CS 299R Independent Study Spring 2023

Exploring Transformer Interpretability

Cynthia Chen *Mentors: Catherine Yeh, Martin Wattenberg, Fernanda Viégas*

Overview

Transformer models are improving at a rapid pace, making it of paramount importance to develop methods to explain, reverse-engineer, and visualize their inner workings. In this work, we study the interpretability of transformer models through a series of experiments divided into two parts: 1) visualization of attention patterns and 2) exploration of induction heads in BERT.

This report presents the methods and results of an independent research study conducted over the course of January to April 2023 at the Harvard Insight and Interaction Lab under the direct mentorship of Professor Martin Wattenberg, Professor Fernanda Viégas, and Catherine Yeh. The code for the experiments in this project can be found at this Github repository.

1 Experiments in Attention Visualization

1.1 Background

AttentionViz. The self-attention mechanism in transformer models plays a critical role in helping the model learn a rich set of relationships between input elements [1]. To assist in our understanding of attention, Yeh et al. developed AttentionViz, a tool that enables the visualization of attention patterns at a more global scale [2]. In particular, AttentionViz introduces a technique for jointly visualizing query and key vectors—two of the core components in computing attention—in a shared embedding space. In AttentionViz, every query and key (originally a 64-dimensional vector) is projected to a 2-dimensional embedding space using t-SNE or UMAP. As shown in Figure 1a, queries and keys are jointly displayed on the same plot, allowing for the visualization of distinct attention patterns among queries and keys.

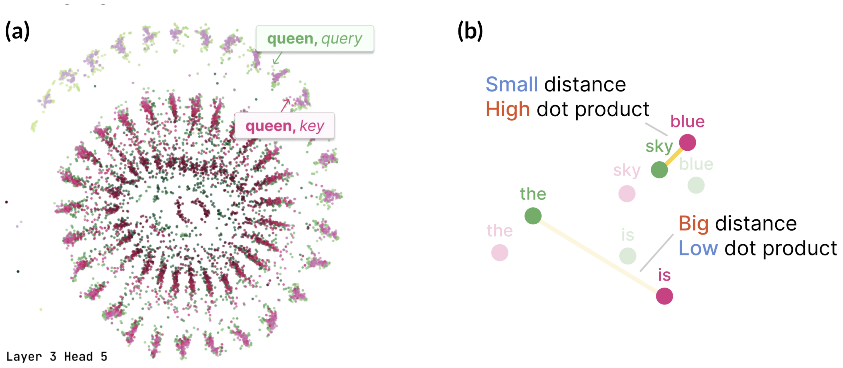

Figure 1: (a) A joint embedding visualization generated by AttentionViz. Queries are represented in green while keys are represented in pink. (b) A depiction of the desired inverse relationship between dot product and distance between queries and keys in the visualization.

Distance as a proxy for attention. A critical idea here is that in the AttentionViz visualizations, we want distance to be an accurate proxy for attention: high-attention query-key pairs should be closer together in the joint embeddings, a relationship depicted in Figure 1b. To optimize for this desired distance-attention relationship, we can take a look at how attention is computed based on the q (query), k (key), and v (value) vectors:

$$\texttt{attention}(q, k, v) = \mathrm{softmax}(\frac{qk^T}{\sqrt{d_k}})v$$

We see that attention directly corresponds to the dot product between the query and key vector (qk^T). Therefore, if we are aiming for *small distance* to be a proxy for *high attention*, then we want the query-key dot product and distance to have a strong, inverse correlation. Put mathematically, we want the correlation between $\texttt{dot-product}(q, k)$ and $\texttt{distance}(q, k)$ to be as close to -1 as possible.

1.2 Optimizing Correlation

How can we optimize the correlation between the dot product and distance between queries and keys without losing the integrity of the attention computation? Luckily, there are two "free parameters" when computing attention: translation and scaling. The operations of *translation* (shifting query and key vectors by a constant vector) and *scaling in opposite directions* (multiplying query vectors by c and dividing key vectors by c) can both be performed without changing the resulting attention value. In the following experiments, we largely focus on scaling and identifying the scaling constant c that provides the best correlation between dot product and distance.

To determine the optimal value of c, we can define a *weighted correlation* metric that places heavier weight on query-key pairs with smaller distances, since we care most about nearby queries and keys in the joint visualization. We first computed a distance threshold d, defined as the 0.5 percentile value of the distance distribution within a specific attention head. For every query-key pair with distance $d_i < d$, we compute the weighted correlation as follows:

$$\texttt{weighted-corr}(x, y, w) = \frac{\mathrm{cov}(x, y; w)}{\sqrt{\mathrm{cov}(x, x; w)\mathrm{cov}(y, y; w)}}$$

The weights w are defined as $(d - d_i)^2$ which assigns more weight to query-key pairs that are closer to one another. We then choose the value of c that gives a weighted correlation closest to -1.

Building off of the weighted correlation metric, we defined a second optimization metric (*weighted correlation, scaled*) as follows. Within each scaling factor, we also kept a count of the number of instances of key-query pairs with distance less than the distance threshold. We then enumerated the number of instances across all the attention heads and normalized all weighted correlations within the scaling factor by this count. Again, we choose a value of c that brings this scaled weighted correlation value closest to -1.

A final metric that we experimented with is the *ratio of the median query norm to the median key norm*. Differences in norm can cause distance and dot product to diverge from one another; as such, we reasoned that standardizing the query and key norms would bring the correlation closer. Rather than maximizing the correlation here, we simply set c to be the square root of the ratio itself, as scaling by c will automatically standardize the query and key norms.

For each attention head, we can thus choose the scale factors c that optimize the three metrics described above. For each of the metrics, we ran experiments with constants $c \in [0.2, 0.4, 0.8, 1, 1.25, 2.5, 5]$. Future work could explore the results of a greater range and granularity of constant values. The optimal scaling constants for each metric are displayed in the heatmaps in Figure 2.

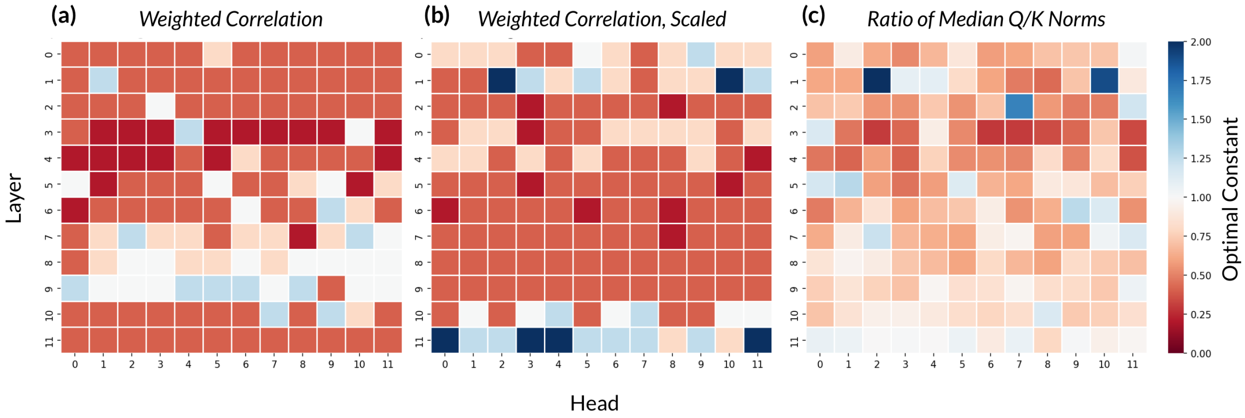

Figure 2: The optimal scaling constants for each attention head, as computed under the three defined metrics—**(a)** weighted correlation, **(b)** weighted correlation, scaled, and **(c)** ratio of Q/K norms—are displayed as heatmaps.

1.3 Scaling Queries and Keys

Here, we show examples of the resulting embedding visualization of keys and queries after they have been scaled. In Figure 3 and 4, we display the joint embeddings for six constants and highlight the visualization with the optimal scaling constant identified by the *weighted correlation (scaled)* metric as shown in Figure 2b.

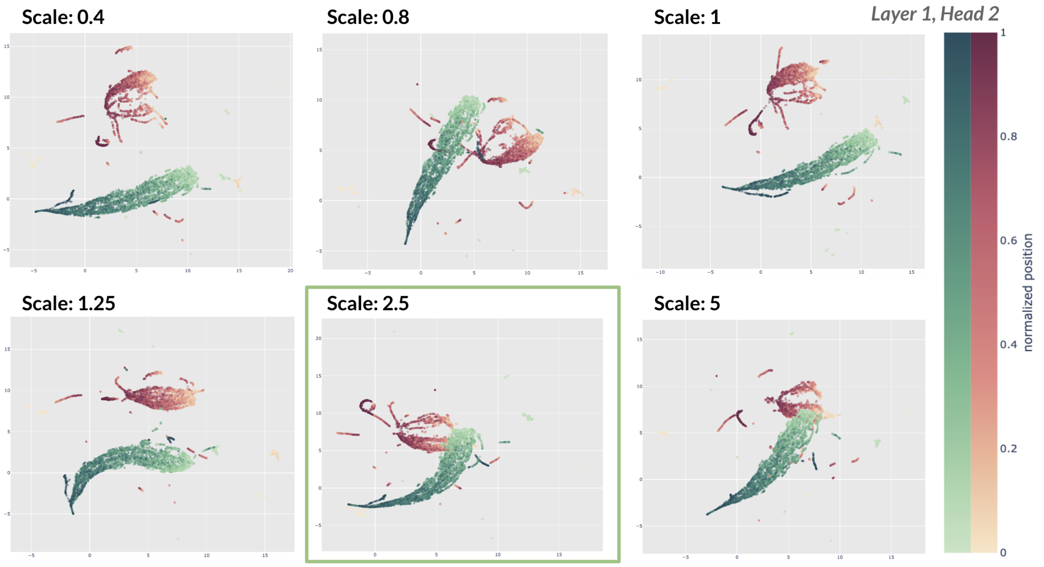

Figure 3: The joint embedding visualization after queries and keys have been scaled and translated, generated for the attention head at layer 1 head 2 in GPT. The optimal constant identified by our methods is $c = 2.5$ as highlighted.

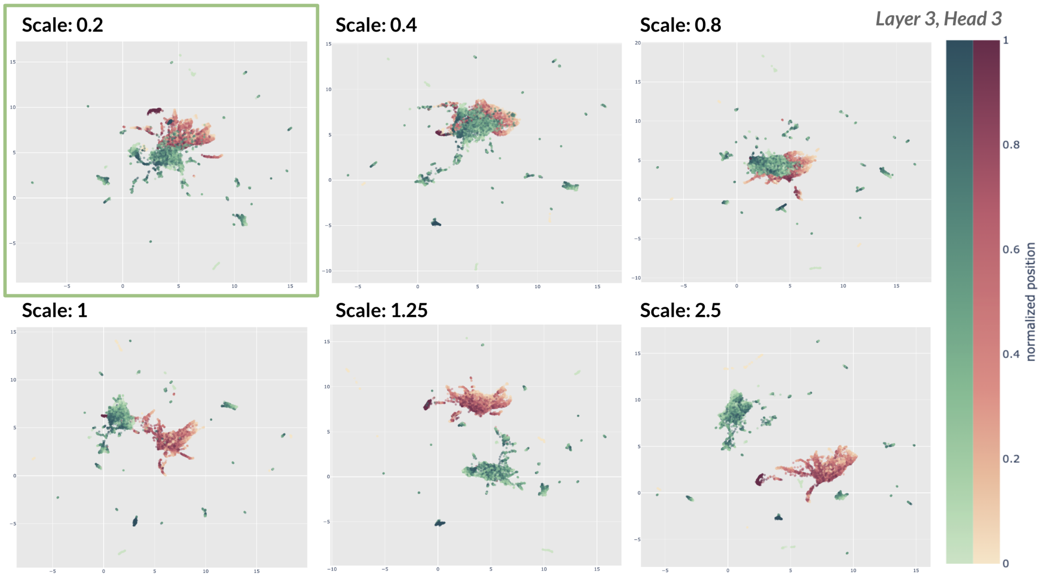

Figure 4: The joint embedding visualization after queries and keys have been scaled and translated, generated for the attention head at layer 3 head 3 in GPT. The optimal constant identified by our methods is $c = 0.2$ as highlighted.

For both of the cases displayed, our method chooses a value of c that yields a strong visualization where the query and key vector clouds are overlapping rather than disjoint. Note that these are visualizations of the query and key embeddings after they have been scaled by the respective constant and then translated so the query and key clouds have the same centroid. Embeddings are generated UMAP using the cosine distance metric.

1.4 Future Directions

There are several directions in which this work could be continued or extended. First, though we focus on optimizing the correlation between dot product and distance, it remains unclear whether this is the best proxy for attention visualization quality. There may well be several other metrics that could be employed (including the ratio of norms, like we explore in the third metric). Furthermore, the current visualizations only show the query and key embeddings and attention patterns at large and do not depict any particular relationships between individual queries and keys. Future work could look into investigating certain patterns in the visualizations at a more zoomed-in level (e.g: Do noun queries attend to pronoun keys? For a given attention head, how does it match keys and queries?).

2 Exploring Induction Heads in BERT

2.1 Motivation: Why study induction heads in BERT?

In-context learning is a phenomenon observed in language models where the models are better at predicting tokens later in the context than earlier ones, even without additional training [3]. In conjunction with observing this phenomenon, previous research has hypothesized that induction heads are the mechanism for the majority of in-context learning [3]. Despite the importance of induction heads, their specific behaviors and why they develop remains a largely unanswered question.

Current literature largely focuses on in-context learning in unidirectional models like GPT. Induction heads have not been previously explored in bidirectional models like BERT; however, the emergent in-context learning behaviors and induction heads found in unidirectional transformer models, as well as cases of prompt-based learning seen in bidirectional models like T5 and BERT [4], point toward the potential existence of induction heads in BERT.

In particular, in this work, we focus here on *behavioral* induction heads (rather than mechanistic ones), defined as attention heads that exhibit prefix matching or copying behaviors observed through attention patterns on out-of-distribution sequences made of repeated random tokens (RRT) [5].

2.2 In-Context Learning in BERT

We first run an initial experiment to check whether BERT exhibits in-context learning. For unidirectional transformer models, Olsson et al. [3] define an in-context learning (ICL) score as the loss of the 500th token in the context minus the loss of the 50th token in the context, averaged over several dataset examples. We replicate a similar heuristic for BERT, where we set the ICL score as the average difference in token-prediction loss for two varying context lengths of 50 and 500:

$$\text{ICL Score} = Loss(\text{50-token context}) - Loss(\text{500-token context})$$

We tokenized the Hugging Face Wikipedia-simple dataset, and selected the first 50 and 500 tokens of an article as the two model contexts. We then chose a random token to mask from each context, used BERT to predict the masked token, and computed the loss. Figure 5 displays the ICL scores computed across 50 trials. The mean difference was 0.23, demonstrating a noticeable difference in performance between the two contexts and a signal of in-context learning in BERT.

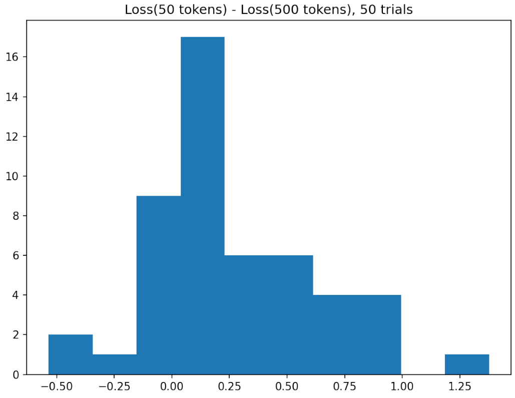

Figure 5: The in-context learning (ICL) score for BERT, defined as the loss difference between a 50-token context and a 500-token context, is shown for 50 trials.

2.3 Methods: Random Repeated Token (RRT) Experiments

To visually explore induction heads in BERT, we drew particular inspiration from the Induction Mosaic [6], a mosaic heatmap of the induction heads in 40 open-source transformer models. In the Induction Mosaic, induction scores were calculated by giving each model a sequence of repeated random tokens and measuring the average attention each head paid to the token after the previous copy of the current token.

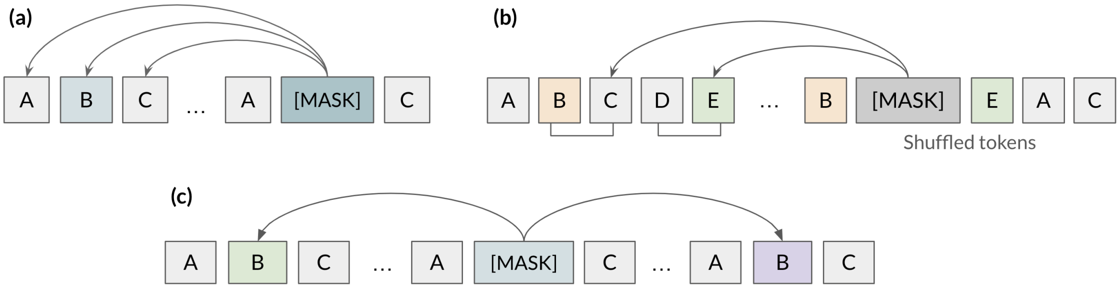

Figure 6: The three RRT experiment setups, with arrows denoting the observed attention values placed on "inductive" tokens by the masked token that were used to compute induction scores. **(a)** The standard RRT experiment setup, with a series of random tokens followed by the same tokens repeated in the same order. **(b)** A series of random tokens, followed by the same tokens but in a shuffled order. **(c)** Bidirectional attention is observed by repeating the same random tokens three times.

We employ a similar experimental setup for exploring potential induction heads in BERT. First, we generate a sequence of 200 random tokens. We then repeat this sequence of tokens, with several experiment variations (standard, shuffled, repeated on both sides) as shown in Figure 6, to generate a set of random repeated tokens (RRT). We randomly select a token from the second set of random tokens to mask and then pass the masked RRT to the BERT model. From here, we observe the output attention placed from the masked tokens to "inductive" areas, such as the previous instance of the masked token and the tokens before/after the previous instance.

For a specific attention head, the average output attention to these inductive tokens, computed across 50 trials, is defined as the induction score. The induction score represents the likelihood for a specific attention head to be an induction head. We generate an "induction map" visualization which displays the induction scores for each attention head as a heatmap.

2.4 Results: Induction Maps

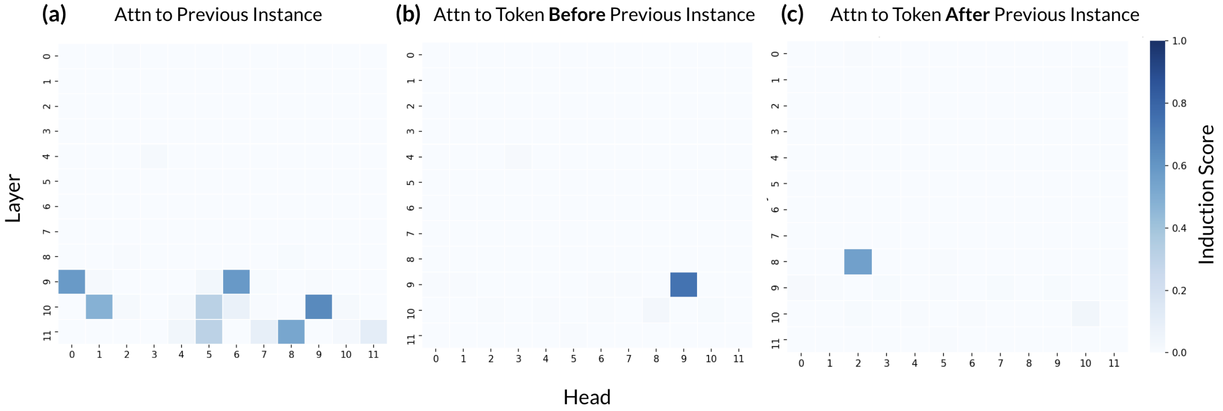

Figure 7: Induction map visualizations for the standard RRT experiment setup, depicted in Figure 6a.

We first generated heatmaps for the standard RRT experiment setup, depicted in Figure 6a, where attention values were observed from the [MASK] token to the previous instance and the tokens before and after the previous instance. The induction maps for these three cases are shown in Figure 7. In the first heatmap, we see that there are several attention heads with strong induction scores. Interestingly, note that all of these high-scoring heads are located in the last three layers. In the cases of observing the attention to the tokens before/after the previous instance, we see in each case that only one head (layer 9 head 9 for the token before and layer 8 head 2 for the token after) gives a strong induction score.

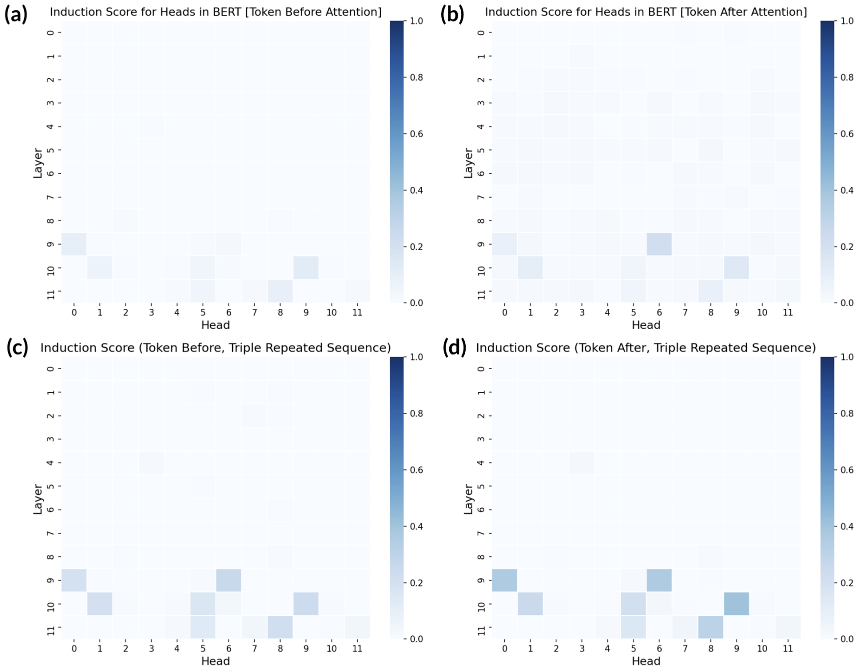

Figure 8: (a, b) Induction map visualizations for the shuffled RRT experiment setup, depicted in Figure 6b.**(c, d)** Induction map visualizations for the bidirectional RRT experiment setup, depicted in Figure 6c.

We also generated induction maps for the shuffled and bidirectional RRT experiment setups. For the shuffled version, we observed attention to the token after the previous instance of the token before the masked token (Figure 8a) and the token before the previous instance of the token after the masked token (Figure 8b). Both yield faint induction scores, but interestingly for the same heads as those in Figure 7a. For the bidirectional attention experiment, we observe attention to the previous and later instances of the masked token, shown in Figures 8c and 8d which again yields the same induction patterns as Figure 7a, but attention is now spread across the two instances.

2.5 Do induction heads always appear in later layers?

We observed that all the attention heads with high induction scores occurred in later layers, and were curious to explore this pattern more. To investigate this further, we ran the same RRT experiment on four BERT model variations that have different sizes: Tiny (2×2), Small (4×8), Medium (8×8), Large (24×16). The induction maps for each of these models are displayed in Figure 9. For the Small, Medium, and Large BERT models, we notice the same pattern that the high-scoring attention heads all occur in the later layers.

Similar trends of induction heads appearing in later layers are also observed in unidirectional models like GPT, as shown in the Induction Mosaic [6]. A potential hypothesis for why this may be the case is that if induction heads mostly serve to directly predict the next tokens, it makes sense that they're in later layers as there is less distance to communicate information down.

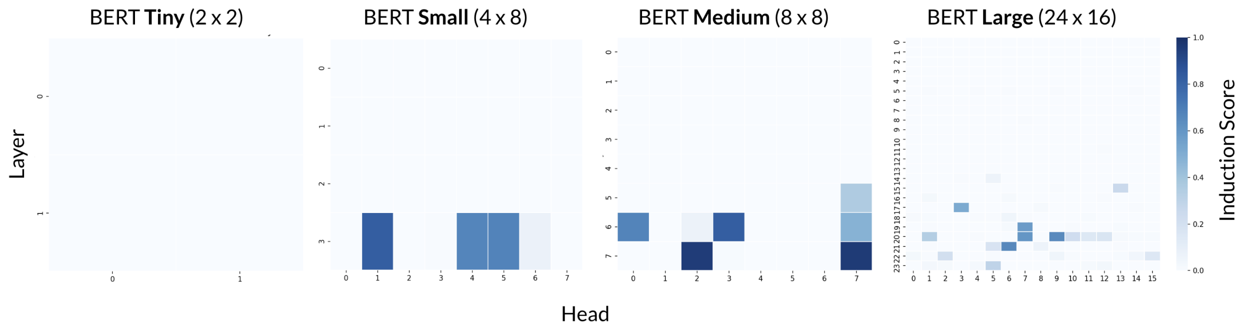

Figure 9: Induction maps generated for the Tiny, Small, Medium, and Large variations of the BERT model. All of these models present induction heads which only occur in later layers.

2.6 Future Directions

The results from our initial investigation into induction heads in BERT pose several interesting questions and possible future avenues of exploration. Future work could look into the effects of model ablation, in particular ablating heads with a high induction score (such as L9H9 and L8H2). An interesting question to explore here would be whether a "backup" induction head would appear if we removed these heads or zeroed out their weights, a behavior which was noted in GPT-2 [7]. Another distinct pattern worth investigating is how induction heads only occur in the latest layers, which could point towards existence of heads performing tasks beyond just direct repetition and copying. Potentially, individual heads could serve unique purposes, such as focusing on language translation, by attending to specific types of tokens. Finally, some other experiments to run would be masking more than one token and observing joint attention, and also exploring the relationship between tokens and words and how attention is distributed amongst them.

3 Conclusion

As a whole, the findings of this work culminate in two main questions: 1) How can we best visualize attention patterns? 2) What is the significance of the induction head behavior exhibited by BERT? The results of this study may serve as a stepping stone for future experiments in these areas to address the unexplained observations and patterns noted.

To conclude, I would like to express gratitude to a few individuals without whom this work would not be possible. I am grateful to Professor Martin Wattenberg and Professor Fernanda Viégas of Harvard University for giving me the unique opportunity to be a part of the Insight and Interaction Lab and for providing guidance and critical insights along the way. I would also like to thank Catherine Yeh, a graduate student at the Insight and Interaction Lab, for her constant mentorship, support, and encouragement. Finally, I am grateful to Neel Nanda for providing inspiration through the Induction Mosaic project and discussing ideas and questions over email.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas, and M. Wattenberg. AttentionViz: A global view of transformer attention. *Preprint*, 2023.

[3] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[4] A. Patel, B. Li, M. S. Rasooli, N. Constant, C. Raffel, and C. Callison-Burch. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*, 2022.

[5] A. Variengien. Some common confusion about induction heads. `https://www.lesswrong.com/posts/nJqftacoQGKurJ6fv/some-common-confusion-about-induction-heads`. Accessed: 2023-04-18.

[6] N. Nanda. Induction Mosaic. `https://www.neelnanda.io/mosaic`. Accessed: 2023-04-18.

[7] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.