Improve Interpretability of Neural Networks via Sparse Contrastive Coding

Junhong Liu[1*], **Yijie Lin**[2*], **Liang Jiang**[1], **Jia Liu**[1], **Zujie Wen**[1], **Xi Peng**[2†]
[1] Ant Financial Services Group, China.
[2] College of Computer Science, Sichuan University, China.
{daniel.ljh, tianxuan.jl, jianiu.lj, zujie.wzj}@antgroup.com
{linyijie.gm, pengx.gm}@gmail.com

Abstract

Although explainable artificial intelligence (XAI) has achieved remarkable developments in recent years, there are few efforts have been devoted to the following problems, namely, i) how to develop an explainable method that could explain the black-box in a model-agnostic way? and ii) how to improve the performance and interpretability of the black-box using such explanations instead of pre-collected important attributions? To explore the potential solution, we propose a model-agnostic explanation method termed as Sparse Contrastive Coding (SCC) and verify its effectiveness in text classification and natural language inference. In brief, SCC explains the feature attributions which characterize the importance of words based on the hidden states of each layer of the model. With such word-level explainability, SCC adaptively divides the input sentences into foregrounds and backgrounds in terms of task relevance. Through maximizing the similarity between the foregrounds and input sentences while minimizing the similarity between the backgrounds and input sentences, SSC employs a supervised contrastive learning loss to boost the interpretability and performance of the model. Extensive experiments show the superiority of our method over five state-of-the-art methods in terms of interpretability and classification measurements. The code is available at https://pengxi.me.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable progress during the past few years. However, relying on stacking somewhat ad-hoc modules, DNNs are often referred to as "black-box" methods that lack understanding of the working mechanisms, thus increasing the risk of applying them into real-world applications (Ribeiro et al., 2016; Rudin, 2019). For example, in medical diagnosis, predictions cannot be acted upon on blind faith. Instead, doctors need to understand the reasons behind the predictions, *e.g.*, which part of the inputs (*e.g.*, chemical index) the model concentrates on.

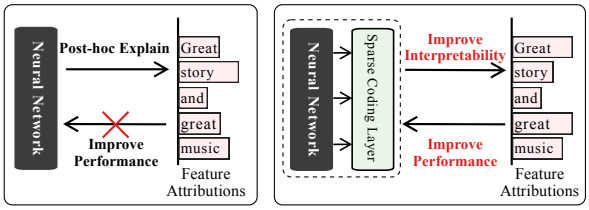


Figure 1: An illustration of the major differences between model explainability and our method. i) Different from model explainability (left) which only gives post-hoc explanations, our method (right) could further improve the interpretability for a given neural network. Such an improvement does not rely on pre-defined important attributions or pre-collected explanations; ii) In addition, unlike most of the existing works which only focus on explainability itself, our method could improve the performance of black-boxes using the obtained explanations.

To understand the working mechanism behind DNNs, explainable artificial intelligence (Chen et al., 2020a, 2018; Lundberg and Lee, 2017) has been devoted in recent and one typical paradigm is explaining the black-boxes from the level of *feature attributions*, *i.e.*, the importance of features w.r.t. the prediction of the network. In general, these studies could be roughly divided into two groups, *i.e.*, *interpretable model* and *model explainability*. To be specific, model explainability (also referred to as post-hoc explanations) (Ribeiro et al., 2016; De Cao et al., 2020; Chen et al., 2021; Sun and Lu, 2020) mainly focuses on explaining the feature attributions through some visualization techniques or agent models. The major merit of the post-hoc explanation is model-agnostic, but it only offers an approximate explainability and cannot improve

[*] The first two authors contributed equally.
[†] Corresponding author.

the interpretability or the performance of the model. On the contrary, interpretable models (Rudin, 2019; Han et al., 2021) try to explain the working mechanism from the model design. In other words, the model could explicitly explain the feature attributions by itself. However, interpretable models are model-specific that cannot be generalized to different neural networks.

Based on the above discussions and observations, this paper aims to study two less-touched problems in XAI. Namely, i) how to develop an explainable method that could explain the black-box in a model-agnostic way? and ii) how to improve the performance and interpretability of the black-box using such explanations instead of pre-collected important attributions. The solution to the problems requires simultaneously enjoying the merits of model explainability and interpretable model to a certain extent. Notably, the answer would be helpful to highlight another perspective of XAI, *i.e.*, XAI should play an important role in improving the model after understanding the model behavior. Notice that, some recent studies have been conducted and proved the effectiveness of XAI in interpretability improvement (Erion et al., 2019; Rieger et al., 2020). However, they often require to collect pre-defined feature attributions, which is labor-intensive and uneconomic.

To explore an effective solution to this problem, this paper proposes a model-agnostic explanation method dubbed sparse contrastive coding (SCC). As shown in Fig. 1, SCC designs a novel sparse coding layer (SCL) which explains the word-level feature attributions based on the hidden states of each layer in the model. To make the explanation faithful and exploit the explainability for model improvement, SCC employs a novel loss function consisting of a sparse coding loss, a contrastive coding loss, and a cross entropy loss. Specifically, the cross entropy loss is enforced between the prediction of texts masked by the feature attributions and the ground-truth to achieve word-level explainability. To make the explanation concise, the sparse coding loss enforces a sparse constraint on the feature attributions so that the foreground words are disentangled from the backgrounds. To further exploit the explainability for improving the model, the contrastive coding loss is enforced on three kinds of input divided by the feature attributions, *i.e.,* the whole texts, foregrounds, and backgrounds. Different from the vanilla methods (He et al., 2020; Chen et al., 2020b; Lin et al., 2021, 2022; Yang et al., 2022), our contrastive coding loss works in a supervised fashion and embraces the properties of negative sample mining and auto data augmentation that could boost the interpretability and performance.

The main contributions and novelties of this paper could be summarized as below: i) we study two less-touched problems in XAI as the aforementioned, *i.e.*, how to develop a model-agnostic method to explain a given black-box and use such explanations to improve its performance and interpretability? To the best of our knowledge, there are few efforts have been devoted so far; ii) we accordingly propose SCC whose basic idea is disentangling the important words from inputs to improve interpretability and discrimination. Extensive experiments on six textual datasets verify the effectiveness of SCC in terms of interpretability and classification metrics.

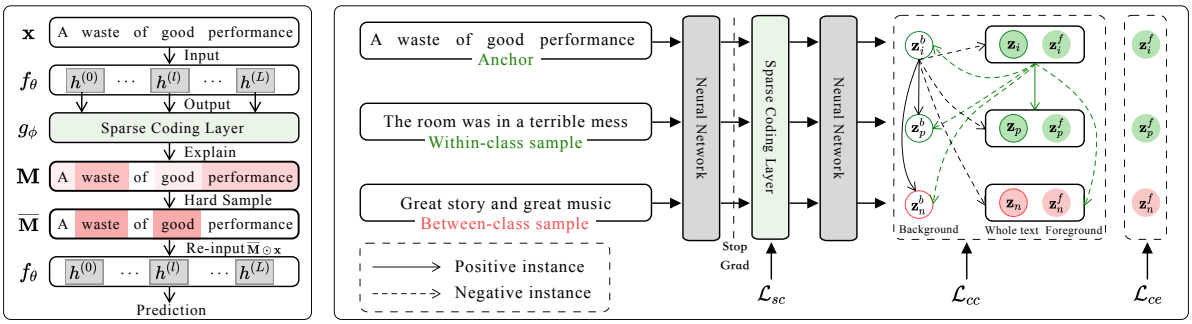
2 Related Work

This work is closely related to model explainability and interpretable models which will be briefly introduced in this section.

2.1 Model Explainability

Model explainability (post-hoc) methods mainly focus on explaining models by detecting feature attributions, *i.e.*, explaining the model by evaluating the contribution of each feature (Guan et al., 2019). For example, LIME (Ribeiro et al., 2016) learns feature attributions by using a local linear model with perturbations to approximate the black-box model. L2X (Chen et al., 2018) aims to reveal the importance of features by maximizing the mutual information between the chosen words and the outputs of the model. KernelSHAP (Lundberg and Lee, 2017) employs different Shapley values to compute feature attributions. Recently, some post-hoc explanation methods enforce the model to focus on pre-defined important features using human-annotated explanations, which achieve remarkable progress. For example, (Rieger et al., 2020) utilizes the contextual decomposition to encode prior knowledge into explanations and (Erion et al., 2019) calculates the expected gradients to make full use of attribution priors.

Although this paper also explores interpretability based on feature attributions, it is different from the aforementioned studies in the given aspects.



(a) Working mechanism of SCL (b) The training framework of SCC

Figure 2: **Overview of the proposed SCC.** (a) Sparse coding layer (SCL) g_ϕ is designed to measure the feature attributions $\mathbf{M}$ based on the output of each hidden layer $h^{(i)}$ in the model f_θ. (b) SCC contains three joint optimization losses, namely, sparse coding loss $\mathcal{L}_{sc}$, contrastive coding loss $\mathcal{L}_{cc}$, and cross entropy loss $\mathcal{L}_{ce}$. Specifically, the cross entropy loss is enforced between the prediction of texts masked by the feature attributions and the ground-truth to achieve word-level explainability. The sparse coding loss is enforced on the feature attributions to distinguish the irrelevant words and make the explanation concise. The contrastive coding loss is enforced on the whole text $\mathbf{z}$, foregrounds (task-relevant words) $\mathbf{z}^f$, and backgrounds (task-irrelevant words) $\mathbf{z}^b$ to boost the interpretability and performance of the model. The subscripts i, p, and n denote the i-th sample and corresponding within-class and between-class samples.

On the one hand, our method does not rely on the pre-defined important attributions or pre-collected explanations (Erion et al., 2019), thus enjoying a more economic solution. On the other hand, our method could not only improve the interpretability but also the performance. In contrast, existing methods may cause the performance drop due to inconsistency between the human explanation and model reason process (Jacovi and Goldberg, 2020).

2.2 Interpretable Models

Instead of generating post-hoc explanations, interpretable models aim to build module-decomposable or algorithm-transparent neural networks. For example, TELL (Xi et al., 2021) proposes an algorithm-transparent clustering network which reformulates the k-means objective as a neural layer. SENN (Alvarez-Melis and Jaakkola, 2018) designs a module-decomposable neural network by progressively stacking a set of linear classifiers. VMASK (Chen and Ji, 2020) utilizes word masks to select important features for building an interpretable neural network.

The major differences between our work and existing works are two-folds. On the one hand, our method is a model-agnostic explanation method which could be applied to explain different black-boxes. In contrast, the interpretability of most existing interpretable models is limited to the original model. On the other hand, most studies achieve interpretability at the cost of performance (Rudin, 2019), whereas our method shows that the interpretability could improve the model performance.

3 Method

This section elaborates on the proposed Sparse Contrastive Coding (SCC) which tries to seek a feasible solution to the aforementioned two problems in XAI, *i.e.*, i) how to develop an explainable method that could explain the black-box in a model-agnostic way? and ii) how to improve the performance and interpretability of the black-box using such explanations?

As illustrated in Fig. 2, SCC explains and improves the black-boxes through three jointly optimizing objectives, namely, sparse coding loss $\mathcal{L}_{sc}$, contrastive coding loss $\mathcal{L}_{cc}$, and cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{sc} + \lambda \mathcal{L}_{cc}, \tag{1}$$

where the balanced factor λ is simply fixed to 0.1 throughout experiments. In the following, we will introduce how the sparse coding layer with $\mathcal{L}_{sc}$ and $\mathcal{L}_{ce}$ is built for embracing explanability in Section 3.1 and how to improve the model performance and the interpretability through $\mathcal{L}_{cc}$ in Section 3.2.

3.1 Explaining Model via Sparse Coding Layer

Without loss of generality, we take text classification as an evaluation task. For an input text

$\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, let $x_i \in \mathbb{R}^d$ $(1 \leq i \leq N)$ represents the embedding of i-th word and N denotes the number of words. The neural network $f_\theta(\cdot)$ aims at predicting the class label $\tilde{y}$ for x through the mapping $f_\theta(\mathbf{x})$.

As shown in Fig. 2(a), to explain the neural networks, we design a sparse coding layer (SCL) g_ϕ that could measure the feature attributions based on the output of each hidden layer in the model. To be specific, let $\mathbf{h} = \langle h^{(0)}, \ldots, h^{(L)} \rangle$ denotes the hidden states of each layer in the neural classifier, where $h^{(0)} = \mathbf{x}$ is the word embedding layer. We identify the important words through g_ϕ: $\mathbf{M} = g_\phi(\mathbf{h}) = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N\}$, where $\mathbf{M}_i \in [0, 1]$ measures the importance of i-th word. In detail, g_ϕ aggregates the information from each layer in a gated form (Chung et al., 2014) with three one-layer MLPs:

$$g^{(i)} = \eta\left(\mathrm{MLP}_{\text{gate}}\left(h^{(i)}\right)\right) \odot \mathrm{MLP}_{\text{rep}}\left(h^{(i)}\right), \tag{2}$$

$$\mathbf{M} = \mathrm{MLP}_{\text{proba}}\left([g^{(0)}; \ldots; g^{(L)}]\right), \tag{3}$$

where η is the Tanh activation function and $[;]$ is the concatenation operation.

To generate an effective explanation, $\mathbf{M}$ is expected to have the following properties: i) removing irrelevant words as many as possible for a concise explanation, ii) and meanwhile correctly selecting relevant words for classification. To achieve the first property, it is encouraged to maximize the sparsity of $\mathbf{M}$ through ℓ_0-norm, *i.e.*, minimizing the number of the non-zeros values,

$$L_0 = \sum_{i=1}^{n} \mathbf{1}_{[\mathbb{R}_{\neq 0}]}\left(\mathbf{M}_i\right). \tag{4}$$

Although ℓ_0-norm is discontinuous and has zero derivative almost everywhere, it is exactly equivalent to ℓ_1-norm under the binary case. Based on this observation, we generate hard masks during training following the reparameterization trick (Maddison et al., 2017) as below

$$\overline{\mathbf{M}}_i = \frac{\exp\left(\left(\log \mathbf{M}_i + G_i^1\right)/\tau\right)}{\sum_{l=0}^{1} \exp\left(\left(w_i^l + G_i^l\right)/\tau\right)}, \tag{5}$$

where $w_i^l = \log\left(l\mathbf{M}_i + (1-l)\left(1-\mathbf{M}_i\right)\right)$, $G_i^l = -\log(-\log u)$ is the l-th Gumbel random variable, $u \sim \mathrm{Uniform}(0, 1)$, and τ is the softmax temperature. In this way, one could surrogate ℓ_0-norm with ℓ_1-norm to achieve sparsity. Furthermore, we introduce a balance term on $\mathbf{M}$ to keep the exploratory in the preliminary training stage. Mathematically, the sparse coding loss for SCL is given by,

$$\mathcal{L}_{sc} = \sum_i \Big(\|\overline{\mathbf{M}}_i\|_1 + \gamma\left(\mathbf{M}_i \log\left(\mathbf{M}_i\right) + \left(1-\mathbf{M}_i\right)\log\left(1-\mathbf{M}_i\right)\right)\Big), \tag{6}$$

where γ is steadily annealed from 1.0 to 0.01 with a decay of 0.099.

To achieve the second property, *i.e.,* selecting most relevant words, we mask the word embeddings based on the measured feature attributions for classification by minimizing the cross entropy loss between the prediction $\tilde{y}$ and the ground-truth y, *i.e.*,

$$\mathcal{L}_{ce}(\tilde{y}, y) = \mathcal{L}_{ce}(f_\theta(\tilde{\mathbf{x}}), y), \tag{7}$$

where

$$\tilde{\mathbf{x}} = \overline{\mathbf{M}} \odot \mathbf{x}, \tag{8}$$

and $\odot$ denotes the element-wise multiplication. As long as the prediction $\tilde{y}$ approximates the ground-truth, we deem $\overline{\mathbf{M}}$ selects the most relevant words.

3.2 Improving Model using Explainability

Most XAI studies are deemed to be important for truthful and safety AI, which somehow ignore another important perspective, *i.e.*, improving the interpretability and performance. After understanding the working mechanism of black-box, it is highly expected not only more trustworthy predictions but also higher performance. To this end, we propose a novel contrastive coding loss which could encourage model improvement using the explainability.

In detail, we first divide the input sentence $\mathbf{x}$ into foreground $\mathbf{x}^f$ and background $\mathbf{x}^b$ through the sparse coding layer,

$$\mathbf{x}^f = \overline{\mathbf{M}} \odot \mathbf{x},\ \mathbf{x}^b = (1 - \overline{\mathbf{M}}) \odot \mathbf{x}, \tag{9}$$

where $\mathbf{x}^f$ denotes the set of important words and $\mathbf{x}^b$ contains all irrelevant words for classification. By passing $\mathbf{x}$, $\mathbf{x}^f$, and $\mathbf{x}^b$ through the neural network f_θ, we could obtain the representations $\mathbf{z}$, $\mathbf{z}^f$, and $\mathbf{z}^b$, accordingly.

For clarity, we first present the general form of our contrastive coding loss and then elaborate on the training details. Let notation t_i marks the i-sample and t_p marks the positive sample for t_i, our

contrastive coding loss is given by

$$\mathcal{L}_{cc} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(t_i \cdot t_p / \tau_2\right)}{\sum_{a \in A(i)} \exp\left(t_i \cdot t_a / \tau_2\right)}, \tag{10}$$

where $P(i)$ denotes the corresponding positive samples set of sample i, $A(i)$ denotes all samples without sample i, and $\tau_2 \in \mathbb{R}^+$ is a scalar temperature parameter.

Notably, one major difference between Eq. (10) and the vanilla contrastive loss (Khosla et al., 2020) lies on the positive/negative construction which is non-trivial as pointed out in (Chen et al., 2020b; Khosla et al., 2020). Specifically, as shown in Fig. 2(b), our contrastive coding loss will deal with three kinds of anchors, *i.e.*, the input sentence $\mathbf{z}_i$, the foreground $\mathbf{z}_i^f$, and the background $\mathbf{z}_i^b$. Mathematically, we have anchor $t_i \in \{\mathbf{z}_i, \mathbf{z}_i^f, \mathbf{z}_i^b\}$ (*i.e.*, t_i could be one of the samples $\mathbf{z}_i$, $\mathbf{z}_i^f$, and $\mathbf{z}_i^b$) which will also determine the choice of positive sample t_p, namely,

- When the anchor is the input sentence or foreground, *i.e.*, $t_i = \mathbf{z}_i$ or $t_i = \mathbf{z}_i^f$, then $t_p \in \{\mathbf{z}_i^f, \mathbf{z}_p, \mathbf{z}_p^f\}$ or $t_p \in \{\mathbf{z}_i, \mathbf{z}_p, \mathbf{z}_p^f\}$ accordingly. In other words, for either of $\mathbf{z}_i$ and $\mathbf{z}_i^f$, the objective aims to minimize its distance with within-class samples $\mathbf{z}_p$ and $\mathbf{z}_p^f$, while maximizing its distance with between-class samples $\mathbf{z}_n$, $\mathbf{z}_n^f$, and all backgrounds $\{\mathbf{z}_i^b, \mathbf{z}_n^b, \mathbf{z}_p^b\}$. The subscripts p and n denote the within-class and between-class samples of $\mathbf{z}_i$ selected by the classification label.
- When the anchor is the background, *i.e.*, $t_i = \mathbf{z}_i^b$, then $t_p \in \{\mathbf{z}_p^b, \mathbf{z}_n^b\}$. In other words, the objective is to ensure that the backgrounds will only contain irrelevant words by pulling all backgrounds together while pushing the other sentences and foregrounds away.

3.3 Discussions

With the above contrastive coding loss, one could maximize the similarity between the foregrounds and the input sentences, while minimizing the similarity between the foregrounds and backgrounds. This strategy could encourage the model to select task-relevant words and throw away irrelevant words, thus boosting the interpretability. By incorporating the explainability into the training process, our contrastive coding loss owns the following desirable properties:

Datasets	C	*#train*	*#dev*	*#test*	L	B
IMDB	2	20K	5K	25K	250	8
SST2	2	67K	872	1.8K	50	16
YELP	2	500K	60K	38K	150	16
TREC	6	5K	452	500	15	16
SUBJ	2	8K	1K	1K	25	16
SciTail	3	24K	1K	2K	50	16

Table 1: **Summary statistics of the datasets.** C is the number of classes, L is the padded sentence length, B is the training batchsize, and # denotes the number of samples in train/dev/test sets.

Negative sample mining. The foregrounds and backgrounds divided by our sparse coding layer could be regarded as augmented positive samples and negative samples. Notably, there are few works that attempt to conduct negative data augmentations since there is no exact definition of negative data augmentations. Through our explainability paradigm, it is reasonable and natural to construct negative pairs using the task-irrelevant samples. As shown in Table 5, the ablation study verifies the effectiveness of such negative sample mining property by discarding the backgrounds contrast.

Auto data augmentation. The huge success of contrastive learning could be partially attributed to effective data augmentation techniques (He et al., 2020; Chen et al., 2020b). Most existing data augmentation methods often resort to hand-crafted approaches, *e.g.*, rotation, flipping, and so on. Different from these methods, our SCC could be regarded as providing an auto data augmentation strategy which utilizes task relevance to filter out the salient words that are negative to the input in the semantic space. In addition, it is worthy to point out the difference between our method and (Gao et al., 2021). In brief, (Gao et al., 2021) randomly removes the fixed-rate words for data augmentation, which is task-irrelevant and the fixed parameter might lead to inferior performance. In contrast, our SCC will select salient words to augment data base on the task relevance, which will be adaptive to different inputs. As shown in Table 3, one could find the superiority of such an auto data augmentation.

4 Experiments

In this section, we carry out experiments on six widely-used textual datasets. For a comprehensive study, we compare SCC with five state-of-the-art approaches on two classification tasks (*i.e.*, sentiment analysis and subjective/objective classifica-

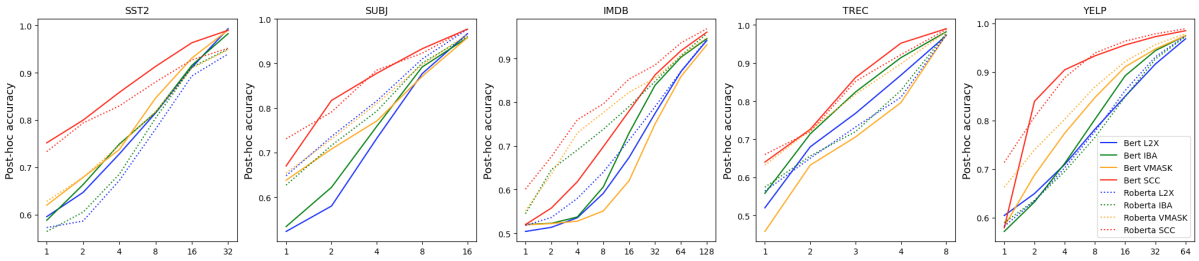


Figure 3: **Post-hoc accuracy. Higher is better.** The vertical axis and horizontal axis denote the post-hoc accuracy and the number of reserved important words, respectively. SCC has achieved significant improvement compared with others.

Models	Methods	AOPC-5					AOPC-10				
		YELP	SST2	IMDB	SUBJ	TREC	YELP	SST2	IMDB	SUBJ	TREC
BERT	BASE	7.23	30.04	5.45	15.08	63.24	**9.80**	32.84	6.57	22.18	66.59
	L2X	6.92	32.28	4.13	15.63	57.29	9.02	32.22	5.06	15.63	57.29
	IBA	6.91	32.04	5.15	15.68	54.64	9.18	32.04	5.90	15.68	54.64
	VMASK	7.47	32.73	5.05	16.43	54.24	9.60	33.73	6.05	16.53	54.24
	SCC	**7.75**	**32.92**	**5.87**	**16.68**	**65.18**	9.64	**36.72**	**6.76**	**23.74**	**68.02**
RoBERTa	BASE	7.66	30.97	4.71	13.80	61.68	9.95	35.59	5.49	**20.37**	65.04
	L2X	6.85	30.07	3.99	13.92	53.09	9.06	35.18	4.80	20.19	57.88
	IBA	7.21	30.43	3.95	13.49	53.92	9.32	35.92	4.73	19.66	58.60
	VMASK	7.10	31.00	6.25	14.03	51.15	9.24	35.75	**7.11**	19.77	56.10
	SCC	**8.01**	**32.02**	**6.39**	**14.69**	**62.48**	**10.06**	**36.31**	7.00	20.12	**65.85**

Table 2: **AOPC scores. Higher is better.** The best and second-best results are highlight in **bold** and underline. SCC focuses on the most important words for prediction compared with baselines thanks to our sparse coding layer and three jointly learning losses.

tion) and natural language inference (NLI) task in terms of the metrics of interpretability and classification performance.

4.1 Experimental Settings

Datasets: Six widely-used datasets are used in our experiments, *i.e.*, YELP reviews dataset (Zhang et al., 2015), movie reviews dataset IMDB (Maas et al., 2011), question classification dataset TREC (Li and Roth, 2002), subjective/objective classification dataset SUBJ (Pang and Lee, 2005), Stanford Sentiment Treebank datasets SST-2 (Socher et al., 2013), and NLI dataset SciTail (Khot et al., 2018). For IMDB and SUBJ datasets, we hold out a portion of the training set as the development set. For the other datasets, we use the original data splits. The statistics of the datasets are given in Table 1.

Implementation Details : The proposed sparse coding layer consists of three one-layer MLPs as shown in Eq. 2 and Eq. 3, *i.e.,* a representation MLP $\mathrm{MLP}_{\mathrm{rep}}$, a gated MLP $\mathrm{MLP}_{\mathrm{gate}}$, and a probability MLP $\mathrm{MLP}_{\mathrm{proba}}$. In detail, $\mathrm{MLP}_{\mathrm{rep}}$ and $\mathrm{MLP}_{\mathrm{gate}}$ project the 768-dimension token representation into 100-dimension, and afterwards, $\mathrm{MLP}_{\mathrm{proba}}$ outputs 1-dimension feature attributions. In the training stage, we optimize the sparse coding layer and the neural classifier in an end-to-end fashion with the aforementioned three objectives. In the testing stage, we only retain the neural classifier and verify its improvement.

To show that SCC could improve the interpretability and the classification performance for a given model, we apply it to two typical neural models, *i.e.*, BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). In detail, we implement SCC in PyTorch 1.7.1 and carry all evaluations on the Red Hat 6.4 OS with a Tesla P100 GPU. To optimize the networks, we adopt AdamW optimizer (Loshchilov and Hutter, 2017) with the default parameters and set the initial learning rate as $1e^{-5}$. The softmax temperature τ and τ_2 are all set to 1.0 and the maximal training epoch is fixed to 10 on all datasets. For fair comparisons, we adopt the best checkpoint on the validation set for all tested methods in terms of accuracy.

Method	Text	Prediction
L2X	Too gory to be a comedy and too silly to be an effective horror film	negative
IBA	Too gory to be a comedy and too silly to be an effective horror film	negative
VMASK	Too gory to be a comedy and too silly to be an effective horror film	negative
SCC	Too gory to be a comedy and too silly to be an effective horror film	negative
L2X	An enjoyable film for the family , amusing and cute for both adults and kids	positive
IBA	An enjoyable film for the family , amusing and cute for both adults and kids	positive
VMASK	An enjoyable film for the family , amusing and cute for both adults and kids	positive
SCC	An enjoyable film for the family , amusing and cute for both adults and kids	positive

Figure 4: **Qualitative evaluation.** The most important four words are highlighted and the color saturation indicates the word attribution. As shown, SCC could capture more precise sentiment words that indicate the same sentiment polarity with the prediction.

Baselines: To show the promising performance of SCC, we compare it with the following methods: i) L2X (Chen et al., 2018) learns the importance of features by maximizing the mutual information between the chosen words and the output of the model. ii) IBA (Schulz et al., 2020) learns the feature attributions based on the information bottleneck theory. iii) VMASK (Chen and Ji, 2020) proposes a learnable mask that removes the irrelevant words and keeps the explanation of the model. iv) SimCSE (Gao et al., 2021) proposes a simple contrastive learning framework whose performance is remarkably benefited from the dropout augmentation. v) Base model which denotes the model is trained by minimizing the cross entropy loss only. Note that, L2X and IBA are proposed for generating post-hoc explanations. To investigate the effectiveness of post-hoc explanations in performance improvement, we integrate L2X and IBA into the model training stage by adding an extra word mask layer as suggested by VMASK.

4.2 Quantitative Evaluation of Interpretability

We adopt two interpretability metrics to evaluate the faithfulness and sufficiency of the model, *i.e.,* AOPC (Nguyen, 2018) and post-hoc accuracy (Chen et al., 2018). In brief, AOPC measures the fidelity by masking the top-scored words and calculating the difference on the predicted probability, while post-hoc accuracy measures the sufficiency of interpretability by keeping the most important words.

AOPC: We calculate the area over the perturbation curve (AOPC) to evaluate the faithfulness of explanations to models. To be specific, AOPC calculates the average change of prediction probability on the predicted class over all test data by deleting top k words. Mathematically,

$$\text{AOPC} = \frac{1}{M}\left(\sum_{i=1}^{M} p(\hat{y} \mid \mathbf{x}_i) - p(\hat{y} \mid \bar{\mathbf{x}}_i^{(k)})\right), \tag{11}$$

where $\hat{y}$ is the predicted label, and M is the number of samples. $\bar{\mathbf{x}}^{(k)}$ is constructed by deleting the top k important words of $\mathbf{x}$ and LIME (Ribeiro et al., 2016) is used to measure the importance of words. Higher AOPC indicates better explanations, *i.e.,* the deleted words are crucial to the model prediction. Note that, due to the over-high computation costs of LIME (Ribeiro et al., 2016), we randomly pick up 2,000 examples from YELP and IMDB, and use the whole SST2, SUBJ, and TREC in the evaluation.

Post-hoc Accuracy: It could evaluate the sufficiency of important words to the model prediction. More specifically, we select the top v words based on feature attributions for classification and compare the performance with the case of the whole text. Note that the importance of words is computed by the word masks (baselines) or the sparse coding layer (SCC). Mathematically, post-hoc accuracy is defined as,

$$\text{post-hoc accuracy} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\left[\tilde{y}_i^{(v)} = \tilde{y}_i\right], \tag{12}$$

where $\tilde{y}_i$ is the predicted label of i-th sample and $\tilde{y}_i^{(v)}$ is that of i-th sample with top v words. Higher values denote better explanations.

Results: Table 2 reports the AOPC scores when the most important 5 and 10 words are deleted (marked as AOPC-5 and AOPC-10). From the

results, one could observe that: i) SCC outperforms all baselines on most datasets in terms of AOPC-5 and AOPC-10. For example, SCC surpasses the best baseline by 1.94% on TREC dataset with BERT in terms of AOPC-5. ii) On the YELP dataset, the AOPC of BASE is even better than L2X and IBA. This phenomenon reveals that assembling post-hoc explanation methods to neural networks is not always encouraging, and proves the effectiveness of our training framework. iii) The AOPC-10 score on TREC is extremely high because the maximum length of sentences is 15 in TREC, *i.e.*, removing the top 10 words would probably exclude most informative words. Figure 3 shows the results of post-hoc accuracy, which shows that SCC significantly outperforms all the tested baselines in all evaluations.

4.3 Qualitative Evaluation of Interpretability

To intuitively investigate the effectiveness of our method, we first present the feature attributions of two examples randomly selected from the SST2 dataset with the BERT backbone. As shown in Figure 4, although all methods have made correct semantic predictions, the interpretability is quite different in such a qualitative evaluation. More specifically, SCC correctly captures the sentiment words “gory" and “silly" in the first example, while the baselines fail in capturing “silly". In the second example, IBA and VMASK ignore “cute", L2X ignores “amusing", while our method captures all three important words. To sum up, SCC could capture more precise sentiment words that indicate the same sentiment polarity with the prediction. More examples are presented in Figure 5.

4.4 Evaluation of Classification

As aforementioned, one major goal of this study is to improve the classification performance by utilizing the explainability. To investigate such a capacity, we compare the tested methods on five datasets in terms of classification accuracy. As shown in Table 3, SCC significantly outperforms baselines by a large performance margin on almost all datasets. It should be pointed out that, SCC is even better than SimCSE (Gao et al., 2021) which is designed for representation learning rather than interpretability.

4.5 Evaluation of Natural Language Inference

Natural language inference is the task of determining whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) given a premise. To verify the universality of our method, we further compare the tested methods on SciTail (Khot et al., 2018) in terms of accuracy and post-hoc accuracy. As shown in Table 4, the remarkable improvement suggests that our SCC could be generalized to different tasks and further improve both the interpretability and performance of the model.

Methods	SST2	TREC	SUBJ	IMDB	YELP
BERT	93.79	96.40	94.80	91.67	96.36
L2X	93.03	96.60	94.85	91.34	96.41
IBA	93.74	96.80	94.90	92.20	96.45
VMASK	93.80	96.80	94.70	92.08	96.48
SimCSE	93.90	**97.00**	94.70	91.72	96.41
SCC	**94.23**	**97.00**	**94.90**	**92.32**	**96.59**
RoBERTa	95.16	96.60	96.00	92.71	96.92
L2X	95.39	96.70	95.90	93.60	96.99
IBA	95.66	96.20	95.70	93.68	97.06
VMASK	95.28	96.40	95.90	90.08	97.07
SimCSE	95.73	96.60	95.90	92.82	97.11
SCC	**96.10**	**97.00**	**96.10**	**93.73**	**97.17**

Table 3: **Classification accuracy.** The top and the bottom six rows denote the results based on BERT and RoBERTa backbones, respectively. As illustrated, SCC outperforms five baselines with two different models on all five datasets.

Methods	Classification Acc	Post-hoc Acc
BERT	91.90	–
L2X	90.73	59.31
IBA	91.44	54.26
VMASK	88.48	58.93
SCC	**92.29**	**62.13**

Table 4: **NLI accuracy.** SCC outperforms four baselines in terms of post-hoc accuracy and classification accuracy.

4.6 Ablation Study

To evaluate our design decisions, we conduct ablation studies on the SST2 dataset. The experiments are designed to isolate the effect of contrastive coding loss and sparse coding loss. Moreover, we also ablate the disentangled backgrounds from contrastive learning to verify the effectiveness of the negative sample mining as discussed in Section 3.3. As shown in Table 5, all objectives are helpful in improving the interpretability and classification performance.

Method	Text	Prediction	Dataset
L2X	Theirs is a simple and heart warming story , full of mirth that should charm all but the most cynical	positive	SST2
IBA	Theirs is a simple and heart warming story , full of mirth that should charm all but the most cynical	positive	
VMASK	Theirs is a simple and heart warming story , full of mirth that should charm all but the most cynical	positive	
SCC	Theirs is a simple and heart warming story , full of mirth that should charm all but the most cynical	positive	
L2X	If you've ever had a mad week-end out with your mates then you'll appreciate this film . excellent fun and a laugh a minute .	positive	IMDB
IBA	If you've ever had a mad week-end out with your mates then you'll appreciate this film . excellent fun and a laugh a minute .	positive	
VMASK	If you've ever had a mad week-end out with your mates then you'll appreciate this film . excellent fun and a laugh a minute .	positive	
SCC	If you've ever had a mad week-end out with your mates then you'll appreciate this film . excellent fun and a laugh a minute .	positive	
L2X	Very bad purchase experience i bought a shirt with a hole covered in the rolled up sleeves but they denied my request to return it i am so angery at this and will never shop their chothes anymore	negative	YELP
IBA	Very bad purchase experience i bought a shirt with a hole covered in the rolled up sleeves but they denied my request to return it i am so angery at this and will never shop their chothes anymore	negative	
VMASK	Very bad purchase experience i bought a shirt with a hole covered in the rolled up sleeves but they denied my request to return it i am so angery at this and will never shop their chothes anymore	negative	
SCC	Very bad purchase experience i bought a shirt with a hole covered in the rolled up sleeves but they denied my request to return it i am so angery at this and will never shop their chothes anymore	negative	

Figure 5: **Qualitative evaluation.** The most important four words are highlighted and the color saturation indicates the word attribution.

Method		Classification Acc	Post-hoc Acc
BERT		93.79	–
SCC	w/o contrastive coding loss	93.96	74.03
	w/o sparse coding loss	94.13	82.27
	w/o background contrast	94.07	84.46
	Full (Ours)	**94.23**	**85.83**

Table 5: **Ablation study.** We select top 4 words for calculating post-hoc accuracy. All loss terms play indispensable roles in SCC.

5 Conclusion

In this paper, we show a feasible solution to solve two less-touched problems in XAI, *i.e.*, how to develop a model-agnostic method to explain the black-box and utilize the explanations to improve the model performance and interpretability. We take text classification and natural language inference as evaluation tasks, and quantitatively and qualitatively show the superiority of our method in terms of interpretability and classification metrics. In the future, we plan to explore the potential of our framework in other applications like medical diagnosis and extend our idea to other data domains like images.

Limitations

The motivation of this work is to highlight another important perspective of explainable AI, *i.e.,* increasing the trustworthiness and performance of black neural network models in decision making. However, we need to retrain the whole network again for improving the black-boxes, which might consume a lot of energy and cause massive CO2 emissions. In addition, there is no need to hide that this paper only considers the word-level explainability (*i.e.*, feature attributions), and it is unclear how to extend this idea to other explainability due to the diversity and rapid development of XAI.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2020YFB1406702; in part by NSFC under Grant U21B2040, U19A2078, and U19A2081; in part by the Sichuan Science and Technology Planning Project under Grant 2022YFQ0014.

References

David Alvarez-Melis and Tommi S Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*.

Hanjie Chen, Song Feng, Jatin Ganhotra, Hui Wan, Chulaka Gunasekara, Sachindra Joshi, and Yangfeng Ji. 2021. Explaining neural network predictions on sentence pairs via learning word-group masks. *NAACL*, pages 3917–3930.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *EMNLP*.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020a. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, pages 5578–5593.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *ICML*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Nicola De Cao, Michael Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. *arXiv:2004.14992*.

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

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2019. Learning explainable models using attribution priors. *arXiv:1902.06787*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *ICML*, pages 2454–2463. PMLR.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted multi-view classification. In *ICLR*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *NeurIPS*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI*, volume 32.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING*.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *CVPR*, pages 11174–11183.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv:1711.05101*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4768–4777.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL*, pages 1069–1078.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *SIGKDD*.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *ICML*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *ICLR*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *ACL*, pages 3418–3428.

Peng Xi, Li Yunfan, Tsang Ivor, Zhu Hongyuan, Lv Jiancheng, and Tianyi Zhou Joey. 2021. Xai beyond classification: Interpretable neural clustering. *JMLR*.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jian Cheng Lv, and Xi Peng. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*.