Selective Attention: Reducing Redundancy in Multi-Head Mechanisms

Ssu-Yung Yeh, Wei Chang, Wei-Chih Chung, Yen-Yun Kuo

1 Introduction

Transformer architecture (Vaswani et al., 2017) has come to dominate various NLP tasks due to its ability to learn long-range dependencies among input tokens. Specifically, attention mechanisms within the Transformer have been widely researched for their potential to interpret what the model has learned. Building on this advancement, large language models like BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and Llama (Dubey et al., 2024) used multiple layers of self-attention to learn syntax and semantics.

A key innovation in these large language models is multi-head attention. Unlike single-head attention, which relies on merely one set of learned parameters, multi-head attention utilizes multiple parallel attention heads to learn distinct patterns from the input, such as long-term dependencies or local relationships. However, this performance comes at a cost: a massive increase in parameters. Each head requires its query, key, and value matrices, as well as an additional layer for head integration. This significantly raises computational costs and memory requirements.

Therefore, our project aims to investigate the role and the meaning of multi-head attention. While current research has focused on several perspectives of multi-head attention, including I/O accelerator (Dao et al., 2022), and model size reduction (Ainslie et al., 2023), our project focuses specifically on model size reduction. Here are two questions we want to explore: Is every parameter in multi-head attention useful? Are there any efficient ways to achieve similar or even better results while reducing parameters in the attention layer?

Previous works have shown that removing certain attention heads does not influence model performance. The success of Group Query Attention (GQA) (Ainslie et al., 2023) further verifies this concept by demonstrating that decreasing the number of heads could maintain considerable performance. Besides, Wu et al. (2024) utilized the concept of information flow to figure out the contribution of each head to the final output, then selectively retrieve the heads that have a larger impact. Our project aims to integrate the concepts of group query attention and retrieval head to reduce the model's complexity without significantly compromising its accuracy.

2 Related Work

2.1 Attention Head

Attention heads are fundamental components of Transformer models, which play a crucial role in capturing dependencies between input tokens. Multi-head attention in the original Transformer architecture (Vaswani et al., 2017) enabled models to focus on different parts of a sequence at the same time. This structure enhances the model's ability to capture various semantic relationships in parallel. Each attention head learns distinct patterns or relationships from the input data independently and then combines them to produce a more comprehensive output.

A study explored the role and necessity of multiple-head attention (Michel et al., 2019). In this study, they discuss the trade-off of multi-head attention, which is the increment of parameter size. Compared to a single-head attention model, multi-head attention has additional learnable parameters but also increases computational cost and memory usage. These findings highlight the balance between utility and computational cost and suggest that optimizing the number of attention heads and parameters could be key to achieving more efficient Transformer models.

While multi-head attention has been widely regarded as essential for parallel information extraction, recent studies have challenged this idea (Liu et al., 2021). Their findings indicated that the key advantage of multi-head attention is not its ability to attend to multiple positions simultaneously but its contribution to training stability. The authors show that using multiple layers of single-head attention has better performance but needs a lot more layers compared to a multi-head configuration. This emphasizes that the main advantage of multi-head attention is its shallow architecture and higher training stability.

Another approach to improve attention mechanisms is Grouped-Query Attention (GQA) (Ainslie et al., 2023). GQA is a technique that clusters heads learning similar patterns and shares parameters across the group. For each group, they will share the same key and value to reduce the number of parameters. This method will decrease the overall memory cost and help to increase the decoding speed. However, GQA also has the risk of quality degradation since fewer key-value representations are used. Therefore, an optimized GQA method is introduced called Asymmetric Grouped-Query Attention (AsymGQA) (Chen et al., 2024). Unlike naive neighbor grouping, AsymGQA adapts the grouping process based on the similarity of attention heads' outputs. This allows the attention heads with closer relationships to share the same key-value pairs for better performance. This method ensures that grouped attention heads focus on related input features and leads to better performance.

2.2 Dataset and task

Multi-NLI (Multi-Genre Natural Language Inference) is the dataset that is used as a benchmark for evaluating natural language inference (NLI) models. Multi-NLI includes 392,000 training samples, which consist of both matched and unmatched tests. Multi-NLI is a key resource for training models in entailment recognition, contradiction detection and semantic understanding problems.

Recent studies have explored how attention mechanisms influence models' performance on the Multi-NLI dataset. Hao et al. (2021) have introduced an approach called Self-Attention Attribution (ATTATTR) to analyze the internal information flow in BERT. They indicated that attention scores are insufficient to explain the model's behavior since the attention connections with the highest attribution scores affect the model's final predictions significantly. In another study, the author used machine translation to re-translate the MNLI dataset into 14 other languages to solve the problems in the original XNLI dataset and create a more accurate cross-lingual evaluation corpus (Upadhyay and Upadhya, 2023). Their goal is to improve cross-lingual understanding and natural language inference (NLI). The experiments showed that the re-translated datasets increased 3% of cross-lingual accuracy compared with the XNLI 2.0 dataset. This shows that the refined dataset enhances the performance of NLI models across multiple low-resource and high-resource languages.

"Needle in a haystack" in NLP means identifying small pieces and specific information from a huge and unstructured dataset. Some common problems in NLP are Information Retrieval, Question Answering and Keyword Extraction. This required the system to extract specific information from huge text corpora.

One of the solutions is incidental bilingualism (Briakou et al., 2023). Their research shows that large multilingual language models like PaLM, can perform zero-shot and few-shot translations without specific training. They found that this ability comes from the model's exposure to bilingual text during training even without a clearly defined purpose. Another study introduced Retrieval Heads (Wu et al., 2024). In this study, they revealed that only less than 5% of attention heads are retrieval heads, which play a crucial role in recalling information from large contexts. These heads have been found in base models like LLaMA-2, and they continue to function effectively even after fine-tuning. The paper also showed that retrieval heads are critical in maintaining accuracy in 'Needle in a Haystack' tasks. Disabling these heads significantly decreases model performance and increases the hallucinations. In contrast, masking non-retrieval heads has only little impact on the model's ability to extract specific information.

These studies highlight how incidental bilingualism and attention-based mechanisms can be used to solve "Needle in a Haystack" problems. These approaches offer Interdependent solutions to the challenge of finding precise details in huge and unstructured corpora.

3 Experiment

We conducted our experiments using the MNLI-m and MNLI-mm datasets, with Llama 3.2 1B Instruct serving as the baseline model. Our experimental design is structured around three core principles:

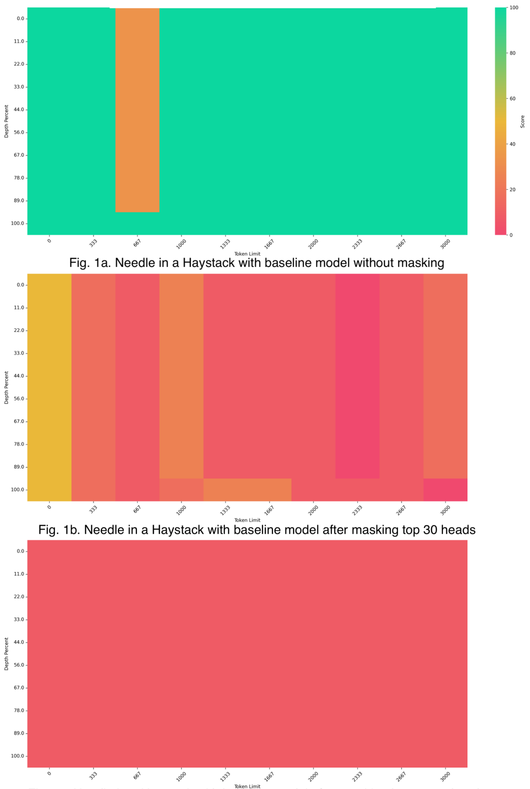

Fig. 1a. Needle in a Haystack with baseline model without masking

Fig. 1b. Needle in a Haystack with baseline model after masking top 30 heads

Fig. 1c. Needle in a Haystack with baseline model after masking bottom 30 heads

Figure 1: Needle in a Haystack experiment with different number of masked attention heads

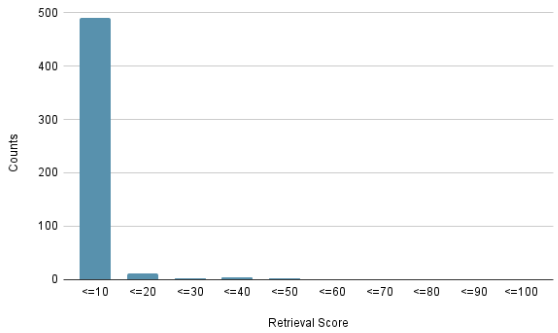

Figure 2: Distribution of retrieval scores

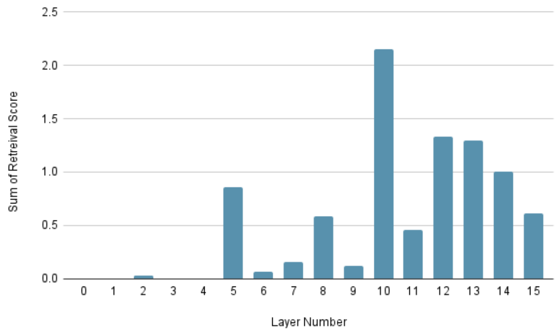

Figure 3: Sum of the retrieval scores in each attention layer

1. Consistency: To ensure all model variants learn the same features effectively, we fine-tuned all models on identical training data for 3 and 5 epochs using LoRA.

2. Comprehensiveness: We explored varying group sizes for GQA and tested multiple thresholds for masking attention heads to evaluate diverse configurations.

3. Stability: To assess the robustness of our approach, we validated the models across different downstream tasks using various datasets.

We have the following four approaches of reducing model size to compare.

1. Baseline model

2. Baseline + random GQA

3. Baseline + retrieval head masking

4. Baseline + random GQA + retrieval head masking

We experimented with group sizes of 4, 8 (baseline), and 16. When increasing the group size to 16, the baseline model's original number of key (K) and value (V) matrices was insufficient, so we stacked the matrices to initialize weights. For a group size of 4, half of the K and V matrices from the baseline model were removed by deleting matrices with even indices.

To implement retrieval head masking, we calculated retrieval scores for each attention head across all layers. Attention heads with the lowest retrieval scores were masked by disabling their query matrices, and the model's accuracy was evaluated. However, our results indicated that retrieval scores were not reliable indicators of attention head importance, prompting further investigation into this metric's limitations.

Our study comprised three sets of experiments:

1. GQA Models: We tested models with different group sizes and fine-tuning epochs to observe the effects on performance.

2. Retrieval Head Masking: We applied retrieval head masking to fine-tuned GQA models to analyze its impact on accuracy.

3. Retrieval Score Analysis: We examined the effectiveness of retrieval scores as a metric for determining attention head importance, exploring potential reasons for their limitations.

4 Result

4.1 GQA models with different group sizes and fine-tuning epochs

The results, summarized in the Tables 1 and 2, reveal that models with modified key and value matrices for group sizes of 4 and 16 initially failed to produce reasonable or grammatical sentences without fine-tuning, resulting in 0% accuracy. However, after just a few epochs of fine-tuning, the accuracies of these modified models improved significantly. Notably, models with adjusted group sizes achieved performance comparable to, or even better than, the baseline model despite having fewer parameters. This highlights the potential of grouping or filtering attention heads using appropriate techniques.

4.2 Fine-tuned GQA models with retrieval head masking

The results of this experiment are presented in Table 3. It was initially expected that attention heads with lower retrieval scores would contribute less to the output tokens. Consequently, masking attention heads with the lowest retrieval scores was anticipated to minimally impact accuracy, with accuracy gradually declining as more attention heads were masked. However, our observations challenge this expectation. The relationship between the number of masked attention heads and accuracy was not consistently inverse. In some cases, accuracy decreased as more attention heads were masked, while in others, it unexpectedly increased. These findings raise several questions and motivate further analysis of the retrieval score mechanism.

4.3 Analysis of retrieval score and its effectiveness

In the previous section, we identified limitations in the retrieval score metric. To explore these further, we analyzed the effectiveness of retrieval scores and their distribution across attention heads.

We began with the "needle in a haystack" experiment, designed to test a model's ability to retrieve critical information from lengthy contexts. The experiments varied both context length and the depth at which the information was embedded, as shown in Figure 1. Results showed that the original model, without masking any attention heads, performed well. However, masking the top 30 attention heads significantly reduced performance, consistent with findings from the original paper. Surprisingly, masking the bottom 30 heads led to an even greater performance drop, where the model scored almost all 0's on different context length and depths. This suggests that attention heads with the lowest retrieval scores may still play a crucial role, contradicting the intended definition of the retrieval score.

Next, we analyzed the distribution of retrieval scores. As shown in Figure 2, over 95% of attention heads scored below 10 on a 100-point scale, making the retrieval score a poor discriminator of attention head importance. Additionally, Figure 3 illustrates that attention heads in lower layers, particularly layers 0 to 4, tend to have extremely low retrieval scores, often close to 0. This highlights a critical limitation of the retrieval score: it inadequately reflects the importance of lower-layer attention heads.

The retrieval score is calculated by counting how often the highest attention score in a matrix corresponds to the current output token. For lower-layer attention heads, their outputted attention matrices are often more related to syntactic structures or neighboring word relations than directly to the output token. As a result, these attention heads tend to have smaller retrieval scores, even though they may still be essential. Masking the bottom 30 attention heads based on retrieval scores nearly eliminates the functionality of layer 0, which likely explains the model's failure to produce coherent sentences under this condition.

5 Future works

5.1 Enhancing Query-Grouping Logic

Instead of tuning all LLMs for downstream tasks, it is more practical to leverage model internal knowledge with an easier tuning method. According to the AsymGQA, changing the pair of attention heads in the group can improve the performance on multiple datasets. Therefore, enhancing the GQA logic by developing systematic methods to map grouped queries to the most relevant keys and values is one of the crucial future works that might reduce the computational and training costs of fine-tuning the pre-train model on downstream tasks.

Two possible methodologies are used to calculate the similarity of attention heads based on the attention layer's weights and attention information, calculated from the inference attention score of the query, key, and value layers. The former method can easily be understood as merging similar parameters into a single head, which might lose the vari-

Group size	**Without fine-tune**	**3 epochs**	**5 epochs**
g = 4	0.00	**0.74**	**0.74**
g = 8 (Original)	0.34	0.66	**0.72**
g = 16	0.00	**0.70**	0.68

Table 1: Accuracy of models with different group sizes of GQA on MNLI-m dataset

Group size	**Without fine-tune**	**3 epochs**	**5 epochs**
g = 4	0.00	0.66	**0.74**
g = 8 (Original)	0.34	0.62	**0.68**
g = 16	0.00	0.46	**0.74**

Table 2: Accuracy of models with different group sizes of GQA on MNLI-mm dataset

ance of two combining heads. The latter indicates the unique information carried on the downstream tasks, which means a higher activation score suggests a more similar attention effect on a given input. This method is more focused on capturing task-specific features. However, neither of these methods shows how to connect query, key, and value heads after merging similar heads. Thus, working on the methodologies of connecting merged queries, keys, and values can potentially reduce the loss of information from the merged head and lead to a lower parameter size.

5.2 Refining Retrieval Mechanisms

To further work on the information flow of retrieval score, our experiments show the limitation of using the correlation between the attention output and the final prediction. Since different layers of attention would capture different levels of information, it is not fair to directly compare each layer's attention score with the final output. For example, the lower layer tends to work on the local patterns and syntactic information, which would not directly affect the final production, causing a lower score in retrieval information. In contrast, the upper final layer tends to encode global semantic information and task-specific features that correlate highly with the model output. Therefore, investigating more precise metrics for identifying the most and least impactful attention heads for different layers is still an unsolved topic.

Two ideas can be discussed to refine the retrieval mechanisms: local-wise and global-wise. Local-wise calculation means comparing each head's output with its attention layer score. This method can ensure all the heads work on the same objective and update or remove the worst attention head in each layer. Global-wise calculation refers to the current methodology with layer-wise adjusted parameters. This mechanism would help reflect a lower attention head output's true impact on the final decision. Moreover, it preserves the intuition that every head needs to use the final out to measure its ability.

5.3 Expanding Applications

In our project result, we only worked on multi-NLI and needle-in-a-haystack tasks to exhibit our experiment idea. To further justify the phenomenon of useless head and misjudged retrieval scores, testing on more diverse datasets and functions is necessary for our research to provide a comprehensive and reliable outcome. Due to the motivation of our project being to reduce the model size on downstream tasks, fine-tuning, and inference, our subsequent research must work on other downstream tasks or multi-modality to see this concept's generalization.

6 Conclusion

In this paper, we discussed the meaning of the attention parameters and their contribution to the final output. We reduced the model size according to the downstream tasks like multi-NLI and needle in a haystack dataset. Our experiments indicated the comparable performance of different sizes of groups in the final layer of GQA, suggesting that lowering the model size with few-shot fine-tuning can recover the performance. Another finding is that the retrieval score calculated in each head would be biased by different layer objectives. With the few shots fine-tuning on a small amount of data, all the parameters are essential to the specific downstream tasks. To achieve the aim of this paper to reduce the model parameters on downstream tasks for fine-tuning and inference, future works

Group size	**0**	**5**	**10**	**15**	**24**	**32**
`g = 4`	0.74	0.72	0.68	0.64	0.72	0.64
`g = 8 (Original)`	0.66	0.40	0.44	0.44	0.44	0.00
`g = 16`	0.70	0.60	0.42	0.64	0.60	0.78

Table 3: Accuracy of the models with different group sizes of GQA and different number of masked heads on MNLI-m dataset. For g=4 and 16, only the attention heads in the final layer are masked

like enhancing query-grouping logic, refining retrieval mechanisms, and expanding applications are proposed for further experiments.

Contribution

Equal contribution

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *arXiv preprint arXiv:2305.10266*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yuang Chen, Cheng Zhang, Xitong Gao, Robert D. Mullins, George A. Constantinides, and Yiren Zhao. 2024. Optimised grouped-query attention mechanism for transformers.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

Liyuan Liu, Jialu Liu, and Jiawei Han. 2021. Multi-head or single-head? an empirical comparison for transformer training. *CoRR*, abs/2106.09650.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *CoRR*, abs/1905.10650.

Ankit Kumar Upadhyay and Harsit Kumar Upadhya. 2023. Xnli 2.0: Improving xnli dataset and performance on cross lingual understanding (xlu). In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.