MESH - Multi-head Efficiency in XTREME with Multilingual Specialization and Pruning Heuristics

Gautam Gupta Ravil Patel Avantika Singh
Department of Data Science and Artificial Intelligence (DSAI)
IIIT Naya Raipur, India
{gautam21102,ravil21102,avantika}@iiitnr.edu.in

Abstract

Meta-learning has gained significant attention for enabling models to learn and adapt quickly to new tasks and languages. In this paper, we set up a multi-task and multilingual environment to test out the applications of meta-learning techniques, specifically focusing on question answering and natural language inferencing tasks across multiple languages. Multi-head attention has been successfully applied to a variety of real-world tasks and even achieved state of the art performance on multiple downstream tasks, however a challenge still persists to determine what mechanisms have been learnt by it.

Our goal is to implement a straightforward multi-task training approach to enhance functional specialization and reduce adverse information transfer in multi-task learning. We achieve this by estimating the importance of heads and then pruning part of the top important heads for each task- a judicious reduction of model parameters based on their importance which is a proven way to mitigate the inference cost incurred from the attention mechanism that serves as an optimization strategy for attention based models. Our results showcase that with strategical selection of hyperparameters while pruning, we can push the boundaries of current state-of-the-art BERT models in multi task settings and improve task association without any compromise on task accuracy.

***Index Terms*—NLP, task-language pairs, attention head pruning, and meta learning**

I. Introduction

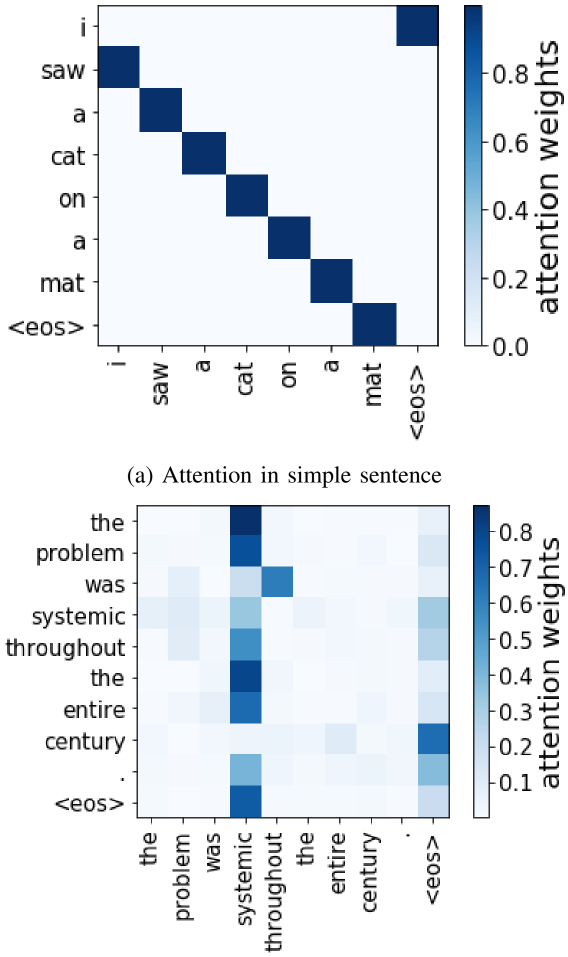
Meta-learning has shown great potential in recent years due to its ability to enable models to rapidly adapt to new tasks and domains. By leveraging meta-learning techniques, we aim to address the challenges associated with multi-task and multilingual modelling in natural language processing (NLP) tasks. In this paper, we implement a meta-learning framework that facilitates effective multi-task learning and multilingual modelling for question answering and natural language inferencing across various languages. Our approach involves leveraging the strengths of existing models and enhancing them to handle multiple tasks and languages simultaneously. We present our methodology, experimental setup, results, and discuss the shortcomings of our approach. This paper aims to address the following question: Could a lightweight multilingual multitask model be produced which leverages information from other task-language pairs, keeping only important attention heads while using much less compute than training one from scratch?

Performing tasks like Question Answering and Natural Language Inference demands a deep semantic understanding and intricate representation of language, requiring models to grasp nuanced relationships and infer meanings from text. In contrast, tasks like Named Entity Recognition, Sentence Classification, and Part of Speech Tagging rely more on surface-level features and patterns within the text rather than intricate semantic understanding. A single model excelling at all these tasks with high accuracy is challenging due to the varying levels of understanding required and the need to balance representations across tasks effectively with limited information. This challenge necessitates careful model design, training strategies, and optimization techniques to achieve desired performance levels across a diverse range of linguistic tasks.

Attention serves as the backbone mechanism which enables neural network models to concentrate on specific and significant chunks of information by calculating their weighted averages during prediction. Multi-headed attention, in particular, drives many cutting-edge natural language processing (NLP) models, such as Transformer-based machine translation models and multilingual models like BERT. These models employ multiple attention mechanisms parallelly, with every attention ”head” focusing on distinct segments of the input, allowing for the expression of complex functions beyond mere weighted averages. In our experimental findings, we come across the surprising observation that a significantly large percentage of attention heads can be pruned at testing for models which have been trained on multiple heads, with an increase on model performance and association with similar tasks. Extending on this, a few layers can be further reduced to singular heads.

Our work is an implementation of semi structured pruning since we only prune attention heads without disturbing any other parts of the architecture. All the other FCN layers and embedding layers are untouched and our findings are restricted to the role of attention module in multitask models We derive two broad sets of findings. The task language pairs we choose initially, we analyze the impact of pruning on the performance of BERT and mBERT within their respective languages. While it might seem that mBERT, with its reduced dedicated attention capacity for English, would suffer more from pruning compared to BERT, our findings reveal otherwise. Surprisingly, mBERT demonstrates a similar resilience to pruning as BERT.

In accordance with the assessment methodology outlined in [Budhraja et al., 2021], we proceed to randomly eliminate k% of attention heads within mBERT as our first baseline and then verify it against eliminating k% important attention heads benchmark. Subsequent to pruning, the model undergoes fine-tuning for a duration of 10 epochs.



(a) Attention in simple sentence

(b) Attention in complex sentence

Fig. 1: Attention visualization

II. Related Works

Attention mechanisms have become a cornerstone in improving the performance of Large Language Models across various natural language processing (NLP) tasks. In recent years, attention head pruning has emerged as a promising technique for enhancing the efficiency and effectiveness, particularly in multi-head meta-learning tasks.

Research in attention head pruning has primarily focused on optimizing the multi-head attention mechanism of Transformer-based architectures. Notably, the work by [Z Zhang et al.] [Dean and Ghemawat, 2004] introduced Important Attention-head Pruning (IAP), which quantifies the importance of attention heads across different tasks and selectively prunes less relevant heads to improve computational efficiency without sacrificing task performance. This approach has been applied in various contexts, including single-task learning and multi-task learning settings.

In the domain of multi-head meta-learning, attention head pruning has garnered significant attention due to its potential to enhance model adaptation and generalization capabilities. [Praboda rajapaksha et al.] explored the use of attention head pruning in multi-head meta-learning tasks, demonstrating its effectiveness in improving meta-learning performance across diverse tasks and domains. By selectively retaining attention heads crucial for meta-learning objectives, models achieve superior adaptation and generalization performance while reducing computational overhead.

Furthermore, attention head pruning has been investigated for its implications on model interpretability and transparency. Studies by [Nicholas Pochinkov et al.] have shown that pruning less relevant attention heads leads to more interpretable attention patterns within LLMs, facilitating clearer functional specialization and task-specific attention distributions.

Our research delves into the distinct functions of individual attention heads within the multi-head attention module, an area of research that has garnered traction in recent years. [Voita et al., 2019] demonstrated the existence of redundant heads in transformers through the elimination of less crucial heads and evaluation of the subsequent performance, a finding independently supported by [Michel et al., 2019]. Examining the linguistic characteristics of sentence representations generated by attention heads, [Jo and Myaeng, 2020] conducted ten linguistic probing tasks. The study conducted by [Hao et al., 2021] only retained focused on retaining only the essential heads in BERT and developing an attribution tree to elucidate the interactions of information within the Transformer. By pruning attention heads, the researchers investigated the specific roles they play in various tasks, rather than merely demonstrating redundancy within the multi-head attention mechanism [Michel et al., 2019].

Our approach is centered on partitioning task-critical modules within shared parameters to alleviate detrimental transfer effects across tasks, distinguishing it from previous methods involving sampling or task-specific adapter mechanisms ([Wu et al., 2020]; [Pilault et al., 2021]). During training, we only need to retain mask variables for each attention head, as opposed to preserving all parameters, leading to a substantial reduction in memory expenses ([Sun et al., 2020], [Lin et al., 2021]; [Xie et al., 2021]; [Liang et al., 2021]).

In this particular section, a thorough examination is presented on multi-head attention [Vaswani et al., 2017] with the aim of establishing a precise technical lexicon essential for the discourse on our methodologies for head pruning. Specifics regarding additional components of the Transformer are excluded, redirecting the audience to the primary publication by [Vaswani et al., 2017] .

Let $z = (z_1, \ldots, z_T)$ be a sequence of T real vectors where $z_t \in \mathbb{R}^d$, and let $q \in \mathbb{R}^d$ be a query vector. An attention mechanism is defined as:

$$att(z, q) = W \sum_{\alpha(q)} W_{v,z,\alpha} T \tag{1}$$

where

$$\alpha_t(q) = \text{softmax}\left(\frac{q^t W_q^t W_k z_t}{\sqrt{d}}\right)_t \tag{2}$$

The projection matrices Wo, Wv, Wq, Wk Rd×d are learnable parameters. In self-attention, query q comes from the same sequence z. A Transformer is composed of L identical layers. In layer different attention mechanisms are applied in parallel; importantly, it is this parallelism that has lead to the rise of the Transformer—it is a more efficient architecture in practice so it can be trained on more data. Each individual attention mechanism is referred to as a head; thus, multi-head attention is the simultaneous application of multiple attention heads in a single architecture. In Vaswani et al. (2017), the multiple heads are combined through summation:

where attlh is the hth attention head in the lth layer. We also implement a gate variable glh that takes values in the interval [0, 1]:

Inserting glh into the multi-head attention enables our pruning approach: setting the gate variable to glh = 0 means the head attlh is pruned away. In the following sections, for the sake of notational simplicity, we ignore the layer structure of heads and label heads with a single index h 1,...,H, where H is the total number of heads in the unpruned model.

Observations by [Budhraja et al., 2021] showcase that with 50% random pruning, the average accuracy drop for mBERT on GLUE tasks is only 2% relative to its base performance. Additionally, they observe that mirroring BERT, mBERT too relies more heavily on the attention heads in the middle layers than those in the top and bottom layers.

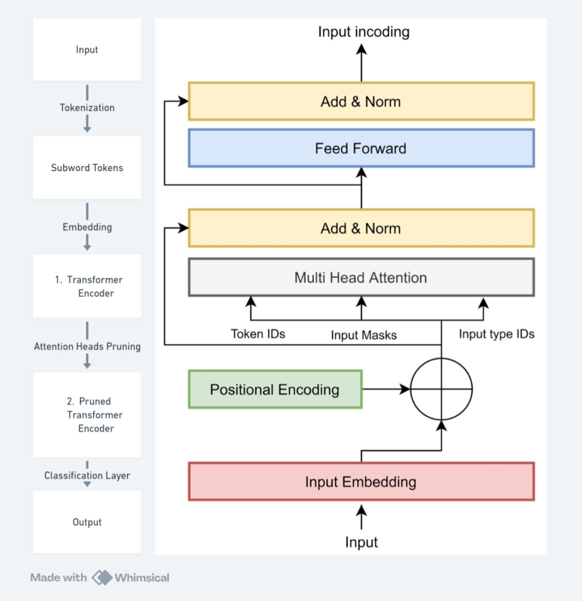


Fig. 2: Attention Mechanism

III. Methodology

A. Head Importance Score

Head Importance Score Michel et al. (2019) proposed an effective method to prune attention heads and evaluate the importance of attention heads for a task. In order to prune the attention head h, they incorporated a mask variable into the attention function:

$$\tilde{A}_h(X) = \mathcal{E}_h \cdot A_h(X) \tag{3}$$

and set it to a zero value. When $\mathcal{E}_h$ equals 1, Equation (3) is the same with the vanilla attention (Eq.(1)). The head importance score I approximated by the expected sensitivity of loss function to the mask variable:

$$I_h^{(i)} = E_{(x,y)\sim D^i} \left| \frac{\partial L^{(i)}(x, y)}{\partial \mathcal{E}_h} \right| \tag{4}$$

is the loss of task T_i on sample (x, y). Different from Michel et al. (2019) which pruned the least important attention heads to prove the redundancy of attention heads, this paper focuses on exploring the functional specialization phenomenon after training, thus we prune the most important heads for each task.

B. Interpreting: Important Attention-head Pruning

We implement a two-step method, namely *Important Attention-head Pruning* (IAP), to quantify the degree of functional specialization in multi-head attention. First, the top $\alpha \in [0, 1]$ percentage important heads H_i^α for task $\mathcal{T}_i$, e.g., the ones circled by dashed lines in Figure 1, are found after dual-task or multi-task training by their head importance scores. Specifically, we calculate the head importance score $I_h^{(i)}$, defined on training samples to approximate the contribution of head h to task $\mathcal{T}_i$.

Second, dissociation experiments are conducted to determine the degree of functional specialization in multi-head attention. Given a model f_θ after dual-task training on tasks $\mathcal{T}_A$ and $\mathcal{T}_B$, for example, the relative performance on $\mathcal{T}_A$ after pruning the top α important attention heads for $\mathcal{T}_B$, denoted by H_B^α, is calculated as follows:

$$RP_A(H_B^\alpha) \quad = \quad \frac{\mathcal{P}\left(f_{\theta \setminus H_B^\alpha}(X_A), Y_A\right)}{\mathcal{P}(f_\theta(X_A), Y_A)} \tag{1}$$

where $\mathcal{P}(\cdot)$ is the performance metric used, e.g., Accuracy, and (X_A, Y_A) is the test sample of Task $\mathcal{T}_A$. Then, we estimate the degree of functional specialization by the relative performance difference after top α important heads for each task are pruned, called dissociation score:

$$\begin{array}{rcl} D_A(\alpha) & = & RP_A(H_B^\alpha) - RP_A(H_A^\alpha), \\ D_B(\alpha) & = & RP_B(H_A^\alpha) - RP_B(H_B^\alpha), \\ D(\alpha) & = & \frac{D_A(\alpha)+D_B(\alpha)}{2} \end{array} \tag{2}$$

where $D_A(\alpha)$ denotes the dissociation score of task $\mathcal{T}_A$, and $D(\alpha)$ is the average dissociation score of this dual-task learning. Given an appropriate α, a larger dissociation score implies a higher degree of functional specialization.

C. Model Working

This work presents a novel meta-learning approach for handling few-shot learning tasks across multiple languages. We leverage a pre-trained multilingual Transformer architecture with a multi-head attention mechanism. To enhance the model's focus on relevant information and potentially reduce computational complexity, we introduce gating and pruning techniques within the attention layers. Trainable gating weights denoted by g are integrated to modulate the importance of each attention head, focusing on the most relevant aspects of the input for the specific task and language. Alternatively, heuristics based on task or language characteristics can be employed to identify and prune redundant attention heads before meta-testing. This pruning is implemented through a binary mask function P that sets elements corresponding to pruned heads to 0.

The algorithm follows a two-stage training process: meta-training and meta-testing. During meta-training, we define a set of few-shot learning tasks T and corresponding languages L. In each iteration, a batch of tasks t_j and languages l_i is sampled. An inner loop then focuses on fine-tuning the model parameters for each sampled pair. The inner loop utilizes the gating function g within the attention mechanism and potentially applies pruning with the mask P. The model parameters are updated using Stochastic Gradient Descent (SGD) to minimize the loss function f based on predictions and true labels in the support set. This adapted model is then used to make predictions on the query set, and the inner loop loss L_inner is calculated. The outer loop leverages the average loss from all inner loops (mean(L_inner) across sampled tasks) to update the meta-learner's parameters , improving its ability to adapt the base model to new languages and tasks.

During meta-testing, a novel task t_new and language l_new not encountered during meta-training are introduced. The meta-learner is then used to adapt the base model's parameters to the new task and language, leveraging the knowledge acquired during meta-training. Finally, the adapted model's performance (_adapted) is assessed on the new few-shot learning task in the unseen language. Additionally, optimization strategies like lower learning rates for meta-training and gradient clipping are employed to improve stability and prevent overfitting. This combination of meta-learning, multilingual capabilities, attention gating/pruning, and optimization strategies has the potential to achieve superior performance in few-shot learning tasks across diverse languages. Future research directions include exploring more sophisticated gating mechanisms and dynamic pruning strategies to further enhance the model's efficiency and effectiveness.

D. Algorithm

Optimization Strategies:
1. Lower learning rate for outer loop compared to inner loop.
2. Implement gradient clipping during meta-update.

Algorithm 1 Inner Loop (Adaptation)

```
[1] Sample Task-Language Pair: $(l_i, t_j) \in L \times T$
Initialize Model: $\theta_{\text{init}} = \theta_{\text{pre}}$
Fine-Tune Model: support set data $(x_s, y_s)$ in task $(l_i, t_j)$ $\theta_{\text{init}} = SGD(\theta_{\text{init}}, f(\text{model}(g(x_s, \theta_{\text{init}}), P(\theta_{\text{init}})), y_s))$
Evaluate on Query Set: Predict on query set data $(x_q, y_q)$: $y_{\text{hat}} = \text{model}(g(x_q, \theta_{\text{init}}), P(\theta_{\text{init}}))$
Calculate inner loop loss: $L_{\text{inner}} = f(y_{\text{hat}}, y_q)$
```

Algorithm 2 Outer Loop (Meta-Update)

```
[1] Average Inner Loop Loss:
$L_{\text{avg}} = \text{mean}(L_{\text{inner}}$ across all sampled tasks)
Update Meta-Learner: $\tau = SGD(\tau, L_{\text{avg}})$
```

Algorithm 3 Meta-Testing

```
[1] Present New Task-Language Pair:
$(l_{\text{new}}, t_{\text{new}}) \notin L \times T$
Adapt Model with Meta-Learner:
$\theta_{\text{adapted}} = \text{model}(g(x_s, \tau(\theta_{\text{pre}})), P(\tau(\theta_{\text{pre}})))$
for support set data
Evaluate on New Task Predict on query set data:
$y_{\text{hat}} = \text{model}(g(x_q, \theta_{\text{adapted}}), P(\theta_{\text{adapted}}))$
Calculate final loss: $L_{\text{final}} = f(y_{\text{hat}}, y_q)$
```

TABLE I: Ablation study of different multi-task methods on XTREME dev set.

Model	Accuracy							
	NLI-en	NLI-es	QA-en	QA-es	NER-en	NER-es	PA-en	PA-es
m-BERT$_{\text{base}}$	$65.51_{\pm 4.0}$	$62.91_{\pm 3.2}$	$53.0_{\pm 1.9}$	$55.28_{\pm 1.3}$	$77.60_{\pm 2.6}$	$74.21_{\pm 1.9}$	$85.33_{\pm 0.5}$	$74.08_{\pm 0.3}$
Prune least imp 30% heads	$69.16_{\pm 4.0}$	$63.16_{\pm 3.2}$	$65.88_{\pm 1.9}$	$62.10_{\pm 1.3}$	$80.50_{\pm 2.6}$	$82.37_{\pm 1.9}$	$83.68_{\pm 0.5}$	$84.25_{\pm 0.3}$
Train random 30% heads	$72.06_{\pm 4.0}$	$68.50_{\pm 3.2}$	$69.72_{\pm 1.9}$	$66.71_{\pm 1.3}$	$82.5_{\pm 2.6}$	$84.03_{\pm 1.9}$	$85.79_{\pm 0.5}$	$86.11_{\pm 0.3}$
Train most imp 30% heads	$75.40_{\pm 4.0}$	$74.11_{\pm 3.2}$	$72.12_{\pm 1.9}$	$73.60_{\pm 1.3}$	$83.10_{\pm 2.6}$	$86.16_{\pm 1.9}$	$86.02^{\pm 0.5}$	$84.55_{\pm 0.3}$

IV. Evaluation

A. Experimental Setup

We selected a diverse dataset covering two languages(French and English) and NLP tasks to evaluate the performance of the mBERT model after applying pruning techniques. After preprocessing the data, we initialized the mBERT model with pretrained weights and architecture suitable for the tasks and languages at hand. Utilizing the SGD optimizer, we configured parameters such as learning rate and momentum and implemented nested loops for training, iterating over task-language pairs to compute individual sample losses. We then aggregated the average loss across sampled pairs and evaluated the pruned mBERT model using metrics like accuracy and F1-score. Employing cross-validation techniques and conducting multiple experimental runs with varying configurations ensured robustness and reproducibility of the results, which were subjected to statistical analysis to compare against baseline performance.

B. Model

Incorporating bidirectional contextual information, mBERT, a pioneering natural language processing (NLP) model presented in a study by [Devlin et al., 2018], leverages the Transformer architecture, which is trained through masked language modeling (MLM) and next sentence prediction (NSP). In the construction of a multi-task meta-learning model that utilizes attention pruning techniques in conjunction with BERT, the first step is to pre-train mBERT on an extensive corpus, followed by fine-tuning it for various NLP tasks concurrently while integrating task-specific layers. The implementation of a meta-learning framework exposes the model to a wide array of tasks during meta-training, thereby enhancing its ability to rapidly adjust to novel tasks during meta-testing. Subsequently, attention pruning methods are incorporated to reduce computational complexity without compromising performance, thereby enhancing the efficiency of the attention mechanisms. This holistic strategy merges mBERT's contextual comprehension with multi-task learning, meta-learning, and attention pruning, producing a versatile and resource-efficient solution for NLP challenges. In all experimental setups involving BERT, the mBERT Base-uncased model is employed, characterized by 12 layers with each layer comprising 12 attention heads, totaling to 144 attention heads.

TABLE II: XTREME Dataset instances used during the training, evaluation and testing for the setup

Task	Language	
	en	es
Natural Language Inference (NLI)	392k	392k
Question Answering (QA)	88.0K	81.8K
Named Entity Recognition (NER)	20K	20K
Paraphrase Identification (PA)	49.4K	49.4K

C. Implementation Details

We adopted the mBERT model as our foundational architecture and implemented pruning techniques, leveraging its proven state-of-the-art performance across diverse NLP tasks spanning multiple languages. In our experimental setup, we employed the SGD optimizer to iteratively update model weights via backpropagation, processing batches comprising various Task-Language pairs. Our model operates within two nested loops: an outer loop and an inner loop. Within the inner loop, the loss for each individual sample is computed, while in the outer loop, we aggregate the average loss across all sampled task-language pairs.

D. Results

Table 2 shows the accuracy scores of various Task-Language pairs in different pruning and retraining settings. Out pruning hyperparameter k is taken as 30, and measured against m-BERT base model with all attention heads. We are able to conclude that pruning least important 30% heads and then training with most important 30% heads improves model performance across deeper language analysis tasks that require deeper semantic representation i.e, QA, and NLI The redundant attention heads are successfully trimmed and model generalization is improved across tasks.

V. Conclusion and Future Scope

We've demonstrated the capacity to derive compact pruned models that match or surpass the performance of larger distilled networks. This approach is applicable during fine-tuning rather than pre-training. It doesn't rely on methods like data augmentation or architecture search and remains effective across various tasks and base models. Additionally, we've noted the emergence of a brain-like functional specialization phenomenon in multi-head attention following dual-task or multi-task learning. Moreover, our experimental findings indicate that the performance and generalization capabilities of multi-task models can be enhanced through the multitask training approach grounded in functional specialization.

Furthermore, similar strategy can be used to optimize tasks like Coreference Resolution, extractive and abstractive Question Answering, textual entailment etc. pur findings can be extended to incorporate languages with high lexical similarity and common scripting style. Experimentation with longform text and Longformer architectures can page way to understanding important functions in attention models.

References

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

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, 2004. URL http://www.usenix.org/events/osdi04/tech/dean.html.

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

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12963–12971, May 2021. doi: 10.1609/aaai.v35i14.17533. URL https://ojs.aaai.org/index.php/AAAI/article/view/17533.

Jae-young Jo and Sung-Hyon Myaeng. Roles and utilization of attention heads in transformer-based neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.311. URL https://aclanthology.org/2020.acl-main.311.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. Super tickets in pre-trained language models: From model compression to improving generalization. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6524–6538, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.510. URL https://aclanthology.org/2021.acl-long.510.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. Learning language specific sub-network for multilingual machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.25. URL https://aclanthology.org/2021.acl-long.25.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Jonathan Pilault, Amine El hattami, and Christopher Pal. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=de11dbHzAMF.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. Learning sparse sharing architectures for multiple tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8936–8943, Apr. 2020. doi: 10.1609/aaai.v34i05.6424. URL https://ojs.aaai.org/index.php/AAAI/article/view/6424.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylzhkBtDB.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. Importance-based neuron allocation for multilingual neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.445. URL https://aclanthology.org/2021.acl-long.445.