BEYOND BINARY SELECTION: EXPLORING SOFT MASKING TECHNIQUES FOR LANGUAGE MODEL EXPLAINABILITY

Steffen Backmann
Department of Computer Science
ETH Zürich
Rämistrasse 101, 8092 Zürich, Schweiz
sbackmann@student.ethz.ch

ABSTRACT

Explainability methods generate various attributions and importance scores, but current evaluation practices often rely on binary transformations of these attributions into explanations, failing to capture the nuances in score variations. This approach can lead to inaccurate assessments of the methods. In response, we propose a novel approach that produces textual explanations reflecting the true output importance scores. This enables a more accurate evaluation and comparison of explainability methods. While our generated explanations may be less faithful to the original model (as they diverge more from the initial input sentences), they better represent the variability among different explainability methods.

1 INTRODUCTION

In recent years, the rapid advancement of machine learning models has been paralleled by an increasing demand for explainability. As models become more complex, understanding the reasoning behind their predictions is crucial for trust, transparency, and debugging. Various explainability methods have emerged, providing attributions and importance scores that indicate the influence of individual features on a model's output. However, the predominant evaluation practices for these methods often simplify the attributions into binary explanations, overlooking the subtle variations in the scores. This binary transformation can result in inaccurate evaluations, as it fails to capture the granularity and nuanced differences among explainability techniques.

To address this limitation, we propose a novel approach that transforms attributions into textual explanations that more accurately reflect the true importance scores. By converting numerical attributions into descriptive text, our method provides a richer and more detailed representation of the model's reasoning process. Although these generated explanations might deviate from the original input sentences, potentially affecting their faithfulness to the initial model, they offer a clearer depiction of the variability and performance of different explainability methods.

This approach not only facilitates a more accurate assessment and comparison of explainability techniques but also enhances the interpretability of model outputs. By focusing on the variability among different methods, our textual explanations provide deeper insights into the strengths and weaknesses of each approach, paving the way for more reliable and comprehensible machine learning models.

2 RETAINING NUANCED ATTRIBUTIONS THROUGH SOFT MASKING FOR TEXTUAL EXPLANATIONS

In this chapter, the approach for developing an explainability method that retains a soft mask, thereby leveraging all the information contained in the attribution scores, is outlined. This method involves the use of *GenericsKB* (Bhakthavatsalam et al., 2020), a text generation dataset and *Gemma* (Team et al., 2024), a publicly available language model to create replacement explanations that maintain

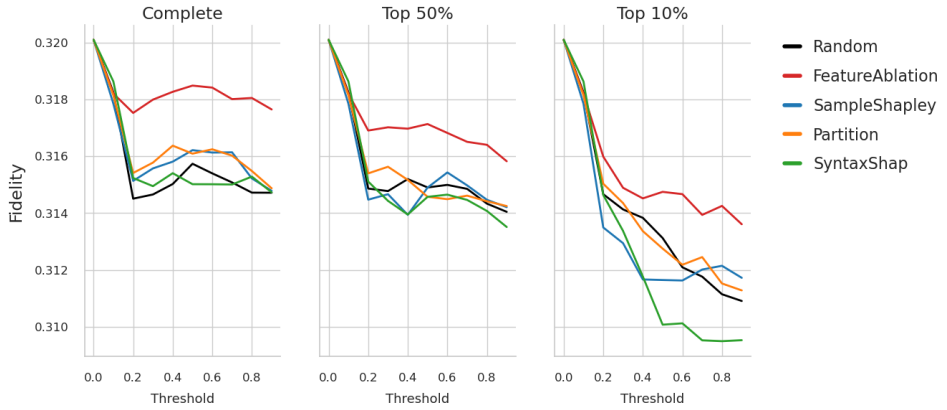

Figure 1: Comparison of sampling from the full embedding space and from the most similar 50 % as well as 10 %.

the nuances in importance scores . We will compare this strategy against traditional binary selection methods to demonstrate its effectiveness. The approach is structured as follows:

2.1 Choosing a Text Generation Dataset and Publicly Available Language Model

The first step in our approach is to select an appropriate text generation dataset and a publicly available language model. The chosen dataset should be diverse and extensive to ensure a comprehensive evaluation of the explainability method. For this, the *GenericsKB* which contains high-quality, semantically complete statements is chosen. For our approach, we use Google's *Gemma* model. *Gemma* is publicly available and provides a large vocabulary, making it well-suited for generating diverse and meaningful word embeddings.

2.2 Obtaining Word Embeddings & Explanations

Word embeddings are obtained from the *Gemma* model. Word embeddings are vector representations of words that capture their semantic meaning and relationships within a text. By using pre-trained language models, we can extract high-quality embeddings that serve as the foundation for calculating distances and sampling replacements. These embeddings will be instrumental in creating the replacement explanations that reflect the importance scores provided by the explainability methods. The embeddings are extracted from both, a lower and a higher layer of the model in order to compare the influence of different levels of contextualization for the replacement techniques.

We then choose a set of explainability techniques, namely SyntaxShap (Amara et al., 2024), FeatureAblation, SampleShapley, Hedge (Chen et al., 2020), and a random baseline, i.e. a normally distributed token importance attribution. By applying these methods, we can generate a set of importance scores for the *GenericsKB* dataset text inputs.

2.3 Creating Replacement Explanations Based on Calculated Distances

Once we have the importance scores, we move to the process of creating replacement explanations that incorporate the nuances of these scores. For each token in the input samples, we calculate the cosine similarity of its embedding to all other token embeddings. We then sample replacement tokens for each input token from the sorted list of token embeddings based on the attribution score assigned by the explainability method. This process ensures that the sampled token replacements maintain coherence relative to their importance. These replacements reflect the soft mask, as they are guided by the importance scores and calculated similarities.

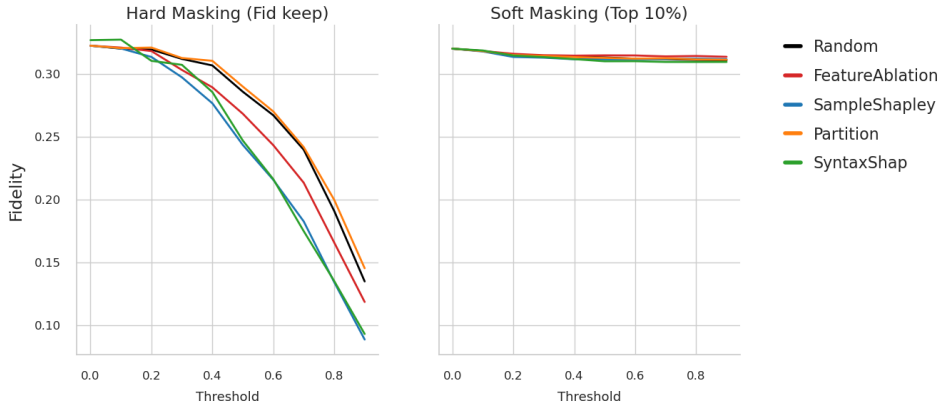

Figure 2: Comparison of traditional hard masking approach and our resampling technique.

3 Experiments & Results

For each experiment, we establish a threshold that determines the proportion of tokens to be resampled based on their attribution scores, while the remaining tokens in the input are masked, i.e., set to zero. For instance, a threshold of 0.6 means that the top 60 % of tokens, as identified by their importance scores from the explainability method, are resampled, whereas the bottom 40 % are masked to zero. We experiment with various resampling techniques that differ primarily in the range of embeddings considered. In the base case, importance scores are mapped across the entire range of embeddings. Variations include restricting the resampling to the top 50 % or 10 % most similar tokens. The results of these experiments are shown in Figure 3. We evaluate the faithfulness of the different resampling strategies by calculating fidelity, a commonly used model-based metric in explainable AI, which examines the top-1 prediction. It is apparent that using the full set and even the closest 50 % of embeddings leads to very little differentiation between the explainability methods and only barely decreasing fidelity values for higher thresholds. One explanation could be that wehen looking at the entire range of embeddings, generally the tokens are too different so that the behavior is similar to sampling a random token, regardless of the attribution scores. This is supported by the fact that we see a clearer trend and better differentiations when limiting the sampling range to the most similar 10 % of tokens. No significant difference could be found when resampling 10 times based on a normal distribution (with the calculated rank position as mean) and taking the average fidelity. This could be rooted in fidelity values being very close to each other when the top-1 prediction probability is low. Here, computing the top-k probability could potentially show better results.

As illustrated in Figure 3, all resampling strategies exhibit significantly higher fidelity levels compared to traditional hard masking approaches. This is to be expected since the resampled (different) tokens can hardly lead to more faithful results than the original input tokens.

References

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. Syntaxshap: Syntax-aware explainability method for text generation, 2024. URL https://arxiv.org/abs/2402.09259.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. Genericskb: A knowledge base of generic statements. *CoRR*, abs/2005.00660, 2020. URL https://arxiv.org/abs/2005.00660.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5578–5593, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.494. URL https://aclanthology.org/2020.acl-main.494.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, and et al. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.