Contrastive Learning with Adversarial Examples for Alleviating Pathology of Language Model

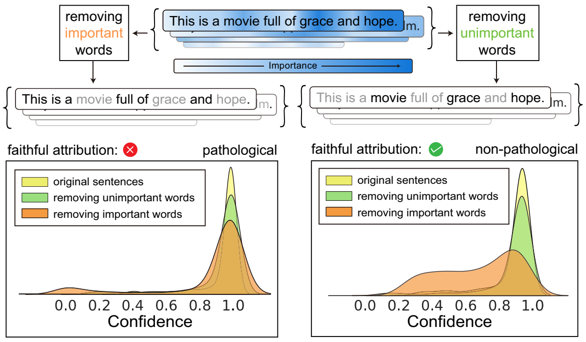
Pengwei Zhan[§‡], **Jing Yang**[§*], **Xiao Huang**[§], **Chunlei Jing**[§], **Jingying Li**[§], **Liming Wang**[§]
[§]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[‡]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{zhanpengwei,yangjing,huangxiao}@iie.ac.cn
{jingchunlei,lijingying,wangliming}@iie.ac.cn

Abstract

Neural language models have achieved superior performance. However, these models also suffer from the pathology of overconfidence in the out-of-distribution examples, potentially making the model difficult to interpret and making the interpretation methods fail to provide faithful attributions. In this paper, we explain the model pathology from the view of sentence representation and argue that the counter-intuitive bias degree and direction of the out-of-distribution examples' representation cause the pathology. We propose a **Con**trastive learning regularization method using **A**dversarial examples for **A**lleviating the **P**athology (ConAAP), which calibrates the sentence representation of out-of-distribution examples. ConAAP generates positive and negative examples following the attribution results and utilizes adversarial examples to introduce direction information in regularization. Experiments show that ConAAP effectively alleviates the model pathology while slightly impacting the generalization ability on in-distribution examples and thus helps interpretation methods obtain more faithful results.

1 Introduction

Neural language models have achieved superior performance in various natural language processing (NLP) domains and are used in many fields to accomplish critical tasks, such as toxic comment classification and rumor detection. However, the drawbacks of NLP models in test-time interpretability pose potential risks to these tasks, as existing interpretation methods always fail to obtain faithful attributions on these models, thereby failing to reveal potential flaws and biases.



(a) Normal (b) ConAAP

Figure 1: Confidence distribution comparison between BERT tuned with normal method and ConAAP. We remove words of different importance on normal examples in testing set (operation is detailed in §3.3). The attribution is obtained by gradient-based method (§3.2). The normally tuned model is pathological, as the confidence distribution after removing important words is similar to after removing unimportant words, indicating that the interpretation method can not obtain faithful attributions. The model tuned with ConAAP is non-pathological, as the model can discriminate between the important and unimportant words in terms of confidence changing, and the attributions are more faithful.

Following Ribeiro et al. (2016), Schwab and Karlen (2019), and Situ et al. (2021), the attribution obtained by a faithful interpretation method should indicate the real contribution of features in terms of model confidence changing. Specifically, the important words marked by a faithful attribution should contribute most to the model prediction, and masking them out from the sentence should greatly decrease model confidence. Conversely, unimportant words should have little impact on prediction and confidence. However, abnormal model behaviors have been widely reported in previous works. For example, Goodfellow et al. (2015) illustrate that a well-trained model will sometime predict pure noise rubbish examples, which should contain only the unimportant features, with high confidence. Feng et al. (2018) also find that model tends to predict meaningless examples with tokens removed with higher confidence than normal examples. We also demonstrate similar abnormal behavior and the unfaithfulness of attribution by showing the confidence distribution on the Movie Review

[*] Corresponding Author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6493–6508
July 9-14, 2023 ©2023 Association for Computational Linguistics

(MR) testing set (Pang and Lee, 2005) of the basic version BERT (Devlin et al., 2019) fine-tuned on MR training set in Figure 1.

According to Guo et al. (2017) and Feng et al. (2018), *model pathology* is a major reason for these abnormal behaviors. They argue that neural language models are overconfident in their prediction as the model overfits the negative log-likelihood loss to produce low-entropy distribution over classes. Thus the model will also be overconfident in examples outside the distribution of training instances, leading to the counter-intuitive model confidence in these abnormal behaviors. Empirically, Feng et al. (2018) also demonstrate the explanation by mitigating the pathology with an entropy regularization that maximizes the uncertainty on out-of-distribution examples. Following their findings, we argue that the interpretation method fails to provide faithful results is mainly due to the drawback of models rather than the drawback of the interpretation method itself, i.e., the unfaithfulness of attribution is due to the model pathology.

In this paper, we explain the model pathology, which potentially makes the model difficult to interpret, from the view of *sentence representation*, and intuitively show *how the pathology leads to unfaithfulness attribution* and *how to alleviate the pathology effectively*. Based on our findings, we also propose a **Con**trastive learning regularization method using **A**dversarial examples for **A**lleviating the **P**athology (ConAAP). We summarize our main contributions as follows:

1. We explain the model pathology and how it causes the unfaithfulness attribution from the view of sentence representation. We argue that the counter-intuitive *bias degree* and *bias direction* of the out-of-distribution examples are two key factors leading to the pathology.

2. We propose ConAAP, a contrastive learning regularization method that calibrates the sentence representation of out-of-distribution examples. ConAAP generates positive and negative examples following the attribution results and utilizes adversarial examples to introduce direction information in regularization.

3. Experiments show that ConAAP effectively alleviates the model pathology while slightly impacting the generalization ability on in-distribution examples and thus helps interpretation methods obtain more faithful results.

2 Related Work

Interpreting the Language Model. To interpret a language model, previous works utilize the gradient-based method (Li et al., 2016; Sundararajan et al., 2017; Ross et al., 2017; Zhan et al., 2022a; Feng et al., 2018; DeYoung et al., 2020), attention scores (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), Occlusion (Gao et al., 2018; Li et al., 2019; Jin et al., 2020; Zhan et al., 2022b; Li et al., 2020), and Shapley values (Lundberg and Lee, 2017) to attribute the model prediction. To quantitatively evaluate the faithfulness of the obtained attribution, metrics including *Reduced Length* (Feng et al., 2018), *Comprehensiveness*, *Sufficiency*, and *Area Over the Perturbation Curve (AOPC)* (DeYoung et al., 2020) are proposed.

Contrastive Learning. Encouraged by the remarkable success of contrastive learning in computer vision (CV) in learning better representation and improving performance on downstream tasks (Chen et al., 2020b,a; Pan et al., 2021), various methods have been proposed for NLP tasks. Limited by the discrete nature of text, instead of generating contrastive pairs by cropping, resizing, and rotating the input like in CV tasks, previous works in NLP are always by back-translating, word deleting, reordering, and substituting (Giorgi et al., 2021; Wu et al., 2020; Gao et al., 2021). It is shown that contrastive learning helps improve sentence representation and model performance on downstream NLP tasks. However, few works focus on model pathology and interpretability.

Adversarial Examples in Contrastive Learning. It is found that using adversarial examples, which can fool the model while being imperceptible to humans (Gao et al., 2018; Li et al., 2019; Jin et al., 2020; Li et al., 2020), in contrastive learning, can produce better sentence representations and increase downstream performance. However, previous works always utilize adversarial examples as challenging examples and focus on the model robustness and performance (Kim et al., 2020; Ho and Vasconcelos, 2020; Meng et al., 2021) rather than the model pathology and interpretability.

3 Method

3.1 Preliminaries

Given a data distribution $\mathcal{D}$ over input text $\boldsymbol{X} \in \mathcal{X}$ and output labels $Y \in \mathcal{Y} = \{1, \ldots, C\}$, a model

$f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ maps the input text to the output softmax probability, which is trained by minimizing the empirical risk $\mathcal{L}_{ce}(\boldsymbol{X}, Y; \boldsymbol{\theta})$ that equals to

$$\mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}[-\log \frac{\exp(w_Y^T r_{\boldsymbol{\theta}}(\boldsymbol{X}))}{\sum_{k=1}^{C} \exp(w_k^T r_{\boldsymbol{\theta}}(\boldsymbol{X}))}] \quad (1)$$

where $\mathcal{W}$ is the classification parameters, $w_Y \in \mathcal{W}$ denotes the classification parameters toward class Y, $\boldsymbol{\theta}$ is the model parameters, and $r_{\boldsymbol{\theta}}(\cdot)$ denotes the sentence representation of input text. Specifically, in classification tasks, BERT always uses the value of [CLS] token as representation, while other models, including LSTM and CNN, always use the average token embedding before the last dense layer. After training, the model correctly classifies text based on the posterior probability:

$$\mathcal{P}(Y_{\textit{true}}|\boldsymbol{X}) = \frac{\exp(w_{\textit{true}}^T r_{\boldsymbol{\theta}}(\boldsymbol{X}))}{\sum_{k=1}^{C} \exp(w_k^T r_{\boldsymbol{\theta}}(\boldsymbol{X}))} \quad (2)$$

where $w_{\textit{true}}$ denotes the classification parameters toward the ground-truth class $Y_{\textit{true}}$. This value is always regarded as the confidence in prediction.

3.2 Faithful Attribution

In this paper, we use the gradient-based method as the basic interpretation method to obtain attribution, which is formally defined as follows:

$$\textit{Attr}(\boldsymbol{X}) = \left(\left\| \frac{\partial\, w_{\textit{true}}^T r_{\boldsymbol{\theta}}(\boldsymbol{X})}{\partial\, \textit{emb}(x_i)} \right\|_2 \right)_{i\in\{1,\dots,N\}} \quad (3)$$

where $\boldsymbol{X} = x_1 x_2 \dots x_N$ is a normal sentence, $\textit{emb}(\cdot)$ denotes the word embedding. To measure the faithfulness of the obtained attribution, previous works always measure the influence of words of different importance on model confidence. We use the *Area Over the Perturbation Curve (AOPC)* form of Comprehensiveness (*Comp.*) and Sufficiency (*Suff.*) metrics (DeYoung et al., 2020; Samek et al., 2017; Nguyen, 2018) to measure the faithfulness. $\textit{AOPC}_{\textit{Comp.}}$ is formulated as

$$\frac{1}{K+1} \sum_{k=1}^{K} \mathcal{P}(Y_{\textit{true}}|\boldsymbol{X}) - \mathcal{P}(Y_{\textit{true}}|t_{/k}^{\textit{imp}}(\boldsymbol{X})) , \quad (4)$$

and $\textit{AOPC}_{\textit{Suff.}}$ is formulated as

$$\frac{1}{K+1} \sum_{k=1}^{K} \mathcal{P}(Y_{\textit{true}}|\boldsymbol{X}) - \mathcal{P}(Y_{\textit{true}}|t_{/k}^{\textit{ump}}(\boldsymbol{X})) , \quad (5)$$

where $t_{/k}^{\textit{imp}}(\cdot)$ means remove the k most important words in a sentence according to attribution, while $t_{/k}^{\textit{ump}}(\cdot)$ means remove the k least important words, K indicates the range of words to be considered. If attribution is faithful, it is expected to have a high $\textit{AOPC}_{\textit{Comp.}}$ value and a low $\textit{AOPC}_{\textit{Suff.}}$ value, indicating that the information in the important words has an overall larger impact on prediction than in unimportant words.

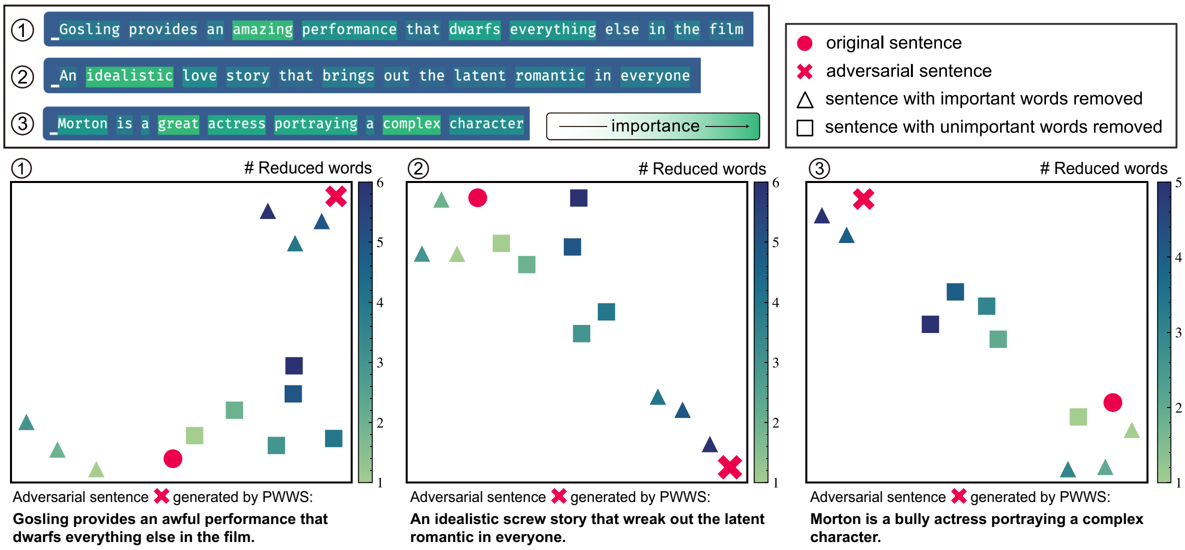
3.3 Model Pathology From the View of Sentence Representation

In this section, we explain the model pathology from the view of sentence representation and try to answer *how does the pathology lead to unfaithfulness attribution?*

Feng et al. (2018) propose an analysis method called input reduction, which iteratively calculates the attribution and removes the least important word in a sentence. By analyzing the model confidence change on the incomplete sentence, they find that when the reduced examples are nonsensical for humans and lack information for supporting the prediction, the models still make the same prediction as the original sentence with high confidence. The counter-intuitive high confidence is attributed to the model overconfidence in such out-of-distribution examples.

To make the analysis process more compatible with the calculation of faithfulness (4) (5), we use a variant reduction method to generate incomplete out-of-distribution examples rather than the one proposed by Feng et al. (2018). Specifically, given a sentence and a well-trained model, we first obtain the attribution of the sentence according to (3), and then *cumulatively* remove the words in the sentence. We remove not only the unimportant words but also the important words. For the important words, we cumulatively remove 50% of words in descending order of the attribution. For the unimportant words, we cumulatively remove 50% of words in ascending order of the attribution. Additionally, we generate adversarial example, which is imperceptible to humans and can mislead the model prediction, from the given normal sentence with PWWS (Ren et al., 2019). Therefore, we have four kinds of examples: (i) the in-distribution normal example, (ii) the out-of-distribution examples with important words removed, (iii) the out-of-distribution examples with unimportant words removed, and (iv) the adversarial example located on the other side and in the vicinity of the decision boundary.

Following these operations, we fine-tune a basic



Adversarial sentence ✖ generated by PWWS: **Gosling provides an awful performance that dwarfs everything else in the film.**

Adversarial sentence ✖ generated by PWWS: **An idealistic screw story that wreak out the latent romantic in everyone.**

Adversarial sentence ✖ generated by PWWS: **Morton is a bully actress portraying a complex character.**

Figure 2: The visualization of sentence representations and the attribution obtained by gradient-based interpretation on MR instances. For the representation visualization, darker △ and □ indicate out-of-distribution examples with more words removed. The out-of-distribution examples closer to the original example are more likely to maintain the same model prediction as the original example, while the examples closer to the adversarial example tend to decrease the confidence in the original class as the adversarial example leads to different predictions and is located in the vicinity of the decision boundary. For the attribution, darker colors indicate higher importance.

BERT on MR training set and obtain the sentence representations of the four kinds of examples derived from the MR testing set instances. We then project the representations to a two-dimensional space with t-SNE (van der Maaten and Hinton, 2008). The visualization of the sentence representation of three MR instances and their attributions according to (3) are shown in Figure 2. We can summarize some counter-intuitive phenomena.

Observation 1: When the most important few words are removed, the representations of such incomplete out-of-distribution examples are still very close to the original sentence. Intuitively, the most important few words should contain the most significant information for supporting the prediction. Losing this information, the model confidence should decrease, and the representation of such incomplete sentences should be close to the adversarial example, which is located on the other side and in the vicinity of the decision boundary. Focusing on instance ①, when the three most important words *amazing* (▲), *dwarfs* (▲), and *everything* (▲) are removed from the instance, the sentence is transformed into *"Gosling provides an ~~amazing~~ performance that ~~dwarfs everything~~ else in the film."*, which is unfathomable to humans and does not contain any information supporting classifying this incomplete sentence into any class (positive or negative). However, the representation of this sentence (▲) is still close to the original sentence (●), indicating that the model still regards it belongs to the original class with high confidence.

Observation 2: When unimportant words are removed, the representations of such incomplete out-of-distribution examples are biased away from the original sentence more than expected. Intuitively, the unimportant words should contain low-impact information to support the prediction. Losing this unimportant information, the model confidence should almost not change, and the representations of such incomplete sentences should still be close to the original sentence. Focusing on instance ①, when the six least important words *else* (■), *film* (■), *Gosling* (■), *an* (■), *in* (■), and *the* (■) are removed from the instance, the sentence is transformed into *"~~Gosling~~ provides ~~an~~ amazing performance that dwarfs everything ~~else in the film~~."*. Even though this sentence is grammatically incorrect, it is still easy for humans to classify it as a positive example. However, the representation of this incomplete sentence (■) is largely biased from the original examples (●) and is even closer to the adversarial example (✖) than the sentence with three important words removed (▲), indicating that the model predicts this out-of-distribution examples with lower confidence.

Similar phenomena can also be observed in instances ② and ③. More results can be found in Appendix B.2. Based on **Observation 1** and **Observation 2**, we can answer the question raised before from the view of sentence representation: When important words are masked out from the sentence, the representations of such out-of-distribution examples are sometimes too close to the original sentence, maintaining the high model confidence, even if such examples do not contain any information supporting the prediction. When unimportant words are masked from the sentence, the representations of such out-of-distribution examples are sometimes largely biased away from the original sentence and are approaching the decision boundary, decreasing the model confidence, even if such examples are still easy for humans to classify.

Appendix B.1 provides further study on the distance between out-of-distribution sentences and the in-distribution normal sentence, which supports our claim on Observation 1 and Observation 2.

3.4 Contrastive Learning with Adversarial Examples for Alleviating the Pathology

In this section, we try to answer *how to alleviate the pathology effectively?* We also detail the proposed ConAAP regularization method. According to our analysis, the model pathology can be explained by the counter-intuitive sentence representation distribution of out-of-distribution examples. Therefore, a natural way to alleviate the pathology is to calibrate their distribution. To calibrate the sentence representation, we should focus on both the *bias degree* and *bias direction*.

For the out-of-distribution examples with *unimportant* words removed, which are always used to measure the $AOPC_{Suff.}$ value, we try to decrease the bias degree of their representation from the original normal example, as most of these examples are still easy to classify. For the out-of-distribution examples with *important* words removed, which are always used to measure the $AOPC_{Comp.}$ value, we try to increase the bias degree of their representation from the original normal example, as these examples are more difficult to classify. However, if they are pushed away from the original example in a direction away from the decision boundary, the counter-intuitive high confidence will still be maintained. Therefore, we also simultaneously force their bias direction toward the decision boundary, which is indicated by the adversarial example.

To achieve the calibration, we reuse the word removal operation we proposed in §3.3 and used in Figure 2. The operation to delete important words is defined as t^{neg}, and the operation to delete unimportant words is defined as t^{pos}. We also define the operation that generates adversarial examples as t^{adv}. To formulate the contrastive loss objective of ConAAP, for convenience, we first define the calculation $\mathcal{S}$:

$$\mathcal{S}_{(i,j)}^{(k,l)} = \exp(\mathrm{sim}[r_{\boldsymbol{\theta}}(\boldsymbol{X}_i^k), r_{\boldsymbol{\theta}}(\boldsymbol{X}_j^l)]/\tau) \tag{6}$$

where sim denotes the cosine similarity, i.e., $\mathrm{sim}[\boldsymbol{r_i}, \boldsymbol{r_j}] = \boldsymbol{r_i^T r_j}/\|\boldsymbol{r_i}\|\|\boldsymbol{r_j}\|$. k, l denotes the example type, and $k, l \in \{neg, pos, adv, \cdot\}$, which respectively indicates the example $\boldsymbol{X}^{neg}$, $\boldsymbol{X}^{pos}$, $\boldsymbol{X}^{adv}$ sampled from the examples generated by the operations t^{neg}, t^{pos}, t^{adv}, and the normal example. i, j are the example indexes. τ is a temperature parameter similar to the normalized temperature-scaled cross-entropy (NT-Xent) loss (Chen et al., 2020a; van den Oord et al., 2018). Therefore, for a normal example in a mini-batch $\{\boldsymbol{X}_i\}_{i=1}^B$, the loss objective of ConAAP can be formulated as:

$$\mathcal{L}_{ConAAP}(\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\substack{\{\boldsymbol{X}_i\}_{i=1}^B \sim \mathcal{D} \\ \boldsymbol{X}_i^{pos} \sim t^{pos}(\boldsymbol{X}_i) \\ \boldsymbol{X}_i^{neg} \sim t^{neg}(\boldsymbol{X}_i) \\ \boldsymbol{X}_i^{adv} \sim t^{adv}(\boldsymbol{X}_i)}} [-\log \frac{\mathcal{S}_{(i,i)}^{(\cdot,pos)} + \mathcal{S}_{(i,i)}^{(neg,adv)}}{\sum_{j=1}^B (\mathcal{S}_{negative})}] \tag{7}$$

where

$$\begin{aligned}\mathcal{S}_{negative} =& \mathcal{S}_{(i,i)}^{(\cdot,neg)} + \mathcal{S}_{(i,i)}^{(\cdot,adv)} \\ &+ \mathbb{1}_{[i \neq j]}[\mathcal{S}_{(i,j)}^{(\cdot,\cdot)} + \mathcal{S}_{(i,j)}^{(\cdot,neg)} + \mathcal{S}_{(i,j)}^{(\cdot,pos)}]\end{aligned}$$

and $\mathbb{1}_{[\cdot]} \in \{0, 1\}$ is an indicator function that equals 1 if $[\cdot]$ is true, B is the batch size.

To reduce the bias degree from the original example of the representation of out-of-distribution examples with *unimportant* words removed, we use the term $\mathcal{S}_{(i,i)}^{(\cdot,pos)}$ in the numerator. This constraint increases the similarity between the representation of the normal example and examples with unimportant words removed, implying that *model should regard the information in the removed unimportant words only slightly impacting the prediction.*

To increase the bias degree from the original example of the representation of out-of-distribution examples with *important* words removed, we use the term $\mathcal{S}_{(i,i)}^{(\cdot,neg)}$ in the denominator. This constraint decreases the similarity between the representation of normal example and examples with important words removed, implying that *model should regard the information in the removed important words significant in prediction.*

We simultaneously use the term $\mathcal{S}_{(i,i)}^{(neg,adv)}$ in the numerator to force the bias direction of out-of-distribution examples with *important* words removed toward the decision boundary indicated by the adversarial example. We also use the term $\mathcal{S}_{(i,i)}^{(\cdot,adv)}$ in the denominator to prevent the representation of normal example and adversarial example from collapsing together, ensuring that the adversarial example can always be utilized as a guide to locate the direction of decision boundary. It should be noted that ConAAP only focuses on alleviating the model pathology, and we leave improving the model robustness to future work.

The terms $\mathcal{S}_{(i,j)}^{(\cdot,\cdot)}+\mathcal{S}_{(i,j)}^{(\cdot,neg)}+\mathcal{S}_{(i,j)}^{(\cdot,pos)}$ in the denominator imply that the model should differentiate the various examples and their derived examples in a mini-batch, as the semantics of different examples should be different.

Finally, we use the $\mathcal{L}_{ConAAP}$ as regularization and combine it with the normal training method, which originally trains the model only with maximum likelihood. The overall objective can thus be formulated as follows:

$$\min_\theta \quad \mathcal{L}_{ce}(\boldsymbol{X}, Y) + \alpha\, \mathcal{L}_{ConAAP}(\boldsymbol{X}) \qquad (8)$$

where α is a parameter balancing the two parts.

4 Experiment

4.1 Metrics

We measure the model pathology and the faithfulness of attribution with metrics *AOPC*$_{Comp.}$ and *AOPC*$_{Suff.}$, and the parameter K in them is both set as the 40% of words for each sentence. We use *AOPC*$_{Diff.}$ to indicate the difference between *AOPC*$_{Comp.}$ and *AOPC*$_{Suff.}$, i.e., the difference between the overall influence of words of different importance on prediction. Based on the *Reduced Length* (Feng et al., 2018), we also use *IR*$\#$ and *UR*$\#$ to measure the influence of **I**mportant and **U**nimportant words on prediction, measuring the number of important and unimportant words removed until the prediction changes. We use $R_{Diff.}$ to indicate the difference between *IR*$\#$ and *UR*$\#$. Larger *AOPC*$_{Diff.}$ and $R_{Diff.}$ are expected for a non-pathological model and faithful attribution. We also use accuracy (*ACC.*) and confidence ($\mathcal{P}(Y|\boldsymbol{X})$) on normal examples to measure the generalization ability of model on in-distribution examples.

4.2 Experiment Setup

Dataset. Focusing on the text classification, our experiments are performed on AG News (Zhang et al., 2015), MR (Pang and Lee, 2005), and IMDB (Maas et al., 2011). More details of datasets are provided in Appendix A.1.

Model. Three models in different architectures are adopted. For TextCNN, we reuse the architecture in (Kim, 2014) while replacing the embedding with the 300-dimensional GloVe (Pennington et al., 2014). For LSTM (Hochreiter and Schmidhuber, 1997), we connect a Bi-LSTM layer with 150 hidden units with a dense layer based on the 300-dimensional GloVe layer. For BERT (Devlin et al., 2019), we use the base uncased version.

Baseline. To show the effectiveness of ConAAP and empirically demonstrate the analysis of the bias degree and bias direction we provide in §3.3 and §3.4, we use the following baselines: (i) *Normal*: using (1) as objective. (ii) *ConAAP*: combining (7) with *Normal* method, using (8) as objective. (iii) *Entropy*: maximizing the model uncertainty on the reduced examples (Feng et al., 2018). Please see Appendix A.3 for more details on *Entropy* method. (iv) *ConAAP w/o imp-dir*: removing $\mathcal{S}_{(i,i)}^{(neg,adv)}$, $\mathcal{S}_{(i,i)}^{(\cdot,adv)}$ in $\mathcal{L}_{ConAAP}$, indicating removing the calibration on the bias direction of out-of-distribution examples with *important* words removed. (v) *ConAAP w/o imp-deg-dir*: removing $\mathcal{S}_{(i,i)}^{(\cdot,neg)}$, $\mathcal{S}_{(i,i)}^{(neg,adv)}$ and $\mathcal{S}_{(i,i)}^{(\cdot,adv)}$ in $\mathcal{L}_{ConAAP}$, indicating removing the calibration on both the bias degree and direction of out-of-distribution examples with *important* words removed.

Implementation Details. The batch size is set as 64. For efficiency, we use a method called *CharDelete* to generate adversarial examples in t^{adv}, which randomly deletes characters in the important words until the attack success. More details of *CharDelete* are in Appendix A.2. We use Adam (Kingma and Ba, 2015) as the optimizer. Most setting of learning rate / α / τ for LSTM, TextCNN, and BERT is 5e-4/1.2/0.1, 5e-4/1.2/0.1, 3e-5/1.2/0.01. All reported results are the average of five independent runs.

4.3 Main Results

ConAAP marginally impacts the generalization performance for in-distribution examples. Table 1 illustrates the accuracy and confidence results for in-distribution examples. Utilizing ConAAP

	LSTM		TextCNN		BERT	
Method	*ACC.* ↑	$\mathcal{P}(Y\|\boldsymbol{X})$	*ACC.* ↑	$\mathcal{P}(Y\|\boldsymbol{X})$	*ACC.* ↑	$\mathcal{P}(Y\|\boldsymbol{X})$
			AG News			
Normal	**91.59**	0.93	91.45	0.94	**94.64**	0.99
Entropy	90.76	0.93	**91.51**	0.92	94.61	0.97
ConAAP	90.71	0.89	91.21	0.92	94.28	0.96
w/o imp-dir	90.32	0.92	91.17	0.92	94.13	0.95
w/o imp-deg-dir	90.31	0.91	91.43	0.93	94.39	0.96
			MR			
Normal	**79.34**	0.85	**79.02**	0.83	86.40	0.97
Entropy	78.05	0.86	78.83	0.84	86.39	0.99
ConAAP	78.23	0.82	78.24	0.88	87.05	0.97
w/o imp-dir	77.89	0.79	77.83	0.80	**87.32**	0.95
w/o imp-deg-dir	77.82	0.81	77.69	0.81	87.14	0.96
			IMDB			
Normal	**78.12**	0.84	77.53	0.82	84.08	0.99
Entropy	75.71	0.78	77.64	0.78	83.39	0.90
ConAAP	77.45	0.82	**77.68**	0.81	**84.09**	0.99
w/o imp-dir	77.46	0.83	77.33	0.82	83.08	0.95
w/o imp-deg-dir	77.52	0.82	77.21	0.82	83.40	0.96

Table 1: The comparisons of model accuracy and confidence on in-distribution normal sentences.

as regularization during training has a minimal impact on the model's behavior for in-distribution examples, as evidenced by the marginally changed model accuracy and confidence. The accuracy difference between ConAAP and Normal training methods is within 1.11%, and the model confidence on normal examples $\mathcal{P}(Y|\boldsymbol{X})$ decreases by at most 0.04 compared to Normal method. These results demonstrate that imposing regularization on the sentence representations of out-of-distribution examples only slightly compromises the model's generalization performance for in-distribution examples. Furthermore, bias degree and direction constraints in ConAAP also have only a minor impact on generalization capabilities.

ConAAP effectively alleviates model pathology. Table 2 illustrates the results on model pathology and attribution faithfulness. ConAAP consistently yields the largest values for $R_{Diff.}$ and $AOPC_{Diff.}$, indicating that the model considers the information in important words to have a more significant impact on predictions than that in unimportant words, and the attributions are more faithful. Moreover, when the calibration on the bias direction of out-of-distribution examples with important words removed (*w/o imp-dir*) is removed, both $R_{Diff.}$ and $AOPC_{Diff.}$ decrease, indicating less faithful attributions and reduced effectiveness in alleviating model pathology. Removing the calibration on both the bias degree and direction of out-of-distribution examples with important words removed (*w/o imp-deg-dir*) leads to further reductions in $R_{Diff.}$ and especially $AOPC_{Diff.}$ values, demonstrating the effectiveness of ConAAP's multi-view objective that simultaneously calibrates the bias degree and direction of the representations of various examples.

		IR#	UR#	$R_{Diff.}$ ↑	$A_{Comp.}$	$A_{Suff.}$	$A_{Diff.}$ ↑
			AG News				
LSTM	Normal	26.59	28.74	2.15	0.07	0.03	0.04
	Entropy	25.40	28.42	3.02	0.06	0.02	0.04
	ConAAP	22.32	27.74	**5.42**	0.18	0.05	**0.13**
	w/o imp-dir	22.97	27.50	4.53	0.18	0.09	0.09
	w/o imp-deg-dir	23.05	27.76	4.71	0.16	0.09	0.07
TextCNN	Normal	19.54	23.45	3.91	0.12	0.04	0.08
	Entropy	19.57	24.50	4.93	0.16	0.03	0.13
	ConAAP	18.68	24.59	**5.91**	0.22	0.04	**0.18**
	w/o imp-dir	18.77	23.88	5.11	0.22	0.07	0.15
	w/o imp-deg-dir	19.38	24.22	4.84	0.17	0.05	0.12
BERT	Normal	27.74	34.72	6.98	0.03	0.01	0.02
	Entropy	27.10	34.85	7.75	0.04	0.01	0.03
	ConAAP	24.09	34.78	**10.68**	0.18	0.01	**0.17**
	w/o imp-dir	27.04	34.68	7.64	0.12	0.02	0.10
	w/o imp-deg-dir	28.08	34.92	6.84	0.05	0.01	0.04
			MR				
LSTM	Normal	9.95	12.68	2.73	0.07	0.03	0.04
	Entropy	9.17	12.03	2.86	0.09	0.04	0.05
	ConAAP	8.94	12.46	**3.52**	0.13	0.03	**0.10**
	w/o imp-dir	9.61	12.76	3.15	0.10	0.02	0.08
	w/o imp-deg-dir	9.13	12.53	3.40	0.10	0.03	0.07
TextCNN	Normal	11.42	13.40	1.98	0.08	0.04	0.04
	Entropy	9.51	11.28	1.77	0.08	0.03	0.05
	ConAAP	6.61	9.90	**3.29**	0.19	0.06	**0.13**
	w/o imp-dir	7.09	9.88	2.79	0.19	0.08	0.11
	w/o imp-deg-dir	7.23	10.28	3.05	0.17	0.07	0.10
BERT	Normal	11.42	15.02	3.60	0.06	0.03	0.03
	Entropy	10.43	14.91	4.48	0.05	0.02	0.03
	ConAAP	10.31	15.08	**4.77**	0.20	0.02	**0.18**
	w/o imp-dir	10.97	15.12	4.15	0.10	0.02	0.08
	w/o imp-deg-dir	10.69	15.09	4.40	0.05	0.02	0.03
			IMDB				
LSTM	Normal	25.85	36.18	10.33	0.05	0.01	0.04
	Entropy	27.63	34.76	7.13	−0.04	0.01	−0.05
	ConAAP	20.33	37.38	**17.05**	0.15	0.05	**0.10**
	w/o imp-dir	20.59	37.03	16.44	0.12	0.05	0.07
	w/o imp-deg-dir	22.51	37.05	14.54	0.13	0.07	0.06
TextCNN	Normal	19.03	25.04	6.01	0.07	0.02	0.05
	Entropy	20.34	24.53	4.19	0.10	0.06	0.04
	ConAAP	16.02	26.69	**10.67**	0.14	0.03	**0.11**
	w/o imp-dir	18.10	26.39	8.29	0.13	0.04	0.09
	w/o imp-deg-dir	18.90	26.58	7.68	0.13	0.04	0.09
BERT	Normal	24.94	37.36	12.42	0.04	0.02	0.02
	Entropy	23.94	37.60	13.66	0.11	0.02	0.09
	ConAAP	22.49	36.34	**13.85**	0.16	0.02	**0.14**
	w/o imp-dir	23.78	36.86	13.08	0.13	0.02	0.11
	w/o imp-deg-dir	24.46	36.91	12.45	0.08	0.03	0.05

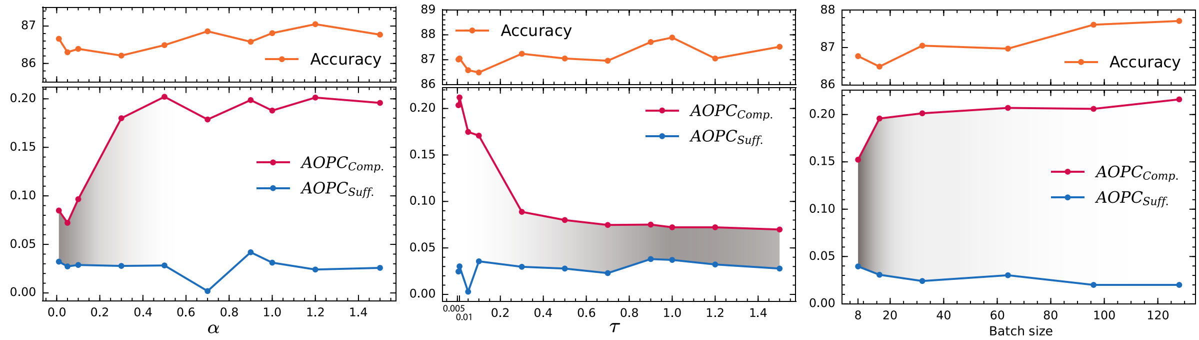
Table 2: The comparisons of the pathology of model and the faithfulness of attribution. *A* is short for *AOPC*.

4.4 Further Analysis and Ablation Study

In this section, we conduct further analysis and ablation study on BERT and MR dataset.

Hyperparameter α. Figure 3(a) illustrates the influence of α. We find that $AOPC_{Diff.}$ begins to increase when $\alpha > 0.05$ and stabilizes for $\alpha > 0.5$. The accuracy is stable and will slightly increase as α continues to increase.

Temperature τ. Figure 3(b) illustrates the influence of τ. We find that ConAAP is sensitive to τ, and an appropriate τ contributes to both model accuracy and the effectiveness in alleviating the pathology. $AOPC_{Diff.}$ reaches its peak when



(a) Hyperparameter α (b) Temperature τ (c) Batch size

Figure 3: Ablation study of hyperparameter α, τ, and batch size . The $AOPC_{Diff.}$ is shown in the distance between $AOPC_{Suff.}$ and $AOPC_{Comp.}$, and a darker color indicates a smaller value.

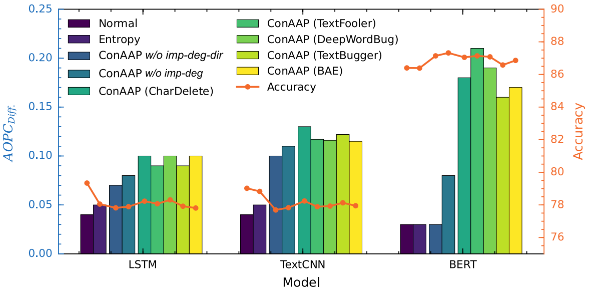


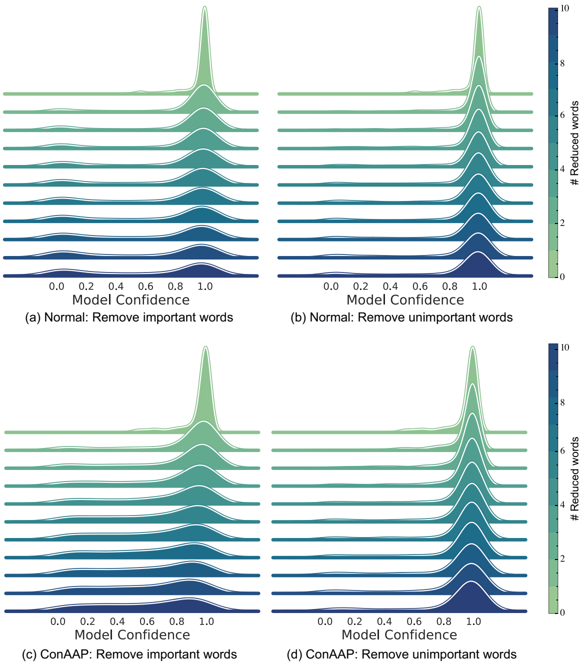
Figure 4: Influence of attack methods. Scatter plot indicate accuracy and bar plot indicate $AOPC_{Diff.}$.

$\tau = 0.01$, while the accuracy fluctuates by no more than 1.4% when τ is assigned different values.

Batch size. Figure 3(c) shows the influence of batch size on ConAAP. Larger batch sizes prove beneficial for ConAAP, as both model accuracy and the ability to alleviate pathology improve with an increase in batch size.

Attack method in t^{adv}. Various attack methods can be utilized in ConAAP (Gao et al., 2018; Garg and Ramakrishnan, 2020; Li et al., 2019; Jin et al., 2020), and the influence of attack methods is shown in Figure 4. ConAAP remains effective in alleviating model pathology when utilizing different attack methods. It should be noted that adversarial examples in ConAAP are used to introduce direction information and are not intended to be nearly imperceptible to humans. Consequently, their quality is not of primary concern, and a fast *CharDelete* method suffices for our purposes.

Confidence Changing with Word Removal. Figure 5 illustrates the impact of word removal on model confidence. As more important words are removed, the confidence of the Normal method remains close to 1, while the label shift induced by word removal causes the model's confidence in the original class to approach 0 (Figure 5(a)). In contrast, the distribution of ConAAP is considerably smoother than the Normal method (Figure 5(c)). When more unimportant words are removed, the confidence for both ConAAP and the Normal method consistently concentrates in a high region (Figure 5(b)(d)).



(a) Normal: Remove important words (b) Normal: Remove unimportant words (c) ConAAP: Remove important words (d) ConAAP: Remove unimportant words

Figure 5: The comparisons of model confidence (range from 0 to 1) density distribution. The results are obtained on the entire MR testing set.

Case study. The case study is shown in Figure 6. For the model trained with the Normal method, various interpretation methods show considerable divergence in word importance. Moreover, the model predicts the sentence with high confidence even after removing the two most important words (e.g.,

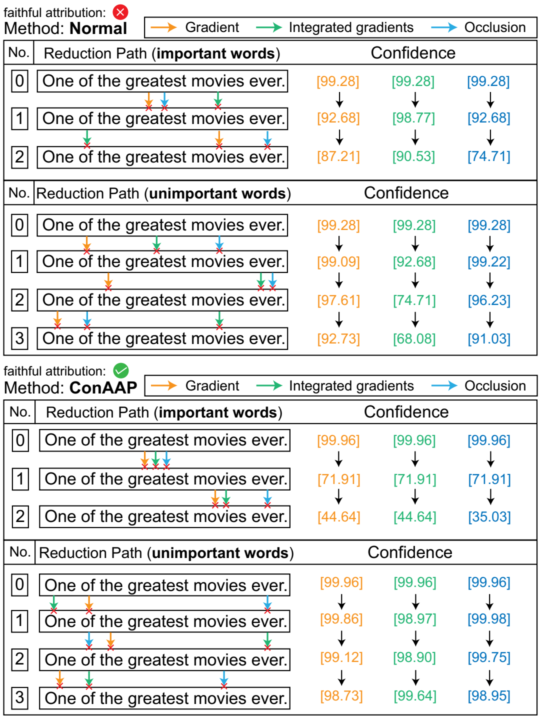

faithful attribution: ✗
Method: Normal
Gradient
Integrated gradients
Occlusion
No.
Reduction Path (important words)
Confidence
0 One of the greatest movies ever. [99.28] [99.28] [99.28]
1 One of the greatest movies ever. [92.68] [98.77] [92.68]
2 One of the greatest movies ever. [87.21] [90.53] [74.71]
No.
Reduction Path (unimportant words)
Confidence
0 One of the greatest movies ever. [99.28] [99.28] [99.28]
1 One of the greatest movies ever. [99.09] [92.68] [99.22]
2 One of the greatest movies ever. [97.61] [74.71] [96.23]
3 One of the greatest movies ever. [92.73] [68.08] [91.03]
faithful attribution: ✓
Method: ConAAP
Gradient
Integrated gradients
Occlusion
No.
Reduction Path (important words)
Confidence
0 One of the greatest movies ever. [99.96] [99.96] [99.96]
1 One of the greatest movies ever. [71.91] [71.91] [71.91]
2 One of the greatest movies ever. [44.64] [44.64] [35.03]
No.
Reduction Path (unimportant words)
Confidence
0 One of the greatest movies ever. [99.96] [99.96] [99.96]
1 One of the greatest movies ever. [99.86] [98.97] [99.98]
2 One of the greatest movies ever. [99.12] [98.90] [99.75]
3 One of the greatest movies ever. [98.73] [99.64] [98.95]

Figure 6: Case study on the instance sentence *"One of the greatest movies ever"*. The *Reduction Path* indicates how the words in the sentence are **cumulatively** removed, *No.* indicate the number of removed words of current-step sentence. The arrows of different colors indicate the most or the least important words in the current-step sentence attributed by different interpretation methods. The confidence values of different colors indicate the model confidence (range from 0 to 100) in *positive* following the reduction path of different attribution methods.

following Gradient attribution, the model predicts the sentence *"One of the ~~greatest~~ ~~movies~~ ever"* as positive with 87.21% confidence). In contrast, for the model trained with ConAAP, different interpretation methods show a more consistent result of word importance (e.g., important words are concentrated in *greatest*, *movie*; unimportant words are concentrated in *one*, *of*, *the*, *ever*), resulting in more faithful attributions. Specifically, when the two most important words are removed, the average confidence across different attributions is 41.43%. Conversely, when unimportant words are removed, the model can still make high-confidence predictions similar to the original examples.

5 Conclusion

In this paper, we argue that the failure of interpretation methods to provide faithful attributions for language models is due to the model pathology that models are overconfident in out-of-distribution examples when making predictions. We explain the model pathology from the perspective of sentence representation and propose ConAAP, a contrastive learning regularization method to calibrate the sentence representation of out-of-distribution examples. Experiments demonstrate the effectiveness of ConAAP in alleviating model pathology, which helps interpretation methods obtain faithful results. We hope that our work will provide a new perspective on research in the field of interpretability.

Limitations

We explain model pathology from a classification perspective, but the pathological nature may exist in language models for performing various tasks, such as reading comprehension, textual entailment, and visual question answering. Although our proposed regularization technique may be applicable to various tasks, we have only investigated its effectiveness in classification problems. Further evaluations are expected to be conducted in future works. The proposed method also leads to more time-consuming training, primarily due to the generation of adversarial examples, while only a minimal amount of time is spent on generating out-of-distribution examples.

Ethics Statement

This paper investigates model pathology from a sentence representation perspective and proposes a regularization technique to alleviate the pathology. It is possible that the proposed method can be used for both benign purposes, such as fixing the potential flaws and biases of models, and malign ones, such as exposing the vulnerabilities of models, which makes it easier for adversaries to generate malicious input. Despite these risks, we argue that studying model pathology openly is essential. Exploring the pathological nature of models will help us effectively control these potential risks and improve our understanding of the mechanics of natural language models. All datasets used in this paper are publicly accessible, and our work fully complies with their respective licenses.

Acknowledgements

The authors would like to thank the anonymous reviewers for their thorough and constructive feedback. This research was supported by National Research and Development Program of China (No.2019YFB1005200).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

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*.

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*.

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

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17.

Chih-Hui Ho and Nuno Vasconcelos. 2020. Contrastive learning with adversarial examples. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

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

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial self-supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. 2021. Self-supervised contrastive learning with adversarial perturbations for robust pretrained language models. *ArXiv preprint*, abs/2107.07610.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*.

Marco Túlio 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*.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11).

Patrick Schwab and Walter Karlen. 2019. *CXPlain: Causal Explanations for Model Interpretation under Uncertainty*.

Xuelin Situ, Ingrid Zukerman, Cécile Paris, Sameen Maruf, and Gholamreza Haffari. 2021. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint*, abs/1807.03748.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *ArXiv preprint*, abs/2012.15466.

Pengwei Zhan, Yang Wu, Shaolei Zhou, Yunjian Zhang, and Liming Wang. 2022a. Mitigating the inconsistency between word saliency and model confidence with pathological contrastive training. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Pengwei Zhan, Chao Zheng, Jing Yang, Yuxiang Wang, Liming Wang, Yang Wu, and Yunjian Zhang. 2022b. PARSE: an efficient search method for black-box adversarial text attacks. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 4776–4787.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*.

A Additional Experimental Details

A.1 Details on Dataset

AG News contains news articles in the areas of World, Sport, Business, and Science/Technology, with 120,000 for training and 7,600 for testing. MR contains movie reviews from Rotten Tomatoes labeled as positive or negative, with 8,530 for training and 1,066 for testing. IMDB contains binary polar movie reviews from the Internet Movie Database, with 25,000 for training and 25,000 for testing.

A.2 Details on CharDelete Attack Method

We use the CharDelete adversarial attack method in t^{adv} to generate adversarial examples in our main experiments. The details of CharDelete are shown in Algorithm 1. ConAAP does not tend to generate high-quality adversarial examples that are imperceptible to humans and only utilizes adversarial examples to introduce direction information into regularization. This attack method meets our requirements, and a complex method is unnecessary.

Algorithm 1: CharDelete Algorithm

```
input  : Original sentence $\boldsymbol{X} = x_1 x_2 \dots x_N$,
         model $f_{\boldsymbol{\theta}}$, true label $Y_{true}$
output : Adversarial example $\boldsymbol{X}^{adv}$
obtain the attribution of all input words $Attr(\boldsymbol{X})$ by
  gradient-based attribution method in (3)
obtain the importance rankings (indexes) of input
  words $R(\boldsymbol{X}) \leftarrow \arg\operatorname{sort}_i Attr(\boldsymbol{x_i})_{i\in\{1,2,\cdots,N\}}$
$\boldsymbol{X'} \leftarrow \boldsymbol{X}$
for $r_i \in R(\boldsymbol{X})$ do
    $\boldsymbol{X'} \leftarrow$ randomly remove the letter in the $r_i$-th
      word of sentence $\boldsymbol{X'}$
    if $\underset{Y\in\mathcal{Y}}{\operatorname{argmax}}\ \mathcal{P}(Y|\boldsymbol{X'}) \neq Y_{true}$ then
        return $\boldsymbol{X}'$ as $\boldsymbol{X}^{adv}$;      /* Success */
return $\boldsymbol{X}$;                                           /* Fail */
```

A.3 Details on Entropy Method

The Entropy training method is proposed by (Feng et al., 2018). They use an entropy of the output distribution as a regularization term in the overall training objective. Specifically, the loss objective of the Entropy method is

$$
\begin{aligned}
\mathcal{L}_{entropy} = & \sum_{(\boldsymbol{X},Y)\in(\mathcal{X},\mathcal{Y})} \log(P(Y|\boldsymbol{X})) \\
& + \lambda \sum_{i\in\{1,\cdots,b\}} \mathbb{H}\left(P\left(Y \mid t^{ump}_{/min}(\boldsymbol{X}_i)\right)\right)
\end{aligned} \tag{9}
$$

where λ is a parameter balancing the two terms, $t^{ump}_{/min}$ generates the sentences with multiple unimportant words reduced to the minimum length that can keep the model predictions by beam search, b is the beam width, $\mathbb{H}$ denotes the entropy. λ is set as 1e-3, in accordance with the original paper.

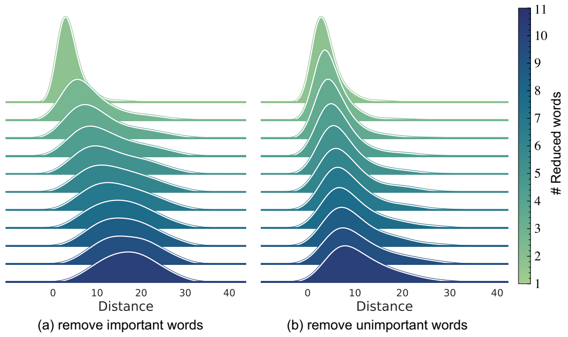
B Additional Experimental Results

B.1 Distance Between Different Examples

We also provide the aggregated results on the distance between out-of-distribution sentences and the in-distribution normal sentence in Figure 7-9. After removing important words, the density distribution of Euclidean distance between such sentence representations and the original sentences becomes smoother, with an increase in the maximum distance. However, most sentence representations remain close to the original ones (with Euclidean distance approaching 0). Intuitively, although the density distribution becomes smoother after important word removal, there is no significant horizontal shift (i.e., minimal distance changes), indicating that information from some important words does not have a sufficient impact on predictions. After removing unimportant words, the change in the density distribution of Euclidean distance between such sentence representations and the original sentence is less pronounced than when important words are removed. However, the representations of some sentences diverge considerably from the originals when only a few unimportant words are removed (e.g., distance greater than 10 in MR when only one unimportant word is removed), indicating that information from some unimportant words may have a much greater influence on predictions than expected.

B.2 Sentence Representation Distribution

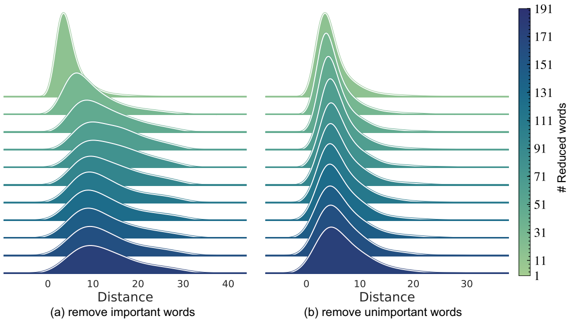
Figure 10-11 show more visualization of the sentence representation and the attribution on instance sentences. Observation 1 and Observation 2 can also be observed in these examples.



(a) remove important words

(b) remove unimportant words

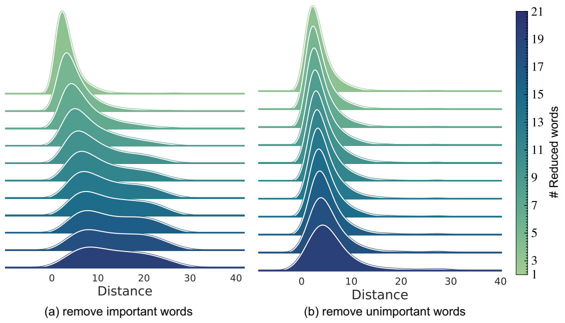
Figure 7: The density distribution of Euclidean distance between the representations of out-of-distribution sentences and in-distribution normal sentences. The results are obtained on the MR test set, with BERT fine-tuned on the MR training set.



(a) remove important words

(b) remove unimportant words

Figure 8: The density distribution of Euclidean distance between the representations of out-of-distribution sentences and in-distribution normal sentences. The results are obtained on the IMDB test set, with BERT fine-tuned on the IMDB training set.



(a) remove important words

(b) remove unimportant words

Figure 9: The density distribution of Euclidean distance between the representations of out-of-distribution sentences and in-distribution normal sentences. The results are obtained on the AG News test set, with BERT fine-tuned on the AG News training set.

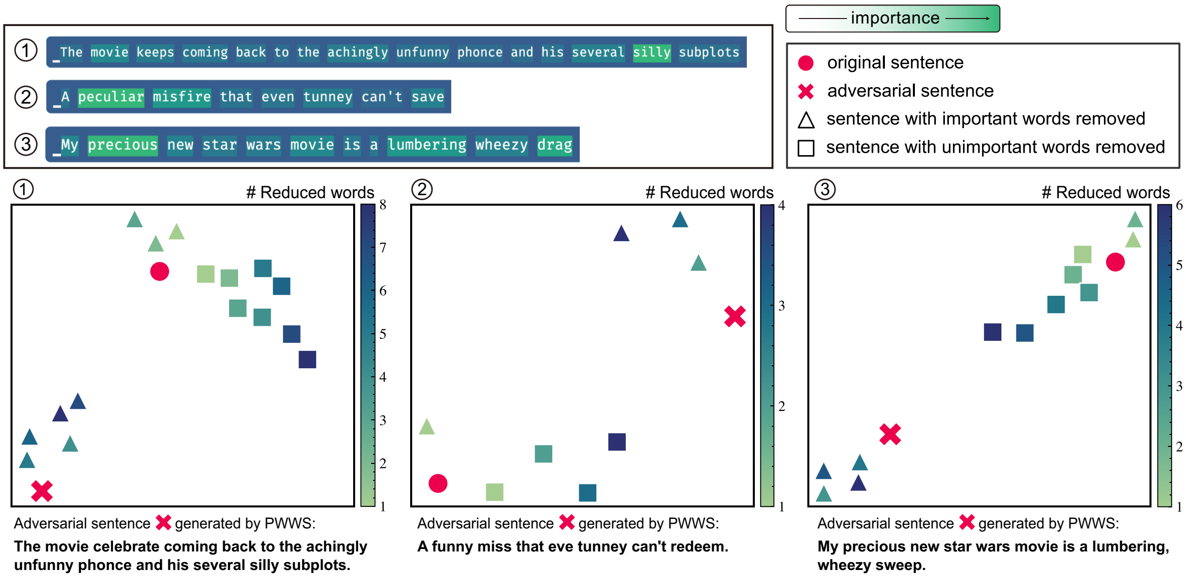


Figure 10: Additional visualization of sentence representations and the attribution obtained by gradient-based interpretation on MR instances. For the representation visualization, darker △ and □ indicate out-of-distribution examples with more words removed. The out-of-distribution examples closer to the original example are more likely to maintain the same model prediction as the original example, while the examples closer to the adversarial example tend to decrease the confidence in original class as the adversarial example leads to different predictions and is located in the vicinity of the decision boundary. For the attribution, darker colors indicate higher importance.

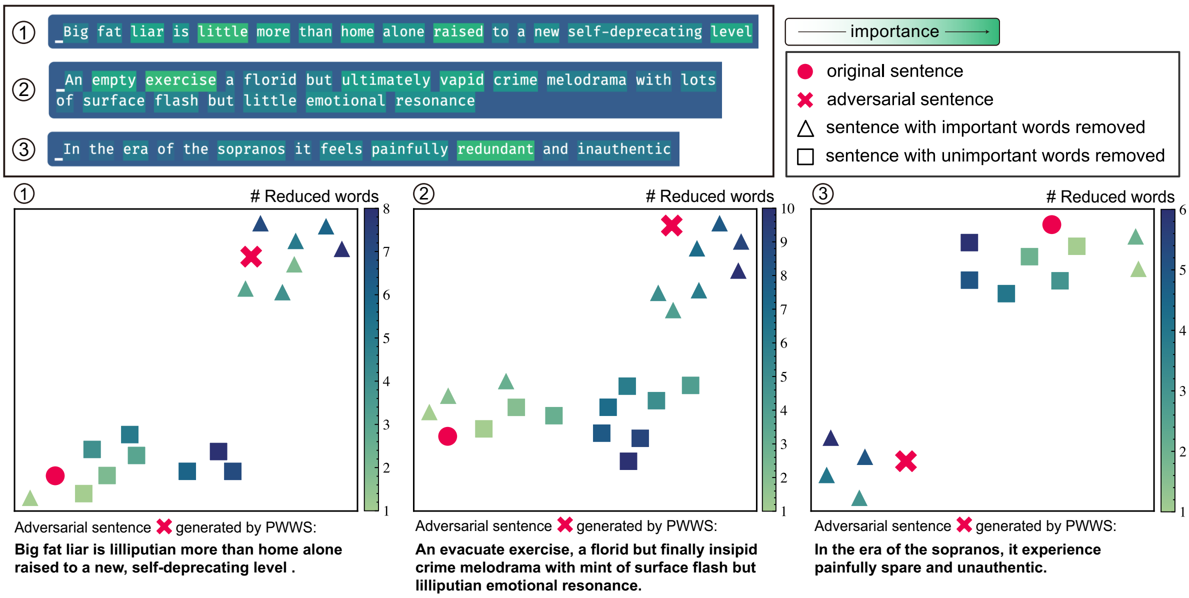


Figure 11: Additional visualization of sentence representations and the attribution obtained by gradient-based interpretation on MR instances. For the representation visualization, darker △ and □ indicate out-of-distribution examples with more words removed. The out-of-distribution examples closer to the original example are more likely to maintain the same model prediction as the original example, while the examples closer to the adversarial example tend to decrease the confidence in original class as the adversarial example leads to different predictions and is located in the vicinity of the decision boundary. For the attribution, darker colors indicate higher importance.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Section Limitations.

☑ A2. Did you discuss any potential risks of your work?
In Section Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In Section 3 and Section 4.

☑ B1. Did you cite the creators of artifacts you used?
In Section 1 and Section 4.2.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section Ethics Statement.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Section Ethics Statement.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All datasets utilized by us are widely adopted benchmark datasets.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 4.1, Section 4.2, and Appendix A.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Appendix A.1.

C ☑ Did you run computational experiments?

In Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 4.2.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.1 and Section 4.2.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 4.2.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section 4.2.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.