From Attention to Understanding: Exploring Transformer Head Behaviors

Yousif Abdulhussein, Margaret Kelley, Alex Warden, Jingtao Zhong
COSC 524: Natural Language Processing, Fall 2025
University of Tennessee, Knoxville

***Abstract*—Transformer models rely on multi-headed self-attention to represent linguistic structure, yet growing evidence suggests that not all attention heads are equally important for downstream tasks. This raises questions about redundancy, parameter efficiency, and where meaningful attention patterns emerge within the model. In this work, we study attention head behavior in the pre-trained BERT-base-uncased model using the Stanford Natural Language Inference (SNLI) dataset, with a focus on head-level variance, cross-sentence semantic attention, and layer-wise differences in representational importance.**

We evaluate head importance using a variance-based metric and conduct parameter reduction experiments to test whether low-variance heads can be removed without degrading classification accuracy. Our findings show that shallow layers contribute little to semantic differentiation across SNLI examples, suggesting a high degree of redundancy. In contrast, deeper layers exhibit stronger and more consistent cross-sentence attention patterns aligned with the semantic reasoning required for natural language inference. Pruning low-variance heads, particularly in early layers, reduces model complexity while maintaining competitive performance.

I. Introduction

Transformer architectures have become the foundation of modern natural language processing (NLP), driving progress in tasks such as translation, summarization, and natural language inference (NLI) [1]. This effectiveness is rooted in the self-attention mechanism. This allows models to compute relationships between each token pair in a given sequence in parallel- rather than step-by-step [2]. This parallel structure allows transformers to capture long-range dependencies and learn complex linguistic patterns in large scales.

An important component of these models is multi-headed self-attention (MHA). Instead of relying on a single attention-type distribution, transformers allocate multiple attention heads to process different aspects of linguistic information in parallel [3]. Prior studies have shown that some heads reliably encode syntactic or positional information, while others capture semantic relationships across sentences [4], [5]. However, many heads remain poorly understood, and the degree to which each one contributes to downstream performance is still an open question.

Recent work has begun to examine not just *what* heads attend to, but *which* heads meaningfully contribute to model performance. This has motivated research into parameter reduction- removing or pruning redundant attention heads to improve efficiency without degrading accuracy. Such results suggest that transformers may allocate more heads than necessary, with only a subset of heads playing a critical and functional role.

Our work builds directly on this perspective. Rather than fully interpreting the behavior of individual heads, we analyze attention head importance through variance-based metrics and parameter reduction experiments performed on the BERT-base-uncased model using the SNLI dataset. We focus on three core questions:

1) Are deeper layers primarily responsible for cross-sentence semantic attention in SNLI?
2) Does head-level variance provide a reliable signal for identifying unimportant heads?
3) To what extent can pruning low-variance heads reduce parameters without harming model accuracy?

In this paper we will demonstrate shallow transformer layers behave similarly across examples and contribute little to semantic differentiation in SNLI. Additionally, that deeper layers demonstrate clearer cross-sentence attention patterns. Using a variance-based head importance metric, we then prune low-variance heads and evaluate the impact on NLI performance. Our results indicate that roughly one-fifth of attention heads (those concentrated in early layers) can be removed with only marginal performance loss- highlighting the structural redundancies all while clarifying where meaningful attention behaviors lie.

II. Background

A growing curiosity is leading to those in the field seeking to understand how transformer models distribute information across layers and attention heads. Early interpretability studies focused on whether the individual heads encode linguistic structure, while more recent work questions the explanatory value of attention weights and emphasizes model efficiency through pruning. Our study aligns with this latter direction by examining head importance through variance metrics and evaluating how pruning affects downstream NLI performance.

A. Attention Head Specialization

Clark et al. (2019) were some of the first ones to preform comprehensive investigations of attention behavior in BERT. Through their qualitative visualization and probing analyses, they find that certain attention heads consistently focus on interpretable linguistic patterns, such as delimiter tokens, local syntactic relations, and coreference-like behavior. These results lead to the idea of individual heads having

specialized functional roles, particularly within deeper layers. This supports the idea that attention patterns sometimes reflect underlying linguistic structure, even if only for a subset of heads. Our work builds on this insight by investigating where cross-sentence semantic attention emerges in BERT's layers and how that relates to model performance on SNLI.

B. Limitations of Attention as Explanation

Jain and Wallace (2019) provide a critical counterpoint to interpretability-through-attention studies. They show that attention weights often correlate weakly with gradient-based feature importance measures. They continue on to highlight that alternative attention distributions can be constructed without changing model predictions. This rests on the assumption that high attention values directly indicate causal influence. In light of this critique, we adopt a conservative stance: rather than treating attention patterns as explanations, we analyze statistical properties- specifically, the head-level variance - to assess which heads differentiate the semantic information across classes.

C. Parameter Reduction and Head Pruning

Another thread of research looks at how much redundancy exists inside Transformer models. Michel et al. (2019) demonstrates that many attention heads can be removed with minimal effect on downstream accuracy, suggesting substantial overparameterization. Similarly, Voita et al. (2019) and Lagunas et al. (2021) introduce methods for scoring head importance and pruning low-significance heads based on gradient or information-theoretic measures. These studies show that model complexity can often be reduced while preserving performance.

Our work aligns with this literature by evaluating whether low-variance heads -specifically, those in those in shallow layers- can be pruned without decreasing the accuracy on SNLI. However, in contrast to the more complex importance measures, our approach relies on a simple variance statistic computed over class-conditional attention scores. We believe this makes the procedure both interpretable and easy to implement.

III. Methodology

Our methodology is designed to examine how attention heads contribute to cross-sentence semantic reasoning in NLI. Additionally, we want to assess whether low-variance heads can be pruned without decreasing the classification performance. We evaluate attention behavior in a pre-trained BERT-base-uncased model and focus on head-level variance across the SNLI dataset.

A. Dataset

We conduct our analysis using the Stanford Natural Language Inference (SNLI) dataset, which contains sentence pairs labeled as *entailment*, *contradiction*, or *neutral*. Each example consists of a *premise* and a *hypothesis*, making SNLI a benchmark for studying cross-sentence attention patterns. This structure not only allows for us to observe how attention operates within each sentence individually, but also track how information flows between them during semantic comparison.

Figure 1 summarizes SNLI statistics, including train/dev/test sizes and class label distributions. The label distribution is relatively balanced, which ensures no single class dominates the training or evaluation corpus.

```
-- Staistics --

Training pairs: 550152
Dev pairs: 10000
Test pairs: 10000
Total pairs: 570152

Train labels: {'entailment': 183416, 'neutral': 182764, '-': 785, 'contradiction': 183187}
Dev labels: {'entailment': 3329, 'neutral': 3235, '-': 158, 'contradiction': 3278}
Test labels: {'entailment': 3368, 'neutral': 3219, '-': 176, 'contradiction': 3237}
```

Fig. 1: SNLI dataset statistics, including train/dev/test pair counts and class label distributions.

B. Experimental Setup

To establish an initial performance benchmark, we evaluate the first 1,000 samples from the SNLI test set using the pre-trained BERT-base-uncased model. Labels are encoded as 0 (entailment), 1 (neutral), and 2 (contradiction). The model achieves an overall accuracy of 89.70%, summarized in the confusion matrix shown in Figure 2.

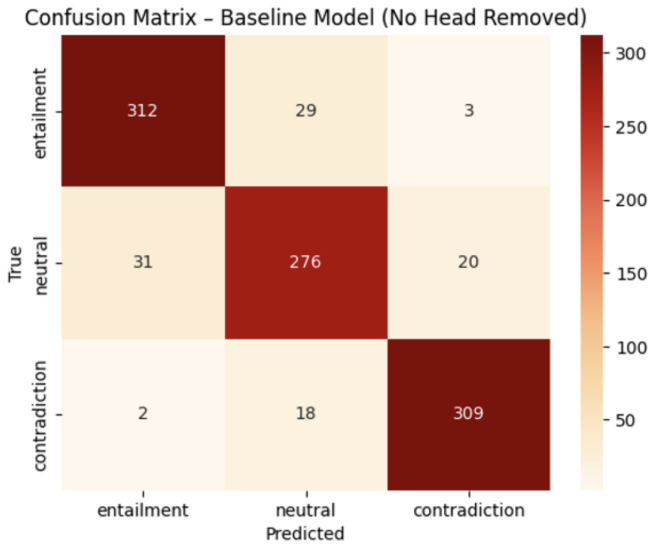


Fig. 2: Confusion matrix for the BERT-base-uncased evaluation on 1,000 SNLI test examples.

BERT-base-uncased consists of 12 Transformer encoder layers, each with 12 attention heads, yielding a total of 144 heads. We specifically avoid fine-tuning the model. By keeping BERT in its pre-trained form, we can better analyze its *inherent* internal behavior rather than adaptations induced by SNLI-specific optimization.

For each of the 144 heads, we extract cross-sentence attention scores across the 1,000 SNLI examples. To ensure that the resulting patterns reflect meaningful reasoning rather than

noise, we restrict our analysis to the 897 examples that the model classifies correctly.

C. Model Architecture Overview

BERT-base-uncased is a bidirectional transformer encoder where each layer consists of a multi-headed self-attention module. This is followed by a position-wise feed-forward network and residual connections. Each attention module operates over 12 parallel heads. Their outputs are concatenated and linearly projected back into a 768-dimensional hidden space. The feed-forward sublayer expands these hidden states to 3,072 dimensions before projecting back. This allows the model to learn nonlinear transformations.

The model processes input as a single packed sequence of tokens containing the premise, a separator token, and the hypothesis:

$$[\texttt{CLS}], \text{Premise Tokens}, [\texttt{SEP}], \text{Hypothesis Tokens}, [\texttt{SEP}].$$

Positional embeddings and token-type embeddings distinguish the two sentences, which enables the model to encode cross-sentence relationships. Because all attention heads in every layer attend to the concatenated sequence, cross-sentence attention flows naturally through the architecture, allowing us to trace which heads contribute to semantic comparison.

This modular structure makes BERT especially suitable for head-level analysis as each attention head constitutes a distinct computational channel whose behavior can be isolated, visualized, and pruned, which is critical for this work.

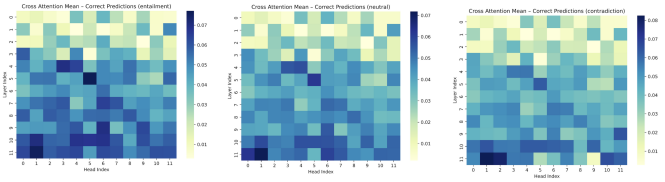
D. Variance-Based Head Importance

To quantify how strongly each head differentiates between the three SNLI classes, we compute a simple variance metric. For each layer L_i and head H_j, we define:

$$\mathrm{Var}(L_i, H_j) = \frac{1}{C} \sum_{k=1}^{C} \left(u_{i,j,k} - u_{i,j}\right)^2, \tag{1}$$

where $C = 3$ is the number of classes, $u_{i,j,k}$ is the mean cross-sentence attention score for head (L_i, H_j) on class k, and $u_{i,j}$ is the mean score across all classes. Heads with high variance respond differently across classes, which suggests a discriminative role. Conversely, low-variance heads exhibit similar attention patterns for all categories, indicating potential redundancy.

Figures 3a–3c present the mean cross-sentence attention scores for each head under the entailment, neutral, and contradiction labels, respectively. As shown, the first three layers exhibit minimal differentiation across classes, implying that their heads capture general features rather than semantic distinctions critical to NLI.



(a) Entailment (b) Neutral (c) Contradiction

Fig. 3: Mean cross-sentence attention scores for correctly predicted SNLI examples under each class label, aggregated across tokens for each head.

E. Cross-Sentence Attention Extraction

To analyze how the model compares the premise and hypothesis, we isolate the portion of each attention map corresponding specifically to cross-sentence interactions. Let P denote the token index range for the premise and H the range for the hypothesis. For each attention head (L_i, H_j) and each example, we extract the attention weights from tokens in P attending to tokens in H and vice versa. This yields two cross-sentence attention matrices:

$$A_{i,j}^{(P \to H)}, \qquad A_{i,j}^{(H \to P)}.$$

Following prior interpretability work, we aggregate cross-sentence attention by averaging across query positions and examples within each class:

$$u_{i,j,k} = \frac{1}{|X_k|} \sum_{x \in X_k} \mathrm{mean}\big(A_{i,j}^{\mathrm{cross}}(x)\big),$$

where X_k is the set of correctly predicted examples for class k and $A_{i,j}^{\mathrm{cross}}(x)$ denotes the relevant cross-sentence portion of the attention matrix for example x. This aggregated value reflects how strongly a given head distinguishes semantic relations across classes. Later, this forms the basis for our variance analysis.

The class-specific means $u_{i,j,k}$ are exactly what we visualize in the heatmaps in Figure 3: for each head (L_i, H_j) and class k, a brighter cell corresponds to a larger cross-sentence attention score. These same class-wise means then feed directly into the variance computation in (1). To summarize- the heatmaps show the raw class-specific means, while the variance metric summarizes how differently each head behaves across the three SNLI labels.

F. Parameter Reduction Experiment

For each head, we compute the class-wise means $u_{i,j,k}$ and the corresponding variance $\mathrm{Var}(L_i, H_j)$ defined in (1). We then use this variance as an importance score that drives our pruning decisions.

To determine whether redundant heads can be removed without harming performance, we conduct pruning experiments using this variance metric. Figure 4 visualizes the variance values across all 144 heads, with the top 20 highest-variance heads highlighted in green and the bottom 20 lowest-variance heads highlighted in yellow. The distribution indicates that shallow layers contribute little to semantic discrimination,

with uniformly low variance across SNLI classes. This suggests that early-layer heads are able to be pruned with limited impact on the accuracy of the classification.

In the experiments that follow, we test several pruning thresholds, including a primary threshold of 1×10^{-6}, which empirically highlights a natural separation between high-variance and low-variance heads. This threshold does not prune during the methodology stage- rather, it defines the criteria that's evaluated later in our experiments.

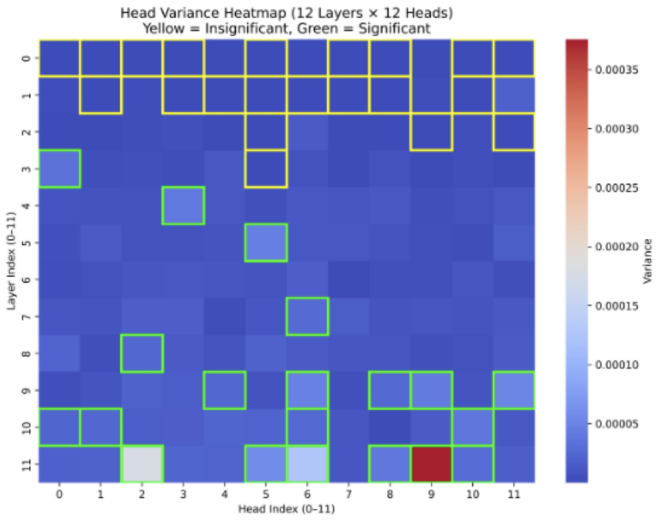


Fig. 4: Variance distribution of all 144 attention heads. Green points indicate the top 20 highest-variance heads; yellow points indicate the bottom 20 lowest-variance heads.

We then remove low-variance heads and evaluate whether BERT maintains competitive performance on SNLI. This analysis allows us to assess both (1) the semantic role of individual heads and (2) the model's tolerance to parameter reduction, directly addressing our third research question.

IV. Experiments

Our experiments aimed to evaluate whether or not variance-based head importance reliably identifies redundant attention heads. Additionally, we aimed to identify how pruning these heads affects SNLI performance. We structure the experiments around two goals: (1) determining where discriminative semantic information emerges within the model (RQ1), and (2) understanding the trade-off between parameter reduction and predictive performance (RQ2 and RQ3).

A. Head Variance Computation

Using the 897 correctly predicted examples from the SNLI test set, we compute cross-sentence attention scores for all 144 heads in BERT-base-uncased. For each head, we calculate the class-wise mean attention scores for entailment, neutral, and contradiction. Then, we apply the variance definition from Section III-D to obtain a single variance score per head. This variance serves as a proxy for head importance as heads with low variance are expected to contribute minimally to task-specific semantic reasoning.

B. Variance-Driven Pruning Strategy

To test the effect of parameter reduction, we implement two complementary pruning strategies:

- **Threshold-based pruning:** which removes all heads whose variance falls below a fixed threshold.
- **Rate-based pruning:** which removes a specified percentage of the lowest-variance heads.

These approaches allow us to evaluate whether the variance metric produces consistent pruning behavior under different selection criteria. Figures 5 and 6 show sensitivity analyses for both strategies.

When applying the primary threshold of 1×10^{-6}, we observe that 32 heads (approximately 22% of the 144 total heads, or just over one-fifth of the attention capacity) fall below this level and are pruned. Notably, the majority of these heads belong to the first three layers, consistent with the trend observed in Figure 4. Later in the report, Figure 7 visualizes which heads remain active after applying the variance threshold, showing that most pruned heads come from early layers.

C. Evaluation of Pruned Models

To determine whether or not pruning affects model performance, we evaluate both the baseline (the unpruned heads) and reduced models on the remaining 8,824 examples from the SNLI test set. We measure accuracy, macro precision, macro recall, and macro F1-score for both configurations. As described in Section IV-D, these metrics are derived from the confusion matrices of the two models. This setup allows us to quantify the performance impact of pruning and examine whether variance-based head removal compromises the model's ability to solve SNLI.

D. Evaluation Metrics

To compare the baseline and pruned models, we compute standard classification metrics for the SNLI task. Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives for a given class. Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Precision and recall are given by:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

The F1-score, which balances both measures, is:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

We report macro-averaged precision, recall, and F1-score by averaging the per-class values over the three SNLI labels (entailment, neutral, contradiction). Using these definitions, we compute accuracy over all 1,000 evaluated examples and macro-averaged metrics over the three classes.

For the baseline model, the confusion matrix in Figure 2 yields the following values:

- **Accuracy:** 89.7%
- **Macro Precision:** 89.7%
- **Macro Recall:** 89.7%
- **Macro F1-score:** 89.7%

These values are calculated directly from the confusion matrix using the equations above. For example, accuracy is given by:

$$\frac{312 + 276 + 309}{1000} = 0.897,$$

where 312, 276, and 309 are the true positives for entailment, neutral, and contradiction, respectively. Macro precision, recall, and F1-score are computed analogously by averaging over the three label-specific scores.

To summarize, these computations show that the baseline model performs consistently well across all three SNLI classes, with nearly identical precision, recall, and F1-scores. This balance indicates that the model is not biased toward any particular label and is able to correctly identify entailment, neutral, and contradiction cases at roughly the same rate. Because all evaluation metrics cluster around 89.7%, they provide a stable reference point for assessing the impact of pruning. Any meaningful degradation would appear as a noticeable drop relative to this compacted grouped baseline.

V. Results

We now examine the effect of pruning low-variance attention heads under the two strategies described in Section IV. The first approach removes all heads whose variance falls below a fixed threshold, while the second removes a specified proportion of the lowest-variance heads. As shown in Figures 5 and 6, both approaches exhibit similar trends, indicating that head importance is captured consistently by the variance metric.

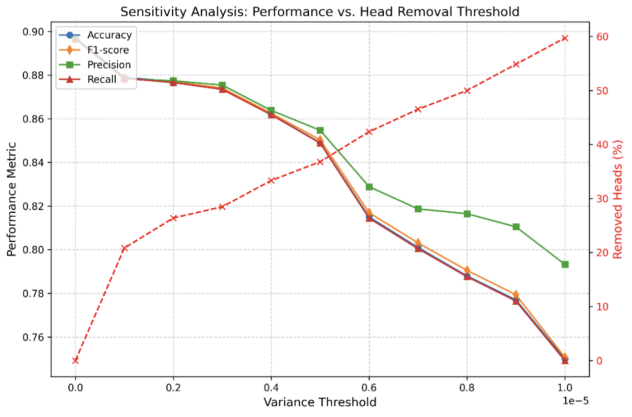


Fig. 5: Sensitivity of model performance as a function of variance threshold. Each point reflects performance after pruning heads whose variance falls below the corresponding threshold.

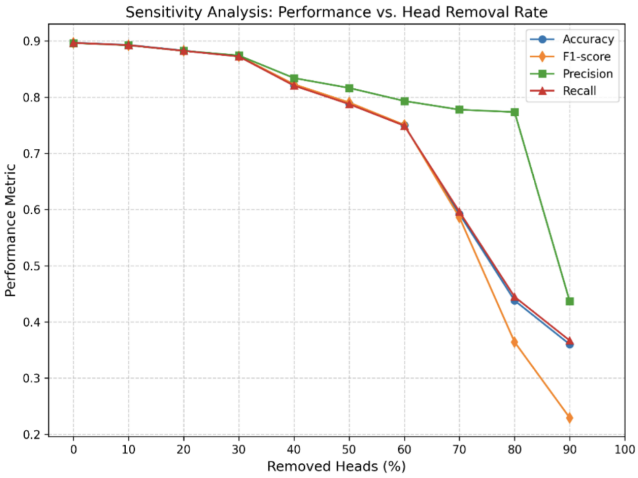


Fig. 6: Sensitivity of model performance as a function of the proportion of lowest-variance heads removed.

Removing heads with variance below 1×10^{-6} results in the elimination of 32 heads, corresponding to roughly 22% of all attention heads. This level of pruning maintains performance across all primary evaluation metrics, suggesting that low-variance heads contribute minimally to the semantic reasoning required for NLI. Figure 7 visualizes the set of remaining heads after pruning, revealing that the majority of removed heads belong to the earliest Transformer layers. Notably, most heads in the first three layers are eliminated, reinforcing the earlier finding that shallow layers provide limited class-specific differentiation.

To validate the impact of pruning, we evaluate the reduced model on the remaining 8,824 examples in the SNLI test set. Figure 7 compares the baseline and pruned models across accuracy, macro precision, macro recall, and macro F1-score. Pruning causes only a marginal decrease in each metric. The overall accuracy of the pruned model differs from the baseline by less than one percentage point, and the macro-averaged precision, recall, and F1-score show similarly small changes.

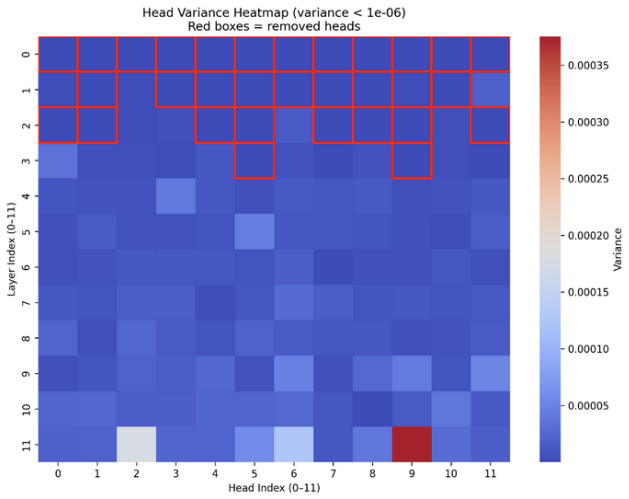


Fig. 7: Comparison of baseline and pruned models on accuracy, macro precision, macro recall, and macro F1-score. Performance differences are below one percentage point for all metrics.

For the baseline model, the confusion matrix in Figure 2 corresponds to 89.7% accuracy and 89.7% macro precision, recall, and F1-score. After pruning 32 low-variance heads (which represents roughly 22% of the total heads) the pruned model exhibits decreases of less than one percentage point across all four metrics. This confirms that the heads removed by the variance thresholding strategy contribute a minuscule amount to the model's final predictions.

From the perspective of our research questions, these results directly address RQ2 and RQ3- variance clearly serves as a useful signal for identifying unimportant heads. Additionally, pruning low-variance heads achieves a substantial parameter reduction, all while preserving SNLI performance.

VI. Analysis

Our analysis of this work centers on understanding how attention heads are able to contribute to semantic differentiation in NLI and how these contributions vary across layers. The variance metric provides a pragmatic way to quantify the degree to which each head responds differently to entailment, contradiction, and neutral inputs. Heads with high variance exhibit clear distinctions across classes, while low-variance heads react similarly- regardless of the underlying semantic relationship between the premise and hypothesis.

The heatmaps in Figure 3 reveal a consistent trend of deeper layers showing stronger and more structured cross-sentence attention patterns. Whereas the early layers exhibit mostly uniform or low-magnitude attention distributions. This observation aligns with previous findings that deeper transformer layers encode more task-specific and semantically rich information, while early layers capture generic lexical or positional features. In terms of RQ1, this provides evidence that deeper layers are indeed primarily responsible for cross-sentence semantic attention in SNLI.

Figure 4 further reinforces this distinction between layers. The high-variance heads are concentrated in middle and deep layers, which suggests these heads play a key role in modeling the semantic interactions required for NLI. However, the lowest-variance heads cluster heavily in the first three layers, which indicates limited task-relevant (semantic) function. Therefore, these low-variance heads become natural candidates for pruning and correspond closely to the heads removed in Figure 7.

From a modeling perspective, this pattern supports the hypothesis that redundancy in BERT is *unevenly* distributed across the architecture. Early layers seem to contain many heads that offer similar, class-agnostic attention behaviors, whereas in later layers contain heads with more specialized and label-sensitive roles. As a result, parameter reduction strategies that target early-layer heads can significantly decrease model size while maintaining semantic capacity.

Our analysis validates variance-based pruning as an effective means of identifying redundant parameters (RQ2) and shows how different parts of the BERT architecture contribute to semantic reasoning on SNLI (RQ1). Additionally, with the performance results in Section V, these findings are also able to answer RQ3: a carefully chosen variance threshold enables non-trivial head pruning with negligible impact on classification accuracy.

VII. Limitations

While our findings provide insight into the redundancy and head importance in BERT, several limitations should be acknowledged.

First, attention variance captures differences in average attention patterns across classes but does not directly measure causal influence on model decisions. As noted in prior work, attention weights may not strictly correspond to model reasoning. Therefore, variance should be interpreted as a proxy for head behavior- rather than being a reliable indicator of evidence how much a head contributes to the model's output.

Secondly, our analysis is restricted to a single dataset (SNLI) and a single architecture (BERT-base-uncased). Semantic reasoning behavior may differ across models such as RoBERTa, DeBERTa, or instruction-tuned large language models. Additionally, this behavior might differ across datasets with longer or noisier inputs. Extending this analysis to such settings would test the solidity of our conclusions.

Thirdly, pruning is applied post-hoc and not integrated into training. The pruned model is not fine-tuned to recover any lost performance, which may understate the potential benefits of structured head pruning. Integrating pruning into training or fine-tuning could yield models that are both smaller and more accurate. Due to time constraints, we were unable to test this theory.

Lastly, our analysis relies on averaging attention scores over many examples. This results in smoothing over instance-specific differences. We hypothesize more granular techniques, such as probing, counterfactual interventions, or token-level

attribution methods, could complement variance-based insights by revealing finer-grained patterns of head behavior.

VIII. Conclusion

This work examined how attention heads in the BERT-base-uncased model contribute to natural language inference and evaluated whether low-variance heads can be pruned without degrading performance. By analyzing cross-sentence attention patterns across the SNLI dataset, we showed that shallow transformer layers exhibit limited semantic differentiation, while deeper layers capture more accurate patterns aligned with entailment, contradiction, and neutrality. These findings support the view that *meaningful* semantic processing in BERT appears primarily in later layers (RQ1).

Using a variance-based head importance metric, we demonstrated that a substantial subset of attention heads -specially those in early layers- contribute minimally to the downstream predictions. Pruning the lowest-variance heads removed roughly 22% of all heads with only minimal reductions in accuracy, macro precision, macro recall, and macro F1-score. This indicates that transformers contain considerable redundancy, meaning, variance can serve as a simple yet effective signal for head importance (RQ2). From a practical standpoint, our results show that *targeted head pruning can reduce model complexity without compromising SNLI performance* (RQ3).

Overall, our study highlights both where linguistic information is encoded in BERT and how model complexity can be reduced while preserving model performance. Future work should integrate pruning directly into training to examine its effect on model stability, generalization, and interpretability, as well as extend these analyses to larger or instruction-tuned transformer architectures and more diverse NLI corpora.

References

[1] G. Tucudean, M. Bucos, B. Dragulescu, and C. Căleanu, "Natural language processing with transformers: a review," *PeerJ Computer Science*, vol. 10, p. e2222, 2024.

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

[3] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5797–5808.

[4] ——, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[5] K. Sun and A. Marasović, "Effective attention sheds light on interpretability," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021, pp. 3572–3585.