SEAT: Stable and Explainable Attention

Lijie Hu[*,1], **Yixin Liu**[*,2], **Ninghao Liu**[3], **Mengdi Huai**[4], **Lichao Sun**[2], **Di Wang**[1,5,6]

[1] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[2] Lehigh University, Bethlehem, Pennsylvania, USA
[3] University of Georgia, Athens, Georgia, USA
[4] Iowa State University, Ames, Iowa, USA
[5] Computational Bioscience Research Center
[6] SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence
{lijie.hu, di.wang}@kaust.edu.sa, {yila22, lis221}@lehigh.edu, ninghao.liu@uga.edu, mdhuai@iastate.edu

Abstract

Attention mechanism has become a standard fixture in many state-of-the-art natural language processing (NLP) models, not only due to its outstanding performance, but also because it provides plausible innate explanations for neural architectures. However, recent studies show that attention is unstable against randomness and perturbations during training or testing, such as random seeds and slight perturbation of embeddings, which impedes it from being a faithful explanation tool. Thus, a natural question is whether we can find an alternative to vanilla attention, which is more stable and could keep the key characteristics of the explanation. In this paper, we provide a rigorous definition of such an attention method named SEAT (**S**table and **E**xplainable **AT**tention). Specifically, SEAT has the following three properties: (1) Its prediction distribution is close to the prediction of the vanilla attention; (2) Its top-k indices largely overlap with those of the vanilla attention; (3) It is robust w.r.t perturbations, i.e., any slight perturbation on SEAT will not change the attention and prediction distribution too much, which implicitly indicates that it is stable to randomness and perturbations. Furthermore, we propose an optimization method for obtaining SEAT, which could be considered as revising the vanilla attention. Finally, through intensive experiments on various datasets, we compare our SEAT with other baseline methods using RNN, BiLSTM and BERT architectures, with different evaluation metrics on model interpretation, stability and accuracy. Results show that, besides preserving the original explainability and model performance, SEAT is more stable against input perturbations and training randomness, which indicates it is a more faithful explanation.

Introduction

As deep neural networks have demonstrated great success in various natural language processing (NLP) tasks (Otter, Medina, and Kalita 2021), to establish trust further, how to interpret these deep models are receiving increasing interest. Recently, a number of interpretation techniques have been developed to understand the decision of deep NLP models (Ribeiro, Singh, and Guestrin 2016; Vaswani et al. 2017; Dong et al. 2019). Among them, *attention mechanism* has become a near-ubiquitous component of modern NLP models. Different from post-hoc interpretation (Du, Liu, and Hu 2020), attention weights are often regarded as providing the "inner workings" of models (Choi et al. 2016; Martins and Astudillo 2016; Lei 2017). For instance, each entry of an attention vector could point us to irrelevant information discarded by the neural network or to relevant elements (tokens) of the input source that have been factored in (Galassi, Lippi, and Torroni 2020).

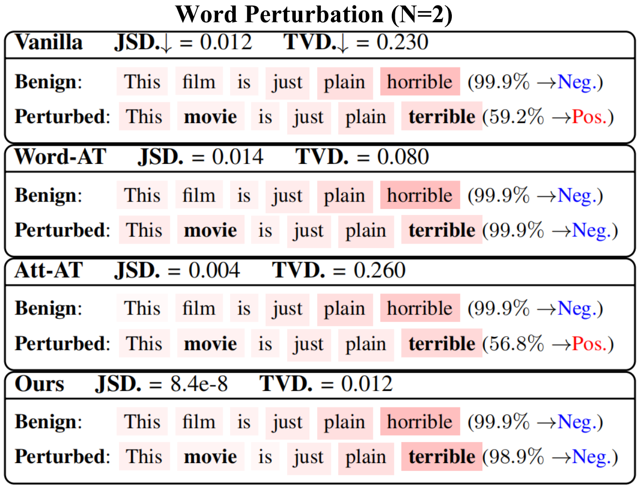

Word Perturbation (N=2)
Vanilla JSD.↓ = 0.012 TVD.↓ = 0.230
Benign: This film is just plain horrible (99.9% →Neg.)
Perturbed: This movie is just plain terrible (59.2% →Pos.)
Word-AT JSD. = 0.014 TVD. = 0.080
Benign: This film is just plain horrible (99.9% →Neg.)
Perturbed: This movie is just plain terrible (99.9% →Neg.)
Att-AT JSD. = 0.004 TVD. = 0.260
Benign: This film is just plain horrible (99.9% →Neg.)
Perturbed: This movie is just plain terrible (56.8% →Pos.)
Ours JSD. = 8.4e-8 TVD. = 0.012
Benign: This film is just plain horrible (99.9% →Neg.)
Perturbed: This movie is just plain terrible (98.9% →Neg.)

Figure 1: An example demonstrates the stability of prediction and attention heat map trained with different methods under word perturbation in the sentiment classification task. There are four methods: Vanilla attention, Word-AT, Att-AT and our SEAT. JSD and TVD are divergences to measure the stability of explainability and prediction distribution (see Experiments section for details). We use the closest synonyms to replace the original word in sentence as word perturbation. N denotes the number of replaced words. We can see explanation (heat map) and prediction are changed in other methods.

Despite its wide adoption, attention mechanism has been questioned as being a **faithful** interpretation. Specifically, Wiegreffe and Pinter (2019) show that attention is unstable, as different model initialization leads to different attention distributions given the same input (see Fig. 3 in Appendix

[*]These authors contributed equally.

for an example). Besides, attention is also fragile to input perturbations during inference. For example, in Fig. 1, we can observe that, after replacing a word with its synonym, the attention changes significantly and may also give wrong predictions. In addition, perturbing word embeddings will also affect the prediction distribution and the explainability of attention (see details in Appendix Fig. 4). Actually, instability has been identified as a common issue of interpretation methods in deep models. Generally, an unstable interpretation makes it easy to be influenced by noises in data, thus impeding users from understanding the inherent rationale behind model predictions. Moreover, instability reduces the reliability of interpretation as a diagnosis tool of models, where small carefully-crafted perturbation on input could dramatically change the interpretation result (Ghorbani, Abid, and Zou 2019; Dombrowski et al. 2019; Yeh et al. 2019). Thus, stability now becomes an important aspect of faithful interpretation. Based on the above facts, a natural question is can we make the attention mechanism more faithful by improving its stability, while keeping the most important explanation and prediction characteristics?

To tackle the problem, we first need to give a rigorous definition of such a "stable attention". Intuitively, a stable attention should have the following three properties: (1) It should produce a similar prediction as the vanilla attention to preserve model utility; (2) The top-k indices of the stable attention and vanilla attention should largely overlap with each other, so that the stable attention inherits the interpretability of vanilla attention; (3) It should be robust to the randomness in training or input perturbations during testing. Based on the above criteria, in this paper, we present a formal definition of SEAT (**S**table and **E**xplainable **At**tention). Specifically, our contributions can be summarized as follows.

- We provide a rigorous mathematical definition of SEAT. Specifically, to keep property (1), SEAT ensures the loss between its prediction distribution (vector) and the prediction distribution (vector) based on attention is sufficiently small. For property (2), we ensure the top-k indices overlap between SEAT and vanilla attention are large enough. For property (3), SEAT guarantees some perturbations on it will not change the prediction distribution too much, which implicitly ensures it is robust to randomness and perturbations during training and testing.
- In the second part of the paper, we propose a method to find a SEAT. Specifically, we present a min-max stochastic optimization problem whose objective function involves three terms, which correspond to the above three properties. However, the main difficulty is that the term induced by property (2) is non-differentiable, which impedes us from using gradient descent based methods. To address this issue, we also propose a surrogate loss function of top-k overlap function.
- Finally, we conduct intensive experiments on four benchmark datasets using RNN, BiLSTM and BERT to verify the above three properties of the SEAT. Particularly, we first demonstrate our SEAT is more stable than other baselines via three different perturbations or randomness: random seeds, embedding vector perturbation, and input token perturbation. We also use three recent evaluation metrics on model interpretability in evaluation. Results reveal our SEAT is a more faithful interpretation. Besides, we compare the F1 score of our SEAT and other baselines, showing that there is almost no accuracy degradation for SEAT compared with vanilla attention.

Related Work

Stability and robustness in attention. There exists some work studying or improving either the stability or the robustness of attention from the explanation perspective. Recently, Kitada and Iyatomi (2021) propose a method to improve the robustness to perturbation of embedding vector for attention. Specifically, they adopt adversarial training during the training process. However, in their method, they do not consider the similarity and closeness between their new attention and the original ones, which means their robust attention loses the prediction performance and explainability of the original attention. Equivalently, while their adversarial training may improve the robustness of attention, it cannot be ensured to be explainable due to the ignorance of the relationship with vanilla attention. Sato et al. (2018) study using adversarial training to improve the robustness and interpretation of the text. However, their work is applied to input embedding space, whose computational cost is high. Moreover, their method still can guarantee the closeness to attention on neither prediction nor explanation, and their method cannot ensure robustness against other randomness such as random seeds. (Mohankumar et al. 2020) explores modifying the LSTM cell with diversity-driven training to enhance the explainability and transparency of attention modules. However, it does not consider the robustness of attention, which makes their method far from a faithful method.

Stability in explanation techniques. Besides attention, there are works on studying stable interpretation. For example, Yeh et al. (2019) theoretically analyzes the stability of post-hoc interpretation approaches and proposes using smoothing to improve interpretation stability. Jacovi and Goldberg (2020) discuss high-level directions of designing reliable interpretation. However, these techniques are designed for post-hoc interpretation, which cannot be directly applied to attention mechanisms. Recently, Yin et al. (2022) introduced two metrics to measure interpretability via sensitivity and stability. They also introduce methods to better test the validity of their evaluation metrics by designing an iterative gradient descent algorithm to get a counterfactual interpretation. But they do not consider how to improve the faithfulness of explainable models. Thus, it is incomparable to our work. And we will use these evaluation metrics in experiments.

Stable and Explainable Attention

Vanilla Attention

We first give a brief introduction to the attention mechanism (Vaswani et al. 2017). Here we follow the notations in (Jain and Wallace 2019). Let $x \in \mathbb{R}^{s\times|V|}$ denote the model input, composed of one-hot encoded words at each position. There is an embedding matrix E with dimension d. After passing through the embedding matrix E, the words have more dense token representations as $x_e \in \mathbb{R}^{s \times d}$.

There is an encoder (**Enc**)-decoder (**Dec**) layer. For the encoder, it takes the embedded tokens in order and produces s number of m-dimensional hidden states after the **Enc** procedure, i.e., $\boldsymbol{h}(x) = \textbf{Enc}(x_e) \in \mathbb{R}^{s \times m}$ with $h_t(x)$ as the word representation for the word at position t in x.

A similarity function ϕ maps $\boldsymbol{h}(x)$ and query $\boldsymbol{Q} \in \mathbb{R}^m$ to scalar scores, and the attention weights are induced by $\boldsymbol{w}(x) = \text{softmax}(\phi(\boldsymbol{h}(x), \boldsymbol{Q})) \in \mathbb{R}^s$. Here we consider two common types of similarity functions: *Additive* $\phi(\boldsymbol{h}(x), \boldsymbol{Q}) = \boldsymbol{v}^T \tanh(\boldsymbol{W_1} \boldsymbol{h}(x) + \boldsymbol{W_2} \boldsymbol{Q})$ (Bahdanau, Cho, and Bengio 2015) and *Scaled Dot-Product* $\phi(\boldsymbol{h}, \boldsymbol{Q}) = \frac{\boldsymbol{h}(x)Q}{\sqrt{m}}$ (Vaswani et al. 2017), where $\boldsymbol{v}$, $\boldsymbol{W_1}$, and $\boldsymbol{W_2}$ are model parameters.

Based on $\boldsymbol{w}$, the model makes predictions after the **Dec** procedure, i.e., $y(x, \boldsymbol{w}) = \sigma(\theta \cdot h_w) \in \mathbb{R}^{|\mathcal{Y}|}$, where $h_w = \sum_{t=1}^{s} \boldsymbol{w}_t(x) \cdot h_t(x)$, σ is an output activation function, θ is a parameter and $|\mathcal{Y}|$ denotes the number of classes.

Stable and Explainable Attention

Motivation. As we mentioned previously, our goal is to find some "stable attention" that keeps the **performance** and **explainability** of attention, while it is more **robust against some randomness and perturbations during training and testing**. Before showing how to find a stable attention, we propose the following three properties for it.

1. A stable attention should preserve the outstanding performance of attention models, i.e., we hope the prediction distribution (vector) based on the stable attention is similar to the distribution (vector) based on vanilla attention for any input x. Mathematically, we can use different divergence metrics to measure the similarity.
2. A stable attention should keep the explainability of vanilla attention. The rank of each entry in the attention vector determines the importance of its associated word token. To keep the order of leading entries, mathematically, we can use the overlaps of top-k indices between stable attention and vanilla attention to measure their similarity on explainability, where k is a hyperparameter.
3. Such a new attention should be stable. Please note that compared with the robustness to adversarial attacks, here our stability definition is more general, i.e., it should be robust against any randomness and perturbations during training and testing. These include random seeds in training, and perturbations on embedding vectors or input tokens during testing. Thus, unlike adversarial training, it is difficult to model the robustness to various randomness or perturbations directly. To resolve the issue, as we mentioned, those randomness and perturbations will cause attention changes dramatically, which could be thought as some noise added to attention will change it significantly. Thus, if the "stable attention" is resilient to any perturbations, then this can indicate that such vector is robust to any randomness and perturbations implicitly. In total, mathematically we can model such robustness via the resilience against perturbations of "stable attention".

Based on the above motivation, in the following we formally give the definition of "stable attention" called Stable and Explainable Attention (SEAT) denoted as $\tilde{\boldsymbol{w}}$. Since we need to use the overlaps of top-k indices to measure the similarity on explainability with attention. We first provide its formal definition.

Definition 1 (Top-k overlaps)**.** *For a vector $x \in \mathbb{R}^d$, we define the set of top-k component $T_k(\cdot)$ as follow,*

$$T_k(x) = \{i : i \in [d] \text{ and } \{|\{x_j \geq x_i : j \in [d]\}| \leq k\}\}.$$

And for two vectors x, x', the top-k overlap function $V_k(x, x')$ is defined by the overlap ratio between the top-k components of two vectors, i.e., $V_k(x, x') = \frac{1}{k}|T_k(x) \cap T_k(x')|$.

Note that in attention, $\boldsymbol{w}$ could be seen as a function of x. Thus, $\tilde{\boldsymbol{w}}$ can also be seen as a function of x. Moreover, since we care about replacing the attention vector, we still follow the original model except for the procedure to produce the vector $\tilde{\boldsymbol{w}}(x)$. We define SEAT as follows.

Definition 2 (**Stable and Explainable Attention**)**.** *We call a vector $\tilde{\boldsymbol{w}}$ is $(D_1, D_2, R, \alpha, \beta, \gamma, V_k)$-Stable and Explainable Attention (SEAT) w.r.t. the vanilla attention $\boldsymbol{w}$ if it satisfies the following properties for any x:*

- *(Closeness of Prediction) $D_1(y(x, \tilde{\boldsymbol{w}}), y(x, \boldsymbol{w})) \leq \gamma$ for $\gamma \geq 0$, where D_1 is some divergence function, $y(x, \tilde{\boldsymbol{w}}) = \sigma(\theta \cdot h_{\tilde{w}}) \in \mathbb{R}^{|\mathcal{Y}|}$ and $y(x, \boldsymbol{w}) = \sigma(\theta \cdot h_w) \in \mathbb{R}^{|\mathcal{Y}|}$;*
- *(Similarity of Explainability) $V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x)) \geq \beta$ for some $1 \geq \beta \geq 0$;*
- *(Stability) $D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta})) \leq \alpha$ for all $\|\boldsymbol{\delta}\| \leq R$, where D_2 is some divergence function, $\|\cdot\|$ is a norm and $R \geq 0$.*

Note that in the previous definition, there are several parameters. Specifically, γ constrains the closeness between the prediction distribution based on $\tilde{\boldsymbol{w}}$ and the prediction distribution based on vanilla attention, where $\tilde{\boldsymbol{w}} = \boldsymbol{w}$ if $\gamma = 0$. Therefore, we hope γ to be as small as possible. The second condition ensures $\tilde{\boldsymbol{w}}$ has similar explainability with vanilla attention. There are two parameters, k and β. The value of k could be decided by prior knowledge, where we hope the top-k attention indices will play the most important role to make the prediction. β measures how much explainability does $\tilde{\boldsymbol{w}}$ inherit from vanilla attention. When $\beta = 1$, it means the top-k of the entries in $\tilde{\boldsymbol{w}}(x)$ is the same as that in vanilla attention. Thus, β should close to 1. The third condition involves two parameters R and α, which correspond to the robust region and the level of stability, respectively. Ideally, if $\tilde{\boldsymbol{w}}$ satisfies this condition with $R = \infty$ and $\alpha = 0$, then $\tilde{\boldsymbol{w}}$ will be stable w.r.t any randomness or perturbations. Thus, in practice we wish R to be as large as possible and α to be sufficiently small. Based on the above discussions, we can see Definition 2 is consistent with our intuitions about "stable attention".

Finding a SEAT

In the last section, we presented a rigorous definition of stable and explainable attention. To find such a SEAT, we propose to formulate a min-max optimization problem that involves the three conditions in Definition 2. Specifically, the formulated optimization problem takes the first condition (closeness of prediction) as the objective, and subjects to the other two conditions as constraints. Specifically, we first have

$$\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x D_1(y(x, \tilde{\boldsymbol{w}}), y(x, \boldsymbol{w})). \tag{1}$$

Equation (1) is the basic optimization goal, where we want to get similar output prediction with vanilla attention for any input x. If there is no further constraint, then we can see the minimizer of Equation (1) is just the vanilla attention w. Thus, we include constraints for this objective function:

$$\forall x \text{ s.t.} \max_{||\delta|| \leq R} D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta})) \leq \alpha, \tag{2}$$

$$V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x)) \geq \beta. \tag{3}$$

Equation (2) is the constraint of stability and Equation (3) corresponds to the condition of similarity of explainability. Combining Equations (1)-(3) and dealing with the constraints using regularization, we can get the following objective:

$$\begin{aligned}\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x[&D_1(y(x, \tilde{\boldsymbol{w}}), y(x, \boldsymbol{w})) + \lambda_2(\beta - V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x))) \\ &+ \lambda_1(\max_{||\boldsymbol{\delta}|| \leq R} D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta})) - \alpha)],\end{aligned} \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters.

From now on, we convert the problem of finding a vector that satisfies the three conditions in Definition 2 to a min-max stochastic optimization problem, where the overall objective is based on the closeness of prediction condition with constraints on stability and top-k overlap.

Next we discuss how to solve the above min-max optimization problem. In general, we can use the stochastic gradient descent based methods to get the solution of outer minimization, and use PSGD (Projected Stochastic Gradient Descent) to solve the inner maximization. However, the main difficulty is that the top-k overlap function $V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x))$ is non-differentiable, which impedes us from using gradient descent. Thus, we need to consider a surrogate loss of $-V_k(\tilde{\boldsymbol{w}}(x), \boldsymbol{w}(x))$ with details as below.

Projected gradient descent to solve $\boldsymbol{\delta}$. Motivated by (Madry et al. 2018), we can interpret perturbation as the attack to $\tilde{w}$ via maximizing δ. Then, δ can be updated by the following procedure in the p-th iteration.

$$\boldsymbol{\delta_p} = \boldsymbol{\delta^*_{p-1}} + \alpha_p \frac{1}{|B_p|} \sum_{x \in B_p} \nabla D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta^*_{p-1}}));$$

$$\boldsymbol{\delta^*_p} = \arg\min_{||\boldsymbol{\delta}|| \leq R} ||\boldsymbol{\delta} - \boldsymbol{\delta_p}||,$$

where α_p is a parameter of step size for PGD, B_p is a batch and $|B_p|$ is the batch size. Using this method, we can derive the optimal δ^* in the t-th iteration for the inner optimization. Specifically, we find a δ as the maximum tolerant of perturbation w.r.t $\tilde{\boldsymbol{w}}$ in the t-th iteration.

Algorithm 1 Finding a SEAT

```
Initialize $\tilde{\boldsymbol{w}}_0$.
for $t = 1, 2, \cdots, T$ do
    Initialize $\boldsymbol{\delta_0}$.
    for $p = 1, 2, \cdots, P$ do
        $\tilde{\boldsymbol{w}}'_{\boldsymbol{t-1}} = \tilde{\boldsymbol{w}}_{\boldsymbol{t-1}} + \boldsymbol{\delta_{p-1}}$
        Update $\boldsymbol{\delta}$ using PGD, where $B_p$ is a batch
            $\boldsymbol{\delta_p} = \boldsymbol{\delta_{p-1}}$
            $+ \alpha_p \frac{1}{|B_p|} \sum_{x \in B_p} \nabla D_2(y(x, \tilde{\boldsymbol{w}}_{t-1}), y(x, \tilde{\boldsymbol{w}}'_{\boldsymbol{t-1}}))$.
        $\boldsymbol{\delta^*_p} = \arg\min_{||\boldsymbol{\delta}|| \leq R} ||\boldsymbol{\delta} - \boldsymbol{\delta_p}||$.
    end for
    Update $\tilde{\boldsymbol{w}}$ using Stochastic Gradient Descent, where
    $B_t$ is a batch
        $\tilde{\boldsymbol{w}}_{\boldsymbol{t}} = \tilde{\boldsymbol{w}}_{\boldsymbol{t-1}} - \frac{\eta_t}{|B_t|} \sum_{x \in B_t} [\nabla D_1(y(x, \tilde{\boldsymbol{w}}_{\boldsymbol{t-1}}), y(x, \boldsymbol{w}))$
        $- \lambda_1 \nabla D_2(y(x, \tilde{\boldsymbol{w}}_{\boldsymbol{t-1}}), y(\tilde{\boldsymbol{w}}_{\boldsymbol{t-1}} + \boldsymbol{\delta^*_P}))$
        $- \lambda_2 \nabla \mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}}_{t-1})]$.
end for
Return: $\tilde{\boldsymbol{w}}^* = \tilde{\boldsymbol{w}}_{\boldsymbol{T}}$.
```

Top-k overlap surrogate loss. Now we seek to design a surrogate loss $\mathcal{L}_{Topk}(\tilde{\boldsymbol{w}}, \boldsymbol{w})$ for $-V_k(\tilde{\boldsymbol{w}}, \boldsymbol{w})$ which can be used in training. To achieve this goal, one possible naive surrogate objective might be some distance (such as ℓ_1-norm) between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$, e.g., $L(\tilde{\boldsymbol{w}}) = ||\tilde{\boldsymbol{w}} - \boldsymbol{w}||_1$. Such a surrogate objective seems like it could ensure the top-k overlap when we obtain the optimal or near-optimal solution (i.e., $w = \arg\min L(\tilde{\boldsymbol{w}})$ and $w \in \arg\min -V_k(\tilde{\boldsymbol{w}}, \boldsymbol{w})$). However, it lacks consideration of the top-k information, which makes it a loose surrogate loss. Since we only need to ensure high top-k indices overlaps between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$, one improved method is minimizing the distance between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ constrained on the top-k entries only instead of the whole vectors, i.e., $||\boldsymbol{w}_{S^k_w} - \tilde{\boldsymbol{w}}_{S^k_w}||_1$, where $\boldsymbol{w}_{S^k_w}, \tilde{\boldsymbol{w}}_{S^k_w} \in \mathbb{R}^k$ is the vector $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ constrained on the indices set S^k_w respectively and S^k_w is the top-k indices set of $\boldsymbol{w}$. Since there are two top-k indices sets, one is for $\tilde{\boldsymbol{w}}$ and the other one is for $\boldsymbol{w}$, here we need to use both of them to involve the top-k indices formation for both vectors. Thus, based on our above idea, our surrogate can be written as follows,

$$\mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \frac{1}{2k}(||\boldsymbol{w}_{S^k_w} - \tilde{\boldsymbol{w}}_{S^k_w}||_1 + ||\tilde{\boldsymbol{w}}_{S^k_{\tilde{w}}} - \boldsymbol{w}_{S^k_{\tilde{w}}}||_1). \tag{5}$$

Note that besides the ℓ_1-norm, we can use other norms. However, in practice we find ℓ_1-norm achieves the best performance. Thus, throughout the paper we only use ℓ_1-norm.

Final objective function and algorithm Based on the above discussion, we can derive the following overall objective function:

$$\begin{aligned}\min_{\tilde{\boldsymbol{w}}} \mathbb{E}_x[&D_1(y(x, \tilde{\boldsymbol{w}}), y(x, \boldsymbol{w})) + \lambda_2 \mathcal{L}_{Topk}(\boldsymbol{w}(x), \tilde{\boldsymbol{w}}(x)) \\ &+ \lambda_1 \max_{||\boldsymbol{\delta}|| \leq R} D_2(y(x, \tilde{\boldsymbol{w}}), y(x, \tilde{\boldsymbol{w}} + \boldsymbol{\delta}))],\end{aligned} \tag{6}$$

where $\mathcal{L}_{Topk}(\boldsymbol{w}, \tilde{\boldsymbol{w}})$ is defined in (5). Based on the previous idea, we propose Algorithm 1 to solve (6).

Model	Method	Emotion			SST			Hate			RottenT		
		JSD↓	TVD↓	F1↑	JSD	TVD	F1	JSD	TVD	F1	JSD	TVD	F1
RNN	Vanilla	0.002	20.145	0.663	0.019	19.566	0.811	0.009	15.576	0.553	0.008	19.139	0.763
	Word-AT	0.028	1.824	0.627	0.016	1.130	0.798	0.026	1.170	0.527	0.037	1.381	0.741
	Word-iAT	0.042	2.691	0.653	0.023	1.277	**0.815**	0.022	1.049	0.523	0.054	1.336	0.766
	Attention-RP	0.025	3.276	0.671	0.028	2.042	0.792	0.025	2.672	0.554	0.009	3.691	**0.770**
	Attention-AT	0.055	2.716	0.665	0.047	2.394	0.782	0.031	2.210	0.528	0.068	4.234	0.755
	Attention-iAT	0.017	3.654	0.645	0.048	2.653	0.746	0.039	2.264	0.533	0.054	1.594	0.753
	SEAT(**Ours**)	**3.81E-08**	**1.750**	**0.672**	**2.75E-07**	**1.099**	0.813	**1.79E-09**	**0.908**	**0.579**	**6.46E-07**	**1.178**	0.763
BiLSTM	Vanilla	0.002	23.447	0.612	0.027	18.640	**0.809**	0.060	15.633	0.524	0.009	20.125	0.764
	Word-AT	0.050	1.927	0.662	0.020	0.810	0.798	0.084	1.537	0.538	0.031	1.071	0.757
	Word-iAT	0.058	1.139	0.640	0.034	1.037	0.802	0.091	1.590	0.530	0.045	1.218	0.765
	Attention-RP	0.031	1.326	0.642	0.034	1.267	0.772	0.052	1.299	0.522	0.066	1.412	0.764
	Attention-AT	0.076	1.541	**0.672**	0.028	1.661	0.779	0.057	1.504	0.523	0.079	2.044	0.766
	Attention-iAT	0.033	1.267	0.651	0.034	1.528	0.801	0.062	2.256	0.525	0.076	1.751	**0.777**
	SEAT(**Ours**)	**1.23E-08**	**0.736**	0.670	**1.80E-08**	**0.777**	0.802	**8.49E-09**	**1.030**	**0.543**	**2.57E-08**	**0.885**	0.771
BERT	Vanilla	0.024	2.127	**0.721**	0.005	2.605	0.912	0.036	1.771	0.493	0.010	2.500	**0.845**
	Word-AT	0.085	0.060	0.694	0.267	0.055	0.900	0.170	0.043	**0.554**	0.510	0.036	0.826
	Word-iAT	0.584	0.029	0.694	0.241	0.054	0.895	0.166	0.049	0.496	0.480	0.049	0.844
	Attention-RP	0.035	0.232	0.657	0.086	0.127	0.893	0.079	0.277	**0.554**	0.078	0.142	0.817
	Attention-AT	0.067	0.119	0.707	0.005	0.156	0.907	0.031	0.230	0.510	0.041	0.189	0.818
	Attention-iAT	0.096	0.222	0.684	0.129	0.200	**0.915**	0.074	0.271	0.512	0.108	0.183	0.831
	SEAT(**Ours**)	**4.70E-07**	**0.002**	0.713	**2.77E-07**	**0.036**	0.907	**2.42E-06**	**0.042**	0.545	**1.68E-08**	**0.003**	0.841

Table 1: Results of evaluating embedding perturbation stability of (modified) attentions under three metrics. The perturbation radius is set as δ_x=1e-3. ↑ means a higher value under this metric indicates better results, and ↓ means the opposite.

Model	Method	IMDB			SST			Hate			RottenT		
		Comp.↑	Suff.↓	Sens.↓	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.	Comp.	Suff.	Sens.
RNN	Vanilla	0.004	0.007	0.131	5.744	7.02E-04	0.090	0.009	0.066	0.138	4.483	0.026	0.090
	Word-AT	2.899	0.018	0.121	5.280	**0.000**	0.081	2.408	0.058	0.137	2.512	0.031	0.093
	Word-iAT	2.060	0.010	0.122	5.452	9.35E-06	0.121	6.069	0.075	0.136	4.534	0.058	0.087
	Attention-RP	0.099	0.073	0.130	6.001	2.87E-04	0.085	3.585	0.080	0.133	4.637	0.052	0.094
	Attention-AT	0.026	0.052	0.131	3.118	**0.000**	0.088	2.493	0.372	0.134	4.713	0.049	0.096
	Attention-iAT	1.994	6.74E-04	0.117	4.788	**0.000**	0.091	5.351	0.126	0.133	2.435	0.039	0.086
	SEAT(**Ours**)	**3.281**	**1.04E-05**	**0.106**	**6.016**	**0.000**	**0.076**	**6.558**	**2.75E-04**	**0.129**	**4.796**	**0.025**	**0.084**
BiLSTM	Vanilla	0.474	0.002	0.129	5.182	0.255	0.086	4.203	0.112	0.142	2.966	0.088	0.092
	Word-AT	1.449	0.015	0.121	5.167	8.44E-04	0.096	5.438	0.207	0.153	3.388	0.062	0.080
	Word-iAT	0.619	0.005	0.127	5.259	3.81E-05	0.087	4.568	0.320	0.145	3.339	0.078	0.082
	Attention-RP	0.561	0.002	0.127	2.865	0.007	0.101	2.248	0.199	0.148	4.073	0.200	0.082
	Attention-AT	1.294	0.051	0.111	2.129	0.004	0.098	3.220	0.065	**0.140**	4.925	0.216	0.082
	Attention-iAT	0.555	0.002	0.127	5.176	0.004	0.083	4.092	0.290	0.141	2.431	0.377	0.083
	SEAT(**Ours**)	**1.502**	**6.41E-04**	**0.098**	**5.435**	**4.37E-06**	**0.076**	**6.240**	**0.025**	**0.140**	**4.941**	**0.046**	**0.077**
BERT	Vanilla	5.07E-04	0.008	0.013	0.003	0.310	0.009	3.20E-04	0.401	0.016	0.001	0.092	0.010
	Word-AT	5.01E-05	0.005	0.016	1.26E-05	4.47E-04	0.005	4.01E-04	0.014	0.017	0.005	0.043	0.016
	Word-iAT	5.47E-04	0.007	0.017	1.51E-05	5.67E-04	0.004	0.003	0.045	0.017	2.93E-04	0.010	0.009
	Attention-RP	0.085	0.086	0.014	0.002	0.010	0.011	0.035	0.034	0.016	0.017	0.003	0.010
	Attention-AT	4.65E-05	0.338	0.016	4.50E-05	0.441	0.006	**0.004**	0.007	0.016	6.26E-04	0.032	0.011
	Attention-iAT	0.002	0.164	0.015	1.19E-04	9.07E-04	0.007	0.001	0.151	0.017	0.003	0.025	0.010
	SEAT(**Ours**)	**0.160**	**0.002**	**0.012**	**0.497**	**0.000**	**0.004**	**0.153**	**0.006**	**0.015**	**0.040**	**0.002**	**0.008**

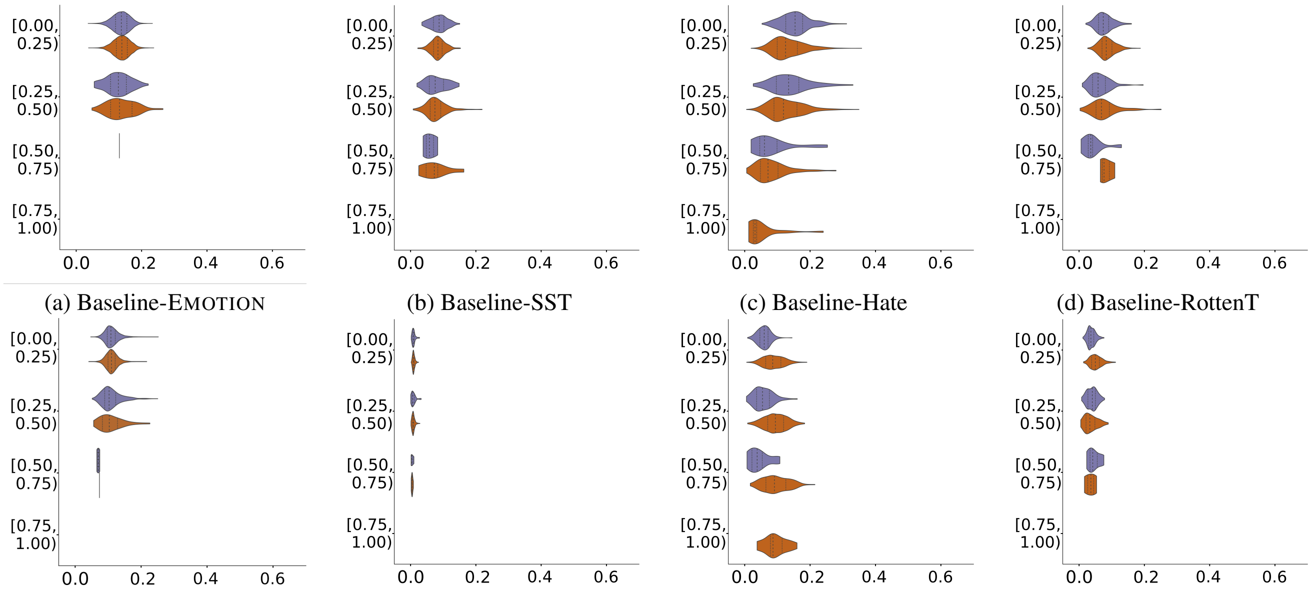
Table 2: Results on evaluating the interpretability of different methods.

Experiments

In our experiments, we conduct extensive experiments to show the performance of our SEAT compared to six baseline methods. We provide a brief introduction to the experimental setup next. **More details are in Appendix.**

Setup. First, we demonstrate the stability of SEAT under three different randomness and perturbations: (1) random seeds during training; (2) embedding perturbation during testing, and (3) word perturbation during testing. For each method, we use Jensen-Shannon Divergence (JSD) between its attention with on perturbation and its attention under perturbation to evaluate the stability of explainability of the learned attention. And the Total Variation Distance (TVD) between the prediction distribution with no perturbation and prediction under perturbation is used to measure prediction stability.

Next, in order to show the explainability of SEAT, we use



(a) Baseline-EMOTION (b) Baseline-SST (c) Baseline-Hate (d) Baseline-RottenT

(e) Ours-EMOTION (f) Ours-SST (g) Ours-Hate (h) Ours-RottenT

Figure 2: Comparison of stability against random seeds for vanilla attention and SEAT. Densities of maximum JS divergences (x-axis) as a function of the max attention (y-axis) in each instance between its base model and models initialized on different random seeds. In each max-attention bin, top (blue) is a negative-label instance, and bottom (red) is a positive-label instance.

Method	Emotion			SST			Hate			RottenT		
	JSD↓	TVD↓	F1↑	JSD	TVD	F1	JSD	TVD	F1	JSD	TVD	F1
Vanilla	0.628	2.847	**0.721**	0.315	3.655	0.912	0.491	2.004	0.493	0.585	3.464	0.845
Word-AT	0.004	0.022	0.694	0.175	0.065	0.910	0.111	0.058	0.546	0.473	0.056	0.836
Word-iAT	0.456	0.059	0.658	0.213	0.046	0.912	0.331	0.044	0.501	0.488	0.048	**0.852**
Attention-RP	0.039	0.235	0.657	0.089	0.128	0.893	0.085	0.278	0.554	0.078	0.143	0.817
Attention-AT	0.082	0.003	0.707	0.006	0.157	0.907	0.035	0.230	0.510	0.049	0.193	0.818
Attention-iAT	0.126	0.228	0.684	0.147	0.204	**0.915**	0.081	0.271	0.512	0.136	0.187	0.831
SEAT(Ours)	**1.72E-09**	**0.001**	0.716	**1.55E-06**	**0.028**	0.907	**8.69E-06**	**0.037**	**0.555**	**1.10E-05**	**0.035**	0.847

Table 3: Results on stability and utility of attention model under word perturbation ($N = 1$) using BERT.

the recent evaluation metrics of model interpretation proposed by (Yin et al. 2022; DeYoung et al. 2020). Specifically, we use three explainability evaluation metrics: *comprehensiveness, sufficiency, and sensitivity*.

Thirdly, we compare the *F1 score* of our SEAT with other baselines to verify the property of closeness of prediction. Finally, we conduct an ablation study to verify the efficiency of our modules (regularizers) in the objective function (6) corresponding to each condition. In Fig. 5 of Appendix, we also test the validity of surrogate loss for the top-k overlap function by comparing the performance of our loss in equation (5) with the true top-k indices overlaps.

Model, Dataset and Baseline. Following (Jain and Wallace 2019) and (Wiegreffe and Pinter 2019), we mainly study the encoder-decoder architectures for binary classification tasks in this paper. For the encoder, we consider three kinds of networks as feature extractors: RNN, BiLSTM, and BERT. For the decoder, we apply one simple MLP followed by a tanh-attention layer (Bahdanau, Cho, and Bengio 2015) and a softmax layer (Vaswani et al. 2017). In all experiments, we use four datasets: Stanford Sentiment Treebank (SST) (Socher et al. 2013), Emotion Recognition (Emotion) (Mohammad et al. 2018), Hate (Basile et al. 2019) and Rotten Tomatoes (RottenT) (Pang and Lee 2005). And we select the Binary Cross Entropy loss as D_1 and D_2 in (6). We compare our method with Vanilla attention (Wiegreffe and Pinter 2019), Word AT (Miyato, Dai, and Goodfellow 2016), Word iAT (Sato et al. 2018), Attention RP (attention weight is trained with random perturbation), Attention AT and Attention iAT (Kitada and Iyatomi 2021).

Stability Evaluation

Random seeds. Here we compare the stability against random seeds for vanilla attention and our SEAT. Specifically, we conduct multiple model training with different random seeds and select one of them as the base model. We visualize the JS divergence of the attention weight distribution between the base model and models trained with different random seeds for different methods. We conduct experiments on several test samples and each testing sample is divided into one of four bins by its maximum attention scores within the sentence. Here we use the RNN architecture.

We can see that Fig. 2 (c) and 2 (d) have heavy tails for the baseline vanilla attention on SST and Hate datasets. The violins cover wider ranges along the x-axis. This can be interpreted as vanilla attention is unstable to random seeds. We can see from Fig. 2 (f)-(h) that while using our SEAT on SST, Hate and Rotten Tomatoes datasets, the violins are more narrow and their tails are lighter, which implies SEAT is more stable. This can be further confirmed by the fact the violins of SEAL are much closer to zero than these of vanilla attention, which means their corresponding JSD values are much smaller.

Embedding perturbation. We compare the stability of our SEAT with other baselines under embedding perturbation. In this setting, we mainly consider two metrics: JSD and TVD, which represents the explainability stability and prediction stability, respectively. Details of our setting are in Appendix. Results are shown in Tab. 1 and Tab. 8 in Appendix. We can see that SEAT outperforms other baselines with RNN, BiLSTM and BERT under JSD and TVD evaluation metrics. Especially we can see that the JSD of all our results is almost zero, which means SEAT is stable to perturbation for explanation. We can see also that the TVD for vanilla attention is large which means vanilla attention is extremely unstable to perturbation for its prediction distribution. However, the TVD of SEAT is small.

Word perturbation. We now aim to evaluate the stability of our proposed method under word perturbation. Following (Yin et al. 2022), we select BERT as our main model in this part and conduct the perturbation in the following process: first, we randomly choose N words from a given sentence and then replace them with the closest synonyms. The distance of words is computed based on gensim (Rehurek and Sojka 2011). We denote the original input and perturbed input as x and x', respectively. Then, similar to the above procedures, we can compute JSD and TVD for each method. Tab. 3 and Tab. 10 in Appendix demonstrate that SEAT achieves SOTA for both JSD and TVD in this setting. Similar to the embedding perturbation case, we can see the JSD of SEAT is much smaller than it of the vanilla attention and its value is quite close to zero in all experiments, which indicates strong explanation stability against word perturbation for SEAT.

Evaluating Interpretability and Utility

In this part, we measure the interpretability of SEAT and other baselines using comprehensiveness, sufficiency and sensitivity. Results are shown in Tab. 2 and Tab. 9 in Appendix. Our results show that SEAT outperforms other baselines on all three evaluation metrics with RNN, BiLSTM and BERT. This further confirms that enhancing stability in attention would derive a more faithful interpretation. Our SEAT improves the model interpretability.

In Tab. 1 and 3 we also compared the F1 score for different methods. We can see that while in some of the results, our method is not the best one. However, among these results, the difference between the best result and ours is quite small, which indicates that there is almost no accuracy deterioration in SEAT. Surprisingly, we can also see that SEAT is better than vanilla attention in most results and it could even achieve SOTA in some cases, such as when the model is RNN and the data is Emotion or Hate in Tab. 1.

Models	**Ablation Setting**		**Metrics**		
	$\mathcal{L}_3$	$\mathcal{L}_{Topk}$	**Suff.↓**	**TVD↓**	**F1↑**
RNN			7.02E-04	21.464	**0.814**
	✓		6.22E-04	1.966	0.804
		✓	2.22E-04	2.997	0.782
	✓	✓	**1.02E-04**	**1.275**	0.813
BiLSTM			0.255	20.398	**0.809**
	✓		0.016	1.214	0.802
		✓	0.004	1.745	0.779
	✓	✓	**4.37E-06**	**1.095**	0.801
BERT			0.310	2.617	**0.912**
	✓		0.280	0.056	0.909
		✓	0.090	0.157	0.907
	✓	✓	**0.019**	**0.028**	0.909

Table 4: Ablation study of SEAT. We evaluate the effectiveness of $\mathcal{L}_3$ and $\mathcal{L}_{Topk}$ in (6). Perturbation on the embedding space (radius $\delta_x = 0.01$) are conducted on SST.

Ablation Study

In the ablation study, we evaluate each module (regularization) in (6). Specifically, we denote $\mathcal{L}_1$ as our main loss in (1), then we consider and evaluate different combinations by deleting $\mathcal{L}_{Topk}$ or/and $\mathcal{L}_3$, where $\mathcal{L}_3$ corresponds to the third term in (6). Note that if there is no $\mathcal{L}_{Topk}$ and $\mathcal{L}_3$, then the model will be the vanilla attention. Tab. 4, Tab. 6 and 7 in Appendix show that each regularizer in our objective function is indispensable and effective. Specifically, we can see the sufficiency will decrease significantly if we add the $\mathcal{L}_{Topk}$ loss. This is due to that $\mathcal{L}_{Topk}$ enforces a large overlap on top-k indices and thus it makes SEAT inherit the explainability of vanilla attention and it makes the model more stable. Moreover, in the case where there is $\mathcal{L}_{Topk}$, adding term $\mathcal{L}_3$ could further decrease sufficiency as it improves stability.

Since $\mathcal{L}_3$ is to make the prediction distribution stable against any randomness and perturbation, from Tab. 6 we can see adding this term could decrease the TVD, which means it improves the stability. Although $\mathcal{L}_{Topk}$ can also help to decrease TVD, we can see it is weaker than $\mathcal{L}_3$. Besides stability, $\mathcal{L}_3$ also can pull back the F1 score to make the model close to vanilla attention. We can see that in the case where there only exists $\mathcal{L}_{Topk}$, the F1 score decreases compared with vanilla attention. And adding $\mathcal{L}_3$ does help minimize the gap. This is because $\mathcal{L}_3$ could improve the model generalization performance by making the model more stable.

Conclusions

In this paper, we provide a first rigorous definition namely SEAT as a substitute for attention to give a more faithful explanation. We also propose a method to get such a SEAT. Results show that SEAT outperforms other baselines and is considered as an effective and more faithful explanation tool.

Acknowledgements

Di Wang and Lijie Hu were supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Di Wang was also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. F. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 3504–3512.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458.

Dombrowski, A.; Alber, M.; Anders, C. J.; Ackermann, M.; Müller, K.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13567–13578.

Dong, Y.; Li, Z.; Rezagholizadeh, M.; and Cheung, J. C. K. 2019. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3393–3402. Florence, Italy: Association for Computational Linguistics.

Du, M.; Liu, N.; and Hu, X. 2020. Techniques for interpretable machine learning. *Commun. ACM*, 63(1): 68–77.

Galassi, A.; Lippi, M.; and Torroni, P. 2020. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10): 4291–4308.

Ghorbani, A.; Abid, A.; and Zou, J. Y. 2019. Interpretation of Neural Networks Is Fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 3681–3688. AAAI Press.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Kitada, S.; and Iyatomi, H. 2021. Attention Meets Perturbations: Robust and Interpretable Attention With Adversarial Training. *IEEE Access*, 9: 92974–92985.

Lei, T. 2017. *Interpretable neural models for natural language processing*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, USA.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Martins, A. F. T.; and Astudillo, R. F. 2016. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1614–1623. JMLR.org.

Miyato, T.; Dai, A. M.; and Goodfellow, I. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 1–17.

Mohankumar, A. K.; Nema, P.; Narasimhan, S.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2020. Towards Transparent and Explainable Attention Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4206–4216.

Otter, D. W.; Medina, J. R.; and Kalita, J. K. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Networks Learn. Syst.*, 32(2): 604–624.

Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 115–124.

Rehurek, R.; and Sojka, P. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. San Diego, California: Association for Computational Linguistics.

Sato, M.; Suzuki, J.; Shindo, H.; and Matsumoto, Y. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4323–4330.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

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

Yeh, C.; Hsieh, C.; Suggala, A. S.; Inouye, D. I.; and Ravikumar, P. 2019. On the (In)fidelity and Sensitivity of Explanations. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 10965–10976.

Yin, F.; Shi, Z.; Hsieh, C.-J.; and Chang, K.-W. 2022. On the Sensitivity and Stability of Model Interpretations in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2631–2647. Dublin, Ireland: Association for Computational Linguistics.