The Heads Hypothesis: A Unifying Statistical Approach Towards Understanding Multi-Headed Attention in BERT

Madhura Pande, Aakriti Budhraja, Preksha Nema, Pratyush Kumar and Mitesh M. Khapra
Department of Computer Science and Engineering, IIT Madras, India
Robert Bosch Center for Data Science and Artificial Intelligence (RBC-DSAI) IIT Madras, India
{mpande, abudhra, preksha, pratyush, miteshk}@cse.iitm.ac.in

Abstract
Multi-headed attention heads are a mainstay in transformer-based models. Different methods have been proposed to classify the role of each attention head based on the relations between tokens which have high pair-wise attention. These roles include syntactic (tokens with some syntactic relation), local (nearby tokens), block (tokens in the same sentence) and delimiter (the special [CLS], [SEP] tokens). There are two main challenges with existing methods for classification: (a) there are no standard scores across studies or across functional roles, and (b) these scores are often average quantities measured across sentences without capturing statistical significance. In this work, we formalize a simple yet effective score that generalizes to all the roles of attention heads and employs hypothesis testing on this score for robust inference. This provides us the right lens to systematically analyze attention heads and confidently comment on many commonly posed questions on analyzing the BERT model. In particular, we comment on the co-location of multiple functional roles in the same attention head, the distribution of attention heads across layers, and effect of fine-tuning for specific NLP tasks on these functional roles. Code is made publicly available at https://github.com/iitmnlp/heads-hypothesis

Introduction
In the short span of two years, the BERT model (Devlin et al. 2019) has become a top contender in many NLP tasks. This success has led to a sub-field of research that tries to analyze why BERT works. This sub-field has become so prominent that it has even received its own name - BERTology. The state of BERTology is well captured by the summary in a recent survey paper: “while BERTology has come a long way it is fair to say we still have more questions than answers about how BERT works” (Rogers, Kovaleva, and Rumshisky 2020). We see these questions as being broadly classified into three types: (a) what does BERT know, (b) what are BERT outputs sensitive to, and (c) what do the attention heads in BERT attend to? For identifying what BERT knows, a typical strategy is to run probing experiments, for instance by evaluating POS tagging with contextual embeddings from a specific layer of BERT (Liu et al. 2019). For understanding how BERT outputs vary, ablation studies such as pruning attention heads and dropping layers are performed (Budhraja et al. 2020; Michel, Levy, and Neubig 2019; Sajjad et al. 2020). Quite separate from these two questions, the third question aims to analyze the intrinsic properties of the attention heads, which are characteristic of Transformer-based models (Vaswani et al. 2017). Analysis and interpretation of attention patterns has been a matter of debate (Jain and Wallace 2019; Wiegreffe and Pinter 2019). However, it is agreed that studying a single attention head at a time enables us to structurally localize linguistic knowledge within the model. In this work, we survey the current approaches to analyze attention heads, identify some deficiencies, and propose a unifying and statistically robust alternative.

Multi-headed self-attention is characterised by attention weights which specify the weights of other tokens when computing the representation of the current token. A natural curiosity is if these attention weights encode patterns that are linguistically or structurally meaningful. More specifically, analysis aims to classify the role of an attention head into functional roles such as local (attending to tokens in a small neighbourhood of the current token), syntactic (attending to tokens which are syntactically related to the current token), delimiter (attending to delimiter tokens such as [CLS] and [SEP]) and block (attending to tokens within the same sentence). Several methods have been proposed for such classification of functional roles of attention heads. However, these methods vary significantly from one work to the other and also amongst the different functional roles. For example, (Kovaleva et al. 2019) use the attention heatmap as an image and a CNN-based classifier to determine whether the image exhibits a local attention pattern. Contrast this with the method used in (Voita et al. 2019) which classifies a head as local if 90% of the attention of the head lies within a small neighbourhood of the current token. Contrast this further with the method used in the same work to identify a syntactic role: A head is labeled as syntactic if the classification accuracy of this head on certain syntactic roles is higher than baseline classifiers. The “higher than” assertion is itself based on the subjectively chosen parameter of 10%.

Thus, existing studies present a diverse set of methods for analysis with their respective parameters which impedes systematic comparison and evolution of ideas. Furthermore, existing studies have an additional challenge: statistical significance. Clearly, as we change the input sequence (one or
2. How are functional roles distributed across layers? Existing works have shown that more syntactic attention heads are found in middle layers as opposed to other layers (Jawahar, Sagot, and Seddah 2019; Hewitt and Manning 2019; Goldberg 2019). Our work with hypothesis testing on sieve scores provides a fresh perspective on these studies. We can define fine-grained functional roles such as separating out the individual syntactic relations and delimiter tokens. Then we can quantify the fraction of heads in each layer that are skilled to perform each of these fine-grained roles. In this work, we quantify such skills across layers and reaffirm existing findings.

3. What is the effect of fine-tuning on functional roles? A common workflow of BERT is to fine-tune the model for specific NLP tasks. Earlier studies (Kovaleva et al. 2019) have shown that only the last few layers specialize during fine-tuning. We can analyse such specialization by quantifying the change in the sieve bias across layers and functional roles. In this paper, we compute these changes in sieve bias across four different NLP tasks from the GLUE benchmark.

In summary, we propose a unifying approach to analyze attention heads which generalizes across functional roles and is statistically significant. The new metrics and visualization tools enable a fresh commentary on existing findings from BERTology.

Related Work
As mentioned earlier, the field of BERTology deals with three main questions: (a) what does BERT know? (Lin, Tan, and Frank 2019a; Tenney et al. 2019; Ethayarajh 2019) (b) what are BERT outputs sensitive to? (Ethayarajh 2019; Wiedemann et al. 2019; Mickus et al. 2019) and (c) what do the attention heads in BERT attend to? (Kovaleva et al. 2019; Clark et al. 2019; Cui et al. 2019). Of these, in this work we focus on the last question. In particular, we are interested in understanding the functional role of heads in BERT. In this section, we review existing work along two different axes.

Types of roles identified: Different works have identified a variety of functional roles for attention heads. These are: local: heads which attend to tokens in a small neighborhood around the input token, typically, previous and/or next tokens (Clark et al. 2019; Kovaleva et al. 2019; Hut et al. 2019). syntactic: heads which attend to tokens which are syntactically related to the input token. The syntactic relations are identified using a dependency parser (Voita et al. 2019; Clark et al. 2019; Hut et al. 2019; Kovaleva et al. 2019; Correia, Niculae, and Martins 2019).
vertical/delimiter: heads which attend to the [SEP] and [CLS] tokens in the input sequence (Clark et al. 2019; Kovaleva et al. 2019). Some works refer to such heads as diagonal heads (Kovaleva et al. 2019) or positional heads (Voita et al. 2019).
block: heads which attend to tokens within the same sentence as opposed to tokens in other sentence before (or after) the [SEP] token (Kovaleva et al. 2019).
interrogative: heads which attend to question marks at the end of a question (Correia, Niculae, and Martins 2019).
In this work, we focus on local, syntactic, block, and delimiter heads which are commonly studied in existing works.

Methods used for classifying a head into a functional role: Once the above functional roles are defined, existing works classify heads into these roles using a scoring function and an appropriate threshold. However, across different works there is no uniformity in the scoring function and the thresholds used for different functional roles. For example, Voita et al. classify a head as local if in at least 90% of the cases, the head pays maximum attention to tokens which are immediately before or after the input token. In contrast, for determining whether a head is syntactic or not, its performance in classifying syntactic roles is compared with a simple position based baseline classifier. A head is then labelled as syntactic if its performance is at least 10% higher than the baseline classifier. Clark et al. also follow a similar strategy wherein they classify a head as syntactic based on its performance on identifying syntactic roles. However, unlike Voita et al. they do not use a threshold of 10% but instead look for heads which substantially outperform a position based baseline classifier. Kovaleva et al. use a CNN based classifier to assign roles to heads based on attention heatmaps. They also propose certain semantic and syntactic tests for attributing roles to heads. For example, they check if certain heads pay more attention to predicates of core frame elements as captured in FrameNet. Existing methods also differ in the treatment of special tokens such as [SEP] and [CLS]. For example, (Htut et al. 2019) and (Lin, Tan, and Frank 2019b) exclude these tokens from their analysis (and thus don’t consider delimiter heads). In contrast, (Clark et al. 2019) show that delimiter heads are prominent in the later layers and thus important. Thus, the mechanisms for determining head’s functional role not only varies from one work to another but also from one functional role to another.

Another limitation with existing work is that they rely on average functional scores of heads across a large number of input sequences. Since the attention patterns (e.g., of syntactic heads) vary sensitively from one input sequence to another, this raises the question of statistical significance. Indeed, in our experiments we find that averages are inflated by outliers sufficiently to incorrectly infer functional roles that are not statistically significant. In this work, we address these deficiencies by proposing a statistically sound unified approach which can be adapted to any functional role.

Unifying Analysis of Attention Heads

The BERT Model

We analyze the BERT model which is a stack of layers each consisting of multi-headed self-attention followed by a fully connected network. The self-attention scheme works as follows: For every token in the input sequence an attention head transforms the token’s input embedding into key, value, and query vectors, which are then linearly combined based on attention weights. The outputs of all heads in a layer are concatenated and then passed through the fully connected network. The embeddings for the next layer are generated by combining the output of a layer with a skip connection from the previous layer. We specifically consider the BERT\textsubscript{BASE} configuration (Devlin et al. 2019) which has 12 layers with 12 attention heads each, and embedding vector of size 64.

Attention Sieves

As noted, attention heads are classified to have different functional roles: local, syntactic, block, and delimiter. Our primary aim is to define a unifying scheme to analyze attention heads across all functional roles. The first observation is that attention heads for different functional roles are expected to attend to different sets of tokens. We characterize this set of tokens for a given attention role as an attention sieve. Formally, attention sieve $S_{I,f}(t)$ for input sequence $I$, functional role $f$, and token $t$ is the set of all tokens of $I$ that should be attended to compute the embedding of $t$ by an attention head classified to be of functional role $f$. We illustrate this for the example sequence shown in Figure 1. A local attention head has a clear sieve that does not depend on the content of the input sequence. The local attention sieve for a token are all tokens around its neighborhood within a distance of say 2, i.e., 5 tokens. For a syntactic attention head we have to first identify all tokens syntactically related to the current token with a dependency parser. These set of tokens would define the syntactic attention sieve. Similarly, we define block and delimiter attention sieves. These attention sieves can be further partitioned to consider sub-types of functional roles. For instance, a delimiter attention sieve can be partitioned to a [CLS] attention sieve and a [SEP] attention sieve. A local attention sieve can be partitioned to a prev-local or a next-local attention sieve that defines locality in only one direction. Thus, for all tokens of an input sequence, we get attention sieves for each functional role.

Sieve Bias Score

For a token $t$ of an input sequence $I$, an attention head $h$ computes an attention weight $\alpha_h(t,t')$ for each token $t' \in I$ as follows:

$$\alpha_h(t,t') = \text{softmax} \left( \frac{q_h^t k_{t'}}{\sqrt{d_t}} \right),$$

where $q_h$ is the query vector for token $t$, and $k_{t'}$ is the key vector for token $t'$ in the sentence. $d_t$ is the size of the embedding. This weight captures the relative contribution of $t'$ to the current token. If we are to posit that an attention head has a specific functional role, then attention weights to tokens in the sieve of that functional role must be higher. Formally, if attention head $h$ is of functional role $f$ then $\sum_{t' \in S_{I,f}(t)} \alpha_h(t,t')$ should be relatively large. This aggregation of attention weights must be normalized to consider both the number of tokens in a sieve and the number of tokens in the input sequence. For example, the sieve of a local attention role may have 5 tokens, while that of a delimiter role may have only 2. Even within the same attention role, different words may be broken into varying number of tokens leading to different sizes of attention sieves. For attention head $h$, current token $t$, input sequence $I$, and functional role $f$, we can capture this normalization with a sieve bias score as:

$$\beta_h(t,I) = \left( \frac{\sum_{t' \in S_{I,f}(t)} \alpha_h(t,t')}{|S_{I,f}(t)|} \right) / \left( \frac{|S_{I,f}(t)|}{|I|} \right)$$

13615
Figure 1: Attention sieves for a sample input sequence: Tokens in boxes show the current token and edges point to other tokens that are in its sieve for various functional roles: (a) block, (b) nsubj (syntactic), (c) local, (d) amod (syntactic) and, (e) delimiter.

Figure 2: Sieve score distribution across sentences for a head. Note the distribution is skewed with large variation.

In the numerator, we have the average attention weight of the head to tokens within the sieve $t_s \in S_{t,f}(t)$. In the denominator, we have the average attention paid by the head to all tokens in the input sequence. Higher this ratio, higher is the attentive bias to the sieve.

A randomly initialised head would pay roughly equal attention weights to all tokens and thus a sieve bias score of 1 for all functional roles. If upon learning, a head becomes more selective, then its sieve bias score would increase for specific sieves. Our concern then is to identify for each head the functional roles for which sieve bias scores are high.

Hypothesis Testing

We have established a relation between a head’s high sieve bias score and its functional role. But the sieve bias scores $\beta^f_h(t, I)$ vary across input sequences. Can we take the average sieve bias score across many input sequences? This would work if the sieve bias scores are symmetrically distributed with a relatively narrow standard deviation across input sequences. On the contrary, we found that the distributions can be skewed and broad, as shown in Figure 2 for a specific head. In the presence of such variation, simple averaging is error-prone, and we instead need robust statistical inferences. For instance, an assertion based on the average being ≥ 3 would assign the functional role to the head in Figure 2. However, hypothesis testing to assert that population mean is more than 3 would fail. To the best of our knowledge, no current work on analyzing attention heads considers either this variation or a statistical treatment.

We define the null hypothesis as the mean of the sieve bias score $\beta^f_h$ across all input sequences is less than or equal to some threshold $\tau$.

This threshold $\tau$ specifies how sharp the attention bias should be to classify the functional role of a head: Higher the value of $\tau$, stricter is the classification rule. We can evaluate this hypothesis by using a one-tailed test of the mean with z-statistics, since population variance is unknown. To this end, we can measure $\beta^f_h$ for a large number $N$ of input sequences as the sample. We can then compute the sample mean of $\beta^f_h$ and derive the p-value using the one-tailed test. As is the convention, we say that the null hypothesis is rejected if the p-value is smaller than 0.05. If the null hypothesis is rejected, it follows that the attention head $h$ is satisfying the functional role $f$. If the p-value is greater than 0.05, then the test is inconclusive, and we do not ascribe the functional role to the head.

In summary, our proposed method involves the following steps: (a) computing attention sieves for all functional roles, (b) computing sieve bias scores for all pairs of attention heads and functional roles, and (c) performing hypothesis testing on the mean sieve bias score across input sequences to ascribe functional roles to each head.

Results and Discussions

In this section, we first describe our experimental setup and the datasets used for our analysis. We then provide details about the assignment of functional roles to heads using our method described earlier. Finally, based on these assigned roles, we discuss the three questions outlined earlier: i) Are functional roles mutually exclusive? ii) How are functional roles distributed across layers? and iii) What is the effect of fine-tuning on functional roles?

Experimental Setup

We pre-train the BERT$_{BASE}$ model using the English Wikipedia corpus and a subset of of the Project Gutenberg corpus released by (Lahiri 2014). We pre-trained the model for around 300K steps with a sequence length of 128 and another 1K steps with a sequence length of 512 and a batch size of 2K tokens (Devlin et al. 2019) on a single Cloud TPU (v3-8). We then tune this model individually for four NLU tasks from the GLUE benchmark (Wang et al. 2018): QNLI (QA Natural Language Inference), QQP (paraphrase detection), MRPC (paraphrase detection) and SST-2 (sentiment analy-
Assigning Functional Roles to Heads

For each task, we feed all the input sequences to the fine-tuned \(\text{BERT}_{\text{BASE}}\) model and compute the sieve bias score \(\beta^f_h(t, I)\) for each input sequence \(I\) as described earlier. This gives us a sample of 1000 sieve bias scores (one score corresponding to each input sequence). Once we compute these scores for a given head \(h\), and a given functional role \(f\), we evaluate the null hypothesis as discussed in the earlier section based on the obtained p-value for each head. If the null hypothesis is rejected then it means that the head indeed performs this specific functional role. Note that we perform this hypothesis test independently for all heads and the four functional roles. This method thus allows multiple roles to be assigned to the same head.

It is important to discuss the choice of the threshold \(\tau\). A large value of \(\tau\) would lead to the null hypothesis being rejected for most heads and thus fewer functional roles assigned to heads. A small value of \(\tau\) would lead to multiple functional roles being assigned to the same head. One approach to set the value of \(\tau\) is to subjectively decide what constitutes attentive bias, i.e., to decide the ratio of attention that a head must pay to tokens in a sieve than to all tokens. For instance, one may suggest that \(\tau = 2\) with double the attention to the sieve tokens than to other tokens is a reasonable choice for assigning functional roles to heads. Any such value of \(\tau\) can be decided and the resulting functional roles can be analysed as we do in the rest of the paper.

In addition to this subjective check, we set the value of \(\tau\) informed by data on the sieve bias scores. Specifically, we plot the cumulative frequency distribution (see Figure 3(a)) of the sieve bias score across all functional roles for 1,000 input sequences, and identify that the mean sieve bias score is around 2.8. We then set \(\tau\) to be the smallest integer larger than this mean, i.e., 3. We then perform the described hypothesis testing and compute the p-values. We plot the individual p-values across all the input sequences (see Figure 3(b)). We observe that this distribution is distinctly bimodal with values either close to 1 or close to 0. Indeed, there are only 1.8% of points in the range (0.05, 0.95). This indicates that the hypothesis testing made clear inferences: Either the null hypothesis was rejected (due to p-values less than 0.05) or the test was inconclusive but with very high p-values. In the former case, we assign the functional role to the head with high confidence, and in the latter case the head is clearly not satisfying that functional role as indicated by the high p-values. Thus, we set the threshold \(\tau = 3\) to meet the two criteria: a subjective criterion of attentive bias and the objective requirement that the hypothesis testing generates clear inferences. Further, note that this threshold is interpretable as it simply means that the attention on the tokens within the sieve is 3 times the attention on tokens outside the sieve. Based on this statistically sound unified approach of classifying heads we now revisit some commonly asked questions while analysing attention heads.

Are Functional Roles Mutually Exclusive?

As mentioned earlier, existing methods typically assign only one functional role to a head. In contrast, as we do the hypothesis test independently for each functional role, the same head can potentially be labeled with multiple functional roles. To find out whether this is really the case, we present a Venn Diagram where heads being labeled with the same role constitute of a set (see Figure 4). We make the following observations from this figure:

**Density of delimiter roles:** Across tasks, there are significant number of delimiter heads (on average 73.43% across the 4 tasks). Also given their high prevalence, they naturally intersect with other functional heads.

**Unskilled heads:** There are very few heads which do not perform any functional role (avg. 9.37% across the 4 tasks). Overlap between local and syntactic roles: There is a significant overlap between local and syntactic heads. More specifically, across the four tasks, 42.1% to 88.8% (with a mean of 70.9%) of syntactic heads overlap with local heads. This observation has an interesting connection to other stud-
ies which show that restricting attention to local patterns can still lead to good performance (Yang et al. 2019). Our findings lend more credence to this observation, as they indicate that by attending to local tokens a head invariably also attends to syntactically related tokens. This is simply because many syntactically related tokens are in a small neighborhood around the input token. To further strengthen this claim, we compute the Spearman correlation between the local and syntactic sieve scores. We observe a very high correlation score of 0.78, 0.81, 0.85, 0.73 scores for QNLI, QQP, MRPC, SST-2 tasks respectively, with a p-value of 0. We further split the syntactic and location heads based on type of syntactic relations and location of attended token (previous or next) respectively. We observe that nsubj heads have a high overlap with next-loc heads and dobj-heads have a high overlap with the prev-loc heads.

**Role of block heads:** There is only a marginal overlap between block heads and syntactic/local heads indicating that these specialisations are unique. Note that there are no block heads in the SST-2 task as it is a single sequence task.

**How Are Functional Roles Distributed across Layers?**

Existing works have shown that the middle layers in BERT are more important as they contain many syntactic heads (Hewitt and Manning 2019; Goldberg 2019; Clark et al. 2019). Our results suggest that while this is largely true, there is significant variance across tasks. To understand this, we refer the reader to Figure 5. Every larger gray colored cell in this figure corresponds to one head and each row corresponds to one layer. Thus, there are a total of 12 × 12 = 144 larger gray squares in each plot. We further divide each of these larger gray squares into 9 smaller squares. Each of these 9 squares corresponds to a different functional role represented by a different color as defined in the legend. For this analysis, we have split the four coarse functional roles into more fine-grained functional roles. In particular, we have split the delimiter role into [SEP] and [CLS] roles. Similarly, we have split the syntactic role into nsubj, dobj, amod and advmod roles. Squares which are predominantly gray correspond to heads which perform very few functional roles. On the other hand, multi-colored squares correspond to multi-skilled heads, i.e., heads performing multiple functional roles. For example, in the first plot in Figure 5 (QNLI), the first head in Layer7 performs eight different functional roles, whereas the last head in the same layer performs only one functional role [SEP]. Also, to aid better visualisation, in each layer we have sorted the heads based on the number of functional roles they perform (hence, the more multi-colored heads appear at the beginning of the row). Based on these plots we make the following observations:

**Multi-skilled heads across layers:** The number of multi-skilled heads is higher in the middle layers (layers 5 to 9) for MRPC, QNLI and SST-2 tasks. However, for QQP the number of multi-skilled heads is relatively higher in the initial layers. This suggests that the role of heads across layers varies from one task to another. Further, these observations may differ since earlier studies have been mostly based on analyzing a single functional role at a time.

**Distribution of delimiter heads:** Across tasks we find a large number of delimiter heads ([SEP] and/or [CLS]) across all layers. In fact, such delimiter heads are often the only functional heads in the last three layers across different tasks. Interestingly, the first layer contains very few (if any) delimiter heads. This may be because the delimiter tokens [CLS] and [SEP] start accumulating meaningful information from the entire sequence only after the first layer.

**Distribution of block heads.** For all the tasks which take two sentences as inputs (MRPC, QNLI and QQP) block heads are prevalent in the zeroth layer. However, their presence in the later layers varies from one task to another. For example, in both QNLI and MRPC, block heads consistently show up in layers 1 to 4, in addition to layer 0. However, in QQP block heads do not show up in the middle layers but are prevalent in the first and last layer. Lastly, for MRPC block heads consistently show up across all layers.

**Distribution of syntactic heads.** Across all tasks there are at least 3 syntactic heads (nsubj, dobj, amod and advmod) for all layers (with very few exceptions). Across tasks, the prevalence of syntactic heads is higher in layers 6 to 8 which is in line with observations made in other studies (Rogers, Kovable, and Rumschisky 2020; Jawahar, Sagot, and Seddah 2019; Liu et al. 2019). However, our findings suggest that the heads in these layers do not just perform syntactic roles but are highly multi-skilled. Lastly, across tasks the last two layers have very few syntactic or multi-functional heads.

Figure 4: Venn diagram with each circle representing a set of heads with a specific role. The intersection between various sets indicate that many heads are multi-functional. Note that the size of the circle is proportional to the cardinality of the set.
What is the Effect of Fine-Tuning on Functional Roles?

Existing works have shown that the final layers of BERT are more task specific. For example, Kovaleva et al. show that the weights in the final layers change a lot during fine-tuning, whereas, Hao et al. show that the performance is less sensitive to weights in the middle layers. To substantiate these findings we measure the change in the average sieve bias scores for the 4 functional roles across initial, middle, and final sets of 4 layers for the 4 tasks (see Figure 6). A positive (negative) value indicates that the average sieve bias score increases (decreases) after fine-tuning. We make the following observations from these plots:

**Functional roles get redistributed in the final layers:** We observe that in the final layers, there is a decrease in the average delimiter bias score \( \langle \beta_h^{\text{delimiter}} \rangle \) across all heads in these layers. On further investigation, we find that this decrease is mainly due to a lower attention on the [SEP] but the attention on the [CLS] token does not change much. This seems to substantiate the hypothesis made in Clark et al. that the [SEP] tokens serve as a no-op indicator and hence the attentive bias to those tokens is reducing as fine-tuning specializes the final layers. We also find that there is an increase in the average block bias score \( \langle \beta_h^{\text{block}} \rangle \) across all heads in these layers. Thus, the later layers are specialising to generally attend to tokens within their sentence without specific syntactic bias.

**Functional roles do not change in the initial and middle layers:** In contrast to the final layers, we observe that average bias scores do not change much in the initial layers that is in line with observations made in other studies. Hao et al.(2019) observe low sensitivity to the weights in middle layers. However, we offer a different perspective that bias scores in initial and middle layers do not change much.

**Conclusion**

In this work, we present a unified and formal approach for analysing the functional roles of attention heads. In particular, the sieve bias score generalizes across functional roles and hypothesis tests ensure statistical significance. This systematises the study of attention heads in BERTology. Based on this analysis, we make several observations. Delimiter heads are prevalent in the network and thus overlap with other functional heads. Syntactic heads often are also local heads which explains claims made in other studies about the sufficiency of local attention patterns (Yang et al. 2018, 2019; Wu et al. 2019). Middle layers have most number of multi-skilled heads, in line with observations made in other studies about the drop in performance while pruning heads in middle layers. Lastly, we are able to carefully study the impact of task-specific fine-tuning on the roles of the attention heads. We find that later layers have the largest change in their roles and specifically the number of delimiter heads drops sharply indicating specialization for task-specific roles. This again supports existing claims about the importance of later layers for specific tasks while providing a fresh perspective based on functional roles. In summary, our formal unified approach to analysing attention heads provides us the right lens to confidently comment on various questions about the functional roles of attention heads.
Acknowledgements

We thank Google for providing us with free TPUs (through TFRC program) which tirelessly ran our code day in and day out. We also thank them for supporting Preksha Nema through their Ph.D. Fellowship program. We are grateful to the Department of Computer Science, RISE Lab, Robert Bosch Center for Data Sciences and Artificial Intelligence (RBC-DSAI), and IIT Madras. We also thank the anonymous reviewers for their valuable suggestions and Madhura Pande acknowledges Maneesha for her support.

References


