e-SNLI: Natural Language Inference with Natural Language Explanations

Oana-Maria Camburu[1] **Tim Rocktäschel**[2] **Thomas Lukasiewicz**[1,3] **Phil Blunsom**[1,4]

{oana-maria.camburu, thomas.lukasiewicz, phil.blunsom}@cs.ox.ac.uk
t.rocktaschel@ucl.ac.uk
[1]Department of Computer Science, University of Oxford
[2]Department of Computer Science, University College London
[3]Alan Turing Institute, London, UK
[4]DeepMind, London, UK

Abstract

In order for machine learning to garner widespread public adoption, models must be able to provide interpretable and robust explanations for their decisions, as well as learn from human-provided explanations at train time. In this work, we extend the Stanford Natural Language Inference dataset with an additional layer of human-annotated natural language explanations of the entailment relations. We further implement models that incorporate these explanations into their training process and output them at test time. We show how our corpus of explanations, which we call e-SNLI, can be used for various goals, such as obtaining full sentence justifications of a model's decisions, improving universal sentence representations and transferring to out-of-domain NLI datasets. Our dataset[1] thus opens up a range of research directions for using natural language explanations, both for improving models and for asserting their trust.

1 Introduction

Humans do not learn solely from labeled examples supplied by a teacher. Instead, they seek a conceptual understanding of a task through both demonstrations and explanations. Machine learning models trained simply to obtain high accuracy on held-out sets often learn to rely heavily on shallow input statistics, resulting in brittle models susceptible to adversarial attacks. For example, Ribeiro et al. [24] present a document classifier that distinguishes between *Christianity* and *Atheism* with an accuracy of 94%. However, on close inspection, the model spuriously separates classes based on words contained in the headers, such as *Posting*, *Host*, and *Re*.

In this work, we introduce a new dataset and models for exploiting and generating explanations for the task of recognizing textual entailment. We argue for free-form natural language explanations, as opposed to formal language, for a series of reasons. First, natural language is readily comprehensible to an end-user who needs to assert a model's reliability. Secondly, it is also easiest for humans to provide free-form language, eliminating the additional effort of learning to produce formal language, thus making it simpler to collect such datasets. Lastly, natural language justifications might eventually be mined from existing large-scale free-form text.

Despite the potential for free-form justifications to improve both learning and transparency, there is currently a lack of such datasets in the machine learning community. To address this deficiency, we have collected a large corpus of human-annotated explanations for the Stanford Natural Language

[1]https://github.com/OanaMariaCamburu/e-SNLI

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

Figure 1: Examples from e-SNLI. Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.

Inference (SNLI) dataset [3]. We chose SNLI because it constitutes an influential corpus for natural language understanding that requires deep assimilation of fine-grained nuances of common-sense knowledge. We call our explanation-augmented dataset e-SNLI, which we collected to enable research in the direction of training with and generation of free-form textual justifications.

In order to demonstrate the efficacy of the e-SNLI dataset, we first show that it is much more difficult to produce correct explanations based on spurious correlations than to produce correct labels. We then implement models that, given a premise and a hypothesis, predict a label and an explanation. We also investigate how the additional signal from explanations received at train time can guide models into learning better sentence representations. Finally, we look into the transfer capabilities of our model to out-of-domain NLI datasets.

2 Background

The task of recognizing textual entailment is a critical natural language understanding task. Given a pair of sentences, called the premise and hypothesis, the task consists of classifying their relation as either (a) *entailment*, if the premise entails the hypothesis, (b) *contradiction*, if the hypothesis contradicts the premise, or (c) *neutral*, if neither entailment nor contradiction hold. The SNLI dataset [3], containing $570K$ data points of human-generated triples (premise, hypothesis, label), has driven the development of a large number of neural network models [25, 21, 22, 6, 19, 5, 7].

Conneau et al. [7] showed that training universal sentence representations on SNLI is both more efficient and more accurate than the traditional training approaches on orders of magnitude larger, but unsupervised, datasets [17, 14]. We take this approach one step further and show that an additional layer of explanations on top of the label supervision brings further improvement.

Recently, Gururangan et al. [13] cast doubt on whether models trained on SNLI are learning to understand language, or are largely fixating on spurious correlations, also called artifacts. For example, specific words in the hypothesis tend to be strong indicators of the label, e.g., *friends, old* appear very often in neutral hypotheses, *animal, outdoors* appear most of the time in entailment hypotheses, while *nobody, sleeping* appear mostly in contradiction hypothesis. They show that a premise-agnostic model, i.e., a model that only takes as input the hypothesis and outputs the label, obtains 67% test accuracy. In section 4.1 we show that it is much more difficult to rely on artifacts to generate explanations than to generate labels.

3 Collecting explanations

We present our collection methodology for e-SNLI, for which we used Amazon Mechanical Turk. The main question that we want our dataset to answer is: *Why is a pair of sentences in a relation of entailment, neutrality, or contradiction?* We encouraged the annotators to focus on the non-obvious

elements that induce the given relation, and not on the parts of the premise that are repeated identically in the hypothesis. For entailment, we required justifications of all the parts of the hypothesis that do not appear in the premise. For neutral and contradictory pairs, while we encouraged stating all the elements that contribute to the relation, we consider an explanation correct, if at least one element is stated. Finally, we asked the annotators to provide self-contained explanations, as opposed to sentences that would make sense only after reading the premise and hypothesis. For example, we prefer an explanation of the form "*Anyone can knit, not just women.*", rather than "*It cannot be inferred they are women.*"

In crowd-sourcing, it is difficult to control the quality of free-form annotations. Thus, we aimed to preemptively block the submission of obviously incorrect answers. We did in-browser checks to ensure that each explanation contained at least three tokens and that it was not a copy of the premise or hypothesis. We further guided the annotators to provide adequate answers by asking them to proceed in two steps. First, we require them to highlight words from the premise and/or hypothesis that they consider essential for the given relation. Secondly, annotators had to formulate the explanation using the words that they highlighted. However, using exact spelling might push annotators to formulate grammatically incorrect sentences, therefore we only required half of the highlighted words to be used with the same spelling. For entailment pairs, we required at least one word in the premise to be highlighted. For contradiction pairs, we required highlighting at least one word in both the premise and the hypothesis. For neutral pairs, we only allowed highlighting words in the hypothesis, in order to strongly emphasize the asymmetry in this relation and to prevent workers from confusing the premise with the hypothesis. We believe these label-specific constraints helped in putting the annotator into the correct mindset, and additionally gave us a means to filter incorrect explanations. Finally, we also checked that the annotators used other words that were not highlighted, as we believe a correct explanation would need to articulate a link between the keywords.

We collected one explanation for each pair in the training set and three explanations for each pair in the validation and test sets. Figure 1 shows examples of collected explanations. There were 6325 workers with an average of 86 explanations per worker and a standard deviation of 403.

Analysis and refinement of the collected dataset In order to measure the quality of our collected explanations, we selected a random sample of 1000 examples and manually graded their correctness between 0 (incorrect) and 1 (correct), giving partial scores of k/n if only k out of n required arguments were mentioned. We also considered an explanation as incorrect if it was uninformative, that is, if the explanation was template-like, extensively repeating details from the premise/hypothesis that are not directly useful for justifying the relation between the two sentences. We observed a few re-occurring templates such as: "*Just because [entire premise] doesn't mean [entire hypothesis]*" for neutral pairs, "*[entire premise] implies [entire hypothesis]*" for entailment pairs, and "*It can either be [entire premise] or [entire hypothesis]*" for contradiction pairs. We assembled a list of templates, which can be found in Appendix A, that we used for filtering the dataset of such uninformative explanations. Specifically, we filtered an explanation if its edit distance to one of the templates was less than 10. We ran this template detection on the entire dataset and reannotated the detected explanations (11% in total).

Our final counts show a total error rate of 9.62%, with 19.55% on entailment, 7.26% on neutral, and 9.38% on contradiction. We notice that entailment pairs were by far the most difficult to obtain proper explanations for. This is firstly due to partial explanations, as annotators had an incentive to provide shorter inputs, so they often only mentioned one argument. A second reason is that many of the entailment pairs have the hypothesis as almost a subset of the premise, prompting the annotators to just repeat that as a statement.

4 Experiments

We first present an experiment which demonstrates that a model which can easily rely on artifacts in SNLI to provide correct labels would not be able to provide correct explanations as easily. We refer to it as PREMISEAGNOSTIC.

We then present a series of experiments to elucidate whether models trained on e-SNLI are able to: (i) predict a label and generate an explanation for the predicted label (referred to as PREDICTANDEXPLAIN), (ii) generate an explanation then predict the label given only the generated explanation

(EXPLAINTHENPREDICT), (iii) learn better universal sentence representations (REPRESENT), and (iv) transfer to out-of-domain NLI datasets (TRANSFER).

Throughout our experiments, our models follow the architecture presented in Conneau et al. [7], as we build directly on top of their code[2]. Therefore, our encoders are 2048-bidirectional-LSTMs [15] with max-pooling, resulting in a sentence representation dimension of 4096. Our label classifiers are 3-layers MLPs with 512 internal size and without non-linearities. For our explanation decoders, we used a simple one-layer LSTM, for which we tried internal sizes of 512, 1024, 2048, and 4096. In order to reduce the vocabulary size for explanation generation, we replaced words that appeared less than 15 times[3] with `<UNK>`. We obtain an output vocabulary of approximately $12K$ words. The preprocessing and optimization were kept the same as in [7].

Whenever appropriate, we run our models with five seeds and provide the average performance with the standard deviation in parenthesis. If no standard deviation is reported, the results are from one experiment with seed 1234.

4.1 PREMISEAGNOSTIC: Generate an explanation given only the hypothesis

Gururangan et al. [13] show that a neural network that only has access to the hypothesis can predict the correct label 67% of the times. We are therefore interested in evaluating how well our explanations can be predicted from hypotheses alone.

Model We train a 2048-bidirectional-LSTM with max-pooling for encoding the hypothesis, followed by a one-layer LSTM for decoding the explanation. The initial state of the decoder is the vector embedding of the hypothesis, which is also concatenated at every timestep of the decoder, to avoid forgetting.

Selection We consider internal sizes of the decoder of 512, 1024, 2048 and 4096. We pick the model that gives the best perplexity on the validation set. We notice that the perplexity strictly decreases when we increase the decoder size. However, for practical reasons, we do not increase the decoder size beyond 4096.

Results We then manually look at the first 100 test examples and obtain that only 6.83[4] were correct. We also separately train the same hypothesis-only encoder for label prediction alone and obtain 66 correct labels in the same first 100 test examples. This validates our intuition that it is much more difficult (approx. 10x for this architecture) to rely on spurious correlations to predict correct explanations than to predict correct labels.

4.2 PREDICTANDEXPLAIN: Jointly predict a label and generate an explanation for the predicted label

In this experiment, we investigate how the typical architecture employed on SNLI can be enhanced with a module that aims to justify the decisions of the entire network.

Model We employ the InferSent [7] architecture, where a bidirectional-LSTM with max-pooling separately encodes the premise, $\boldsymbol{u}$, and hypothesis, $\boldsymbol{v}$. The vector of features $\boldsymbol{f} = [\boldsymbol{u}, \boldsymbol{v}, |\boldsymbol{u}-\boldsymbol{v}|, \boldsymbol{u}\odot\boldsymbol{v}]$ is then passed to the MLP classifier that outputs a distribution over the 3 labels. We add a one-layer LSTM decoder for explanations, which takes the feature vector $\boldsymbol{f}$ both as an initial state and concatenated to the word embedding at each time step.

In order to condition the explanation also on the label, we prepend the label as a word (*entailment, contradiction, neutral*) at the beginning of the explanation. At training time, the gold label is provided, while at test time, we use the label predicted by the classifier. This architecture is depicted in Figure 2.

Loss We use negative log-likelihood for both classification and explanation losses. The explanation loss is much larger in magnitude than the classification loss, due to the summation of negative log-likelihoods over the words in the explanations. To account for this difference during training, we

[2]https://github.com/facebookresearch/InferSent. We fixed the issue raised in https://github.com/ facebookresearch/InferSent/issues/51 that the max-pooling was taken over paddings.

[3]Counted among premises, hypothesis, and explanations.

[4]Partial scoring as explained in Section 3.

use a weighting coefficient $\alpha \in [0, 1]$. Hence, our overall loss is:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{label}} + (1 - \alpha)\mathcal{L}_{\text{explanation}} \quad (1)$$

Selection We consider α values from 0.1 to 0.9 with a step of 0.1 and decoder internal sizes of 512, 1024, 2048, and 4096. For this experiment, we choose as model selection criterion the accuracy on the SNLI validation set, because we want to investigate how well a model can generate justifications without sacrificing accuracy. As future work, one can inspect different trade-offs between accuracy and explanation generation. We found $\alpha = 0.6$ and the decoder size of 512 to produce the best validation accuracy, of 84.37%, while InferSent with no explanations produced 84.30% validation accuracy. We call our model e-INFERSENT, since it freezes the InferSent architecture and training procedure, and only adds the explanations decoder.

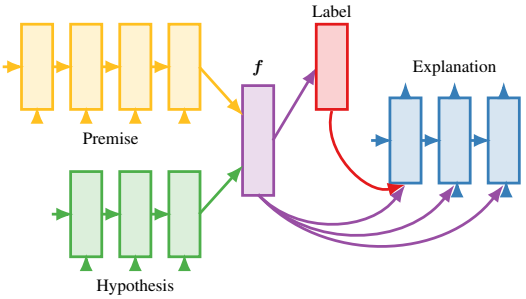

Figure 2: Overview of the e-INFERSENT architecture.

Results The average test accuracy that we obtain when training InferSent[7] on SNLI with five seeds is 84.01% (0.25). Our e-INFERSENT model obtains essentially the same test accuracy, of 83.96% (0.26), which shows that one can get additional justifications without sacrificing label accuracy. For the generated explanations, we obtain a perplexity of $10.58(0.4)$ and a BLEU-score of $22.40(0.7)$. Since we collected 3 explanations for each example in the validation and test sets, we compute the inter-annotator BLEU-score of the third explanation with respect to the first two, and obtain 22.51. For consistency, we used the same two explanations as the only references when computing the BLEU-score for the predicted explanations. Given the low inter-annotator score and the fact that generated explanations almost match the inter-annotator BLEU-score, we conclude that this measure is not reliable for our task, and we further rely on human evaluation. Therefore, we manually annotated the first 100 datapoints in the test set (we used the same partial scoring as in Section 3). Since the explanation is conditioned on the predicted label, for incorrect labels, the model would not produce a correct explanation. Therefore, we provide as correctness score the percentage of correct explanations in the subset of the first 100 examples where the predicted label was correct (80 in this experiment). We obtain a percentage of 34.68% correct explanations. While this percentage is low, we keep in mind that the selection criteria was only the accuracy of the label classifier and not the perplexity of the explanation. In the next experiment, we show how training (and selecting) only for generating explanations results in higher quality explanations.

4.3 EXPLAINTHENPREDICT: Generate an explanation then predict a label

In PREDICTANDEXPLAIN, we conditioned the explanation on the label predicted by the MLP, because we wanted to see how the typical architecture used on SNLI can be adapted to justify its decisions in natural language. However, a more natural approach for solving inference is to think of the explanation first and based on the explanation to decide a label. Therefore, in this experiment, we first train a network to generate an explanation given a pair of (premise, hypothesis), and, separately, we train a network to provide a label given an explanation. This is a sensible decomposition for our dataset, due to the following key observation: In our dataset, in the large majority of the cases, one can easily detect for which label an explanation has been provided. We highlight that this is not the case in general, as the same explanation can be correctly arguing for different labels, depending on the premise and hypothesis. For example, the explanation *"A woman is a person"* would be a correct explanation for the entailment pair *("A woman is in the park", "A person is in the park")* as well for the contradiction pair *("A woman is in the park", "There is no person in the park")*. However, there are multiple ways of formulating an explanation. In our example, for the contradiction pair, one could also explain that *"There cannot be no person in the park if a woman is in the park"*, which read alone would allow one to infer that the pair was a contradiction. To support our observation, we train a neural network that given only an explanation predicts a label. We use the same bidirectional encoder and MLP-classifier as above. We obtain an accuracy of 96.83% on the test set of SNLI.

Table 1: Summary of the performance of the models PREDICTANDEXPLAIN, EXPLAINTHENPREDICT. The averages are over five seeds and the standard deviations are in parenthesis. The baseline InferSent-SNLI-ours is the InferSent model [7] that we re-run with 5 seeds. ExplCorrect@100 is the score of correctness for the generated explanations, that we manually annotated for the first 100 data points in the SNLI test set for one seed.

Model	Label Accuracy	Perplexity	BLEU	ExplCorrect@100
InferSent-SNLI-ours	84.01 (0.25)	-	-	-
e-INFERSENT	83.96 (0.26)	10.58 (0.4)	22.40 (0.7)	34.68
EXPLAINTHENPREDICTSEQ2SEQ	81.59 (0.45)	8.95 (0.03)	24.14 (0.58)	49.8
EXPLAINTHENPREDICTATTENTION	81.71 (0.36)	6.1 (0)	27.58 (0.47)	64.27

Models For predicting an explanation given a pair of (premise, hypothesis), we first train a simple seq2seq model that we call EXPLAINTHENPREDICTSEQ2SEQ. Essentially, we keep the architecture in e-INFERSENT, where we eliminate the classifier by setting $\alpha = 0$, and we decode the explanation without prepending the label. Secondly, we train an attention model, which we refer to as EXPLAINTHENPREDICTATTENTION. Attention mechanisms in neural networks brought consistent improvements over the non-attention counter-parts in various areas, such as computer vision [27], speech [4], or natural language processing [12, 2]. We use the same encoder and decoder as in EXPLAINTHENPREDICTSEQ2SEQ, and we add two identical but separate attention modules, over the tokens in the premise and hypothesis. For details of the attention modules, see Appendix B.

Selection Our only hyper-parameter is internal sizes for the decoder of 512, 1024, 2048, and 4096. Our model selection criterion is the perplexity on the validation set of SNLI. We obtain the best configuration for both EXPLAINTHENPREDICTSEQ2SEQ and EXPLAINTHENPREDICTATTENTION to have an internal size of 1024.

Results With the described setup, the SNLI test accuracy drops from 83.96% (0.26) in PREDICTANDEXPLAIN to 81.59% (0.45) in EXPLAINTHENPREDICTSEQ2SEQ and $81.71\%(0.36)$ in EXPLAINTHENPREDICTATTENTION. However, when we again manually annotate the first 100 generated explanations in the test set, we obtain significantly higher percentages of correct explanations: 49.8% for EXPLAINTHENPREDICTSEQ2SEQ and 64.27% for EXPLAINTHENPREDICTATTENTION. We note that the attention mechanism indeed significantly increases the quality of the explanations. The perplexity and BLEU-score are $8.95(0.03)$ and $24.14(0.58)$ for EXPLAINTHENPREDICTSEQ2SEQ, and $6.1(0)$ and $27.58(0.47)$ for EXPLAINTHENPREDICTATTENTION. Our experiment shows that, while sacrificing a bit of performance, we get a better trust that when EXPLAINTHENPREDICT predicts a correct label, it does so for the right reasons. We summarize the results in table 1.

Qualitative analysis of explanations In Table 2, we provide examples of generated explanations from the test set from (a) PREDICTANDEXPLAIN, (b) EXPLAINTHENPREDICTSEQ2SEQ, and (c) EXPLAINTHENPREDICTATTENTION. At the end of each explanation, we give in brackets the score that we manually allocated as explained in section 3. We notice that the explanations are mainly on topic for all the three models, with minor exceptions, such as the mention of "*camouflage*" in (1c). We also notice that even when incorrect, they are sometimes frustratingly close to being correct, for example, explanation (2b) is only one word (out of its 20 words) away from being correct. It is also interesting to inspect the explanations provided when the predicted label is incorrect. For example, in (1a), we see that the network omitted the information of "*facing the camera*" in the premise and therefore classified the pair as neutral, which is backed up by an otherwise correct explanation in itself. We also see that model EXPLAINTHENPREDICTSEQ2SEQ correctly classifies this pair as entailment, however, it only motivates 1 out the 3 reasons why it is so, and it also picks arguably the easiest reason. Interestingly, the attention model (1c) points to the correct evidence but argues that "*standing*" and "*facing a camera*" is not enough to conclude "*posing for a picture*".

4.4 REPRESENT: Universal sentence representations

Obtaining universal sentence representations is the task of learning an encoder to provide semantically meaningful fixed-length representations of phrases/sentences. These representations are further used as features in other downstream tasks, particularly for tasks where labeled training data are scarce. In computer vision, pretrained ImageNet-based encoders [10] provide standard image feature extractors

Table 2: Examples of predicted labels and generated explanations from (a) PREDICTANDEXPLAIN, (b) EXPLAINTHENPREDICTSEQ2SEQ, and (c) EXPLAINTHENPREDICTATTENTION. In brackets our scores of correctness.

(1) Premise: 3 young man in hoods standing in the middle of a quiet street facing the camera. Hypothesis: Three hood wearing people pose for a picture. Gold label: entailment		
(a) Predicted label: neutral Explanation: Just because the men are in the middle of a street doesn't mean they are posing for a picture. [0]	(b) Predicted label: entailment Explanation: three young men are people. [0.33]	(c) Predicted label: neutral Explanation: Just because three young man in camouflage standing in the middle of a quiet street facing the camera does not mean they pose for a picture. [0]
(2) Premise: Three firefighter come out of subway station. Hypothesis: Three firefighters putting out a fire inside of a subway station. Gold label: neutral		
(a) Predicted label: contradiction Explanation: The firefighters can not be putting out a fire station and putting out a fire at the same time. [0]	(b) Predicted label: neutral Explanation: The fact that three firemen are putting out of a subway station doesn't imply that they are putting out a fire. [0]	(c) Predicted label: neutral Explanation: The firefighters may not be putting out a fire inside of the subway station. [1]
(3) Premise: A blond-haired doctor and her African American assistant looking threw new medical manuals. Hypothesis: A man is eating pb and j. Gold label: contradiction		
(a) Predicted label: contradiction Explanation: A man is not a woman. [1]	(b) Predicted label: contradiction Explanation: One can not be looking and eating simultaneously. [0]	(c) Predicted label: contradiction Explanation: A person can not be looking at a medical and a book at the same time. [0]

for other downstream tasks. However, in natural language processing, there is still no consensus on general-purpose sentence encoders. It remains an open question on which task and dataset should such an encoder be trained. Traditional approaches make use of very large unsupervised datasets [17, 14], taking weeks to train. Conneau et al. [7] showed that training only on NLI is both more accurate and more time-efficient than training on orders of magnitude larger but unsupervised datasets. Their results constitute a previous state-of-the-art for universal sentence representations and encourage the idea that supervision can be more beneficial than larger but unsupervised datasets. We hypothesize that an additional layer of supervision in the form of natural language explanations should further improve learning of universal sentence representations.

Model We use our e-INFERSENT model already trained in PREDICTANDEXPLAIN. While we compare our model with InferSent that has not been trained on explanations, we want to ensure that eventual improvements are not purely due to the addition of a language model in the decoder network. We therefore introduce a second baseline, INFERSENTAUTOENC, where instead of decoding explanations, we decode the premise and hypothesis separately from each sentence representation using one shared decoder.

Evaluation metrics Typically, sentence representations are evaluated by using them as fixed features on top of which shallow classifiers are trained for a series of downstream tasks. Conneau et al. [7] provide an excellent tool, called SentEval, for evaluating sentence representations on 10 diverse tasks: movie reviews (**MR**), product reviews (**CR**), subjectivity/objectivity (**SUBJ**), opinion polarity (**MPQA**), question-type (**TREC**), sentiment analysis (**SST**), semantic textual similarity (**STS**), paraphrase detection (**MRPC**), entailment (**SICK-E**), and semantic relatedness (**SICK-R**). We refer to their work for a more detailed description of each of these tasks and of SentEval, which we use for comparing the quality of the sentence embeddings obtained by additionally providing our explanations on top of the label supervision.

Results In Table 3, we report the average results and standard deviations of e-INFERSENT, our retrained InferSent model, and the additional INFERSENTAUTOENC baseline on the downstream tasks mentioned above. To test if the differences in performance of INFERSENTAUTOENC and e-INFERSENT relative to the InferSent baseline are significant, we performed Welch's t-test.[5] We mark with * the results that appeared significant under the significance level of 0.05.

[5]Using the implementation in `scipy.stats.ttest_ind` with `equal_var=False`.

Table 3: Transfer results on downstream tasks. For MRPC we report accuracy/F1 score, for STS14 we report the Person/Spearman correlations, for SICK-R the Person correlation, and for all the rest their accuracies. Results are the average of 5 runs with different seeds. The standard deviations is shown in brackets, and the best result for every task is indicated in bold. * indicates significant difference at level 0.05 with respect to the InferSent baseline.

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	SICK-E	SICK-R	STS14
InferSent-SNLI-ours	**78.18**	81.28	**92.46**	88.46	**82.12**	89.32	74.82 / 82.74	**85.96**	0.887	0.65 / 0.63
	(0.25)	(0.15)	(0.15)	(0.21)	(0.22)	(0.5)	(0.66 / 0.27)	(0.32)	(0.002)	(0 / 0)
INFERSENTAUTOENC	75.94*	79.26*	91.72*	88.16	80.9*	**90.52***	**76.2*** / 82.48	85.58	0.88*	0.5* / 0.5*
	(0.18)	(0.36)	(0.28)	(0.26)	(0.48)	(0.52)	(0.93 / 1.23)	(0.33)	(0)	(0.02 / 0.02)
e-INFERSENT	77.76	**81.3**	92.14*	**88.78***	81.84	90	75.56 / **83.24***	85.92	**0.89***	**0.68 / 0.65***
	(0.44)	(0.16)	(0.21)	(0.22)	(0.4)	(0.51)	(0.62 / 0.24)	(0.52)	(0)	(0.01 / 0.01)

We notice that INFERSENTAUTOENC is performing significantly worse than InferSent on 6 tasks and significantly outperforms this baseline on only 2 tasks. This indicates that just adding a language generator can harm performance. Instead, e-INFERSENT significantly outperforms InferSent on 4 tasks, while it is significantly outperformed only on 1 task. Therefore, we conclude that training with explanations helps the model to learn overall better sentence representations.

4.5 TRANSFER: Transfer without fine-tuning to out-of-domain NLI

Transfer without fine-tuning to out-of-domain entailment datasets is known to exhibit poor performance. For example, Bowman et al. [3] obtained an accuracy of only 46.7% when training on SNLI and evaluating on SICK-E [20]. We test how our explanations affect the direct transfer in both label prediction and explanation generation by looking at SICK-E [20] and MultiNLI [26]. The latter includes a diverse range of genres of written and spoken English, as well as test sets for cross-genre transfer.

Model We again use our already trained e-INFERSENT model from PREDICTANDEXPLAIN.

Results In Table 4, we present the performance of e-INFERSENT and our 2 baselines when evaluated without fine-tuning on SICK-E and MultiNLI. We notice that the accuracy improvements obtained with e-INFERSENT are very small. However, e-INFERSENT additionally provides explanations, which could bring insight into the inner workings of the model. We manually annotated the first 100 explanations of the test sets. The percentage of correct explanations in the subset where the label was predicted correctly was 30.64% for SICK-E and only 1.92% for MultiNLI. We also noticed that the explanations in SICK-E, even when wrong, were generally on-topic and valid statements, while the ones in MultiNLI were generally nonsense or off-topic. Therefore, transfer learning for generating explanations in out-of-domain NLI would constitute challenging future work.

5 Related work

Interpretability One main direction in interpretability for neural networks is providing extractive justifications, i.e., explanations consisting of subsets of the raw input, such as words or image patches. Extractive techniques can be divided into post-hoc (applied after training) and architecture-incorporated (guiding the training). For example, Ribeiro et al. [24] introduce a post-hoc extractive technique, LIME, that explains the prediction of any classifier via a local linear approximation around the prediction. Alvarez-Melis and Jaakkola [1] introduces a similar approach but for structured prediction, where a variational autoencoder provides relevant perturbations of the inputs that are then used to infer pairs of input-output tokens that are causally related. While these models provide valuable insight for detecting biases, further model and dataset refinements would have to be made on a case-by-case basis. For example, Gururangan et al. [13] identified a set of biases in SNLI, but noted that their attempts to remove them would give rise to other biases.

Table 4: The average performance over 5 seeds of e-INFERSENT and the 2 baselines on SICK-E and MultiNLI with no fine-tuning. Standard deviations are in parenthesis.

Model	SICK-E	MultiNLI
InferSent-SNLI-ours	53.27 (1.65)	57 (0.41)
INFERSENTAUTOENC	52.9 (1.77)	55.38 (0.9)
e-INFERSENT	**53.54** (1.43)	**57.16** (0.51)

Attention-based models, such as [2, 25], offer some degree of interpretability and have been shown to also improve performance on downstream tasks. However, soft attention, the most prominent attention model, often does not learn to single out human-interpretable inputs.

Neither extractive nor attention-based techniques can provide full-sentence explanations of a model's decisions. Moreover, they cannot capture fine-grained relations and asymmetries, especially in a task like recognizing textual entailment. For example, if the words *person, woman, mountain, outdoors* are extracted as justification, one may not know whether the model correctly learned that *A woman is a person* and not that *A person is a woman*, let alone that the model correctly paired (woman, person) and (mountain, outside).

Natural language explanations In our work, we have taken a step further and built a neural network that is able to directly provide full-sentence natural language justifications. There has been little work on incorporating and outputting natural language free-form explanations, mostly due to the lack of appropriate datasets. In this direction, and very similar to our approach, is the recent work by Park et al. [23], who introduce two datasets of natural language explanations for the tasks of visual question-answering and activity recognition. Another work in this direction is that of Ling et al. [18], who introduced a dataset of textual justifications for solving math problems and formulate the task in terms of program execution. Nonetheless, their setup is specific to the task of solving math problems, and thus hard to transfer to more general natural understanding tasks. Jansen et al. [16] provided a dataset of natural language explanation graphs for elementary science questions. However, with only $1,680$ pairs of questions and explanations, their corpus is orders of magnitude smaller than e-SNLI.

Breaking natural language inference Recently, an increasing amount of analysis has been carried out on the SNLI dataset and on the inner workings of different models trained on it. For example, Dasgupta et al. [9] assembled a dataset to test whether inference models actually capture compositionality beyond word level. They showed that InferSent sentence embeddings [7] indeed do not exhibit significant compositionality and that downstream models using these sentence representations largely rely on simple heuristics that are ecologically valid in the SNLI corpus. For example, high overlap in words between premise and hypothesis usually predicts entailment, while most contradictory sentence pairs have no or very little overlap of words. Negation words would also strongly indicate a contradiction. Glockner et al. [11] introduce a toy dataset, BreakingNLI, to test whether natural language inference models capture world knowledge and generalize beyond statistical regularities. To construct BreakingSNLI, they modified some of the original SNLI sentences such that they differ by at most one word from the sentences in the training set. Glockner et al. show that models achieving high accuracies on SNLI, such as [21, 22, 6], show dramatically reduced performance on this simpler dataset, while the model of Chen et al. [5] is more robust due to incorporating external knowledge. As the explanations in e-SNLI are mostly self-contained, our dataset provides the precise external knowledge that one requires in order to solve the SNLI inference task. It is therefore a perfect testbed for developing models that incorporate external knowledge from free-form natural language.

6 Conclusions and future work

We introduced e-SNLI, a large dataset of natural language explanations for an influential task of recognizing textual entailment. To demonstrate the usefulness of e-SNLI, we experimented with various ways of using these explanations for outputting human-interpretable full-sentence justifications of classification decisions. We also investigated the usefulness of these explanations as an additional training signal for learning better universal sentence representations and the transfer capabilities to out-of-domain NLI datasets. In this work, we established a series of baselines using straight-forward recurrent neural network architectures for incorporating and generating natural language explanations. We hope that e-SNLI will be valuable for future research on more advanced models that would outperform our baselines.

Finally, we hope that the community will explore the dataset in other directions. For example, we also recorded the highlighted words, which we release with the dataset. Similar to the evaluation performed for visual question answering in Das et al. [8], our highlighted words could provide a source of supervision and evaluation for attention models [25, 22] or post-hoc explanation models where the explanation consists of a subset of the input.

Acknowledgments This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. We would also like to thank Jakob Foerster for the valuable discussions.

References

[1] Alvarez-Melis, D. and Jaakkola, T. S. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. *CoRR*, abs/1707.01943.

[2] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[3] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

[4] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). Listen, attend and spell. *CoRR*, abs/1508.01211.

[5] Chen, Q., Zhu, X., Ling, Z., Inkpen, D., and Wei, S. (2017). Natural language inference with external knowledge. *CoRR*, abs/1711.04289.

[6] Chen, Q., Zhu, X., Ling, Z., Wei, S., and Jiang, H. (2016). Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

[7] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

[8] Das, A., Agrawal, H., Zitnick, C. L., Parikh, D., and Batra, D. (2016). Human attention in visual question answering: Do humans and deep networks look at the same regions? *CoRR*, abs/1606.03556.

[9] Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings.

[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*.

[11] Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proc. of ACL*.

[12] Gong, Y., Luo, H., and Zhang, J. (2017). Natural language inference over interaction space. *CoRR*, abs/1709.04348.

[13] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proc. of NAACL*.

[14] Hill, F., Cho, K., and Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. *CoRR*, abs/1602.03483.

[15] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[16] Jansen, P. A., Wainwright, E., Marmorstein, S., and Morrison, C. T. (2018). Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *CoRR*, abs/1802.03052.

[17] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought vectors. *CoRR*, abs/1506.06726.

[18] Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. *CoRR*, abs/1705.04146.

[19] Liu, P., Qiu, X., and Huang, X. (2016). Modelling interaction of sentence pair with coupled-lstms. *CoRR*, abs/1605.05573.

[20] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., bernardi, R., and Zamparelli, R. (2014). A sick cure for the evaluation of compositional distributional semantic models.

[21] Nie, Y. and Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45. Association for Computational Linguistics.

[22] Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

[23] Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1802.08129.

[24] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[25] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

[26] Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

[27] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

Appendices

A List of templates to filter uninformative explanations

General templates

"<premise>"

"<hypothesis>"

"<hypothesis> <premise>"

"<premise> <hypothesis>"

"Sentence 1 states <premise>. Sentence 2 is stating <hypothesis>"

"Sentence 2 states <hypothesis>. Sentence 1 is stating <premise>"

"There is <hypothesis>"

"There is <premise>"

Entailment templates

"<premise> implies <hypothesis>"

"If <premise> then <hypothesis>"

"<premise> would imply <hypothesis>"

"<hypothesis> is a rephrasing of <premise>"

"<premise> is a rephrasing of <hypothesis>"

"In both sentences <hypothesis>"

"<premise> would be <hypothesis>"

"<premise> can also be said as <hypothesis>"