ATTENTION IS NOT *ALL* YOU NEED

PURE ATTENTION LOSES RANK DOUBLY EXPONENTIALLY WITH DEPTH

Authored by Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas

Presented by Scott Holtshousen and Kevin Mills

CS 886 – Learning Theory for Modern AI

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Introduction

- Empirical results show that attention-based architectures are highly effective

- Limited theoretical grounding to justify/understand effectiveness

Introduction

- There are major mathematical problems with deep, pure self-attention networks

- Transformers seem to overcome these problems. Why?

Standard Mathematical Framework for Transformers

$$X \in \mathbb{R}^{n \times d_{\text{in}}} \quad (n \text{ tokens}, d_{\text{in}} \text{ embedding dimension})$$

$$Q_h = XW_{Q,h},\ K_h = XW_{K,h},\ V_h = XW_{V,h}$$

Standard Mathematical Framework for Transformers

$$X \in \mathbb{R}^{n \times d_{\text{in}}} \quad (n \text{ tokens}, d_{\text{in}} \text{ embedding dimension})$$

$$Q_h = XW_{Q,h},\ K_h = XW_{K,h},\ V_h = XW_{V,h}$$

$$P_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_{qk}}}\right) \in \mathbb{R}^{n \times n}$$

$$\text{Head}_h(X) = P_h X W_{V,h} + \mathbf{1} b_{V,h}^\top$$

Standard Mathematical Framework for Transformers

$$X \in \mathbb{R}^{n \times d_{\text{in}}} \quad (n \text{ tokens}, d_{\text{in}} \text{ embedding dimension})$$

$$Q_h = XW_{Q,h},\ K_h = XW_{K,h},\ V_h = XW_{V,h}$$

$$P_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_{qk}}}\right) \in \mathbb{R}^{n \times n}$$

$$\text{Head}_h(X) = P_h X W_{V,h} + \mathbf{1} b_{V,h}^\top$$

$$\text{MHA}(X) = [\text{Head}_1(X), \ldots, \text{Head}_H(X)]\, W_O + \mathbf{1} b_O^\top$$

$$X^{(l)} = \text{MLP}\Big(\text{LN}\big(X^{(l-1)} + \text{MHA}(X^{(l-1)})\big)\Big)$$

Standard Mathematical Framework for Transformers

$$X \in \mathbb{R}^{n \times d_{\text{in}}} \quad (n \text{ tokens}, d_{\text{in}} \text{ embedding dimension})$$

$$Q_h = XW_{Q,h},\ K_h = XW_{K,h},\ V_h = XW_{V,h}$$

$$P_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_{qk}}}\right) \in \mathbb{R}^{n \times n}$$

$$\text{Head}_h(X) = P_h X W_{V,h} + \mathbf{1} b_{V,h}^\top$$

$$\text{MHA}(X) = [\text{Head}_1(X), \ldots, \text{Head}_H(X)]\, W_O + \mathbf{1} b_O^\top$$

$$X^{(l)} = \text{MLP}\Big(\text{LN}\big(X^{(l-1)} + \text{MHA}(X^{(l-1)})\big)\Big)$$

$$X^{(L)} = (T_L \circ T_{L-1} \circ \cdots \circ T_1)(X)$$

Path Decomposition of a Pure Self-Attention Network

$$\mathrm{SA}_h(\boldsymbol{X}) = \boldsymbol{P}_h \boldsymbol{X} \boldsymbol{W}_{V,h} + \mathbf{1}\boldsymbol{b}_{V,h}^\top.$$

1. Output of a single self-attention head

$$\begin{aligned}\mathrm{SA}(\boldsymbol{X}) &= \mathbf{1}[\boldsymbol{b}_{O,1}^\top, \cdots, \boldsymbol{b}_{O,H}^\top] + [\mathrm{SA}_1(\boldsymbol{X}), \cdots, \mathrm{SA}_H(\boldsymbol{X})]\ [\boldsymbol{W}_{O,1}^\top, \cdots, \boldsymbol{W}_{O,H}^\top]^\top \\ &= \sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X} \boldsymbol{W}_h + \mathbf{1}\boldsymbol{b}_O^\top,\end{aligned}$$

2. Output of multiple self-attention heads (which are concatenated together).

$$\begin{aligned}\boldsymbol{X}^L &= \sum_{h\in[H]} \boldsymbol{P}_h^L \boldsymbol{X}^{L-1} \boldsymbol{W}_h^L \\ &= \sum_{h\in[H]} \boldsymbol{P}_h^L \left(\sum_{h'\in[H]} \boldsymbol{P}_{h'}^{L-1} \boldsymbol{X}^{L-2} \boldsymbol{W}_{h'}^{L-1} \right) \boldsymbol{W}_h^L\end{aligned}$$

3. Formula for recursion: express the output of the next layer in terms of the output of the previous layer.

Bias terms are temporarily being ignored

$$\boldsymbol{X}^L = \sum_{h_1,\ldots,h_L\in[H]^L} (\boldsymbol{P}_{h_L}^L \cdots \boldsymbol{P}_{h_1}^1)\, \boldsymbol{X}\, (\boldsymbol{W}_{h_1}^1 \cdots \boldsymbol{W}_{h_L}^L).$$

4. Recursive formula for the L-th layer output if we (temporarily) ignore biases.

Mathematical Framework for Transformers in This Paper

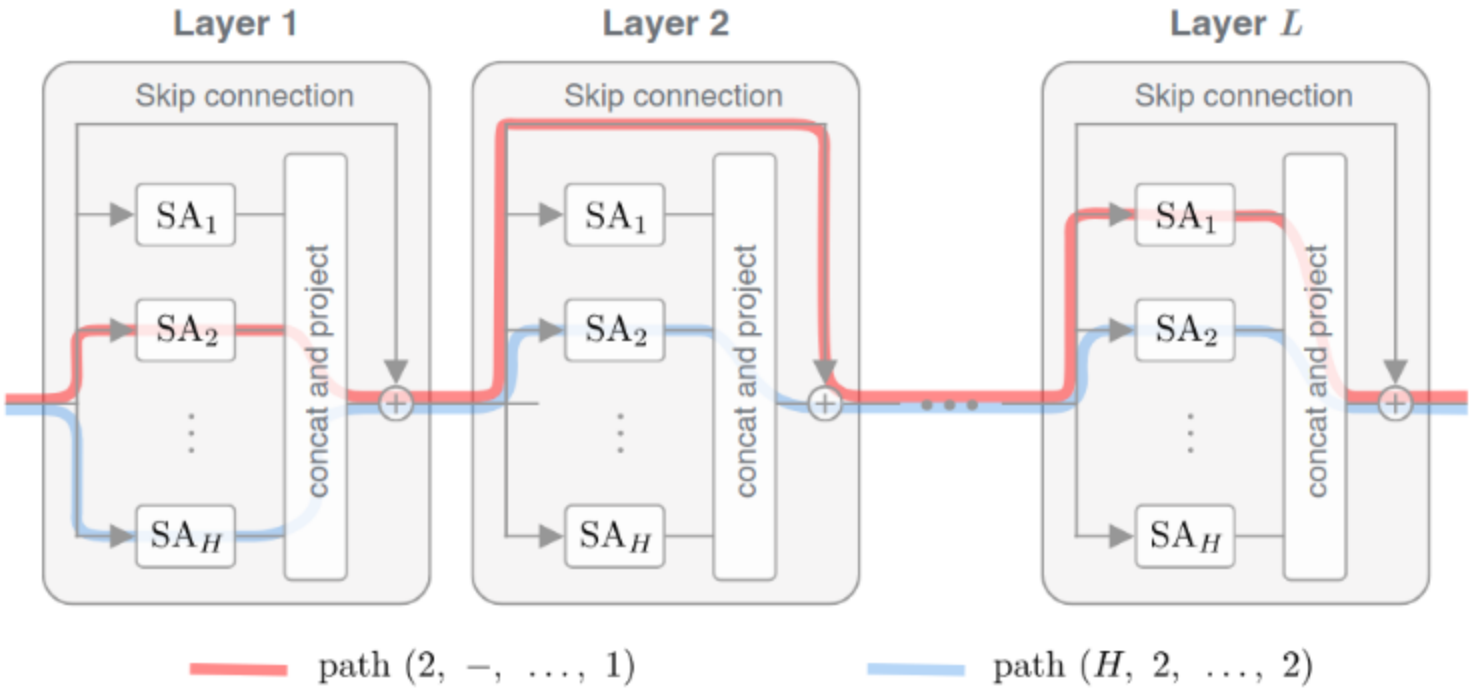


Figure 1: Two paths in a deep Self-Attention Network (SAN) with H heads and L layers. At each layer, a path can go through one of the heads or bypass the layer. Adding an MLP block after each attention layer forms the transformer architecture.

Legend: Path Decomposition of a Pure Self-Attention Network

$$\boldsymbol{P}_h = \mathrm{softmax}\big(d_{qk}^{-\frac{1}{2}}(\boldsymbol{X}\boldsymbol{W}_{Q,h} + \mathbf{1}\boldsymbol{b}_{Q,h}^\top)(\boldsymbol{X}\boldsymbol{W}_{K,h} + \mathbf{1}\boldsymbol{b}_{K,h}^\top)^\top\big)$$

$$= \mathrm{softmax}(d_{qk}^{-\frac{1}{2}}(\boldsymbol{X}\boldsymbol{W}_{QK,h}\boldsymbol{X}^\top + \mathbf{1}\boldsymbol{b}_{Q,h}^\top\boldsymbol{W}_{K,h}^\top\boldsymbol{X}^\top)),$$

$$\boldsymbol{W}_{QK,h} = \boldsymbol{W}_{Q,h}\boldsymbol{W}_{K,h}^\top$$

$$\mathrm{SA}_h(\boldsymbol{X}) = \boldsymbol{P}_h\boldsymbol{X}\boldsymbol{W}_{V,h} + \mathbf{1}\boldsymbol{b}_{V,h}^\top.$$

$$\boldsymbol{W}_h = \boldsymbol{W}_{V,h}\,\boldsymbol{W}_{O,h}^\top \text{ and } \boldsymbol{b}_O = \textstyle\sum_h \boldsymbol{b}_{O,h}.$$

Why can we ignore the biases?

$$
\begin{aligned}
X^L &= \sum_{h_1 \in [H]} P_{h_1}^L \left(X^{L-1}\right) W_{h_1}^L + 1 b_{O,L}^T \\
X^{L+1} &= \sum_{h_2 \in [H]} P_{h_2}^{L+1} \left(X^L\right) W_{h_2}^{L+1} + 1 b_{O,L+1}^T \\
&= \sum_{h_2 \in [H]} P_{h_2}^{L+1} \left(\sum_{h_1 \in [H]} P_{h_1}^L X^{L-1} W_{h_1}^L + 1 b_{O,L}^T \right) W_{h_2}^{L+1} + 1 b_{O,L+1}^T \\
&= \sum_{h_2 \in [H]} P_{h_2}^{L+1} \left(\sum_{h_1 \in [H]} P_{h_1}^L X^{L-1} W_{h_1}^L \right) W_{h_2}^{L+1} + \sum_{h_2 \in [H]} P_{h_2}^{L+1} \left(1 b_{O,L}^T \right) W_{h_2}^{L+1} + 1 b_{O,L+1}^T \\
&= \sum_{h_2 \in [H]} P_{h_2}^{L+1} \left(\sum_{h_1 \in [H]} P_{h_1}^L X^{L-1} W_{h_1}^L \right) W_{h_2}^{L+1} + \sum_{h_2 \in [H]} \left(1 b_{O,L}^T \right) W_{h_2}^{L+1} + 1 b_{O,L+1}^T
\end{aligned}
$$

tl;dr: the biases work out to be a constant that doesn't depend on the input, because P is row-stochastic.

Path Decomposition of a Pure Self-Attention Network

Theorem 2.1 (Path decomposition of SAN). *The output of a depth L self-attention network with H heads per layer (including biases and skip connections) is given by*

$$SAN(\boldsymbol{X}) = \sum_{path \in [H]^L} \boldsymbol{P}_{path} \boldsymbol{X} \boldsymbol{W}_{path} + \boldsymbol{1}\boldsymbol{b}^\top, \tag{1}$$

where $\boldsymbol{P}_{path} = \boldsymbol{P}^L_{h_L} \cdots \boldsymbol{P}^1_{h_1}$ is an input-dependent stochastic matrix, whereas $\boldsymbol{W}_{path} = \boldsymbol{W}^1_{h_1} \cdots \boldsymbol{W}^L_{h_L}$ and $\boldsymbol{b}$ do not depend on the input.

But path decomposition is actually of limited importance for the paper's proofs. It's not required to prove the fundamental bounds, and its not required for the recursive analysis used in the bounding.

$$\|res(SA(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4H\gamma\beta}{\sqrt{d_{qk}}} \|res(\boldsymbol{X})\|^3_{1,\infty},$$

$$\begin{aligned}
\|\mathrm{res}(\boldsymbol{X}^L)\|_{1,\infty} &\leq \frac{4\gamma\beta}{\sqrt{d_{qk}}} \|\mathrm{res}(\boldsymbol{X}^{L-1})\|^3_{1,\infty} && (17)\\
&\leq \frac{4\gamma\beta}{\sqrt{d_{qk}}} \left(\frac{4\gamma\beta}{\sqrt{d_{qk}}} \|\mathrm{res}(\boldsymbol{X}^{L-2})\|^3_{1,\infty}\right)^3 && (18)\\
&= \frac{4\gamma\beta}{\sqrt{d_{qk}}} \left(\frac{4\beta}{\sqrt{d_{qk}}}\right)^3 \|\mathrm{res}(\boldsymbol{X}^{L-2})\|^{3^2}_{1,\infty} && (19)\\
&\leq \dots && (20)\\
&\leq \prod_{l=1}^{L} \left(\frac{4\gamma\beta}{\sqrt{d_{qk}}}\right)^{3^{l-1}} \|\mathrm{res}(\boldsymbol{X})\|^{3^L}_{1,\infty} = \left(\frac{4\gamma\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|\mathrm{res}(\boldsymbol{X})\|^{3^L}_{1,\infty}, && (21)
\end{aligned}$$

Residual of Token Matrix

- We want to determine how the residual

$$\text{res}(\boldsymbol{X}) = \boldsymbol{X} - \mathbf{1}\boldsymbol{x}^\top, \quad \text{where} \quad \boldsymbol{x} = \text{argmin}_{\boldsymbol{x}} \|\boldsymbol{X} - \mathbf{1}\boldsymbol{x}^\top\|$$

  changes during the forward pass

- Specifically, we want to find a bound on the residual of the output of SA

Visualization of Matrices in the Residual

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,d} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,d} \end{bmatrix} \qquad \mathbf{1}\mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \\ x_1 & x_2 & \cdots & x_d \\ \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & \cdots & x_d \end{bmatrix}$$

$$\operatorname{res}(X) = \begin{bmatrix} X_{1,1} - x_1 & X_{1,2} - x_2 & \cdots & X_{1,d} - x_d \\ X_{2,1} - x_1 & X_{2,2} - x_2 & \cdots & X_{2,d} - x_d \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} - x_1 & X_{n,2} - x_2 & \cdots & X_{n,d} - x_d \end{bmatrix}$$

A Bound on One Layer of Single-Head Self-Attention

$$\|res(SA(X)\|_{1,\infty} \leq C\|res(X)\|_{1,\infty}^3 = \frac{1}{\sqrt{C}}(\sqrt{C}\|res(X)\|_{1,\infty})^3, \text{ where } C = \frac{4\gamma\beta}{\sqrt{d_{qk}}}$$

- Standard theory says that multiplying stochastic matrices converges linearly
- Why do we see cubic convergence here?

Proving the Bound on One Layer of Single-Head Self-Attention

Represent the single-head SA layer as: $\mathrm{SA}(X) = P_h X W_V$

Use the residual representation of X: $X = 1x^\top + R$,
where $1 \in \mathbb{R}^{n\times 1}$, $x \in \mathbb{R}^{d_{\text{in}}\times 1}$, $R = \mathrm{res}(X) \in \mathbb{R}^{n\times d_{\text{in}}}$, and $W_{QK} \in \mathbb{R}^{d_{\text{in}}\times d_{\text{in}}}$

Ignoring biases, the probabilities are: $P_h = \mathrm{softmax}\left(\frac{X W_{QK} X^\top}{\sqrt{d_{qk}}}\right)$

Proving the Bound on One Layer of Single-Head Self-Attention

Expanding the product $XW_{QK}X^\top$ gives:

$$\begin{aligned} XW_{QK}X^\top &= (1x^\top + R)W_{QK}(1x^\top + R) \\ &= 1x^\top W_{QK}x1^\top + RW_{QK}x1^\top + 1x^\top W_{QK}R^\top + RW_{QK}R^\top \\ &= RW_{QK}R^\top + 1x^\top W_{QK}R^\top + (1x^\top W_{QK}x + RW_{QK}x)1^\top, \end{aligned}$$

where $v = (1x^\top W_{QK}x + RW_{QK}x)$ is an $n \times 1$ column vector, so:

$$v\mathbf{1}^\top = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} v_1 & v_1 & \cdots & v_1 \\ v_2 & v_2 & \cdots & v_2 \\ \vdots & \vdots & \ddots & \vdots \\ v_n & v_n & \cdots & v_n \end{bmatrix}$$

Proving the Bound on One Layer of Single-Head Self-Attention

Softmax is shift-invariant:

$$\text{softmax}_j(z + c\mathbf{1}^\top) = \frac{e^{z_j+c}}{\sum_{k=1}^{n} e^{z_k+c}} = \frac{e^c \cdot e^{z_j}}{e^c \sum_{k=1}^{n} e^{z_k}} = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} = \text{softmax}_j(z)$$

Therefore:

$$\begin{aligned}\text{softmax}(XW_{QK}X^\top) &= \text{softmax}(RW_{QK}R^\top + 1x^\top W_{QK}R^\top + v1^\top) \\ &= \text{softmax}(RW_{QK}R^\top + 1x^\top W_{QK}R^\top).\end{aligned}$$

Proving the Bound on One Layer of Single-Head Self-Attention

$$P_h = \text{softmax}\left(\frac{RW_{QK}R^\top}{\sqrt{d_{qk}}} + 1\left(\frac{x^\top W_{QK}R^\top}{\sqrt{d_{qk}}}\right)\right)$$
$$= \text{softmax}\left(E + 1r^\top\right),$$

where $r^\top$ is a row vector, so $1r^\top$ is a matrix with identical rows.

Observe that if $E = 0$, $\text{softmax}(1r^\top) = 1q^\top$ has identical rows.

The term preventing collapse is E, which is $\mathcal{O}(\|R\|^2)$; as $E \to 0$, P_h converges to a rank-1 matrix with a rate that is $\mathcal{O}(\|R\|^2)$.

Proving the Bound on One Layer of Single-Head Self-Attention

Now consider $P_h X = P_h(1x^\top + R_X) = 1x^\top + P_h R_X$, where we have used the fact that $P_h 1 = 1$, owing to row-stochasticity of P_h.

Let $P_h = 1q^\top + R_{P_h}$, then:

$$\begin{aligned} P_h X &= 1x^\top + P_h R_X \\ &= 1x^\top + (1q^\top + R_{P_h})R_X \\ &= 1(x^\top + q^\top R_X) + R_{P_h} R_X \\ &= 1u^\top + R_{P_h} R_X, \end{aligned}$$

where $u^\top$ is a row vector, so $1u^\top$ is a matrix with identical rows.

Proving the Bound on One Layer of Single-Head Self-Attention

Recall that $\text{res}(Y) = Y - 1y^\top$, where $y^\top = \text{argmin}_y \|Y - 1y^\top\|$. Then,
$\|\text{res}(P_h X)\| \leq \|P_h X - 1u^\top\| = \|R_{P_h} R_X\| \leq \|R_{P_h}\| \|R_X\|$.

The authors then show (via Lemma A.3) that:

$$\|R_{P_h}\| = \|P_h - 1q^\top\| \leq 2\|D1q^\top\|,$$

where D is diagonal and $\|D\| \leq \mathcal{O}(\|E\|) = \mathcal{O}(\|R_X\|^2) \implies \|R_{P_h}\| \leq \mathcal{O}(\|R_X\|^2)$.

The proof concludes with

$$\|\text{res}(P_h X)\| \leq \|R_{P_h}\| \|R_X\| = \mathcal{O}(\|R_X\|^2)\|R_X\| = \mathcal{O}(\|R_X\|^3).$$

A Quick Note on Constants

$$\|\mathrm{res}(\mathrm{SA}(X))\|_{1,\infty} \leq C \cdot \|\mathrm{res}(X)\|_{1,\infty}^3 \qquad C = \frac{4\gamma \|W_{QK}\|_1 \|W_V\|_{1,\infty}}{\sqrt{d_{qk}}} \leq \frac{4\gamma\beta}{\sqrt{d_{qk}}}$$

$\sqrt{d_{qk}}$ comes from attention scaling inside softmax: $\mathrm{softmax}\left(\frac{XW_{QK}X^\top}{\sqrt{d_{qk}}}\right)$.

β is an upper bound on the norms of the weight matrices.

γ is a measure of the variation in attention scores in $XW_{QK}X^\top$.

Recursion Over Layers and Theorem 2.2

Let $\|R_i\| = \|R_i\|_{1,\infty}$ denote the residual following i layers of self-attention. Then,

$$\|R_1\| \leq C\|R_0\|^3 \Rightarrow \|R_2\| \leq C\|R_1\|^3 = C\|C\|R_0\|^3\|^3 = C^4\|R_0\|^9 \Rightarrow \|R_L\| \leq C^{\frac{3^L-1}{2}}\|R_0\|^{3^L}$$

Recursion Over Layers and Theorem 2.2

Let $\|R_i\| = \|R_i\|_{1,\infty}$ denote the residual following i layers of self-attention. Then,

$$\|R_1\| \leq C\|R_0\|^3 \Rightarrow \|R_2\| \leq C\|R_1\|^3 = C\|C\|R_0\|^3\|^3 = C^4\|R_0\|^9 \Rightarrow \|R_L\| \leq C^{\frac{3^L-1}{2}}\|R_0\|^{3^L}$$

Theorem 2.2 (Simplified). *For any single-head SAN consisting of L layers with $\|\boldsymbol{W}_{QK}^l\|_1\|\boldsymbol{W}_V^l\|_{1,\infty} \leq \beta$ and for a term γ that depends on the attention entries, we have that*

$$\|res(SAN(\boldsymbol{X}))\|_{1,\infty} \leq \left(\frac{4\gamma\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|_{1,\infty}^{3^L},$$

which amounts to a doubly exponential convergence to a rank-1 matrix.

Why Networks Meet the Convergence Condition in Practice

- Softmax of attention scores yields convex combinations of tokens

- Healthy training dynamics avoid extreme values of the constants γ, β, and d_{qk}

Multi-Head Self-Attention Also Converges

Theorem 2.3 (Simplified). *Consider a depth-L and width-H self-attention network without skip connections. Suppose that $\|\boldsymbol{W}_{QK,h}^{l}\|_1 \|\boldsymbol{W}_{h}^{l}\|_{1,\infty} \leq \beta$ for all heads $h \in [H]$ and layers $l \in [L]$, and let γ be a term that depends on the attention entries. We have*

$$\|res(SAN(\boldsymbol{X}))\|_{1,\infty} \leq \left(\frac{4\gamma\beta H}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|_{1,\infty}^{3^L},$$

which amounts to a doubly exponential rate of convergence.

Proof for Convergence of Multi-Head Self-Attention

Multi-head output is the sum of H heads:

$$\mathrm{SA}(X) = \sum_{h=1}^{H} P_h X W_h.$$

Applying the triangle inequality (justified by Lemma A.2):

$$\|\mathrm{res}(\mathrm{SA}(\mathrm{X}))\| \leq \sum_{h=1}^{H} \|\mathrm{res}(\mathrm{SA}_h(X))\| \leq H \cdot \max_{h \in [H]} \|\mathrm{res}(\mathrm{SA}_h(X))\|.$$

Plugging in the bound we found for single heads, we have:

$$\|\mathrm{res}(\mathrm{SA}(X))\|_{1,\infty} \leq \left(\frac{4\gamma\beta H}{\sqrt{d_{qk}}} \right) \|\mathrm{res}(X)\|_{1,\infty}^{3}.$$

Width Cannot Outrun Depth

Number of paths is H^L, or $(H+1)^L$ with skip connections. Theorem 2.2 showed that each path through the network converges doubly exponentially, i.e., $\mathcal{O}(c^{3^L})$.

Width (number of heads) can slow down collapse, but will not prevent it.

What About Other Attention Architectures (XFormers)?

- Low-rank and kernel-based attention approximations accelerate convergence
- Randomized attention can delay convergence
- All variations still converge eventually
- So, why do Transformers work?

Path Decomposition With Skip Connections

- Add a additional "virtual head" (ID 0) with the properties $\boldsymbol{P}_0 = \boldsymbol{I}$ and $\boldsymbol{W}_0 = \boldsymbol{I}$.

- Write the path decomposition formula as

$$\begin{aligned}\boldsymbol{X}^L &= \sum_{h \in [H] \cup \{0\}} \boldsymbol{P}_h^L \boldsymbol{X}^{L-1} \boldsymbol{W}_h^L \\ &= \dots \\ &= \sum_{h_1, \dots, h_L \in ([H] \cup \{0\})^L} (\boldsymbol{P}_{h_L}^L \cdots \boldsymbol{P}_{h_1}^1)\, \boldsymbol{X}\, (\boldsymbol{W}_{h_1}^1 \cdots \boldsymbol{W}_{h_L}^L)\end{aligned}$$

Paths of a Specific Length

- Let the length of a path be the number of non-virtual heads the path traverses

- The number of paths of length l is given by

$$|\mathcal{P}_l| = \binom{L}{l} H^l$$

Paths of a Specific Length

- Let the length of a path be the number of non-virtual heads the path traverses

- The number of paths of length l is given by

$$|\mathcal{P}_l| = \binom{L}{l} H^l$$

Number of ways to choose a subset of l layers

Number of head choice combinations for each path of length l

Skip Connection Hypothesis

- In theorems 2.2 and 2.3, there were only paths of length L
- Hypothesis: the presence of short paths ($l < L$) prevents rank collapse
- Motivated by the empirical observation that SANs with skip connections do not collapse

Existence of Parameterizations Preventing Rank Collapse

Claim 3.1. *Consider a depth-L and width-H self-attention network with skip connections. There exist infinitely many parameterizations for which $\|res(\boldsymbol{X}^L)\| \geq \|res(\boldsymbol{X})\|$. The preceeding holds even for $L \to \infty$ and β arbitrarily small.*

Proof: Set $\boldsymbol{W}_V^l = 0$ for every $l \in [L]$, in which case $\|\text{res}(\boldsymbol{X}^L)\| = \|\text{res}(\boldsymbol{X})\|$ for any $\boldsymbol{W}_Q^l, \boldsymbol{W}_K^l$.

The conclusion is that rank collapse is no longer mandatory

What Happens in Practice?

- In practice the value weights are not going to be zero
- The residual will naturally decrease because attention mixes tokens
- Skip connections have the capacity to prevent complete collapse
- Deep multi-head SANs behave like ensembles of shallow single-head SANs

3.2: Multi-layer Perceptrons help counteract convergence

$$\boldsymbol{X}^{l+1} = f_l\left(\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h\right)$$

Lipschitz continuity: $||f(x) - f(y)|| \leq \lambda||x - y||$ (for some $\lambda \in \mathbb{R}$).

$$\left\|\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h - \mathbf{1}(\boldsymbol{r}')^\top\right\|_p \leq 2H \max_{h\in[H]} \|\boldsymbol{D}_h \mathbf{1}\,\mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h\|_p,$$

for $p \in [1, \infty]$.

$$\begin{aligned}
\|\mathrm{res}(\mathrm{SAN}(\boldsymbol{X}))\|_p &= \left\|f\left(\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h\right) - \mathbf{1}r''^\top\right\|_p \\
&= \left\|f\left(\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h\right) - f(\mathbf{1}(r')^\top)\right\|_p && \triangleright\ f \text{ preserves constancy-across-rows.} \\
&\leq \lambda_{l,p}\left\|\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h - \mathbf{1}(r')^\top\right\|_p && \triangleright\ \text{By definition of Lipschitz constant.} \\
&\leq 2\lambda_{l,p} H \max_{h\in[H]} \|\boldsymbol{D}_h \mathbf{1}\,\mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h\|_p && \triangleright\ \text{By Eq 27.}
\end{aligned}$$

Intuitively: ||F(A + R) - F(A)|| <= c||A + R - A|| = c||R||, where F(A) is also a rank 1 matrix.

If you ignore the MLP, you're converging to a rank 1 matrix doubly exponentially. If you apply an MLP to a layer's output, which can be represented as a rank 1 matrix + residual, your residual from a rank 1 matrix will be a linear factor of your residual without the MLP. So you have (at most) a linear scaling factor for your distance from a rank 1 matrix competing with a doubly exponential contraction.

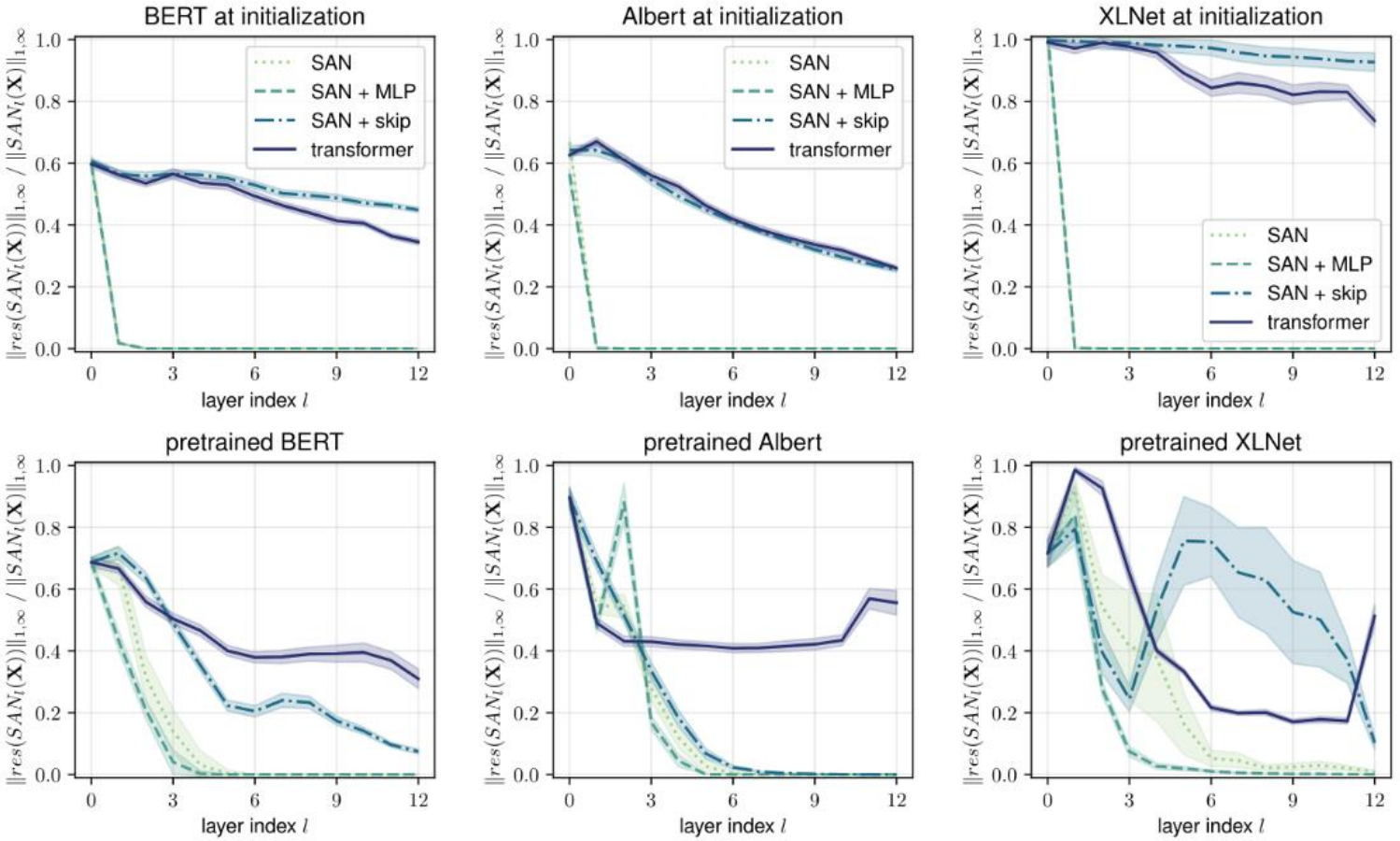
3.2: Multi-layer Perceptrons help counteract convergence

Corollary 3.2 (Simplified). *Consider a depth-L and width-H SAN with MLP. Suppose that $\|\boldsymbol{W}^l_{QK,h}\|_1 \|\boldsymbol{W}^l_h\|_{1,\infty} \leq \beta$ for all $h \in [H]$ and $l \in [L]$, let γ be a term that depends on the attention entries, and fix $\lambda_{l,1,\infty} \leq \lambda$. We have that*

$$\|res(\boldsymbol{X}^L)\|_{1,\infty} \leq \left(\frac{4\,\gamma\,\beta\,H\,\lambda}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|^{3^L}_{1,\infty}, \tag{3}$$

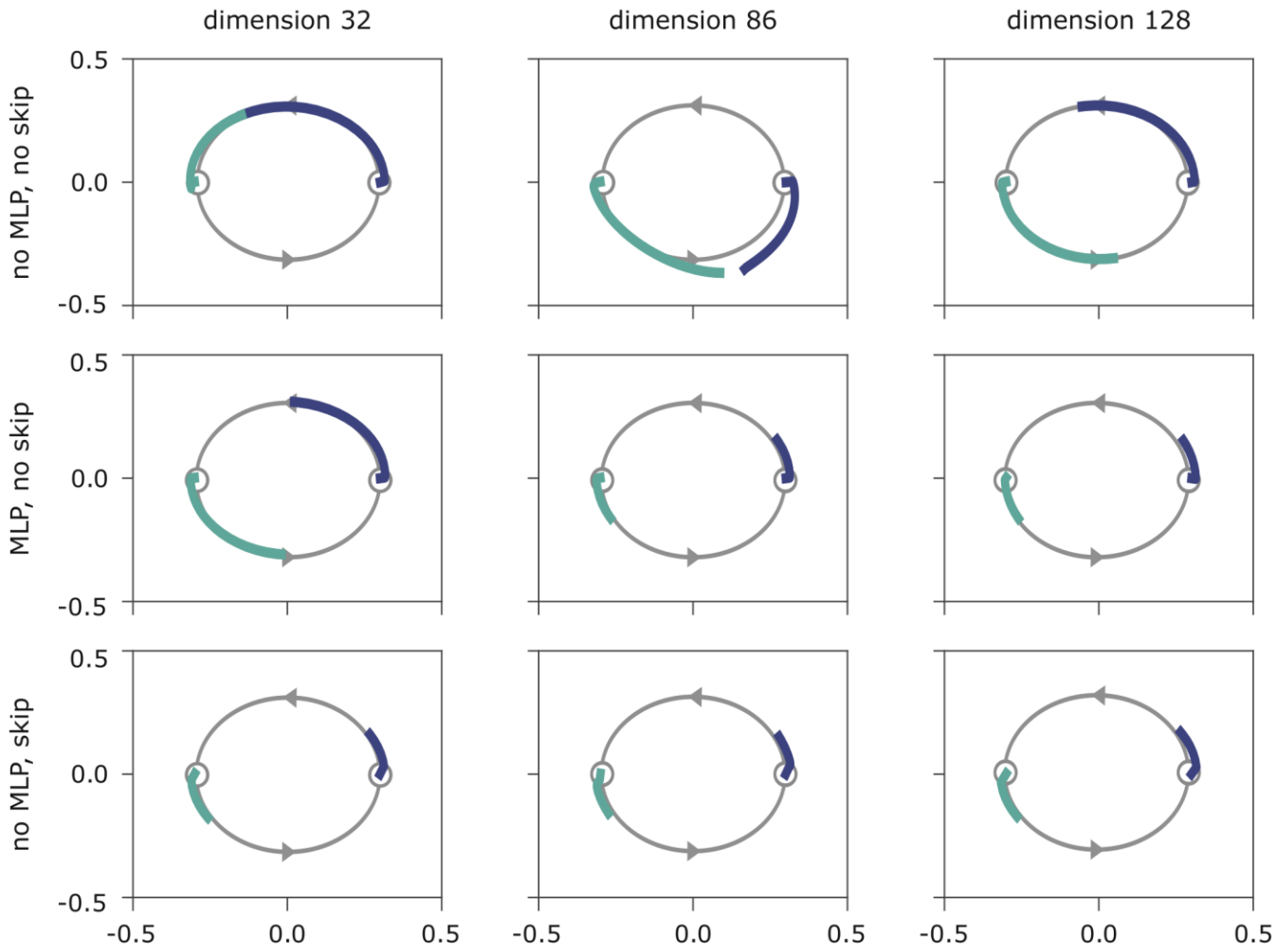
which amounts to a doubly exponential rate of convergence.

Empirical Observations of Rank Collapse (Token Uniformity)



(a) Bert (b) Albert (c) XLNet

Rank Collapse in a 2d Prediction Task



Task: track 2 opposite points on a circle that move with a constant (equal) orbit.

Learning: used teacher-forcing (network gets ground truth at each step), trained until network "could perfectly memorize the next step ... with near-zero loss".

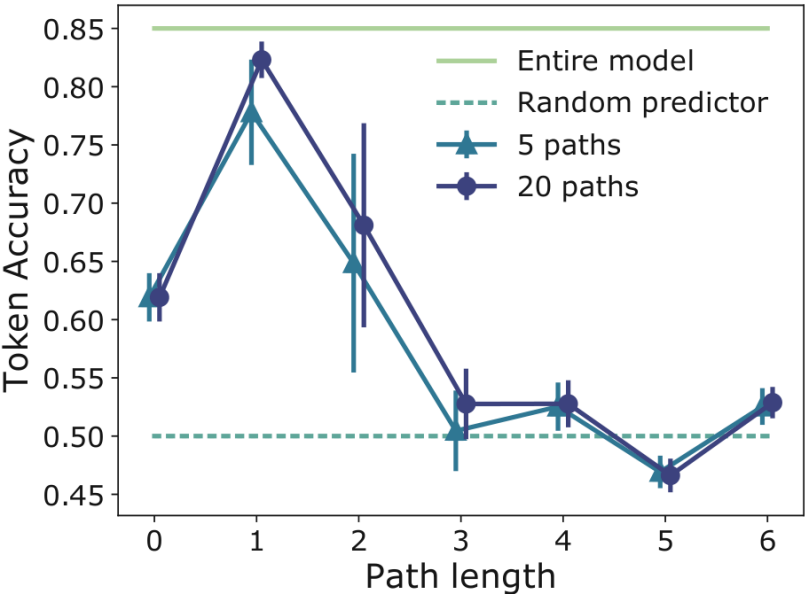
Test: ran the network recurrently (output becomes input, simulating multiple layers) *without* teacher forcing (so, errors can compound).

Findings: low dimension, no MLP, no skip exhibits rank collapse. Adding MLP and skip helps prevent this. Increasing dimension helps prevent this (remember that \beta is a term in the upper bound, and \beta increases as the dimension increases since $\beta \geq \|\boldsymbol{W}_{QK}^{l}\|_1 \|\boldsymbol{W}_{V}^{l}\|_{1,\infty}$ and the terms on the right use sums of absolute values).

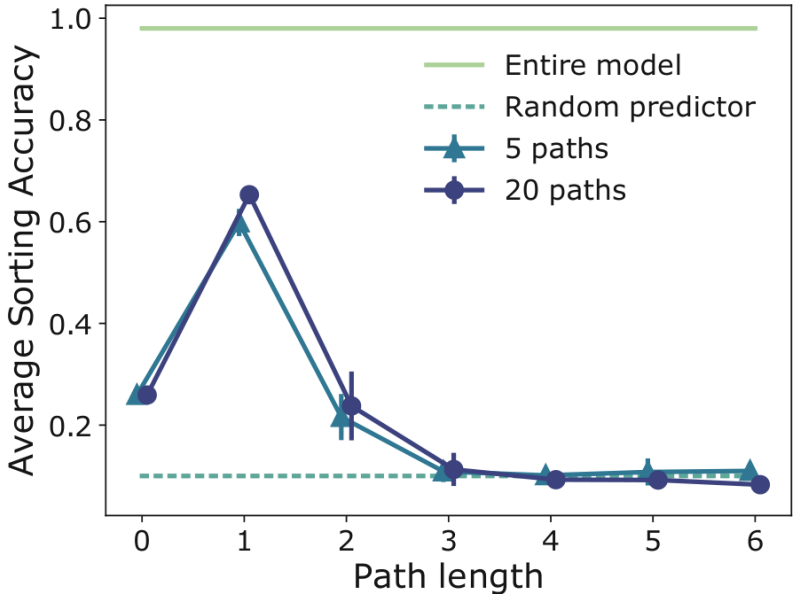
$$\|res(\boldsymbol{X}^{L})\|_{1,\infty} \leq \left(\frac{4\,H\,\gamma\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^{L}-1}{2}} \|res(\boldsymbol{X})\|_{1,\infty}^{3^{L}},$$

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

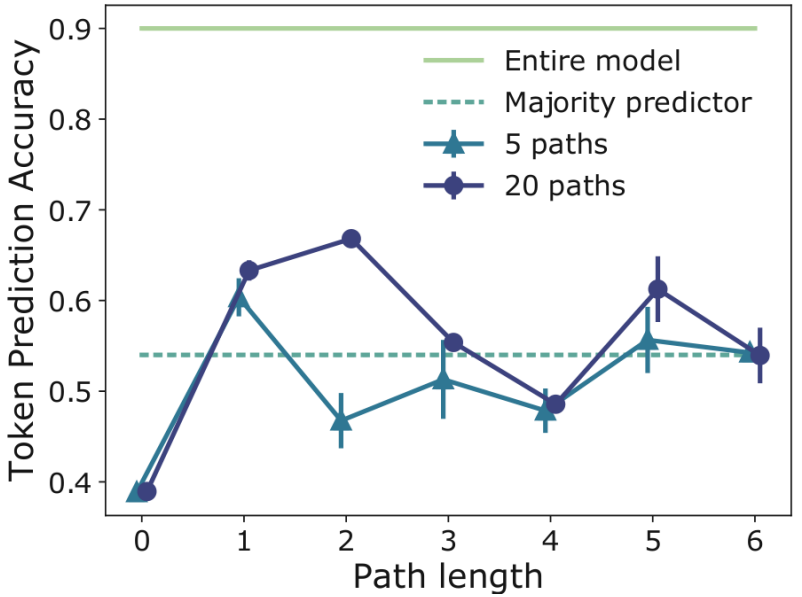
Empirical Observations of Short Path Effectiveness



(a) Memorization



(b) Sorting

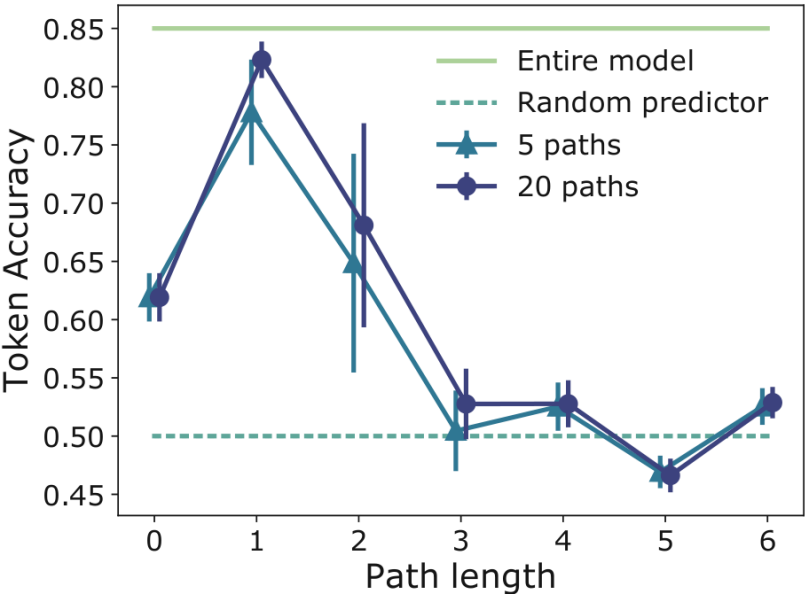


(c) Convex hull

Model trained to solve one of three tasks. At inference time, they isolate the weights and output of an individual path for any given sequence of heads. They evaluate task performance based solely on the normalized sum of this subset of paths.

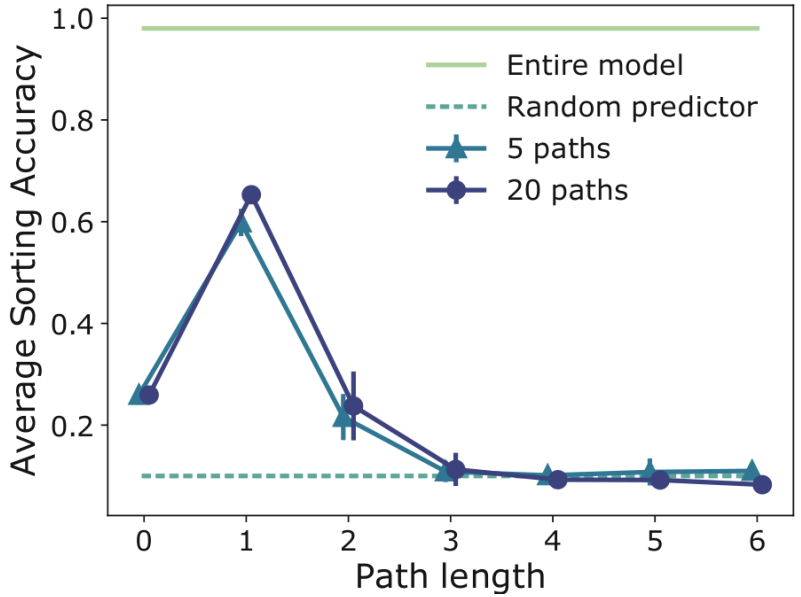
$$SAN(\boldsymbol{X}) = \sum_{path \in [H]^L} \boldsymbol{P}_{path} \boldsymbol{X} \boldsymbol{W}_{path} + \boldsymbol{1}\boldsymbol{b}^\top,$$

Empirical Observations of Short Path Effectiveness



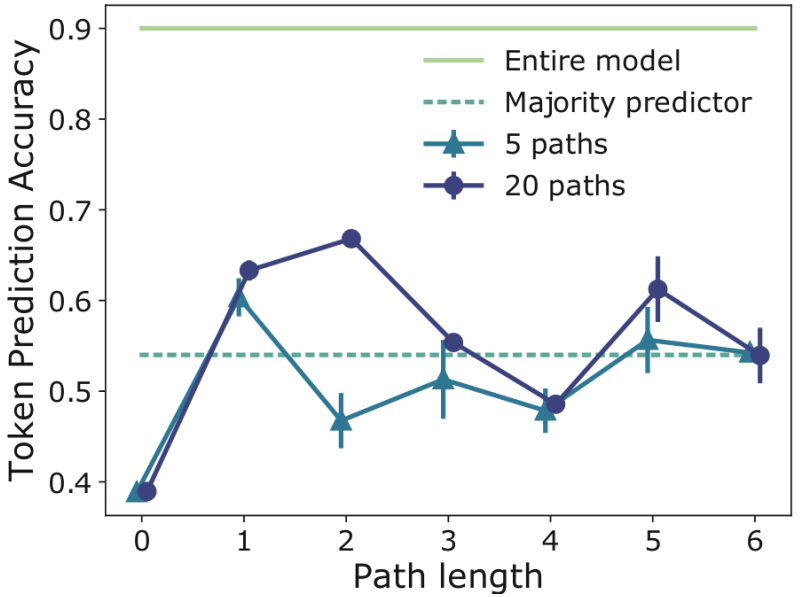
(a) Memorization

Memorize a pre-determined mapping from natural language sentences to random binary sequences of the same token length. "The models tested are trained to minimize the cross entropy loss between predicted and the ground truth labels."



(b) Sorting

Given in input sequence of length 8 from an alphabet of size 10, sort the sequence in alphabetical order. "... we experimented with a variety of settings (adjusting the model depth, number of heads, and ... the alphabet size and sequence length) and observed consistent behavior."



(c) Convex hull

"Given a sequence of N points uniformly distributed in [0, 1] × [0, 1] and shifted by a random bivariate standard normal ... for each point in the set, the model predicts whether it's part of the convex hull. The training set consists of 10,000 sequences of points in [0, 1] × [0, 1], each of length 10."