Beyond cross-entropy: learning highly separable feature distributions for robust and accurate classification

Beyond cross-entropy: learning highly separable feature distributions for robust and accurate classification

Arslan Ali, Andrea Migliorati, Tiziano Bianchi, Enrico Magli
Department of Electronics and Telecommunications
Politecnico di Torino (Turin, Italy)
name.surname@polito.it

***Abstract*—Deep learning has shown outstanding performance in several applications including image classification. However, deep classifiers are known to be highly vulnerable to adversarial attacks, in that a minor perturbation of the input can easily lead to an error. Providing robustness to adversarial attacks is a very challenging task especially in problems involving a large number of classes, as it typically comes at the expense of an accuracy decrease. In this work, we propose the Gaussian class-conditional simplex (GCCS) loss: a novel approach for training deep robust multiclass classifiers that provides adversarial robustness while at the same time achieving or even surpassing the classification accuracy of state-of-the-art methods. Differently from other frameworks, the proposed method learns a mapping of the input classes onto target distributions in a latent space such that the classes are linearly separable. Instead of maximizing the likelihood of target labels for individual samples, our objective function pushes the network to produce feature distributions yielding high inter-class separation. The mean values of the distributions are centered on the vertices of a simplex such that each class is at the same distance from every other class. We show that the regularization of the latent space based on our approach yields excellent classification accuracy and inherently provides robustness to multiple adversarial attacks, both targeted and untargeted, outperforming state-of-the-art approaches over challenging datasets.**

I. Introduction

In recent years, deep neural networks have reached accuracy comparable with or even greater than that of humans in visual tasks such as recognizing traffic signs [1], handwritten digits [2], and faces [3]. Also, they have shown excellent performance at learning complex mappings [4] and addressing difficult classification tasks [5]. However, as their integration in contemporary society grows, they become ever more subject to the action of malicious *adversaries*.

In fact, despite the success of deep neural networks, many obstacles still hinder their use in fields where security is essential, such as systems for autonomous driving and medical diagnostics [6, 7]. A major threat is represented by adversarial perturbations, a set of techniques used to tamper with the inputs of a neural network. The modifications are often invisible to the human eye but may still be able to disrupt the algorithm operation and cause unexpected, undesired outputs. Malicious attackers could exploit such fallacies to cause malfunctions in systems, and the attack would be very hard to detect.

Although many countermeasures have been proposed, an effective defence mechanism against the broad spectrum of adversarial perturbations is not available yet. In particular, a downside of deep learning techniques is that the learned decision boundaries in the feature space are highly complex and non-linear [8]. Works addressing this specific problem [8, 9] concluded that most of the mass of the data points gathers close to the decision boundaries and this may strongly affect the robustness of the classifier against perturbations. Recent techniques tackling this problem can be found in [10] and [11], where logit regularization and curvature regularization methods are deployed as adversarial defense respectively, and also in [12] and [13], where theoretical insight is given on the effect of the use of unlabeled data and noise injection at inference time, respectively. At the same time, new techniques are also being developed to craft more successful adversarial attacks [14]. In the present paper, adversarial training is not considered as it entails the cost of generating and training on a substantial amount of additional input samples; moreover, adversarial training typically provides robustness against a specific type of attack, whereas we are interested in tackling the robustness problem with a more general approach.

In order to improve the robustness of a classifier in the presence of adversarial perturbation of the inputs, we propose a novel classifier design that goes beyond the cross-entropy loss function. The proposed method employs a new objective function enabling learning of features that maximize inter-class separation and decision variables exhibiting simple and well-defined distributions that are linearly separable in the latent space. The proposed objective function provides state-of-the-art classification while at the same time ensures robustness against adversarial attacks as it is. To correctly evaluate the robustness against adversarial examples, we follow the methodological foundations established in [15] and [16].

The resulting classifier employs simple threshold-based decisions in the regularized latent space. This design provides several benefits: on one hand, the accuracy is typically improved with respect to cross-entropy even in the case of no attacks. On the other hand, such classifier exhibits remarkably improved robustness against adversarial attacks; indeed, due to the uniformity of the distributions of the features in the latent space and the lack of a short path towards a neighboring deci-

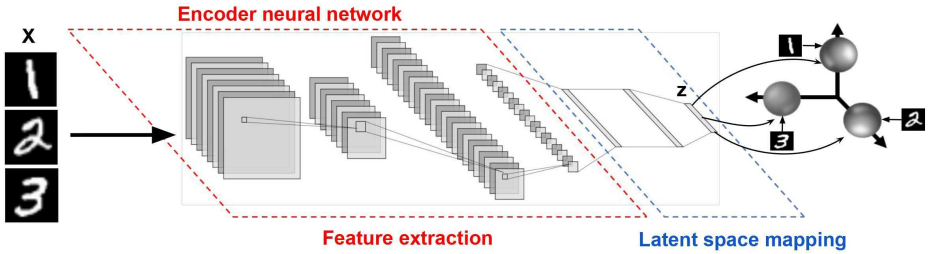


Fig. 1: The GCCS architecture takes input data and learns discriminative features that are mapped onto Gaussian target distributions in the latent space.

sion region, the attack strength must be much larger in order to generate a misclassification. Finally, the proposed method can be easily applied to an existing pre-trained cross-entropy based classifier, by continuing the training of the features and classification stage using our proposed loss function.

This paper presents a detailed assessment and analysis of the proposed method in several image classification problems, providing accuracy results on well-known datasets such as MNIST [17], FMNIST [18], SVHN [19], as well as more challenging datasets such as CIFAR10 and CIFAR100 [20]. In particular, we show that our loss function is inherently more robust than cross-entropy. We support our claim by following state-of-the-art robustness evaluation frameworks [15]. We validate our approach comparing it to state-of-the-art techniques for adversarial robustness and show that GCCS outperforms those methods under both targeted (PGD [21]) and untargeted attacks (JSMA [22], TGSM [23]).

II. Related works

The concept of adversarial perturbation was first introduced for spam email detection [24, 25]. In the following years, Szegedy et al. [26] showed how neural networks can easily be tricked into wrong classification if fed with specifically altered inputs produced considering the sign of the loss function gradients with respect to the inputs. In works such as [27–29] adversarial samples are used in the training phase as a particular form of data augmentation in order to improve robustness. However, such adversarial training does not prevent adversaries to effectively tamper with the final classification stage [16]. Rather, it has been proven that universal adversarial perturbations can be crafted so as to induce wrong classification with high probability independently of the used dataset [30], and also to generalize well over different network structures [23, 31, 32]. Recent theoretical works have also demonstrated that the robustness to adversarial attacks for a classification problem is bounded by limits that cannot be escaped by any classifier since they are dependent on the used datasets, the strength of the attack, and the way perturbations are measured [33].

The authors in [34] investigated how the effectiveness of adversarial attacks can transfer to models other than the targeted one, and they showed that adversarial examples that are generated to fool a specific model are likely to impact all the models that are trained on the same dataset. Also, [23] concluded that adversarial-generated images are misclassified even when printed out on paper and digitally re-acquired, proving that the phenomenon is relevant in both the digital and the physical domains. Further, [35] showed that deep learning methods for face recognition may wrongly classify faces when users are wearing ad-hoc designed adversarial glasses. Finally, [36] described a method for generating image patches to be placed on input target images in order to cause the neural network to output the desired class. This kind of attack is constructed and performed without knowledge of the targeted image, and it potentially allows the adversarial patch to be widely used with malicious intent after it is distributed over the Internet.

Several papers have also investigated defense techniques against attacks. The authors in [34] propose the input gradient regularization method, which is employed during the training phase to force the model to have smooth gradients. They hypothesize that a model trained with gradients that exhibits fewer extreme values is more resistant to adversarial perturbations and that its behavior in response to those attacks is also more easily interpretable. Moreover, [37] calculates instance-specific lower bounds on the norm of the input perturbation necessary to alter the decision of the classifier, providing a formal characterization of its robustness. The article also introduces the Cross-Lipschitz regularization functional which forces the differences of the classifier functions at the data points to be constant. Jakubovitz et al. [38] instead suggest a low-complexity regularization technique that uses the Frobenius norm of the Jacobian of the network, which is applied to already trained models as post-processing, robustness-improving step. In particular, while not being an active defence method, the proposed GCCS method ensures improved robustness against adversarial perturbations as it is.

If standard approaches focus on learning the classification boundaries, the proposed GCCS approach instead learns a mapping of the input classes onto target distributions in the latent space. Specifically, an encoder maps features of each class onto Gaussian distributions on a simplex for an arbitrary number of classes, maximizing inter-class separability. Other papers also propose to learn a mapping onto a regularized space, such as [39] and [40] that respectively introduce techniques based on adversarial and variational autoencoders. In [41] Stuhlsatz et al. present an approach to feature extraction that generalizes the classical Linear Discriminant Analysis (LDA) employing neural networks. The authors in [42] nonlinearly extend LDA by putting it on top of a deep neural network

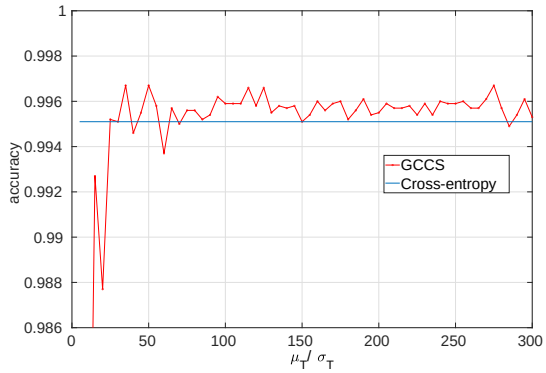


Fig. 2: Classification accuracy (%) for GCCS and cross-entropy on MNIST with ResNet-18.

and maximize the eigenvalues of LDA on the last hidden representation. The primary objective of most discriminant analysis methods is dimensionality reduction [43]. One of the shortcomings of these methods is that they tend to maximize the distance of the classes that are already well separated, at the expense of poorly-separated neighboring classes, leading to a nonhierarchical pattern in terms of inter-class separability. Another relevant work is RegNet [44], a deep learning technique for biometric authentication that deals with the one-vs-all classification problem of separating authorized users from non-authorized ones. This technique regularizes a two-dimensional latent space through a loss function based on a simplified equation for the Kullback–Leibler divergence; however, this approach is not suitable for high-dimensional classification problems as addressed by GCCS.

III. Proposed Method

The proposed method is based on the architecture shown in Fig. 1. Labeled training data $\mathbf{X}$ for a D-class classification problem is given as input to a neural network that is composed of a feature extractor and a latent space mapper. The goal of the feature extractor is to learn nonlinear transformations from arbitrary data distributions and extract distinctive and highly separable features. The latent space mapper consists of one or more fully connected layers with the goal of mapping the output $\mathbf{z}$ onto specific target distributions in a D-dimensional latent space (i.e. as many dimensions as the number of classes); no non-linear activation function is employed in the last layer of the mapper. It is worth noting that the proposed method does not depend on a specific feature extraction architecture, so existing state-of-the-art architectures can be used for this task.

In order to achieve the desired target, the proposed method needs to define three main components: a target model for the distribution of features in the latent space; a loss function to achieve that distribution; finally, a decision rule. The details are as follows.

A. Model for the target distributions

GCCS aims to learn the most discriminative features and maximize the inter-class separability by finding a nonlinear projection of high-dimensional observations onto a lower-dimensional space. This is obtained by regularizing the latent space to D different statistical distributions, where D is the number of classes the data belongs to. Let us first define the desired target distribution $\mathbb{P}_i$ for class $\mathcal{C}_i$, $i = 1, \dots D$, as a D-variate Gaussian distribution, i.e. $\mathbb{P}_i = \mathcal{N}(\boldsymbol{\mu}_{Ti}, \boldsymbol{\Sigma}_T)$, with $\boldsymbol{\mu}_{Ti} = \mu_T \mathbf{e}_i$ and $\boldsymbol{\Sigma}_T = \sigma_T^2 \mathbb{I}_D$, where $\mathbf{e}_i$ is the ith standard unit vector and $\mathbb{I}_D$ is the $D \times D$ identity matrix. μ_T and σ_T are user-defined parameters that are related to inter-class separation and are discussed later in the manuscript. Here, it should be noted that in order to have separable distributions we should have $\mu_T/\sigma_T > \sqrt{2D}$, otherwise as D grows the classes will inevitably mix.

Since each distribution $\mathbb{P}_i$ has mean value proportional to $\mathbf{e}_i$, the statistical distributions are centered on the vertices of a regular $(D-1)$-simplex at $\mu_T \mathbf{e}_i$, as shown in Fig. 1. This target model has several advantages. First, this choice guarantees that each class is at the same distance from all other classes. Due to the uniformity of the feature distributions in the latent space and the consequent lack of a short path, the attack strength must be much larger in order to generate a misclassification, leading to improved robustness. Moreover, since the distributions are Gaussian, the decision boundaries are straightforward to compute. This is in contrast with the typical behavior of neural networks, which tend to yield very complex boundaries, and it promotes accuracy as well as adversarial robustness.

B. Loss function

In order to train the network, we need to introduce a loss function that allows us to minimize a suitable distance metric between the distributions of the output latent variables and the target distributions.

Let us refer to the output of the encoding neural network as $\mathbf{z} = H(\mathbf{x})$, where $[z_1, \dots, z_D] \in \mathbb{R}^D$ indicates the latent mapping, and $\mathbf{x} \in \mathbb{R}^n$ denotes the input data belonging to D different classes. The goal is to learn an encoding function of the input $\mathbf{z} = H(\mathbf{x})$ such that $\mathbf{z} \sim \mathbb{P}_i$ if $\mathbf{x} \in \mathcal{C}_i$.

During the training phase, the network is given as input a batch of samples $\mathbf{X} \in \mathbb{R}^{b \times n}$, where b is the batch size, and it computes the encoded outputs $\mathbf{Z} \in \mathbb{R}^{b \times D}$. We are interested in their first and second order statistics, which can be estimated as sample mean $\boldsymbol{\mu}_{Oi}$ and sample covariance $\boldsymbol{\Sigma}_{Oi}$ for each class. Considering that the target statistics are known and the sample statistics for the batch have been computed, we can proceed to define a suitable loss to measure how far the distributions are from each other. More in detail, we employ the *Kullback-Leibler* divergence (KL).

For the sample distribution of any class $\mathcal{C}_i$, the KL divergence with respect to the Gaussian target distribution can be written as:

$$\mathcal{L}_i = \log \frac{|\boldsymbol{\Sigma}_T|}{|\boldsymbol{\Sigma}_{Oi}|} - D + \operatorname{tr}(\boldsymbol{\Sigma}_T^{-1} \boldsymbol{\Sigma}_{Oi}) + (\boldsymbol{\mu}_{Ti} - \boldsymbol{\mu}_{Oi})^\intercal \boldsymbol{\Sigma}_T^{-1} (\boldsymbol{\mu}_{Ti} - \boldsymbol{\mu}_{Oi}) \quad (1)$$

Method	MNIST ResNet-18	FMNIST ResNet-18	SVHN ResNet-18	CIFAR-10 ResNet-18	CIFAR-10 Shake-Shake-96	CIFAR-100 Shake-Shake-112
GCCS - regular training	99.58	92.69	94.20	82.97	96.19	76.53
GCCS - fine tuning	**99.64**	**93.83**	**95.58**	**81.52**	**97.06**	**77.48**
No Defense - cross-entropy	99.35	91.91	94.12	78.59	95.78	76.30
Jacobian Reg. - regular training [38]	98.99	91.79	94.11	70.09	-	-
Jacobian Reg. - fine-tuning[38]	98.53	92.43	93.54	82.09	-	-
Input Gradient Reg. - regular training [34]	97.98	88.45	93.77	78.32	96.50	74.89
Input Gradient Reg. - fine-tuning [34]	99.11	92.55	93.17	76.15	96.90	75.68
Cross Lipschitz regular training [37]	96.78	92.54	91.42	80.10	-	-
Cross Lipschitz - fine-tuning [37]	98.77	92.41	93.50	79.39	-	-

TABLE I: Maximum test accuracy obtained through *regular training* vs *fine-tuning* over different benchmark datasets with different competing techniques in the case in which no adversarial attack is performed.

We consider the cumulative loss $\mathcal{L} = \sum_{i=1}^{D} \mathcal{L}_i$. This loss reaches its minimum when the sample statistics of the D encoded distributions match the target ones. However, for a small batch size, it can be difficult to control the behavior of the tails of the obtained distributions relying only on KL. Hence, we also consider the Kurtosis $\mathcal{K}_{i,j}$, [45] of the jth component of the ith target distribution, defined as $\mathcal{K}_{i,j} = \left(\frac{z_{i,j} - \boldsymbol{\mu}_{Oi,j}}{\sigma_{Oi,j}}\right)^4$.

In the case of multiple i.i.d. univariate normal distributions such as those we are enforcing at training, the target Kurtosis for each class is $\mathcal{K}_{i,j} = 3$. This can be added to the cumulative loss, obtaining the loss $\mathcal{L}^{\mathrm{GCCS}}$ as follows:

$$\mathcal{L}^{\mathrm{GCCS}} = \sum_{i=1}^{D} \left[\mathcal{L}_i + \lambda(\mathcal{K}_i - 3)\right], \tag{2}$$

where $K_i = 1/D \sum_j K_{i,j}$ and λ determines the strength of the Kurtosis term and is set to $\lambda = 0.2$.

C. Decision Rule

Once the preconditions are fulfilled, GCCS allows to define optimal decision boundaries in the resulting latent space. For the given target distributions, the optimal boundaries are obtained by partitioning the space into Voronoi regions such that all points in a region are closer to the respective centroid (the mean vector $\boldsymbol{\mu}_{Ti}$) than to any other centroid in the $(D-1)$-simplex. The resulting decision rule consists of computing the distance of the feature point from all centers and choose the class with the minimum distance. To determine which class a test image belongs to, the following decision rule is employed:

$$\widehat{y} = \arg\max_i z_i, \tag{3}$$

which returns the index of the predicted class for the test image.

IV. Experiments

A. Datasets and Training Parameters

The performance of classifiers trained using the GCCS loss was tested on MNIST [17], FMNIST [18], SVHN [19], CIFAR-10 and CIFAR-100 [20]. For less complex datasets such as MNIST, FMNIST, and SVHN, the experiments were conducted using ResNet-18 [46] as the feature extraction network. For the more challenging CIFAR-10 and CIFAR-100 datasets, the Shake-Shake-96 and Shake-Shake-112 [47] regularization networks have been employed respectively, using a widen factor equal to 6 for the former and 7 for the latter. The encoder's last layer is followed by a fully-connected layer that outputs a vector with dimension D. We trained each network for a total of 1800 epochs. For better network convergence, we employed cosine learning rate decay [48] with an initial value of 0.01 as well as weight decay with a rate set to 0.001. Finally, dropout regularization [49] with a 0.8 keep probability value was applied to all the fully connected layers in the network.

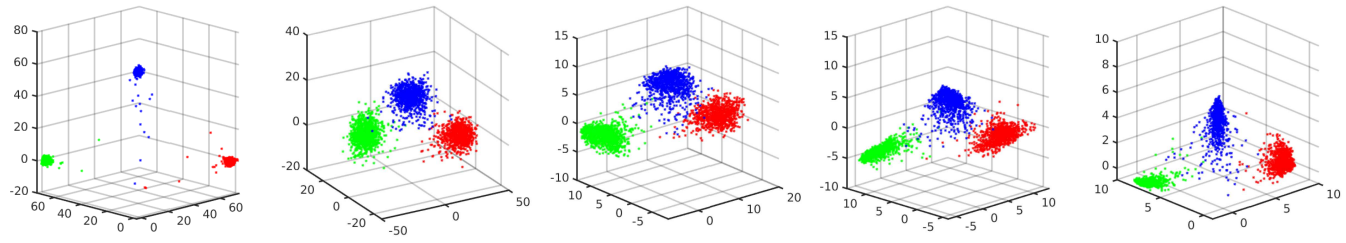
1) Target Distributions Parameters: In this section, we perform an experiment to explore the behavior of the target distributions for different mean and variance values. Since we fix the mean μ_T and variance σ_T values for the target distributions so that they are centered on the vertices of a regular $(D-1)$-simplex, the only parameter affecting our design is the μ_T/σ_T ratio, i.e., how far apart the distributions are with respect to the chosen variance.

In this experiment we set $\sigma_T = 1$, so that the target distributions are $\mathbb{P}_i = \mathcal{N}(\mu_T \mathbf{e}_i, \mathbb{I}_D)$; then, we compute the classification accuracy as a function of $\mu_T \in [0.5, 300]$. Fig. 2 shows the accuracy as a function of μ_T/σ_T for MNIST-10 dataset. It can be observed that in the $\mu_T \geq 20$ region the accuracy is even higher than that obtained with the traditional cross-entropy loss.

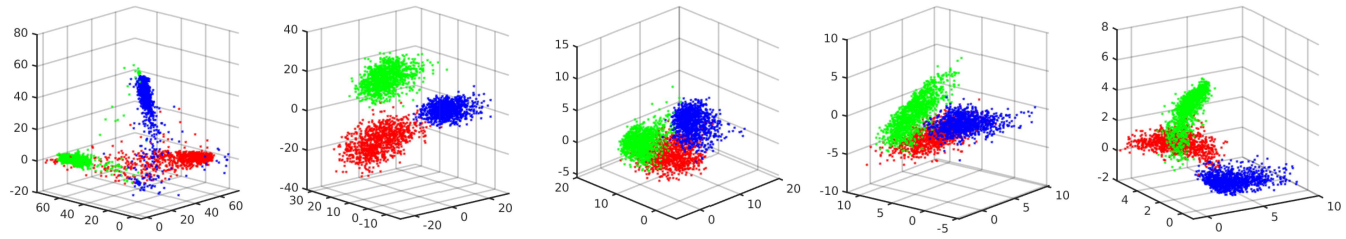
In the following, we choose $\mu_T = 70$ and $\sigma_T = 1$. This choice ensures that we operate in that region, and also that the target distributions are sufficiently far apart from each other.

B. Classification accuracy

As a first experiment, we compared the classification accuracy of GCCS with that obtained by an equivalent network trained with cross-entropy loss (no defense) and with state-of-the-art defense techniques such as Jacobian Regularization [38], Input Gradient Regularization [34], and Cross Lipschitz regularization [37], in the case in which no adversarial attack is performed. As shown in Table I, GCCS yields high classification accuracy both when the networks are trained from scratch (*regular training*) and when they are first trained using regular cross-entropy loss and then fine-tuned with either GCCS loss or the other defense techniques (*fine-tuning*). In particular, Table I shows that the proposed technique outperforms the standard cross-entropy loss and other existing approaches [38], [34], and [37] over the considered datasets. In more detail, it can be seen that other techniques generally cause a small decrease in classification accuracy with respect to the standard cross-entropy loss function, whereas GCCS provides



(a) GCCS (b) No defense (c) Jacobian Reg. (d) Input Reg. (e) Cross Lipschitz



(f) GCCS (g) No defense (h) Jacobian Reg. (i) Input Reg. (j) Cross Lipschitz

Fig. 3: **(a-e)** Visual representation of latent space output distributions on MNIST for regular training in the case that no adversarial attack is applied. For better visualization of the separability, only three classes are shown, and an appropriate scale is used for each plot. (a) GCCS; (b) standard cross-entropy; (c) Jacobian Regularization [38]; (d) Input Gradient Regularization [34]; (e) Cross Lipschitz Regularization [37]. **(f-j)** Visual representation of latent space output distributions on MNIST for TGSM (5 steps, $\epsilon = 2e^{-3}$) is applied. For better visualization, only three classes are shown. (f) GCCS; (g) standard cross-entropy; (h) Jacobian Regularization [38]; (i) Input Gradient Regularization [34]; (j) Cross Lipschitz Regularization [37].

an improvement in testing accuracy, especially for challenging datasets such as CIFAR-10 and CIFAR-100.

The higher classification accuracy yielded by GCCS is due to the high separability of the target distributions in the latent space, as opposed to the other methods. To better highlight this, we refer to Fig. 3 in which the output distributions for three different MNIST classes [0, 1 and 9] are reported. Looking at Fig. 3-a against Fig. 3-b, Fig. 3-c, Fig. 3-d, and Fig. 3-e, one can immediately observe that the output distributions of the three classes are less spread out and more separated than the other cases.

Also, Fig. 3 shows that GCCS provides lighter distribution tails, compared to the other methods.

C. Robustness Evaluation

In this section, we evaluate how the classification accuracy of GCCS and the other competing techniques degrades under both targeted attacks (TGSM, JSMA) and non-targeted attack (PGD). The accuracy is evaluated as a function of a tunable parameter ϵ that indicates how strong the applied attack is.

Namely, the noise vector $\mathbf{n}$ added by the attack to the input signal $\mathbf{x}$ satisfies $\|\mathbf{n}\|_\infty / \|\mathbf{x}\|_\infty \leq \epsilon$.

1) Non-targeted Attacks: We start by evaluating the performance of all methods when subjected to the non-targeted PGD attack on the MNIST, SVHN, CIFAR-10, and CIFAR-100 datasets. Projected Gradient Descent (PGD) [21], is an iterative version of FGSM in which noise is added in multiple steps. In particular, PGD is the strongest adversarial attack that exploits first-order local information about the trained model. In this work, for PGD we apply a 5-iterations attack, i.e. PGD-5 as done in [33, 50, 51].

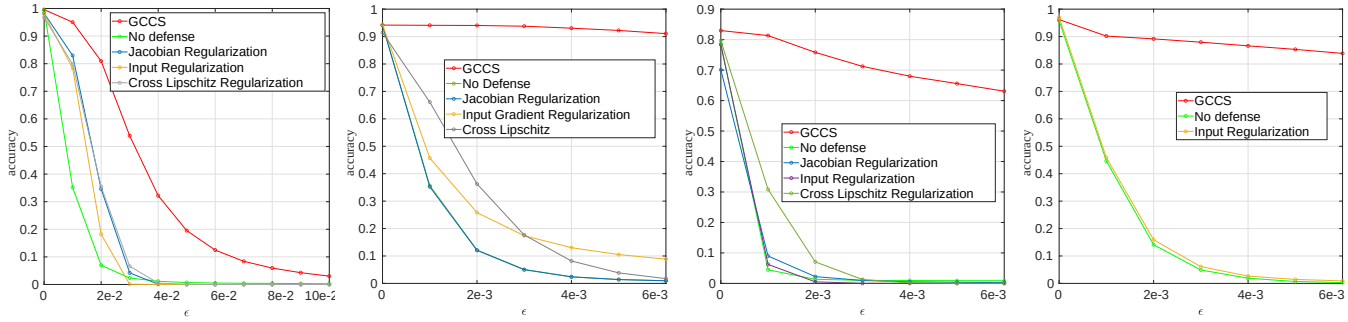
For MNIST, we set $0 \leq \epsilon \leq 10e^{-2}$, while for SVHN, CIFAR-10, and CIFAR-100 we set $0 \leq \epsilon \leq 6e^{-3}$, since MNIST is, in general, a less challenging dataset. As illustrated in Fig. 4, GCCS outperforms by a large amount the competing approaches on all the considered datasets. Our approach proves to be much more robust than the others, especially for stronger attacks. The performance gap is particularly evident in the case of PGD, which is indeed the strongest adversarial attack utilizing the local first-order network information.

2) Targeted Attacks: We also consider targeted adversarial attacks such as TGSM and JSMA. Similarly to Sec. IV-C1, we present curves of the classification accuracy against the attack strength ϵ.

TGSM Attack: In TGSM [23] the input samples are perturbed by adding noise in the direction of the negative gradient with respect to a selected target class. Fig. 5 presents the results for TGSM-5, a 5-iterations TGSM attack, over the MNIST, SVHN, CIFAR-10, and CIFAR-100 datasets. In this attack, the targeted output class is y_{l+1} when the true class is y_l.

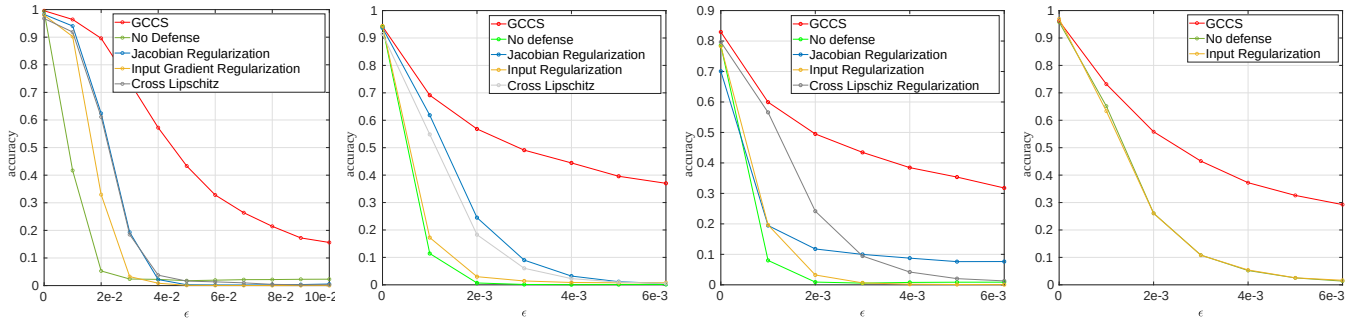
It can be observed from Fig. 5 that GCCS yields significantly higher performance compared to the other methods, throughout different datasets and with different attack strength ϵ. In order to gain a better understanding of why the proposed method works much better than the others, in Fig. 3 we show a visual representation of the target distributions in the latent space *after* the TGSM-5 attack $\epsilon = 2e^{-3}$ has been performed.

Fig. 3-g shows clearly the effectiveness of the attack when



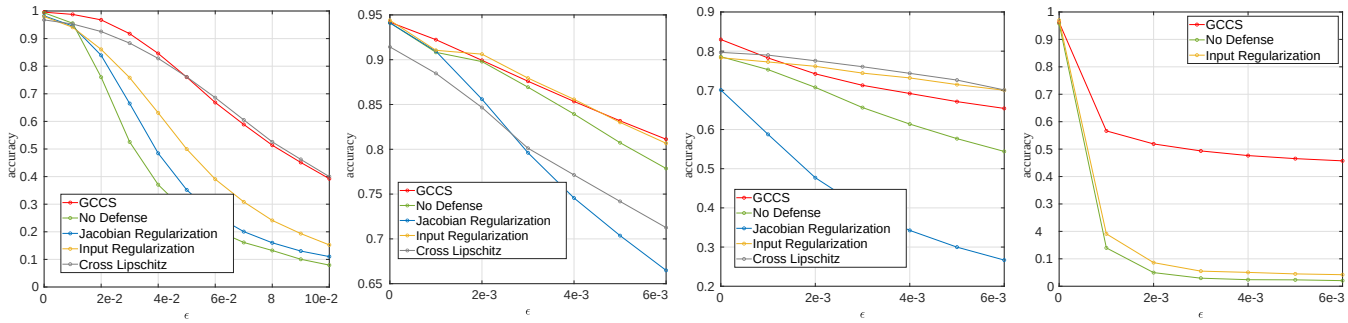
(a) MNIST@ ResNet-18 (b) SVHN@ ResNet-18 (c) Cifar10@ ResNet-18 (d) Cifar10@ Shake-Shake

Fig. 4: Test accuracy for PGD (5 steps) attack on (a) ([MNIST, ResNet-18]); (b) ([SVHN, ResNet-18]); (c) ([CIFAR-10, ResNet-18]); (d) ([CIFAR-10, Shake-Shake-96]) for different values of ϵ.



(a) MNIST@ ResNet-18 (b) SVHN@ ResNet-18 (c) Cifar10@ ResNet-18 (d) Cifar10@ Shake-Shake

Fig. 5: Test accuracy when applying the TGSM attack (5 steps) for (a) ([MNIST, ResNet-18]) ; (b) ([SVHN, ResNet-18]); (c) ([CIFAR-10, ResNet-18]) (d) ([CIFAR-10, Shake-Shake-96]), for different values of ϵ.



(a) MNIST@ ResNet-18 (b) SVHN@ ResNet-18 (c) Cifar10@ ResNet-18 (d) Cifar10@ Shake-Shake

Fig. 6: Test accuracy when applying the JSMA attack (200 steps, 1 pixel) on (a) ([MNIST, ResNet-18]); (b) ([SVHN, ResNet-18]); (c) ([CIFAR-10, ResNet-18]); (d) ([CIFAR-10, Shake-Shake-96]), for different values of ϵ.

no defense mechanism is employed, in the sense that the output distributions are shifted so as to replace the output distribution of the next class. Fig. 3-h, Fig. 3-i, and Fig. 3-j report the output distributions under TGSM in the case of Jacobian, Input Gradient, and Cross-Lipschitz regularization respectively, showing that, despite the defense mechanism, the distributions still tend to move their position in the latent space towards the adjacent classes, causing a very important drop in classification accuracy as seen in Fig. 5. In the GCCS case instead (Fig. 3-f), even if the tails of the output distributions become heavier, their positions are not swapped with the neighboring classes, allowing for better separability and hence improved classification accuracy and robustness.

JSMA Attack: The other targeted attack we consider is JSMA [22], which consists in iteratively computing the Jacobian matrix of the network function to form a saliency map; this map is used at every iteration to choose which pixels to tamper with so that the likelihood of changing the output class towards a selected one is increased. In our case, we consider JSMA-200 with a 1-pixel saliency map. Similarly to the TGSM case, Fig. 6 shows the classification accuracy for increasing attack strength ϵ. The proposed method confirms its robustness even to JSMA attack, achieving better robustness than other methods especially on the challenging CIFAR-10 dataset.

V. Conclusions

We have presented an approach that goes beyond cross-entropy, employing a loss function that promotes class separability and robustness by learning a mapping of the decision variables onto Gaussian distributions. Our work was motivated by the idea that mapping the centroids of the distributions on the vertices of a simplex could lead to the uniformity of the feature distributions in the latent space and the lack of a short path towards a neighboring decision region. Experiments on different multi-class datasets show excellent performance of the classifiers trained using the GCCS loss both in terms of accuracy and robustness of the classifier against adversarial attacks, outperforming existing state-of-the-art methods, both when used to train a network from scratch and when applied as a fine-tuning step on pre-trained networks. The performance is analyzed both for targeted and non-targeted adversarial attacks. We have shown that regularizing the latent space onto target distributions significantly increases the robustness against adversarial perturbations. Indeed, an analysis of the distributions in the latent space for the proposed GCCS method shows that the different classes tend to remain separated even in the presence of targeted attacks, whereas a similar attack strength invariably mixes the distributions achieved by competing methods.

VI. Acknowledgment

This work results from the research cooperation with Sony R&D Center Europe Stuttgart Laboratory 1.

References

[1] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[2] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066, 2013.

[3] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[4] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[7] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[8] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Classification regions of deep neural networks. *arXiv preprint arXiv:1705.09552*, 2017.

[9] A. Fawzi, S. Moosavi-Dezfooli, and P. Frossard. The robustness of deep networks: A geometrical perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, Nov 2017.

[10] Cecilia Summers and Michael J Dinneen. Improved adversarial robustness via logit regularization methods. *arXiv preprint arXiv:1906.03749*, 2019.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019.

[12] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[13] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, pages 11838–11848, 2019.

[14] Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2253–2260, 2019.

[15] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[16] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[19] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[22] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[24] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

[25] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan

Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[27] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[28] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[31] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[32] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[33] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

[34] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[35] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016.

[36] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[37] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.

[38] Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.

[39] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[41] Andre Stuhlsatz, Jens Lippel, and Thomas Zielke. Feature extraction with deep neural networks by a generalized discriminant analysis. *IEEE transactions on neural networks and learning systems*, 23(4):596–608, 2012.

[42] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015.

[43] Jieping Ye and Shuiwang Ji. Discriminant analysis for dimensionality reduction: An overview of recent developments. *Biometrics: Theory, Methods, and Applications. Wiley-IEEE Press, New York*, 2010.

[44] Matteo Testa, Arslan Ali, Tiziano Bianchi, and Enrico Magli. Learning mappings onto regularized latent spaces for biometric authentication. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[45] DN Joanes and CA Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[48] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[50] Tianhang Zheng, Changyou Chen, and Kui Ren. Is pgd-adversarial training necessary? alternative training via a soft-quantization network with noisy-natural samples only. *arXiv preprint arXiv:1810.05665*, 2018.

[51] Todor Davchev, Timos Korres, Stathi Fotiadis, Nick Antonopoulos, and Subramanian Ramamoorthy. An empirical evaluation of adversarial robustness under transfer learning. *arXiv preprint arXiv:1905.02675*, 2019.