WHY IS ATTENTION NOT SO INTERPRETABLE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention-based methods have played an important role in model interpretations, where the calculated attention weights are expected to highlight the critical parts of inputs (e.g., keywords in sentences). However, recent research points out that attention-as-importance interpretations often do not work as well as we expect. For example, learned attention weights sometimes highlight less meaningful tokens like "`[SEP]`", "`,`", and "`.`", and are frequently uncorrelated with other feature importance indicators like gradient-based measures. Finally, a debate on the effectiveness of attention-based interpretations has been raised. In this paper, we reveal that one root cause of this phenomenon can be ascribed to the *combinatorial shortcuts*, which stands for that in addition to the highlighted parts, the attention weights themselves may carry extra information which could be utilized by downstream models of attention layers. As a result, the attention weights are no longer pure importance indicators. We theoretically analyze the combinatorial shortcuts, design one intuitive experiment to demonstrate their existence, and propose two methods to mitigate this issue. Empirical studies on attention-based interpretation models are conducted, and the results show that the proposed methods can effectively improve the interpretability of attention mechanisms on a variety of datasets.

1 INTRODUCTION

Interpretation for machine learning models has increasingly gained interest and becomes a necessity as the industry rapidly embraces machine learning technologies. Model interpretation explains how models make decisions, which is particularly essential in mission-critical domains where the accountability and transparency of the decision-making process are crucial, such as medicine (Wang et al., 2019), security (Chakraborti et al., 2019), and criminal justice (Lipton, 2018).

Attention mechanisms have played an important role in model interpretations and have been widely adopted for interpreting neural networks (Vaswani et al., 2017) and other black-box models (Chen et al., 2018). In this paper, similar to Vaswani et al. (2017), we assume that we have the query Q and the key-value pairs $\langle K, V \rangle$, then attention mechanisms work in the following way,

$$\text{Attention}(Q, K, V) = \text{Mask}(Q, K) \odot V^{[1]},$$

where $\text{Mask}(\cdot, \cdot)$ maps the query and keys to the attention weights (denoted as *masks* in this paper), and then the masks filter the information of V. Intuitively, the *masks* are expected to represent the importance of different parts of V (e.g., words of a sentence, pixels of an image) and highlight those the models should focus on to make decisions. Many researchers directly use the masks to provide interpretability of models (Choi et al., 2016; Vaswani et al., 2017; Wang et al., 2016).

However, recent research suggests that attention mechanisms' highlighted parts do not necessarily correlate with greater importance on the final predictions. For example, Clark et al. (2019) find that a surprisingly large amount of BERT's attention focuses on less meaningful tokens like "`[SEP]`", "`,`", and "`.`". Moreover, many researchers provide evidence to support or refute the interpretability of the attention mechanisms, and a debate on the effectiveness of attention-based interpretations has been raised (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019).

[1] Oftentimes a sum pooling operator is applied after the Hadamard product operator to obtain a single fixed-length representation. However, sometimes models other than simple pooling are applied (Bang et al., 2019; Chen et al., 2018; Zhu et al., 2017). So we use the most general form here.

In this paper, we suggest a root cause that hinders the interpretability of attention mechanisms, which we refer to as ***combinatorial shortcuts***. As mentioned earlier, we expect that the results of attention mechanisms mainly contain information from the highlighted parts of V, which is a critical assumption for attention-based interpretations' effectiveness. However, as the results are products of the masks and V, we find that the *masks* themselves can carry extra information other than the highlighted parts of V, which could be utilized by the down-stream parts of models. As a result, the calculated masks may work as another kind of *"encoding layers"* rather than providing pure importance measures. For an extreme example, in a (binary) text classification task, the attention mechanisms could highlight the first word for positive cases and highlight the last word for negative cases, regardless of the words. The downstream parts of attention layers could then predict the label by checking whether the first or the last word is highlighted. It may give good accuracy scores, while completely fail at providing interpretability[2].

We further study the effectiveness of attention-based interpretations and dive into the combinatorial shortcut problem. From the perspective of causal effect estimations, we firstly analyze the difference between ordinary attention mechanisms and definitive interpretations theoretically, and then show the existence of combinatorial shortcuts through a representative experiment. Based on this, we propose two practical methods to mitigate the issue, i.e., random attention pretraining and instance weighting for mask-neutral learning. Without loss of generality, we examine the effectiveness of proposed methods upon an end-to-end attention-based model-interpretation method, i.e., L2X (Chen et al., 2018), which can select a given number of input components to explain arbitrary black-box models. Experimental results show that the proposed methods can successfully mitigate the adverse impact of combinatorial shortcuts and improve explanation performance.

2 Related Work

Attention mechanisms for model interpretations Attention mechanisms have been widely adopted in natural language processing (Bahdanau et al., 2015; Vinyals et al., 2015), computer vision (Fu et al., 2016; Li et al., 2019), recommendations (Bai et al., 2020; Zhang et al., 2020b) and so on. Attention mechanisms are believed to explain how models make decisions by exhibiting the importance distribution over inputs (Choi et al., 2016; Martins & Astudillo, 2016; Wang et al., 2016), which we can regard as a kind of model-specific interpretations. Besides, there are also attention-based methods for model-agnostic interpretations. For example, L2X (Chen et al., 2018) is a hard attention model (Xu et al., 2015) that employs Gumbel-softmax (Jang et al., 2017) for instancewise feature selection. VIBI (Bang et al., 2019) improved L2X to encourage the briefness of the learned explanation by adding a constraint for the feature scores to a global prior. Liang et al. (2020) and Yu et al. (2019) improved attention-style model interpretation methods through adversarial training to encourage the gap between the predictability of selected/unselected features.

However, there has been a debate on the interpretability of attention mechanisms recently. Jain & Wallace (2019) suggested that "attention is not explanation" by finding that the attention weights are frequently uncorrelated with other feature importance indicators like gradient-based measures. On the other side, Wiegreffe & Pinter (2019) argued that "attention is not not-explanation" by challenging many assumptions underlying (Jain & Wallace, 2019) and suggested that they did not disprove the usefulness of attention mechanisms for explainability. Serrano & Smith (2019) applied a different analysis based on intermediate representation erasure and found that while attention noisily predicts input components' overall importance to a model, it is by no means a fail-safe indicator.

In this work, we take another perspective on this problem called *combinatorial shortcuts*, and show that it can provide one root cause of the phenomenon. Liang et al. (2020) and Yu et al. (2019) also mentioned this phenomenon, and used adversarial training to indirectly mitigate the combinatorial shortcut problem. However, they did not analyze why combinatorial shortcuts exist, and the adversarial scheme may change the behavioral tendency of models. For example, considering a sample with features $\{a, a', b, b', \text{noisy features}\}$, where a' is collinear with a, b' is collinear with b, and a is relatively more predictive than b while they are somehow complementary, adversarial methods that

[2]One may argue that for the most ordinary practice where sum pooling is applied, we lose the positional information, and as a result, the intuitive case described above may not hold. However, since (1) the distributions of different positions are not the same, (2) positional encodings (Vaswani et al., 2017) have been used widely, it is still possible for attention mechanisms to utilize the positional information with sum pooling.

encourage the gap between the predictability of selected/unselected features may tend to select a and a' to maximum the gap, and fail to select a and b which might be more interpretable.

Causal effect estimations Causal effect is an important concept for quantitative empirical analyses. The causal effect of one treatment, E, over another, C, is defined as the difference between what would have happened if a particular unit had been exposed to E and what would have happened if the unit had been exposed to C (Rubin, 1974). Randomized experiments, where the experimental units across the treatment groups are randomly allocated, play a critical role in causal inference. However, when randomized experiments are infeasible, researchers have to resort to nonrandomized data from surveys, censuses, and administrative records (Winship & Morgan, 1999), and there may be some other variables controlling the treatment allocation process in such data. For example, consider the causal inference between uranium mining and health. Ideally, the treatment (uranium mining) should be randomly allocated. However, mining workers are usually stronger people among all humans in the world. If some appropriate measures are absent, we may draw biased conclusions like "uranium mining has no adverse health impact because the average life span of uranium mine workers is not shorter than that of ordinary people". For recovering causal effects form nonrandomized data, instance weighting-based approaches have been used widely (Ertefaie & Stephens, 2010; Rosenbaum & Rubin, 1983; Winship & Morgan, 1999).

3 Combinatorial Shortcuts

In this section, by showing the difference between attention mechanisms and definitive explanations, we analyze why attention mechanisms become less interpretable from a perspective of causal effect estimations, and conduct an experiment to demonstrate the existence of combinatorial shortcuts.

3.1 The difference between attention mechanisms and definitive explanations

Firstly, we analyze what definitive explanations would be. Assume that we have samples drawn independently and identically distributed (i.i.d.) from a distribution with domain $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature domain, and $\mathcal{Y}$ is the label domain[3]. Additionally, we assume that the mask is drawn from a distribution with domain $\mathcal{M}$. Usually, $\mathcal{M}$ is under some constraints, for example, only being able to select a fixed number of features, or being non-negative and summing to 1. Given any sample $\langle x, y\rangle \sim \mathcal{X} \times \mathcal{Y}$, for $m_1 \sim \mathcal{M}$ and $m_2 \sim \mathcal{M}$, if $\mathcal{L}\big(\mathbb{E}(Y|x \odot m_1), y\big) < \mathcal{L}\big(\mathbb{E}(Y|x \odot m_2), y\big)$ where $\mathcal{L}(\cdot)$ is the loss function and $\mathbb{E}(\cdot)$ calculates the expectation, we say that for this sample, m_1 is superior to m_2 in term of interpretability. If an unbiased estimation of $\mathbb{E}(Y|X \odot M)$ is available, the best mask for sample $\langle x, y\rangle$ that can select the most informative features can be obtained by solving $\operatorname{argmin}_m \mathcal{L}\big(\mathbb{E}(Y|x \odot m), y\big)$. As a conclusion, under the definitions above, the definitive explanation for sample $\langle x, y\rangle$ is $\operatorname{argmin}_m \mathcal{L}\big(\mathbb{E}(Y|x \odot m), y\big)$ with an unbiased estimation of $\mathbb{E}(Y|X \odot M)$.

In practice, we often need to train models to estimate $\mathbb{E}(Y|X \odot M)$. Ideally, if the data (combinations of X and M, as well as the label Y) is exhaustive and the model is consistent, we can train a model to obtain an unbiased estimation of $\mathbb{E}(Y|X \odot M)$ following the empirical risk minimization principle (Fan et al., 2005; Vapnik, 1992). Nevertheless, in reality, it is not possible to exhaust all combinations of X and M. Taking a step back, from the perspective of causal effect estimations, we could consider different m as different treatment, and randomized combinations for X and M can still be proven to give unbiased estimations on expectations (Rubin, 1974).

However, attention mechanisms do not work in this way. Considering the downstream part of attention models, i.e., the part estimating the function $\mathbb{E}(Y|X \odot M)$, we can find that it receives highly selective combinations of X and M. The used mask M during the training procedure is a mapping from query and keys, making the used mask for samples highly related to the feature X (and Y as well). Therefore, the training procedure of attention mechanism produces a nonrandomized experiment (Shadish et al., 2008). Thus the model cannot learn unbiased $\mathbb{E}(Y|X \odot M)$. In turn, the attention mechanism will try to select features to adapt the biased estimations, and thus fail to highlight the essential features. As a result, the attention mechanism and downstream part may cooperate and find unexpected ways to fit the data, e.g., highlighting the first word for positive cases

[3]Note that here we use $\mathcal{X}$ to represent the features for the convention. X is the same as the value V introduced in the Introduction. Besides, the labels could be either from the real world for explaining the real world, or from some specific models for explaining given black-box models.

and the last word for negative cases, ultimately failing to provide interpretability. In this paper, we denote the effects of nonrandomized combination for X and M as *combinatorial shortcuts*, which hinders the interpretability of attention mechanisms.

3.2 EXPERIMENTAL DEMONSTRATION FOR THE COMBINATORIAL SHORTCUTS

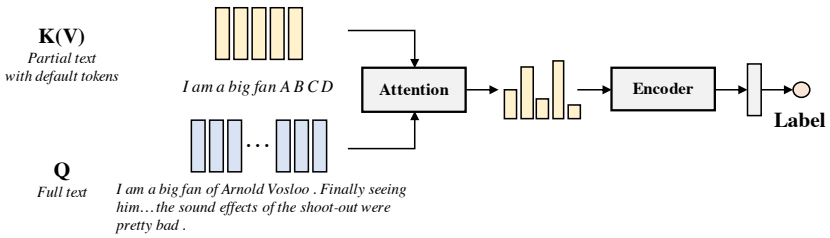

Figure 1: The structure of the model used in this demo experiment. The attention is restrict to the first five tokens of sentences, as well as four default tokens which appear in all samples.

To intuitively demonstrate the existence of combinatorial shortcuts, we design a simple experiment on text classification tasks. As Figure 1 shows, we train attention models, where the query of attention is encoded with the whole sentences. However, we only allow attention to highlight keywords among the first five tokens and four default tokens, i.e., "A", "B", "C", and "D". Since the default tokens appear in all samples and do not carry any useful information, if the attention mechanism can indeed highlight the critical parts of the inputs, little attention should be paid to them. Moreover, we could check whether the models put differential attention to the default tokens for different classes, to show if the information is encoded in the mask due to combinatorial shortcuts.

We use the real-world IMDB dataset (Maas et al., 2011) for experiments and examined different settings regarding the encoder in Figure 1, i.e., whether the encoder was a simple sum pooling or a trainable neural network model, i.e., recurrent convolutional neural network (RCNN) proposed by Lai et al. (2015). We use pre-trained GloVe word embeddings (Pennington et al., 2014) and kept them fixed to prevent shortcuts through word embeddings. We trained soft attention models for 25 epochs with RMSprop optimizers using default parameters and recorded the averaged results of 10 runs. The results are reported in Table 1.

Table 1: Experiments about how much attention is put to default tokens out of all attention weights. Note that we report the results on the training set to demonstrate how models fit the data.

No.	Encoder	Label	Attention to default tokens				Total
			A	B	C	D	
(1)	Pooling	pos	68.0%	0.1%	0.0%	0.2%	68.3%
(2)		neg	0.1%	36.6%	38.3%	18.5%	93.5%
(3)	RCNN	pos	16.1%	0.2%	0.1%	0.6%	17.0%
(4)		neg	1.3%	20.2%	50.4%	13.3%	85.2%

As we can see in Table 1, the attention models place more than half of the attention weights to the default tokens, and the weights for different classes are significantly different. Taking Pooling encoders as an example, the models put up to 68.0% attention weights to default token "A" for positive samples, and put 36.6%, 38.3%, and 18.5% attention weights to default token "B", "C", and "D" respectively for negative samples, summing up to 93.4% in total. As for RCNN encoders which is position-aware, the results are similar. The reason for observing repeatable results on the default tokens with different initialization may be their slight asymmetry in the GloVe embedding space. These results suggest that the attention mechanism may not work as expected to highlight the critical parts of inputs and provide interpretability. Instead, they learn to work as another kind of "encoding layers", and utilize the default tokens to fit the data through combinatorial shortcuts.

4 METHODS FOR MITIGATING COMBINATORIAL SHORTCUTS

In this section, based on the perspective of causal effect estimations introduced in Section 3.1, we come up with two practical methods, *random attention pretraining* and *mask-neutral learning with instance weighting*, to mitigate combinatorial shortcuts for better interpretability.

4.1 RANDOM ATTENTION PRETRAINING

We first propose a simple and straightforward method to address the issue. As analyzed in Section 3.1, the fundamental reason for combinatorial shortcuts of attention mechanisms is the biased estimation of $\mathbb{E}(Y|X \odot M)$, and random combinations of X and M can give unbiased results in theory. Inspired by this idea, we can first generate the masks completely at random and train the downstream part of the attention model. And then, we fix the downstream part, replace the random attention with a trainable attention layer, and train the attention layer only. As the downstream parts of neural networks are trained unbiasedly and fixed, training the attention layers solely is solving $\operatorname{argmin}_m \mathcal{L}\big(\mathbb{E}(Y|x \odot m), y\big)$ with an unbiased estimation of $\mathbb{E}(Y|X \odot M)$. Thus the interpretability is guaranteed.

In theory, this method is complete. However, it may be practically incompetent because there are countless viable cases of the combinations of X and M. It could be challenging to estimate $\mathbb{E}(Y|X \odot M)$ well, especially when the dimension of input features is high. Under such cases, the pretraining procedure may become less efficient as it needs to explore all possible masks evenly, even if most of the masks are worthless. In conclusion, the model may fail to estimate $\mathbb{E}(Y|X \odot M)$ well in some cases and thus limiting the interpretability.

4.2 MASK-NEUTRAL LEARNING WITH INSTANCE WEIGHTING

The second method is designed as a supplementary solution to address the shortcomings of random attention pretraining. This method is based on instance weighting, which has been successfully applied for mitigating sample selection bias (Zadrozny, 2004; Zhang et al., 2019), social prejudices bias (Zhang et al., 2020a), and also for recovering the causal effects (Ertefaie & Stephens, 2010; Winship & Morgan, 1999). The core idea of this method is that instead of learning a biased $\mathbb{E}(Y|X \odot M)$, with instance weighting, we could recover a *mask-neutral distribution* where the masks are unrelated to the labels. Thus the combinatorial shortcuts can be partially mitigated.

Generation of biased distributions from mask-neutral distributions We first define the mask-neutral distribution and its relationship with the biased distribution to which ordinary attention mechanisms are trained. Considering the downstream part of the attention layers, which is estimating $\mathbb{E}(Y|X \odot M)$, we assume that there is a mask-neutral distribution $\mathscr{Q}$ with domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{M} \times \mathcal{S}$, where $\mathcal{X}$ is the feature space, $\mathcal{Y}$ is the (binary) label space[4], $\mathcal{M}$ is the feature mask space and $\mathcal{S}$ is the binary sampling indicator space. During the training procedure, the selective combination of masks and features result in the combinatorial shortcuts. We assume for any given sample (x, y, m, s) drawn independently from $\mathscr{Q}$, it will be selected to appear in the training of attention mechanisms if and only if $s = 1$, which results in the biased distribution $\mathscr{P}$. We use $P(\cdot)$ to represent probabilities of the biased distribution $\mathscr{P}$, and $Q(\cdot)$ for the mask-neutral distribution $\mathscr{Q}$, then we have

$$P(\cdot) = Q(\cdot|S = 1)\,. \tag{1}$$

Ideally, we should have $M \perp (X, Y)$ on $\mathscr{Q}$ to obtain unbiased $\mathbb{E}(Y|X \odot M)$ as discussed in Section 3.1. However, when both sides are vectors, it will be intractable. Therefore, we take a step back and only assume $Y \perp M$ on $\mathscr{Q}$, i.e.,

$$Q(Y|M) = Q(Y)\,. \tag{2}$$

If S is completely at random, $\mathscr{P}$ will be consistent with $\mathscr{Q}$. However, the attention layers are highly selective, making that only some combinations of X and M are visible to the downstream model. To further simplify the problem, we assume that M and Y control S. And for any given Y and M, the probability of selection is greater than 0, defined as

$$Q(S = 1|X, Y, M) = Q(S = 1|Y, M) > 0\,. \tag{3}$$

[4]We focus on binary classification problems in this paper, but the proposed methodology can be easily extended to multi-class classifications.

Additionally, we assume that the selection does not change the marginal probability of M and Y, i.e.,

$$P(M) = Q(M)\,, P(Y) = Q(Y)\,. \tag{4}$$

In other words, we assume that although S is dependent on the combination of M and Y in $\mathscr{Q}$, it is independent on either M or Y only, i.e., $Q(S|M) = Q(S)$ and $Q(S|Y) = Q(S)$.

The unbiased expectation of loss with instance weighting We show that, by adding proper instance weights, we can obtain an unbiased estimation of the loss on the mask-neutral distribution $\mathscr{Q}$, with only the data from the biased distribution $\mathscr{P}$.

Fact 1 (Unbiased Loss Expectation)**.** *For any function $f = f(x \odot m)$, and for any loss $\mathcal{L} = \mathcal{L}\big(f(x \odot m), y\big)$, if we use $w = \frac{P(y)}{P(y|m)}$ as the instance weights, then*

$$\mathbb{E}_{x,y,m\sim\mathscr{P}}\Big[w\mathcal{L}\big(f(x \odot m), y\big)\Big] = \mathbb{E}_{x,y,m\sim\mathscr{Q}}\Big[\mathcal{L}\big(f(x \odot m), y\big)\Big]\,.$$

Fact 1 shows that, by a proper instance-weighting method, the the downstream part of the attention model can learn on the mask-neutral distribution $\mathscr{Q}$, where $Q(Y|M) = Q(Y)$. Therefore, the independence between M and Y is encouraged, then it will be hard for the classifier to approximate Y solely by M. Thus, the classifier will have to use useful information from X, and have the combinatorial shortcuts problem mitigated.

We present the proof for Fact 1 as follows.

Proof. We first present an equation with the weight w,

$$\begin{aligned} w =& \frac{P(y)}{P(y|m)} = \frac{Q(y)}{Q(y|m, S=1)} = \frac{Q(y)}{Q(S=1|y,m)Q(y|m)/Q(S=1|m)} \\ =& \frac{Q(S=1)}{Q(S=1|y,m)} = \frac{Q(S=1)}{Q(x,y,m|S=1)Q(S=1)/Q(x,y,m)} = \frac{Q(x,y,m)}{P(x,y,m)}\,. \end{aligned}$$

Then we have

$$\begin{aligned} \mathbb{E}_{x,y,m\sim\mathscr{P}}\Big[w\mathcal{L}\big(f(x\odot m), y\big)\Big] &= \int \frac{Q(x,y,m)}{P(x,y,m)}\mathcal{L}\big(f(x \odot m), y\big)dP(x,y,m) \\ &= \int \mathcal{L}\big(f(x \odot m), y\big)dQ(x,y,m) = \mathbb{E}_{x,y,m\sim\mathscr{Q}}\Big[\mathcal{L}\big(f(x \odot m), y\big)\Big]\,. \end{aligned}$$

□

Mask-neutral learning With Fact 1, we now propose mask-neutral learning for better interpretability of attention mechanisms. As shown, by adding instance weight $w = \frac{P(y)}{P(y|m)}$ to the loss function, we can obtain unbiased loss of the mask-neutral distribution. As distribution $\mathscr{P}$ is directly observable, estimating $P(\cdot)$ is possible. In practice, we could train a classifier to estimate $P(Y|M)$ along with the training of the attention layer, and optimize it and the attention layers, as well as the other parts of models alternatively.

Compared with the random attention pretraining method, the instance weighting-based approach concentrates more on the useful masks. Thus it will suffer less from the efficiency problem. Nevertheless, the effectiveness of the instance weighting method relies on the assumptions as shown in Equation (1)–(4). However, in some cases, the assumptions may not hold. For example, in Equation (3), we assume that given Y and M, S is independent on X. In other words, X controls S only through Y. This assumption is necessary for simplifying the problem, while may not sometimes hold when given Y and M, X can still influence S. Besides, the effectiveness of the method also relies on an accurate estimation of $P(Y|M)$, which may require careful tuning as the probability $P(Y|M)$ is dynamically changing along the training process of attention mechanisms.

5 Experiments

In this section, we present the experimental results of the proposed methods. For simplicity, we denote *random attention pretraining* as **Pretraining** and *mask-neutral learning with instance weighting* as **Weighting**. Firstly, we analyze the effectiveness of mitigating combinatorial shortcuts. Then, we examine the effectiveness of improving interpretability.

5.1 EXPERIMENTS FOR MITIGATING COMBINATORIAL SHORTCUTS

We applied the proposed methods to the experiments introduced in Section 3.2 to check whether we can mitigate the combinatorial shortcuts. We summarize the results in Table 2.

Table 2: Effectiveness of the proposed methods for mitigating the combinatorial shortcuts.

No.	Method	Encoder	Label	Attention to default tokens A	B	C	D	Total
(1)	Pretraining	Pooling	pos	0.0%	4.6%	2.3%	0.0%	6.9%
(2)			neg	0.0%	0.4%	20.2%	0.0%	20.6%
(3)		RCNN	pos	0.2%	1.1%	1.0%	0.3%	2.6%
(4)			neg	2.3%	4.3%	6.3%	2.8%	15.7%
(5)	Weighting	Pooling	pos	1.3%	2.5%	0.5%	2.2%	6.5%
(6)			neg	3.3%	0.7%	1.5%	1.2%	6.7%
(7)		RCNN	pos	6.7%	1.4%	10.3%	2.1%	20.5%
(8)			neg	4.6%	1.8%	12.5%	1.9%	20.8%

As presented, after applying Pretraining and Weighting, the percentage of attention weights assigned to the default tokens were significantly reduced. Since that the default tokens do not provide useful information but only serve as carriers for combinatorial shortcuts, the results reveal that our methods have mitigated the combinatorial shortcuts successfully.

5.2 EXPERIMENTS FOR IMPROVING INTERPRETABILITY

In this section, using L2X (Chen et al., 2018) as an example and basis, we present the effectiveness of mitigating combinatorial shortcuts for better interpretability. We first introduce the evaluation scheme, then show the experimental results and discussions.

5.2.1 EVALUATION SCHEME

Here we present the evaluation scheme. Due to space constraints, We present details in Appendix A.

Evaluation protocol Our evaluation scheme was the same as L2X (Chen et al., 2018). L2X is an instancewise feature selection model using hard attention that employs the Gumbel-softmax trick. It tries to select a certain number of input components to approximate the model's output to be explained with attention mechanisms. As discussed before, such a setting suffers from combinatorial shortcuts. Thus the interpretability may be limited. Also, similar with (Liang et al., 2020), to further enrich the information for the model explanation, we incorporate the original model's outputs to be explained, i.e.$\hat{y}$, as part of the query for feature selection. This trick can make it easier for the explanation model to select the best features. As obtaining the outputs requires no further information apart from samples' features and the model to be explained, it does not hurt the model-agnostic property of explanation methods nor require additional annotations. We adopt binary feature-attribution masks to select features, i.e., top k values of the mask were set to 1, others are set to 0, then we treat $X \odot M$ as the selected features (Chen et al., 2018).

Evaluation metrics The same as Chen et al. (2018) and Liang et al. (2020), we performed a predictive evaluation that evaluates how accurate the original given model can approximate the original model-outputs using the selected features, and we report the post-hoc accuracy. We repeat ten times with different initialization for each method on each dataset and report the averaged results.

Datasets We report evaluations on four datasets: IMDB (Maas et al., 2011), Yelp P. (Zhang et al., 2015), MNIST (LeCun et al., 1998), and Fashion-MNIST (F-MNIST) (Xiao et al., 2017). IMDB and Yelp P. are two text classification datasets. IMDB is with 25,000 train examples and 25,000 test examples. Yelp P. contains 560,000 train examples and 38,000 test examples. MNIST and F-MNIST are two image classification datasets. For MNIST, following Chen et al. (2018), we collected a binary classification subset by choosing images of digits 3 and 8, with 11,982 train examples and 1,984 test examples. For F-MNIST, following Liang et al. (2020), we selected the data of Pullover and Shirt with 12,000 train examples and 2,000 test examples.

Models to be explained The same as Chen et al. (2018) for IMDB and Yelp P., we implemented CNN-based models and selected 10 and 5 words, respectively, for explanations. For MNIST and

F-MNIST, we used the same CNN model as (Chen et al., 2018) and selected 25 and 64 pixels, respectively (Liang et al., 2020).

Baselines We considered state-of-the-art model-agnostic baselines: LIME (Ribeiro et al., 2016), CXPlain (Schwab & Karlen, 2019), L2X (Chen et al., 2018), VIBI (Bang et al., 2019), and AIL (Liang et al., 2020). We also compared with model-specific baselines, i.e., Gradient (Simonyan et al., 2013). Our methods follow the same paradigm as L2X, VIBI, and AIL. A brief introduction to the baseline methods can be found in Appendix A.4.

5.2.2 Experimental results

Following the aforementioned evaluation scheme, we report the results in Table 3.

Table 3: Effectiveness of the proposed methods for improving interpretability. We report the post-hoc accuracy scores with different methods.

No.	Method		IMBD	Yelp P.	MNIST	F-MNIST
(1)	Gradient (Simonyan et al., 2013)		85.6%	82.3%	98.2%	58.6%
(2)	LIME (Ribeiro et al., 2016)		89.8%	87.4%	80.4%	75.6%
(3)	CXPlain (Schwab & Karlen, 2019)		90.6%	97.7%	99.4%	59.7%
(4)	L2X (Chen et al., 2018)		89.2%	88.2%	91.4%	77.3%
(5)	VIBI (Bang et al., 2019)		90.8%	94.4%	98.3%	84.1%
(6)	AIL (Liang et al., 2020)†		**98.5%**	**99.3%**	99.0%	**97.8%**
(7)		– –	48.8%	77.8%	94.9%	85.3%
(8)	L2X with $\hat{y}$	Pretraining	**97.1%**	**99.0%**	66.3%	89.4%
(9)		Weighting	94.3%	87.7%	**99.8%**	**95.4%**

†AIL utilizes additional information about the models to be explained, i.e., their gradients.

From the table, we find that directly adding $\hat{y}$ to the query did not always improve the performance by comparing Row (4) and (7). Interestingly, for the text classification datasets, adding $\hat{y}$ led to decreased performance, and meanwhile, Pretraining outperformed Weighting. For the image classification datasets, we had the exact opposite conclusion. We ascribe this phenomenon to the inherent differences between the two tasks. Firstly, a single word in a sentence is much more informative than a single pixel in an image. Secondly, the importance of words is more "continuous", and in contrast, the importance of pixels is more "discrete" and co-adapting. Intuitively, the function of $\mathbb{E}(Y|X \odot M)$ is smoother and more comfortable to learn for text classification tasks than for image classification tasks. As a result, as discussed in Section 4.1, it may be hard for Pretraining to learn reasonable estimations of $\mathbb{E}(Y|X \odot M)$ efficiently for images. Thus the performance of interpretability is limited, especially for MNIST, where the digital numbers are placed randomly compared with F-MNIST, where the items are aligned better.

By comparing with the baselines (especially L2X with $\hat{y}$), we find that despite the simplicity, Pretraining and Weighting can outperform most of the baselines and give comparable results with AIL, which utilizes additional information (i.e., their gradients) of the models to be explained. We conclude that mitigating the combinatorial shortcuts can effectively improve the interpretability. We present visualization examples of explanations in Appendix B.

6 Conclusion

Attention-based model interpretations have been popular for their convenience to integrate with neural networks. However, many researchers find that attention sometimes yields non-interpretable results, and there has been a debate on the interpretability of the attention mechanisms. In this paper, we propose that the combinatorial shortcuts are one of the root causes hindering attention mechanisms' interpretability. We analyze the combinatorial shortcuts theoretically and design experiments to show their existence. Furthermore, we propose two practical methods to mitigate combinatorial shortcuts for better interpretability. Experiments show that the proposed methods effectively mitigate combinatorial shortcuts and improve the interpretability of attention mechanisms. The results presented in this paper may help us better understand how attention mechanisms work.

REFERENCES

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

Bing Bai, Guanhua Zhang, Ye Lin, Hao Li, Kun Bai, and Bo Luo. CSRN: Collaborative sequential recommendation networks for news retrieval. *arXiv preprint arXiv:2004.04816*, 2020.

Seojin Bang, Pengtao Xie, Wei Wu, and Eric Xing. Explaining a black-box using deep variational information bottleneck approach. *arXiv preprint arXiv:1902.06918*, 2019.

Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29(1), pp. 86–96, 2019.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pp. 883–892, 2018.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512, 2016.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

Ashkan Ertefaie and David A Stephens. Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics*, 6(2), 2010.

Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip S Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pp. 605–608, 2005.

Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2321–2334, 2016.

Sarthak Jain and Byron C Wallace. 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)*, pp. 3543–3556, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations*, 2017.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2267–2273, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7026–7035, 2019.

Jian Liang, Bing Bai, Yuren Cao, Kun Bai, and Fei Wang. Adversarial infidelity learning for model interpretation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 286–296, 2020.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-volume 1*, pp. 142–150. Association for Computational Linguistics, 2011.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pp. 1614–1623, 2016.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, pp. 10220–10230, 2019.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019.

William R Shadish, Margaret H Clark, and Peter M Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American statistical association*, 103(484):1334–1344, 2008.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pp. 831–838, 1992.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pp. 2773–2781, 2015.

Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Annals of Internal Medicine*, 2019.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 606–615, 2016.

Sarah Wiegreffe and Yuval Pinter. 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)*, pp. 11–20, 2019.

Christopher Winship and Stephen L Morgan. The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706, 1999.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4085–4094, 2019.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pp. 114, 2004.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. Selection bias explorations and debias methods for natural language sentence matching datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4418–4429, 2019.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4134–4145, 2020a.

Junqi Zhang, Bing Bai, Ye Lin, Jian Liang, Kun Bai, and Fei Wang. General-purpose user embeddings based on mobile app usage. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2831–2840, 2020b.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pp. 649–657, 2015.

Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5513–5522, 2017.