Adversarial Machine Learning

Oğuz Kaan Yüksel **Advisor.** Assist. Prof. İnci Meliha Baytaş

June 2020, Boğaziçi CmpE

TL;DR

Randomized Gradient Step *scales sign gradients pixel-wise in order to improve state of the art Projected Gradient Descent adversarial training to have higher diversity in adversarial generation thereby improve robustness and generalization of deep neural networks.*

Keywords: Adversarial Machine Learning, Robustness

Outline

- What is Adversarial Machine Learning?
- Background on Adversarial Attacks and Defences
- Previous Work on Understanding and Regularizing Adversarials
- Improving Adversarial Generation: **Randomized Gradient Step**
- Future directions to explore

What is Adversarial Machine Learning?

Adversarial Samples

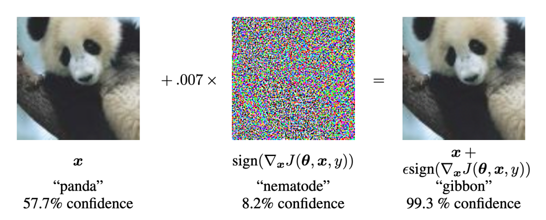


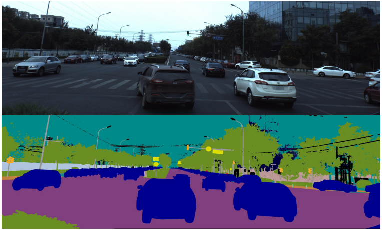
Figure: Adversarial panda of ImageNet on GoogLeNet [1].

"Deliberately" constructed samples that include smart *imperceptible* perturbations to *deceive* ML models.

- Deep learning models are **incredibly** vulnerable to adversarial samples.
- Adversarial Machine Learning is a branch of ML studying *adversarial*
  - Attacks (adversarial sample generation)
  - Detection (of adversarially crafted samples)
  - Defense (developing robust models and training procedures)
  - Theory (understanding vulnerability of Deep Learning to adversarials)

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

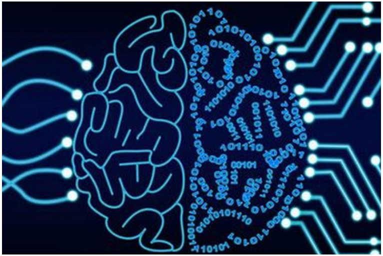
Why Adversarial Machine Learning is Important?



Practical Concerns

- Adversarial samples are pervasive across models and data domains.
- Perturbations used are almost imperceptible by humans as well!
- How can we use ML models robustly in real-life especially considering issues such as malicious users, security, fairness?

Why Adversarial Machine Learning is Important?



Theoretical Concerns

- State of the art computer vision is very impressive **but not** in an adversarial setting. Did (or how much) we really solved vision?
- Black-box deep neural networks work but evidently in a very counter-intuitive fashion! What are we exactly "learning"?
- Are we using the correct metrics, optimization objectives and procedures for ML?

Adversarial Attacks

- **Fast Gradient Sign Method.** Perturbs input in the direction of gradient to maximize model loss.
$$x_{adv} \leftarrow x + \epsilon * sign(\nabla_x L(\theta, x, y))$$
- **Projected Gradient Descent.** Iterative variant that search samples inside a potential adversarial space S (usually ℓ_∞-ball around sample).
$$x^{k+1} \leftarrow \Pi_{x+S}(x^k + \alpha * sign(\nabla_x L(\theta, x^k, y)))$$
- Both can be applied with ℓ_1 and ℓ_2 by taking a fixed step in the gradient direction (taking sign is just *normalization* operation for ℓ_∞).

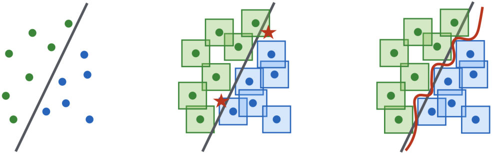


Figure: Illustration of standard and adversarial decision boundary [2].

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

Adversarial Training

- **Saddle Point Formulation:** $\min_{\theta} \max_{\delta \in S} L(\theta, x + \delta, y)$,
  Adversarial attack | defense $\longleftrightarrow$ approximates *inner max.* | *outer min.*
- **Adversarial Training.** Above formulation is non-convex $\rightarrow$ hard to solve! Use simple attacks in training as approximation.

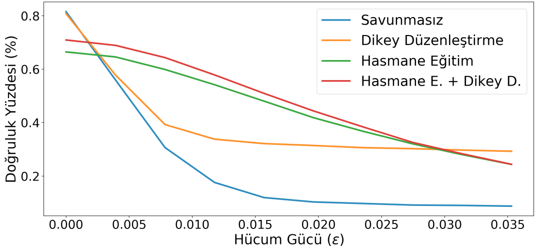


Figure: PGD-10 performance of VGG-16 network on CIFAR10 dataset. Accuracy vs perturbation (ϵ). Blue. *Standard tr.* Orange. *Orthogonal reg.* Green. *Adversarial tr.* Red. *Ortho. reg. & Adv tr.*

Figure: Adversarial images of CIFAR10 with varying ϵ.

Note that $\epsilon = 0.03$ corresponds to just $8/255$ change in pixel values. Adversarial performance of standard model is **below** random (0.1) after just $4/255$ change [3].

- Numerically study activation values in different neurons & layers to understand effect of adversarial perturbations.
  - Extract activation statistics, count “outlier” activations
  - Try to identify “ill-behaving” neurons and portions of network
  - Develop a learnable “clipping” layer to avoid perturbation effects
- Find network regularizations to dampen effect of adv. perturbations.
  - Gradient Reg. $L_R(\theta, x, y) = ||\nabla_x L(\theta, x, y))||_2$
  - Gradient Diff. Reg. $L_R(\theta, x, y) = ||\nabla_x L(\theta, x_{adv}, y)) - \nabla_x L(\theta, x, y))||_2$
  - Orthogonality Reg. $L_R(\theta, x, y)) = \sum_{W \in \theta} ||W^T W - I_n||_F^2$ [3]
- We refer to CMPE 491 Poster for illustrations and results.

[3] Yüksel, Oğuz K. and İnci Meliha Baytaş. "Orthogonality for Adversarial Robustness." accepted to SIU 2020, will be published at IEEE Xplore.

- FGSM attack can exhibit "catastrophic overfitting"
- Using a randomized start within ϵ-ball helps [4].
- Does PGD exhibit similar but more subtle overfitting?

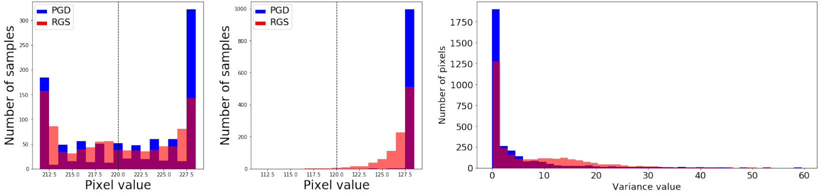
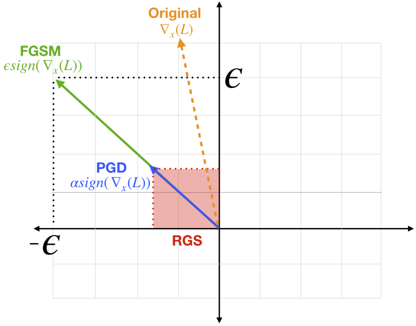


Figure: *Left & Middle*. Histogram of a selected pixel's values in 1000 replication of adversarial generation process on the same sample (middle line is value at original image). *Left*. PGD fixes step size $\alpha = 2.0$ causes generation to miss certain pixel values. *Middle*. Most of the time PGD stagnates at $\epsilon = 8.0$ border. *Right*. Histogram of variances per pixel values in adversarial generation. [5]

[4] Wong, Eric, Leslie Rice, and J. Zico Kolter. "Fast is better than free: Revisiting adversarial training." arXiv preprint arXiv:2001.03994 (2020).

- RGS uses a randomized step *in each pixel* to increase generation variability and avoid subtle correlations and corner clustering.
  $\forall$ pixel, perturbation $\sim Uniform(0, min(\alpha, d_{boundary}))$



Model	PGD Acc (%)	RGS Acc (%)	Natural Acc (%)
PGD	48.34	56.30	87.01
RGS	**49.38**	**63.53**	86.72

Figure: RGS ($\alpha = 8.0$) vs PGD ($\alpha = 2.0$) with k=10, $\epsilon = 8.0$ CIFAR10 training [5].

- RGS performs slightly better than PGD in CIFAR10, MNIST, FMNIST.
- Both overfits to RGS samples more quickly than PGD samples.
- Adversarial generation variability might be a key factor in robustness.

[5] Yüksel, Oğuz K. and İnci Meliha Baytaş. "Randomized Gradient Adversarial Training." under review at ECCV 2020.

Future Directions

- **RGS improvements.** RGS inject high randomness $\rightarrow$ possibly reducing convergence to "high adversarial" regions.
  - Optimize convergence with reducing step size $\rightarrow$ so far we failed.
  - Reduce step of pixels that gradient sign change occurs?
- **PGD Overfitting.** Scope of overfitting in adv. training?
  - Significance of variance in adversarial generation?
  - Does PGD training causes PGD adv. generation to have low variety?
- **Meta Adversarial.** Use variation capable meta-learning algos. to tackle adversarial problem.
  - Use different attacks with various hyperparams as different tasks
  - Try to learn a model that can be tuned to robust in each task
  - Employ universal adversarial perturbation in meta-training?
  - **Current work:** An adaptation of [6] didn't work.

---

[6] Sun, Qianru, et al. "Meta-transfer learning for few-shot learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

Thank you for listening!

kaan.yuksel@boun.edu.tr
inci.baytas@boun.edu.tr

Checkout GitHub repository for project poster and more details:
https://github.com/okyksl/cmpe491-492

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] Yüksel, Oğuz K. and İnci Meliha Baytaş. "Orthogonality for Adversarial Robustness." accepted to SIU 2020, will be published at IEEE Xplore.

[4] Wong, Eric, Leslie Rice, and J. Zico Kolter. "Fast is better than free: Revisiting adversarial training." arXiv preprint arXiv:2001.03994 (2020).

[5] Yüksel, Oğuz K. and İnci Meliha Baytaş. "Randomized Gradient Adversarial Training." under review at ECCV 2020.

[6] Sun, Qianru, et al. "Meta-transfer learning for few-shot learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

Image Source Articles (excluding referenced ones)

- *Slide 5 Left.* `https://www.kaggle.com/c/cvpr-2018-autonomous-driving`
- *Slide 5 Right.* `https://unbabel.com/blog/gender-bias-artificial-intelligence/`
- *Slide 6.* `https://www.eweek.com/innovation/predictions-2019-how-ai-machine-learning-continue-to-impact-us`