On the Expressive Flexibility of Self-Attention Matrices

Valerii Likhosherstov[1*], **Krzysztof Choromanski**[2*], **Adrian Weller**[1,3]

[1]University of Cambridge
[2]Google Brain
[3]The Alan Turing Institute
vl304@cam.ac.uk

Abstract

Transformer networks are able to capture patterns in data coming from many domains (text, images, videos, proteins, etc.) with little or no change to architecture components. We perform a theoretical analysis of the core component responsible for signal propagation between elements, i.e. the self-attention matrix. We ask the following questions: **Can a self-attention matrix approximate arbitrary patterns? How small is the query dimension d required for such approximation?** Our first result shows that the task of deciding whether approximation of a given pattern is possible or not is **NP-hard** for a fixed $d > 1$. In practice, the self-attention matrix typically exhibits two properties: it is sparse, and it changes dynamically depending on the input to the module. Motivated by this observation, we show that the self-attention matrix can provably approximate sparse matrices. While the parameters of self-attention are fixed, various sparse matrices can be approximated by only modifying the inputs. Our proof is based on the random projection technique and uses the seminal Johnson-Lindenstrauss lemma. In particular, we show that, in order to approximate any sparse matrix up to a given precision defined in terms of preserving matrix element ratios, $\boldsymbol{d}$ **grows only logarithmically with the sequence length** $\boldsymbol{n}$ (i.e. $\boldsymbol{d = O(\log n)}$).

Introduction

Transformer networks have demonstrated strong performance in the area of large-scale deep learning, coming close to or beating the state of the art in a wide range of tasks. Initially proposed in the context of neural machine translation (Vaswani et al. 2017), Transformers were found to generalize well across a variety of natural language processing tasks when pretrained on large text corpora (Devlin et al. 2019; Radford et al. 2019; Brown et al. 2020). These successes facilitated the application of Transformers in other domains. For instance, in biology, Transformers pretrained on large corpora of proteins were shown to predict proteins' structure and function (Elnaggar et al. 2019; Rives et al. 2021), and to generate protein sequences with specific properties (Madani et al. 2020). Another exciting advancement was the emergence of Vision Transformers (Dosovitskiy et al. 2021) and, later, Video Vision Transformers (Arnab et al. 2021). Thus, Transformers appear domain-agnostic and can learn any priors once a suitable large-scale dataset is provided. Finally, Transformers were recently shown to be applicable for end-to-end training on large-scale multimodal data of images with textual annotations extracted from the Internet (Radford et al. 2021; Jia et al. 2021). The resulting models are highly generalizable and perform very well in zero-shot classification from scratch, and when fine-tuned on standard benchmarks of a smaller scale.

The omnivorous nature of these models suggests that Transformers and their core component, self-attention, have an inherent ability to capture useful patterns in data regardless of the domain. A thorough analysis is required to gain a deeper understanding of this remarkable phenomenon. We take a step in this direction by analyzing the expressiveness of the self-attention module.

In self-attention, dependencies between elements of the input are propagated via a self-attention matrix, which can be thought of as an input-dependent linear projection applied to the input. By the definition, this right stochastic matrix (i.e. having nonnegative elements with rows summing up to 1) encodes input-dependent patterns in the data. Therefore, we aim to analyze the expressiveness of this matrix, to understand how flexible these input-dependent patterns can be. Importantly, we consider the setup when **the query dimension of self-attention $\boldsymbol{d}$ is much smaller than the sequence length $\boldsymbol{n}$** aiming to characterize relationships between the two. We focus on the case when $d < n$ because it is typical in practice, e.g. $d = 64, n = 512$ in BERT (Devlin et al. 2019). Also, the case $d \geq n$ can model any right stochastic matrix as shown in (Bhojanapalli et al. 2020). Finally, smaller d facilitates computational efficiency of Transformers, which are notorious for their high compute demand and CO_2 footprint (Strubell, Ganesh, and McCallum 2019).

Our first contribution shows that understanding which right stochastic matrices can be approximated by a self-attention matrix is a challenging task. Namely, we show that the decision task which accepts the right stochastic matrix and outputs whether approximation is possible or not for a given $d > 1$ is **NP-hard**. We further show that the case $d = 1$ is tractable and give an algorithm for the decision problem. Our derivations are inspired by the theory of dot product graphs (Kang and Müller 2012).

Since it is hard to tackle the general setup of approximat-

[*]These authors contributed equally.

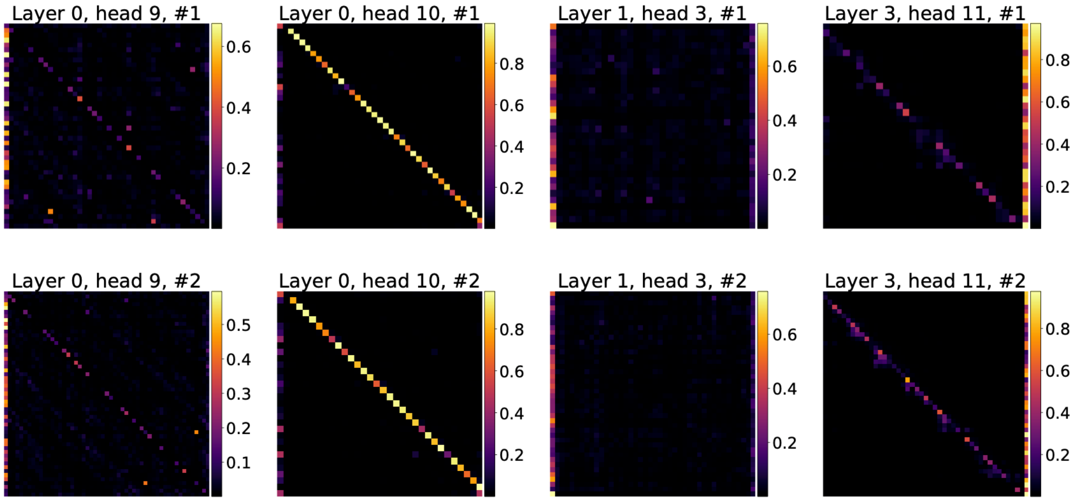

Figure 1: Examples of self-attention matrices appearing in the trained DistilBERT model (Sanh et al. 2019). Each column corresponds to a randomly chosen self-attention module defined by layer and attention head in the model. Cells in each column correspond to realisations of self-attention matrix for randomly sampled input sentences from the text corpus. As we see, self-attention matrices tend to be sparse.

ing *any* possible right stochastic matrix with self-attention, we narrow down the scope by only considering *sparse* matrices, meaning that most of the elements of the matrix are near zero. In other words, each token of the output only depends on a small number of input tokens. For instance, in neural machine translation, output words usually depend on a short context near the word they translate. This assumption typically holds in practice, as illustrated in Figure 1. The sparsity assumption, in particular, has provided insight for a series of results related to fast computation of self-attention for long sequences (Kitaev, Kaiser, and Levskaya 2020; Vyas, Katharopoulos, and Fleuret 2020; Roy et al. 2020). Consequently, two questions of interest are: **Can a self-attention module approximate arbitrary sparse patterns depending on the input? How small is the query dimension d required for such approximation?**

We make progress in addressing these questions by theoretically showing that there exist self-attention weights such that, when the precision of approximation is fixed, d **grows only logarithmically with the sequence length** n to approximate any sparse matrix by only changing the input to the module. Here, the approximation precision is defined in terms of preserving attention weight ratios, and sparsity is characterized by the bounded number of nonzero elements in each row and column. Our proof uses random projection techniques and the seminal Johnson-Lindenstrauss lemma.

We commence by defining the self-attention module and self-attention matrix. After that, we present NP-hardness results and approximation results and proceed with the proof. Finally, we present experimental simulations, discuss related work and make concluding remarks.

Prerequisites: Self-Attention Module

Let n be the length of a processed sequence and d_{hid} be the size of a hidden representation passed through the neural network. We define the *unnormalized self-attention matrix* as a parametrized mapping from the current hidden state $X \in \mathbb{R}^{n \times d_{hid}}$ into $\mathbb{R}^{n \times n}$. The mapping depends on two learnable parameter matrices $W_{\mathcal{Q}}, W_{\mathcal{K}} \in \mathbb{R}^{d_{hid} \times d}$, $d \leq d_{hid}$, where $d \leq d_{hid}$ is the *query dimension*. The mapping is defined as

$$\mathrm{USAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) = \exp[X W_{\mathcal{Q}} W_{\mathcal{K}}^\top X^\top], \quad (1)$$

where $\exp[\cdot]$ is an elementwise exponent. Next, we define the *(normalized) self-attention matrix* as

$$\mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) = \mathrm{RN}(\mathrm{USAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})). \quad (2)$$

Here, $\mathrm{RN}(\cdot)$ (*row normalization*) divides each row by the sum of its elements, so that the result is *right stochastic* meaning that all rows are nonnegative and sum to 1.

Finally, *self-attention* is defined as a parametrized mapping from X into $\mathbb{R}^{n \times d}$ with parameters $W_{\mathcal{Q}}, W_{\mathcal{K}}, W_{\mathcal{V}} \in \mathbb{R}^{d_{hid} \times d}$. It has the form:

$$\mathrm{SA}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}, W_{\mathcal{V}}) = \mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) X W_{\mathcal{V}}. \quad (3)$$

Self-attention has the form of a differentiable dictionary, where the output at each position $1 \leq i \leq n$ is a sum of all values $W_{\mathcal{V}}^\top X_{i'}$, $1 \leq i' \leq n$, weighted proportionally to exponentiated dot products of the query $W_{\mathcal{Q}}^\top X_i$ and the key vectors $W_{\mathcal{V}}^\top X_{i'}$. Usually (Vaswani et al. 2017), these dot product are also divided by $\sqrt{d}$, since this empirically facilitates stable training. Without loss of generality, we do not include this linear scaling factor in our definition (1), since it can be fused into one of the matrices $W_{\mathcal{Q}}$ or $W_{\mathcal{K}}$.

NP-Hardness of Approximation

Below we define a notion of $\mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ *weakly approximating* a right stochastic matrix $A = \mathrm{RN}(S)$, where $S \in \{0,1\}^{n \times n}$ is a matrix of zeros and ones with at least one 1 in every row so that its row normalization is well-defined. The goal of this notion is to capture a minimal set of conditions which make it possible to say that $\mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ approximates A:

Definition 0.1. Suppose $S \in \{0,1\}^{n \times n}$ has at least one 1 in every row and $X \in \mathbb{R}^{n \times d_{hid}}$, $W_{\mathcal{Q}}, W_{\mathcal{K}} \in \mathbb{R}^{d_{hid} \times d}$, $d_{hid} \geq d$, are some matrices. $M = \mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ weakly approximates $A = \mathrm{RN}(S)$ when

$$\forall 1 \leq i \leq n : \min_{j, A_{i,j} > 0} M_{i,j} > \max_{j, A_{i,j} = 0} M_{i,j}, \tag{4}$$

where we assume that $\max$ over an empty set is $-\infty$. If such $X, W_{\mathcal{Q}}, W_{\mathcal{K}}$ exist for A, then we say that A can be (d, d_{hid})-weakly SAM-approximated.

According to this definition, M approximates A if in each row, the values of M corresponding to zero positions in A are smaller than values corresponding to nonzero positions in A. For integers $d_{hid} \geq d \geq 1$, by (d, d_{hid})-*WEAK-SAM-APPROX* we denote the algorithmic problem of deciding, given an integer $n > 0$ and a matrix $S \in \{0,1\}^{n \times n}$ vertices with at least one 1 in each row, whether $\mathrm{RN}(S)$ can be (d, d_{hid})-weakly SAM-approximated or not. Consequently, (d, d_{hid})-WEAK-SAM-APPROX returns either True if the approximation is possible or False otherwise. Our first contribution is as follows:

Theorem 0.2. *(d, d_{hid})-WEAK-SAM-APPROX is NP-hard for all $d_{hid} \geq d > 1$.*

Proof. See Appendix. □

Theorem 0.2 states that the approximation task with a small d (since the size of the graph n can be arbitrarily large in (d, d_{hid})-WEAK-SAM-APPROX) is very hard and possibly unsolvable in polynomial time. This supports our initial claim that the problem of right stochastic matrix approximation by self-attention is a challenging task. The proof is inspired by a similar result in the theory of dot product graphs (Kang and Müller 2012).

For completeness, we also show that the case of $(1, d_{hid})$-WEAK-SAM-APPROX, $d_{hid} \geq 1$, has a positive characterization:

Theorem 0.3. *Algorithm 1 solves $(1, d_{hid})$-WEAK-SAM-APPROX in the time which is polynomial of the input's size.*

Proof. See Appendix. □

Since the general case of right stochastic matrix approximation by self-attention is very hard and potentially unsolvable, in the next section we focus on a special case of sparse matrices. For this class of matrices, we derive a lower bound on d which gives a provable approximation.

Algorithm 1: Algorithm for solving $(1, d_{hid})$-WEAK-SAM-APPROX ($\mathcal{V}$ denotes $\{1, \ldots, n\}$.). **Input:** $S \in \{0,1\}^{n \times n}$ with at least one 1 in each row. **Output:** True or False.

```
For $1 \leq i \leq n$, set $I^{(i)} = \{1 \leq j \leq n \,|\, M_{i,j} = 1\}$.
Set $K_0 = \{1 \leq i \leq n \,|\, I^{(i)} \neq \{1, \ldots, n\}\}$. If $K_0 = \varnothing$, return True.
Let $H$ be a graph on $K_0$ as vertices, such that there is an edge between $v, w$ if $I^{(v)} \subseteq I^{(w)}$ or $I^{(w)} \subseteq I^{(v)}$.
If $H$ doesn't consist of two (possibly empty) connected components $K_1, K_2 \subseteq K_0$, return False.
For $1 \leq i \leq n$, set $I'^{(i)} = I^{(i)}$ if $i \notin K_2$ and $I'^{(i)} = \mathcal{V} \setminus I^{(i)}$ if $i \in K_2$.
If there exist $i, j \subseteq K_0$ such that neither $I'^{(i)} \subseteq I'^{(j)}$, nor $I'^{(j)} \subseteq I'^{(i)}$, return False.
If $d_{hid} > 1$, return True.
Let $i_1, \ldots, i_n$ be an ordering of $1, \ldots, n$ such that $I'^{(i_1)} \subseteq \cdots \subseteq I'^{(i_n)}$.
If $K_2 \subseteq I'^{(i_1)}$, or $I'^{(i_n)} \subseteq K_2$ or there is $1 \leq r < n$ such that $I'^{(i_r)} \subseteq K_2 \subseteq I'^{(i_{r+1})}$, return True.
If $K_2 \subseteq \mathcal{V} \setminus I'^{(i_n)}$, or $\mathcal{V} \setminus I'^{(i_1)} \subseteq K_2$, or there is $1 \leq r < n$ such that $\mathcal{V} \setminus I'^{(i_{r+1})} \subseteq K_2 \subseteq \mathcal{V} \setminus I'^{(i_r)}$, return True.
Return False.
```

Approximating Sparse Matrices by Self-Attention Matrix

The Main Result

We will call the square matrix *k-nonzero-bounded* if for each row or column of the matrix, the total number of nonzero elements is no more than k.

Apart from the notion of the bounded number of nonzero elements, we also define matrices with elements of a bounded variation. For $\gamma \geq 1$, we call the matrix $A \in \mathbb{R}^{n \times n}$ with nonnegative elements *γ-variation-bounded*, if for every row $1 \leq i \leq n$ and every column indices $1 \leq j_1, j_2 \leq L$ such that $A_{i,j_1}, A_{i,j_2} \neq 0$,

$$\gamma^{-1} \leq \frac{A_{i,j_1}}{A_{i,j_2}} \leq \gamma. \tag{5}$$

For instance, all nonzero entries of a 1-variation-bounded matrix are the same for each row of the matrix.

The following theorem is the main result about approximation of sparse matrices by self-attention:

Theorem 0.4. *Let $n > 1, k, d_{hid}, d \leq \min(d_{hid}, 2n)$ be natural numbers, $0 < \epsilon_1 < 1$, $\epsilon_2 > 0$, $\gamma \geq 1$ be real numbers,*

$$d = 2\left\lceil 16k^2 \left(\frac{\log \gamma - \log \epsilon_1}{\epsilon_2} + 1 \right)^2 \right.$$
$$\left. \times (2 \log n + \log(n-1) + \log 2) \right\rceil. \tag{6}$$

Then there exist $W_{\mathcal{Q}}, W_{\mathcal{K}} \in \mathbb{R}^{d_{hid} \times d}$, such that for any right stochastic, k-nonzero-bounded, γ-variation-bounded matrix $A \in \mathbb{R}^{n \times n}$, there is $X \in \mathbb{R}^{n \times d_{hid}}$ and $M = \mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ satisfying

1. *For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq n$ such that $A_{i,j_1} = 0, A_{i,j_2} \neq 0$, it holds that*

$$\frac{M_{i,j_1}}{M_{i,j_2}} < \epsilon_1; \tag{7}$$

2. *For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq n$ such that $A_{i,j_1} \neq 0, A_{i,j_2} \neq 0$, it holds that*

$$\frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(-\epsilon_2) < \frac{M_{i,j_1}}{M_{i,j_2}} < \frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(\epsilon_2). \qquad (8)$$

$W_{\mathcal{Q}}, W_{\mathcal{K}}$ can be constructed in $O(d_{hid} \cdot d)$ time. For any A, X and M can be computed in randomized time polynomial in n, d_{hid}, k.

Informally, Theorem 0.4 states that for hidden sizes d_{hid}, d **growing only logarithmically** with the sequence length n when the sparsity and variability parameters (k, γ) are fixed, there exist **fixed** parameter matrices $W_{\mathcal{Q}}, W_{\mathcal{K}}$ such that for any nonzero-bounded matrix A there is a self-attention input X such that $M = \mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ approximates A very well. The quality of approximation is characterized by upper and lower bounds on ratios of elements located in the same row of M:

1. Equation (23) means that zero elements of A are approximated by elements of M which are small compared to nonzero elements of the same row when ϵ_1 is chosen small. By definition M is a strictly positive matrix, therefore in principle we can only approximate zero elements of A by very small positive numbers.
2. Equation (24) means that ratios of nonzero elements of the same row in M are in a close multiplicative neighborhood of the corresponding ratios in A when ϵ_2 is chosen small. Since rows of both matrices A and M sum up to 1, similar enough ratios of element pairs also imply element similarity in terms of their absolute magnitude.

Remark 0.5. In the statement of Theorem 0.4, M weakly approximates A for any $0 < \epsilon_1 < 1$, $\epsilon_2 > 0$ which can be checked by Definition 0.1 (Equation 23).

Finally, as the proof is constructive, we will obtain an algorithm for computing $W_{\mathcal{Q}}, W_{\mathcal{K}}$, which turn out to be matrices of a simple structure. For any A from the theorem statement, the probabilistic algorithm induced by the proof enables X and M to be computed in randomized polynomial time in n, d_{hid}, k.

In the rest of the section we describe the detailed proof and intuition behind it.

Proof of Theorem 0.4: Matrix B and the Intuition Behind the Proof

Define vector $\alpha \in \mathbb{R}^n$ so that for each row index $1 \leq i \leq n$, the minimal nonzero element in this row is α_i. Define matrix $B \in \mathbb{R}^{n \times n}$ as follows. For all $1 \leq i, j \leq n$,

$$B_{i,j} = \begin{cases} 0 & \text{if } A_{i,j} = 0; \\ \log(A_{i,j}/(\epsilon_1 \alpha_i)) + \epsilon_2 & \text{otherwise.} \end{cases} \qquad (9)$$

Observe, that $C = \epsilon_1 \exp(-\epsilon_2) \mathrm{diag}(\alpha) \exp(B)$ can be thought of as an approximation of A:

$$C = \epsilon_1 \exp(-\epsilon_2) \mathrm{diag}(\alpha) \exp(B) \approx A. \qquad (10)$$

Indeed, for any $1 \leq i, j \leq n$ such that $A_{i,j} \neq 0, C_{i,j} = A_{i,j}$ by definition of B and C (Equations 9, 10). On the other hand, when $A_{i,j} = 0$, $C_{i,j} = \epsilon_1 \exp(-\epsilon_2) \alpha_i \leq \epsilon_1$, where we use $\epsilon_2 > 0$ and $\alpha_i \leq 1$, since A is a right stochastic matrix. Hence, smaller ϵ_1 yields a better approximation (10).

Suppose we find $X, W_{\mathcal{Q}}, W_{\mathcal{K}}$ such that

$$B \approx X W_{\mathcal{Q}} W_{\mathcal{K}}^\top X^\top. \qquad (11)$$

Intuitively, if the approximation (11) is sufficiently good then

$$\begin{aligned} \mathrm{USAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) &= \exp(X W_{\mathcal{Q}} W_{\mathcal{K}}^\top X^\top) \\ &\approx \exp(B) = \epsilon_1^{-1} \exp(\epsilon_2) \mathrm{diag}(\alpha)^{-1} C \\ &\approx \epsilon_1^{-1} \exp(\epsilon_2) \mathrm{diag}(\alpha)^{-1} A, \end{aligned} \qquad (12)$$

where in the last transition we use (10). Since the unnormalized self-attention matrix $\mathrm{USAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ is a good approximation for A with rescaled rows (recall $\epsilon_1^{-1} \exp(\epsilon_2) \mathrm{diag}(\alpha)^{-1}$ multipliers), the normalized self-attention matrix $\mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ should be a good approximation for A, which is itself row-normalized (right stochastic):

$$\mathrm{SAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) \approx A. \qquad (13)$$

Next, we formally construct such $X, W_{\mathcal{Q}}, W_{\mathcal{K}}$ and derive tight error bounds for the approximation (13) in terms of matrix element ratios (23,24).

Proof of Theorem 0.4: Construction of $X, W_{\mathcal{Q}}, W_{\mathcal{K}}$ through Random Projections

Consider a singular value decomposition (SVD, Trefethen and Bau 1997) of the matrix B: $B = U \Sigma V^\top$, where $U, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \ldots \sigma_L \geq 0$ are singular values of B. Define $D = U \Sigma$, then B can be decomposed as $B = D V^\top$.

We will use random projections to compress D and V into matrices of shape $n \times d/2$ (d is even according to (22)). Namely, let $Y \in \mathbb{R}^{n \times d/2}$ be a random matrix sampled from a uniform distribution (Haar measure) on a set of Stiefel matrices[1] $\{\Omega \in \mathbb{R}^{n \times d/2} | \Omega^\top \Omega = I_{d/2}\}$. Here, $I_{d/2}$ is a $(d/2) \times (d/2)$ identity matrix. Then we set $X^{(1)} = (2n/d)^{1/2} D Y \in \mathbb{R}^{n \times d/2}$, $X^{(2)} = (2n/d)^{1/2} V Y \in \mathbb{R}^{n \times d/2}$. $X^{(1)}, X^{(2)}$ can be considered as compressions of D, V since $X^{(1)} X^{(2)\top}$ is an unbiased approximation of $B = D V^\top$:

$$\begin{aligned} \mathbb{E} X^{(1)} X^{(2)\top} &= D \times \mathbb{E}\left[(2n/d) \cdot Y Y^\top\right] \times V^\top \\ &= D \times \mathbb{E}\left[n \cdot Y_{:,1} Y_{:,1}^\top\right] \times V^\top = D V^\top = B, \end{aligned} \qquad (14)$$

where we use the fact that columns of Y are marginally uniformly distributed on $\mathcal{S}^{n-1}$. See Figure 2a for an illustration. We set $X, W_{\mathcal{Q}}, W_{\mathcal{K}}$ as

$$X = \left[X^{(1)} \quad X^{(2)} \quad \mathbf{0}_{L \times (d_{hid} - d)}\right], \qquad (15)$$

$$W_{\mathcal{Q}} = \left[I_d \ \ \mathbf{0}_{d \times (d_{hid} - d)}\right]^\top, \qquad (16)$$

[1] While Y can be defined as a matrix with i.i.d. sub-Gaussian entries (Kaban 2015), in general, orthogonal projections outperform unstructured ones in theory and practice (Choromanski, Rowland, and Weller 2017; Choromanski et al. 2021; Lin et al. 2020). We also manage to obtain better dot product concentration results for Stiefel projections compared to unstructured ones (see discussion in Appendix).

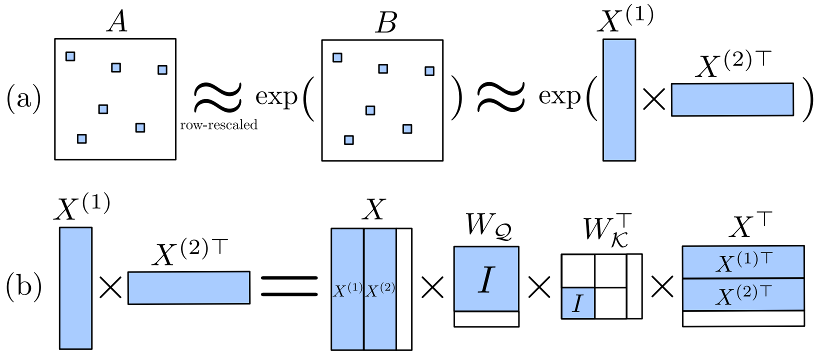

Figure 2: Illustration of the approximation scheme. **(a)** $\exp(B)$ is a row-rescaled approximation of A (Equation 10), whereas $X^{(1)}X^{(2)\top}$ is an unbiased approximation to B (14). **(b)** Representation of $X^{(1)}X^{(2)\top}$ as $XW_{\mathcal{Q}}W_{\mathcal{K}}X^{\top}$ according to (15,16,17).

$$W_{\mathcal{K}} = \left[\begin{bmatrix}\mathbf{0}_{d/2\times d/2} & I_{d/2} \\ \mathbf{0}_{d/2\times d/2} & \mathbf{0}_{d/2\times d/2}\end{bmatrix} \;\; \mathbf{0}_{d\times(d_{hid}-d)}\right]^{\top}, \qquad (17)$$

where $\mathbf{0}_{\ldots\times\ldots}$ denotes a zero matrix of the corresponding shape. It is easy to see that in this case $XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top} = X^{(1)}X^{(2)\top}$ (see Figure 2b).

Our next step is to prove that with a nonzero probability, differences of elements in $XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top}$ concentrate near the same differences in B:

Lemma 0.6. *With probability greater than $(n+2)^{-1}$ it holds that*

$$\forall 1 \le i, j_1, j_2 \le n, j_1 \ne j_2 : |(XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_1} - (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_2} - B_{i,j_1} + B_{i,j_2}| < \epsilon_2. \qquad (18)$$

The proof (in Appendix) uses a corollary of the seminal Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984) about inner product preservation under random projections (Kaban 2015). Two crucial observations are that

- $W_{\mathcal{Q}}$ and $W_{\mathcal{K}}$ do not depend on A by construction (16,17);
- according to Lemma 0.6, X satisfying (18) can be found with any probability by redrawing Y $O(n)$ times.

Suppose that (15,16,17) hold. Then for any $1 \le i, j_1, j_2 \le n$, $j_1 \ne j_2$:

$$B_{i,j_1} - B_{i,j_2} - \epsilon_2 < (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_1} - (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_2} < B_{i,j_1} - B_{i,j_2} + \epsilon_2. \qquad (19)$$

By definition of B (9), whenever $A_{i,j_1} = 0, A_{i,j_2} \ne 0$, the right hand side inequality in (19) is rewritten as

$$(XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_1} - (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_2} < -\log A_{i,j_2} + \log\alpha_i + \log\epsilon_1 \le \log\epsilon_1. \qquad (20)$$

Here we also used $A_{i,j_2} \ge \alpha_i$. (20) is equivalent to (23) after exponentiating, since exponents of $(XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_1}$ and $(XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_2}$ are M_{i,j_1} and M_{i,j_2} rescaled by the same factor.

Similarly to (20), whenever $A_{i,j_1}, A_{i,j_2} \ne 0$, by expanding B's definition, (19) is rewritten as

$$\log A_{i,j_1} - \log A_{i,j_2} - \epsilon_2 < (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_1} - (XW_{\mathcal{Q}}W_{\mathcal{K}}^{\top}X^{\top})_{i,j_2} < \log A_{i,j_1} - \log A_{i,j_2} + \epsilon_2, \qquad (21)$$

which is equivalent to (24) after exponentiating. This concludes the proof of Theorem 0.4. □

The Case of Causal Self-attention

Another popular type of self-attention mechanism is *causal self-attention*, when each position i only attends to elements $j \le i$. This modification is required for autoregressive language modelling (Radford et al. 2019; Brown et al. 2020) when each token is modelled as depending only on previous tokens in the sequence. We define the causal self-attenion matrix CSAM and causal self-attention CSA as

$$\begin{gathered}\mathrm{CSAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}) = \mathrm{diag}(\mathcal{M}'\mathbf{1}_n)^{-1}\mathcal{M}', \\ \mathcal{M}' = \mathrm{tril}(\mathrm{USAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})), \\ \mathrm{CSA}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}}, W_{\mathcal{V}}) = \\ \mathrm{CSAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})XW_{\mathcal{V}},\end{gathered}$$

where $\mathrm{tril}(\cdot)$ is the *lower triangular part* of the argument matrix, meaning that it zeroes out all elements strictly above the main diagonal.

A natural question is whether the analog of Theorem 0.4 holds for causal self-attention matrices. Since these matrices are lower triangular, we should only attempt to approximate lower-triangular right-stochastic matrices A. In fact, we obtain the following result (differences with Theorem 0.4 are highlighted in bold).

Corollary 0.7. *Let $n > 1, k, d_{hid}, d \le \min(d_{hid}, 2n)$ be natural numbers, $0 < \epsilon_1 < 1$, $\epsilon_2 > 0$, $\gamma \ge 1$ be real numbers,*

$$d = 2\left\lceil 16k^2\left(\frac{\log\gamma - \log\epsilon_1}{\epsilon_2} + 1\right)^2 \times (2\log n + \log(n-1) + \log 2)\right\rceil. \qquad (22)$$

*Then there exist $W_{\mathcal{Q}}, W_{\mathcal{K}} \in \mathbb{R}^{d_{hid}\times d}$, such that for any **lower triangular**, right stochastic, k-nonzero-bounded, γ-variation-bounded matrix $A \in \mathbb{R}^{n\times n}$, there is $X \in \mathbb{R}^{n\times d_{hid}}$ and $M = \mathbf{CSAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ satisfying*

1. *For all row indices $1 \le i \le n$ and column indices $\boldsymbol{1 \le j_1, j_2 \le i}$ such that $A_{i,j_1} = 0, A_{i,j_2} \ne 0$, it holds*

that

$$\frac{M_{i,j_1}}{M_{i,j_2}} < \epsilon_1; \tag{23}$$

2. *For all row indices $1 \leq i \leq n$ and column indices* $\boldsymbol{1 \leq j_1, j_2 \leq i}$ *such that $A_{i,j_1} \neq 0, A_{i,j_2} \neq 0$, it holds that*

$$\frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(-\epsilon_2) < \frac{M_{i,j_1}}{M_{i,j_2}} < \frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(\epsilon_2). \tag{24}$$

$W_{\mathcal{Q}}, W_{\mathcal{K}}$ can be constructed in $O(d_{hid} \cdot d)$ time. For any A, X and M can be computed in randomized time polynomial in n, d_{hid}, k.

Proof. The proof is unchanged compared to the proof of Theorem 0.4 with the only change that j_1, j_2 are considered in the range $1 \leq j_1, j_2 \leq i$ when computing difference bounds (20,21). Cases when column indices j_1 or j_2 are bigger than the row index i are redundant, since both A and $M = \mathrm{CSAM}(X; d, d_{hid}, W_{\mathcal{Q}}, W_{\mathcal{K}})$ have zero entries above the main diagonal. □

We conclude that the same logarithmic dependence $d = O(\log n)$ holds for causal self-attention.

Experiments

Theorem 0.4 suggests an upper bound (r.h.s. in Equation 22) for the $d_{\min}(\epsilon_1, \epsilon_2)$ – i.e. the minimal d which results in M satisfying (23,24) for fixed ϵ_1, ϵ_2. A question which we address in the experimental section is, therefore, "**What is the actual $\boldsymbol{d_{\min}(\epsilon_1, \epsilon_2)}$ in practice? Does it satisfy the logarithmic law $\boldsymbol{d = O(\log n)}$**?"

To answer this question, we perform the following simulation. We select a range of $(k, \gamma, \epsilon_1, \epsilon_2)$ parameters. For each set of parameters, we iterate over n on a uniform grid from 512 to 3072 with a step size 256. For each n we sample the matrix A and iterate over a uniform grid of d values in ascending order until we find such d which results in M satisfying (23,24). We sample nonzero positions of A by taking a union over k random permutation matrices. The nonzero value is set to either 1 or γ by a coin flip.

To check whether for the current d there is M satisfying (23,24), we construct Y, $X^{(1)}$, $X^{(2)}$ and M using the algorithm implied by the proof of Theorem 0.4. To sample Stiefel matrices Y, we use the algorithm based on QR decompositions of random Gaussian matrices from (Stewart 1980). We redraw the Y matrix Qn times, $Q = 1$, in the spirit of Lemma 0.6 suggesting that $O(n)$ redraws should be enough to find the right $Y, X^{(1)}, X^{(2)}, X$ with a constant probability (when d is big enough).

Figure 3 illustrates the results. A remarkable observation is that, although empirical $d_{\min}(\epsilon_1, \epsilon_2)$ (red circles) grows slower than the theoretical upper bound (blue dashed line shows the angle of this line), it nicely fits the logarithmic curve $d = O(\log n)$ (black dotted line) in all twelve evaluated setups. The fact that the true $d_{\min}(\epsilon_1, \epsilon_2)$ grows slower than (22) is natural, since (22) is an upper bound on it. Though, as experiments reveal, both curves differ only by a constant multiplicative factor.

We run an additional experiment to reveal how $d_{\min}(\epsilon_1, \epsilon_2)$ depends on the number of samples Qn used to find M satisfying (23,24). We take 2 setups and try a range of Q values from 0.1 to 5.0. Results are shown in Figure 4. We observe that $d_{\min}(\epsilon_1, \epsilon_2)$ does not depend a lot on the choice of Q and is roughly unchanged. Therefore, we conclude that our findings regarding the behaviour of empirical $d_{\min}(\epsilon_1, \epsilon_2)$ do not depend on Q much and $Q = 1$ is a reasonable choice. Additional experimental details and results (plots for bigger k values) can be found in Appendix.

Related Work

Expressivity of Transformers. As Transformers gained popularity, more theoretical results have emerged to explain their expressivity. Transformers were shown to be universal approximators (Yun et al. 2020a), Turing-complete (Bhattamishra, Patel, and Goyal 2020) and able to recognize counter languages (Bhattamishra, Ahuja, and Goyal 2020). Furthermore, Transformer modifications such as BigBird (Zaheer et al. 2020), Transformers with hard attention (Pérez, Marinković, and Barceló 2019) and sparse Transformers (Yun et al. 2020b) were shown to be universal approximators. Note that (Yun et al. 2020a; Bhattamishra, Patel, and Goyal 2020) rely on multilayer constructions, whereas we consider a single self-attention module, and (Zaheer et al. 2020; Pérez, Marinković, and Barceló 2019; Yun et al. 2020b) analyze nonconventional forms of self-attention. Dong, Cordonnier, and Loukas (2021) analyze limitations of a pure self-attention Transformer, i.e. without feedforward blocks and skip connections. Cordonnier, Loukas, and Jaggi (2020) show that multi-head self-attention can provably model any image convolution layer. Bhojanapalli et al. (2020) show that for *large* d ($d \geq n$) and fixed inputs, there exist $W_{\mathcal{Q}}, W_{\mathcal{K}}$ which approximate any positive right stochastic matrix via self-attention. In contrast, we analyze expressivity when d is very small ($d = O(\log n)$).

Random projections and Johnson-Lindenstrauss lemma. Our proof techniques rely on the seminal Johnson-Lindenstrauss tranformation (JLT, Johnson and Lindenstrauss 1984) used for dimensionality reduction (Fedoruk et al. 2018). A random projection approach similar to ours was used by Frankl and Maehara (1987) to lower-bound graph sphericity – a characteristic which is NP-hard to compute in general. A related *random features* technique, relying on random projections, was originally introduced to improve efficiency of kernel SVMs (Rahimi and Recht 2008), but recently found application in speeding up long-sequence Transformers (Choromanski et al. 2020, 2021). We use Stiefel matrices as random projections, which in general result in tighter approximations than unconstrained projections (Lin et al. 2020). Ensembles of orthogonal random projections were shown to provide much better concentration results for the estimators relying on them in various other contexts, in particular: kernel approximation (Choromanski and Sindhwani 2016; JLT can be considered a special instantiation with a dot-product kernel), estimation of the gradients of Gaussian smoothings with evolution strategy methods (Choromanski et al. 2018), kernel ridge

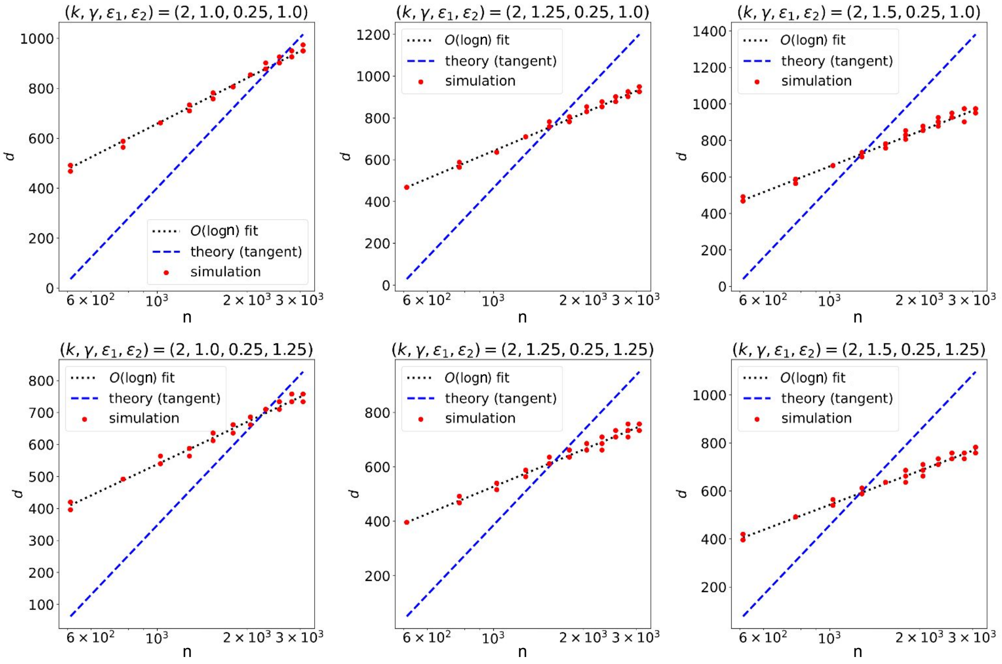

Figure 3: Finding empirical dependency of $d_{\min}(\epsilon_1, \epsilon_2)$ on n given the fixed set of parameters $(k, \gamma, \epsilon_1, \epsilon_2)$. Each plot corresponds to one out of six sets of tested parameters. Red circles correspond to simulation results: we redraw matrix A 5 times for each n, resulting in 5 red circles per n. The horizontal n axis is in a logarithmic scale, so that $O(\log n)$ corresponds to a straight line. The black dotted line corresponds to a $O(\log n)$ fit for the dots (linear when x axis scale is logarithmic). The blue dashed line indicates the *slope* of the theoretical upper bound on $d_{\min}(\epsilon_1, \epsilon_2)$ (right hand side in Equation 22). We experiment with k from $\{1, 2\}$ (results for bigger values of k can be found in Appendix). The *slope* of the observations is lower than for the theoretical blue line, confirming our theoretical result.

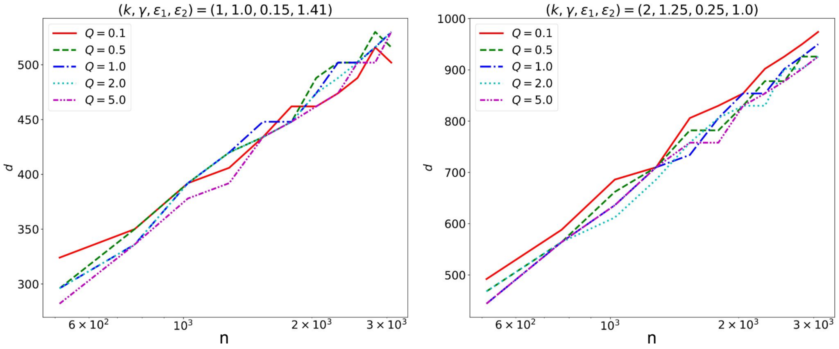

Figure 4: Empirical dependence of $d_{\min}(\epsilon_1, \epsilon_2)$ on Q – the factor defining the number of samples Qn (rounded to an integer) used to find the right matrix M. For each Q we repeat the procedure to generate empirical values of $d_{\min}(\epsilon_1, \epsilon_2)$ (red circles from Figure 3) and connect them into a line for the better visualization.

regression (Choromanski, Downey, and Boots 2018), sliced Wasserstein distance (Rowland et al. 2019) and more.

Conclusion

We have analyzed the expressive flexibility of the self-attention matrix as a mechanism to approximate sparse patterns. First, we show that even deciding whether the approximation is possible or not for a given pattern is NP-hard in general. Then we prove that weights of self-attention can be constructed in such a way that any sparse matrix can be approximated with certain input to the self-attention module. We show that, when error and other parameters are fixed, d grows only logarithmically with the sequence length n, i.e. $d = O(\log n)$ when other matrix parameters are fixed. We hope our work will facilitate further in-depth theoretical analysis of self-attention and Transformers to understand better their remarkable performance across a variety of tasks.

Acknowledgements

V. L. acknowledges support from the Cambridge Trust and DeepMind. V. L. was part-time employed by Google while a PhD student. A.W. acknowledges support from a Turing AI Fellowship under EPSRC grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; and Schmid, C. 2021. ViViT: A Video Vision Transformer. *CoRR*, abs/2103.15691.

Bhattamishra, S.; Ahuja, K.; and Goyal, N. 2020. On the Ability of Self-Attention Networks to Recognize Counter Languages. *CoRR*, abs/2009.11264.

Bhattamishra, S.; Patel, A.; and Goyal, N. 2020. On the Computational Power of Transformers and Its Implications in Sequence Modeling. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 455–475. Online: Association for Computational Linguistics.

Bhojanapalli, S.; Yun, C.; Rawat, A. S.; Reddi, S.; and Kumar, S. 2020. Low-Rank Bottleneck in Multi-head Attention Models. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 864–873. PMLR.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.

Choromanski, K.; Downey, C.; and Boots, B. 2018. Initialization matters: Orthogonal Predictive State Recurrent Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Belanger, D.; Colwell, L.; and Weller, A. 2020. Masked language modeling for proteins via linearly scalable long-context transformers. *arXiv preprint arXiv:2006.03555*.

Choromanski, K.; Rowland, M.; Sindhwani, V.; Turner, R. E.; and Weller, A. 2018. Structured Evolution with Compact Architectures for Scalable Policy Optimization. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 969–977. PMLR.

Choromanski, K.; and Sindhwani, V. 2016. Recycling Randomness with Structure for Sublinear time Kernel Expansions. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 2502–2510. JMLR.org.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; Belanger, D. B.; Colwell, L. J.; and Weller, A. 2021. Rethinking Attention with Performers. In *International Conference on Learning Representations*.

Choromanski, K. M.; Rowland, M.; and Weller, A. 2017. The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 219–228.

Cordonnier, J.-B.; Loukas, A.; and Jaggi, M. 2020. On the Relationship between Self-Attention and Convolutional Layers. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dong, Y.; Cordonnier, J.; and Loukas, A. 2021. Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth. *CoRR*, abs/2103.03404.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; and Rost, B. 2019. End-to-end multitask learning, from protein language to protein features without alignments. *bioRxiv*.

Fedoruk, J.; Schmuland, B.; Johnson, J.; and Heo, G. 2018. Dimensionality Reduction via the Johnson—Lindenstrauss Lemma: Theoretical and Empirical Bounds on Embedding Dimension. *J. Supercomput.*, 74(8): 3933–3949.

Frankl, P.; and Maehara, H. 1987. The Johnson-Lindenstrauss Lemma and the Sphericity of Some Graphs. *J. Comb. Theory Ser. A*, 44(3): 355–362.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *CoRR*, abs/2102.05918.

Johnson, W.; and Lindenstrauss, J. 1984. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26: 189–206.

Kaban, A. 2015. Improved Bounds on the Dot Product under Random Projection and Random Sign Projection. In *Pro-*

ceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, 487–496. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Kang, R.; and Müller, T. 2012. Sphere and Dot Product Representations of Graphs. *Discrete Computational Geometry*, 47: 548–568.

Kitaev, N.; Kaiser, L.; and Levskaya, A. 2020. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lin, H.; Chen, H.; Choromanski, K. M.; Zhang, T.; and Laroche, C. 2020. Demystifying Orthogonal Monte Carlo and Beyond. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Madani, A.; McCann, B.; Naik, N.; Keskar, N. S.; Anand, N.; Eguchi, R. R.; Huang, P.-S.; and Socher, R. 2020. ProGen: Language Modeling for Protein Generation. arXiv:2004.03497.

Pérez, J.; Marinković, J.; and Barceló, P. 2019. On the Turing Completeness of Modern Neural Network Architectures. In *International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8): 9.

Rahimi, A.; and Recht, B. 2008. Random Features for Large-Scale Kernel Machines. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; and Fergus, R. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).

Rowland, M.; Hron, J.; Tang, Y.; Choromanski, K.; Sarlós, T.; and Weller, A. 2019. Orthogonal Estimation of Wasserstein Distances. In Chaudhuri, K.; and Sugiyama, M., eds., *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, 186–195. PMLR.

Roy, A.; Saffar, M.; Vaswani, A.; and Grangier, D. 2020. Efficient Content-Based Sparse Attention with Routing Transformers. *arXiv*, 2003.05997.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Stewart, G. W. 1980. The Efficient Generation of Random Orthogonal Matrices with an Application to Condition Estimators. *SIAM Journal on Numerical Analysis*, 17(3): 403–409.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. *CoRR*, abs/1906.02243.

Trefethen, L.; and Bau, D. 1997. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics. ISBN 9780898713619.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vyas, A.; Katharopoulos, A.; and Fleuret, F. 2020. Fast Transformers with Clustered Attention. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21665–21674. Curran Associates, Inc.

Yun, C.; Bhojanapalli, S.; Rawat, A. S.; Reddi, S.; and Kumar, S. 2020a. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*.

Yun, C.; Chang, Y.-W.; Bhojanapalli, S.; Rawat, A. S.; Reddi, S.; and Kumar, S. 2020b. O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13783–13794. Curran Associates, Inc.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; and Ahmed, A. 2020. Big Bird: Transformers for Longer Sequences. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 17283–17297. Curran Associates, Inc.