Self-Guided Attention for Explainable Natural Language Inference

Phil Gu (fg53)
https://github.com/alwaysstrive2024/selfGuided.git

Abstract

Explainable Artificial Intelligence (XAI) aims to make machine learning models transparent and interpretable, particularly in high-stakes decision-making tasks. In this work, I investigate the role of human-guided intervention in transformer-based models to improve both interpretability and predictive performance in Natural Language Inference (NLI). I propose Self-Guided Attention, a mechanism that incorporates human-provided explanations into the attention layers of BERT. By biasing attention toward tokens highlighted by annotators, I aim to align the model's internal reasoning with human rationales while enhancing classification accuracy. Experiments on the e-SNLI dataset demonstrate that guided attention significantly improves alignment with human explanations, measured by top-k attention Intersection-over-Union (IoU), and can also improve model performance compared to unguided baselines. This study demonstrates that human-intervention in attention can be a practical XAI strategy for improving both explainability and effectiveness. My implementation is publicly available at https://github.com/alwaysstrive2024/selfGuided.git.

1 Introduction

Explainable Artificial Intelligence (XAI) is a rapidly growing research area that seeks to make machine learning models interpretable and accountable, especially in tasks with potential real-world impact. One fundamental challenge in XAI is understanding how model reasoning aligns with human rationale, and whether explicit interventions can improve both interpretability and model performance. Natural Language Inference (NLI) is particularly suitable for this study, as it requires models to determine the logical relationship between a premise and a hypothesis, a task for which human reasoning can be explicitly represented via annotated explanations.

Transformer-based models, such as BERT, have achieved remarkable accuracy on NLI tasks. However, their attention mechanisms, often used as a proxy for interpretability, may not faithfully reflect human-relevant reasoning, limiting the transparency of model predictions. To address this, I investigate the hypothesis that explicitly guiding model attention using human-provided explanations can simultaneously improve interpretability and predictive performance. This work focuses on the design, implementation, and evaluation of a Self-Guided Attention mechanism, which directly intervenes in the model's internal attention distribution to prioritize tokens highlighted in human explanations. By doing so, I aim to demonstrate that targeted intervention can serve as an effective XAI strategy for enhancing both model reasoning and outcomes.

2 Related Work

Transformer-based models, particularly BERT, have become the standard for NLI tasks due to their strong contextualized representations and ability to capture long-range dependencies [2]. These models achieve state-of-the-art performance on SNLI and MNLI benchmarks, demonstrating remarkable accuracy in capturing semantic relationships between premise-hypothesis pairs. Despite these successes, such models remain largely opaque, and their internal reasoning is not directly interpretable. Previous studies have shown that naive attention weights, often used as a proxy for interpretability, do not

consistently reflect meaningful or human-aligned reasoning [3]. This limitation motivates research on methods that explicitly integrate human rationales into model architectures to improve both transparency and trustworthiness.

The e-SNLI dataset [1] extends SNLI by providing human-written natural language explanations for each labeled example, indicating which tokens or phrases are critical for the inference decision. Prior work using e-SNLI has primarily focused on post-hoc evaluation of interpretability, analyzing whether model attention or gradient-based attributions align with human explanations. While such analyses provide insights into model behavior, they do not directly influence training or improve model reasoning. Some recent approaches have attempted to incorporate explanation supervision as an auxiliary loss, using human rationales to guide predictions indirectly, but these methods often require additional tuning and do not intervene explicitly in the attention mechanism.

Guided attention mechanisms have been explored in both computer vision and NLP, where human-provided annotations are used to bias the model's focus toward relevant regions or tokens. In NLP, prior studies have applied guided attention to tasks such as text classification, sentiment analysis, and reading comprehension, showing that integrating human knowledge can enhance interpretability and, in some cases, improve performance. However, systematic application of guided attention to NLI, particularly with transformer architectures like BERT, remains limited. Existing works either intervene in a single attention layer or impose soft regularization constraints on attention distributions, rather than directly modifying the attention probabilities in multiple layers where semantic reasoning is concentrated.

My work differs from previous studies in several key aspects. First, I implement Self-Guided Attention, which replaces the self-attention modules in the last four encoder layers of BERT with guided attention layers that directly incorporate human explanation masks. This explicit intervention biases the model to focus on tokens identified as important by annotators, rather than merely adding auxiliary loss terms. Second, I investigate the impact of this intervention on both interpretability and predictive performance, systematically comparing *vanilla* fine-tuning, guided attention, and random attention baselines. Third, by aligning human explanations at the token level and integrating them into multiple attention layers, my approach captures richer reasoning patterns than previous methods that rely on single-layer guidance or post-hoc evaluation. Finally, I provide quantitative metrics for explainability, using top-k attention Intersection-over-Union (IoU) to measure alignment with human rationales, along with standard classification accuracy, demonstrating that guided attention can simultaneously improve interpretability and performance.

Overall, my approach extends prior research by demonstrating that direct, human-guided intervention in BERT's attention mechanism is an effective strategy for explainable NLI. It provides a practical framework for integrating human rationales into transformer-based models, highlighting a promising direction for XAI research where interventions are used not only for interpretability but also to enhance model performance.

3 Proposed Method

In this work, I focus on integrating human explanations into transformer-based models to improve both interpretability and predictive performance in NLI. My approach, Self-Guided Attention, intervenes in BERT's attention mechanism to bias model focus toward tokens identified as important by human annotators. In this section, I describe the standard attention mechanism, the guided attention formulation, gold mask construction, training objectives, and alternative intervention strategies considered.

3.1 Attention Formulation and Intervention

BERT's standard multi-head self-attention can be expressed as:

$$\text{Attn}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{n \times d_k}$ are the query, key, and value matrices for a sequence of length n, and d_k is the hidden dimension per head. The resulting attention distribution

$$\text{attn\_probs} = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big) \tag{2}$$

represents the focus over tokens for each query position. In standard BERT, this attention is learned implicitly and is not constrained by human explanations.

To incorporate explanations, I define a binary gold mask $M \in \{0, 1\}^n$, where tokens appearing in human-provided rationales are assigned 1, and all others 0. I investigated several ways to integrate this mask into attention:

1. Soft multiplicative biasing:

$$\text{attn\_probs}^{\text{guided}} = \text{softmax}\Big(\text{attn\_scores} \odot (1 + \lambda M)\Big) \tag{3}$$

   where $\odot$ denotes element-wise multiplication and λ is a hyperparameter controlling guidance strength.

2. Additive biasing:

$$\text{attn\_probs}^{\text{guided}} = \text{softmax}(\text{attn\_scores} + \lambda M) \tag{4}$$

3. Hard replacement:

$$\text{attn\_probs}^{\text{guided}} = \frac{M}{\sum_i M_i} \tag{5}$$

The corresponding attention output after intervention is:

$$H^{\text{guided}} = \text{attn\_probs}^{\text{guided}} V \tag{6}$$

After experimentation, I adopted the soft multiplicative biasing (Eq. 3) because it balances human guidance with model flexibility, provides stable training, and allows the model to learn additional patterns beyond the annotated tokens.

3.2 Gold Mask Construction

The gold mask M is generated by aligning human-written explanations to tokenized input sequences. For a premise-hypothesis pair (p, h) with explanation e, the input text is concatenated as

$$x = [p \text{ [SEP] } h] \tag{7}$$

Using the BERT tokenizer's offset mapping, each token is mapped to its character span in x. A token i is marked as 1 in the gold mask if it overlaps with any phrase in the explanation e, and 0 otherwise. Tokens corresponding to special tokens ([CLS], [SEP], padding) are automatically set to 0. Lowercasing and punctuation stripping are applied to improve alignment robustness.

3.3 Training Objective

The model is trained with the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{8}$$

where C is the number of classes, y the one-hot ground truth label, and $\hat{y}$ the predicted probability. In guided mode, attention layers are intervened as in Eq. 3, while in *vanilla* mode, no intervention is applied. A *random* mode assigns a randomly generated mask of the same sparsity as the gold mask, serving as a control.

3.4 Alternative Intervention Strategies Considered

During preliminary experiments, I explored:

- Layer-specific guidance: Applying intervention to only the last one or two layers. This reduced computational cost but showed smaller gains in accuracy and interpretability compared to guiding the last four layers.
- Gradient-based attention regularization: Adding an auxiliary loss to encourage overlap between attention and the gold mask:
$$\mathcal{L}_{\text{attn}} = \text{KL}\big(\text{softmax}(A) \,||\, M\big) \tag{9}$$
where A represents attention logits. This method was less stable and required careful tuning.
- Hard replacement (Eq. 5): Directly replacing attention with the normalized gold mask, which improved alignment but restricted model flexibility, resulting in lower overall performance.

Based on these evaluations, I selected soft multiplicative biasing applied to the last four layers for all experiments, balancing guidance strength, stability, and performance.

4 Experimental Setup

4.1 Dataset

Experiments are conducted on the e-SNLI dataset, which contains 549,367 training examples, 9,842 validation examples, and 9,824 test examples. Each instance includes a premise, hypothesis, label, and optionally a human-written explanation. I preprocess and cache tokenization and alignment to enable efficient experimentation.

4.2 Implementation Details

I fine-tune BERT-base-uncased with guided attention applied to the last four encoder layers. Training employs the AdamW optimizer with a learning rate of 2×10^{-5}, batch size 16, and maximum sequence length of 128 tokens. Mixed precision is used for efficient GPU training. The top-k Intersection-over-Union (IoU) between predicted attention and gold masks is calculated to evaluate alignment with human explanations.

4.3 Evaluation Metrics

I evaluate the model using two metrics: classification accuracy to assess predictive performance, and attention IoU to quantify alignment with human rationales:

$$\text{IoU} = \frac{|\text{Top-k Attention Tokens} \cap \text{Gold Mask Tokens}|}{|\text{Top-k Attention Tokens} \cup \text{Gold Mask Tokens}|} \tag{10}$$

The IoU metric captures how effectively the intervention aligns the model's internal reasoning with human explanations.

5 Results and Discussion

The experimental results demonstrate that incorporating human-guided attention into BERT significantly improves both predictive performance and interpretability in NLI tasks. Table 1 summarizes the outcomes of three experimental settings: the Guided Model, in which attention in the last four encoder layers is biased using human explanation masks; the Vanilla Model, which is trained with standard fine-tuning; and the Random Model, which uses randomly generated attention masks as a control. Across all

metrics, the Guided Model consistently outperforms the baselines. In terms of predictive performance, the Guided Model achieves a validation accuracy of 0.9067, slightly higher than the Vanilla Model at 0.9011 and substantially higher than the Random Model at 0.8926. This indicates that structured attention guidance, derived from human rationales, can contribute to improved model generalization even on a task where BERT already demonstrates strong baseline performance.

Attention alignment, measured using top-k Intersection-over-Union (IoU) with the gold masks, exhibits a notable increase under the Guided Model, reaching 0.3784, compared to 0.2665 for the Vanilla Model and 0.3181 for the Random baseline. This substantial improvement confirms that the proposed guided attention mechanism effectively directs the model's focus toward semantically relevant tokens, aligning its internal reasoning more closely with human explanations. Further supporting this conclusion, the Guided Model achieves the highest Area Under the Precision-Recall Curve (AUPRC) at 0.6561, reflecting better discrimination in attention-based token importance relative to human annotations. Explanation-based metrics, including Comprehensiveness and Sufficiency, also indicate superior alignment with human rationales: the Guided Model attains values of 0.3199 and 0.4480 respectively, surpassing both the Vanilla and Random Models. High Comprehensiveness suggests that removing the tokens highlighted by the model leads to a significant drop in confidence, indicating that the model relies on semantically meaningful evidence. Similarly, high Sufficiency demonstrates that the tokens prioritized by guided attention are sufficient to support accurate predictions, providing quantitative validation of interpretability.

The Random Model, although occasionally achieving moderate IoU, performs substantially worse on AUPRC and explanation metrics, confirming that improvements in the Guided Model are attributable to structured human intervention rather than incidental attention patterns. Qualitative inspection of attention distributions further corroborates these findings. Visualization of attention heatmaps reveals that the Guided Model consistently emphasizes tokens that are critical for inference, including negations, causal connectors, and named entities, closely mirroring the focus observed in human explanations. In contrast, the Vanilla Model exhibits more dispersed and less semantically coherent attention, and the Random Model frequently highlights irrelevant tokens, underscoring the importance of structured guidance.

Collectively, these results demonstrate that targeted human intervention in attention serves as a practical strategy for explainable artificial intelligence. By simultaneously improving predictive performance and attention alignment with human rationales, the proposed Self-Guided Attention mechanism validates the hypothesis that integrating human explanations into model reasoning enhances both interpretability and task effectiveness. The combination of quantitative metrics and qualitative analysis provides a comprehensive assessment, highlighting the potential of guided attention as a generalizable XAI technique for transformer-based models.

Table 1: Comparison of Guided, Vanilla, and Random Models on e-SNLI. Metrics reported include training loss, validation accuracy, attention IoU, AUPRC, Comprehensiveness, Sufficiency.

Model	Train Loss	Val Accuracy	Val IoU	AUPRC	Comprehensiveness	Sufficiency
Guided	0.2761	0.9067	0.3784	0.6561	0.3199	0.4480
Vanilla	0.2809	0.9011	0.2665	0.4881	0.2135	0.3636
Random	0.2801	0.8926	0.3181	0.3053	0.1535	0.3413

6 Conclusion

I propose Self-Guided Attention, a human-intervention-based mechanism to improve both explainability and performance in transformer-based NLI models. By biasing attention toward tokens highlighted in human explanations, I demonstrate that targeted guidance can enhance interpretability without sacrificing accuracy. Experiments on e-SNLI show that guided attention improves alignment with human

rationales and can also lead to higher classification performance compared to unguided models. This study highlights the potential of using human interventions as a practical XAI strategy, and the full implementation is publicly available at https://github.com/alwaysstrive2024/selfGuided.git.

References

[1] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[3] Sarthak Jain and Byron C. Wallace. 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, 2019.