On the Existence of Hidden Subnetworks Within a Randomly Weighted Multi-Head Attention Mechanism

Hikari Otsuka OTSUKA.HIKARI@ARTIC.IIR.ISCT.AC.JP
Institute of Science Tokyo

Yasuyuki Okoshi OKOSHI.YASUYUKI@ARTIC.IIR.ISCT.AC.JP
Institute of Science Tokyo

Daichi Fujiki DFUJIKI@ARTIC.IIR.ISCT.AC.JP
Institute of Science Tokyo

Susumu Takeuchi SUSUMU.TAKEUCHI@NTT.COM
NTT Corporation

Masato Motomura MOTOMURA@ARTIC.IIR.ISCT.AC.JP
Institute of Science Tokyo

Daiki Chijiwa DAIKI.CHIJIWA@NTT.COM
NTT Corporation

Abstract

The strong lottery ticket hypothesis (SLTH) conjectures that high-performing subnetworks are hidden in randomly initialized neural networks. Although recent theoretical works have established the existence of such subnetworks across various neural architectures, the existence of SLTs in transformer architectures has only been observed empirically and lacks theoretical understanding. In particular, the current SLTH theory does not yet account for the multi-head self-attention (MHA) mechanism, a core component of transformers. To address this gap, we introduce a theoretical analysis of the existence of SLTs within the attention mechanism. Given H heads, we prove that an arbitrary target MHA can be approximated by suitably pruning the randomly initialized MHA with the key and value dimensions $O(d\log(Hd^{3/2}))$, where d is the dimension of the input and output. We further empirically validate our theoretical findings, demonstrating that an SLT within a random MHA of logarithmically wider hidden dimensions can approximate the performance of trained counterparts.

1. Introduction

The *lottery ticket hypothesis* [7]—overparameterized networks should contain subnetworks that achieve comparable accuracy to fully trained networks even if trained in isolation—presented new possibilities for compact and high-performing models in deep neural networks. Subsequent works [18, 23] proposed a stronger claim, which is formally defined as the *strong lottery ticket hypothesis* (SLTH) [13]: overparameterized networks should contain subnetworks that achieve high accuracy comparable to the trained dense network even without any training. Unraveling these hypotheses is important for a deeper understanding of the intrinsic nature of overparameterized neural networks.

The first rigorous proof for the existence of such subnetworks was given by Malach et al. [13]. They proved that, given a fully-connected network (a *target network*), there exists an SLT that ap-

© H. Otsuka, Y. Okoshi, D. Fujiki, S. Takeuchi, M. Motomura & D. Chijiwa.

proximates the target network, in a randomly-weighted fully-connected network of the sufficient width and depth (a *source network*). Afterwards, this overparameterization requirement for the source network has been relaxed [2, 15, 17]. Pensia et al. [17] concluded that the logarithmic overparameterization is approximately optimal in the case of fully-connected networks by utilizing the subset-sum approximation [12]. Then, the SLTH has been extended both theoretically and empirically to more complex architectures such as convolutional, residual, and equivariant networks [1, 3, 4, 6].

However, its applicability to *transformers*, which form the basis of modern language models, has been only empirically observed [11, 16, 20] and remains unexplored theoretically. By the definition of Vaswani et al. [21], a transformer is mainly constructed with residual connections, fully-connected networks, and attention mechanisms. Based on previous studies on residual and fully-connected networks [1, 17], we can say that the SLTH is already partially established in transformers; thus, the last piece left to prove the existence of SLTs in transformers is the attention mechanism. While attention mechanisms require second-order computation via inner product between input vectors, the existing SLTH theories have discussed only first-order computation (i.e., operations without any product between input vectors, such as linear operations), which makes it non-trivial whether SLTs theoretically exist within transformers. This gap motivates our key research question: *how can we extend the theory of the SLTH to the attention mechanisms?*

This work addresses this open problem by providing a proof of the existence of SLTs within attention mechanisms. The main idea for our proof is to view the inner product in the target attention as a quadratic form. This form contains the single matrix defined by the key and query projection matrices of the target attention. Consequently, we can approximate this single matrix by the product of (suitably pruned) key and query projection matrices in the source attention, leading to the approximation of the target attention by the pruned source attention. As shown in the following statement, we prove that any multi-head self-attention mechanism can be approximated by pruning a randomly initialized attention mechanism of the logarithmically wider hidden dimensions (i.e., the query, key, value, and output dimensions).

Theorem 1 (informal) *Given T tokens as inputs, a suitably pruned randomly initialized attention mechanism of H heads and hidden dimension $O(d\log(Hd^{3/2}/\epsilon))$ can approximate any multi-head attention of H heads and hidden dimension d_{K}, with probability at least $1-\epsilon$.*

We also empirically demonstrate the justification of Theorem 1, namely, the relationship between the hidden dimension and the approximation error holds on the order of $O(d\log(Hd^{3/2}/\epsilon))$.

Our contributions are summarized as follows:

- We provide the first theoretical proof that SLTs exist within the source attention mechanism based on the formula transformation of the target attention.
- We then empirically validate the justification of Theorem 1. Our experiment shows that the approximation error ϵ decreases exponentially with increasing the source hidden dimensions.

2. Background: Strong Lottery Ticket Hypothesis

The strong lottery ticket hypothesis (SLTH) [13, 18] conjectured that a randomly initialized network inherently contains subnetworks that achieve high accuracy comparable to trained dense networks, without any weight updates. The first theoretical results of the SLTH was by Malach et al. [13], who

proved the existence of such SLTs in fully connected ReLU networks. Subsequent works [2, 15, 17] relaxed the architectural requirements for containing such subnetworks. In particular, Pensia et al. [17] showed that SLTs exist in networks of the double depth and logarithmically wider width relative to the target function by applying a subset-sum approximation technique [12].

Lemma 2 *Given $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$, let $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$ and $g(\boldsymbol{x}) = \tilde{\boldsymbol{W}}_2 \mathrm{ReLU}(\tilde{\boldsymbol{W}}_1(\boldsymbol{x}))$ be target and source networks, respectively. Assume that $\|\boldsymbol{W}\| \leq 1$, $\|\boldsymbol{x}\| \leq 1$, and each entry of $\tilde{\boldsymbol{W}}_1 \in \mathbb{R}^{n \times d_1}$ and $\tilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_2 \times n}$ is drawn i.i.d. from $U[-1, 1]$. Then, if the intermediate dimension n satisfies $n \geq d_1 C \log(2d_1 d_2/\epsilon)$, with probability at least $1 - \epsilon$, there exist binary masks $\boldsymbol{M}_1 \in \{0,1\}^{n \times d_1}$ and $\boldsymbol{M}_2 \in \{0,1\}^{d_2 \times n}$ such that*

$$\|\boldsymbol{W}\boldsymbol{x} - (\tilde{\boldsymbol{W}}_2 \odot \boldsymbol{M}_2)\mathrm{ReLU}((\tilde{\boldsymbol{W}}_1 \odot \boldsymbol{M}_1)\boldsymbol{x})\| \leq \epsilon.$$

This approach, which approximates a single matrix using two matrices by the subset-sum approximation, is now foundational for theoretical SLTH analyses across architectural variants [1, 3, 4, 6].

3. SLT Existence Within Attention Mechanisms

This section analyzes the existence of SLTs within multi-head self-attention mechanisms (MHAs). For a notation and detailed proof, see Appendix A.

3.1. Setups for Main Theorem

Input Tokens: We consider MHAs with inputs of length T. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times d_1}$ denote the input token embedding matrix, where each token embedding vector $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ satisfies $\|\boldsymbol{x}_i\| \leq \alpha$ for some constant $\alpha > 0$. For each token, let $\boldsymbol{a}_i \in \{0,1\}^T$ denote its attention mask, where $a_{i,j} = 1$ indicates that the i-th token attends to the j-th token. Throughout this paper, we assume that each token has to attend to at least one other token, i.e., $\|\boldsymbol{a}_i\|_1 \geq 1$.

Multi-Head Attention Mechanism: An MHA with H heads, denoted by $\mathrm{Attn}(\text{-})$, is defined as

$$\begin{aligned}
\mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)}) :=& \left[\mathrm{head}_i^{(1)} \dots \mathrm{head}_i^{(H)}\right] \boldsymbol{W}_{\mathrm{O}} = \sum_{j=1}^{H} \mathrm{head}_i^{(j)} \boldsymbol{W}_{\mathrm{O}}^{(j)} \in \mathbb{R}^{1 \times d_2}, \\
\mathrm{head}_i^{(j)} :=& \sigma\left(\frac{1}{\sqrt{d_{\mathrm{K}}}}(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}}^{(j)})(\boldsymbol{X}\boldsymbol{W}_{\mathrm{K}}^{(j)})^\top; \boldsymbol{a}_i\right) \boldsymbol{X}\boldsymbol{W}_{\mathrm{V}}^{(j)} \in \mathbb{R}^{1 \times d_{\mathrm{V}}}, \\
\sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)_j :=& \frac{a_{i,j} \exp(x_{i,j})}{\sum_{k=1}^{T} a_{i,k} \exp(x_{i,k})},
\end{aligned}$$

where $\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)} \in \mathbb{R}^{d_1 \times d_{\mathrm{K}}}$, and $\boldsymbol{W}_{\mathrm{V}}^{(j)} \in \mathbb{R}^{d_1 \times d_{\mathrm{V}}}$ are query, key, and value weights for the j-th head. σ is the softmax function with an attention mask. We decompose the output weight $\boldsymbol{W}_{\mathrm{O}} \in \mathbb{R}^{Hd_{\mathrm{V}} \times d_2}$ into $\boldsymbol{W}_{\mathrm{O}} = [\boldsymbol{W}_{\mathrm{O}}^{(1)\top} \dots \boldsymbol{W}_{\mathrm{O}}^{(H)\top}]^\top$, where $\boldsymbol{W}_{\mathrm{O}}^{(j)} \in \mathbb{R}^{d_{\mathrm{V}} \times d_2}$. We denote the all weights as the set $\boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)} := \{\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)}, \boldsymbol{W}_{\mathrm{V}}^{(j)}, \boldsymbol{W}_{\mathrm{O}}^{(j)}\}_{j=1}^{H}$.

Target and Source Attention Mechanisms: To validate the existence of SLTs for attention mechanisms, we consider two MHAs: a target MHA Attn_T with arbitrary tuned weights, and a suitably pruned source MHA Attn_S with randomly initialized weights, denoted as follows:

$$\mathrm{Attn}_\mathrm{T}(\boldsymbol{x}_i) = \mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)}) \quad \text{and} \quad \mathrm{Attn}_\mathrm{S}(\boldsymbol{x}_i) = \mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, (\tilde{\boldsymbol{W}} \odot \boldsymbol{M})_{\mathrm{Q:O}}^{(1:H)}). \tag{1}$$

Here, $\tilde{\boldsymbol{W}}_\mathrm{Q}^{(j)}, \tilde{\boldsymbol{W}}_\mathrm{K}^{(j)} \in \mathbb{R}^{d_1 \times n_\mathrm{K}}, \tilde{\boldsymbol{W}}_\mathrm{V}^{(j)} \in \mathbb{R}^{d_1 \times n_\mathrm{V}}$, and $\tilde{\boldsymbol{W}}_\mathrm{O}^{(j)} \in \mathbb{R}^{n_\mathrm{V} \times d_2}$ are the query, key, value, and output weights of Attn_S for the j-th head, respectively, and $\boldsymbol{M}_\mathrm{Q}^{(j)}, \boldsymbol{M}_\mathrm{K}^{(j)}, \boldsymbol{M}_\mathrm{V}^{(j)}$ and $\boldsymbol{M}_\mathrm{O}^{(j)}$ are their corresponding binary masks. We define the set of pruned weights as

$$(\tilde{\boldsymbol{W}} \odot \boldsymbol{M})_{\mathrm{Q:O}}^{(1:H)} := \{\tilde{\boldsymbol{W}}_\mathrm{Q}^{(j)} \odot \boldsymbol{M}_\mathrm{Q}^{(j)}, \tilde{\boldsymbol{W}}_\mathrm{K}^{(j)} \odot \boldsymbol{M}_\mathrm{K}^{(j)}, \tilde{\boldsymbol{W}}_\mathrm{V}^{(j)} \odot \boldsymbol{M}_\mathrm{V}^{(j)}, \tilde{\boldsymbol{W}}_\mathrm{O}^{(j)} \odot \boldsymbol{M}_\mathrm{O}^{(j)}\}_{j=1}^{H}.$$

Note that the target and source MHAs have different hidden dimensions: $d_\mathrm{K}, d_\mathrm{V}$ for the target and $n_\mathrm{K}, n_\mathrm{V}$ for the source. We assume that $\alpha \geq \max(\sqrt{d_1}, \sqrt{d_2})$ for the input tokens, and $\|\boldsymbol{W}_\mathrm{Q}^{(j)}\|, \|\boldsymbol{W}_\mathrm{K}^{(j)}\|, \|\boldsymbol{W}_\mathrm{V}^{(j)}\|, \|\boldsymbol{W}_\mathrm{O}^{(j)}\| \leq 1$ for each head of the target MHA. The source MHA are initialized such that each entry in $\tilde{\boldsymbol{W}}_\mathrm{Q}, \tilde{\boldsymbol{W}}_\mathrm{K}$ is drawn i.i.d. from $U[-n_\mathrm{K}^{1/4}, n_\mathrm{K}^{1/4}]$ and each entry in $\tilde{\boldsymbol{W}}_\mathrm{V}, \tilde{\boldsymbol{W}}_\mathrm{O}$ from $U[-1, 1]$.

3.2. Main Theorem: Existence of Strong Lottery Tickets within Attention Mechanisms

Now, we prove the following SLT existence theorem:

Theorem 3 *Let Attn_S and Attn_T be as defined in Equation (1). Then, with probability at least $1-\epsilon$, there exists a choice of binary masks $\boldsymbol{M}_\mathrm{Q}^{(j)}, \boldsymbol{M}_\mathrm{K}^{(j)}, \boldsymbol{M}_\mathrm{V}^{(j)}, \boldsymbol{M}_\mathrm{O}^{(j)}$ that satisfy*

$$\max_{i \in [T]} \|\mathrm{Attn}_\mathrm{S}(\boldsymbol{x}_i) - \mathrm{Attn}_\mathrm{T}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the source hidden dimensions satisfy

$$n_\mathrm{K} \geq d_1 C \log\left(\frac{8H\alpha^3 d_1^{3/2}}{\epsilon}\right) \quad \textit{and} \quad n_\mathrm{V} \geq d_1 C \log\left(\frac{2H\alpha d_1 \sqrt{d_2}}{\epsilon}\right),$$

for some universal constant $C > 0$.

To prove this theorem, it is necessary to approximate the inner product $(\boldsymbol{x}_i^\top \boldsymbol{W}_\mathrm{Q}^{(j)})(\boldsymbol{X}\boldsymbol{W}_\mathrm{K}^{(j)})^\top$ of the target MHA by the inner product $(\boldsymbol{x}_i^\top(\tilde{\boldsymbol{W}}_\mathrm{Q} \odot \boldsymbol{M}_\mathrm{Q}))(\boldsymbol{X}(\tilde{\boldsymbol{W}}_\mathrm{K} \odot \boldsymbol{M}_\mathrm{K}))^\top$ of the pruned source MHA. If we consider naively applying the existing approximation theory to each matrix $\boldsymbol{W}_\mathrm{Q}$ and $\boldsymbol{W}_\mathrm{K}$, it requires a two-layer structure (fully-connected network) for the source key or query projection, respectively. However, since our source MHA has a single-layer projection for the query or key ($(\tilde{\boldsymbol{W}}_\mathrm{Q} \odot \boldsymbol{M}_\mathrm{Q})$ or $(\tilde{\boldsymbol{W}}_\mathrm{K} \odot \boldsymbol{M}_\mathrm{K})$), we cannot naively apply it in our setting. Instead, to overcome this situation, we propose to merge the two target matrices $\boldsymbol{W}_\mathrm{Q}$ and $\boldsymbol{W}_\mathrm{K}$ into a single matrix as

$$\boldsymbol{W}_{\mathrm{QK}}^{(j)} := \frac{1}{\sqrt{d_k}} \boldsymbol{W}_\mathrm{Q}^{(j)} (\boldsymbol{W}_\mathrm{K}^{(j)})^\top. \tag{2}$$

Now, these operations enable us to apply the standard approximation approach, which approximate a single matrix by two matrices, to the composite matrices $\boldsymbol{W}_{\mathrm{QK}}^{(j)}$ by the inner product

$(\boldsymbol{x}_i^\top(\tilde{\boldsymbol{W}}_{\mathrm{Q}} \odot \boldsymbol{M}_{\mathrm{Q}}))(\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{K}} \odot \boldsymbol{M}_{\mathrm{K}}))^\top$ of the pruned source MHA. Similarly, we can approximate the matrix multiplication of the value and output projections in the target attention by defining the merged martix as $\boldsymbol{W}_{\mathrm{VO}}^{(j)} := \boldsymbol{W}_{\mathrm{V}}^{(j)}\boldsymbol{W}_{\mathrm{O}}^{(j)}$.

Lemma 4 *Let $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ be a target matrix with $\|\boldsymbol{W}\| \leq 1$, and let $\tilde{\boldsymbol{W}}_1 \in \mathbb{R}^{n \times d_1}$ and $\tilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_2 \times n}$ be source matrices whose entries are independently drawn from $U[-1, 1]$. For any $0 < \epsilon < 1$, suppose that $n \geq d_1 C \log(d_1 d_2/\epsilon)$ for some universal constant $C > 0$. Then, with probability at least $1 - \epsilon$, there exist binary masks $\boldsymbol{M}_1, \boldsymbol{M}_2$ such that*

$$\left\| \boldsymbol{W} - (\tilde{\boldsymbol{W}}_2 \odot \boldsymbol{M}_2)(\tilde{\boldsymbol{W}}_1 \odot \boldsymbol{M}_1) \right\|_{\max} \leq \frac{\epsilon}{d_1 d_2}.$$

Note that this Lemma 4 can be seen as a variant of Lemma 2 without ReLU activation, since our approximation here does not involve any nonlinearity between matrices.

Next, we analyze the behavior of the attention output when the query and key weights are approximated and passed through the softmax function. Although one may consider using the fact that the softmax function is 1-Lipschitz [9], such an approximation results in rough upper bounds. Indeed, if we employ the Lipschitz continuity in our argument, we obtain an upper bound of softmax error growing with the number of input tokens, while the difference between softmax functions should never diverge. In our setting, we can assume the perturbation within the softmax input is bounded by a small finite constant, leading to the following tighter analysis than Lipschitz continuity.

Lemma 5 *Let $\boldsymbol{\epsilon} \in \mathbb{R}^{d_1}$ be an error vector with $\|\boldsymbol{\epsilon}\|_{\max} \leq \epsilon_{\max}$ for some $0 \leq \epsilon_{\max} \leq 1/2$. Then,*

$$\max_{i \in [T]} \|\sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i + \boldsymbol{\epsilon}; \boldsymbol{a}_i)\boldsymbol{X}\| \leq 4\sqrt{d_1}\alpha\epsilon_{\max}.$$

Finally, using these two lemmas, we can prove Theorem 3. (See Appendix A.3 for the full proof.)

Proof Sketch of Theorem 3: First, as shown in Equation (2), we reformulate the approximation task by combining the target weights into two merged matrices: one for query-key and one for value-output. By applying Lemma 4 to these matrices, we can prune the source MHA so that the source inner product approximates the target inner product. Next, we analyze how the approximation error in $\boldsymbol{W}_{\mathrm{QK}}^{(j)}$ affects the attention output via the softmax function. Lemma 5 gives a bound on how the softmax output changes corresponding to small perturbations via query-key approximation. This bound shows that the output difference is linear with respect to the perturbation magnitude and independent of the number of inputs. Putting everything together, we conclude that if the source hidden dimensions n_{K} and n_{V} are sufficiently large, then there exist binary masks that approximate the target attention mechanism with an error no greater than ϵ. Finally, we ensure the overall probability of successful approximation is at least $1 - \epsilon$ by applying a union bound across all steps.

4. Experimental Results

This section empirically validates our SLTH theorem by approximating a trained MHA via pruning a randomly initialized MHA. We evaluate the approximation by using a synthetic toy dataset designed for an angular velocity estimation task. To identify the SLTs, we apply a subset-sum technique via Gurobi's mixed integer program solver [10]. For more details, see Appendix C.

We vary the key and value source dimensions n_{K} and n_{V}, and observe the approximation error ϵ. For simplicity, these parameters are set equal. As shown in Figure 1, the error decreases rapidly as the hidden dimensions increase. Since the results can be fitted by $\epsilon = 0.77\exp(-0.055n_{\mathrm{K}})$, it provides the empirical support for our theoretical claim: given a target MHA, each source hidden dimension requires $O(\log(1/\epsilon))$ for the existence of SLTs.

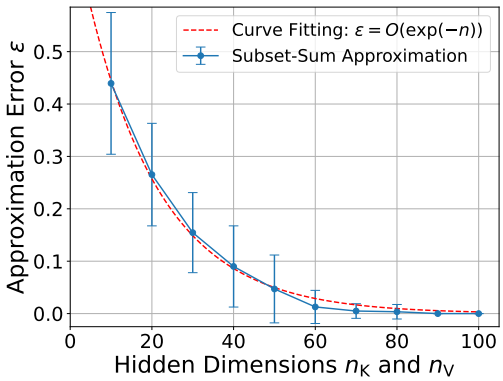

Figure 1: Approximation error ϵ of SLTs for source hidden dimensions $n_{\mathrm{K}} = n_{\mathrm{V}}$. This result shows that the error holds $\epsilon = O(\exp(-n))$, as shown in Theorem 3.

5. Conclusion

This work investigated the existence of SLTs within a multi-head self-attention mechanism (MHA). We extended the existing SLTH theory and found that, if the source MHA has sufficiently large hidden dimensions, SLTs exist in the model. We hope that our findings will contribute to developing efficient network architectures in the future.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number JP23H05489, JP25K03092, JP25KJ1236, and JST-ALCA-Next Japan Grant # JPMJAN24F3.

References

[1] Rebekka Burkholz. Convolutional and residual networks provably contain lottery tickets. In *International Conference on Machine Learning*, pages 2414–2433. PMLR, 2022.

[2] Rebekka Burkholz. Most activation functions can win the lottery without excessive depth. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=NySDKS9SxN.

[3] Arthur Da Cunha and Francesco d'Amore. Polynomially over-parameterized convolutional neural networks contain structured strong winning lottery tickets. *Advances in Neural Information Processing Systems*, 36:25929–25957, 2023.

[4] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[6] Damien Ferbach, Christos Tsirigotis, Gauthier Gidel, and Joey Bose. A general framework for proving the equivariant strong lottery ticket hypothesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vVJZtlZB9D.

[7] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

[8] Advait Harshal Gadhikar, Sohom Mukherjee, and Rebekka Burkholz. Why random pruning is all we need to start sparse. In *International Conference on Machine Learning*, pages 10542–10570. PMLR, 2023.

[9] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[10] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL https://www.gurobi.com.

[11] Hiroaki Ito, Jiale Yan, Hikari Otsuka, Kazushi Kawamura, Masato Motomura, Thiem Van Chu, and Daichi Fujiki. Uncovering strong lottery tickets in graph transformers: A path to memory efficient and robust graph learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=B1q9po4LPl.

[12] George S Lueker. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures & Algorithms*, 12(1):51–62, 1998.

[13] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020.

[14] Emanuele Natale, Davide Ferre', Giordano Giambartolomei, Frédéric Giroire, and Frederik Mallmann-Trenn. On the sparsity of the strong lottery ticket hypothesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=aBMESB1Ajx.

[15] Laurent Orseau, Marcus Hutter, and Omar Rivasplata. Logarithmic pruning is all you need. *Advances in Neural Information Processing Systems*, 33:2925–2934, 2020.

[16] Hikari Otsuka, Daiki Chijiwa, Ángel López García-Arias, Yasuyuki Okoshi, Kazushi Kawamura, Thiem Van Chu, Daichi Fujiki, Susumu Takeuchi, and Masato Motomura. Partially frozen random networks contain compact strong lottery tickets. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=xpnPYfufhz.

[17] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in neural information processing systems*, 33:2599–2610, 2020.

[18] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11893–11902, 2020.

[19] Sheng Shen, Alexei Baevski, Ari Morcos, Kurt Keutzer, Michael Auli, and Douwe Kiela. Reservoir transformers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4294–4309, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.331. URL https://aclanthology.org/2021.acl-long.331/.

[20] Sheng Shen, Zhewei Yao, Douwe Kiela, Kurt Keutzer, and Michael Mahoney. What's hidden in a one-layer randomly weighted transformer? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2914–2921, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.231. URL https://aclanthology.org/2021.emnlp-main.231/.

[21] 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.

[22] Ziqian Zhong and Jacob Andreas. Algorithmic capabilities of random transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=plH8gW7tPQ.

[23] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in neural information processing systems*, 32, 2019.

Appendix A. Proofs of Main Theorems

This section presents the detailed proofs of the main theorems in the manuscript. We first introduce several lemmas that are used to prove the main theorem. Then, leveraging these lemmas, we prove the existence of SLTs in attention mechanisms.

Notation: In this paper, scalars, vectors, and matrices are denoted by lowercase, bold lowercase, and bold uppercase letters, respectively. We use the norm of matrices and vectors $\|\cdot\|$ as the spectral norm unless otherwise specified by a subscript. We denote the uniform distribution on $[a, b]$ by $U[a, b]$. "$\odot$" represents an element-wise multiplication, Hadamard product.

A.1. Weight Approximation

To prove our main theorem, we require approximating a target weight matrix using the product of two random matrices, modified only via pruning. Pensia et al. [17] have shown that a two-layer ReLU network can approximate any real matrix with high probability. Our setting can be viewed as a simplified version of their construction, in which the ReLU nonlinearity is omitted. We follow their proof strategy and adapt it to the linear (non-activated) case.

Lemma 6 (Weight Approximation) *Let $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ be a target matrix with entries in $[-1, 1]$. Let $\tilde{\boldsymbol{W}}_1 \in \mathbb{R}^{n \times d_1}$ and $\tilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_2 \times n}$ be source random matrices whose entries are independently drawn from the uniform distribution $U[-1, 1]$. For any $0 < \epsilon < 1$, suppose that $n \geq d_1 C \log(d_1 d_2/\epsilon)$ for some universal constant $C > 0$. Then, with probability at least $1 - \epsilon$, there exist binary masks $\boldsymbol{M}_1 \in \{0, 1\}^{n \times d_1}$ and $\boldsymbol{M}_2 \in \{0, 1\}^{d_2 \times n}$ such that*

$$\left\| \boldsymbol{W} - (\tilde{\boldsymbol{W}}_2 \odot \boldsymbol{M}_2)(\tilde{\boldsymbol{W}}_1 \odot \boldsymbol{M}_1) \right\|_{\max} \leq \frac{\epsilon}{d_1 d_2}.$$

Proof *This result follows directly from Corollary 3.3 of Lueker [12], which provides an exponentially good approximation guarantee for subset-sum problems. While Pensia et al. [17] apply a similar argument within a ReLU-activated setting, our linear setup allows us to invoke the original result without modification.* ■

A.2. Spectral Norm of Softmax Difference

In addition to approximating weights, we also analyze the stability of the softmax output under small perturbations in the query vector, with respect to the spectral norm of the resulting attention-weighted output.

Lemma 7 (Spectral Norm Bound for Softmax Output Perturbation) *Let $\boldsymbol{\epsilon} \in \mathbb{R}^{d_1}$ be a perturbation vector such that $\|\boldsymbol{\epsilon}\|_{\max} \leq \epsilon_{\max}$ for some $\epsilon_{\max} \geq 0$. Then,*

$$\|\sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i + \boldsymbol{\epsilon}; \boldsymbol{a}_i)\boldsymbol{X}\| \leq \sqrt{d_1}\alpha \left(\exp(2\epsilon_{\max}) - 1\right).$$

Proof *Let $\boldsymbol{p} = \sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)$ and $\boldsymbol{p}' = \sigma(\boldsymbol{x}_i + \boldsymbol{\epsilon}; \boldsymbol{a}_i)$. Then for each coordinate j,*

$$p'_j = p_j \cdot \frac{\exp(\epsilon_j)}{Z}, \quad \textit{where } Z = \sum_{k=1}^{T} p_k \exp(\epsilon_k).$$

By the assumption $\|\boldsymbol{\epsilon}\|_{\max} \leq \epsilon_{\max}$*, we have*

$$\left|1 - \frac{\exp(\epsilon_j)}{Z}\right| \leq \exp(2\epsilon_{\max}) - 1.$$

Now, bounding the spectral norm for the i*-th token:*

$$\begin{aligned}
\left\|\boldsymbol{p}\boldsymbol{X} - \boldsymbol{p}'\boldsymbol{X}\right\| &\leq \sqrt{d_1} \cdot \left\|\boldsymbol{p}\boldsymbol{X} - \boldsymbol{p}'\boldsymbol{X}\right\|_{\max} \\
&\leq \sqrt{d_1} \cdot \max_{i\in[d_1]} \left|\sum_{j=1}^{T} (p_j - p'_j) x_{j,i}\right| \\
&\leq \sqrt{d_1} \cdot \max_{i\in[d_1]} \sum_{j=1}^{T} |x_{j,i}| \cdot |p_j - p'_j| \\
&\leq \sqrt{d_1} \cdot \alpha \sum_{j=1}^{T} |p_j - p'_j| \\
&\leq \sqrt{d_1} \cdot \alpha \sum_{j=1}^{T} p_j \left|1 - \frac{\exp(\epsilon_j)}{Z}\right| \\
&\leq \sqrt{d_1} \cdot \alpha \left(\exp(2\epsilon_{\max}) - 1\right) \sum_{j=1}^{T} p_j \\
&= \sqrt{d_1} \cdot \alpha \left(\exp(2\epsilon_{\max}) - 1\right).
\end{aligned}$$

Since the final upper bound is independent of i*, then the upper bound of* $\max_{i\in[T]} \|\boldsymbol{p}\boldsymbol{X} - \boldsymbol{p}'\boldsymbol{X}\|$ *is the same as that final bound.* ■

A.3. SLT Existence within Attention Mechanisms

Finally, we prove the following main theorem:

Theorem 8 (SLT Existence within MHA) *Let* $\mathrm{Attn_S}$ *and* $\mathrm{Attn_T}$ *be as defined in Equation (1). Assume* $\alpha \geq \max(\sqrt{d_1}, \sqrt{d_2})$ *for the input tokens. Then, with probability at least* $1 - \epsilon$*, there exists a choice of binary masks* $\boldsymbol{M}_{\mathrm{Q}}^{(j)}, \boldsymbol{M}_{\mathrm{K}}^{(j)}, \boldsymbol{M}_{\mathrm{V}}^{(j)}, \boldsymbol{M}_{\mathrm{O}}^{(j)}$ *that satisfy*

$$\max_{i\in[T]} \|\mathrm{Attn_S}(\boldsymbol{x}_i) - \mathrm{Attn_T}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the source dimensions satisfy

$$n_1 \geq d_1 C \log\left(\frac{8H\alpha^3 d_1^{3/2}}{\epsilon}\right), \quad n_2 \geq d_1 C \log\left(\frac{2H\alpha d_1 \sqrt{d_2}}{\epsilon}\right),$$

for some universal constant $C > 0$*.*

Proof *We divide the proof into three key steps.*

Step 1: Representation Alignment. *The target MHA weights are merged as*

$$\boldsymbol{W}_{\mathrm{Q}K}^{(j)} := \frac{1}{\sqrt{d_k}} \boldsymbol{W}_{\mathrm{Q}}^{(j)} (\boldsymbol{W}_{\mathrm{K}}^{(j)})^\top, \quad \boldsymbol{W}_{\mathrm{V}O}^{(j)} := \boldsymbol{W}_{\mathrm{V}}^{(j)} \boldsymbol{W}_{\mathrm{O}}^{(j)},$$

which allows us to express each head of $\mathrm{Attn}_{\mathrm{T}}$ *as*

$$\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)}) = \sum_{j=1}^{H} \sigma\left(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}K}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i\right) \boldsymbol{X} \boldsymbol{W}_{\mathrm{V}O}^{(j)}.$$

By the norm assumption, we have $\|\boldsymbol{W}_{\mathrm{Q}K}^{(j)}\| \le 1/\sqrt{d_k}$ *and* $\|\boldsymbol{W}_{\mathrm{V}O}^{(j)}\| \le 1$.

Step 2: Weight Approximation. *From Lemma 6, for any* $0 < \epsilon < 1$, *if*

$$n_{\mathrm{K}} \ge d_1 C \log\left(\frac{8H\alpha^3 d_1^{3/2}}{\epsilon}\right),$$

then with probability at least $1 - \frac{\epsilon}{8H\alpha^3\sqrt{d_1}}$, *there exist binary masks* $\boldsymbol{M}_{\mathrm{Q}}^{(j)}, \boldsymbol{M}_{\mathrm{K}}^{(j)}$ *such that*

$$\left\| \boldsymbol{W}_{\mathrm{Q}K}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)}\right)^\top \right\|_{\max} \le \frac{\epsilon}{8H\alpha^3 d_1^{3/2}}.$$

We can also bound the infinity-norm inside the softmax as follows:

$$\begin{aligned}
&\left\| \boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}K}^{(j)} \boldsymbol{X}^\top - \boldsymbol{x}_i^\top \left(\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)}\right)^\top \boldsymbol{X}^\top \right\|_\infty \\
&= \max_{k\in[T]} \left| \boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}K}^{(j)} \boldsymbol{x}_k - \boldsymbol{x}_i^\top \left(\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)}\right)^\top \boldsymbol{x}_k \right| \\
&\le \alpha^2 \left\| \boldsymbol{W}_{\mathrm{Q}K}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)}\right)^\top \right\| \\
&\le \alpha^2 d_1 \left\| \boldsymbol{W}_{\mathrm{Q}K}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)}\right)^\top \right\|_{\max} \\
&\le \alpha^2 d_1 \frac{\epsilon}{8H\alpha^3 d_1^{3/2}} \\
&= \frac{\epsilon}{8H\alpha\sqrt{d_1}}.
\end{aligned}$$

Now, for each token $i \in [T]$ *and each head* $j \in [H]$, *we define*

$$\boldsymbol{p}_i^{(j)} := \sigma\left(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}K}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i\right), \quad \boldsymbol{p}_i'^{(j)} := \sigma\left(\boldsymbol{x}_i^\top (\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)})^\top \boldsymbol{X}^\top; \boldsymbol{a}_i\right).$$

Then, applying Lemma 7, we obtain

$$\begin{aligned}
\left\| \boldsymbol{p}_i^{(j)} \boldsymbol{X} - \boldsymbol{p}_i'^{(j)} \boldsymbol{X} \right\| &\le \sqrt{d_1}\alpha \left(\exp\left(\frac{\epsilon}{4H\alpha\sqrt{d_1}}\right) - 1\right) \\
&\le \frac{\epsilon}{2H}, \qquad \text{(since } 0 < \tfrac{\epsilon}{4H\alpha\sqrt{d_1}} < 1.\text{)}
\end{aligned}$$

Step 3: Output Approximation. *Similarly, from Lemma 6, if*

$$n_{\mathrm{V}} \geq d_1 C \log\left(\frac{2H\alpha d_1\sqrt{d_2}}{\epsilon_2}\right),$$

then with probability at least $1 - \frac{\sqrt{d_2}\epsilon}{2H\alpha}$*, there exist binary masks* $\boldsymbol{M}_{\mathrm{V}}^{(j)}, \boldsymbol{M}_{\mathrm{O}}^{(j)}$ *such that*

$$\left\|\boldsymbol{W}_{\mathrm{V}O}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\|_{\max} \leq \frac{\epsilon}{2H\alpha d_1\sqrt{d_2}}.$$

The difference between the target and pruned source attention outputs is

$$\begin{aligned}
\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i)\| &= \left\|\sum_{j=1}^{H}\left(\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}O}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right)\right\| \\
&\leq \sum_{j=1}^{H}\left\|\left(\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}O}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right)\right\|
\end{aligned}$$

We apply the triangle inequality and break the spectral norm into two terms for each head j*:*

$$\begin{aligned}
&\left\|\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}O}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\| \\
&\quad\leq \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}O}^{(j)}\right\| + \left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\boldsymbol{W}_{\mathrm{V}O}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\|.
\end{aligned}$$

Since $\|\boldsymbol{W}_{\mathrm{VO}}^{(j)}\| \leq 1$*, the first term is bounded as follows.*

$$\begin{aligned}
\left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}O}^{(j)}\right\| &\leq \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\right\| \left\|\boldsymbol{W}_{\mathrm{V}O}^{(j)}\right\| \\
&\leq \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\right\| \\
&\leq \frac{\epsilon}{2H}. \qquad \text{(From the result of Step 2.)}
\end{aligned}$$

For the second term, since the Euclidean norm of each token is α *or below, we have*

$$\begin{aligned}
&\left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\boldsymbol{W}_{\mathrm{V}O}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\| \\
&\quad\leq \sqrt{d_1}\left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}\right\|_{\infty}\sqrt{d_1 d_2}\left\|(\boldsymbol{W}_{\mathrm{V}O}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\|_{\max} \\
&\quad\leq d_1\sqrt{d_2}\alpha\left\|(\boldsymbol{W}_{\mathrm{V}O}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\|_{\max} \\
&\quad\leq d_1\sqrt{d_2}\alpha\frac{\epsilon}{2H\alpha d_1\sqrt{d_2}} \\
&\quad= \frac{\epsilon}{2H}.
\end{aligned}$$

These results do not depend on the token index i*; thus, we obtain*

$$\max_{i\in[T]}\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i)\| \leq \sum_{j=1}^{H}\left(\frac{\epsilon}{2H} + \frac{\epsilon}{2H}\right) = \epsilon.$$

Since we assume $\alpha \geq \max(\sqrt{d_1}, \sqrt{d_2})$, from the union bound, the probability that the two weight approximations of Steps 2 and 3 succeed simultaneously is as follows.

$$1 - \frac{\epsilon}{8H\alpha^3\sqrt{d_1}} - \frac{\sqrt{d_2}\epsilon}{2H\alpha} \geq 1 - \epsilon.$$

■

Appendix B. Related Works

Strong Lottery Tickets: Zhou et al. [23] and Ramanujan et al. [18] empirically found the subnetworks that achieve high accuracy without any weight training. This existence of high-performing subnetworks is called as the strong lottery ticket hypothesis (SLTH), and Malach et al. [13] provided the first theoretical proof in fully-connected ReLU networks: given a target network with arbitrary weights, the randomly initialized network (source network) contain SLTs that approximate the target network if the source network has sufficient width and double depth to the target. Later, some works [2, 15, 17] relaxed the architectural requirements for containing such subnetworks in the scenario of fully-connected networks. In particular, Pensia et al. [17] introduced the subset-sum approximation [12] to approximate the target weights, and proved that the logarithmic overparameterization of the source network is sufficient for the existence of SLTs.

Based on these pioneering studies, subsequent works have extended the SLTH in three main directions. The first direction involves introducing additional flexibility into the source network. In this context, it has been demonstrated that iterative randomization or small perturbations to the source weights can reduce the required width of the source network. The second direction, in contrast, imposes additional constraints on the source network. These studies in this context has established the existence of SLTs in scenarios involving the sparse [8], partially frozen [16], and sparsity-constrained networks [14]. The third direction expands the SLTH to various architectures, including binarized networks, non-ReLU activation functions, networks with random biases, convolutional networks, residual networks, and equivariant networks. Our work contributes to this third direction about architectural expansion by proving the existence of SLTs within attention mechanisms, a core component of transformer architectures. (For our theorem, see Section 3.)

Randomly Weighted Transformers: Several studies have empirically investigated the capabilities of random transformers—the transformer architectures with randomly initialized weights. Shen et al. [19] demonstrated that a transformer with a few randomly weighted layers achieves accuracy comparable to fully trained models on translation and language understanding tasks. Zhong and Andreas [22] found that random transformers can solve toy tasks with high accuracy as the hidden dimension increases. Shen et al. [20] experimentally showed the existence of SLTs within random transformers. Our work provides a theoretical explanation for the improved performance of random transformers as the hidden dimension increases, particularly in scenarios where pruning is used for optimization. Moreover, our analysis offers theoretical support of SLTH for the empirical findings by Shen et al. [20].

Appendix C. Detailed Experimental Settings

Dataset: We evaluate the approximation by using a synthetic toy dataset designed for an angular velocity estimation task. Each input token encodes a two-dimensional coordinate on a unit circle,

updated according to a predefined angular velocity. We use a regression token—similar to the CLS token in BERT [5]—to estimate the angular velocity. We generate 10,000 samples of training, validation, and test data, respectively.

Models: From the dataset setting, we set the input and output dimensions of both target and source MHA as two and one, respectively. Both target and source MHA have a single head, and we define the trained MHA with hidden dimension 8 is used as the target function. The target MHA is trained for 25 epochs using the AdamW and MSE loss function with the batch size of 1024 and learning rates of 0.05, 0.1, 0.5, and 1.0, respectively. Our experiment employs the best setting of the learning rate of 0.5. To identify the SLTs that achieve the best approximation, we apply a subset-sum technique via Gurobi's mixed integer program solver [10]. We approximate the target MHA with 100 randomly initialized source MHAs, and report the mean and standard deviation of the approximation error.