Supplementary Material of Enhancing Intrinsic Adversarial Robustness via Feature Pyramid Decoder

Guanlin Li[1,*] Shuya Ding[2,*] Jun Luo[2] Chang Liu[2]

[1]Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan)

[2]School of Computer Science and Engineering, Nanyang Technological University

leegl@sdas.org {di0002ya,junluo,chang015}@ntu.edu.sg

A. Experimental Results on MNIST

Notation and Implementation Details Firstly, let us define the following notations for accurate description: $\mathcal{F}$ represents the enhanced CNN; $\mathcal{O}$ is the original CNN; $\mathcal{F}_{\mathrm{PGD}}$ and $\mathcal{O}_{\mathrm{PGD}}$ is adversarial trained by L_∞-PGD (on MNIST: ϵ=0.3, step=100 and step length=0.01); $\mathcal{F}_{\mathrm{FGSM}}$ and $\mathcal{O}_{\mathrm{FGSM}}$ is adversarial trained by L_∞-FGSM (on MNIST: ϵ=0.3). All results are achieved with the batch size 100, running on the RTX Titan.

We focus on applying FPD to ResNet-101 on MNIST, attacked by black-box attacks. As stated in the main paper, we mainly concentrate on two performance metrics: classification accuracy and attack time. As stated in [1], most of black-box attacks setting is that the attacker calls the target model with limited queries. Limited queries can be a result of attacking time limit and monetary limit if the attacker incurs a cost for each query. In this perspective, we believe that longer attacking time (more queries) may result in the excess of time and monetary limit. The attacker may surrender the attack. Therefore, we state that attackers spend more time attacking networks, which may protect the networks from another perspective.

In black-box attacks, instead of relying on gradients of the network, we directly attack a corresponding substitute of the network to generate the adversarial samples. For each $\mathcal{F}$-based network (i.e. $\mathcal{F}$, $\mathcal{F}_{\mathrm{FGSM}}$ and $\mathcal{F}_{\mathrm{PGD}}$), we select four networks $\mathcal{O}$, $\mathcal{O}_{\mathrm{FGSM}}$, $\mathcal{O}_{\mathrm{PGD}}$, $\mathcal{F}$ as the substitute. Similarly, $\mathcal{F}$, $\mathcal{F}_{\mathrm{FGSM}}$, $\mathcal{F}_{\mathrm{PGD}}$, $\mathcal{O}$ are used as the substitute of each $\mathcal{O}$-based network (i.e. $\mathcal{O}$, $\mathcal{O}_{\mathrm{FGSM}}$ and $\mathcal{O}_{\mathrm{PGD}}$). Afterwards, we attack each substitute with two different norms L_2 and L_∞. For each norm, we utilize two attack approaches: FGSM and PGD. We set ϵ=1.5 and 0.3 to bound the permutations for L_2 and L_∞ norm. Both L_2-PGD and L_∞-PGD are set to attack for 100 iterations and each step length is 0.1.

Results Overall results of all six networks that attacked by black-box attack with L_2 and L_∞ norm are reported in Table 1 and Table 2, respectively. For L_2 norm, compared with $\mathcal{O}$-based networks on average, our $\mathcal{F}$-based networks can achieve comparable results. As for L_∞ norm, the average accuracy of $\mathcal{F}$-based networks is higher than $\mathcal{O}$-based networks as well. Particularly, $\mathcal{F}_{\mathrm{FGSM}}$ is more robust than $\mathcal{O}_{\mathrm{FGSM}}$ around 10%. Moreover, it is noticeable that $\mathcal{F}$-based networks are substituted by $\mathcal{O}$-based network and vice-versa. The average results reveal that the computational time of attacking $\mathcal{F}$-based substituted network is longer than $\mathcal{O}$-based substituted network around 20 min (L_2) and 26 min (L_∞). Taking into account the time complexity, the result has demonstrated that attackers have to spend more time generating adversarial examples for $\mathcal{F}$-based networks, which may protect the networks from another perspective.

B. Proof

Theorem 1 (the constraint on Lipschitz constant for fully–connected network)**.** *Let* $\mathrm{NN}_{\mathrm{FC}}$ *be a* K*-way-*L*-layer-fully-connected network,* $\mathrm{NN}_{\mathrm{FC}}(x)_k$ *be the* k*-th component of the network output given input* x*,* w_i *be the weight matrix of the* i*-th layer of the network, and* b_i *be a bias matrix of the same layer. Given a noise vector* ξ*, we can bound the variation* $\mathcal{V}$ *component-wisely from above by:*

$$\mathcal{V}_k = |\mathrm{NN}_{\mathrm{FC}}(x)_k - \mathrm{NN}_{\mathrm{FC}}(x+\xi)_k| \le \frac{e^{\theta_k|_x}(e^{\eta} - e^{-\eta})}{\sum_p e^{\theta_p|_{x+\xi}}},$$

where $\theta_k|_x$ *is the* k*-th component of the input to* Softmax *given input* x*. Given* Softmax *function as the activation function of the output layer, we denote the activation function of earlier layers by* f*,* f*'s* Lipschitz constant *by* C*, and let* $\eta = \max_{k=1,\cdots,K}\{[w_L C^{L-1}\,|w_{L-1}w_{L-2}\dots w_1\xi|+b_L]_k\}$.

[*]These authors contributed equally to this work.

[†]The author is also affiliated with Shandong Computer Science Center, Shandong Academy of Sciences, School of Cyber Security, Qilu University of Technology, China

Substitute	Attack	$\mathcal{F}$	$\mathcal{F}_{\text{FGSM}}$	$\mathcal{F}_{\text{PGD}}$	$\mathcal{F}$-Average	$\mathcal{O}$	$\mathcal{O}_{\text{FGSM}}$	$\mathcal{O}_{\text{PGD}}$	$\mathcal{O}$-Average
$\mathcal{O}$	FGSM	(99%,0.5)	(99%,1.47)	(100%,2.1)	(99.33%,1.36)	✗	(100%,0.32)	(99%,0.32)	(99.5%,0.32)
	PGD	(99%,62.23)	(99%,57.82)	(100%,66.9)	(99.33%,62.32)	✗	(100%,42.87)	(98%,42.57)	(99%,42.72)
$\mathcal{F}$	FGSM	✗	(98%,2.82)	(100%,1.9)	(99%,2.36)	(100%,2.05)	(100%,2.8)	(99%,4.8)	(99.67%,3.22)
	PGD	✗	(98%,55.23)	(100%,108.7)	(99%,81.97)	(100%,167.03)	(100%,127.58)	(99%,79.12)	(99.67%,124.57)
$\mathcal{O}_{\text{FGSM}}$	FGSM	(99%,1.8)	(99%,1.32)	(100%,1.18)	(99.33%,1.43)	✗	✗	✗	✗
	PGD	(99%,42.95)	(99%,71.25)	(100%,69.22)	(99.33%,61.14)	✗	✗	✗	✗
$\mathcal{F}_{\text{FGSM}}$	FGSM	✗	✗	✗	✗	(100%,2.38)	(100%,2.39)	(98%,2.52)	(99.33%,2.43)
	PGD	✗	✗	✗	✗	(100%,117.68)	(100%,158.17)	(98%,159.27)	(99.33%,145.04)
$\mathcal{O}_{\text{PGD}}$	FGSM	(99%,1.1)	(99%,1.67)	(100%,1.62)	(99.33%,1.46)	✗	✗	✗	✗
	PGD	(98%,47.75)	(98%,80.9)	(100%,59.1)	(98.67%,62.58)	✗	✗	✗	✗
$\mathcal{F}_{\text{PGD}}$	FGSM	✗	✗	✗	✗	(100%,2.67)	(100%,2.42)	(98%,3.18)	(99.33%,2.76)
	PGD	✗	✗	✗	✗	(100%,158.07)	(100%,116.53)	(98%,86.93)	(99.33%,120.51)
Average		(98.83%,26.06)	(98.63%,34.06)	(100%,38.84)	(99.17%,34.33)	(100%,68.02)	(100%,56.64)	(98.38%,47.34)	(**99.40%,55.20**)

Table 1: L_2 Metrics: robustness evaluation results (Accuracy(%), Attack Time(min)) in thwarting the black-box attacks with ResNet-101 on MNIST. $\mathcal{F}$-Average ($\mathcal{O}$-Average) measure the average performance of $\mathcal{F}$, $\mathcal{F}_{\text{FGSM}}$, $\mathcal{F}_{\text{PGD}}$ ($\mathcal{O}$, $\mathcal{O}_{\text{FGSM}}$, $\mathcal{O}_{\text{PGD}}$).

Substitute	Attack	$\mathcal{F}$	$\mathcal{F}_{\text{FGSM}}$	$\mathcal{F}_{\text{PGD}}$	$\mathcal{F}$-Average	$\mathcal{O}$	$\mathcal{O}_{\text{FGSM}}$	$\mathcal{O}_{\text{PGD}}$	$\mathcal{O}$-Average
$\mathcal{O}$	FGSM	(50%,0.4)	(87%,1.77)	(100%,2.08)	(79%,1.42)	✗	(51%,0.58)	(98%,0.58)	(74.5%,0.58)
	PGD	(65%,44.23)	(87%,50.42)	(99%,75.95)	(83.67%,56.87)	✗	(67%,18.13)	(99%,13.13)	(83%,15.63)
$\mathcal{F}$	FGSM	✗	(80%,2.95)	(94%,1.95)	(87%,2.45)	(47%,1.3)	(58%,2.82)	(95%,4.82)	(66.67%,2.98)
	PGD	✗	(84.7%,128.25)	(98.96%,113.78)	(91.83%,121.02)	(69.07%,167.9)	(77.05%,90.92)	(99%,80.5)	(81.71%,113.11)
$\mathcal{O}_{\text{FGSM}}$	FGSM	(57%,1.82)	(90%,1.23)	(100%,1.23)	(82.33%,1.43)	✗	✗	✗	✗
	PGD	(72%,44.2)	(87%,87.38)	(100%,76.73)	(86.33%,69.44)	✗	✗	✗	✗
$\mathcal{F}_{\text{FGSM}}$	FGSM	✗	✗	✗	✗	(42%,2.37)	(58%,2.47)	(96%,2.63)	(65.33%,2.49)
	PGD	✗	✗	✗	✗	(59.66%,122.22)	(61.78%,163.13)	(98%,166.03)	(73.15%,150.46)
$\mathcal{O}_{\text{PGD}}$	FGSM	(39%,1.1)	(54%,1.7)	(90%,0.87)	(61%,1.22)	✗	✗	✗	✗
	PGD	(33%,48.49)	(58%,82.98)	(87%,66.72)	(59.33%,66.06)	✗	✗	✗	✗
$\mathcal{F}_{\text{PGD}}$	FGSM	✗	✗	✗	✗	(53%,4.22)	(79%,4.27)	(94%,4.23)	(75.33%,4.24)
	PGD	✗	✗	✗	✗	(55.34%,168.58)	(93.84%,167.87)	(90.99%,167.7)	(80.06%,168.05)
Average		(52.67%,23.37)	(78.46%,44.59)	(96.12%,42.41)	(**75.75%,36.79**)	(54.35%,77.77)	(68.21%,56.27)	(96.25%,54.95)	(72.94%,62.99)

Table 2: L_∞ Metrics: robustness evaluation results (Accuracy(%), Attack Time(min)) in thwarting the black-box attacks with ResNet-101 on MNIST. $\mathcal{F}$-Average ($\mathcal{O}$-Average) measure the average performance of $\mathcal{F}$, $\mathcal{F}_{\text{FGSM}}$, $\mathcal{F}_{\text{PGD}}$ ($\mathcal{O}$, $\mathcal{O}_{\text{FGSM}}$, $\mathcal{O}_{\text{PGD}}$).

Proof. *For K classify mission, input has n dim, fixed active function is f for each layer, and the network has $L+1$ layers. So we will have a table 3 below to show the structure of the network (for convenience, we assume that the network is a fully connected network):*

Assuming x normalized, we have $x \in [0,1]^{n*1}$. And we sample a noise vector ξ^{n*1} adding to x. We assume that $x+\xi$ will not be too big or small for 1 or 0.

For the convenience of proof, let $L = 2$, and we have:

$$Layer_2\left(x\right) = f(w_2 f\left(w_1 x + b_1\right) + b_2)$$

$$Layer_2\left(x+\xi\right) = f(w_2 f\left(w_1 (x+\xi) + b_1\right) + b_2)$$

$$Layer_3(x)_k = \frac{\exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_k)}{\sum_i \exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_i)}$$

$$Layer_3(x+\xi)_k = \frac{\exp(f(w_2 f\left(w_1 (x+\xi) + b_1\right) + b_2)_k)}{\sum_i \exp(f(w_2 f\left(w_1 (x+\xi) + b_1\right) + b_2)_i)}.$$

Define D_k as the k-th component of the variation:

$$D_k = Layer_3(x)_k - Layer_3(x+\xi)_k$$
$$= \frac{\exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_k) * \sum \exp(f\left(w_2 f\left(w_1\left(x+\xi\right) + b_1\right) + b_2\right)_i)}{\sum \exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_i) * \sum \exp(f\left(w_2 f\left(w_1\left(x+\xi\right) + b_1\right) + b_2\right)_i)}$$
$$- \frac{\exp(f\left(w_2 f\left(w_1\left(x+\xi\right) + b_1\right) + b_2\right)_k) * \sum \exp(f\left(w_2 f\left(w_1 x + b_1\right) + b_2\right)_i)}{\sum \exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_i) * \sum \exp(f\left(w_2 f\left(w_1\left(x+\xi\right) + b_1\right) + b_2\right)_i)},$$
$$\mathcal{V}_k = \left|D_k\right|.$$

Firstly, we need to analyse the relationship between $Layer_2\left(x\right)$ and $Layer_2\left(x+\xi\right)$. We assume that active function has property satisfies:

$$\left|f\left(x\right) - f(y)\right| \leq C\left|x-y\right|,$$

as known as C-Lipschitz constant.
Then we have:

$$\left|f\left(w_1 x + b_1\right) - f\left(w_1(x+\xi) + b_1\right)\right| \leq C\left|w_1 \xi\right|.$$

Defining the symbol $\leq$ means every corresponding element from LH and RH satisfies the inequality. Then we can

layer	input	Output
0	x^{n*1}	/
1	x^{n*1}	$f\left(w_1^{n_1*n}x{+}b_1^{n_1*1}\right) = Layer_1^{n_1*1}$
2	$f\left(w_1^{n_1*n}x{+}b_1^{n_1*1}\right) = Layer_1^{n_1*1}$	$f\left(w_2^{n_2*n_1}Layer_1^{n_1*1}{+}b_2^{n_2*1}\right) = Layer_2^{n_2*1}$
$\cdots$	$\cdots$	$\cdots$
i	$f\left(w_{i-1}^{n_{i-1}*n_{i-2}}Layer_{i-2}^{n_{i-2}*1}{+}b_{i-1}^{n_{i-1}*1}\right) = Layer_{i-1}^{n_{i-1}*1}$	$f\left(w_i^{n_i*n_{i-1}}Layer_{i-1}^{n_{i-1}*1}{+}b_i^{n_i*1}\right) = Layer_i^{n_i*1}$
$\cdots$	$\cdots$	$\cdots$
L	$f\left(w_{L-1}^{n_{L-1}*n_{L-2}}Layer_{L-2}^{n_{L-2}*1}{+}b_{L-1}^{n_{L-1}*1}\right) = Layer_{L-1}^{n_{L-1}*1}$	$f\left(w_L^{k*n_{L-1}}Layer_{L-1}^{n_{L-1}*1}{+}b_L^{k*1}\right) = Layer_L^{k*1}$
L+1	$f\left(w_L^{K*n_{L-1}}Layer_{L-1}^{n_{L-1}*1}{+}b_L^{K*1}\right) = Layer_L^{K*1}$	$\text{Softmax}\left(Layer_L^{K*1}\right) = Layer_{L+1}^{K*1}$

Table 3: The structure of the network

use this to more layers:

$$\begin{aligned}
&\left|Layer_2\left(x+\xi\right) - Layer_2\left(x\right)\right| \\
&= \left|f\left(w_2 f\left(w_1\left(x+\xi\right)+b_1\right)+b_2\right) - f(w_2 f\left(w_1 x + b_1\right)+b_2)\right| \\
&\leq C\left|w_2\left(f\left(w_1\left(x+\xi\right)+b_1\right) - f\left(w_1 x + b_1\right)\right)\right| \leq C^2\left|w_2 w_1 \xi\right|.
\end{aligned}$$

Defining $\zeta = \max_{i=1,2,\ldots,n_2}\{[C^2\left|w_2 w_1 \xi\right|]_i\}$ means the biggest absolute value of element in vector $C^2 w_2 w_1 \xi$. Now we can estimate the upper and lower bound of D_k:

$$\begin{aligned}
D_k = &\frac{\exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_k) * \sum \exp(Layer_2\left(x+\xi\right)_i)}{\sum \exp(Layer_2(x)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(f\left(w_2 f\left(w_1\left(x+\xi\right)+b_1\right)+b_2\right)_k) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2(x)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
\leq &\frac{\exp(Layer_2\left(x\right)_k) * \sum \exp(Layer_2\left(x\right)_i + C^2\left|w_2 w_1 \xi\right|)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(Layer_2\left(x\right)_k - C^2\left|w_2 w_1 \xi\right|) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
\leq &\frac{\exp(Layer_2\left(x\right)_k) * \sum \exp(Layer_2\left(x\right)_i + \zeta)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(Layer_2\left(x\right)_k - \zeta) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
= &\frac{\exp(Layer_2\left(x\right)_k) * (\exp\left(\zeta\right) - \exp\left(-\zeta\right))}{\sum \exp(Layer_2\left(x+\xi\right)_i)}
\end{aligned}$$

and

$$\begin{aligned}
D_k = &\frac{\exp(f(w_2 f\left(w_1 x + b_1\right) + b_2)_k) * \sum \exp(Layer_2\left(x+\xi\right)_i)}{\sum \exp(Layer_2(x)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(f\left(w_2 f\left(w_1\left(x+\xi\right)+b_1\right)+b_2\right)_k) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2(x)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
\geq &\frac{\exp(Layer_2\left(x\right)_k) * \sum \exp(Layer_2\left(x\right)_i - C^2\left|w_2 w_1 \xi\right|)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(Layer_2\left(x\right)_k + C^2\left|w_2 w_1 \xi\right|) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
\geq &\frac{\exp(Layer_2\left(x\right)_k) * \sum \exp(Layer_2\left(x\right)_i - \zeta)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
&- \frac{\exp(Layer_2\left(x\right)_k + \zeta) * \sum \exp(Layer_2\left(x\right)_i)}{\sum \exp(Layer_2\left(x\right)_i) * \sum \exp(Layer_2\left(x+\xi\right)_i)} \\
= &\frac{\exp(Layer_2\left(x\right)_k) * (\exp\left(-\zeta\right) - \exp\left(\zeta\right))}{\sum \exp(Layer_2\left(x+\xi\right)_i)}.
\end{aligned}$$

The limitation of $\mathcal{V}_k$:

$$\lim_{\zeta \to 0} \mathcal{V}_k \; = 0.$$

To estimate the bound of $|w_2 w_1 \xi|$, firstly we fill the weights matrix (except for the w_1 and w_L) by zero to make them have same shape $m * m$:

$$w_1 = \left[\begin{array}{ccc} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,m} \end{array}\right],$$

$$\xi = \left[\begin{array}{c} \xi_1 \\ \vdots \\ \xi_n \end{array}\right],$$

$$w_2 = \left[\begin{array}{ccc} w'_{1,1} & \cdots & w'_{1,m} \\ \vdots & \ddots & \vdots \\ w'_{m,1} & \cdots & w'_{m,m} \end{array}\right],$$

$$w_1\xi = \begin{bmatrix} \sum_{j=1}^{n} w_{1,j} * \xi_j \\ \vdots \\ \sum_{j=1}^{n} w_{m,j} * \xi_j \end{bmatrix}$$

and

$$w_2 w_1 \xi = \begin{bmatrix} \sum_{i=1}^{m} (w'_{1,i} * \sum_{j=1}^{n} w_{i,j} * \xi_j) \\ \vdots \\ \sum_{i=1}^{m} (w'_{m,i} * \sum_{j=1}^{n} w_{i,j} * \xi_j) \end{bmatrix}.$$

We can roughly estimate the bound by using the biggest element of weight matrix:

$$\begin{aligned} w_1\xi &\le \begin{bmatrix} \max_{i=1,2,\ldots,n}\{[\xi]_i\} \; * \sum_{j=1}^{n} w_{1,j} \\ \vdots \\ \max_{i=1,2,\ldots,n}\{[\xi]_i\} \; * \sum_{j=1}^{n} w_{m,j} \end{bmatrix} \\ &\le \begin{bmatrix} \max_{i=1,2,\ldots,n}\{[\xi]_i\} \; * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{n} w_{i,j}]_i\} \\ \vdots \\ \max_{i=1,2,\ldots,n}\{[\xi]_i\} \; * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{n} w_{i,j}]_i\} \end{bmatrix} \end{aligned}$$

and

$$w_2 w_1 \xi \le \begin{bmatrix} \max_{i=1,2,\ldots,n}\{[\xi]_i\} \\ * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{n} w_{i,j}]_i\} * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{m} w'_{i,j}]_i\} \\ \vdots \\ \max_{i=1,2,\ldots,n}\{[\xi]_i\} \\ * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{n} w_{i,j}]_i\} * \max_{i=1,2,\ldots,m}\{[\sum_{j=1}^{m} w'_{i,j}]_i\} \end{bmatrix}.$$

Define $\rho_i = \max_{q=1,2,\ldots,m}\{[\sum_j w_{i,k,j}]_q\}$ *means the biggest one in layer* i *weight matrix summing up by row, similar as the* 1*-matrix norm but without using element absolute value. So for normal situation, we have:*

$$w_L w_{L-1} \ldots w_1 \xi \le \begin{bmatrix} \sum_{i=1}^{L} \rho_i * \max_{j=1,2,\ldots,n}\{[\xi]_j\} \\ \vdots \\ \sum_{i=1}^{L} \rho_i * \max_{j=1,2,\ldots,n}\{[\xi]_j\} \end{bmatrix}.$$

So

$$\begin{aligned} &|Layer_L\,(x+\xi) - Layer_L\,(x)| \\ &\le (C * \max_{j=1,2,\ldots,L}\{[\rho_i]_j\}\,)^L * \max_{j=1,2,\ldots,n}\{[\xi]_j\}. \end{aligned}$$

If we have $C < \dfrac{1}{\max_{j=1,2,\ldots,L}\{[\rho_i]_j\})}$, *then* $\lim_{C\to 0}\mathcal{V}_k\; = 0$ *or equally* $\lim_{L\to\infty}\mathcal{V}_k\; = 0$.

We assume that for every matrix element satisfies $w_{i,j,k} \sim P(mean = \mu_i, stddev = \sigma_i)$*, we calculate the expectation and variance of* $\sum_j w_{i,k,j}$*,*

$$E\left[\sum_j w_{i,k,j}\right]_{w_{i,k,j}\sim P(\mu_i,\sigma_i)} = m\mu_i,$$

and

$$Var\left[\sum_j w_{i,k,j}\right]_{w_{i,k,j}\sim P(\mu_i,\sigma_i)} = m\sigma_i^2.$$

Now we have:

$$\frac{\sum_j w_{i,k,j} - m\mu_i}{m\sigma_i^2} \sim N(0,1).$$

that means ρ_i *will not be too big than* $m\mu_i$ *if* σ_i^2 *is small.*

Comparing with a network that does not have an activation function before Softmax at layer L, such as

$$\text{Softmax}(w_L Layer_{L-1}(x) + b_L)$$

and

$$\theta_k|_x = [w_L f(w_{L-1}(\ldots(f(w_1 x + b_1))) + b_{L-1}) + b_L]_k,$$

$$\theta_k|_{x+\xi} = [w_L f(w_{L-1}(\ldots(f(w_1(x{+}\xi){+}b_1))){+}b_{L-1}){+}b_L]_k,$$

we have the bound:

$$\mathcal{V}_k \le \frac{e^{\theta_k|_x}(e^{\eta} - e^{-\eta})}{\sum_p e^{\theta_p|_{x+\xi}}},$$

and

$$\eta = \max_{k=1,2,\ldots,K}\{[w_L C^{L-1}\,|w_{L-1}w_{L-2}\ldots w_1\xi|{+}b_L]_k\}.$$

Define $\tau = \sqrt[(L-1)]{\max_{k=1,2,\ldots,K}\{[|w_{L-1}w_{L-2}\ldots w_1\xi|]_k\}}$. So we have three situations:

1. $$\lim_{0\le C\tau<1,\; 0<K<M,\; L\to\infty} \eta = \max_{k=1,2,\ldots,K}\{[b_L]_k\}$$

2. $$\lim_{C\tau=1,\; 0<K<M,\; L\to\infty} \eta = \max_{k=1,2,\ldots,K}\{[\rho_L + b_L]_k\}$$

3. $$\lim_{C\tau>1,\; 0<K<M,\; L\to\infty} \eta \to \infty$$

So when $C\tau \le 1$, w_L and b_L will have a bigger influence than $C^{L-1}\,|w_{L-1}w_{L-2}\ldots w_1\xi|$.

References

[1] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *Proc. of the ICML*, 2018. 1