Reproducing "Are Sixteen Heads Really Better than One?"

Stephen Zhao *
stephen.zhao@mail.utoronto.ca

Sherry Yuan *
shanli.yuan@mail.utoronto.ca

Abstract

Multi-headed attention (MHA) is crucial to several modern NLP models. In "Are Sixteen Heads Really Better than One?", Michel et al., 2019 [3] aim to improve our understanding of when and where MHA is important through a series of experiments involving pruning attention heads. In this paper, we reproduce the authors' experiments. Our results are broadly supportive of their conclusions; many attention heads can be ablated without a noticeable impact on performance, the encoder-decoder attention mechanism benefits the most from MHA, and important heads are determined in the early stages of training.

1 Introduction

Multi-headed attention (MHA) is a key component of transformer-based and BERT-based models which have shown state of the art performance on a variety of NLP tasks. MHA has been shown to improve performance and help with subject-verb agreement [8]. In "Are Sixteen Heads Really Better than One?" (hereafter, "the original paper"), Michel et al., 2019 [3] (hereafter, "the authors") aim to build upon our understanding of how MHA contributes to performance by conducting a series of experiments involving pruning attention heads, to shed light on where MHA is most important and how much performance is impacted when heads are removed. Surprisingly, the authors find that most attention heads can be individually removed without having a significant impact on performance, and some layers can even be reduced to a single head. The authors found MHA is most important in the encoder-decoder attention layers in the transformer model, as that was most sensitive to pruning. Also, the authors demonstrate the interaction between multi-headedness and training dynamics by showing that the distinction between important and unimportant heads increases as training progresses. In this work, we reproduce the ablation and pruning results, and provide extended results and discussion on different datasets and models. In general, our work supports the conclusions of the original paper.

2 Background: Attention, Multi-headed Attention, and Masking

In this section, we recap the notation used by the authors and also provide further background on attention.

2.1 Single-headed Attention

Attention was introduced as a mechanism to improve performance by allowing models to place more emphasis on certain relevant parts of the input sequence. At a high level, attention takes three vectors - a query, key, and value - and assigns a weight to each value depending on the compatibility of the query with the corresponding key. In this case, the compatibility is determined by a dot product, and the keys and values are functions of the inputs. That is, given a sequence of n d-dimensional

*Equal Contribution

Preprint. Under review.

input vectors $\mathbf{x} = x_1, x_2..., x_n \in R^d$, and a query vector $q \in \mathbb{R}$, the attention layer parameterized by $W_k, W_q, W_v, W_o \in \mathbb{R}^{d \times d}$ computes the weighted sum:

$$Att_{W_k,W_q,W_v,W_o}(\mathbf{x}, q) = W_o \sum_{i=1}^{n} \alpha_i W_v x_i$$

$$\text{where } \alpha_i = \text{softmax}\left(\frac{q^T W_q^T W_k x_i}{\sqrt{d}}\right)$$

and $\sqrt{d}$ is a normalizing constant. In self-attention, each input x_i is also used as the query q.

2.2 Multi-headed Attention (MHA)

Multi-headed attention applies single-headed attention N_h times independently, allowing the model to more finely divide its attention; at each position, separate heads can focus on different input locations. That is, we apply N_h independently parameterized attention layers in parallel:

$$MHAtt(\mathbf{x}, q) = \sum_{h=1}^{N_h} Att_{W_k^h,W_q^h,W_v^h,W_o^h}(\mathbf{x}, q)$$

where $W_k^h, W_q^h, W_v^h \in \mathbb{R}^{d_h \times d}$ and $W_o^h \in \mathbb{R}^{d \times d_h}$. When $d_h = d$, MHA is strictly more expressive than single-headed attention. However, d_h is typically set to $\frac{d}{N_h}$ to keep the number of parameters constant; this results in each head projecting onto a lower dimensional subspace.

The transformer model then applies a non-linear feed-forward network over MHA's output to allow for interaction between different attention heads (Vaswani et al., 2017 [9]).

2.3 Masking Attention Heads

To identify the contributions of each attention head, the authors experiment by ablating heads with mask variables $\xi_h \in \{0, 1\}$. The MHAtt formula with the mask applied is as follows:

$$MHAtt(x, q) = \sum_{h=1}^{N_h} \xi_h Att_{W_k^h,W_q^h,W_v^h,W_o^h}(x, q)$$

where head h is masked when $\xi_h = 0$.

3 Are All Attention Heads Important?

In this section, we reproduce the experiments run by the authors that test the change in performance when one or all but one heads are removed from trained architectures at test time.

3.1 Experimental Setup

The models used throughout the original paper and our reproduction are BERT, WMT, and IWSLT.

BERT is a single transformer architecture consisting of 12 layers with 12 attention heads per layer, using self-attention in each layer. We follow the original authors in using the pre-trained model of Devlin et al., 2018 [2], and we evaluate performance on MNLI (sentence understanding [11]), CoLA (grammatical acceptability [10]), and SST-2 (sentiment analysis on movie reviews [7]). We test for statistical significance using a binomial test comparing the accuracy after ablation with the base accuracy, over the number of data points.

WMT is a "large" transformer architecture consisting of 6 layers with 16 heads per layer (Vaswani et al., 2017 [9]). We use the pretrained model of Ott et al., 2018 [6], trained on the WMT2014 English to French corpus, and report BLEU scores on the newstest2014 test set, taken from the same source.[2] Statistical significance is tested via paired bootstrap resampling using compare-mt[3] (Neubig et al.,

[2]Source, same as the original paper: `https://github.com/pytorch/fairseq/tree/master/examples/translation`. We used the newstest2014 dataset instead of newstest2013 as that dataset came with the model from the same source.

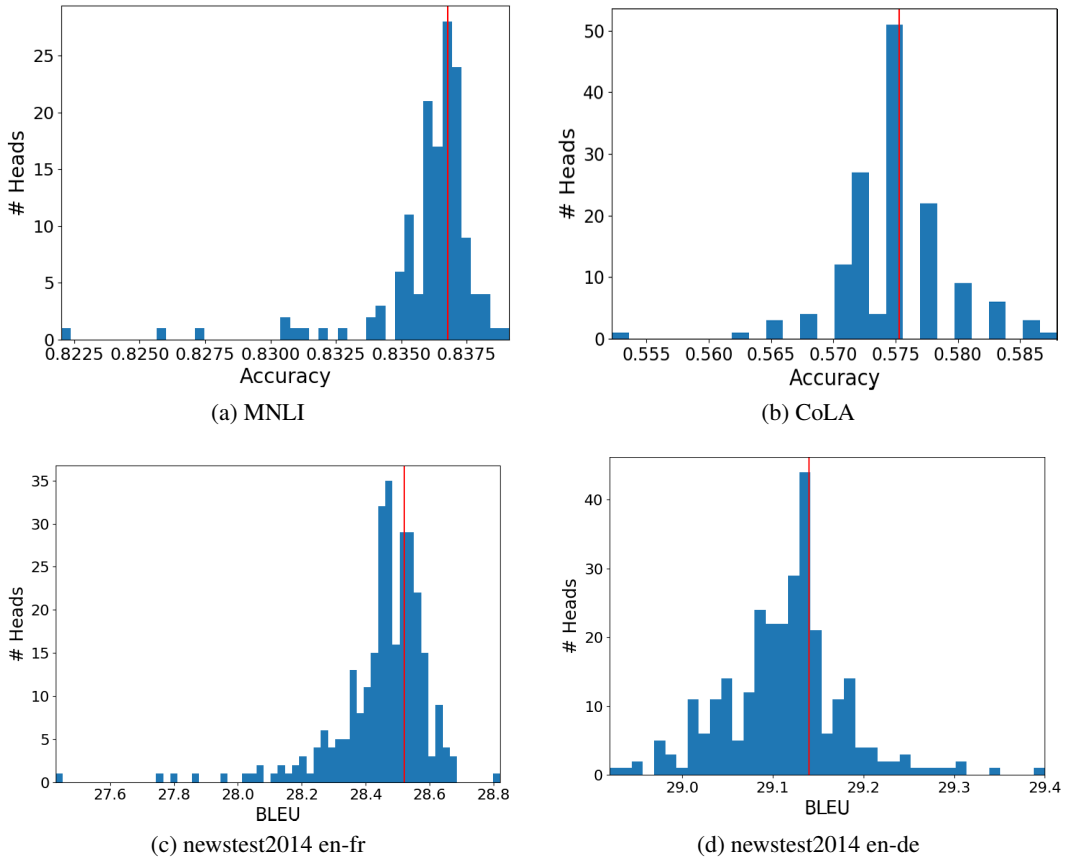
[3]`https://github.com/neulab/compare-mt`

2019 [5]) with 1000 resamples, as done in the original paper. The WMT model has 3 multi-headed attention mechanisms - encoder self-attention (Enc-Enc), encoder-decoder attention (Enc-Dec), and decoder self-attention (Dec-Dec).

IWSLT is a smaller version of the WMT model, with 6 layers and 8 heads per layer. The model is trained for German-to-English translation on the IWSLT 2014 dataset [1].[4]

3.2 Ablating One Head

The authors claim that, for most attention heads, removing the head individually does not cause a large change in performance. Our results in the graphs and tables below support this claim. In Figure 1, we provide histograms showing the distribution of scores after ablating individual heads. In general, the change in accuracy as individual heads are ablated tends to be close to zero for the majority of heads. Figure 1 (a) and (c) correspond to Figure 1a and 1b in the original paper. We also ran experiments on additional datasets, verifying that the results hold across different tasks. For BERT, we found similar results on CoLA and SST-2; we show the histogram for CoLA in Figure 1 (b). For WMT, we also show results on English-German translation in Figure 1 (d).



(a) MNLI

(b) CoLA

(c) newstest2014 en-fr

(d) newstest2014 en-de

Figure 1: (a) and (b) show head ablation results for BERT; (c) and (d) show results for WMT. Red lines represent base accuracy/BLEU score.

The following tables provide a more detailed breakdown. Table 1 shows the change in performance for ablating each individual head of BERT on MNLI. Removal results in a statistically significant change in performance for only 2% of the heads (3 out of 144), and for 41% of the heads, accuracy improves when the head is removed.

[4]Model and data source: `https://github.com/pytorch/fairseq/tree/master/examples/translation`

Results for WMT are shown in Tables 2-4. Table 2 shows results for the encoder self-attention mechanism, and corresponds to Table 1 in the original paper. For further detail, we also include results for the encoder-decoder attention (Table 3) and decoder self-attention mechanisms (Table 4). We find similar results for WMT; only about 2% of heads are statistically significant, and for around one-third of heads, removal results in a positive change in BLEU score. Our results are generally in line with the numbers given in the original paper, supporting the authors' claim that most heads are redundant given the rest of the model at test time.

Layer \ Head	1	2	3	4	5	6	7	8	9	10	11	12
1	-0.04%	0.06%	-0.07%	-0.16%	-0.07%	0.23%	0.02%	0.04%	-0.27%	-0.02%	-0.09%	0.09%
2	-0.01%	0.01%	0.06%	0.01%	-0.07%	-0.03%	-0.06%	-0.18%	-0.29%	-0.06%	0.05%	-0.01%
3	-0.19%	-0.08%	-0.12%	-0.17%	-0.01%	0.14%	-0.03%	-0.04%	-0.03%	-0.16%	-0.12%	0.00%
4	-0.62%	-0.16%	0.15%	-0.16%	0.03%	-0.03%	-0.05%	-0.04%	-0.08%	-0.02%	-0.07%	0.05%
5	-0.06%	-0.26%	0.08%	-0.62%	0.04%	0.05%	-0.12%	-0.03%	-0.06%	-0.17%	-0.01%	0.03%
6	-0.13%	0.01%	0.11%	0.05%	0.16%	-0.16%	-0.06%	0.05%	-0.56%	0.05%	-0.07%	0.01%
7	**-1.48%**	0.00%	0.07%	0.04%	0.02%	-0.11%	-0.14%	-0.15%	-0.17%	-0.02%	0.02%	-0.05%
8	-0.46%	-0.08%	0.01%	0.06%	0.02%	-0.06%	-0.17%	-0.02%	-0.16%	-0.09%	0.03%	0.00%
9	0.05%	**-0.96%**	-0.09%	-0.02%	-0.40%	0.04%	0.01%	-0.27%	-0.01%	0.01%	0.01%	0.08%
10	0.10%	0.06%	-0.03%	-0.04%	0.00%	0.01%	0.03%	-0.08%	**-1.08%**	-0.06%	-0.03%	-0.29%
11	0.01%	0.02%	-0.03%	-0.07%	-0.05%	-0.04%	-0.04%	0.07%	0.15%	0.10%	-0.15%	-0.59%
12	0.03%	0.13%	0.05%	-0.06%	-0.01%	0.03%	0.01%	-0.01%	-0.13%	0.01%	0.06%	0.05%

Table 1: Difference in accuracy of BERT when ablating each head individually. Underlined numbers indicate that the change is statistically significant with $p < 0.01$. The base accuracy is 83.68%.

Enc-Enc: Layer \ Head	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-0.14	-0.13	0.08	0.00	-0.10	**-0.65**	0.15	**-1.09**	0.04	-0.04	-0.04	0.11	-0.05	-0.23	-0.05	-0.13
2	-0.10	-0.08	0.03	-0.14	0.16	-0.05	-0.05	0.00	0.05	0.00	-0.02	0.04	-0.16	0.03	0.04	0.03
3	-0.05	-0.08	-0.06	-0.08	-0.05	-0.09	0.04	-0.03	0.00	-0.06	-0.02	0.02	-0.26	0.02	-0.05	-0.03
4	-0.03	-0.12	-0.04	-0.01	-0.04	0.02	-0.12	0.09	-0.09	-0.07	-0.06	-0.07	-0.05	0.03	0.06	-0.04
5	-0.19	-0.22	-0.10	-0.06	0.04	-0.09	0.07	-0.09	-0.05	0.05	-0.12	-0.07	0.01	-0.04	0.01	0.02
6	-0.06	0.00	-0.24	0.06	-0.37	-0.01	0.05	-0.04	-0.17	-0.14	0.07	-0.02	-0.01	0.05	-0.10	0.10

Table 2: Difference in BLEU score for each head of the encoder's self-attention mechanism in the WMT model. For tables 2-4, underlined numbers indicate that the change is statistically significant with $p < 0.01$. The base BLEU is 28.52.

Enc-Dec: Layer \ Head	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-0.05	-0.02	0.05	**-0.56**	-0.07	0.06	-0.04	-0.01	-0.05	-0.08	-0.06	-0.15	-0.11	-0.11	0.06	-0.08
2	-0.08	0.11	**-0.71**	-0.08	0.01	-0.09	0.00	-0.07	0.07	-0.01	-0.18	0.05	-0.08	-0.11	-0.05	0.03
3	0.01	-0.48	0.05	-0.25	-0.04	0.14	-0.31	0.04	0.01	-0.11	-0.31	-0.10	-0.02	-0.03	-0.06	0.07
4	-0.15	-0.01	-0.20	-0.09	0.02	-0.15	0.13	0.04	-0.17	-0.09	-0.19	-0.01	-0.14	-0.50	-0.25	-0.22
5	0.00	-0.33	0.07	0.00	-0.38	0.30	-0.06	-0.25	-0.28	-0.08	0.16	-0.03	0.00	-0.28	-0.15	-0.40
6	-0.44	-0.22	-0.19	-0.04	**-0.76**	-0.29	-0.21	-0.28	-0.34	-0.32	-0.21	0.00	-0.45	-0.07	-0.38	-0.17

Table 3: Difference in BLEU score for each head of the encoder-decoder attention mechanism.

Dec-Dec: Layer \ Head	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.01	-0.09	-0.06	0.01	-0.01	0.00	-0.05	0.02	-0.04	-0.02	-0.01	0.00	-0.12	0.04	0.00	0.02
2	0.02	-0.07	0.05	0.00	-0.01	-0.05	-0.06	0.12	-0.04	-0.05	0.01	0.12	-0.12	0.06	0.03	-0.06
3	-0.02	-0.08	-0.03	0.07	-0.04	-0.04	0.05	-0.04	-0.07	0.13	-0.06	-0.11	-0.01	0.04	-0.03	0.06
4	0.07	0.01	-0.15	0.05	0.07	-0.04	-0.21	0.13	0.08	-0.07	-0.16	-0.10	-0.09	-0.14	-0.15	-0.05
5	-0.18	-0.28	-0.01	0.11	-0.15	0.01	-0.02	0.02	0.04	-0.01	0.00	-0.14	0.10	-0.16	0.06	-0.25
6	0.01	0.04	0.10	-0.01	-0.05	0.01	-0.12	-0.03	0.05	0.03	-0.20	-0.04	0.10	-0.06	-0.04	0.03

Table 4: Difference in BLEU score for each head of the decoder's self-attention mechanism.

3.3 Ablating All Heads but One

Given the minimal impact of ablating most heads individually, the authors then asked what would happen if we ablated all but one head in each layer. In Tables 5 and 6, we reproduce the authors' work on ablating all but one head, reporting results here for WMT and BERT on newstest2014 and MNLI respectively. Our results are generally in line with the original authors' results; the change in accuracy is surprisingly small for most heads, and layers that were most impacted are the same as in the original paper. There are some minor differences, including one statistically significant drop in accuracy for BERT, and fewer layers that had improved performance after ablating to only one head.

Layer	Enc-Enc	Enc-Dec	Dec-Dec
1	**-1.19**	-0.16	-0.03
2	-0.53	0.04	-0.03
3	-0.31	-0.14	0.25
4	-0.50	-0.50	0.09
5	-0.22	**-1.56**	0.09
6	-0.51	**-9.96**	-0.04

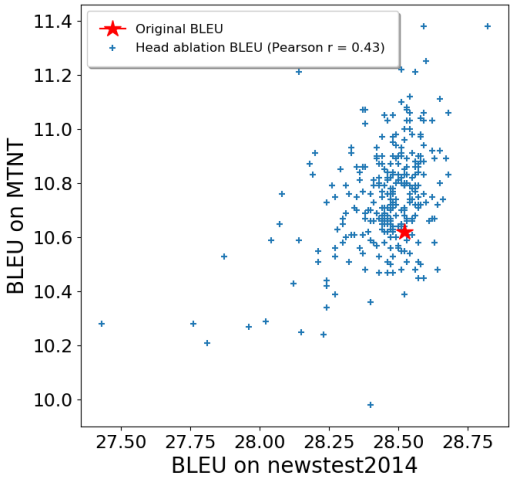
(a) Table 5: Best change in BLEU by layer when only one head is kept in the WMT model. Underlined numbers indicate that the change is statistically significant with p < 0:01. The base BLEU is 28.52.

Layer		Layer	
1	-0.31%	7	-0.31%
2	-0.28%	8	-0.61%
3	-0.46%	9	**-1.13%**
4	-0.69%	10	-0.59%
5	-0.61%	11	-0.41%
6	-0.43%	12	-0.03%

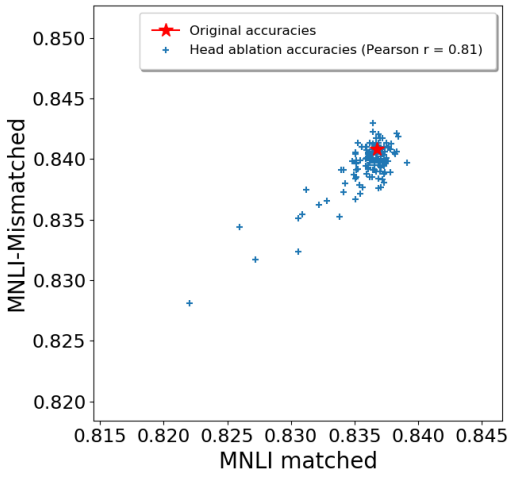
(b) Table 6: Best difference in accuracy for BERT when ablating all but one head in each layer. Underlined numbers indicate that the change is statistically significant with p < 0.01. The base accuracy is 83.68%

3.4 Are Important Heads the Same Across Datasets?

The authors then investigated whether the heads with the highest effect on performance were the same across different domains. In Figure 2 we replicate the authors' work demonstrating that, even on out-of-domain test sets (MTNT [4] for WMT and the MNLI-mismatched validation set for BERT), there is a correlation between the effect of removing any particular head on the performance on each dataset. That is, using the head ablation scores from section 3.2, we plot, for each head, the score on both datasets when the head is ablated.



(a) BLEU on newstest2014 and MTNT when individual heads are removed from WMT. The ranges are not the same on the X and Y axis; there is much more variation on MTNT.



(b) Accuracies on MNLI-matched and -mismatched when individual heads are removed from BERT. Here the scores remain in the same approximate range of values.

Figure 2: Cross-task analysis of effect of pruning on accuracy

We also compared BERT on MNLI versus CoLA and SST2, and found no significant correlation ($r \leq 0.08$), which confirms our expectations, given the differences in the tasks.

4 Iterative Pruning of Attention Heads

Since section 3 focused on pruning heads within a single layer, the authors then turned to sequentially pruning multiple heads across different layers. To determine the order of pruning, since combinatorial search would be prohibitively time-consuming, the authors develop a proxy score for head importance.

4.1 Head Importance Score for Pruning

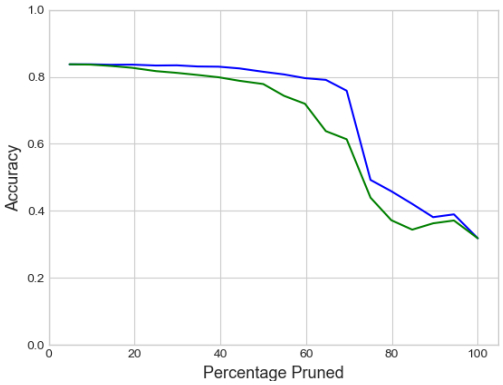
The authors use the expected sensitivity of the model to the mask variables ξ_h defined in 2.3 as a proxy score: $I_h = \mathbb{E}_{x \sim X} \left| \frac{\partial \mathcal{L}(x)}{\partial \xi_h} \right|$, where X is the data distribution. The absolute value is taken to identify heads that have the greatest effect on the model's predictions, in terms of the change in loss if the head would be removed, preventing highly negative and positive values from canceling each other out. Heads are pruned in increasing order from the smallest to the largest value of I_h.

Applying the chain rule, the authors also develop the alternate formulation:

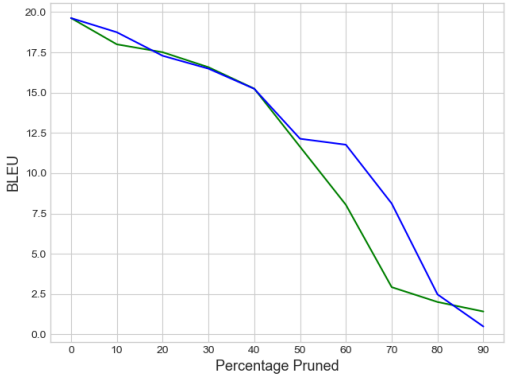
$$I_h = \mathbb{E}_{x \sim X} \left| Att_h(x)^T \frac{\partial \mathcal{L}(x)}{\partial Att_h(x)} \right|$$

In practice, the expectation is computed over a subset of the training data, and since I_h can be estimated from a forward and backward pass, the incremental time cost is no more than the time spent on training.

4.2 Effect of Pruning on BLEU/Accuracy



(a) BERT on MNLI dataset. Heads are pruned 5% at a time sequentially based on the authors' importance score (blue line) and a static evaluation of accuracy change for ablating each individual head as calculated in section 3.2 (green line).



(b) BLEU score on IWSLT14 German-to-English translation when heads are pruned according to I_h (blue line) and BLEU difference (green line). Here, we prune up to 90% of heads in increments of 10%.

Figure 3: Evolution of BLUE/accuracy by number of heads pruned

In Figure 3 (a) we reproduce the original paper's results using BERT on MNLI, incrementally removing 5% of heads based on increasing I_h. Our graph is very similar to Figure 3 (b) in the original paper, and supports the claim that 40% of the heads for BERT can be pruned without a noticeable drop in model performance.

We also plot results when pruning based on the score difference calculated in section 3.2 for each individual head; our results support the use of the authors' importance metric as a better method of pruning.

In Figure 3 (b), we present results on IWSLT instead of WMT in the original paper. Similar to the authors, we found the language translation task to be more sensitive to pruning (compared to BERT on MNLI). Our results appear to be more sensitive to pruning compared to the authors' results in section 6 of their paper; we have a larger, though still small, decrease in performance when pruning a small percentage of total heads. If we had more time, we would investigate how this compares to WMT, and whether the smaller number of heads in IWSLT means pruning has a larger impact on it. We note that comparing Figures 3 and 5 in the original paper suggests that there is no such difference, and both WMT and IWSLT appear to be similarly affected by sequential pruning.

5 When Are More Heads Important? The Case of Machine Translation

To better understand which mechanisms benefit most from MHA, the authors explore pruning heads from only one type of attention layer while keeping all other heads. We repeat the authors' experiment, using IWSLT instead of WMT. Figure 4 shows our results when we prune up to 90% of heads within each type of attention. The graph is similar to the one in the original paper using the WMT model. We find that the BLEU score drops most rapidly when heads from the Encoder-Decoder layer are pruned, supporting the claim in the original paper that the Encoder-Decoder layer benefits the most from multi-headed attention.

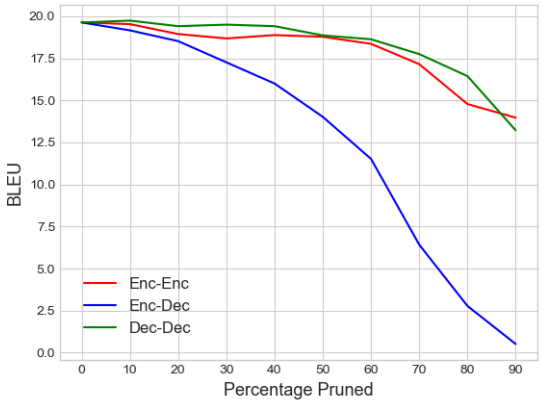


Figure 4: BLEU when incrementally pruning heads from each attention type in the IWSLT model. Here we prune up to 90% of heads in increments of 10%.

6 Dynamics of Head Importance During Training

Since the previous sections all focused on pruning heads in trained models, the authors then explore the effect of pruning during training, repeating the pruning experiment of section 4.2 on IWSLT at every epoch. In Figure 5 below, we repeat the same experiment, showing results for selected epochs. Similar to the original paper, in early epochs, performance decreases roughly proportional to the rate of pruning, suggesting that most heads are equally important during early training. As training progresses, the lines towards the top gather closer together, suggesting heads can be pruned with less impact to the model. This supports the authors' claim that important heads are determined early, but not immediately during the training process. However, our results appear to be less pronounced than those reported in the original paper; we are able to prune about 20-30% rather than 40% of heads while remaining within 85-90% of the original BLEU score.

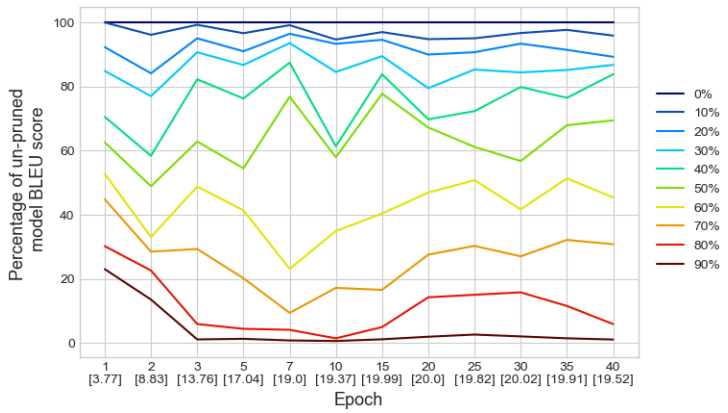


Figure 5: Relationship between the percentage of heads pruned and relative score decrease during the training of the IWSLT model. Epochs are reported on a non-constant scale. The y-axis plots the performance of the pruned model as a percentage of the un-pruned model. The BLEU score of the original, un-pruned model is indicated in brackets. The legend shows the color scheme for the percentage of heads pruned from 0 - 90%.

7 Reproduction Process and Discussion

Overall, the authors did a good job in making their work reproducible, as shown by the similar conclusions we reached in this paper. The authors made their code publicly accessible[5] and provided instructions to run heads ablation and pruning. The authors also provided links to repositories for tools, datasets, and pretrained models.

That said, we encountered several difficulties during the reproduction process, perhaps due to our relative inexperience compared to the authors. For some experiments, documentation was explicit and easy to follow, whereas for other experiments with less explicit documentation, we had to sift through the source code to find the right files to use, what arguments to call, and how to call them. We spent a fair bit of time on issues such as file structure and argument use that might be simple for more experienced researchers. For the translation tasks using WMT and IWSLT, while information on data preprocessing was provided, it was at a relatively high level and we had to infer the details. This may be a reason why our conclusions differ slightly from the authors', and why our base BLEU scores in certain experiments are different; we may have incorrectly inferred or missed certain details on preprocessing steps and configurations for data, models, and experiments.

For several tasks, we also had to modify or extend existing code. We believe some issues to be related to differences in computer architecture or environment setup, requiring small fixes to avoid errors. In other cases, we had difficulty finding the code with our desired functionality. Fortunately, we were able to relatively easily work off of existing code and add our custom code. We include our modified code that we used to generate the results and graphs in this paper.[6]

8 Conclusion

In this paper, we reproduced the main results of the original paper. While we had some differences in experimental setup, our conclusions are broadly supportive of those presented in the original paper; most heads can be individually ablated without a significant impact on model performance, many layers can even be ablated to a single head, around 40% of the total heads in BERT can be pruned with no noticeable impact on performance on MNLI, encoder-decoder attention heads benefit the most from multi-head attention, and important heads are determined in the early stages of training. While we believe the authors did a good job making the experiments repeatable, additional details in the documentation would make the code more easily reproducible for a wider audience.

[5] `https://github.com/pmichel31415/are-16-heads-really-better-than-1`

[6] We provide the code we used here: `https://github.com/Silent-Zebra/reproduce/tree/master/are-16-heads-really-better-than-1`

References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. Report on the 11th iwslt evaluation campaign, iwslt 2014.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3] P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*, 2019.

[4] P. Michel and G. Neubig. Mtnt: A testbed for machine translation of noisy text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[5] G. Neubig, Z.-Y. Dou, J. Hu, P. Michel, D. Pruthi, and X. Wang. compare-mt: A tool for holistic comparison of language generation systems. *Proceedings of the 2019 Conference of the North*, 2019.

[6] M. Ott, S. Edunov, D. Grangier, and M. Auli. Scaling neural machine translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

[7] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[8] G. Tang, M. Müller, A. Rios, and R. Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[10] A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, Mar 2019.

[11] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.