reduction of dense layers from o (n 2) to o (nlogn) with cyclic sparsely connected layers," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2019, pp. 1–6.

[7] H. Sayadi et al., "Machine learning-based approaches for energy-efficiency prediction and scheduling in composite cores architectures," in *2017 IEEE International Conf. on Computer Design (ICCD)*, 2017, pp. 129–136.

[8] K. Neshatpour et al., "Design space exploration for hardware acceleration of machine learning applications in mapreduce," in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2018, pp. 221–221.

[9] A. Mirzaeian *et al.*, "Nesta: Hamming weight compression-based neural proc. engine," in *Proceedings of the 25th Asia and South Pacific Design Automation Conference*, 2020.

[10] Y. Chen et al., "Eyeriss v2: A flexible and high-performance accelerator for emerging deep neural networks," *CoRR*, vol. abs/1807.07928, 2018. [Online]. Available: http://arxiv.org/abs/1807.07928

[11] A. Mirzaeian *et al.*, "Tcd-npe: A re-configurable and efficient neural processing engine, powered by novel temporal-carry-deferring macs," in *2020 International Conference on ReConFigurable Computing and FPGAs (ReConFig)*, Jan 2020.

[12] A. Mehrabi *et al.*, "Bayesian optimization for efficient accelerator synthesis," *ACM Trans. Archit. Code Optim.*, vol. 18, no. 1, Dec. 2021. [Online]. Available: https://doi.org/10.1145/3427377

[13] M. Daneshtalab et al., *Hardware Architectures for Deep Learning*. Materials, Circuits and Device, 2020.

[14] S. R. Faraji *et al.*, "Hbucnna: Hybrid binary-unary convolutional neural network accelerator," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct 2020.

[15] A. Mehrabi *et al.*, "Prospector: Synthesizing efficient accelerators via statistical learning," in *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020, pp. 151–156.

[16] A. Vakil *et al.*, "Lasca: Learning assisted side channel delay analysis for hardware trojan detection," *arXiv preprint arXiv:2001.06476*, 2020.

[17] H. Sayadi *et al.*, "2smart: A two-stage machine learning-based approach for run-time specialized hardware-assisted malware detection," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 728–733.

[18] A. Vakil *et al.*, "Learning assisted side channel delay test for detection of recycled ics," *Asia and South Pacific Design Automation Conf.*, 2021.

[19] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *International Conf. on Learning Representations*, vol. abs/1312.6199, 2013.

[20] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," in *International Conf. on Learning Representations*, vol. abs/1412.6572, 2014.

[21] N. Papernot *et al.*, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[22] N. Papernot et al., "Practical black-box attacks against machine learning," in *ASIA CCS '17*, 2016.

[23] G. F. Elsayed *et al.*, "Adversarial examples that fool both computer vision and time-limited humans," in *NeurIPS*, 2018.

[24] F. Behnia et al., "Code-bridged classifier (cbc): A low or negative overhead defense for making a cnn classifier robust against adversarial attacks," in *International Symposium on Quality Electronic Design (ISQED) 2020*, 2020.

[25] A. Ilyas *et al.*, "Adversarial examples are not bugs, they are features," *ArXiv*, vol. abs/1905.02175, 2019.

[26] B. Biggio et al., "Evasion attacks against machine learning at test time," in *Joint European Conf. on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[27] S. Amberkar *et al.*, "Efficient utilization of adversarial training towards robust machine learners and its analysis," in *Proceedings of the International Conf. on Computer-Aided Design*. ACM, 2018, p. 78.

[28] F. Tramèr and D. Boneh, "Adversarial training and robustness for multiple perturbations," *ArXiv*, vol. abs/1904.13000, 2019.

[29] N. Papernot *et al.*, "Distillation as a defense to adversarial perturbations against desep neural networks," *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2015.

[30] N. Carlini et al., "Defensive distillation is not robust to adversarial examples," *arXiv preprint arXiv:1607.04311*, 2016.

[31] A. Athalye and others., "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conf. on Machine Learning*, vol. 80, 2018, pp. 274–283.

[32] I.-T. Chen and B. Sirkeci-Mergen, "A comparative study of autoencoders against adversarial attacks," in *Proc. of the Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2018, pp. 132–136.

[33] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[34] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[35] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach *et al.*, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] J. Rauber *et al.*, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," *arXiv preprint arXiv:1707.04131*, 2017.

[37] S. Mannor *et al.*, "The cross entropy method for classification," in *ICML*, 2005.

[38] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[39] S.-M. o. Moosavi-Dezfooli, "Universal adversarial perturbations," in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, 2017.

[40] N. Carlini et al., "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[41] F. Yu *et al.*, "Interpreting adversarial robustness: A view from decision surface in input space," *arXiv preprint arXiv:1810.00144*, 2018.

[42] A. S. Ross et al., "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[43] T. Pang et al., "Improving adversarial robustness via promoting ensemble diversity," *arXiv preprint arXiv:1901.08846*, 2019.

[44] A. Kurakin *et al.*, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.