Limits to Depth Efficiencies of Self-Attention

Yoav Levine
The Hebrew University of Jerusalem
yoavlevine@cs.huji.ac.il

Noam Wies
The Hebrew University of Jerusalem
noam.wies@cs.huji.ac.il

Or Sharir
The Hebrew University of Jerusalem
or.sharir@cs.huji.ac.il

Hofit Bata
The Hebrew University of Jerusalem
hofit.bata@cs.huji.ac.il

Amnon Shashua
The Hebrew University of Jerusalem
amnons@cs.huji.ac.il

Abstract

Self-attention architectures, which are rapidly pushing the frontier in natural language processing, demonstrate a surprising depth-**in**efficient behavior: Empirical signals indicate that increasing the internal representation (network width) is just as useful as increasing the number of self-attention layers (network depth). In this paper, we theoretically study the interplay between depth and width in self-attention. We shed light on the root of the above phenomenon, and establish two distinct parameter regimes of depth efficiency and inefficiency in self-attention. We invalidate the seemingly plausible hypothesis by which widening is as effective as deepening for self-attention, and show that in fact stacking self-attention layers is so effective that it quickly saturates a capacity of the network width. Specifically, we pinpoint a "depth threshold" that is logarithmic in the network width: for networks of depth that is below the threshold, we establish a double-exponential depth-efficiency of the self-attention operation, while for depths over the threshold we show that depth-inefficiency kicks in. Our predictions accord with existing empirical ablations, and we further demonstrate the two depth-(**in**)efficiency regimes experimentally for common network depths of 6, 12, and 24. By identifying network width as a limiting factor, our analysis indicates that solutions for dramatically increasing the width can facilitate the next leap in self-attention expressivity.

1 Introduction

The golden age of deep learning has popularized the depth-efficiency notion: From an expressiveness standpoint, increasing a neural network's size by adding more layers (deepening) is advantageous relatively to other parameter increase alternatives, such as increasing the dimension of the internal representation (widening). Beyond overwhelming empirical signals for this notion [Simonyan and Zisserman, 2014, He et al., 2016], depth-efficiency was theoretically supported from a variety of angles [Cohen et al., 2016, Eldan and Shamir, 2016, Raghu et al., 2017, Daniely, 2017].

Diminishing returns in the case of very deep networks were mainly attributed to optimization issues, and indeed the alleviation of these issues has allowed network depths to mount from 10s to 100s and beyond [He et al., 2016], enabling deep convolutional networks (ConvNets) to advance the state-of-the-art in computer vision applications. However, as the field matured, a more nuanced perspective emerged. Empirical [Zagoruyko and Komodakis, 2016, Wu et al., 2019] and theoretical [Lu et al., 2017] studies suggest that the interplay between depth and width may be more subtle. Recently, a heuristic method for increasing width and depth in tandem has lead to the current state-of-the-art on

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

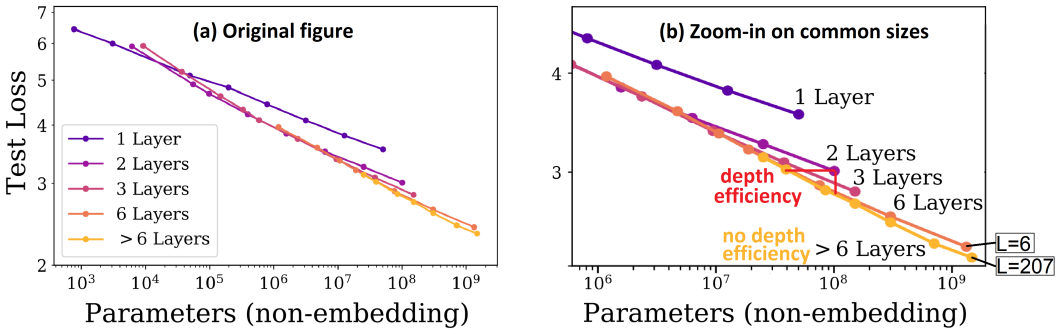

Figure 1: An ablation by Kaplan et al. [2020], examining the perplexity scores on the language modeling task in an extended version of the WebText dataset [Radford et al., 2019], attained when training autoregressive self-attention networks of varying depths and widths. **(a)** Original figure. The perplexity is reported as a function of the overall network size, excluding embedding parameters **(b)** A zoom-in on a parameter regime fitting common widths of $d \geq 200$, which are shown to be sufficient for the task of language modeling. Experiments on the $L > 6$ curve (yellow) include self-attention networks of depths $L = 12, 24, 36, 48, 207$, all approximately obeying the same improvement trend which depends only on the number of network parameters and not on the depth to width ratio ("depth **in**efficiency"). For $L \leq 6$, depth-efficiency is clearly demonstrated.

ImageNet to be set by a ConvNet using a fraction of the parameters used by previous leaders [Tan and Le, 2019].

Since the introduction of the Transformer [Vaswani et al., 2017], along with its encoder-only variant, BERT [Devlin et al., 2019], self-attention based deep learning architectures have taken over the field of natural language processing [Liu et al., 2019, Radford et al., 2019, Yang et al., 2019, Raffel et al., 2019, Clark et al., 2020]. However, in contrast to the depth "arms race" that took place in the ConvNet case, the leading self-attention networks are not much deeper than the original BERT model. In fact, even the strongest self-attention models trained to date, which increased the parameter count of BERT-large by factors of 100s, from 0.3B to 11B [Raffel et al., 2019] and 175B [Brown et al., 2020], have only increased its depth by factors of 2-4. The remaining size increase stems from an increase in layer widths, clearly countering the depth-efficiency notion.

A recent empirical ablation study by Kaplan et al. [2020] provides support for the above signal. Figure 1(a), taken from this study, shows that the overall (non-embedding) network size, given by $12 \cdot L \cdot d_x^2$ where L is the number of self-attention layers (network depth) and d_x is the hidden representation dimension (network width), is the main predictor of performance regardless of the depth to width ratio. This suggests that depth does not play as crucial a role in self-attention networks as it does in convolutional networks. A question may arise whether this phenomenon is not rooted in expressivity but in optimization, which has been shown to be delicate in Transformers [Huang et al.].

In this paper, we theoretically address the above question of the depth to width interplay in self-attention network expressivity, and reveal fundamental subtleties in the above picture. We analyze self-attention networks in which all non-linear activations and normalization operations are removed. Otherwise, the analyzed class (presented in section 2) has the regular deep multi-headed Key/Query/Value structure of common self-attention. After presenting this class in detail, we point to recent studies which demonstrate that normalization and position-wise activations are much less pertinent to the ability of self-attention to correlate inputs than its core connectivity, described in full by our analyzed model. More generally, removing non-linearities for analysis of deep network connectivity traits is commonly done: results on expressiveness and optimization of fully-connected [Saxe et al., 2013, Kawaguchi, 2016, Hardt and Ma, 2016], convolutional [Cohen et al., 2016], and recurrent [Khrulkov et al., 2018, Levine et al., 2018a] networks have been attained via this technique.

Theoretical results on Transformers include a proof that they are universal approximators of sequence to sequence functions [Yun et al., 2019], an examination of their robustness [Shi et al., 2020], a comparison between a single self-attention layer and a single convolutional layer [Cordonnier et al., 2019], and an analysis of the low-rank constraint caused by the multi-headed mechanism [Bhojanapalli et al., 2020]. A different empirical trend demonstrated in Kaplan et al. [2020] was recently addressed theoretically in Sharma and Kaplan [2020], which shed light on the scaling exponent of the loss with

model size in neural models. To the best of our knowledge, our analysis is the first to address the question of parameter allocation between depth and width in self-attention networks.

We employ the tool of a function's separation rank with respect to subsets of its inputs, which quantifies its ability to model input dependencies (presented in section 3). The separation rank was employed for attaining theoretical insights on the dependencies modeled by convolutional and recurrent networks [Cohen and Shashua, 2017, Levine et al., 2018a].

Rather than reinforcing the seemingly plausible hypothesis for the trend in figure 1, by which widening a self-attention network is as effective as deepening it, we confirm the contrary. We show that the operation of stacking self-attention layers is so effective that it quickly saturates a capacity of the network's width. We establish in section 4 the existence of a depth threshold which depends logarithmically on the width d_x, denoted $L_{\text{th}}(d_x) = \log_3(d_x)$. Below the threshold, we prove that depth-efficiency takes place in self-attention networks: a network of depth $L \leq L_{\text{th}}(d_x)$ cannot be replicated by a shallower network, unless the latter's width grows double-exponentially with L. We prove the above by showing that the separation rank of functions realized by self-attention networks grows double-exponentially with depth, but only polynomially with width, shedding light on the effectiveness of the self-attention mechanism in modeling input interactions when recursively repeated. However, we show that this overwhelming advantage of depth is quickly replaced by a balanced growth. We prove that for self-attention networks with $L > L_{\text{th}}(d_x)$ the ability to model input dependencies, as modeled by the separation rank, increases similarly with depth and width.

A closer look at the experiment of Kaplan et al. [2020], displayed in figure 1(b), reveals an agreement with our theoretical indications. For two networks with the same parameter count but of different depths $L_1 < L_2$ and widths $d_2 < d_1$: (1) the performance is the same when the dimension of the deeper network d_2 is too small (our theory indicates that the width caps the benefit of the added layers of depths $L_1 + 1, ..., L_2$), but (2) the deeper network outperforms the shallower one when its width d_2 is large enough such that the added layers are in the depth efficiency regime. Thus, even though a depth **in**efficiency of self-attention was concluded from this experiment, it shows traces of the more nuanced phenomenon predicted by our theory. The experiments in figure 1 show the depth efficiency regime for networks of depths $L \leq 6$. In section 5 we demonstrate empirically that depth efficiency**/in**efficiency regimes affect more commonly used self-attention depths of $L = 6, 12, 24$. Following the presentation of our results, we discuss in section 6 practical outcomes derived from our theoretical insights.

2 The self-attention mechanism

Differentiable attention models in which the output attends over all LSTM-based input representations have been introduced in the context of machine translation [Bahdanau et al., 2014]. Self-attention (also referred to as intra-attention), which relates different inputs to each other, was first employed for machine reading [Cheng et al., 2016], and soon thereafter shown to be useful for a variety of language applications when operating over LSTM-based representations [Parikh et al., 2016, Paulus et al., 2017, Lin et al., 2017]. Vaswani et al. [2017] were the first to demonstrate that a model based solely on attention, the Transformer, can be better than LSTM based networks. The Transformer's encoder, BERT [Devlin et al., 2019], based entirely on self-attention, has demonstrated unprecedented performance across natural language understanding tasks.

2.1 The Transformer encoder architecture

We begin by describing the self-attention operation of the original Transformer, and then in the next subsection we present the modifications made in our analyzed model. Each layer $l \in [L] := \{1, ..., L\}$ of a depth-L Transformer encoder is comprised of two sub-layers. The H-headed self-attention sublayer of layer l computes the following function at position $i \in [N]$, over its N inputs $\{\mathbf{x}^{l,j} \in \mathbb{R}^{d_x}\}_{j=1}^{N}$:

$$\mathbf{f}_{\text{SA}}^{l,i}\left(\mathbf{x}^{l,1,}, ..., \mathbf{x}^{l,N}\right) = \sum_{j=1}^{N}\sum_{h=1}^{H} SM_j\left\{1/\sqrt{d_a}\left\langle W^{\text{Q},l,h}\mathbf{x}^{l,i}, W^{\text{K},l,h}\mathbf{x}^{l,j}\right\rangle\right\} W^{\text{O},l,h}W^{\text{V},l,h}\mathbf{x}^{l,j} \qquad (1)$$

where $SM_j\left\{f(j)\right\} := e^{f(j)}/\sum_{j'} e^{f(j')}$ is the softmax operation and $\forall h \in [H]$ the learned weights matrices $W^{\text{K},l,h}, W^{\text{Q},l,h}, W^{\text{V},l,h} \in \mathbb{R}^{d_a \times d_x}$ convert the representation from its dimension d_x into the attention dimension $d_a = {}^{d_x}/_{H}$, creating Key, Query, and Value representations, respectively. The learned weights matrix $W^{\text{O},l,h} \in \mathbb{R}^{d_x \times d_a}$ converts the attention result back into the representation

dimension. The multi-headed self-attention sublayer output in eq. (1), followed by a residual connection and layer-norm [Ba et al., 2016], is inserted into a position-wise feed-forward + ReLU sublayer, such that each layer's output at position $i \in [N]$ is:

$$\mathbf{f}^{l,i}_{\text{Layer}}\left(\mathbf{x}^{l,1},...,\mathbf{x}^{l,N}\right) = W^{\text{FF},2}ReLU\left(W^{\text{FF},1}LayerNorm\left(\mathbf{f}^{l,i}_{\text{SA}} + \mathbf{x}^{l,i}\right)\right), \tag{2}$$

where the feed-forward matrices are usually taken to be $W^{\text{FF},1} \in \mathbb{R}^{4d_x \times d_x}, W^{\text{FF},2} \in \mathbb{R}^{d_x \times 4d_x}$, such that the parameter count for an entire layer is $12 \cdot d_x^2$. Finally, the depth-L multi-headed self-attention operation of the Transformer encoder is obtained by a composition of L such layers, *i.e.*, when setting $\forall l \in \{2,...,L\}, j \in [N]: \ \mathbf{x}^{l,j} = LayerNorm\left(\mathbf{f}^{l-1,j}_{\text{Layer}}\right)$, with $\mathbf{x}^{1,j}$ denoting the input to the deep self-attention network at position j.[1]

2.2 The analyzed architecture

We analyze a deep multi-headed self-attention network variant which excludes the layer-norm operation, the softmax normalization, and the ReLU activation (see a thorough discussion on the effect of these relaxations in the next subsection). For cleanliness of presentation, we defer the analysis of the residual connection to the appendix (it bears insignificant impact on our bounds). Specifically, in the analyzed network, each layer $l \in [L]$ computes the following function at position $i \in [N]$ over its inputs $\{\mathbf{x}^{l,j} \in \mathbb{R}^{d_x}\}_{j=1}^{N}$:

$$\mathbf{y}^{l,i}\left(\mathbf{x}^{l,1},...,\mathbf{x}^{l,N}\right) = \sum_{j=1}^{N}\sum_{h=1}^{H}\left\langle W^{\text{Q},l,h}\mathbf{x}^{l,i}, W^{\text{K},l,h}\mathbf{x}^{l,j}\right\rangle W^{\text{O},l,h}W^{\text{V},l,h}\mathbf{x}^{l,j}, \tag{3}$$

where the Feed-Forward matrices can be now effectively embedded within $W^{\text{O},l,h}$. Our analysis below treats a deep multi-headed self-attention network that is attained by a concatenation of L such layers. Importantly, the resultant "linearized" network form, where activations and normalizations are removed, is by no means a linear mapping over the network input – every layer integrates 3 copies of its input in the above non-linear fashion.

By recursively applying eq. (3) L times we attain the analyzed depth-L self-attention network. We denote the function realized by a network with embedding dimension d_x and H attention heads per layer at output location $i \in [N]$ by:

$$\mathbf{y}^{i,L,d_x,H,\Theta}\left(\mathbf{x}^1,...,\mathbf{x}^N\right) := \sum_{j_1,...,j_C=1}^{N}\mathbf{g}^L\left(\mathbf{x}^i,\mathbf{x}^{j_1},...,\mathbf{x}^{j_C}\right), \tag{4}$$

where Θ denotes all $4LH$ learned weight matrices: $\forall(l,h) \in [L] \otimes [H]: W^{\text{K},l,h}, W^{\text{Q},l,h}, W^{\text{V},l,h} \in \mathbb{R}^{d_a \times d_x}$, and $W^{\text{O},l,h} \in \mathbb{R}^{d_x \times d_a}$, and the function $\mathbf{g}^L$ is a placeholder, fully detailed in the appendix, which integrates $C = \frac{3^L-1}{2}$ different input vectors. Network connectivity implies that the number of summed position indices is also C. Comparing the form of eq. (4) to the operation of a single layer in eq. (3), it can be seen schematically that while a single layer mixes the output position i with every input position j once and aggregates the result, depth brings forth an exponential enhancement to the amount of inputs mixed at once as well as to the amount of summed terms. In section 4, we quantify this effect and analyze the limitations posed by the dimension of the internal representation (the width) on the network's ability to make use of this exponential growth with depth. In the following subsection, we comment on the differences between the Transformer encoder architecture described in eqs. (1) and (2) and the self-attention architecture presented in eqs. (3) and (4).

2.3 Relaxations

Empirical evidence indicates that while the ReLU activations and softmax normalization contribute to performance (layer-norm mainly contributes to optimization), the basic mechanism in eqs. (3) and (4) above captures the defining self-attention characteristic of integrating the inputs with each other in a flexible manner:

The ReLU activation relaxation: Press et al. [2019] demonstrate that a "self-attention first" BERT variant that first performs all of the self-attention operations (eq. (1)) consecutively, and only then

[1] Focusing on the self-attention operation, we omit a description of the input embedding matrix, as well as of the positional embeddings added at the input, which do not affect our analysis given realistic vocabulary sizes.

performs all of the position-wise feed-forward+ReLU operations, achieves comparable language modeling performance relatively to the Baseline, which takes the regular approach of interleaving these functionalities (*i.e.*, concatenating the BERT's layer described in eq. (2)). They report that the interleaved Baseline achieves a perplexity score of 18.63 ± 0.26 on the WikiText-103 test [Merity et al., 2016] when averaged over 5 random seeds, while the "self-attention first" model achieves a perplexity score of 18.82 on this test set. The best pre-Transformer perplexity result on the WikiText-103 test, reported by an LSTM-based architecture, was 29.2 [Rae et al., 2018]. Since ReLU and feed-forward do not mix different locations, this outcome directly implies that the self-attention mechanism itself provides all of the elaborate input integration which differentiates BERT from previous architectures.

The softmax normalization relaxation: Initially, an intuitive interpretation of attention as distributing "fractions" of an overall attention budget among inputs was given to its actual operation of dynamically linking input and output locations. The intuitive interpretation, tightly linked to the need to transform the Key/Query similarity score into a distribution, has been recently challenged, as a growing body of work shows that the attention weights distribution does not directly correlate with predictions [Jain and Wallace, 2019, Pruthi et al., 2019, Brunner et al., 2020]. Moreover, Richter and Wattenhofer [2020] recently point out undesirable traits of the softmax operation, demonstrating that its property of confining the outcome to the convex hull of its inputs unnecessarily limits the expressibility of the self-attention mechanism. They experiment on a suite of synthetic tasks with a BERT variant in which the softmax normalization is removed, and find it to perform on par on almost all examined tasks. When replacing the softmax with other normalizations they report improvements. Finally, completely linearized attention (softmax removed) was employed on real tasks as means of reducing costs, since the softmax operation cost scales with the input size [de Brébisson and Vincent, 2016, Wang et al., 2020].

The goal of the above points is not to advocate modifications in BERT's non-linearity or normalization operations (we leave that to other works), but to note that while these are under examination and are susceptible for alteration, the connectivity of self-attention, manifested by eqs. (3) and (4) , is the core mechanism driving its functionality. Our results, to be presented in section 4, demonstrate how conclusions drawn by directly analyzing this mechanism accord with the operation of commonly employed self-attention networks.

3 A measure of capacity for modeling input dependencies

In this section, we introduce the separation rank of the function realized by a self-attention network as a measure that quantifies its ability to model dependencies between subsets of its variable set $\{\mathbf{x}^j\}_{j=1}^N$. We will use this measure in order to establish the two depth efficiency/ inefficiency regimes in self-attention. The separation rank, introduced in Beylkin and Mohlenkamp [2002] for high-dimensional numerical analysis, was employed for various applications, *e.g.*, chemistry [Harrison et al., 2003], particle engineering [Hackbusch, 2006], and machine learning [Beylkin et al., 2009]. Importantly, the separation rank has been established as a measure of dependencies modeled by deep convolutional and recurrent networks w.r.t. their inputs [Cohen and Shashua, 2017, Levine et al., 2018a,b].

Let (A, B) be a partition of the input locations, *i.e.*, A and B are disjoint subsets of $[N]$ whose union gives $[N]$. The separation rank of a function $y(\mathbf{x}^1, \ldots, \mathbf{x}^N)$ w.r.t. partition (A, B), is the minimal number of summands that together sum up to equal y, where each summand is *multiplicatively separable w.r.t.* (A, B), *i.e.*, is equal to a product of two functions – one that intakes only inputs from one subset $\{\mathbf{x}^j : j \in A\}$, and another that intakes only inputs from the other subset $\{\mathbf{x}^j : j \in B\}$. Formally, the *separation rank* of $y : (\mathbb{R}^{d_x})^N \to \mathbb{R}$ w.r.t. the partition (A, B) is defined as follows:

$$sep(y; A, B) := \min \Big\{ R \in \mathbb{N} \cup \{0\} : \exists g_1 \ldots g_R : (\mathbb{R}^{d_x})^{|A|} \to \mathbb{R}, g'_1 \ldots g'_R : (\mathbb{R}^{d_x})^{|B|} \to \mathbb{R} \;\; s.t. \\ y\left(\mathbf{x}^1, \ldots, \mathbf{x}^N\right) = \sum\nolimits_{r=1}^{R} g_r\left(\{\mathbf{x}^j : j \in A\}\right) g'_r\left(\{\mathbf{x}^j : j \in B\}\right) \Big\} \tag{5}$$

If the separation rank of a function w.r.t. a partition of its input is equal to 1, the function is separable, meaning it cannot take into account consistency between the values of $\{\mathbf{x}^j\}_{j\in A}$ and those of $\{\mathbf{x}^j\}_{j\in B}$. In a statistical setting, if y is a probability density function, this would mean that $\{\mathbf{x}^j\}_{j\in A}$ and $\{\mathbf{x}^j\}_{j\in B}$ are statistically independent. The higher $sep(y; A, B)$ is, the farther y is from this situation, *i.e.* the more it models dependency between $\{\mathbf{x}^j\}_{j\in A}$ and $\{\mathbf{x}^j\}_{j\in B}$, or equivalently, the stronger the correlation it induces between the inputs indexed by A and those indexed by B.

The fixed connectivity of ConvNets has been shown to yield high separation ranks w.r.t. partitions which separate neighboring inputs (*e.g.*, where all odd positions are in A and all even positions are in B), while suffering from low separation ranks w.r.t. partitions which separate distant inputs (*e.g.*, where $A = 1, ..., N/2$ and $B = N/2 + 1, ..., N$). Our analysis establishes a qualitatively different trait for self-attention networks, which treat all balanced partitions alike:

Proposition 1. *For $p \in [d_x]$, let $y_p^{i,L,d_x,H,\Theta}$ be the scalar function computing the pth entry of an output vector at position $i \in [N]$ of the depth-L self-attention network with embedding dimension d_x and H attention heads per layer, defined in eqs.* (3) *and* (4)*. Then, its separation rank w.r.t. balanced partitions, which obey $A \cup B = [N], |A| , |B| = N/2$, is invariant to the identity of the partition,* i.e.*, $\forall A \cup B = [N], \tilde{A} \cup \tilde{B} = [N], \;\; s.t. \;\; |A| , |B| , |\tilde{A}|, |\tilde{B}| = N/2$:*

$$sep(y_p^{i,L,d_x,H,\Theta}; A, B) = sep(y_p^{i,L,d_x,H,\Theta}; \tilde{A}, \tilde{B}) \tag{6}$$

Accordingly, we will omit the specification of the partition in future uses, denoting $sep(y_p^{i,L,d_x,H,\Theta})$ as the separation rank of $y_p^{i,L,d_x,H,\Theta}$ w.r.t. any balanced partition of the inputs.

This result accords with the intuition regarding the flexibility of the attention mechanism – it does not integrate the input in a predefined pattern like convolutional networks, but dynamically learns to correlate any inter-dependent subsets of the inputs. Natural text exhibits non-smooth non-local dependency structures, as correlations between input segments can abruptly rise and decay with distance. The fact that self-attention facilitates all correlation patterns equally poses it as a more natural architecture for language modeling related tasks. Convolutional networks, with their local connectivity, may have the right inductive bias for imagery data, but partitions unfavored by them may reflect more erratic correlations that are nonetheless relevant for natural language inputs.

However, the above property of indifference to the input partition is not enough for succeeding at tasks with elaborate input dependencies, since a function with equally low separation ranks for all input partitions has limited ability to model such dependencies. In the following section, we analyze how different architectural parameters affect the ability of self-attention networks to correlate their inputs, and by bounding their separation ranks, we establish the different depth-efficiency regimes in self-attention networks.

4 The effect of depth in self-attention networks

In this section, we present tight bounds on the separation rank of self-attention networks, which reveal two qualitatively different regimes. In the first regime of $L < \log_3(d_x)$, analyzed in subsection 4.1, we establish that deepening is clearly preferable to widening. In the second regime of $L > \log_3(d_x)$, analyzed in subsection 4.2, we show that deepening and widening play a similar role in enhancing the expressiveness self-attention networks.

4.1 Depth efficiency in self-attention

The recursive structure of deep self-attention hints at an exponential increase of input mixing with depth: The output of each layer is introduced 3 times into the Key/Query/Value computation made by the subsequent layer. In this subsection, we formalize this intuition for self-attention networks of sufficient width, $d_x > 3^L$. Theorem 1 below bounds the separation rank of such networks. Subsequent to its statement and brief outline of its proof, we explicitly show in corollary 1 the implied double-exponential requirement from a bounded depth network attempting to replicate a deeper one.

Theorem 1. *For $p \in [d_x]$, let $y_p^{i,L,d_x,H,\Theta}$ be the scalar function computing the pth entry of an output vector at position $i \in [N]$ of the depth-L self-attention network with embedding dimension d_x and H attention heads per layer, defined in eqs.* (3) *and* (4)*. Let $sep(y_p^{i,L,d_x,H,\Theta})$ be its separation rank (section 3). If L, d_x obey $L < \log_3 (d_x)$, then the following holds almost everywhere in the network's learned parameter space,* i.e. *for all values of the weight matrices (represented by Θ) but a set of Lebesgue measure zero:*

$$3^{L-2} \left(\log_3 \left(d_x - H\right) + a\right) \leq \log_3 \left(sep(y_p^{i,L,d_x,H,\Theta})\right) \leq \frac{3^L - 1}{2} \log_3 \left(d_x + H\right) \tag{7}$$

with $a = -L + [2 - \log_3 2]$. (note that $\log_3 \left(d_x - H\right) + a > 0$ in this regime of $L < \log_3(d_x)$).

We provide below a short proof sketch of the lower bound in the above theorem. The derivation of the upper bound is more straightforward, and is left for the appendix, along with a formal proof of the lower bound.

Proof sketch for the lower bound in theorem 1: We make use of grid tensor based function discretization [Hackbusch, 2012] – The function realized by a self-attention network is evaluated for a set of points on an exponentially large grid in the input space, and the outcomes are stored in a matrix $\mathcal{M}\left(y_p^{i,L,d_x,H,\Theta}\right)$, which we prove upholds: rank $\left[\mathcal{M}\left(y_p^{i,L,d_x,H,\Theta}\right)\right] \leq sep(y_p^{i,L,d_x,H,\Theta})$, *i.e.*, its rank lower bounds the separation rank. Since the entries of $\mathcal{M}\left(y_p^{i,L,d_x,H,\Theta}\right)$ vary polynomially with the self-attention network's weights, we show that it suffices to find a single network weights assignment Θ for which the rank of the matrix is greater than the desired lower bound, in order to prove the case for almost all of the configurations of the network's learned weights (but a set of measure zero). Thus, we prove the lower bound in theorem 1 by choosing a simple weight assignment that still represents the self-attention connectivity, and showing that for this value of Θ, rank $\left[\mathcal{M}\left(y_p^{i,L,d_x,H,\Theta}\right)\right]$ achieves the lower bound, in turn lower bounding the separation rank. □

Theorem 1 bounds the separation rank of a deep self-attention network of sufficient width between two functions that grow double-exponentially with depth and polynomially with width, tightly describing its behavior w.r.t. depth and width. Because equivalence cannot hold between two functions of different separation ranks, the above result implies a double-exponential requirement from the width of a shallow network attempting to replicate the deep one:

Corollary 1. *With probability* 1*, the function realized upon randomization of the weights of a deep self-attention network defined in eqs.* (3) *and* (4) *with depth* L^{deep} *and width* $d_x^{deep} > 3^{L^{deep}}$*, may only be realized by a shallower network with depth* $L^{shallow} = {}^{L^{deep}}/_{d}$ *and width* $d_x^{shallow} = w d_x^{shallow}$*, where* $d > 1, w > 1$ *(*i.e.*, the deep network is deeper by a factor of* d *and the shallow network is wider by a factor of* w*), if the following holds:*

$$w \propto \exp(\exp(d)).$$

The above requirement implies clear-cut (double-exponential) depth-efficiency: the shallow network must grow impractically large to match the deeper one. For example, for BERT-large parameters of $d_x^{\text{deep}} = 1000, H = 16$, by taking the deep network under the depth-efficiency threshold $L^{\text{deep}} = 6$, the width of a depth $L^{\text{shallow}} = 2$ network has to be $d_x^{\text{shallow}} \simeq 2 \cdot 10^{17}$ and the width of a depth $L^{\text{shallow}} = 3$ network has to be $d_x^{\text{shallow}} \simeq 2 \cdot 10^5$ to match the deep network's operation. These numbers were attained by numerically equating the upper bound in eq. (7) for the shallow network and the lower bound in eq. (7) for the deep network, *i.e.*, by asking when the upper bound on the shallow network is larger than the lower bound on the deep network.

4.2 Depth in-efficiency in self-attention

Beyond establishing depth-efficiency in early self-attention layers, the above analysis sheds light on the contribution of a self-attention network's depth to its ability to correlate input subsets. The separation rank (w.r.t. any partition) of a single layer, given by eq. (3), is only linear in H and d_x, showcasing a limitation of the class of functions realized by single self-attention layers to model elaborate input dependencies. Theorem 1 quantifies the double exponential growth of this capacity measure with the number of stacked self-attention layers. The following theorem shows that this growth is capped by the dimension of the internal representation:

Theorem 2. *For* $y_p^{i,L,d_x,H,\Theta}$ *as defined in theorem 1, if* $L > \log_3(d_x)$*, then the following holds almost everywhere in the network's learned parameter space,* i.e. *for all values of the weight matrices (represented by* Θ*) but a set of Lebesgue measure zero:*

$$\frac{1}{2} d_x \cdot L + b_1 + b_2 \leq \log_3\left(sep(y_p^{i,L,d_x,H,\Theta})\right) \leq 2d_x \cdot L + c_1 + c_2 \tag{8}$$

with corrections on the order of L*:* $b_1 = -L\left(\frac{H}{2} + 1\right)$*,* $c_1 = L$*, and on the order of* $d_x \log_3(d_x)$*:* $b_2 = -d_x\left(1 + \frac{1}{2}\log_3\left(\frac{d_x - H}{2}\right)\right)$*,* $c_2 = -2d_x \cdot \log_3 {}^{d_x}/_{2\sqrt{2e}} + \log_3 d_x$*.*

We provide below a proof sketch of the upper bound in the above theorem. The formal proof, along with the proof of the lower bound, which is similar to the one illustrated above for the lower bound in theorem 1, are left for the appendix.

Proof sketch for the upper bound in theorem 2: By observing that $y_p^{i,L,d_x,H,\Theta}$ is a polynomial of degree $2C = 3^L - 1$ (C is introduced in eq. (4)), we find a kernel $\psi\left(\mathbf{x}^1, ..., \mathbf{x}^N\right)$ that maps the input into a space where each of the output monomials is a linear functional. We find a basis for the subspace V spanned by the output monomials, and bound the separation rank of each element in that basis by a constant. The dimension of V is exponential in Nd_x and polynomial in $3^L - 1$,

providing equal groundings for depth and width. A careful analysis that exploits the sums over the indices $j_1, ..., j_C$ in eq. (4), removes the dependence on N. $\square$

Theorem 2 states that when the network's depth passes a width dependent threshold, the separation rank turns from increasing polynomially with width and double-exponentially with depth to increasing-exponentially with width and depth together. Thus, while an increase in network size increases its capacity to model input dependencies, our result shows that there is no longer a clear cut advantage of depth in this respect:

Corollary 2. *Let $\mathbf{y}^{deep}$ denote the function realized by a deep self-attention network at any output location $i \in [N]$, defined in eqs.* (3) *and* (4) *with depth and width denoted L^{deep}, d_x^{deep} such that $L^{deep} > \log_3 d_x^{deep}$. Denote $\beta_1 := \frac{\log_3 d_x^{deep}}{L^{deep}} < 1$. Then, there exists $\beta_2 = O(log(H) \cdot log(d_x^{deep}) \cdot log(L^{deep}))$ such that the function realized by a network of depth: $L^{shallow} = \beta_1 \cdot L^{deep} + \beta_2$, and width: $d_x^{shallow} = 3^{\beta_2} d_x^{deep}$, denoted $\mathbf{y}^{shallow}$, has higher separation rank,* i.e.*:*

$$sep(y_p^{shallow}) > sep(y_{p'}^{deep}) \quad ; \quad \textit{where } p, p' \in [d_x] \tag{9}$$

Corollary 2, which follows from theorems 1 and 2, shows that the separation rank of a function realized by a self-attention network of arbitrary depth $L > \log_3(d_x)$ can be surpassed by a shallower network of polynomial width, contrarily to the established behavior for networks of depth $L < \log_3(d_x)$.

We leave it as an open conjecture that a polynomially sized shallower network can exactly replicate the operation of a deeper network in this regime. With that, we point out that a variety of results which directly bound different complexity measures of deep networks have been put forward, shedding light on their operation [Montufar et al., 2014, Bianchini and Scarselli, 2014, Raghu et al., 2017, Serra et al., 2017, Inoue, 2019]. Bounds on the separation rank have been used to explain the operation of more veteran architectures, and we find them to be particularly relevant in the case of self-attention: this complexity measure quantifies the amount of input inter-dependency induced by the network, directly reflecting a widespread intuition on the success behind the self-attention mechanism.

5 Depth efficiency regimes in common self-attention networks

While we proved the existence of the two different depth efficiency regimes for a simplified version of self-attention networks (described in section 2), our theoretical predictions are manifested in common self-attention networks. Kaplan et al. [2020] emphasize the depth **in**efficiency trait of self-attention [figure 1(a)], but the depth efficiency regime is clearly demonstrated in their experiments for $L < 6$ [figure 1(b)]. To show that the predicted phenomenon occurs for networks of more practical depths, we conducted a similar experiment which focuses on depths $L = 6, 12, 24$.

Specifically, we trained decoder-only (unidirectional) self-attention architectures of varying depths and widths, while optimizing the autoregressive log-likelihood. Importantly, our experiments were conducted over common self-attention architectures which include all nonlinearity and normalization operations that were omitted in our theoretical analysis. Our training set was English Wikipedia, tokenized using byte-pair encoding with a vocabulary size of 1000. Autoregressive models were shown to work well even on character level vocabularies [Peters et al., 2018]; we used a small vocabulary size so that the embedding parameters would constitute a small fraction of the learned parameters for all data points. The remainder of the training details are given in the appendix.

Figure 2 shows that the two depth efficiency/**in**efficiency regimes impact common self-attention architectures. When comparing depths $L^{\text{shallow}} = 6$ to $L^{\text{deep}} = 12$, or depths $L^{\text{shallow}} = 12$ to $L^{\text{deep}} = 24$, a qualitatively different depth efficiency behavior is observed as the network size varies. For smaller network sizes, the shallow and deep networks perform comparably. Our theoretical analysis predicts this, showing that when the width of the deeper network is not large enough it can not use its excess layers efficiently. However, when the network size in increased by widening, a clear advantage of depth is demonstrated: for the same parameter budget a deeper network performs better.

6 Discussion

An apparent "depth-**in**efficiency" of self-attention networks was pointed out by prior works [Kaplan et al., 2020] – in contrast to other successful deep learning architectures, in the case of self-attention there does not seem to be a clear advantage to deepening vs. widening. Our theoretical analysis clearly reflects this behavior in one parameter setting, but suggests an important nuance regarding its origins, while predicting a separate "depth-efficiency" regime in another parameter setting. Rather than an

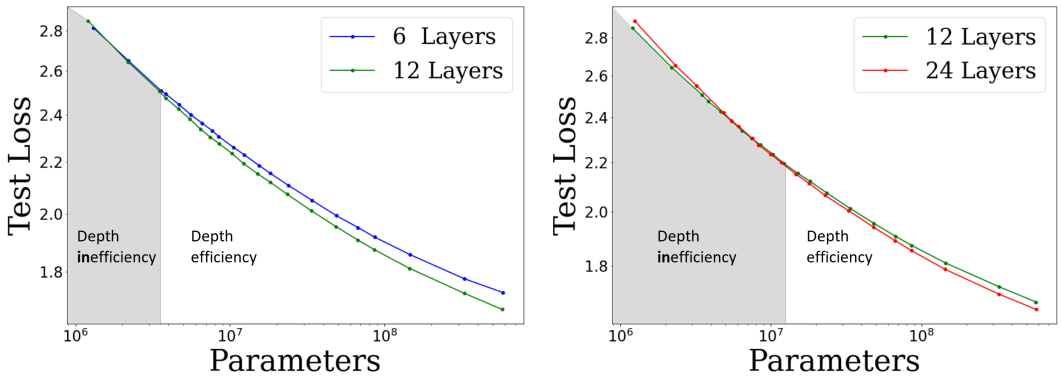

Figure 2: An experimental validation of the existence of the two depth efficiency/**in**efficiency regimes for common self-attention networks. The number of the non-embedding parameters is $12 \cdot L \cdot d_x^2$ (network widths d_x are given in the appendix). In the right plot (depths 12 and 24), the transition between regimes occurs in larger network sizes than in the left plot (depths 6 and 12), in agreement with our theory.

obvious explanation for the observed depth inefficiency, by which the self-attention mechanism does not benefit much from the operation of compounding, our analysis strongly points at the converse: self-attention is so effective at integrating its inputs, that it very quickly reaches saturation in the amount of dependencies that can be supported by the representation dimension.

Thus, for early self-attention compounding, we prove a rapid growth in expressiveness with depth, and specifically in the ability to flexibly correlate between any input locations, which can not be accounted for by any reasonable widening. However, our analysis pinpoints a transition in which the capacity of width to support the above rapid growth exhausts. Thus, when the width of a self-attention network is not large enough, the above depth efficiency disappears – deepening and widening become equivalent in terms of expressiveness.

We did not find a result which directly upper bounds depth-efficiency in other architecture classes. Works by Sharir and Shashua [2018], Levine et al. [2019] show an exponential growth with depth of a measure related to the separation rank in certain classes of convolutional networks. Comparing this with the double-exponential growth shown in theorem 1 for early self-attention layers, it may be conjectured that convolutional networks seemingly benefit more from depth than self-attention does because their separation rank grows less rapidly, so they do not saturate some width dependent threshold as quickly as self-attention does. We leave these investigations for future work.

Our analysis yields practical implications. On the one hand, the proved depth efficiency suggests always to exploit any parameter budget such that depth does not fall below a width related threshold. In this case, we have shown a clear theoretical disadvantage in the expressiveness of shallower networks, reinforced by the experiments in figure 2. On the other hand, by indicating the network width as the limiting factor for depth-efficiency, our analysis encourages the development of methods for significantly increasing network width. GPT3, the deepest self-attention network trained to date with 96 layers, has matched this depth with an unprecedented width of 12K [Brown et al., 2020]. Perhaps, given the right theoretical motivation, width can be increased even more drastically.

For example, we point at the concept of ShuffleNet [Ma et al., 2018] for increasing the representation dimension while using only a fraction of it for computation in each layer. This way, the computation costs are contained, but the theoretical limitations posed by our work are relaxed. Similarly, alternative methods for efficiently increasing the representation dimension are also supported by our analysis [Bengio et al., 2013, Shazeer et al., 2017]. Generally, width increases have greater potential for speeding up network training and inference because it can be parallelized [Shoeybi et al., 2019], as opposed to depth which yields a sequential computation. A theoretical indication that the contribution of depth and width is indeed on the same order, and that width constrains depth from contributing further, motivates the development of more extensive model parallelism methods for Transformers. Indeed, we view our work as part of an effort to provide timely theoretical interpretations as feedback for the tremendous empirical pull in our field.

Broader Impact

Our work aims at providing fundamental guidelines which can assist all fields that employ Transformer-based architectures to use more efficient models. This way, these fields can achieve their goals while consuming less resources. Additionally, this work made an effort to provide a theoretical interpretation by examining the (many) empirical signals already published by others, while providing only a required minimum of further experimentation. This was done under the belief that while experiments are crucial for the advancement of the field, it is important not to conduct them superfluously as they incur an environmental price [Schwartz et al., 2019].

Acknowledgments

We thank Daniel Jannai for assistance in the experiments, and Jared Kaplan for the permission to use the figure in Kaplan et al. [2020]. This research was supported by the ERC (European Research Council) and the ISF (Israel Science Foundation). Experiments were performed with Cloud TPUs and supported by Google's TensorFlow Research Cloud (TFRC). Yoav Levine was supported by the Israel Academy of Sciences Adams fellowship.

Arash Amini, Amin Karbasi, and Farokh Marvasti. Low-rank matrix approximation using point-wise operators. *IEEE Transactions on Information Theory*, 58(1):302–310, 2012.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Gregory Beylkin and Martin J Mohlenkamp. Numerical operator calculus in higher dimensions. *Proceedings of the National Academy of Sciences*, 99(16):10246–10251, 2002.

Gregory Beylkin, Jochen Garcke, and Martin J Mohlenkamp. Multivariate regression and machine learning with sums of separable functions. *SIAM Journal on Scientific Computing*, 31(3):1840–1857, 2009.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. *arXiv preprint arXiv:2002.07028*, 2020.

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(8): 1553–1565, 2014.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Gino Brunner, Yang Liu, Damian Pascual Ortiz, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. 2020.

Richard Caron and Tim Traynor. The zero set of a polynomial. *WSMR Report 05-02*, 2005.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=r1xMH1BtvB`.

Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. *Conference On Learning Theory (COLT)*, 2016.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

Amit Daniely. Depth separation for neural networks. *arXiv preprint arXiv:1702.08489*, 2017.

Alexandre de Brébisson and Pascal Vincent. A cheap linear attention mechanism with fast lookups and fixed-size representations. *arXiv preprint arXiv:1609.05866*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

Wolfgang Hackbusch. On the efficient evaluation of coalescence integrals in population balance models. *Computing*, 78(2):145–159, 2006.

Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer Science & Business Media, 2012.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

Robert J Harrison, George I Fann, Takeshi Yanai, and Gregory Beylkin. Multiresolution quantum chemistry in multiwavelet bases. In *Computational Science-ICCS 2003*, pages 103–110. Springer, 2003.

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.

Xiao Shi Huang, Felipe Pérez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization.

K. Inoue. Expressive numbers of two or more hidden layer relu neural networks. In *2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW)*, pages 129–135, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1357. URL `https://doi.org/10.18653/v1/n19-1357`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Yoav Levine, Or Sharir, Alon Ziv, and Amnon Shashua. Benefits of depth for long-term memory of recurrent networks. *(ICLR 2018) International Conference on Learning Representations workshop*, 2018a.

Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua. Deep learning and quantum entanglement: Fundamental connections with implications to network design. In *6th International Conference on Learning Representations (ICLR)*, 2018b.

Yoav Levine, Or Sharir, Nadav Cohen, and Amnon Shashua. Quantum entanglement in deep learning architectures. *Phys. Rev. Lett.*, 122:065301, Feb 2019. doi: 10.1103/PhysRevLett.122.065301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.122.065301`.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Ofir Press, Noah A Smith, and Omer Levy. Improving transformer models by reordering their sublayers. *arXiv preprint arXiv:1911.03864*, 2019.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Chris Dyer, Peter Dayan, and Timothy P Lillicrap. Fast parametric learning with activation memorization. *arXiv preprint arXiv:1803.10049*, 2018.

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. *arXiv preprint arXiv:1910.10683*, 2019.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org, 2017.

Oliver Richter and Roger Wattenhofer. Normalized attention without probability cage. *arXiv preprint arXiv:2005.09561*, 2020.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *arXiv preprint arXiv:1907.10597*, 2019.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. *arXiv preprint arXiv:1711.02114*, 2017.

Or Sharir and Amnon Shashua. On the expressive power of overlapping architectures of deep learning. In *6th International Conference on Learning Representations (ICLR)*, 2018.

Or Sharir, Ronen Tamari, Nadav Cohen, and Amnon Shashua. Tractable generative convolutional arithmetic circuits. 2016.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold, 04 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *ICLR*, 2020.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need`.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in neural information processing systems*, pages 550–558, 2016.

Chengwei Wang, Tengfei Zhou, Chen Chen, Tianlei Hu, and Gang Chen. Off-policy recommendation system without exploration. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 16–27. Springer, 2020.

Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019. URL `http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding`.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.