A Fine-grained Interpretability Evaluation Benchmark for Neural NLP

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao,
Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, Haifeng Wang
Baidu Inc, Beijing, China
{wanglijie,shenyaozong@baidu.com}

Abstract

While there is increasing concern about the interpretability of neural models, the evaluation of interpretability remains an open problem, due to the lack of proper evaluation datasets and metrics. In this paper, we present a novel benchmark to evaluate the interpretability of both neural models and saliency methods. This benchmark covers three representative NLP tasks: sentiment analysis, textual similarity and reading comprehension, each provided with both English and Chinese annotated data. In order to precisely evaluate the interpretability, we provide token-level rationales that are carefully annotated to be sufficient, compact and comprehensive. We also design a new metric, i.e., the consistency between the rationales before and after perturbations, to uniformly evaluate the interpretability on different types of tasks. Based on this benchmark, we conduct experiments on three typical models with three saliency methods, and unveil their strengths and weakness in terms of interpretability. We will release this benchmark[1] and hope it can facilitate the research in building trustworthy systems.

1 Introduction

In the last decade, deep learning (DL) has been rapidly developed and has greatly improved various artificial intelligence tasks in terms of accuracy (Deng and Yu, 2014; Litjens et al., 2017; Pouyanfar et al., 2018). However, as DL models are black-box systems, their inner decision processes are opaque to users. This lack of transparency makes them untrustworthy and hard to be applied in decision-making applications in fields such as health, commerce and law (Fort and Couillault, 2016). Consequently, there is a growing interest in explaining the predictions of DL models (Simonyan et al., 2014; Ribeiro et al., 2016; Alzantot et al., 2018; Bastings et al., 2019; Jiang et al., 2021). Accordingly, many evaluation datasets are constructed and the corresponding metrics are designed to evaluate related works (DeYoung et al., 2020; Jacovi and Goldberg, 2020).

In order to accurately evaluate model interpretability[2] with human-annotated rationales[3] (i.e., evidence that supports the model prediction), many researchers successively propose the properties that a rationale should satisfy, e.g., sufficiency, compact-

Sentiment Analysis (SA)
Instance^o: although it bangs a very cliched drum at times, this crowd-pleaser's fresh dialogue, energetic music, and good-natured spunk are often infectious. **Sentiment label**: positive
Instance^p: although it bangs a very cliched drum at times, this crowd-pleaser's novel dialogue, vigorous music, and good-natured spunk are often infectious. **Sentiment label**: positive
Semantic Textual Similarity (STS)
Instance1^o: Is there a reason why we should travel alone? **Instance2^o**: What are some reasons to travel alone? **Similarity**: same
Instance1^p: Is there any reason why we travel alone? **Instance2^p**: List some reasons to travel alone? **Similarity**: same
Machine Reading Comprehensive (MRC)
Question: What part of France were the Normans located? **Article^o**: ...and customs to synthesize a unique "Norman" culture in the north of France. ... **Answer**: north
Question: Where in France were the Normans located? **Article^p**: ...and customs to synthesize a unique "Norman" culture in the north of France. ... **Answer**: north

Table 1: Examples from our benchmark. In each instance, colored tokens are rationales, and tokens in the same color constitute an independent rationale set. Each perturbed example (^{p}) is created on an original example (^{o}), where underlined tokens in the original example have been altered. The consistency of rationales under perturbations is used to evaluate interpretability.

[1] https://www.luge.ai/#/luge/task/taskDetail?taskId=15

[2] Despite fine-grained distinctions between "interpretability" and "explainability", we use them interchangeably.

[3] In this paper, we focus on highlight-based rationales, which consist of input elements, such as words and sentences, that play a decisive role in the model prediction.

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 70 - 84
December 7-8, 2022 ©2022 Association for Computational Linguistics

ness and comprehensiveness (see Section 3.3 for their specific definitions) (Kass et al., 1988; Fischer et al., 1990; Lei et al., 2016; Yu et al., 2019). However, the existing datasets are designed for different research aims with different metrics, and their rationales do not satisfy all properties needed, as shown in Table 2, which makes it difficult to track and facilitate the research progress of interpretability. In addition, all existing datasets are in English.

Meanwhile, many studies focus on designing guidelines and metrics for interpretability evaluation, where plausibility and faithfulness are proposed to measure interpretability from different perspectives (Herman, 2017; Alvarez Melis and Jaakkola, 2018; Yang et al., 2019; Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020). Plausibility measures how well the rationales provided by models align with human-annotated rationales. With different annotation granularities, token-level and span-level F1-scores are proposed to measure plausibility (DeYoung et al., 2020; Mathew et al., 2021). Faithfulness measures to what extent the provided rationales influence the corresponding predictions. Some studies (Yu et al., 2019; DeYoung et al., 2020) propose to compare the model's prediction on the full input to its prediction on input masked according to the rationale and its complement (i.e., non-rationale). However, it is difficult to apply this evaluation method to non-classification tasks, such as machine reading comprehension. Furthermore, the model prediction on the non-rationale has gone beyond the standard output scope, e.g., the prediction label on the non-rationale should be neither positive nor negative in the sentiment classification task. Thus the metric provided by this method can not generally and may not precisely evaluate the interpretability.

In order to address the above problems, we release a new interpretability evaluation benchmark which provides fine-grained rationales for three tasks and a new evaluation metric for interpretability. Our contributions include:

- Our benchmark contains three representative tasks in both English and Chinese, i.e., sentiment analysis, semantic textual similarity and machine reading comprehension. Importantly, all annotated rationales meet the requirements of sufficiency, compactness and comprehensiveness by being organized in the set form.

- To precisely and uniformly evaluate the interpretability of all tasks, we propose a new evaluation metric, i.e., the consistency between the rationales provided on examples before and after perturbation. The perturbations are crafted in a way that will not change the model decision mechanism. This metric measures model fidelity under perturbations and could help to find the relationship between interpretability and other metrics, such as robustness.

- We give an in-depth analysis based on three typical models with three popular saliency methods, as well as a comparison between our proposed metrics and the existing metrics. The results show that our benchmark can be used to evaluate the interpretability of DL models and saliency methods. Meanwhile, the results strongly indicate that the research on interpretability of NLP models has much further to go, and we hope our benchmark will do its bit along the way.

2 Related Work

As our work provides a new interpretability evaluation benchmark with human-annotated rationales, in this section, we mainly introduce saliency methods for the rationale extraction, interpretability evaluation datasets and metrics.

Saliency Methods In the post-hoc interpretation research field, saliency methods are widely used to interpret model decisions by assigning a distribution of importance scores over the input tokens to represent their impacts on model predictions (Simonyan et al., 2014; Ribeiro et al., 2016; Murdoch et al., 2018). They are mainly divided into four categories: gradient-based, attention-based, erasure-based and linear-based. In gradient-based methods, the magnitudes of the gradients serve as token importance scores (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017). Attention-based methods use attention weights as token importance scores (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). In erasure-based methods, the token importance score is measured by the change of output when the token is removed (Li et al., 2016; Feng et al., 2018). Linear-based methods use a simple and explainable linear model to approximate the evaluated model behavior locally and use the learned token weights as importance scores (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017). These methods have their own advantages and limitations from aspects of computational efficiency, interpretability performance and so on (Nie

Datasets	Granularity	Properties		
		Sufficiency	Compactness	Comprehensiveness
e-SNLI* (Camburu et al., 2018)	word	✗	✓	✗
HUMMINGBIRD (Hayati et al., 2021)	word	✓$^{-}$	✗	–
HateXplain (Mathew et al., 2021)	word	✓$^{-}$	–	✓
Movie Reviews* (Zaidan and Eisner, 2008)	snippet	✓	✗	✗
CoS-E* (Rajani et al., 2019)	snippet	✓$^{-}$	✗	✓
Evidence Inference* (Lehman et al., 2019)	snippet	✓	✗	✗
BoolQ* (DeYoung et al., 2020)	snippet	✓	✗	✓
WikiQA (Yang et al., 2015)	sentence	✓	✗	–
MultiRC* (Khashabi et al., 2018)	sentence	✓	✗	✓
HotpotQA (Yang et al., 2018)	sentence	✓	✗	✓
FEVER* (Thorne et al., 2018)	sentence	✓	✗	–
SciFact (Wadden et al., 2020)	sentence	✓	✗	–
Ours	word	✓	✓	✓

Table 2: Statistics of existing datasets with highlight-based rationales. The datasets marked with * are collected and modified by ERASER (DeYoung et al., 2020). ERASER manually reviews and constructs snippet-level rationales to make them satisfy sufficiency and comprehensiveness. ✓$^{-}$ represents the rationale contains key words, but does not contain enough information for the prediction. The value '-' represents the property is not mentioned in the paper.

et al., 2018; Jain and Wallace, 2019; De Cao et al., 2020; Sixt et al., 2020).

Interpretability Datasets Many datasets with human-annotated rationales have been published for interpretability evaluation, e.g., highlight-based rationales (DeYoung et al., 2020; Mathew et al., 2021), free-text rationales (Camburu et al., 2018; Rajani et al., 2019) and structured rationales (Ye et al., 2020; Geva et al., 2021). To create high-quality highlight-based rationales, many studies give their views on the properties that a rationale should satisfy. Kass et al. (1988) propose that a rationale should be understood by humans. Lei et al. (2016) point that a rationale should be compact and sufficient, i.e., it is short and contains enough information for a prediction. Yu et al. (2019) introduce comprehensiveness as a criterion, requiring all rationales to be selected, not just a sufficient set. Although the above criteria have been proposed for highlight rationales, the existing datasets in Table 2 are built with part of them, as they are conducted on different tasks with individual aims.

Interpretability Metrics For highlight-based rationales, plausibility and faithfulness are often used to measure interpretability from the aspects of human cognition and model fidelity (Arras et al., 2017; Mohseni et al., 2018; Weerts et al., 2019). DeYoung et al. (2020) propose to use IOU (Intersection-Over-Union) F1-score and AUPRC (Area Under the Precision-Recall curve) score to measure plausibility of snippet-level rationales. Mathew et al. (2021) use token F1-score to evaluate plausibility of token-level rationales. Jacovi and Goldberg (2020) provide concrete guidelines for the definition and evaluation of faithfulness. DeYoung et al. (2020) propose to evaluate faithfulness from the perspectives of sufficiency and comprehensiveness of rationales (Equation 4). However, this evaluation manner is only applicable to classification tasks and brings uncontrollable factors to interpretability evaluation. Thus Yin et al. (2022) propose sensitivity and stability as complementary metrics for faithfulness. Ding and Koehn (2021) evaluate faithfulness of saliency methods on natural language models by measuring how consistent the rationales are regarding perturbations.

In this work, we provide a new interpretability evaluation benchmark, containing fine-grained annotated rationales, a new evaluation metric and the corresponding perturbed examples.

3 Evaluation Data Construction

As illustrated in Figure 1, the construction of our datasets mainly consists of three steps: 1) data collection for each task; 2) perturbed data construction; 3) iterative rationale annotation and checking. We first introduce the annotation process, including the annotation criteria for perturbations and rationales. Then we describe our data statistics. In addition, we show other annotation details in Appendix A.

3.1 Data Collection

In order to provide a general and unified interpretability evaluation benchmark, we construct

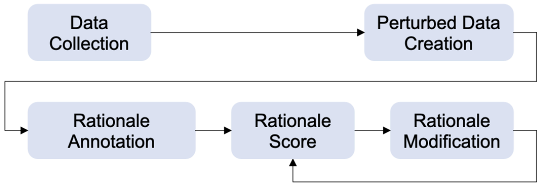


Figure 1: The construction workflow of our datasets.

evaluation datasets for three representative tasks, i.e., sentiment analysis, semantic textual similarity, and machine reading comprehension. Meanwhile, we create both English and Chinese evaluation datasets for each task.

Sentiment Analysis (SA), a single-sentence classification task, aims to predict a sentiment label for the given instance. For English, we randomly select 1,500 instances from Stanford Sentiment Treebank (SST) (Socher et al., 2013) dev/test sets, and 400 instances from Movie Reviews (Zaidan and Eisner, 2008) test set. For Chinese, we randomly sample 60,000 instances from the logs of an open SA API[4] with the permission of users. The annotators select instances for annotation (see Appendix A for details) and label a sentiment polarity for each unlabeled instance. Then 2,000 labeled instances are chosen for building evaluation set.

Semantic Textual Similarity (STS), a sentence-pair similarity task, is to predict the similarity between two instances. We randomly select 2,000 pairs from Quora Question Pairs (QQP) (Wang et al., 2018) and LCQMC (Liu et al., 2018) to build English and Chinese evaluation data respectively.

Machine Reading Comprehension (MRC), a long-text comprehension task, aims to extract an answer based on the question and the corresponding passage. We randomly select 1,500 triples with answers and 500 triples without answers from SQUAD2.0 (Rajpurkar et al., 2018) and DuReader (He et al., 2018) for building English and Chinese evaluation set respectively.

3.2 Perturbed Data Creation

Recent studies (Jacovi and Goldberg, 2020; Ding and Koehn, 2021) claim that a saliency method is faithful if it provides similar rationales for similar inputs and outputs. Inspired by them, we propose to evaluate the model faithfulness via measuring how consistent its rationales are regarding perturbations that are supposed to preserve the same model decision mechanism. In other words, under perturbations, a model is considered to be faithful if the change of its rationales is consistent with the change of its prediction. Consequently, we construct perturbed examples for each original input.

Perturbation Criteria Perturbations should not change the model internal decision mechanism. We create perturbed examples from two aspects: 1) perturbations do not influence model rationales and predictions; 2) perturbations cause the alterations of rationales and may change predictions. Please note that the influence of perturbations comes from human's basic intuition on model's decision-making mechanism. Based on the literature (Jia and Liang, 2017; McCoy et al., 2019; Ribeiro et al., 2020), we define three perturbation types.

- **Alteration of dispensable words**. Insert, delete and replace words that should have no effect on model predictions and rationales, e.g., the sentence "*what are* some reasons to travel alone" is changed to "*list* some reasons to travel alone".

- **Alteration of important words**. Replace important words which have an impact on model predictions with their synonyms or related words, such as "i dislike you" instead of "i hate you". In this situation, the model prediction and rationale should change with perturbations.

- **Syntax transformation**. Transform the syntax structure of an instance without changing its semantics, e.g., "the customer commented the hotel" is transformed into "the hotel is commented by the customer". In this case, the model prediction and rationale should not be affected.

For each original input, the annotator first selects a perturbation type, then creates a perturbed example according to the definition of this perturbation type. Please note that the annotators can select more than one perturbation type for an original input. We ask the annotator to create at least one perturbed example for each original input. And they need to create at least 100 perturbed examples for each perturbation type. For each task, we have two annotators to create perturbed examples and label golden results for these examples, i.e., sentiment label for SA, similarity label for STS and answer for MRC. According to the perturbation criteria, most of the perturbed examples have the same results as their original ones. Then we ask the other

[4] https://ai.baidu.com/tech/nlp_apply/sentiment_classify. Due to the diversity of these logs, we choose instances from these logs for annotation.

two annotators to review and modify the created examples and their corresponding results. Since the annotation task in this step is relatively easy, the accuracy of created examples after checking is more than 95%.

3.3 Iterative Rationale Annotation Process

Given an input and the corresponding golden result, the annotators highlight important input tokens that support the prediction of golden result as the rationale. Then we introduce the rationale criteria and the annotation process used in our work.

Rationale Criteria As discussed in recent studies (Lei et al., 2016; Yu et al., 2019), a rationale should satisfy the following properties.

- **Sufficiency**. A rationale is sufficient if it contains enough information for people to make the correct prediction. In other words, people can make the correct prediction only based on tokens in the rationale.
- **Compactness**. A rationale is compact if all of its tokens are indeed required in making a correct prediction. That is to say, when any token is removed from the rationale, the prediction will change or become difficult to make.
- **Comprehensiveness**. A rationale is comprehensive if its complements in the input can not imply the prediction, that is, all evidence that supports the output should be labeled as rationales.

Annotation Process To ensure the data quality, we adopt an iterative annotation workflow, consisting of three steps, as described in Figure 1.

Step 1: rationale annotation. Based on human's intuitions on the model decision mechanism, given the input and the corresponding golden result, the ordinary annotators who are college students majoring in languages label all critical tokens to guarantee the rationale's comprehensiveness. Then they organize these tokens into several sets, each of which should be sufficient and compact. That is to say, each set can support the prediction independently. As described in Table 1, the first example contains three rationale sets, and tokens in the same color belong to the same set. Based on this set form, the rationale satisfies the above three criteria.

Step 2: rationale scoring. Our senior annotators[5] double-check the annotations by scoring the given rationales according to the annotation criteria. For each rationale set, the annotators rate their confidences for sufficiency and compactness. The confidences for **sufficiency** consist of three classes: *can not support result (1)*, *not sure (2)* and *can support result (3)*. And the confidences for **compactness** compose of four classes: *include redundant tokens (1)*, *include disturbances (2)*, *not sure (3)* and *conciseness (4)*. Then based on all rationale sets for each input, the annotators rate their confidences for **comprehensiveness** on a 3-point scale including *not be comprehensive (1)*, *not sure (2)*, *be comprehensive (3)*.

Tasks	English			Chinese		
	Size	RLR	RSN	Size	RLR	RSN
SA	1,999	20.1%	2.1	2,160	27.6%	1.4
STS	2,248	50.4%	1.0	2,146	66.6%	1.0
MRC	1,969	10.4%	1.0	2,315	9.8%	1.0

Table 3: Overview of our datasets. "Size" shows the number of original/perturbed pairs. "RLR" represents the ratio of rationale length to its input length. "RSN" represents the number of rationale sets in an input. We report the average RLR and RSN over all data.

A rationale is considered to be of high-quality if its average score on sufficiency, compactness and comprehensiveness is equal to or greater than 3.0, 3.6, 2.6. That is to say, at least two-thirds of the annotators give the highest confidence, and less than one-third of the annotators give the confidence of "*not sure*". Then all unqualified data whose average score on a property is lower than the corresponding threshold goes to the next step.

Step 3: rationale modification. Low-quality rationales are shown to the ordinary annotators again. The annotators correct the rationales to meet the properties with scores below the threshold.

Then the corrected rationales are scored by senior annotators again. The unqualified data after three loops is discarded. This iterative annotation-scoring process can ensure the data quality.

Other annotation details, such as annotator information, annotation training and data usage instructions, are described in Appendix A.

3.4 Data Statistics

We give a comparison between our benchmark and other existing datasets, as shown in Table 2. Compared with existing datasets, our benchmark contains three NLP tasks with both English and Chinese annotated data. Compared with ERASER which collects seven existing English datasets in

[5] They are full-time employees, and have lots of experience in annotating data for NLP tasks.

Models	SA		STS		MRC	
	Acc^f	Acc^r	Acc^f	Acc^r	F1^f	F1^r
English						
LSTM	78.2	**86.2**	74.6	69.8	54.4	53.4
RoBERTa-base	93.8	92.4	92.7	89.3	71.7	**80.8**
RoBERTa-large	95.4	91.5	93.2	88.8	76.0	**76.7**
Chinese						
LSTM	60.0	**70.4**	75.2	**80.7**	66.4	**82.2**
RoBERTa-base	59.8	**77.0**	85.5	**88.1**	65.8	**89.3**
RoBERTa-large	62.6	**80.6**	86.0	**87.4**	67.8	**83.3**

Table 4: Model performance on the original full input (Acc^f) and human-annotated rationale (Acc^r).

its benchmark and provides snippet-level rationales to satisfy sufficiency and comprehensiveness, our benchmark provides token-level rationales and satisfies all three primary properties of rationales.

Table 3 shows the detailed statistics of our benchmark. We can see that the length ratio and the number of rationales vary with datasets and tasks, where the length ratio affects the interpretability performance on plausibility, as shown in Table 6.

Meanwhile, we evaluate the sufficiency of human-annotated rationales by evaluating model performance on rationales, as shown in Table 4. Despite the input construction based on rationales has destroyed the distribution of original inputs, model performance on human-annotated rationales is competitive with that on full inputs, especially on MRC task and Chinese datasets. We can conclude that human-annotated rationales are sufficient. Meanwhile, we give more data analysis in Table 7, such as model performance on non-rationales, sufficiency and comprehensiveness scores.

4 Metrics

Following existing studies (DeYoung et al., 2020; Ding and Koehn, 2021; Mathew et al., 2021), we evaluate interpretability from the perspectives of plausibility and faithfulness. Plausibility measures how well the rationales provided by the model agree with human-annotated ones. And faithfulness measures the degree to which the provided rationales influence the corresponding predictions.

Different from existing work, we adopt **token-F1** score for plausibility and propose a new metric **MAP** for faithfulness.

Token F1-score is defined in Equation 1, which is computed by overlapped rationale tokens. Since an instance may contain multiple golden rationale sets, for the sake of fairness, we take the set that has the largest F1-score with the predicted rationale as the ground truth for the current prediction.

$$\begin{aligned} \texttt{Token-F1} &= \frac{1}{N}\sum_{i=1}^{N}(2\times\frac{P_i \times R_i}{P_i + R_i}) \\ \texttt{where}\quad P_i &= \frac{|S_i^p \cap S_i^g|}{|S_i^p|} \texttt{ and } R_i = \frac{|S_i^p \cap S_i^g|}{|S_i^g|} \end{aligned} \tag{1}$$

where S_i^p and S_i^g represent the rationale set of i-th instance provided by models and human respectively; N is the number of instances.

MAP (Mean Average Precision) measures the consistency of rationales under perturbations and is used to evaluate faithfulness. According to the original/perturbed input pair, MAP aims to calculate the consistency of two token lists sorted by token importance score, as defined in Equation 2. The high MAP indicates the high consistency.

$$\texttt{MAP} = \frac{\sum_{i=1}^{|X^p|}(\sum_{j=1}^{i} G(x_j^p, X_{1:i}^o))/i}{|X^p|} \tag{2}$$

where X^o and X^p represent the sorted rationale token list of the original and perturbed inputs, according to the token important scores assigned by a specific saliency method. $|X^p|$ represents the number of tokens in X^p. $X_{1:i}^o$ consists of top-i important tokens of X^o. The function $G(x, Y)$ is to determine whether the token x belongs to the list Y, where $G(x, Y) = 1\,\texttt{iff}\,x \in Y$.

Meanwhile, we also report results of metrics proposed in DeYoung et al. (2020), i.e., IOU F1-score for plausibility, and the joint of sufficiency and comprehensiveness for faithfulness.

IOU F1-score is proposed on span-level rationales, which is the size of token overlap in two sets divided by the size of their union, as shown by S_i in Equation 3. A rationale is considered as a match if its S_i is equal to or greater than 0.5, as illustrated by the $Greater$ function.

$$\begin{aligned} \texttt{IOU-F1} &= \frac{1}{N}\sum_{i=1}^{N} Greater(S_i, 0.5) \\ \texttt{where}\quad S_i &= \frac{|S_i^p \cap S_i^g|}{|S_i^p \cup S_i^g|} \end{aligned} \tag{3}$$

The joint of **sufficiency** (`Score-Suf`) and **comprehensiveness** (`Score-Com`) is shown in Equation 4. A lower sufficiency score implies the rationale is more sufficient and a higher comprehensiveness score means the rationale is more influential in the prediction. A faithful rationale should have a low sufficiency score and a high comprehensiveness

Models	**SA** (Acc)		**STS** (Acc)		**MRC** (F1)	
	Ori	Ours	Ori	Ours	Ori	Ours
English						
LSTM	78.6	78.2	78.6	74.6	58.6	54.4
RoBERTa-base	92.1	93.8	91.5	92.7	78.4	71.7
RoBERTa-large	91.3	95.4	91.4	93.2	83.8	76.0
Chinese						
LSTM	86.7	60.0	77.4	75.2	75.0	66.4
RoBERTa-base	95.1	59.8	88.1	85.5	74.4	65.8
RoBERTa-large	95.0	62.6	88.1	86.0	77.8	67.8

Table 5: Conventional performance of base models on three tasks, where "Acc" is short for accuracy. The "Ori" dev/test set comes from the same dataset as training set. "Ours" represents our evaluation datasets.

score.

$$\begin{aligned}\texttt{Score-Suf} &= \frac{1}{N}\sum_{i=1}^{N}(F(x_i)_j - F(r_i)_j)\\ \texttt{Score-Com} &= \frac{1}{N}\sum_{i=1}^{N}(F(x_i)_j - F(x_i \setminus r_i)_j)\end{aligned} \tag{4}$$

where $F(x_i)_j$ represents the prediction probability provided by the model F for class j on the input x_i; r_i represents the rationale of x_i, and $x_i \setminus r_i$ represents its non-rationale.

5 Experiments

5.1 Experiment Settings

We implement three widely-used models and three saliency methods. We give brief descriptions of them and leave the implementation details to Appendix B. The source code will be released with our evaluation datasets.

Saliency Methods We adopt integrated gradient (**IG**) method (Sundararajan et al., 2017), attention-based (**ATT**) method (Jain and Wallace, 2019) and linear-based (**LIME**) (Ribeiro et al., 2016) method in our experiments. IG assigns importance score for each token by integrating the gradient along the path from a defined input baseline to the original input. ATT uses attention weights as importance scores, and the acquisition of attention weights depends on the specific model architecture. LIME uses the token weights learned by the linear model as importance scores.

For each saliency method, we take the top-k^d important tokens to compose the rationale for an input, where k^d is the product of the current input length and the average rationale length ratio of a dataset d, as shown by *RLR* in Table 3.

Comparison Models For each task, we re-implement three typical models with different network architectures and parameter sizes, namely LSTM (Hochreiter and Schmidhuber, 1997), RoBERTa-base and RoBERTa-large (Liu et al., 2019). Based on these backbone models, we then fine-tune them with commonly-used datasets of three specific tasks. For SA, we select training sets of SST and ChnSentiCorp[6] to train models for English and Chinese respectively. For STS, training sets of QQP and LCQMC are used to train English and Chinese models. For MRC, SQUAD2.0 and DuReader are used as training sets for English and Chinese respectively. For each task, we select the best model on the original dev set.

In order to confirm the correctness of our implementation, Table 5 shows model performances on both original dev/test and our evaluation datasets. We can see that our re-implemented models output close results reported in related works (Liu et al., 2018; WANG and JIANG; Liu et al., 2019). Meanwhile, the results of Chinese SA and MRC tasks decrease significantly on our evaluation sets. This may be caused by the poor generalization and robustness of the model, as our evaluation datasets contain perturbed examples and Chinese data for SA is not from the ChnSentiCorp dataset.

5.2 Evaluation Results

Table 6 shows the evaluation results of interpretability from the plausibility and faithfulness perspectives. Within the scope of baseline models and saliency methods used in our experiments, there are three main findings. First, based on all models and saliency methods used in our experiments, our metrics for interpretability evaluation, namely token-F1 score and MAP, are more fine and generic, especially MAP, which applies to all three tasks. Second, IG method performs better on plausibility and ATT method performs better on faithfulness. Meanwhile, ATT method achieves best performance in sentence-pair tasks. Third, with all three saliency methods, in these three tasks, LSTM model is comparable with transformer model (i.e., RoBERTa based model in our experiments) on interpretability, though LSTM performs worse than transformer in term of accuracy. We think that the generalization ability of LSTM model is weak, leading to low accuracy, even with relatively reasonable rationales.

In the following paragraphs, we first give a comparison between our proposed metrics and those

[6] https://github.com/pengming617/bert_classification

Models + Methods	SA					STS					MRC		
	Plausibility		Faithfulness			Plausibility		Faithfulness			Plausibility		Faithfulness
	Token-F1↑	IOU-F1↑	MAP↑	Suf↓	Com↑	Token-F1	IOU-F1	MAP	Suf	Com	Token-F1	IOU-F1	MAP
LSTM + IG	36.9	12.1	67.2	-0.025	0.708	54.1	17.3	69.0	0.048	0.441	40.7	11.0	72.3
RoBERTa-base + IG	37.4	10.4	64.1	0.059	0.392	52.9	24.2	65.3	0.153	0.478	42.1	11.0	66.9
RoBERTa-large + IG	35.0	7.9	40.6	0.130	0.260	52.7	35.9	49.7	0.224	0.400	18.0	0.1	18.0
LSTM + ATT	36.6	12.4	67.8	0.123	0.298	49.6	11.8	76.0	0.221	0.313	19.9	0.4	88.3
RoBERTa-base + ATT	33.2	9.4	69.2	0.267	0.128	66.5	54.2	73.6	0.185	0.337	22.6	2.6	55.0
RoBERTa-large + ATT	23.3	3.1	75.9	0.301	0.095	56.8	35.9	75.4	0.136	0.399	26.6	1.3	76.0
LSTM + LIME	36.6	11.3	63.2	-0.040	0.762	54.5	19.2	60.0	0.134	0.311	-	-	-
RoBERTa-base + LIME	41.5	13.8	61.0	0.032	0.568	58.7	34.9	70.5	0.064	0.509	-	-	-
RoBERTa-large + LIME	41.4	14.3	62.9	0.053	0.505	61.2	42.3	71.8	0.019	0.524	-	-	-

Table 6: Interpretability evaluation results on English datasets of three tasks. The metric with ↑ means the higher the score, the better the performance. Conversely, ↓ means a low score represents a good performance. As LIME is specially designed for classification tasks, we have not applied it to MRC. Meanwhile, the sufficiency score (`Suf`) and the comprehensiveness score (`Com`) are also only suitable for classification tasks, as shown in Equation 4. Thus we do not report these two scores on MRC.

used in related studies. Then we give a detailed analysis about the interpretability results of three saliency methods and three evaluated models.

Comparison between Evaluation Metrics We report results of token-F1 and IOU-F1 scores for plausibility. The higher the scores, the more plausible the rationales. It can be seen that the two metrics have the similar trends in all three tasks with all three saliency methods. But token-F1 is much precise than IOU-F1, as the IOU-F1 score of a rationale is 1 only if its overlap with ground truth is no less than 0.5 (Equation 3). However, in all three tasks, overlaps of most instances are less than 0.5, especially in the task with a low *RLR*. Thus IOU-F1 is too coarse to evaluate token-level rationales. Instead, token-F1 focuses on evaluating token impact on model predictions, so as to be more suitable for evaluating compact rationales.

For faithfulness evaluation, we report results of MAP, sufficiency and comprehensiveness scores. We can see that our proposed MAP is an efficient metric for faithfulness evaluation. Specifically, it applies to most tasks, especially non-classification tasks. Moreover, in the two classification tasks (i.e., SA and STS), with IG and LIME methods, MAP has the same trend as the other two metrics over all three models, which further indicates that MAP can well evaluate the faithfulness of rationales. With ATT method, there is no consistent relationship between these three metrics. We think this is because the calculations of sufficiency and comprehensiveness scores with ATT method are not accurate and consistent enough. For example, in the SA task, from the comparison of three saliency methods with LSTM model, we can see that the rationales extracted by these methods have similar plausibility scores, but the sufficiency score with ATT method is much higher than that with the other two methods. Please note that a low sufficiency score means a sufficient rationale. Similarly, in the STS task with RoBERTa-base model, the rationales extracted by ATT method have a higher plausibility score, as well as a higher sufficiency score. Finally, we believe that other metrics can be proposed based on our benchmark.

Evaluation of Saliency Methods LIME, which uses a linear model to approximate a DL classification model, is model-agnostic and task-agnostic. It obtains the highest performance on token-F1 and sufficiency scores in SA and STS tasks, as the rationales extracted by it more accurately approximate the decision process of DL models. But how to better apply LIME to more NLP tasks is very challenging and as the future work.

When comparing IG and ATT, we find ATT performs better on faithfulness and sentence-pair tasks. In SA and MRC, IG performs better on plausibility and ATT method achieves better results on faithfulness, which is consistent with prior works (Jain and Wallace, 2019; DeYoung et al., 2020). In STS, ATT method achieves higher results both on plausibility and faithfulness than IG method. We think this is because the cross-sentence interaction attentions are more important for sentence-pair tasks. Interestingly, on all three tasks, there is a positive correlation between MAP (faithfulness) and token-F1 (plausibility) with IG method.

Evaluation of Models While analyzing interpretability of model architectures, we mainly focus on IG and ATT methods, as LIME is model-agnostic. We find that interpretability of model architectures vary with saliency methods and tasks.

Compared with transformer models, based on IG method, LSTM is competitive on plausibility and performs better on faithfulness in all three tasks. On the contrary, based on ATT method, transformer models outperform LSTM on plausibility and are competitive on faithfulness in STS and MRC tasks. As discussed above, the interaction between inputs is more important in these two tasks.

From the comparison between two transformer models with different parameter sizes, i.e., RoBERTa-base and RoBERTa-large, we find that RoBERTa-base outperforms RoBERTa-large on plausibility with these two saliency methods. Interestingly, for faithfulness evaluation, RoBERTa-base performs better than RoBERTa-large with IG method, and RoBERTa-large performs better than RoBERTa-base with ATT method.

We believe these findings are helpful to the future work on interpretability.

6 Limitation Discussion

We provide a new interpretability evaluation benchmark which contains three tasks with both English and Chinese annotated data. There are three limitations in our work.

- How to evaluate the quality of human-annotated rationales is still open. We have several annotators to perform quality control based on human intuitions and experiences. Meanwhile, we compare model behaviors on full inputs and human-annotated rationales to evaluate the sufficiency and comprehensiveness of rationales, as shown in Table 4 and Table 7. However, this manner has damaged the original input distribution and brings uncontrollable factors on model behaviors. Therefore, how to automatically and effectively evaluate the quality of human-annotated rationales should be studied in the future.

- We find that the interpretability of model architectures and saliency methods vary with tasks, especially with the input form of the task. Thus our benchmark should contain more datasets of each task type (e.g., single-sentence task, sentence-pair similarity task and sentence-pair inference task) to further verify these findings. And we will build evaluation datasets for more tasks in the future.

- Due to space limitation, there is no analysis of the relationships between metrics, e.g., the relationship between plausibility and accuracy, and the relationship between faithfulness and robustness. We will take these analyses in our future work.

Finally, we hope more evaluation metrics and analyses are proposed based on our benchmark. And we hope our benchmark can facilitate the research progress of interpertability.

7 Conclusion

We propose a new fine-grained interpretability evaluation benchmark, containing token-level rationales, a new evaluation metric and corresponding perturbed examples for three typical NLP tasks, i.e., sentiment analysis, textual similarity and machine reading comprehension. The rationales in this benchmark meet primary properties that a rationale should satisfy, i.e., sufficiency, compactness and comprehensiveness. The experimental results on three models and three saliency methods prove that our benchmark can be used to evaluate interpretability of both models and saliency methods. We will release this benchmark and hope it can facilitate progress on several directions, such as better interpretability evaluation metrics and causal analysis of NLP models.

Acknowledgements

We are very grateful to our anonymous reviewers for their helpful feedback on this work. This work is supported by the National Key Research and Development Project of China (No.2018AAA0101900).

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems, 31.

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, Brussels, Belgium. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. " what is relevant in a text document?": An interpretable machine learning approach. PloS one, 12(8):e0181142.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. Advances in Neural Information Processing Systems, 31.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3243–3255, Online. Association for Computational Linguistics.

Li Deng and Dong Yu. 2014. Deep learning: methods and applications. Foundations and trends in signal processing, 7(3–4):197–387.

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, Online. Association for Computational Linguistics.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5034–5052, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Gerhard Fischer, Thomas Mastaglio, Brent Reeves, and John Rieman. 1990. Minimalist explanations in knowledge-based systems. In Twenty-Third Annual Hawaii International Conference on System Sciences, volume 3, pages 309–317. IEEE.

Karën Fort and Alain Couillault. 2016. Yes, we care! results of the ethics and natural language processing surveys. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1593–1600, Portorož, Slovenia. European Language Resources Association (ELRA).

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 9:346–361.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In Proceedings of the Workshop on Machine Reading for Question Answering, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. NIPS 2017 Symposium on Interpretable Machine Learning.

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

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, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment rationale for natural language inference. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5372–5387, Online. Association for Computational Linguistics.

Robert Kass, Tim Finin, et al. 1988. The need for user models in generating expert system explanations. International Journal of Expert Systems, 1(4).

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 107–117, Austin, Texas. Association for Computational Linguistics.

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

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. Medical image analysis, 42:60–88.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1952–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In AAAI.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sina Mohseni, Jeremy E Block, and Eric D Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075.

W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In International Conference on Learning Representations.

Weili Nie, Yang Zhang, and Ankit Patel. 2018. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In International Conference on Machine Learning, pages 3809–3818. PMLR.

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5):1–36.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Leon Sixt, Maximilian Granz, and Tim Landgraf. 2020. When explanations lie: Why many modified bp attributions fail. In International Conference on Machine Learning, pages 9046–9057. PMLR.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

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, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3319–3328.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shuohang WANG and Jing JIANG. Machine comprehension using match-lstm and answer pointer.(2017). In ICLR 2017: International Conference on Learning Representations, Toulon, France, April 24-26: Proceedings, pages 1–15.

Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of shap for alert processing. Proceedings of KDD workshop on Explainable AI 2019 (KDD-XAI).

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1599–1615, Online. Association for Computational Linguistics.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2631–2647, Dublin, Ireland. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

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, Honolulu, Hawaii. Association for Computational Linguistics.

A Other Details of Our Datasets

Other Annotation Details We give more details about data collection, annotator information, annotation training and payment, and instructions for data usage.

Data collection. Except for Chinese data of SA, the annotated instances for other datasets are collected from the existing datasets, as described in Section 3.1. In the process of collection, we ask annotators to discard instances that contain: 1) offensive content, 2) information that names or uniquely identifies individual people, 3) discussions about politics, guns, drug abuse, violence or pornography.

Annotator information. We have two ordinary annotators for each task, and three senior annotators for all tasks. The ordinary annotators annotate the rationales and modify the rationales according to the scores from the senior annotators. They are college students majoring in languages. Our senior annotators are full-time employees, and perform quality control. Before this work, they have lots of experience in annotating data for NLP tasks.

Annotation training and payment. Before real annotation, we train all annotators for several times so that they understand the specific task, rationale criteria, etc. During real annotation, we have also held several meetings to discuss common mistakes and settle disputes. Our annotation project for each task lasts for about 1.5 month. And we cost about 15.5 RMB for the annotation of each instance.

Instructions of data annotation and usage. Before annotation, we provide a full instruction to all annotators, including the responsibility for leaking data, disclaimers of any risks, and screenshots of annotation discussions. Meanwhile, our datasets are only used for interpretability evaluation. And we will release a license with the release of our benchmark.

Data Analysis We report sufficiency and comprehensiveness scores of human-annotated rationales, as shown in Table 7. The sufficiency scores of human-annotated rationales are lower than those of rationales provided by transformer models or extracted by IG and ATT methods. We can conclude that our human-annotated rationales are sufficient. However, with IG and LIME methods, the comprehensiveness scores of human-annotated rationales are lower than those of rationales provided by models. As discussed before, the model performance on non-rationales is not accurate enough, as shown by Acc^{nr}, which achieves about 50% on non-rationales. How to effectively evaluate the quality of human-annotated rationales should be studied in the future.

B Implementations Details

B.1 Implementations of Evaluated Models

We utilize HuggingFace's Transformer (Wolf et al., 2019) to implement RoBERTa based models for three tasks. Please refer to their source codes[7] for more details. The LSTM model architectures for three tasks are shown in Figure 2.

B.2 Implementations of Saliency Methods

We first describe experimental setups for three saliency methods. Then we introduce implementation details of attention-based method. Finally, we illustrate the limitations of LIME in STS and MRC tasks.

Experimental setup. In IG-based method, token importance is determined by integrating the gradient along the path from a defined baseline x_0 to the original input. In the experiments, a sequence of all zero embeddings is used as the baseline x_0. And the step size is set to 300.

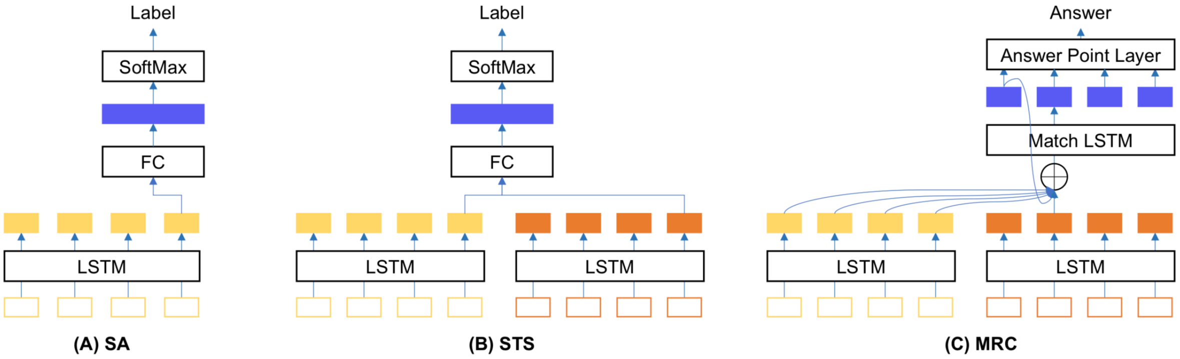
LIME uses the token weight learned by the linear model as the token's importance score. For each original input, N perturbed samples which contains K tokens of it are created. Then the weighted square loss is used to optimize the selection of tokens that are useful for the model prediction. In the experiments, we set N to $5,000$ and K to 10. In the STS task, an input is a pair of two instances. Each perturbed sample for an input consists of a perturbed example for one instance and the original input for the other instance.

ATT method on LSTM models. Figure 2 shows the architectures of LSTM models in three tasks. In the SA task, given the input instance Q, an LSTM encoder is used to get the representation for each token, denoted as h_i^Q. And a full connected layer (FC) is used to get the instance representation based on the last hidden representation. We use h^{fc} to represent the representation after the FC layer. Then the instance representation h^{fc} is fed into the softmax layer to get the predicted label. The attention weight for token i in Q is calculated by $\frac{h^{fc} \cdot h_i^Q}{\sum_{j=1}^{|Q|} h^{fc} \cdot h_j^Q}$, where $|Q|$ represents the number of tokens in Q. Then the attention weight of the

[7] https://huggingface.co/transformers/

Models	SA					STS					MRC	
	Acc^f	Acc^r	Acc^{nr}	Suf	Com	Acc^f	Acc^r	Acc^{nr}	Suf	Com	F1^f	F1^r
English												
LSTM	78.2	**86.2**	60.7	0.151	0.217	74.6	69.8	61.3	0.152	0.291	54.4	53.4
RoBERTa-base	93.8	92.4	70.6	0.084	0.251	92.7	89.3	54.8	0.075	0.418	71.7	**80.8**
RoBERTa-large	95.4	91.5	74.4	0.086	0.234	93.2	88.8	53.9	0.085	0.420	76.0	**76.7**
Chinese												
LSTM	60.0	**70.4**	48.7	0.172	0.135	75.2	**80.7**	51.2	0.083	0.339	66.4	**82.2**
RoBERTa-base	59.8	**77.0**	50.2	0.252	0.207	85.5	**88.1**	48.8	0.048	0.399	65.8	**89.3**
RoBERTa-large	62.6	**80.6**	47.6	0.212	0.147	86.0	**87.4**	48.9	0.051	0.433	67.8	**83.3**

Table 7: Model performance on the original full input (Acc^f), human-annotated rationale (Acc^r), and non-rationale (Acc^{nr}) by removing human-annotated rationale from the original full input. Suf and Com represent the sufficiency score and comprehensiveness score of the human-annotated rationales, as shown in Equation 4. We do not report F1^{nr} on the MRC task, as the golden answer is not from the non-rationale.



(A) SA

(B) STS

(C) MRC

Figure 2: LSTM model architectures for three tasks.

token is used as its importance score for the model prediction.

Similarly, in the STS task, the model architecture is mostly the same as that of SA. The main difference is that the input of STS consists of two instances, denoted as Q and P, and the concatenation of their last hidden representations is fed into an FC layer. Then, referring to the attention weight calculation of Q, the attention weight for the token in P is calculated by $\frac{h^{fc}\cdot h_i^P}{\sum_{j=1}^{|P|} h^{fc}\cdot h_j^P}$, where $|P|$ represents the number of tokens in P. For each instance in a pair, we select top-k^d important tokens as the rationale.

In the MRC task, the input also consists of two sequences: the question Q and the passage P. We adopt the match-LSTM model (WANG and JIANG) as our baseline model. The match-LSTM model uses two LSTMs to encode the question and passage respectively. Then it uses the standard word-by-word attention mechanism to obtain the attention weight for each token in the passage. And the final representation of each token in the passage is obtained by combining a weighted version of the question. We use $\bar{h}_i^P$ to represent the representation of i-th token in the passage. Then the importance score of j-th token is calculated by Equation 5.

$$a_j = \frac{\sum_{i=1}^{|Q|} e_{ij}}{|Q|} \qquad e_{ij} = \frac{h_i^Q \cdot \bar{h}_j^P}{\sum_{k=1}^{|Q|} h_i^Q \cdot \bar{h}_k^P} \tag{5}$$

where a_j is used as the importance score of token j.

ATT method on pre-trained models. Following related studies (Jain and Wallace, 2019; DeYoung et al., 2020), on transformer-based pre-trained models, attention scores are taken as the self-attention weights induced from the [CLS] token index to all other indices in the last layer. As the pre-trained model uses wordpiece tokenization, we sum the self-attention weights assigned to its constituent pieces to compute a token's score. Meanwhile, as the pre-trained model has multi-heads, we average scores over heads to derive a final score. In the MRC task, for each token in the passage, importance score is taken as the average self-attention weights induced from this token index to all indices of the question in the last layer.

Limitations of LIME. Given an input, LIME constructs a token vocabulary for it and aims to assign an important score for each token in this vocabulary. That is to say, for the token that appears multiple times, LIME neglects its position

Models + Methods	SA					STS					MRC		
	Plausibility		Faithfulness			Plausibility		Faithfulness			Plausibility		Faithfulness
	Token-F1↑	IOU-F1↑	MAP↑	Suf↓	Com↑	Token-F1	IOU-F1	MAP	Suf	Com	Token-F1	IOU-F1	MAP
LSTM + IG	38.2	9.8	60.6	-0.131	0.707	68.2	61.5	58.6	0.336	0.419	19.9	0.6	87.1
RoBERTa-base + IG	35.2	12.5	51.5	0.118	0.489	71.9	71.4	62.1	0.139	0.470	34.0	9.1	67.9
RoBERTa-large + IG	37.9	12.9	43.6	0.123	0.381	71.8	72.0	58.1	0.251	0.547	25.2	1.7	61.9
LSTM + ATT	24.0	9.8	72.6	0.171	0.225	72.7	72.1	77.3	0.110	0.359	2.7	0.0	79.6
RoBERTa-base + ATT	25.7	6.0	69.5	0.191	0.320	67.2	55.4	71.3	0.201	0.399	28.5	5.3	61.4
RoBERTa-large + ATT	30.7	8.2	67.9	0.173	0.248	68.0	59.8	67.0	0.251	0.547	28.5	5.5	48.8
LSTM + LIME	38.6	10.1	59.4	-0.130	0.701	74.8	79.0	65.9	-0.015	0.411	-	-	-
RoBERTa-base + LIME	37.3	14.3	56.6	0.051	0.660	77.3	83.2	74.8	-0.041	0.494	-	-	-
RoBERTa-large + LIME	39.0	14.5	53.0	-0.013	0.653	76.8	82.9	74.3	-0.024	0.562	-	-	-

Table 8: Interpretability evaluation results on Chinese datasets of three tasks.

information and only assigns one score for it. However, in STS and MRC, the position of a token is very important. Therefore, It can not guarantee the effectiveness of evaluation on these two tasks with LIME. In addition, as LIME is designed for classification models, it is difficult to apply it to the MRC task.

C Interpretability Evaluation on Chinese Datasets

We report interpretability results of three baseline models with three saliency methods on Chinese evaluation datasets in Table 8. It can be seen that interpretability results on Chinese datasets have the similar trends as those on English datasets. Different from the conclusions on English datasets, on all three tasks, IG-based method outperforms ATT-based method on plausibility. And ATT method performs better than IG on faithfulness in SA and STS tasks.