Geometric Universality of Adversarial Examples in Deep Learning

Haosheng Zou [1] **Hang Su** [1] **Tianyu Pang** [1] **Jun Zhu** [1]

Abstract

We consider the problem of adversarial examples in deep learning and attempt to provide geometric insights on their universality. Specifically, we define *adversarial directions* and prove relevant results towards universality of adversarial examples with few theoretical assumptions. Our results raise attention to fully-connected layers as the last layer of most neural networks, which may be prone to adversarial examples, demanding further research in this regard. A longer version with full proofs and discussions is provided with the submission email and also here.

Consider the softmax regression layer at the end of many popular neural networks for visual classification tasks (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2016; He et al., 2016) and the hidden space of the input neurons to the softmax layer. Denote the hidden space of the input to softmax layer $\mathcal{H} \subseteq \mathbb{R}^l$, and let $\boldsymbol{h} \in \mathcal{H}$ be this input vector, and l is the number of neurons in the final hidden layer. We further denote m the number of classes. We define softmax function $\mathbb{S}(\boldsymbol{z}) : \mathbb{R}^m \mapsto \mathbb{R}^m$ as $\mathbb{S}(\boldsymbol{z})$ where $\boldsymbol{z}$ is the logits. Then the overall softmax layer could be denoted $\mathbb{S}(W^T\boldsymbol{h} + \boldsymbol{b})$. The neural network classifier first maps input images $\boldsymbol{x}$ to the hidden representation $\boldsymbol{h}$ with the complex multi-layer non-linear function g: $\mathcal{X} \mapsto \mathcal{H}$, $\boldsymbol{h} = g(\boldsymbol{x})$, and then perform softmax regression to obtain a predicted label $y = \arg\max_{i\in[m]} \mathbb{S}(W^T\boldsymbol{h} + \boldsymbol{b})_i$. We only show results with the case $\mathcal{H} = \mathbb{R}^l$ here.

Definition 1. *(Adversarial Direction) An* **adversarial direction** *is defined on any $\boldsymbol{h} \in \mathcal{H}$ as a direction $\boldsymbol{d}$ such that $\forall\theta \in \mathbb{R}$, $\mathbb{S}(W^T(\boldsymbol{h} + \theta\boldsymbol{d}) + \boldsymbol{b}) = \mathbb{S}(W^T\boldsymbol{h} + \boldsymbol{b})$, i.e., arbitrarily traversing along $\boldsymbol{d}$ preserves the softmax output.*

Such directions are adversarial in that no output difference could be observed with input manipulation to any degree along them, which opens a wide range in $\mathcal{H}$ for potential adversarial examples. We further assume $l > m$, which is the case in almost all top-performing neural networks on ImageNet (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2016; He et al., 2016) with $l = 4096$ (or $l = 2048$) and $m = 1000$ for the 1000 classes.

[1]Department of Computer Science and Technology, Tsinghua University, Beijing. Correspondence to: Haosheng Zou <zouhs16@mails.tsinghua.edu.cn>.

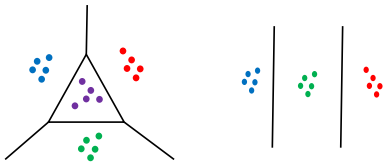

Figure 1. Left: illustration of a decision region (purple ones) that does not extend to ∞ when $l < m$. We show that in popular architectures, $l > m$ and *all* decision regions extend to ∞, facilitating adversarial examples. Right: "parallel" softmax layers where two of the three decision regions determined by the pair of parallel hyperplanes are not adjacent (red and blue ones), making adversarial examples between the two classes harder to find.

Theorem 1. *$\forall\boldsymbol{h} \in \mathcal{H}$, $\exists V \subseteq \mathbb{R}^l$ with $\dim(V) \geq l - m$, s.t. $\forall\boldsymbol{d} \in V$, $\forall\theta \in \mathbb{R}$, $\mathbb{S}(\boldsymbol{h} + \theta\boldsymbol{d}) = \mathbb{S}(\boldsymbol{h})$.*

Theorem 2. *(Universality of Adversarial Directions) For any region in $\mathcal{H}$ that's classified as a certain class, there always exists at least one direction, infinitely far along which the points are still classified as the same class with identical output probabilities (contrary to Fig. 1, left).*

For most g and softmax layers, we may even conjecture we could find adversarial examples for *any* data pair.

Definition 2. *A softmax layer is* **parallel** *if at least one pair of its decision boundries is parallel (Fig. 1, right).*

Conjecture. *(Universality of Adversarial Examples) For most multi-layer non-linear mappings $g : \mathcal{X} \mapsto \mathcal{H}$ and non-"parallel" softmax layers $(W, \boldsymbol{b})$,* **for any** *data pair $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ where $y \neq y'$, there exists an imperceptible adversarial example $\boldsymbol{x}^*$ with $\|\boldsymbol{x}^* - \boldsymbol{x}\|_p \leq \epsilon$ for an imperceptible ϵ and $g(\boldsymbol{x}^*) = g(\boldsymbol{x}') + \theta\boldsymbol{d}'$, where $\theta \in \mathbb{R}$ and $\boldsymbol{d}'$ is an* adversarial direction *of class y' in space $\mathcal{H}$.*

Significance and Implications: We provide a deeper understanding of softmax regression widely used without question in neural networks. The decision boundary of softmax regression is piece-wise linear, and the decision region for each class is convex, which makes softmax regression simple and expectedly robust enough. However, we show that the decision regions are generally unconstrained, probably leading to the universality of adversarial examples combined with the non-linear preceding layers. Little work has been done on the final classification layer. Serving as theoretical evidence for some preliminary work already looking for substitute classification layers (Pang et al., 2018), this paper raises attention to softmax layers on adversarial robustness.

References

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *CVPR*, 2016.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Pang, Tianyu, Du, Chao, and Zhu, Jun. Max-mahalanobis linear discriminant analysis networks. *CoRR*, abs/1802.09308, 2018.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jonathon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.