Self-Attention Limits Working Memory Capacity of Transformer-Based Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work on Transformer-based large language models (LLMs) has revealed striking limits in their working memory capacity, similar to what has been found in human behavioral studies. Specifically, these models' performance drops significantly on *N*-back tasks as *N* increases. However, there is still a lack of mechanistic interpretability as to why this phenomenon would arise. Inspired by the executive attention theory from behavioral sciences, we hypothesize that the self-attention mechanism within Transformer-based models might be responsible for their working memory capacity limits. To test this hypothesis, we train vanilla decoder-only transformers to perform *N*-back tasks and find that attention scores gradually aggregate to the *N*-back positions over training, suggesting that the model masters the task by learning a strategy to pay attention to the relationship between the current position and the *N*-back position. Critically, we find that the total entropy of the attention score matrix increases as *N* increases, suggesting that the dispersion of attention scores might be the cause of the capacity limit observed in *N*-back tasks.

1 Introduction

In cognitive science, working memory is defined as the ability of humans to temporarily maintain and manipulate task-relevant information for flexible behaviors [1]. Recent advancements in Transformer-based LLMs have sparked interest in evaluating their cognitive abilities, including working memory capacity [9]. By designing multiple variants of *N*-back tasks (Figure 1a) [11, 10] and employing different instructional strategies, it has been found that LLMs consistently perform worse as *N* increases (Figure 1b), which is reminiscent of the capacity limit of human working memory [2, 15, 17].

However, due to the black-box nature of LLMs, we still lack mechanistic insights as to why the observed capacity limit would emerge, especially given the fact that the length of *N*-back task sequences (e.g., 24 letters in [9]) is well within the context length of these models [16]. To answer this question, we were inspired by the executive attention theory [7, 5, 6] in human working memory research. The executive attention theory proposes that working memory requires executive attention to maintain access to information in the face of interference. suggesting that it is the scarcity of attentional resources [12, 14], but not memory storage itself, that is responsible for working memory capacity limits. In Transformer-based LLMs, the self-attention mechanism computes the importance of each element in the input sequence relative to other elements. While this approach allows the model to focus on relevant information, as *N* increases in the *N*-back task, it could be increasingly hard to maintain focus between distant positions. Therefore, we hypothesize that self-attention might be the cause of working memory capacity limits in Transformer-based models.

In the current study, we train causal Transformers on *N*-back tasks and observe that as *N* increases, the model presents a decline in its prediction accuracy. We further find that the prediction accuracy at position i is positively correlated with the attention score at position $i - N$. Furthermore, the model's

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

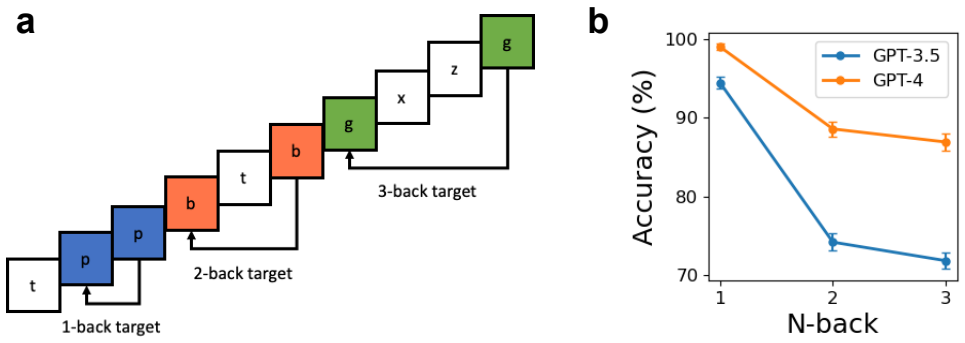

Figure 1: **(a)**: N-back task schematic. Participants (humans or LLMs) are instructed to give a response (humans: press a button; LLMs: output "`m`") when the current letter is matched with the letter N step(s) ago, and withhold responses (humans: do nothing; LLMs: output "`-`") if it's a nonmatch. N is fixed for a given task sequence, and here we put $\{1, 2, 3\}$-back in the same schematic for illustration purposes only. **(b)**: performance of GPT-3.5 and GPT-4 on this task, reproduced from results in [9]. Error bars represent ± 1 standard error of the mean.

performance is negatively correlated with the total entropy of the attention score matrix. Our findings suggest that model's inability to aggregate most of its attention to the target position leads to the decline in its prediction accuracy as N increases.

2 Methods

Dataset. We use the same procedure described by Gong et al. [9] to generate a dataset of N-back tasks consisting of task sequences and correct answers. Each task sequence contains 24 letters sampled from an alphabet commonly used in the behavioral literature ("`bcdfghjklnpqrstvwxyz`"), and the correct answers always consist of 8 matches and 16 nonmatches, mimicking the setup in some human studies. For $N \in \{1, 2, 3, 4, 5, 6\}$, we generate 800 sequences for training and 200 sequences for testing, while our analyses mostly focus on $N \in \{1, 2, 3\}$ to compare with previous studies.

Model. We use vanilla Transformers in order to facilitate interpretability, as done in prior work aiming to better understand computations in Transformers in more controlled task settings [4, 13]. We mainly focus our analysis on a causal Transformer containing one decoder layer with only one attention head (Figure 6 in Appendix), although we also test a few architectural variants in the number of decoder layers (L) and number of attention heads per layer (H) for comparisons (see Section 3 for details). The decoder layer contains masked self-attention so that for each position in the sequence the model can only attend to the current and previous positions. No multi-layer feed-forward networks or layer normalization are applied. The decoder layer is then followed by an unembedding layer to project the decoder outputs to two logits (representing match and nonmatch) for each position.

Training and Evaluation. We train 50 independent models for each N. We choose to train each model for 10 epochs because empirically the model converges after around 10 epochs of training (see Figure 7 in Appendix for details). Cross-entropy loss is computed between the output logits and the correct answers at each position.

3 Results

Model accuracy decreases as N increases. For $\mathsf{L} \in \{1, 2\}$ and $\mathsf{H} \in \{1, 2, 4\}$, we train models on the N-back task (Figures 2a) and find a significant decline in model performance as N increases for the 1-layer 1-head model (Kruskal-Wallis test: H-statistic = 38.517, $p < .001$, $\epsilon^2 = 0.248$; see Table 1 in Appendix for post-hoc comparisons using Mann-Whitney U tests[1]). To further confirm this pattern, we extend the task to $N = 6$, and find a significant logarithmic decline in the test accuracy as N

[1]We use nonparametric Kruskal-Wallis and Mann-Whitney tests instead of F and t tests because the data do not conform to the assumptions of parametric tests (normality and homogeneity of the variance).

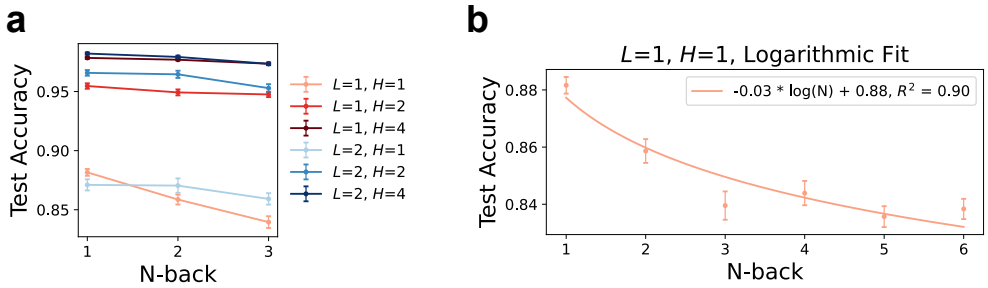

Figure 2: **(a)**: N-back task performance of Transformers with different number of decoder layers and attention heads per layer. **(b)**: for the 1 -layer 1-head Transformer model, task performance drops logarithmically as N increases. Error bars represent ± 1 standard error of the mean.

increases (Figure 2b). For models with a larger L or H, most of them achieved over 95% accuracy on all N-back tasks. However, they still present slight declines in test accuracy as N increases, suggesting that the working memory capacity limit does exist in the nature of transformer models.

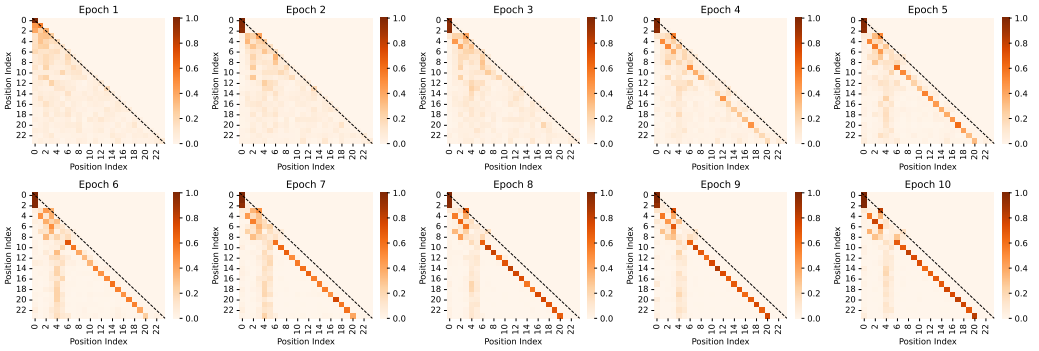

Figure 3: the model learns to attend target locations over training epochs. Here we show attention maps of a 1-layer, 1-head Transformer model trained on the 3-back task as an example. See Appendix for attention maps in the 1-back and 2-back tasks.

Attention scores during training reflect the trajectory of learning. To investigate how the self-attention mechanism influences model performance, we visualize attention maps after each training epoch (Figures 3, 8 and 10). For each position, we also plot the trajectory of attention scores over training epochs (Figures 9, 11, and 12) to see with more granularity how the model learns to perform the task. Starting with almost uniformly distributed attention scores in each row, attention scores gradually aggregate to a line corresponding to the N-back positions. For each position in the sequence, attention scores gradually aggregate to the N-back position over training epochs and attention scores converge faster for positions earlier in sequence (Figures 9, 11, and 12). This shows that the Transformer model learns to master the N-back task by increasing the attention score between the current position and the N-back position.

Attention score at position $i-N$ increases with test accuracy at position i. To further investigate the relationship between attention scores and test accuracy, we plot accuracy at position i against the attention score at the position $i-N$ over training epochs ($i \in \{1..24\}$, $N \in \{1, 2, 3\}$). The accuracy at position i is defined as the percentage of the model making a correct prediction at position i. Over training epochs, we find that the attention score at position $i-N$ increases along with the accuracy at position i (Figure 4a-c). We reason that in order to produce an accurate prediction at position i, the Transformer model needs to learn to put most attention on the $i-N$ position and reduce dispersion of attention to other positions. To better visualize dispersion of attention scores across positions, we use the same data in Figure 4a-c but assign colors to the dots according to which position each dot belongs to (Figure 4d-f). This reveals a clear pattern that attention scores get dispersed at later locations, suggesting that more interference is caused when there are more preceding positions.

Total entropy of attention scores increases as N increases. Building up from the results above, we take a step further to investigate the overall characteristic of attention scores as N increases.

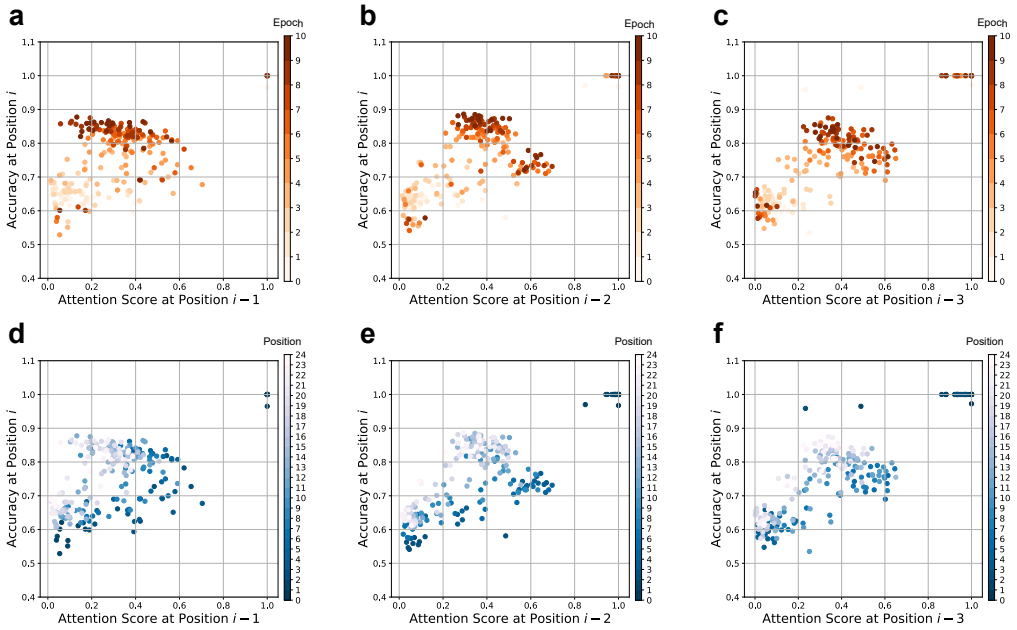

Figure 4: **(a)-(c)**: the relationship between test accuracy at position i and the attention score at position $i - N$. Colors represent different epochs each dot belongs to. **(d)-(f)**: same as **(a)-(c)** but colors represent different position each dot belongs to.

To measure the dispersion of attention scores for each *N*, we define the total entropy H_N of each attention score matrix $A \in \mathbb{R}^{24\times 24}$ as:

$$H_N(A) = -\sum_{i=1}^{24}\sum_{j=1}^{i} A_{i,j}\log\left(A_{i,j}\right) \tag{1}$$

where

$$A_{i,j} = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})_{i,j} \tag{2}$$

The entropy H_N is well-defined as $\{A_{i,1}, A_{i,2}, ..., A_{i,i}\}$ gives a probability distribution with $\sum_{j=1}^{i} A_{i,j} = 1$ thanks to the Softmax function and causal masking.

We find that H_N increases as *N* increases, leading to the decrease in test accuracy (Figure 5). We infer that as *N* increases, it gets harder for the model to learn to attend to the *N*-back letter and the model is less confident about which letter is important, leading to higher entropy and lower accuracy.

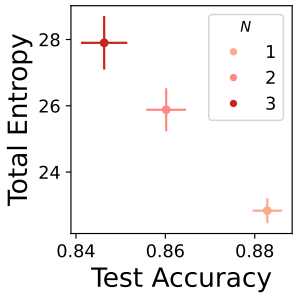

Figure 5: H_N increases as the test accuracy decreases with larger *N*. Error bars represent ± 1 standard error of the mean.

4 Discussion

The current study provides important insights for the mechanistic interpretability of working memory capacity limits observed in Transformer-based LLMs [9]. The self-attention mechanism is critical for the model to achieve good performance in the *N*-back task, but also limits its capacity on the other hand. This is analogous to the mechanism of selective attention in the human brain, which prioritizes relevant information and filter out the rest to ensure effective task performance, but also restricts our information processing by imposing neural and cognitive bottlenecks [3]. Future work should explore a more formal mathematical proof as to why capacity limits might naturally emerge in complex intelligent systems [8, 18].

References

[1] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[2] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

[3] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[4] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022.

[5] Randall W. Engle. Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1):19–23, 2002.

[6] Randall W. Engle and Michael J. Kane. Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation*, 44:145–199, 2003.

[7] Randall W. Engle, Michael J. Kane, and Stephen W. Tuholski. *Individual Differences in Working Memory Capacity and What They Tell Us About Controlled Attention, General Fluid Intelligence, and Functions of the Prefrontal Cortex*, page 102–134. Cambridge University Press, 1999.

[8] Steven M Frankland, Taylor Webb, and Jonathan D Cohen. No coincidence, george: Capacity-limits as the curse of compositionality, 2021.

[9] Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of chatgpt: An empirical study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10048–10056, 2024.

[10] Michael J. Kane and Randall W. Engle. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, 9(4):637–671, December 2002.

[11] Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[12] Peter Lennie. The cost of cortical computation. *Current Biology*, 13(6):493–497, 2003.

[13] Yuxuan Li and James McClelland. Representations and computations in transformers that support generalization on structured tasks. *Transactions on Machine Learning Research*, 2023.

[14] Grace W Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:516985, 2020.

[15] Klaus Oberauer, Simon Farrell, Christopher Jarrold, and Stephan Lewandowsky. What limits working memory capacity? *Psychological Bulletin*, 142(7):758–799, July 2016.

[16] Saurav Pawar, SM Tonmoy, SM Zaman, Vinija Jain, Aman Chadha, and Amitava Das. The what, why, and how of context length extension techniques in large language models–a detailed survey. *arXiv preprint arXiv:2401.07872*, 2024.

[17] Oliver Wilhelm, Andrea Hildebrandt, and Klaus Oberauer. What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4, 2013.

[18] Yudi Xie, Yu Duan, Aohua Cheng, Pengcen Jiang, Christopher J Cueva, and Guangyu Robert Yang. Natural constraints explain working memory capacity limitations in sensory-cognitive models. *bioRxiv*, pages 2023–03, 2023.

Appendix

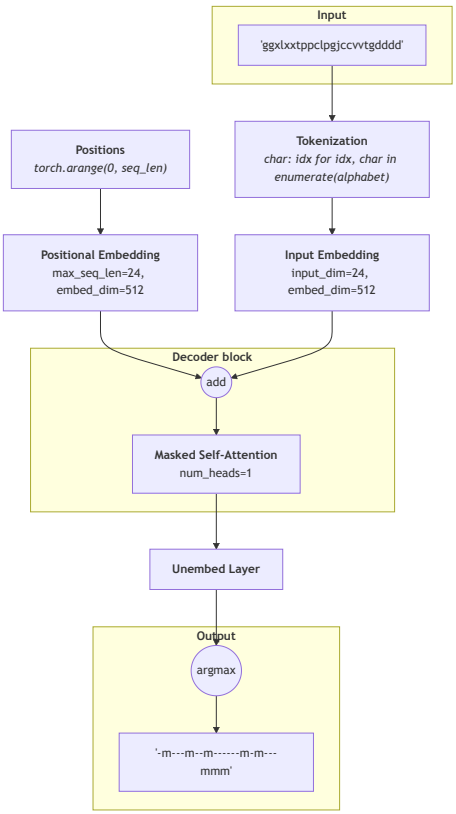

Figure 6: The architecture of the 1-layer 1-head Transformer.

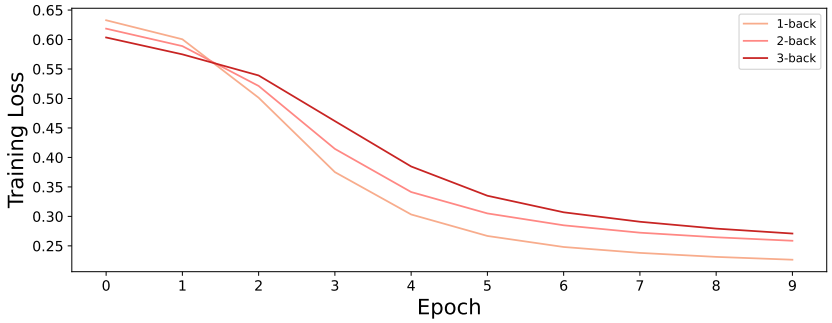

Figure 7: Training loss of the 1-layer 1-head Transformer converges after 10 epochs.

Table 1: Post-hoc Mann-Whitney U test results for the 1-layer 1-head model.

N-back	U	p	r
1 vs 2	1825.0000	0.0002	-0.4600
1 vs 3	2096.0000	0.0000	-0.6768
2 vs 3	1665.0000	0.0128	-0.3320

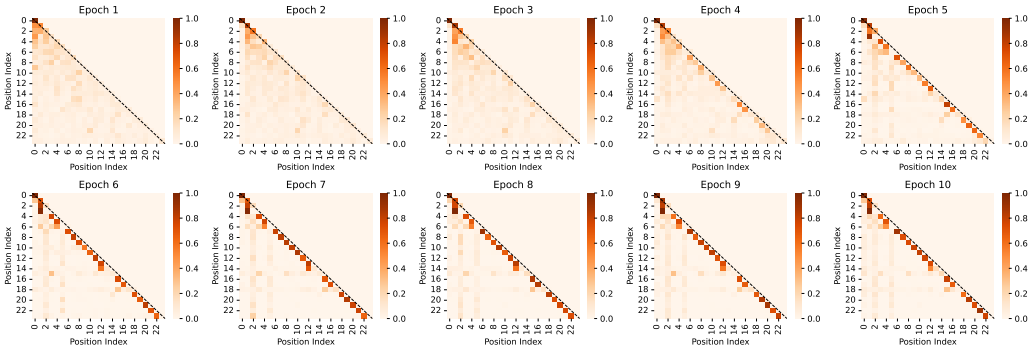

Figure 8: Attention maps over training epochs for a 1-layer 1-head Transformer trained on the 1-back task.

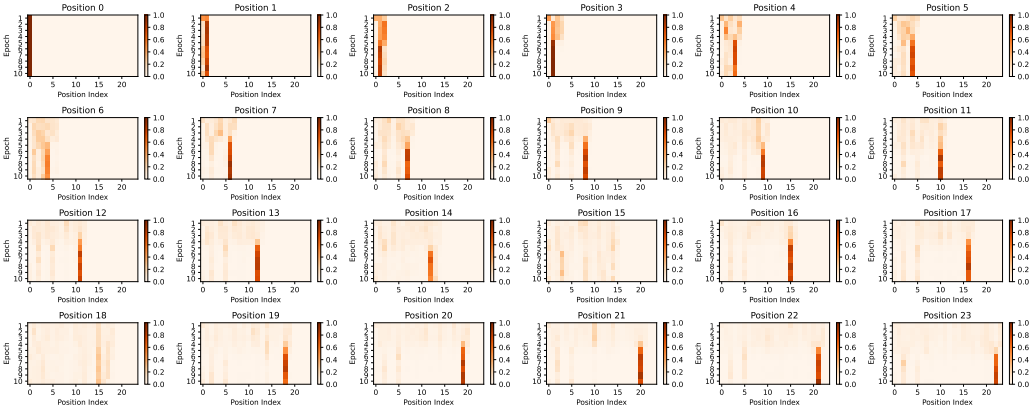

Figure 9: Training trajectory of attention scores over 10 epochs for the 1-back task.

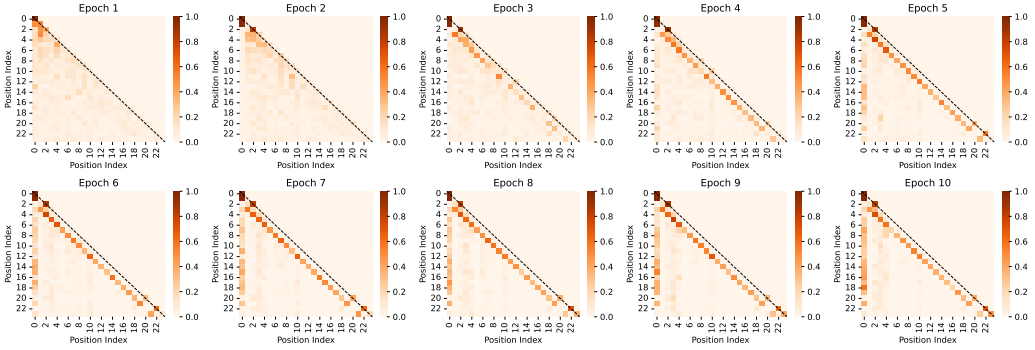

Figure 10: Attention maps over training epochs for a 1-layer 1-head Transformer trained on the 2-back task.

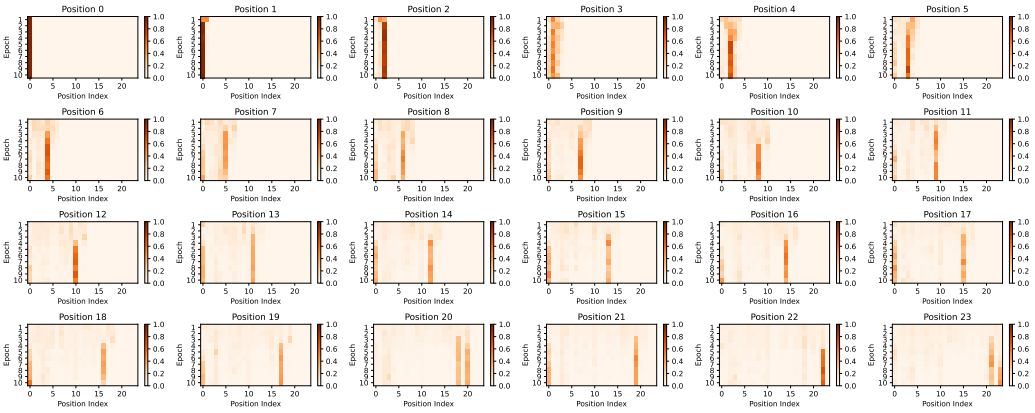

Figure 11: Training trajectory of attention scores over 10 epochs for the 2-back task.

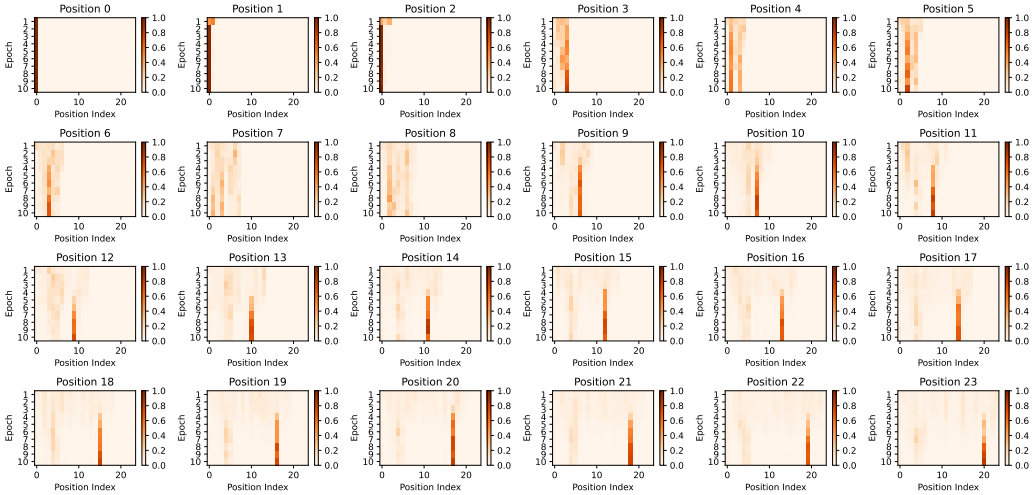

Figure 12: Training trajectory of attention scores over 10 epochs for the 3-back task.