Supplementary Materials

Xuanqing Liu, Minhao Cheng, Huan Zhang and Cho-Jui Hsieh

University of California, Davis

About Generalization Bound

This result is a direct outcome of [4](Appendix B), and we repeat it for the sake of completeness (Note also that [4](Appendix B) is further influenced by [2](Corollary 5)). The relation between empirical risk and population risk is given by uniform convergence theory:

$$\left|\frac{1}{N}\sum_{(x,y)\in\mathcal{D}}\ell(w;x,y)-\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(w;x,y)]\right|\leq E(N,\rho,\delta), \tag{1}$$

where E is some error bound, $\mathcal{P}$ is the data distribution, $\mathcal{D}$ is the set of samples.

How to Attack an Ensemble of Models

Suppose we have an ensemble of models $\mathcal{M}=\{f_1,f_2,\ldots,f_n\}$, or more generally, it could be a group of infinite models parameterized by some random variables like our RSE model: $\mathcal{M}=\{f_\epsilon|\epsilon\sim N(0,\sigma^2)\}$. For an input pair (x,y) the model group makes decision by voting or aggregation: $\hat{y}=\text{max\_idx}\{\frac{1}{n}\sum_{i=1}^n f_i(x)\}$.

Under this setting, the goal of attack is to find an adversarial image $x'\approx x$ such that this ensemble predicts differently $\hat{y'}\neq y=\hat{y}$. One of the most direct way to achieve this goal to maximize the prediction loss so that $\frac{1}{n}\sum_{i=1}^n\ell(f_i(x),y)$ is large enough. However, for infinite ensemble case (where $\mathcal{M}=\{f_\epsilon|\epsilon\sim N(0,\sigma^2)\}$), our objective function is $\mathbb{E}_\epsilon[\ell(f_\epsilon(x'),y)]$. More formally,

$$\delta^*=\arg\max_{\|\delta\|\leq D}\mathbb{E}_\epsilon[\ell(f_\epsilon(x+\delta),y)], \tag{2}$$

where D is the predefined constraint. Equivalently, we can formulate the above problem by adding an explicit regularizer

$$\delta^*=\arg\max_{\|\delta\|}\mathbb{E}_\epsilon[\ell(f_\epsilon(x+\delta),y)]+\frac{\lambda}{2}\|\delta\|^2. \tag{3}$$

In our implementation, we solve (3) by 300 steps of Adam [3] iteration.

† Note that this attacking method is also mentioned in [1], our experiment result indicates that even if the attacker is aware of the defense method, it still cannot easily find suitable adversarial examples.

Additional experiment

We have additional experiments on the 143-classes subset of ILSVRC-12, specifically, we extract all images within class IDs `np.range(151, 294)` and then resize them to 64×64 pixels. The result is shown in Fig. 1.

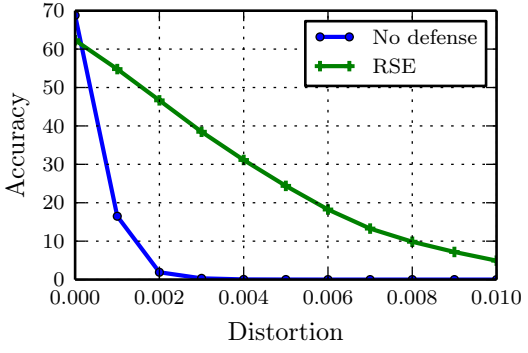

Fig. 1. Accuracy under attack on ILSVRC-12 subset. We only compare our RSE with plain networks. As we can see, our RSE significantly increases the robustness of classifier.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. 35th International Conference on Machine Learning (ICML) (2018)
2. Kakade, S.M., Sridharan, K., Tewari, A.: On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In: Advances in neural information processing systems. pp. 793–800 (2009)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Steinhardt, J., Koh, P.W.W., Liang, P.S.: Certified defenses for data poisoning attacks. In: Advances in Neural Information Processing Systems. pp. 3520–3532 (2017)