On Trace of PGD-Like Attacks

Mo Zhou Vishal M. Patel

Johns Hopkins University

arXiv Github

Introduction

Adversarial attacks (e.g., PGD) pose security concerns to deep learning applications, but their characteristics are under-explored.

Problem setting: to identify the trace of (PGD-like) adversarial attacks, given an already-trained deep neural network, a tiny set (e.g., 50) training data, without any change in network architecture or weights, nor any auxiliary deep networks. (more practical in realistic scenarios, such as federated learning where original training data is inaccessible.)

Assumption: Strong adversarial attacks leaves a "strong trace" in the adversarially attacked image, based on the observation that PGD-like attacks manifest a "local linearity" characteristics.

Contributions: We present the Adversarial Response Characteristics (ARC) features to identify and characterize the uniqe trace (named Sequel Attack Effect, or SAE) of PGD-like attacks from adversarial examples. Our proposed method survives the tough problem setting.

Experiments: ResNet-18 on CIFAR-10, ResNet-152/SwinT-B on ImageNet.

Our Method

Based on FGSM paper, adversarial attack triggers the "local linearity" of the neural network around the input. So we approximate the network $\boldsymbol{f}(\cdot)$ with the first-order Taylor expansion around the input $\tilde{\boldsymbol{x}} \triangleq \boldsymbol{x} + \boldsymbol{r}$:

$$f_n(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}) \approx f_n(\tilde{\boldsymbol{x}}) + \boldsymbol{\delta}^T \nabla f_n(\tilde{\boldsymbol{x}}), \forall n \in \{1, 2, \dots, N\},$$

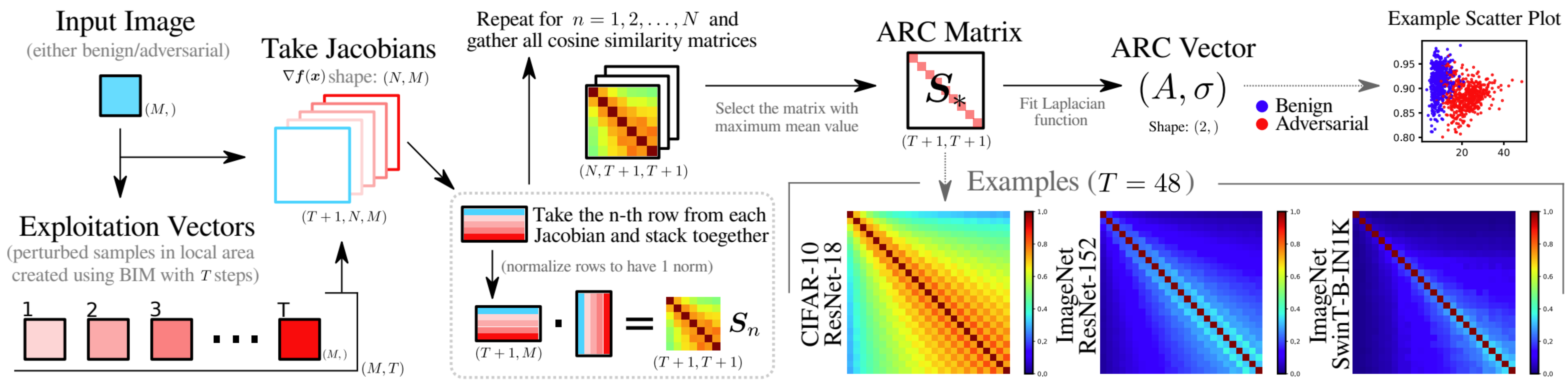
where $f_n(\cdot)$ is the n-th element of the vector function $\boldsymbol{f}(\cdot)$.

Adversarial Response Characteristics (ARC): Given an arbitrary input, either benign or adversarially attacked. We adversarially perturb this image with BIM, and then measure the model's gradient direction consistency along the trajectory with cosine similarity as ARC matrix:

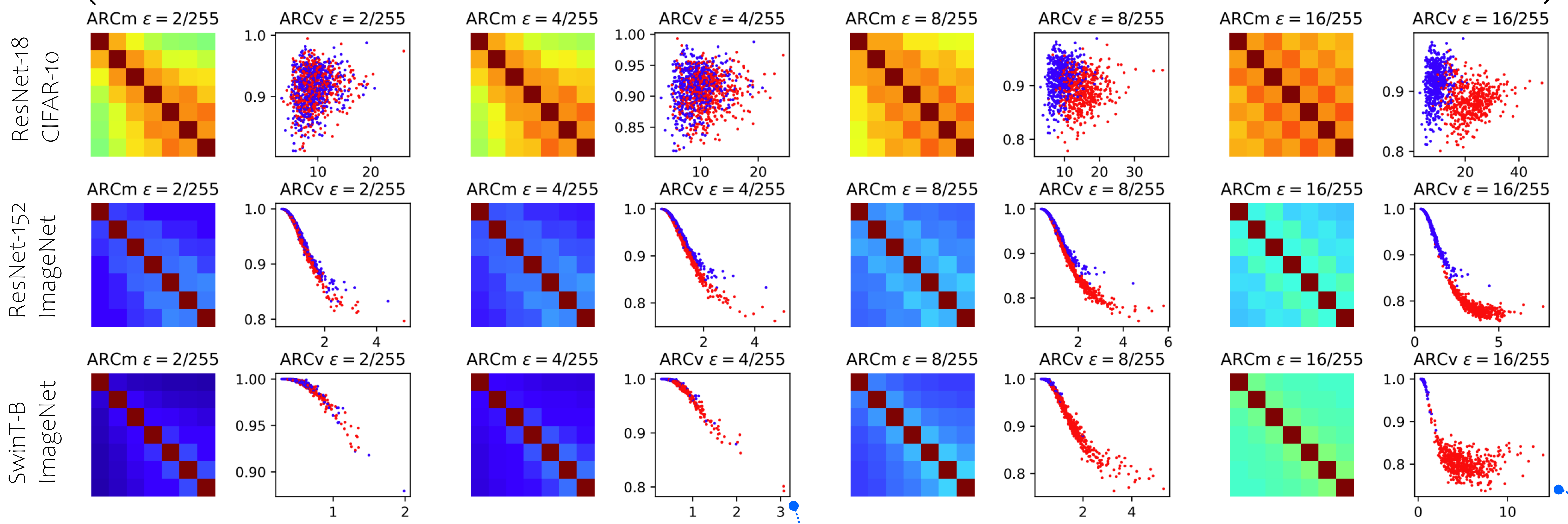
$$\boldsymbol{\delta}_{t+1} \leftarrow \mathrm{Clip}_\Omega\Big(\boldsymbol{\delta}_t + \alpha\, \mathrm{sign}[\nabla L_{\mathrm{CE}}(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}_t, \check{c}(\boldsymbol{x}))]\Big),$$

$$s_n^{(i,j)} = \cos\left[\nabla f_n(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}_i), \nabla f_n(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}_j)\right].$$

Sequel Attack Effect (SAE): if input is benign, the attack turns the model from "non-linear" to "linear". If the input is adversarially attacked, the model is turned from "linear" to "very linear".



ARC Features of BIM Attack



Weak adversarial attacks: with the attack step increasing (from top-left to bottom-right), the model turns from "non-linear" to moderately "linear" (reflected by higher cosine similarity). Weak attacks are not very separatable in the ARC vector space.

Strong adversarial attacks: with the attack step increasing (from top-left to bottom-right), the model turns from "linear" to "very linear" (reflected by higher cosine similarity). Strong attacks make the ARC feature clearly separatable in the ARC vector space.

ARC Features of Other PGD-Like Attacks