Supplementary Material: Attention is not *all* you need, pure attention loses rank doubly exponentially with depth

The supplementary material is structured as follows: §A contains the proofs of the theorems presented in the main text. §B contains additional experimental details and results, including the path length distribution of commonly used attention-based architectures and the circle experiment.

A SAN convergence results

We build our argument step by step, by first considering a single-head self-attention layer in §A.1 and then moving to deeper networks with single and multiple heads in §A.3 and §A.4. The results are extended to take into account skip connections and MLPs in §A.5 and §A.6

A.1 Single-layer and single-head

We consider a single-head self-attention layer:

$$\boldsymbol{X}' = \mathrm{SA}(\boldsymbol{X}) = \boldsymbol{P}\boldsymbol{X}\boldsymbol{W}_V$$

We focus in particular on how the residual changes. As discussed previously, the value bias can be safely ignored since it does not contribute to the residual.

The following is proved:

Lemma A.1. *The residual abides to:* $\|res(SA(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4\,\|\boldsymbol{W}_{QK}\|_1\,\|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}}\,\|res(\boldsymbol{X})\|_{1,\infty}^3$.

The unscaled attention scores are computed as follows,

$$\boldsymbol{A} = (\boldsymbol{X}\boldsymbol{W}_Q + \boldsymbol{1}\boldsymbol{b}_Q^\top)(\boldsymbol{X}\boldsymbol{W}_K + \boldsymbol{1}\boldsymbol{b}_K^\top)^\top \tag{1}$$

and following Cordonnier et al. (2020), we can use the softmax shift invariance property to prune the terms constant over the columns and obtain,

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{W}_{QK}\boldsymbol{X}^\top + \boldsymbol{1}\boldsymbol{b}_{QK}^\top\boldsymbol{X}^\top \tag{2}$$

with $\boldsymbol{W}_{QK} = \boldsymbol{W}_Q\boldsymbol{W}_K^\top$ and $\boldsymbol{b}_{QK} = \boldsymbol{W}_K\boldsymbol{b}_Q$.

We use the shorthand notation $\boldsymbol{R} := \mathrm{res}(\boldsymbol{X})$ and $\boldsymbol{R}' := \mathrm{res}(\boldsymbol{X}')$.

The attention matrix can be written as

$$\begin{aligned} A &= (\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R})\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}(\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R})^\top + \mathbf{1}\frac{\boldsymbol{b}_{QK}^\top}{\sqrt{d_{qk}}}(\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R})^\top \\ &= \left(\frac{\boldsymbol{x}^\top \boldsymbol{W}_{QK}\boldsymbol{x}}{\sqrt{d_{qk}}}\mathbf{1} + \boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{x} + \mathbf{1}\frac{\boldsymbol{b}_{QK}^\top}{\sqrt{d_{qk}}}\boldsymbol{x}\right)\mathbf{1}^\top + \mathbf{1}\boldsymbol{x}^\top\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top + \boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top + \mathbf{1}\frac{\boldsymbol{b}_{QK}^\top}{\sqrt{d_{qk}}}\boldsymbol{R}^\top \end{aligned}$$

Using the shift-invariance property of the softmax operator, the first term above can be safely ignored since it is constant across columns. We therefore have that

$$\boldsymbol{P} = \mathrm{softmax}\left(\boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{r}^\top\right),$$

where we have set $\boldsymbol{r} := \boldsymbol{R}\frac{\boldsymbol{W}_{QK}^\top}{\sqrt{d_{qk}}}\boldsymbol{x} + \boldsymbol{R}\frac{\boldsymbol{b}_{QK}}{\sqrt{d_{qk}}}$.

Setting $\boldsymbol{E} = \boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top$ and $\tilde{\boldsymbol{A}} = \mathbf{1}\boldsymbol{r}^\top$, the input reweighted by the attention probibilities $\boldsymbol{PX}$ is given by

$$\begin{aligned} \boldsymbol{PX} &= \boldsymbol{P}(\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R}) && (3) \\ &= \mathbf{1}\boldsymbol{x}^\top + \boldsymbol{PR} && (4) \\ &= \mathbf{1}\boldsymbol{x}^\top + \mathrm{softmax}(\mathbf{1}\boldsymbol{r}^\top + \boldsymbol{E})\boldsymbol{R} && (5) \\ &\leq \mathbf{1}\boldsymbol{x}^\top + (\boldsymbol{I} + 2\boldsymbol{D})\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{R} && (6) \\ &= \mathbf{1}(\boldsymbol{x}^\top + \mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{R}) + 2\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{R} && (7) \end{aligned}$$

where the inequality above is entry-wise and follows from Lemma A.3. Similarly $\boldsymbol{PX} \geq \mathbf{1}(\boldsymbol{x}^\top + \mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{R}) - \boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{R}$, where we again invoke Lemma A.3.

Therefore, the (entry-wise) distance of the output of the self-attention layer $SA(\boldsymbol{X}) = \boldsymbol{PXW}_V$ from being constant across tokens is at most:

$$|[SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}')^\top]_{ij}| \leq 2\,|[\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{RW}_V]_{ij}|, \tag{8}$$

where $\boldsymbol{r}' = (\boldsymbol{x} + \boldsymbol{R}^\top \mathrm{softmax}(\boldsymbol{r}))\boldsymbol{W}_V$.

Now we bound the right hand side of the above inequality. For the ℓ_1 norm we obtain:

$$\|\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top \boldsymbol{RW}_V\|_1 \leq \|\boldsymbol{D}\mathbf{1}\|_\infty\,\|\boldsymbol{R}\|_1\|\boldsymbol{W}_V\|_1, \tag{9}$$

where the last step is due to Hölder's inequality, the fact that $\|\mathrm{softmax}(\boldsymbol{r})\|_1 = 1$, and $\|\boldsymbol{AB}\|_1 \leq \|\boldsymbol{A}\|_1\|\boldsymbol{B}\|_1$. Moreover, by the definition of $\boldsymbol{D}$ as in Lemma A.3, $\|\boldsymbol{D}\mathbf{1}\|_\infty$ can be bounded as:

$$\begin{aligned} \|\boldsymbol{D}\mathbf{1}\|_\infty = \max_{i,j,j'} |\boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})| \leq 2\max_{ij} |E_{ij}| &\leq 2\,\|\boldsymbol{E}\|_1 \\ &= 2\,\|\boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top\|_1 \\ &\leq \frac{2}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1\|\boldsymbol{W}_{QK}\|_1\|\boldsymbol{R}^\top\|_1 \\ &= \frac{2}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1\|\boldsymbol{W}_{QK}\|_1\|\boldsymbol{R}\|_\infty, \end{aligned}$$

implying

$$\|SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}')^\top\|_1 \leq \frac{4}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1^2\|\boldsymbol{R}\|_\infty\,\|\boldsymbol{W}_{QK}\|_1\,\|\boldsymbol{W}_V\|_1.$$

On the other hand, an analogous argument gives the following bound on the ℓ_∞ norm of the residual:

$$\begin{aligned}\|SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}')^\top\|_\infty &\leq 2\|\boldsymbol{D}\,\mathbf{1}\|_\infty \|\text{softmax}(\boldsymbol{r})^\top \boldsymbol{R}\boldsymbol{W}_V\|_\infty \\ &\leq 2\|\boldsymbol{D}\,\mathbf{1}\|_\infty \|\boldsymbol{R}\|_\infty \|\boldsymbol{W}_V\|_\infty \\ &\leq \frac{4}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1 \|\boldsymbol{R}\|_\infty^2 \|\boldsymbol{W}_{QK}\|_1\,\|\boldsymbol{W}_V\|_\infty.\end{aligned}$$

Combining the two norms we obtain:

$$\|\boldsymbol{R}'\|_{1,\infty} = \sqrt{\|\boldsymbol{R}'\|_1 \|\boldsymbol{R}'\|_\infty} \leq \frac{4\,\|\boldsymbol{W}_{QK}\|_1 \|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}}\,(\sqrt{\|\boldsymbol{R}\|_1 \|\boldsymbol{R}\|_\infty})^3 = \frac{4\,\|\boldsymbol{W}_{QK}\|_1 \|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_{1,\infty}^3$$

which is equivalent to the main claim.

A.2 Multiple-heads and single-layer

Lemma A.2. *The residual of the output of a H-heads attention layer abides to:*

$$\|res(SA(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4H\beta}{\sqrt{d_{qk}}}\,\|res(\boldsymbol{X})\|_{1,\infty}^3\,, \tag{10}$$

where $\|\boldsymbol{W}_{QK,h}\|_1 \|\boldsymbol{W}_h\|_{1,\infty} \leq \beta$ for all heads $h \in [H]$.

Proof. The output of a multi-head attention layer is

$$SA(\boldsymbol{X}) = \sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}\boldsymbol{W}_h\ = \sum_{h\in[H]} SA_h(\boldsymbol{X}), \tag{11}$$

where $\boldsymbol{W}_h := \boldsymbol{W}_{V,h}\boldsymbol{W}_{O,h}$ as in the main text and $\boldsymbol{P}_h$ is computed using the heads parameters $\boldsymbol{W}_{QK,h}$ and $\boldsymbol{b}_{QK,h}$. The proof proceeds similarly to Section A.2 until eq. 8,

$$|[SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}'')^\top]_{ij}| \leq 2\left|\left[\sum_h \boldsymbol{D}_h\,\mathbf{1}\,\text{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h\right]_{ij}\right|, \tag{12}$$

where $\boldsymbol{r}'' = \sum_h(\boldsymbol{x} + \boldsymbol{R}^\top \text{softmax}(\boldsymbol{r}_h))\boldsymbol{W}_h$.

The elementwise inequality implies inequalities for ℓ_1 and ℓ_∞ norms and applying the triangle inequality on the sum, we obtain

$$\|SA^H(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}'')^\top\|_1 \leq 2\sum_{h\in[H]} \|\boldsymbol{D}_h\,\mathbf{1}\,\text{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h\|_1 \leq 2H \max_{h\in[H]} \|\boldsymbol{D}_h\,\mathbf{1}\,\text{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h\|_1$$

and a similar expression for the ℓ_∞ norm. The rest of the proof proceeds similarly as the single head proof. □

A.3 Single-head and multiple-layers

We next consider how the residual changes after L layers of the form: $\boldsymbol{X}^l = \text{SA}_1^l(\boldsymbol{X}^{l-1})$.

Corollary 2.2. *For any single-head SAN consisting of L layers with $\|\boldsymbol{W}_{QK,1}^l\|_1 \leq \beta$ for every $l \in [L]$, the residual is bounded by*

$$\|res(SAN(\boldsymbol{X}))\|_{1,\infty} \leq \left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}}\ \|res(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{13}$$

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

Proof. Unfolding the recursion backwards from the last layer to the first and applying Lemma A.1 we obtain:

$$\begin{aligned}
\|\text{res}(\boldsymbol{X}^L)\|_{1,\infty} &\leq \frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X}^{L-1})\|_{1,\infty}^3 && (14)\\
&\leq \frac{4\,\beta}{\sqrt{d_{qk}}}\left(\frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X}^{L-2})\|_{1,\infty}^3\right)^3 && (15)\\
&= \frac{4\,\beta}{\sqrt{d_{qk}}}\left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^3\,\|\text{res}(\boldsymbol{X}^{L-2})\|_{1,\infty}^{3^2} && (16)\\
&\leq \dots && (17)\\
&\leq \prod_{l=1}^{L}\left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^{3^{l-1}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^{3^L} = \left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^{3^L}, && (18)
\end{aligned}$$

matching the theorem statement. □

A.4 Multiple-head and multiple-layers

Corollary 2.3 (mutli-head multi-layer)**.** *Consider a depth-L SAN with H heads per layer. Fix* $\|\boldsymbol{W}_{QK,h}^l\|_1\|\boldsymbol{W}_h^l\|_{1,\infty} \leq \beta$ *for all* $h \in [H]$ *and* $l \in [L]$*. The output residual is bounded by*

$$\|\mathit{res}(\boldsymbol{X}^L)\|_{1,\infty} \leq \left(\frac{4\,H\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}}\,\|\mathit{res}(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{19}$$

which indicates that the output convergences to a rank-1 matrix doubly exponentialy.

Proof. The proof procceeds recursively as for Theorem 2.2 in the single head case but using the bound on single-layer multi-heads residuals from Lemma A.2. □

A.5 SAN with skip connections

As noted in the main text, a lower bound on the residual better aligns with practice, where SANs with skip connections do not suffer rank collapse. For consistency with the other analyses and as one way to illustrate residual growth, we provide a (vacuously large) upper bound on the residual for SANs with skip connections.

Corollary 3.1 (SAN with skip connections)**.** *Consider a depth-L SAN with H heads per layer and skip connections. Fix* $\|\boldsymbol{W}_{QK,h}^l\|_1\|\boldsymbol{W}_h^l\|_{1,\infty} \leq \beta$ *for all heads* $h \in [H]$ *and layers* $l \in [L]$*. The output residual is bounded by*

$$\|\mathit{res}(\boldsymbol{X}^L)\|_{1,\infty} \leq \max_{0\leq l\leq L}\left(\frac{8\,\beta\,H}{\sqrt{d_{qk}}}\right)^{\frac{3^l-1}{2}}\,(2H)^{3^l(L-l)}\|\mathit{res}(\boldsymbol{X})\|_{1,\infty}^{3^l},$$

which does not indicate convergence.

Proof. For a SAN with skip connections, the residual bound for a single-head single-layer SAN from lemma A.1 now becomes:

$$\|\text{res}(\text{SAN}(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4\,\|\boldsymbol{W}_{QK,h}\|_1\|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^3 + \|\text{res}(\boldsymbol{X})\|_{1,\infty} \tag{20}$$

To obtain a multi-layer bound, we unfold the recursion backwards.

Let us consider a single head model first and fix $\|\boldsymbol{W}^l_{QK,h}\|_1 \|\boldsymbol{W}^l_h\|_{1,\infty} \leq \beta$ for all $l \in [L]$. We have that:

$$\begin{aligned}\|\mathrm{res}(\boldsymbol{X}^L)\|_{1,\infty} &\leq \frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\mathrm{res}(\boldsymbol{X}^{L-1})\|^3_{1,\infty} + \|\mathrm{res}(\boldsymbol{X}^{L-1})\|_{1,\infty} \\ &\leq 2\max(\frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\mathrm{res}(\boldsymbol{X}^{L-1})\|^3_{1,\infty},\ \|\mathrm{res}(\boldsymbol{X}^{L-1})\|_{1,\infty}) \end{aligned} \tag{21}$$

Now we unroll this bound across layers to write it in terms of $\mathrm{res}(\boldsymbol{X})$. At the k^{th} step of unrolling, the max is one of the two terms in Eq 21: either $\frac{4\beta}{\sqrt{d_{qk}}}\,\|\mathrm{res}(\boldsymbol{X}^{L-k})\|^3_{1,\infty}$ or $\|\mathrm{res}(\boldsymbol{X}^{L-k})\|_{1,\infty}$, i.e. we make a binary choice. Thus unrolling through all L layers corresponds to a path from the root to the maximum leaf in a depth-L complete binary tree. Each leaf has the form $\left(\frac{8\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^l-1}{2}} 2^{3^l(L-l)}\|\mathrm{res}(\boldsymbol{X})\|^{3^l}_{1,\infty}$, where l indicates the number of times the term $\frac{4\beta}{\sqrt{d_{qk}}}\,\|\mathrm{res}(\boldsymbol{X}^{L-k})\|^3_{1,\infty}$ is chosen as the max. Note the ordering of these choices does not matter, only the number of times a term is chosen. Consequently, the residual bound is the maximum amongst such leaf terms:

$$\|\mathrm{res}(\boldsymbol{X}^L)\|_{1,\infty} \leq \max_{0\leq l\leq L} \left(\frac{8\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^l-1}{2}} 2^{3^l(L-l)}\|\mathrm{res}(\boldsymbol{X})\|^{3^l}_{1,\infty}.$$

We now apply this bound to H heads, we use Lemma A.2, which for a single layer gives:

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

Therefore, accounting for the factor of H in above, we obtain a residual bound for a depth-L width-H SAN with skip connections:

$$\|\mathrm{res}(\boldsymbol{X}^L)\|_{1,\infty} \leq \max_{0\leq l\leq L} \left(\frac{8\,\beta\,H}{\sqrt{d_{qk}}}\right)^{\frac{3^l-1}{2}} (2H)^{3^l(L-l)}\|\mathrm{res}(\boldsymbol{X})\|^{3^l}_{1,\infty},$$

which concludes the proof. □

A.6 SAN with MLP

We now study how using an MLP affects the residual. Recall we focus on SANs with layers written as

$$\boldsymbol{X}^{l+1} = f_l\left(\sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h\right). \tag{22}$$

Note that, to keep the notation compact, we use f_l to encompass both the MLP as well as the output bias.

In our subsequent analysis, we use $\lambda_{l,1,\infty}$ to denote the Lipschitz constant of f_l with respect to $\ell_{1,\infty}$ norm.

The proof proceeds the same way as in §A.1. For clarity, we point out the differences with proof in §A.1 without repeating details that remain the same.

Theorem 3.2 (SAN with MLP)**.** *Consider a depth-L and width-H SAN with MLP. Moreover, let $\|\boldsymbol{W}^l_{QK,h}\|_1 \|\boldsymbol{W}^l_h\|_{1,\infty} \leq \beta$ for all $h \in [H]$ and $l \in [L]$ and fix $\lambda_{l,1,\infty} \leq \lambda$. We then have that*

$$\|res(\boldsymbol{X}^L)\|_{1,\infty} \leq \left(\frac{4\,\beta\,H\,\lambda}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|^{3^L}_{1,\infty}, \tag{23}$$

which amounts to a doubly exponential rate of convergence. with respect to the $\ell_{1,\infty}$ norm.

Proof. With an MLP as formulated in Eq 22, we have $\boldsymbol{W}_h := \boldsymbol{W}_V \boldsymbol{W}_O$ in place of just the value weight $\boldsymbol{W}_V$, as defined in the main text. As before, let $\boldsymbol{R}$ denote $\operatorname{res}(\boldsymbol{X})$.

The proof proceeds the same way as in Lemma A.1, until Eq 8, where we handle the multi-head case the same way as in Eq A.2 to obtain the entrywise inequality:

$$\left| \left[\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h - \boldsymbol{1}(\boldsymbol{r}')^\top \right]_{ij} \right| \leq 2 \left| \left[\sum_h \boldsymbol{D}_h \, \boldsymbol{1} \operatorname{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \right]_{ij} \right|, \tag{24}$$

As in the proof of A.2, this elementwise inequality implies the corresponding inequality in matrix norms ℓ_1 and ℓ_∞, to each of which we apply the triangle inequality to yield:

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

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

We now use the fact that $f(\boldsymbol{1} r'^\top)$ also takes the form $\boldsymbol{1} r''^\top$ for some vector r''. Indeed, f encompasses weight matrix multiplications, bias addition, and entrywise nonlinearities, all of which preserve the fact that $f(\boldsymbol{1} r'^\top)$ is constant across rows. Therefore,

$$\begin{aligned}
\|\operatorname{res}(\mathrm{SAN}(\boldsymbol{X}))\|_p &= \left\| f\left(\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h \right) - \boldsymbol{1} r''^\top \right\|_p \\
&= \left\| f\left(\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h \right) - f(\boldsymbol{1}(r')^\top) \right\|_p && \triangleright\ f \text{ preserves constancy-across-rows.} \\
&\leq \lambda_{l,p} \left\| \sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h - \boldsymbol{1}(r')^\top \right\|_p && \triangleright \text{By definition of Lipschitz constant.} \\
&\leq 2\lambda_{l,p} \, H \max_{h \in [H]} \| \boldsymbol{D}_h \, \boldsymbol{1} \operatorname{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_p && \triangleright \text{ By Eq 24.}
\end{aligned}$$

Subsequently, just like for the single-head single-layer proof, we bound $\| \boldsymbol{D}_h \, \boldsymbol{1} \operatorname{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_p$ in the above by

$$\| \boldsymbol{D}_h \, \boldsymbol{1} \operatorname{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_1 \leq \| \boldsymbol{D}_h \boldsymbol{1} \|_\infty \, \|\boldsymbol{R}\|_1 \|\boldsymbol{W}_h\|_1, \tag{25}$$

$$\| \boldsymbol{D}_h \, \boldsymbol{1} \operatorname{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_\infty \leq \| \boldsymbol{D}_h \boldsymbol{1} \|_\infty \, \|\boldsymbol{R}\|_\infty \|\boldsymbol{W}_h\|_\infty. \tag{26}$$

Since $\|\boldsymbol{D}_h \boldsymbol{1}\|_\infty$ can be bounded above by $\frac{2}{\sqrt{d_{qk}}} \|\boldsymbol{R}\|_1 \|\boldsymbol{W}_{QK,h}\|_1 \|\boldsymbol{R}\|_\infty$, applying this to both Eq 25 and Eq 26, and combining the two as in Lemma A.1, yields the bound:

$$\|\operatorname{res}(\mathrm{SAN}(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4 \, H \, \lambda_{l,1,\infty} \|\boldsymbol{W}_{QK,h}\|_1 \|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}} \|\operatorname{res}(\boldsymbol{X})\|_{1,\infty}^3$$

Finally, we recursively unroll the bound across layers to obtain a residual bound in terms of $\operatorname{res}(\boldsymbol{X})$:

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

which concludes the proof. □

A.7 A technical lemma

Lemma A.3. *Suppose that $\boldsymbol{P}$ is the row-stochastic matrix associated with $\boldsymbol{A}$ and let $\tilde{\boldsymbol{P}}$ be the one associated with $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{E}$ for some matrix $\boldsymbol{E}$. Then*

$$(\boldsymbol{I} - \boldsymbol{D})\,\tilde{\boldsymbol{P}} \leq \boldsymbol{P} \leq (\boldsymbol{I} + 2\boldsymbol{D})\,\tilde{\boldsymbol{P}}$$

with the diagonal matrix $\boldsymbol{D}$ having $D_{ii} = \max_j |\boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})|$, with the inequality taken entry-wise.

Proof. Let us start by the definition of the row-stochastic matrix:

$$P_{ij} = [\mathrm{softmax}(\boldsymbol{A})]_{ij} = [\mathrm{softmax}(\tilde{\boldsymbol{A}} + \boldsymbol{E})]_{ij} = \frac{\exp\left(\tilde{A}_{ij} + E_{ij}\right)}{\sum_{t=1}^{n} \exp\left(\tilde{A}_{it} + E_{it}\right)} = \frac{\exp\left(\tilde{A}_{ij}\right)\,\exp\left(E_{ij}\right)}{\sum_{t=1}^{n} \exp\left(\tilde{A}_{it}\right)\,\exp\left(E_{it}\right)}$$

The above, implies that for every i, j we have:

$$\min_{j'} \exp\left(E_{ij} - E_{ij'}\right) \tilde{P}_{ij} \leq P_{ij} \leq \tilde{P}_{ij} \,\max_{j'} \exp\left(E_{ij} - E_{ij'}\right),$$

which by the Taylor expansion of exp can be further relaxed to

$$(1 - \min_{j'}(E_{ij} - E_{ij'}))\,\tilde{P}_{ij} \leq P_{ij} \leq \tilde{P}_{ij}\,(1 + 2\max_{j'}(E_{ij} - E_{ij'})).$$

Notice also

$$\max_{j'}(E_{ij} - E_{ij'}) = \max_{j} \boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'}) \quad \text{and} \quad \min_{j'}(E_{ij} - E_{ij'}) = \max_{j'} \boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_{j'} - \boldsymbol{\delta}_j),$$

both of which are at most $\max(\max_{j'} \boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})) = \max_{j'} |\boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})|$, from which the claim follows. □

B Additional results

B.1 The path length distribution of transformers

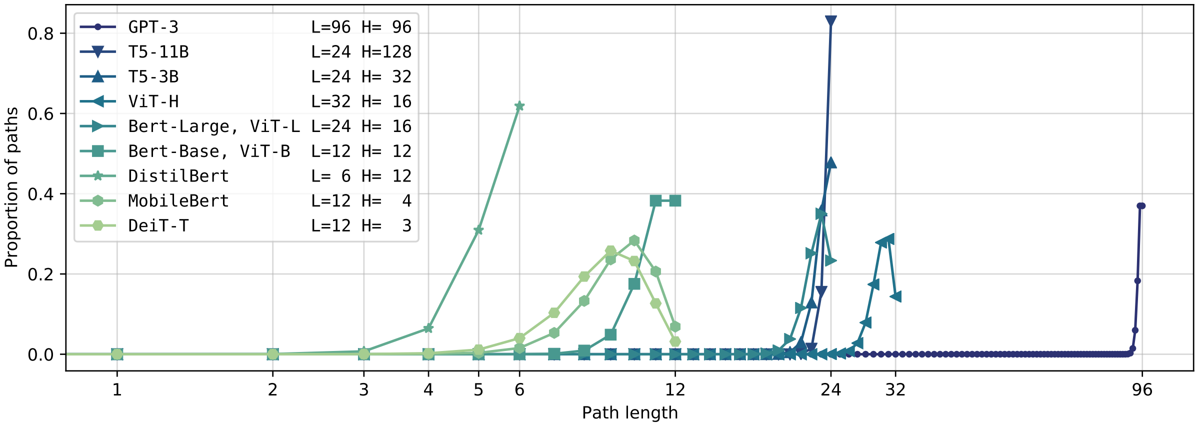
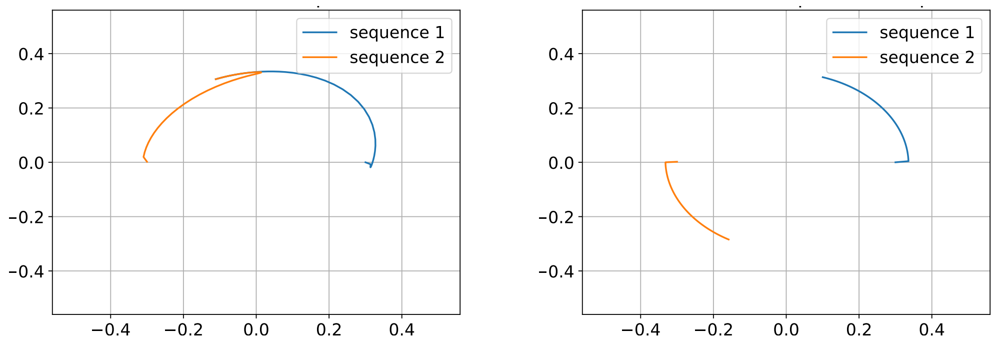


Figure 1: Distribution of the path length for a diverse selection of transformer architectures (encoder only) with different depths and widths. The legends are sorted by the total number of heads in the architecture L×H. We provide the following architecture: GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), Bert (Devlin et al., 2018), ViT (Dosovitskiy et al., 2021), DistilBert (Sanh et al., 2019), MobileBert (Sun et al., 2020).



(a) Learned trajectories converge.

(b) No convergence with larger β.

Figure 2: Convergence behavior of learned trajectories in the circle experiment with different β values. Both plots illustrate models with hidden dimension 40, with neither skip connections nor MLPs. The only difference between (a) and (b) is that the weight matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are scaled up 20-fold for (b), leading to larger β. As shown, the increased β leads to the same diminishing convergence behavior between the two trajectories — the same observation as for increasing β through increasing hidden dimension, and consistent with the theory prediction.

As we saw in §2.1, transformers can be viewed as an interdependent ensemble of simpler networks (or paths) each of different depth (or length). Aiming to gain more insight about the ensemble structure in practice, Fig 1 visualizes the path length distribution in various commonly-used architectures.

Based on the exponential decay of path effectiveness result, we hypothesize that models that focus overwhelmingly on long paths are less efficient than models with a more diverse path distribution. The long-paths models are furthermore likely to be less robust, as they require larger MLP Lipschitz constants to counteract the token-uniformity inductive bias caused by self-attention, as described in §3. It is perhaps no coincidence that the intentionally more efficient models, such as DistilBert or MobileBert, have some of the most diverse path distributions; and that for the most extreme long-paths-focused model, GPT3, studies found that its model size can be reduced by several orders of magnitude and achieve similar performance (Schick & Schütze, 2020). We leave these exciting directions for future work.

B.2 Circle experiment additional discussion

We elaborate further on the circle experiment designed to study the inductive biases of different architectural variants. Recall that we train a single-layer transformer to sequentially predict two circular arcs in $\mathbb{R}^2$, each directed counter-clockwise and consisting of 1000 points.

This task is designed to be a reconstruction task that is simple to learn. Indeed, the model only needs to learn a rotaion by a *fixed* angle, i.e. an $O(2)$ (orthogonal group) action, given by multiplication with $\begin{bmatrix} \cos(\pi/1000) & -\sin(\pi/1000) \\ \sin(\pi/1000) & \cos(\pi/1000) \end{bmatrix}$, where $\pi/100 = 2\pi/2000$ arises from the the fact that there are 2000 points total in the two arcs. This means that in theory, 4 parameters suffice to perfectly learn the task. The simplicity ensures that, the models of different β values evaluated, even the ones with the smallest number of parameters, can easily learn the task, greatly reducing the possibility that the observed results are artifacts of the larger models overfitting the training data.

In addition, we test the effects of increasing β on the learned trajectories independent of increasing the hidden dimension: by scaling up the weight matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$. As observed in Figure 2, an increased β diminishes the convergence behavior between the two trajectories — the same observation as for increasing β through increasing hidden dimension, and consistent with the theory prediction.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. 2020. URL arXiv:2005.14165.

Cordonnier, J.-B., Loukas, A., and Jaggi, M. Multi-head attention: Collaborate instead of concatenate. 2020. URL arXiv:2006.16362.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018. URL arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Schick, T. and Schütze, H. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2158–2170. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.195. URL https://doi.org/10.18653/v1/2020.acl-main.195.