Randomized Gradient Adversarial Training

Oğuz Kaan Yüksel, Advisor: Asist. Prof. İnci Meliha Baytaş
CmpE 492 Project - kaan.yuksel@boun.edu.tr

Adversarial Samples

Deliberately constructed samples that attempt to deceive machine learning models.

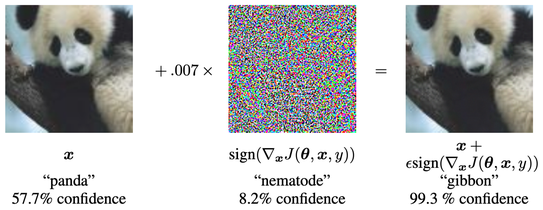

Figure 1: Adversarial panda of ImageNet on GoogLeNet[1]

- 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*:** Multi-step variant considered as *generic first-order adversary*.

$$x^{k+1} \leftarrow \Pi_{x+S}(x^k + \epsilon * sign(\nabla_x L(\theta, x, y)))$$

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

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

- Attacks can be also defined for ℓ_1 and ℓ_2 norms.
- Allowed perturbations set (S) is usually ϵ-ball around the natural sample.
- $\lambda = 0$ with 10-40 step PGD is state of the art.

Problem Statement

Investigate the connection between diversity in adversarial generation and robustness provided by adversarial training, and use adversarial variability to facilitate robustness and generalizability of deep neural networks.

Recent Findings

- FGSM adversarial training may exhibit *catastrophic overfitting* where performance in PGD samples suddenly drops to 0.
- One hypothesized cause: FGSM *can only* generate boundary samples (Figure 2) [3].
- Initial random perturbation within ϵ-ball fixes overfitting and greatly improves performance [3].

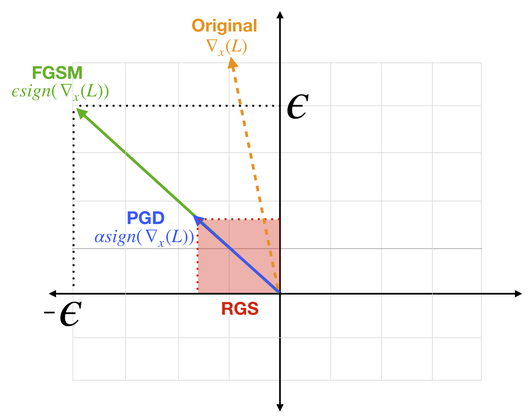

Figure 2: Illustration of single steps in adversarial generation

Question: Does PGD exhibit a similar but more subtle overfitting?

PGD Observations

- Fixed α *for all* steps. *e.g.* pixels can move ± 2 at each step, no way of reaching $\pm(0, 2)$ interval.
- Fixed perturbation *for all* pixels with ℓ_∞. Might create subtle correlations in generation.
- Significant clustering around ϵ boundary (more perturbation $\rightarrow$ worse result). Might caused by fixed step size and projections to boundary.

Randomized Gradient Step

- Use different perturbation scales for each pixel.
- Avoid clustering at boundaries with upper bound.
- Still use gradient sign to determine direction.

$x^{k+1} \leftarrow x^k + U(0, max(\alpha, d_b)) \times sign(\nabla_x L_\theta(x^k, y)))$

where d_b is pixel-wise distance to ϵ boundary.

CIFAR10 Results

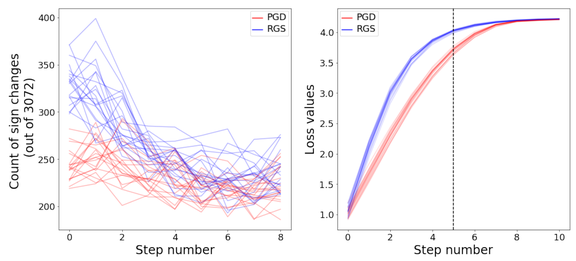

Figure 3: **(a)** Number of sign changes, **(b)** associated loss values in 10 replication of adversarial generation (same sample).

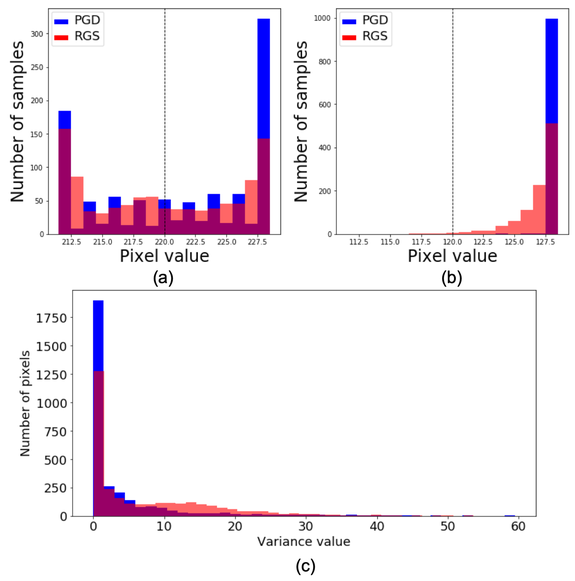

Figure 4: **(a)** & **(b)** Histogram of a selected sample and single pixel's values in 1000 replication of adversarial generation. **(c)** Histogram of variances in all pixel's values for the same sample.

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

RGS vs PGD

- Low amount of sign change indicate attack moves in the same direction i.e. costly gradient calculations can be saved if higher step sizes used. RGS shows higher variation and presents a better trade-off in high step size conditions.
- We observe clusterings seperated with $\alpha = 2$ and significant corner clustering (high freq. in many pixels) in PGD. RGS shows much higher variation and less clustering.
- RGS trained network surpassed PGD trained network **both** on RGS and PGD adversarial samples. We also observe similar and better performance in RGS on MNIST and FMNIST.
- Using random step hinders convergence of RGS adversarial samples greatly. However, PGD has also limited convergence due to fixed α.

Future Work

- **PGD Overfitting.** RGS performance of PGD training networks starts to deteriorate after time. What is the extent of overfitting in adversarial training and its relation with variability?
- **RGS Improvements.** Increase convergence by diminishing step sizes in case of sign change. Simply decreasing at each step didn't help.
- **Meta Adversarial.** Use variation capable meta learning. So far, adapted [4] with multiple attacks as different tasks but failed.

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

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

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

GitHub repository: https://github.com/okyksl/cmpe491-492