EXPLAIN THIS, PRUNER! The Effect of Zero-Order Pruning on LLM Explainability and Curvature

Joseph Bejjani, Camilo Brown-Pinilla, David Ettel
Harvard College '26

Large Language Models (LLMs) excel in language understanding and generation tasks but have significant memory and computation requirements. In addition, the size and complexity of LLMs pose challenges in XAI, an emerging field in ML concerned with the problem of explaining how a model arrives at its outputs. Model compression techniques such as pruning can be effective in reducing resource requirements and enabling more efficient inference in downstream tasks. However, it is not well understood if and how pruning of LLMs affects their explainability. Our work investigates this open problem. We identify faithfulness of explanations as a necessary metric in determining a model's explainability. We then evaluate the faithfulness of SHapley Additive exPlanations (SHAP) and Integrated Gradients (IG) explanations of variously pruned and non-pruned DistilBERT and RoBERTa models trained on the IMDb and Yelp Polarity datasets for binary sentiment classification. We find that while magnitude-based pruning does not significantly affect explanation faithfulness, random pruning can degrade explainability. Furthermore, our results indicate that explainability is primarily influenced by model architecture. We investigate the underlying geometry of the models to explain our results and find that depending on pruning method and target sparsity, high-curvature regions can emerge, potentially undermining explanation faithfulness. Our code is available at https://github.com/camilobrownpinilla/Explain-This-Pruner.

Introduction

As our ability to compute has, and continues, to dramatically increase thanks to Moore's law, there has been an increased interest in machine learning, the branch of computer science that allows us to learn complex patterns from data. Among the many tools stemming from this rich field, Large Language Models (LLMs) are arguably the most powerful and well known. LLMs are trained on vast amounts of text in order to understand and generate natural language. These models can answer questions, summarize information, and assist across a wide variety of language-based tasks, proving to be a useful tool to humanity at large.

As LLMs grow in size, they have become the target of model compression techniques aiming to reduce computational demands while preserving model performance. In particular, recent work has focused on pruning, the class of methods involving the removal a subset of network parameters according to precise criteria, leaving a sparse model with more manageable resource requirements and minimal accuracy degradation (Kwon et al., 2022; Sun, Liu, Bair, & Kolter, 2024; Dery et al., 2024; J. Li, Dong, & Lei, 2024; Ma, Fang, & Wang, 2023; Frantar & Alistarh, 2023; Kurtic et al., 2022).

With LLM-usage becoming more prevalent—accelerated in part by model compression techniques making them more resource-friendly—the challenge of explaining models has become more pressing (Zhao et al., 2023). That is to say, while we can train models to produce very accurate outputs, we do not know precisely how these outputs are produced. Being able to explain how an LLM arrives at a particular output has important implications for uncovering and debugging model bias, building user trust, and enabling transparency in decision-making. Accordingly, the field of eXplainable AI (XAI) has emerged, seeking to address these concerns, with recent work in the XAI literature focusing on developing and evaluating methods for explaining LLMs. Of the many explanation methods proposed, feature-based attribution methods refer to the class of methods that seek to explain model outputs in terms of the inputs to the model (Sanyal & Ren, 2021; Volkov & Averkin, 2024; Hao, Dong, Wei, & Xu, 2021)(Enguehard, 2023; Lyu, Apidianaki, & Callison-Burch, 2024).

Despite the focus on each issue in isolation, the relationship between LLM-pruning and LLM-explainability has not received much attention in the literature. As model compression techniques improve and become more widely used before model deployment, it is important to understand if and how the explainability of the deployed model is affected.

By investigating the effect of pruning on the explainability of LLMs, our work aims to make progress in building a bridge between developments in model compression and XAI. We hypothesize that by reducing model complexity through the extraction of high-performing subnetworks, pruning yields models whose prediction function has lower curvature, increasing LLM explainability.

We test our hypothesis through experiments with two closely related encoder-only models, DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020) and RoBERTa (Liu et al., 2019), trained on the IMDb (Maas et al., 2011) and Yelp Polarity (Zhang, Zhao, & LeCun, 2015) datasets for sequence classification. Following XAI literature, we employ SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and Integrated Gradients (IG) (Sundararajan, Taly, & Yan, 2017), feature-attribution-based explanation methods, and identify the faithfulness of explanations as a necessary condition for a model to be explainable (Lyu et al., 2024). We choose these models, datasets, and explanation methods because of their ubiquity in the literature and in practical applications. As both models are variants of BERT, their architectural similarities allow us to isolate the effects if pruning on explainability more precisely than what would be possible with models that differ more substantially. Additionally, the relatively small size of these models (on the order

of 100 million parameters) lets us run several experiments that would otherwise be prohibitively expensive.

We select zero-order pruning methods, including unstructured and structured magnitude-based pruning and random pruning, and prune the models with each method to varying degrees of sparsity. We evaluate the faithfulness of explanations of each pruned subnetwork against the faithfulness of explanations of the unpruned network. We also evaluate each pruned subnetwork against a network of equivalent size that is trained from a random initialization—this gives insight into whether the particular pruning method affects explanation faithfulness, or if effects on explainability should be attributed primarily to the reduction in network size.

Finally, we investigate the local geometry of each model to understand how pruning affects the faithfulness of local explanation methods. We analyze network geometry through the lens of local curvature, since SHAP operates on the assumption of local linearity[1]. In particular, we approximate the global average of the local curvature for training samples. We estimate the local curvatures using an approximation of the Hessian diagonals computed through a variation of Hutchinson's trace estimator (Yao et al., 2021).

We find that magnitude-based pruning does not significantly affect explanation faithfulness, and importantly, does not hurt explainability while maintaining test accuracy comparable to an unpruned model. However, we find that Random Unstructured pruning can degrade faithfulness of explanations and argue that this occurs due to the emergence of high-curvature regions that violate linearity assumptions of the explanation methods.

Methods

In this section, we describe and motivate the models, datasets, methods, and metrics used in our experiments. We then detail our approach.

2.1 Models

We conduct experiments using DistilBERT and RoBERTa, high-performing language models based on the transformerencoder architecture (Devlin, Chang, Lee, & Toutanova, 2019)(Liu et al., 2019). These models are commonly used in related literature for benchmarking of language modeling tasks and are prevalent in practical applications for Natural Language Processing (NLP), making them suitable choices for studying the topic of this work.

2.2 Datasets

We train and evaluate our models on the IMDb (Maas et al., 2011) and Yelp Polarity (Zhang et al., 2015) datasets for the task of binary sentiment classification. These datasets are commonly used in conjunction with DistilBERT and RoBERTa in the literature for experimentation in the NLP domain.

2.3 Pruning Methods

We prune our trained models using Random Unstructured, L1 Unstructured, and L1 Structured pruning. Random Unstructured pruning removes a random subset of parameters constituting a specified percentage of the network. The latter two are magnitude-based methods, which are the foundation of both classic and more recent pruning techniques (Han, Pool, Tran, & Dally, 2015; Sun et al., 2024). For example, pruning 80% of a network with L1 Unstructured pruning removes the smallest 80% of parameters ordered by L1-norm. In contrast, L1 Structured pruning removes entire channels with the lowest L1-norm.

2.4 Explanation Methods

We generate explanations of the model's sentiment classifications using the feature attribution methods SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017). These methods assign importance scores to input features for a particular model prediction.

SHAP computes feature attributions based on Shapley values from cooperative game theory, approximating each feature's marginal contribution to the prediction by considering different subsets of features (Lundberg & Lee, 2017). IG approximates the integral of gradients of the model's output with respect to the input features along the straight-line path from a baseline input to the given input (Sundararajan et al., 2017). Following standard practice, we use the zero vector as the baseline input.

We follow previous work in adopting these methods for explaining language models. For example, Mosca et al. review SHAP-based methods applied to NLP tasks (Mosca, Szigeti, Tragianni, Gallagher, & Groh, 2022). Hao et al. (Hao et al., 2021) and Janizek et al. (Janizek, Sturmfels, & Lee, 2020) adapt IG to explain token importance in sentiment classification.

2.5 Geometry

To assess the local geometry of the functions represented by the models of interest, we consider the average local curvature of the logit function with respect to the embeddings. For each model, we compute the local curvature around the predictions for 10% of the training set and take the average.

Computing second order derivatives for machine learning models with high parameter count is prohibitively expensive (Elsayed, Farrahi, Dangel, & Mahmood, 2024). Therefore, we adopt compute-efficient approximations of local curvature. There is a consensus in the literature that the Hessian diagonal is a good proxy for this task (Elsayed et al., 2024)(Yao et al., 2021). In particular, we compute an approximation of the diagonal of the Hessian around a given prediction using a modification of Hutchinson's method for trace estimation, modelled after the implementation by Yao et al. (Yao et al., 2021). This method makes use of the fact that

$$\text{diagonal}(H) = \mathbb{E}[v \odot (Hv)]$$

where $\odot$ is pointwise multiplication and the expectation can be taken over a Rademacher or Gaussian distribution (Meyer & Avron, 2023). We choose the latter. The expectation is approximated by averaging over vectors v randomly drawn from the distribution.

Given the approximate Hessian diagonals, we then proceed with averaging them over a 10% subset of the training set.

We further distill the resulting average Hessian diagonal by looking at the largest absolute value achieved by its elements as well as their average absolute value, as the diagonals themselves are too large for human interpretation.

[1] We do not discuss magnitudes of gradients in the body of this paper, as we do not expect SHAP nor IG to be influenced by them. We did, however, collect gradient-magnitude data and find that gradient magnitude is not correlated with faithfulness of explanations. The data can be found in figure 13.

2.6 Evaluation Metrics

We use the following evaluation metrics to study the effect of pruning on model explainability.

Accuracy. We consider the accuracy of each model variation on the test dataset in order to verify the usefulness of the pruned models; in particular, we aim to investigate whether pruning affects explainability while avoiding sacrifices in model accuracy. In a practical application, a highly pruned model would not be particularly useful if it suffered from significant accuracy degradation, even if it saw improvements in explainability.

Faithfulness. Following Lyu et al., we identify faithfulness as the most important principle for evaluating an explanation (Lyu et al., 2024). Accordingly, we use the faithfulness of explanations as a metric for the explainability of our models.

Faithfulness is the degree to which an explanation accurately reflects how a model made a prediction; an unfaithful explanation, then, does not accurately describe a model's decision-making process and therefore is not much of an explanation (Lyu et al., 2024). In the context of feature attribution explanations, faithfulness refers to the extent to which an explanation correctly captures which features of an input the model uses to generate its corresponding output (Lyu et al., 2024).

Previous work has proposed various methods for measuring the faithfulness of explanations by determining how well-aligned feature attributions are with true model behavior. Two faithfulness metrics frequently used in the XAI literature are Infidelity (INFD) (Yeh, Hsieh, Suggala, Inouye, & Ravikumar, 2019) and Faithfulness Correlation (FCor) (Bhatt, Weller, & Moura, 2020). These metrics rely on perturbing input features, measuring corresponding changes in model output, and comparing these changes to the importance scores of the perturbed features.

Yeh et al. (Yeh et al., 2019) define Infidelity as

$$\mathrm{INFD}(\Phi, f, \mathbf{x}) = \mathbb{E}_{\mathbf{I}\sim\mu_I}\left[\left(\mathbf{I}^T\Phi(f,\mathbf{x}) - (f(\mathbf{x}) - f(\mathbf{x}-\mathbf{I}))\right)^2\right].$$

Here, f is a black-box model, and Φ is an explanation functional. $\mathbf{I} \in \mathbb{R}^d$ is a random variable, where μ_I represents input perturbations of interest. A typical perturbation $\mathbf{I}$ is to replace a feature in $\mathbf{x}$ with some baseline value, such as 0. For a faithful explanation, we would expect the model output to change by an amount proportional to the sum of the importance scores of the perturbed features (Decker, Bhattarai, Gu, Tresp, & Buettner, 2024).

Bhatt et al. (Bhatt et al., 2020) measure faithfulness with correlation. For a model f, explanation functional Φ, input $\mathbf{x} \in \mathbb{R}^d$, baseline value $\bar{\mathbf{x}}_s$, and subset size $|S|$, FCor defines the faithfulness of Φ to f at $\mathbf{x}$ as

$$\mathrm{FCor}(f, \Phi, \mathbf{x}) = \mathrm{corr}\left(\sum_{S\in\binom{[d]}{|S|}} \Phi(f,\mathbf{x})_i,\ f(\mathbf{x}) - f\left(\mathbf{x}_{[\mathbf{x}_s = \bar{\mathbf{x}}_s]}\right)\right)$$

Here, corr is the Pearson correlation. Faithful explanations should have an FCor close to 1, indicating a strong positive correlation between the attribution scores given by Φ for an input $\mathbf{x}$ and the changes in the predictions of f under corresponding perturbations to $\mathbf{x}$.

We select the FCor metric because it provides an interpretable score on a standard scale in the range [–1, 1], facilitating comparison across explanation method.

2.7 Our Approach

Model Generation. Given a model architecture and target sparsity, we randomly initialize a model to serve as the unpruned, 'base' network. We then create a second model with its own random initializations and perform Random Unstructured pruning to the target sparsity. This second model serves as an independent, 'smaller' network.

We choose to start from random initializations to simulate the process of constructing and training a model from end to end. Additionally, starting with random initializations allows us to create and test the 'smaller' model; starting from pretrained weights would make the 'smaller' model just a pruned version of 'base,' rather than a smaller version with an independent set of parameters.

After the initialization of 'base' and 'smaller,' we train both for 3 epochs on the given dataset. Then, for each pruning method, we make a copy of 'base' and prune the parameters to the target sparsity. Following standard practice, we train each pruned model for an additional epoch to allow for accuracy recovery.

FCor Approximation. Following training and pruning, we evaluate the explainability of each model as follows. Due to resource constraints, we select 3% of the test dataset and generate an explanation for the model's output on each test sample. For each test sample $\mathbf{x}$, we compute an FCor value. We first fix $|S|$ in Eq. 3 with a hyperparameter k. We then randomly sample 100 subsets $\mathbf{x}_s$ of size k from $\mathbf{x}$ and set them to the [MASK] token. This gives an estimate of FCor for that data point, since we do not see all $\binom{[d]}{|S|}$ subsets in the calculation, which is computationally prohibitive for long input sequences. In future work, we plan to evaluate faithfulness over more samples and with more perturbations per sample in order to achieve more accurate approximations of faithfulness.

	DistilBERT		RoBERTa	
	IMDb	Yelp	IMDb	Yelp
base	88.93	94.10	87.33	93.68
40pct	89.08	93.99	86.65	93.65
60pct	88.72	93.90	74.76	93.54
80pct	88.06	93.85	73.37	93.36

Figure 1. Test accuracies(%) across sparsities. Accuracies of pruned models are averaged across the RandUnstruct, L1Unstruct, and L1Struct methods. Accuracies of base models are averaged across the three 'base' models trained for each model, dataset combination.

We then take the mean of the local FCor estimates, giving a measure of the average faithfulness of explanations for that particular combination of model, dataset, explanation method, and k.

We select the [MASK] token as our baseline value for input perturbations during FCor computation because it aligns with DistilBERT's pre-training objective of masked language modeling (Devlin et al., 2019). Moreover, the SHAP library uses the [MASK] token when perturbing inputs to estimate feature importance, so it is a natural choice for faithfulness evaluation (Lundberg & Lee, 2017).

k **Hyperparameter**. We ran experiments with different *k* to study whether masking different numbers of tokens at a time has an effect on FCor; in particular, we were interested in observing whether token importance scores are independent, or if explanations vary in their faithfulness when we consider groups of tokens and their summed importance. Initial tests showed negligible differences in FCor scores across $k = 1, 2, 3, 5$, suggesting that explanations are similarly faithful for each of these feature subset sizes. Future work will consider greater k to study if faithfulness is impacted. We fix $k = 3$ for our remaining experiments, due to resource constraints.

Experiments and Results

We conduct experiments on DistilBERT and base RoBERTa, trained on IMDb and Yelp Polarity. We prune each model using Random Unstructured, L1 Unstructured, and L1 Structured pruning, to sparsities of 40%, 60%, and 80%. We elect these percentages to balance computational constraints and the minimal effects initially observed when pruning under 40%.

We evaluate the faithfulness of SHAP and IG explanations for each model using FCor with $k = 3$ and plot the distribution of the scores across test samples. We plot the distribution of FCor scores for SHAP on all different models and report the average FCor scores in **Figure 2**. We find no significant pattern in our results, but note that scores were much more consistent across all DistilBERT experiments compared to RoBERTa, indicating that explainability may rely more on model architecture than sparsity or training data.

To quantify the global average of a model's local curvature, we use 10% of its training data to approximate the Hessian Diagonal for each sample using the variation of Hutchinson's trace estimation described in 2.5. For each sample, we use 3 directional vectors drawn from a standard normal distribution, to reduce computational costs. When computing the maximum value along this diagonal, we consistently find extreme outliers among the models pruned via random unstructured pruning, indicating this method may produce regions of high curvature in the underlying model's geometry. Further, we average over many test samples to help mitigate variance in local approximations caused by a low number of directional vectors (**Figure 4**). We note that, in this case, randomly pruned models still result in the largest values. Increasing the number of directional vectors may result in a more accurate estimation for each sample, and is left for future work.

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb | 40pct	**0.45**	0.35	0.43	0.43	0.42	0.16	-0.07	**0.33**	0.27	0.03
IMDb | 60pct	**0.44**	0.41	**0.44**	0.43	0.38	**0.32**	0.09	0.14	0.13	-0.19
IMDb | 80pct	0.43	0.31	**0.44**	**0.44**	0.42	0.19	0.24	0.22	-0.07	**0.30**
Yelp | 40 pct	**0.63**	0.55	0.62	**0.63**	**0.63**	0.45	0.46	0.32	0.47	**0.49**
Yelp | 60 pct	0.64	0.60	0.59	**0.65**	0.61	0.47	0.32	0.50	**0.54**	0.34
Yelp | 80 pct	**0.62**	0.59	0.59	0.61	**0.62**	0.46	-0.13	**0.47**	0.33	0.00

Figure 2. Average FCor Scores (**max** in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb | 40pct	0.81	12.68	1.02	**0.80**	49.68	**0.36**	3.86	0.44	0.42	7.86
IMDb | 60pct	**0.75**	15.99	1.19	1.28	43.52	0.56	**0.06**	0.80	1.36	41.86
IMDb | 80pct	**0.83**	739.25	1.06	1.09	184.78	0.50	**0.03**	0.58	0.80	0.28
Yelp | 40 pct	**0.96**	17.11	1.20	1.11	38.36	2.26	4.95	3.29	**0.67**	10.67
Yelp | 60 pct	1.37	24.28	**1.24**	1.28	13.69	1.01	0.92	1.29	**0.49**	6.18
Yelp | 80 pct	1.53	61.92	1.23	**0.80**	72.80	0.73	14090.04	0.66	1.22	**0.08**

Figure 3. Maximum absolute value of Hessian Diagonal (**min** in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb | 40pct	**0.07**	0.08	0.09	**0.07**	0.08	0.03	0.04	0.03	**0.02**	0.06
IMDb | 60pct	0.07	0.10	0.08	**0.05**	0.12	0.04	**0.00**	0.05	0.03	0.28
IMDb | 80pct	0.08	0.22	**0.06**	**0.06**	0.18	0.03	**0.00**	0.04	0.02	**0.00**
Yelp | 40 pct	0.05	**0.04**	0.05	**0.04**	**0.04**	0.04	**0.03**	0.07	**0.03**	**0.03**
Yelp | 60 pct	0.05	**0.04**	0.05	**0.04**	0.05	0.03	**0.02**	0.03	**0.02**	0.03
Yelp | 80 pct	0.05	0.06	0.04	**0.03**	0.05	0.03	13.72	0.02	0.02	**0.00**

Figure 4. Average absolute value of Hessian Diagonal (**min** in bold, values rounded to 2 decimal places).

Discussion

We discuss the results of our experiments, highlight key findings, and explain our findings in terms of local geometry.

We observe in all our experiments that magnitude-based pruning does not significantly degrade accuracy on the test set. This agrees with the pruning literature, which has found that a prune-retrain approach can achieve high levels of sparsity with competitive accuracy (Han et al., 2015; Frankle & Carbin, 2019).

4.1 Standard IG is ill-suited for the language domain

IG assumes that **(1)** input features are independent, and **(2)** there exists a zero-information baseline from which a straight-line path integral yields faithful feature attributions (Sundararajan et al., 2017). IG was developed with the image classification domain in mind, where it was designed to operate on pixels as input features. In this domain, assumptions **(1)** and **(2)** are intuitively reasonable due to the continuous nature of image data. However, these assumptions do not hold for language models: language is inherently context dependent, invalidating **(1)**; and **(2)** does not hold because points interpolated along a straight-line path from, say, a baseline zero-vector to the input embedding cannot be assumed to represent valid text data, since the word embedding space is discrete (Sanyal & Ren, 2021). Indeed, our experiments give evidence for the unfaithfulness of IG explanations for language models: across all models, pruning methods, and sparsities, the distribution of FCor scores of IG explanations are approximately normal with mean 0 (**Figures S1, S2, S3, S4**). This implies that there is no correlation between the IG-assigned token importance scores and the actual behavior of the model.

While assumption **(1)** is unavoidable due to the nature of natural language, variations of IG have been developed to correct **(2)** for language settings by considering semantically plausible non-linear paths from a baseline embedding to the input embedding (Enguehard, 2023; Sanyal & Ren, 2021).

We observe significantly higher FCor scores for SHAP explanations, suggesting that the failure of assumption **(2)** underlies IG's unfaithfulness. SHAP still assumes **(1)** (Lundberg & Lee, 2017), but the FCor scores indicate a moderate to strong positive correlation between claimed feature importances and model behavior (**Figure 2**). We leave further investigation of these claims and evaluation using improved IG methods (Enguehard, 2023; Sanyal & Ren, 2021) to future work.

4.2 Magnitude-based pruning does not affect explainability, but Random Unstructured pruning may hurt it

Our results do not show a significant effect of pruning on faithfulness of explanations for magnitude-based methods. In our experiments (**Figures 5, 6, 8**), we see that for a particular choice of architecture and dataset, the distribution of FCor scores does not vary significantly between the 'base' model and the models pruned with L1Unstructured and L1Structured methods. A notable exception is RoBERTa trained on IMDb (**Figure 7**), which we discuss further below. Moreover, the data do not show a consistent relationship between target sparsity and FCor score with remaining variables held constant (**Figure 2**), suggesting that changes in explanation faithfulness are primarily due to other factors, particularly model architecture and dataset.

However, we observe that explanations of Random Unstructured pruned models generally underperform in faithfulness. In particular, Random Unstructured pruning never gives the highest FCor for a particular model, dataset, and sparsity. Moreover, it gives the lowest FCor of the pruning methods in all but 3 experiments, where RandUnstruct has the second lowest FCor by only 0.01-2 (**Figure 2**). These findings suggest that Random Unstructured pruning may negatively affect model explainability by undermining faithfulness of explanations. To understand why Random Unstructured pruning may negatively affect model explainability, we consider the effect of pruning on the local geometry of a model's decision function.

4.3 Random Unstructured pruning creates highly curved regions

We observe that Random Unstructured pruned models have the largest Maximum Absolute Value of Hessian Diagonal (MAVHD) across the pruning methods for all but 3 experiments (**Figure 3**). Furthermore, the average MAVHD across all Random Unstructured pruned models is about 1069x and 1323x the average MAVHD for L1Unstructured pruned models and L1Structured pruned models, respectively. Note the outliers 739.25 and 14090.04 in the entries for DistilBERT-IMDb-80pct and RoBERTa-Yelp-80pct, respectively. We discuss a possible explanation below.

These findings suggest that Random Unstructured pruning destroys local linearity of the models' underlying functions. A pruned model can be imagined as a fewer-parameter approximation of a base function. Removing weights at random has the potential to significantly modify the geometry of the function, creating regions with jagged decision boundaries and increased local curvature. By contrast, magnitude-based methods prune the weights that contribute least to the output, reducing the capacity for altering the behavior of the function. We hypothesize that, over models with large numbers of parameters (~ 100M), there is a low probability of creating a high-curvature region through random pruning, resulting in a similar average local curvature despite lower explainability due to a few diabolical regions.

The presence of such highly curved regions can be verified by considering the MAVHD. The data show that MAVHD for RandUnstruct is substantially higher than other methods in most cases, despite the average value being very similar (**Figures 3, 4**).

The data also suggest that the probability of high-curvature regions emerging depends on the target sparsity, with 80% Random Unstructured pruning resulting in extreme MAVHD values in some cases.

We also find that the 'smaller' models have large maximum local curvature. Recall that we create the smaller model by randomly removing weights in an initialization independent from the 'base,' running the same risk of creating curved regions as the RandUnstruct method. However, the smaller model trains for longer at that level of sparsity and therefore has more opportunities to smooth curved regions during training. The data reflect this, as the largest MAVHD for the smaller models is substantially less than that of the random unstructured models (14090.04 vs 184.78).

Precisely characterizing the mechanism by which random pruning produces sharply curved regions is an interesting direction for future work.

4.4 Highly curved regions make SHAP less faithful

The SHAP explanation method operates on the assumption of a locally linear model (Lundberg & Lee, 2017). A high MAVHD indicates the presence of a region with high local curvature,

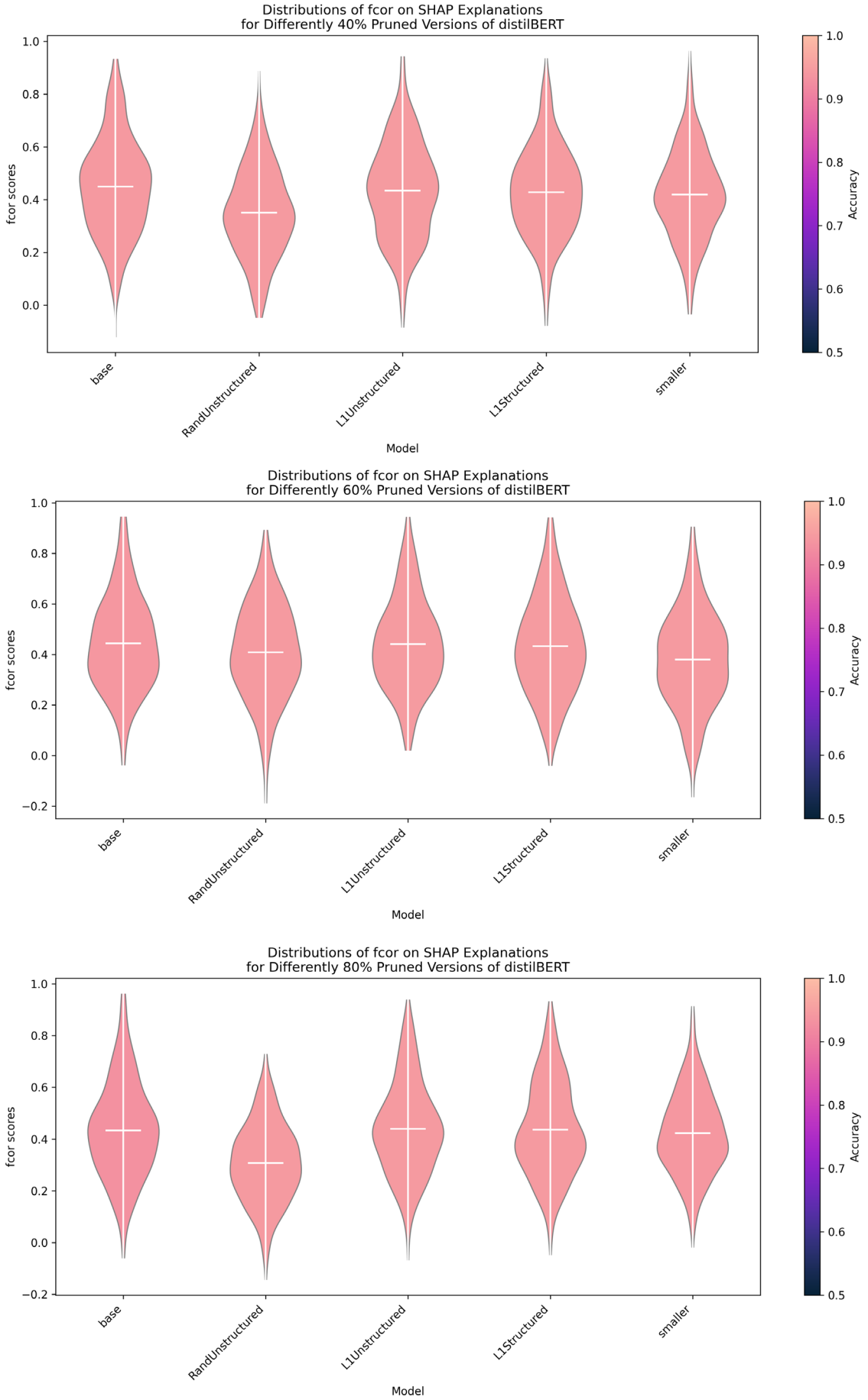

Figure 5. Distributions of FCor scores for SHAP on DistilBERT on IMDb.

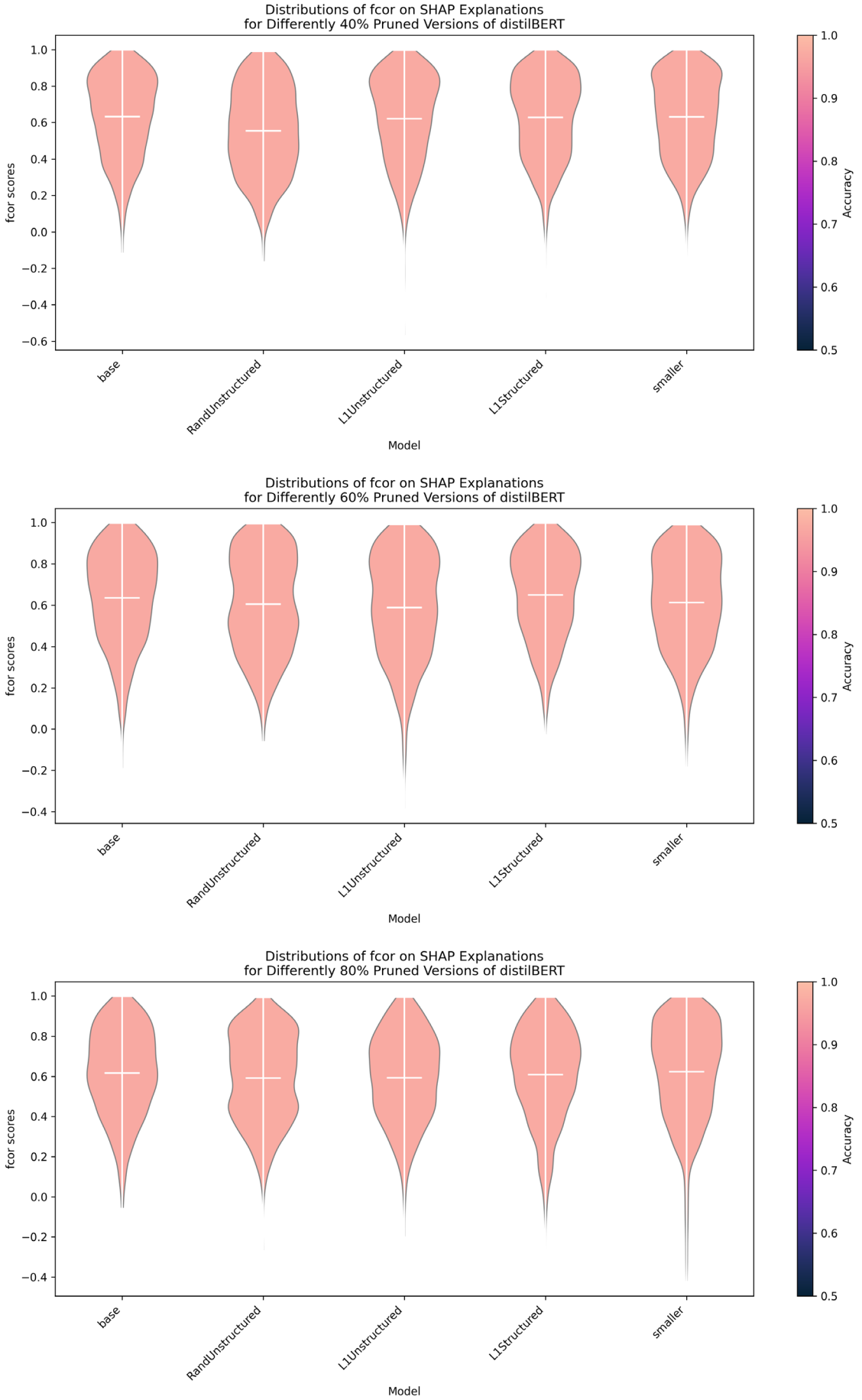

Figure 6. Distributions of FCor scores for SHAP on DistilBERT on Yelp.

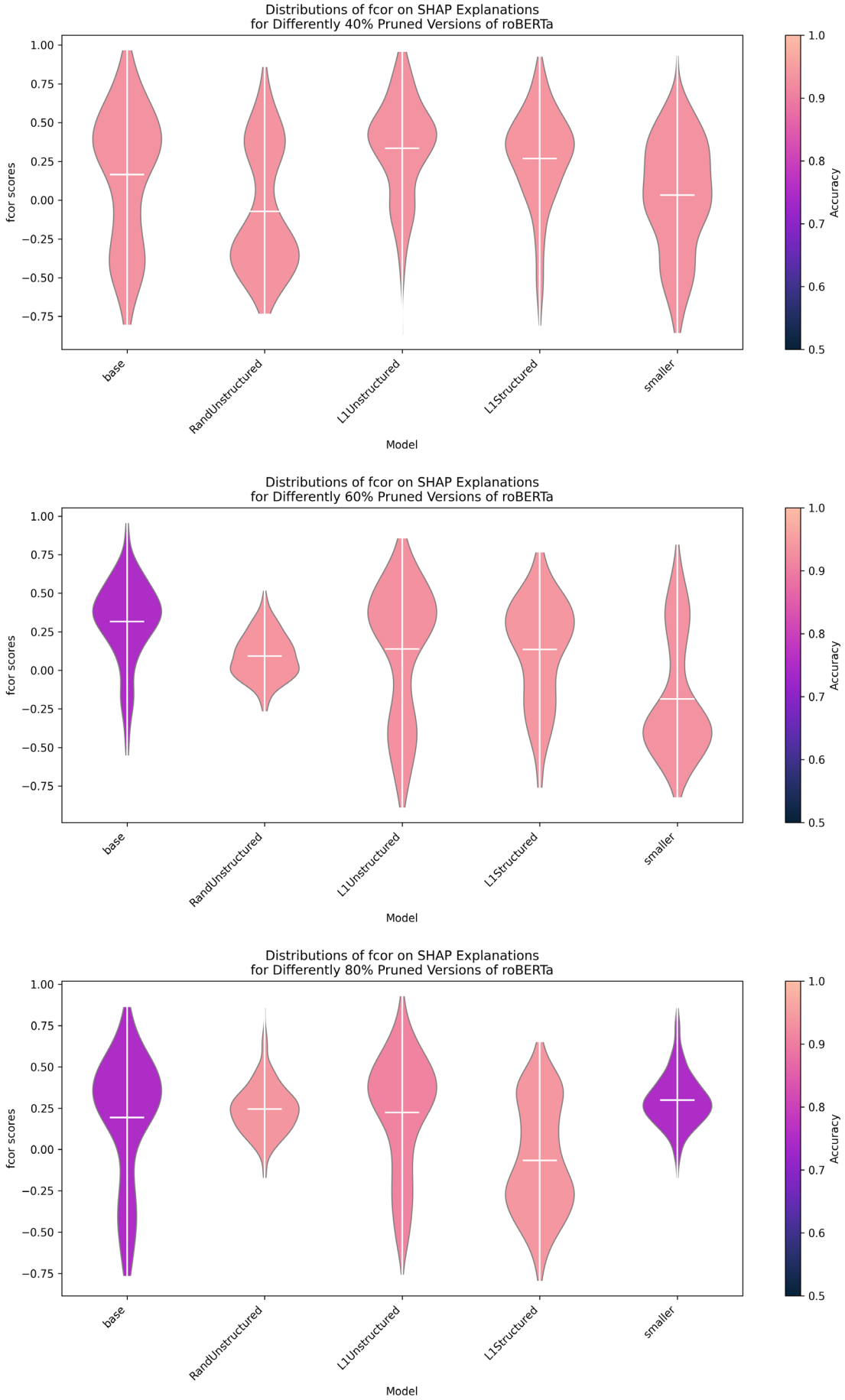

Figure 7. Distributions of FCor scores for SHAP on RoBERTa on IMDB.

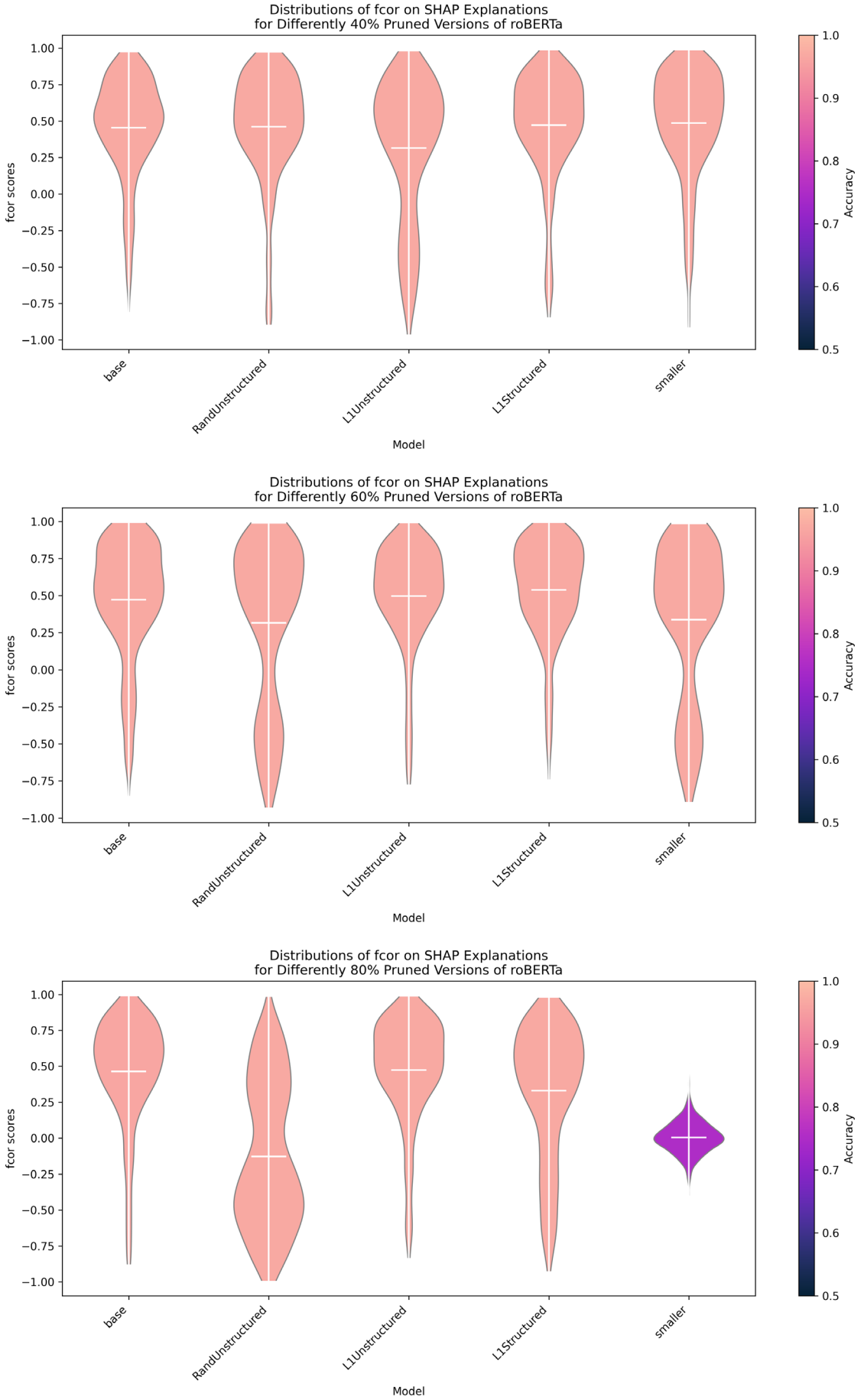

Figure 8. Distributions of FCor scores for SHAP on RoBERTa on Yelp.

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb | 40pct	**0.24**	1.35	**0.24**	0.30	2.07	0.29	0.27	**0.02**	0.07	0.16
IMDb | 60pct	0.24	**0.21**	0.31	0.45	1.73	0.25	**0.01**	0.04	0.09	0.07
IMDb | 80pct	0.31	0.44	0.32	0.38	0.78	**0.27**	**0.00**	0.02	0.06	**0.00**
Yelp | 40 pct	0.14	0.70	0.17	0.16	2.19	**0.04**	0.22	**0.02**	0.10	0.42
Yelp | 60 pct	**0.17**	0.84	0.21	0.28	0.27	0.05	0.05	**0.01**	0.06	0.08
Yelp | 80 pct	**0.18**	0.83	**0.18**	0.25	0.38	0.07	2.11	0.08	0.10	**0.00**

Figure 9. Maximum absolute value of gradient (**min** in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb | 40pct	0.03	0.03	0.02	0.03	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
IMDb | 60pct	0.03	0.03	0.03	0.03	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
IMDb | 80pct	0.02	0.04	0.03	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Yelp | 40 pct	0.01	0.01	0.01	0.01	0.01	**0.00**	**0.00**	**0.00**	**0.00**	0.01
Yelp | 60 pct	0.01	0.01	0.01	0.01	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Yelp | 80 pct	0.01	0.01	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.10	0.03

Figure 10. Average absolute value of gradient (**min** in bold, values rounded to 2 decimal places).

implying local non-linearity; therefore, we expect high MAVHD to correspond with low FCor scores, since a model that is locally non-linear undermines the faithfulness of SHAP explanations. Indeed, we find a strong negative correlation between FCor scores and the MAVHD, with $r = -0.76$, $r = -0.13$, $r = -0.65$, and $r = -0.72$ for DistilBERT-IMDb, DistilBERT-Yelp, RoBERTa-IMDb, and RoBERTaYelp, respectively. Note the outlying weak correlation for DistilBERT trained on Yelp; we hypothesize that due to using only 3% and 10% of the training sets for FCor and Hessian computation, respectively, our FCor and Hessian approximations are not equal in their accuracy and coverage of the model's decision landscape, potentially explaining outliers in results. Future work will address these issues.

4.5 Explainability is architecture-sensitive

We observe that explanation faithfulness is significantly impacted by model architecture. This is a very intuitive result, as one could imagine the trivial case of a constant function which is maximally explainable. However, our results give some insight into how explainability is impacted by the architecture of complex language models.

First, we observe that the distribution of FCor scores has a characteristic shape that varies primarily with architecture, holding all else constant. This suggests that the explainability of a model and its approximations (i.e. pruned models) is highly sensitive to choice of architecture (**Figures 5, 6** vs. **7, 8**).

Second, we observe more variation in FCor across pruning methods when applied to RoBERTa. This suggests that, in addition to influencing faithfulness of explanations of a model, choice of architecture also impacts a model's sensitivity to pruning with respect to explainability. Framing this in the context of the pruning-curvature hypothesis discussed in section 4.3, it is possible that the probability of creating a curved region from a random pruning event is determined by model architecture. For example, **Figures 7** and **8** demonstrate that the distributions of FCor for the RandUnstruct and 'smaller' models (which we hypothesize to be the most susceptible to the creation of high-curvature regions) tend to vary the most dramatically from the 'base' distribution. In contrast, the DistilBERT distributions exhibit negligible changes in their characteristic shape (**Figures 5, 6**).

To explain the differences in explainability across architectures, we observe that **(1)** RoBERTa has twice as many layers as DistilBERT, and **(2)** RoBERTa makes use of a pooling layer, which is designed to aggregate and capture all of the information contained within an encoded, variable-length input sequence and compress it into a single, fixed-length vector (Liu et al., 2019).

We hypothesize that the increased sensitivity of RoBERTa to pruning compared to DistilBERT stems from the inclusion of the pooling layer. Intuitively, the process of condensing an entire variable-length representation sequence into a fixed-length vector may result in a very dense and uninterpretable representation vector that is used directly to compute the model's output. Future work will take a modular approach to explainability and consider the effects of particular architectural choices on a model's explainability.

Related Work

There is some existing literature at the intersection of XAI and neural network pruning. Weber et al. study the effect of pruning on CNN explainability, finding that magnitude-based pruning methods are effective in reducing network complexity and thereby improving explainability of image classification models (Weber, Merkle, Schöttle, & Schlögl, 2023). Khalifa et al. use tree-based pruning methods to transform Random Forest models into explainable models without sacrificing accuracy (Khalifa, Abdelkader, & Elsaid, 2024).

In addition, there has been work on explainability-aware pruning methods, which seek to use explainability criteria to

determine which parameters or filters of a model to prune (Yu & Xiang, 2023; Z. Li & Song, 2024).

However, to the best of our knowledge, previous work has not investigated the effect of pruning on the explainability of LLMs. The impact of pruning on the local geometry of the network has not been well-studied. While there have been investigations on the effect of pruning on the geometry of the loss landscape or on the decision boundaries, no such work has been conducted for the impact of pruning on the local geometry of the function represented by a neural network (Cai et al., 2023; Tran, Fioretto, Kim, & Naidu, 2022).

Conclusion

This work investigates the effect of zero-order pruning methods on the explainability of DistilBERT and RoBERTa, as measured by the FCor scores of SHAP and IG explanations. We initially find that IG is ill-suited for the language domain due to the discrete nature of natural language, while SHAP gives much more faithful explanations for the sentiment analysis task.

We do not find that magnitude-based pruning affects explainability; however, it preserves accuracy as expected. In contrast, Random Unstructured pruning had a negative effect on explainability on average. We explain this finding by showing that Random Unstructured pruning can create highly curved regions in a network's decision function, undermining SHAP faithfulness by violating the local linearity assumption. Finally, we observe that explanation faithfulness is highly dependent on model architecture, and offer an explanation based on RoBERTa's pooling layer.

Limitations and Future Work. We experiment with relatively small language models by current standards. Future work will experiment with larger models and more varied architectures to study how the relationship between pruning and explainability is affected.

Both the IMDb and Yelp Polarity datasets used in this work represent the task of binary sentiment classification. Future work will investigate in more depth the effect of varying dataset task, size, and complexity on the trained model's explainability. This is an especially interesting line of future work, since our results show that both accuracy and FCor increase across the board when comparing models trained on IMDb to models trained on Yelp, holding all else constant (**Figures 1, 2**). We hypothesize that the improvement in accuracy is due to the Yelp dataset's larger size, but the effect of the choice of dataset on model explainability is unclear.

Additionally, we recognize that the paradigm in state-of-the-art language model training favors the fine-tuning of highperforming foundation models to specific tasks, contrasting with our reinitialization and train from scratch approach. Future work will investigate if our results remain consistent across training schemes.

There also remains much to explore with regard to other pruning methods. While this work selects classic, magnitude based methods as a starting point for investigating the effect of pruning on explainability, recent work has developed pruning methods tailored for LLMs, including structured and higher-order methods (Kwon et al., 2022; Sun et al., 2024; Dery et al., 2024; J. Li et al., 2024; Ma et al., 2023; Frantar & Alistarh, 2023; Kurtic et al., 2022). These methods may vary in their effect on the evaluation metrics. Additionally, our sparsity levels are not exhaustive, and there remains much to learn on how pruning below 40% and beyond 80% affects model accuracy, explanation faithfulness, and network geometry.

Future work will investigate the effect of pruning on other metrics in the explainability literature such as robustness (Chen, Subhash, Havasi, Pan, & Doshi-Velez, 2024).

Furthermore, it will be interesting to perform a more fine-grained analysis of the models' local geometries and to develop theoretical guarantees for the effects of different pruning methods on the geometry of the network.

Finally, all of the future work mentioned so far will be helpful for developing an explainability-optimizing pruning method that does not significantly impact accuracy.

Additional Materials

Supplemental figures S1-S4 can be found online at thurj.org.

References

Bhatt, U., Weller, A., & Moura, J. M. F. (2020, May). *Evaluating and Aggregating Feature-based Model Explanations*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2005.00631 (arXiv:2005.00631 [cs]) doi: 10.48550/arXiv.2005.00631

Cai, J., Nguyen, K.-N., Shrestha, N., Good, A., Tu, R., Yu, X., . . . Serra, T. (2023, January). *Getting Away with More Network Pruning: From Sparsity to Geometry and Linear Regions*. arXiv. Retrieved 2024-11-08, from http://arxiv.org/abs/2301.07966 (arXiv:2301.07966 [cs])

Chen, Z., Subhash, V., Havasi, M., Pan, W., & Doshi-Velez, F. (2024). *What makes a good explanation?: A harmonized view of properties of explanations*. Retrieved from https://arxiv.org/abs/2211.05667

Decker, T., Bhattarai, A. R., Gu, J., Tresp, V., & Buettner, F. (2024, June). *Provably Better Explanations with Optimized Aggregation of Feature Attributions*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2406.05090 (arXiv:2406.05090 [cs]) doi: 10.48550/arXiv.2406.05090

Dery, L., Kolawole, S., Kagy, J.-F., Smith, V., Neubig, G., & Talwalkar, A. (2024, February). *Everybody Prune Now: Structured Pruning of LLMs with only Forward Passes*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2402.05406 (arXiv:2402.05406 [cs]) doi: 10.48550/arXiv.2402.05406

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/1810.04805 (arXiv:1810.04805 [cs]) doi: 10.48550/arXiv.1810.04805

Elsayed, M., Farrahi, H., Dangel, F., & Mahmood, A. R. (2024). Revisiting scalable hessian diagonal approximations for applications in reinforcement learning. *ArXiv, abs/2406.03276*. Retrieved from https://api.semanticscholar.org/CorpusID:270258083

Enguehard, J. (2023, May). *Sequential Integrated Gradients: a simple but effective method for explaining language models*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2305.15853 (arXiv:2305.15853 [cs]) doi: 10.48550/arXiv.2305.15853

Frankle, J., & Carbin, M. (2019, March). *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/1803.03635 (arXiv:1803.03635 [cs]) doi: 10.48550/arXiv.1803.03635

Frantar, E., & Alistarh, D. (2023, March). *SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot*. arXiv. Retrieved 2024-12-07, from http://arxiv.org/abs/2301.00774 (arXiv:2301.00774 [cs]) doi: 10.48550/arXiv.2301.00774

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015, October). *Learning both Weights and Connections for Efficient Neural Networks*. arXiv. Retrieved 2024-12-07,

from http://arxiv.org/abs/1506.02626 (arXiv:1506.02626 [cs]) doi: 10.48550/arXiv.1506.02626

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021, February). *Self-Attention Attribution: Interpreting Information Interactions Inside Transformer*. arXiv. Retrieved 2024-12-07, from http://arxiv.org/abs/2004.11207 (arXiv:2004.11207 [cs]) doi: 10.48550/arXiv.2004.11207

Janizek, J. D., Sturmfels, P., & Lee, S.-I. (2020, June). *Explaining Explanations: Axiomatic Feature Interactions for Deep Networks*. arXiv. Retrieved 2024-12-07, from http://arxiv.org/abs/2002.04138 (arXiv:2002.04138 [cs]) doi: 10.48550/arXiv.2002.04138

Khalifa, F. A., Abdelkader, H. M., & Elsaid, A. H. (2024). An analysis of ensemble pruning methods under the explanation of random forest. *Information Systems*, 120, 102310. Retrieved from https://www.sciencedirect.com/science/article/pii/S0306437923001461 doi: https://doi.org/10.1016/j.is.2023.102310

Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., . . . Alistarh, D. (2022). *The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models*. arXiv. Retrieved 2024-11-08, from https://arxiv.org/abs/2203.07259 (Version Number: 3) doi: 10.48550/ARXIV.2203.07259

Kwon, W., Kim, S., Mahoney, M. W., Hassoun, J., Keutzer, K., & Gholami, A. (2022). *A Fast Post-Training Pruning Framework for Transformers*. arXiv. Retrieved 2024-11-08, from https://arxiv.org/abs/2204.09656 (Version Number: 2) doi: 10.48550/ARXIV.2204.09656

Li, J., Dong, Y., & Lei, Q. (2024, July). *Greedy Output Approximation: Towards Efficient Structured Pruning for LLMs Without Retraining*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2407.19126 (arXiv:2407.19126 [cs]) doi: 10.48550/arXiv.2407.19126

Li, Z., & Song, Z. (2024, April). Structured Pruning Strategy Based on Interpretable Machine Learning. In *2024 5th International Conference on Computer Engineering and Application (ICCEA)* (pp. 801–804). Hangzhou, China: IEEE. Retrieved 2024-12-03, from https://ieeexplore.ieee.org/document/10603526/ doi: 10.1109/ICCEA62105.2024.10603526

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, July). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/1907.11692 (arXiv:1907.11692 [cs]) doi: 10.48550/arXiv.1907.11692

Lundberg, S., & Lee, S.-I. (2017, November). *A Unified Approach to Interpreting Model Predictions*. arXiv. Retrieved 2024-12-07, from http://arxiv.org/abs/1705.07874 (arXiv:1705.07874 [cs]) doi: 10.48550/arXiv.1705.07874

Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024, January). *Towards Faithful Model Explanation in NLP: A Survey*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2209.11326 (arXiv:2209.11326 [cs]) doi: 10.48550/arXiv.2209.11326

Ma, X., Fang, G., & Wang, X. (2023, September). *LLM-Pruner: On the Structural Pruning of Large Language Models*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2305.11627 (arXiv:2305.11627 [cs]) doi: 10.48550/arXiv.2305.11627

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1015

Meyer, R. A., & Avron, H. (2023). Hutchinson's estimator is bad at kronecker-trace-estimation. *ArXiv, abs/2309.04952*. Retrieved from https://api.semanticscholar.org/CorpusID:261681808

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022, October). SHAP-based explanation methods: A review for NLP interpretability. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2022.coling-1.406

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv. Retrieved 2024-12-08, from http://arxiv.org/abs/1910.01108 (arXiv:1910.01108 [cs]) doi: 10.48550/arXiv.1910.01108

Sanyal, S., & Ren, X. (2021, August). *Discretized Integrated Gradients for Explaining Language Models*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2108.13654 (arXiv:2108.13654 [cs]) doi: 10.48550/ arXiv.2108.13654

Sun, M., Liu, Z., Bair, A., & Kolter, J. Z. (2024, May). *A Simple and Effective Pruning Approach for Large Language Models*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2306.11695 (arXiv:2306.11695 [cs]) doi: 10.48550/arXiv.2306.11695

Sundararajan, M., Taly, A., & Yan, Q. (2017, June). *Axiomatic Attribution for Deep Networks*. arXiv. Retrieved 2024-12- 07, from http://arxiv.org/abs/1703.01365 (arXiv:1703.01365 [cs]) doi: 10.48550/arXiv.1703.01365

Tran, C., Fioretto, F., Kim, J.-E., & Naidu, R. (2022, October). *Pruning has a disparate impact on model accuracy*. arXiv. Retrieved 2024-11-29, from http://arxiv.org/abs/2205.13574 (arXiv:2205.13574 [cs]) doi: 10.48550/arXiv.2205.13574

Volkov, E. N., & Averkin, A. N. (2024, May). Local Explanations for Large Language Models: a Brief Review of Methods. In *2024 XXVII International Conference on Soft Computing and Measurements (SCM)* (pp. 189–192). Saint Petersburg, Russian Federation: IEEE. Retrieved 2024-12-03, from https://ieeexplore.ieee.org/document/10554222/ doi: 10.1109/SCM62608.2024.10554222

Weber, D., Merkle, F., Schöttle, P., & Schlögl, S. (2023, February). *Less is More: The Influence of Pruning on the Explainability of CNNs*. arXiv. Retrieved 2024-12-08, from http://arxiv.org/abs/2302.08878 (arXiv:2302.08878 [cs]) doi: 10.48550/arXiv.2302.08878

Yao, Z., Gholami, A., Shen, S., Mustafa, M., Keutzer, K., & Mahoney, M. W. (2021, April). *ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning*. arXiv. Retrieved 2024-12-09, from http://arxiv.org/abs/2006.00719 (arXiv:2006.00719 [cs]) doi: 10.48550/arXiv.2006.00719

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019, November). *On the (In)fidelity and Sensitivity for Explanations*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/1901.09392 (arXiv:1901.09392 [cs]) doi: 10.48550/arXiv.1901.09392

Yu, L., & Xiang, W. (2023, June). *X-Pruner: eXplainable Pruning for Vision Transformers*. arXiv. Retrieved 2024-12-03, from http://arxiv.org/abs/2303.04935 (arXiv:2303.04935 [cs]) doi: 10.48550/arXiv.2303.04935

Zhang, X., Zhao, J., & LeCun, Y. (2015, September). Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., . . . Du, M. (2023, November). *Explainability for Large Language Models: A Survey*. arXiv. Retrieved 2024-12-08, from http://arxiv.org/abs/2309.01029 (arXiv:2309.01029 [cs]) doi: 10.48550/arXiv.2309.01029