Generating Token-Level Explanations for Natural Language Inference

James Thorne
University of Cambridge
jt719@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Christos Christodoulopoulos
Amazon
chrchrs@amazon.co.uk

Arpit Mittal
Amazon
mitarpit@amazon.co.uk

Abstract

The task of Natural Language Inference (NLI) is widely modeled as supervised sentence pair classification. While there has been a lot of work recently on generating explanations of the predictions of classifiers on a single piece of text, there have been no attempts to generate explanations of classifiers operating on pairs of sentences. In this paper, we show that it is possible to generate token-level explanations for NLI without the need for training data explicitly annotated for this purpose. We use a simple LSTM architecture and evaluate both LIME and Anchor explanations for this task. We compare these to a Multiple Instance Learning (MIL) method that uses thresholded attention make token-level predictions. The approach we present in this paper is a novel extension of zero-shot single-sentence tagging to sentence pairs for NLI. We conduct our experiments on the well-studied SNLI dataset that was recently augmented with manually annotation of the tokens that explain the entailment relation. We find that our white-box MIL-based method, while orders of magnitude faster, does not reach the same accuracy as the black-box methods.

1 Introduction

Large-scale datasets for Natural Language Inference (NLI) (Bowman et al., 2015; Williams et al., 2018) have enabled the development of many deep-learning models (Rocktäschel et al., 2016; Peters et al., 2018; Radford et al., 2018). The task is modeled as 3-way classification of the entailment relation between a pair of sentences. Model performance is assessed through accuracy on a held-out test set. While state-of-the-art models achieve high accuracy, their complexity makes it difficult to interpret their behavior.

Explaining the predictions made by classifiers has been of increasing concern (Doshi-Velez and Kim, 2017). It has been studied in natural language processing through both black-box analysis, and through modifications to the models under investigation; we refer to the latter approaches as *white-box*. Common black-box techniques generate explanations of predictions through training meta-models by perturbing input tokens (Ribeiro et al., 2016; Nguyen, 2018; Ribeiro et al., 2018) or through interpretation of model sensitivity to input tokens (Li et al., 2016; Feng et al., 2018). White-box methods induce new features (Aubakirova and Bansal, 2016), augment models to generate explanations accompanying their predictions (Lei et al., 2016; Camburu et al., 2018), or expose model internals such as magnitude of hidden states (Linzen et al., 2016), gradients (as a proxy for model sensitivity to input tokens (Li et al., 2016)) or attention (Bahdanau et al., 2014; Xu et al., 2015).

Premise:	Children smiling and waving at a camera
Hypothesis:	The kids are frowning
Label:	Contradiction

Figure 1: Example of token-level highlights from the e-SNLI dataset (Camburu et al., 2018). Annotators were provided a premise and hypothesis and asked to highlight words considered essential to explain the label.

Model explanations typically comprise a list of features (such as tokens) that contributed to the prediction and can serve two distinct purposes: acting either as a diagnostic during model development or to allow for a rationale to be generated for a system user. While methods for explaining predictions may output what was salient to the model, there is no guarantee these will correspond to the features that users deem important.

In this paper we introduce a white-box method that thresholds the attention matrix of a neural entailment model to induce token-level explanations.

Proceedings of NAACL-HLT 2019, pages 963–969
Minneapolis, Minnesota, June 2 - June 7, 2019.

To encourage the model's prediction of salient tokens to correspond better to the tokens users would find important, our approach uses Multiple Instance Learning (MIL) (Maron and Lozano-Pérez, 1998) to regularize the attention distributions.

We compare this against two black-box methods: LIME (Ribeiro et al., 2016) and Anchor Explanations (Ribeiro et al., 2018); both white- and black-box methods are applied to a simple neural architecture relying on independent sentence encoding with cross-sentence attention, and thus could also be applied to more complex architectures of the same family. Finally, we also compare against a fully supervised baseline trained to jointly predict entailment relation and token-level explanations. Our experiments are conducted on e-SNLI (Camburu et al., 2018), a recently introduced extension to SNLI (Bowman et al., 2015), containing human-selected highlights of which words are required to explain the entailment relation between two sentences (see Fig. 1).

Our experimental results indicate that regularizing the model's attention distributions encourages the explanations generated to be better aligned with human judgments (even without our model having explicit access to the labels which tokens annotators found important). Compared to the baseline thresholded attention mechanism, our method provides an absolute increase in token-level precision and recall by 6.68% and 28.05% respectively for the hypothesis sentence for e-SNLI explanations.

We also found that attention based explanations are not as reliable as black-box model explanation techniques, as indicated by higher F_1 scores for both LIME and Anchor Explanations. This is consistent with findings of contemporaneous work by Jain and Wallace (2019). However, we do show that, if generating explanations from a model is a requirement, incorporating an explicit objective in training can be beneficial. This can be particularly useful in practicw due to the computational cost of black-box model explanations, which in empirical evaluation we found to be orders of magnitude slower (0.01 seconds vs 64 seconds per instance).

2 NLI Model

The model we use for both white- and black-box experiments is based on an architecture widely adopted for sentence-pair classification (Lan and Xu, 2018). It comprises the following:

Word Embeddings We use pretrained GloVe embeddings (Pennington et al., 2014) that were fixed during training.

Sentence Encoding Both the premise and hypothesis are independently encoded with the same LSTM (Hochreiter and Schmidhuber, 1997), yielding $\mathbf{h}^p$ and $\mathbf{h}^h$ respectively.

Attention A matrix of soft alignments between tokens in the premise sentence and the hypothesis sentence is computed using attention (Bahdanau et al., 2014) over the encodings. Like Parikh et al. (2016), we project the encoded sentence representations using a feed-forward network, f_{attend}, $(u_i = f_{attend}(h_i^p),\ v_j = f_{attend}(h_j^h))$ before computing the inner product: $\tilde{A}_{ij} = u_i^T v_j$. Given a premise of length m, the attention distribution for the hypothesis sentence is $\mathbf{a}^h = \text{normalize}(\tilde{A}_{m,*})$ where linear normalization is applied ($\text{normalize}(w) = \frac{w}{\|w\|_1}$). Likewise for the corresponding hypothesis of length n, the premise attention distribution is $\mathbf{a}^p = \text{normalize}(\tilde{A}_{*,n})$.

Output Classifier We predict the class label through a feed-forward neural network, f_{cls}, where both attended encodings of the premise and hypothesis final hidden states are concatenated as input: $f_{cls}([a_m^p h_m^p; a_n^h h_n^h])$. The logits are normalized using the softmax function, yielding a distribution over class labels $\hat{y}$.

Training The model is trained in a supervised environment using cross-entropy loss between the predicted class labels for an instance $\hat{y}$ and the labeled value in the dataset, formally defined in Section 3.

3 Generating Token-Level Explanations

Let $\mathbf{x}^p = (x_1^p, \ldots, x_m^p)$ and $\mathbf{x}^h = (x_1^h, \ldots, x_n^h)$ be sequences of tokens of length m and n respectively for the input premise and hypothesis sentences. Let y represent an entailment relation between $\mathbf{x}^p$ and $\mathbf{x}^h$ where $y \in \{\text{entails}, \text{contradicts}, \text{neutral}\}$. Labeled training data is provided of the form $\{(\mathbf{x}_k^p, \mathbf{x}_k^h, y_k)\}_{k=1}^K$. For each instance, the model must generate an explanation $\mathbf{e}$ defined as a subset of zero or more tokens from both the premise and hypothesis sentences: $\mathbf{e}^p \in \mathcal{P}(\mathbf{x}^p)$, $\mathbf{e}^h \in \mathcal{P}(\mathbf{x}^h)$.

We generate token-level explanations by thresholding token attention weights. Concretely, we select all tokens, x, with a weight greater than

a threshold. While similar to Rei and Søgaard (2018), we incorporate a re-scaling using the tanh function: $\mathbf{e}^p = \{x_i^p | \tilde{a}_i^p \in \tilde{A}_{*,n} \wedge \tanh(\tilde{a}_i^p) \geq \tau\}$ and likewise for the hypothesis.

3.1 Multiple Instance Learning

Thresholding the attention distributions from our model will give an indication of which tokens the model is weighting strongly for the entailment task. However, as mentioned in the introduction, there is no guarantee that this method of explaining model behavior will correspond with tokens that humans judge as a reasonable explanation of entailment. To better align the attention-based explanations with the human judgments, we model the generation of explanations as Multiple Instance Learning (MIL) (Maron and Lozano-Pérez, 1998). In training the model sees labeled "bags" (sentences) of unlabeled features (tokens) and learns to predict labels both for the bags and the features. In MIL, this is often achieved by introducing regularizers when training. To encourage our NLI model to predict using sparser attention distributions (which we expect to correspond more closely with human token-level explanations), we introduce the following regularizers into the loss function:

R_1**:** This entropy-based term allows us to penalize a model that uniformly distributes probability mass between tokens.

$$\begin{aligned} R_1 &= \sum_{k=1}^{K} \left(\mathbb{H}(\mathbf{a}_k^p) + \mathbb{H}(\mathbf{a}_k^h)\right) \\ &= -\sum_{k=1}^{K}\left(\sum_{i=1}^{m} a_{k,i}^p \log a_{k,i}^p + \sum_{j=1}^{n} a_{k,j}^h \log a_{k,j}^h\right) \end{aligned} \tag{1}$$

R_2**:** This term, adapted from a loss function for zero-shot tagging on single sentences (Rei and Søgaard, 2018), penalizes the model for breaking the assumption that at least one token must be selected from both premise and hypothesis sentences to form an explanation. The only exception is that, following the e-SNLI dataset annotation by Camburu et al. (2018), if the neutral entailment is predicted, no tokens are selected from the premise.

$$\begin{aligned} R_2 = \sum_{k=1}^{K} \big((\max_i a_{k,i}^p - \mathbb{I}[k_c \neq \text{neutral}])^2 \\ +(\max_j a_{k,j}^h - 1)^2\big) \end{aligned} \tag{2}$$

R_3**:** This term, also adapted from Rei and Søgaard (2018), encodes the assumption that not all tokens must be selected in the explanation. This is achieved by penalizing the smallest non-zero attention weight, which has the effect of encouraging at least one weight to be close to zero.

$$R_3 = \sum_{k=1}^{K} \left((\min_i a_{k,i}^p)^2 + (\min_j a_{k,j}^h)^2\right) \tag{3}$$

The loss function used for training of our proposed model incorporating the regularizers which are controlled with hyperparameters is:

$$L = \sum_{k=1}^{K}\sum_{c \in C} y_{k,c} \log \hat{y}_{k,c} + \alpha R_1 + \beta R_2 + \gamma R_3 \tag{4}$$

4 Alternative Models

4.1 Black-box explanations of NLP models

We use two established black-box model explanation techniques for generating token-level explanations: LIME (Ribeiro et al., 2016) and Anchors (Ribeiro et al., 2018). Both techniques probe a classifier by making perturbations to a single input and modeling which of these perturbations influence the classification. To adapt these for use in NLI, we make a simple modification that runs the analysis twice: once for the premise sentence and once for the hypothesis sentence on the NLI model described in Section 2.

LIME Generates local explanations for a classifier through the introduction of a simple meta-model that is trained to replicate a local decision boundary of an instance under test. The training data is generated through observing the impact on classification when removing tokens from the input string.

Anchor Explanations Considers the distribution of perturbed instances in the neighborhood of the instance under test through word substitution to identify a rule (a set of tokens in our case) for which the classification remains unchanged.

4.2 Supervised Model

For a supervised model we build upon the model discussed in Section 2, adding components to support LSTM-CRF-based tagging (Lample et al., 2016). We use the following architecture:

Model	Runtime (s) per instance	Token Explanation (%)					
		Premise			Hypothesis		
		P	R	F1	P	R	F1
Fully Supervised LSTM-CRF	0.02	86.91	40.98	55.70	81.16	54.79	65.41
Thresholded Attention (Linear)	**0.01**	19.96	19.67	19.56	46.70	34.92	39.89
+ MIL Regularizers (R1)	-	16.59	15.67	16.12	50.02	42.44	46.01
+ MIL Regularizers (R2 + R3)	-	18.19	20.18	19.13	51.29	50.73	51.00
+ MIL Regularizers (R1 + R2 + R3)	-	19.23	26.21	22.18	53.38	62.97	57.78
LIME	64	**60.56**	**48.28**	**53.72**	**57.04**	**66.92**	**61.58**
Anchors	10	42.06	20.04	27.14	53.12	63.94	58.03

Table 1: Token-level scores for human-selected explanations of NLI using the e-SNLI dataset. The select-all baseline precision for the premise is 18.5% and 35.2% for the hypothesis.

Context Encoding We use the same pretrained GloVe embeddings (Pennington et al., 2014) that were fixed during training. The premise and hypothesis sentence were independently encoded with the same LSTM (Hochreiter and Schmidhuber, 1997) yielding $\mathbf{h}^p$ and $\mathbf{h}^h$ respectively and attended to as per the description in Section 2.

Outputs The model is jointly trained with two output objectives: a labeling objective and a tagging objective. During training, the losses for both tasks are equally weighted. The first output objective is the three-way SNLI classification over the pair of sentences. This is the same component as the model presented in Section 2.

The second objective is a binary tagging objective over the highlighted token-level explanations. We use a jointly-trained LSTM-CRF decoder architecture (Lample et al., 2016) which operates a CRF over encoded representations for each token. In our model, we independently decode the premise and hypothesis sentences. The inputs to our CRF are the attended premise and hypothesis: $\mathbf{a}^p \odot \mathbf{h}^p$ and $\mathbf{a}^h \odot \mathbf{h}^h$ respectively (where $\odot$ is the point-wise multiplication between the attention vector and the encoded tokens).

5 Experiments

We evaluate the generated explanations through evaluation of token-level F_1 scores comparing them against tokens selected by humans to explain the entailment relation using the e-SNLI dataset (Camburu et al., 2018). The development split of the e-SNLI dataset is used for hyperparameter selection and we report results on the test split. Where multiple annotations are available for a sentence pair, the union of the annotations is taken. We also report average runtime per sentence in seconds measured using 1 thread on an AWS c4.xlarge instance.

Implementation Details The model is implemented in AllenNLP (Gardner et al., 2018) and we optimized our model with Adagrad (Duchi et al., 2011), selecting the models which attained high hypothesis F_1 without greatly affecting the accuracy of entailment task (approx 81% for the thresholded attention model). The cell state and hidden dimension was 200 for the LSTM sentence encoder. The projection for attention, f_{attend}, was a single layer 200 dimension feed forward network with ReLU activation. The final feed forward classifier, f_{cls}, dimension was $(200, 200, 3)$ and ReLU activation over the first 2 layers. For the comparison against black-box explanation mechanisms, we use the code made public by the authors of the respective works setting any hyperparameters to the default values or those suggested in the papers.

Results Our experimental results (Table 1) indicate that the LIME black-box explanation technique over the model described in Section 2 provides token-level explanations that are more similar to human judgments than thresholding the attention distributions. We show that the addition of MIL regularizers for generating explanations using thresholded attention improved precision and recall hypothesis explanations. However, similar improvements were not realized for the premise sentence. While the black-box methods generated better explanations than thresholded attention, they were 3 orders of magnitude slower.

Only LIME was able to generate good token-level explanations for the premise. This is in contrast to the attention-based explanations of the premise (in the model that LIME was run on) which could not generate satisfactory explanations (see row 2 of Table 1). This supports findings in recent works (Jain and Wallace, 2019) that indicate that attention does not always correspond to other measures of feature importance. We also found that the black-box model explanation methods behave differently given the same model under test: the premise explanation generated by the Anchors method was more in line with what the model attended to, reflected by the lower recall.

The fully supervised model had high precision yet (relatively) low recall. We observed it has a bias towards predicting common words that often appear in highlights (e.g. 'man', 'woman', 'dog', 'people') for both premise and hypothesis sentences rather than highlighting keywords that would form an instance-specific explanation. This behaviour is also more pronounced in the premise sentence highlights rather than the hypothesis. We reason that this due to how the SNLI dataset was constructed: a premise sentence was used to generate 3 hypothesis sentences (entailed, contradicted and neutral). This is corroborated by a survey of 250 instances from the SNLI dataset, where we found that all or part of the subject noun phrase remained unchanged between the premise and hypothesis sentences 60% of the time. While the supervised model correctly captured commonly occurring text patterns, as demonstrated by the high F_1 scores, this behaviour alone was not sufficient to identify tokens that correlated with the entailment label. We found that most of the commonly predicted tokens by our supervised model did not appear in lists of features highly correlated with the entailment label (Poliak et al., 2018; Gururangan et al., 2018).

6 Conclusions

In this paper we explored how to generate token-level explanations from NLI models. We compared the LIME and Anchors black-box methods against a novel, white-box Multiple Instance Learning (MIL) method and a fully supervised baseline. The explanations generated by LIME were more similar to the human judgments of the tokens that justify an entailment relation than the attention thresholding approach. This corroborates contemporaneous work (Jain and Wallace, 2019) indicating a lack of correspondence between attention and other measures of feature importance.

The MIL method we introduced steered the attention distributions over tokens in our model to correspond closer to the human judgments allowing better explanations to be generated. Even though, when considering the token-level F_1 score, the attention-based explanations were not as good as the black-box techniques we evaluated, they were orders of magnitude faster.

The attention thresholding model we tested did not generate satisfactory explanations had low F_1 for the premise sentences. A possible explanation for the poor performance is what is found by Rei and Søgaard (2018) who show that MIL regularizers performed better when there is a higher degree of association between the sentence-level label and the token-level labels. Our model used *independent* encodings of the premise and hypothesis but in NLI there is a strong dependence between the two sentences; thus the entailment prediction should be explained through pairwise token comparisons (e.g. synonyms, upward entailment, etc.). In future work we plan to address this by adding explicit cross-sentence semantic knowledge (Joshi et al., 2018).

Acknowledgements

This work was conducted while James Thorne was an intern at Amazon. Andreas Vlachos is supported by the EU H2020 SUMMA project (grant agreement number 688139).

References

Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. pages 2035–2041.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Oana-maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI : Natural Language Inference with Natural Language Explanations. In *32nd Conference on Neural Information Processing Systems (NIPS)*, pages 1–13.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Shi Feng, Eric Wallace, Alvin Grissom, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2018. pair2vec: Compositional word-pair embeddings for cross-sentence inference. *arXiv preprint arXiv:1810.08854*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902. 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. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. *arXiv preprint arXiv:1612.08220*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Oded Maron and Tomás Lozano-Pérez. 1998. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576.

Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. *Naacl*, pages 1069–1078.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. pages 2249–2255.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, Louisiana.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. (1):180–191.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *arXiv*, pages 1–12.

Marek Rei and Anders Søgaard. 2018. Zero-shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens. pages 293–302.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 39(2011):117831.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proc. of 32nd Conference on Artificial Intelligence (AAAI)*, pages 1527–1535.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about Entailment with Neural Attention. *Iclr*, (2015):1–9.

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. Association for Computational Linguistics.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org.