Improving Adversarial Robustness to Sensitivity and Invariance Attacks With Deep Metric Learning (Student Abstract)

Anaelia Ovalle*, Evan Czyzycki *, Cho-Jui Hsieh

University of California, Los Angeles
{anaelia, eczy, chohsieh}@cs.ucla.edu

Abstract

Intentionally crafted adversarial samples have effectively exploited weaknesses in deep neural networks. A standard method in adversarial robustness assumes a framework to defend against samples crafted by minimally perturbing a sample such that its corresponding model output changes. These *sensitivity* attacks exploit the model's sensitivity toward task-irrelevant features. Another form of adversarial sample can be crafted via *invariance* attacks, which exploit the model underestimating the importance of relevant features. Previous literature has indicated a tradeoff in defending against both attack types within a strictly ℓ_p bounded defense. To promote robustness toward both types of attacks beyond existing adversarial training methods, we propose circumventing Euclidean norms with a regularized angular loss function to better distinguish between natural samples and adversarial samples. Our preliminary results indicate that regularizing over invariant perturbations in our framework improves combined invariant and sensitivity defense.

Introduction

Adversarial robustness can be motivated by the canonical comparison between two seemingly identical images of a panda. One image, however, contains an imperceptibly small perturbation that results in a completely different image classification (Goodfellow, Shlens, and Szegedy 2015). Such adversarial attacks exploit a model's sensitivity to features that it considers highly important to the learning task but are actually of little significance (Tramèr et al. 2020). However, a less studied class of adversarial samples exploits a model's invariance to relevant features.

Ensuring safety in machine learning algorithms requires the field of adversarial robustness to be keen on examining new forms of attack and subsequent mitigation strategies. Tramèr et al. (2020) explore invariance attacks and observe a fundamental tradeoff in defending sensitivity and invariance attacks when considering the neural networks as presented in Jacobsen et al. (2020). Furthermore, Tramèr et al. (2020) find that using common adversarial training frameworks that rely on ℓ_p perturbations to improve robustness toward sensitivity attacks necessarily worsens robustness toward invariance attacks. Augmenting training solely with these constraints in Euclidean space causes the model to become increasingly invariant towards task-relevant features.

Navigating this trade-off has not been explored in non-Euclidean spaces. In this study, we propose an adversarial framework that deviates from creating perturbations under Euclidean norms. To do this, our model learns a geometrically-meaningful distance metric using an angular loss. This approach allows us to produce adversarial samples anchored outside a Euclidean ℓ_p bounded ball, which allows us to simultaneously regularize over both invariance and sensitivity attacks.

Sensitivity and Invariance Adversarial Attacks

Let us consider a classification task with samples $(x, y) \in \mathbb{R}^d x\{1, ..., C\} \sim D$. Let us also consider a ground truth labeling oracle $\mathcal{O} : \mathbb{R}^d \rightarrow \{1, ..., C\}$.

Definition 1 (Sensitivity Adversarial Example) *Given some classifier f, and a correctly classified input $(s, y) \sim D$, an ϵ-bound sensitivity adversarial example is an input $x^* \in \mathbb{R}^d$ such that:*

1. $f(x^*) \neq f(x)$.
2. $||x^* - x|| \leq \epsilon$.

Definition 2 (Invariance Adversarial Example) *Given some classifier f, and a correctly classified input $(s, y) \sim D$, an ϵ-bound invariance adversarial example is an input $x^* \in \mathbb{R}^d$ such that:*

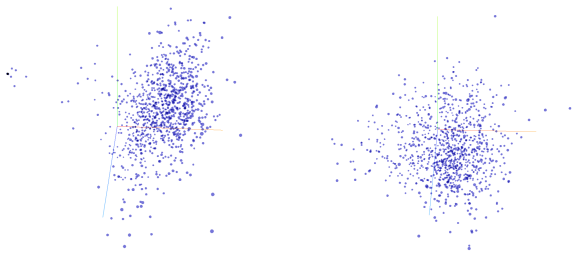
1. $f(x^*) = f(x)$.
2. $\mathcal{O}(x^*) \neq \mathcal{O}(x)$ *and* $\mathcal{O}(x^*) \neq \perp$.
3. $||x^* - x|| \leq \epsilon$.

Note that the above formulation and definitions mirror those found in Tramèr et al. (2020). A significant assumption required for Definition 1 is that for all x and associated perturbations x^*, if $||x^* - x|| \leq \epsilon$, then $\mathcal{O}(x^*) = \mathcal{O}(x)$. Informally, perturbations of magnitude less than ϵ preserve the oracle's labelling. As shown in Tramèr et al. (2020), it is precisely the violation of this assumption that results in a fundamental tradeoff between robustness toward these two types of adversarial attacks.

*These authors contributed equally.

Adversarial Training Method	A_{Orig}	A_{SA}	A_{IA}
FGSM (Baseline)	99.02	98.95	85.67
Baseline + ML_{SA}	99.38	**99.17**	82.49
Baseline + ML_{SA} + ML_{IA}	99.09	98.98	**87.80**

Table 1: Average Accuracy (A) over MNIST data across three randomly seeded runs. ML is our metric learning norm that uses different adversarial samples. SA and IA indicate sensitivity or invariance samples, respectively. Regularizing with angular triplet loss for both sensitivity and invariance attacks improves accuracy over invariance samples with minimal impact on sensitivity accuracy.



(a) No Invariance (b) With Invariance

Figure 1: PCA plots showing difference in FGSM-perturbed images with invariance regularizer.

Adversarial Metric Learning Framework

We employ a metric learning framework that learns a distance measure over embeddings in angular space. Previous results reflect a more flexible adversarial optimization framework (Duan et al. 2018). While this study investigates the use of metric learning in traditional adversarial defense against sensitivity attacks, our study builds upon this framework to defend against invariance attacks.

Our loss function, L_t, is defined as the classic triplet loss found in (Mao et al. 2019), which creates a fixed margin between the differences in the anchor sample and positive and negative examples respectively. We define the distance, $D(\cdot)$, as the angular distance between two samples in order to encode the information in the angular metric space.

$$D(h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_{p,n}^{(i)})) = 1 - \frac{|h(\mathbf{x}_a^{(i)}) \cdot h(\mathbf{x}_{p,n}^{(j)})|}{\|h(\mathbf{x}_a^{(i)})\|_2 \|h(\mathbf{x}_{p,n}^{(j)})\|} \quad (1)$$

We derive an adversarial training framework using the below loss term by considering the anchor sample to be a natural image x_a, a positive example to be a perturbed image x_p, and a negative sample to be an image x_n from a different class. We construct our loss function by including a sensitivity triplet loss regularization term, an invariance triplet loss regularization term, and a feature norm. This forces adversarial and natural samples closer together in learned space.

$$\begin{aligned} L_{all} = & \sum_{i}^{N} L_{ce}(f(\mathbf{x}_a^{(i)}), y^{(i)}) \\ & + \lambda_1 L_{sa}((h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_p^{(i)}), h(\mathbf{x}_n^{(i)})) \\ & + \lambda_2 L_{ia}((h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_q^{(i)}), h(\mathbf{x}_n^{(i)})) \\ & + \sum_{t \in S} \|h(\mathbf{x}_t)\|_2 \end{aligned} \quad (2)$$

L_{ce} is cross entropy loss, L_{sa} uses sensitivity attacks for positive class, L_{ia} uses invariance attacks for positive class, x_p and x_q are sensitivity and invariance perturbed samples respectively, and $\lambda_1, \lambda_2, \lambda_3$ are coefficients $\in \mathbb{R}^+$.

Experiments and Discussion

In our experiments we generate sensitivity attacks using FGSM (Goodfellow, Shlens, and Szegedy 2015) and invariance attacks using the method described in Tramèr et al. (2020). We generate a single sensitivity attack and a single invariance attack for each sample in the MNIST dataset.

Our results in Table 1 indicate that regularizing with sensitivity and invariance attacks using an angular triplet loss can improve performance against invariance attacks with minimal loss in accuracy over sensitivity attacks. Our model trained with sensitivity and invariance regularization outperforms the adversarial baseline which uses ℓ_p-bound norms.

After each model is trained, we extract the penultimate layer and examine the learned embedding space with PCA. Figure 1a shows the distribution of adversarial images without invariance regularization. In comparison, when the invariance triplet regularizer is added, it is shown to be more tightly grouped and circular (1b). This indicates our model's ability to better identify perturbed samples because they're grouped together. Overall these results imply that models may be trained to resist both sensitivity and invariance attacks without significantly sacrificing performance in one or the other. For future work, we plan to expand this analysis to more datasets and adversarial attacks.

References

Duan, Y.; Zheng, W.; Lin, X.; Lu, J.; and Zhou, J. 2018. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2780–2789.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

Jacobsen, J.-H.; Behrmann, J.; Zemel, R.; and Bethge, M. 2020. Excessive Invariance Causes Adversarial Vulnerability. arXiv:1811.00401.

Mao, C.; Zhong, Z.; Yang, J.; Vondrick, C.; and Ray, B. 2019. Metric Learning for Adversarial Robustness. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tramèr, F.; Behrmann, J.; Carlini, N.; Papernot, N.; and Jacobsen, J.-H. 2020. Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations. arXiv:2002.04599.