Scaleable input gradient regularization for adversarial robustness

Chris Finlay & Adam M Oberman

Department of Mathematics & Statistics, McGill University

christopher.finlay@mail.mcgill.ca, adam.oberman@mcgill.ca

Overview

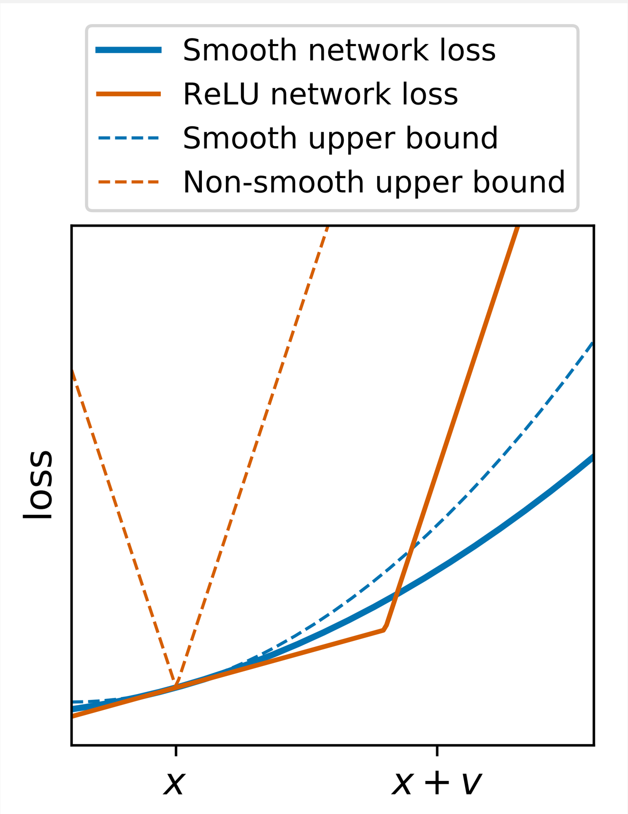
- Neural networks used in computer vision are vulnerable to perturbations of their input specially crafted to cause misclassification, called *adversarial attacks*. These perturbations are invisible to the human eye [1]
- To date the most popular and effective defence against adversarial attacks is to train networks with adversarial images, called *adversarial training* (AT) [2]. However AT has not scaled well to very large networks and datasets, such as on ImageNet-1k.
- Instead we advocate training with input-gradient regularization, in which networks are penalized for having large gradients. We motivate input gradient regularization with theoretical lower bounds on the minimum distance necessary to adversarially perturb an image.
- When implemented with finite differences, input gradient regularization scales readily to larger regimes, avoiding 'double backprop'.

Theoretical motivation: *attack bounds from Taylor expansion*

Adversarial attacks are found by minimizing the perturbation v about an image x, such that the image is misclassified. If a network and loss $\ell(x)$ are L-Lipschitz, then the loss can be bounded above by

$$\ell(x+v) \leq \ell(x) + L\|v\| \qquad (1)$$

For some losses, there is a constant ℓ_0 that determines whether or not the classification is correct.



Suppose then, there is a minimum adversarial perturbation $\ell(x+v) = \ell_0$. Using (1), the minimum adversarial distance is bounded below by

$$\|v\| \geq \frac{\max\{\ell_0 - \ell(x), 0\}}{L} \qquad (L\text{-bound})$$

Unfortunately the Lipschitz constant of a network is a global quantity, and hard to estimate in general. If instead the network is differentiable, we can derive a tighter bound using *local* gradient information, provided we can estimate the maximum curvature C:

$$\|v\|_2 \geq \frac{1}{C}\left(-\|\nabla \ell(x)\|_2 + \sqrt{\|\nabla \ell(x)\|_2^2 + 2C\max\{\ell_0 - \ell(x), 0\}}\right) \qquad (C\text{-bound})$$

Interpretation: *use input gradient regularization*

What do (L-bound) and (C-bound) say heuristically?

- The loss gap, $\ell_0 - \ell(x)$, between the misclassification threshold and the loss at the image, should be large. This is exactly what standard already does.
- The gradient of the loss, *with respect to the input image*, should be small. Since the Lipschitz constant of a network L is the maximum gradient over all inputs, small gradients should help bound (L-bound). Moreover *locally* small gradients help bound (C-bound).
- The network's maximum curvature C should be small. This is hard to penalize directly, however we shall see that finite difference approximations of the gradient regularizer can implicitly penalize for large curvature.

This suggests training a NN $f(x;w)$ with input gradient regularization

$$\min_w \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}}\left[\mathcal{L}(f(x;w),y) + \frac{\lambda}{2}\|\nabla_x \mathcal{L}(f(x;w),y)\|_*^2\right]$$

Input gradient regularization is not new to the NN community: it is commonly used in both training of autoencoders and GANs. It has been attempted in the adversarial robustness community in the past but experimental results were mixed.

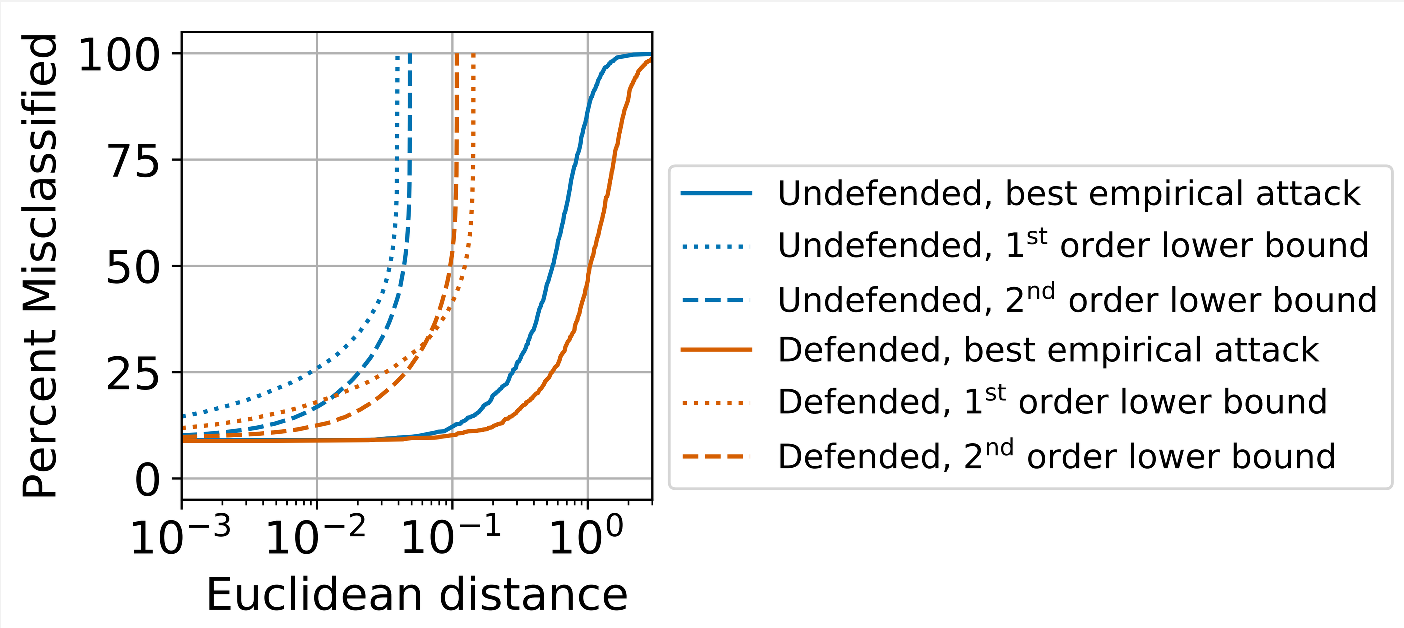


Figure: Theoretical minimum lower bound on adversarial distance for ImageNet-1k, on networks with 'smooth ReLU' activation functions. Defended networks trained with $\lambda = 0.1$, penalized with squared ℓ_2 norm gradient.

Implementation: *finite differences are fast*

We approximate the gradient regularization term with finite differences rather than using double backprop. Let d be the normalized input gradient direction: $d = \nabla_x \ell(x)/\|\nabla_x \ell(x)\|_2$. Then

$$\|\nabla_x \ell(x)\|_2^2 \approx \left(\frac{\ell(x+hd) - \ell(x)}{h}\right)^2$$

The error of this approximation is proportional to the curvature; thus a finite difference approximation *also penalizes curvature* implicitly.

Experimental results: *adversarial robustness that scales*

We train with gradient regularization, and attack models with a host of adversarial attacks [3]

- We obtain similar robustness results compared to the current state-of-the-art on CIFAR-10 [2]
- However unlike other reported methods, ours scales to ImageNet-1k: we can train adversarially robust models in just over a day on four consumer grade GPUs
- Experimental results show that as implemented here, gradient regularization *does not lead to "gradient obfuscation"* [4]

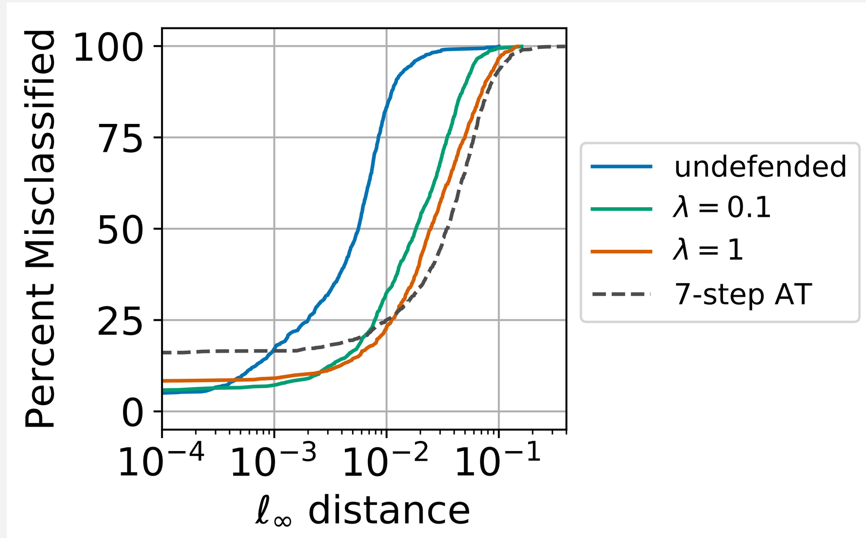


Figure: ℓ_∞ norm adversarial attacks on CIFAR-10

Table: Adversarial robustness statistics, measured in the ℓ_∞ norm. Top1 error is reported on CIFAR-10; Top5 error on ImageNet-1k. *We report statistics using the best adversarial attack on a per-image basis.*

	% clean error	% error at $\varepsilon = \frac{2}{255}$	$\varepsilon = \frac{8}{255}$	mean distance	training time (hours)
CIFAR-10					
Undefended	**4.36**	70.82	98.94	6.62e−3	2.06
Madry et al (7-step AT)	16.33	22.86	**46.02**	**4.07e−2**	12.10
squared ℓ_1 norm, $\lambda = 0.1$	6.45	24.92	70.41	2.35e−2	5.22
squared ℓ_1 norm, $\lambda = 1$	9.02	**18.47**	58.69	3.34e−2	5.15
ImageNet-1k					
Undefended	**6.94**	90.21	98.94	3.94e−3	20.30
squared ℓ_2 norm, $\lambda = 0.1$	7.66	70.56	97.53	7.96e−3	32.60
squared ℓ_2 norm, $\lambda = 1$	10.26	**52.79**	**95.93**	**9.95e−3**	33.87

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[3] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0.8.0: A Python toolbox to benchmark the robustness of machine learning models. *CoRR*, abs/1707.04131, 2017.

[4] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 3–14, 2017.

[5] Chris Finlay and Adam M. Oberman. Scaleable input gradient regularization for adversarial robustness. *CoRR*, abs/1905.11468, 2019.