Exploring the Impact of Multi-Headed vs. Single-Headed Attention in Transformer-based NLP Models

Yuvraj Kapoor
Northeastern University
Email: kapoor.y@northeastern.edu

I. Abstract

Transformers have become foundational in Natural Language Processing (NLP), largely due to their attention mechanisms, particularly multi-headed attention. While multi-headed attention enables models to represent various linguistic features and dependencies in parallel, it remains unclear how critical each individual head is to downstream performance, and what impact reducing the model to a single-head attention configuration would have. This study examines these questions through controlled experiments with two popular pre-trained transformer-based models, GPT-2 (an autoregressive decoder-only model) and DistilBERT (a compact encoder-only model). I measure the importance of individual attention heads by analyzing gradient-based metrics and then selectively disable heads identified as most or least important, as well as random heads, followed by further fine-tuning.

Results on the Stanford Sentiment Treebank (SST-2) binary classification task show that disabling a small number of the least important heads can often be done with minimal performance degradation, and in some cases (for GPT-2) even slightly improve final accuracy after re-fine-tuning. Conversely, disabling the most important heads results in consistent performance drops. Further, I examine models that use only a single attention head per layer. While single-headed variants reach moderate accuracy levels, they generally underperform their multi-headed counterparts and show limited improvement with extended training.

These findings highlight that while certain attention heads are crucial for maintaining high accuracy, others are redundant or even potentially detrimental. I also observe that multi-headed attention is particularly beneficial, and reducing to a single head may limit representational capacity. My results inform model interpretability and efficiency research, suggesting that targeted head pruning or adjustments can serve as an effective technique for reducing complexity.

II. Introduction

The transformer architecture [1] has revolutionized NLP, enabling models to handle complex language understanding tasks. Central to the transformer's success is its self-attention mechanism, which allows the model to focus on relevant parts of the input sequence. A distinguishing factor of the transformer is *multi-headed attention*, wherein multiple attention heads run in parallel, each potentially capturing distinct linguistic phenomena or dependencies.

Despite the widespread use of multi-headed attention, questions remain about the necessity of all these heads. Are some heads superfluous or even noisy? If certain heads are less important, can we prune them without harming performance? Conversely, what happens if we remove too many heads or the most critical ones? Additionally, how does a model behave if forced to rely on a single attention head per layer?

This work addresses these questions by conducting experiments on two representative models:

- **DistilBERT** [2]: A distilled, efficient encoder-only transformer closely related to BERT [3].
- **GPT-2** [4]: A decoder-only generative transformer widely adapted for various tasks via fine-tuning.

I focus on the Stanford Sentiment Treebank (SST-2) dataset [5], a binary sentiment classification task. By fine-tuning these models, computing head importance metrics, selectively disabling heads, and observing the effects on accuracy, I gain insights into the roles attention heads play. I also experiment with single-headed versions of these models to examine how performance deteriorates without multi-headed diversity.

A. Motivation and Hypotheses

Motivation: Investigating attention mechanisms in transformer models can provide deeper insights into model interpretability, efficiency, and representational capacity. By focusing on DistilBERT (encoder-only) and GPT-2 (decoder-only), I aim to understand whether findings about head importance and configuration generalize across different transformer architectures. I seek to determine if certain heads are essential, and if removing unimportant heads can simplify models without sacrificing accuracy.

Hypothesis 1: Disabling the most important attention heads will lead to a significant degradation in model performance compared to disabling the least important or randomly selected heads.

Hypothesis 2: Models with multi-headed attention will outperform single-headed attention models in terms of accuracy on the sentiment analysis task. However, single-headed attention models may exhibit higher interpretability, as their simpler attention structure is easier to analyze.

III. Background

The original Transformer architecture [1] uses an encoder-decoder structure composed of a stack of identical layers. Each layer includes:

- **Multi-Head Self-Attention Mechanism:** Instead of computing a single attention distribution, multiple heads run in parallel. Each head can focus on different parts of the input, capturing diverse syntactic and semantic dependencies.
- **Feed-Forward Neural Network (FFN):** After the attention step, a position-wise FFN is applied, increasing the representational power of the layer.
- **Residual Connections and Layer Normalization:** These ensure stable training and allow gradients to flow more easily, aiding in deeper architectures.
- **Positional Encodings:** Since transformers lack recurrence or convolutional structure, positional encodings provide information about the order of tokens in the input sequence.

While the original Transformer introduced a full encoder-decoder stack, variants such as GPT-2 use only the decoder stack (autoregressive modeling for language generation), and models like DistilBERT are derived from the BERT encoder stack and are thus encoder-only.

A. Encoder-only vs. Decoder-only Models

Encoder-only Models (e.g., DistilBERT): These models are designed for tasks requiring rich contextual understanding of entire input sequences (e.g., classification, sentence similarity). The encoder processes input tokens in parallel, integrating contextual information at multiple layers. DistilBERT [2] is a compressed version of BERT [3], retaining much of the representational power while being more efficient.

Decoder-only Models (e.g., GPT-2): Decoder-only architectures, like GPT-2 [4], predict the next token given previously seen context. They leverage masked self-attention to avoid attending to future tokens, making them suitable for generative tasks. However, they can also be adapted for classification tasks by appending classification heads and fine-tuning.

IV. Related Work

Past research has probed the necessity of multiple attention heads. Michel et al. [6] showed that pruning heads can be done with minimal performance loss in some cases. Voita et al. [7] linked certain heads to specific linguistic functions and found that many heads are redundant.

My work complements these efforts by experimenting on DistilBERT and GPT-2, assessing both top-ranked and bottom-ranked heads, random head removal, and examining the extreme case of single-headed attention. This comparative approach across different architectures adds to the understanding of head importance and model complexity.

V. Single vs. Multi-Headed Attention

Multi-headed attention is considered a key innovation of the Transformer architecture. By having multiple heads, models can:

- Capture a broader range of dependencies and linguistic phenomena.
- Distribute representational capacity across multiple learned projections, potentially improving performance on complex tasks.

Single-headed attention, by contrast, forces all representational power into a single distribution. While simpler, it may lack the flexibility and richness of multi-headed attention. Studying single-headed variants provides insights into the necessity of multiple heads and whether the same performance can be achieved through a single head with additional training.

VI. Dataset: Stanford Sentiment Treebank (SST-2)

The Stanford Sentiment Treebank (SST-2) [5] is a well-established benchmark for binary sentiment classification. It consists of movie review snippets labeled as positive or negative. We choose this dataset because:

- Sentiment classification is a straightforward, widely studied task, making results and comparisons easier to interpret.
- The dataset size and complexity are manageable, allowing for multiple rounds of fine-tuning and head manipulations without excessive computational costs.

Due to computational constraints, we use a subset:

- **Training Set:** 5,000 samples from the original training split.
- **Validation Set:** 500 samples from the original validation split.

VII. Experimental Setup

A. Models and Tokenization

DistilBERT: I use `distilbert-base-uncased`, an encoder-only model distilled from BERT. Its smaller size and efficiency make it a practical choice for repeated experiments.

Model Size: Approximately forty percent smaller than BERT-base, with around 66 million parameters compared to BERT-base's 110 million.

Layers: Consists of 6 transformer encoder layers, half the number in BERT-base.

Hidden Size: Maintains a hidden size of 768 dimensions, the same as BERT-base.

Attention Heads: Each encoder layer includes 12 attention heads, identical to BERT-base.

Feed-Forward Networks (FFN): Each layer contains an FFN with two linear transformations and a GELU (Gaussian Error Linear Unit) activation function.

GPT-2: I use `gpt2`, originally a decoder-only generative model. After adapting it with a classification head, we fine-tune it for sentiment classification.

Model Sizes: GPT-2 comes in various sizes; for practical purposes, the small model with 117 million parameters is often used.

Layers: Consists of 12 transformer decoder layers.

Hidden Size: Has a hidden size of 768 dimensions.

Attention Heads: Each decoder layer includes 12 attention heads.

Feed-Forward Networks (FFN): Similar to DistilBERT, each layer contains an FFN with two linear transformations and a GELU activation.

Each model is fine-tuned using the dataset described above.

B. Training Procedure

For all experiments:

- **Optimizer:** AdamW
- **Learning Rate:** 5×10^{-5}
- **Batch Size:** 16
- **Epochs:** 5 (for the initial fine-tuning and subsequent re-fine-tuning)

I first fine-tune each model on the SST-2 subset to establish a baseline accuracy.

C. Head Importance Computation

I compute head importance using gradients:

1) Compute the loss and backpropagate to obtain gradients.
2) Accumulate absolute gradient values from attention outputs.
3) Average these gradients to rank heads by importance.

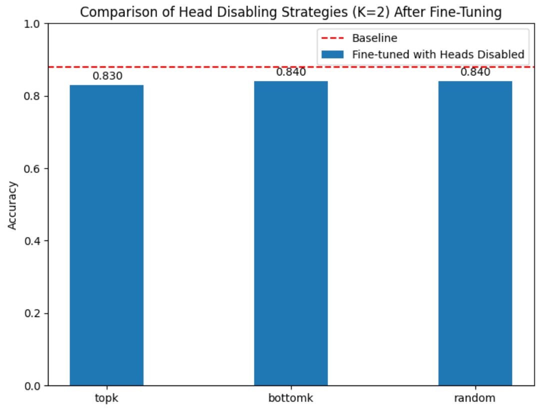

Fig. 1. DistilBERT: Comparison of Head Disabling Strategies (K=2) After Fine-Tuning. Baseline ≈0.88.

D. Selective Head Disabling

Once heads are ranked, I create head masks to disable specific heads:

- **Top-K:** The K most important heads are disabled.
- **Bottom-K:** The K least important heads are disabled.
- **Random-K:** K heads are chosen at random from each layer and disabled.

I evaluate model performance with these disabled heads immediately and then re-fine-tune the model with these heads disabled to see if it can recover or even improve performance.

E. Single-Headed Variants

I replace all multi-headed attention with a single-head attention mechanism in DistilBERT and GPT-2 to understand the drop in representational capacity. I fine-tune these single-headed models and measure their validation accuracy over multiple training epochs.

VIII. Results and Analysis

A. DistilBERT Results

Baseline: After 5 epochs of fine-tuning, DistilBERT achieves approximately 0.88 accuracy on the validation set.

1) K=2 Head Disabling

As shown in Fig. 1, after re-fine-tuning:

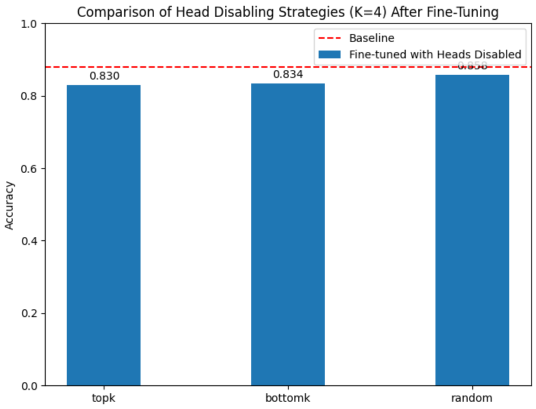

Fig. 2. DistilBERT: Comparison of Head Disabling Strategies (K=4) After Fine-Tuning. Baseline ≈0.88.

- **Top-2 disabled:** Accuracy ≈0.830 (a drop from baseline).
- **Bottom-2 disabled:** Accuracy ≈0.840 (closer to baseline).
- **Random-2 disabled:** Accuracy ≈0.840.

Removing top-k heads harms performance more than removing bottom-k or random heads, which the model can nearly recover from.

2) K=4 Head Disabling

For K=4 (Fig. 2):

- **Top-4 disabled:** Accuracy ≈0.830
- **Bottom-4 disabled:** Accuracy ≈0.834
- **Random-4 disabled:** Accuracy ≈0.830

Even with more heads disabled, bottom-k removal remains less detrimental than removing top-k heads. Random removal yields similar accuracy to top-k removal, suggesting that when disabling more heads, random selection can be nearly as harmful as removing the most important ones.

3) Single-Headed DistilBERT

As seen in Fig. 3, the single-headed DistilBERT only achieves about 0.53 accuracy and does not improve over multiple epochs. This demonstrates that multiple heads are crucial for DistilBERT to capture nuanced features.

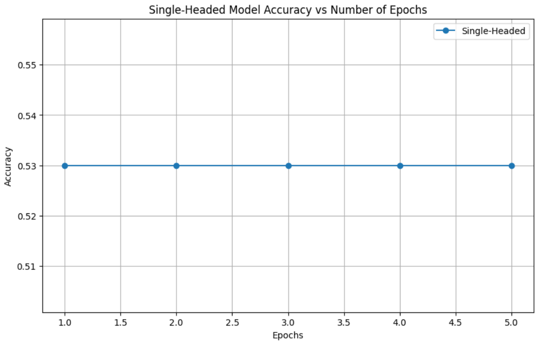

Fig. 3. DistilBERT: Single-Headed Model Accuracy vs. Number of Epochs.

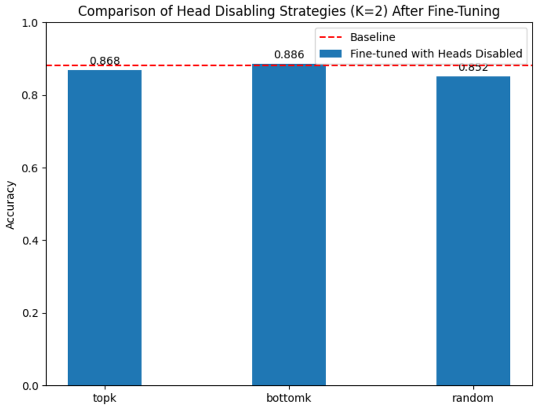

Fig. 4. GPT-2: Comparison of Head Disabling Strategies (K=2) After Fine-Tuning. Baseline ≈0.88.

B. GPT-2 Results

Baseline: After fine-tuning, GPT-2 achieves about 0.88 accuracy on the validation set.

1) K=2 Head Disabling

From Fig. 4:

- **Top-2 disabled:** ≈0.868
- **Bottom-2 disabled:** ≈0.886 (slightly better than baseline)
- **Random-2 disabled:** ≈0.852

Interestingly, disabling the least important heads (bottom-2) can slightly improve GPT-2's performance after re-fine-tuning, possibly removing noisy or unhelpful heads.

2) K=4 Head Disabling

For K=4 (Fig. 5):

- **Top-4 disabled:** ≈0.858
- **Bottom-4 disabled:** ≈0.824

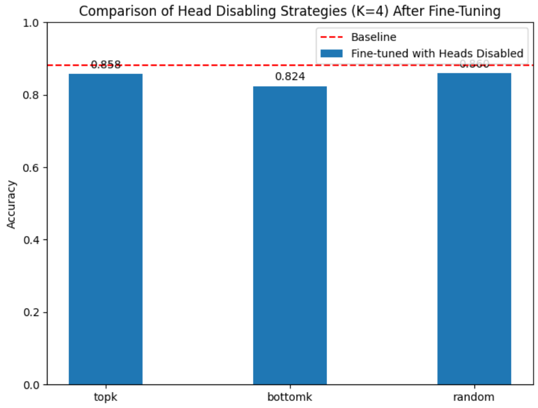

Fig. 5. GPT-2: Comparison of Head Disabling Strategies (K=4) After Fine-Tuning. Baseline ≈0.88.

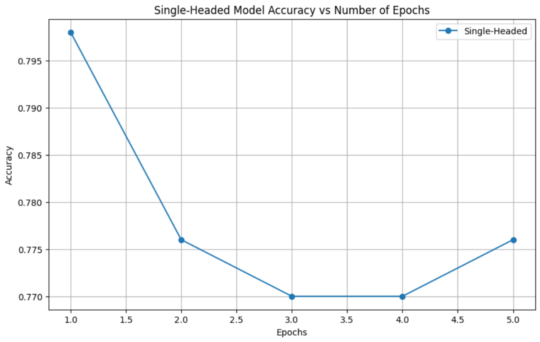

Fig. 6. GPT-2: Single-Headed Model Accuracy vs. Number of Epochs.

- **Random-4 disabled:** ≈0.850

Removing more heads at once makes the model more sensitive. With K=4, disabling bottom-k heads now results in lower accuracy than top-k removal. Random removal yields an intermediate accuracy (0.850). This suggests that while removing a few unimportant heads can help, removing too many might disrupt the model's balance.

3) *Single-Headed GPT-2*

Fig. 6 shows that single-headed GPT-2 starts around 0.798 accuracy but declines over successive training intervals, stabilizing around 0.77-0.78. It never reaches the multi-headed baseline, illustrating that multiple heads are beneficial for GPT-2 as well.

IX. Discussion

My results highlight several key points:

1) **Most Important Heads Matter:** Disabling top-ranked heads generally harms accuracy, indicating these heads capture crucial information.
2) **Less Important Heads Can Be Removed:** For DistilBERT with K=2 and GPT-2 with K=2, removing bottom-ranked heads has little negative effect; GPT-2 even shows a mild improvement.
3) **Impact of Larger K:** When disabling 4 heads, performance drops become more pronounced. For GPT-2, removing bottom-4 heads is no longer beneficial. This suggests that the number of heads removed and their selection strategy interact in complex ways.
4) **Single-Headed Models Underperform:** Both DistilBERT and GPT-2 single-headed variants underperform their multi-headed baselines, confirming that multiple heads offer representational richness.

X. Conclusion and Future Work

I investigated head importance in DistilBERT and GPT-2 using SST-2 sentiment classification. Our experiments confirm that not all heads are equally important and that selectively disabling less critical heads often does minimal harm. In some cases, it can even lead to slight improvements. However, removing too many heads or the most critical ones degrades performance. Single-headed attention variants are consistently weaker, reaffirming the value of multi-headed designs.

Future work could:

- Expand to other tasks and datasets, exploring if head importance patterns generalize.
- Investigate iterative or dynamic head pruning to find optimal subsets of heads.
- Link head importance more explicitly to linguistic phenomena, improving interpretability.

This study contributes to a better understanding of attention heads, guiding future model pruning and interpretability research.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” in *NeurIPS*, 2017.

[2] V. Sanh, L. Debut, J. Chaumond, T. Wolf, “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” in *NeurIPS EMC2 Workshop*, 2019.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL-HLT*, 2019.

[4] A. Radford, J. Wu, R. Child, et al., “Language Models are Unsupervised Multitask Learners,” OpenAI Technical Report, 2019.

[5] R. Socher, A. Perelygin, J. Wu, et al., “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,” in *EMNLP*, 2013.

[6] P. Michel, O. Levy, G. Neubig, “Are Sixteen Heads Really Better than One?” in *NeurIPS*, 2019.

[7] E. Voita, D. Talbot, R. Mohr, I. Titov, “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned,” in *ACL*, 2019.