Theory

Carl-Johann et al.[1] established the connection between adversarial training and regularization in detail:

The actual updated loss function of adversarial training can be expressed as follows (Aleksander Madry et al.[2] proposed that PGD method does not use normal example, but TRADES et al.[3] pointed out that using normal example can increase the accuracy of the model. Here adopts the loss function of adding normal example.):

$$\tilde{L}(x,y) := \frac{1}{2}(L(x,y) + L(x+\delta,y))$$

where $L(x,y)$ represents the loss function of the training process. δ represents the perturbation superimposed on the input. y is the label of the example. If the perturbation is relatively small (most of the perturbations used in adversarial training are relatively small at present), the first-order Taylor expansion (It is assumed that the loss function is locally linear near the input, the same as that of FGSM adversarial training.) can be used to approximate it:

$$\tilde{L}(x,y) \approx \frac{1}{2}(L(x,y) + L(x,y) + \delta\cdot\partial_x L(x,y)) = L(x,y) + \frac{1}{2}\delta\cdot\partial_x L(x,y)$$

where the second term is the change of loss function caused by perturbation:

$$\delta\cdot\partial_x L(x,y) = \underset{\delta:\|\delta\|_p\le\epsilon}{max}|L(x+\delta,y) - L(x,y)| \approx \underset{\delta:\|\delta\|_p\le\epsilon}{max}|\partial_x L\cdot\delta| = \epsilon\|\partial_x L\|_q$$

where $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$, defined as:

$$\|\mathbf{z}\|_q = sup\{\mathbf{z}^\top\mathbf{x} \mid \|\mathbf{x}\|_p \le 1\}$$

when $p = 1$ is $q = \infty$, there is usually $\frac{1}{p} + \frac{1}{q} = 1$.

Substitute the result in Formula 3 into Formula 2:

$$\tilde{L}(x,y) \approx L(x,y) + \frac{\epsilon}{2}\|\partial_x L\|_q$$

Adversarial training becomes adding a special regularization term to the loss function, which is also proved by the method of enhancing the robustness of the model based on local linear regularization (Local Linear Regularization[4], Curvature Regularization[5]).

The regularization method used in PGD is L2 regularization. However, the solutions obtained by L2 regularization are usually not sparse and do not guarantee to reduce the complexity of the model. To alleviate this problem, our proposed NP-GD takes the L1 regularization method into account, which first performs L1 regularization on the vectors and then applies L2 regularization to the generated vectors. NP-GD has the advantage of first using L1 regularization to reduce the effect of large values on the vectors, and then applying L2 regularization to ensure that the resulting vectors have a consistent length and sum to 1. NP-GD can improve the stability of the regularization process while retaining the advantages of L1 and L2 regularization.

References

[1] Simon-Gabriel, C. J., Ollivier, Y., Bottou, L., Schölkopf, B., & Lopez-Paz, D. (2019, May). First-order adversarial vulnerability of neural networks and input dimension. In International conference on machine learning (pp. 5809-5817). PMLR.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018.

[3] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. ICML 2019.

[4] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In NeurIPS, 2019.

[5] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, Pascal Frossard. Robustness via Curvature Regularization, and Vice Versa. CVPR 2019.