# Head Ablation Study in Cross-Language settings with mBERT

**Ruilin Yang**
University of Potsdam
ruilin.yang@uni-potsdam.de

## 1 Introduction

Transformer models rely on attention mechanisms that consist of multiple heads to process information. However, these models often have billions of parameters, and thus are too computation-intensive to suit low-capability devices or applications with strict latency requirements. Various studies proposed methods of compression transformer model such as BERT (Ganesh et al., 2021). While structured pruning studies have been conducted on monolingual BERT, there is limited understanding of head importance in multilingual models across different languages. This study aims to fill this gap by evaluating the significance of individual attention heads in multilingual BERT model and examining their generalizability across languages.

## 2 Previous Works

Pruning techniques such as head ablation (Michel et al., 2019) have provided valuable insights into the inner workings of monolingual BERT models, revealing that a high percentage of heads can be pruned without significantly affecting the model performance. Budhraja et al. (2021) compared the effects of pruning on BERT and multilingual BERT. However, the specific roles of individual attention heads in multilingual BERT remain underexplored. While multilingual BERT models like mBERT have demonstrated impressive cross-lingual zero-shot transfer learning capabilities (Wu and Dredze, 2019), to which extent its transfer learning ability is associated with individual heads remains unknown.

This study aims to apply the systematic pruning approach of (Michel et al., 2019) on mBERT, to analyze head importance in multilingual in-language learning settings, as well as in zero-shot cross-lingual transfer learning.

## 3 Experiment Setup

### 3.1 Model

We will use the *bert-base-multilingual-uncased* model (Devlin et al., 2018) as the starting point of subsequent fine-tuning and experiments. With 12 layers of encoders and 12 attention heads within each encoder layer, it has in total 168M parameters.

### 3.2 Dataset

We will fine-tune the selected models on the *XNLI dataset* (Conneau et al., 2018), which includes pairs of sentences of NLI task in 15 languages. Evaluation will be conducted on the whole test set as well as language-specific subsets of the test set.

### 3.3 Ablation Procedure

**Head Ablation**: We will systematically ablate one head at a time and measure the performance impact.
**Layer Ablation**: For each layer, we will ablate all heads but one to evaluate the layer-wise importance of individual heads.

## 4 Research Questions

**RQ1**: What is the overall and per language baseline performance of mBERT, after being fine-tuned on the XNLI dataset?
**RQ2**: After each of the ablation methods, how does the overall and per language performance drift?
This research question could be break down into multiple sub-questions: Is there any relationship among the changes across different languages? Do languages from the same language family tend to have similar pattern of performance drift when different heads are pruned? Does the finding of Michel et al. (2019) still hold? Are there any specific heads or layers that have a significant impact on performance, when ablated? And are these impact the same across languages, or are they language-specific?

**RQ3**: Does Ablation affect cross-lingual zero-shot transfer learning?
For this question, we fine-tune mBERT on English data only, and evaluate its performance on every other language after each of the ablation methods. Do the patterns found in RQ2 (if any) still hold?

## 5 Evaluation and Analysis

Performance metrics, such as accuracy and F1 score, will be used to assess the impact of head ablation across languages. The results will be compared to identify patterns and correlations in head importance between languages. Statistical methods will be employed to correlate head importance across different languages, providing insights into the generalizability of heads.

## 6 Personal Learning Goals

1. Get hands-on experience of fine-tuning a pretrained model, and tweaking the model source code in some way.
2. Brush up the statistical methods that are commonly used in NLP reasearch to analyze the evaluation data.

## 7 Conclusion

This proposal outlines a cross-language head ablation study to evaluate the importance of individual attention heads in multilingual BERT model. By systematically analyzing head importance across languages, this study will provide valuable insights into the cross-lingual capabilities of multilingual models and inform future advancements in multilingual NLP.

### 7.1 References

## References


Aakriti Budhraja, Madhura Pande, Pratyush Kumar, and Mitesh M Khapra. 2021. On the prunability of attention heads in multilingual bert. *arXiv preprint arXiv:2109.12683*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, 9:1061–1080.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.