Alignment Rationale for Natural Language Inference

Zhongtao Jiang[1,2], **Yuanzhe Zhang**[1,2], **Zhao Yang**[1,2], **Jun Zhao**[1,2] and **Kang Liu**[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{zhongtao.jiang, yzzhang, zhao.yang, jzhao, kliu}@nlpr.ia.ac.cn

Abstract

Deep learning models have achieved great success on the task of Natural Language Inference (NLI), though only a few attempts try to explain their behaviors. Existing explanation methods usually pick prominent features such as words or phrases from the input text. However, for NLI, alignments among words or phrases are more enlightening clues to explain the model. To this end, this paper presents AREC, a post-hoc approach to generate alignment rationale explanations for co-attention based models in NLI. The explanation is based on *feature selection*, which keeps few but sufficient alignments while maintaining the same prediction of the target model. Experimental results show that our method is more faithful and readable compared with many existing approaches. We further study and re-evaluate three typical models through our explanation beyond accuracy, and propose a simple method that greatly improves the model robustness.[1]

1 Introduction

Natural Language Inference (NLI) is a fundamental task in Natural Language Processing (NLP) which is to determine if a hypothesis entails a premise. Recently, with the introduction of large-scale annotated datasets (Bowman et al., 2015; Williams et al., 2018), deep learning models are adopted to solve the task in a supervised manner (Conneau et al., 2017; Chen et al., 2017; Devlin et al., 2019) and achieve great success, while inner mechanisms of these methods are still opaque due to high computational complexities.

Towards interpretability, explaining the model behavior has gained increasing attention. Lots of approaches are based on *feature attribution* which assigns saliency scores for input features (Bahdanau et al., 2015; Lundberg and Lee, 2017; Thorne et al., 2019; Kim et al., 2020), and *feature selection* or *rationale* that keeps a subset of features sufficient for the prediction (Lei et al., 2016; Bastings et al., 2019; De Cao et al., 2020; DeYoung et al., 2020). Figure 1 (a) and (b) present a text attribution explanation by LIME (Ribeiro et al., 2016) and a text rationale explanation from Li et al. (2016) of an NLI sentence pair. Both explanations provide insights of which input words are responsible for the prediction. However, NLI is a cross-sentence task requiring a system to reason over alignments[2] (MacCartney and Manning, 2009). Intuitively, it is more sensible to explain NLI systems in the way of

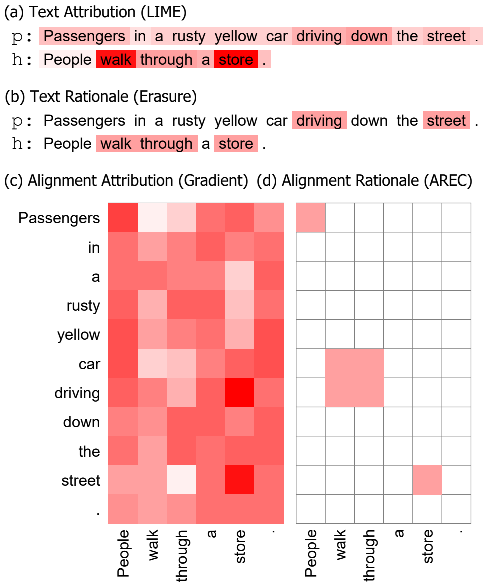


Figure 1: Different post-hoc explanations. For attribution explanations, features with deeper colors are considered more important.

[1] Our code is available at https://github.com/changmenseng/arec

[2] In machine translation, alignments refer to bilingual text pairs with identical meanings. But for NLI, the semantics of two sentences may be different, it is more suitable to define alignments as any text pairs related lexically or logically, etc.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5372–5387
August 1–6, 2021. ©2021 Association for Computational Linguistics

alignments instead of isolated words/phrases. For the example in Figure 1, the contradicted phrase pair `street – store` is one of the key alignments responsible for the correct prediction.

To explain NLI models over alignments, the literature usually looks at co-attention weights (Parikh et al., 2016; Pang et al., 2016; Chen et al., 2017), which is a dominant way to implicitly align word pairs (Wang et al., 2017; Gong et al., 2018; Devlin et al., 2019). However, attention is argued not as explainable as expected (Jain and Wallace, 2019; Serrano and Smith, 2019; Bastings and Filippova, 2020). Moreover, co-attention assigns scores among words thus forbids us to observe phrase-level alignments, which is a flaw that generally exists for attribution explanations as shown in Figure 1 (c). Other works build hard alignments resorting sparse attention (Yu et al., 2019; Bastings et al., 2019; Swanson et al., 2020). But their self-explanatory architectures pay for the interpretability at a cost of performance dropping on accuracy (Molnar, 2020). Meanwhile, these techniques are unable to analyze well-trained models.

To resolve above problems, this paper proposes AREC, a post-hoc local approach to generate **A**lignment **R**ationale **E**xplanation for **C**o-attention based models. Analogous with Lei et al. (2016), our *alignment rationale* is a set that contains text pairs from the NLI sentence pair with two requirements. First, the explanation is supposed to be faithful to the predictive model, where selected text pairs must alone suffice for the original prediction. Second, the explanation should be human-friendly or readable (Miller, 2019), which means the pairs are few to promote compact rationales, and extracted continuously to make phrase-level rationales as far as possible (Lei et al., 2016; Bastings et al., 2019). Figure 1 (d) presents an example of AREC explanation. It shows that the model reaches the right prediction reasonably: it identifies `People–Passengers`, `walk through–car driving` and `store – street` to make up the alignment rationale. AREC is flexible to apply on any co-attention architectures, allowing us for deep investigations of well-trained models.

With the proposed AREC, we study three typical co-attention based models Decomposable Attention (DA) (Parikh et al., 2016), Enhanced LSTM (ESIM) (Chen et al., 2017) and BERT (Devlin et al., 2019) on four benchmarks including SNLI (Bowman et al., 2015), ESNLI (Camburu et al., 2018), BNLI (Glockner et al., 2018) and HANS (McCoy et al., 2019). Experimental results show that our method could generate more faithful and readable explanations. Moreover, we employ our proposed AREC to analyze these models deeply from the aspect of alignments. Based on our explanations, we further present a simple improvement strategy that greatly increases robustness of different models without modifying their architectures or retraining. This proves that our method could factually reflect how models work.

Our contributions are summarized as follows:

1) We come up with AREC, a post-hoc local explanation method to extract the alignment rationale for co-attention based models. We compare AREC with other explanation methods, illustrating its advantages on faithfulness and readability.

2) We diagnose three typical co-attention based models using AREC by re-evaluating them in a more fine-grained alignment level beyond accuracy. Experimental results could reveal potential improvement solutions. To the best of our knowledge, we are the first to study existing models with alignment exhaustively.

2 Related Works

Natural Language Inference

Natural Language Inference has been studied for years. Despite lots of works construct representations for the input two sentences individually (Bowman et al., 2015; Mueller and Thyagarajan, 2016; Conneau et al., 2017), the task actually requires a system to recognize alignments (MacCartney and Manning, 2009). In early days, alignment detection is sometimes formed as an independent task (Chambers et al., 2007; MacCartney et al., 2008), or a component of a pipeline system (MacCartney et al., 2006). Currently deep learning methods seek to model alignments implicitly through co-attention mechanism (Parikh et al., 2016; Pang et al., 2016; Chen et al., 2017; Wang et al., 2017; Gong et al., 2018; Joshi et al., 2019; Devlin et al., 2019). The technique is first proposed in machine translation (Bahdanau et al., 2015), and soon dominates in many applications including NLI. However why models with co-attention layers are effective is still called for answers.

Explaining Models in NLP

Explaining model behaviors has attracted much interests. Existing studies include opening the component of models (Murdoch et al., 2018), assigning word importance scores (Ribeiro et al., 2016; Li et al., 2016; Kim et al., 2020), extracting predictive related input pieces, referred as sufficient input subset (Carter et al., 2019) or rationale (Lei et al., 2016; Bastings et al., 2019), building hierarchical explanations (Chen et al., 2020; Zhang et al., 2020), and generating natural language explanations (Camburu et al., 2018; Kumar and Talukdar, 2020). However, they usually explain the model on the granularity of words/phrases. Such ways are sufficient for text classification but not suitable for NLI, since atom features in the task are alignments.

Co-attention itself is often viewed as an explanation. Indeed, co-attention is a key proxy to model alignments, where perturbing its weights has a significant impact (Vashishth et al., 2019). Yet recently, attention is argued to be not explainable as expected (Jain and Wallace, 2019; Serrano and Smith, 2019; Grimsley et al., 2020; Bastings and Filippova, 2020). Secondly, co-attention along with *feature attribution* explanations just assigns scores among words, which is infeasible to observe phrase-level alignments. Furthermore, for models with multiple attentions (Vaswani et al., 2017), it's hard to acquire a global understanding of alignments. Other approaches include Yu et al. (2019), who adopts *generator-encoder* architecture (Lei et al., 2016) to generate corresponded rationales. But their approach is unable to extract more fine-grained alignments (e.g., one-to-one continuous alignments). Bastings et al. (2019); Swanson et al. (2020) design sparse attention for hard alignments. However, these methods trade performance for interpretability, and are immutable to analyze well-trained models.

3 Method

In this section, we describe our AREC in details. As mentioned before, AREC is a post-hoc approach for explaining co-attention based models. Thus we first introduce the co-attention layer, then depict the propose AREC.

3.1 Background: Co-Attention in NLI Models

In our notation, we have an instance including a premise $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_{|p|}] \in \mathbb{R}^{d \times |p|}$ and a hypothesis $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_{|h|}] \in \mathbb{R}^{d \times |h|}$, where $|p|/|h|$ is the length of the premise/hypothesis, and $\mathbf{p}_i/\mathbf{h}_j \in \mathbb{R}^d$ denotes corresponding word embedding (fixed or contextual). Co-attention layer accepts $\mathbf{P}$ and $\mathbf{H}$ as input and outputs alignment enhanced word representations $\bar{\mathbf{P}} \in \mathbb{R}^{d \times |p|}$ and $\bar{\mathbf{H}} \in \mathbb{R}^{d \times |h|}$. At the first step, we compute a similarity matrix $\mathbf{S} \in \mathbb{R}^{|p| \times |h|}$

$$\mathbf{S}_{i,j} = \phi(\mathbf{p}_i, \mathbf{h}_j) \tag{1}$$

where ϕ is a similarity function, ordinarily a vector dot product (Chen et al., 2017). Then $\mathbf{S}$ is normalized to compute soft alignment scores for every word in a sentence w.r.t all the words in its partner

$$\begin{aligned} \mathbf{AP}_{i,:} &= \mathrm{softmax}(\mathbf{S}_{i,:}) \\ \mathbf{AH}_{:,j} &= \mathrm{softmax}(\mathbf{S}_{:,j}) \end{aligned} \tag{2}$$

Here $\mathbf{AP}$ and $\mathbf{AH}$ are so-called co-attention matrices, each element inside indicates the matching degree of the corresponding word pair. Next, we obtain soft alignments features for every word in the premise/hypothesis by averaging word embeddings in the hypothesis/premise weighted by the soft alignment scores

$$\begin{aligned} \bar{\mathbf{P}} &= \mathbf{H} \cdot \mathbf{AP}^T \\ \bar{\mathbf{H}} &= \mathbf{P} \cdot \mathbf{AH} \end{aligned} \tag{3}$$

Now $\bar{\mathbf{P}}/\bar{\mathbf{H}}$ is a richer representation of $\mathbf{P}/\mathbf{H}$ enhanced by $\mathbf{H}/\mathbf{P}$ and fed to following modules, such as a classifier which outputs probabilities of candidate categories, i.e., *entailment*, *contradiction* and *neutral* in NLI task.

3.2 Problem Formation

The proposed AREC relies on *feature selection*, keeping few but sufficient alignments while maintaining the original prediction. Thus to restrict the model to only consider some specific alignments, we intuitively mask co-attention matrices $\mathbf{AP}$ and $\mathbf{AH}$ following Serrano and Smith (2019); Pruthi et al. (2020). Let $\mathbf{Z} \in \{0,1\}^{|p| \times |h|}$ be a binary mask indicating the presence or absence of every word pair alignment, and M be a model with co-attention layers. Then the masking process is simply Hadamard product between mask $\mathbf{Z}$ and co-attention matrices $\mathbf{AP}$ and $\mathbf{AH}$. An alignment rationale is obtained by an optimistic problem

$$\tilde{\mathbf{Z}} = \arg\min_{\mathbf{Z}} \lambda_0 \mathcal{L}_0 + \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 \tag{4}$$

The loss contains three terms ($\mathcal{L}_0$, $\mathcal{L}_1$ and $\mathcal{L}_2$) to satisfy faithfulness and readability as mentioned in Section 1. λ_0, λ_1 and λ_2 are hyper-parameters standing for loss weights. Every rectangular region in $\tilde{\mathbf{Z}}$ represents a text alignment in the alignment rationale.

We now describe loss terms. The first term $\mathcal{L}_0$ is about *fidelity*, asking that the model prediction is maintained after masking (Molnar, 2020). *Fidelity* ensures faithfulness, making the derived explanation depict the true profile of how the model works. We choose the euclidean distance between logits as this loss term, i.e.,

$$\mathcal{L}_0 := \|\mathrm{M}_l(\mathbf{P},\mathbf{H}) - \mathrm{M}_l^{\mathbf{Z}}(\mathbf{P},\mathbf{H})\|_2 \qquad (5)$$

where $\mathrm{M}_l(\mathbf{P},\mathbf{H})$ and $\mathrm{M}_l^{\mathbf{Z}}(\mathbf{P},\mathbf{H}) \in \mathbb{R}^3$ are original output logits and output logits when applying the mask $\mathbf{Z}$ respectively. Compared to commonly used KL divergence (De Cao et al., 2020) or label equality (Feng et al., 2018), the euclidean distance between logits is a stricter constraint that narrows down the solution space and would lead to more faithful explanations[3].

Secondly, an explanation ought to be readable (Molnar, 2020). That requirement contains *compactness* and *contiguity* under the context of alignment explanation. *Compactness* draws intuition from the philosophy that a good explanation should be short or selective (Miller, 2019), which encourages fewer alignments to be selected. *Compactness* loss is simply the L1 norm of the mask $\mathbf{Z}$

$$\mathcal{L}_1 := |\mathbf{Z}|_1 = \sum_{i,j} z_{i,j} \qquad (6)$$

where $z_{i,j}$ is an element in $\mathbf{Z}$. *Contiguity* encourages continuous phrase-level alignments[4] (Zenkel et al., 2020), which is helpful for human understandings. Concretely, *contiguity* prefers $\mathbf{Z}$ with rectangular clusters. Thus, we have

$$\mathcal{L}_2 := \sum_{i,j} \mathbb{1}\left(\sum_{z \in \mathrm{W}_{i,j}^z} z = 3\right) \qquad (7)$$

where $\mathbb{1}(\cdot)$ is the indicator function and $\mathrm{W}_{i,j}^z = \{z_{i,j}, z_{i,j+1}, z_{i+1,j}, z_{i+1,j+1}\}$ is a 2×2 window at the position. The loss is based on the observation that if there are three 1s in the window, there must be a non-rectangle region nearby, as marked by red boxes in Figure 2.

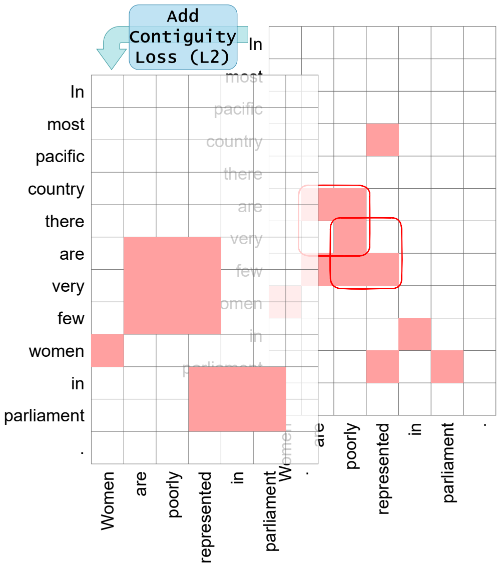


Figure 2: The contiguity loss $\mathcal{L}_2$ could encourage the algorithm to extract phrase alignments, i.e., penalises $\mathbf{Z}$ with non-rectangular clusters, as marked by red boxes.

3.3 Optimization

Searching the exponential huge $(2^{|p||h|})$ solution space of $\mathbf{Z}$ straightforwardly is impracticable. To use the gradient-based method, we relax binary $\mathbf{Z}$ to be a stochastic matrix $\boldsymbol{Z}$, and optimize loss expectation over it. Specifically, we assume that every element $Z_{i,j}$ in $\boldsymbol{Z}$ is an independent random variable satisfying HardConcrete distribution (Louizos et al., 2018a). HardConcrete variables are allowed to be exactly discrete 0 and 1, while having continuous and differential probability densities on the open interval $(0, 1)$. Additionally, HardConcrete distribution accommodates *reparameterization*, permitting us to obtain a HardConcrete sample z by transforming a parameter-less unit uniform sample u, i.e., $z = g(u; \alpha)$, where g is differential. Details are shown in Appendix A.

Under this setting, we turn to optimize the expectation of the objective. For $\mathcal{L}_0$, we have

$$\begin{aligned}\mathcal{L}_0 &= \mathbb{E}_{\boldsymbol{U}}[\|\mathrm{M}_l(\mathbf{P},\mathbf{H}) - \mathrm{M}_l^{g(\boldsymbol{U};\boldsymbol{\alpha})}(\mathbf{P},\mathbf{H})\|_2] \\ &\simeq \frac{1}{n}\sum_{i=1}^{n} \|\mathrm{M}_l(\mathbf{P},\mathbf{H}) - \mathrm{M}_l^{g(\mathbf{U}_i;\boldsymbol{\alpha})}(\mathbf{P},\mathbf{H})\|_2\end{aligned} \qquad (8)$$

Here, $\boldsymbol{U}$ is a random matrix filled with i.i.d unit uniform variables, $\boldsymbol{\alpha} \in \mathbb{R}_+^{|p| \times |h|}$ is the parameter of $\boldsymbol{Z}$. The second line is a Monte-Carlo approximation of the expectation, where n is the sample size, and $\mathbf{U}_i$ is the i-th sample of $\boldsymbol{U}$.

[3] If we use label equality (Feng et al., 2018), which the prediction is only maintained in terms of the label, there are many explanations satisfying the constraint. Using a strict fidelity constraint ensures uniqueness or less variety, making the explanation more faithful.

[4] Following Lei et al. (2016) and Bastings et al. (2019), a phrase could be any continuous span in a sentence, which may not be a syntactical phrase.

For $\mathcal{L}_1$ and $\mathcal{L}_2$, we have

$$\begin{aligned}\mathcal{L}_1 &= \sum_{i,j} \mathbb{E}(Z_{i,j}) \leq \sum_{i,j} \mathrm{P}(Z_{i,j} \neq 0; \alpha_{i,j}) \\ \mathcal{L}_2 &= \sum_{i,j} \mathbb{E}\left[\mathbb{1}\left(\sum_{Z \in \mathrm{W}^{Z}_{i,j}} \lceil Z \rceil = 3\right)\right] \\ &= \sum_{i,j} \sum_{Z \in \mathrm{W}^{Z}_{i,j}} \mathrm{P}(Z = 0; \alpha) \\ &\qquad \prod_{Z' \in \mathrm{W}^{Z}_{i,j} \setminus \{Z\}} \mathrm{P}(Z' > 0; \alpha')\end{aligned} \tag{9}$$

where $\lceil Z \rceil$ is the up round of Z and $\mathrm{P}(\cdot; \alpha)$ is the probability over the parameter α. Now, all the losses are differential over $\boldsymbol{\alpha}$, making gradient descent feasible. Derivation details are presented in Appendix B.

After training, we obtain the alignment rationale as follows

$$\tilde{z}_{i,j} = \arg\max_{v \in \{0,1\}} \mathrm{P}(Z_{i,j} = v; \alpha_{i,j}) \tag{10}$$

4 Experiments

Our experiments include two parts. First, we quantitatively compare the proposed AREC with several typical explanation methods (Section 4.1) to prove the effectiveness of our method. Second, by means of AREC, we study and re-evaluate different models from the aspect of alignment beyond accuracy, revealing potential improvements (Section 4.2).

Datasets

We use four datasets SNLI (Bowman et al., 2015), ESNLI (Camburu et al., 2018), BNLI (Glockner et al., 2018) and HANS as our testbeds. SNLI is a traditional NLI benchmark, while ESNLI extends it by annotating text rationales. BNLI and HANS are stress testing sets to test lexical inference and overlap heuristics respectively.

Models

We choose three typical co-attention based NLI models DA[5] (Parikh et al., 2016), ESIM (Chen et al., 2017) and BERT (base version) (Devlin et al., 2019) for our discussion. DA applies the co-attention directly on word embeddings. ESIM further incorporates order information by putting two LSTMs before and after the co-attention layer (Hochreiter and Schmidhuber, 1997) to boost the performance. Differently, BERT concatenates the input sentence pair with a template "`[CLS]` p `[SEP]` h `[SEP]`" and uses global self-attention (Vaswani et al., 2017). All the models are trained on SNLI training set and tested across datasets.

Implementation

We mask attention matrices for DA and ESIM as described in Section 3.2 since they are directly formed by co-attention. For BERT, we use a single mask to mask co-attention corresponded sub-matrices[6] of all the attention matrices identically, no matter of their layers or attention heads.

We consider that faithfulness has a higher priority than readability. Correspondingly, we adjust weights in the loss dynamically, based on fidelity of current mask. To this end, weights are set as

$$\lambda_0 = 1, \lambda_1 = \lambda_2 = 0.15 \times \mathrm{SpAc} \tag{11}$$

where SpAc is the accuracy of current sampled masks

$$\mathrm{SpAc} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\mathrm{M}_y(\mathbf{P}, \mathbf{H}) = \mathrm{M}_y^{g(\mathbf{U}_i;\boldsymbol{\alpha})}(\mathbf{P}, \mathbf{H})] \tag{12}$$

Here, $\mathrm{M}_y^{\mathbf{Z}}$ is the model predicted label under mask $\mathbf{Z}$. Thus terms related to readability are controlled by the explanation faithfulness. This simple dynamic weight strategy is similar to the approach in Platt and Barr (1988) and highly improves the explanation quality and the algorithm stability.

4.1 Explanation Evaluation

In this section, we aim to evaluate the faithfulness and readability of different explanations.

4.1.1 Baselines

We select feature attribution baselines including co-attention itself, perturbation-based approaches LEAVEONEOUT (Li et al., 2016), LIME (Ribeiro et al., 2016), BACKSELECT (Carter et al., 2019), gradient-based approaches GRADIENT (Simonyan et al., 2014) and INTEGRATGRAD (Sundararajan et al., 2017) and a feature selection method DIFFMASK (De Cao et al., 2020). The original DIFFMASK is applied on text level, we derive an alignment variant for comparison in Appendix C.

[5] Following Glockner et al. (2018), in our implementation, we discard the optional intra-sentence attention and achieve simlar and comparable accuracy performance.

[6] For a BERT attention map $\mathbf{A} \in \mathbb{R}^{(|p|+|h|+3)\times(|p|+|h|+3)}$, $\mathbf{A}_{2:|p|+1,|p|+3:|p|+|h|+2}$ and $\mathbf{A}_{|p|+3:|p|+|h|+2,2:|p|+1}$ are co-attention corresponded sub-matrices.

4.1.2 Metrics

Inspired by DeYoung et al. (2020), we use Area Over Reservation Curve (AORC) to evaluate faithfulness[7] as follows

$$\mathrm{AORC} = \sum_{k=0}^{K} \|\mathrm{M}_l(\mathbf{P},\mathbf{H}) - \mathrm{M}_l^{\mathbf{Z}^{(k)}}(\mathbf{P},\mathbf{H})\|_2 \tag{13}$$

where $\mathbf{Z}^{(k)}$ is the mask that reserves top k% co-attention weights from an attribution explanation. Though AREC belongs to feature selection explanations, its parameter $\boldsymbol{\alpha}$ also provides importance scores. We also report fidelity defined in Equation (5) as a measure of faithfulness.

For readability evaluation, we report compactness and contiguity defined in Equation (6) and Equation (7) respectively. We also conduct human evaluations on random sampled 300 examples from SNLI testing test to directly measure readability. We let 2 annotators to rate how easy the explanation is to read and understand the model's decision-making process along alignments from 1 to 5 points and report the average scores[8].

We admit that metrics including fidelity, compactness and contiguity are that AREC optimizes. Actually it's hard to unitedly evaluate different explanations since their contexts and techniques are usually completely different. If we only follow definitions of those metrics, we consider they are reasonable. Note that these metrics are not compatible for feature attribution explanations. For fair comparison, we follow Carter et al. (2019) to induce alignment rationales by thresholding[9] for feature attribution baselines. That is, we sequentially remain co-attention weights according to attribution scores until the fidelity loss is lower than the pre-defined threshold.

4.1.3 Results

Automatic evaluation and readability human evaluation results are shown in Table 1 and Table 2 respectively. We obtain the following findings:

1) AREC is quite faithful with the lowest AORC and fidelity value in most cases. Perturbation-based methods are equally matched with moderate performances, while gradient-based ones have the least faithfulness. Surprisingly, co-attention is a very strong baseline to indicate important alignments for NLI, surpassing most other baselines on AORC, extremely for ESIM. This result is of accordance with Vashishth et al. (2019) that attention is more faithful in cross-sentence tasks compared with single-sentence tasks.

2) AREC is quite readable which achieves the lowest compactness value and contiguity value in most cases for automatic evaluation. AREC is also the most readable explanation according to human evaluation. As a contrast, feature attribution methods are unable to induce readable alignment rationales. They reserve too much co-attention weights, usually half of which, to ensure similar fidelity with AREC rather than satisfying compactness and contiguity. Appendix E shows some examples for intuitive feelings of different explanations' readabilities.

3) Compared to rationale explanation DIFFMASK, AREC is far more promising that outperforms it with huge gaps on fidelity while maintains equivalent or better compactness and contiguity. In our knowledge, DIFFMASK is to *globally* learn to explain *local* instances: the explainer is trained on a training set which may contain artifacts and biases (Gururangan et al., 2018; Tsuchiya, 2018; Poliak et al., 2018). Therefore this architecture leverages data information. It is susceptible to over-fitting and generate data-relevant biased explanations as a result, leading to poor fidelity when facing held-out data (BNLI and HANS) as shown in Table 1. Moreover, we believe that a faithful explanation is a profile of a model. Correspondingly, an explanation method should only access knowledge from the model instead of from the data. That is an appealing theoretical advantage of our method.

4.2 Beyond Accuracy: Behavior Testing of NLI Models with AREC

Diverse evaluations are pursued to understand models profoundly (Ribeiro et al., 2020). Beyond accuracy, in this section, we analyze DA, ESIM and BERT resorting to our proposed AREC by re-evaluating them from the more fined-grained aspect of alignment. For a model, we first generate its alignment rationales using AREC, then

[7] We don't use Area Over Perturbation Curve (AOPC) (DeYoung et al., 2020) because our method is to reserve features (i.e., alignments) that keep the prediction, it is fitter to utilize reservation curve.

[8] Both annotators are well-educated postgraduates major in computer science. We conduct human evaluation on randomly sampled 300 examples in SNLI testing set.

[9] The threshold is set to $\mathcal{L}_0 + 0.1$ of AREC to obtain alignment rationales with similar fidelity for fair comparison. We don't use fix size constraint to construct rationales as done in Jain et al. (2020) because we think the size of a rationale depends on the instance.

Models	Explanations	SNLI				BNLI				HANS			
		Faithfulness		Readability		Faithfulness		Readability		Faithfulness		Readability	
		AORC	FIDE	COMP	CONT	AORC	FIDE	COMP	CONT	AORC	FIDE	COMP	CONT
DA	Co-Attention	0.60	0.45*	42.46	131.30	0.46	0.39*	30.93	59.85	**0.48**	0.56*	22.88	41.90
	LeaveOneOut	1.12	0.43*	57.78	70.91	1.23	0.34*	64.67	65.02	0.95	0.58*	66.30	125.06
	BackSelect	1.15	0.43*	57.05	67.08	1.34	0.34*	65.19	55.88	1.07	0.58*	71.61	137.85
	LIME	0.99	0.43*	52.80	90.81	1.22	0.34*	63.01	71.95	0.81	0.57*	48.32	124.71
	Gradient	1.42	0.42*	65.65	135.09	1.73	0.35*	74.80	155.50	1.76	0.55*	65.69	194.50
	IntegratGrad	1.83	0.35*	63.87	49.76	2.31	0.25*	81.60	44.76	2.37	0.38*	70.98	80.43
	DiffMask	0.54	1.28	**2.77**	**0.21**	0.62	1.30	6.86	1.36	0.71	0.97	6.46	1.39
	AREC (Ours)	**0.47**	**0.36**	6.23	1.40	**0.42**	**0.32**	**6.83**	**1.12**	0.60	**0.50**	**6.07**	**0.23**
ESIM	Co-Attention	**0.24**	0.29*	8.72	4.43	**0.55**	0.15*	15.46	6.555	**0.51**	0.42*	14.40	1.36
	LeaveOneOut	1.01	0.25*	42.88	17.80	1.05	0.16*	53.15	23.38	1.05	0.43*	56.37	30.76
	BackSelect	0.90	0.25*	41.08	15.73	1.08	0.16*	52.32	16.12	0.98	0.43*	50.88	27.52
	LIME	0.94	0.27*	52.46	72.29	1.52	0.16*	76.52	57.85	1.29	0.42*	73.68	179.10
	Gradient	2.84	0.20*	73.37	109.19	3.51	0.10*	83.60	78.83	5.15	0.22*	91.05	111.14
	IntegratGrad	2.99	0.21*	80.32	33.21	3.80	0.15*	89.68	13.91	4.45	0.38*	91.38	55.63
	DiffMask	0.51	1.21	**3.94**	**0.26**	0.71	2.62	**9.77**	2.00	0.79	1.89	**8.34**	1.06
	AREC (Ours)	0.40	**0.23**	4.86	0.70	0.60	**0.15**	11.02	**0.62**	0.73	**0.36**	12.43	**0.41**
BERT	Co-Attention	0.52	0.45*	27.91	58.20	0.65	0.34*	26.81	46.40	0.61	0.50*	29.60	57.68
	LeaveOneOut	1.00	0.44*	45.50	50.05	0.64	0.36*	39.82	66.35	0.93	0.48*	43.51	58.19
	BackSelect	0.92	0.45*	41.32	42.08	0.69	0.37*	40.08	60.90	0.98	0.48*	40.94	55.80
	LIME	0.82	0.44*	39.69	57.69	0.62	0.36*	44.01	96.05	0.99	0.46*	50.47	92.14
	Gradient	1.77	0.39*	75.58	127.92	4.63	0.16*	90.35	74.64	3.59	0.26*	90.93	132.30
	IntegratGrad	1.45	0.42*	59.82	56.57	1.21	0.32*	54.30	70.37	2.52	0.31*	74.26	90.15
	DiffMask	0.62	1.00	14.40	7.41	1.61	2.67	19.43	20.17	0.70	0.95	18.95	10.26
	AREC (Ours)	**0.43**	**0.36**	**6.05**	**2.18**	**0.47**	**0.28**	**8.30**	**2.65**	**0.53**	**0.44**	**8.56**	**0.79**

Table 1: Evaluation results of explanations across datasets. FIDE, COMP and CONT denote fidelity, compactness and contiguity respectively. We report COMP in % and CONT in ‰ for convenience. Numbers marked by * are fidelity of attribution induced rationales and are at the same level with AREC's fidelity for fair comparison.

Explanations	Models		
	DA	ESIM	BERT
Co-Attention	2.70	3.75	2.19
LeaveOneOut	2.42	2.67	2.47
BackSelect	2.60	2.71	2.74
LIME	2.40	2.13	2.42
Gradient	1.68	1.42	1.31
IntegratGrad	2.14	1.69	2.28
DiffMask	3.98	3.92	3.08
AREC (Ours)	**4.07**	**4.03**	**3.98**

Table 2: Human evaluation results of readability.

we evaluate its alignment plausibility (Jacovi and Goldberg, 2020): how well do its alignment rationales agree with human judgments (DeYoung et al., 2020). Since it is established in Section 4.1 that our method is faithful, thus alignment plausibility reflects a model's power of alignment detection, i.e., whether it makes a prediction with right alignments. Figure 3 illustrates the evaluation process.

Firstly, let's have a look at Table 3 that shows the accuracy performances of various models across datasets. Both DA, ESIM and BERT achieve high and tied accuracy performances on SNLI. However, they are distinguished on lexical reasoning, where BERT surpasses others significantly on BNLI. Additionally, neither of them is robust against overlap heuristic, as their performances are extremely poor on *non-entailment* instances. We seek to uncover the behind reasons (Section 4.2.2) and try to make improvements (Section 4.2.3) using our AREC.

4.2.1 Metrics

We define different metrics to measure alignment plausibility (or equally speaking, alignment rationale agreements with humans) in various datasets.

For eSNLI, since it's annotated in the text level, we simply collect corresponding words to convert an alignment rationale to a text rationale for comparison. We adopt IOU-F1 and Token-F1 from DeYoung et al. (2020), and only use a subset of eSNLI whose instances are labeled *contradiction* for our evaluation[10].

In BNLI, each sentence pair differs by a single word or phrase. Naturally this pair forms up an annotation, which should be counted in a golden alignment rationale. Further, We reasonably presume this pair is the most essential alignment in its corresponding alignment rationale. Thus, three metrics are defined: 1) Max-F1: we remain the alignment with max score from the alignment rationale outputted by AREC according to LeaveOneOut. Max-F1 is the F1 measure comparing remaining ones and annotations. 2) Exact-Inc: The

[10] In eSNLI, every *contradiction* instance selects words in both the premise and the hypothesis to make up text rationale, fitting with AREC explanations.

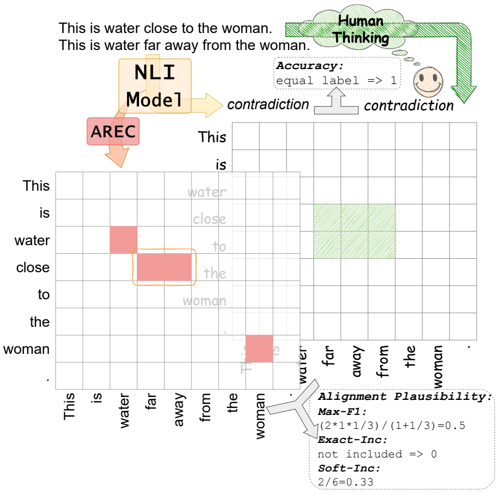


Figure 3: An illustration of evaluating instance-wise accuracy and alignment plausibility for a BNLI instance. Both evaluations compare model outputs and human outputs. Alignment in the orange box is remained for computing Max-F1. Human thinking outputs include annotated labels and rationales which could be annotated text rationales (ESNLI), annotated essential alignments (BNLI) and any other forms. If there are no annotated rationales, we apply human evaluations (HANS) to directly judge the agreements.

metric is the proportion that the alignment rationale includes the annotated alignment. 3) Soft-Inc: It is a loosed version of Exact-Inc, which is the average recall comparing alignment rationales and annotations. Details are shown in Figure 3.

We carry out human evaluations on HANS because it is not annotated in any form of rationales. We ask 2 human annotators if (yes/no) the decision process observed by AREC is agreed with them and report averaged agreed ratio[11] (see Appendix D for details).

4.2.2 Results

Table 3 shows alignment plausibility results, where we obtain the following findings:

1) Across datasets, alignment plausibilities are consistent with the accuracy performances in different degrees. Especially on BNLI, where BERT surpasses other competitors on all metrics substantially, quantitatively revealing that the alignment detection ability is important and distinguishes NLI models. We also discover that modeling order information explicitly is also useful for NLI, where ESIM achieves a better accuracy even with a poorer alignment plausibility on SNLI compared to DA.

Testsets	Metrics	DA	ESIM	BERT
SNLI	Accuracy	85.04	87.78	**90.27**
ESNLI$_\text{C}$	IOU-F1	27.62	20.44	**30.24**
	Token-F1	54.45	44.57	**60.52**
BNLI	Accuracy	48.82	67.09	**95.40**
	Max-F1	35.04	49.90	**64.05**
	Exact-Inc	66.58	83.11	**89.50**
	Soft-Inc	71.86	89.01	**93.11**
HANS$_\text{E}$	Accuracy	96.94	99.35	**99.56**
	Human	41.67	91.33	**94.00**
HANS$_\text{N}$	Accuracy	2.47	1.51	**16.59**
	Human	9.33	24.00	**27.33**

Table 3: Re-evaluation results of different models including rationale plausibility besides accuracy. ESNLI$_\text{C}$ is the *contradiction* labeled subset of ESNLI. HANS$_\text{C}$ and HANS$_\text{N}$ are *entailment* and *non-entailment* labeled subsets of HANS respectively.

Combining the two factors makes BERT an effective approach for NLI.

2) Our explanation method is helpful to detect artifacts or biases leveraged by the model. For example, though obtaining high accuracy on HANS$_\text{E}$, DA's low alignment plausibility suggests it usually makes a right prediction with wrong alignments (see Appendix D for examples). Further, all the models are brittle on catching reasonable alignments when facing *non-entailment* instances in HANS. As we will discuss next, they tend do shallow literal lexical matching, which we conjecture the reason why they also fail on accuracy.

In summary, the ability to capture correct alignments is closely related to accuracy performance in NLI. This conclusion is often discussed qualitatively in previous works. But we are the first to illustrate and prove this point exhaustively via quantitative evaluation.

4.2.3 Improving Robustness against Overlap Heuristics

With our AREC, we find that both three models tend to align overlapped words between the sentence pair no matter their syntactical or semantic roles, causing wrong predictions in HANS. Figure 4 presents an example, where the model mistakenly matches identical words. However, `president` in the premise and `doctor` in the hypothesis are subjects of the same predicate `advised`, they should be aligned, and so do `doctor` in the premise and `president` in the hypothesis.

To remedy this, we turn to Semantic Role Labeling (SRL), the task to recognize arguments for a predicate and assign semantic role labels to them,

[11] The human evaluation is conducted on randomly selected 300 examples, 10 examples per heuristic.

Methods	HANS						
	Entailment			Non-Entailment			Avg
	Lex	Sub	Cons	Lex	Sub	Cons	
DA	97.18	96.02	97.62	2.66	1.76	3.00	49.71
ESIM	**99.68**	98.76	99.60	0.18	0.12	4.22	50.43
BERT	98.82	**100.00**	**99.86**	43.02	2.94	3.82	58.08
DA$_{\text{SRL\_GUID}}$	93.66	96.64	96.36	88.24	25.88	3.28	67.34
ESIM$_{\text{SRL\_GUID}}$	93.94	96.76	99.42	**99.10**	**32.28**	5.30	**71.13**
BERT$_{\text{SRL\_GUID}}$	96.24	99.36	99.74	96.26	29.44	0.24	70.21
BERT$^{\ddagger}_{\text{SRL\_MTL}}$	91.00	98.00	95.00	71.00	13.00	**25.00**	66.00

Table 4: Accuracy performances of different models across different datasets. Lex, Sub and Cons are different overlap heuristics in HANS (McCoy et al., 2019). BERT$^{\ddagger}_{\text{SRL\_MTL}}$ is reported from Cengiz and Yuret (2020) that utilizes NLI and SRL multi-task learning and just for reference since they use different resources.

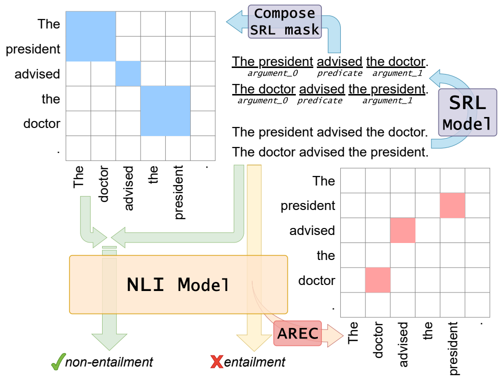

Compose SRL mask
The president advised the doctor.
argument_0 predicate argument_1
The doctor advised the president.
argument_0 predicate argument_1
SRL Model
The president advised the doctor.
The doctor advised the president.
The
president
advised
the
doctor
.
NLI Model
AREC
✓non-entailment
✗entailment

Figure 4: An illustration of using SRL to guide alignments. A NLI model fails on highly overlapped non-entailed examples (yellow path) because it mistakenly aligns overlapped words. To relief this problem, we use SRL to guide alignments by masking co-attention with a SRL mask (green path).

to guide alignments for NLI models. In particular, we employ an off-the-shelf BERT-based SRL model (Shi and Lin, 2019) to extract predicates and their corresponding arguments from the premise and the hypothesis in advance. Then we limit the model to only align identical predicates and phrases with identical semantic roles by applying a corresponding co-attention mask (SRL mask), as presented in Figure 4. In this way the semantic role information is injected into the model. Note that there is no need to modify the model architecture or design new training protocol, contrary to Cengiz and Yuret (2020) who jointly train NLI and SRL in a multi-task learning (MTL) manner.

We report model accuracy performances when alignments are guided by SRL masks (subscripted with SRL_GUID) in Table 4. The results show that without obvious performance drops on *entailment* instances, applying SRL masks gains significant improvements on *non-entailment* instances, especially for lexical heuristic. Nevertheless, it doesn't boost model performances for constituent heuristic. We speculate that is because constituent heuristic instances are accompanied with restrictions such as prepositions, which is unable to handle only with alignments. Overall, the results show that guiding alignments is a potential promising way to incorporate useful information. Additionally, this also proves that our method is faithful towards models from another point of view.

5 Conclusions

In this work, we propose AREC, a new post-hoc method to generate alignment rationale for co-attention based NLI models. Experimental results show that our explanation is faithful and readable. We study typical models using our method and shed lights on potential improvements. We believe our method and findings are illuminating for NLI. For future works, we plan to explore model-agnostic alignment explanations, and analyze models in other NLP tasks.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No.2020AAA0105200), the National Natural Science Foundation of China (No. 61922085, 61831022, 61906196). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006) and independent research project of National Laboratory of Pattern Recognition.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David K. Gifford. 2019. What made you do this? understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 567–576. PMLR.

Cemil Cengiz and Deniz Yuret. 2020. Joint training with semantic role labeling for better generalization in natural language inference. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 78–88, Online. Association for Computational Linguistics.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 165–170, Prague. Association for Computational Linguistics.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France. European Language Resources Association.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel Weld, and Luke Zettlemoyer. 2019. pair2vec: Compositional word-pair embeddings for cross-sentence inference. 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)*, pages 3597–3608, Minneapolis, Minnesota. Association for Computational Linguistics.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018a. Learning sparse neural networks through l_0 regularization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018b. Learning sparse neural networks through l_0 regularization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu, Hawaii. Association for Computational Linguistics.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 41–48, New York City, USA. Association for Computational Linguistics.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Citeseer.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations,* *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38.

Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu. com.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2786–2792. AAAI Press.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2793–2799. AAAI Press.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

John Platt and Alan Barr. 1988. Constrained differential optimization. In *Neural Information Processing Systems*. American Institute of Physics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

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

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. 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)*, pages 963–969, Minneapolis, Minnesota. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *CoRR*, abs/1909.11218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150. ijcai.org.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn. 2019. Learning the Dyck language with attention-based Seq2Seq models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 138–146, Florence, Italy. Association for Computational Linguistics.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

Die Zhang, Huilin Zhou, Xiaoyi Bao, Da Huo, Ruizhao Chen, Xu Cheng, Hao Zhang, Mengyue Wu, and Quanshi Zhang. 2020. Interpreting hierarchical linguistic interactions in dnns.

A The HardConcrete Distribution

The HardConcrete distribution (Louizos et al., 2018b) is derived from the binary Concrete distribution (Maddison et al., 2017) using *stretch and rectify*, assigning probability densities on the close unit interval $[0, 1]$. The Concrete distribution is a continuous relaxation of Categorical distribution and submissive for reparameterization (Gumbel-Softmax trick) (Maddison et al., 2017). We only introduce the special binary case here for conciseness.

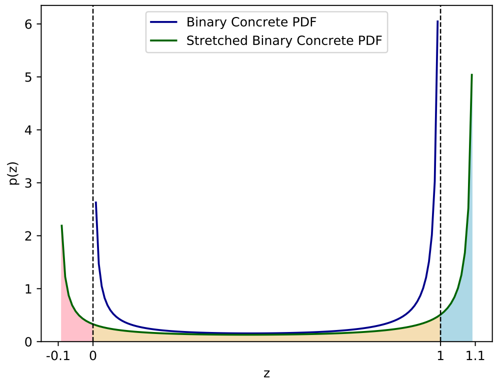


Figure 5: *Stretch and rectify* process of binary Concrete distribution. The binary Concrete PDF is stretched from (0,1) to (-0.1, 1,1). Red and blue regions are probability masses that the binary HardConcrete variable equals 0 and 1 separately.

A binary Concrete variable $\hat{Z}$ could be sampled by first sampling $U \sim \mathcal{U}(0, 1)$, and conducting the following transformations

$$\begin{aligned} L &= \log U - \log(1 - U) \\ \hat{Z} &= \sigma(\log \alpha + L)/\tau) \end{aligned} \tag{14}$$

where σ is sigmoid function, α and τ are parameters of $\hat{Z}$, where the latter one is called temperature controling the sharpness. In practice, $\log \alpha$ is usually the logit outputted by a classifier, e.g., a neural network. The probability density function (PDF) and the cumulative distribution function (CDF) of Z is

$$\begin{aligned} p_{\hat{Z}}(z) &= \frac{\tau \alpha z^{-\tau-1}(1 - z)^{-\tau-1}}{(\alpha z^{-\tau} + (1 - z)^{-\tau})^2} \\ Q_{\hat{Z}}(z) &= \sigma((\log z - \log(1 - z))\tau - \log \alpha) \end{aligned} \tag{15}$$

However, we are about to generate binary masks as our rationales, implying word alignment appearances. That is, we require Z remains some discrete properties, allowing us to sample the exact 0 and 1. For this purpose, Louizos et al. (2018b) introduces *stretch and rectify* strategy. As illustrated in Figure 5, the binary Concrete PDF is first stretched to support (γ, ζ), where $\gamma < 0$ and $\zeta > 1$, via a scaling transformation, then we rectify densities on the close unit interval

$$Z = \min(1, \max(0, \gamma + (\zeta - \gamma)\hat{Z})) \tag{16}$$

where γ, ζ and τ are hyperparameters and we set -0.1, 1.1 and 0.2 respectively. Transformations in

Equation (14) and Equation (16) compose g in Equation (8). Now, we have

$$\begin{aligned}
\mathrm{P}(Z=0) &= \mathrm{P}\left(0 < \hat{Z} \le \frac{\gamma}{\gamma-\zeta}\right) \\
&= Q_{\hat{Z}}\left(\frac{\gamma}{\gamma-\zeta}\right) \\
&= \sigma\left(\tau \log\left(-\frac{\gamma}{\zeta}\right) - \log\alpha\right)
\end{aligned} \tag{17}$$

and

$$\begin{aligned}
\mathrm{P}(Z=1) &= \mathrm{P}\left(\frac{1-\gamma}{\zeta-\gamma} \le \hat{Z} < 1\right) \\
&= 1 - Q_{\hat{Z}}\left(\frac{1-\gamma}{\zeta-\gamma}\right) \\
&= \sigma\left(\log\alpha - \tau \log\left(\frac{1-\gamma}{\zeta-\gamma}\right)\right)
\end{aligned} \tag{18}$$

B Loss Derivation

According to the above basis, for $\mathcal{L}_1$, we have

$$\begin{aligned}
\mathcal{L}_1 &= \sum_{i,j} \mathbb{E}(Z_{i,j}) \\
&= \sum_{i,j} \mathrm{P}(Z_{i,j}=1) + \int_0^1 z p_{Z_{i,j}}(z) dz \\
&\le \sum_{i,j} \mathrm{P}(Z_{i,j}=1) + \int_0^1 f_{Z_{i,j}}(z) dz \\
&= \sum_{i,j} (1 - \mathrm{P}(Z_{i,j}=0)) \\
&= \sum_{i,j} \sigma\left(\log\alpha_{i,j} - \tau \log\left(-\frac{\gamma}{\zeta}\right)\right)
\end{aligned} \tag{19}$$

Note that we optimize $\mathcal{L}_1$'s upper bound instead of itself. For $\mathcal{L}_2$, we have

$$\begin{aligned}
\mathcal{L}_2 &= \sum_{i,j} \mathbb{E}\left[\mathbb{1}\left(\sum_{Z \in \mathrm{W}_{i,j}} \lceil Z \rceil = 3\right)\right] \\
&= \sum_{i,j} \mathrm{P}\left[\mathbb{1}\left(\sum_{Z \in \mathrm{W}_{i,j}} \lceil Z \rceil = 3\right)\right] \\
&= \sum_{i,j} \sum_{Z \in \mathrm{W}^Z_{i,j}} \mathrm{P}(Z=0) \\
&\qquad \prod_{Z' \in \mathrm{W}^Z_{i,j} \setminus \{Z\}} (1 - \mathrm{P}(Z'=0)) \\
&= \sum_{i,j} \sum_{\alpha \in \mathrm{W}^\alpha_{i,j}} \sigma\left(\tau \log\left(-\frac{\gamma}{\zeta}\right) - \log\alpha\right) \\
&\qquad \prod_{\alpha' \in \mathrm{W}^\alpha_{i,j} \setminus \{\alpha\}} \sigma\left(\log\alpha' - \tau \log\left(-\frac{\gamma}{\zeta}\right)\right)
\end{aligned} \tag{20}$$

Optimizing $\mathcal{L}_1$ and $\mathcal{L}_2$ is directly since we don't need to sample. Now the loss functions are differential about $\boldsymbol{\alpha}$, allowing us to process gradient descent. In the implementation, we actually optimize over $\log\boldsymbol{\alpha}$ because it's a free variable.

C Alignment DiffMask Baseline

DiffMask utilizes a neural network to obtain $\log\boldsymbol{\alpha}$ on input representations, and optimizes the neural network on a training set. In the original implementation (De Cao et al., 2020), the neural network is feed with word vectors from different layers. To make it be on alignment level, $\log\boldsymbol{\alpha}$ is computed on alignment features

$$\log\alpha_{i,j} = \mathrm{FFN}([\mathbf{p}_i;\ \mathbf{h}_j; \mathbf{p}_i - \mathbf{h}_j; \mathbf{p}_i \odot \mathbf{h}_j]) \tag{21}$$

where FFN is a feed forward neural network with one hidden layer and ; means concatenation. Word representations $\mathbf{p}_i$ and $\mathbf{h}_j$ are the input incontextualized word vectors. The subsequent steps are similar to Arec, except that DiffMask is trained on a traning set, leveraging data knowledge.

D Alignment Plausibility Human Evaluation

The principle of manual evaluation is that the decision process observed by Arec is agreed with humans when it includes complete alignment information for the correct prediction. Thus, an alignment rationale could not agree with humans even instruct

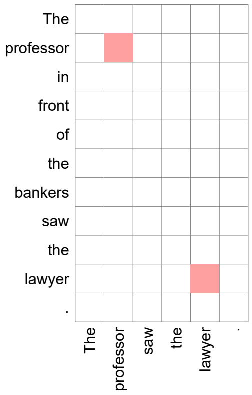
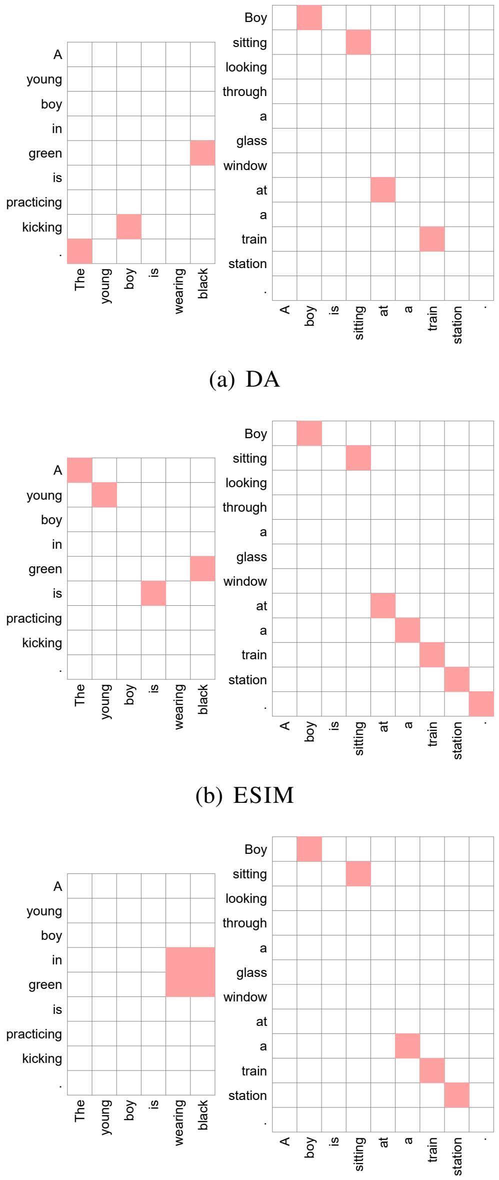


Figure 6: An example labeled *entailment* in HANS, where the alignment rationale is extracted from DA. The alignment rationale is not agreed with humans while allowing humans to reach the correct prediction.

humans to arrive the correct prediction. This is different from Human Accuracy (Jain et al., 2020). Figure 6 presents an example. From the alignment rationale, a human is able to predict *entailment* with identical nouns `professor – professor` and `lawyer – lawyer`. However, as a human, we also need to identify the predicate pair `saw – saw` for complete semantics. Thus, we consider alignment rationales like in Figure 6 are not agreed with human justifications.

E Visualization

We plot a few examples of AREC explanations in Figure 7. We also present examples of different alignment explanations in Figure 8. It's clear that our proposed AREC explanation is the most readable one.

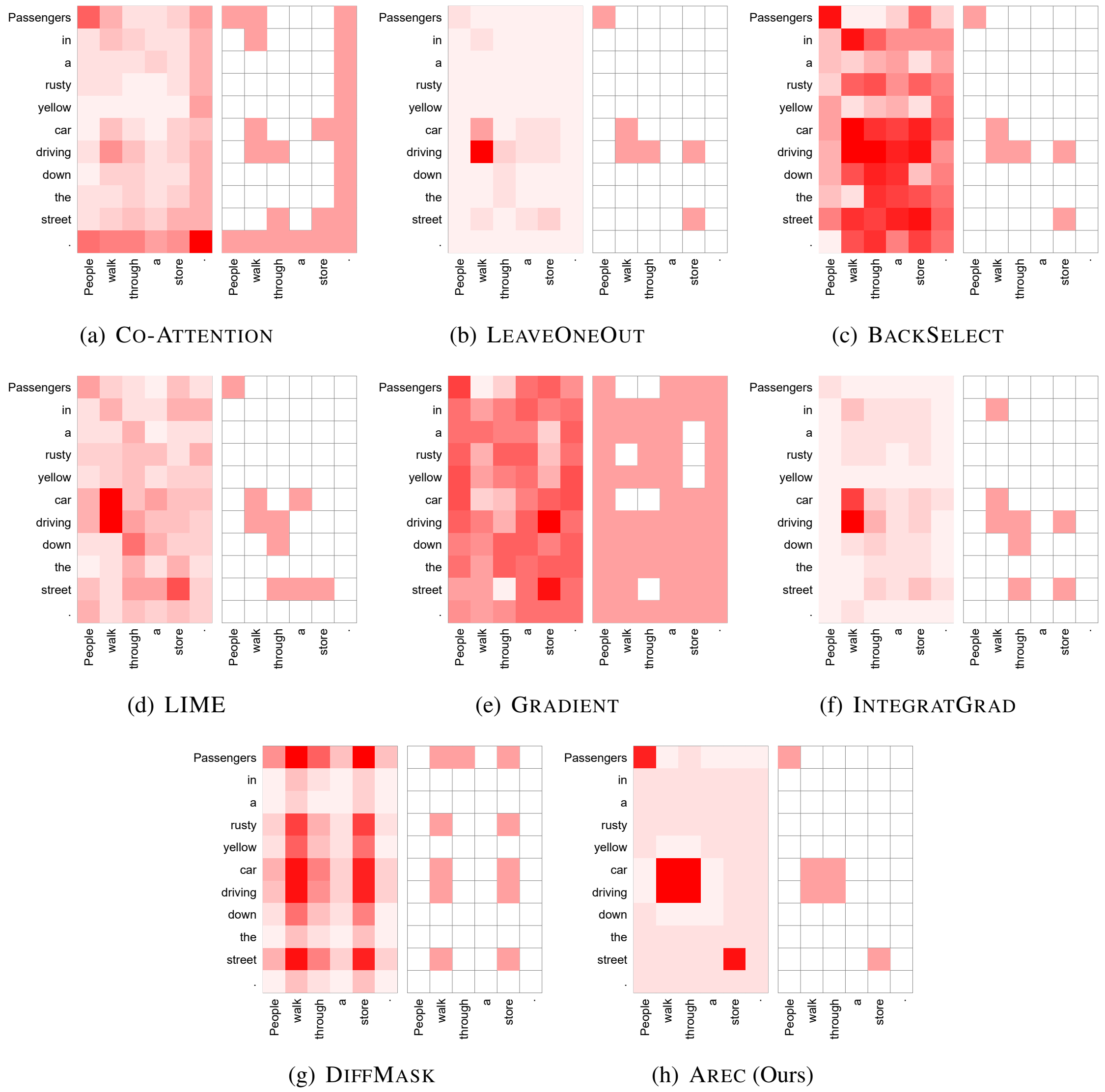


(a) DA

(b) ESIM

(c) BERT

Figure 7: AREC Explanation examples.



(a) CO-ATTENTION (b) LEAVEONEOUT (c) BACKSELECT

(d) LIME (e) GRADIENT (f) INTEGRATGRAD

(g) DIFFMASK (h) AREC (Ours)

Figure 8: Visualization of different alignment explanations. All the explanations are generated from BERT. For attribution explanations (a) - (f), we plot attribution maps (left) and induced rationales (right). For rationale explanations (g) and (h), we plot parameters α (left) and rationales (right).