Mitigating Transformer Overconfidence via Lipschitz Regularization (Supplementary Material)

Wenqian Ye[1,4] **Yunsheng Ma**[2,4] **Xu Cao**[3,4] **Kun Tang**[5]

[1]Department of Computer Science, University of Virginia, Charlottesville, VA, USA
[2]College of Engineering, Purdue University, West Lafayette, IN, USA
[3]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA
[4]AI Lab, Shenzhen Children's Hospital, Shenzhen, China
[5]T Lab, Tencent, Beijing, China

A PROOF FOR THE LIPSCHITZ CONSTANT OF LAYERNORM

The LayerNorm operation [Ba et al., 2016] used in LRFormer can be expressed as:

$$\mathrm{LN}(\mathbf{x}) = \frac{\mathbf{x} - \mu(\mathbf{x})}{\sqrt{\sigma^2(\mathbf{x}) + \epsilon}} * \boldsymbol{\gamma} + \boldsymbol{\beta}$$

where $\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^N$, $\mu(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\sigma^2(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(\mathbf{x}))^2$.

WLOG, assume $N > 2$ and not all x_i are equal.

The derivatives of μ and σ^2 w.r.t x:

$$\frac{\partial \mu}{\partial \mathbf{x}} = \frac{1}{N}\mathbb{1}^\top$$

$$\frac{\partial \sigma^2}{\partial \mathbf{x}} = \frac{2}{N}(\mathbf{x} - \mu)^\top$$

Take the derivative of $\mathrm{LN}(\mathbf{x})_i$, the ith element of $\mathrm{LN}(\mathbf{x})$, with respect to $\mathbf{x}$ is:

$$\frac{\partial \mathrm{LN}(\mathbf{x})_i}{\partial \mathbf{x}} = \gamma_i(\sigma^2 + \epsilon)^{-\frac{1}{2}}\left[(\mathbf{e}_i - \frac{1}{N}\mathbb{1})^\top - \frac{1}{N}(\sigma^2 + \epsilon)^{-1}(x_i - \mu)(\mathbf{x} - \mu)^\top\right]. \tag{1}$$

where $\mathbf{e}_I \in \mathbb{R}^N$ is a one-hot vector with 1 at the ith element. Therefore,

$$\frac{\partial \mathrm{LN}(\mathbf{x})}{\partial \mathbf{x}} = (\sigma^2 + \epsilon)^{-\frac{1}{2}}\left[\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{N}\boldsymbol{\gamma}\mathbb{1}^\top - \frac{1}{N}(\sigma^2 + \epsilon)^{-1}\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top\right].$$

$$\left\|\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{N}\boldsymbol{\gamma}\mathbb{1}^\top\right\|_\infty = \frac{2(N-1)}{N}\max_i |\gamma_i|, \tag{2}$$

Take the infinity-norm on both sides, we have:

$$\begin{aligned}
\left\|\frac{\partial \mathrm{LN}(\mathbf{x})}{\partial \mathbf{x}}\right\|_\infty &= (\sigma^2 + \epsilon)^{-\frac{1}{2}}\left\|\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{N}\boldsymbol{\gamma}\mathbb{1}^\top - \frac{1}{N}(\sigma^2 + \epsilon)^{-1}\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top\right\|_\infty \\
&\leq \epsilon^{-\frac{1}{2}}\left(\frac{2(N-1)}{N}\max_i |\gamma_i| + \frac{1}{N}\max_i |\gamma_i| N(N-2)\right) \\
&\leq \epsilon^{-\frac{1}{2}}\max_i |\gamma_i| N.
\end{aligned}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

B PROOF FOR THE LIPSCHITZ CONSTANT OF LRSA

The pair-wise LRSA function is expressed as:

$$S_{ij} = -\frac{\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2^2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}} \tag{3}$$

$$P_i = S_i(X)$$

$$P_{ij} = \frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}} \leq 1$$

To take the derivative P_{ij}, there are two cases.

When $t = j$:

$$\begin{aligned} \frac{\partial P_{ij}}{\partial S_{it}} = \frac{\partial P_{ij}}{\partial S_{ij}} = \frac{\partial}{\partial S_{ij}}\left(\frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}}\right) &= \frac{e^{S_{ij}}(\sum_{t=1}^{n} e^{S_{it}}) - (e^{S_{ij}})^2}{(\sum_{t=1}^{n} e^{S_{it}})^2} \\ &= \frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}}\left(1 - \frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}}\right) = P_{ij}(1 - P_{ij}) \end{aligned} \tag{4}$$

When $t \neq j$:

$$\frac{\partial P_{ij}}{\partial S_{it}} = \frac{\partial}{\partial S_{it}}\left(\frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}}\right) = -\frac{e^{S_{ij}}}{\sum_{t=1}^{n} e^{S_{it}}} \frac{e^{S_{it}}}{\sum_{t=1}^{n} e^{S_{it}}} = -P_{ij}P_{it}$$

$$\frac{\partial P_{ij}}{\partial x_k} = \sum_{t=1}^{n} \frac{\partial P_{ij}}{\partial S_{it}} \frac{\partial S_{it}}{\partial x_k} = P_{ij}(1 - P_{ij}) \frac{\partial S_{ij}}{\partial x_k} - \sum_{t=1, t\neq j}^{n} P_{ij}P_{it} \frac{\partial S_{it}}{\partial x_k} = P_{ij} \frac{\partial S_{ij}}{\partial x_k} - P_{ij} \sum_{t=1}^{n} P_{it} \frac{\partial S_{it}}{\partial x_k} \tag{5}$$

Take the infinity-norm on S_{it}, we get:

$$\begin{aligned} \left\| \frac{\partial S_{it}}{\partial x_k} \right\|_\infty &= \left\| \frac{\partial}{\partial x_k}\left(-\frac{\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2^2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}} \right) \right\|_\infty \\ &= \left\| -\frac{2\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}} \frac{\partial \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2}{\partial x_k} + \frac{\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2^2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}^2} \frac{\partial \left\| X^\top \right\|_{(\infty,2)}}{\partial x_k} \right\|_\infty \\ &\leq \left\| \frac{2\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}} \frac{\partial \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2}{\partial x_k} \right\|_\infty + \left\| \frac{\alpha \left\| x_i^\top W_Q - x_j^\top W_K \right\|_2^2}{\|Q\|_F \, \|X^\top\|_{(\infty,2)}^2} \frac{\partial \left\| X^\top \right\|_{(\infty,2)}}{\partial x_k} \right\|_\infty \\ &\leq \frac{2\alpha}{\|Q\|_F} \frac{\left\| x_i^\top W_Q \right\|_2 + \left\| x_j^\top W_K \right\|_2}{\|X^\top\|_{(\infty,2)}} \left(\frac{\partial \left\| x_j^\top W_Q \right\|_2}{\partial x_k} + \frac{\partial \left\| x_j^\top W_K \right\|_2}{\partial x_k} \right) + \frac{\alpha}{\|Q\|_F} \left(\frac{\left\| x_i^\top W_Q \right\|_2 + \left\| x_j^\top W_K \right\|_2}{\|X^\top\|_{(\infty,2)}} \right)^2 \\ &\leq \frac{2\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} + \frac{\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} \\ &= \frac{3\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} \end{aligned}$$

Thus,

$$\begin{aligned} \left\| \frac{\partial P_{ij}}{\partial x_k} \right\|_\infty &= \left\| P_{ij} \frac{\partial S_{ij}}{\partial x_k} - P_{ij} \sum_{t=1}^{n} P_{it} \frac{\partial S_{it}}{\partial x_k} \right\|_\infty \leq P_{ij} \frac{3\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} + P_{ij} \sum_{t=1}^{n} P_{it} \frac{3\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} \\ &\leq \frac{6\alpha(\|W_Q\|_2 + \|W_K\|_2)^2}{\|Q\|_F} \leq \frac{6\alpha}{\|X\|_F} \cdot \frac{(\|W_Q\|_2 + \|W_K\|_2)^2}{\|W_Q\|_F} \end{aligned}$$

C GAUSSIAN PROCESS LAYER

As an optional module in LRFormer, Gaussian Process (GP) with an RBF kernel following SNGP [Liu et al., 2020] is capable of perserving the distance awareness between input test sample and previously seen training data. This approach makes sure the model returns a uniform distribution over output labels when the input sample is OOD.

To make it end-to-end trainable, the Gaussian Process layer can be implemented a two-layer network:

$$\mathrm{logits}(x) = \Phi(x)\beta, \quad \Phi(x) = \sqrt{\frac{2}{M}} * \cos(Wx + b) \tag{6}$$

Here, x is the input, and W and b are frozen weights initialized randomly from Gaussian and uniform distributions, respectively. $\Phi(x)$ is Random Fourier Features (RFF) [Williams and Rasmussen, 2006]. β is the learnable kernel weight similar to that of a Dense layer. The layer outputs the class prediction $\mathrm{logits}(x) \in \mathbb{R}_{\mathrm{NumClasses}}$.

D EXPERIMENTAL DETAILS

In Table 1, we provide the training details used for reproducing the main results in Tables above. The $Depth = 12$ (pretraining) is the experimental setup of the ImageNet1K dataset pretraining. The other hyperparameters follows the same setting from DeiT III [Touvron et al., 2022].

Table 1: Hyperparameters for LRFormer Training.

Hyperparameters	$Depth = 6$	$Depth = 12$	$Depth = 12$ (pretraining)
Layer depth	6	12	12
Input size	224×224	224×224	224×224
Batch size	128	32	32
Warm-up steps	5	5	5
Optimizer	SGD	AdamW	AdamW
Learning rate	0.01	0.006	0.004
Weight decay	0.05	0.05	0.05
Learning rate scheduler	cosine	cosine	cosine
Training epochs	100	100	100

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *ArXiv*, abs/2006.10108, 2020.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Herv'e J'egou. Three things everyone should know about vision transformers. *ArXiv*, abs/2203.09795, 2022.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.