Jacobian Regularization for Mitigating Universal Adversarial Perturbations

Kenneth T. Co[1,2](✉), David Martinez Rego[2], and Emil C. Lupu[1]

[1] Imperial College London, London SW7 2AZ, UK
{k.co,e.c.lupu}@imperial.ac.uk
[2] DataSpartan, London EC2Y 9ST, UK
david@dataspartan.com

Abstract. Universal Adversarial Perturbations (UAPs) are input perturbations that can fool a neural network on large sets of data. They are a class of attacks that represents a significant threat as they facilitate realistic, practical, and low-cost attacks on neural networks. In this work, we derive upper bounds for the effectiveness of UAPs based on norms of data-dependent Jacobians. We empirically verify that Jacobian regularization greatly increases model robustness to UAPs by up to four times whilst maintaining clean performance. Our theoretical analysis also allows us to formulate a metric for the strength of shared adversarial perturbations between pairs of inputs. We apply this metric to benchmark datasets and show that it is highly correlated with the actual observed robustness. This suggests that realistic and practical universal attacks can be reliably mitigated without sacrificing clean accuracy, which shows promise for the robustness of machine learning systems.

Keywords: Adversarial machine learning · Universal adversarial perturbations · Computer vision · Jacobian regularization

1 Introduction

Neural networks have been the algorithm of choice for many applications such as image classification [15], real-time object detection [21], and speech recognition [11]. Although they appear to be robust to noise, their accuracy can rapidly deteriorate in the face of adversarial examples – inputs that appear similar to genuine data, but have been maliciously designed to fool the model [1,25]. Thus, it is important to ensure that neural networks are robust to such attacks, especially in safety-critical applications, as this can greatly undermine the performance and trust in these models.

A concerning subset of attacks on neural networks come in the form of Universal Adversarial Perturbations (UAPs), where a single adversarial perturbation can cause a model to misclassify a large set of inputs [18]. These present a systemic risk, as many practical and physically realizable adversarial attacks are

Kenneth T. Co is supported in part by the DataSpartan research grant DSRD201801.

I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12894, pp. 202–213, 2021.
https://doi.org/10.1007/978-3-030-86380-7_17

based on UAPs. These attacks can take the form of adversarial patches for image classification [2], person recognition [26], camera-based [7,8] and LiDAR-based object detection [3,9,10,28]. In the digital domain, UAPs have been shown to facilitate realistic attacks on perceptual ad-blockers for web pages [27] and machine learning-based malware detectors [16]. Furthermore, an attacker can utilize UAPs to perform query-efficient black-box attacks on neural networks [4,6].

In the literature, existing defenses to adversarial attacks focus primarily on input-specific ("per-input") attacks–where adversarial perturbations need to be crafted *for each single input*. In contrast to universal attacks, input-specific attacks fool the model on only *one input*. However, the practicality of input-specific attacks suffers in realistic settings, as the perturbations need to be constantly modified to match the current input. In contrast, defences against UAPs have not been thoroughly investigated, even if they are potentially more dangerous and should intuitively be easier to defend against because the same perturbation needs to be shared across many inputs. These are the main focus of this paper.

A number of studies have investigated the use of Jacobian regularization to improve the stability of model predictions to small changes to the input, but up to this point, studies have only considered input-specific perturbations [12, 13,20,22,24,29]. In this work, we expand the theoretical formulation of Jacobian regularization to UAPs and derive upper bounds on the effectiveness of UAPs based on the properties of Jacobian matrices for individual inputs. Our work shows that for inputs to strongly share adversarial perturbations, their Jacobians need to share singular vectors.

We empirically verify our theoretical findings by applying Jacobian regularization to neural networks trained on popular benchmark datasets: MNIST [17], Fashion-MNIST [30] and then evaluating their robustness to various UAPs. Our results show that even a small amount of Jacobian regularization drastically improves model robustness against many universal attacks with negligible downsides to clean performance. To summarize, we make the following contributions:

- We extend theoretical formulations for universal adversarial perturbations and are the first to show that the effectiveness of UAPs is bounded above by the norms of data-dependent Jacobians.
- We empirically verify our theoretical results and show that even a minimal amount of Jacobian regularization reduces effectiveness of UAPs by up to 4-times, whilst leaving clean accuracy relatively unaffected.
- We propose the use of cosine similarity for Jacobians of inputs to measure the strength of shared adversarial perturbations between distinct inputs. Our empirical evaluations on benchmark datasets demonstrate that this similarity measure is an effective proxy for measuring robustness to UAPs.

The rest of this paper is organized as follows. Section 2 introduces adversarial examples, universal adversarial perturbations, and Jacobian regularization. Section 3 formulates Jacobian regularization for UAPs and derives our key propositions. Section 4 evaluates the robustness of models trained with Jacobian regularization to various UAP attack. Finally, Sect. 5 discusses implications of our results and summarizes our findings.

2 Background

2.1 Universal Adversarial Perturbations

Let $f : \mathcal{X} \subset \mathbb{R}^n \rightarrow \mathbb{R}^d$ denote the logits of a piece-wise linear classifier which takes as input $\mathbf{x} \in \mathcal{X}$. The output label assigned by this classifier is defined by $F(\mathbf{x}) = \arg\max(f(\mathbf{x}))$. Let $\tau(\mathbf{x})$ denote the true class label of an input $\mathbf{x}$).

An *adversarial example* $\mathbf{x}'$ is an input that satisfies $F(\mathbf{x}') \neq \tau(\mathbf{x})$, despite $\mathbf{x}'$ being close to $\mathbf{x}$ according to some distance metric (implicitly, $\tau(\mathbf{x}) = \tau(\mathbf{x}')$). The difference $\delta = \mathbf{x}' - \mathbf{x}$ is referred to as an adversarial perturbation and its norm is often constrained to $\|\delta\|_p < \varepsilon$, for some ℓ_p-norm and small $\varepsilon > 0$ [25].

Universal Adversarial Perturbations (UAP) can come in targeted or untargeted forms depending on the attacker's objective. An untargeted UAP is an adversarial perturbation $\delta \in \mathbb{R}^n$ that satisfies $F(\mathbf{x} + \delta) \neq \tau(\mathbf{x})$ for sufficiently many $\mathbf{x} \in \mathcal{X}$ and with $\|\delta\|_p < \varepsilon$ [18]. Untargeted UAPs are generated by maximizing the loss $\sum_i \mathcal{L}(\mathbf{x}_i + \delta)$ with an iterative stochastic gradient descent algorithm [5,19,23,27]. Here, $\mathcal{L}$ is the model's training loss, $\{\mathbf{x}_i\}$ are batches of inputs, and δ are small perturbations that satisfy $\|\delta\|_p < \varepsilon$. Updates to δ are done in mini-batches in the direction of $-\sum_i \nabla\mathcal{L}(\mathbf{x}_i + \delta)$. Targeted UAPs for a class c are adversarial perturbations δ that satisfy $F(\mathbf{x} + \delta) = c$ for sufficiently many $\mathbf{x} \in \mathcal{X}$ and with $\|\delta\|_p < \varepsilon$. To generate this type of attack, we use the same stochastic gradient descent as in the untargeted case, but modify the loss to be minimized when all resulting inputs $\mathbf{x}_i + \delta$ are classified as c.

2.2 Jacobian Regularization

Given that $f(\mathbf{x})$ is the logit output of the classifier for input $\mathbf{x}$, we write $\mathbf{J}_f(\mathbf{x})$ to denote the input-output Jacobian of f at $\mathbf{x}$. We can linearise f within a neighbourhood around $\mathbf{x}$ as follows using the Taylor series expansion:

$$f(\mathbf{x} + \delta) = f(\mathbf{x}) + \mathbf{J}_f(\mathbf{x})\delta + O(\delta^2) \tag{1}$$

For a sufficiently small neighbourhood $\|\delta\|_p \leq \varepsilon$ with $\varepsilon > 0$, the higher order terms of δ can be neglected and the stability of the prediction is determined by the Jacobian.

$$f(\mathbf{x} + \delta) \simeq f(\mathbf{x}) + \mathbf{J}_f(\mathbf{x})\delta \tag{2}$$

and equivalently, for any q-norm, we have:

$$\|f(\mathbf{x} + \delta) - f(\mathbf{x})\|_q \approx \|\mathbf{J}_f(\mathbf{x})\delta\|_q \tag{3}$$

For a small ε, we want the δ that maximizes the right hand side of Eq. 3 in order to sufficiently change the original output and fool the model. With constraint $\|\delta\|_p \leq \varepsilon$, this is equivalent to finding the (p, q) singular vector for $\mathbf{J}_f(\mathbf{x})$ [14].

To improve the stability of model outputs to small perturbations δ, existing works have proposed regularizing the Frobenius norm [12,13,20] or the Spectral norm [22,24,29] of this data-dependent Jacobian $\mathbf{J}_f(\mathbf{x})$ for each input. Additionally, [22] show that the input-specific adversarial perturbations align with the dominant singular vectors of these Jacobian matrices.

Although [14] considered Jacobians in the context of UAPs, they only focused on the computation of δ as an attack and did not perform any theoretical or empirical analysis for mitigating the effects of UAPs. Prior studies that explore Jacobian regularization focused solely on improving robustness to single-input perturbations and did not explain nor consider the effectiveness of Jacobian regularization for UAPs. Thus, we extend these formulations [14,22] to have a more concrete theoretical understanding for how Jacobian regularization mitigates UAPs.

3 Jacobians for Universal Adversarial Perturbations

When computing a universal adversarial perturbation δ that uniformly generalizes across multiple inputs $\{\mathbf{x}_i\}_{i=1}^N$, one would optimize:

$$\max_{\delta:\|\delta\|_p=1} \sum_{i=1}^{N} \|\mathbf{J}_f(\mathbf{x}_i)\delta\|_q \tag{4}$$

This extends the intuition from Eq. 3 to many inputs, and due to the homogeneity of the norm, it is sufficient to solve this for $\|\delta\|_p = 1$ [14]. The solution to δ for Eq. 4 is equivalent to finding the (p, q) singular vector for the **stacked Jacobian** matrix $\overline{\mathbf{J}}_N$, the matrix formed by vertically stacking the Jacobians of the first N inputs.

$$\max_{\delta:\|\delta\|_p=1} \|\overline{\mathbf{J}}_N\delta\|_q \quad \text{where} \quad \overline{\mathbf{J}}_N = \begin{bmatrix} \mathbf{J}_f(\mathbf{x}_1) \\ \mathbf{J}_f(\mathbf{x}_2) \\ \vdots \\ \mathbf{J}_f(\mathbf{x}_N) \end{bmatrix} \tag{5}$$

3.1 Upper Bounds for the Stacked Jacobian

To obtain an upper bound for the (p, q)-operator norm shown in Eq. 5, note that it is bounded above by its Frobenius norm denoted by $\|\overline{\mathbf{J}}_N\|_F$:

$$\|\overline{\mathbf{J}}_N\delta\|_q \leq \|\overline{\mathbf{J}}_N\|_F\|\delta\| \tag{6}$$

Thus, mitigating the effectiveness of a UAP across multiple inputs can be achieved by limiting the Frobenius norm of the stacked Jacobian $\|\overline{\mathbf{J}}_N\|_F$.

Before proceeding, let us define the inner product induced by the Frobenius norm for two real matrices. Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, let the inner product in $\mathbb{R}^{m\times n}$ be defined as:

$$\langle\mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}'\mathbf{B}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij} \tag{7}$$

where $\mathbf{A}'$ denotes the transpose of $\mathbf{A}$, the lowercase letters a_{ij} are the entries of the matrix $\mathbf{A}$, and $\mathrm{Tr}(\cdot)$ is the trace. This inner product is associated with the Frobenius norm $\|\cdot\|_F$. Now we introduce the following proposition.

Proposition 1. *For matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$*, we have:*

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_F \tag{8}$$

with equality if and only if $\mathbf{A}$ *and* $\mathbf{B}$ *share singular directions and their singular values satisfy* $\sigma_i(\mathbf{A}) = s \cdot \sigma_i(\mathbf{B})$ *for all i for a constant scalar* $s > 0$*, where* $\sigma_i(\cdot)$ *is the singular value that corresponds to the i-th largest singular value.*

Proof. Consider the singular value decomposition of $\mathbf{A} = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A'$ and $\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B'$, where $\mathbf{U}_A, \mathbf{U}_B, \mathbf{V}_A, \mathbf{V}_B$ are orthogonal matrices and $\mathbf{\Sigma}_A, \mathbf{\Sigma}_B$ are diagonal matrices whose diagonal entries $\sigma_i(\mathbf{A})$ and $\sigma_i(\mathbf{B})$ are non-negative and in descending order. Let $r = \max(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B}))$.

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \rangle &= \mathrm{Tr}(\mathbf{A}'\mathbf{B}) \\ &= \mathrm{Tr}(\mathbf{V}_A \mathbf{\Sigma}_A' \mathbf{U}_A' \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B') \\ &= \mathrm{Tr}(\mathbf{V}_B' \mathbf{V}_A \mathbf{\Sigma}_A' \mathbf{U}_A' \mathbf{U}_B \mathbf{\Sigma}_B) \qquad \text{cyclic property of trace} \end{aligned}$$

Note that since $\mathbf{U}_A, \mathbf{U}_B, \mathbf{V}_A, \mathbf{V}_B$ are all orthogonal matrices, $\|\mathbf{U}_A'\mathbf{U}_B\|_2 \leq \|\mathbf{U}_A'\|_2 \|\mathbf{U}_B\|_2 = 1$, and in a similar way, $\|\mathbf{V}_B'\mathbf{V}_A\|_2 \leq 1$.

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \rangle &= \mathrm{Tr}(\mathbf{V}_B' \mathbf{V}_A \mathbf{\Sigma}_A' \mathbf{U}_A' \mathbf{U}_B \mathbf{\Sigma}_B) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} z_{ij} \cdot \sigma_i(\mathbf{A}) \sigma_j(\mathbf{B}) \qquad \text{where } \sum_{i=1}^{r} |z_{ij}| \leq 1, \sum_{j=1}^{r} |z_{ij}| \leq 1 \\ &\leq \sum_{i=1}^{r} \sigma_i(\mathbf{A})\, \sigma_i(\mathbf{B}) \qquad \text{equality} \iff z_{ij} \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \\ &\leq \left(\sum_{i=1}^{r} \sigma_i^2(\mathbf{A}) \right)^{\frac{1}{2}} \left(\sum_{i=1}^{r} \sigma_i^2(\mathbf{B}) \right)^{\frac{1}{2}} \qquad \text{Cauchy-Schwarz Inequality} \\ &= \|\mathbf{A}\|_F \|\mathbf{B}\|_F \qquad \square \end{aligned}$$

The equality conditions for the above requires $z_{ii} = 1, \forall i$ as the σ_i are in descending order. This implies that $\mathbf{U}_A'\mathbf{U}_B$ and $\mathbf{V}_B'\mathbf{V}_A$ are identity matrices, which requires $\mathbf{U}_A = \mathbf{U}_B$ and $\mathbf{V}_A = \mathbf{V}_B$, i.e. $\mathbf{A}$ and $\mathbf{B}$ share the same singular vectors. Equality under Cauchy-Schwarz requires the singular values to be scalars of one another: $\sigma_i(\mathbf{A}) = s \cdot \sigma_i(\mathbf{B})$ for the same scalar $s > 0, \forall i$.

This proposition is significant as it gives us upper bounds for the inner product and equality conditions to achieve this upper bound. Applying this result to the stacked Jacobian matrix $\overline{\mathbf{J}}_N$ gives us the following:

$$\begin{aligned}
\|\overline{\mathbf{J}}_N\|_F^2 &= \mathrm{Tr}(\overline{\mathbf{J}}_N'\overline{\mathbf{J}}_N) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{J}_f(\mathbf{x}_i)'\mathbf{J}_f(\mathbf{x}_j)\right) \\
&= \sum_{i,j}\mathrm{Tr}(\mathbf{J}_f(\mathbf{x}_i)', \mathbf{J}_f(\mathbf{x}_j)) \\
&= \sum_{i,j}\langle\mathbf{J}_f(\mathbf{x}_i), \mathbf{J}_f(\mathbf{x}_j)\rangle && \text{Frobenius inner product} \\
&\leq \sum_{i,j}\|\mathbf{J}_f(\mathbf{x}_i)\|_F\|\mathbf{J}_f(\mathbf{x}_j)\|_F && \text{Proposition 1}
\end{aligned}$$

With equality if and only if, for all pairs of inputs $(\mathbf{x}_i, \mathbf{x}_j)$, we have $\mathbf{J}_f(\mathbf{x}_i)$ and $\mathbf{J}_f(\mathbf{x}_j)$ sharing singular vectors and their corresponding singular values are constant up to a fixed scalar $s > 0$.

Our result can be summarized with the following equation:

$$\|\overline{\mathbf{J}}_N\|_F \leq \left(\sum_{i,j}\|\mathbf{J}_f(\mathbf{x}_i)\|_F\|\mathbf{J}_f(\mathbf{x}_j)\|_F\right)^{\frac{1}{2}} \tag{9}$$

From a defense perspective, this shows that regularizing the Frobenius of the Jacobian for the $\mathbf{x}_i$ decreases the total Frobenius norm of the stacked Jacobian and hinders the overall effectiveness of a UAP. Thus, data-dependent Jacobian regularization across inputs should make it significantly more difficult to generate effective UAPs.

3.2 Measuring Alignment of Jacobians

To measure the alignment between Jacobians of two distinct inputs, we use the **cosine similarity** between their respective Jacobians under the inner product induced by the Frobenius norm:

$$\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle\mathbf{J}_f(\mathbf{x}_i), \mathbf{J}_f(\mathbf{x}_j)\rangle}{\|\mathbf{J}_f(\mathbf{x}_i)\|_F\|\mathbf{J}_f(\mathbf{x}_j)\|_F} \leq 1 \tag{10}$$

This is precisely the formula given in Proposition 1, with the above ratio equal to one if and only if the singular vectors of their Jacobians are the same. This shows to us that alignment of Jacobians can be evaluated with this similarity measure. Also, combining this with our findings from Eq. 9, this ratio allows us to measure how strongly two inputs share adversarial perturbations.

Although the Jacobian is a first-order derivative, we show in later sections that our Jacobian similarity measure correlates with vulnerability to iterative UAP attacks. Thus, demonstrating that it is an effective measure to determine the "universality" of adversarial vulnerability even against iterative adversaries.

Having a similarity measure like this is beneficial as this allows us to easily determine if two inputs are likely to share adversarial perturbations. This is more advantageous than manually generating adversarial perturbations for each pair of inputs as one would have to consider many additional attack parameters when generating adversarial attacks, including the ε bounds, chosen ℓ_p-norm, step size, number of attack iterations, and so on.

4 Experiments

4.1 Experimental Setup

Models & Datasets. We consider the benchmark datasets MNIST [17] and Fashion-MNIST [30]. These are widely-used image classification datasets, each with 10 classes, whose images are 28 by 28 pixels, and their pixel values range from 0 to 1. For the neural network architecture, we use a modernized version of LeNet-5 [17] as detailed in [12] as it is a commonly used benchmark neural network. We refer to this model as LeNet.

Jacobian Regularization. For training with Jacobian regularization (JR), we optimize the following joint loss and use the algorithm as proposed by [12]:

$$\mathcal{L}_{\text{joint}}(\theta) = \mathcal{L}_{\text{train}}(\{\mathbf{x}_i, \mathbf{y}_i\}_i, \theta) + \frac{\lambda_{\text{JR}}}{2}\left(\frac{1}{B}\sum_i \|\mathbf{J}(\mathbf{x}_i)\|_F^2\right) \tag{11}$$

where θ represent the parameters of the model, $\mathcal{L}_{\text{train}}$ is the standard cross-entropy training loss, $\{\mathbf{x}_i, \mathbf{y}_i\}$ are input-output pairs from the mini-batch, and B is the mini-batch size. This optimization uses a regularization parameter λ_{JR}, which lets us adjust the trade-off between regularization and classification loss.

UAP Attacks. We evaluate the robustness of these models to UAPs generated via iterative stochastic gradient descent with 100 iterations and a batch size of 200. Perturbations are applied under ℓ_∞-norm constraints. The ε we consider in our attacks for this norm are from 0.1 to 0.3, this perturbation magnitude is equivalent to 10%–30% of the maximum total possible change in pixel values.

We generate untargeted and targeted attacks. For targeted UAPs, we generate one UAP for each of 10 classes of each dataset. Clean and UAP evaluations are done on the entire 10,000 sample test sets.

Robustness Metrics. The effectiveness of untargeted attacks are measured using the *Universal Evasion Rate (UER)*, defined as the proportion of inputs that are misclassified. Targeted UAPs for class c are evaluated according to their *Targeted Success Rate (TSR)*, the proportion of inputs classified as class c.

4.2 Jacobian Regularization Mitigates UAPs

Regular training without JR (i.e. $\lambda_{\text{JR}} = 0$) achieves 99.08% and 90.84% test accuracy on MNIST and Fashion-MNIST respectively. Figure 1 shows that increasing

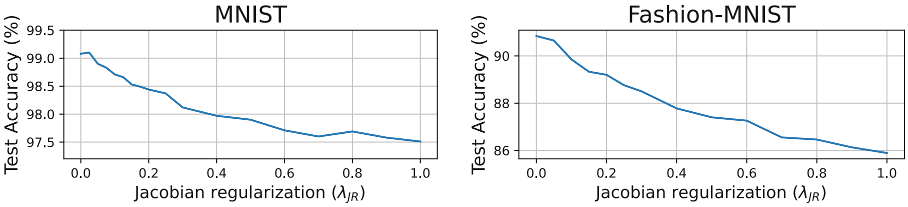

Fig. 1. Test accuracy of LeNet on MNIST (left) and Fashion-MNIST (right) for various Jacobian regularization strengths λ_{JR}.

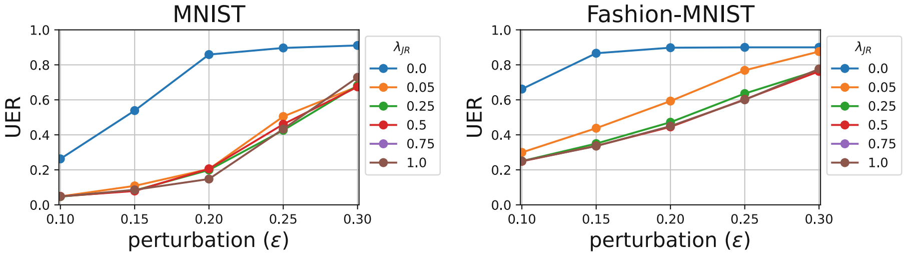

Fig. 2. Effectiveness of untargeted UAPs for various ℓ_∞-norm perturbation constraints ε. Plots are shown for various models with different degrees of Jacobian regularization.

the weight of JR decreases the resulting model's test accuracy. Note, however, that this decrease appears to be negligible for very small $\lambda_{\mathrm{JR}} \leq 0.1$.

Untargeted UAPs. Figure 2 presents the effectiveness of our untargeted UAP attacks on different LeNet with varying JR strengths. The regularly trained model is especially vulnerable to UAP attacks on both datasets, with untargeted UAPs achieving above 80% UER for $\varepsilon \geq 0.2$ on both datasets.

On MNIST, UAP attacks seem to gain reasonable success only after $\varepsilon \geq 0.25$. This is permissible as the adversary perturbs the input by 25% of its maximum possible value in this case, which entails an enormous change. What is striking is that JR has a protective effect for $\varepsilon \leq 0.2$, even for small amounts of regularization at $\lambda_{\mathrm{JR}} = 0.05$. Here, UAP effectiveness is down from 80% to 20% at $\varepsilon = 0.2$. Increasing the strength of the regularization likely has diminishing returns for robustness as stronger regularization also begins to damage clean accuracy, and thus the model's generalization. Fashion-MNIST can be seen to be less robust since it begins with a lower clean accuracy at around 91%. This means that the model is overall less robust to begin with than the model trained for MNIST, so we can expect it to be less robust to UAP attacks in general. Nonetheless, we still see a protective effect from JR for $\varepsilon \leq 0.15$ even with only a minor degree of regularization $\lambda_{\mathrm{JR}} = 0.05$.

Targeted UAPs. Figure 3 shows our results for the effectiveness of targeted UAPs. These plots follow a similar trend as with untargeted UAPs, suggesting that JR is able to improve model robustness against a diverse array of UAP attacks and not only against untargeted UAPs.

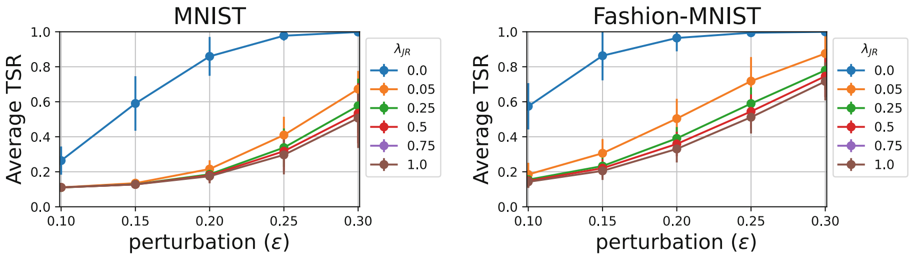

Fig. 3. Average Targeted Success Rate (TSR) of targeted UAPs generated for each class, with error bars showing standard deviation across UAPs for different classes. Plots are shown for various models with different degrees of Jacobian regularization.

Even a minor amount of regularization in $\lambda_{\text{JR}} = 0.05$ provides up to a 4-times decrease in effectiveness of UAPs while maintaining the model's performance on the clean test set, as seen in Table 1.

Comparison with Adversarial Training. We compare JR with the current state-of-the-art defense against universal attacks: Universal Adversarial Training (UAT) [23], where adversarial training is done on UAPs. UAT models in Table 1 are trained on $\varepsilon = 0.2$ and $\varepsilon = 0.15$ adversaries for MNIST and Fashion-MNIST respectively. Although UAT improves robustness to UAPs compared to standard training, it doubles the test error on both clean datasets. In contrast, JR achieves better robustness than UAT without damaging clean accuracy.

Adversarial training relies on training against specific UAP perturbations. The heuristic quality of UAT makes improving robustness against all possible perturbations computationally difficult. Our results show that regularizing a more general property of the model, in the norm of the Jacobian, leads to better robustness while maintaining accuracy.

Table 1. Performance metrics (in %) of LeNet. Jacobian regularization (JR) uses $\lambda_{\text{JR}} = 0.05$. UAP evaluations are for ℓ_∞-norm attacks at $\varepsilon = 0.2$ for MNIST and $\varepsilon = 0.15$ for Fashion-MNIST. Lowest values indicate the best robustness and are highlighted.

	MNIST			Fashion-MNIST		
	Standard	UAT [23]	JR	Standard	UAT [23]	JR
Test Error	0.92	1.81	**0.90**	9.16	16.66	**9.15**
Untargeted UER	85.88	27.49	**20.47**	86.63	34.10	**29.96**
Average TSR	85.94	24.05	**21.57**	86.33	**26.64**	30.59

4.3 Jacobian Alignment of Input Pairs

We now investigate how the cosine similarity of input Jacobians as introduced in Eq. 10 correlates with the models' robustness to UAPs. We consider LeNet with

Jacobian regularization ($\lambda_{\text{JR}} = 0.05$) and without ($\lambda_{\text{JR}} = 0.0$). The performance of the models on the test sets is the same as the ones in Table 1. For each dataset, we take a random subset of 1,000 test set images with a uniform distribution on the output classes. Thus, we measure the similarity for a million input pairs.

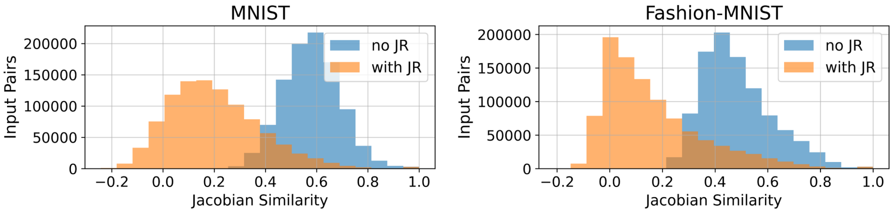

Fig. 4. Jacobian similarity for pairs of inputs on MNIST (left) and Fashion-MNIST (right) for LeNet with and without Jacobian regularization (JR). Median similarity values on MNIST are 0.18 and 0.58; and on Fashion-MNIST are 0.11 and 0.46 with and without JR respectively.

Figure 4 shows the histogram of the similarity values for the generated random pairs (cosine similarity is bounded in $[-1, 1]$). We observe that Jacobian regularization significantly reduces the median of the distributions by around 0.35. Although the Jacobian is only a first-order derivative, this greatly correlates with the models' robustness even for iterative stochastic gradient descent UAP attacks. This shows that observing the similarity measure we introduced can help to analyze the strength of shared adversarial perturbations, allowing defenders to better evaluate model robustness against UAPs.

5 Conclusion

In this work, we are the first to derive upper bounds on the impact of UAPs, we theoretically show and then empirically verify that data-dependent Jacobian regularization significantly reduces the effectiveness of UAPs, and finally we propose cosine similarity of Jacobians to measure the strength of shared adversarial perturbation between inputs.

In contrast to input-specific adversarial examples which have been shown to be difficult to defend against and often incur a notable decline in accuracy to achieve robustness, we show that Jacobian regularization can greatly mitigate the effectiveness of UAPs whilst maintaining clean performance through theoretical bounds and comprehensive empirical results.

These results give us confidence that applying Jacobian regularization to existing models significantly improves robustness to practical and realistic universal attacks at minimal cost to clean accuracy. Additionally, the proposed similarity metric for Jacobians can be used to further diagnose and analyze the vulnerability of models by identifying subsets of inputs with shared adversarial perturbations. Overall, these enable us to put defenses for neural networks against realistic and systemic UAP attacks on a more practical footing.

References

1. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 387–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_25
2. Brown, T.B., Mané, D.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Cao, Y., et al.: Adversarial sensor attack on lidar-based perception in autonomous driving. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 2267–2281 (2019)
4. Co, K.T., Muñoz González, L., de Maupeou, S., Lupu, E.C.: Procedural noise adversarial examples for black-box attacks on deep convolutional networks. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 275–289. CCS 2019 (2019). https://doi.org/10.1145/3319535.3345660
5. Co, K.T., Muñoz-González, L., Kanthan, L., Glocker, B., Lupu, E.C.: Universal adversarial robustness of texture and shape-biased models. arXiv preprint arXiv:1911.10364 (2019)
6. Co, K.T., Muñoz-González, L., Lupu, E.C.: Sensitivity of deep convolutional networks to gabor noise. arXiv preprint arXiv:1906.03455 (2019)
7. Eykholt, K., et al.: Physical adversarial examples for object detectors. In: 12th USENIX Workshop on Offensive Technologies (*WOOT* 18) (2018)
8. Eykholt, K., et al.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1625–1634 (2018)
9. Hau, Z. Co, K.T., Demetriou, S., Lupu, E.C.: Object removal attacks on lidar-based 3d object detectors. arXiv preprint arXiv:2102.03722 (2021)
10. Hau, Z., Demetriou, S., Muñoz-González, L., Lupu, E.C.: Ghostbuster: Looking into shadows to detect ghost objects in autonomous vehicle 3d sensing. arXiv preprint arXiv:2008.12008 (2020)
11. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Magazine **29**(6), 82–97 (2012)
12. Hoffman, J., Roberts, D.A., Yaida, S.: Robust learning with jacobian regularization. arXiv preprint arXiv:1908.02729 (2019)
13. Jakubovitz, D., Giryes, R.: Improving DNN robustness to adversarial attacks using jacobian regularization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 514–529 (2018)
14. Khrulkov, V., Oseledets, I.: Art of singular vectors and universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8562–8570 (2018)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1097–1105 (2012)
16. Labaca-Castro, R., Muñoz-González, L., Pendlebury, F., Rodosek, G.D., Pierazzi, F., Cavallaro, L.: Universal adversarial perturbations for malware. arXiv preprint arXiv:2102.06747 (2021)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)

18. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1765–1773 (2017)
19. Mummadi, C.K., Brox, T., Metzen, J.H.: Defending against universal perturbations with shared adversarial training. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4928–4937 (2019)
20. Novak, R., Bahri, Y., Abolafia, D.A., Pennington, J., Sohl-Dickstein, J.: Sensitivity and generalization in neural networks: an empirical study. In: International Conference on Learning Representations (2018)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
22. Roth, K., Kilcher, Y., Hofmann, T.: Adversarial training is a form of data-dependent operator norm regularization. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
23. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S., Goldstein, T.: Universal adversarial training. arXiv preprint arXiv:1811.11304 (2018)
24. Sokolić, J., Giryes, R., Sapiro, G., Rodrigues, M.R.: Robust large margin deep neural networks. IEEE Trans. Signal Process. **65**(16), 4265–4280 (2017)
25. Szegedy, C., et al.: Intriguing properties of neural networks. In: Proceeding of the International Conference on Learning Representations (ICLR) (2014)
26. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: CVPRW: Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (2019)
27. Tramèr, F., Dupré, P., Rusak, G., Pellegrino, G., Boneh, D.: Adversarial: Perceptual ad blocking meets adversarial machine learning. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, pp. 2005–2021 (2019). https://doi.org/10.1145/3319535.3354222
28. Tu, J., et al.: Physically realizable adversarial examples for lidar object detection. arXiv preprint arXiv:2004.00543 (2020)
29. Varga, D., Csiszárik, A., Zombori, Z.: Gradient regularization improves accuracy of discriminative models. arXiv preprint arXiv:1712.09936 (2017)
30. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)