Numerical Aspects of Adversarial Machine Learning

Oğuz Kaan Yüksel, Advisor: İnci Meliha Baytaş
CmpE 491 Project - kaan.yuksel@boun.edu.tr

Adversarial Samples

Deliberately constructed samples that attempt to deceive machine learning models.

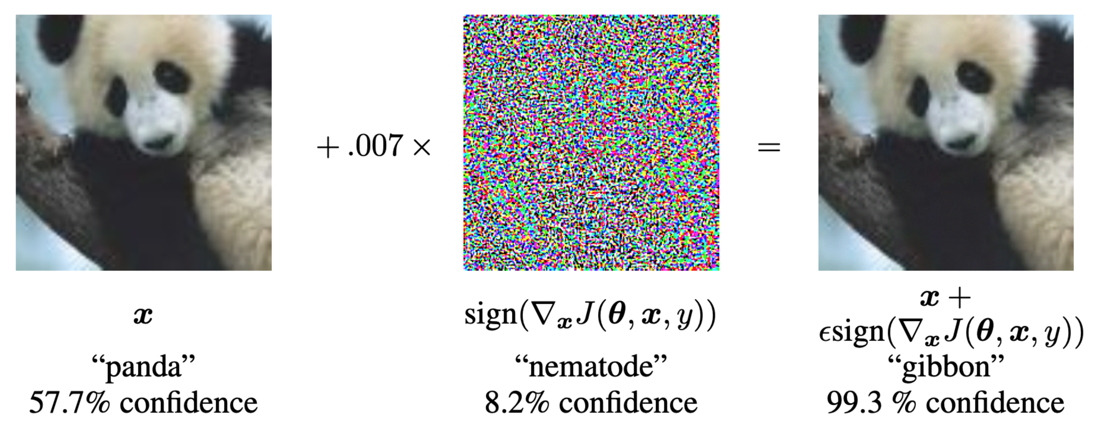

Figure 1: Adversarial panda of ImageNet on GoogLeNet[1]

- Pervasive *across* models and data domains.
- Smart *imperceptible* perturbations are enough.
- Challenges *robustness* of ML in real-world usage.
- Theoretically begs the question 'what is learning'.

Adversarial Machine Learning

Systematic study of adversarial *detection*, *attack*, *defense* and *theory* of adversarials. Saddle point formulation below, captures objectives of adversarial attack and defense together [2]:

$$min_\theta \; max_{x'} \; J(\theta, x', y) \; s.t. \; \|x' - x\|_l < \epsilon$$

where θ is model parameters, J is loss function, x is data sample, y is true label, x' is adversarial sample.

***Fast Gradient Sign Method*:** One step attack that maximizes loss linearly.

$$x' \leftarrow x + \epsilon * sign(\nabla_x J(\theta, x, y))$$

***Projected Gradient Descent*:** Iterative attack considered as *generic first-order adversary.*

Algorithm 1: *PGD* with norm $l = 0$

$x' \leftarrow x + \delta * \mathcal{N}(0, 1)$
repeat K **times**
  $x' \leftarrow FGSM(x', y, \epsilon)$
  $x' \leftarrow clip(x', [x - \epsilon_{\max}, x + \epsilon_{\max}])$
end

***Adversarial Training*:** Adding term for adversarially perturbated version of original samples.

$$J_{adv}(\theta, x, y) = \lambda J(\theta, x, y) + (1 - \lambda) J(\theta, x', y)$$

Problem Statement

Study adversarial *vulnerability* of artificial neural networks by analyzing *distinctive internal behaviors* of adversarials and develop defense heuristics against *white-box* adversaries.

Computational Regime hypothesis: Each neuron of artificial neural networks receives from and outputs to certain *numerical ranges* where operation is *meaningful*. Outside of that range, neuron might be rather *injecting noise* to the network rather than providing any information.

Activation Studies

- Analyze hidden neuron and layer activation statistics of regular vs. adversarial samples.
- Count *outlier* neuron activations per layer

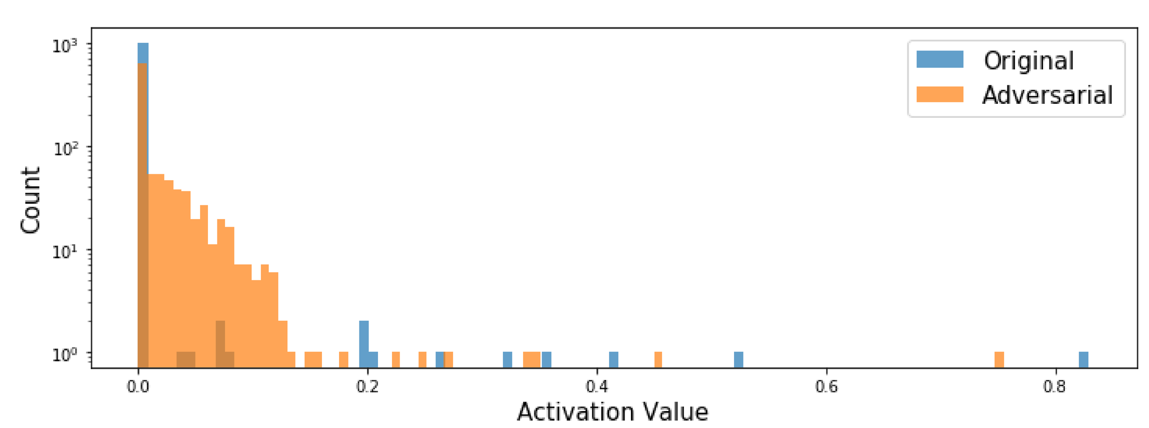

Figure 2: Activation distributions of a hidden neuron

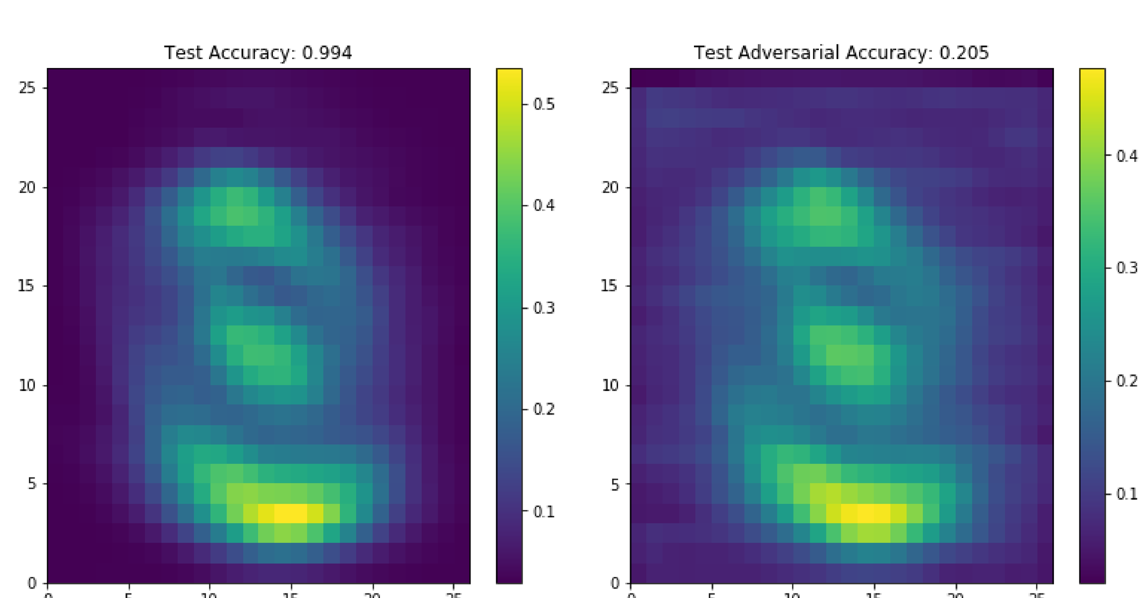

Figure 3: Mean activations of a hidden layer

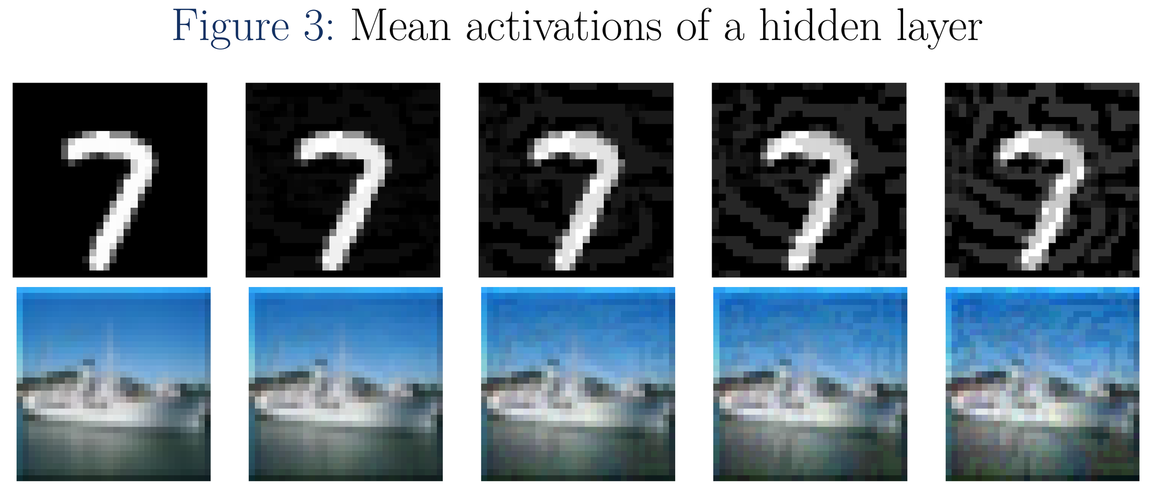

Figure 4: Examples of adversarial samples

Regularizing Adversarials

***Clipping*:** For each neuron n of layer l, determine a safe activation range $[min_n, max_n]$ possibly by activation outlier analysis. Later, this range can be fine-tuned by freezing the rest of the network and running a few more training steps. Use these ranges to form a *double-sided ReLU* after layer l.

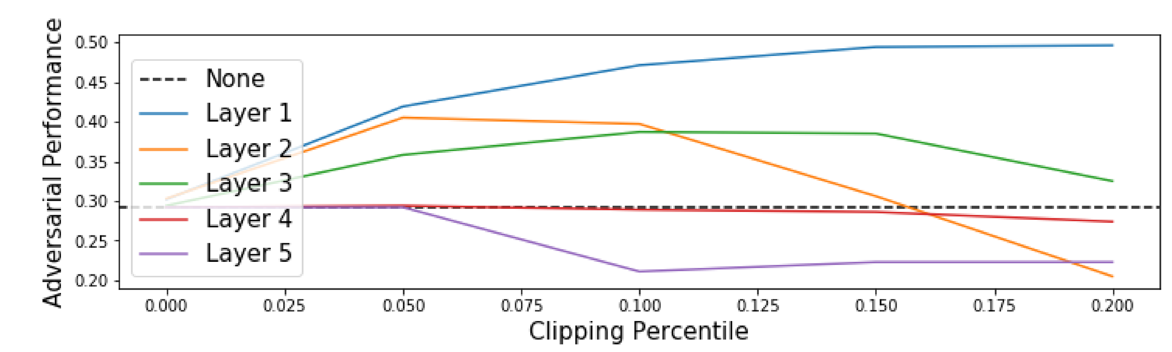

Figure 5: Clipping layer on *MNIST*

***Orthogonality Regularization*[3]:** Orthogonal weights might result in harder adversarial sample generation as perturbation **need** to propagate through less correlated hidden representations.

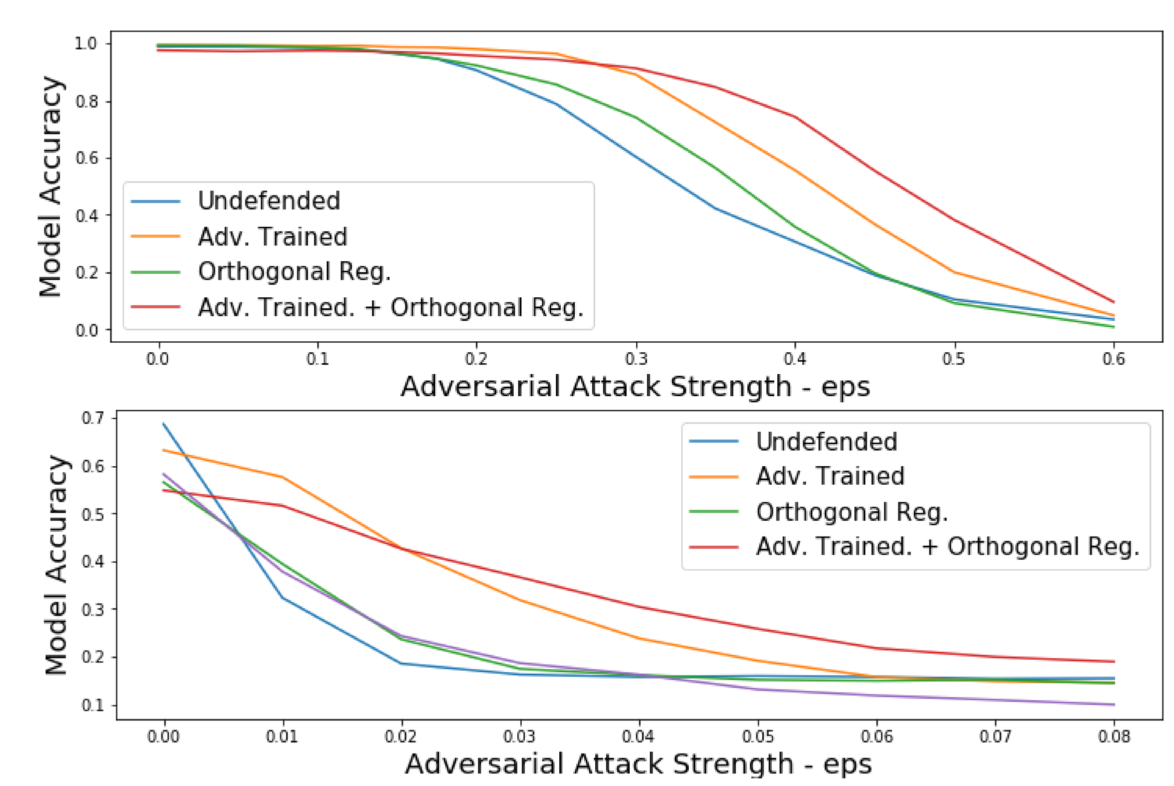

Figure 6: Orthogonal Reg. results on *MNIST* and *CIFAR10*.

***Gradient Difference Regularization*:** We add an extra penalty in adversarial training for second-order input gradients as opposed to first-order [4].

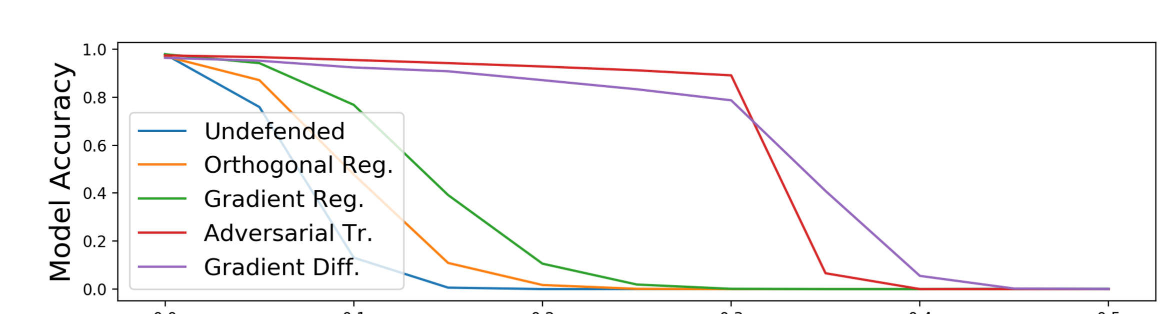

Figure 7: Gradient Difference Reg. results on *MNIST*

Results

- Activation studies support *Computational Regime* hypothesis but we are far from properly accounting all the data gathered.
- Even though Clipping is successful in *MNIST* with small *CNN*s, it **does not** scale to *CIFAR10* and more complex networks.
- Orthogonality Regularization is a much faster (no **back-propagation**!) but a much weaker defense than Adversarial Training.
- Gradient Difference Regularization extends robustness to high-adversarial settings but loses low-adversarial performance.

Future Work

- Employ *feature attribution* methods to quantify impact of deviations found in activation studies.
- Inquire whether *auto-correlation* and *internal covariate bias* in neural networks have any relation with *orthogonality* and adversarial vulnerability. More specifically, study if *orthonormality* is a feasible constraint for *CNN*s.
- Investigate full power of *Gradient Difference Regularization* and recently developed ideas such as *Adversarial Ball Training* and *orthogonal adversarial generation.*

References

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 'Explaining and harnessing adversarial examples.' arXiv preprint arXiv:1412.6572 (2014).

[2] Madry, Aleksander, et al. 'Towards deep learning models resistant to adversarial attacks.' arXiv preprint arXiv:1706.06083 (2017).

[3] Bansal, Nitin, Xiaohan Chen, and Zhangyang Wang. 'Can we gain more from orthogonality regularizations in training deep CNNs?.' Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., 2018.

[4] Ross, Andrew Slavin, and Finale Doshi-Velez. 'Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients.' Thirty-second AAAI conference on artificial intelligence. 2018.