Always Skip Attention (Supplementary Material)

Yiping Ji[1,2], Hemanth Saratchandran[1], Peyman Moghadam[2,3], Simon Lucey[1]
[1]Adelaide University, [2]CSIRO, [3]Queensland University of Technology
{yiping.ji, hemanth.saratchandran, simon.lucy}@adelaide.edu.au,
{yiping.ji, peyman.moghadam}@csiro.au

A. Theoretical framework

In this section, we give the proof of our Proposition 4.1.

Proof. By the properties of 2-norm condition number $\kappa(\mathbf{AB}) \leq \kappa(\mathbf{A}) \cdot \kappa(\mathbf{B})$, then we can rewrite left sides of Eq.(5) as follows:

$$\kappa(\mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}_{\mathrm{V}}) \leq \kappa(\mathbf{X}\mathbf{W}_{\mathrm{Q}})\cdot\kappa(\mathbf{X}\mathbf{W}_{\mathrm{K}})\cdot\kappa(\mathbf{X}\mathbf{W}_{\mathrm{V}}). \tag{10}$$

By definition we have:

$$\kappa(\mathbf{X}\mathbf{W}_{\mathrm{Q}}) \leq \kappa(\mathbf{X}) \cdot \kappa(\mathbf{W}_{\mathrm{Q}}). \tag{11}$$

Since $\mathbf{W}_{\mathrm{Q}}$ does not depend on data $\mathbf{X}$, we denote $\kappa(\mathbf{W}_{\mathrm{Q}}) = C_{\mathrm{Q}}$. Then we have

$$\kappa(\mathbf{X}\mathbf{W}_{\mathrm{Q}}) \leq C_{\mathrm{Q}} \cdot \kappa(\mathbf{X}). \tag{12}$$

Similarly, this holds for $\kappa(\mathbf{X}\mathbf{W}_{\mathrm{K}})$ and $\kappa(\mathbf{X}\mathbf{W}_{\mathrm{V}})$ as well. Since all weight matrices stem from a zero mean i.i.d. with high statistical likelihood $C = C_Q \cdot C_K \cdot C_V$ will tend towards unity. This completes the proof. □

Next, we give the proof of our Proposition 4.2

Proof. By the properties of 2-norm condition number $\kappa(\mathbf{AB}) \leq \kappa(\mathbf{A}) \cdot \kappa(\mathbf{B})$, then we can rewrite left sides of Eq.(7) as follows:

$$\begin{aligned} &\kappa(\mathbf{XM} + \mathbf{X}) && (13)\\ =&\kappa(\mathbf{X}(\mathbf{M} + \mathbf{I})) && (14)\\ \leq&\kappa(\mathbf{X}) \cdot \kappa(\mathbf{M} + \mathbf{I}). && (15)\end{aligned}$$

Then we take the notation $C = C_Q \cdot C_K \cdot C_V$ and we have

$$\kappa(\mathbf{M}) \leq \frac{C \cdot \sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} \tag{16}$$

$$\kappa(\mathbf{M} + \mathbf{I}) \leq \frac{C \cdot \sigma_{\text{max}}^2 + 1}{\sigma_{\text{min}}^2 + 1} \tag{17}$$

Where $C = C_Q \cdot C_K \cdot C_V$ σ_{max} and σ_{min} represents maximal and minimal singular value of $\mathbf{X}$. Then we have

$$\kappa(\mathbf{X}) \cdot \kappa(\mathbf{M}) \leq \frac{C \cdot \sigma_{\text{max}}^3}{\sigma_{\text{min}}^3} \tag{18}$$

$$\kappa(\mathbf{X}) \cdot \kappa(\mathbf{M} + \mathbf{I}) \leq \frac{C \cdot \sigma_{\text{max}}^3 + \sigma_{\text{max}}}{\sigma_{\text{min}}^3 + \sigma_{\text{min}}} \tag{19}$$

For a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ whose entries are i.i.d with mean zero, it holds with high probability $\sigma_{\text{min}} < 1$ and $\sigma_{\text{max}} > 1$. Then we have

$$C \cdot \sigma_{\text{max}}^3 + \sigma_{\text{max}} \approx C \cdot \sigma_{\text{max}}^3 \tag{20}$$

$$\sigma_{\text{min}}^3 + \sigma_{\text{min}} > \sigma_{\text{min}}^3 \tag{21}$$

Therefore, $\kappa(\mathbf{X}) \cdot \kappa(\mathbf{M} + I) \ll \kappa(\mathbf{X}) \cdot \kappa(\mathbf{M})$
This completes our proof for Proposition 4.2.

□

B. ConvMixer

In this section, we demonstrate how removing skip connections in Convolutional Neural Networks (CNNs) impacts performance. ConvMixer, an extremely simple model inspired by Vision Transformers (ViTs), was introduced in [28] and remains widely used within the research community. ConvMixer consists of LL ConvMixer blocks, formally defined as follows:

$$\mathbf{X}_l^{'} = \text{BN}(\sigma\{\text{ConvDepthwise}(\mathbf{X}_l)\} + \mathbf{X}_l \tag{22}$$

$$\mathbf{X}_{l+1} = \text{BN}(\sigma\{\text{ConvPointwise}(\mathbf{X}_l^{'})\} \tag{23}$$

where $\mathbf{X} \in \mathbb{R}^{h\times n/p\times n/p}$, BN is batch normalization, and σ is activation function. h is feature dimension, n is image height and width and p is patch size.

For the ConvDepthwise transformation, we analyze the condition number in the linear case and in the absence of batch normalization. We have:

$$\kappa(\mathbf{W}_{\text{CD}}\mathbf{X}_l) \leq \kappa(\mathbf{W}_{\text{CD}}) \cdot \kappa(\mathbf{X}_l) \leq C_{\text{CD}} \cdot \kappa(\mathbf{X}_l) \tag{24}$$

Since $\mathbf{W}_{\text{CD}}$ does not depend on data, we denote $\kappa(\mathbf{W}_{\text{CD}}) = C_{\text{CD}}$. Compared to the Self-Attention Mechanism, the ConvDepthwise transformation demonstrates better conditioning with a lower bound on condition numbers.

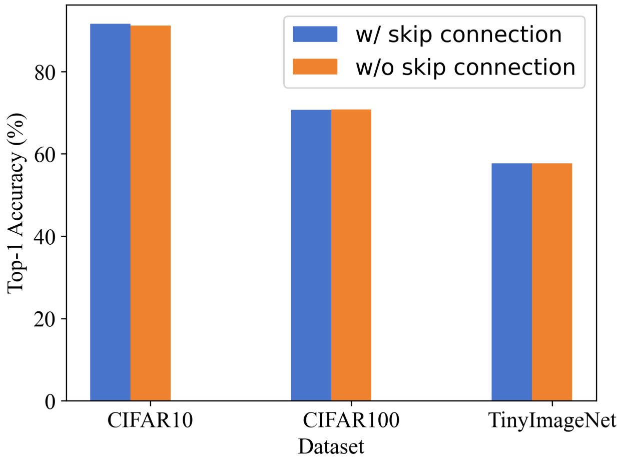

Figure 6. Top-1 Accuracy using the ConvMixer-Tiny model on three different datasets with and without skip connections.

In Fig. 6, we demonstrate that the performance of the ConvMixer Tiny model does not degrade when skip connections are removed. This observation contrasts with Vision Transformers (ViTs), where the absence of skip connections in the self-attention mechanism leads to a noticeable performance drop.

C. Empirical Neural Tangent Kernel

The Neural Tangent Kernel (NTK) describes the evolution of deep neural networks during training by gradient descent. In this paper, we propose to measure the condition of the self-attention output embeddings as a proxy for its Jacobian condition, since analyzing the transformer model NTK is computationally expensive and unrealistic. However, in Fig. 7 and Fig. 8, we empirically demonstrate the Jacobian of the self-attention mechanism using SVDTG and DCTTG. Intrinsically, the self-attention exhibits a disproportionately ill-conditioned spectrum, and using our TG methods, we observe an improvement in this regard.

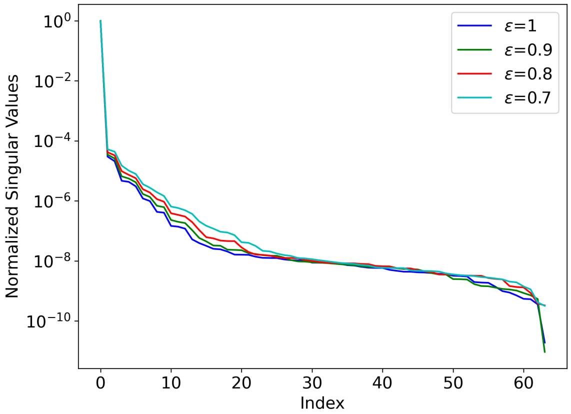

Figure 7. Spectrum of Jacobian of one SAB using **SVDTG**. Lower ϵ represents to better input token condition.

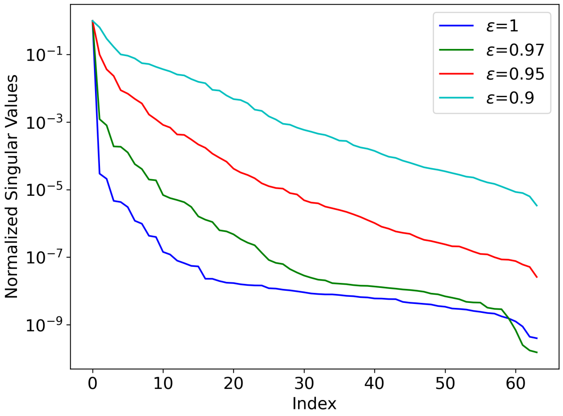

Figure 8. Spectrum of Jacobian of one SAB using **DCTTG**. Lower ϵ represents better input token condition.