ICML
International Conference
On Machine Learning

On the Nonlinearity of Layer Normalization

Yunhao Ni, Yuxin Guo, Junlong Jia, Lei Huang*

Beihang University
Beijing, China.

复杂关键软件环境全国重点实验室
State Key Laboratory of Complex & Critical Software Environment

北京航空航天大学人工智能研究院
Institute of Artificial Intelligence, Beihang University

Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

Background

Input ↓

Linear Layer → **store** the main **parameters** of the neural network

↓

Non-Linear Layer → contains the main **expressive power**

↓

Normalization Layer →

Previous:

→ **stablizing** training and **accelerating** optimization

BN: $\hat{x} = \dfrac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$ → Optimization > Fitting

But?

↓

……

↓

Normalization Layer

↓

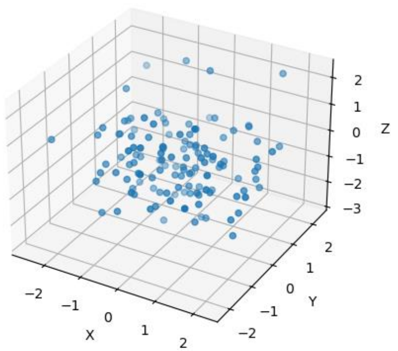
Linear Layer

Output ↓

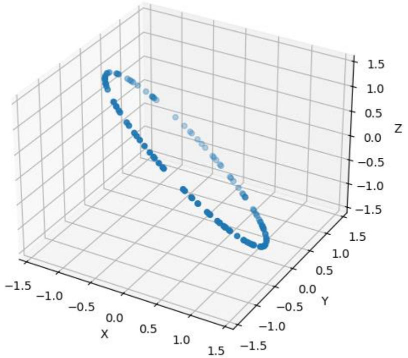
Background

ICML International Conference On Machine Learning

Layer Normalization



$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

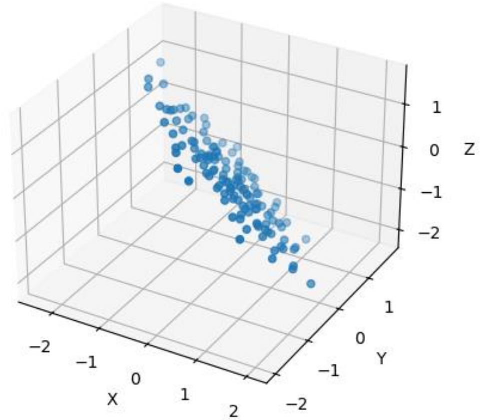


Centering

$$\bar{\boldsymbol{x}} = \boldsymbol{x} - \frac{\mathbf{1}}{\boldsymbol{d}}(\mathbf{1}_d^{\top} \boldsymbol{x}) \cdot \mathbf{1}_d$$

Projected onto the Hyperplane

$$\{\boldsymbol{x} \in \mathbb{R}^d : x_1 + \cdots + x_d = 0\}$$

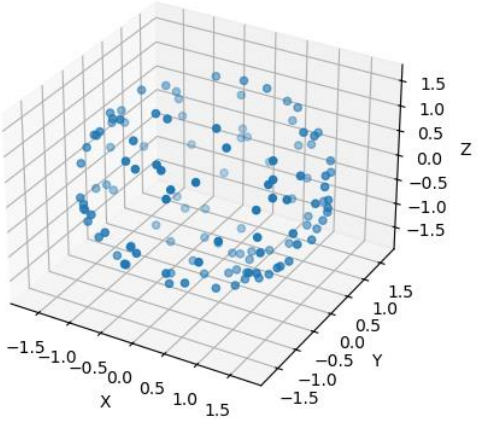


Scaling

$$\hat{\boldsymbol{x}} = \sqrt{d}\,\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|_2}$$

Projected onto the Hypersphere

$$\{\boldsymbol{x} \in \mathbb{R}^d : [x_1]^2 + \cdots + [x_d]^2 = \mathrm{d}\}$$



Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

The Existence of Nonlinearity in LN

ICML International Conference On Machine Learning

Find a linearly invariant index.

$$SS(\boldsymbol{X}_c) = \sum_{i=1}^{m} \|\boldsymbol{x}_{ci} - \overline{\boldsymbol{x}}_c\|^2$$

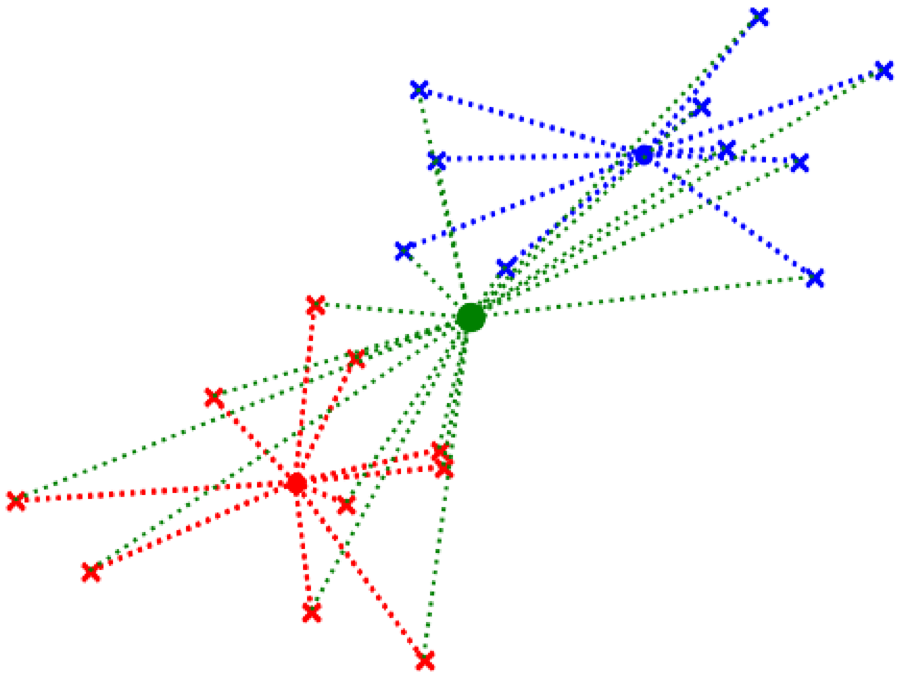
Intra distance

$$SSR(\boldsymbol{X}_1,\boldsymbol{X}_2) = \frac{\underline{SS(\boldsymbol{X}_1)} + SS(\boldsymbol{X}_2)}{\underline{SS([\boldsymbol{X}_1,\boldsymbol{X}_2])}}$$

Total distance

$$LSSR(\boldsymbol{X}_1,\boldsymbol{X}_2) = \inf_{\underline{\varphi \in \mathbb{D}_\varphi(d)}} SSR(\varphi(\boldsymbol{X}_1),\varphi(\boldsymbol{X}_2))$$

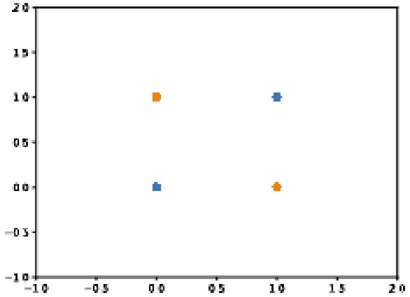
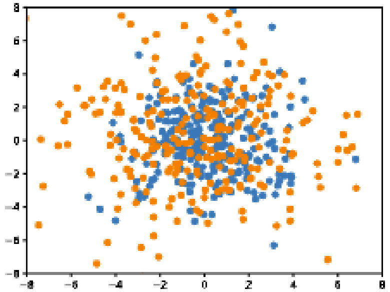
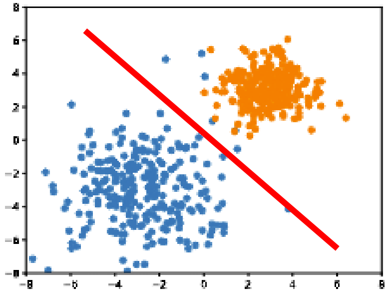
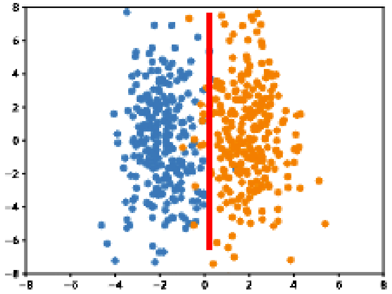
Linear Transformations



The Existence of Nonlinearity in LN

SSR and LSSR

*XOR data

Data				
Linearly separable?	No	No	Yes	Yes
SSR	0.9963	0.9929	0.2304	**0.7365**
LSSR	0.9929	0.9859	**0.1312**	0.2157

LSSR is a better index to describe linear separability than SSR.

The Existence of Nonlinearity in LN

Proof Method

Taylor's Expansion $f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0)$

A brief equation $$\underbrace{SSR\left(v^T\hat{X}_1, v^T\hat{X}_2\right)}_{\text{Linear+Scaling}} = \underbrace{LSSR(X_1, X_2)}_{\text{Lower Bound}} \boxed{-\frac{2(T_1 + T_2)}{T_3}t + o(t)}$$

$< \mathbf{0}$

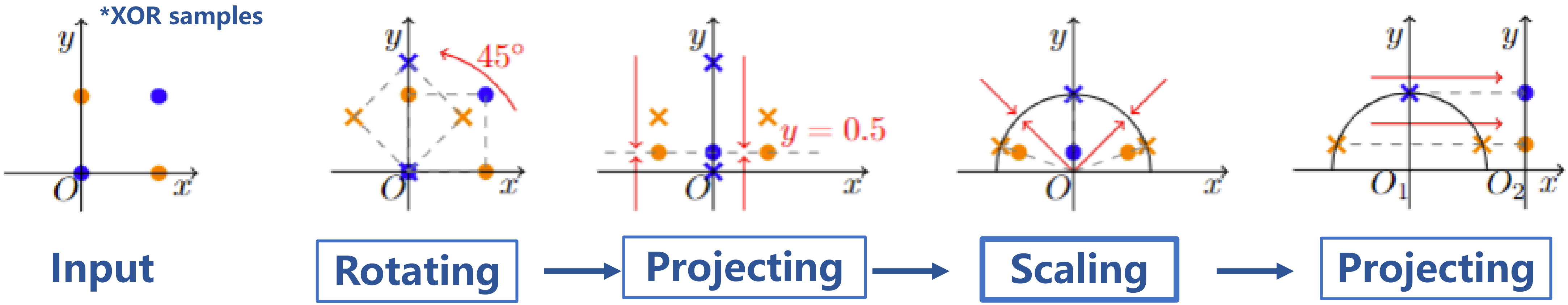
Linear transformations with LN can break LSSR.

Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

Capacity of a Network with LN

XOR



Classify XOR samples with linear transformations and scaling only.

※ **Hint:** Scaling can be represented by LN and linear transformations only.

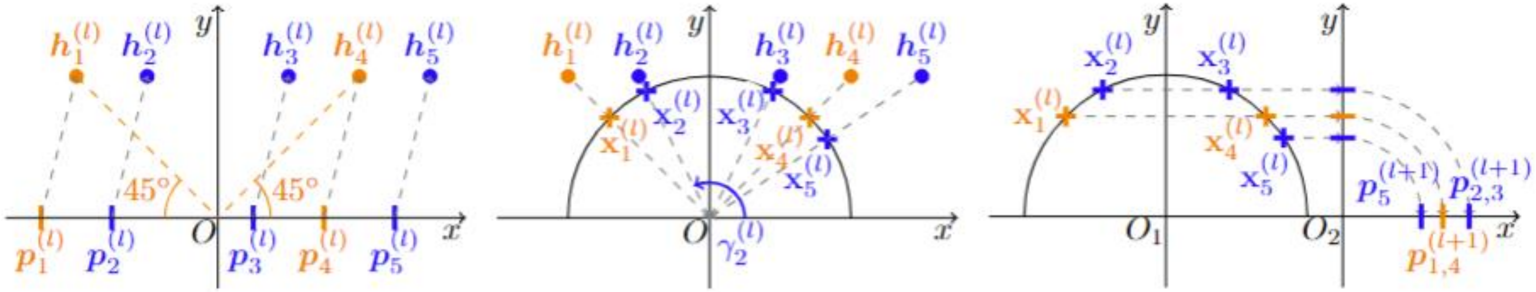
Capacity of a Network with LN

Binary Classification

Algorithm 1 Projection Merge Algorithm

input The initial input $\boldsymbol{P}^{(1)}$.
output The final output $\boldsymbol{P}^{(L)}$.

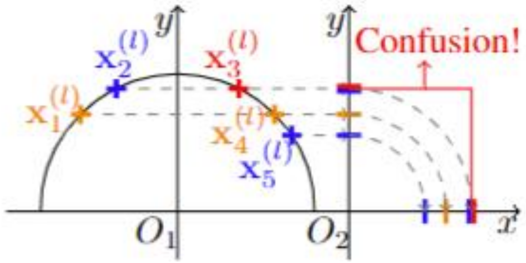
$l \leftarrow 1$;
$\mathbb{P} \leftarrow \{\boldsymbol{p}_1^{(l)}, \boldsymbol{p}_2^{(l)}, \cdots, \boldsymbol{p}_m^{(l)}\}$;
while $\mathbb{P} \neq \emptyset$ **do**
  $i \leftarrow \arg\min_k \{p_k^{(l)} : \boldsymbol{p}_k^{(l)} \in \mathbb{P}\}$;
  $\mathbb{J}_i \leftarrow \{\boldsymbol{p}_j^{(l)} \in \mathbb{P} : \boldsymbol{p}_j^{(l)} \neq \boldsymbol{p}_i^{(l)}, y_j = y_i\}$;
  if $\mathbb{J}_i \neq \emptyset$ **then**
    $j \leftarrow \arg\min_k \{p_k^{(l)} : \boldsymbol{p}_k^{(l)} \in \mathbb{J}_i\}$;
    for $k \leftarrow 1$ to m **do**
      $\boldsymbol{h}_k^{(l)} \leftarrow \boldsymbol{p}_k^{(l)} - \begin{bmatrix} p_i^{(l)} + p_j^{(l)} \\ p_i^{(l)} - p_j^{(l)} \end{bmatrix} / 2$;
      $\mathbf{x}_k^{(l)} \leftarrow \boldsymbol{h}_k^{(l)} / \|\boldsymbol{h}_k^{(l)}\|$;
      $\boldsymbol{p}_k^{(l+1)} \leftarrow \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}_k^{(l)}$;
    end for
    $l \leftarrow l + 1$;
    $\mathbb{P} \leftarrow \{\boldsymbol{p}_1^{(l)}, \boldsymbol{p}_2^{(l)}, \cdots, \boldsymbol{p}_m^{(l)}\}$;
  else
    remove $\boldsymbol{p}_j^{(l)}$ from $\mathbb{P}$, as long as $\boldsymbol{p}_j^{(l)} = \boldsymbol{p}_i^{(l)}$;
  end if
end while
return $\boldsymbol{P}^{(l)}$;



Projecting → **Scaling** → **Projecting**

Binarily classify any m samples with linear transformations and scaling only.

Multiclass Classification



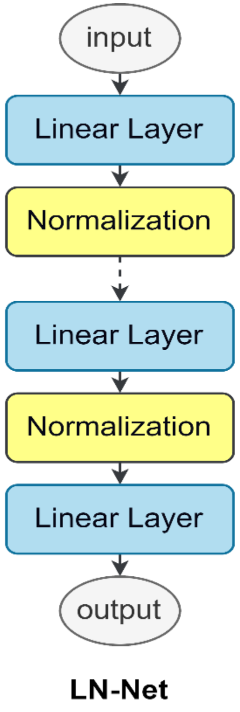
Difference: Confusion is possible.

Solution: Breaking Parallelization.

Capacity of a Network with LN

▶ The first to discuss about the expressive power of layer normalization.

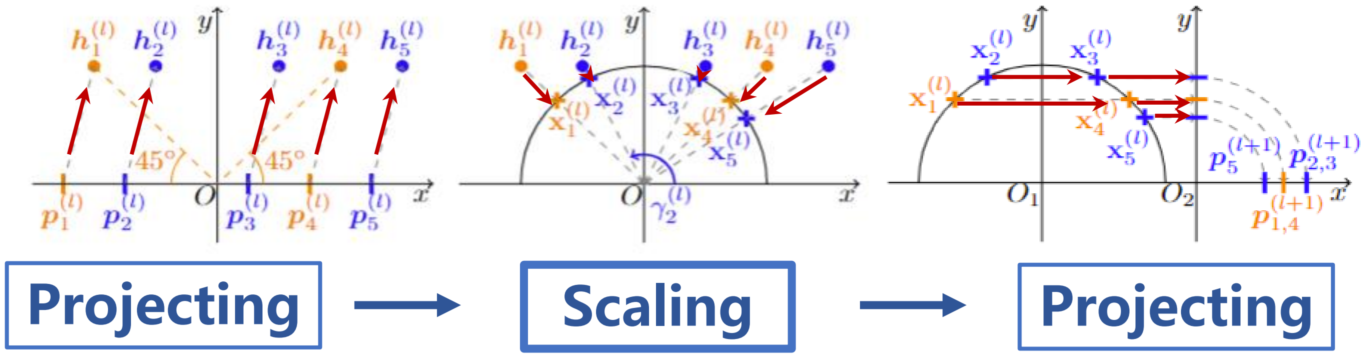
Universal Approximation Theory of LN-Net



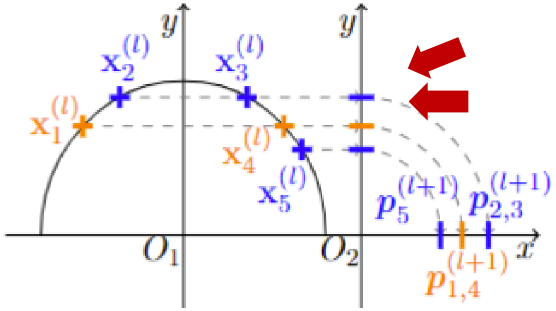
An infinitely deep LN-Net can classify any given m samples correctly.

Computation Operation ⇒ **Merging Operation**

Universal Approximation ⇒ **Universal Classification**



VC dimension



Samples with the same label: at least two are merged with each LN.
Samples with different labels: none are merged.
Width: at least 3.

▶ Given an LN-Net $f_\theta(\cdot)$ with width 3 and depth L its VC dimension $VCdim(f_\theta(\cdot))$ is lower bounded by $L+2$.

Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

Amplify and Exploit the Nonlinearity of LN

ICML International Conference On Machine Learning

Group based LN (LN-G) has stronger nonlinearity than LN

Measurement of Nonlinearity

Hessian: $\mathcal{H}(f;x) = \sum_{i=1}^{d} \left\| \left| \frac{\partial^2 y_i}{\partial x^2} \right| \right\|_F^2$

Noted that $\mathcal{H}(f;x) \geq 0$, and $\mathcal{H}(f;x) = 0$ if and only if f is linear.

$\rightarrow$ We ***assume*** that the larger $\mathcal{H}(f;x)$, the more nonlinearity f contains.

Amplifying Nonlinearity by Group

Proposition: Given $g \leq d/3$, we have $\frac{\mathcal{H}(\psi_G(g;\cdot);x)}{\mathcal{H}(\psi_L(\cdot);x)} \geq 1$. When $g = d/4$, $\frac{\mathcal{H}(\psi_G(g;\cdot);x)}{\mathcal{H}(\psi_L(\cdot);x)} \geq \frac{d}{8}$.

※ $\psi_G(g;\cdot)$ denotes LN-G on $\mathbb{R}^d$ with group number g, $\psi_L(\cdot)$ denotes LN on $\mathbb{R}^d$.

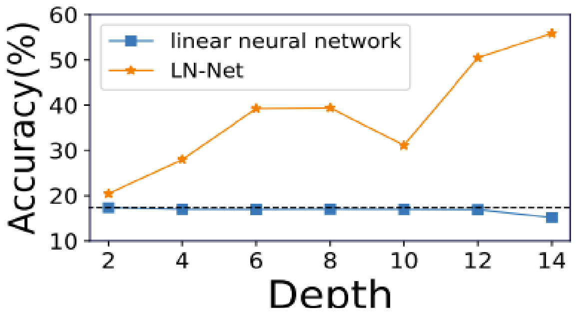
$\rightarrow$ LN-G can amplify the nonlinearity of LN by using appropriated group number.

Amplify and Exploit the Nonlinearity of LN

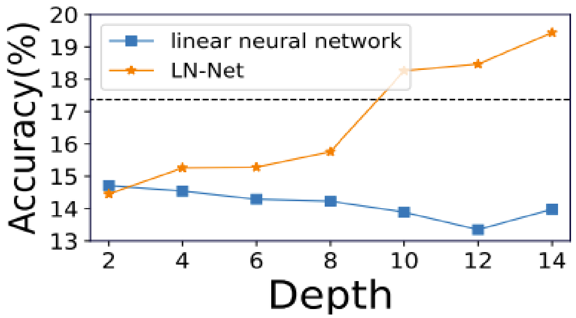
ICML International Conference On Machine Learning

Comparison of Representation Capacity by Fitting Random Labels

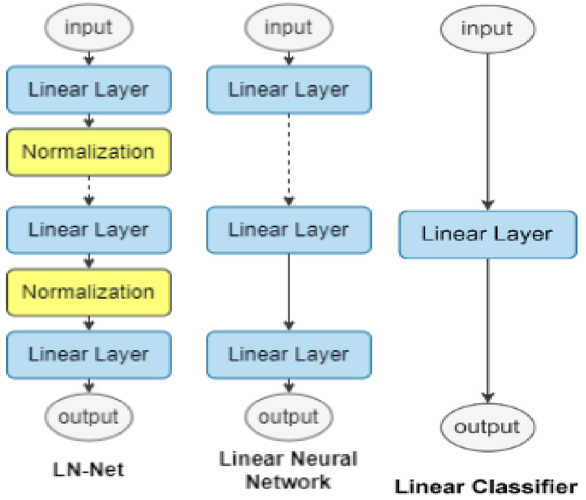
LN-Net & linear neural network & linear classifier



(a) CIFAR-10-RL.

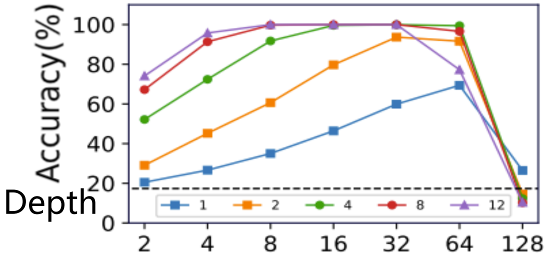


(b) MNIST-RL.

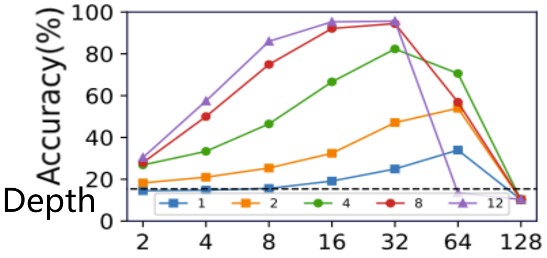


→ LN can break the bound of linearity.

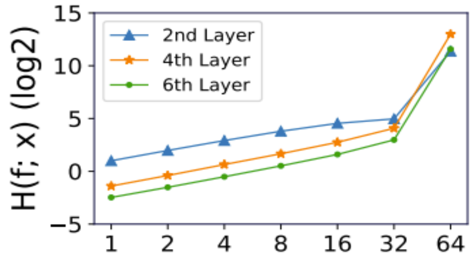
LN-Net (LN-G)



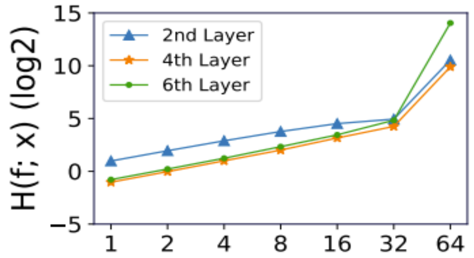
(a) Accuracy on CIFAR-10-RL.



(b) Accuracy on MNIST-RL.



(c) $\mathcal{H}(f;\mathbf{x})$ on CIFAR-10-RL.



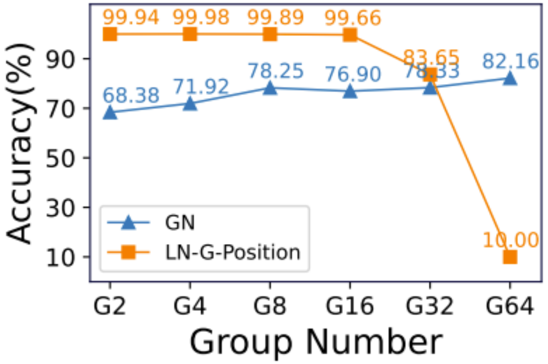
(d) $\mathcal{H}(f;\mathbf{x})$ on MNIST-RL.

Amplify and Exploit the Nonlinearity of LN

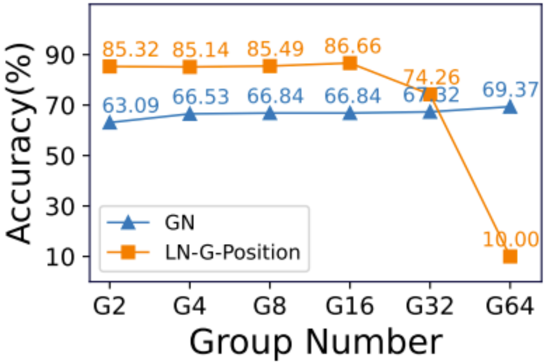
ICML
International Conference
On Machine Learning

Inspiration for Neural Architecture Design

ResNet without ReLU on CIFAR-10



(a) Training.



(b) Test.

Normalization methods	Train Acc(%)	Test Acc(%)
IN	10	10
BN	36.0	39.3
LN	59.5	62.85
GN	82.16	69.37
LN-G-Position	99.66	86.66

Transformer

fairseq-py on IWSLT14 De-EN:(BLEU) LN: 35.01 ± 0.10 ; LN-G: 35.23 ± 0.07

Tiny-ViT on CIFAR-10*: (test Acc)* LN: 88.81% ; LN-G: 89.26%

Outline

- Background
- The Existence of Nonlinearity in LN
- Capacity of a Network with LN
- Amplify and Exploit the Nonlinearity of LN
- Conclusion

Conclusion

- Mathematically demonstrated that LN is a nonlinear transformation.
- Theoretically showed the representation capacity of an LN-Net in correctly classifying samples with any label assignment.
- Call for reconsidering the analyses of the representation capacity of a network with normalization layer.

Thanks for your attention!