Shapley Head Pruning: Identifying and Removing Interference in Multilingual Transformers

William Held
wheld3@gatech.edu

Diyi Yang
diyiy@stanford.edu

Abstract

Multilingual transformer-based models demonstrate remarkable zero and few-shot transfer across languages by learning and reusing language-agnostic features. However, as a fixed-size model acquires more languages, its performance across all languages degrades. Those who attribute this interference phenomenon to limited model capacity address the problem by adding additional parameters, despite evidence that transformer-based models are overparameterized. In this work, we show that it is possible to reduce interference by instead identifying and pruning language-specific attention heads. First, we use Shapley Values, a credit allocation metric from coalitional game theory, to identify attention heads that introduce interference. Then, we show that pruning such heads from a fixed model improves performance for a target language on both sentence classification and structural prediction. Finally, we provide insights on language-agnostic and language-specific attention heads using attention visualization.[1]

1 Introduction

Cross-lingual transfer learning aims to utilize a natural language processing system trained on a source language to improve results for the same task in a different target language. The core goal is to maintain relevant learned patterns from the source while disregarding those which are inapplicable to the target. Multilingual pretraining of transformer language models has recently become a widespread method for cross-lingual transfer; demonstrating remarkable zero and few shot performance across languages when finetuned on monolingual data (Pires et al., 2019; Conneau et al., 2019; Xue et al., 2021).

However, adding languages beyond a threshold begins to harm cross-lingual transfer in a fixed-size model as shown in prior work (Conneau et al., 2019; Xue et al., 2021). This phenomenon, termed interference, has been addressed with additional parameters, both language-specific (Pfeiffer et al., 2020) and broadly (Conneau et al., 2019; Xue et al., 2021). Wang et al. (2020) justifies this by showing that competition over limited capacity drives interference. This seems to contradict the lottery ticket hypothesis, which has shown that pretrained language models are highly overparameterized (Frankle and Carbin, 2019; Chen et al., 2020).

We offer an alternate hypothesis that interference is caused by components that are specialized to language-specific patterns and introduce noise when applied to other languages. To test this hypothesis, we introduce a methodology that selectively removes noisy components to improve language-specific performance without updating or adding additional language-specific parameters. Our work builds on prior research studying monolingual models that shows they can be pruned aggressively (Michel et al., 2019; Voita et al., 2019).

We leverage Shapley Values, the mean marginal contribution of a player to a collaborative reward, to identify attention heads that cause interference. Unlike prior methods, Shapley Values map each head to positive and negative values in a way that abides by all axioms of fair attribution (Ali et al., 2022). Therefore, negative values soundly mark interfering heads where removal will improve performance. We approximate Shapley Values in a computationally tractable but functionally accurate manner using truncation and multi-armed bandit sampling following prior work in computer vision (Ghorbani and Zou, 2020). We contribute the following:

1. **Attention Head Language Affinity:** Even when computed from aligned sentences, Attention Head Shapley Values vary based on the language of input. This highlights that a subset of attention heads has language-specific importance, while others are language-agnostic as shown in Figure 1.

[1] We release code to compute Shapley Values on GitHub

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2416–2427

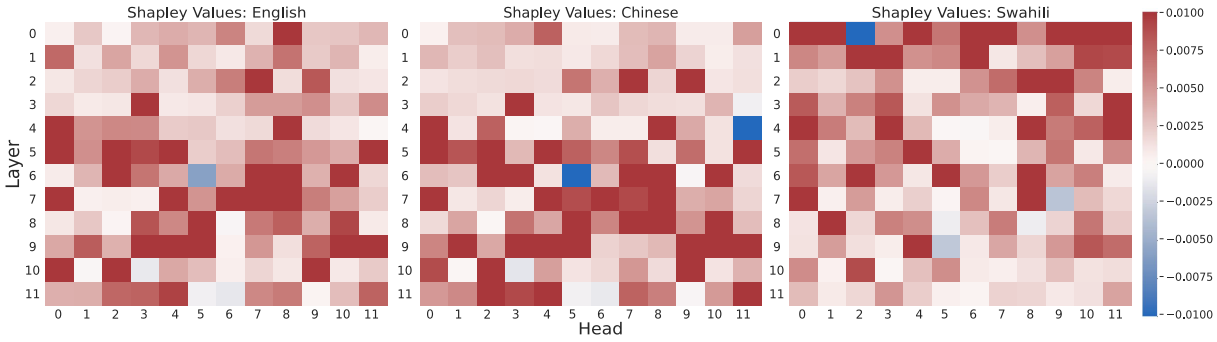

Figure 1: Attention Head Shapley Values for 3 Languages computed from 512 aligned examples for XLM-R finetuned on English XNLI. Each value represents the mean marginal effect an attention head has on accuracy for the test set in that language. The set of harmful heads changes for language, with the most distinct set for Swahili.

2. **Improving through Pruning:** Model pruning according to Shapley Values improves performance without updating parameters on the Cross-Lingual Natural Language Inference corpus (Conneau et al., 2018) and the Universal Dependencies Part-of-Speech corpus (Nivre et al., 2020). This opens a path of work to reduce interference through pruning rather than scaling.

3. **Interpreting Multilingual Heads:** In a qualitative study, we find that the most language-agnostic heads identified have a visible language-agnostic function, while language differences can be measured meaningfully for language-specific heads.

2 Related Work

2.1 Multilingual Learning

A large amount of work has studied both the theoretical underpinnings of learning common structures for language and their applications to cross-lingual transfer. Early works exploited commonality through the use of pivot representations, created either by translation (Mann and Yarowsky, 2001; Tiedemann et al., 2014; Mayhew et al., 2017) or language-agnostic task formulations (Zeman, 2008; McDonald et al., 2011).

As NLP has increasingly used representation learning, dense embedding spaces replaced explicit pivots. This led to methods that identified the commonalities of embedding spaces and ways to align them (Joulin et al., 2018; Artetxe et al., 2018; Artetxe and Schwenk, 2019). Recently, many works (Pires et al., 2019; Conneau et al., 2019; Liu et al., 2020; Xue et al., 2021; Hu et al., 2021) have trained multilingual transformer models as the basis for cross-lingual transfer. These models both implicitly and explicitly align the embedding space across languages, although they empirically achieve stronger alignment between closely related languages (Artetxe et al., 2020; Conneau et al., 2020).

With language-specific data, further work has studied how to reduce interference by adding a small number of language-specific parameters. These works adapt a model for the target language by training only Adapters (Wang et al., 2020; Pfeiffer et al., 2020; Ansell et al., 2021), prompts (Zhao and Schütze, 2021), or subsets of model parameters (Ansell et al., 2022).

Ma et al. (2021) previously investigated pruning in multilingual models using gradient-based importance metrics to study variability across attention heads. However, they used a process of iterative pruning and language-specific finetuning. This iterative process is not consistent since there are many trainable subnetworks within large models (Prasanna et al., 2020). Our method is the first to address interference and improve cross-lingual performance purely by pruning, without updating or adding additional language-specific parameters.

2.2 Model Pruning

Model pruning has largely been focused on reducing the onerous memory and computation requirements of large models. These techniques are broken into two approaches: structured and unstructured pruning. *Unstructured pruning* aims to remove individual parameters, which allows for more fine-grained removal. This process often has minimal effects even at extremely high degrees of sparsity. To efficiently prune a large number of parameters, many techniques propose using gradients

or parameter magnitude (Sundararajan et al., 2017; Lee et al., 2019; Frankle and Carbin, 2019; Chen et al., 2020) as importance metrics.

Structured pruning, or removing entire structural components, is motivated by computational benefits from hardware optimizations. In the case of Transformers, most of this pruning work targets removal of attention heads, either through static ranking (Michel et al., 2019) or through iterative training (Voita et al., 2019; Prasanna et al., 2020; Xia et al., 2022). These pruning methods have also been used to study model behavior, but methods with iterative finetuning are not consistent as many sub-networks can deliver the same level of performance once trained (Prasanna et al., 2020).

Our work studies pruning without updating model parameters, which aligns with Michel et al. (2019) which was able to remove up to 40% of total attention heads without impacting accuracy on English Natural Language Inference. However, their gradient-based importance metric does not meet key criteria of efficiency in fair allocation, which states that the sum of the metric across all heads should sum to the model's total performance (Ali et al., 2022). Furthermore, Kovaleva et al. (2019) found that pruning attention heads could sometimes improve model performance without further finetuning. We build on this to develop a methodology for consistently identifying pruned models which improve performance.

3 Methods

To identify and remove interference, we need a metric that can separate harmful, unimportant, and beneficial attention heads. Prior work (Michel et al., 2019; Ma et al., 2021) utilized the magnitude of gradients as an importance metric. However, this metric measures the sensitivity of the loss function to the masking of a particular head. Defined in this way, importance will spike indiscriminately for both harmful and beneficial heads. Therefore, we develop a simple yet effective method to separate these classes.

Shapley Values (Shapley, 1953) have often been applied in model interpretability since they are the only attribution method to abide by the theoretical properties of local accuracy, missingness, and consistency laid out by Lundberg and Lee (2017). In our setting, Shapley Values have two advantages over gradient-based importance metrics. Firstly, gradient-based approaches require differentiable relaxations of evaluation functions and masking, but Shapley Values do not. Therefore, we can instead use the evaluation functions and binary masks directly. Secondly, Shapley Values are meaningfully signed which allows them to distinguish beneficial, unimportant, and harmful heads rather than just important and unimportant heads. This latter property is essential for our goal of identifying interference.

We apply Shapley Values to the task of structural pruning. In order to compute Shapley Values for each head, we first formalize the forward pass of a Transformer as a coalitional game between attention heads. Then, we describe a methodology to efficiently approximate Shapley Values using Monte Carlo simulation combined with truncation and multi-armed bandit search. Finally, we propose a pruning algorithm using the resulting values to evaluate the practical utility of this theoretically grounded importance metric.

3.1 Attention Head Shapley Values

We formalize a Transformer performing a task as a coalitional game. Our set of players A are attention heads of the model. In order to remove self-attention heads from the game without retraining, we follow Michel et al. (2019) which augments multi-headed attention with an added gate $G_h = \{0, 1\}$ for each head Att_h in a layer with N_h heads as follows:

$$\text{MHAtt}(x, q) = \sum_{h=1}^{N_h} G_h \underset{h}{\text{Att}}(x, q) \quad (1)$$

With $G_h = 0$, that attention head does not contribute to the output of the transformer and is therefore considered removed from the active coalition.

Our characteristic function $V(A)$ is the task evaluation metric $M_v(A)$ over a set of validation data within a target language, adjusted by the evaluation metric with all heads removed to abide by the $V(\emptyset) = 0$ property of coalitional games:

$$V(A) = M_v(A) - M_v(\emptyset) \quad (2)$$

With these established, the Shapley Value φ_h for an attention head Att_h is the mean performance improvement from switching gate G_h from 0 to 1 across all P permutations of other gates:

$$\varphi_h = \frac{1}{|P|} \sum_{A \in P} V(A \cup h) - V(A) \quad (3)$$

3.2 Approximating Shapley Values

The exact computation of Shapley Values for N attention heads requires 2^N evaluations of our valida-

tion metric, which is intractable for the number of heads used in most architectures. The computation becomes more tractable with Monte Carlo simulation as an approximation (Castro et al., 2009). This replaces the full permutation set P in Equation 3 a randomly sampled subset of permutations.

Computing low-variance Shapley Value estimates with Monte Carlo simulation alone is computationally expensive and provides no clear metric for convergence. Therefore, we follow Ghorbani and Zou (2020) to accelerate our computations. We add a truncation heuristic using priors about the behavior of neural networks and formulate estimation as a multi-armed bandit problem of separating harmful heads from all others. We show in Section 4.3 that this approximation not only reduces the number of samples but explicitly converges to a consistent set of harmful heads across runs, showing consistency even across languages.

Truncation Heuristics Truncation stops sampling the marginal contributions from the rest of a permutation of features once a stopping criterion is reached for that permutation of the Monte Carlo simulation. Prior work selects stopping criterion based on either total performance (Ghorbani and Zou, 2020) or marginal improvements (Ghorbani and Zou, 2019). To avoid tailoring a threshold to each dataset, we instead choose to truncate based on the percentage of remaining attention heads. For all experiments, we truncate when less than 50% of attention heads remain in the coalition. This biases our estimations towards the effect of heads when the majority of the full network is present.

Multi-Armed Bandit Sampling The multi-armed bandit optimization stops sampling the marginal contributions of a particular player once a stopping criterion has been reached according to the variance of that player. Our stopping criterion is based on Empirical Bernstein Bounds (Maurer and Pontil, 2009), a confidence interval based on variance estimation. For t samples with observed variance σ_t and a maximum variance range of R, there is a probability of $1-\delta$ that the difference between the observed mean $\hat{\mu}$ and true mean μ abides by the following inequality formulated by Mnih et al. (2008):

$$|\hat{\mu} - \mu| \leq \sigma_t \sqrt{\frac{2\log(3/\delta)}{t}} + \frac{3R\log(3/\delta)}{t} \quad (4)$$

We stop sampling for a particular head once this bound is less than $|\mu - 0|$, meaning that we have identified the Shapley Value as positive or negative with probability $1-\delta$. This saves us significant computation while confidently separating heads into helpful and harmful buckets. For all experiments, we use $R = 1$ since the model's worst-case performance is zero and $\delta = 0.1$ to give a 95% confidence lower and upper bound.

3.3 Importance-Based Structured Pruning

Our pruning procedure works with any signed importance metric. Specifically, we test the utility of the Shapley Values metric for removing interference and helping multilingual models generalize to unseen test data.

Our hypothesis is that attention heads with negative Shapley Values introduce interference. Our pruning method reflects this by using the sign of our approximation directly. We remove all attention heads whose Shapley Value is negative with probability $1-\delta$ by the Empirical Bernstein inequality from Equation 4. This is a parameter-free approach for deciding the number of heads to preserve. This approach is consistent, with the same set of negative heads identified for pruning across 3 separate runs.

Alternatively, once Shapley Values are computed the model could be pruned to any sparsity level. Unlike prior pruning approaches besides Michel et al. (2019), we do not perform any weight updates following or during pruning and leave all parameters fixed. This provides constant time pruning to the desired size. We evaluate performance in this configurable pruning setting in 4.6.

4 Experiments

4.1 Datasets

We evaluate our methodology on the Cross Lingual Natural Language Inference (XNLI) and Universal Dependencies Part-Of-Speech (UDPOS) tasks. These allow us to analyze the applicability of Attention Head Shapley Values to both sequence classification and structured prediction. We provide a description of dataset sizes in Table 1.

Cross-Lingual Natural Language Inference (XNLI) We use the Cross Lingual Natural Language Inference (XNLI) Benchmark (Conneau et al., 2018). This dataset is aligned which allows us to control for possible confounding semantic variation in the content. Given a premise and a hypothesis and tasks, XNLI is the task of classifying

Dataset	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH
XNLI	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
UDPOS	5.4	1.7	1.1	22.4	2.8	3.1	9.5	2.7	11.3	N/A	N/A	4.8	0.5	0.8	5.5

Table 1: Size of the test sets for the datasets in thousands of sentence pairs and sentences respectively. We use a 512-example subset of the released development sets to compute Shapley Values in all languages for all datasets.

whether the hypothesis is entailed by the premise, contradicted by the premise, or neither.

Universal Dependencies Part-of-Speech (UDPOS) For structured prediction, we evaluate on the Part-of-Speech (POS) tags from the Universal Dependencies (UD) v2 corpus (Nivre et al., 2020), which has the largest cross-lingual gap in the XTREME benchmark (Hu et al., 2020). The authors suspect that structured prediction requires more language-specific knowledge than many classification tasks.

For direct comparison with our experiments on XNLI, we only retain the 13 languages from UDPOS which have a development and test split, which also exist in XNLI. Unlike XNLI, each language in UDPOS hasa different number of examples which are not aligned across languages.

4.2 Experimental Setup

As the basis for our experiments, we finetune XLM-R Base (Conneau et al., 2019) using the Transformers library (Wolf et al., 2020) on only English data. Evaluation is done using the Datasets library (Lhoest et al., 2021) implementation of the accuracy metrics. Finetuning and Shapley Value computation were both done on a single NVIDIA GeForce 12GB RTX 2080 Ti. We finetune the following hyper-parameter tuning procedures from prior work: using Hu et al. (2020) for XNLI and de Vries et al. (2022) for UDPOS.

For all tasks and languages, we use the accuracy on 512 examples of the development set as the characteristic function for our coalitional game. Our pruning baselines include the gradient-based importance metric of Michel et al. (2019) and the average of 10 randomly pruned networks. We prune the same number of heads pruned by our method for all strategies since our baselines require selecting the number of heads to prune.

4.3 Language Affinity

First, we analyze the Attention Head Shapley values for XNLI. We focus only on the role of the source language by using an aligned sample from

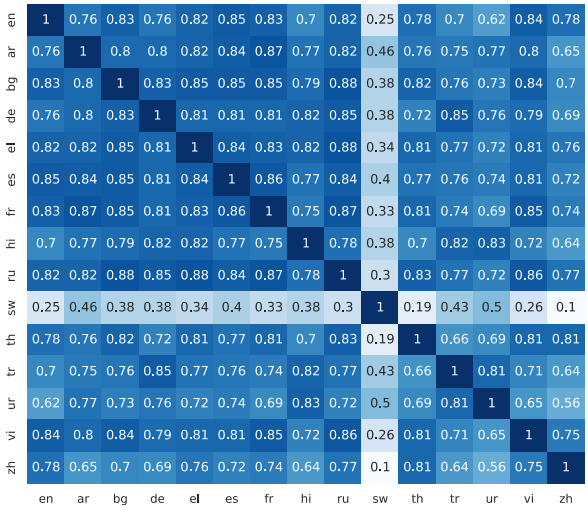

Figure 2: Spearman ρ of Attention Head Shapley Values across languages in XNLI using XLM-R finetuned on the English training split.

XNLI to control our results for differences independent from language variation. In Figure 1, we visualize the results across English, Chinese, and Swahili. As expected from prior work (Michel et al., 2019; Voita et al., 2019), many heads have low magnitude Shapley Values indicating that they play no significant role in the final network. We compare the similarity of Shapley Values learned across languages using Spearman's ρ in Figure 2 and find that Shapley Values are heavily correlated between all languages but Swahili, which is a major outlier. This cross-lingual consistency across languages is juxtaposed with inconsistency of methods that utilize finetuning as shown by Prasanna et al. (2020).

Despite this consistency, we find some attention heads demonstrate high language-specificity. Most notably, the fifth attention head in layer six is positive for Swahili but strongly negative for all other 14 languages. This indicates that this head serves a function specific to Swahili within the model. We investigate the behavior of language-specific and language-agnostic heads further in Section 5.

It is worth noting that the outlier, Swahili, is the language with the fewest number of examples in the

	XNLI Accuracy														
Pruning Strategy	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH
No Pruning	84.1	70.6	76.7	76.8	75.4	79.8	77.7	70.0	74.7	63.4	70.6	71.9	65.9	73.3	73.5
Random	81.5$^-$	67.2$^-$	72.7$^-$	72.7$^-$	71.3$^-$	75.5$^-$	73.0$^-$	66.3$^-$	70.5$^-$	63.5	67.4$^-$	68.4$^-$	61.6$^-$	69.7$^-$	70.8$^-$
Michel et al. (2019)	84.3	71.0	77.3	77.4	72.8$^-$	80.2	78.4	71.5$^+$	75.2	63.1	70.7	71.7	66.9$^+$	73.3	**77.2**$^+$
Shapley Value (φ_i)	**85.1**$^+$	**72.0**$^+$	**77.8**$^+$	**78.3**$^+$	**76.3**	**80.6**	**79.7**$^+$	**71.5**$^+$	**76.5**$^+$	**63.8**	**73.3**$^+$	**73.2**$^+$	**67.6**$^+$	**75.3**$^+$	**77.2**$^+$
Pruned Heads (K)	4	6	6	5	4	5	5	7	5	5	7	5	6	6	9

	UDPOS Accuracy														
Pruning Strategy	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH
No Pruning	**95.7**	75.1	**90.9**	**88.8**	71.5	**89.8**	81.3	73.9	**88.2**	-	-	**78.7**	67.3	**66.3**	50.2
Random	**95.7**	74.3$^-$	**90.9**	**88.8**	71.8	**89.8**	81.4	73.7	**88.2**	-	-	**78.7**	67.5	**66.3**	55.3$^+$
Michel et al. (2019)	**95.7**	75.1	**90.9**	**88.8**	71.1	**89.8**	81.1	73.8	**88.2**	-	-	**78.7**	67.3	**66.3**	48.9$^-$
Shapley Value (φ_i)	**95.7**	**76.6**$^+$	**90.9**	**88.8**	**72.8**$^+$	**89.8**	**82.6**$^+$	**75.6**$^+$	**88.2**	-	-	**78.7**	**69.5**$^+$	**66.3**	**62.6**$^+$
Pruned Heads (K)	0	4	0	0	4	0	2	2	0	-	-	0	4	0	18

Table 2: Accuracy for UDPOS and XNLI after pruning according to importance metrics. For all metrics, we remove the Bottom-K heads ($K = |\{H_i \mid \varphi_i < 0\}|$) according to that metric. $^+$ and $^-$ indicate significant ($P < 0.05$) improvement and harm by a pairwise bootstrap test. Model parameters remain fixed for all methods.

data used in the pretraining of XLM-R. Whether the large variation between Swahili and all other languages is induced by linguistic features or the training dynamics of low-resource languages within multilingual models is unclear. We leave this to be explored further in future work.

4.4 Targeted Pruning

To understand the practical applicability of the resulting Shapley Values, we evaluate models before and after pruning all attention heads with negative Shapley Values as described in Section 3.3.

Each resulting language-specific model can be represented with only the 144 mask parameters which indicate whether each attention head is removed or kept. Therefore, this pruning can be seen alternatively as a parameter-efficient learning method, using $1 \cdot 10^{-6}\%$ of the parameters it would require to finetune the model for each language[2].

XNLI In Table 2, we report the accuracy of models after targeted pruning across all languages for both XNLI and UDPOS. For XNLI, we see that targeted pruning improves performance by an average of +1.59 across all 15 languages with the maximum improvement being in Chinese (+3.78) and the minimum improvement in Swahili (+0.37). We might expect that languages closely related to our finetuning language of English would benefit less from pruning, even closely related languages such as French (+1.97) and German (+1.53) are improved significantly.

[2]144 parameters compared to $1.25 \cdot 10^8$ for full finetuning.

UDPOS Improvements in UDPOS vary to a higher degree. Only 6 out of 13 languages improve after pruning, with the rest identical with no negative Shapley Values. Surprisingly, this indicates that attention heads do not introduce interference for these languages. We hypothesize that interference for these languages may instead lie largely within the Transformer feed-forward layers, which we do not study in this work. The largest improvement is again in Chinese (+12.4) and the smallest in French (+1.3). In the case of Chinese, this is a 24.7% improvement purely by removing attention heads. Across the languages which were pruned, the average improvement is 3.4 – reducing the cross-lingual gap (Hu et al., 2020) by 0.7.

Comparison to Baselines Randomly pruning is ineffectual or harms performance in both tasks, indicating that pruning alone is not the source of our improvement. Pruning according to the gradient-based metric proposed by Michel et al. (2019) maintains rather than improves performance. This supports our hypothesis that methods that use the magnitude of gradients largely identify non-impactful heads as opposed to harmful heads.

4.5 Zero-Shot Pruning

Given the high rank correlation between many of the languages, we evaluate transferability by using the Shapley Values for English to prune the model for all languages. We report results in Table 3.

XNLI On XNLI, surprisingly, this transferred pruning across languages has similar benefits to our targeted pruning results despite only being learned

Pruning Strategy	EN	AR	BG	DE	EL	ES	FR	HI	RU	SW	TH	TR	UR	VI	ZH
No Pruning	84.1	70.6	76.7	76.8	75.4	79.8	77.7	70.0	74.7	**63.4**	70.6	71.9	65.9	73.3	73.5
Random	81.7$^-$	67.1$^-$	72.3$^-$	72.9$^-$	71.1$^-$	75.1$^-$	73.5$^-$	65.7$^-$	71$^-$	60.7$^-$	67$^-$	68.3$^-$	61$^-$	69.7$^-$	70.7$^-$
Michel et al. (2019)	84.3	70.3	76.7	77.1	75.9	80.1	77.9	70.1	75.1	62.9	71.6	72.5	66.1	74.7$^+$	74.5$^+$
Shapley Value (φ_i)	**85.1**$^+$	**72.0**$^+$	**77.8**$^+$	**79.4**$^+$	**76.3**	**80.6**	**79.7**$^+$	**71.5**$^+$	**76.5**$^+$	63.3	**73.1**$^+$	**73.1**$^+$	**68.4**$^+$	**75.2**$^+$	**76.3**$^+$

Table 3: Accuracy for XNLI after pruning using importance metrics from English. For all metrics, we remove the Bottom-K heads ($K = |\{H_i \mid \varphi_i < 0\}|$) according to that metric. $^+$ and $^-$ indicate significant ($P < 0.05$) improvement and harm by a pairwise bootstrap test.

for English. Two languages (Urdu and German) achieve better results in the zero-shot pruning than they did in the targeted pruning, five achieve worse results, and the remaining eight are equivalent.

It is likely that the strength of zero-shot transfer is largely due to the removal of the fifth head of layer six, which is one of the top 2 most negative heads for all languages barring Swahili. Interestingly, the Attention Head Shapley Values for Swahili also have the lowest rank correlation with English of any language.

UDPOS However, UDPOS highlights the major shortcoming of zero-shot pruning: all attention heads receive a positive Shapley Value for English for UDPOS. This means that no zero-shot pruning is performed despite targeted pruning finding benefits for languages shown in Table 2.

4.6 Iterative Pruning of Attention Heads

Finally, we evaluate the effectiveness of Shapley Values as a ranking methodology for the iterative pruning evaluation performed by Michel et al. (2019). Iterative pruning evaluates how well each importance ranking captures the combinatorial effects of removing attention heads at different compute budgets. We compare random pruning, the gradient-based approach from Michel et al. (2019), and Shapley Values computed through plain Monte Carlo simulation and Shapley Values using Truncation and Multi-Armed Bandit optimization (TMAB). We plot results in Figure 3.

Averaged across all levels of sparsity, our method outperforms the Random baseline (+5.8), Monte Carlo Shapley Values (+1.6), and the Gradient baseline (+0.6). At different stages. Depending on the target sparsity of interest however, Shapley Values and Gradient-based pruning have different levels of sparsity. Our method is the only method that identifies strongly harmful heads, with performance improving compared to the unpruned model for the first 6 heads removed. Our method achieves

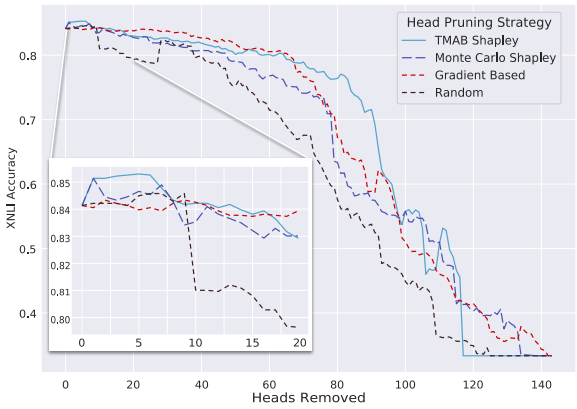

Figure 3: Evolution of XNLI Accuracy as Heads are removed according to different pruning strategies.

the largest performance gap at 44% of model capacity outperforming the Gradient baseline, Monte Carlo Shapley Values, and the Random Baseline by +12.2, +15.1, and +20.9 respectively. However, the gradient baseline outperforms our method when more than 80% of heads are pruned, although neither method performs well above chance at this sparsity.

5 Qualitative Attention Analysis

In order to provide intuition into the function of attention heads, prior work has turned to attention visualization as the basis for qualitative analysis of the inner workings of transformer models. Clark et al. (2019) and Hoover et al. (2020) both explore patterns within attention heads.

We visualize the attention patterns of outlier attention heads using BertViz (Vig, 2019) from our model to give a qualitative understanding of the attention head patterns associated with language-agnostic and language-specific heads.

5.1 Language-Agnostic Heads

We define the set of language-agnostic heads as the intersection of the the top 20 attention heads for each language. In Figure 4, we visualize the

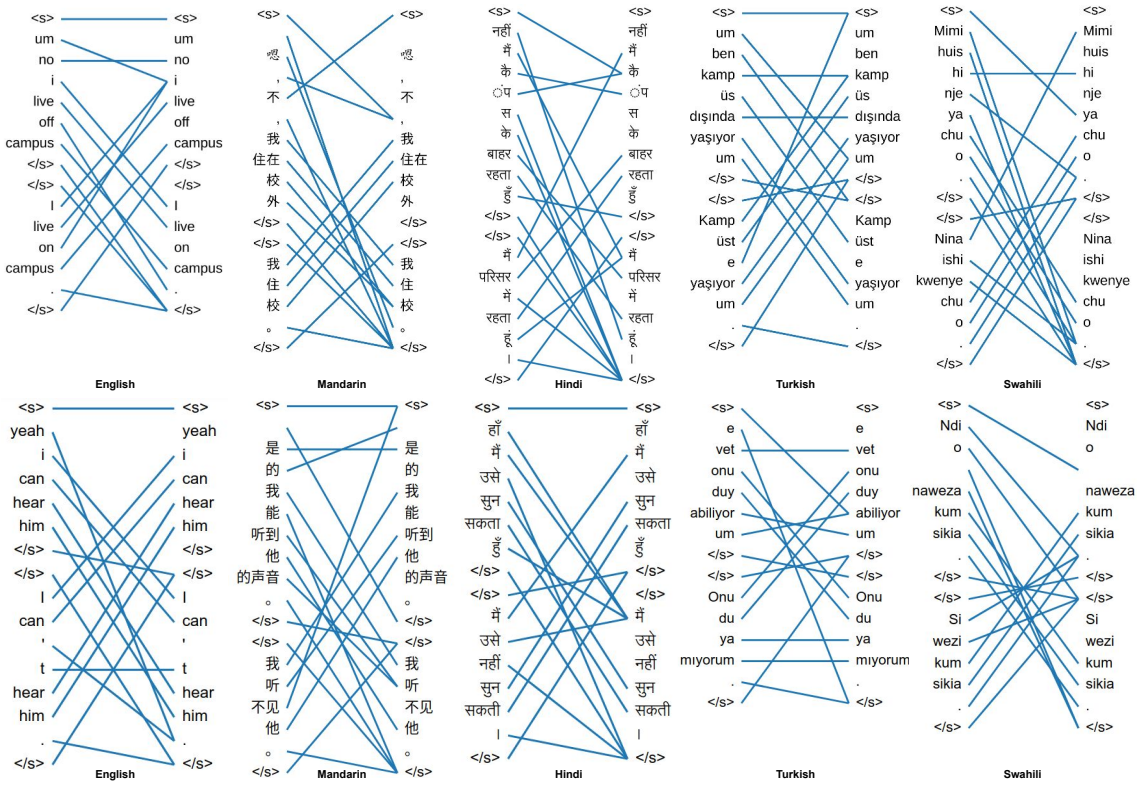

Figure 4: Attention of Layer 2, Head 9 of our XNLI model which is identified as language-agnostic. The attention pattern links synonyms in the premise and hypothesis for all languages. For clarity, we connect the left token to the token on the right which receives the largest attention weight.

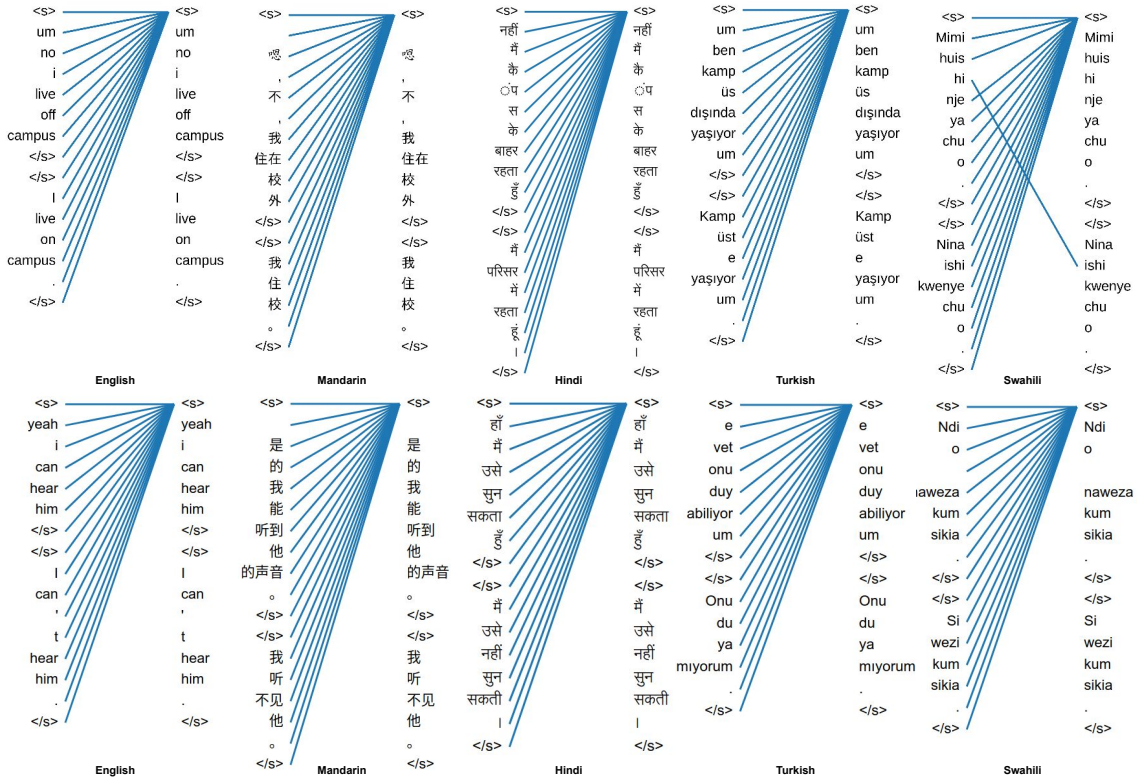

Figure 5: Attention of Layer 6, Head 5 of our XNLI model which is identified as language-specific to Swahili. Unlike the language-agnostic head, there is no obvious pattern in visualisation. However, in Section 5.2 we measure a significant difference in Swahili's attention to separator tokens compared to other languages.

attention pattern of the highest-ranked of the 4 heads which meet this criterion. The visualization highlights the same attention pattern across all languages: words from the premise are matched to near-synonyms in the hypothesis and vice versa.

The synonym-matching pattern clearly applies to NLI, where synonyms critically participate in commonalities and contradictions between the premise and hypothesis. Synonym linking is possible via token semantics and the separator tokens, so this pattern does not require any knowledge of language-specific syntax or morphology.

The visualization reveals a meaningful language-agnostic pattern which may explain why the positive Shapley Value across all languages. This usage highlights that while we utilize Shapley Values to remove harmful learned patterns, they also can direct mechanistic interpretability work to understand the effectiveness of transformers for a particular task (Wang et al., 2022).

5.2 Language-Specific Heads

As highlighted in Section 4.3, the fifth head of layer six has a positive Shapley Value only for Swahili. In Figure 5, we see that this head sometimes exhibits unique behavior for Swahili, connecting the incorrectly tokenized "ishi" suffix of "*Mimi Huishi*" and "*Ninaishi*" meaning "*I live*" in the Habitual and Present tense respectively. However, this use is not found frequently in our Swahili examples, as shown in the second example.

Therefore, we aim to understand whether the head functions in a measurably different fashion for Swahili across our entire dataset rather than on specific examples. Using the hypothesis from Clark et al. (2019) that attention to separator tokens indicates an inapplicable learned pattern, we look at the percentage of sentences where all tokens attend primarily to separators. This criterion is true in 56% Swahili XNLI inputs, but only 41% of non-Swahili inputs on average ($\sigma = 4.3\%$).

The frequency of separator attention combined with the minimal negative performance impact from removing this head for Swahili in Section 4.5 supports the idea that this head supports a rare pattern, perhaps stemming from poor tokenization. However, the relatively low rate of separator attention indicates that this head does impact other languages often, introducing noise.

6 Conclusions & Future Work

In this work, we developed a simple yet effective approach to measure the impact of individual attention heads on task performance by leveraging Shapley Values. We used this to identify language-specific and language-agnostic structural components of multilingual transformer language models. We demonstrated that the resulting values exhibit language affinity, varying across languages. We then applied these Attention Head Shapley Values to improve cross-lingual performance through pruning for both sequence classification and structured prediction. Finally, we performed provided insights on language-agnostic and language-specific attention heads using attention visualization.

We believe that attention head Shapley Values have strong potential to systematically inform future studies of multilingual models and transformers broadly. Future work should explore the relationship between linguistic features, training data volume, and the language-specificity of attention heads. Additionally, the benefits of removing heads motivates work that reduces cross-lingual interference introduced by language-specific components during pre-training, such as pruning during pre-training or utilizing sparsely activated networks.

7 Limitations

Even with our optimizations, using Shapley Values as an importance metric requires a significant computational cost compared to gradient-based methods: gradient-based methods take approximately 3.33e14 FLOPs and our optimized Shapley Values take approximately 3.27e16 FLOPs to converge. While the computation is parallelizable, it took several days on a single GPU to compute accurate estimates. This expense is reasonable for understanding the behavior of base models more deeply but limits the use of this method as a rapid iteration tool. For those looking to reduce this computational cost further, we recommend first using gradient-based methods to identify a set of heads to which the output is sensitive and then using Shapley Values to interpret the direction of the effect. While this may miss some harmful heads, it is likely to find the most harmful heads for a reduced cost.

Additionally, we rely on analysis of attention patterns to help ground our findings. However, there is debate as to whether analysis of attention patterns is a sound analytical tool (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

8 Acknowledgements

We are thankful to Caleb Ziems, Chris Hidey, Yanzhe Zhang, Hongxin Zhang, and our anonymous reviewers for their feedback. This work is supported in part by Cisco, an Amazon Faculty Research Award, and NSF grant IIS-2144562.

References

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. XAI for transformers: Better explanations through conservative propagation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained bert networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 15834–15846. Curran Associates, Inc.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR.

Amirata Ghorbani and James Y Zou. 2020. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems*, volume 33, pages 5922–5932. Curran Associates, Inc.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. 2019. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *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 1956–1966, Online. Association for Computational Linguistics.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Andreas Maurer and Massimiliano Pontil. 2009. Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 62–72.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. 2008. Empirical bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 672–679, New York, NY, USA. Association for Computing Machinery.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Lloyd S. Shapley. 1953. *A value for n-person games*, page 31–40. Cambridge University Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. Association for Computational Linguistics.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1513–1528. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.