Mitigating the Inconsistency Between Word Saliency and Model Confidence with Pathological Contrastive Training

Pengwei Zhan[§‡], **Yang Wu**[§‡], **Shaolei Zhou**[§‡], **Yunjian Zhang**[§‡], **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,wuyang0419}@iie.ac.cn
{zhoushaolei,zhangyunjian,wangliming}@iie.ac.cn

Abstract

Neural networks are widely used in various NLP tasks for their remarkable performance. However, the complexity makes them difficult to interpret, i.e., they are not guaranteed right for the right reason. Besides the complexity, we reveal that the model pathology - the inconsistency between word saliency and model confidence, further hurts the interpretability. We show that the pathological inconsistency is caused by the representation collapse issue, which means that the representation of the sentences with tokens in different saliency reduced is somehow collapsed, and thus the important words cannot be distinguished from unimportant words in terms of model confidence changing. In this paper, to mitigate the pathology and obtain more interpretable models, we propose **P**athological **C**ontrastive **T**raining (PCT) framework, which adopts contrastive learning and saliency-based samples augmentation to calibrate the sentences representation. Combined with qualitative analysis, we also conduct extensive quantitative experiments and measure the interpretability with eight reasonable metrics. Experiments show that our method can mitigate the model pathology and generate more interpretable models while keeping the model performance. Ablation study also shows the effectiveness.

1 Introduction

Neural networks have achieved remarkable success in various NLP tasks, while the extremely high complexity of such models makes them difficult to interpret. Complex models may learn significantly different attributions with similar accuracy during training as datasets are often full of ambiguities (Ross et al., 2017). If a model is deployed without ensuring that it is right for the right reason, it may completely fail to make reliable predictions on new data, which is very dangerous. For example, some models will counter-intuitively consider prepositions to have extremely high saliency in rumor detection tasks. Interpretable models can ensure that the attribution of model prediction is consistent with human intuition, allowing the model to be trusted in critical applications.

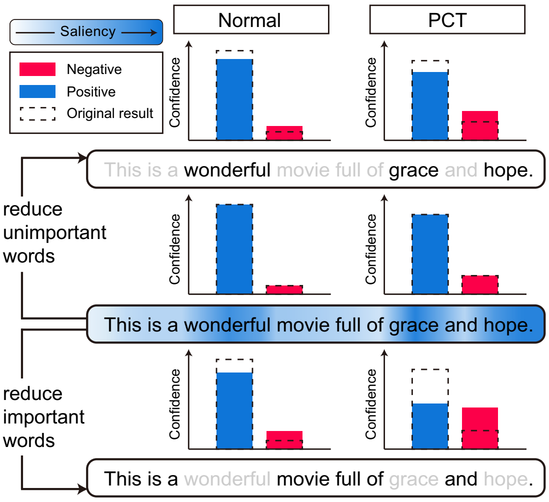


Figure 1: Word saliency and model confidence on case sentence. Normal model can not distinguish well between the influence of important and unimportant words, and the confidence on *Positive* class always focuses on a high region. Our method mitigates the pathology.

In addition to the complexity, the pathology also makes models more difficult to interpret (Feng et al., 2018). Neural networks are more linear than expected, leading models to overfit the negative log-likelihood loss to output low-entropy distributions over classes, and thus models will be overconfident on examples outside the training data distribution (Goodfellow et al., 2015). This consequently leads to the models giving counter-intuitive high confidence predictions on meaningless rubbish examples, and the word saliency will drastically change with even unimportant words reduced.

The model pathology indicates that words with low saliency actually have a more significant impact on prediction than expected. We further extend the pathology to a more general definition: the

[*]Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2226 - 2244
May 22-27, 2022 ©2022 Association for Computational Linguistics

Saliency and Confidence is inconsistent. Specifically, we show that the important words (with high saliency) are actually not so important to the model prediction and the unimportant words (with low saliency) are actually not so unimportant in normal models, as the representation of the text with tokens in different saliency reduced are somehow collapsed. The model prediction confidence will only slightly change when words are reduced, and the important words cannot be distinguished from unimportant words in terms of model confidence changing. Traditional methods usually train models with additional supervision, i.e., annotation on rationales, to force models better distinguish the influence between important words and unimportant words. However, human annotation is costly and often unavailable.

In this paper, to mitigate the pathology, i.e., the inconsistency between saliency and confidence, and train a more interpretable model while avoiding the dependence on extra labeled data, we propose a model-agnostic training method called **P**athological **C**ontrastive **T**raining (PCT). Inspired by contrastive learning, we encourage the original text to be closer to the text with unimportant words reduced while keeping away from the text with important words reduced. Our method can generate more interpretable models while keeping model performance. An example of model pathology and the effectiveness of our method is shown in Figure 1. The major contributions of this paper are summarized as follows:

1. We reveal the model representation collapse issue and the model pathology: the inconsistency between Saliency and Confidence.

2. We propose PCT that can mitigate the pathology by contrastive learning with saliency-based samples augmentation.

3. Extensive experiments show that our method can generate more interpretable models, while keeping the performance.

2 Related Work

Training interpretable model. A common method to obtain interpretable model is to let the model learn from the human-labeled rationales (Zhang et al., 2016; Ross et al., 2017; Rajani et al., 2019; Strout et al., 2019). However, the labeled data is costly. Other works try to assign interpretable properties to model through unsupervised regularization. Feng et al. (2018) train model with an objective containing an entropy regularization term to mitigate the model pathology that the confidence remained almost constant and sometimes increased when unimportant words are reduced.

Evaluating rationales. Lack of unified metrics for the interpretability of NLP models, many previous works measure the quality of the prediction rationales directly by human study, e.g., by visualizing the attribution through a saliency heatmap (Li et al., 2016; Sundararajan et al., 2017) and asking humans to give the quality of rationales provided by the model (Strout et al., 2019; Nguyen, 2018). To reduce the human work in the rationales evaluating, DeYoung et al. (2020) propose automatic metrics including *Comprehensiveness* and *Sufficiency*. Feng et al. (2018) utilized *Reduced Length* to measure the pathology of the model.

Contrastive learning. Contrastive learning is first applied to unsupervised computer vision tasks (Hadsell et al., 2006; Zhuang et al., 2019; Chen et al., 2020b), while the discrete nature of the text makes methods designed for continuous images fail to construct textual contrastive pairs. Previous works propose various textual data augmentation methods for construing textual contrasts, e.g., by generating overlapping or contained spans (Giorgi et al., 2021), by randomly performing word deletion, span deletion, reordering, and synonym substitution (Wu et al., 2020), by using back-translation (Fang et al., 2020), and by performing adversarial attacks, shuffling, cutoff and dropout on the embedding (Yan et al., 2021).

3 Model Pathology Analysis

3.1 Common Notation

Let $X = (x_1, \dots, x_N)$ denotes an input sentence with N words. To define text classification task, let $\mathcal{Y} = \{y_j | j \in [1, T]\}$ be the set with T possible class labels, i.e., the output space, let $\mathcal{X} = \{X_j | j \in [1, D]\}$ be the input space, and D represent the size of training dataset, thus $\{(X_j, Y_j) \, | j \in [1, D]\}$ is the training dataset, noted that Y_j is the label of j-th input sentence X_j. A target model is defined as $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, which maps input feature space to output space.

3.2 Gradient-based Attribution

Gradient-based attribution is a kind of faithful post-hoc explanation method (Smilkov et al., 2017; Sundararajan et al., 2017; Ross et al., 2017) that can measure the word saliency in the input without changing the original model. This method was first proposed in computer vision task, and it assumes that the model is fully differentiable (Papernot et al., 2016). However, because of the discrete nature of text, these methods instead calculate the word saliency on the embedding, rather than input text, in language models. Formally, generating word saliency with gradient-based method has the following steps. First compute the forward derivative:

$$\nabla\mathcal{F}(X) = \frac{\partial\mathcal{F}(X)}{\partial e(X)} = \left[\frac{\partial\mathcal{F}_j(X)}{\partial e(x_i)}\right]_{i\in 1..N, j\in 1..T} \qquad (1)$$

and the saliency of word x_i is defined as:

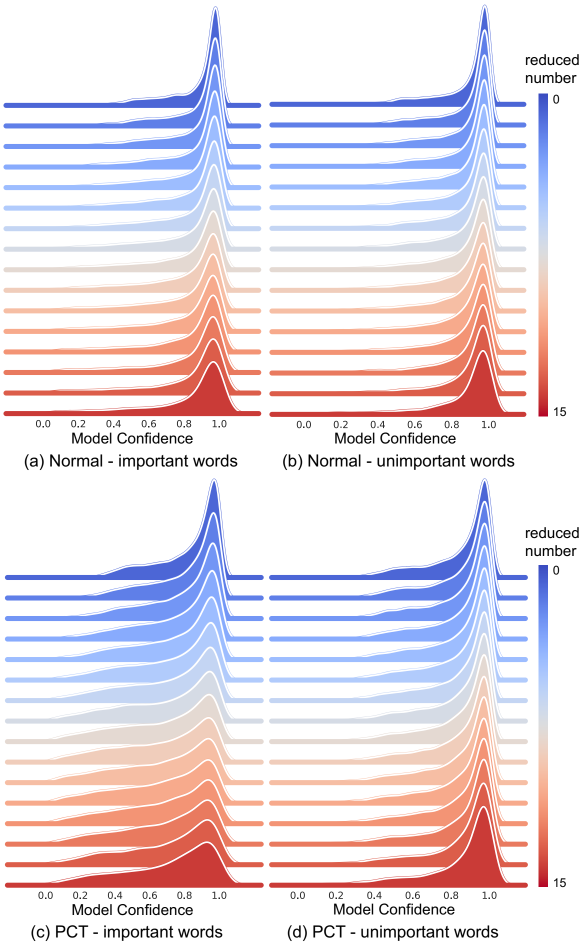
$$S(x_i) = \frac{\partial\mathcal{F}_{true}(X)}{\partial e(x_i)} \qquad (2)$$

where, $\mathcal{F}_j(\cdot)$ is the output w.r.t. class j, $true$ means the ground truth class, $e(\cdot)$ denotes the embedding.

3.3 Inconsistency Between Confidence and Saliency Damaging Interpretability

To demonstrate the inconsistency between saliency and confidence, we trained a Bi-LSTM model that consists of a 300-dimensional embedding layer, and a Bi-directional LSTM layer composed of 150 units, in a normal manner using cross-entropy loss on the AG News dataset (Zhang et al., 2015). Then we calculate the word saliency of all text on the test set with the gradient-based method, then generate two sentence sets from the original text by (i) cumulatively reducing high saliency words (important words) and (ii) cumulatively reducing low saliency words (unimportant words). We use the model to predict the two sentence sets containing text with words in different saliency are reduced. Figure 2(a)(b) shows the confidence density distributions of the normally trained model on reduced inputs. To give a better understanding of the pathology and demonstrate the effectiveness of our method at mitigating the pathology, we also provide the results of the model trained with PCT (Figure 2(c)(d)). See Figure 13-18 in the Appendix for more comparisons on confidence distribution.

After removing the important and unimportant words, the confidence distributions of the normal model are extremely similar, both concentrate in an extremely high region (0.8-1.0) that are similar to the results on original text (first line in Figure 2(a)(b)). With the increase of reduced number, the distribution of confidence after removing important words is only slightly smoother than after removing unimportant words. Even after removing 15 important words (the last line in Figure 2(a), the confidence is still concentrate above 0.8. It indicates that the influence of words with high saliency on the prediction confidence is too close to the words with low saliency, which is not distinguishable, and the model is not interpretable. While for the model trained with PCT, the confidence change tendency of as different types of tokens are reduced is much distinguishable (Figure 2(c)(d)). The words with high saliency have a greater impact on the prediction confidence, reducing which the confidence will relatively decrease, and the distribution becomes much smoother. Meanwhile, the confidence distribution only slightly changes when unimportant words are reduced, proving that our method can provide asymmetric regularization and can mitigate the pathology of inconsistency.



(a) Normal - important words

(b) Normal - unimportant words

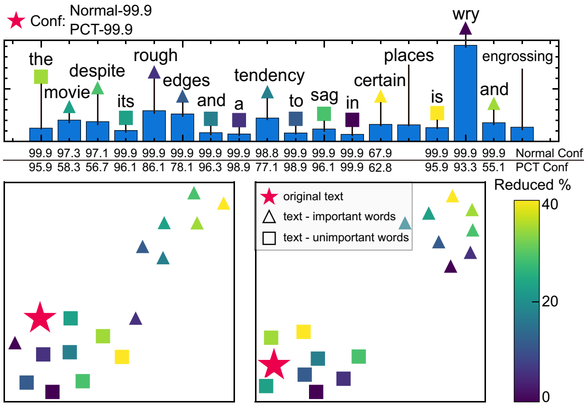
(c) PCT - important words

(d) PCT - unimportant words

Figure 2: Confidence density distribution of LSTM trained with normal and PCT methods on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.

3.4 Representation Collapse Deteriorate the Pathology of Inconsistency

To show that the inconsistency is somehow caused by the representation collapse issue, we fine-tune a BERT (Devlin et al., 2019) with normal method and PCT, respectively. Figure 3 shows the t-SNE (van der Maaten and Hinton, 2008) visualization, word saliency, and confidence on the sentence representation of a normal sample and the reduced samples from IMDB (Maas et al., 2011) dataset.



(a) Normal (b) PCT

Figure 3: Illustration of representation collapse issue. Cumulatively reducing words in Positive instance *the movie, despite its rough edges and a tendency to sag in certain places, is wry and engrossing*. Bar plot indicates word saliency obtained with gradient method, scatter plot indicates the t-SNE visualization of sentence representation. *Conf* is short for confidence.

For the normally tuned BERT, the sentence representation of text with important words *wry* and *rough* deleted are collapse with original text and the text with unimportant words reduced (e.g., *in,a,to*), even the saliency of word *wry* is leading other words. When important words are reduced, the confidence hardly decreases. While for the model tuned with PCT, the sentence representation of text with different types of words reduced are better separated, and the confidence intuitively decreased, which has a better interpretability. See Figure 5-12 and Table 6-9 in the Appendix for more illustrations on representation collapse issue.

3.5 Quantitative Analysis of Interpretability

In the previous section, we qualitatively analyzed the inconsistency between saliency and confidence, but we also need to quantify the extent of this inconsistency to better evaluate the pathology and interpretability of the model. How to quantify the pathology and interpretability is an open question. Besides the accuracy, we used seven extra metrics to measure the pathology. The following gives our analysis on interpretable model and these metrics:

Confidence on normal text ($\mathcal{F}(X)$). This metric measures how confident the model is in making predictions on normal sentences. The words with high saliency in the original text should have enough impact on confidence, and the confidence value should be at a high level.

Comprehensiveness (*Comp*) (DeYoung et al., 2020). This metric measures the influence of **important** words on confidence, i.e., the change in confidence after the removal of important words:

$$Comp = \mathcal{F}(X) - \mathcal{F}(\hat{X}^{imp}) \tag{3}$$

where $\hat{X}^{imp}$ is the text with important words reduced. A higher *Comp* value indicates that important words are influential in the prediction, and thus the model has better interpretability. If *Comp* value is low, or even negative, the saliency and confidence is inconsistent, rationales cannot be used to explain the model, and the model is not interpretable.

Sufficiency (*Suff*) (DeYoung et al., 2020). This metric measures the influence of **unimportant** words on confidence, i.e., the change in confidence after the removal of unimportant words:

$$Suff = \mathcal{F}(X) - \mathcal{F}(\hat{X}^{ump}) \tag{4}$$

where $\hat{X}^{ump}$ is the text with unimportant words reduced. The influence of unimportant words on confidence should be slight. However, these unimportant words also provide information about the context, thus we take it reasonable when $Suff \in (0, Comp)$. And a larger gap between *Suff* and *Comp* indicates a more interpretable model.

Reduced number (Feng et al., 2018). The number of important (***IR#***) / unimportant (***UR#***) words deleted until the label is changed. A smaller *IR#* indicates that important words have a greater impact on the prediction. A higher *UR#* indicates that unimportant words have a smaller impact on the prediction. Thus, we consider it reasonable when $IR\# < UR\#$. An small or even negative value of $UR\# - IR\#$ indicates the pathology of the model.

Saliency variance. We propose this as the variance of saliency rank after removing important (***I-Var***) / unimportant (***U-Var***) words. These metrics

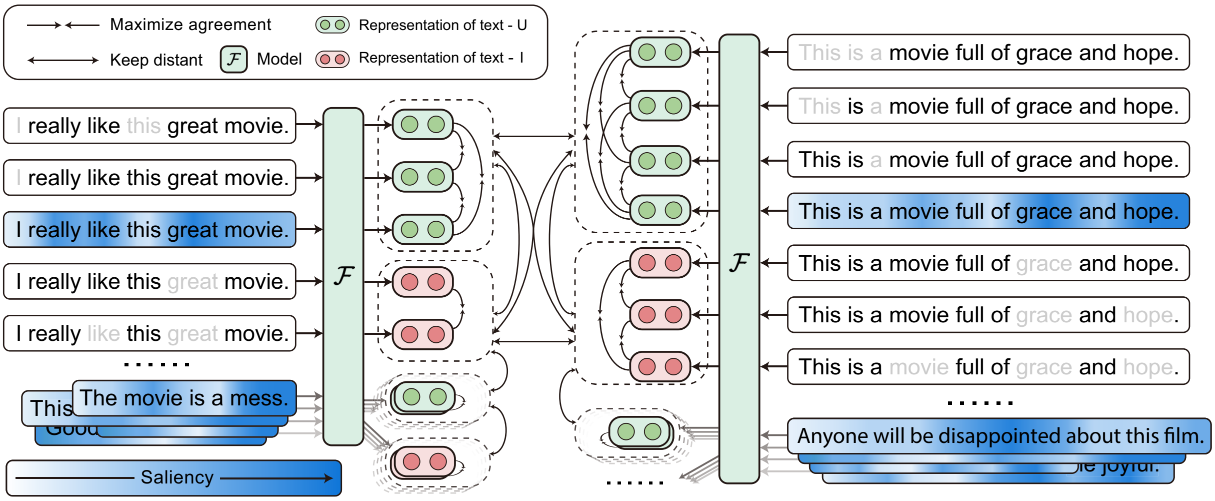

Maximize agreement
Keep distant
$\mathcal{F}$ Model
Representation of text - U
Representation of text - I
I really like this great movie.
I really like this great movie.
I really like this great movie.
I really like this great movie.
I really like this great movie.
......
This
The movie is a mess.
Good
Saliency
$\mathcal{F}$
$\mathcal{F}$
This is a movie full of grace and hope.
This is a movie full of grace and hope.
This is a movie full of grace and hope.
This is a movie full of grace and hope.
This is a movie full of grace and hope.
This is a movie full of grace and hope.
This is a movie full of grace and hope.
......
Anyone will be disappointed about this film.
le joyful.
......

Figure 4: General Framework of PCT. For each sentence in a mini-batch, we compute the word saliency through the gradient-based method, then augment the normal sentence to two sets: text with **I**mportant / **U**nimportant words cumulatively reduced. The sentence representations in all sets are encoded by target model $\mathcal{F}$, and the sentences with **important / unimportant** words reduced are takes as *negative* / *positive* pairs for the original sentence.

measure the influence of a word on the saliency of the other words. Formally:

$$Var = \frac{1}{N-1}\sum_{i=1}^{N-1}\left(d_i - d_i'\right)^2 \qquad (5)$$

where d_i is the index of i-th important word on the original text, and d_i' is the index of i-th important word on the text with one word reduced. The unimportant words should have less impact on both the final confidence and the word saliency of other words, while the impact of important words on the saliency should be greater than unimportant words, so we consider it reasonable when *I-Var* > *U-Var*.

4 Pathological Contrastive Training

According to the above analysis, the representation collapse characteristic of neural networks causes the influence of high saliency words and low saliency words on prediction to be indistinguishable. To mitigate this issue, we propose PCT that utilizes saliency-based samples augmentation for contrasting learning. The key idea of our method can be summarized as: the original normal text are encouraged to be closer to the derived text with unimportant words reduced while keeping away from the derived text with important words reduced. As shown in Figure 4, the framework comprises the following three major components:

Data augmentation module. We limit the contrast scope to within a mini-batch rather than the entire training set, as the latter is extremely computationally expensive. The data augmentation module will generate positive and negative samples in a self-supervised manner before the new mini-batch is sent to the model. Suppose there are K normal examples in a mini-batch, for each sample in the batch, we first use gradient-based attribution method to obtain the saliency of the normal input sentence $S(X_i)$, and define m words with the highest saliency and lowest saliency as important words and unimportant words, respectively. We then cumulatively reduce the important words in a descending order of saliency value to generate a text set containing text with multiply important words are reduced $\{\hat{X}_i^{imp}\}_{i=1}^{m}$. Parallelly, we generate the text set $\{\hat{X}_i^{ump}\}_{i=1}^{m}$ by cumulatively reducing unimportant words. Adding the original text X_i to $\{\hat{X}_i^{ump}\}_{i=1}^{m}$, we have the positive set $\mathbf{X}_i^+ = X_i \cup \{\hat{X}_i^{ump}\}_{i=1}^{m}$ derived from X_i, with no ambiguity, we denote $\{\hat{X}_i^{imp}\}_{i=1}^{m}$ as $\mathbf{X}_i^-$, the negative set derived from X_i. There are $2K$ text set after processed by data augmentation module. For each text set, there are at most m sentences if not considering X_i. Thus, a sentence has at most m positive pairs and $(2K-1)m$ negative pairs.

Target model $\mathcal{F}$. Model is utilized as an encoder that extracts representations for both the original text and the augmented text. Our method does not impose restrictions on the type of model. Specifically, for BERT, we use the representation of [CLS] token at the last hidden layer as sentence representations. For other models (e.g., CNN, LSTM), we use the average pooling of the token embedding at the layer before the last dense layer as sentence representation.

Model-agnostic contrastive loss objective. This loss objective controls the representation distances of the samples in a mini-batch. To mitigate the representation collapse issue, we maximize the agreement of representation from the same set and keep distance of representation from different sets, the loss function for a sample X_p involving in set $\mathbf{X}_i$ (same for both $\mathbf{X}_i^-$ and $\mathbf{X}_i^+$) is defined as:

$$\mathcal{L}_{con} = -\log \frac{\sum_{X_j \in \mathbf{X}_i} \mathbb{1}_{[j \neq p]} e^{(\text{sim}(r(X_p), r(X_j))/\tau)}}{\sum_{X_j \notin \mathbf{X}_i} e^{(\text{sim}(r(X_p), r(X_j))/\tau)}} \tag{6}$$

Where $r(\cdot)$ is the sentence representation, $\mathbb{1}_{[j \neq p]} \in \{1, 0\}$ is an indicator function for excluding the sample itself, $sim(\boldsymbol{r_i}, \boldsymbol{r_j}) = \boldsymbol{r_i}^\top \boldsymbol{r_j} / \|\boldsymbol{r_i}\| \|\boldsymbol{r_j}\|$, i.e., the cosine similarity, τ is a temperature parameter. The final loss $\mathcal{L}_{con}$ for contrastive learning is computed by averaging the loss on every sample in each text set in a mini-batch. This loss is a generalization of the NT-Xent (the normalized temperature-scaled cross-entropy loss)(Chen et al., 2020a), as more than one positive pairs for each sample are considered .

Besides the contrastive part, we also incorporate supervised information in the final loss objective $\mathcal{L}_{PCT}$ for optimizing on both model performance and interpretability:

$$\mathcal{L}_{PCT} = \underbrace{\mathcal{L}_{sup}}_{\text{Performance}} + \underbrace{\alpha \mathcal{L}_{con}}_{\text{Interpretability}} \tag{7}$$

Where $\mathcal{L}_{sup}$ is the supervised loss objective (e.g., cross-entropy loss), α is a parameter balancing the two objectives. The joint training objective ensures that the accuracy of model is not hurt while addressing the representation collapse issue.

5 Evaluation

To verify the effectiveness of our method, we evaluate PCT with two other baselines on three popular datasets involving four different models.

5.1 Experiment Setup

Dataset. Our experiments are conducted on three datasets. **AG News** (Zhang et al., 2015), a topic classification dataset containing news articles in the World, Sport, Business, and Sci/Tech area, 120,000 for training and 7,600 for testing. **MR** (Pang and Lee, 2005), a polar samples dataset that contains movie reviews from Rotten Tomatoes, 8,530 for training, and 1,066 for testing. **IMDB** (Maas et al., 2011), a binary sentiment classification dataset that contains 25,000 polar movie reviews for training, and 25,000 for testing.

Model. Four models with different structures and complexities are adopted. **TextCNN** (Kim, 2014): This model has a 300-dimensional embedding layer (Pennington et al., 2014), a convolutional layer with 3 window sizes $(3, 4, 5)$ and 150 filters for each window size, and a dense layer. **LSTM**: This model has a 300-dimensional embedding layer, a Bi-directional LSTM layer composed of 150 units, and a dense layer. **BERT**: This model is a transformer model pretrained on a large corpus of language data. **DistilBERT** (Sanh et al., 2019) : This model is a small, fast Transformer model with 40% less parameters than *bert-base-uncased*.

Baselines. As few works have been devoted to addressing the model pathology, we compare PCT with two training methods. **Normal**: This method trains or fine-tunes the model with the cross-entropy loss objective $\mathcal{L}_{sup}$. **Entropy** (Feng et al., 2018): This method trains or fine-tunes the model to simultaneously maximize the log-likelihood on normal examples and the entropy on the samples with unimportant words reduced. See Appendix for the details on baselines.

Implementation Details. The max sequence length is set as 64. The batch size is set as 64. We use the *bert-base-uncased* as the basic BERT model, and the *distilbert-base-uncased* as the basic DistilBERT model. We set 10% of words with highest and lowest saliency in a sentence as important ($p_i = 0.1$) or unimportant ($p_u = 0.1$) words rather than using a fixed number m. We adopt Adam (Kingma and Ba, 2015) as optimizer. Most setting in learning rate / parameter α / parameter τ for TextCNN, LSTM, BERT, DistilBERT: 5e-4 / 0.1 / 0.7, 5e-4 / 0.1 / 0.7, 3e-5 / 1.2 / 0.7, 3e-5 / 0.15 / 0.15. Parameter λ in Entropy is set as 1e-3 which is the same as the original paper. All reported results are the average of three individual runs. Accuracy and $\mathcal{F}(X)$ are computed on all original text, while others are computed on all reduced samples.

5.2 Main Results

Model accuracy is not impaired. Interpretability is often inconsistent with the model perfor-

		AG News				MR				IMDB			
		ACC	$\mathcal{F}(X)$	Comp	Suff	ACC	$\mathcal{F}(X)$	Comp	Suff	ACC	$\mathcal{F}(X)$	Comp	Suff
LSTM	Normal	91.59	0.93	0.07	0.03	79.64	0.94	0.07	0.06	**78.12**	0.84	0.05	0.01
	Entropy	90.76	0.93	0.05	0.03	80.02	0.92	0.10	0.07	75.71	0.78	-0.04	0.01
	PCT	**92.09**	0.95	0.33 $\overset{0.14}{\leftrightarrow}$	0.19	**80.39**	0.83	0.08 $\overset{0.04}{\leftrightarrow}$	0.04	77.78	0.83	0.13 $\overset{0.06}{\leftrightarrow}$	0.07
TextCNN	Normal	89.49	0.92	0.02	0.02	79.02	0.83	0.08	0.04	75.34	0.80	0.03	0.02
	Entropy	89.59	0.91	0.03	0.02	78.83	0.84	0.07	0.03	77.84	0.78	0.10	0.06
	PCT	**92.18**	0.94	0.10 $\overset{0.06}{\leftrightarrow}$	0.04	**79.74**	0.92	0.12 $\overset{0.08}{\leftrightarrow}$	0.04	**77.94**	0.85	0.10 $\overset{0.06}{\leftrightarrow}$	0.04
DistilBERT	Normal	94.50	1.00	0.01	0.01	84.62	0.99	0.04	0.02	82.30	1.00	0.04	0.02
	Entropy	**94.63**	0.97	0.03	0.02	**85.65**	1.00	0.05	0.02	**82.44**	1.00	0.05	0.02
	PCT	93.59	0.92	0.09 $\overset{0.08}{\leftrightarrow}$	0.01	85.12	0.91	0.09 $\overset{0.05}{\leftrightarrow}$	0.04	82.36	0.90	0.12 $\overset{0.10}{\leftrightarrow}$	0.02
BERT	Normal	**95.16**	0.98	0.01	0.01	**86.40**	1.00	0.03	0.02	**84.30**	1.00	0.03	0.02
	Entropy	94.61	1.00	0.02	0.01	86.39	0.99	0.04	0.02	83.80	0.92	0.07	0.03
	PCT	94.88	0.96	0.08 $\overset{0.04}{\leftrightarrow}$	0.04	86.37	0.97	0.08 $\overset{0.04}{\leftrightarrow}$	0.04	83.78	0.91	0.08 $\overset{0.05}{\leftrightarrow}$	0.03

Table 1: The comparison on accuracy, confidence, comprehensiveness, and sufficiency of PCT with baselines. **Bold** indicates the best accuracy (in %). All $\mathcal{F}(X)$ results are at an acceptable high region. The $\leftrightarrow$ between *Comp* and *Suff* indicates the largest gap between the two values, which means the influence of important and unimportant words are the most distinguishable. A small or negative value of $(\mathit{Comp} - \mathit{Suff})$ indicates the model pathology.

		AG News				MR				IMDB			
		IR#	UR#	I-Var	U-Var	IR#	UR#	I-Var	U-Var	IR#	UR#	I-Var	U-Var
LSTM	Normal	26.59	28.74	51.35	33.22	13.35	15.48	9.20	7.18	28.06	37.83	52.12	43.31
	Entropy	27.78	28.47	18.26	16.29	12.88	15.13	9.04	6.58	27.63	34.76	49.60 $\overset{9.56}{\Leftrightarrow}$	40.04
	PCT	17.74 $\overset{8.93}{\leftrightarrow}$	26.67	52.31 $\overset{25.44}{\Leftrightarrow}$	26.87	12.62 $\overset{3.27}{\leftrightarrow}$	15.89	10.07 $\overset{3.13}{\Leftrightarrow}$	6.94	25.64 $\overset{10.54}{\leftrightarrow}$	36.18	40.51	36.76
TextCNN	Normal	24.17	24.19	65.67	61.74	11.42	13.40	10.28	6.60	23.53	24.32	51.03	55.77
	Entropy	24.06	24.08	40.48	51.09	9.51	11.28	11.38	7.19	20.34 $\overset{4.19}{\leftrightarrow}$	24.53	64.67 $\overset{16.89}{\Leftrightarrow}$	47.78
	PCT	23.10 $\overset{0.50}{\leftrightarrow}$	23.60	51.41 $\overset{3.07}{\Leftrightarrow}$	48.34	9.25 $\overset{2.61}{\leftrightarrow}$	11.86	10.86 $\overset{4.31}{\Leftrightarrow}$	6.55	21.61	24.71	54.23	52.51
DistilBERT	Normal	33.15	35.22	27.73	26.06	14.89	17.55	9.85	8.92	30.83	39.21	45.84	41.73
	Entropy	33.20	35.44	25.27	22.84	14.61	17.90	10.56	9.16	30.87	39.20	46.14	42.22
	PCT	31.17 $\overset{2.62}{\leftrightarrow}$	33.79	32.10 $\overset{6.75}{\Leftrightarrow}$	25.35	14.24 $\overset{3.55}{\leftrightarrow}$	17.79	11.60 $\overset{3.32}{\Leftrightarrow}$	8.28	29.31 $\overset{9.95}{\leftrightarrow}$	39.26	53.25 $\overset{14.17}{\Leftrightarrow}$	39.08
BERT	Normal	34.01	35.57	27.58	27.10	15.18	17.86	10.94	11.77	33.01	39.96	47.24	47.37
	Entropy	33.62	35.41	27.16	27.24	15.21	18.14	10.70	10.99	32.80	40.08	46.54	44.94
	PCT	33.52 $\overset{2.07}{\leftrightarrow}$	35.59	25.89 $\overset{1.01}{\Leftrightarrow}$	24.88	14.27 $\overset{3.74}{\leftrightarrow}$	18.01	12.16 $\overset{1.68}{\Leftrightarrow}$	10.48	32.20 $\overset{7.91}{\leftrightarrow}$	40.11	54.18 $\overset{10.03}{\Leftrightarrow}$	44.15

Table 2: The comparison on reduced number and saliency variance of PCT with baselines. The $\leftrightarrow$ and the $\Leftrightarrow$ indicate the largest values of (*UR#* − *IR#*) and (*I-Var* − *U-Var*), which means the most distinguishable influence of important and unimportant words. A model is pathology if *IR#* < *UR#*, and if *I-Var* < *U-Var*.

mance, as complex models tend to have better performance, while simple models are more interpretable. We report the accuracy of models trained with different methods on three datasets in Table 1.

Our method does not hurt the performance, which meets our basic expectation, but can also slightly improve LSTM and TextCNN. This result indicates that the regularization brought by the contrastive part of our method helps mitigate the overfitting of the unpre-trained model. On the pre-trained models (BERT, DistilBERT), our model is guaranteed to have only a slight impact on performance.

Saliency is more consistent with Confidence. The confidence related results are illustrated in Table 1. For normal samples, our method ensures that the model confidence is sufficiently high ($\mathcal{F}(X) > 0.83$), indicating that on unperturbed samples, the model can adequately consider the influence of important words. While the *Comp* value shows a large decrease when the important words are reduced, indicating that the important words are influential in decision. The *Suff* value will also slightly decrease, i.e., $\mathit{Suff} < \mathit{Comp}$, which is interpretable as we analyzed in Section 3.5 that unimportant words also contain context information while they should not be focused much on. It should be noted that, for the Normal model, *Comp* value is very close to the *Suff* value (average $\mathit{Comp} - \mathit{Suff}$, Normal: 0.015, Entropy: 0.019, PCT: 0.067), which quantitatively demonstrates the inconsistency. The effectiveness of Entropy is weaker than our method.

Important words are more influential in shifting label. The results on reduced number are reported in Table 2. Our method can effectively decrease *IR#*, indicating that important words are actually influential for the prediction, and it is intuitive that the labels will change with fewer important words reduced. As for *UR#*, our method ensures that *IR#* $<$ *UR#*, indicating the influence of unimportant words are lower than important words, and more words reduction are needed to shift the label. The average gap between *IR#* and *UR#* of our method is 4.89, while for Entropy is 3.48, for Normal is 3.26, which indicates that the model is more interpretable when regularization is imposed both on important and unimportant words.

Unimportant words have less impact on saliency stability. The results on saliency variance are reported in Table 2. Our method ensures *U-Var* decrease, indicating that the unimportant words have a slighter impact on the saliency of other words. The average *U-Var* of Normal is 30.89, while of 27.19 for Entropy and 27.51 for PCT. Meanwhile, our method enlarge the gap between *U-Var* and *I-Var* (average, Normal: 3.18; Entropy: 2.79; PCT: 6.54), which demonstrate that important words have broader impact on the saliency of other words than unimportant words. Entropy ensures *U-Var* decrease, while fail to enlarge the gap.

5.3 Ablation Study

In this section, we conduct ablation study on batch size, reduced percentage, parameter τ and α.

Batch size. The influence of batch size is shown in Table 3. We find that the model tend to get better accuracy and interpretability with a larger batch size, as more contrastive samples are generated.

Batch Size	ACC	$\mathcal{F}(X)$	Comp	Suff	IR#	UR#	I-Var	U-Var
4	77.67	0.92	0.10	0.06	8.63	10.47	10.43	7.37
8	77.76	0.83	0.11	0.05	9.15	11.36	10.73	7.33
16	77.86	0.79	0.11	0.04	9.51	11.45	10.18	6.95
32	78.51	0.87	0.11	0.04	8.87	11.19	10.56	7.23
64	79.74	0.92	0.12	0.04	9.25	11.86	10.86	6.55
96	79.17	0.80	0.13	0.04	9.36	11.58	10.43	6.78
128	79.36	0.86	0.13	0.04	9.28	11.33	10.38	7.11

Table 3: Influence of batch sizes when TextCNN trained with PCT on MR.

Reduced percentage. The influence of reduced percentage is shown in Table 4. We find that the reduced percentage hardly affects the model accuracy. The *Suff* value will decrease effectively when the positive contrasts (p_u) are added, while the negative contrasts (p_i) tend to enlarge the gap between *Comp* and *Suff*. Our method is not sensitive to the reduced percentage when both positive and negative pairs are considered.

	ACC	$\mathcal{F}(X)$	Comp	Suff
Normal Model	79.02	0.83	0.08	0.04
$+p_u = 0.1$	79.17	0.87	0.08	0.03
$+p_u = 0.3$	79.04	0.79	0.09	0.02
$+p_i = 0.1$	79.26	0.87	0.12	0.06
$+p_i = 0.3$	78.93	0.82	0.12	0.06
$+p_u = 0.1, +p_i = 0.1$	79.74	0.92	0.12	0.04
$+p_u = 0.3, +p_i = 0.3$	79.12	0.83	0.12	0.04

Table 4: Influence of reduced percentage p_i and p_u when TextCNN trained with PCT on MR. $+p_u = 0.1$ means only generate positive pairs by reducing 10% unimportant words, $+p_i$ means only generate negative pairs, $+p_i, +p_u$ means generate both.

Parameter τ and α. The influence of temperature τ and α is shown in Table 5. We find that model accuracy is slightly affected by τ, and the gap between *Comp* and *Suff* is guaranteed with different τ, while the values will slightly fluctuate. Our method is sensitive to α, as a over large α will hurt model performance and interpretability, while a proper α will benefits them both.

τ	ACC	$\mathcal{F}(X)$	Comp	Suff	α	ACC	$\mathcal{F}(X)$	Comp	Suff
0.05	78.51	0.82	0.10	0.04	0.05	79.34	0.80	0.10	0.03
0.10	78.14	0.77	0.12	0.03	0.10	79.74	0.92	0.12	0.04
0.15	78.05	0.76	0.11	0.04	0.15	78.42	0.74	0.09	0.01
0.30	78.05	0.88	0.13	0.06	0.30	78.14	0.79	0.12	0.07
0.50	78.61	0.74	0.09	0.01	0.50	75.42	0.83	0.13	0.09
0.70	79.74	0.92	0.12	0.04	0.70	74.20	0.78	0.14	0.12
0.90	78.71	0.78	0.08	0.02	0.90	72.89	0.78	0.14	0.13
1.00	78.71	0.84	0.09	0.04	1.00	72.61	0.78	0.14	0.13
1.20	77.77	0.78	0.08	0.02	1.20	71.58	0.74	0.14	0.13

Table 5: Influence of parameter τ and α when TextCNN trained with PCT on MR.

6 Conclusion

In this paper, we propose PCT, a contrastive learning framework for addressing the representation collapse issue and mitigating the inconsistency between word saliency and model confidence for natural language models. We construct the contrastive pairs with saliency-based word reduction. Our model-agnostic method can generate more interpretable models without extra data and changes to the model. Extensive quantitative and qualitative evaluations demonstrate that our method can mitigate the model pathology while keeping model performance. We hope the analysis and the method proposed in our paper will provide a new perspective on model interpretability.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This research was supported by National Research and Development Program of China (No.2019YFB1005200).

Ethical and Societal Impact

In this paper, we reveal the model pathology on the inconsistency between word saliency and model confidence and present a contrastive learning framework for mitigating the model pathology. It is possible that the method of measuring the model pathology can be utilized for benign purposes like ensuring the attribution of model prediction is consistent with human intuition and malign ones such as discovering and exploiting model vulnerabilities. The method may also amplify safety and security concerns in critical domains such as toxic comment classification and rumor detection. However, we argue that it is necessary to study the model pathology and interpretability openly if we want the security risks to be better controlled. We believe that the research on model pathology and interpretability will also motivate the community to pursue models with higher reliability and trustworthiness, rather than just the models with better performance and efficiency. The proposed framework is a possible solution to mitigate security risks for these untrustworthy models. All the datasets we use in this paper are publicly available. No demographic or identity characteristics are used in this paper.

References

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, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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, NeurIPS 2020, December 6-12, 2020, virtual*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 3719–3728. Association for Computational Linguistics.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 879–895. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the NAACL-HLT*, pages 1069–1078. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4932–4942. Association for Computational Linguistics.

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, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Julia Strout, Ye Zhang, and Raymond J. Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 56–62. 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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. 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, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5065–5075. Association for Computational Linguistics.

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, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Ye Zhang, Iain James Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 795–804. The Association for Computational Linguistics.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. 2019. Local aggregation for unsupervised learning of visual embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6001–6011. IEEE.

Appendix

Additional Experiential Details

Details on Baselines. We detail the baseline methods in the main text:

- **Entropy** (Feng et al., 2018) **.** This method fine-tune the existing model to simultaneously maximize the log-likelihood on regular examples and the entropy on reduced examples:

$$\mathcal{L}_{ent} = \sum_{(X,Y)\in(\mathcal{X},\mathcal{Y})} \log(f(Y|X)) + \lambda \sum_{X^-\in\mathbf{X}^-} \mathbb{H}(f(Y|X^-)) \tag{8}$$

  where $f(Y|X)$ is the probability of the model predicting Y given X, $\mathbb{H}(\cdot)$ is the entropy, λ is a hyperparameter controlling the strength of entropy regularization, X^- is an sample with unimportant words reduce from the set $\mathbf{X}^-$.

Additional Experiential Results

Confidence Distribution Change with Epoch. Besides the confidence distribution comparisons we report in the Figure 2, we give more results that involving more models and the detailed effect in training process on the confidence distribution. The results of confidence distribution change with epoch are shown in Figure 13-18.

Additional Case Study

t-SNE Visualization of Sentence Representation. We give more case study of representation collapse issue in Figure 5-12. The instance sentences are randomly picked from MR or IMDB dataset. Same as in the main text, BERT is used as the basic model.

Input Reduction Comparisons To demonstrate the effectiveness of our method, we give more case study of input reduction in Table 6-9. The instance sentences are randomly picked from MR or IMDB dataset.

IR#	Sentence
🟩 0	leigh's film is full of **memorable** performances from top to bottom
🟩 1	leigh's film is full of **performances** from top to bottom
🟩 2	leigh's film is full of from top to **bottom**
🟩 3	leigh's **film** is full of from top to
🟩 4	leigh's is full of from **top** to
🟩 5	leigh's is **full** of from top to
🟩 6	leigh's **is** of from top to
🟥 7	leigh's of **from** top to

UR#	Sentence
🟩 0	leigh's film is full of memorable performances **from** top to bottom
🟩 1	leigh's film is full **of** memorable performances top to bottom
🟩 2	**leigh's** film is full memorable performances top to bottom
🟩 3	film is full memorable performances top **to** bottom
🟩 4	film **is** full memorable performances top bottom
🟩 5	film **full** memorable performances top bottom
🟩 6	**film** memorable performances top bottom
🟩 7	memorable performances **top** bottom
🟩 8	memorable **performances** bottom
🟥 9	memorable **bottom**

Table 6: Case 1, Performing input reduction on instance sentence *leigh's film is full of memorable performances from top to bottom*, the illustration of prediction made by model trained with *Normal* method. Green number indicate the *Positive* label predicted by model, and red number indicate *Negative*. **Bold** indicate the word with highest / lowest saliency in the IR# / UR# setting.

IR#	Sentence
0	leigh’s film is full of **memorable** performances from top to bottom
1	leigh’s film is full of **performances** from top to bottom
2	leigh’s film is full of from top to **bottom**

UR#	Sentence
0	leigh’s film is full **of** memorable performances from top to bottom
1	leigh’s film is full memorable performances **from** top to bottom
2	leigh’s film **is** full memorable performances top to bottom
3	leigh’s film full memorable performances top **to** bottom
4	**leigh’s** film full memorable performances top bottom
5	film **full** memorable performances top bottom
6	**film** memorable performances top bottom
7	memorable performances **top** bottom

Table 7: Case 1, Performing input reduction on instance sentence *leigh’s film is full of memorable performances from top to bottom*, the illustration of prediction made by model trained with *PCT* method. Green number indicate the *Positive* label predicted by model, and red number indicate *Negative*. **Bold** indicate the word with highest / lowest saliency in the IR# / UR# setting.

IR#	Sentence
0	a work of **astonishing** delicacy and force
1	a work of delicacy and **force**
2	a work of **delicacy** and
3	a **work** of and
4	a **of** and
5	a **and**

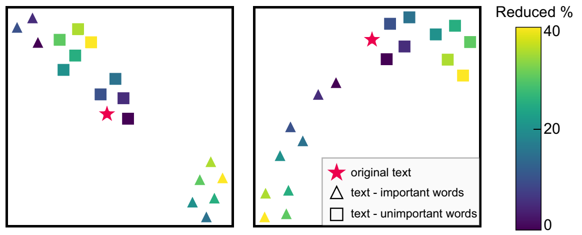
UR#	Sentence
0	a work **of** astonishing delicacy and force
1	a work astonishing **delicacy** and force
2	**a** work astonishing and force
3	work astonishing **and** force
4	work astonishing **force**
5	**work** astonishing
6	**astonishing**

Table 8: Case 2, Performing input reduction on instance sentence *a work of astonishing delicacy and force*, the illustration of prediction made by model trained with *Normal* method. Green number indicate the *Positive* label predicted by model, and red number indicate *Negative*. **Bold** indicate the word with highest / lowest saliency in the IR# / UR# setting.

IR#	Sentence
0	a work of **astonishing** delicacy and force
1	a work of delicacy and **force**
2	a work of **delicacy** and
3	a **work** of and

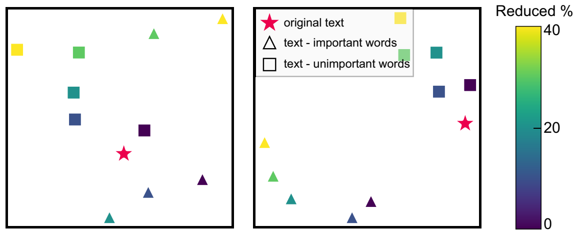
UR#	Sentence
0	a work **of** astonishing delicacy and force
1	a work astonishing **delicacy** and force
2	**a** work astonishing and force
3	work astonishing **and** force
4	**work** astonishing force
5	astonishing **force**
6	**astonishing**

Table 9: Case 2, Performing input reduction on instance sentence *a work of astonishing delicacy and force*, the illustration of prediction made by model trained with *PCT* method. Green number indicate the *Positive* label predicted by model, and red number indicate *Negative*. **Bold** indicate the word with highest / lowest saliency in the IR# / UR# setting.



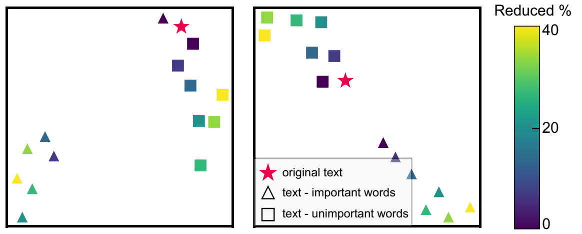
(a) Normal (b) PCT

Figure 5: t-SNE visualization of sentence representation. Cumulatively reducing words in Positive instance *reign of fire never comes close to recovering from its demented premise , but it does sustain an enjoyable level of ridiculousness.*



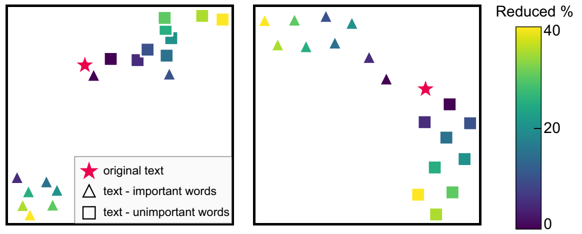
(a) Normal (b) PCT

Figure 6: t-SNE visualization of sentence representation. Cumulatively reducing words in Negative instance *the movie tries to be ethereal , but ends up seeming goofy.*



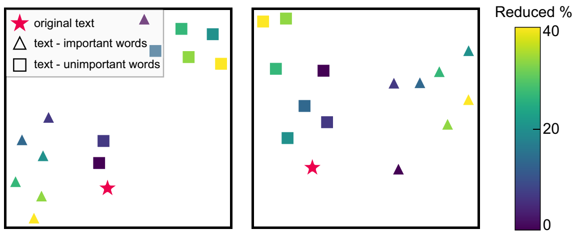
(a) Normal (b) PCT

Figure 7: t-SNE visualization of sentence representation. Cumulatively reducing words in Negative instance *the script is a tired one , with few moments of joy rising above the stale material.*



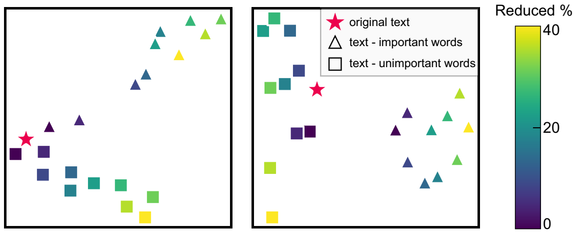
(a) Normal (b) PCT

Figure 8: t-SNE visualization of sentence representation. Cumulatively reducing words in Positive instance *it seems like i have been waiting my whole life for this movie and now i can't wait for the sequel.*



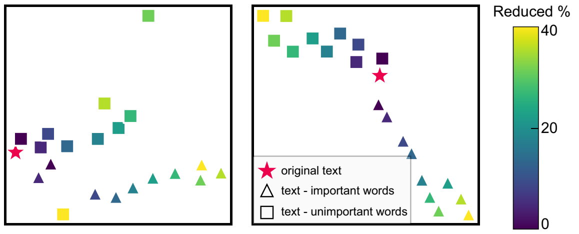
(a) Normal (b) PCT

Figure 9: t-SNE visualization of sentence representation. Cumulatively reducing words in Negative instance *supposedly authentic account of a historical event that 's far too tragic to merit such superficial treatment.*



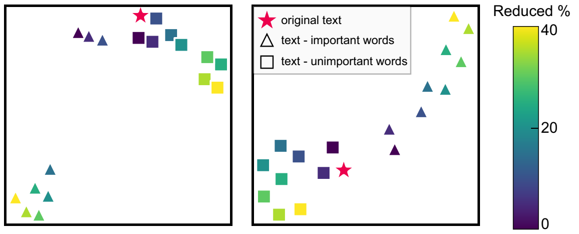
(a) Normal (b) PCT

Figure 10: t-SNE visualization of sentence representation. Cumulatively reducing words in Negative instance *not at all clear what it 's trying to say and even if it were i doubt it would be all that interesting.*



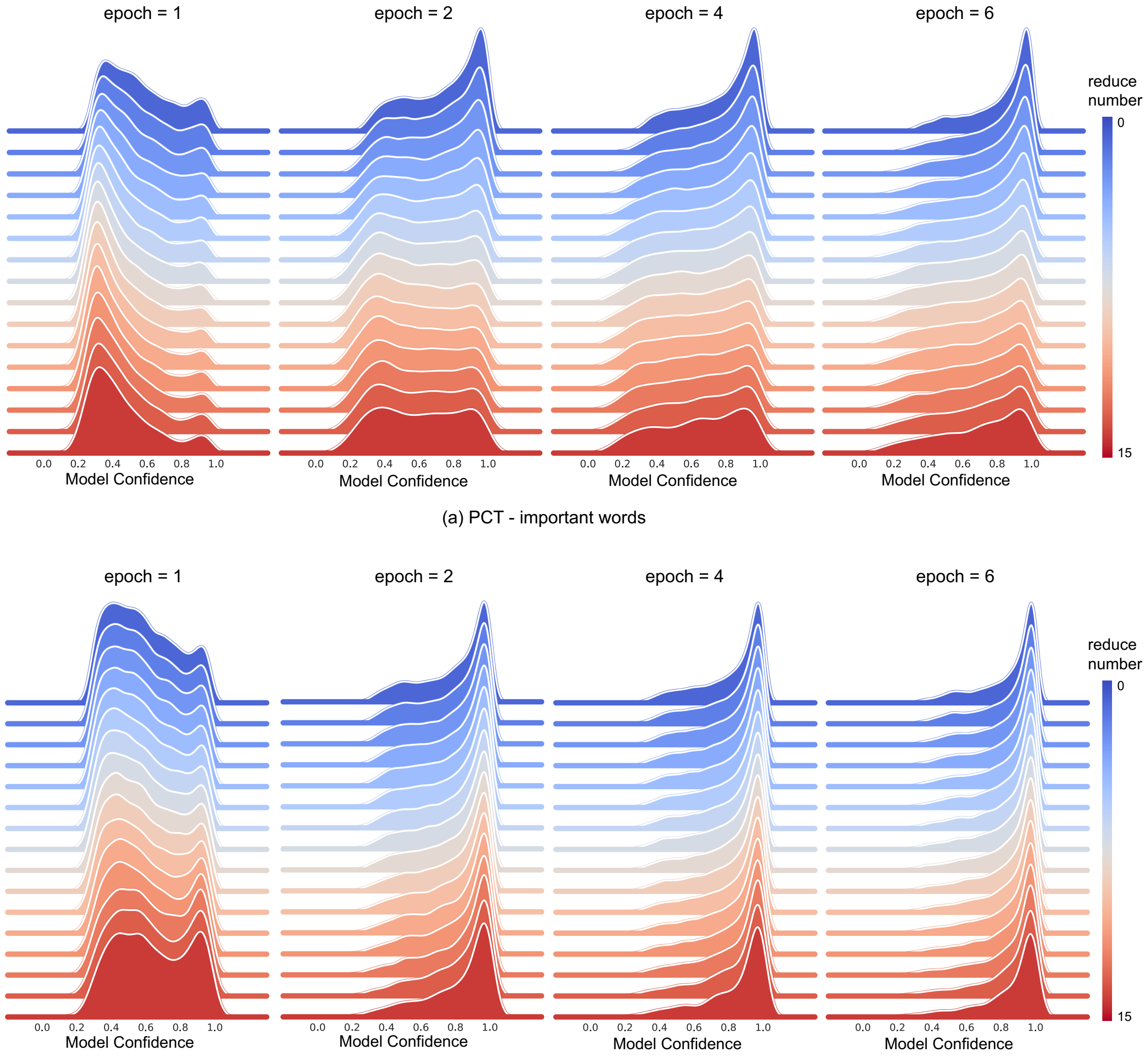
(a) Normal (b) PCT

Figure 11: t-SNE visualization of sentence representation. Cumulatively reducing words in Positive instance *that the real antwone fisher was able to overcome his personal obstacles and become a good man is a wonderful thing that he has been able to share his story so compellingly with us is a minor miracle.*



(a) Normal (b) PCT

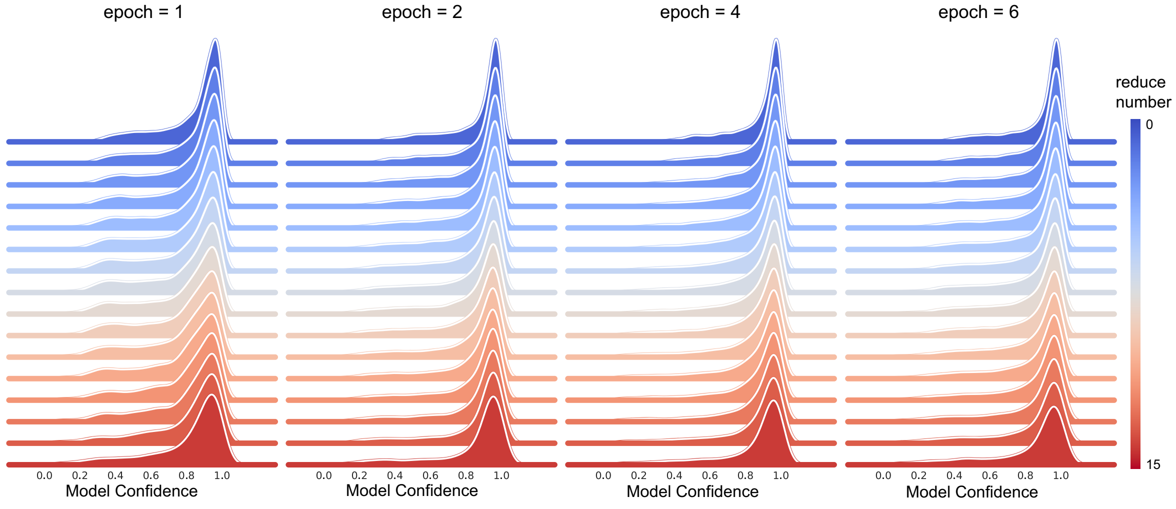
Figure 12: t-SNE visualization of sentence representation. Cumulatively reducing words in Positive instance *tentertaining despite its one joke premise with the thesis that women from venus and men from mars can indeed get together.*



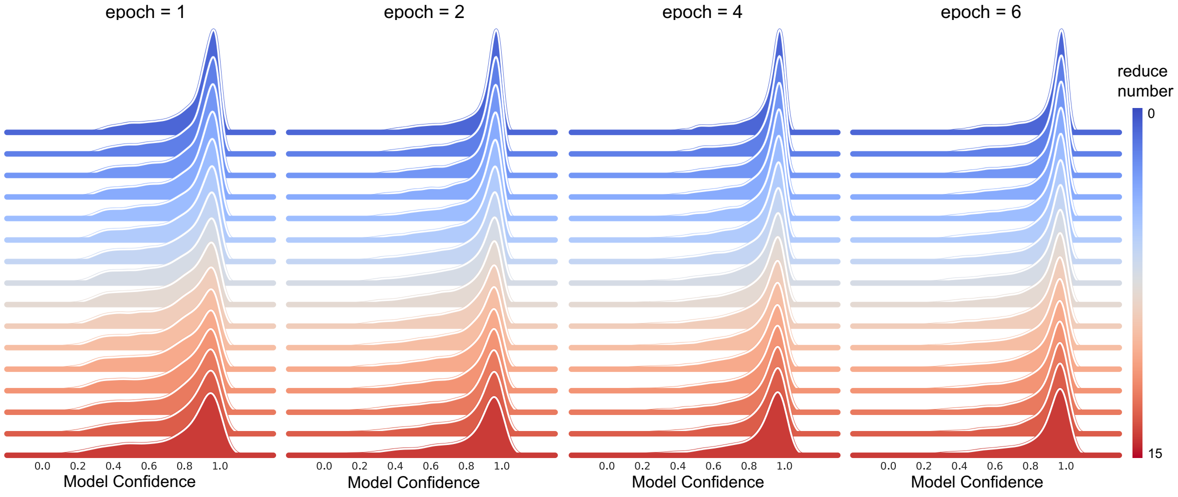
(a) PCT - important words

(b) PCT - unimportant words

Figure 13: The confidence density distribution change with epoch of LSTM trained with *PCT* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.

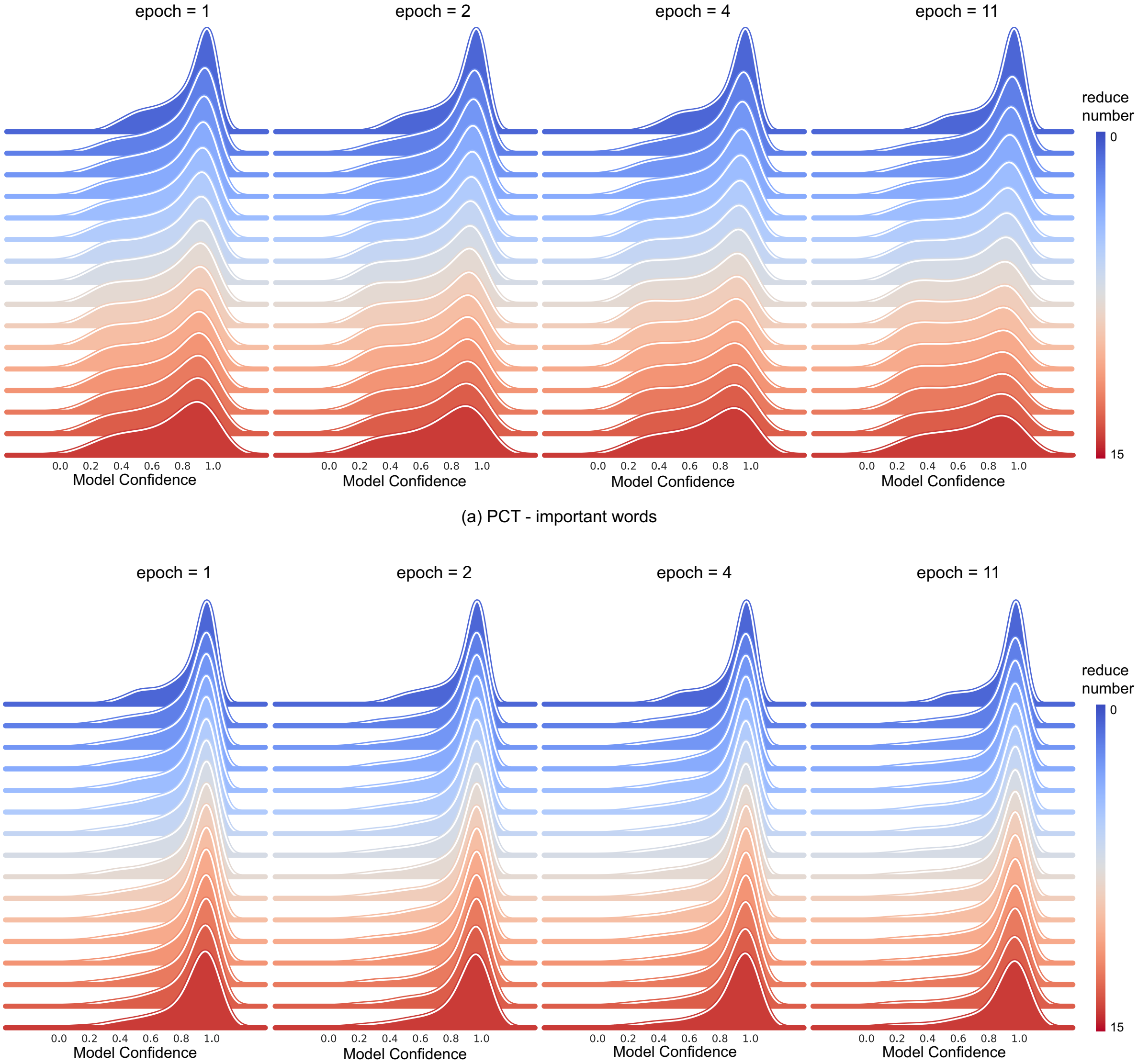


(a) Normal - important words



(b) Normal - unimportant words

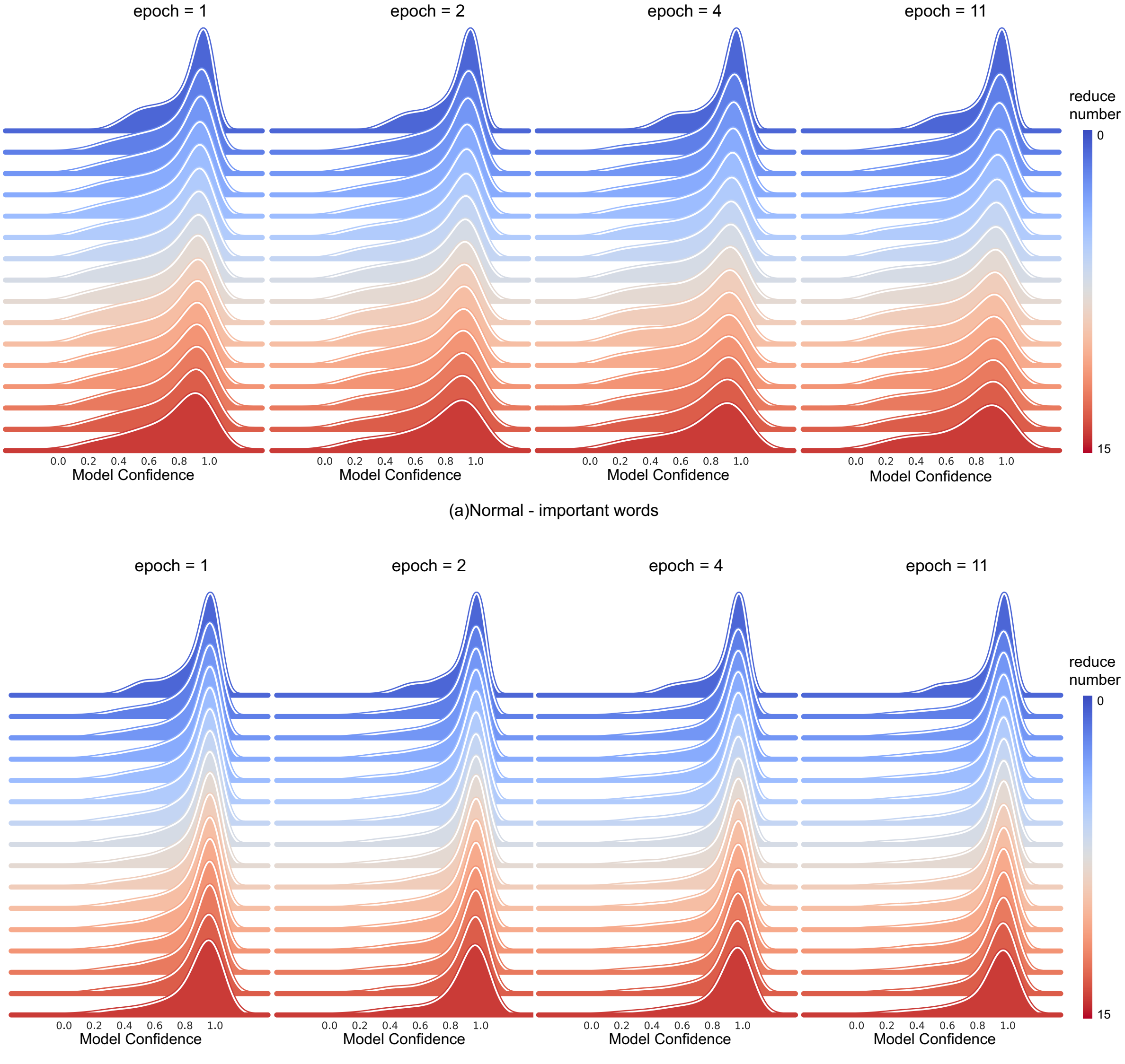
Figure 14: The confidence density distribution change with epoch of LSTM trained with *Normal* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.



(a) PCT - important words

(b) PCT - unimportant words

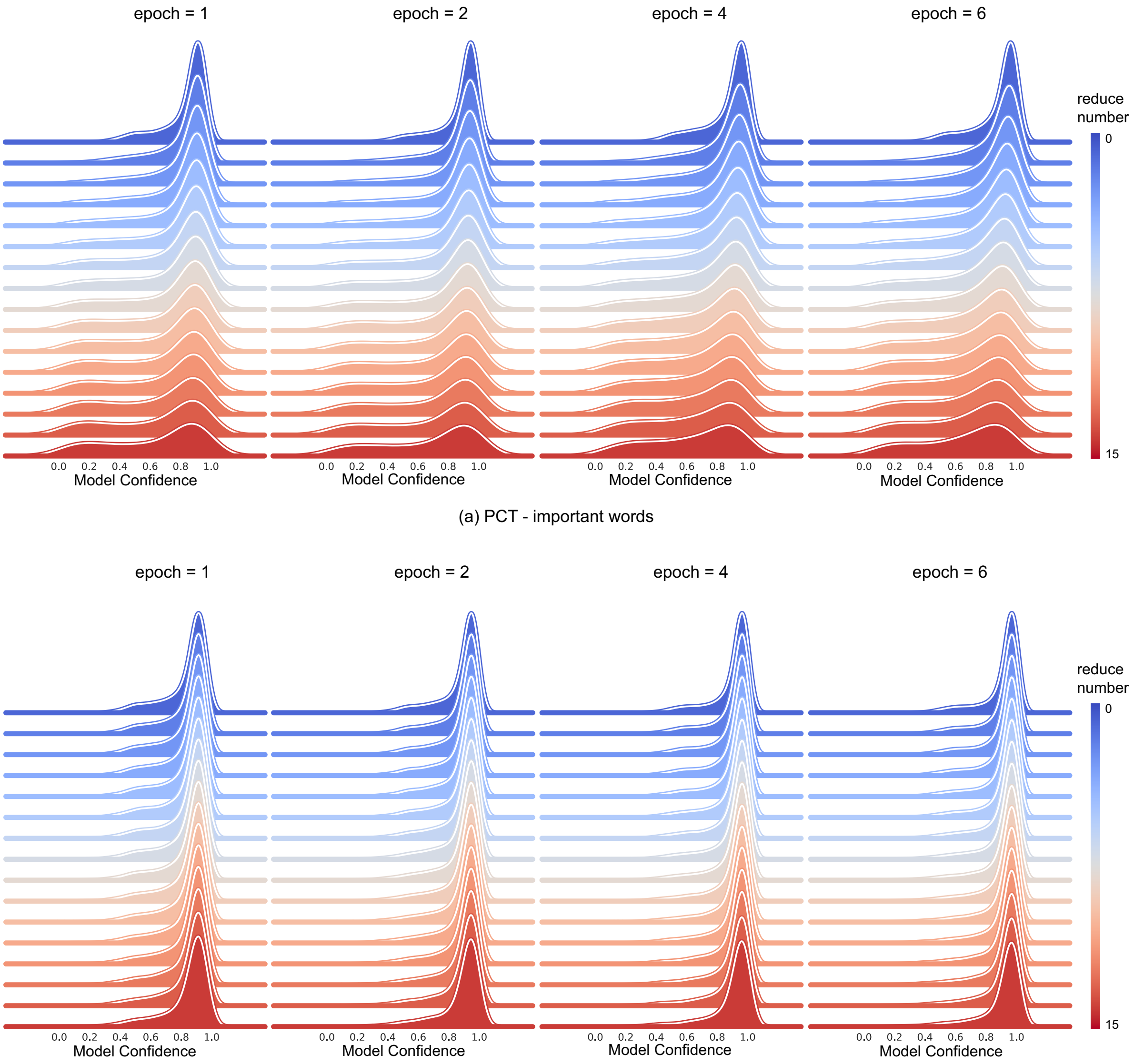
Figure 15: The confidence density distribution change with epoch of TextCNN trained with *PCT* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.



(a)Normal - important words

(b) Normal - unimportant words

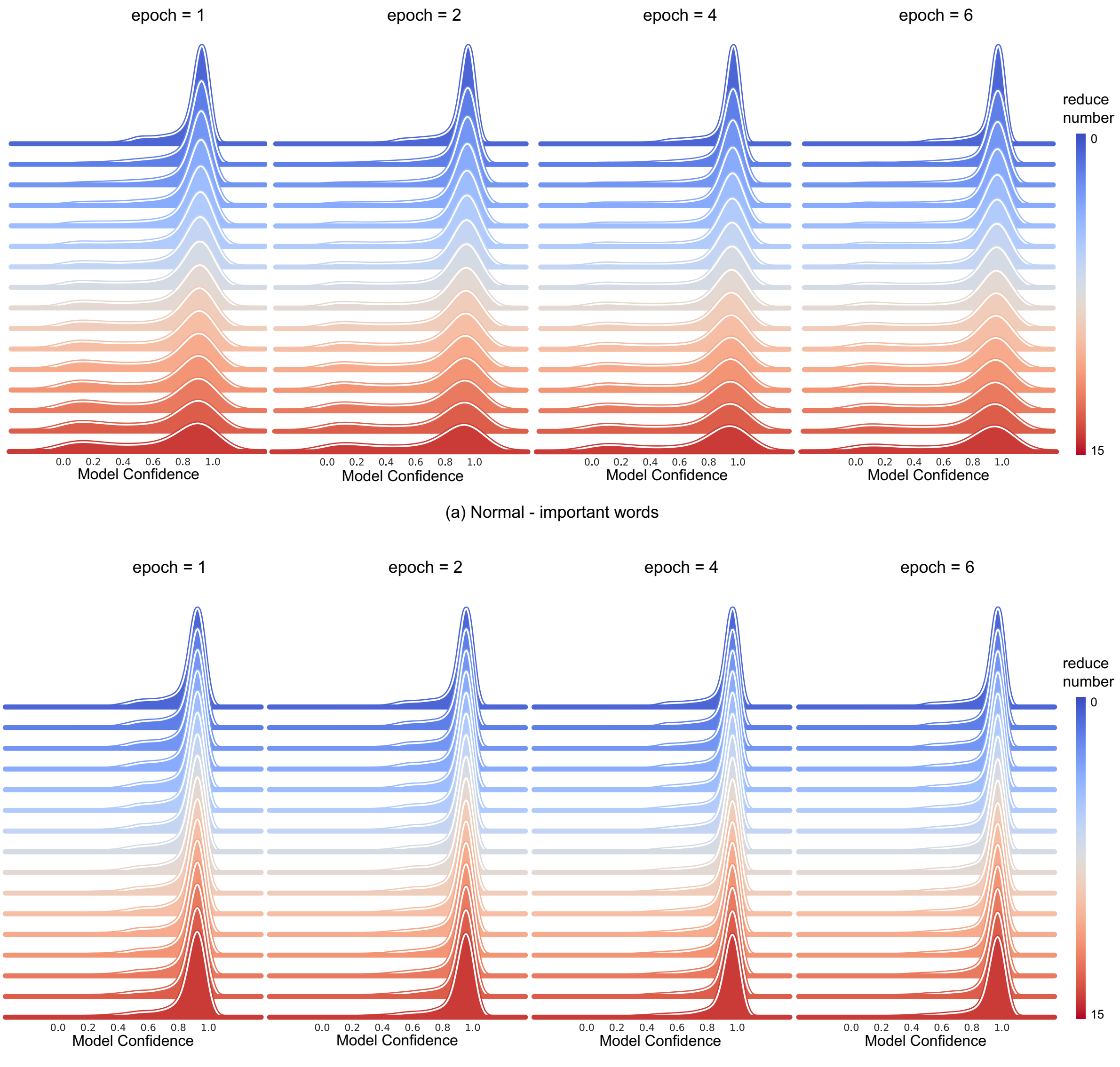
Figure 16: The confidence density distribution change with epoch of TextCNN trained with *Normal* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.



(a) PCT - important words

(b) PCT - unimportant words

Figure 17: The confidence density distribution change with epoch of DistilBERT trained with *PCT* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.



(a) Normal - important words

(b) Normal - unimportant words

Figure 18: The confidence density distribution change with epoch of DistilBERT trained with *Normal* method on the text with important words and unimportant words reduced on AG News testing set. The reduced number is limited to $[0, 15]$. Color indicates the reduced number.