On the Faithfulness Measurements for Model Interpretations

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang
University of California, Los Angeles
{fanyin20, zshi, chohsieh, kwchang}@cs.ucla.edu;

Abstract

Recent years have witnessed the emergence of a variety of post-hoc interpretations that aim to uncover how natural language processing (NLP) models make predictions. Despite the surge of new interpretations, it remains an open problem how to define and quantitatively measure the faithfulness of interpretations, i.e., to what extent they conform to the reasoning process behind the model. To tackle these issues, we start with three criteria: the removal-based criterion, the sensitivity of interpretations, and the stability of interpretations, that quantify different notions of faithfulness, and propose novel paradigms to systematically evaluate interpretations in NLP. Our results show that the performance of interpretations under different criteria of faithfulness could vary substantially. Motivated by the desideratum of these faithfulness notions, we introduce a new class of interpretation methods that adopt techniques from the adversarial robustness domain. Empirical results show that our proposed methods achieve top performance under all three criteria. Along with experiments and analysis on both the text classification and the dependency parsing tasks, we come to a more comprehensive understanding of the diverse set of interpretations.

1 Introduction

As complex NLP models being widely deployed in various real-life applications, there is an increasing interest in opening the black-box and understanding why these models come to certain decisions. To this end, the line of research on interpretation techniques grows rapidly, facilitating a broad range of model understanding, from building user trust on models (Ribeiro et al., 2016; Hase and Bansal, 2020) to exposing subtle biases (Zhao et al., 2017; Doshi-Velez and Kim, 2017).

In this paper, we focus on post-hoc interpretations in NLP. Given a trained model and a specific input text, post-hoc interpretations assign an importance score to each token in the input to indicate its contribution to the model output. Current methods in this direction can be roughly divided into three categories: gradient-based methods (Simonyan et al., 2014; Li et al., 2016); reference-based methods (Sundararajan et al., 2017; Shrikumar et al., 2017); and perturbation-based methods (Zeiler and Fergus, 2014; Ribeiro et al., 2016).

Despite the emergence of new techniques, one critical issue is that there is little consensus on how to define and evaluate the faithfulness of these techniques, i.e., whether they generate explanations that reflect the real reasoning process behind the model. A widely employed criterion, especially in NLP, is the removal-based criterion (Zaidan and Eisner, 2008; DeYoung et al., 2020), which removes or only preserves important tokens given by an interpretation and observe how much the model prediction would change. However, this criterion has obvious limitations as pointed out in prior work (Bastings and Filippova, 2020; Ancona et al., 2018): the corrupted version of an input produced by this criterion might fall out of the distribution that models are trained on, and thus results in an inaccurate measurement of faithfulness. It also has an intrinsic preference to interpretations that directly optimize its objective. Therefore, we complement the removal-based criterion with two other quantitative criteria, the sensitivity of interpretations and the stability of interpretations, which are previously overlooked in NLP. The three criteria represent different notions and desired properties of faithfulness.

Sensitivity is based on the notion that model predictions should be easily changed when adding noise to the set of important tokens identified by a faithful explanation. In contrast to the removal-based criterion, which induces global perturbations, we add small but adversarial perturbations under a
local region in the embedding space. This is a natural improvement that avoids generating incomplete texts when conducting evaluations. This criterion is recently discussed in Hsieh et al. (2020), mainly for computer vision applications, while ours provides a comprehensive analysis on NLP models and tasks.

Stability assumes that a faithful interpretation should not produce substantially different explanations for two inputs that the model finds similar. There are several attempts to generate such a pair of inputs. The most relevant work to ours is Ghorbani et al. (2019). However, their method is only applicable to differentiable interpretations. Our work proposes a new paradigm that employs a black-box algorithm that generates a semantically related neighbor of the original input, which is especially designed for NLP tasks and applicable to all interpretation techniques.

Experimental results with the three criteria suggest that there is a big gap between the behavior of gradient-based methods and reference-based methods though both of them rely on the back-propagation process of a model. Inspired by the pros and cons of these two classes, we propose a new class of interpretations using the projected gradient descent (PGD) attack (Madry et al., 2018) and a robustness verification approach (Shi et al., 2020; Xu et al., 2020) from the adversarial robustness domain. Our methods consider a larger region than gradient-based methods but not “too large” that hurts the sensitivity criterion as reference-based methods do. We empirically demonstrate that they achieve remarkable performance among all criteria.

Besides experiments on three text classification datasets, we provide an initial attempt to apply interpretation methods and criteria to the structured prediction task in NLP. We propose a new paradigm that focuses on the prepositional phrase (PP) attachment ambiguity to interpret the dependency parsing task. Our experiments involve two popular model architectures: BERT (Devlin et al., 2019) and Bidirectional-LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997).

Our contributions can be summarized as follows:

1. We consider three notions of faithfulness for interpretation methods, which are previously overlooked in NLP. We propose quantitative criteria and conduct systematical experiments to evaluate interpretation methods under the three notions.

2. We propose a new class of interpretation methods, which draws the connection between the adversarial robustness domain and the interpretation domain. We demonstrate that the new methods are effective under all three criteria.

3. We propose a new paradigm to evaluate model interpretations on the dependency parsing task.

2 Faithfulness Evaluation Criteria

By definition, a faithful post-hoc interpretation should identify the most important part of the input that a model truly relies on when making its decision. Formally, let \( x = [x_1; x_2; \ldots; x_n] \) be a sequence of tokens. \( e(\cdot) \) denotes the token embedding function. A neural NLP model \( f : X \rightarrow Y \) takes \( x \) as input and provides its prediction \( f(x) = y \) at this single data point. We use \( s(e(x)) \) to denote the output score of \( f \) on \( y \) from the token embeddings. An interpretation method \( \phi \), given the model \( f \), the input \( x \), and the prediction \( y \), generates a list of importance scores \( \phi(x) = [\phi(x_1); \phi(x_2); \phi(x_3); \ldots; \phi(x_n)] \) that indicates how important each token is to the model decision. We use these notations in all later sections.

We first define the three criteria for quantitatively evaluating different notions of interpretation faithfulness. The first criterion is well-established in previous literature of interpretations in NLP like (DeYoung et al., 2020). We briefly review it here and emphasize its relation to the two criteria sensitivity and stability, for which we propose novel paradigms to adapt them to various NLP tasks.

Removal-based Criterion A well-established property of a faithful interpretation is that the presence of a set of important tokens identified by a faithful interpretation should have more meaningful influence on the model’s decision than an arbitrary set of tokens. The removal-based criterion is proposed to quantify this notion by measuring how much the model performance would drop after the set of the “most relevant” tokens in an explanation is removed. Formally, we use \( r_k \) to denote the sequence of the top \( k \) most relevant tokens and \( x \setminus r_k \) to denote the input after \( r_k \) is removed. To avoid choosing an unjustified size of \( r_k \), we vary \( k \) across a list of candidate thresholds, denoted as \( K \).

\footnote{Following common setups, we replace an important word by a reference token as a proxy for removing it.}
and summarize the corresponding score on each threshold. The metric is computed as:

\[
\frac{1}{|K|} \sum_{k} \left( s(e(x)) - s(e(x \setminus r_k)) \right).
\]

(1)

Following DeYoung et al. (2020), we call this metric the comprehensiveness (comp.) score. A higher comprehensiveness score means the set of tokens are more influential to the model prediction, and thus a more faithful explanation. On the other hand, we can also solely preserve the most relevant tokens - according to the interpretation - and observe to what extent the original model performance is maintained. This is referred to as the sufficiency (suff.) score in previous literature:

\[
\frac{1}{|K|} \sum_{k} \left( s(e(x)) - s(e(r_k)) \right).
\]

(2)

A lower sufficiency score suggests a more faithful explanation. Note that completely masking out parts of the input produces incomplete sentences. Such a large and global perturbation could lead to several issues as pointed out by prior studies (Feng et al., 2018; Bastings and Filippova, 2020).

**Sensitivity** Instead of removing relevant tokens, the sensitivity criterion adds local but adversarial noise to the embedding vectors of the relevant tokens and measures the magnitude of the noise needed to change the model prediction. This is inspired by the notion that models should be more sensitive to perturbations being added to the most relevant tokens identified by a faithful interpretation compared to a random or irrelevant set of tokens. From the adversarial robustness perspective (Goodfellow et al., 2015), this notion requires that we can reach the local decision boundary of a model with the minimum perturbation on the most relevant tokens.

We call the metric that quantify this property sensitivity (sens.). Given the relevant tokens \( r_k \) and its corresponding embedding vector \( e(r_k) \), we compute the minimal perturbation magnitude on \( e(r_k) \), denoted by \( \epsilon_{r_k} \), that changes the model prediction at this data point:

\[
\epsilon_{r_k} = \min \| \delta_{r_k} \|_2 \quad \text{s.t.} \quad f(x + \delta_{r_k}) \neq y, \quad (3)
\]

where \( \| \cdot \|_2 \) is the magnitude under \( L_2 \) norm, and \( \delta_{r_k} \) denotes the perturbation being added to \( e(r_k) \), which is concatenated to a vector when computing its \( L_2 \) norm. Since the exact computation of \( \epsilon_{r_k} \) is intractable, we instead use the PGD attack (Madry et al., 2018) with binary search to approximate \( \epsilon_{r_k} \), which is an upper bound of the real value. A lower \( \epsilon_{r_k} \) means a more faithful interpretation. In practice, we vary the size of \( r_k \), compute multiple \( \epsilon_{r_k} \), and summarize them with the area under the curve (AUC) score.

**Stability** Another desired property of faithfulness is that a faithful interpretation should not give substantially different token importance orders for two input points that the model finds similar. To construct a pair of similar inputs, we propose a novel paradigm to generate a contrast example to the original example under some similarity constraints. Then, we measure the difference in token importance ranks between this pair of inputs.

Specifically, given the input text \( x \) and an importance order list of \( x \) derived from \( \phi(x) \), denoted as \( m(x) \) here, we intend to find a contrast example of \( x, \tilde{x} \), that (1) has at most \( k \) tokens different from \( x \); (2) the model output score at \( \tilde{x} \) changes by no more than \( \tau \) compared to that score on \( x \); (3) leads to the largest rank difference \( D \) between \( m(x) \) and the alternated importance order \( m(\tilde{x}) \):

\[
\arg \max_{\tilde{x}} D(m(x), m(\tilde{x})),
\]

\[
s.t. \quad |s(e(x)) - s(e(\tilde{x}))| \leq \tau \quad \text{and} \quad \|x - \tilde{x}\|_0 \leq k. \quad (4)
\]

To achieve this, we first extract synonyms for each token \( x_i \) in \( x \) following Alzantot et al. (2018). Then, in the decreasing order of \( m(x) \), we greedily search for a substitution of each token that induces the most substantial change in \( m(x) \) and repeat this process until the model output score changes by more than \( \tau \) or the pre-defined constraint \( k \) is reached. Finally, we measure the difference between two token importance ranks using Spearman’s rank order correlation (Spearman, 1961). We call this criterion stability (stah.). A higher score indicates a more faithful interpretation.

Note that instead of using the gradient information of interpretations to find contrast examples like Ghorbani et al. (2019), our algorithm treats interpretations as black-box, which makes it applicable to all kinds of interpretations, including non-differentiable ones. The paradigm of generating similar textual inputs through semantically equivalent word substitutions is also natural for NLP applications (Alzantot et al., 2018).
3 Adversarial Robustness Techniques as Interpretations

In this section, we define a new class of interpretation methods by adopting techniques in the adversarial robustness domain and describe its relation to existing interpretations. To start with, we give a brief review of existing interpretations and then introduce our new methods.

We roughly divide existing techniques into three categories: gradient-based methods, reference-based methods, and perturbation-based methods, and discuss representatives of them here.

Gradient-based methods This first class of methods leverage local gradient information to approximate the contribution of some parts of the input to the model output. We consider the Vanilla Gradient (VaGrad) method (Simonyan et al., 2014) and the Gradient×Input (GradInp) method (Li et al., 2016). Following the notations in the last section, VaGrad represents the importance score for each token by the $L_2$ norm of the (signed) partial derivatives of the model output score $s(e(x))$ w.r.t each token embedding:

$$\text{VaGrad}(x)_i = \| \frac{\partial s(e(x))}{\partial e(x_i)} \|_2. \quad (5)$$

GradInp instead multiplies the gradient of the model output w.r.t the input embedding with the input embedding vector itself to represent importance scores:

$$\text{GradInp}(x)_i = e(x_i) \cdot \frac{\partial s(e(x))}{\partial e(x_i)}. \quad (6)$$

Reference-based methods This class of methods computes the importance scores by distributing the difference between the model output on a user-selected reference point and on the point of interest to the model. We consider the Integrated Gradient (IngGrad) method (Sundararajan et al., 2017) and the DeepLIFT (Shrikumar et al., 2017) method in this class. IngGrad computes the average gradient while the input varies along a linear path from a user-selected reference point $\overline{x}$ to the original input point:

$$\text{IngGrad}(x)_i = (e(x_i) - e(\overline{x}_i)) \cdot \int_{\alpha=0}^{1} \frac{\partial s(e(\overline{x}) + \alpha (e(x) - e(\overline{x})))}{\partial e(x_i)} d\alpha. \quad (7)$$

DeepLIFT decomposes the relative effect of an inner neuron that activated at the original input point compared to a reference point proportionally and back-propagates the effect to each input token as its importance score. We use DeepLIFT with the Rescale rule as in (Ancona et al., 2018). Notice that DeepLIFT diverges from IngGrad when encounters multiplicative interactions. See discussion in (Ancona et al., 2018).

Perturbation-based methods Methods in this class query model outputs on perturbed inputs. We choose the Occlusion method (Zeiler and Fergus, 2014) and LIME (Ribeiro et al., 2016). Occlusion replaces one token at a time by a reference value and uses the corresponding drop on model performance to represent the importance score of each token. LIME uses a linear model to fit model outputs on the neighborhood of input $x$ and represents token importance by the weights in the trained linear model.

Robustness-based methods As described before, IngGrad leverages the average gradient in a segment while gradient-based methods consider the gradient at a single point. We believe such a difference leads to the gap between their behaviors: gradient-based methods saturate and generate unreasonable explanations at some specific inputs. On the other hand, IngGrad is likely to neglect local properties as desired by the sensitivity criterion. Our experimental results on Section 4.2 also demonstrate this.

To address this issue, we propose a new class of interpretations that enlarges the scope of gradient-based methods while not “too large” that hurts local properties. We use approaches in the adversarial robustness domain: the PGD attack (Madry et al., 2017) and the certifying robustness method (Xu et al., 2020) as interpretations.

Instead of solely considering the gradient information at the current input, PGD takes several “mini-steps” within a pre-defined vicinity of the input to search for an adversarial example. We represent the input text $x$ by the concatenation of its embedding vectors for each token and perform $t$ iterations of the standard PGD procedure starting from $e(x^{(0)}) = e(x)$:

$$e(x^{(j)}) = \mathcal{P} \left( e(x^{(j-1)}) - \alpha \nabla s(e(x^{(j-1)})) \right). \quad (8)$$

$\mathcal{P}$ represents the operation that projects the new instance at each step back to the vicinity of $e(x)$. The final importance score for each token $x_i$ is represented by $\| e(x_i^{(t)}) - e(x_i) \|_2$. Note that different
from the PGD attack we use for approximating the sensitivity criterion, we manually decide the magnitude of the vicinity of \( e(x) \) instead of using a binary search. We add perturbations to the whole sentence at the same time and the final \( x^{(t)} \) doesn’t necessarily change the model prediction. See Appendix B for details. We call this interpretation method Vanilla PGD (VaPGD). Similar to GradInp, we can also represent the contribution of each token \( x_i \) by multiplying that with the input embedding itself \( \left( e(x_i) - e(x^{(t)}) \right) \cdot e(x_i) \). We call the later one PGD \( \times \) Input (PGDInp).

Certifying robustness algorithms aim to provide guaranteed lower and upper bounds of a model output in a vicinity of the original input. We use the linear relaxation based perturbation analysis (LiRPA) as discussed in (Shi et al., 2020; Xu et al., 2020). LiRPA looks for two linear functions that bound the model. Specifically, LiRPA computes \( w, w', b, b' \) that satisfy \( w e(x') + b \leq s(e(x')) \leq w'e(x') + b' \) for any point \( e(x') \) that lies within the \( L_2 \) ball of \( e(x) \) with size \( \delta \). We use the IBP+backward method in Xu et al. (2020), which uses Interval Bound Propagation (Gowal et al., 2018) to compute bounds of internal neurons of the model and then constructs the two linear functions with a bound back-propagation process. Finally, the importance score of the i-th token in the input is represented as \( w_i \cdot e(x_i) \), where \( w_i \) is the i-th row of \( w \). We call this method Certify.

4 Experiments on Text Classification

In this section, we present the results of interpretation methods on text classification tasks under the three criteria. Results verify the effectiveness of our new interpretation methods across all three criteria, especially the sensitivity and the stability criteria. We also find that the performance of an interpretation can be very inconsistent given the varying desideratum of the three criteria. Experiments on a structured prediction task are presented in the next section.

4.1 Experimental Setup

Datasets We conduct experiments on three text classification datasets: SST-2 (Socher et al., 2013), Yelp (Zhang et al., 2015), and AGNews (Zhang et al., 2015) following Jain and Wallace (2019)’s preprocessing approach. All of them are converted to binary classification tasks. SST-2 and Yelp are sentiment classification tasks that models predict whether a review is negative (0) or positive (1). AGNews is to discriminate between world (0) and business (1) articles. See Appendix A for statistics of the three datasets. When evaluating interpretation methods, for each dataset, we select 100 random samples (50 samples from class 0 and 50 samples from class 1) from the test set.

Models For text classification tasks, we consider two model architectures: BERT (Devlin et al., 2019) and BiLSTM (Hochreiter and Schmidhuber, 1997). Training and implementation details are presented in Appendix A.

Interpretation Methods As discussed in section 3, besides our novel robustness-based interpretations PGDInp, VaPGD, and Certify, we experiment with six interpretations from three existing categories: VaGrad, GradInp (gradient-based); IngGrad, DeepLIFT (reference-based); and Occlusion, LIME (perturbation-based). We also include a random baseline Random, where we randomly assign important scores. Comparisons with Random make sure all interpretations are extracting meaningful information. Their implementation and evaluation details are presented in Appendix B and C.

4.2 Results and Discussion

Table 1 shows the results of interpretations for BERT while Table 2 shows the results for BiLSTM. Generally, the results verify the effectiveness of robustness-based interpretations, especially VaPGD, which achieves the best performance under the sensitivity and the stability criteria for both BERT and BiLSTM compared to all other methods (except the stability with AGNews, BiLSTM). Robustness-based methods also outperform their gradient-based counterparts under the removal-based criterion, further demonstrating the strength of considering both local and global properties. Specifically, we find that PGDInp overcomes the failure of GradInp for BERT on SST-2 and AGNews.

For other interpretation methods, IngGrad performs well for BiLSTM under the removal-based criterion, which gives four out of six best numbers on three datasets. On the other hand, LIME performs well for BERT under the removal-based criterion. Occlusion sometimes achieves the best results for both BiLSTM and BERT as it directly optimizes the measurement. However, unlike robustness-based methods, which perform reasonably well
under all three criteria, these methods are inconsistent under different measurements. For example, while IngGrad achieves the best performance for BiLSTM under the removal-based criterion, it has very limited performance under the sensitivity metric for BiLSTM on SST-2 and Yelp datasets. Thus, when deploying interpretation methods, we advocate for a careful selection based on the underlying faithfulness notion you want to evaluate.

Qualitatively, given an example, we observe that the most relevant or irrelevant sets of words identified by different interpretations are highly overlapped for BiLSTM, although the exact order of importance scores might be different. Whereas for BERT, they are more diverse. We showcase an example from SST-2 in Figure 5. The BiLSTM classifier assigns a positive label to this instance. See Appendix D for the same example for BERT. A deeper red color means the token is identified as more important to the model output by an interpretation while a deeper blue color stands for a less important token. In the case of BiLSTM, the most important tokens are converged to words like snappy or clever among all interpretations. Notice that different from other interpretations, VaPGD (also VaGrad) identifies the token hate as an important token. We suspect this is because VaPGD, which leverages the $L_2$ norm of the gradient information at token embeddings, will identify both negatively and positively pertinent tokens as important tokens to model predictions. The word hate could negatively impact the model’s decision in this positive example. This also explains why VaPGD or VaGrad have a relatively limited performance under the removal-based criterion but a good performance under the sensitivity criterion: removing positively and negatively pertinent tokens at the same time would neutralize the output score. How-

Table 1: Results of evaluating different interpretation methods for BERT under three criteria and three text classification datasets. ↑ means a higher number under this metric indicates a better performance while ↓ means the opposite. The best performance across all interpretations is bolded. Certify is missed here since current certifying robustness approaches cannot be scaled to complex Transformer-based models like BERT. We verify that robustness-based methods perform the best under the sensitivity and the stability criteria.

| Methods | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Random  | 0.202  | 0.431  | 0.909  | -0.323 | 0.189  | 0.405  | 1.450  | -0.244 | 0.039  | 0.269  | 1.790  | -0.392 |
| VaGrad  | 0.391  | 0.267  | 0.608  | 0.834  | 0.288  | 0.316  | 0.965  | 0.814  | 0.204  | 0.099  | 1.145  | 0.825  |
| GradInp | 0.260  | 0.364  | 0.892  | 0.294  | 0.284  | 0.353  | 1.227  | 0.568  | 0.073  | 0.206  | 1.496  | 0.313  |
| Occlusion | 0.550  | 0.178  | 0.695  | 0.574  | 0.541  | 0.203  | 1.104  | 0.682  | 0.177  | 0.183  | 1.530  | 0.594  |
| LIME    | 0.617  | 0.174  | 0.669  | 0.426  | 0.532  | 0.253  | 1.168  | 0.006  | 0.389  | 0.051  | 1.287  | 0.095  |
| IngGrad | 0.466  | 0.258  | 0.794  | 0.680  | 0.489  | 0.217  | 1.178  | 0.798  | 0.209  | 0.186  | 1.365  | 0.733  |
| DeepLIFT | 0.260  | 0.364  | 0.892  | 0.294  | 0.284  | 0.353  | 1.227  | 0.568  | 0.073  | 0.206  | 1.496  | 0.313  |
| PGD     | 0.414  | 0.264  | 0.629  | 0.590  | 0.303  | 0.335  | 1.020  | 0.610  | 0.200  | 0.099  | 1.106  | 0.553  |
| VaPGD   | 0.397  | 0.254  | 0.607  | 0.835  | 0.293  | 0.310  | 0.945  | 0.838  | 0.223  | 0.092  | 1.062  | 0.877  |

Table 2: Results of evaluating different interpretation methods for BiLSTM under three criteria and three text classification datasets.

| Methods | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ | Comp.↑ | Suff.↑ | Sens.↑ | Stab.↑ |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Random  | 0.132  | 0.240  | 5.150  | -0.316 | 0.054  | 0.227  | 13.973 | -0.242 | 0.059  | 0.170  | 13.408 | -0.378 |
| VaGrad  | 0.191  | 0.208  | 3.348  | 0.870  | 0.148  | 0.093  | 9.037  | 0.882  | 0.058  | 0.151  | 9.558  | 0.836  |
| GradInp | 0.496  | 0.013  | 4.391  | 0.662  | 0.551  | -0.049 | 11.576 | 0.807  | 0.431  | -0.019 | 12.512 | 0.830  |
| Occlusion | 0.593  | -0.028 | 4.528  | 0.760  | 0.748  | -0.056 | 11.568 | 0.811  | 0.513  | -0.022 | 11.966 | 0.845  |
| LIME    | 0.609  | -0.023 | 4.412  | 0.601  | 0.391  | 0.004  | 12.121 | 0.122  | 0.556  | -0.023 | 11.441 | 0.440  |
| IngGrad | 0.599  | -0.029 | 4.630  | 0.772  | 0.776  | -0.055 | 12.304 | 0.837  | 0.590  | -0.023 | 12.512 | 0.825  |
| DeepLIFT | 0.534  | -0.015 | 4.535  | 0.632  | 0.600  | -0.052 | 11.588 | 0.810  | 0.464  | -0.025 | 11.193 | 0.751  |
| PGD     | 0.538  | -0.002 | 4.320  | 0.681  | 0.631  | -0.052 | 11.185 | 0.814  | 0.499  | -0.012 | 10.787 | 0.834  |
| VaPGD   | 0.227  | 0.162  | 3.344  | 0.884  | 0.155  | 0.087  | 8.659  | 0.906  | 0.081  | 0.123  | 9.169  | 0.827  |
| Certify | 0.493  | 0.014  | 4.381  | 0.678  | 0.551  | -0.047 | 11.588 | 0.800  | 0.436  | -0.019 | 11.188 | 0.836  |
Table 3: Evaluating interpretation methods for Deep-Biaffine under the comprehensiveness metric and the sensitivity metric on the dependency parsing task. We discard the sufficiency score as it is unreasonable to remove a large proportion of tokens on a structured prediction task. We also discard the stability metric as there is little consensus on how to attack a structured model.

Table 3: Evaluating interpretation methods for Deep-Biaffine under the comprehensiveness metric and the sensitivity metric on the dependency parsing task. We discard the sufficiency score as it is unreasonable to remove a large proportion of tokens on a structured prediction task. We also discard the stability metric as there is little consensus on how to attack a structured model.

Figure 1: An example of interpreting BiLSTM using five interpretation methods. A deeper red color means the token is identified as more important to the model output while a deeper blue color stands for a less important token. The comprehensiveness and the sensitivity scores for each interpretation are shown. We find that the important tokens identified by different methods are highly overlapped for BiLSTM.

5.1 Experimental Setup
Following Dozat and Manning (2017), we use DeepBiaffine, a graph-based dependency parser as the target model, which consists of a BiLSTM encoder and a Biaffine classifier. We use the English Penn Treebank converted to Stanford Dependencies version 3.5.0 (PTB-SD). We consider the same interpretation methods for text classification tasks. See Appendix A, B, and C for data statistics, model, and evaluating details.

5.2 Evaluation Paradigm
As a new paradigm to evaluate interpretation methods on the dependency parsing task, we focus on the PP attachment ambiguity. A dependency parser needs to determine either the preposition in PP attaches to the preceding noun phrase NP (NP-attachment) or the verb phrase VP (VP-attachment) (Hindle and Rooth, 1993), which involves not only syntactic considerations but some semantics as well. The basic structure of ambiguity is VP –
NP – PP. For example, in the sentence *I saw a cat with a telescope*, a parser needs the semantics of the noun phrase *a telescope* to predict the head of *with*, which is *saw*. If we change *a telescope* to *a tail*. The head of *with* would become the preceding noun *cat*. We will later call nouns in PPs like *telescope* “disambiguating nouns”, as they provide important semantic information for a parser to disambiguate PP attachment ambiguity. It is thus natural to expect that disambiguating nouns being considered important by a faithful interpretation method.

Specifically, we extract 100 examples from PTB-SD that contain the PP attachment ambiguity to evaluate interpretations. Interpretation methods will assign an importance score to each token in the sentence to indicate how much it impacts the model prediction on PP attachment arcs. \( s(e(x)) \) in this task is the unlabeled arc score between the preposition and its head after Softmax in the graph-based dependency parser. We test the faithfulness of the attributions by employing the comprehensiveness metric and the sensitivity metric. Note that for DeepBiaffine, each input token is represented by the concatenation of its word embedding and its part-of-speech tag embedding. When applying the interpretation methods and the evaluation metrics, we only modify the word embeddings but keep the part-of-speech tag embeddings unchanged.

5.3 Results and Discussion

Results are shown in Table 3. Similar to the results on text classification tasks, we find that perturbation-based methods like LIME and Occlusion perform well on the comprehensiveness metric, while VaPGD performs the best under the sensitivity metric. PGDInp and Certify are slightly better than GradInp under both the two metrics. Generally, we find that common relevant tokens for a PP-attachment arc extracted by interpretation methods are the preposition itself, the preceding noun or verb, and the disambiguating noun. This is close to our expectations. An example is shown in Figure 4.

Disambiguating Noun Analysis Humans would expect the disambiguating nouns to be identified as important signals by faithful interpretations. We conduct an experiment to explore that. For the 100 test cases, we summarize how many times the disambiguating nouns are positioned as the top-k most important words by those interpretation methods, where k varies in 10%, 20%, ..., 100% of total tokens in an example. Results are shown in Figure 3. We observe that there are obvious patterns in different categories. This demonstrates that interpretation methods from the same category do have high correlations regarding extracting disambiguating nouns. For example, interpretation methods like VaGrad and VaPGD, which leverage the gradient information only, tend to position the disambiguating nouns on the top of their importance lists. From this perspective, they are more conformed with human judgments. However, the pattern changes for GradInp and PGDInp that multiply the gradient information with the input itself. Further, the two perturbation-based methods, Occlusion and LIME, also put the disambiguation words to very similar positions.

6 Related Work

Interpretation methods Various post-hoc interpretation methods are proposed to explain the behaviors of black-box models. The most prominent category is the gradient-based methods, which is first introduced by Simonyan et al. (2014) in the
Figure 3: Where do interpretations place the disambiguating nouns. The results demonstrate obvious patterns in different categories. X-axis is the top-k threshold, where k varies in 10%, 20%, ..., 100%. Y-axis is the number of examples where an interpretation method ranks the disambiguating noun in each example as the top-k most important tokens.

computer vision domain and later adopted to the NLP domain by Li et al. (2016). Other categories of methods are also proposed to address several limitations of gradient-based methods. For example Shrikumar et al. (2017); Sundararajan et al. (2017) propose to back-propagate the model output difference between the original point and a reference point, which are classified as reference-based methods in this paper. Ribeiro et al. (2016); Zeiler and Fergus (2014) are perturbation-based methods that model local accuracy of model output on perturbed data. Lundberg and Lee (2017) unified several works leveraging the Shapley value in game theory. In our work, we propose a new class of interpretation methods that adopt the techniques in the adversarial robustness domain and bridge the gap between the gradient-based methods and the reference-based methods. We empirically find that the new methods achieve remarkable performance on different quantitative metrics.

Evaluate interpretation methods

One line of studies explores what is the best way to evaluate interpretation methods. Several works test the faithfulness of interpretations, i.e., whether the explanations reflect the real reasoning process of models. A large proportion of faithfulness measurements consider automatic ways that perturb or occlude tokens identified as important by interpretations and measure the changes on model output (DeYoung et al., 2020; Jain and Wallace, 2019; Hsieh et al., 2020; Zaidan and Eisner, 2008; Serrano and Smith, 2019). Some other works propose to evaluate the faithfulness of interpretations by checking to what extent they satisfy some desired axioms (Ancona et al., 2018; Sundararajan et al., 2017; Shrikumar et al., 2017). Besides, Alvarez-Melis and Jaakkola (2018); Ghorbani et al. (2019); Kindermans et al. (2019) reveal limitations in interpretation faithfulness through finding contrast examples that induce similar model outputs but a substantial difference in the interpretation outputs.

Another group of studies focus on the plausibility of interpretations, i.e., whether the explanations conform with human judgments (Doshi-Velez and Kim, 2017; Ribeiro et al., 2016; Lundberg and Lee, 2017), or assist in predicting model behaviors on new data (Hase and Bansal, 2020).

Note that although there exist many hybrid works that evaluate both the faithfulness and the plausibility of interpretations by proposing a suite of diagnostic tests (DeYoung et al., 2020; Atanasova et al., 2020; Adebayo et al., 2018; Liu et al., 2020), Jacovi and Goldberg (2020) advocate to explicitly distinguish between the two distinct aspects. In this paper, we focus on the faithfulness of interpretations but consider three plausible notions of faithfulness, and thus provide a more systematical evaluation to different interpretation methods. We also propose a new paradigm to extend these metrics to the structured prediction task.
7 Conclusion

In this paper, we studied three different criteria of faithfulness: the removal-based criterion, the sensitivity, and the stability, and conducted systematical experiments to evaluate existing interpretations under the three criteria. We found that interpretations have very inconsistent performance regarding different criteria. We proposed a new class of interpretation methods leveraging robustness techniques, which achieve the best performance under the sensitivity and the stability criteria. We further proposed a novel paradigm to evaluate interpretations on the dependency parsing task, which moves beyond text classification tasks in previous literature. Our study shed a light on understanding the property of interpretations.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31:9505–9515.

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2890–2896.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lima, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3256–3274.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 149–155.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4443–4458.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. Arxiv.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3719–3728.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2019. Interpretation of neural networks is fragile. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 3681–3688. AAAI Press.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5540–5552.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. Computational Linguistics, 19(1):103–120.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.
Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seunyong Kim, Sanjiv Kumar, and Cho-Jui Hsieh. 2020. Evaluations and methods for explanation through robustness analysis. CoRR, abs/2006.00442.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205.

Sarthak Jain and Byron C. Wallace. 2019. 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.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un)reliability of saliency methods. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pages 267–280. Springer.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In ICLR (Poster).

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 681–691. The Association for Computational Linguistics.

Ninghao Liu, Yunsong Meng, Xia Hu, Tie Wang, and Bo Long. 2020. Are interpretations fairly evaluated? a definition driven pipeline for post-hoc interpretability. arXiv preprint arXiv:2009.07494.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30:4765–4774.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 97–101.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2931–2951.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 3145–3153. PMLR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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.

Charles Spearman. 1961. The proof and measurement of association between two things.

Mukund Sundararajan, Ankur Taly, and Qi Qi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 3319–3328. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. 2020. Automatic perturbation analysis for scalable certified robustness and beyond. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.
Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 31–40.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, volume 8689 of Lecture Notes in Computer Science, pages 818–833. Springer.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, pages 649–657.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989.
A Dataset and Model Details

Datasets Statistics of the three text classification datasets and the dependency parsing dataset is presented in Table 4.

| Dataset  | Train/Dev/Test | Avg Len |
|----------|----------------|---------|
| SST-2    | 67.3k/0.8k/1.8k | 19.2    |
| Yelp     | 447.9k/112.0k/1.2k | 119.8   |
| AGNews   | 51.0k/9.0k/3.8k  | 35.5    |
| PTB-SD   | 39.8k/1.7k/2.4k  | 23.5    |

Table 4: Data Statistics

Models All models are implemented based on the PyTorch library.

For BERT, we use the bert-base-uncased model from the pytorch-transformers library (Wolf et al., 2020). We fine-tune BERT model on each dataset, using a unified setup: dropout rate 0.1, Adam (Kingma and Ba, 2015) with an initial learning rate of 1e-4, batch size 128, and no warm-up steps. We set the maximum number of fine-tuning to be 3 and deploy interpretation methods to explain the model after 3 epochs of fine-tuning. The fine-tuned BERT achieves 90.7, 95.4, and 96.9 accuracy on SST-2, Yelp and AGNews, respectively. Note that each token is transformed to three embedding vectors after BERT’s embedding layer, namely its word embedding vector, its position embedding vector, and its segment embedding vector. We only consider the contribution of word embeddings to the model output.

For BiLSTM classifier, we use a one-layer BiLSTM encoder with a linear classifier. The embedding is initialized with the 100-dimensional pre-trained GloVe word embedding while the part-of-speech tag embeddings are initialized to all zero. The encoder hidden size is 100. The arc and dependency relation hidden size are both 500. We get an UAS of 95.1 with our model.

B Interpretation Methods Details

For VaGrad, GradInp, VaPGD, PGDInp, and IngGrad, we leverage the automatic differentiation mechanism of PyTorch. For LIME, we modify the code from the original implementation of Ribeiro et al. (2016). For DeepLIFT, we use the implementation in Captum. For Certify, we modify the code in the autoLiRPA library.

For the two reference-based methods IngGrad and DeepLIFT, we use all zero word embeddings as the reference point. To approximate the integral in IngGrad, we sum up 50 points along the linear path from the reference point to the current point. For the perturbation-based methods LIME and Occlusion, we also set the word embedding of a token to an all zero embedding when it is perturbed. Specifically, we find 500 perturbed examples around the original point for LIME to fit a linear regression model. Finally, for PGDInp and VaPGD, we set the maximum magnitude of perturbations $\epsilon$ as 0.5 for BERT and 2.2 for BiLSTM classifier, the number of iterations as 50, and the step size as $\epsilon/25$.

C Evaluation Criteria Details

Sensitivity Details As mentioned in Section 2, we use PGD with a binary search for the minimal perturbation magnitude. In practice, we set the number of iterations to be 100 and the step size to be 1.0. Then, we conduct a binary search to estimate the smallest vicinity of the original point which contains an adversarial example that changes the model prediction.

Stability Details The synonyms in the stability metrics come from (Alzantot et al., 2018), where they extract nearest neighbors in the GloVe embeddings space with a counter-fitting method to filter out antonyms. We allow at most four tokens changed to their synonyms for each input text and at most 0.1 change in model output probability of the original prediction for BERT and 0.2 for BiLSTM.

---

1 https://pytorch.org/
2 https://github.com/pytorch/captum
3 https://github.com/KaidiXu/auto_LiRPA
Steers turns in a snappy screenplay that curls at the edges; it’s so clever you want to hate it.

Steers turns in a snappy screenplay that curls at the edges; it’s so clever you want to hate it.

Steers turns in a snappy screenplay that curls at the edges; it’s so clever you want to hate it.

Steers turns in a snappy screenplay that curls at the edges; it’s so clever you want to hate it.

Steers turns in a snappy screenplay that curls at the edges; it’s so clever you want to hate it.

Figure 5: An example of interpreting BERT with five interpretation methods.

**Thresholds** To compute the removal-based metrics and the AUC of sensitivity for text classification tasks, we vary the number of tokens being removed (preserved) or perturbed to be 10%, 20%, ..., 50% of the total number of tokens in the input text. For the dependency parsing task, the corresponding thresholds are 10%, 20% and 30%.

**D BERT Example**

An example of interpreting BERT classification model on SST-2. BERT assigns a postive label for this instance.

**E Examples for the Stability Criterion**

**E.1 SST-2 Examples**

Table 5 shows some contrast examples constructed for the stability criterion on SST-2.

**E.2 AGNews Examples**

Table 6 shows some contrast examples constructed for the stability criterion on AGNews

**E.3 Yelp Examples**

Table 7 shows some contrast examples constructed for the stability criterion on Yelp.

---

**VaPGD, BERT on SST-2**

| Rank correlation | Model change |
|------------------|--------------|
| 0.346            | 0.00         |

**Original** This is a film well worth seeing, talking and singing heads and all.

**Contrast** This is a films well worth staring, talking and singing heads and entirety.

**IngGrad, BERT on SST-2**

| Rank correlation | Model change |
|------------------|--------------|
| 0.645            | 0.15         |

**Original** Ray Liotta and Jason Patric do some of their best work in their underwritten roles, but do n’t be fooled: Nobody deserves any prizes here.

**Contrast** Ray Liotta and Jason Patric do certain of their best collaborate in their underwritten roles, but do n’t be fooled: Nobody deserves any awards here.

**LIME, BiLSTM on SST-2**

| Rank correlation | Model change |
|------------------|--------------|
| 0.425            | 0.05         |

**Original** Nearly surreal, dabbling in French, this is no simple movie, and you ’ll be taking a risk if you choose to see it.

**Contrast** Almost surreal, dabbling in French, this is no simple cinematography, and you ’ll be taking a risk if you choose to see it.

Table 5: Generated contrast examples for evaluating the stability criterion on SST-2. Modified words are underlined. Spearman’s rank correlation between a pair of examples and the performance difference of a model on the pair of examples are shown above each pair.
| Method          | Rank correlation | Model change | Original                                                                                                                                                                                                 | Contrast                                                                                                                                                                                                 |
|-----------------|------------------|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Erasure, BERT   | 0.689            | 0.08         | Supporters and rivals warn of possible fraud; government says chavez’s defeat could produce turmoil in world oil market.                                                                             | Supporters and rivals warn of possible fraud; government says chavez’s defeat could produce disorder in planet oil trade.                                                                              |
| DeepLIFT, BERT  | 0.317            | 0.00         | Mills corp. agreed to purchase a qqq percent interest in nine malls owned by general motors asset management corp. for just over qqq billion, creating a new joint venture between the groups. | Mills corp. agree to purchase a qqq percent interest in nine malls owned by comprehensive motors asset management corp. for just over qqq trillion, creating a new joint venture between the groups. |
| VaGrad, BERT    | 0.970            | 0.12         | London (reuters) - oil prices surged to a new high of qqq a barrel on wednesday after a new threat by rebel militia against iraqi oil facilities and as the united states said inflation had stayed in check despite rising energy costs. | London (reuters) - oil prices surged to a new high of qqq a canon on wednesday after a new menace by rebel militia against iraqi oil facilities and as the united states said inflation had stayed in check despite rising energy costs. |
| PGD, BiLSTM on Yelp | 0.530          | 0.00         | Love this beer distributor. They always have what I’m looking for. The workers are extremely nice and always willing to help. Best one I’ve seen by far.                                                 | Love this beer distributor. They repeatedly have what I’m seeking for. The workers are extremely nice and always loan to help. Best one I’ve seen by far.                                                    |
| Certify, BiLSTM on Yelp | 0.633          | 0.01         | Last summer I had an appointment to get new tires and had to wait a super long time. I also went in this week for them to fix a minor problem with a tire they put on. They "fixed" it for free, and the very next morning I had the same issue. I called to complain, and the "manager" didn’t even apologize!!! So frustrated. Never going back. They seem overpriced, too. | Last summer I took an appoints to get new tires and had to wait a super long time. I also went in this week for them to fix a minor problem with a tire they put on. They "fixed" it for free, and the very impending morning I had the same issue. I called to complain, and the "manager" didn’t even apologize!!! So frustrated. Never going back. They seem overpriced, too. |

Table 6: Generated contrast examples for evaluating the stability criterion on AGNews.

Table 7: Generated contrast examples for evaluating the stability criterion on Yelp.