ONLY LARGE WEIGHTS (AND NOT SKIP CONNECTIONS) CAN PREVENT THE PERILS OF RANK COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention mechanisms lie at the heart of modern large language models (LLMs). Straightforward algorithms for forward and backward (gradient) computation take quadratic time, and a line of work initiated by [Alman and Song NeurIPS 2023] and [Alman and Song NeurIPS 2024] has shown that quadratic time is necessary unless the model weights are small, in which case almost linear time algorithms are possible. In this paper, we show that large weights are necessary to avoid a strong preclusion to representational strength we call layer collapse, which means that the entire network can be approximated well by a network with only a single layer. This means that transformers with small weights are shockingly weak, and that the quadratic running time of attention is unavoidable for expressive transformers.

The notion of layer collapse that we introduce is a variant on the notion of rank collapse from the work of [Dong, Cordonnier, and Loukas ICML 2021]. They showed that in Self Attention Networks with small weights and with skip connections, rank collapse must occur. This is typically interpreted as justifying the necessity of skip connections in expressive networks. However, our result shows that even with skip connections, if the weights are small, then layer collapse still occurs. Thus, only large weights, and not skip connections, can prevent these representational weaknesses.

1 INTRODUCTION

The rapid progress of large language models, text-to-image and text-to-video models like Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2018), GPT-4 (OpenAI, 2023), Llama 3 (Llama Team, 2024), and Gemini 2.0 (Google, 2025), has enabled powerful language modelling abilities. These models take advantage of large-scale pretraining on massive textual data, which equips them with strong abilities to interpret the complex patterns of natural language. These LLMs have a broad range of applications, influencing domains such as human-computer interaction, multilingual translation, language comprehension, text generation, and rapid prototyping of software.

The major architecture behind the success of all these language models is the attention mechanism. Specifically, attention computes pairwise similarities by calculating inner products between vectorized representations of words, with input sequences represented as vectors. Formally, softmax attention can be formulated as follows:

Definition 1.1 (Self-Attention with Softmax Units)**.** *Let $A \in \mathbb{R}^{n \times d}$ and weights $Q, K, V \in \mathbb{R}^{d \times d}$. Let g represent the entry-wise exponentiation function, i.e., for $z \in \mathbb{R}$ we have $g(z) = \exp(z)$, and for a matrix W we have $g(W)_{i,j} = g(W_{i,j})$. The attention computation can be defined as*

$$\mathsf{SAtt}(X, Q, K, V) = \underbrace{D^{-1}}_{n \times n} \underbrace{g(XQK^\top X^\top)}_{n \times n} \underbrace{X}_{n \times d} \underbrace{V}_{d \times d}$$

where $D := \mathrm{diag}(g(XQK^\top X^\top)\mathbf{1}_n)$, and where $\mathbf{1}_n \in \mathbb{R}^n$ is a length-n vector whose entries are all 1.

Small Coefficients are Needed for Fast Algorithms However, the straightforward algorithm for computing self-attention results in a quadratic $O(n^2 d)$ running time, where n is the length of the input token and d is the hidden dimension. Under popular complexity-theoretic assumptions, there is no better, subquadratic time algorithm to compute attention, even approximately (Alman & Song, 2023). Therefore, models based on attention may face difficulties when they handle long contexts.

In fact, a key observation of this line of work on the computational complexity of attention is that attention can be computed (or tightly approximated) faster if one restricts to small weights, i.e., an upper bound on how large the entries of Q, K, V can be in Definition 1.1 above. Indeed, a line of work (Alman & Song, 2023; 2024a;b; 2025) has shown that small weights are both necessary and sufficient for a faster algorithm: If the weights are large, then the aforementioned complexity-theoretic result shows that there is no subquadratic time algorithm. However, if the weights are small, then attention can be approximated to low error in *almost linear time*! Their algorithm is based on low-rank approximations of the $n \times n$ attention matrix (the matrix $g(XQK^\top X^\top)$ in Definition 1.1 above).

This type of observation is also frequently used in practice; many LLM implementations have enforced bounds on the weights, often using techniques like approximation or quantization, and then used this for substantial speedups. For some examples, see (Zafrir et al., 2019; Katharopoulos et al., 2020b; Frantar et al., 2022; Perez et al., 2023; Dettmers et al., 2023; Egashira et al., 2024; Liu et al., 2024b; Xu et al., 2024a; Lin et al., 2025; Chen et al., 2025b; Liu et al., 2025; Ouyang et al., 2025; Deng et al., 2025; Hu et al., 2025c; Fu et al., 2025; Hu et al., 2025b; Park et al., 2025; Zeng et al., 2025; Yu et al., 2025; Wei et al., 2025).

In this paper, we investigate the representational strength of transformers with small weights. Our main result will show a limitation, that without large weights, a transformer cannot take advantage of more than a single layer. In other words, we will show that in order to take advantage of the full expressive power of the transformer model, large weights are necessary.

Rank Collapse and Skip Connections We will crucially build on the approach of Dong, Cordonnier, and Loukas (Dong et al., 2021), who studied the representational strength of different variants on the transformer architecture through the lens of a notion called *rank collapse*. We say that a model experiences rank collapse if, on any input, the output must always be close to a rank 1 matrix. (See Definition 3.4 below for the precise meaning.) Beyond being unable to represent complex concepts, models with rank collapse also have numerous other issues in both training and evaluation (Noci et al., 2022; Roth & Liebig, 2024; Naderi et al., 2024; Nguyen et al., 2024; Heo & Choi, 2024; Yuan & Xu, 2024; Barbero et al., 2025; Bonino et al., 2025).

The work of (Dong et al., 2021) highlights *skip connections* (or residual connections) in a transformer network as crucial for avoiding rank collapse. They show that in a Self-Attention Network without skip connections, rank collapse occurs with a doubly exponential rate of convergence. More precisely, if β is a bound on the ℓ_1 norm of the weight matrices of the network, and the network has L layers, then they show the distance to a rank-1 matrix shrinks as

$$O(\beta)^{\frac{3^L-1}{2}}. \tag{1}$$

Meanwhile, they observe that networks with skip connections may experience no rank collapse at all. For instance, it is not hard to simulate the *identity* function as a Self-Attention Network with skip connections (simply set all value weights to 0, so that only the skip connections are output). In this case, any input which is far from rank-1 will result in an output which is also far from rank-1. They study other mechanisms in transformer networks as well, including multi-layer perceptrons and layer normalization, but find that only skip connections prevents the rank collapse of Equation (1). This result is frequently cited in the literature as evidence of the importance of skip connections (Ma et al., 2021; Noci et al., 2022; Sander et al., 2022; Guo et al., 2023; Li et al., 2023; Kim et al., 2023; Geshkovski et al., 2023; Kim et al., 2024; Ji et al., 2025).

The Importance of Large Weights and Layer Collapse We begin with a simple observation: in order for Equation (1) to be shrinking as L grows, it is necessary that β is small, i.e., that the weights of the network are small. In other words, the result of (Dong et al., 2021) really says that:

To avoid rank collapse, one needs either skip connections *or* large weights.

In this paper, we prove that Self Attention Networks with skip connections, but with small weights, must suffer from a phenomenon similar to rank collapse which we call *layer collapse*. We say that an L-layer Self Attention Network S has layer collapse if there is a nearly equivalent Self Attention Network S' which only has a single layer. In other words, although S' only has one layer, it is still as expressive as S, since on any input X, the outputs $S(X)$ and $S'(X)$ differ in each entry by at most a small error parameter.

When combined with (Dong et al., 2021), our result implies:

> To avoid rank and layer collapse, one needs large weights (skip connections do not suffice).

This challenges the previous popular interpretation of (Dong et al., 2021), that skip connections were crucial for the representational strength of the model.

The connection between layer collapse and rank collapse may not be evident from the definitions, but it will become clear in our proofs below. At a high level, we will find that the attention mechanisms in lower layers of the Self Attention Network must exhibit rank collapse (regardless of skip connections), and can thus be removed from the network without substantially changing the output. We will show

Theorem 1.2 (Main result, informal)**.** *If S is a Self Attention Network whose weight matrices have ℓ_∞ norm bounded by η, then there is a Self Attention Network S' with only one layer, such that on any input X with $\|X\|_\infty \leq O(1)$, we have $\|S(X) - S'(X)\|_\infty \leq O(\eta)$.*

In fact, the example from (Dong et al., 2021) of the identity network with skip connections heavily inspired our definition of layer collapse. That network indeed does not have rank collapse, so we could not hope to prove a version of Theorem 1.2 with rank collapse instead of layer collapse. On the other hand, it is essentially not making use of its attention mechanisms; they could be removed without changing the output of the network. Our key idea is to show that, more generally, the attention mechanisms with small weights can be removed from any Self Attention Network, with skip connections, without changing the output of the network by very much.

Our η in Theorem 1.2 is a bound on the ℓ_∞ norm of the weight matrices (maximum magnitude of an entry), whereas the prior result in Eq. (1) above uses parameter β, which is a bound on the ℓ_1 norm (sum of magnitudes of all entries). Our η could thus be quite a bit smaller (by a factor of d^2 for $d \times d$ weight matrices), and there are thus networks without skip connections where (Dong et al., 2021) does not imply rank collapse (since $\beta \gg 1$ is too big) but our Theorem 1.2 still implies layer collapse (since $\eta \ll 1$ is smaller).

We also note that both our informal statement of Theorem 1.2 and our presentation of the main result of (Dong et al., 2021) in Eq. (1), are given assuming that the Self Attention Network has a constant number of heads and layers. The more complete statement in terms of the number of heads and layers is presented in Theorem D.1 in the appendix. Both results have modest assumptions on the relationships between η (or β), $\|X\|_\infty$, and the numbers of heads and layers, and we emphasize that these assumptions are nearly identical in both results; see Remark D.3 for more details.

Roamdap. In Section 2, we present the related work. In Section 3, we introduce several basic notations and definitions. In Section 4, we study perturbation properties of several functions, such as softmax. In Section 5, we provide several major rank collapse results. In Section 6, we provide the conclusion of this paper.

2 Related Work

Low-rank Approximations Low rank approximation is a fundamental topic in numerical linear algebra (Clarkson & Woodruff, 2013; Nelson & Nguyên, 2013; Song et al., 2023b;a). Many problems require either computationally or analytically finding a low-rank approximation under different settings such as linear and kernel SVMs (Gu et al., 2025), tensor regression (Song et al., 2021b; Reddy et al., 2022; Diao et al., 2018; 2019), low rank approximation with Frobenious norm (Clarkson & Woodruff, 2013; Nelson & Nguyên, 2013), weighted low rank approximation (Razenshteyn et al., 2016; Gu et al., 2024; Li et al., 2025a; Song et al., 2025), general norm column subset selection (Song et al., 2019a), entrywise ℓ_1 norm low rank approximation (Song et al., 2017; 2019b),

tensor low rank approximation (Song et al., 2019c), tensor power method (Deng et al., 2023b), and matrix CUR decomposition (Boutsidis & Woodruff, 2014; Song et al., 2017; 2019c). Rank collapse and other techniques we use here build on this line of work.

Algorithmic Result for Attention Computations The quadratic time complexity of attention mechanisms (Vaswani et al., 2017) has posed significant computational challenges for long sequences. In response to this problem, a wide range of works have been proposed to reduce computational cost and enhance the scalability of attention mechanisms, including sparsification (Child et al., 2019; Zaheer et al., 2020; Beltagy et al., 2020; Hubara et al., 2021; Shi et al., 2023a; Kurtic et al., 2023; Frantar & Alistarh, 2023; Li et al., 2024b; Liang et al., 2024a; Han et al., 2024), kernel-based approaches (Liu & Zenke, 2020; Charikar et al., 2020; Zandieh et al., 2023; Deng et al., 2023a; Liang et al., 2024b), and low-rank methods (Li et al., 2016; Razenshteyn et al., 2016; Hu et al., 2022; 2024b; Zeng & Lee, 2024). Additionally, another promising line of research is linear attention (Tsai et al., 2019; Katharopoulos et al., 2020a; Schlag et al., 2021; Deng et al., 2023c; Sun et al., 2023; Zhang et al., 2023b; Ahn et al., 2024; Li et al., 2024a; Shi et al., 2023c; Zhang et al., 2024), which significantly accelerates traditional softmax attention. Other relevant works have explored important aspects of attention mechanisms, covering topics such as circuit complexity (Chen et al., 2024a;c; Li et al., 2025b), model pruning (Frantar & Alistarh, 2023; Shen et al., 2024; Sun et al., 2024; Liang et al., 2025), privacy protection (Liang et al., 2024d; Gao et al., 2024), regression (Gao et al., 2023b), half-space reporting (HSR) (Jiang et al., 2021; Chen et al., 2024b), and quantum computation (Gao et al., 2023c; Zhao et al., 2024).

Polynomial Kernels for Attention Acceleration With the assumption that model weights are small, polynomial kernels (Alman & Song, 2023; 2024b) are powerful tools for approximating attention computation in almost linear time complexity, providing promising acceleration for both training and inference of a single attention layer. This approach can be further extended to a wide range of applications. For instance, polynomial kernels can provide insights into novel attention mechanisms and model designs, such as modern Hopfield models (Hu et al., 2024a), Diffusion Transformers (DiTs) (Shen et al., 2025; Hu et al., 2024d), multi-layer Transformers (Liang et al., 2024c), and tensor attention mechanisms (Liang et al., 2024e; Alman & Song, 2024a). These polynomial kernel methods also contribute to efficient and model-utility-preserving fine-tuning of foundation models, such as model adapters (Hu et al., 2022; Zhang et al., 2023a; Shi et al., 2023b; Gao et al., 2023a), multi-task fine-tuning (Gao et al., 2021; Oswald et al., 2023; Xu et al., 2024b), black-box model tuning (Sun et al., 2022), and instruction tuning (Li & Liang, 2021; Chung et al., 2022; Mishra et al., 2022). Other promising applications include privacy protection in attention computation (Liang et al., 2024d), CoT reasoning (Khattab et al., 2022; Wei et al., 2022; Yao et al., 2023; Zheng et al., 2024), and model calibration (Zhao et al., 2021; Zhou et al., 2023). Very recently, (Gupta et al., 2025) further extends the work of (Alman & Song, 2023) to almost all the regimes of parameter d (see definition of d in Defintion 1.1).

Regression Models The unprecedented energy consumption in training large-scale ML models has necessitated the development of scalable and efficient ML models (Venkataramani et al., 2015; Bender et al., 2021; McDonald et al., 2022). As a simple yet powerful approach to solving various machine learning problems (Bubeck, 2015; Brand et al., 2021; Song et al., 2024b; Subrahmanya & Shin, 2009), simple regression models have raised significant concerns in model acceleration, with recent advances from different perspectives, including sketching (Song & Yu, 2021; Reddy et al., 2022; Song et al., 2023a) and pre-conditioning (Yang et al., 2018; Kelner et al., 2022; Song et al., 2024a). Our work discusses low-rank approximations in attention mechanisms, while our general insight can be extended to other low-rank method applications, such as accelerated regression models.

Diffusion Models Diffusion models and score-based generative models have achieved remarkable success in generating human-preference-aligned and high-quality visual content (Ho et al., 2020; Song et al., 2021a; Blattmann et al., 2023). These advances not only benefit vision tasks but also enhance the performance of other applications, such as language modeling (Lin et al., 2023; Sahoo et al., 2024), chemical design (Xu et al., 2023; Wen et al., 2024), and e-commerce (Yang et al., 2023; Wang et al., 2023; Liu et al., 2024a). Relevant works have discussed the theoretical guarantee that diffusion models can be approximated efficiently (Hu et al., 2024d; 2025a; 2024c;

Gong et al., 2025). Empirical approaches to accelerate diffusion models have addressed various aspects, such as shortcuts (Frans et al., 2024; Dao et al., 2024; Chen et al., 2025a), parameter pruning (Castells et al., 2024; Ma et al., 2024), and lazy computation (Nitzan et al., 2024; Shen et al., 2025). With these acceleration techniques, diffusion models can be trained on larger-scale data, overcoming inherent limitations such as counting (Hui et al., 2024; Cao et al., 2025; Guo et al., 2025a), text rendering (Chen et al., 2023; Tuo et al., 2024; Guo et al., 2025c), and adherence to physical constraints (Motamed et al., 2025; Guo et al., 2025b; Bansal et al., 2025). Most diffusion models leverage Transformer backbones for enhanced modelling capability. Our work accelerates attention mechanism computations, significantly benefiting a wide range of diffusion models.

Graph ML Models Relational data is prevalent in many real-world scenarios, where graph neural networks (GNNs) are the powerful solutions for mining effective patterns from such relations (Kipf & Welling, 2017; Hamilton et al., 2017; Wu et al., 2019). Recent scalability approaches have widely adopted low-rank approximations, such as sketching (Ding et al., 2022; Chamberlain et al., 2023) and vector quantization (Ding et al., 2021; Wang et al., 2025), which can take insights from this paper. These accelerations empower a wide range of applications, including misleading information mitigation (Xu et al., 2022; Chang et al., 2024), social network prediction (Fan et al., 2019; Zhang et al., 2022), and human action recognition (Peng et al., 2020; Li et al., 2021; Fu et al., 2021), while also inspiring advances in multiple aspects of graph learning, such as differential privacy (Lin et al., 2022; Mueller et al., 2022), robustness (Geisler et al., 2021; Dai et al., 2022; Zeng et al., 2022), and sensitive data removal (Chien et al., 2023; Zhang, 2024; Yi & Wei, 2025). A recent work (Zhang, 2024) proposes an efficient framework for empowering sensitive data impact removal from trained GNNs with partial retraining, leveraging model utility-aware data partitioning and contrastive submodel aggregation.

3 PRELIMINARIES

In Section 3.1, we provide basic notation, definitions and facts. In Section 3.2 and Section 3.3, we define the Res function and balanced matrix notation which will appear prominently in our constructions. In Section 3.4, we provide the definition of a multi-layer multi-head Self Attention Network which we study here.

3.1 BASIC NOTATION AND FACTS

For an arbitrary positive integer n, we use $[n]$ to represent the set $\{1, 2, \cdots, n\}$. We define $\mathbf{1}_n$ as a length-n vector where all entries are ones. For any $x \in \mathbb{R}^n$, we use $\exp(x) \in \mathbb{R}^n$ to represent a length-n vector whose i-th entry is $\exp(x_i)$. For any vector $x \in \mathbb{R}^n$, we use $x^\top$ to denote its transpose. For a vector x, the vector ℓ_2 norm is denoted by $\|x\|_2$, i.e., $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$. For a vector x, we use $\|x\|_\infty$ to denote its ℓ_∞ norm, i.e., $\|x\|_\infty := \max_{i=1}^n |x_i|$. For a vector x, we use $\|x\|_1$ to denote its entrywise ℓ_1 norm, i.e., $\|x\|_1 := \sum_{i=1}^n |x_i|$. For a matrix, we use $\|A\|_1$ to denote its ℓ_1 norm, i.e., $\|A\|_1 = \sum_{j,l} |A_{j,l}|$. We use $\|A\|_\infty$ to denote its ℓ_∞ norm, i.e., $\|A\|_\infty := \max_{j,l} |A_{j,l}|$. For a vector $x \in \mathbb{R}^n$, we use $\mathrm{diag}(x)$ to denote a diagonal matrix where i, i-th entry on diagonal is x_i for all $i \in [n]$.

Definition 3.1. *For a vector $x \in \mathbb{R}^n$, we define $\alpha(x) := \langle \exp(x), \mathbf{1}_n \rangle$. We define $\mathsf{softm}(x)$ as $\mathsf{softm}(x) := \alpha(x)^{-1} \exp(x)$. For a matrix A, we use the notation $\mathsf{softm}(A)$ to denote that we apply softm to each row of A individually.*

Fact 3.2 (Shift-invariance property of softmax)**.** *For any vector $x \in \mathbb{R}^n$ and for any fixed scalar $a \in \mathbb{R}$, we have $\mathsf{softm}(x) = \mathsf{softm}(x + a\mathbf{1}_n)$.*

Fact 3.3 (Norm inequality)**.** *For any matrices A, B we have (1) $\|AB\|_1 \leq \|A\|_1 \cdot \|B\|_1$, (2)$\|AB\|_\infty \leq \|A\|_\infty \cdot \|B\|_\infty$, (3) $\|AB\|_1 \leq \|A\|_1 \cdot \|B\|_\infty$.*

3.2 DEFINITIONS OF Res

Definition 3.4 (Res)**.** *Let $Z \in \mathbb{R}^{n \times d}$ denote any matrix, we define function the $\mathsf{Res} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ as $\mathsf{Res}(Z) := Z - \mathbf{1}_n y^\top$ where $y := \arg\min_{y \in \mathbb{R}^d} \|Z - \mathbf{1}_n y^\top\|_\infty$.*

Res is the key definition behind the notion of rank collapse from prior work (Dong et al., 2021); we will use it here to study layer collapse as well, although we use the ∞ norm here in contrast to prior work which uses a $1, \infty$ norm.

3.3 θ-BALANCE

We also need a measure of how balanced a matrix is.

Definition 3.5 (θ-balance)**.** *Given a matrix $E \in \mathbb{R}^{n \times n}$, we define a corresponding matrix $D \in \mathbb{R}^{n \times n}$ to be the diagonal matrix with $D_{i,i} := \max_{j,l \in [n]} |E_{i,j} - E_{i,l}|$. We say E is θ-balanced, if $\|D\|_\infty \leq \theta$.*

3.4 SELF-ATTENTION NETWORK

Definition 3.6. *Let g denote the entry-wise exponentiation function, i.e., for $z \in \mathbb{R}$ we have $g(z) = \exp(z)$, and for a matrix W we have $g(W)_{i,j} = g(W_{i,j})$. Given $A \in \mathbb{R}^{n \times d}$ and weights $Q, K, V \in \mathbb{R}^{d \times d}$, the attention computation can be defined as*

$$\mathsf{SAtt}_H(X) := \sum_{h=1}^{H} \underbrace{D_h^{-1}}_{n \times n} \underbrace{g(XQ_hK_h^\top X^\top)}_{n \times n} \underbrace{X}_{n \times d} \underbrace{V_h}_{d \times d}$$

where $D_h := \mathrm{diag}(g(XQ_hK_h^\top X^\top)\mathbf{1}_n)$, and where $\mathbf{1}_n \in \mathbb{R}^n$ is a length-n vector whose entries are all 1.

Definition 3.7. *Let L, H denote fixed constants, where L represents the number of layers of the network, and H represents the number of heads per layer. Let SAtt_H denote the multi-heads version of SAtt where H is the number of heads. For each $\ell \in [L]$, $X_\ell \in \mathbb{R}^{n \times d}$ denote the ℓ-th layer input of self-attention network, then we have $X_{\ell+1} = \mathsf{SAtt}_H(X_\ell) + X_\ell$.*

4 PERTURBATION PROPERTY

We now move on to our main proof of layer collapse. We begin by showing that the relevant measure of matrices to not change much when their inputs are perturbed. We will ultimately show that layer collapse occurs because lower layers of the network can be seen as slightly perturbing their inputs. We study the Res function in Section 4.1, the α function in Section 4.2.

4.1 PERTURBATION PROPERTY OF RES FUNCTION

Lemma 4.1. *Let $\mathsf{Res}()$ be defined as Definition 3.4. If $\|A - B\|_\infty \leq \epsilon$, then $\|\mathsf{Res}(A) - \mathsf{Res}(B)\|_\infty \leq \epsilon$.*

Proof. Let $y \in \mathbb{R}^d$ be the vector such that $\mathsf{Res}(B) = B - \mathbf{1}_n y^\top$. Then, $\|\mathsf{Res}(A) - \mathsf{Res}(B)\|_\infty \leq \|A - \mathbf{1}_N y^\top - \mathsf{Res}(B)\|_\infty \leq \|B - \mathbf{1}_N y^\top - \mathsf{Res}(B)\|_\infty + \|B - A\|_\infty \leq 0 + \epsilon$. □

4.2 PERTURBATION PROPERTY OF EXP FUNCTION

Lemma 4.2. *If the following conditions hold: Let $a, b \in \mathbb{R}^n$. Let $\|b\|_\infty \leq \epsilon$. Then, we can show*

- $|\exp(a_i + b_i) - \exp(a_i)| \leq (e^\epsilon - 1) \cdot \exp(a_i)$.
- $|\exp(a_i + b_i) - \exp(a_i)| \leq (e^\epsilon - 1) \cdot \exp(a_i + b_i)$.
- $|\alpha(a + b) - \alpha(a)| \leq (e^\epsilon - 1) \cdot \alpha(a)$.
- $|\alpha(a + b) - \alpha(a)| \leq (e^\epsilon - 1) \cdot \alpha(a + b)$.

Proof. It is easy to see that

$$\max\{|\exp(-b_i) - 1|, |\exp(b_i) - 1|\} \leq e^\epsilon - 1 \tag{2}$$

We can show

$$|\exp(a_i + b_i) - \exp(a_i)| = \exp(a_i)|\exp(b_i) - 1| \leq \exp(a_i) \cdot (e^\epsilon - 1) \tag{3}$$

where the first step follows from simple algebra, the second step follows from Eq. (2).

Thus, we have

$$|\alpha(a+b) - \alpha(a)| = |\langle \exp(a+b), \mathbf{1}_n \rangle - \langle \exp(a), \mathbf{1}_n \rangle| \leq \sum_{i=1}^n |\exp(a_i + b_i) - \exp(a_i)|$$

$$\leq \sum_{i=1}^n \exp(a_i) \cdot (e^\epsilon - 1) = (e^\epsilon - 1)\alpha(a)$$

where the second step follows from triangle inequality, the third step follows from Eq. (3), the last step follows from definition of $\alpha(\cdot)$ function.

Similarly, we can show

$$|\exp(a_i + b_i) - \exp(a_i)| = \exp(a_i + b_i)|\exp(-b_i) - 1| \leq \exp(a_i + b_i) \cdot (e^\epsilon - 1) \tag{4}$$

where the first step follows from simple algebra, the second step follows from Eq. (2).

Then, we have

$$|\alpha(a+b) - \alpha(a)| = |\langle \exp(a+b), \mathbf{1}_n \rangle - \langle \exp(a), \mathbf{1}_n \rangle| \leq \sum_{i=1}^n |\exp(a_i + b_i) - \exp(a_i)|$$

$$\leq \sum_{i=1}^n \exp(a_i + b_i) \cdot (e^\epsilon - 1) = (e^\epsilon - 1)\alpha(a+b)$$

where the first step follows from definition of α (Definition 3.1), the second step follows from triangle inequality, the third step follows from Eq. (4), and the last step follows from definition of α (Definition 3.1).

Thus, we complete the proof. □

5 Rank Collapse Property

In Section 5.1, we present a Lemma which connects $\mathsf{Res}(\mathsf{SAtt}())$ and $\mathsf{Res}()$. In Section 5.2, we present our key lemma, a perturbation theorem for a layer of a Transformer. In Section 5.3, we present our main result and proof sketch.

5.1 The Connection Between $\mathsf{Res}(\mathsf{SAtt}())$ and $\mathsf{Res}()$

We next establish the relationship between $\mathsf{Res}(\mathsf{SAtt}())$ and $\mathsf{Res}()$ in terms of the balance of the inputs.

Lemma 5.1. *If the following conditions hold: Let $X \in \mathbb{R}^{n \times d}$ denote the input of attention layer. Let $\widetilde{X} = \mathsf{SAtt}(X)$ (see Definition 1.1 for function SAtt). Let $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ be the weight matrices of SAtt. Let $W = W_q W_k^\top$. Let $E = \beta \mathsf{Res}(X) W \mathsf{Res}(X)^\top$. Suppose that E is a θ-balanced matrix (see Definition 3.5). Let $\beta := 1/\sqrt{d_0}$ denote the normalization factor. Let $K := (e^\theta - 1)\|W_v\|_\infty$. Then, we have $\|\mathsf{Res}(\mathsf{SAtt}(X))\|_\infty \leq K \cdot \|\mathsf{Res}(X)\|_\infty$.*

Proof. The unscaled attention scores are computed as follows $A = (XW_q + \mathbf{1}_n b_q^\top) \cdot (XW_k + \mathbf{1}_n b_k^\top)^\top$. Recall that $W = W_q W_k^\top$. For notational convenience, we define $b := W_k b_q$.

We can use the softmax shift invariance property to remove terms which are constant over the columns and obtain, $A = \underbrace{X}_{n \times d} \underbrace{W}_{d \times d} \underbrace{X^\top}_{d \times n} + \underbrace{\mathbf{1}_n}_{n \times 1} \underbrace{b^\top}_{1 \times d} \underbrace{X^\top}_{d \times n}$.

We define $\widetilde{R} := \mathsf{Res}(\widetilde{X}) \in \mathbb{R}^{n \times d}$ (Recall the definition of the function $\mathsf{Res}()$ in Definition 3.4).

In next equation, we will use the definition of R to simplify A. The attention matrix can be written as

$$\begin{aligned} A &= \beta \cdot (\mathbf{1}_n x^\top + R) W (\mathbf{1}_n x^\top + R)^\top + \beta \cdot \mathbf{1}_n b^\top (\mathbf{1}_n x^\top + R)^\top \\ &= \beta \cdot (x^\top W x \mathbf{1}_n + RWx + \mathbf{1}_n b^\top x) \mathbf{1}_n^\top + \beta \cdot (RWR^\top + \mathbf{1}_n x^\top W R^\top + \mathbf{1}_n b^\top R^\top) \end{aligned} \tag{5}$$

Using Fact 3.2, we can remove the first term in the above equation since it is constant across columns. We thus have that the following equation for $P = \mathsf{softm}(A) \in \mathbb{R}^{n \times n}$

$$P = \mathsf{softm}(\beta RWR^\top + \mathbf{1}_n r^\top) = \mathsf{softm}(E + \mathbf{1}_n r^\top) \tag{6}$$

where the first step follows from $r = \beta R(W^\top x + b) \in \mathbb{R}^n$, the second step follows from setting $E = \beta RWR^\top \in \mathbb{R}^{n \times n}$.

To continue the proof, we also set $\widetilde{A} = \mathbf{1}_n r^\top \in \mathbb{R}^{n \times n}$, the input reweighted by the attention probabilities PX will be entry-wisely upper bounded as follows

$$\begin{aligned} PX = P(\mathbf{1}_n x^\top + R) &= \mathbf{1}_n x^\top + PR \\ &= \mathbf{1}_n x^\top + \mathsf{softm}(\mathbf{1}_n r^\top + E) R \\ &\leq \mathbf{1}_n x^\top + (I + e^D - I) \mathbf{1}_n \mathsf{softm}(r)^\top R \\ &= \mathbf{1}_n (x^\top \mathsf{softm}(r)^\top R) + (e^D - I) \mathbf{1}_n \mathsf{softm}(r)^\top R \end{aligned} \tag{7}$$

where the first step follows from definition of R, the second step follows from $P\mathbf{1}_n = \mathbf{1}_n$, the third step follows from Eq. (6), the forth step follows from Lemma B.3 and e^D is diagonal matrix where the i, i-th entry on diagonal is $e^{D_{i,i}}$.

Therefore, the entry-wise distance of the output of the self-attention layer $\mathsf{SAtt}(X) = PXW_v$ from being constant across token is at most:

$$|[\mathsf{SAtt}(X) - \mathbf{1}_n \widetilde{r}^\top]_{i,j}| = |[PXW_v - \mathbf{1}_n \widetilde{r}^\top]_{i,j}| \leq (e^\theta - 1) \cdot |[D\mathbf{1}_n \mathsf{softm}(r)^\top R W_v]_{i,j}|$$

where the second step follows from $\widetilde{r} = (x + R^\top \mathsf{softm}(r)) W_v$ and Eq. (7).

Now we bound the right hand side of the above inequality.

For $\|\cdot\|_\infty$, we can show

$$\begin{aligned} \|\mathsf{SAtt}(X) - \mathbf{1}_n (r')^\top\|_\infty &\leq (e^\theta - 1) \|D\mathbf{1}_n\|_\infty \cdot \|\mathsf{softm}(r)^\top R W_v\|_\infty \\ &\leq (e^\theta - 1) \|D\mathbf{1}_n\|_\infty \|R\|_\infty \|W_v\|_\infty \leq (e^\theta - 1) \|R\|_\infty \cdot \|W_v\|_\infty, \end{aligned} \tag{8}$$

where the last step follows from Definition 3.5.

Note that $R' = \mathsf{Res}(\mathsf{SAtt}(X))$ and $R = \mathsf{Res}(X)$ and using the definition of K in Lemma statement, we can show $\|\mathsf{Res}(\mathsf{SAtt}(X))\|_\infty \leq K \cdot \|\mathsf{Res}(X)\|_\infty$. Thus, we complete the proof. □

5.2 Perturbation of One Transformer Layer

Lemma 5.2 (Single Head). *Let $X \in \mathbb{R}^{n \times d}$. Let $A = \mathsf{softm}_1(X)$ (Recall that $\mathsf{softm}()$ function is defined as Definition 3.1. Note that softm_1 and softm_2 are two different instantiations with different W_k, W_q, W_v weights). Let $B = X + A$. Suppose $\|\mathsf{Res}(A)\|_\infty \leq K \cdot \|\mathsf{Res}(X)\|_\infty \leq \epsilon$. (We remark that this condition will hold due to Lemma 5.1; here K is as defined in Lemma 5.1). Let $g(\epsilon) := 2(e^\epsilon - 1)$ and let $\epsilon_0 = 2g(2\epsilon)$. Then we can show*

$$\|\mathsf{softm}_2(B) - \mathsf{softm}_2(X)\|_\infty \leq \epsilon_0.$$

Proof. Let $R_X = \mathsf{Res}(X)$ so that $X = R_X + y_X \mathbf{1}_d^\top$ for some vector $y_X \in \mathbb{R}^n$. Using Fact 3.2, we can show that

$$\mathsf{softm}_1(X) = \mathsf{softm}_1(R_X). \tag{9}$$

Let $R_A = \mathsf{Res}(A)$ so that $A = R_A + y_A \mathbf{1}_d^\top$ for some vector $y_A \in \mathbb{R}^n$. Using Fact 3.2, we can show that $\mathsf{softm}(A) = \mathsf{softm}(R_A)$.

Let $R_B = \mathsf{Res}(B)$ so that $B = R_B + y_B \mathbf{1}_d^\top$ for some vector $y_B \in \mathbb{R}^n$. Using Fact 3.2, we can show that $\mathsf{softm}_2(B) = \mathsf{softm}_2(R_B)$.

Next we will show that $\| \mathsf{softm}(R_X + R_A) - \mathsf{softm}(R_X) \|_\infty$ is small.

Let us consider the vector version $\| \mathsf{softm}(a + b) - \mathsf{softm}(a) \|_\infty$. Note that if $|b_i| \leq \epsilon$, then using Lemma B.2, we can show $\| \mathsf{softm}(a + b) - \mathsf{softm}(a) \|_\infty \leq g(\epsilon)$.

Thus, as long as $\|R_A\|_\infty \leq \epsilon$, then using Lemma B.2, we have

$$\| \mathsf{softm}_2(R_X + R_A) - \mathsf{softm}_2(R_X) \|_\infty \leq g(\epsilon) \tag{10}$$

We can show $R_X = \mathsf{Res}(X) = \mathsf{Res}(B - A) = \mathsf{Res}(B - R_A)$. Then, we know $\|R_X - R_B\|_\infty \leq \| \mathsf{Res}(B - R_A) - \mathsf{Res}(B) \|_\infty \leq \|R_A\|_\infty \leq \epsilon$.

Recall $B = X + A$, and $\|R_A\|_\infty \leq \epsilon$ then we know

$$\|R_X + R_A - R_B\|_\infty \leq \|R_X - R_B\|_\infty + \|R_A\|_\infty \leq 2\|R_A\|_\infty \leq 2\epsilon.$$

Since $\|R_X + R_A - R_B\|_\infty \leq 2\epsilon$, then using Lemma B.2, we have

$$\| \mathsf{softm}_2(R_X + R_A) - \mathsf{softm}_2(R_B) \|_\infty \leq g(2\epsilon) \tag{11}$$

Then, we can show

$$\begin{aligned}
& \| \mathsf{softm}_2(B) - \mathsf{softm}_2(X) \|_\infty \\
= & \| \mathsf{softm}_2(B) - \mathsf{softm}_2(R_X) \|_\infty \\
= & \| \mathsf{softm}_2(R_B) - \mathsf{softm}_2(R_X) \|_\infty \\
\leq & \| \mathsf{softm}_2(R_B) - \mathsf{softm}_2(R_X + R_A) \|_\infty + \| \mathsf{softm}_2(R_X + R_A) - \mathsf{softm}_2(R_X) \|_\infty \\
\leq & g(2\epsilon) + g(\epsilon) \leq 2g(2\epsilon)
\end{aligned}$$

where the first step follows from $\mathsf{softm}(X) = \mathsf{softm}(R_X)$, the second step follows from $\mathsf{softm}(R_B) = \mathsf{softm}(B)$, the third step follows from triangle inequality, the forth step follows from Eq. (10) and Eq. (11), and the last step follows from g is monotone. □

5.3 Putting it All Together

Proof Sketch of Theorem 1.2. We'll show what to do to delete one layer, then repeat that $L - 1$ times to get down to one layer. When we delete the first layer, Lemma C.1 (which is the version of Lemma 5.2 which deals with multiple heads) says that the output of the second layer will differ by at most $O(\eta \cdot \epsilon_0)$, where $\epsilon_0 = O(1) \cdot \|X\|_\infty$ is the constant from Lemma 5.1 and Lemma 5.2, and X is the input of first layer of network. Therefore, by applying Lemma B.2 iteratively to each layer, it follows that the outputs of all subsequent layers will also change by at most $O(\eta \cdot \epsilon_0)$. In particular, the final output will differ by at most $O(\eta \cdot \epsilon_0)$. We finally repeat this $L - 1$ times to remove all but one layer and get the final error. We defer further proof details to the Appendix due to space limitations. □

6 Conclusion

We have shown that Self Attention Networks must experience layer collapse unless they have large attention weights, even if they have skip connections. Our result proves that two different common notions in the literature are actually misconceptions.

The first misconception is the common interpretation of the prior work (Dong et al., 2021) that skip connections are the key to the expressive power of Self Attention Networks. We extend their result and show that even with skip connections, large weights are needed to prevent layer collapse.

The second misconception is that Self Attention Networks with smaller weights may still have reasonable expressive power. Indeed, although it is intuitive that bounding the magnitudes of weights must limit the expressive power to some extent, there is nonetheless a long line of work on trying to use networks with small weights, weight quantization, or similar approaches. This work is (presumably) hoping that the limit is only modest. We show that the limit is severe: networks with small weights cannot take advantage of more than one layer! This is the first theoretical limitation result on networks with small weights to our knowledge.

ETHIC STATEMENT

This paper does not involve human subjects, personally identifiable data, or sensitive applications. We do not foresee direct ethical risks. We follow the ICLR Code of Ethics and affirm that all aspects of this research comply with the principles of fairness, transparency, and integrity.

REPRODUCIBILITY STATEMENT

We ensure reproducibility of our theoretical results by including all formal assumptions, definitions, and complete proofs in the appendix. The main text states each theorem clearly and refers to the detailed proofs. No external data or software is required.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

Josh Alman and Zhao Song. Fast attention requires bounded entries. *NeurIPS*, 2023.

Josh Alman and Zhao Song. How to capture higher-order correlations? generalizing matrix softmax attention to kronecker computation. In *ICLR*, 2024a.

Josh Alman and Zhao Song. The fine-grained complexity of gradient computation for training large language models. In *NeurIPS*, 2024b.

Josh Alman and Zhao Song. Fast rope attention: Combining the polynomial method and fast fourier transform. In *arXiv preprint arXiv:2505.11892*, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *Workshop on Video-Language Models @ NeurIPS 2024*, 2025. URL https://openreview.net/forum?id=xMlYKYFd03.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Razvan Pascanu, et al. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Matteo Bonino, Giorgia Ghione, and Giansalvo Cirrincione. The geometry of bert. *arXiv preprint arXiv:2502.12033*, 2025.

Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing (STOC)*, pp. 353–362, 2014.

Jan van den Brand, Binghui Peng, Zhao Song, and Omri Weinstein. Training (overparametrized) neural networks in near-linear time. *ITCS*, 2021.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yuefan Cao, Xuyang Guo, Jiayan Huo, Yingyu Liang, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Zhen Zhuang. Text-to-image diffusion models cannot count, and prompt refinement cannot help. *arXiv preprint arXiv:2503.06884*, 2025.

Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, 2024.

Benjamin Paul Chamberlain, Sergey Shirobokov, Emanuele Rossi, Fabrizio Frasca, Thomas Markovich, Nils Yannick Hammerla, Michael M. Bronstein, and Max Hansmire. Graph neural networks for link prediction with subgraph sketching. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m1oqEOAozQU.

Ya-Ting Chang, Zhibo Hu, Xiaoyu Li, Shuiqiao Yang, Jiaojiao Jiang, and Nan Sun. Dihan: A novel dynamic hierarchical graph attention network for fake news detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 197–206, 2024.

Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 172–183. IEEE, 2020.

Bo Chen, Xiaoyu Li, Yingyu Liang, Jiangxuan Long, Zhenmei Shi, and Zhao Song. Circuit complexity bounds for rope-based transformer architecture. *arXiv preprint arXiv:2411.07602*, 2024a.

Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Hsr-enhanced sparse attention acceleration. *arXiv preprint arXiv:2410.10165*, 2024b.

Bo Chen, Chengyue Gong, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. High-order matching for one-step shortcut diffusion models. *arXiv preprint arXiv:2502.00688*, 2025a.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 28306–28315, 2025b.

Yifang Chen, Jiayan Huo, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Fast gradient computation for rope attention in almost linear time. *arXiv preprint arXiv:2412.17316*, 2024c.

Eli Chien, Chao Pan, and Olgica Milenkovic. Efficient model updates for approximate unlearning of graph-structured data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fhcu4FBLciL.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. In *STOC*, 2013.

Enyan Dai, Wei Jin, Hui Liu, and Suhang Wang. Towards robust graph neural networks for noisy graphs with sparse labels. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 181–191, 2022.

Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *European Conference on Computer Vision*, pp. 176–192. Springer, 2024.

Juncan Deng, Shuaiting Li, Zeyu Wang, Hong Gu, Kedong Xu, and Kejie Huang. Vq4dit: Efficient post-training vector quantization for diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39:15, pp. 16226–16234, 2025.

Yichuan Deng, Zhao Song, Zifan Wang, and Han Zhang. Streaming kernel pca algorithm with small space. *arXiv preprint arXiv:2303.04555*, 2023a.

Yichuan Deng, Zhao Song, and Junze Yin. Faster robust tensor power method for arbitrary order. *arXiv preprint arXiv:2306.00406*, 2023b.

Yichuan Deng, Zhao Song, and Tianyi Zhou. Superiority of softmax: Unveiling the performance edge over linear attention. *arXiv preprint arXiv:2310.11685*, 2023c.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems (NeurIPS)*, 36:10088–10115, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Huaian Diao, Zhao Song, Wen Sun, and David Woodruff. Sketching for kronecker product regression and p-splines. In *International Conference on Artificial Intelligence and Statistics*, pp. 1299–1308. PMLR, 2018.

Huaian Diao, Rajesh Jayaram, Zhao Song, Wen Sun, and David Woodruff. Optimal sketching for kronecker product regression and low rank approximation. *Advances in neural information processing systems*, 32, 2019.

Mucong Ding, Kezhi Kong, Jingling Li, Chen Zhu, John Dickerson, Furong Huang, and Tom Goldstein. Vq-gnn: A universal framework to scale up graph neural networks using vector quantization. *Advances in Neural Information Processing Systems*, 34:6733–6746, 2021.

Mucong Ding, Tahseen Rabbani, Bang An, Evan Wang, and Furong Huang. Sketch-gnn: Scalable graph neural networks with sublinear training complexity. In *Advances in Neural Information Processing Systems*, 2022.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting llm quantization. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:41709–41732, 2024.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pp. 417–426, 2019.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Minghao Fu, Hao Yu, Jie Shao, Junjie Zhou, Ke Zhu, and Jianxin Wu. Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4462–4472, 2025.

Ziwang Fu, Feng Liu, Jiahao Zhang, Hanyang Wang, Chengyi Yang, Qing Xu, Jiayin Qi, Xiangling Fu, and Aimin Zhou. Sagn: semantic adaptive graph network for skeleton-based human action recognition. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pp. 110–117, 2021.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023a.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Yeqi Gao, Sridhar Mahadevan, and Zhao Song. An over-parameterized exponential regression. *arXiv preprint arXiv:2303.16504*, 2023b.

Yeqi Gao, Zhao Song, Xin Yang, and Ruizhe Zhang. Fast quantum algorithm for attention computation. *arXiv preprint arXiv:2307.08045*, 2023c.

Yeqi Gao, Zhao Song, Xin Yang, and Yufa Zhou. Differentially private attention computation. In *Neurips Safe Generative AI Workshop 2024*, 2024.

Simon Geisler, Tobias Schmidt, Hakan Sirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. In *NeurIPS*, 2021.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Chengyue Gong, Yekun Ke, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. On computational limits of flowar models: Expressivity and efficiency. *arXiv preprint arXiv:2502.16490*, 2025.

Google. Gemini 2.0 is now available to everyone, 2025. URL `https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/`.

Yuzhou Gu, Zhao Song, Junze Yin, and Lichen Zhang. Low rank matrix completion via robust alternating minimization in nearly linear time. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR'24, 2024.

Yuzhou Gu, Zhao Song, and Lichen Zhang. Faster algorithms for structured linear and kernel support vector machines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=DDNFTaVQdU`.

Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

Xuyang Guo, Zekai Huang, Jiayan Huo, Yingyu Liang, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Can you count to nine? a human evaluation benchmark for counting limits in modern text-to-video models. *arXiv preprint arXiv:2504.04051*, 2025a.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025b.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vtextbench: A human evaluation benchmark for textual control in video generation models. *arXiv preprint arXiv:2505.04946*, 2025c.

Shreya Gupta, Boyang Huang, Barna Saha, Yinzhan Xu, and Christopher Ye. Subquadratic algorithms and hardness for attention with any temperature. In *arXiv preprint arXiv:2505.14840*, 2025.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Insu Han, Rajesh Jayaram, Amin Karbasi, Vahab Mirrokni, David Woodruff, and Amir Zandieh. Hyperattention: Long-context attention in near-linear time. In *The Twelfth International Conference on Learning Representations*, 2024.

DongNyeong Heo and Heeyoul Choi. Generalized probabilistic attention mechanism in transformers. *arXiv preprint arXiv:2410.15578*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jerry Yao-Chieh Hu, Thomas Lin, Zhao Song, and Han Liu. On computational limits of modern hopfield models: A fine-grained complexity analysis. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a.

Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (lora) for transformer-based models. *arXiv preprint arXiv:2406.03136*, 2024b.

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably optimal memory capacity for modern hopfield models: Tight analysis for transformer-compatible dense associative memories. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024c.

Jerry Yao-Chieh Hu, Weimin Wu, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). *arXiv preprint arXiv:2407.01079*, 2024d.

Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Weiming Hu, Haoyan Zhang, Cong Guo, Yu Feng, Renyang Guan, Zhendong Hua, Zihan Liu, Yue Guan, Minyi Guo, and Jingwen Leng. M-ant: Efficient low-bit group quantization for llms via mathematically adaptive numerical type. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 1112–1126. IEEE, 2025b.

Xing Hu, Yuan Cheng, Dawei Yang, Zhixuan Chen, Zukang Xu, JiangyongYu, XUCHEN, Zhihang Yuan, Zhe jiang, and Sifan Zhou. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025c. URL `https://openreview.net/forum?id=rAcgDBdKnP`.

Itay Hubara, Brian Chmiel, Moshe Island, Ron Banner, Joseph Naor, and Daniel Soudry. Accelerated sparse neural training: A provable and efficient method to find n: m transposable masks. *Advances in neural information processing systems*, 34:21099–21111, 2021.

Xiaofei Hui, Qian Wu, Hossein Rahmani, and Jun Liu. Class-agnostic object counting with text-to-image diffusion model. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Yiping Ji, Hemanth Saratchandran, Peyman Moghaddam, and Simon Lucey. Always skip attention. *arXiv preprint arXiv:2505.01996*, 2025.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. A faster algorithm for solving general lps. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 823–832, 2021.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020a.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020b.

Jonathan A Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 550–561. IEEE, 2022.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.

Bo-Kyeong Kim, Geonmin Kim, Tae-Ho Kim, Thibault Castells, Shinkook Choi, Junho Shin, and Hyoung-Kyu Song. Shortened llama: A simple depth pruning for large language models. *arXiv preprint arXiv:2402.02834*, 11, 2024.

Jihwan Kim, Miso Lee, and Jae-Pil Heo. Self-feedback detr for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10286–10296, 2023.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Eldar Kurtic, Denis Kuznedelev, Elias Frantar, Michael Goin, and Dan Alistarh. Sparse finetuning for inference acceleration of large language models. *arXiv preprint arXiv:2310.06927*, 2023.

Chenyang Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Exploring the frontiers of softmax: Provable optimization, applications in diffusion model, and beyond. *arXiv preprint arXiv:2405.03251*, 2024a.

Chenyang Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. When can we solve the weighted low rank approximation problem in truly subquadratic time? In *AISTATS*, 2025a.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3316–3333, 2021.

Miaoyu Li, Ying Fu, and Yulun Zhang. Spatial-spectral transformer for hyperspectral image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37:1, pp. 1368–1376, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021.

Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. A tighter complexity analysis of sparsegpt. *arXiv preprint arXiv:2408.12151*, 2024b.

Xiaoyu Li, Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, and Zhen Zhuang. Simulation of hypergraph algorithms with looped transformers. *arXiv preprint arXiv:2501.10688*, 2025b.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pp. 2358–2367. PMLR, 2016.

Yingyu Liang, Heshan Liu, Zhenmei Shi, Zhao Song, and Junze Yin. Conv-basis: A new paradigm for efficient attention inference and gradient computation in transformers. *arXiv preprint arXiv:2405.05219*, 2024a.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Differential privacy mechanisms in neural tangent kernel regression. *arXiv preprint arXiv:2407.13621*, 2024b.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Yufa Zhou. Multi-layer transformers gradient can be approximated in almost linear time. *arXiv preprint arXiv:2408.13233*, 2024c.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Differential privacy of cross-attention with provable guarantee. *arXiv preprint arXiv:2407.14717*, 2024d.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Tensor attention training: Provably efficient learning of higher-order transformers. *arXiv preprint arXiv:2405.16411*, 2024e.

Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, and Yufa Zhou. Beyond linear approximations: A novel pruning approach for attention matrix. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Computing and Communications*, 28(4):12–17, 2025.

Wanyu Lin, Baochun Li, and Cong Wang. Towards private learning on decentralized graphs with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:2936–2946, 2022.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pp. 21051–21064. PMLR, 2023.

Chengyi Liu, Jiahao Zhang, Shijie Wang, Wenqi Fan, and Qing Li. Score-based generative diffusion models for social recommendations. *arXiv preprint arXiv:2412.15579*, 2024a.

Han Liu, Haotian Gao, Xiaotong Zhang, Changya Li, Feng Zhang, Wei Wang, Fenglong Ma, and Hong Yu. Septq: A simple and effective post-training quantization paradigm for large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1 (KDD)*, pp. 812–823, 2025.

Tianlin Liu and Friedemann Zenke. Finding trainable sparse networks through neural tangent transfer. In *International Conference on Machine Learning*, pp. 6336–6347. PMLR, 2020.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024b.

AI @ Meta Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wenchi Ma, Tianxiao Zhang, and Guanghui Wang. Miti-detr: Object detection based on transformers with mitigatory self-attention convergence. *arXiv preprint arXiv:2112.13310*, 2021.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15762–15772, 2024.

Joseph McDonald, Baolin Li, Nathan Frey, Devesh Tiwari, Vijay Gadepally, and Siddharth Samsi. Great power, great responsibility: Recommendations for reducing energy for training language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1962–1970, 2022.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.

Tamara T Mueller, Johannes C Paetzold, Chinmay Prabhakar, Dmitrii Usynin, Daniel Rueckert, and Georgios Kaissis. Differentially private graph neural networks for whole-graph classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7308–7318, 2022.

Alireza Naderi, Thiziri Nait Saada, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in transformers. *arXiv preprint arXiv:2410.07799*, 2024.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

Tam Minh Nguyen, César A Uribe, Tan Minh Nguyen, and Richard Baraniuk. Pidformer: Transformer meets control theory. In *Forty-first International Conference on Machine Learning*, 2024.

Yotam Nitzan, Zongze Wu, Richard Zhang, Eli Shechtman, Daniel Cohen-Or, Taesung Park, and Michaël Gharbi. Lazy diffusion transformer for interactive image editing. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Von Johannes Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*. PMLR, 2023.

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. Low-bit quantization favors undertrained llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 32338–32348, 2025.

Yeonhong Park, Jake Hyun, Hojoon Kim, and Jae W Lee. {DecDEC}: A systems approach to advancing {Low-Bit}{LLM} quantization. In *19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25)*, pp. 803–819, 2025.

Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.

Sergio P Perez, Yan Zhang, James Briggs, Charlie Blake, Josh Levy-Kramer, Paul Balanca, Carlo Luschi, Stephen Barlow, and Andrew William Fitzgibbon. Training and inference of large language models using 8-bit floating point. *arXiv preprint arXiv:2309.17224*, 2023.

Ilya Razenshteyn, Zhao Song, and David P Woodruff. Weighted low rank approximations with provable guarantees. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 250–263, 2016.

Aravind Reddy, Zhao Song, and Lichen Zhang. Dynamic tensor product regression. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4791–4804, 2022.

Andreas Roth and Thomas Liebig. Rank collapse causes over-smoothing and over-correlation in graph neural networks. In *Learning on Graphs Conference*, pp. 35–1. PMLR, 2024.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*. PMLR, 2021.

Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Jing Liu, Ruiyi Zhang, Ryan A Rossi, Hao Tan, Tong Yu, Xiang Chen, et al. Numerical pruning for efficient autoregressive models. *arXiv preprint arXiv:2412.12441*, 2024.

Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Yanyu Li, Yifan Gong, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, et al. Lazydit: Lazy learning for the acceleration of diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023a.

Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. Why larger language models do in-context learning differently? In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023c.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=PxTIG12RRHS.

Zhao Song and Zheng Yu. Oblivious sketching-based central path method for linear programming. In *International Conference on Machine Learning*, pp. 9835–9847. PMLR, 2021.

Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 688–701, 2017.

Zhao Song, David Woodruff, and Peilin Zhong. Towards a zero-one law for column subset selection. *Advances in Neural Information Processing Systems*, 32, 2019a.

Zhao Song, David Woodruff, and Peilin Zhong. Average case column subset selection for entrywise ℓ_1-norm loss. *Advances in Neural Information Processing Systems*, 32, 2019b.

Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2772–2789. SIAM, 2019c.

Zhao Song, David Woodruff, Zheng Yu, and Lichen Zhang. Fast sketching of polynomial kernels of polynomial degree. In *International Conference on Machine Learning*, pp. 9812–9823. PMLR, 2021b.

Zhao Song, Mingquan Ye, Junze Yin, and Lichen Zhang. A nearly-optimal bound for fast regression with ℓ_∞ guarantee. In *International Conference on Machine Learning*, pp. 32463–32482. PMLR, 2023a.

Zhao Song, Junze Yin, and Ruizhe Zhang. Revisiting quantum algorithms for linear regressions: Quadratic speedups without data-dependent parameters. *arXiv preprint arXiv:2311.14823*, 2023b.

Zhao Song, Junze Yin, and Lichen Zhang. Solving attention kernel regression problem via preconditioner. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–216. PMLR, 2024a.

Zhao Song, Lichen Zhang, and Ruizhe Zhang. Training multi-layer over-parametrized neural network in subquadratic time. *ITCS*, 2024b.

Zhao Song, Mingquan Ye, Junze Yin, and Lichen Zhang. Efficient alternating minimization with applications to weighted low rank approximation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rvhu4V7yrX.

Niranjan Subrahmanya and Yung C Shin. Sparse multiple kernel learning for signal processing applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):788–798, 2009.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*. PMLR, 2022.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ezBH9WE9s2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Swagath Venkataramani, Anand Raghunathan, Jie Liu, and Mohammed Shoaib. Scalable-effort classifiers for energy-efficient machine learning. In *Proceedings of the 52nd annual design automation conference*, pp. 1–6, 2015.

Limei Wang, Kaveh Hassani, Si Zhang, Dongqi Fu, Baichuan Yuan, Weilin Cong, Zhigang Hua, Hao Wu, Ning Yao, and Bo Long. Learning graph quantized tokenizers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oYSsbY3G4o.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 832–841, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jianyu Wei, Shijie Cao, Ting Cao, Lingxiao Ma, Lei Wang, Yanyong Zhang, and Mao Yang. T-mac: Cpu renaissance via table lookup for low-bit llm deployment on edge. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, pp. 278–292, 2025.

Yibo Wen, Chenwei Xu, Jerry Yao-Chieh Hu, and Han Liu. Alignab: Pareto-optimal energy alignment for designing nature-like antibodies. *arXiv preprint arXiv:2412.20984*, 2024.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. Pmlr, 2019.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pp. 2501–2510, 2022.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. Onebit: Towards extremely low-bit large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:66357–66382, 2024a.

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. Towards few-shot adaptation of foundation models via multitask finetuning. In *The Twelfth International Conference on Learning Representations*, 2024b.

Jiyan Yang, Yin-Lam Chow, Christopher Ré, and Michael W Mahoney. Weighted sgd for ell_p regression with randomized preconditioning. *Journal of Machine Learning Research*, 18(211): 1–43, 2018.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. *Advances in Neural Information Processing Systems*, 36:24247–24261, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lu Yi and Zhewei Wei. Scalable and certifiable graph unlearning: Overcoming the approximation error barrier. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pPyJyeLriR.

Hao Yu, Yang Zhou, Bohua Chen, Zelan Yang, Shen Li, Yong Li, and Jianxin Wu. Treasures in discarded weights for llm quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22218–22226, 2025.

Shen Yuan and Hongteng Xu. Towards better multi-head attention via channel-wise sample permutation. *arXiv preprint arXiv:2410.10914*, 2024.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pp. 36–39. IEEE, 2019.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Amir Zandieh, Insu Han, Majid Daliri, and Amin Karbasi. Kdeformer: Accelerating transformers via kernel density estimation. In *ICML*. arXiv preprint arXiv:2302.02451, 2023.

Chao Zeng, Songwei Liu, Yusheng Xie, Hong Liu, Xiaojian Wang, Miao Wei, Shu Yang, Fangmin Chen, and Xing Mei. Abq-llm: Arbitrary-bit quantized inference acceleration for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22299–22307, 2025.

Yuansong Zeng, Zhuoyi Wei, Zixiang Pan, Yutong Lu, and Yuedong Yang. A robust and scalable graph neural network for accurate single-cell classification. *Briefings in Bioinformatics*, 23:2: bbab570, 2022.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiahao Zhang. Graph unlearning with efficient partial retraining. In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 1218–1221, 2024.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *ICLR*, 2024.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023b.

Yanfu Zhang, Shangqian Gao, Jian Pei, and Heng Huang. Improving social network embedding via new second-order continuous graph neural networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2515–2523, 2022.

Ren-Xin Zhao, Jinjing Shi, and Xuelong Li. Qksan: A quantum kernel self-attention network. *TPAMI*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 2021.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Step-back prompting enables reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Appendix

Roadmap. In Section A, we provide several simple definitions. In Section B, we show more perturbation properties for the softmax matrix. In Section C, we provide the proofs related to network layers with multiple attention heads. In Section D, we prove our main Theorem. In Section E, we provide the broader impact of this work. In Section F, we discuss the LLM usage.

A Preliminaries

Note that $\mathsf{softm}(X)$ function defined above is ignoring the effect of the weights W_v. Here we incorporate them in another function $\mathsf{softmv}(X)$ (which is also usually called self-attention).

Definition A.1. *Let W_q, W_k be weights being used in* softm. *Let W_v denote the extra weights that will be used in* softmv. *We define* $\mathsf{softmv}(X)$ *as follows*

$$\mathsf{softmv}(X) := \mathsf{softm}(X) X W_v.$$

Next we define a very useful parameter ϵ_ℓ which captures the Lipschitz and layer norm property of every layer.

Definition A.2. *Let all layers' weights are bounded, i.e, $\|W_q\|_\infty, \|W_k\|_\infty, \|W_v\|_\infty \leq \eta$. Let X_0 denote the first layer input of entire neural network and it is bounded $\|X_0\|_\infty \leq \phi_0$. Let H denote the number of heads. For each layer $\ell \in [L]$, we define a parameter $\epsilon_\ell := 2\eta\phi_0(1+H\eta)^\ell$.*

Definition A.3. *We define function $g(\epsilon) := 2(e^\epsilon - 1)$.*

B Perturbation Property of Softmax Matrix

Lemma B.1. *If the following conditions hold: Let $a, b \in \mathbb{R}^n$. Let $\|b\|_\infty \leq \epsilon$. Then we can show*

- $|\alpha(a+b)^{-1} - \alpha(a)^{-1}| \leq (e^\epsilon - 1)\alpha(a)^{-1}$
- $|\alpha(a+b)^{-1} - \alpha(a)^{-1}| \leq (e^\epsilon - 1)\alpha(a+b)^{-1}$

Proof. We can show that

$$\begin{aligned} |\alpha(a+b)^{-1} - \alpha(a)^{-1}| &= \alpha(a+b)^{-1}\alpha(a)^{-1}|\alpha(a+b) - \alpha(a)| \\ &\leq \alpha(a+b)^{-1}\alpha(a)^{-1} \cdot (e^\epsilon - 1)\alpha(a+b) \\ &= (e^\epsilon - 1)\alpha(a)^{-1} \end{aligned}$$

where the first step follows from simple algebra, the second step follows from Lemma 4.2.

Similarly, we can also show $|\alpha(a+b)^{-1} - \alpha(a)^{-1}| \leq (e^\epsilon - 1)\alpha(a+b)^{-1}$. □

Lemma B.2. *Let $a, b \in \mathbb{R}^n$. If $|b_i| \leq \epsilon$ for all $i \in [n]$, then, we can show that*

$$\|\mathsf{softm}(a+b) - \mathsf{softm}(a)\|_\infty \leq 2(e^\epsilon - 1)$$

Proof. For each $i \in [n]$, we can show

$$\begin{aligned} &|\alpha(a+b)^{-1}\exp((a+b)_i) - \alpha(a)^{-1}\exp(a_i)| \\ =\ &|\alpha(a+b)^{-1}\exp((a+b)_i) - \alpha(a+b)^{-1}\exp(a_i) + \alpha(a+b)^{-1}\exp(a_i) - \alpha(a)^{-1}\exp(a_i)| \\ \leq\ &\alpha(a+b)^{-1}|\exp(b_i+a_i) - \exp(a_i)| + \exp(a_i)\cdot|\alpha(a+b)^{-1} - \alpha(a)^{-1}| \\ :=\ &A_1 + A_2 \end{aligned}$$

where the second step follows from the triangle inequality.

We can upper bound A_1 as

$$\begin{aligned} A_1 &= \alpha(a+b)^{-1} \cdot |\exp(a_i+b_i) - \exp(a_i)| \\ &\leq \alpha(a+b)^{-1} \cdot (e^\epsilon - 1)\exp(a_i+b_i) \end{aligned}$$

$$\leq (e^\epsilon - 1)$$

where the second step follows from Lemma 4.2, the third step follows from $\alpha(x)^{-1}\exp(x_i) \in (0,1)$ for any x and i.

We can upper bound A_2 as

$$\begin{aligned} A_2 = &\exp(a_i) \cdot |\alpha(a+b)^{-1} - \alpha(a)^{-1}| \\ \leq &\exp(a_i) \cdot (e^\epsilon - 1)\alpha(a)^{-1} \\ \leq &e^\epsilon - 1 \end{aligned}$$

where the second step follows from Lemma B.1, and the third step follows from $\alpha(x)^{-1}\exp(x_i) \in (0,1)$ for any x and i.

Putting everything together, we can show

$$\begin{aligned} \|\mathsf{softm}(a+b) - \mathsf{softm}(a)\|_\infty = &\max_{i\in[n]} |\alpha(a+b)^{-1}(a+b)_i - \alpha(a)^{-1}a_i| \\ \leq &2(e^\epsilon - 1). \end{aligned}$$

Thus, we complete the proof. □

Lemma B.3. *If the following conditions hold*

- *Let $P = \mathsf{softm}(A)$ (see Definition 3.1 for function $\mathsf{softm}()$).*
- *Let $\widetilde{A} = A - E$.*
- *Let $\widetilde{P} = \mathsf{softm}(\widetilde{A})$.*
- *Let D be defined as Definition 3.5, i.e., $D_{i,i} := \max_{j,l\in[n]} |E_{i,j} - E_{i,l}|$*

Then we can show, for all $i \in [n], j \in [n]$

$$e^{-D_{i,i}}\widetilde{P}_{i,j} \leq P_{i,j} \leq e^{D_{i,i}}\widetilde{P}_{i,j}.$$

Proof. Let us start by the definition of P, for each $i \in [n]$, $j \in [n]$

$$\begin{aligned} P_{i,j} = &(\mathsf{softm}(A))_{i,j} \\ = &(\mathsf{softm}(\widetilde{A} + E))_{i,j} \\ = &\frac{\exp(\widetilde{A}_{i,j} + E_{i,j})}{\sum_{l=1}^n \exp(\widetilde{A}_{i,l} + E_{i,l})} \\ = &\frac{\exp(\widetilde{A}_{i,j}))}{\sum_{l=1}^n \exp(\widetilde{A}_{i,l})\exp(E_{i,l} - E_{i,j})} \end{aligned}$$

where the first step follows from definition of P, the second step follows from definition of A, the third step follows from definition of softmax (Definition 3.1), and the last step follows from property of $\exp$.

We define $D_{i,i} := \max_{j,l\in[n]} |E_{i,j} - E_{i,l}|$. We have that

$$P_{i,j} \in [\widetilde{P}_{i,j}\exp(-D_{i,i}), \widetilde{P}_{i,j}\exp(D_{i,i})]$$

Thus, we complete the proof. □

Lemma B.4. *If the following conditions hold*

- *Let W_q, W_k, W_v be the matrix that $\|W_q\|_\infty, \|W_k\|_\infty, \|W_v\|_\infty \leq \eta$.*
- *Let $W = W_q W_k^\top$.*
- *Let $E = \beta\,\mathsf{Res}(X) W\,\mathsf{Res}(X)^\top$.*

- *Let β satisfy that $\beta \leq 1/(\| \mathsf{Res}(X)\|_\infty^2 \eta^2)$.*

Then, we have E is θ-balanced with $\theta = 1$.

Proof. First, note that $\|W\|_\infty \leq \|W_q\|_\infty \cdot \|W_k\|_\infty \leq \eta^2$.

We can show

$$\begin{aligned} \|E\|_\infty &\leq \beta \cdot \| \mathsf{Res}(X)\|_\infty^2 \cdot \|W\|_\infty \\ &\leq \beta \cdot \| \mathsf{Res}(X)\|_\infty^2 \cdot \eta^2 \\ &\leq 1 \end{aligned}$$

Thus, we complete the proof. □

C Multiple Heads

Here, we generalize the proof of Lemma 5.2 to multiple heads. Note that Lemma 5.2 presented a simplified proof by ignoring the effects of XW_v, and thus automatically assuming $n = d$. In this section we remove that condition and prove the result for general n and d. In Section C.1, our goal is to prove Lemma C.1, which is the multiple heads version of Lemma 5.2. In Section C.2, we show that several required conditions in Lemma C.1 are satisfied.

C.1 Multiple Heads For Skipping One Layer

In the next Lemma C.1, we will put the effect of softmv back. We remark that the major idea of the proof remains the same as Lemma 5.2.

Lemma C.1 (Multiple Heads version of Lemma 5.2)**.** *If the following conditions hold,*

- *Let H denote the number of heads.*
- *Note that softm_1 and softm_2 are two different instantiations with different W_k, W_q, W_v weights.*
- *Let $X \in \mathbb{R}^{n \times d}$.*
- *$A_i = \mathsf{softmv}_{1,i}(X) \in \mathbb{R}^{n \times d}$ for $i \in [H]$. (Let softmv be defined as Definition A.1)*
- *$B = X + \sum_{i=1}^H A_i \in \mathbb{R}^{n \times d}$.*
- *$\| \mathsf{Res}(A_i)\|_\infty \leq K \cdot \| \mathsf{Res}(X)\|_\infty \leq \epsilon$ for all $i \in [H]$. (We remark that this condition will hold due to Lemma 5.1; here K is as defined in Lemma 5.1)*
- *Let $g(\epsilon) := 2(e^\epsilon - 1)$ (see Definition A.3).*
- *Let W_v satisfy that $\|(B - X)W_v\|_\infty \leq H \cdot \epsilon$ and $\|XW_v\|_\infty \leq 1$*
- *Let $\epsilon_0 = 3g(2H\epsilon)$*

Then we can show

- **Part 1.** $\| \mathsf{softm}_2(B) - \mathsf{softm}_2(X)\|_\infty \leq \epsilon_0$
- **Part 2.** $\| \mathsf{softmv}_2(B) - \mathsf{softmv}_2(X)\|_\infty \leq \epsilon_0$

Proof. **Proof of Part 1.**

Let $R_X = \mathsf{Res}(X)$ so that $X = R_X + y_X \mathbf{1}_d^\top$ for some vector $y_X \in \mathbb{R}^n$.

Using Fact 3.2, we can show that

$$\mathsf{softm}_1(X) = \mathsf{softm}_1(R_X). \tag{12}$$

To notataionaly help in our proof, we define the prefix sums of matrices $A_0, A_1, \cdots, A_i \in \mathbb{R}^{n \times d}$ as

$$A_{[i]} := \sum_{j=0}^{i} A_i$$

where A_0 is an artificial matrix that has 0 everywhere.

For each $i \in [H]$, let $R_{A_{[i]}} = \mathsf{Res}(A_{[i]})$ so that $A_{[i]} = R_{A_{[i]}} + y_{A_{[i]}} \mathbf{1}_d^\top$ for some vector $y_{A,[i]} \in \mathbb{R}^n$. Using Fact 3.2, we can show that

$$\mathsf{softm}(A_{[i]}) = \mathsf{softm}(R_{A_{[i]}}).$$

Let $R_B = \mathsf{Res}(B)$ so that $B = R_B + y_B \mathbf{1}_d^\top$ for some vector $y_B \in \mathbb{R}^n$. Using Fact 3.2, we can show that

$$\mathsf{softm}_2(B) = \mathsf{softm}_2(R_B). \tag{13}$$

Let us consider the vector version $\| \mathsf{softm}(a+b) - \mathsf{softm}(a) \|_\infty$. Note that if $\|b\|_\infty \leq \epsilon$, then using Lemma B.2, we can show

$$\| \mathsf{softm}(a+b) - \mathsf{softm}(a) \|_\infty \leq g(\epsilon)$$

Since $\|R_{A_{[i]}} - R_{A_{[i-1]}}\|_\infty \leq \epsilon$ for all $i \in [H]$, then using Lemma B.2, we have: for each $i \in [H]$

$$\| \mathsf{softm}_2(R_X + R_{A_{[i]}}) - \mathsf{softm}_2(R_X + R_{A_{[i-1]}}) \|_\infty \leq g(\epsilon) \tag{14}$$

We can show $R_X = \mathsf{Res}(X) = \mathsf{Res}(B - A_{[H]}) = \mathsf{Res}(B - R_{A_{[H]}})$.

Then, we know

$$\begin{aligned} \|R_X - R_B\|_\infty &= \| \mathsf{Res}(B - R_{A_{[H]}}) - \mathsf{Res}(B) \|_\infty \\ &\leq \|R_{A_{[H]}}\|_\infty \end{aligned} \tag{15}$$

where the first step follows from $R_X = \mathsf{Res}(B - R_{A_{[H]}})$ and $R_B = \mathsf{Res}(B)$.

Recall $B = X + A$, and $\|R_A\|_\infty \leq \epsilon$ then we know

$$\begin{aligned} \|R_X + R_A - R_B\|_\infty &\leq \|R_X - R_B\|_\infty + \|R_{A_{[H]}}\|_\infty \\ &\leq 2\|R_{A_{[H]}}\|_\infty \\ &\leq 2H\epsilon, \end{aligned}$$

where the first step follows from triangle inequality, the second step follows from $\|R_X - R_B\|_\infty \leq \|R_{A_{[H]}}\|_\infty$, and the last step follows from $\|R_{A_{[H]}}\|_\infty \leq H\epsilon$.

Since $\|R_X + R_{A,H} - R_B\|_\infty \leq 2H\epsilon$, then using Lemma B.2, we have

$$\| \mathsf{softm}_2(R_X + R_{A,H}) - \mathsf{softm}_2(R_B) \|_\infty \leq g(2H\epsilon) \tag{16}$$

Then, we can show

$$\begin{aligned} & \| \mathsf{softm}_2(B) - \mathsf{softm}_2(X) \|_\infty \\ = & \| \mathsf{softm}_2(B) - \mathsf{softm}_2(R_X) \|_\infty \\ = & \| \mathsf{softm}_2(R_B) - \mathsf{softm}_2(R_X) \|_\infty \\ \leq & \| \mathsf{softm}_2(R_B) - \mathsf{softm}_2(R_X + R_{A,H}) \|_\infty \\ & + \sum_{i=1}^{H-1} \| \mathsf{softm}_2(R_X + R_{A,i}) - \mathsf{softm}_2(R_X + R_{A,i-1}) \|_\infty \\ \leq & g(2H\epsilon) + H \cdot g(\epsilon) \\ \leq & 2g(2H\epsilon) \end{aligned} \tag{17}$$

where the first step follows from $\mathsf{softm}(X) = \mathsf{softm}(R_X)$ (see Eq. (12)), the second step follows from $\mathsf{softm}(R_B) = \mathsf{softm}(B)$ (see Eq. (13)), the third step follows from triangle inequality, the forth step follows from Eq. (14) and Eq. (16), and the last step follows from property of function g.

Proof of Part 2.

We can show that

$$\begin{aligned}
& \| \mathsf{softmv}_2(B) - \mathsf{softmv}_2(X) \|_\infty \\
= & \| \mathsf{softm}_2(B) B W_v - \mathsf{softm}_2(X) X W_v \|_\infty \\
\leq & \| \mathsf{softm}_2(B) B W_v - \mathsf{softm}_2(B) X W_v \|_\infty + \| \mathsf{softm}_2(B) X W_v - \mathsf{softm}_2(X) X W_v \|_\infty \\
\leq & \| \mathsf{softm}_2(B) \|_\infty \cdot \| (B - X) W_v \|_\infty + \| \mathsf{softm}_2(B) - \mathsf{softm}_2(X) \|_\infty \cdot \| X W_v \|_\infty \\
\leq & H\epsilon + g(2H\epsilon) + H \cdot g(\epsilon) \\
\leq & 2g(2H\epsilon)
\end{aligned}$$

where the second step follows from triangle inequality, the third step follows from Fact 3.3, and the forth step follow from Eq. (17) , where the last step follows from $\epsilon \leq g(\epsilon)$ and $2Hg(\epsilon) \leq g(2H\epsilon)$. □

C.2 Conditions in Lemma C.1 are Satisfied

Here we will show that the three conditions in Lemma C.1 will be satisfied for each layer ℓ.

- $\| \mathsf{Res}(A_i) \|_\infty \leq K \cdot \| \mathsf{Res}(X) \|_\infty \leq \epsilon$ (where $K := (e^\theta - 1)\|W_v\|_\infty$, definition of K recall Lemma 5.1). Here $\theta = 1$ due to Lemma B.4
- $\|(B - X)W_v\|_\infty \leq H \cdot g(\epsilon)$
- $\|XW_v\|_\infty \leq 1$

Lemma C.2. *If the following conditions hold*

- $\epsilon_\ell := 2\eta\phi_0(1 + H\eta)^\ell$. *(see Definition A.2)*
- *Let* $\eta \in (0, 1]$.
- *Let* $\epsilon_\ell \in (0, 1)$.

Then, we can show

- *Part 1.* $K \cdot \| \mathsf{Res}(X_\ell) \|_\infty \leq \epsilon_\ell$
- *Part 2.* $\|(B - X_\ell)W_v\|_\infty \leq H\epsilon_\ell$
- *Part 3.* $\|X_\ell W_v\|_\infty \leq 1$

Proof. **Proof of Part 1.**

We can show that

$$\begin{aligned}
K \cdot \| \mathsf{Res}(X_\ell) \|_\infty & \leq 2\eta \| \mathsf{Res}(X_\ell) \|_\infty \\
& \leq 2\eta \|X_\ell\|_\infty \\
& \leq 2\eta\phi_0 \cdot (1 + H\eta)^\ell \\
& = \epsilon_\ell
\end{aligned}$$

where the first step follows from $\theta = 1$ and $\|W_v\|_\infty \leq \eta$, the third step follows from Lemma D.8, and the last step follows from definition ϵ_ℓ

Proof of Part 2.

$$\begin{aligned}
\|(B - X_\ell)W_v\|_\infty & \leq \eta \cdot \|B - X_\ell\|_\infty \\
& = \eta \cdot \sum_{i=1}^{H} \| \mathsf{softmv}_i(X_\ell) \|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\leq \eta \cdot \sum_{i=1}^{H} \| \mathsf{softm}_i(X_\ell) X_\ell W_{v,i} \|_\infty \\
&\leq \eta \cdot \sum_{i=1}^{H} \| \mathsf{softm}_i(X_\ell) \|_\infty \cdot \|X_\ell\|_\infty \cdot \|W_{v,i}\|_\infty \\
&\leq \eta^2 H \cdot \|X_\ell\|_\infty \\
&\leq 0.5\eta H \cdot \epsilon_\ell \\
&\leq H\epsilon_\ell
\end{aligned}
$$

where the first step follows from $\|W_v\|_\infty \eta$, the second step follows from definition of B, the third step follows from definition of softmv, the forth step follows from Fact 3.3, the fifth step follows from $\| \mathsf{softm}() \|_\infty \leq 1$ and $\|W_{v,i}\|_\infty \eta$, and the last step follows from $\eta \leq 1$.

Proof of Part 3.

We can show

$$
\begin{aligned}
\|X_\ell W_v\|_\infty &\leq \eta \|X_\ell\|_\infty \\
&\leq \eta \phi_0 (1 + H\eta)^\ell \\
&\leq \epsilon_\ell \\
&\leq 1
\end{aligned}
$$

where the second step follows from Lemma D.8, the third step follows from choice of ϵ_ℓ last step follows from Lemma statement condition.

□

D Multiple Layers

In Section D.1, we provide the proof of our main theorem. In Section D.2, we provide the Lipshitz property of several key functions being used in our proofs. In Section D.3, we prove the Lipschitz property for each layer of our Self Attention Network. Finally, in Section D.4, prove the norm of each layer in the Self Attention Network is not increasing much.

D.1 Proof of Theorem D.1

Theorem D.1 (Formal version of Theorem 1.2)**.** *Suppose S is a SAtt with residuals, with the property that for every attention head in every one of its layers, the weight matrices $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ all have the bound $\|W_q\|_\infty, \|W_k\|_\infty, \|W_v\|_\infty \leq \eta$. Let H denote the number of heads. Let L denote the number of layers. Assume $\eta \leq A \cdot \min\{1/(HL), 1/\phi_0\} \leq 1$ for some parameter $A \leq O(1)$. Then, there exists a SAtt with residuals S' with just one layer so that, for any bounded $X \in \mathbb{R}^{n \times d}$ with $\|X\|_\infty \leq \phi_0$, we have $\|S(X) - S'(X)\|_\infty \leq O(A/L)$.*

Proof. We define

$$
X_\ell^{\ell_0} = B_\ell^{\ell_0}
$$

Then, we define

$$
B_\ell^{\ell_0} = \begin{cases} X_{\ell-1}^{\ell_0} + \sum_{i=1}^{H} A_{\ell-1,i}^{\ell_0}, & \text{if} \ell \leq \ell_0; \\ \sum_{i=1}^{H} A_{\ell-1,i}^{\ell_0} & \text{otherwise.} \end{cases}
$$

Let $\mathsf{softmv}()$ function be defined as Defintion A.1. We define

$$
A_{\ell-1,i}^{\ell_0} = \mathsf{softmv}_{\ell-1,i}^{\ell_0}(X_{\ell-1}^{\ell_0})
$$

Note that the notation B_L^0 means we have residual in every layer, whereas the notation $B_L^{\ell_0}$ means we don't have a residual connection from layer ℓ_0 to layer L.

Let ϵ_ℓ be defined as Definition A.2. Let $\delta := \max_{\ell \in [L]} 2g(2H\epsilon_\ell)$. Using Lemma C.1, we can show for all $\ell \in [L]$,

$$\| \mathsf{softmv}_\ell(B_\ell^{\ell-1}) - \mathsf{softmv}_\ell(B_\ell^\ell) \|_\infty \leq \delta$$

Let $C := \max_{\ell \in [L]} 3\eta(\epsilon_\ell^2 + 1)$.

Then we can show

$$\begin{aligned}
& \| \mathsf{softmv}_2(B_2^0) - \mathsf{softmv}_2(B_2^2) \|_\infty \\
\leq & \| \mathsf{softmv}_2(B_2^0) - \mathsf{softmv}_2(B_2^1) \|_\infty + \| \mathsf{softmv}_2(B_2^1) - \mathsf{softmv}_2(B_2^2) \|_\infty \\
\leq & C \cdot \| \mathsf{softmv}_1(B_1^0) - \mathsf{softmv}_1(B_1^1) \|_\infty + \| \mathsf{softmv}_2(B_2^1) - \mathsf{softmv}_2(B_2^2) \|_\infty \\
\leq & (C+1)\delta
\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from the fact that one layer of the network is C-Lipschitz (see Lemma D.6), and the last step follows from merging the errors.

For three layers, we have

$$\begin{aligned}
& \| \mathsf{softmv}_3(B_3^0) - \mathsf{softmv}_3(B_3^3) \|_\infty \\
\leq & \| \mathsf{softmv}_3(B_3^0) - \mathsf{softmv}_3(B_3^1) \|_\infty \\
& + \| \mathsf{softmv}_3(B_3^1) - \mathsf{softmv}_3(B_3^2) \|_\infty \\
& + \| \mathsf{softmv}_3(B_3^2) - \mathsf{softmv}_3(B_3^3) \|_\infty \\
\leq & C^2 \cdot \| \mathsf{softmv}_1(B_1^0) - \mathsf{softmv}_1(B_1^1) \|_\infty \\
& + C \cdot \| \mathsf{softmv}_2(B_2^1) - \mathsf{softmv}_2(B_2^2) \|_\infty \\
& + \| \mathsf{softmv}_3(B_3^2) - \mathsf{softmv}_3(B_3^3) \|_\infty \\
\leq & C^2\delta + C\delta + \delta \\
= & (C^2 + C + 1)\delta
\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from one layer of network is C-Lipshitz (see Lemma D.6), and the forth step follows from Lemma C.1, and the last step follows from merging the errors.

Therefore for L layers we have

$$\| \mathsf{softmv}_L(B_L^0) - \mathsf{softmv}_L(B_L^L) \|_\infty \leq (C^L + \cdots + C + 1)\delta$$

Thus we complete the proof.

Now, we are ready to analyze the final bound, Recall that above we have $\epsilon_\ell = 2\phi_0\eta(1+H\eta)^\ell \in (0,1)$, $\delta = \max_\ell 2g(2H\epsilon_\ell)$, and $C = \max_\ell 3\eta(\epsilon_\ell^2+1)$.

Recall that $\eta \leq A \cdot \min\{1/(HL), 1/\phi_0\} \leq 1$ for some parameter $A \leq O(1)$. Then we can show $\epsilon_\ell = 2\phi_0\eta(1+H\eta)^\ell = 2\phi_0\eta e^{\ell H\eta} \leq 2\phi_0\eta e^A < 1$. Next, we can show that compute $C \leq 3\eta(\epsilon_\ell^2+1) \leq 3\eta \cdot 2 < 0.5$. Thus $C^L + \cdots + C + 1 \leq 2$.

Note that $2H\epsilon_\ell \in (0, 0.5)$ $\delta = 2(e^{2H\epsilon_\ell} - 1) \leq 8H\epsilon_\ell \leq AH \cdot \min\{1/HL, 1/\phi_0\}$

The final is $2\delta \leq 2AH \cdot \min\{1/HL, 1/\phi_0\} \leq 2A/L$.

□

Remark D.2. *We remark that our proof can be straightforwardly generalized to the situation where the Self Attention Network also has MLP layers, similar to Section 3.2 in (Dong et al., 2021), by defining $X_\ell^{\ell_0} = f(B_\ell^{\ell_0})$ where f is the MLP layer. Note that the Lipshitz property of f will appear correspondingly in the final bound.*

Remark D.3. *Note that in our Theorem D.1, we assume that $\eta \leq O(\min\{1/(HL), 1/\phi_0\})$. Meanwhile, the prior work (Dong et al., 2021, Corollary 2.3) similarly requires $\beta \leq O(\sqrt{d}/(H\phi_0))$.*

Recall that β is in terms of the ℓ_1 norm, and so may be up to a factor of d^2 larger than our η; their factor of $\sqrt{d}$ only modestly helps with this.

(Their result statement is in terms of $\|\mathsf{Res}(X)\|_\infty$ rather than $\|X\|_\infty$, but we could state ours in terms of $\mathsf{Res}(X)$ instead; we simply bound $\|\mathsf{Res}(X)\|_\infty \leq \|X\|_\infty$ in the proof of our Lemma C.2.)

D.2 Lipschitz Property

We state a simple application of Lemma B.2.

Corollary D.4. *Let $a, b \in \mathbb{R}^n$. Then, we can show that*

$$\|\mathsf{softm}(a+b) - \mathsf{softm}(a)\|_\infty \leq 2(e^{\|b\|_\infty} - 1)$$

Proof. The proof is same as Lemma B.2. □

Lemma D.5. *Let $a, b \in \mathbb{R}^n$. If $\|b\|_\infty \leq 1$, then we have*

$$\|\mathsf{softm}(a+b) - \mathsf{softm}(b)\|_\infty \leq 4\|b\|_\infty$$

Proof. Note that for $x \in (0, 1]$, we know $e^x - 1 \leq 2x$.

Thus, we know

$$\begin{aligned} \|\mathsf{softm}(a+b) - \mathsf{softm}(b)\|_\infty &\leq 2(e^{\|b\|_\infty} - 1) \\ &\leq 4\|b\|_\infty \end{aligned}$$

where the first step follows from Corollary D.4, the second step follows from $e^x - 1 \leq 2x$. □

Lemma D.6. *If the following conditions hold*

- *Let W_q, W_k, W_v denote weight matrices.*
- *Let $W = W_q W_k^\top$.*
- *Let Y satisfy that $\|Y - X\|_\infty \leq 2\|X\|_\infty$*
- *$K_1 := 12\|X\|_\infty \|W\|_\infty$.*
- *$K_2 := K_1 \|X\|_\infty \|W_v\|_\infty + \|W_v\|_\infty$*

Then, we can show

- **Part 1.**

$$\|\mathsf{softm}(X) - \mathsf{softm}(Y)\|_\infty \leq K_1 \cdot \|X - Y\|_\infty$$

- **Part 2.**

$$\|\mathsf{softmv}(X) - \mathsf{softmv}(Y)\|_\infty \leq K_2 \cdot \|X - Y\|_\infty$$

Proof. **Proof of Part 1.** We can show

$$\begin{aligned} & \|\mathsf{softm}(X) - \mathsf{softm}(Y)\|_\infty \\ \leq & 4\|XWX^\top - YWY^\top\|_\infty \\ \leq & 4\|XWX^\top - XWY^\top\|_\infty + 4\|XWY^\top - YWY^\top\|_\infty \\ \leq & 4\|XW\|_\infty \cdot \|X - Y\|_\infty + 4\|WY^\top\|_\infty \cdot \|X - Y\|_\infty \\ \leq & 4 \cdot (\|WX\|_\infty + \|WY\|_\infty) \cdot \|X - Y\|_\infty \\ \leq & 12\|W\|_\infty \|X\|_\infty \|X - Y\|_\infty \end{aligned}$$

where the first step follows from Lemma D.5, the second step follows triangle inequality, the third step follows from Fact 3.3, the last step follows from Fact 3.3.

Proof of Part 2. We can show

$$\begin{aligned}
&\| \mathsf{softmv}(X) - \mathsf{softm}(Y) \|_\infty \\
= &\| \mathsf{softm}(X) X W_v - \mathsf{softm}(Y) Y W_v \|_\infty \\
\leq &\| \mathsf{softm}(X) X W_v - \mathsf{softm}(Y) X W_v \|_\infty + \| \mathsf{softm}(Y) X W_v - \mathsf{softm}(Y) Y W_v \|_\infty \\
\leq &\| \mathsf{softm}(X) - \mathsf{softm}(Y) \|_\infty \cdot \|X W_v\|_\infty + \|(X-Y) W_v\|_\infty \\
\leq &K_1 \|X W_v\|_\infty \|X - Y\|_\infty + \|W_v\|_\infty \|X - Y\|_\infty
\end{aligned}$$

where the first step follows from definition, the second step follows from triangle inequality, the third step follows from Fact 3.3, and the last step follows from Part 1 and Fact 3.3. □

D.3 Instantiating an Instance for Each Layer Lipschitiz Property

Lemma D.7. *If the following conditions hold*

- *Let X_ℓ denote ℓ-th layer output*
- *Let $\|W_q\|_\infty, \|W_k\|_\infty, \|W_v\|_\infty \leq \eta$*
- *Let Y satisfy that $\|Y - X_\ell\|_\infty \leq 2\|X_\ell\|_\infty$*
- *$\epsilon_\ell := 2\eta\phi_0(1+H\eta)^\ell$.*

Then, we can show

- $\| \mathsf{softmv}(X_\ell) - \mathsf{softmv}(Y) \|_\infty \leq 3\eta(\epsilon_\ell^2 + 1)$

Proof. We can show

$$\| \mathsf{softmv}(X_\ell) - \mathsf{softmv}(Y) \|_\infty \leq K_2 \cdot \|X - Y\|_\infty$$

We just need to upper bound K_2

$$\begin{aligned}
K_2 = &K_1 \|X_\ell\|_\infty \|W_v\|_\infty + \|W_v\|_\infty \\
\leq &12 \|X_\ell\|_\infty^2 \|W\|_\infty \|W_v\|_\infty + \|W_v\|_\infty \\
\leq &12 \|X_\ell\|_\infty^2 \eta^3 + \eta \\
\leq &12(\phi_0 \cdot (1+H\eta))^2 \eta^3 + \eta \\
= &3\eta(\epsilon_\ell^2 + 1)
\end{aligned}$$

where the first step follows from the definition of K_2, the forth step follows from Lemma D.8, and the fifth step follows from the definition of ϵ_ℓ. □

D.4 Each Layer Norm is not Increasing Much

Lemma D.8. *If the following conditions hold*

- *Let X_0 denote the input of first layer of neural network, and satisfy $\|X_0\|_\infty \leq \phi_0$*
- *For $\ell \in [L]$, we use X_ℓ to denote the ℓ-th layer output*
- *Let $\|W_v\|_\infty \leq \eta$*

Then, we can show

- *Part 1. For any ℓ, $\|X_{\ell+1}\|_\infty \leq \|X_\ell\|_\infty \cdot (1 + H\eta)$*
- *Part 2. For any ℓ, $\|X_\ell\|_\infty \leq \phi_0 \cdot (1+H\eta)^\ell$*

Proof. **Proof of Part 1.**

For any ℓ, we have

$$\begin{aligned}\|X_{\ell+1}\|_\infty &= \|X_\ell + \sum_{i=1}^H \mathsf{softmv}_i(X_\ell)\|_\infty \\ &\leq \|X_\ell\|_\infty + H \cdot \|\mathsf{softmv}_i(X_\ell)\|_\infty \\ &= \|X_\ell\|_\infty + H \cdot \|\mathsf{softm}_i(X_\ell) X_\ell W_{v,i}\|_\infty \\ &\leq \|X_\ell\|_\infty + H \cdot \|\mathsf{softm}_i(X_\ell)\|_\infty \cdot \|X_\ell\|_\infty \cdot \|W_{v,i}\|_\infty \\ &\leq \|X_\ell\|_\infty (1 + H\eta)\end{aligned}$$

where the first step follows from definition of X_1, the second step follows from triangle inequality, the third step follows from definition of softmv, the forth step follows from Fact 3.3, and the last step follows from $\|\mathsf{softm}()\|_\infty \leq 1$ and $\|W_{v,i}\|_\infty \leq 1$. (Here $W_{v,i} \in \mathbb{R}^{n \times d}$ denotes the weight matrix, W_v, for the i-th head.)

Proof of Part 2.

We can show

$$\begin{aligned}\|X_\ell\|_\infty &\leq \|X_{\ell-1}\|_\infty (1 + H\eta) \\ &\leq \cdots \\ &\leq \|X_0\|_\infty (1 + H\eta)^\ell \\ &\leq \phi_0 \cdot (1 + H\eta)^\ell\end{aligned}$$

where the first step follows from Part 1, the third step follows from recursively applying Part 1, and the last step follows from $\|X_0\|_\infty \leq \phi_0$.

Therefore, we complete the proof. □

E Broader Impact

Our results offer new theoretical insights into the expressiveness of attention mechanisms in transformers. These findings may guide the future design of large language models toward more expressive architectures. We do not foresee any potential negative societal impacts from this work.

F LLM Usage Disclosure

LLMs were used only to polish language, such as grammar and wording. These models did not contribute to idea creation or writing, and the authors take full responsibility for this paper's content.