Is AmI (Attacks Meet Interpretability) Robust to Adversarial Examples?

Nicholas Carlini (*Google Brain*)

***Abstract*—No.**

I. Attacking "Attacks meet Interpretability"

AmI (Attacks meet Interpretability) is an "attribute-steered" defense [3] to detect [1] adversarial examples [2] on face-recognition models. By applying interpretability techniques to a pre-trained neural network, AmI identifies "important" neurons. It then creates a second augmented neural network with the same parameters but increases the weight activations of important neurons. AmI rejects inputs where the original and augmented neural network disagree.

We find that this defense (presented at at NeurIPS 2018 as a spotlight paper—the top 3% of submissions) is completely ineffective, and even *defense-oblivious*[1] attacks reduce the detection rate to 0% on untargeted attacks. That is, AmI is no more robust to untargeted attacks than the undefended original network. Figure 1 shows selected adversarial examples that fool the AmI defense. We are incredibly grateful to the authors for releasing their source code[2] which we build on[3]. We hope that future work will continue to release source code by publication time to accelerate progress in this field.

A. Evaluation

We assume familiarity with prior work (specifically [1]–[3]).

Unfortunately, the defense paper [3] does not contain a threat model or make any *specific* claims about robustness, making it difficult to perform a proper security evaluation. The authors state the bound was meant to be 0.01 ($3\times$ *lower* than the $8/255 \approx 0.031$ bound used in most prior work); this extremely low distortion bound is never given in the paper.

We generate adversarial examples by completely ignoring the defense and generating high-confidence adversarial examples on the original neural network. This approach, while simple, has proven surprisingly successful in the past when attacking detection-based defenses [1]. We choose an (incorrect) target label at random and generate a high-confidence targeted adversarial example for that target using *only* the original network. We then test to see if the resulting image by chance happens to be adversarial on the combined defended model (i.e., is misclassified the same way by both networks). If it is not (and would therefore be rejected), we repeat the process and try again until we succeed. The median number of attempts is 25.

This naïve attack is successful 100% of the time: the detector has a 0% true-positive rate (*lower* than the 9.9% false positive rate); Figure 1 contains successful adversarial examples.

Fig. 1. (left) Original images; (right) adversarial examples defeating AmI.

II. Conclusion

"Attacks Meet Interpretability" [3] is not robust to untargeted adversarial examples with ℓ_∞ bound 0.01, even when the attacker is oblivious to existence of the defense. While our attack is not efficient, we believe an adaptive attack that specifically targeted the defense would be much more efficient while remaining 100% successful at evading detection.

We implore researchers who propose defenses to investigate *why* attacks fail before declaring a proposed defense effective; and similarly implore those reading or reviewing papers to think critically about why the attacks could have failed before believing the claimed defense results. Feynman wrote "the first principle [of research] is that you must not fool yourself": this is true especially in security, where we must always consider the possibility that attacks fail not because the defense is effective, but for some other unintended reason.

[1] We mount a defense-oblivious attack because it shows that even under this incredibly weak threat model the defense is ineffective. (The defense is also written in Caffe which the author did not want to have to use.) Future defenses should *not* argue security only under this threat model.

[2] https://github.com/AmIAttribute/AmI

[3] https://github.com/carlini/AmI

References

[1] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," *AISec*, 2017.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[3] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 7728–7739.