Variance-Based Integrated Gradients Regularization for Robust Natural Language Inference

Author
Julian Weaver
jiw446
julian.weaver@utexas.edu

Abstract

Pre-trained language models can achieve high accuracy on standard natural language inference (NLI) tasks while still relying heavily on dataset artifacts rather than learning genuine semantic reasoning. In this work, I employ Integrated Gradients (IG) to examine the attributions of an ELECTRA-small model's predictions on the Stanford Natural Language Inference (SNLI) dataset, uncovering its dependence on superficial patterns. To mitigate this issue, I introduce a variance-based penalty on token-level IG attributions, encouraging the model to distribute importance more evenly across relevant tokens and thereby reduce its reliance on spurious correlations. Quantitatively, this Integrated Gradients Variance model shows modest improvements in accuracy and ROC AUC over a baseline model. Qualitatively, it demonstrates more stable, context-sensitive attributions, better handling of complex examples, and improved calibration. Although the performance gains are incremental, these findings highlight the potential of attribution-guided regularization to foster more reliable, reasoning-oriented NLI models.

1 Introduction

The exploration of pre-trained language models in natural language inference (NLI) tasks has shown both their strengths and limitations. While these models achieve impressive performance on benchmark datasets, their success often stems from exploiting superficial patterns or dataset artifacts rather than engaging in the deeper semantic reasoning the task requires (Bowman et al., 2015). This issue is pressing, as it leads models to frequently fail to generalize to adversarial or out-of-domain examples, highlighting the fragility of their learned representations.

To better understand these shortcomings, this project builds on previous work with explainable AI techniques. Specifically, I employed Integrated Gradients (IG) (Sundararajan et al., 2017), a method for attributing a neural network's predictions to its input features by integrating gradients over a path from a baseline to the input. IG excels in providing reliable, feature-level insights into model behavior, making it a natural choice for diagnosing the artifacts driving an NLI model's decisions. Using the Captum library (Kokhlikyan et al., 2020), I conducted an in-depth error analysis of an ELECTRA-small classifier, uncovering how it over-relies on spurious correlations in the dataset (Clark et al., 2020).

Additionally, I propose a novel regularization technique. This method penalizes the variance of token-level attributions across inputs, aiming to guide the model toward learning stable, distributed, contextually meaningful representations rather than brittle shortcuts. This approach builds on existing regularization strategies by directly incorporating attribution-based insights into the training process.

In this report, I evaluate this method both quantitatively, by measuring accuracy and ROC AUC on challenging subsets of the data, and qualitatively, by examining how attributions and predictions change after applying the IG-based variance penalty. The results show promising improvements on examples where the baseline previously failed due to artifact exploitation. In doing so, I highlight the potential of attribution-guided regularization to foster more reliable, reasoning-oriented behavior in NLI models.

2 Methodology

2.1 Data

All experiments were conducted on the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), a standard benchmark for evaluating models' abilities to determine the logical relationship between a premise and a hypothesis.

Each example in SNLI is a tuple of two sentences: a premise and a hypothesis. The task is to classify the pair into one of three relationship categories: entailment, contradiction, or neutral. Prior to training, any examples that were missing valid labels were filtered out. The example sequences were then tokenized using the ELECTRA tokenizer (?).

2.2 Integrated Gradients

I utilized Integrated Gradients (IG) through the Captum library (Kokhlikyan et al., 2020), an open-source framework for model interpretability in PyTorch. IG efficiently computes token-level attributions with respect to a model's predictions by integrating gradients over a defined path from a baseline input to the actual input. For this project, the baseline IG input consisted of all-zero embeddings. To compute the attributions, I passed the token embeddings through a modified forward function that accounted for both embeddings and attention masks, ensuring compatibility with the model's internal attention mechanisms.

Attributions were calculated with 10 integration steps, approximating the integral path. After computing IG for an example, I summed the attributions over the embedding dimension to yield a single scalar attribution per token. This one-dimensional attribution vector provided a clear measure of each token's contribution to the model's predictions and allowed the calculation of the variance-based penalty.

2.3 Variance-Based Loss Regularization

A central contribution of this project was the introduction of a variance-based penalty on IG attributions, designed to address the model's over-reliance on superficial dataset artifacts. The key motivation was to encourage more balanced and contextually-informed token attributions, fostering robust semantic reasoning rather than reliance on spurious correlations. During training, the model's standard classification loss, computed via a cross-entropy objective, was augmented with an additional penalty term based on the variance of token-level attributions. High variance in attributions indicated that the model was disproportionately focusing on a small subset of tokens, possibly suggesting artifact exploitation. The penalty pushed the model to distribute attributions more evenly across tokens, promoting more stable and distributed representations.

The loss function was defined as:

$$\text{Total Loss} = \mathcal{L}_{\text{CE}} + \lambda \cdot \tanh(\alpha \cdot \text{Var}_{\text{attr}})$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, Var_{attr} is the mean variance of token attributions per input, $\alpha = 10$ scales the variance for meaningful attribution differences, and $\lambda = 0.1$ controls the regularization strength, ensuring the penalty shapes the model's learning trajectory without overwhelming the primary classification objective. The penalty term was passed through a hyperbolic tangent function to stabilize its magnitude, preventing extreme contributions during training.

2.4 Experiments

The experiments were conducted in two phases. In the first phase, a baseline model was fine-tuned without any attribution-based penalty to establish a reference point. This model followed standard NLI training procedures for the `google/electra-small-discriminator` model, using a learning rate of 5×10^{-5}, a batch size of 16, and early stopping based on validation performance. In the second phase, the variance penalty term was introduced into the loss function. Training was performed using a custom `IGTrainer` class that overrode the default loss computation to integrate IG attributions and their variance-based penalty. At each training step, the model optimized both for classification accuracy and for reducing patterns of attribution that were overly concentrated on a few tokens.

2.5 Evaluation

To assess the model's performance, I monitored both quantitative and qualitative indicators. For quantitative evaluation, accuracy and ROC AUC were computed on the validation set to measure classification performance. For qualitative analysis, token attributions were inspected for specific examples to evaluate how variance-based regularization affected the model's reasoning patterns. This analysis highlighted changes in the model's reliance on specific tokens and its ability to move away from dataset artifacts.

3 Results

As shown in Table 1, the Integrated Gradients (IG) Variance model achieved slightly higher accuracy (0.8963 vs. 0.8918) and ROC AUC (0.9578 vs.

0.9571) than the Baseline model, though this improvement was not statistically significant. By contrast, the Baseline model had a noticeably faster runtime (17.9s vs. 70.7s), reflecting the computational overhead introduced by the variance penalty. Overall, while the IG Variance model's gains in performance metrics were modest, Figure 1 suggests that its elevated training loss was due to the additional penalty, indicating potential for further tuning to amplify the benefits of this approach.

Metric	Baseline	IG Variance
Accuracy	0.8918	**0.8963**
ROC AUC	0.9571	**0.9578**
Runtime (s)	**17.9134**	70.7104

Table 1: Evaluation metrics for the Baseline model and the IG Variance model. The IG Variance model slightly outperforms the Baseline on accuracy and ROC AUC, but the Baseline is more efficient.

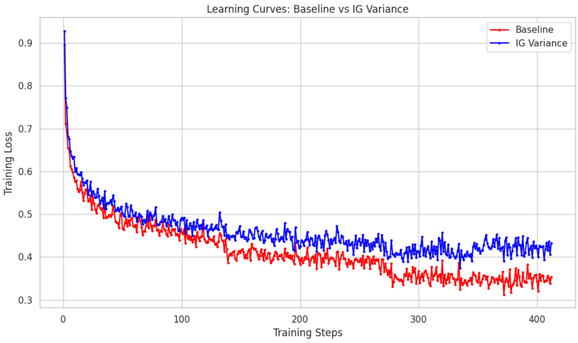

Figure 1: Training loss over time for both models. The IG Variance model maintains a higher overall loss due to the added variance penalty, though the gap is relatively stable, suggesting the model can still learn effectively.

A deeper understanding of these results emerges when examining how each model assigns attributions and handles challenging examples. Figures 2 and 3 present attributions for a neutral case where the Baseline model mistakenly predicted contradiction. While the Baseline model overemphasized certain tokens (e.g., "grocery"), leading it astray, the IG Variance model distributed attributions more evenly and correctly identified the example as neutral. This aligns with our quantitative findings: although the overall performance improvement is small, the IG Variance model demonstrates a qualitatively richer and more consistent attribution pattern, suggesting it is less prone to spurious correlations.

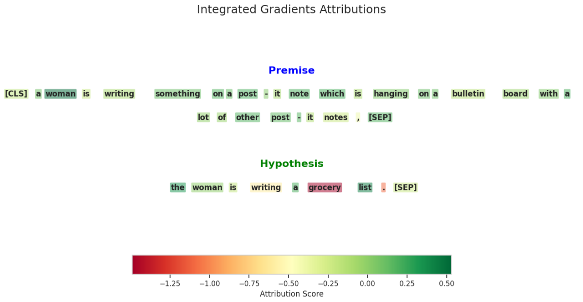

Figure 2: Baseline attributions for the premise "A woman is writing something on a post-it note..." and the hypothesis "The woman is writing a grocery list." The Baseline incorrectly predicted contradiction, attributing excessive importance to "grocery" and showing a skewed distribution of attributions.

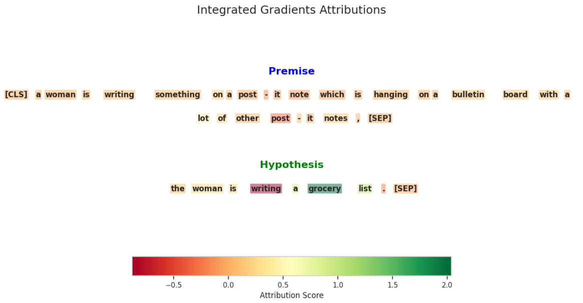

Figure 3: IG Variance attributions for the same neutral example as in Figure 2. The IG Variance model correctly predicted neutral, displaying more balanced attributions across tokens and avoiding the strong misattribution seen in the Baseline.

Similarly, Figures 4 and 5 show a case of contradiction where both premise and hypothesis share words like "boy" and "inflatable ride." The Baseline model, shown in Figure 4, struggled to recognize the semantic mismatch and misclassified the example as neutral. In contrast, the IG Variance model (Figure 5) successfully highlighted the critical token "knife" in the hypothesis, correctly identifying it as a contradiction. This example illustrates that the IG Variance model is less swayed by surface-level lexical overlap and more attentive to key semantic cues, reflecting the intended effect of the variance penalty.

Longer, more complex premises presented another challenge (Figures 6 and 7). While the Baseline model often became confused when faced with lengthy, detail-rich examples, misclassifying an entailment as contradiction, the IG Variance model handled the complexity more gracefully. Figure 8 confirms that error rates rise with length for both models, but the IG Variance model's smoother

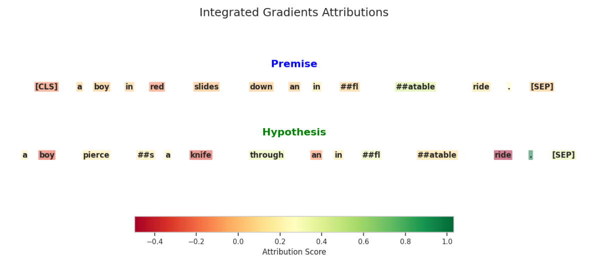

Figure 4: Baseline attributions for the premise “A boy in red slides down an inflatable ride.” and the hypothesis “A boy pierces a knife through an inflatable ride.” The Baseline, distracted by shared terms, wrongly predicted neutral.

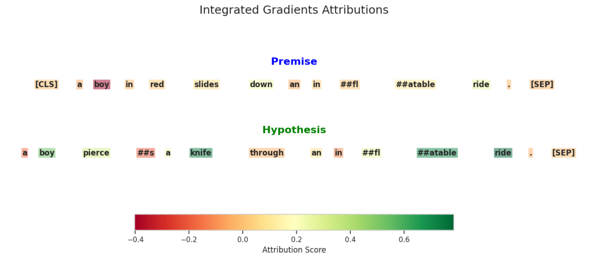

Figure 5: IG Variance attributions for the same contradiction example as in Figure 4. The IG Variance model identified “knife” as the critical token, properly classifying the relationship as a contradiction.

trend suggests that it manages complexity more consistently.

Beyond complexity and logical distinctions, confusion matrices in Figure 9 show that the IG Variance model slightly outperforms the Baseline across all classes. Furthermore, Figure 10 demonstrates that the IG Variance model’s confidence aligns more closely with its accuracy, suggesting improved calibration. This implies that reducing attribution variance may also promote more reliable confidence estimates, a key goal in building robust, trustworthy models.

In summary, the IG Variance model’s quantitative improvement, while small, is supported by qualitative evidence of more stable, context-sensitive attributions and better handling of complex and challenging cases. These results suggest that the variance-based penalty nudges the model toward more semantically meaningful reasoning processes, ultimately laying the groundwork for future enhancements through hyperparameter tuning and more sophisticated attribution-guided interventions.

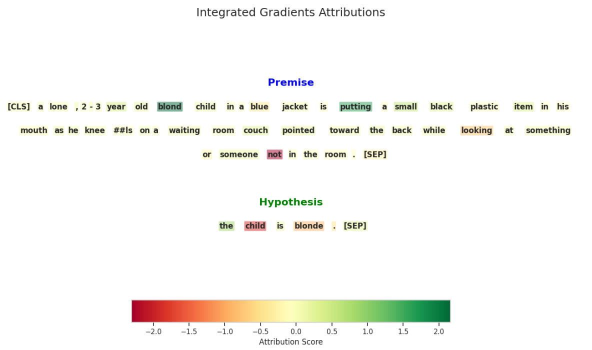

Figure 6: Baseline attributions for a complex entailment example involving a blond child. The model’s attributions are erratic, leading it to predict a contradiction.

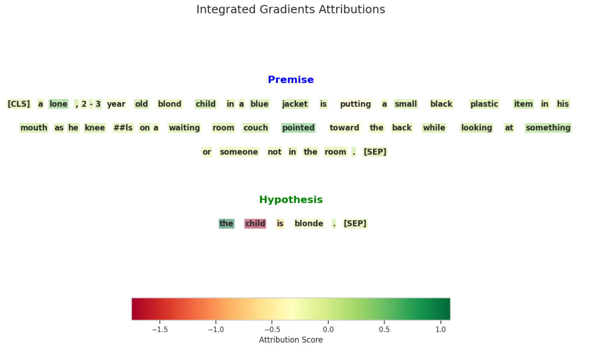

Figure 7: IG Variance attributions for the same complex example as in Figure 6. The model’s more stable attribution distribution allowed it to correctly identify the entailment.

4 Conclusion

This work explored how integrating attribution-based regularization into the training process of an NLI model can mitigate its reliance on superficial dataset artifacts. By penalizing the variance of token-level attributions derived from Integrated Gradients, the model was nudged toward more stable and semantically meaningful representations. Although the observed improvements in accuracy and ROC AUC were modest, the qualitative analyses demonstrated clearer, context-sensitive attributions and more robust reasoning, particularly in challenging scenarios such as complex premises or subtle logical distinctions.

However, there are several limitations to acknowledge. First, the introduced variance penalty increased computational costs, as reflected in longer runtimes. This may limit the feasibility of the approach in resource-constrained environments. Additionally, while the penalty showed promise, it did not produce substantial overall performance gains, suggesting that further hyperparameter tun-

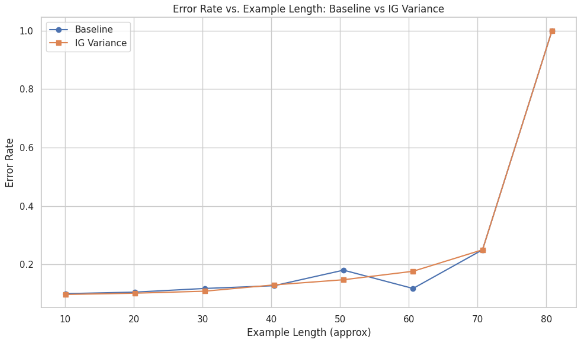

Figure 8: Error rate as a function of example length. Both models struggle with longer examples, but the IG Variance model's error curve is smoother, indicating that the variance penalty helps maintain more stable performance.

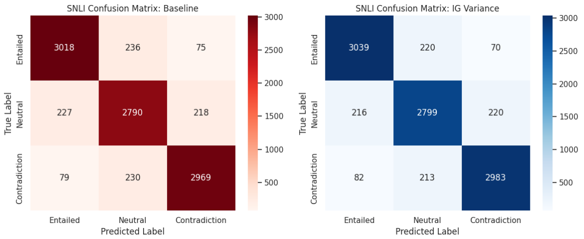

Figure 9: Comparison of confusion matrices. The IG Variance model improves on every class, albeit slightly, reflecting a more balanced understanding of the data.

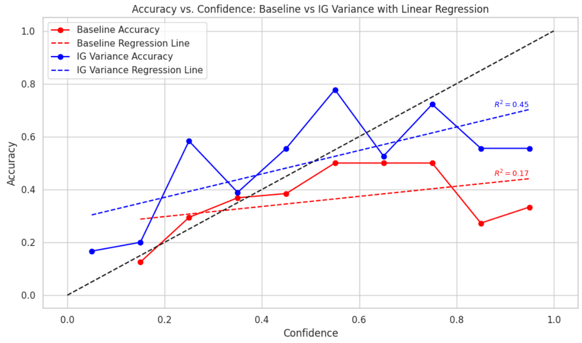

Figure 10: Accuracy vs. prediction confidence. The IG Variance model shows a higher R^2 value, indicating that its confidence is more predictive of accuracy, thus better calibrated.

ing or additional techniques may be necessary to fully realize its potential. For example, combining this method with other debiasing strategies, adversarial training, or contrastive learning may yield more pronounced improvements.

Finally, the focus here was on a single dataset (SNLI) and one model architecture (ELECTRA-small). Future work should investigate whether the variance penalty generalizes well across a range of NLI benchmarks and model sizes, and explore more principled methods for setting the penalty's hyperparameters. Extending the approach to other tasks beyond NLI could also clarify the broader utility of attribution-guided regularization. Overall, while incremental, the results presented lay the groundwork for using integrated attribution signals not only as diagnostic tools but also as a means to steer models toward more reliable linguistic reasoning.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *Preprint*, arXiv:2009.07896.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.