Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Logan Engstrom* Andrew Ilyas* Anish Athalye*

LabSix, Massachusetts Institute of Technology

Adversarial Logit Pairing

Robust optimization, as in Madry et al., solves:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in S} L(\theta, x+\delta, y) \right]$$

L is a loss function, $\mathcal{D}$ is the data distribution, and $\mathcal{S}$ defines the allowed perturbations.

In **Kannan et al.**, the **Adversarial Logit Pairing** method solves:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(\theta,x,y) + \lambda D\left(f(\theta,x), f(\theta, x+\delta^*)\right)\right]$$

$$\text{where } \delta^* = \arg\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)$$

D is a distance function, f maps parameters and inputs to logits, and λ is a hyperparameter.

Attack

Run PGD [Madry et al.] until convergence. For each input (x,y) we solve, in the untargeted case (S is an ϵ-norm ℓ_∞ box):

$$\max_{\delta\in S} L(x+\delta, y),$$

In the targeted case, given a target label y', we solve:

$$\min_{\delta\in S} L(x+\delta, y')$$

Evaluation

Source Defense ($\epsilon = 16$)	Kannan et al. Claimed Accuracy	this work Accuracy	this work Attacker Success
Madry et al.	1.5%	–	–
Kannan et al.	27.9%	**0.6%**	**98.6%**

Table 1: Claimed robustness of Adversarial Logit Pairing against targeted attacks on ImageNet, from Kannan et al., compared to the lower bound on attacker success rate from this work. Attacker success rate is the percentage of times an attacker successfully induces the adversarial target class; accuracy is the percentage of times the classifier outputs the *correct* class. We calculate accuracy as in Kannan et al., i.e. correct classification rate under targeted adversarial attack.

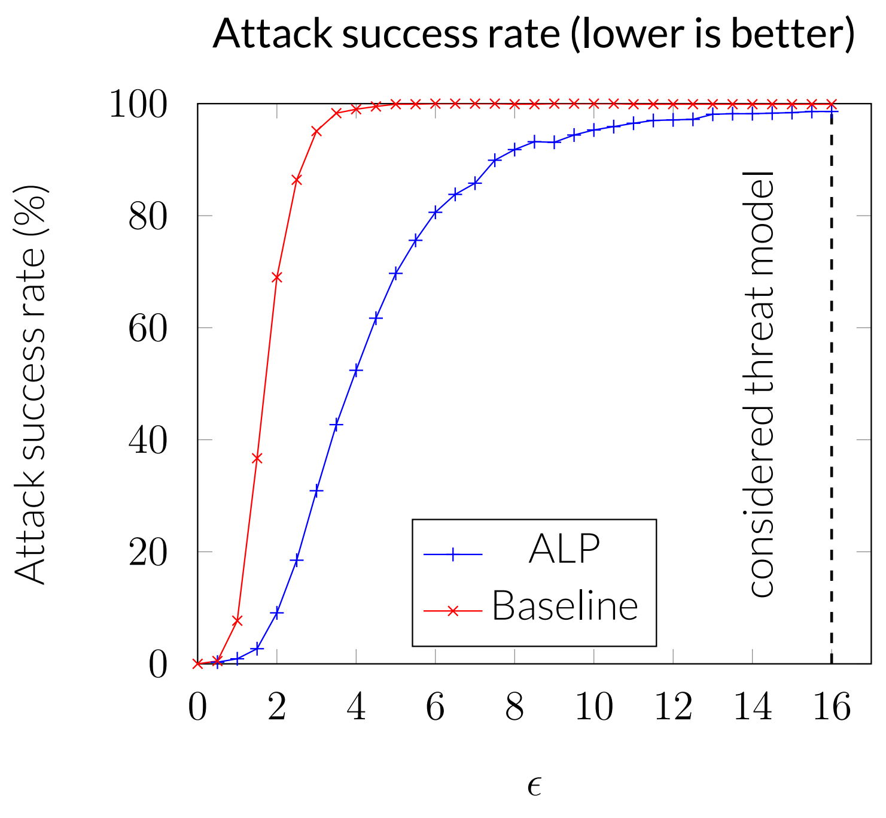

Figure 1: Comparison under targeted attack. Our attack reaches 98.6% success rate (and 0.6% correct classification rate) at $\epsilon = 16/255$.

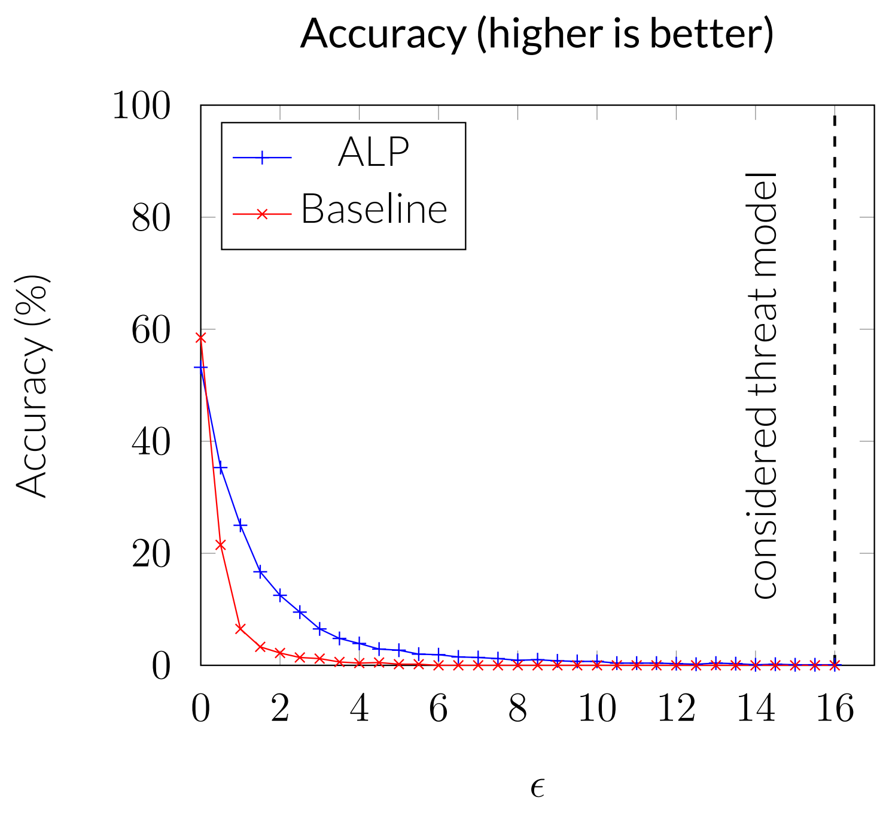

Figure 2: Comparison under untargeted attack. The ALP-trained model achieves 0.1% accuracy at $\epsilon = 16/255$.

Takeaways

Run attacks to convergence

Similar to how models are trained to convergence, attacks should be run to convergence. While there is minimal increase in robustness at $\epsilon = 16/255$, PGD can take longer to converge against ALP-trained models.

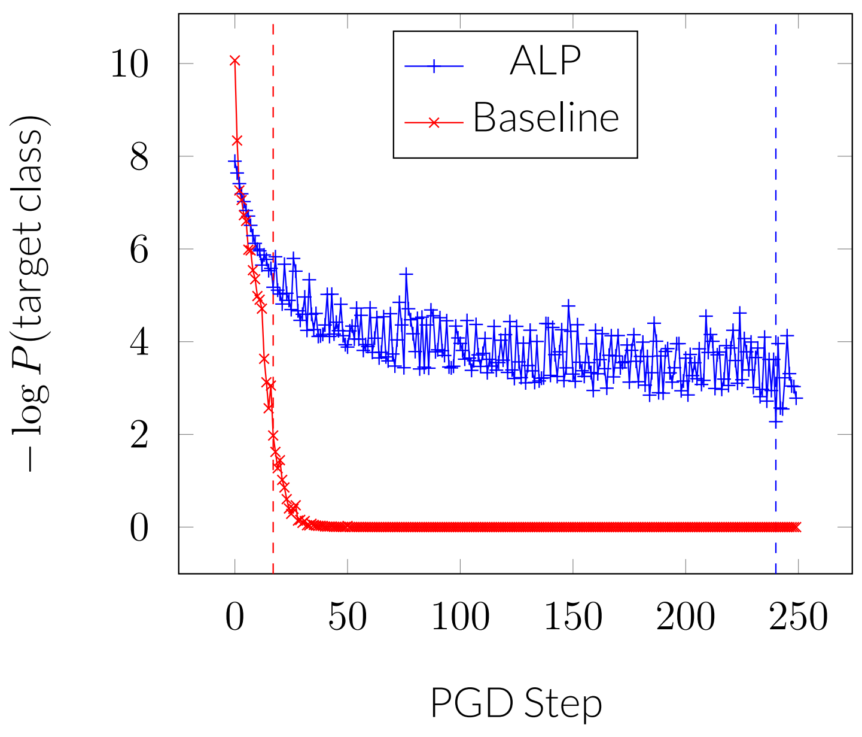

Figure 3: Targeted attack

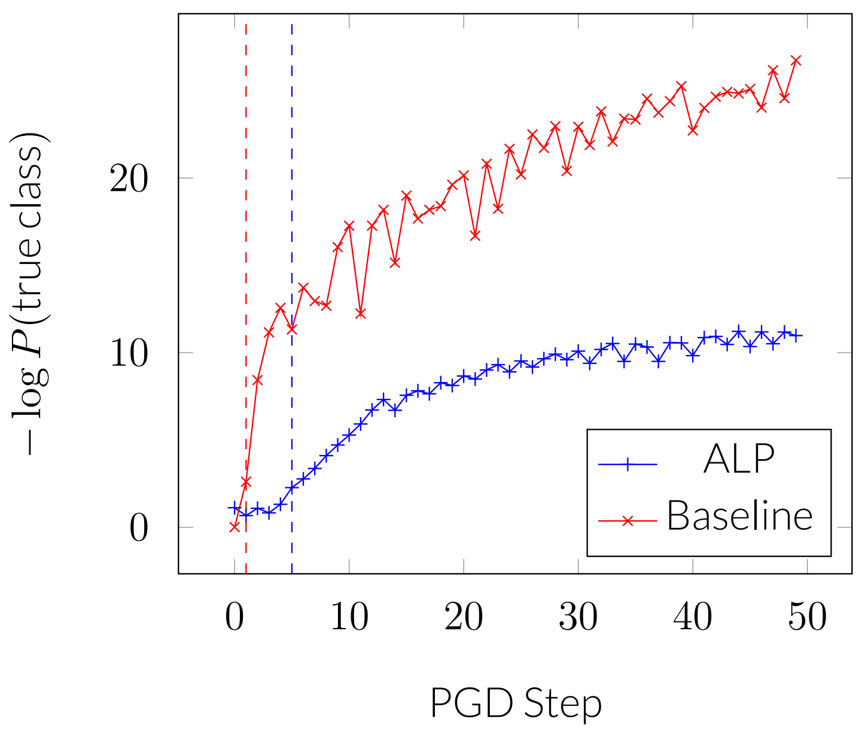

Figure 4: Untargeted attack

Examine loss landscape

Inspecting the loss landscape induced by ALP gives some insight into why a small number of PGD steps are not sufficient to find adversarial examples.

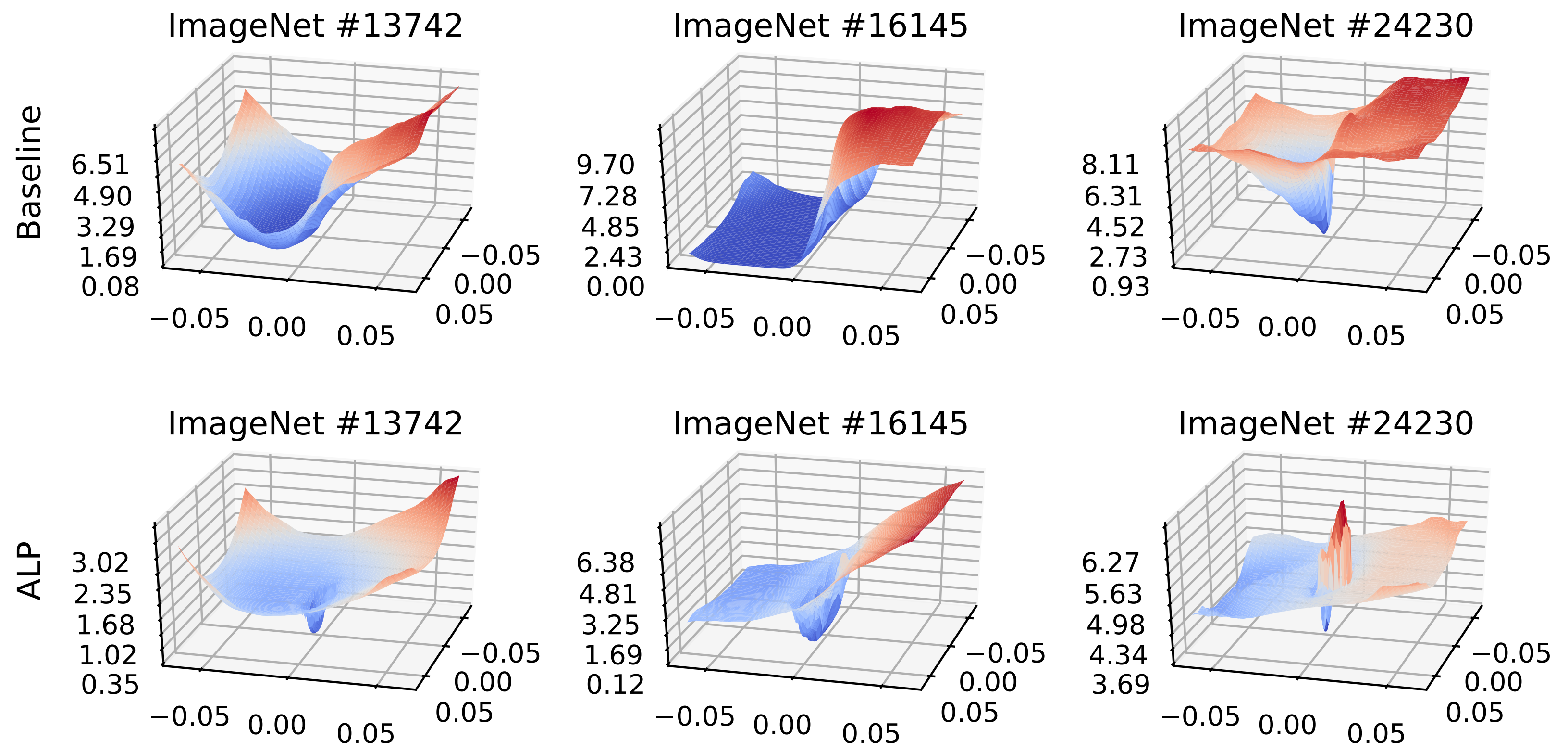

Figure 5: Comparison of loss landscapes of ALP-trained model and baseline model.

ALP sometimes induces decreased loss locally, and gives a "bumpier" optimization landscape.

Code

Source code for our analysis is available at `https://github.com/labsix/adversarial-logit-pairing-analysis`.

References

H. Kannan, A. Kurakin, and I. Goodfellow. Adversarial logit pairing. *arXiv preprint*. URL `https://arxiv.org/abs/1803.06373`.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*. URL `https://arxiv.org/abs/1706.06083`.