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

Yihe Dong
Google
yihed@google.com

Jean-Baptiste Cordonnier
EPFL
jean-baptiste.cordonnier@epfl.ch

Andreas Loukas
EPFL
andreas.loukas@epfl.ch

Abstract

Attention-based architectures have become ubiquitous in machine learning. Yet our understanding of the reasons for their effectiveness remains limited. This work proposes a new way to understand self-attention networks: we show that their output can be decomposed into a sum of smaller terms, each involving the operation of a sequence of attention heads across layers. Using this decomposition, we prove that self-attention possesses a strong inductive bias towards "token uniformity". Specifically, without skip connections or multi-layer perceptrons (MLPs), the output converges doubly exponentially to a rank-1 matrix. On the other hand, skip connections and MLPs stop the output from degeneration. Our experiments verify the identified convergence phenomena on different variants of standard transformer architectures[1].

1 Introduction

The attention mechanism [BCB15] was initially developed to better learn long-range sequential knowledge, and found effective use in transformer networks [VSP+17]. Since then, attention-based architectures have permeated across data domains machine learning applications, such as in natural language processing [DCLT18], speech recognition [LZL+20], and computer vision [RPV+19, BZV+19]. As such, it is vital to develop tools to understand the inner workings of transformers and attention in general, both to shed light on existing models, and to design more effective future models.

This work provides new insights about the operation and inductive bias of networks built by stacking multiple self-attention layers. Surprisingly, we find that *pure* self-attention networks (SANs), i.e., transformers with skip connections and multi-layer perceptrons (MLPs) disabled, lose expressive power *doubly exponentially* with respect to network depth. More specifically, we prove that the output converges with a cubic rate to a rank one matrix that has identical rows. While we derive the convergence bounds in part by using properties of stochastic matrices, our results go beyond what one would expect based on standard results. In particular, by leveraging the cascading effects of specifically stacking self-attention modules, we show exponentially faster convergence than what standard theory prescribes. Furthermore, while previous studies have considered the rank of individial self-attention matrices [WLK+20, KVPF20, CLJ20a], our results are the first to address conditions under which the *entire* network converges to rank *one*.

This raises the question, why do transformers work? Our analysis indicates that skip connections play a key role in mitigating rank collapse, and MLPs can slow down the convergence by increasing their Lipschitz constant. We characterize these counteracting forces by proving upper and lower bounds of this convergence behavior under SAN architectural variants that resemble transformers. Our results reveal a previously unknown vital utility of skip connections, beyond facilitating optimization and gradient flow [HZRS16a, BFL+18].

---

[1]Our code is publicly available at https://github.com/twistedcubic/attention-rank-collapse
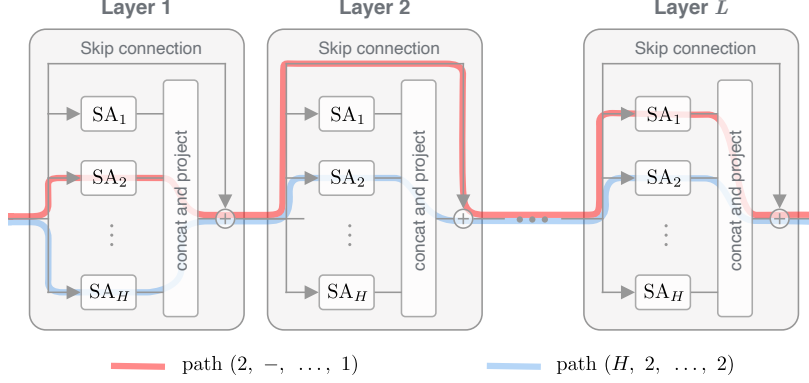
Figure 1: Two paths in a deep Self-Attention Network (SAN) with H heads and L layers. At each layer, a path can go through one of the heads or bypass the layer. Adding an MLP block after each attention layer forms the transformer architecture.

In the process, we develop a new *path decomposition* to study self-attention networks. Namely, we decompose a SAN into a linear combination of weakly-interdependent *paths*, where each 'path' corresponds to a deep single-head SAN. Intuitively, one can view the self-attention heads in each layer of the original network as different gateways, and a path follows a sequence of gateway choices, one gateway per layer (Figure 1). Coupled with the rank collapse analysis, our results suggest that deep SANs with skip connections behave like an ensemble of weakly-dependent shallow networks.

Our main contributions are as follows: (1) We present a systematic study of building blocks of the transformer, revealing opposing impacts between self-attention and the counteracting forces: skip connections and MLP, in contributing and preventing a *rank collapse* in transformers. As a corollary, this reveals a previously unknown vital effect of skip connections beyond facilitating optimization. (2) We propose a new method for analyzing SANs via a *path decomposition*, revealing SANs as an ensemble of shallow networks. (3) We verify our theory with experiments on common transformer architectures.

Notation. In the sequel, bold-face letters denote vectors and matrices. We denote the ℓ_1, ℓ_∞-composite norm of a matrix $\boldsymbol{X}$ as $\|\boldsymbol{X}\|_{1,\infty} = \sqrt{\|\boldsymbol{X}\|_1 \|\boldsymbol{X}\|_\infty}$. We note that $\ell_{1,\infty}$ is not a proper norm as it does not satisfy the triangle inequality, though it is absolutely homogeneous and positive definite. We also use the shorthand notation $[H] = (1, \cdots, H)$.

2   Attention doubly exponentially loses rank

We start by studying self-attention networks (SANs) built exclusively out of multi-head self-attention layers. We prove that SANs converge exponentially (with depth) to a rank-1 matrix that makes all tokens identical.

Our analysis in §2.1 relies on an unconventional way to express the output of a multi-head SAN as a sum of single-head networks. We refer to the latter as *paths*, where each path is denoted by a sequence of attention heads (see Figure 1). A proof sketch of why rank collapse occurs is given in §2.2, whereas the main rank collapse result is presented in §2.3.

2.1 The path decomposition argument

Let $\boldsymbol{X}$ be a $n \times d_{in}$ input matrix consisting of n tokens. A SAN is built out of L multi-head self-attention layers, each having H heads. The output of the h-th self-attention head can be written as

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

Above, $\boldsymbol{W}_{V,h}$ is a $n \times d_v$ value weight matrix and the $n \times n$ *row-stochastic* matrix $\boldsymbol{P}_h$ is given by

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

where (1) the key and query weight matrices $\boldsymbol{W}_{K,h}$ and $\boldsymbol{W}_{Q,h}$ are of size $d_{in} \times d_{qk}$, (2) $\boldsymbol{W}_{QK,h} = \boldsymbol{W}_{Q,h}^\top \boldsymbol{W}_{K,h}$, and (3) the softmax operates independently on each row of its input. We obtain the final equation by noting that softmax is shift-invariant and disregarding terms that provide a constant contribution across rows [CLJ20a].

The output of each SAN layer is formed by concatenating the individual outputs of all H attention heads (along the last dimension) and linearly projecting them onto a subspace of appropriate size:

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

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

Let $\boldsymbol{X}^l$ be the output of the l-th layer and fix $\boldsymbol{X}^0 = \boldsymbol{X}$. As is common practice, we let all layers consist of the same number of heads.

Excluding biases $\mathbf{1}\boldsymbol{b}_{O,h}^\top$, the SAN output is given by

$$\boldsymbol{X}^L = \sum_{h \in [H]} \boldsymbol{P}_h^L \boldsymbol{X}^{L-1} \boldsymbol{W}_h^L$$
$$= \sum_{h \in [H]} \boldsymbol{P}_h^L \bigg(\sum_{h' \in [H]} \boldsymbol{P}_{h'}^{L-1} \boldsymbol{X}^{L-2} \boldsymbol{W}_{h'}^{L-1} \bigg) \boldsymbol{W}_h^L = \sum_{h_L, h_{L-1} \in [H]^2} \boldsymbol{P}_{h_L}^L \boldsymbol{P}_{h_{L-1}}^{L-1} \boldsymbol{X}^{L-2} \boldsymbol{W}_{h_{L-1}}^{L-1} \boldsymbol{W}_{h_L}^L,$$

which, after unrolling the recursion backwards, yields:

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

The above equations have a clear interpretation if we think of the SAN as a directed acyclic graph, with nodes corresponding to self-attention heads and directed edge connecting heads of consecutive layers.

We formalize this intuition in the following:

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*

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

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

Proof. The proof follows from the fact that the set of row-stochastic matrices is closed under multiplication (i.e., $\boldsymbol{P}_{h_L}^L \cdots \boldsymbol{P}_{h_i}^i$ is row-stochastic) and, moreover, for any row-stochastic matrix $\boldsymbol{P}$, we have $\boldsymbol{P}\mathbf{1} = \mathbf{1}$. □

Each term in (1) describes a path of length L across heads of different layers

$$path = (h_1, \ldots, h_L), \quad \text{where} \quad h_l \in (0, 1, \ldots, H)$$

and there are a total of H^L such paths.

The path decomposition thus describes the action of a multi-head SAN as the combination of simpler single-head networks. To gain intuition on path interdependence, it helps to split the operations performed into two types: those that act across tokens (multiplication from left) and those that apply independently on each token (multiplication from right). As seen, though paths can interact through token mixing (since P_{path} matrices jointly depend on X), token-wise operations are independent. We can also notice that biases are not particularly meaningful: their total contribution amounts to the single term $\mathbf{1}\boldsymbol{b}^\top$ independently of the number of layers or heads used.

In the following we show that each path converges rapidly (as a function of length) to a rank-1 matrix with identical rows. Interestingly, this convergence is so dominant that adding more layers to the SAN does not help: though the number of paths is increased exponentially, each path degenerates doubly exponentially, leading also to a rank-1 output.

2.2 Convergence of single-head SAN

Before tackling the full SAN, it is instructive to consider the behavior of each path separately. We examine, in particular, 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.

As the following theorem shows, the residual norm converges to zero surprisingly quickly (doubly exponentially with a cubic rate):

Theorem 2.2. *For any single-head SAN consisting of L layers with $\|\boldsymbol{W}_{QK}^l\|_1 \|\boldsymbol{W}_V^l\|_{1,\infty} \leq \beta$, we have that*

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

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

Note that the bound in Eq 2 guarantees $\|\text{res}(SAN(\boldsymbol{X}))\|_{1,\infty}$ convergence for all inputs of small residual whenever $4\beta < \sqrt{d_{qk}}$. In practice, our experiments imply that the region for convergence can be much greater.

The identified cubic rate of convergence is significantly faster than what would be expected when analyzing products of stochastic matrices (linear rate). As a rule of thumb, to achieve a decline of three orders of magnitude, say from 1000 to 1, one could expect a linear rate of convergence to require roughly a dozen iterations, whereas a cubic rate can do so in just two or three iterations. The reason why we get a cubic rate is that the rank of attention matrices depends also on the rank of the input. As we show, the self-attention heads mix tokens faster when formed from a low-rank matrix. This phenomenon becomes stronger as we build deeper SANs, leading to a cascading effect.

We provide a proof sketch bellow. Detailed proofs can be found in the Appendix.

4

Proof sketch. To analyze how the formation of $\boldsymbol{P}_h$ is affected by the rank of the input, we start by writing $\boldsymbol{X} = \boldsymbol{1}\boldsymbol{x}^\top + \boldsymbol{R}$ for $\boldsymbol{R} = \mathrm{res}(\boldsymbol{X})$ and expanding the attention matrix accordingly:

$$\boldsymbol{X}\boldsymbol{W}_{QK}\boldsymbol{X}^\top = \left(\boldsymbol{1}\boldsymbol{x}^\top + \boldsymbol{R}\right)\boldsymbol{W}_{QK}\left(\boldsymbol{1}\boldsymbol{x}^\top + \boldsymbol{R}\right)^\top$$

Invoking once more the shift-invariance property of the softmax, the above can be simplified to

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

for some appropriate $\boldsymbol{r}$. Observe that if the matrix within the softmax was $\boldsymbol{1}\boldsymbol{r}^\top$, then $\boldsymbol{P}_h$ would also degenerate to a rank-1 matrix: $\mathrm{softmax}(\boldsymbol{1}\boldsymbol{r}^\top) = \boldsymbol{1}\boldsymbol{q}^\top$ and the convergence would happen instantly.

The proof builds on this observation by showing that if $\boldsymbol{E} = \boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top$ is small then $\boldsymbol{P}_h$ is almost rank-1:

$$\|\boldsymbol{P}_h - \boldsymbol{1}\boldsymbol{q}^\top\| \le 2\,\|\boldsymbol{D}\,\boldsymbol{1}\,\boldsymbol{q}^\top\|,$$

where $\boldsymbol{D}$ is diagonal and $D_{ii} = \max_j |\boldsymbol{\delta}_i^\top \boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})|$. Thus, we have

$$\boldsymbol{P}_h\boldsymbol{X} = \boldsymbol{P}_h(\boldsymbol{1}\boldsymbol{x}^\top + \boldsymbol{R}) = \boldsymbol{1}\boldsymbol{x}^\top + \mathrm{softmax}(\boldsymbol{1}\boldsymbol{r}^\top + \boldsymbol{E})\boldsymbol{R}$$

and, moreover, $\|\mathrm{res}(\boldsymbol{P}_h\boldsymbol{X})\| \le 2\,\|\boldsymbol{D}\boldsymbol{1}\,\boldsymbol{q}^\top\boldsymbol{R}\|$. The proof concludes by bounding the above term and applying the argument recursively over successive layers. □

2.3 Exponential convergence for attention networks

We now move on to analyse the convergence of SANs with *multiple* heads per layer.

Our main result is as follows:

Theorem 2.3. *Consider a depth-L and width-H self-attention network without skip connections. Supposing that $\|\boldsymbol{W}_{QK,h}^l\|_1\|\boldsymbol{W}_h^l\|_{1,\infty} \le \beta$ for all heads $h \in [H]$ and layers $l \in [L]$, we have*

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

which amounts to a doubly exponential rate of convergence.

The bound guarantees convergence of SAN($\boldsymbol{X}$) to rank one when $4\beta H < \sqrt{d_{qk}}$. Our experiments show that this is a rather pessimistic estimate, as, in practice, we observe widespread convergence of output to rank-1.

Remark 1. Implications for Xformers. There has been a surge of architectural variants – that we collectively refer to as Xformers – aimed to improve the vanilla transformer [VSP+17] by reducing the quadratic self-attention complexity. The rank collapse result of Theorem 2.3 carries interesting implications for these architectures. One such variant relies on low-rank or kernel-based approximations to the full attention matrix [KVPF20, WLK+20, CLD+20], in which case the paths likely converge even faster to rank one due to the imposed low-rankedness. Another variant only computes a subset of the attention matrix entries using particular patterns [ZGD+20, CGRS19], such as random patterns, in which case one expects the paths to converge more slowly, as randomization tends to increase the rank of the output.

3 Mechanisms that counteract rank collapse

Our findings raise a pertinent question—*why do attention-based networks work in practice if attention degenerates to a rank-1 matrix doubly exponentially with depth?* Aiming to obtain a deeper understanding, we focus on the transformer architecture [VSP$^+$17] and expand our analysis by incorporating the three important components of transformers that SANs lack: *skip connections*, *multi-layer perceptrons*, and *layer normalization.*

We adopt a methodical approach where the modifications to the SAN architecture are introduced one at a time. For each case, we re-derive the convergence bounds and discuss the observed effect.

3.1 Skip connections are crucial

A simple modification to the path decomposition argument for SAN suffices to take into account skip connections. Specifically, we indicate the event that a path has skipped a layer by setting $h = 0$ on the corresponding notation:

$$\boldsymbol{X}^L = \sum_{h \in [H] \cup \{0\}} \boldsymbol{P}_h^L \boldsymbol{X}^{L-1} \boldsymbol{W}_h^L$$

$$= \ldots$$

$$= \sum_{h_1, \ldots, h_L \in ([H] \cup \{0\})^L} (\boldsymbol{P}_{h_L}^L \cdots \boldsymbol{P}_{h_1}^1) \, \boldsymbol{X} \, (\boldsymbol{W}_{h_1}^1 \cdots \boldsymbol{W}_{h_L}^L),$$

where we have fixed $\boldsymbol{P}_0 = \boldsymbol{I}$ and $\boldsymbol{W}_0 = \boldsymbol{I}$.

As observed, skip connections dramatically diversify the path distribution. Denote by $\mathcal{P}_l$ the set of paths of length l. With skip connections enabled, we have

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

paths of length l (whereas before we had only length L paths). We hypothesize that it is the presence of short paths that stops SAN from degenerating to rank-1.

While we can derive an upper bound for the residual similar to above (which we do in the Appendix for completeness) such an upper bound is vacuously large. Indeed, it is more informative to have a *lower* bound on the residual, to align with practice, where SANs with skip connections do not suffer rank collapse. We present the following simple lower bound:

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.*

The proof is elementary: by the path decomposition, there is always a path that skips all layers, i.e. the path with length 0, preserving the residual. It then follows that, for any parametrization that renders the contribution of the SAN layers orthogonal to the input, we will have $\|res(\boldsymbol{X}^L)\| \geq \|res(\boldsymbol{X})\|$. A simple example of such as parametrization can be recovered by setting $\boldsymbol{W}_V^l = 0$ for every $l \in [L]$, in which case $\|res(\boldsymbol{X}^L)\| = \|res(\boldsymbol{X})\|$.

A tight lower bound to the residual in the presence of skip connections is highly nontrivial, and we pose it as an open challenge to the community.

Remark 2. SANs as ensembles of shallow networks. It can be deduced from Theorem 2.3 that the SANs with skip connections enabled heavily rely on short paths (since the residual rapidly declines as the

path length becomes larger). In other words, SANs behave like ensembles of shallow single-head self-attention networks. The phenomenon was previously identified for ResNets [VWB16a] (though the latter study didn't study the rank-collapse phenomenon). Here, the components of this ensemble are inter-dependent, as each attention head participates in many paths of different lengths. Experimental results in §4 support this implication. The supplementary material also provides a study of the paths distribution across several common architectures.

3.2 Multi-layer perceptrons (MLP) help

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

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

Note that, to keep the notation compact, we use f_l to denote 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. Note that, though finding the exact constant can be NP-hard even for shallow MLPs [SV18], since f_l comprises of linear transformations with Lipschitz nonlinearities, f_l is generally Lipschitz.

Corollary 3.2 (SAN with MLP). *Consider a depth-L and width-H SAN with MLP. Supposing that $\|\boldsymbol{W}_{QK,h}^l\|_1 \|\boldsymbol{W}_h^l\|_{1,\infty} \le \beta$ for all $h \in [H]$ and $l \in [L]$ and fix $\lambda_{l,1,\infty} \le \lambda$, we have that*

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

which amounts to a doubly exponential rate of convergence.

As seen, though the effect of MLP is less drastic than that of skip connections, the convergence rate in Cor 3.2 can be controlled by the Lipschitz constants $\lambda_{f,1,\infty}$ of the MLPs: the more powerful the MLPs are the slower the convergence becomes. This reveals a tug-of-war between the self-attention layers and the the MLPs, which due to their nonlinearity can increase the rank. §4 shows that indeed MLPs counteract convergence in experiments.

We should emphasize that using MLPs to counteract the rank-collapse is not without drawbacks: While increasing the Lipschitz constants slows down residual convergence, it also renders the model less sensitive/robust to input perturbations [CKSN18]. Larger Lipschitz constants may also pose greater challenges to optimization, as they lead to larger gradient variance.

3.3 Layer normalization plays no role

Layer normalization is accomplished by rescaling and shifting the input across the feature dimension:

$$\mathrm{LN}(\mathrm{SA}(\boldsymbol{X})) = \mathrm{LN}\left(\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X} \boldsymbol{W}_h + \mathbf{1}\boldsymbol{b}_O^\top \right) = \left(\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X} \boldsymbol{W}_h + \mathbf{1}\boldsymbol{b}_O^\top - \mathbf{1}\boldsymbol{b}_{LN}^\top \right) \boldsymbol{D}_{LN}^{-1},$$

where $\boldsymbol{b}_{\mathrm{LN}}$ is the mean of each column $\mathrm{SA}(\boldsymbol{X})$ and $\boldsymbol{D}_{LN}$ is a diagonal matrix with entries corresponding to the (possibly scaled or shifted) standard deviation of each column $\mathrm{SA}(\boldsymbol{X})$.
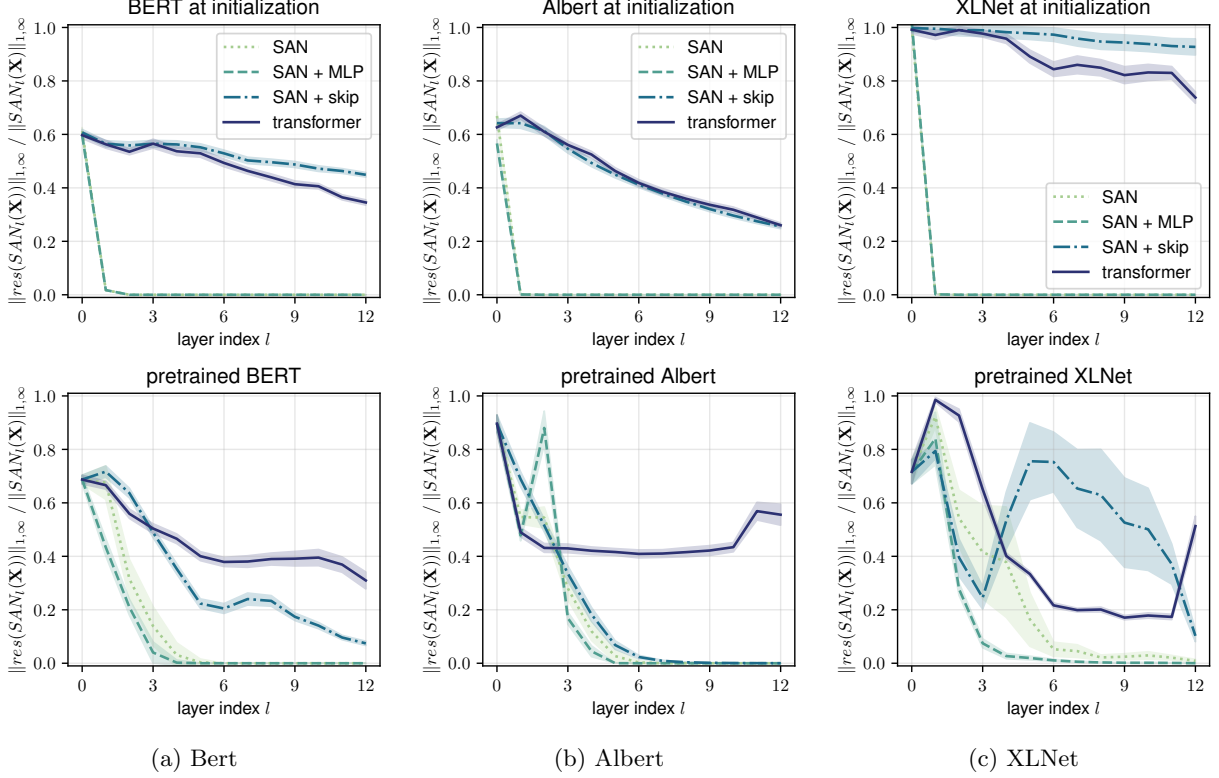
Figure 2: Relative norm of the residual along the depth for three models before and after training. Pure attention (SAN) converges rapidly to a rank-1 matrix. Adding MLP blocks and skip connection gives a transformer. Skip connections play a critical role in mitigating rank collapse (i.e., a zero residual).

By setting $\tilde{\boldsymbol{W}}_h = \boldsymbol{W}_h \boldsymbol{D}_{LN}^{-1}$ and $\tilde{\boldsymbol{b}}_O = \boldsymbol{b}_O - \boldsymbol{b}_{LN}$, the above is re-written as

$$\mathrm{LN}(\mathrm{SA}(\boldsymbol{X})) = \sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X} \tilde{\boldsymbol{W}}_h + \mathbf{1}\tilde{\boldsymbol{b}}_O^\top,$$

which is identical to the equation before layer normalization was applied, though now $\tilde{\boldsymbol{W}}_h$ and $\tilde{\boldsymbol{b}}_O$ are input dependent. Since right multiplication cannot increase the rank of a matrix, we conclude that layer normalization does not mitigate the rank collapse.

4   Experiments

Our experiments first test the rank collapse phenomenon in several well-known transformers architectures (§4.1). We also visually illustrate the inductive bias of some architectural variants of transformers with a toy example in §4.2 and test the paths effectiveness with respect to length in §4.3. Additional results can be found in the Appendix.

4.1   Rank collapse in real architectures

To verify our theoretical predictions, we examine the residual of three well-known transformer architectures: BERT [DCLT18], Albert [LCG+19], and XLNet [YDY+19]. Figure 2 plots the relative residual $\|\mathrm{res}(\mathrm{SAN}(\boldsymbol{X}^l)\|_{1,\infty}/\|\mathrm{SAN}(\boldsymbol{X}^l)\|_{1,\infty}$, of each layer's output before and after the networks have been trained.
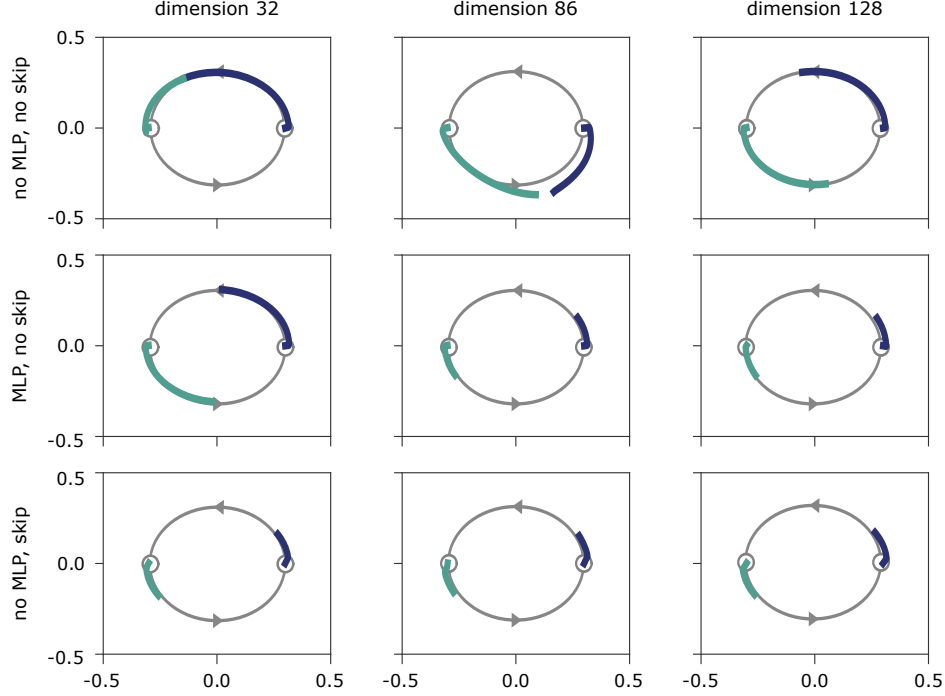
Figure 3: Applying a trained single-layer transformer module recurrently, to models of increasing hidden dimension (horizontal direction) and across architectural variants (vertical direction). The two light background paths illustrate the two training trajectories, for which the starting points are $(-0.3, 0)$ and $(0.3, 0)$. Each figure contains the same number of steps. Consistent with the theory in §3, convergence slows down or stops as the dimension increases (since $\beta \geq \|\boldsymbol{W}_{QK}^l\|_1 \|\boldsymbol{W}_V^l\|_{1,\infty}$ is generally larger), as well as when either MLP or skip connections are added.

To compute these ratios we ran the network on 32 samples of 128 tokens excerpts of biographies from Wikipedia [LGA16] and display the mean and standard deviation.

The experiments confirm that, as soon as the skip connections are removed, all networks exhibit a rapid rank collapse. Though MLPs do not seem to help in the mitigation of convergence, we caution that the observation is not an accurate portrayal of how trained transformers behave: removing the skip connections introduces a drastic distribution shift in the MLP input. We expect that the convergence will slow down if the network is retrained.

4.2 Visualizing the bias of different architectures

To empirically investigate the inductive bias of the different components of the transformer architecture, we study the behavior of a single-layer transformer when applied *recurrently* (akin to the universal transformer [DGV+19]) to predict a simple 2D circular sequence.

Specifically, we train a single-layer transformer to sequentially predict two circular arcs in $\mathbb{R}^2$ of radius 0.3, starting at $(-0.3, 0)$ and $(0.3, 0)$, respectively, each directed counter-clockwise and consisting of 1000 points (illustrated as gray trajectories). An input sample consists of a sequence of two opposing points on the circle, one from the top arc and the other from the bottom arc. We apply teacher-forcing at each step, meaning we give the network the ground truth coordinates of the two current points, and train it to predict the next two points. The model attempts to minimize the MSE loss between the predicted points and the ground truth points on the trajectories. At inference time, we don't apply teacher-forcing, and simply feed the model output as input for the next step.

Since this recurrent application of a single-layer transformer can be reparametrized to be equivalent to a multi-layer transformer without skip connections, *we hypothesize that at inference time the predicted trajectories of the two arcs will converge to the same point (indicating a rank collapse)*, rather than following the training trajectories. Note that the setting has also been intentionally constructed to enable training even without skip connections (by using teacher forcing) and thus to disentangle the two distinct benefits of skip connections: their ability to improve optimization and their mitigation of rank collapse.

We trained the network until it could perfectly memorize the next step on the circular trajectories with near-zero loss. Figure 3 demonstrates the trajectories predicted at inference time (i.e., without teacher forcing). As seen on the top row, without MLP or skip connections the network exhibits rank collapse. Theorem 2.2 predicts that the convergence is slower when $\beta \geq \|\boldsymbol{W}_{QK}^l\|_1 \|\boldsymbol{W}_V^l\|_{1,\infty}$ increases. Indeed, as the hidden dimension increases from 32 to 128 (leading to larger β at initialization), the convergence slows down, becoming hardly observable for dimension 128.

We conclude that, in accordance to our analysis, adding MLP or skip connections either stops or drastically slows down rank collapse. As observed, skip connections tend to slow down points from moving. The latter phenomenon is because in this setting skip connections introduce a bias towards remaining in the same position. On the other hand, adding MLPs does not exhibit the same bias.

4.3 Path effectiveness

SANs can be seen as ensembles of paths of different lengths (from 0 to L), each involving a different sequence of self-attention heads. Our analysis of SAN with skip connections indicates that path expressivity decreases with path length, even if the number of non-linear operations involved increases. To test this hypothesis, we isolate paths of different lengths and evaluate their predictive power.

Tasks. We considered the following three tasks to test path effectiveness with respect to length:

- *Sequence memorization.* To solve this task, a model needs to memorize a pre-determined mapping from natural language sentences and random label sequences of the same length. We use random tokens (rather than actual labels) to make this purely a test of *expressiveness* of a network by way of *memorizing* training data, rather than confounding effects such as generalizability. The models tested are trained to minimize the cross entropy loss between predicted and the ground truth labels. The training data consist of 500 English sentences from Wikipedia and News sources [DGM06, WSM+19], which are tokenized using the SentencePiece tokenizer [KR18] into a vocabulary of size 30522 with 128 tokens per sequence. Each sequence is mapped to a random binary sequence of the same length.

- *Learning to sort.* Given an input sequence of letters, this task learns to sort the letters in alphabetical ordering (similar task have been studied before [FOŠ19]). Specifically, the model's output for each input letter is used to determine the position of that letter in the predicted ordering. Each input sequence, of length 8, is created by sampling uniformly randomly, with replacement, from an alphabet of size 10. The training and test sets consist of 1000 and 200 sequences, respectively. To ensure robustness with respect to hyperparameters, we experimented with a variety of settings (adjusting the model depth, number of heads, and the difficulty of the task by changing the alphabet size and sequence length) and observed consistent behavior.

- *Convex hull prediction.* This task was inspired by the work of [VFJ15]. Given a sequence of N points uniformly distributed in $[0,1] \times [0,1]$ and shifted by a random bivariate standard normal, this task predicts the convex hull of these points. Specifically, 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 $[0,1] \times [0,1]$, each of length 10.

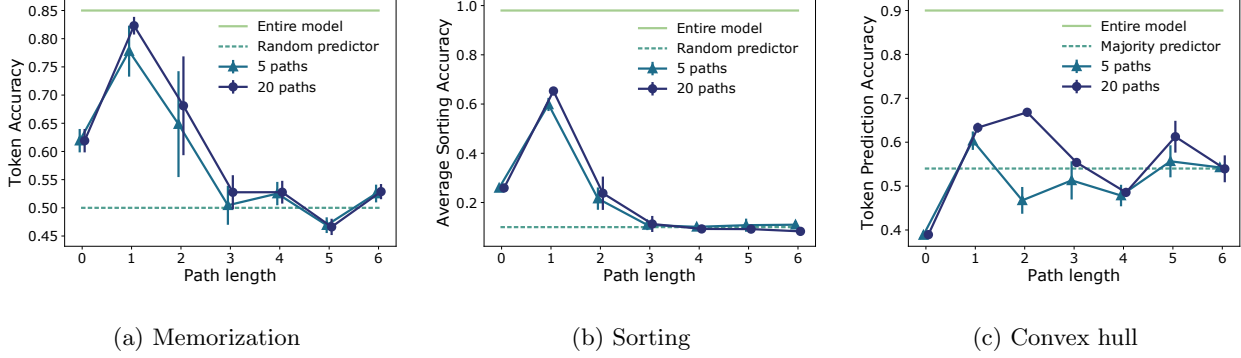In all three tasks, we report the test-set per-token label prediction accuracy as the evaluation metric.

| (a) Memorization | (b) Sorting | (c) Convex hull |

Figure 4: To determine how much of the expressive power can be attributed to short vs long paths, we examine the performance of subsets of paths of different lengths (rather than of the entire SAN). Performance can be seen to consistently deteriorate with respect to path length, supporting our hypothesis that short paths are responsible for the majority of the expressive power.

Path effectiveness test. We measure the effectiveness of individual paths by a 'path disentanglement' procedure that we apply at inference time: the procedure isolates the weights involved and the output of an individual path $(\boldsymbol{P}_{h_L}^L \cdots \boldsymbol{P}_{h_1}^1) \boldsymbol{X} (\boldsymbol{W}_{h_1}^1 \cdots \boldsymbol{W}_{h_L}^L)$ for *any* given sequence of heads $h_1, \cdots, h_L \in [H \cup 0]^L$. After the transformer has been successfully trained to solve each task (without modifications), we use this procedure to determine the output of a randomly sampled set of paths of a given length. We then evaluate the task performance based solely on the normalized sum of this subset of paths (rather than from all paths). Note that the training remains unaltered and uses all heads simultaneously, therefore ensuring that each path learns to its full effectiveness.

Figure 4 illustrates the resulting performance across all three tasks. We test different subset sizes and report the average and standard deviation of five repetitions. For reference, we also plot the accuracy of a naive classifier as well as of the entire trained model (i.e., before the path decomposition). As observed, short paths carry predictive power, with length-1 paths attaining accuracy above 0.8,0.6, and, 0.65 in the memorization, sorting, and convex hull tasks, respectively. On the other hand, the output of longer paths is not much better than a random guess (red horizontal lines). We note that, since there is a class imbalance in the convex hull task, we use a majority class predictor to obtain a random baseline. Though the difference in accuracy between short and long paths is less pronounced for the convex hull task, we observe that the variance of the long paths is significantly larger, rendering them not much better than a random guess. Length zero paths attain very small variance, but contain no useful information about the task (likely because they do not exploit global information).

The depths (L), number of heads (H), and hidden dimensions (d) for the three models are: L:6, H:2, d :250 for memorization, L:6, H:2, d:48 for sorting, and L:6, H:3, d:84 for convex hull. It's important to note that for all three tasks, while higher *peak* accuracies are attainable with increased model capacity and training time, our focus is to study the effects of path length on performance. Indeed, the trend for degenerating performance as path length increases stayed consistent across model sizes in all experiments.

The rapidly diminishing effectiveness of paths with respect to length indicates that the transformer relies almost exclusively on short paths. In other words, the transformer behaves like an ensemble of *shallow* networks. Furthermore, the results indicate that there is *underutilized* capacity in long paths, and suggest that one way to make them, and hence the transformer, more effective, is to prevent the long paths from losing rank.

5   Related works

Skip connections were first introduced in ResNets [HZRS16a], ever since, it has been used to facilitate optimization in deep networks [HZRS16b, VWB16b, BFL⁺18]. In particular, skip connections tackle the vanishing gradient problem, by allowing the gradient to flow bypass the skipped layers during backpropagation. The original motivation of using skip connections in transformers follow the same reasoning on facilitating optimization [VSP⁺17]. With the paths decomposition for transformers, we discover an additional surprising importance of skip connections: they prevent the transformer output from degenerating to rank one exponentially quickly with respect to network depth.

Veit et al ([VWB16b]) introduced an analogous interpretation for residual networks as a collection of paths of varying lengths, and found that the length of the effective paths in deep residual networks are much shorter than the total network depth, due to the gradients used for parameter updates coming overwhelmingly from these short paths. Our finding suggests that SANs rely on short paths to avoid rank collapse.

Some recent works have approximated the attention matrix with low-rank factorizations [WLK⁺20, TBM⁺20] or kernel methods [KVPF20, CLD⁺20], to reduce the quadratic self-attention complexity. Are work is orthogonal to these works, by studying the rank of the network's output (rather than of the attention matrix).

There have been other recent advances in understanding the theory behind transformers: [PMB19, DGV⁺19] proved Turing universality, [CLJ20b] provided necessary and sufficient conditions for attention to simulate convolution. A linearized form of self-attention was also found to exhibit a depth phase transition [LWS⁺20]; and the Lipschitz constant of self-attention was analyzed by [KPM20].

Perhaps the convergence to rank one of a path should come as no surprise: each path component contains row-stochastic matrices as a result of the softmax attention, and [AT77] showed the exponential convergence of products of stochastic matrices to rank one. While the intuition behind stochastic matrices driving convergence still applies, in deep attention networks these matrices interact in more complex ways than what classical analyses consider. As we show, because of these interactions the rank collapses much faster than what would be expected based on classical analyses (cubic vs linear rate).

6   Conclusion

This work exposes competing forces over rank collapse in self-attention networks, namely self-attention vs skip connections and MLPs. In the process, we develop a path decomposition for SANs, which modularizes the study of self-attention and is of independent interest to additional applications. These results open the door for many exciting future directions. For instance, how can one leverage the token-uniformity inductive bias revealed to design more effective networks, perhaps better at utilizing long paths? What are some practical implications for width-depth trade-off? How do we prove meaningful lower bounds of residue convergence for transformers? Answering these questions has broad implications in advancing the state of the art in deep learning.

References

[AT77]   Jac M. Anthonisse and Henk Tijms. Exponential convergence of products of stochastic matrices. In *Journal of Mathematical Analysis and Applications*, 1977.

[BCB15]   Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[BFL+18]   David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question?, 2018.

[BMR+20]   Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

[BZV+19]   Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. In *International Conference on Computer Vision*, 2019.

[CGRS19]   Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. 2019.

[CKSN18]   Zac Cranko, Simon Kornblith, Zhan Shi, and Richard Nock. Lipschitz networks and distributional robustness. *arXiv preprint arXiv:1809.01129*, 2018.

[CLD+20]   Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Jared Davis, Tamas Sarlos, David Belanger, Lucy Colwell, and Adrian Weller. Masked language modeling for proteins via linearly scalable long-context transformers. 2020.

[CLJ20a]   Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. 2020.

[CLJ20b]   Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020.

[DBK+21]   Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[DCLT18]   Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

[DGM06]   Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, 2006.

[DGV+19]   Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

[FOŠ19]   Karlis Freivalds, Emīls Ozoliņš, and Agris Šostaks. Neural shuffle-exchange networks - sequence processing in o(n log n) time. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 6630–6641. Curran Associates, Inc., 2019.

[HZRS16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[HZRS16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[KPM20]   Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. *arXiv preprint arXiv:2006.04710*, 2020.

[KR18]   Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Empirical Methods in Natural Language Processing*, 2018.

[KVPF20] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. 2020.

[LCG+19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[LGA16]   Rémi Lebret, David Grangier, and Michael Auli. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771, 2016.

[LWS+20] Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. Limits to depth efficiencies of self-attention. *arXiv preprint arXiv:2006.12467*, 2020.

[LZL+20] Haoneng Luo, Shiliang Zhang, Ming Lei, , and Lei Xie. Simplified self-attention for transformer-based end-to-end speech recognition. In *CoRR*, 2020.

[PMB19]   Jorge Perez, Javier Marinkovic, and Pablo Barcelo. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019.

[RPV+19] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Neural Information Processing Systems*, 2019.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[SS20]   Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.

[SV18]   Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *arXiv preprint arXiv:1805.10965*, 2018.

[SYS+20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2158–2170. Association for Computational Linguistics, 2020.

[TBM+20] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, , and Che Zheng. Synthesizer: Rethinking self-attention in transformer models. 2020.

[VFJ15]   Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2692–2700, 2015.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[VWB16a] Andreas Veit, Michael Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 550–558, 2016.

[VWB16b] Andreas Veit, Michael Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, 2016.

[WLK+20] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self attention with linear complexity. 2020.

[WSM+19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

[YDY+19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763, 2019.

[ZGD+20] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, 2020.

A   Deffered Proofs

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,

$$\boldsymbol{A} = (\boldsymbol{X}\boldsymbol{W}_Q + \mathbf{1}\boldsymbol{b}_Q^\top)(\boldsymbol{X}\boldsymbol{W}_K + \mathbf{1}\boldsymbol{b}_K^\top)^\top \tag{4}$$

and following [CLJ20a], we can use the softmax shift invariance property to prune the terms constant over the columns and obtain,

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{W}_{QK}\boldsymbol{X}^\top + \mathbf{1}\boldsymbol{b}_{QK}^\top\boldsymbol{X}^\top \tag{5}$$

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

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

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

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

$$\boldsymbol{PX} = \boldsymbol{P}(\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R}) \tag{6}$$

$$= \mathbf{1}\boldsymbol{x}^\top + \boldsymbol{PR} \tag{7}$$

$$= \mathbf{1}\boldsymbol{x}^\top + \mathrm{softmax}(\mathbf{1}\boldsymbol{r}^\top + \boldsymbol{E})\boldsymbol{R} \tag{8}$$

$$\leq \mathbf{1}\boldsymbol{x}^\top + (\boldsymbol{I} + 2\boldsymbol{D})\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top\boldsymbol{R} \tag{9}$$

$$= \mathbf{1}(\boldsymbol{x}^\top + \mathrm{softmax}(\boldsymbol{r})^\top\boldsymbol{R}) + 2\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top\boldsymbol{R} \tag{10}$$

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:

$$|[SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}')^\top]_{ij}| \leq 2\,|[\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top\boldsymbol{RW}_V]_{ij}|, \tag{11}$$

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:

$$\|\boldsymbol{D}\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r})^\top\boldsymbol{RW}_V\|_1 \leq \|\boldsymbol{D1}\|_\infty\|\boldsymbol{R}\|_1\|\boldsymbol{W}_V\|_1, \tag{12}$$

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, $\|\boldsymbol{D1}\|_\infty$ can be bounded as:

$$\|\boldsymbol{D1}\|_\infty = \max_{i,j,j'}|\boldsymbol{\delta}_i^\top\boldsymbol{E}(\boldsymbol{\delta}_j - \boldsymbol{\delta}_{j'})| \leq 2\max_{ij}|E_{ij}| \leq 2\,\|\boldsymbol{E}\|_1$$

$$= 2\,\|\boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top\|_1$$

$$\leq \frac{2}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1\|\boldsymbol{W}_{QK}\|_1\|\boldsymbol{R}^\top\|_1$$

$$= \frac{2}{\sqrt{d_{qk}}}\,\|\boldsymbol{R}\|_1\|\boldsymbol{W}_{QK}\|_1\|\boldsymbol{R}\|_\infty,$$

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:

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

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:*

$$
\|res(SA(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4H\beta}{\sqrt{d_{qk}}}\,\|res(\boldsymbol{X})\|_{1,\infty}^3\,, \tag{13}
$$

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

$$
SA(\boldsymbol{X}) = \sum_{h\in[H]} \boldsymbol{P}_h \boldsymbol{X}\boldsymbol{W}_h = \sum_{h\in[H]} SA_h(\boldsymbol{X}), \tag{14}
$$

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. 11,

$$
|[SA(\boldsymbol{X}) - \mathbf{1}(\boldsymbol{r}'')^\top]_{ij}| \leq 2 \left| \left[\sum_h \boldsymbol{D}_h\,\mathbf{1}\,\mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R}\boldsymbol{W}_h \right]_{ij} \right|, \tag{15}
$$

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

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

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

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

A.3 Single-head and multiple-layers

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

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*

$$
\|res(SAN(\boldsymbol{X}))\|_{1,\infty} \leq \left(\frac{4\,\beta}{\sqrt{d_{qk}}} \right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{16}
$$

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

17

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

$$\|\text{res}(\boldsymbol{X}^L)\|_{1,\infty} \leq \frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X}^{L-1})\|_{1,\infty}^3 \tag{17}$$

$$\leq \frac{4\,\beta}{\sqrt{d_{qk}}}\,\left(\frac{4\,\beta}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X}^{L-2})\|_{1,\infty}^3\right)^3 \tag{18}$$

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

$$\leq \ldots \tag{20}$$

$$\leq \prod_{l=1}^{L}\left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^{3^{l-1}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^{3^L} = \left(\frac{4\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{21}$$

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*

$$\|res(\boldsymbol{X}^L)\|_{1,\infty} \leq \left(\frac{4\,H\,\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}}\,\|res(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{22}$$

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*

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

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:

$$\|\text{res}(\text{SAN}(\boldsymbol{X}))\|_{1,\infty} \leq \frac{4\,\|\boldsymbol{W}_{QK,h}\|_1\|\boldsymbol{W}_V\|_{1,\infty}}{\sqrt{d_{qk}}}\,\|\text{res}(\boldsymbol{X})\|_{1,\infty}^3 + \|\text{res}(\boldsymbol{X})\|_{1,\infty} \tag{23}$$

18

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

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

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

$$\leq 2\max(\frac{4\beta}{\sqrt{d_{qk}}} \|\mathrm{res}(\boldsymbol{X}^{L-1})\|_{1,\infty}^3, \|\mathrm{res}(\boldsymbol{X}^{L-1})\|_{1,\infty}) \tag{24}$$

Now we unroll this bound across layers to write it in terms of $\mathrm{res}(\boldsymbol{X})$. At the kth step of unrolling, the max is one of the two terms in Eq 24: either $\frac{4\beta}{\sqrt{d_{qk}}} \|\mathrm{res}(\boldsymbol{X}^{L-k})\|_{1,\infty}^3$ 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 $\left(\frac{8\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^l-1}{2}} 2^{3^l(L-l)} \|\mathrm{res}(\boldsymbol{X})\|_{1,\infty}^{3^l}$, where l indicates the number of times the term $\frac{4\beta}{\sqrt{d_{qk}}} \|\mathrm{res}(\boldsymbol{X}^{L-k})\|_{1,\infty}^3$ 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:

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

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

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

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

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

which concludes the proof. $\square$

A.6 SAN with MLP

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

$$\boldsymbol{X}^{l+1} = f_l \left(\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h \right). \tag{25}$$

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 $\|\boldsymbol{W}_{QK,h}^l\|_1 \|\boldsymbol{W}_h^l\|_{1,\infty} \leq \beta$ for all $h \in [H]$ and $l \in [L]$ and fix $\lambda_{l,1,\infty} \leq \lambda$. We then have that*

$$\|res(\boldsymbol{X}^L)\|_{1,\infty} \leq \left(\frac{4\beta H \lambda}{\sqrt{d_{qk}}}\right)^{\frac{3^L-1}{2}} \|res(\boldsymbol{X})\|_{1,\infty}^{3^L}, \tag{26}$$

19

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 25, 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 $\mathrm{res}(\boldsymbol{X})$.

The proof proceeds the same way as in Lemma A.1, until Eq 11, where we handle the multi-head case the same way as in Eq A.2 to obtain the entrywise inequality:

$$\left| \left[\sum_{h \in [H]} \boldsymbol{P}_h \boldsymbol{X}^l \boldsymbol{W}_h - \boldsymbol{1}(\boldsymbol{r}')^\top \right]_{ij} \right| \le 2 \left| \left[\sum_h \boldsymbol{D}_h \boldsymbol{1} \, \mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \right]_{ij} \right|, \tag{27}$$

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:

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

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,

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

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

$$\| \boldsymbol{D}_h \boldsymbol{1} \, \mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_1 \le \| \boldsymbol{D}_h \boldsymbol{1} \|_\infty \| \boldsymbol{R} \|_1 \| \boldsymbol{W}_h \|_1, \tag{28}$$

$$\| \boldsymbol{D}_h \boldsymbol{1} \, \mathrm{softmax}(\boldsymbol{r}_h)^\top \boldsymbol{R} \boldsymbol{W}_h \|_\infty \le \| \boldsymbol{D}_h \boldsymbol{1} \|_\infty \| \boldsymbol{R} \|_\infty \| \boldsymbol{W}_h \|_\infty. \tag{29}$$

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 28 and Eq 29, and combining the two as in Lemma A.1, yields the bound:

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

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

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

which concludes the proof. $\qquad\square$

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

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

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. $\square$

B Additional results

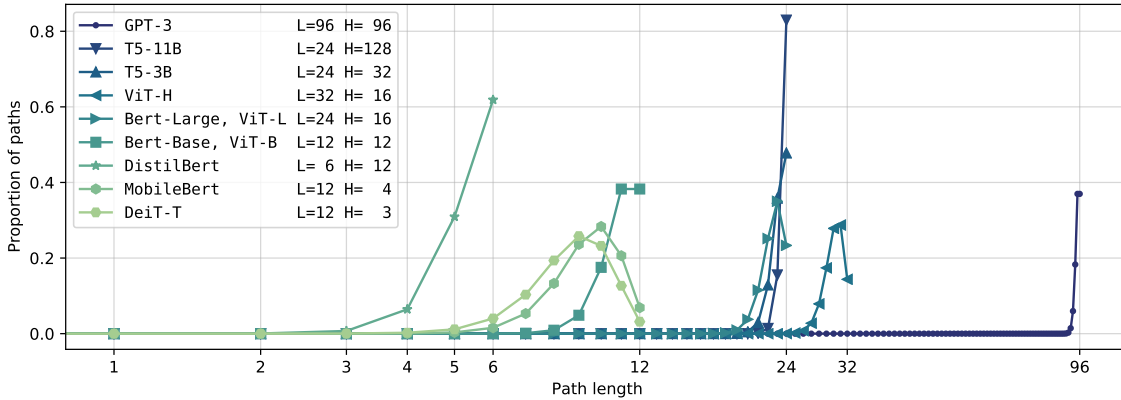## B.1 The path length distribution of transformers

Figure 5: 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 [BMR+20], T5 [RSR+20], Bert [DCLT18], ViT [DBK+21], DistilBert [SDCW19], MobileBert [SYS+20].

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 5 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 [SS20]. We leave these exciting directions for future work.