RELATED WORK - NON-DATA-RELATED REASONS BEHIND EVASION ATTACKS

There has been a variety of hypotheses regarding the reasons behind adversarial vulnerability of ML systems, particularly for evasion attacks. These include data-related properties extensively discussed in this survey, as well as reasons related to the models themselves, computational resources, and feature learning procedures. We discuss these below.

Model. When Szegedy et al. [9] first discovered adversarial examples for visual models, they suspected that the high non-linearity of DNNs resulted in low probability 'pockets' of adversarial examples in the learned representation manifold. They hypothesize that while these pockets can be found through attack algorithms, the samples residing in these pockets have different distributions compared to normal samples and are thus subsequently harder to find when randomly sampling from the input space. Instead, Goodfellow et al. [4] hypothesize that the linearity from activation functions, like ReLU and sigmoid found in high-dimensional neural networks, induce vulnerability towards adversarial perturbations. To support their claim, they present the attack method FGSM that exploits the linearity of the target classifier. Fawzi et al. [3] also argue against the hypothesis of high non-linearity as the cause for adversarial examples. They show that all classifiers are susceptible to adversarial attacks and claim that it is the low flexibility of the classifier compared to the complexity of the classification task that results in vulnerability. The lack of consensus on the primary causes of model vulnerability invites more studies on this topic.

Singla et al. [8] show that enforcing invariance to circular shifts (e.g., rotation) in neural networks induces decision boundaries with a smaller margin than normal, fully connected networks, which, in turn, reduces the adversarial robustness of the model. Moosavi-Dezfooli et al. [7] introduce universal, input-agnostic perturbations to mislead the classifier and hypothesize that the vulnerability of a multi-class classifier to such perturbations is related to the shape of its decision boundaries, e.g., linear classifiers with decision boundaries that are parallel to each other and nonlinear classifier with decision boundaries that are curved in a similar way tend to be less robust as perturbations in one direction can change the prediction label for a different class.

Tanay and Griffin [10] conjecture that the decision boundary learned by the classifier being too close to (or 'tilted towards') the data manifold instead of being perpendicular to it, results in small perturbations being sufficient to move samples across the decision boundary for misclassification.

Computational Resources. Bubeck et al. [1] use computational hardness theory to show that the time complexity for learning a robust model is exponential to the size of input data and thus is computationally intractable. Hence, they attribute adversarial vulnerability to computational limitations of current learning algorithms. Degwekar et al. [2] further extend this work and also show the impossibility of efficiently training robust classifiers.

Feature Learning. Ilyas et al. [5] show that adversarial vulnerability can be a consequence of a model exploiting well-generalizing but non-robust features, i.e., features that are spurious and sometimes incomprehensible to humans; when constraining the model to use robust features, the adversarial robustness increases together with the interpretability of the learned features. However, Tsipras et al. [11] note that, as the features for achieving high accuracy may be different from the ones for achieving high robustness, robustness may be at odds with standard accuracy.

Instead of seeing adversarial vulnerability as a product of classifiers being overly sensitive to changes in spurious features, Jacobsen et al. [6] hypothesize that classifiers can rather be overly insensitive to relevant semantic information, e.g., images with drastically different content can share similar latent representations. The authors introduce a new type of adversarial examples that exploit such insensitivity, where the content of images is altered without changing the resulting prediction label.

While all these works propose possible reasons for adversarial vulnerabilities, they are orthogonal to our survey, which focuses particularly on the influence of training data.

REFERENCES

[1] Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. 2019. Adversarial Examples from Computational Constraints. In *International Conference on Machine Learning (ICML)*. 831–840.

[2] Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. 2019. Computational Limitations in Robust Classification and Win-Win Results. In *Conference on Learning Theory (COLT)*. 994–1028.

[3] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2015. Fundamental Limits on Adversarial Robustness. In *ICML Workshop on Deep Learning*.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

[5] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Mądry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems (NeurIPS)*. 125–136.

[6] Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. 2019. Excessive Invariance Causes Adversarial Vulnerability. In *International Conference on Learning Representations (ICLR)*.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. 2018. Robustness of Classifiers to Universal Perturbations: A Geometric Perspective. In *International Conference on Learning Representations (ICLR)*.

[8] Vasu Singla, Songwei Ge, Ronen Basri, and David Jacobs. 2021. Shift Invariance Can Reduce Adversarial Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1858–1871.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations (ICLR)*.

[10] Thomas Tanay and Lewis D. Griffin. 2016. A Boundary Tilting Perspective on the Phenomenon of Adversarial Examples. *ArXiv* (2016).

[11] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations (ICLR)*.