Lipschitz Continuity Retained Binary Neural Network

Yuzhang Shang[1], Dan Xu[2], Bin Duan[1],
Ziliang Zong[3], Liqiang Nie[4], and Yan Yan[1]⋆

[1] Illinois Institute of Technology, USA
[2] Hong Kong University of Science and Technology, Hong Kong
[3] Texas State University, USA
[4] Harbin Institute of Technology, Shenzhen, China
{yshang4, bduan2}@hawk.iit.edu, danxu@cse.ust.hk, ziliang@txstate.edu,
nieliqiang@gmail.com, and yyan34@iit.edu

1 Supplemental Material

1.1 Proofs.

Lemma 1. If a function $f : \mathbb{R}^n \longmapsto \mathbb{R}^m$ is a locally Lipschitz continuous function, then f is differentiable almost everywhere. Moreover, if f is Lipschitz continuous, then

$$\|f\|_{Lip} = \sup_{\mathbf{x}\in\mathbb{R}^n} \|\nabla_{\mathbf{x}} f\|_2 \tag{1}$$

where $\|\cdot\|_2$ is the L2 matrix norm.
Proof. Based on Rademacher's theorem, for the functions restricted to some neighborhood around any point is Lipschitz, their Lipschitz constant can be calculated by their differential operator.
Lemma 2. Let $\mathbf{W} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ and $T(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}$ be an linear function. Then for all $\mathbf{x} \in \mathbb{R}^n$, we have

$$\nabla g(\mathbf{x}) = \mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{x} \tag{2}$$

where $g(\mathbf{x}) = \frac{1}{2}\|f(\mathbf{x}) - f(\mathbf{0})\|_2^2$.
Proof. By definition, $g(\mathbf{x}) = \frac{1}{2}\|f(\mathbf{x}) - f(\mathbf{0})\|_2^2 = \frac{1}{2}\|(\mathbf{W}\mathbf{x} + \mathbf{b}) - (\mathbf{W}\mathbf{0} + \mathbf{b})\|_2^2 = \frac{1}{2}\|\mathbf{W}\mathbf{x}\|_2^2$, and the derivative of this equation is the desired result.
Theorem 1. If a matrix $\mathbf{U}$ is an orthogonal matrix, such that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$, where $\mathbf{I}$ is a unit matrix, the largest eigenvalues of $\mathbf{U}^\mathsf{T}\mathbf{H}\mathbf{U}$ and $\mathbf{H}$ are equivalent:

$$\sigma_1(\mathbf{U}^\mathsf{T}\mathbf{H}\mathbf{U}) = \sigma_1(\mathbf{H}), \tag{3}$$

where the notation $\sigma_1(\cdot)$ indicates the largest eigenvalue of a matrix.
Proof. Because for $\mathbf{U}^{-1}$, we have

$$(\mathbf{U}^{-1})^\mathsf{T}(\mathbf{U}^\mathsf{T}\mathbf{H}\mathbf{U})(\mathbf{U}^{-1}) = (\mathbf{U}\mathbf{U}^{-1})^\mathsf{T}\mathbf{H}(\mathbf{U}\mathbf{U}^{-1}) = \mathbf{H}. \tag{4}$$

⋆ Corresponding author.

Thus matrix $(\mathbf{U}^\top \mathbf{H}\mathbf{U})$ and matrix $(\mathbf{H})$ are similar. The Theorem 1 can be proven by this matrix similarity.

Exact Lipschitz constant computation is NP-Hard. We take a 2-layer fully-connected neural network with ReLU activation function as an example to demonstrate that Lipschitz computation is not achievable in polynomial time. As we denoted in Method Section, this 2-layer fully-connected neural network can be represented as

$$f(\mathbf{W}^1, \mathbf{W}^2; \mathbf{x}) = (\mathbf{W}^2 \circ \sigma \circ \mathbf{W}^1)(\mathbf{x}), \tag{5}$$

where $\mathbf{W}^1 \in \mathbb{R}^{d_0 \times d_1}$ and $\mathbf{W}^2 \in \mathbb{R}^{d_1 \times d_2}$ are matrices of first and second layers of neural network, and $\sigma(x) = \max\{0, x\}$ is the ReLU activation function.

Proof. To prove that computing the exact Lipschitz constant of Networks is NP-hard, we only need to prove that deciding if the Lipschitz constant $\|f\|_{Lip} \leq L$ is NP-hard.

From a clearly NP-hard problem:

$$\max\min \Sigma_i (\mathbf{h}_i^\top \mathbf{p})^2 = \mathbf{p}^\top \mathbf{H}\mathbf{p} \tag{6}$$
$$s.t. \quad \forall k, 0 \leq p_k \leq 1, \tag{7}$$

where matrix $\mathbf{H} = \Sigma_i \mathbf{h}_i \mathbf{h}_i^\top$ is positive semi-definite with full rank. We denote matrices W_1 and W_2 as

$$\mathbf{W}_1 = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{d_1}), \tag{8}$$

$$\mathbf{W}_2 = (\mathbf{1}_{d_1 \times 1}, \mathbf{0}_{d_1 \times d_2 - 1})^\top, \tag{9}$$

so that we have

$$\mathbf{W}_2 \mathrm{diag}\,(\mathbf{p})\, \mathbf{W}_1 = \begin{bmatrix} \mathbf{h}_1^\top \mathbf{p} \; 0 \ldots 0 \\ \vdots \quad \vdots \; \ddots \\ \mathbf{h}_n^\top \mathbf{p} \; 0 \quad\; 0 \end{bmatrix}^\top \tag{10}$$

The spectral norm of this 1-rank matrix is $\Sigma_i (\mathbf{h}_i^\top \mathbf{p})^2$. We prove that Eq. 6 is equivalent to the following optimization problem

$$\max\min \|\mathbf{W}_2 \mathrm{diag}\,(\mathbf{p})\, \mathbf{W}_1\|_2^2 \tag{11}$$
$$s.t. \quad \mathbf{p} \in [0, 1]^n\,. \tag{12}$$

Because H is full rank, W_1 is subjective and all $\mathbf{p}$ are admissible values for $\nabla g(\mathbf{x})$ which is the equality case. Finally, ReLU activation units take their derivative within $\{0, 1\}$ and Eq. 11 is its relaxed optimization problem, that has the same optimum points. So that our desired problem is NP-hard.

Algorithm 1 Compute Spectral Norm using Power Iteration

```
Require: Targeted matrix $\mathbf{RM}$ and stop condition $res_{stop}$.
Ensure: The spectral norm of matrix $\mathbf{RM}$, i.e., $\|\mathbf{RM}\|_{SN}$.
Initialize $\mathbf{v}_0 \in \mathbb{R}^m$ with a random vector.
while $res \geq res_{stop}$ do
    $\mathbf{v}_{i+1} \leftarrow \mathbf{RM}\mathbf{v}_i / \|\mathbf{RM}\mathbf{v}_i\|_2$
    $res = \|\mathbf{v}_{i+1} - \mathbf{v}_i\|_2$
end while
return $\|\mathbf{RM}\|_{SN} = \mathbf{v}_{i+1}^{\mathsf{T}}\mathbf{RM}\mathbf{v}_i$
```

1.2 Power Iteration Algorithm

1.3 Detailed derivation of the gradient.

The derivative of the loss function $\mathcal{L}$ w.r.t $\mathbf{W}_B^k$ is:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}_B} &= \frac{\partial(\mathcal{L}_{CE})}{\partial \mathbf{W}_B} + \frac{\partial(\mathcal{L}_{Lip})}{\partial \mathbf{W}_B^k} \\
&= \mathbf{M} - \lambda \sum_{k=1}^{L-1} \beta^{k-L} \left(\frac{\|\mathbf{RM}_F^k\|_{SN}}{\|\mathbf{RM}_B^k\|_{SN}}\right) \frac{\partial \|\mathbf{RM}_B^k\|_{SN}}{\partial \mathbf{W}_B^k} \\
&\approx \mathbf{M} - \lambda \sum_{k=1}^{L-1} \beta^{k-L} \left(\frac{\|\mathbf{RM}_F^k\|_{SN}}{\|\mathbf{RM}_B^k\|_{SN}}\right) \frac{\partial \|\mathbf{W}_B^k\|_{SN}}{\partial \mathbf{W}_B^k} \\
&\approx \mathbf{M} - \lambda \sum_{k=1}^{L-1} \beta^{k-L} \left(\frac{\|\mathbf{RM}_F^k\|_{SN}}{\|\mathbf{RM}_B^k\|_{SN}}\right) \mathbf{u}_1^k (\mathbf{v}_1^k)^{\mathsf{T}},
\end{aligned} \tag{13}$$

For the third equation:

$$\mathbf{M} - \lambda \sum_{k=1}^{L-1} \beta^{k-L} \left(\frac{\|\mathbf{RM}_F^k\|_{SN}}{\|\mathbf{RM}_B^k\|_{SN}}\right) \frac{\partial \|\mathbf{W}_B^k\|_{SN}}{\partial \mathbf{W}_B^k} \approx \mathbf{M} - \lambda \sum_{k=1}^{L-1} \beta^{k-L} \left(\frac{\|\mathbf{RM}_F^k\|_{SN}}{\|\mathbf{RM}_B^k\|_{SN}}\right) \mathbf{u}_1^k (\mathbf{v}_1^k)^{\mathsf{T}}, \tag{14}$$

we provide the core proof in here, *i.e.* the first pair of left and right singular vectors of $\mathbf{W}_B$ can reconstruct $\frac{\partial \|\mathbf{W}_B\|_{SN}}{\partial \mathbf{W}_B}$ precisely. For $\mathbf{W}_B \in \mathbb{R}^{m \times n}$, the spectral norm $\|\mathbf{W}_B\|_{SN} = \sigma_1(\mathbf{W}_B)$ stands for its biggest singular value, $\mathbf{u}_1$ and $\mathbf{v}_1$ are correspondingly left and singular vectors. The SVD of $\mathbf{W}_B$ is $\mathbf{W}_B = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Therefore $\|\mathbf{W}_B\|_{SN} = \mathbf{e}_1^T \mathbf{U}^T (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)\mathbf{V}\mathbf{e}_1$, where $\mathbf{e}_1$ is the largest eigenvalue of matrix $\mathbf{W}_B^T \mathbf{W}_B$. Hence $\|\mathbf{W}_B\|_{SN} = \mathbf{u}_1^T \mathbf{W}_B \mathbf{v}_1$. Thus the derivative of spectral norm can be evaluated in the direction $\mathbf{H}$: $\frac{\partial \|W\|_{SN}}{\partial \mathbf{W}_B}(\mathbf{H}) = \mathbf{u}_1^T \mathbf{H} \mathbf{v}_1 = \text{trace}(\mathbf{u}_1^T \mathbf{H} \mathbf{v}_1) = \text{trace}(\mathbf{v}_1 \mathbf{u}_1^T \mathbf{H})$. The gradient is $\frac{\partial \|\mathbf{W}_B\|_{SN}}{\partial \mathbf{W}_B} = \mathbf{v}_1 \mathbf{u}_1^T$, which supports the Eq.13.

1.4 ImageNet-C

Sample Visualization of ImageNet-C. In Section 4.4 we evaluate methods on a common image corruptions benchmark (ImageNet-C) to demonstrate the effectiveness of *LCR* from the perspective of model robustness. As illustrated in Section 4.4, ImageNet-C [1] consists of 19 different types of corruptions with five levels of severity from the noise, blur, weather and digital categories applied to the validation images of ImageNet (see Fig. 1). As the figure presented, it is natural to introduce the ImageNet-C to measure the semantic robustness of models. Recently, ImageNet-C indeed has became the most widely acknowledged dataset for measuring the robustness of models.

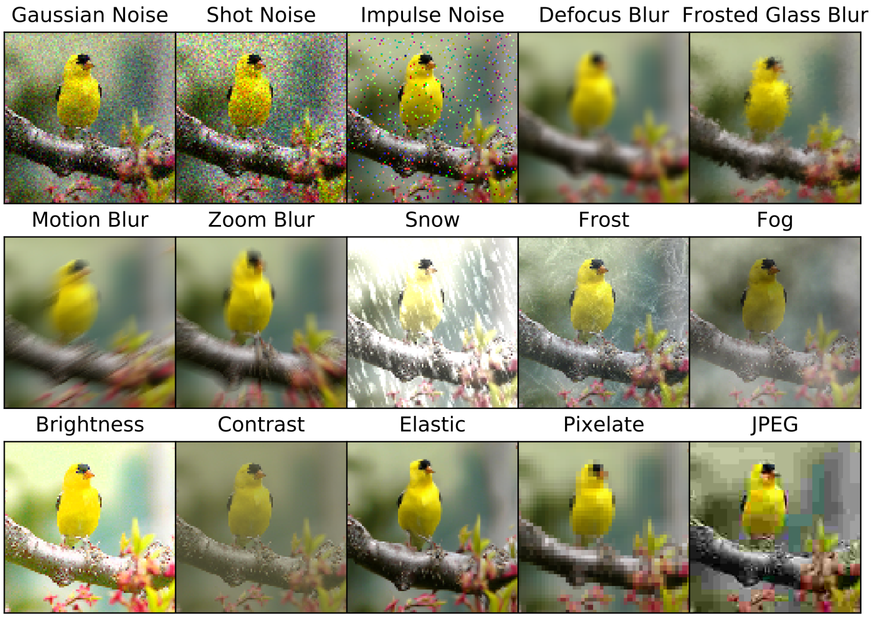

Fig. 1. Examples of each corruption type in the image corruptions benchmark. While synthetic, this set of corruptions aims to represent natural factors of variation like noise, blur, weather, and digital imaging effects. This figure is reproduced from Hendrycks & Dietterich (2019).

References

1. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019) 4