Identifying the Source of Vulnerability in Fragile Interpretations: A Case Study in Neural Text Classification

Anonymous ACL submission

Abstract

Prior works mainly used input perturbation methods for testing stability of post-hoc interpretation methods and observed fragile interpretations. However, different works show conflicting results on the primary source of fragile interpretations because input perturbation can cause potential effects on the model and the interpretation methods. Instead, this work proposes a simple *output perturbation* method that circumvents models' potential effects by slightly modifying the prediction probability. We evaluate the proposed method using two popularly-used post-hoc interpretation methods (LIME and Sample Shapley), and CNN, LSTM, and BERT as the neural classifiers. The results show that post-hoc methods produce only slightly different interpretations under output perturbation. It suggests that the black-box model is the primary source causing fragile interpretations.

1 Introduction

Interpretation methods have been developed to generate interpretations to provide insights into the model decision-making process because of the opacity of neural networks models' prediction process (Du et al., 2019). Recent works have raised concerns on interpretation robustness (Ghorbani et al., 2019; Slack et al., 2020; Yeh et al., 2019). Slack et al. (2020) show post-hoc interpretation methods are not stable to input perturbation due to the observation of fragile interpretations. Yet, Ghorbani et al. (2019) conclude that input perturbation may cause fragile interpretations, but the perturbation might not influence post-hoc interpretation methods. Two conflicting conclusions raise the concern about how to assess the stability of post-hoc interpretation methods. In this work, we propose to analyze the stability of post-hoc interpretation methods by tracing the primary source of fragile interpretations.

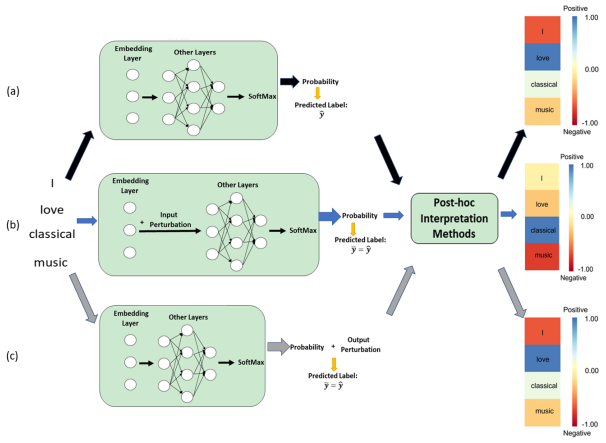

Figure 1: The pipeline to generate interpretations by a post-hoc method (a) with no perturbation applied; (b) with input perturbation applied; (c) with output perturbation applied. We apply colors to represent sentiment polarities in interpretations. We use blue for positive and red for negative.

Prior work assesses the stability of post-hoc interpretation methods by applying perturbation on model inputs, which is not sufficient for concluding interpretation methods are not stable. As illustrated in Figure 1(a) and (b), we demonstrate examples of the process that generate interpretations with no perturbation and with input perturbation applied (at the input text embedding), respectively. Figure 1(a) shows the standard pipeline of using post-hoc interpretation methods, and Figure 1(b) shows the one of applying input perturbation for assessing stability. One intuitive thing is that the random input perturbation has to go through both the model and the interpretation method. Therefore, when interpretations are inconsistent due to the random input perturbation, it is difficult to tell whether it is the model or the interpretation method that causes the inconsistency.

To circumvent the potential influence on interpretations from the model's vulnerability in this work, we design an output perturbation method

by slightly modifying the prediction probability of black-box models, shown in Figure 1(c). We propose to apply a small noise directly to the prediction probability of the model in order to isolate the influence from classification models.

First, we conduct an experiment of perturbation comparison to assess the stability of post-hoc interpretation methods by identifying the primary source causing fragile interpretations. Our analysis shows that the primary reason for fragile interpretations is the model vulnerability, and the stability of interpretation methods only has some minimal contributions. Based on the previous result, we design an experiment of evaluating the effects of different perturbations. It provides a potential explanation that input perturbation might enlarge models' vulnerability that reflects in prediction probability and indirectly causes fragile interpretations.

2 Output Perturbation Method

For text classification, $\boldsymbol{x}$ denotes the input text consisting of N words, $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}]$, with each component $\boldsymbol{x}^{(n)} \in R^d$ representing the n-th word embedding. We define a classifier as $f(\cdot)$ and the output probability of a given $\boldsymbol{x}$ on label k is $P(y = k \mid \boldsymbol{x}) = f_k(\boldsymbol{x})$, where $k \in \{1, \ldots, C\}$ and C is the total number of classes.

Let $I(\boldsymbol{x}, \hat{y}, P)$ denote the interpretation for the model prediction on $\boldsymbol{x}$, where $\hat{y}$ is the predicted label and P represents output probabilities. Let $I(\bar{\boldsymbol{x}}, \hat{y}, \bar{P})$ denote the interpretation for the model prediction on the perturbed input $\bar{\boldsymbol{x}}$, where the perturbed input $\bar{\boldsymbol{x}} = \boldsymbol{x} + \varepsilon_{input}$. $\bar{P}$ represents output probabilities that are caused by the perturbed input $\bar{\boldsymbol{x}}$. Similarly, we apply $I(\boldsymbol{x}, \hat{y}, \widetilde{P})$ as the interpretation for the perturbed output probabilities $\widetilde{P} = P + \varepsilon_{output}$. Let the interpretation change between $I(\bar{\boldsymbol{x}}, \hat{y}, \bar{P})$ and $I(\boldsymbol{x}, \hat{y}, P)$ denote δ_1. Let the interpretation change between $I(\boldsymbol{x}, \hat{y}, \widetilde{P})$ and $I(\boldsymbol{x}, \hat{y}, P)$ denote δ_2.

Previous work applies the input perturbation method to access the stability of post-hoc interpretation methods. Fragile δ_1 is observed and post-hoc interpretation methods are considered unstable. But with different inputs, the model may behave differently, e.g. $P \neq \bar{P}$, which would reflect in post-hoc interpretations, though may not in prediction labels, e.g. $\hat{y} = \bar{y}$. We propose a method to evaluate δ_2 by directly adding perturbation to model output probabilities. The method can circumvent the potential model influence.

Output Perturbation. Given an example $\boldsymbol{x}$, we add a small perturbation to model output probabilities $\{P(y = k \mid \boldsymbol{x}) + \varepsilon_{output}\}_{k=1}^{C}$, where each $\varepsilon_{output} \sim \mathcal{N}(0, \sigma^2)$, σ^2 is the variance of Gaussian distribution. To guarantee the modified $\{P(y = k \mid \boldsymbol{x}) + \varepsilon_{output}\}_{k=1}^{C}$ are still legitimate probabilities, we further normalize them as

$$\widetilde{P}(y = k \mid \boldsymbol{x}) = \frac{P(y = k \mid \boldsymbol{x}) + \varepsilon_{output}}{\sum_{i=1}^{C} P(y = i \mid \boldsymbol{x}) + \varepsilon_{output}} \tag{1}$$

Similar to finding an interpretation of the original prediction, the interpretation of perturbed outputs is computed based on $\widetilde{P}(y = \hat{y} \mid \boldsymbol{x})$.

Example: Output perturbation in LIME. To give an example of the proposed perturbation idea, we would like to represent $\boldsymbol{r'}$ as the bag-of-words representations of the original input texts. A simplified version [1] of LIME is equivalent to finding a solution of the following linear equation:

$$\boldsymbol{w}_{\hat{y}}^T \boldsymbol{r'} = \widetilde{\boldsymbol{p}}_{\hat{y}} \tag{2}$$

where $\widetilde{\boldsymbol{p}}_{\hat{y}} = [\widetilde{P}(y = \hat{y} \mid \boldsymbol{x}), \widetilde{P}(y = \hat{y} \mid \boldsymbol{z}_1), \ldots, \widetilde{P}(y = \hat{y} \mid \boldsymbol{z}_L)]^T$ are the perturbed probabilities on the label $\hat{y}$, and $\boldsymbol{w}_{\hat{y}}^T$ is the weight vector, where each element measures the contribution of an input word to the prediction $\hat{y}$. A typical interpretation from LIME consists of top important words according to $\boldsymbol{w}_{\hat{y}}$. Essentially, our proposed output perturbation is similar to the perturbation analysis in linear systems (Golub and Van Loan, 2013), which aims to identify the stability of these systems. Despite the simple formulation in Equation 2, a similar linear system can also be used to explain the Shapley-based interpretation methods (e.g., Sample Shapley (Strumbelj and Kononenko, 2010)). We leave the detailed description out due to the page limitation.

3 Experiments and Results

3.1 Experimental Setup

Models. We apply three neural network models, Convolutional Neural Network (Kim, 2014, CNN), Long Short Term Memory network (Hochreiter and Schmidhuber, 1997, LSTM), and Bidirectional Encoder Representations from Transformers (Devlin et al., 2018, BERT).

[1] Without the example weight computed from a kernel function and the regularization term of interpretation complexity.

Datasets. We adopt four text classification datasets: IMDB movie reviews dataset (Maas et al., 2011, IMDB), AG's news dataset (Zhang et al., 2015, AG's News), Stanford Sentiment Treebank dataset with binary labels (Socher et al., 2013, SST-2), and 6-class questions classification dataset TREC (Li and Roth, 2002, TREC).

Post-hoc Interpretation Methods. We adopt two post-hoc interpretation methods, LIME (Ribeiro et al., 2016) and Sample Shapley (Strumbelj and Kononenko, 2010). LIME and Sample Shapley are additive feature attribution methods. The additive feature method provides a feature importance score on every feature for each text input based on model prediction. The summary statistics of datasets are shown in Table 2 in Appendix refapppendix:table1.

In the experiment of perturbation comparison, we apply all models, datasets, and post-hoc interpretation methods listed for evaluation. In the experiment of evaluating the effects of different perturbations experiment, we apply the CNN model, the SST-2 dataset, and the LIME method.

Evaluation Metrics. In the experiment of perturbation comparison (subsection 3.2), we apply Kendall's Tau order rank correlation score, and the top-k important words overlap score as two evaluation metrics. We use the first metric to evaluate the discrepancy in feature scores on each feature in the same text input between different perturbation levels. We use the second metric to evaluate the discrepancy on the ordered feature indices (rank by order of feature scores) on the same text input between different perturbation levels. In this work, we set $k = 5$.

When evaluating the effects of different perturbations (subsection 3.3), We apply the relative entropy between the original probability distribution and the perturbed probability distribution as the evaluation metric. Let $D_{KL}(P||\widetilde{P})$ denote the relative entropy between the original probability distribution and the output perturbed probability distribution. Let the relative entropy between the original probability distribution and the input perturbed probability distribution denote as $D_{KL}(P||\bar{P})$. We calculate the mean and the standard deviation of relative entropy to represent the situation of the entire dataset.

Dataset	CNN	LSTM	BERT
IMDB	86.30	86.60	92.00
SST-2	82.00	81.05	91.43
AG's News	89.80	90.25	94.80
TREC	90.80	92.40	97.20

Table 1: Prediction accuracy(%) of three models on the four benchmark datasets.

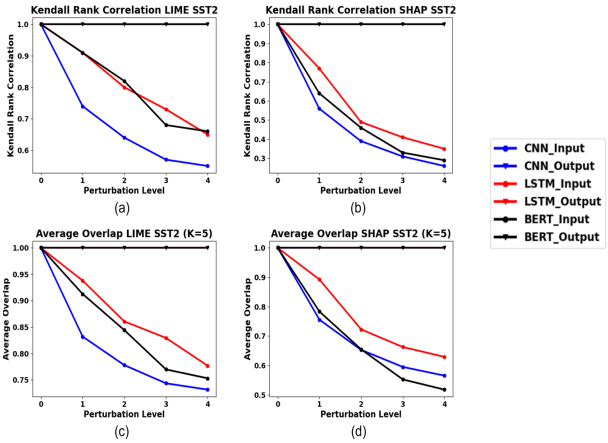

Figure 2: Results on SST-2 dataset in comparison experiments. (a), (b) are results of Kendall's Tau order rank correlation score in. (c), (d) are results of top-k word importance score.

3.2 Perturbation Comparison

To explore the primary source causing interpretation vulnerability, we conduct comparison experiments to evaluate and compare δ_1, and δ_2. δ_1 represents the interpretation discrepancy of $I(\boldsymbol{x}, \hat{y}, P)$ and $I(\bar{\boldsymbol{x}}, \hat{y}, \bar{P})$. δ_2 represents the interpretation discrepancy of $I(\boldsymbol{x}, \hat{y}, P)$ and $I(\bar{\boldsymbol{x}}, \hat{y}, \widetilde{P})$. If δ_1 and δ_2 have an obvious gap, it illustrates that the black-box model is more likely the primary source causing fragile interpretations.

We train the three models on the four benchmark datasets to ensure the performance on all models is acceptable. Prediction accuracy is recorded in Table 1. In this experiment, we select an input perturbation method that is directly adding random noise to the input word embeddings (Liu et al., 2020), as shown in Figure 1(b). We apply the Gaussian distribution to generate perturbation and control the level by modifying the variance of Gaussian distribution σ^2_{input}. Detailed values of input perturbation levels and output perturbation levels applied are shown in Table 3 in Appendix subsection B.2.

We display plots of results on the SST-2 dataset in Figure 2. Kendall's Tau order rank correlation

score plots are shown in Figure 2(a) and (b). Top-k important words overlap score plots are shown in Figure 2(c) and (d). Due to page limitations, we include plots on other datasets in Figure 4, Figure 5 and Figure 6 in Appendix B, which are indicating a similar tendency and conclusion.

Kendall's Tau order rank correlation score results. Kendall's Tau order rank correlation score results indirectly illustrate the stability of post-hoc interpretation methods. In Figure 2(a) and (b), plots display apparent discrepancies between δ_1 against δ_2 on the evaluation metric in LIME and Sample Shapley. For output perturbation method results, it is noticeable that the values of Kendall's Tau order rank correlation scores remain the same with the increased perturbation levels. This tendency indicates that, for a given input, if $\boldsymbol{x}$ and $\hat{y}$ stay unchanged, the output perturbation ε_{output} is unlikely to influence interpretations generated by post-hoc interpretation methods. In other words, the interpretation vulnerability is unlikely caused by post-hoc interpretation methods. Meanwhile, for input perturbation method results, it is notifiable that the values of Kendall's Tau order rank correlation scores decrease obviously with the increase of input perturbation intensity levels. It means that the black-box model becomes more vulnerable to the input perturbation and causes fragile interpretations. Compared to the previous result, the black-box model is more likely to be the primary source causing fragile interpretations.

Top-k word importance score results. Top-k word importance score plots reflect the same result: the model is the primary source causing fragile interpretations. In Figure 2(c) and (d), δ_1 against δ_2 displays an obvious discrepancy in both post-hoc interpretation methods as well. For output method results, δ_2 displays no change on the overlap score of the k most important words. The result in this metric indicates that the black-box model is more likely to be the source that causes interpretation vulnerability compared to interpretation methods.

3.3 Effects of Different Perturbations

Based on the previous results, input perturbation causes vulnerability of the black-box models to generate fragile interpretations indirectly . It is natural to consider the potential models' effects of different perturbations. Unfortunately, there is no direct answer to this question. The reason is that

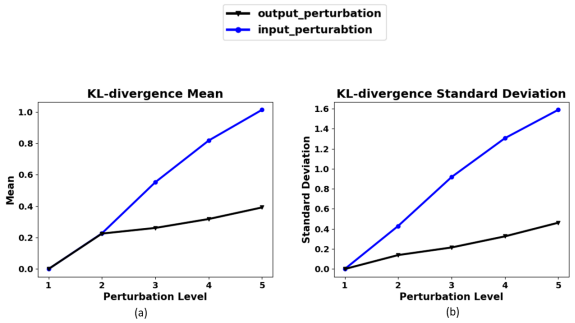

Figure 3: Effects of different perturbations experiment results. (a) the mean of the relative entropy. (b) the standard deviation of the relative entropy.

we cannot directly compare the effects of perturbations that are applied at different positions of the pipeline. Therefore, we propose an experiment to provide simple insights into the effects of different perturbations. The basic idea is to compare the relative entropy of prediction probability under output perturbation with the relative entropy of the prediction probability under input perturbation.

Results show that the mean and the standard deviation of $D_{KL}(P||\widetilde{P})$ are remarkably lower than those of $D_{KL}(P||\bar{P})$, shown in Figure 3(a) and (b). It indicates that the average value and dispersion degree of prediction probability of the input perturbation are higher than those of prediction probability of output perturbation. It provides slight insight that the vulnerability of black-box models may enlarge the influence of input perturbation and deteriorate the potential models' effects, which reflects in the prediction probability. When a post-hoc interpretation method uses the influenced prediction probability, it is possible to generate fragile interpretations. The experiment result provides a potential explanation for the observation in prior works.

4 Conclusion

In this work, we propose an output perturbation method by slightly modifying the prediction probabilitiy of black-box models. The major contribution of the proposed method helps to identify the primary source causing non-robust interpretations is the black-box model. Also, we provide slight insights into the stability of post-hoc interpretation methods. Our method provides a new focus on the research of interpretations robustness in post-hoc interpretation methods with some fine-grained experiment design in future works.

References

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

Reza Ghaeini, Xiaoli Z Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *arXiv preprint arXiv:1808.03894*.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688.

Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*, volume 3. JHU press.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. *arXiv preprint arXiv:1611.07634*.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. *arXiv preprint arXiv:1902.02041*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2017. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Lloyd S Shapley. 2016. *17. A value for n-person games*. Princeton University Press.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Erik Strumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18.

Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. 2019. Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 2020–2029.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of nlp models is manipulable. *arXiv preprint arXiv:2010.05419*.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. 2019. On the (in) fidelity and sensitivity for explanations. *arXiv preprint arXiv:1901.09392*.

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cédric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the lack of robust interpretability of neural text classifiers. *arXiv preprint arXiv:2106.04631*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

A Related Works

Post-hoc interpretations. Most work focuses on explaining neural network models from the post-hoc manner, especially generating a local interpretation for each model prediction. The white-box interpretation methods, such as gradient-based interpretations (Hechtlinger, 2016) and attention-based interpretations (Ghaeini et al., 2018), either require additional information (e.g. gradients) from the model or incur much debates regarding their faithfulness to model predictions (Jain and Wallace, 2019). Another line of work focuses on explaining black-box models in a model-agnostic way. For example, Li et al. (2016) proposed a perturbation-based explanation method, Leave-one-out, that attributes feature importance to model predictions by erasing input features one by one. Ribeiro et al. (2016) proposed to estimate feature contributions locally via linear approximation based on pseudo examples. Some other works proposed the variants of Shapley value (Shapley, 2016) to measure feature importance, such as Sample Shapley (Strumbelj and Kononenko, 2010), KernelSHAP (Lundberg and Lee, 2017), and L/C-Shapley (Chen et al., 2018). In this work, we focus on two well-adopted black-box interpretation methods, LIME (Ribeiro et al., 2016) and Sample Shapley (Strumbelj and Kononenko, 2010), and their robustness to input perturbation.

Model robustness. Recent works have shown the vulnerability of model prediction robustness to adversarial attacks (Szegedy et al., 2013; Zhao et al., 2017). Adversarial examples are similar to original examples but can quickly flip model predictions (Jia et al., 2019). In the text domain, a common way to generate adversarial examples is by heuristically manipulating the input text, such as replacing words with their synonyms (Ren et al., 2019; Jin et al., 2020), inserting/removing words (Liang et al., 2017), or concatenating triggers (Wallace et al., 2019). Unlike these works that modify original input texts, we add noise at model outputs and disentangle the sources of fragile interpretations.

Interpretation robustness. Previous work explored interpretation robustness by either perturbing the inputs (Ghorbani et al., 2019; Subramanya et al., 2019; Zhang et al., 2020; Heo et al., 2019) or manipulating the model (Wang et al., 2020; Slack et al., 2020; Zafar et al., 2021). For example, Slack et al. (2020) fooled post-hoc interpretation methods by hiding the bias for black-box models based on the proposed novel scaffolding technique. However, all of these works cannot disentangle the sources that cause fragile interpretations. Differently, our proposed method mitigates the influence of model to the interpretations by perturbing model outputs.

B More Tables and Figures

B.1 Table of Dataset Summary Statistics

Table 2 displays the summary statics on four datasets, IMDB movie reviews dataset (Maas et al., 2011, IMDB), AG's news dataset (Zhang et al., 2015, AG's News), Stanford Sentiment Treebank dataset with binary labels (Socher et al., 2013, SST-2), and 6-class questions classification dataset TREC (Li and Roth, 2002, TREC) .

Dataset	C	L	$\#train$	$\#dev$	$\#test$	vocab	threshold	length
IMDB	2	268	20K	5K	25K	29571	5	250
SST-2	2	19	6920	872	1821	16190	0	50
AG's News	4	32	114K	6K	7.6K	21838	5	50
TREC	6	10	5000	452	500	8026	0	15

Table 2: Summary statistics for the datasets where C is the number of classes, L is the average sentence length, # counts the number of examples in train/dev/test sets, vocab is the vocab size, threshold is low-frequency threshold, and length is mini-batch sentence length.

B.2 Table of Detailed Value of σ^2

Table 3 displays the detailed value of σ^2_{input} and σ^2_{output} that represents the detailed perturbation levels we applied in interpretation robustness experiments.

Dataset	**Model**	**Input Perturbation Level (σ^2_{input})**					**Output Perturbation Level (σ^2_{output})**				
		0	1	2	3	4	0	1	2	3	4
IMDB	CNN	0	0.05	0.14	0.18	0.2	0	0.25	0.50	0.75	1
	LSTM	0	0.05	0.13	0.16	0.18	0	0.25	0.50	0.75	1
	BERT	0	0.25	0.50	0.85	1	0	0.25	0.50	0.75	1
SST-2	CNN	0	0.16	0.26	0.33	0.38	0	0.25	0.50	0.75	1
	LSTM	0	0.05	0.13	0.18	0.24	0	0.25	0.50	0.75	1
	BERT	0	0.25	0.50	0.75	0.85	0	0.25	0.50	0.75	1
AG's News	CNN	0	0.08	0.14	0.18	0.24	0	0.25	0.38	0.50	0.75
	LSTM	0	0.05	0.13	0.16	0.18	0	0.25	0.38	0.50	0.75
	BERT	0	0.25	0.50	0.75	1.25	0	0.25	0.38	0.50	0.75
TREC	CNN	0	0.04	0.05	0.09	0.11	0	0.25	0.33	0.41	0.50
	LSTM	0	0.08	0.12	0.19	0.21	0	0.25	0.33	0.41	0.50
	BERT	0	0.25	0.50	0.75	0.85	0	0.25	0.50	0.75	1

Table 3: Perturbation levels applied on four datasets. Input Perturbation Level (σ^2_{input}) represents the input perturbation applied. Output Perturbation Level (σ^2_{output}) represents the output perturbation applied.

B.3 Figures of IMDB Dataset

Figure 4 displays results of IMDB dataset. Kendall's Tau order rank correlation score results are shown in Figure 4(a) and (b). Top-k important words overlap score results are shown in Figure 4(c) and (d).

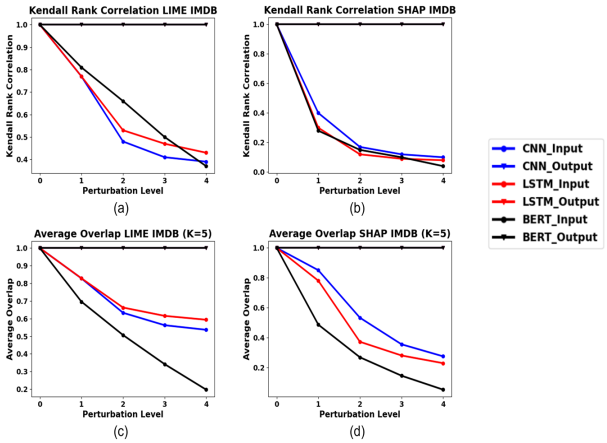

Figure 4: Results on IMDB dataset in comparison experiments. (a), (b) are results of Kendall's Tau order rank correlation score in. (c), (d) are results of top-k word importance score.

B.4 Figures of AG's News Dataset

Figure 5 displays results of AG's News dataset. Kendall's Tau order rank correlation score results are shown in Figure 5(a) and (b). Top-k important words overlap score results are shown in Figure 5(c) and (d).

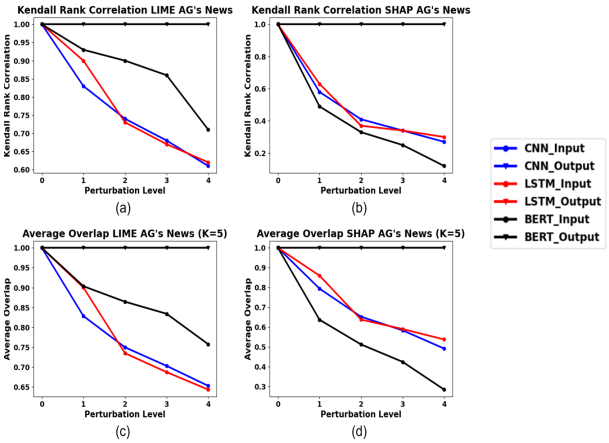

Figure 5: Results on AG's News dataset in comparison experiments. (a), (b) are results of Kendall's Tau order rank correlation score in. (c), (d) are results of top-k word importance score.

B.5 Figures of TREC Dataset

Figure 6 displays results of TREC dataset. Kendall's Tau order rank correlation score results are shown in Figure 6(a) and (b). Top-k important words overlap score results are shown in Figure 6(c) and (d).

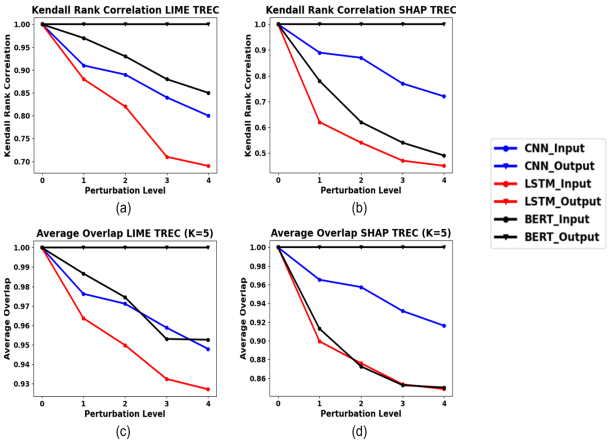

Figure 6: Results on TREC dataset in comparison experiments. (a), (b) are results of Kendall's Tau order rank correlation score in. (c), (d) are results of top-k word importance score.