DNT: a Deeply Normalized Transformer

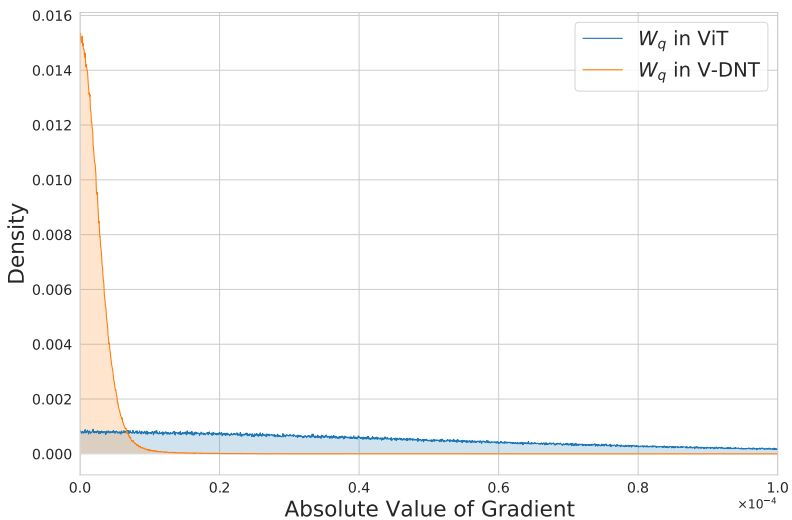
That Can Be Trained by Momentum SGD

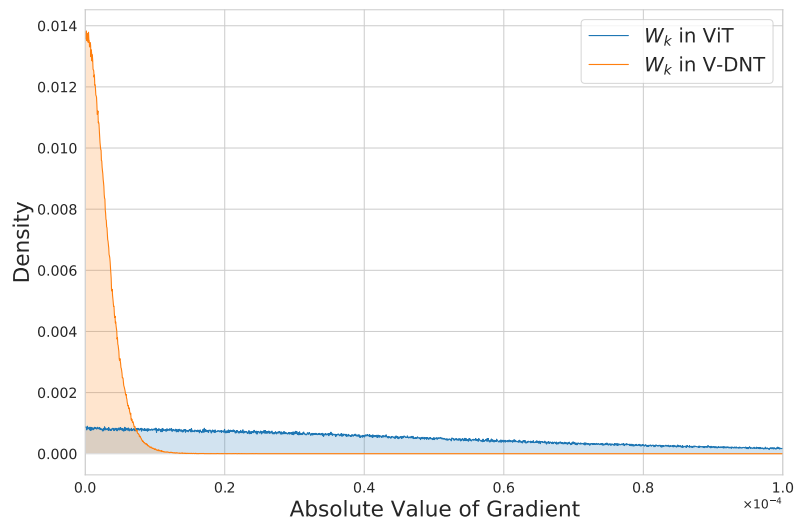
Xianbiao Qi[1], Marco Chen[2], Wenjie Xiao[3], Jiaquan Ye[1],
Yelin He[1], Chun-Guang Li*[4], Zhouchen Lin*[5,6]
[1]Intellifusion Inc., [2]Tsinghua University, [3]JHU,
[4]BUPT, [5]Peking University

ICLR 2026

Motivation: Why Can't SGD Train Transformers?

- CNNs can be trained with vanilla momentum SGD (mSGD)
- Transformers **require** adaptive optimizers like AdamW
- **Root cause:** Gradients in Transformers exhibit heavy-tailed distributions





Standard ViT: gradients spread in $[0, 10^{-4}]$ vs V-DNT: concentrated around $[0, 10^{-5}]$

Root Cause: Jacobian Matrix Perspective

For a forward layer $\boldsymbol{x}^l = \boldsymbol{W}^l \boldsymbol{x}^{l-1}$, the gradient is:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^l} = \underbrace{\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}^{l+1}}}_{\text{upstream grad}} \cdot \underbrace{\frac{\partial \boldsymbol{x}^{l+1}}{\partial \boldsymbol{x}^l}}_{\text{Jacobian}} \cdot \boldsymbol{x}^{l-1\top}$$

Key insight: Heavy-tail gradients $\Leftarrow$ diverse singular values in the Jacobian matrix

Causes of diverse singular values

1. Weight matrices $\boldsymbol{W}$ have very diverse singular values
2. Activations span widely $\Rightarrow$ uneven Jacobian singular values
3. Large condition number $\Rightarrow$ stretches input differently along directions

Solution: Constrain Jacobian singular values via proper normalization

Five Normalization Settings

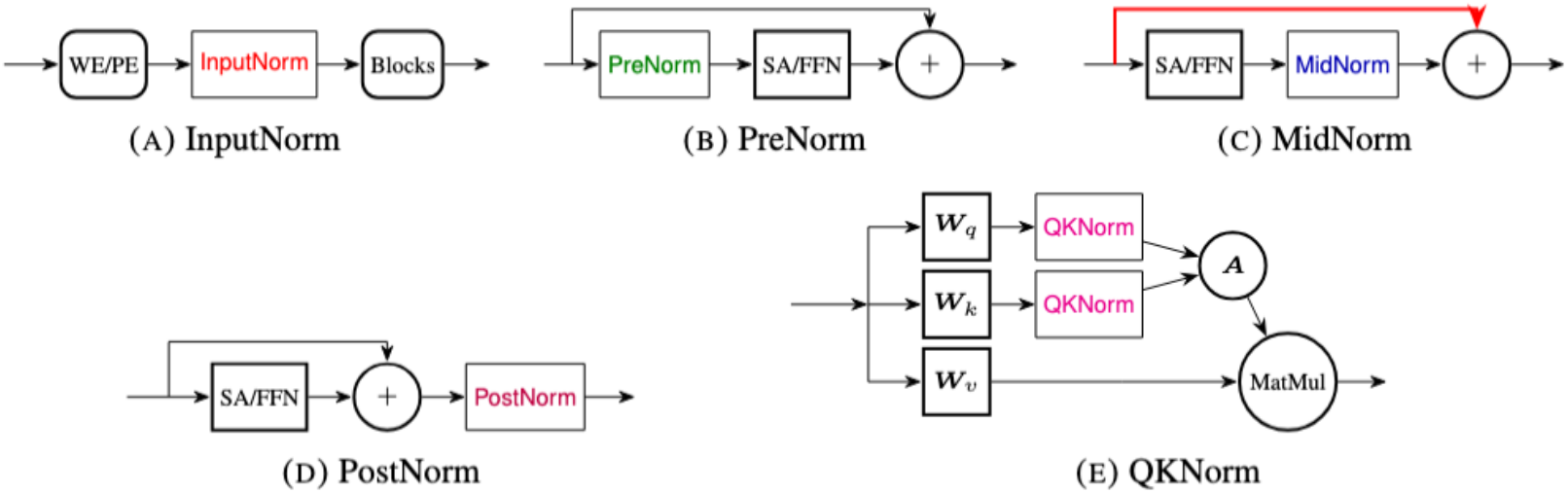


Figure: Five different normalization methods. The only difference between them is the position of normalization. In (A), "WE/PE" indicates Word Embedding (WE) and Patch Embedding (PE).

DNT Architecture

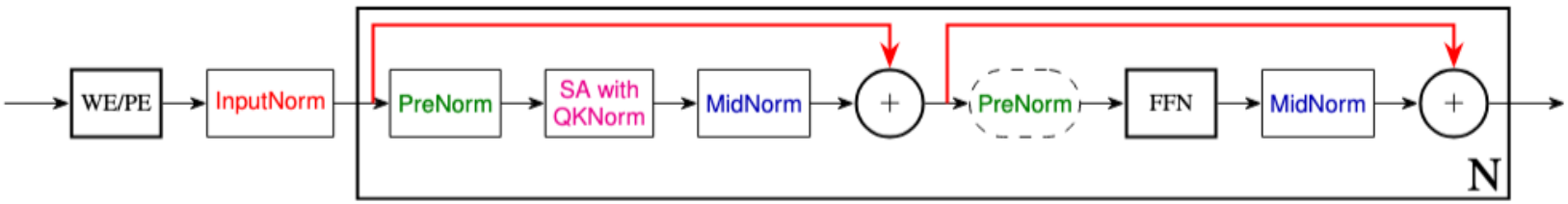


Figure: DNT architecture. The second PreNorm marked with dashed and rounded corners is optional. By default, we do not use the second PreNorm.

Why InputNorm?

For a residual network: $\boldsymbol{x}^{l+1} = \boldsymbol{x}^0 + f(\boldsymbol{x}^0) + f(\boldsymbol{x}^1) + \cdots + f(\boldsymbol{x}^l)$

In high dimensions, random vectors are almost orthogonal, so:

$$\|\boldsymbol{x}^{l+1}\|_2 \asymp \sqrt{\|\boldsymbol{x}^0\|_2^2 + \|f(\boldsymbol{x}^0)\|_2^2 + \cdots + \|f(\boldsymbol{x}^l)\|_2^2}$$

The Jacobian of RMSNorm is:

$$\frac{\partial \,\mathrm{RMSN}(\boldsymbol{x}^{l+1})}{\partial \boldsymbol{x}^{l+1}} = \frac{\sqrt{d}}{\sqrt{\|\boldsymbol{x}^{l+1}\|_2^2 + \epsilon}} \,\mathrm{diag}(\gamma) \left(\boldsymbol{I} - \frac{\boldsymbol{x}^{l+1} \boldsymbol{x}^{l+1\top}}{\|\boldsymbol{x}^{l+1}\|_2^2 + \epsilon} \right)$$

Why PreNorm?

Self-attention: $\boldsymbol{Y} = \boldsymbol{W}_v\boldsymbol{X}\boldsymbol{A}$, where $\boldsymbol{A} = \text{softmax}(\frac{\boldsymbol{X}^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k \boldsymbol{X}}{\sqrt{d_q}})$

The Jacobian matrix of self-attention w.r.t. $\boldsymbol{X}$:

$$\frac{\partial \text{vec}(\boldsymbol{Y})}{\partial \text{vec}(\boldsymbol{X})} = (\boldsymbol{A}^\top \otimes \boldsymbol{W}_v) + (\boldsymbol{I}_n \otimes \boldsymbol{W}_v\boldsymbol{X})\frac{\boldsymbol{J}}{\sqrt{d_q}}\Big((\boldsymbol{X}^\top \boldsymbol{W}_k^\top \boldsymbol{W}_q \otimes \boldsymbol{I}_n)\boldsymbol{C} + (\boldsymbol{I}_n \otimes \boldsymbol{X}^\top \boldsymbol{W}_q^\top \boldsymbol{W}_k)\Big)$$

Why MidNorm?

For FFN: $\boldsymbol{z} = \boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x})$, followed by MidNorm
$\boldsymbol{y} = \mathrm{MidNorm}(\boldsymbol{z})$

The Jacobian of FFN + MidNorm:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \sqrt{d}\,\mathrm{diag}(\gamma)\left(\boldsymbol{I} - \frac{\boldsymbol{z}\boldsymbol{z}^\top}{\|\boldsymbol{z}\|_2^2}\right) \cdot \underbrace{\frac{\boldsymbol{W}_2\,\mathrm{diag}\left(\mathbf{1}(\boldsymbol{W}_1\boldsymbol{x} > \mathbf{0})\right)\boldsymbol{W}_1}{\|\boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x})\|_2}}_{\boldsymbol{M}}$$

Why **NOT** PostNorm?

PostNorm normalizes *after* the residual connection:

$$\boldsymbol{x}^{l+1} = \text{PostNorm}(\boldsymbol{z}^{l+1}), \quad \text{where} \quad \boldsymbol{z}^{l+1} = \boldsymbol{x}^{l} + f(\boldsymbol{x}^{l}; \boldsymbol{W}^{l+1})$$

The Jacobian:

$$\frac{\partial \boldsymbol{x}^{l+1}}{\partial \boldsymbol{z}^{l+1}} = \frac{\sqrt{d}}{\|\boldsymbol{z}^{l+1}\|_2} \operatorname{diag}(\gamma) \left(\boldsymbol{I} - \frac{\boldsymbol{z}^{l+1} \boldsymbol{z}^{l+1}{}^{\top}}{\|\boldsymbol{z}^{l+1}\|_2^2} \right)$$

PostNorm is **unstable** — we do **not** use it in DNT.

Why QKNorm?

Normalize queries and keys:

$$\boldsymbol{q}'_i = \sqrt{d_h}\,\mathrm{diag}(\gamma_q)\frac{\boldsymbol{W}_q\boldsymbol{x}_i}{\|\boldsymbol{W}_q\boldsymbol{x}_i\|_2}, \quad \boldsymbol{k}'_j = \sqrt{d_h}\,\mathrm{diag}(\gamma_k)\frac{\boldsymbol{W}_k\boldsymbol{x}_j}{\|\boldsymbol{W}_k\boldsymbol{x}_j\|_2}$$

The gradient of the logit $P'_{ij} = \boldsymbol{q}'^{\top}_i \boldsymbol{k}'_j$:

$$\frac{\partial P'_{ij}}{\partial \boldsymbol{x}} = \sqrt{d_h}\,\mathrm{diag}(\gamma_q)\,\boldsymbol{k}'^{\top}_j \left(\boldsymbol{I} - \frac{\boldsymbol{q}'_i\boldsymbol{q}'^{\top}_i}{\|\boldsymbol{q}'_i\|_2^2}\right)\frac{\boldsymbol{W}_q}{\|\boldsymbol{W}_q\boldsymbol{x}_i\|_2} + \sqrt{d_h}\,\mathrm{diag}(\gamma_k)\,\boldsymbol{q}'^{\top}_i \left(\boldsymbol{I} - \frac{\boldsymbol{k}'_j\boldsymbol{k}'^{\top}_j}{\|\boldsymbol{k}'_j\|_2^2}\right)\frac{\boldsymbol{W}_k}{\|\boldsymbol{W}_k\boldsymbol{x}_j\|_2}$$

Experimental Results

Optimizer	Model	**ImageNet (Acc. ↑)** 307M	632M	**OpenWebText (Val Loss ↓)** 124M	774M	1436M
AdamW	ViT/GPT2	81.7	80.8	2.867	2.492	2.435
AdamW	V-DNT/L-DNT	**82.1**	**81.9**	**2.863**	**2.481**	**2.396**
mSGDW	ViT/GPT2	78.2	73.5	2.906	2.544	2.472
mSGDW	V-DNT/L-DNT	**81.5**	**81.2**	**2.849**	**2.503**	**2.408**

Conclusion

1. **Theoretical analysis:** heavy-tail gradients stem from diverse Jacobian singular values
2. Each normalization has a **clear theoretical justification**
3. **DNT architecture:** four strategically placed normalizations to constrain Jacobians