Optimal transport as a defense against adversarial attacks

Quentin Bouniot, Romaric Audigier, Angelique Loesch

Optimal Transport as a Defense Against Adversarial Attacks

Quentin Bouniot, Romaric Audigier, Angelique Loesch
Université Paris-Saclay, CEA, List,
F-91120, Palaiseau, France
{quentin.bouniot, romaric.audigier, angelique.loesch}@cea.fr

***Abstract*—Deep learning classifiers are now known to have flaws in the representations of their class. Adversarial attacks can find a human-imperceptible perturbation for a given image that will mislead a trained model. The most effective methods to defend against such attacks trains on generated adversarial examples to learn their distribution.**

Previous work aimed to align original and adversarial image representations in the same way as domain adaptation to improve robustness. Yet, they partially align the representations using approaches that do not reflect the geometry of space and distribution. In addition, it is difficult to accurately compare robustness between defended models. Until now, they have been evaluated using a fixed perturbation size. However, defended models may react differently to variations of this perturbation size.

In this paper, the analogy of domain adaptation is taken a step further by exploiting optimal transport theory. We propose to use a loss between distributions that faithfully reflect the ground distance. This leads to SAT (Sinkhorn Adversarial Training), a more robust defense against adversarial attacks.

Then, we propose to quantify more precisely the robustness of a model to adversarial attacks over a wide range of perturbation sizes using a different metric, the Area Under the Accuracy Curve (AUAC). We perform extensive experiments on both CIFAR-10 and CIFAR-100 datasets and show that our defense is globally more robust than the state-of-the-art.

I. Introduction

Deep learning models have become the state-of-the-art in many computer vision tasks. Yet, recent work [1][2][3] has shown that deep neural network can be easily fooled by *adversarial examples*, a human-imperceptible noise added to an image. Misclassification of adversarial images shows that neural network still have issues generalizing their class representation. These adversarial attacks represent a major issue for the widespread adoption of deep learning in practical applications.

Currently, training on adversarial examples generated is the most effective defense against adversarial attacks [2][4][5], which means learning the distribution of adversarial examples. This distribution is far from the original distribution as they lie in low probability regions [6], *i.e.* they are assigned low probability in the original distribution. Defending against adversarial attacks can be seen as aligning the representations of adversarial and original images, similar to a domain adaptation problem [7][8]. Previous work aligns the moments of the distributions or explicitly constrain the cluster of classes [8][9]. Yet, aligning the moments is an imprecise approach to define a distance. It reflects neither the geometry of space nor that of the distribution and can lead to instabilities during training. As for the other case, enforcing images to their class clusters does not take into account the whole distribution. These approaches result in a partial alignment of the representations.

Besides, defenses are tested against a fixed adversarial perturbation size to compare robustness to attacks. However, defended models may react differently to changes in the size of the perturbation. Some may be more robust at low perturbations and others at high perturbations, differences that will not be visible when evaluating with a fixed size.

In this article, we propose to directly constrain the representations of adversarial and original images to remain close after attack using a least-effort approach. To do this, we propose to use a discrepancy between distributions based on the theory of optimal transport. This allows us to consider the whole distribution with a loss that reflects geometric properties rather than an alignment of moments. We minimize the ground distance between representations to align and bring closer the distributions.

Then, we propose a metric to measure the robustness of a model to adversarial attacks. To do this, we use a wide range of sizes and take into account the evolution of performance over this range. Thus, we can compare the behaviour of different models for several perturbation sizes at the same time.

Our contributions can be summarized as follow :

- We propose **SAT** (Sinkhorn Adversarial Training), a new defense against adversarial attacks using a **Sinkhorn Divergence**, a loss based on the theory of optimal transport.
- We propose a **different evaluation metric for robustness** against adversarial examples: **Area Under the Accuracy Curve (AUAC)**.

In the following, we introduce related works on adversarial attacks and defenses in Section II. After preliminaries on the theory of optimal transport, we propose our defense in Section III and our evaluation metric in Section IV. Finally, we compare our defense with the state-of-the-art in Section V.

II. Related Work

In this section, we present an overview of previous work on adversarial attacks and defenses.

A. Adversarial Attacks

After Szegedy *et al.* [1] shed light on the effect of their attack on a standard classifier, several more efficient ways of computing adversarial examples have been discovered. They can be grouped into three families:

- *Unbounded attacks* [10][11][12] solve a constrained optimization problem, and seek the smallest perturbation that will misclassify a given image. The perturbation size can be different for two images. These attacks are the most effective since the perturbation size is not bounded. Given enough iterations and time, they always find an adversarial example.
- *Bounded attacks* [2][4][5][13] use the gradients of the model for a given image and perform several steps of projected gradient descent. The size of the perturbation found is fixed and the same for each images.
- *Gradient Reconstitution attacks* [14][15][3] search the input space by following an approximation of the Jacobian or by a random walk. These greedy attacks are time consuming but they can bypass defenses based on obfuscating gradients [16].

Having efficient attacks, *i.e.* that find quickly good adversarial examples, is useful to improve the defense. Indeed, when generating adversarial examples for training, the use of efficient attacks leads to better examples and more robust models. Furthermore, efficient attacks provide an useful evaluation of defenses.

To this end, we consider in our experiments the most efficient Bounded and Unbounded attacks, namely Projected Gradient Descent (PGD, a.k.a. Basic Iterative Method or Iterative Fast Gradient Sign Method) [4][5] and Trust-Region Attack (TR) [11].

B. Adversarial Defenses

There have been multiple attempts to protect against adversarial examples or to increase robustness of deep learning models.

a) Obfuscating gradients: A first type of approaches performs preprocessing [17][18][6] on images to detect or to mitigate the effect of the perturbation. Similarly, some methods complexify models [19][20] to make attacks more difficult to achieve. However, all these methods lead to obfuscated gradients and provide limited robustness. They are still vulnerable to more specific adversarial attacks [16][21].

b) Adversarial training: As previously stated, the most successful defense against adversarial attacks is *adversarial training*. Goodfellow *et al.* [2] considered a mixed batch of adversarial examples (generated during training) and original images whereas Madry *et al.* [5] proposed to use only adversarial examples in the batch. Tramèr *et al.* [7] introduced several pre-trained models to generate different adversarial examples during training. Adversarial training aims to learn a single distribution containing both original and adversarial versions of training samples.

In contrast, Song *et al.* [8] framed the defense as a domain adaptation problem. They considered the original and adversarial training distributions as distinct and aimed to bring their representations in the logit space closer. They improved adversarial training by aligning the means with a Maximum Mean Discrepancy (MMD) [22] norm and covariance with a Correlation Alignment (CORAL) [23]. Similarly, Mustafa *et al.* [9] explicitly ensured a large distance between class prototypes and a small intra-class distance.

Bringing the distributions closer in the logit space seems to have a significant effect for robustness. However aligning the moments of representations results in an unstable training and does not faithfully reflect the distance between the two distributions. In practice, this is reflected by artifacts (*vanishing* or *exploding gradients*) next to the extreme points of the distribution [24]. Constraining the distance between and within clusters of classes does not take into account the distribution of images as a whole.

Based on the theory of optimal transport, we propose to use Sinkhorn Divergences to consider the discrepancy between adversarial and original representations integrally and with more accurate geometric properties. We aim to minimize the ground distance between the representations.

III. Sinkhorn Adversarial Training

In this section, we describe our training protocol to increase robustness to adversarial examples. First, we introduce necessary background on Optimal Transport.

A. Optimal Transport

The theory of Optimal Transport aims to find the minimal cost (in terms of distance) to move simultaneously several items (or a continuous distribution of items in the most extreme case) from one configuration onto another. In particular, it can be used for computing distances between probability distributions.

In practice, solving the optimal transport problem is *costly* and suffers from the *curse of dimensionality*. Therefore, we consider its *entropic regularization* [25][26]:

$$W_{2,\sigma}(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \int_{\Omega\times\Omega} ||x-y||_2^2 d\pi(x,y) + \sigma \mathrm{KL}(\pi|\alpha\otimes\beta)$$

with $\mathrm{KL}(\pi|\alpha\otimes\beta) = \int_{\Omega\times\Omega} \log\left(\frac{d\pi}{d\alpha d\beta}\right) d\pi$, the Kullback-Leibler divergence. α and β are two probability measures with finite second moment, $\Pi(\alpha,\beta)$ is the set of probability measures over the product set $\Omega\times\Omega$ with marginals α and β and $\sigma > 0$ is the entropy regularization parameter. With this regularization, the problem can be solved efficiently on GPU [27].

Finally, *Sinkhorn Divergences* [28][24] are defined as

$$S_\sigma(\alpha,\beta) = W_{2,\sigma}(\alpha,\beta) - \frac{1}{2}W_{2,\sigma}(\alpha,\alpha) - \frac{1}{2}W_{2,\sigma}(\beta,\beta).$$

S_σ interpolates between an *Optimal Transport loss* and an *MMD loss* depending on σ. The effect of the entropy will be studied in Section V-C. We will use this divergence to align the original and adversarial representations.

Algorithm 1 SAT

Input : Model f, training set $\mathbf{D}^{tr}$ and labels $\mathbf{Y}^{tr}$, number of epochs T, sinkhorn entropy σ

Output : Defended model f

for $t = 1$ to T **do**
  for batch $\mathbf{D}_b^{tr} \in \mathbf{D}^{tr}$ and labels $\mathbf{Y}_b^{tr} \in \mathbf{Y}^{tr}$ **do**
    Use the current state of f to generate an adversarial batch of images $\hat{\mathbf{D}}_b^{tr}$;
    Compute $\mathcal{L}_{SAT_\sigma}(\mathbf{D}_b^{tr}, \hat{\mathbf{D}}_b^{tr}, \mathbf{Y}_b^{tr}, f)$ (Eq. 1) and update parameters of f by backpropagation;
  end for
end for

B. Training with a Sinkhorn Divergence

Given a training set $\mathbf{D}^{tr} = \{x_i^{tr}\}$ with associated labels $\mathbf{Y}^{tr}$, adversarial examples $\hat{\mathbf{D}}^{tr} = \{\hat{x}_i^{tr}\}$ are generated to create a large shift in the model representations from a small perturbation. We aim to reduce this shift by using optimal transport.

Our *SAT* (Sinkhorn Adversarial Training) combines adversarial training with a Sinkhorn Divergence as a loss between the distributions of original and adversarial representations. We constrain not only the moments, but the totality of the distributions by making sure to respect the underlying geometrical properties in the representation space. We minimize the ground distance, *i.e.* the L_2 distance, between representations of original images and their adversarial counterpart to bring them closer.

The loss function used to train a defended classifier with SAT is

$$\mathcal{L}_{SAT_\sigma}(\mathbf{D}^{tr}, \hat{\mathbf{D}}^{tr}, \mathbf{Y}^{tr}, f) = \mathcal{L}_{CE}(\hat{\mathbf{D}}^{tr}, \mathbf{Y}^{tr}, f) + S_\sigma(f(\mathbf{D}^{tr}), f(\hat{\mathbf{D}}^{tr})), \quad (1)$$

with $\mathcal{L}_{CE}$ the cross entropy used for classification, $f(\mathbf{D}^{tr}) = \{f(x_i^{tr})\}$ and $f(\hat{\mathbf{D}}^{tr}) = \{f(\hat{x}_i^{tr})\}$ the representations of original and adversarial training images respectively. The full training algorithm is described in Algorithm 1.

During training, for each new batch of images, we generate adversarial examples by attacking the current state of the model being trained with a state-of-the-art adversarial attack (*e.g.* PGD). Then, we use the original and adversarial batch of images as separate distributions to compute the loss $\mathcal{L}_{SAT_\sigma}$ using Equation 1. The size of the perturbation ϵ_{train} used by the attack during training is fixed.

Having defined our defense, we now want to evaluate its robustness against adversarial examples. We want to assess performance both on original data and on varying degrees of adversarial data (within the limit of the definition). Currently, defenses are evaluated against an attack with a given perturbation size ϵ. In general, the perturbation size used for evaluation is the same as the one chosen for adversarial training. However, this protocol does not allow an overall estimation of the robustness of a model.

IV. Measuring robustness

The section below raises concerns on the current evaluation protocol against adversarial examples, and proposes a different and more accurate metric.

A. Defining perturbation size

As pointed out in the introduction to this paper, performance of models significantly differs when the size of the perturbation ϵ varies.

Each attack has a different way to generate adversarial examples and target a suitable loss function $\mathcal{L}$ (*e.g.* cross-entropy). The perturbation size ϵ can be variable between attacked images or not. It is defined as the difference for a given norm (usually L_∞ or L_2) between an image x with label y and the corresponding adversarial example $\hat{x}$ found by a given attack: $\epsilon = ||x - \hat{x}||$.

On the one hand, with a *Bounded attack*, *e.g.* PGD [4][5], the size of the perturbation is explicitly defined and remains constant between attacked images:

$$\hat{x}_0 = x + \eta$$
$$\hat{x}_{n+1} = \Pi_x^\epsilon \left(\hat{x}_n + \alpha \cdot \text{sign}(\nabla_{\hat{x}_n} \mathcal{L}(\hat{x}_n, y, f)\right)$$

with ϵ the size of the perturbation, η a random initial noise, $\hat{x}_n$ the resulting adversarial image after the n-th iteration and α the iteration step size. Π_x^ϵ is the clip function, which ensures that $||\hat{x}_n - x|| \leq \epsilon$ and that $\hat{x}_n$ is a valid image, *i.e.* the pixels are in the range $[0, 1]$.

On the other hand, with other attacks, *e.g.* TR attack [11], the size of the perturbation can be variable between images. The attack find the smallest perturbation that fool the model for a given image:

$$\arg\min_{||\Delta x||} \arg\max \mathcal{L}(x + \Delta x, y, f)$$

with Δx the perturbation found for image x and $\epsilon = ||\Delta x||$.

Some models are more robust to high perturbations and other to low perturbations, which makes comparison difficult. Choosing the most robust model involves a trade-off in terms of accuracy.

B. A metric for robustness

We noted a recurring trend to consider a fixed perturbation size ϵ when comparing defenses [6][8][19][9]. A perturbation size that is often identical to the one used to generate the adversarial examples during training. However, in a realistic scenario, the model can be attacked with a wide range of perturbation sizes. In addition, even if the perturbation is high and visible to a human, the class may still be recognized. In this case, for generalization purposes, the model should still have relatively good performance. Therefore, we recommend evaluating over a wide range of perturbation sizes.

In binary classification problems, the Receiver Operating Characteristic (ROC) curve is computed depending on the discrimination threshold. Then models are evaluated using the Area Under the Curve (AUC) to take into account both accuracy and recall. Similarly, we propose to compute the

accuracy for a wide range of perturbation sizes and then consider the *Area Under the Accuracy Curve (AUAC)*.

Given a testing set $\mathbf{D}^{ts} = \{x_i^{ts}\}$ with associated labels $\mathbf{Y}^{ts} = \{y_i^{ts}\}$, we note $\ell_f(x_i^{ts})$ the label attributed to the test datum x_i^{ts} by a given model f. We define the AUAC up to ϵ_{max}:

$$Acc(f, \epsilon, \mathbf{D}^{ts}) = \frac{\mathbb{1}_{\{x_i^{ts} \in \mathbf{D}^{ts} | \ell_f(\hat{x}_i^{ts}) = y_i^{ts}, ||x_i^{ts} - \hat{x}_i^{ts}|| \leq \epsilon\}}}{|\mathbf{D}^{ts}|}$$

$$AUAC_{\epsilon_{max}}(f) = \frac{1}{\epsilon_{max}} \int_{\epsilon=0}^{\epsilon_{max}} Acc(f, \epsilon, \mathbf{D}^{ts}) d\epsilon.$$

$Acc(f, \epsilon, \mathbf{D}^{ts})$ is the *accuracy* of f on the test set $\mathbf{D}^{ts}$ with perturbations of size up to ϵ. ϵ_{max} is the highest perturbation considered for evaluation. In practice, to evaluate the accuracy for a given ϵ, we count the number of perturbations found below that ϵ.

Models must be defended against unknown attacks, so the evaluation cannot be fixed on a given ϵ. It is important to know if the defended model will be resistant to all kinds of perturbation sizes. AUAC evaluates the performance trade-off for a range of perturbation sizes. It takes into account the performance of a model from low to high perturbations. The closer the AUAC is to 1, the more robust the classifier is. It offers a more accurate and fair way to compare defended models.

Having discussed how to compare and evaluate defenses, we have to find an upper bound ϵ_{max} for the perturbation size. In the following, we present our choice for this bound and our experimental results.

V. Experiments

In this section we present an ablative study of the σ parameter of our *SAT* as well as a performance comparison with the state-of-the-art. First, we detail our experimental settings.

A. Experimental settings

We consider two popular datasets, namely, CIFAR-10 (6000 32×32 RGB examples of each of 10 classes) and CIFAR-100 (600 32×32 examples of each of 100 classes) [29]. For all experiments, the pixel values and perturbation size ϵ are normalized to [0,1] by dividing 255. We compare our *SAT* to normal training as well as several state-of-the-art defenses:

- **Standard Training (Standard)** trains with a cross entropy loss only on original training examples.
- **Madry Adversarial Training (Madry)** [5] trains with a cross entropy loss only on adversarial examples generated at each iteration.
- **Mixed Adversarial Training (Mixed)** [2][4] uses a mixed batch of adversarial examples (generated at each iteration) and original images with a cross entropy loss.
- **Adversarial Training with Domain Adaptation (ATDA)** [8] combines Mixed Adversarial Training with MMD and CORAL losses between adversarial and original images. We do not include their Supervised Domain Adaptation loss since it does not lead to significant improvement according to their ablative study.

For each method, we reproduce their protocol and train a Resnet20 [30] or WideResnet 28-10 [31] on each dataset until convergence to provide a fair comparison with the same model architecture.

To compute the Sinkhorn Divergences, we use the iterative Sinkhorn algorithm [27] with 50 iterations.

For both models, we use a Stochastic Gradient Descent, with an initial learning rate of 0.1, a weight decay of $5 \cdot 10^{-4}$ and a batch size of 128. For Resnet20, all the models converged after 60 epochs of training with a multiplicative factor of 0.1 on the learning rate at epoch 20 and 40. For WideResnet 28-10, the models converged after 200 epochs of training for Standard, Madry, Mixed and ATDA, and after 400 epochs for SAT. We add a multiplicative factor of 0.2 on the learning rate at epoch 60, 120, 160, 210, 250, 300, 350.

We perform 7 iterations of L_∞-PGD to generate our adversarial examples during training. For the evaluation, we apply 10 iterations for PGD or 2000 iterations for TR attack. All experiments are implemented on a single Titan X GPU.

Before evaluating the defenses, we have to define an upper bound ϵ_{max} on the perturbation size as mentioned on Section IV-B.

B. Choosing ϵ_{max}

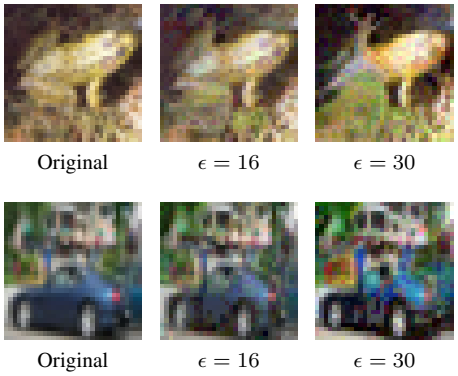

Fig. 1. Examples of images from CIFAR-10 dataset perturbed with PGD at $\epsilon = 16$ and $\epsilon = 30$.

To compute our AUAC, we have to choose an upper bound ϵ_{max}. This bound may depend on the format of images and datasets. In our case, we consider the same two possible choices for this bound on both datasets: $\epsilon_{max} = 16$ or $\epsilon_{max} = 30$.

As a first bound, we choose $\epsilon_{max} = 16$ since beyond that the perturbation becomes visible to a human. For $\epsilon > 16$, we consider that the perturbation is out of the scope of adversarial examples, yet small enough for humans to recognize the images.

As a second bound, we choose $\epsilon_{max} = 30$ since beyond that the class of the image becomes severely altered. In this

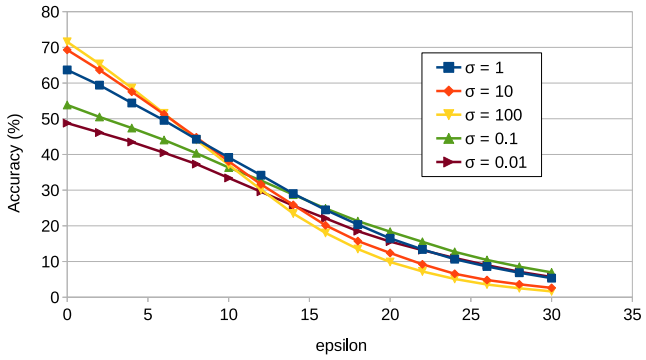

Fig. 2. Accuracy of SAT on CIFAR-10 for different entropy σ against adversarial examples of varying size ϵ generated with PGD. Defenses use a Resnet20 and $\epsilon_{train} = 8$.

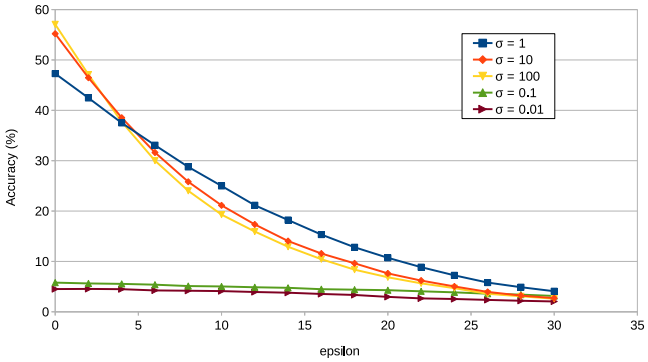

Fig. 3. Accuracy of SAT on CIFAR-100 for different entropy σ against adversarial examples of varying size ϵ generated with PGD. Defenses use a WideResnet28-10 and $\epsilon_{train} = 8$.

case, for $\epsilon > 30$, a human would no longer recognize the class of the image. The perturbed images do not belong in their original cluster anymore.

Figure 1 gives examples from CIFAR-10 for both bounds. We can see that when $\epsilon = 16$, the perturbation is barely detectable, but when $\epsilon = 30$, the class of the image is severely altered and hard to recognize.

With the upper bounds defined, we can evaluate and compare defenses for both of them.

C. Ablative study on the entropy parameter

Given that the properties of Sinkhorn Divergences vary depending on the value of the entropy parameter σ, we examine its impact on robustness.

Figures 2 and 3 compare the accuracy of SAT for different values of σ and perturbation size ϵ on CIFAR-10 and CIFAR-100. What can clearly be seen in these figures is that the performance depends on the accuracy on the original images, and how the accuracy decreases when ϵ increases. The original accuracy gives a good approximation of the performance for small ϵ, but not for high perturbations. Furthermore, comparing the performance with single ϵ does not give a good overview of the robustness of the defense. For instance, on CIFAR-100, for $\epsilon = 2$, SAT_1 and SAT_{100} have respectively 42.46% and 47.09% accuracy but for $\epsilon = 8$, the same models have respectively 28.77% and 24.02% accuracy. SAT_1 has the best performance when $\epsilon > 5$ whereas SAT_{100} is more robust when $\epsilon < 5$. This emphasizes the need for a more accurate evaluation metric for robustness.

Table I provides AUAC up to $\epsilon_{max} = 16$ or $\epsilon_{max} = 30$ depending on the entropy parameter σ for Resnet20 on CIFAR-10 and WideResnet28-10 on CIFAR-100. We can see that SAT is the most robust overall when $\sigma = 1$. For instance, on CIFAR-100 with $\sigma = 1$, SAT has an AUAC of 29.69% up to $\epsilon_{max} = 16$ and 19.83% when $\epsilon_{max} = 30$.

In summary, these results show that using $\sigma = 1$ results in a more robust defense over a wide range of perturbation sizes. This entropy represents a hinge value between an MMD and an Optimal Transport loss [32]. In the following, we will always consider SAT with an entropy of 1.

Dataset / Architecture	σ	AUAC@16 (%)	AUAC@30 (%)
CIFAR-10 / Resnet20	1	44.26	**29.69**
	10	**44.69**	28.09
	100	44.38	27.11
	0.1	39.94	28.16
	0.01	36.44	25.32
CIFAR-100 / WideResnet28-10	1	**29.69**	**19.83**
	10	28.55	18.08
	100	27.59	17.30
	0.1	5.18	4.59
	0.01	4.16	3.48

TABLE I
COMPARISON OF AUAC UP TO $\epsilon_{max} = 16$ AND $\epsilon_{max} = 30$ AGAINST PGD FOR DIFFERENT VALUES OF σ IN SAT.

Let us now compare the performance of SAT with other state-of-the-art defenses.

D. Comparison with other defenses

In the following experiments, we compare the performance of our SAT (with $\sigma = 1$) to other state-of-the-art defenses, first against PGD then against TR attack.

Figure 4 provides an overview of the accuracy on CIFAR-10 with $\epsilon_{train} = 8$ against PGD over a wide range of perturbation sizes, from $\epsilon = 0$ (original accuracy) to $\epsilon = 30$. In addition to the methods presented in Section V-A, we report the results for the Prototype Conformity [9] (PC) using the weights publicly available. Interestingly, this defense is less robust to low perturbations than other methods. On the other hand, other defenses have similar performance at low perturbations but they do not react in the same way when the size of the perturbation increases. ATDA has the lowest performance when $\epsilon \geq 12$ and SAT is the most robust method overall. For instance, for $\epsilon = 8$, our SAT has an accuracy of 49.31% whereas Mixed has 46.46%, Madry 46.26%, ATDA 42.1%, PC 32.32% and the Standard method 0.02%.

Similarly, Figure 5 presents a comparison of the performance on CIFAR-100 against PGD for the same range of per-

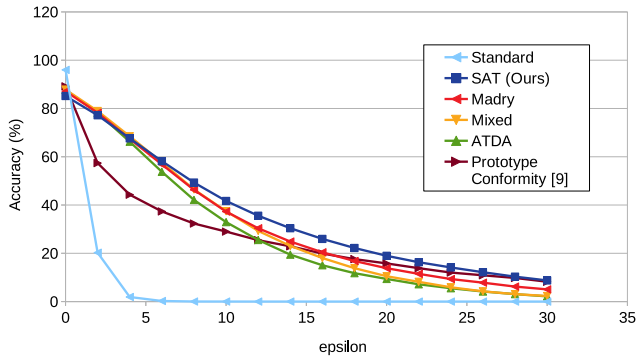

Fig. 4. Performance of the defenses on CIFAR-10 against PGD for perturbation sizes up to $\epsilon = 30$. Our SAT and the other reproduced defenses use a WideResnet28-10 and adversarial examples generated with $\epsilon_{train} = 8$. Prototype Conformity defense uses a Resnet110 with a uniformly sampled ϵ_{train} between 3 and 13.

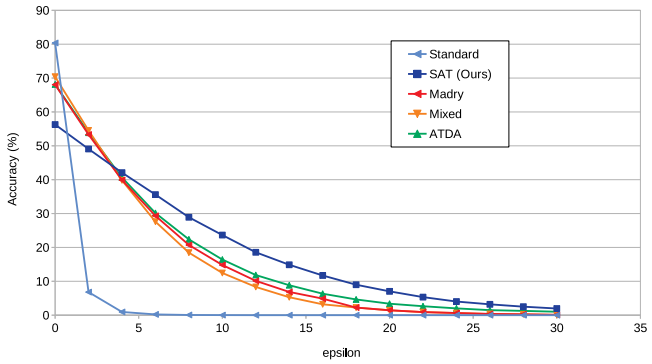

Fig. 5. Performance of the defenses on CIFAR-100 against PGD for perturbation sizes up to $\epsilon = 30$. Our SAT and the other reproduced defenses use a WideResnet28-10 and adversarial examples generated with $\epsilon_{train} = 4$.

turbation size but with $\epsilon_{train} = 4$. In this case, it can be seen that all defenses except SAT have very similar performance. SAT is significantly more robust to large perturbations than the other methods but not as robust to small perturbations. SAT achieves better robustness overall but with lower original accuracy. For instance, for $\epsilon = 4$, our SAT has an accuracy of 42.09%, whereas Mixed has 39.8%, Madry 39.91%, ATDA 40.64% and the Standard method 0.95%.

The results of the AUAC against PGD on both datasets are provided in Table II. SAT is the most robust defense over all perturbations used during evaluation. For instance, on CIFAR-10 with $\epsilon_{train} = 8$, SAT has an AUAC up to $\epsilon = 16$ of 51.93% whereas ATDA has 46.19%.

Figures 6 and 7 compares the performance of the defenses against TR attack. On CIFAR-10, we can see that all the defenses have similar performance against small perturbations ($\epsilon \leq 5$) but SAT is significantly more robust to higher perturbations. On CIFAR-100, SAT has lower performance on small perturbations than the other defenses. However, SAT and ATDA are the most robust to high perturbations.

Table III provides the AUAC against TR attack on both

Dataset	Archi.	Model	AUAC (%)	
			$\epsilon_{max} = 16$	$\epsilon_{max} = 30$
CIFAR-10	Resnet20	Standard	5.79	3.09
		Madry	44.18	26.53
		Mixed	40.68	22.73
		ATDA	35.58	21.63
		SAT (Ours)	**44.26**	**29.69**
	Resnet110	PC [9]	37.89	26.47
	WideResnet28-10	Standard	8.8	4.69
		Madry	49.37	31.54
		Mixed	49.27	30.01
		ATDA	46.19	27.94
		SAT (Ours)	**51.93**	**35.12**
CIFAR-100	WideResnet28-10	Standard	6.03	3.22
		Madry	27.27	16.14
		Mixed	27.80	16.13
		ATDA	28.59	17.11
		SAT (Ours)	**29.69**	**19.83**

TABLE II
COMPARISON OF AUAC (IN %) UP TO $\epsilon_{max} = 16$ OR $\epsilon_{max} = 30$ AGAINST PGD ON BOTH CIFAR-10 AND CIFAR-100. OUR SAT AND OTHER DEFENSES REPRODUCED ARE TRAINED WITH ADVERSARIAL EXAMPLES GENERATED WITH $\epsilon_{train} = 8$.

datasets when training on adversarial examples generated with $\epsilon_{train} = 4$ or $\epsilon_{train} = 8$. On CIFAR-10, our SAT is significantly more robust than the other defenses. On CIFAR-100, despite good robustness on high perturbations, the lower accuracy on small perturbations leads to a lower AUAC for SAT than ATDA. However, SAT still has competitive AUAC compared to other defenses.

Note that when increasing ϵ_{train}, the defended models become more robust to high perturbations but the accuracy on original images decreases. However, the accuracy on high perturbations is also bounded by the accuracy on original images. Thus, the AUAC can help finding a good trade-off on ϵ_{train} between accuracy on original images and robustness to high perturbations.

In this section, we provided an experimental analysis of our SAT. First, we presented an ablative study on the entropy parameter σ which resulted in fixing $\sigma = 1$. It can be interpreted as using a Sinkhorn Divergence at the boundary between an MMD loss and an Optimal Transport loss. Then, we compared our SAT to other state-of-the-art defenses and showed that it is a more robust defense on a wide range of perturbation sizes.

VI. Conclusion

Defending against adversarial attacks is a critical issue for practical applications using deep learning. Images can easily be altered in such a way that undefended models make mistakes, which can lead to security breaches.

In this work, we follow the analogy with domain adaptation and consider separate distributions of original and adversarial images. We aim to improve the robustness of models to adversarial attacks by aligning the representations of both distributions using a least effort approach. In addition, we raise

Dataset	Archi.	Model	AUAC (%)			
			$\epsilon_{max} = 16$		$\epsilon_{max} = 30$	
			$\epsilon_{train} = 4$	$\epsilon_{train} = 8$	$\epsilon_{train} = 4$	$\epsilon_{train} = 8$
CIFAR-10	WideResnet28-10	Standard	11.23		6.00	
		Madry	52.56	52.05	34.93	36.07
		Mixed	51.50	52.22	33.44	34.89
		ATDA	49.43	50.94	31.04	33.54
		SAT (Ours)	**54.80**	**56.29**	**35.99**	**41.85**
CIFAR-100	WideResnet28-10	Standard	5.34		2.85	
		Madry	26.76	27.25	15.55	16.64
		Mixed	25.76	27.95	14.56	16.71
		ATDA	**30.82**	**30.15**	**19.40**	**19.64**
		SAT (Ours)	29.71	27.80	18.99	19.27

TABLE III
COMPARISON OF AUAC (IN %) UP TO $\epsilon_{max} = 16$ OR $\epsilon_{max} = 30$ AGAINST TR ATTACK ON BOTH CIFAR-10 AND CIFAR-100. DEFENSES ARE TRAINED WITH ADVERSARIAL EXAMPLES GENERATED WITH $\epsilon_{train} = 4$ OR $\epsilon_{train} = 8$.

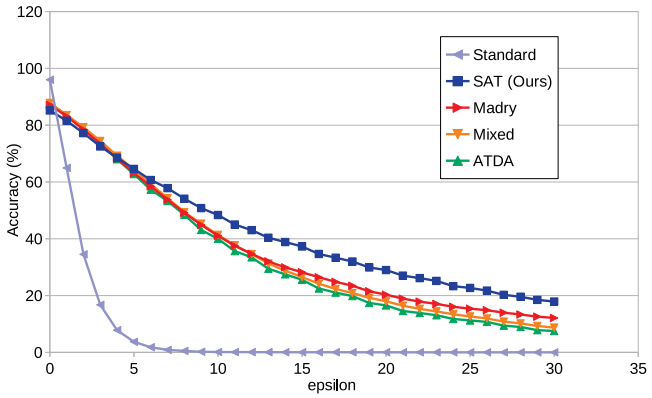

Fig. 6. Performance of the defenses on CIFAR-10 against TR attack for perturbation sizes up to $\epsilon = 30$. Our SAT and the other reproduced defenses use a WideResnet28-10 and adversarial examples generated with $\epsilon_{train} = 8$.

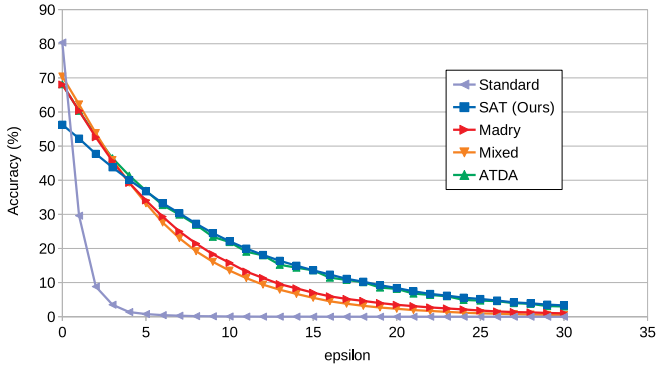

Fig. 7. Performance of the defenses on CIFAR-100 against TR attack for perturbation sizes up to $\epsilon = 30$. Our SAT and the other reproduced defenses use a WideResnet28-10 and adversarial examples generated with $\epsilon_{train} = 4$.

concerns on the recurring trend to evaluate defenses with a single perturbation size similar to the one use during training. It does not reflect the variations in robustness depending on the size of the perturbations.

First, we propose SAT (Sinkhorn Adversarial Training), a new defense using the theory of optimal transport with Sinkhorn Divergences. We minimize the ground distance between representations to bring closer the distributions and take into account the geometry of space. Then, for a fair evaluation of robustness, we propose the Area Under the Accuracy Curve (AUAC) to compare defenses. The AUAC integrates the performance on a wide range of perturbation sizes to quantify robustness.

Finally, we perform a thorough analysis on CIFAR-10 and CIFAR-100 datasets and show that SAT outperforms other defenses by comparing accuracy curves and our AUAC.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2014. 1, 2, 4

[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *European Symposium on Security and Privacy (EuroS&P)*, 2016. 1, 2

[4] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," in *International Conference on Learning Representations, ICLR*, 2017. 1, 2, 3, 4

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3, 4

[6] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3

[7] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[8] C. Song, K. He, L. Wang, and J. E. Hopcroft, "Improving the generalization of adversarial training with domain adaptation," in *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 3, 4

[9] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial defense by restricting the hidden space of deep neural networks," in *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 5, 6

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Symposium on Security and Privacy (S&P)*, 2017. 2

[11] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. W. Mahoney, "Trust region based adversarial attack on neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[12] J. Rony, L. G. Hafemann, L. S. Oliveira, I. Ben Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[13] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[14] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations (ICLR)*, 2018. 2

[15] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR), Workshop*, 2017. 2

[16] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning (ICML)*, 2018. 2

[17] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations (ICLR)*, 2018. 2

[18] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[19] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *International Conference on Machine Learning (ICML)*, 2019. 2, 3

[20] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," in *International Conference on Learning Representations (ICLR)*, 2018. 2

[21] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *arXiv:2002.08347*, 2020. 2

[22] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 2

[23] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Conference on Artificial Intelligence (AAAI)*, 2016. 2

[24] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré, "Interpolating between optimal transport and mmd using sinkhorn divergences," in *Proceedings of Machine Learning Research (PMLR)*, 2019. 2

[25] A. Genevay, "Entropy-regularized optimal transport for machine learning," Ph.D. dissertation, 2019. 2

[26] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, 2019. 2

[27] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 2, 4

[28] A. Genevay, G. Peyré, and M. Cuturi, "Learning generative models with sinkhorn divergences," 2017. 2

[29] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. 4

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[31] S. Zagoruyko and N. Komodakis, "Wide residual networks," *British Machine Vision Conference (BMVC)*, 2016. 4

[32] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. 5