

e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations

Virginie Do¹, Oana-Maria Camburu^{1,2}, Zeynep Akata³, and Thomas Lukasiewicz^{1,2}

¹University of Oxford, ²Alan Turing Institute, London, ³University of Tuebingen,
virginiedo@gmail.com, firstname.lastname@cs.ox.ac.uk,
zeynep.akata@uni-tuebingen.de

Abstract

The recently proposed SNLI-VE corpus for recognising visual-textual entailment is a large, real-world dataset for fine-grained multimodal reasoning. However, the automatic way in which SNLI-VE has been assembled (via combining parts of two related datasets) gives rise to a large number of errors in the labels of this corpus. In this paper, we first present a data collection effort to correct the class with the highest error rate in SNLI-VE. Secondly, we re-evaluate an existing model on the corrected corpus, which we call SNLI-VE-2.0¹, and provide a quantitative comparison with its performance on the non-corrected corpus. Thirdly, we introduce e-SNLI-VE¹, which appends human-written natural language explanations to SNLI-VE-2.0. Finally, we train models that learn from these explanations at training time, and output such explanations at testing time.

1. Introduction

Inspired by textual entailment [2], Xie *et al.* [13] introduced the visual-textual entailment (VTE)² task, which considers semantic entailment between a premise image and a textual hypothesis. Semantic entailment consists in determining if the hypothesis can be concluded from the premise, and assigning to each pair of (premise image, textual hypothesis) a label among *entailment*, *neutral*, and *contradiction*. In Figure 1, the label for the first image-sentence pair is *entailment*, because the hypothesis states that “a bunch of people display different flags”, which can be clearly derived from the image. On the contrary, the second image-

sentence pair is labelled as *contradiction*, because the hypothesis stating that “people [are] running a marathon” contradicts the image with static people.

Xie *et al.* also propose the SNLI-VE dataset as the first dataset for VTE. SNLI-VE is built from the textual entailment SNLI dataset [2] by replacing textual premises with the Flickr30k images that they originally described [15]. However, images contain more information than their descriptions, which may entail or contradict the textual hypotheses (see Figure 1). As a result, the neutral class in SNLI-VE has substantial labelling errors. Vu *et al.* [12] estimated ~31% errors in this class, and ~1% for the contradiction and entailment classes.

In this work, we first focus on decreasing the error in the neutral class by collecting new labels for the neutral pairs in the validation and test sets of SNLI-VE, using Amazon Mechanical Turk (MTurk). To ensure high quality annotations, we used a series of quality control measures, such as in-browser checks, inserting trusted examples, and collecting three annotations per instance. Secondly, we re-evaluate current image-text understanding systems, such as the bottom-up top-down attention network (BUTD) [1] on VTE using our corrected dataset, which we call SNLI-VE-2.0.

Thirdly, we introduce the e-SNLI-VE corpus, which we form by appending human-written natural language explanations to SNLI-VE-2.0. These explanations were collected in e-SNLI [3] to support textual entailment for SNLI. For the same reasons as above, we re-annotate the explanations for the neutral pairs in the validation and test sets, while keeping the explanations from e-SNLI for all the rest. Finally, we extend a current VTE model with the capacity of learning from these explanations at training time and out-

IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision, 2020

¹We updated the dataset in our ICCV 2021 paper [6]. It is available at <https://github.com/maximek3/e-ViL>.

²Xie *et al.* [13] introduced the VTE task under the name of “visual entailment”, which could imply recognizing entailment between images only. This paper prefers to follow Suzuki *et al.* [11] and call it “visual-textual entailment” instead, as it involves reasoning on image-sentence pairs.


Premise	Hypothesis	Label
<p>SNLI Premise: Young individuals in uniforms are gathering under the flags of several different countries.</p> 	A bunch of people display different flags.	Entailment
	A group of people running a marathon.	Contradiction
	A group of people outdoors.	Neutral Entailment
	They have an American flag.	Neutral* (ambiguous)

Figure 1. Examples from SNLI-VE-2.0. (a) In red, the *neutral* label from SNLI-VE is wrong, since the picture clearly shows that the crowd is outdoors. We corrected it to *entailment* in SNLI-VE-2.0. (b) In green, an ambiguous instance. There is indeed an American flag in the background but it is very hard to see, hence the ambiguity between *neutral* and *entailment*, and even *contradiction* if one cannot spot it. Further, it is not clear whether “they” implies the whole group or the people visible in the image.

putting an explanation for each predicted label at testing time.

2. SNLI-VE-2.0

The goal of VTE is to determine if a textual hypothesis H_{text} can be concluded, given the information in a premise image P_{image} [13]. There are three possible labels:

- **Entailment:** if there is enough evidence in P_{image} to conclude that H_{text} is true.
- **Contradiction:** if there is enough evidence in P_{image} to conclude that H_{text} is false.
- **Neutral:** if neither of the earlier two are true.

The SNLI-VE dataset proposed by Xie *et al.* [13] is the combination of Flickr30k, a popular image dataset for image captioning [15] and SNLI, an influential dataset for natural language inference [2]. Textual premises from SNLI are replaced with images from Flickr30k, which is possible, as these premises were originally collected as captions of these images (see Figure 1).

However, in practice, a sensible proportion of labels are wrong due to the additional information contained in images. This mostly affects neutral pairs, since images may contain the necessary information to ground a hypothesis for which a simple premise caption was not sufficient. An example is shown in Figure 1. Vu *et al.* [12] report that the label is wrong for $\sim 31\%$ of neutral examples, based on a random subset of 171 neutral points from the test set. We also annotated 150 random neutral examples from the test set and found a similar percentage of 30.6% errors.³

³Our annotations are available at https://github.com/virginie-do/e-SNLI-VE/tree/master/annotations/gt_labels.csv

2.1. Re-annotation details

In this work, we only collect new labels for the neutral pairs in the validation and test sets of SNLI-VE. While the procedure of re-annotation is generic, we limit our re-annotation to these splits as a first step to verify the difference in performance that current models have when evaluated on the corrected test set as well as the effect of model selection on the corrected validation set. We leave for future work re-annotation of the training set, which would likely lead to training better VTE models. We also chose not to re-annotate entailment and contradiction classes, as their error rates are much lower ($<1\%$ as reported by Vu *et al.* [12]).

The main question that we want our dataset to answer is: “What is the relationship between the image premise and the sentence hypothesis?”. We provide workers with the definitions of entailment, neutral, and contradiction for image-sentence pairs and one example for each label. As shown in Figure 2, for each image-sentence pair, workers are required to (a) choose a label, (b) highlight words in the sentence that led to their decision, and (c) explain their decision in a comprehensive and concise manner, using at least half of the words that they highlighted. The collected explanations will be presented in more detail in Section 3.2, as we focus here on the label correction. We point out that it is likely that requiring an explanation at the same time as requiring a label has a positive effect on the correctness of the label, since having to justify in writing the picked label may make workers pay an increased attention. Moreover, we implemented additional quality control measures for crowdsourced annotations, such as (a) collecting three annotations for every input, (b) injecting trusted annotations into the task for verification [10], and (c) restricting to workers with at least 90% previous approval rate.

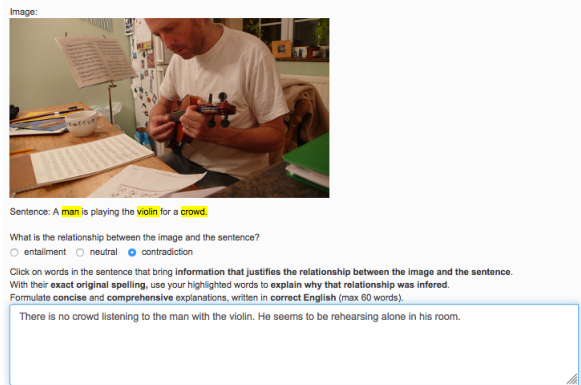


Figure 2. MTurk annotation screen. (a) The label *contradiction* is chosen, (b) the evidence words “man”, “violin”, and “crowd” are highlighted, and (c) an explanation is written with these words.

First, we noticed that some instances in SNLI-VE are

[virginie-do/e-SNLI-VE/tree/master/annotations/gt_labels.csv](https://github.com/virginie-do/e-SNLI-VE/tree/master/annotations/gt_labels.csv)

ambiguous. We show some examples in Figure 1 and in Appendix 5.3. In order to have a better sense of this ambiguity, three authors of this paper independently annotated 100 random examples. All three authors agreed on 54% of the examples, exactly two authors agreed on 45%, and there was only one example on which all three authors disagreed. We identified the following three major sources of ambiguity:

- mapping an emotion in the hypothesis to a facial expression in the image premise, e.g., “*people enjoy talking*”, “*angry people*”, “*sad woman*”. Even when the face is seen, it may be subjective to infer an emotion from a static image (see Figure 9 in Appendix 5.3).
- personal taste, e.g., “*the sign is ugly*”.
- lack of consensus on terms such as “*many people*” or “*crowded*”.

To account for the ambiguity that the neutral labels seem to present, we considered that an image-sentence pair is too ambiguous and not suitable for a well-defined visual-textual entailment task when three different labels were assigned by the three workers. Hence, we removed these examples from the validation (5.2%) and test (5.5%) sets.

To ensure that our workers are correctly performing the task, we randomly inserted trusted pairs, i.e., pairs among the 54% on which all three authors agreed on the label. For each set of 10 pairs presented to a worker, one trusted pair was introduced at a random location, so that the worker, while being told that there is such a test pair, cannot figure out which one it is. Via an in-browser check, we only allow workers to submit their answers for each set of 10 instances only if the trusted pair was correctly labelled. Other in-browser checks were done for the collection of explanations, as we will describe in Section 3.2. More details about the participants and design of the Mechanical Turk task can be found in Appendix 5.2.

After collecting new labels for the neutral instances in the validation and testing sets, we randomly select and annotate 150 instances from the validation set that were neutral in SNLI-VE. Based on this sample, the error rate went down from 31% to 12% in SNLI-VE-2.0. Looking at the 18 instances where we disagreed with the label assigned by MTurk workers, we noticed that 12 were due to ambiguity in the examples, and 6 were due to workers’ errors. Further investigation into potentially eliminating ambiguous instances would likely be beneficial. However, we leave it as future work, and we proceed in this work with using our corrected labels, since our error rate is significantly lower than that of the original SNLI-VE.

Finally, we note that only about 62% of the originally neutral pairs remain neutral, while 21% become contradiction and 17% entailment pairs. Therefore, we are now

facing an imbalance between the neutral, entailment, and contradiction instances in the validation and testing sets of SNLI-VE-2.0. The neutral class becomes underrepresented and the label distributions in the corrected validation and testing sets both become E / N / C: 39% / 20% / 41%. To account for this, we compute the *balanced accuracy*, i.e., the average of the three accuracies on each class.

2.2. Re-evaluation of Visual-Textual Entailment

Since we decreased the error rate of labels in the validation and test set, we are interested in the performance of a VTE model when using the corrected sets.

Model. To tackle SNLI-VE, Xie *et al.* [13] used EVE (for “Explainable Visual Entailment”), a modified version of the BUTD architecture, the winner of the Visual Question Answering (VQA) challenge in 2017 [1]. Since the EVE implementation is not available at the time of this work, we used the original BUTD architecture⁴, with the same hyperparameters as reported in [13].

BUTD contains an image processing module and a text processing module. The image processing module encodes each image region proposed by FasterRCNN [9] into a feature vector using a bottom-up attention mechanism. In the text processing module, the text hypothesis is encoded into a fixed-length vector, which is the last output of a recurrent neural network with 512-GRU units [4]. To input each token into the recurrent network, we use the pretrained GloVe vectors [8]. Finally, a top-down attention mechanism is used between the hypothesis vector and each of the image region vectors to obtain an attention weight for each region. The weighted sum of these image region vectors is then fused with the text hypothesis vector. The multimodal fusion is fed to a multilayer perceptron (MLP) with tanh activations and a final softmax layer to classify the image-sentence relation as entailment, contradiction, or neutral.

We use the original training set from SNLI-VE. To see the impact of correcting the validation and test sets, we do the following three experiments:

1. model selection as well as testing are done on the original uncorrected SNLI-VE.
2. model selection is done on the uncorrected SNLI-VE validation set, while testing is done on the corrected SNLI-VE-2.0 test set.
3. model selection as well as testing are done on the corrected SNLI-VE-2.0.

Models are trained with cross-entropy loss optimized by the Adam optimizer [7] with batch size 64. The maximum

⁴Using the implementation from <https://github.com/claudiogreco/coling18-gte>.

BUTD	val-original	val-corrected
test-original	73.02%	N/A
test-corrected	73.18%	72.52%

Table 1. Accuracies obtained with BUTD on SNLI-VE (val-original, test-original) and SNLI-VE-2.0 (val-corrected, test-corrected).

number of training epochs is set to 100, with early stopping when no improvement is observed on validation accuracy for 3 epochs. The final model checkpoint selected for testing is the one with the highest validation accuracy.

Results. The results of the three experiments enumerated above are reported in Table 1. Surprisingly, we obtained an accuracy of 73.02% on SNLI-VE using BUTD, which is better than the 71.16% reported by Xie *et al.* [13] for the EVE system which meant to be an improvement over BUTD. It is also better than their reproduction of BUTD, which gave 68.90%.

The same BUTD model that achieves 73.02% on the uncorrected SNLI-VE test set, achieves 73.18% balanced accuracy when tested on the corrected test set from SNLI-VE-2.0. Hence, for this model, we do not notice a significant difference in performance. This could be due to randomness. Finally, when we run the training loop again, this time doing the model selection on the corrected validation set from SNLI-VE-2.0, we obtain a slightly worse performance of 72.52%, although the difference is not clearly significant.

Finally, we recall that the training set has not been reannotated, and hence approximately 31% image-sentence pairs are wrongly labelled as neutral, which likely affects the performance of the model.

3. Visual-Textual Entailment with Natural Language Explanations

In this work, we also introduce e-SNLI-VE, a dataset combining SNLI-VE-2.0 with human-written explanations from e-SNLI [3], which were originally collected to support textual entailment. We replace the explanations for the neutral pairs in the validation and test sets with new ones collected at the same time as the new labels. We extend a current VTE model with an explanation module able to learn from these explanations at training time and generate an explanation for each predicted label at testing time.

3.1. e-SNLI-VE

e-SNLI [3] is an extension of the SNLI corpus with human-annotated natural language explanations for the ground-truth labels. The authors use the explanations to train models to also generate natural language justifications for their predictions. They collected one explanation for

each instance in the training set of SNLI and three explanations for each instance in the validation and testing sets.

We randomly selected 100 image-sentence pairs in the validation set of SNLI-VE and their corresponding explanations in e-SNLI and examined how relevant these explanations are for the VTE task. More precisely, we say that an explanation is relevant if it brings information that justifies the relationship between the image and the sentence. We restricted the count to correctly labelled inputs and found that 57% explanations were relevant. For example, the explanation for entailment in Figure 3 (“*Cooking in his apartment is cooking*”) was counted as irrelevant in our statistics, because it would not be the best explanation for an image-sentence pair, even though it is coherent with the textual pair. We investigate whether these explanations improve a VTE model when enhanced with a component that can process explanations at train time and output them at test time.

To form e-SNLI-VE, we append to SNLI-VE-2.0 the explanations from e-SNLI for all except the neutral pairs in the validation and test sets of SNLI-VE, which we replace with newly crowdsourced explanations collected at the same time as the labels for these splits (see Figure 3). Statistics of e-SNLI-VE are shown in Appendix 5.1, Table 3.

3.2. Collecting Explanations

As mentioned before, in order to submit the annotation of an image-sentence pair, three steps must be completed: workers must choose a label, highlight words in the hypothesis, and use at least half of the highlighted words to write an explanation for their decision. The last two steps thus follow the quality control of crowd-sourced explanations introduced by Camburu *et al.* [3]. We also ensured that workers do not simply use a copy of the given hypothesis as explanation. We ensured all the above via in-browser checks before workers’ submission. An example of collected explanations is given in Figure 3.

To check the success of our crowdsourcing, we manually assessed the relevance of explanations among a random subset of 100 examples. A marking scale between 0 and 1 was used, assigning a score of k/n when k required attributes were given in an explanation out of n . We report an 83.5% relevance of explanations from workers.

We note that, since our explanations are VTE-specific, they were phrased differently from the ones in e-SNLI, with more specific mentions to the images (e.g., “*There is no labcoat in the picture, just a man wearing a blue shirt.*”, “*There are no apples or oranges shown in the picture, only bananas.*”). Therefore, it would likely be beneficial to collect new explanations for all SNLI-VE-2.0 (not only for the neutral pairs in the validation and test sets) such that models can learn to output convincing explanations for the task at hand. However, we leave this as future work, and we

show in this work the results that one obtains when using the explanations from e-SNLI-VE.


Premise	Hypothesis	Label	Explanation
 SNLI premise: An Asian man kept company by his cat is cooking in his apartment.	An Asian man is cooking.	Entailment	Cooking in his apartment is cooking.
	An Asian man is cooking dumplings.	Neutral	The man is cooking but the picture isn't clear enough to see he's making dumplings.

Figure 3. Two image-sentence pairs from e-SNLI-VE with (a) at the top, an uninformative explanation from e-SNLI, (b) at the bottom, an explanation collected from our crowdsourcing. We only collected new explanations for the neutral class (along with new labels). The SNLI premise is not included in e-SNLI-VE.

3.3. VTE Models with Natural Language Explanations

This section presents two VTE models that generate natural language explanations for their own decisions. We name them PAE-BUTD-VE and ETP-BUTD-VE, where PAE (resp. ETP) is for PREDICTANDEXPLAIN (resp. EXPLAINTHENPREDICT), two models with similar principles introduced by Camburu *et al.* [3]. The first system learns to generate an explanation conditioned on the image premise, textual hypothesis, and predicted label. In contrast, the second system learns to first generate an explanation conditioned on the image premise and textual hypothesis, and subsequently makes a prediction solely based on the explanation.

3.3.1 Predict and Explain

PAE-BUTD-VE is a system for solving VTE and generating natural language explanations for the predicted labels. The explanations are conditioned on the image premise, the text hypothesis, and the predicted label (ground-truth label at train time), as shown in Figure 4.

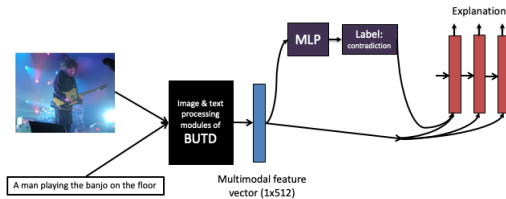


Figure 4. PAE-BUTD-VE. The generation of explanation is conditioned on the image premise, textual hypothesis, and predicted label.

Model. As described in Section 2.2, in the BUTD model, the hypothesis vector and the image vector were fused in

a fixed-size feature vector \mathbf{f} . The vector \mathbf{f} was then given as input to an MLP which outputs a probability distribution over the three labels. In PAE-BUTD-VE, in addition to the classification layer, we add a 512-LSTM [5] decoder to generate an explanation. The decoder takes the feature vector \mathbf{f} as initial state. Following Camburu *et al.* [3], we prepend the label as a token at the beginning of the explanation to condition the explanation on the label. The ground truth label is provided at training time, whereas the predicted label is given at test time.

At test time, we use beam search with a beam width of 3 to decode explanations. For memory and time reduction, we replaced words that appeared less than 15 times among explanations with “#UNK#”. This strategy reduces the output vocabulary size to approximately 8.6k words.

Loss. The training loss is a weighted combination of the classification loss and the explanation loss, both computed using softmax cross entropy: $\mathcal{L} = \alpha \mathcal{L}_{label} + (1 - \alpha) \mathcal{L}_{explanation}$; $\alpha \in [0, 1]$.

Model selection. In this experiment, we are first interested in examining if a neural network can generate explanations at no cost for label accuracy. Therefore, only balanced accuracy on label is used for the model selection criterion. However, future work can investigate other selection criteria involving a combination between the label and explanation performances.

We performed hyperparameter search on α , considering values between 0.2 and 0.8 with a step of 0.2. We found $\alpha = 0.4$ to produce the best validation balanced accuracy of 72.81%, while BUTD trained without explanations yielded a similar 72.58% validation balanced accuracy.

Results. As summarised in Table 2, we obtain a test balanced accuracy for PAE-BUTD-VE of 73%, while the same model trained without explanations obtains 72.52%. This is encouraging, since it shows that one can obtain additional natural language explanations without sacrificing performance (and eventually even improving the label performance, however, future work is needed to conclude whether the difference 0.48% improvement in performance is statistically significant).

Camburu *et al.* [3] mentioned that the BLEU score was not an appropriate measure for the quality of explanations and suggested human evaluation instead. We therefore manually scored the relevance of 100 explanations that were generated when the model predicted correct labels. We found that only 20% of explanations were relevant. We highlight that the relevance of explanations is in terms of whether the explanation reflects ground-truth reasons supporting the correct label. This is not to be confused with whether an explanation is correctly illustrating the inner

	Label	Expl.	Expl. (Camburu <i>et al.</i>)
PAE-BUTD-VE	73%	20%	34.68%
ETP-BUTD-VE	69.40%	35%	49.8%

Table 2. Label balanced accuracies and explanation relevance rates of our two explanatory systems on e-SNLI-VE. Comparison with their counterparts in e-SNLI [3]. Without the explanation component, the balanced accuracy on SNLI-VE-2.0 is 72.52%

working of the model, which is left as future work. It is also important to note that on a similar experimental setting, Camburu *et al.* report as low as 34.68% correct explanations, training with explanations that were actually collected for their task. Lastly, the model selection criterion at validation time was the prediction balanced accuracy, which may contribute to the low quality of explanations. While we show that adding an explanation module does not harm prediction performance, more work is necessary to get models that output trustable explanations.

3.3.2 Explain Then Predict

When assigning a label, an explanation is naturally part of the decision-making process. This motivates the design of a system that explains itself before deciding on a label, called ETP-BUTD-VE. For this system, a first neural network is trained to generate an explanation given an image-sentence input. Separately, a second neural network, called EXPLToLABEL-VE, is trained to predict a label from an explanation (see Figure 5).

Model. For the first network, we set $\alpha = 0$ in the training loss of the PAE-BUTD-VE model to obtain a system that only learns to generate an explanation from the image-sentence input, without label prediction. Hence, in this setting, no label is prepended before the explanation.

For the EXPLToLABEL-VE model, we use a 512-LSTM followed by an MLP with three 512-layers and ReLU activation, and softmax activation to classify the explanation between entailment, contradiction, and neutral.

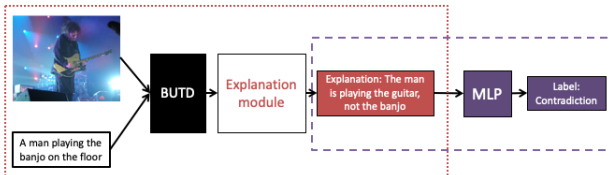


Figure 5. Architecture of ETP-BUTD-VE. Firstly, an explanation is generated, secondly the label is predicted from the explanation. The two models (in separate dashed rectangles) are **not** trained jointly.

Model selection. For EXPLToLABEL-VE, the best model is selected on balanced accuracy at validation time. For ETP-BUTD-VE, perplexity is used to select the best model parameters at validation time. It is computed between the explanations produced by the LSTM and ground truth explanations from the validation set.

Results. When we train EXPLToLABEL-VE on e-SNLI-VE, we obtain a balanced accuracy of 90.55% on the test set.

As reported in Table 2, the overall PAE-BUTD-VE system achieves 69.40% balanced accuracy on the test set of e-SNLI-VE, which is a 3% decrease from the non-explanatory BUTD counterpart (72.52%). However, by setting α to zero and selecting the model that gives the best perplexity per word at validation, the quality of explanation significantly increased, with 35% relevance, based on manual evaluation. Thus, in our model, generating better explanations involves a small sacrifice in label prediction accuracy, implying a trade-off between explanation generation and accuracy.

We note that there is room for improvement in our explanation generation method. For example, one can implement an attention mechanism similar to Xu *et al.* [14], so that each generated word relates to a relevant part of the multi-modal feature representation.

3.3.3 Qualitative Analysis of Generated Explanations

We complement our quantitative results with a qualitative analysis of the explanations generated by our enhanced VTE systems. In Figures 6 and 7, we present examples of the predicted labels and generated explanations.

Figure 6 shows an example where the ETP-BUTD-VE model produces both a correct label and a relevant explanation. The label is contradiction, because in the image, the students are playing with a soccer ball and not a basketball, thus contradicting the text hypothesis. Given the composition of the generated sentence (“*Students cannot be playing soccer and baseball at the same time.*”), EXPLToLABEL-VE was able to detect a contradiction in the image-sentence input. In comparison, the explanation from e-SNLI-VE is not correct, even if it was valid for e-SNLI when the text premise was given. This emphasizes the difficulty that we are facing with generating proper explanations when training on a noisy dataset.

Even when the generated explanations are irrelevant, we noticed that they are on-topic and that most of the time the mistakes come from repetitions of certain sub-phrases. For example, in Figure 7, PAE-BUTD-VE predicts the label neutral, which is correct, but the explanation contains an erroneous repetition of the n-gram “*are in a car*”. However, it appears that the system learns to generate a sentence in the form “*Just because ... doesn’t mean ...*”, which is


Premise	Text premise (not included): In the Fall, college students return to their schools.	
	Hypothesis: College students at the basketball game.	
e-SNLI-VE-2.0	PaE-BUTD-VE	ETP-BUTD-VE
Label Contradiction	Label Contradiction	Label Contradiction
Explanation The college students can not be at a basketball game when they are returning to school	Explanation The students are not necessarily students.	Explanation Students can not be playing soccer and basketball at the same time.

Figure 6. Both systems PAE-BUTD-VE and ETP-BUTD-VE predict the correct label, but only ETP-BUTD-VE generates a relevant explanation.


Premise	Text premise (not included): Two Asian women hug one another while smiling outside of a white Sports Utility Vehicle	
	Hypothesis: Asian women have just purchased a car	
e-SNLI-VE-2.0	PaE-BUTD-VE	ETP-BUTD-VE
Label Neutral	Label Neutral	Label Neutral
Explanation Nothing to suggest or confirm they have just purchased the car in the background	Explanation Just because the women are in a car doesn't mean they are in a car	Explanation Just because the women are taking a picture of a car does not mean they are going to the car

Figure 7. Both systems PAE-BUTD-VE and ETP-BUTD-VE predict the correct label, but generate irrelevant explanations.

frequently found for the justification of neutral pairs in the training set. The explanation generated by ETP-BUTD-VE adopts the same structure, and the EXPLTO LABEL-VE component correctly classifies the instance as neutral. However, even if the explanation is semantically correct, it is not relevant for the input and fails to explain the classification.

4. Conclusion

In this paper, we first presented SNLI-VE-2.0, which corrects the neutral instances in the validation and test sets of SNLI-VE. Secondly, we re-evaluated an existing model on the corrected sets in order to update the estimate of its performance on this task. Thirdly, we introduced e-SNLI-VE, a dataset which extends SNLI-VE-2.0 with natural language explanations. Finally, we trained two types of models that learn from these explanations at training time, and output such explanations at test time, as a stepping stone in explainable artificial intelligence. Our work is a jumping-off point for both the identification and correction of SNLI-VE, as well as in the extension to explainable VTE. We hope that the community will build on our findings to create more robust as well as explainable multimodal systems.

Acknowledgements. This work was supported by the Oxford Internet Institute, a JP Morgan PhD Fellow-

ship 2019-2020, an Oxford-DeepMind Graduate Scholarship, the Alan Turing Institute under the EPSRC grant EP/N510129/1, and the AXA Research Fund, as well as DFG-EXC-Nummer 2064/1-Projektnummer 390727645 and the ERC under the Horizon 2020 program (grant agreement No. 853489).

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *Proc. CVPR*, pages 6077–6086, 2018. 1, 3
- [2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proc. EMNLP*, 2015. 1, 2
- [3] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549, 2018. 1, 4, 5, 6
- [4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3
- [5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5
- [6] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. *arXiv preprint arXiv:2105.03761*, 2021. 1
- [7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3
- [8] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, 2014. 3
- [9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 3
- [10] Alexander Sorokin and David Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proc. CVPR Workshops*, pages 1–8, 2008. 2
- [11] Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. Multimodal logical inference system for visual-textual entailment. *arXiv preprint arXiv:1906.03952*, 2019. 1
- [12] Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. *arXiv preprint arXiv:1806.05645*, 2018. 1, 2

- [13] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1, 2, 3, 4
- [14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 6
- [15] M. H. Peter Young, Alice Lai, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1, 2

5. Appendix

5.1. Statistics of e-SNLI-VE

e-SNLI-VE is the combination of SNLI-VE-2.0 with explanations from either e-SNLI or our crowdsourced annotations where applicable. The statistics of e-SNLI-VE are shown in Table 3.

	Training	Validation	Testing
#Images	29,783	1,000	1,000
#Entailment	176,932	6,913	6,903
#Neutral	176,045	3,453	3,537
#Contradiction	176,550	7,181	7,134
#Total			
#Explanations from e-SNLI	529,527	11,888	11,898
#Explanations from our data collection	None	5,659	5,676
Vocabulary size ⁵	41,230	8,963	9,197

Table 3. Summary of e-SNLI-VE (= SNLI-VE-2.0 + explanations). Image-sentence pairs labelled as neutral in the training set have not been corrected.

5.2. Details of the Mechanical Turk Task

We used Amazon Mechanical Turk (MTurk) to collect new labels and explanations for SNLI-VE. 2,060 workers participated in the annotation effort, with an average of 1.98 assignments per worker and a standard deviation of 5.54. We required the workers to have a previous approval rate above 90%. No restriction was put on the workers’ location.

Each assignment consisted of a set of 10 image-sentence pairs. For each pair, the participant was asked to (a) choose a label, (b) highlight words in the sentence that led to their decision, and (c) explain their decision in a comprehensive and concise manner, using a subset of the words that they highlighted. The instructions are shown in Figure 8. Workers were also guided with three annotated examples, one for each label.

For each assignment of 10 questions, one trusted annotation with gold standard label was inserted at a random position, as a measure to control the quality of label annotation.

⁵Including text hypotheses and explanations.

Instructions


- Your task is to **find the relationship between the image and the sentence**.
 - **Entailment**: There is enough evidence in the image to conclude that the sentence is true
 - **Contradiction**: There is enough evidence in the image to conclude that the sentence is false
 - **Neutral**: The evidence in the image is insufficient to draw a conclusion about the sentence
- You should first highlight the **most important words** in the sentence that led to your conclusion about its relationship to the image. Click on the words to highlight them. To unhighlight a word, click on it again.
- Use your highlighted words with their **exact original spelling** to **explain why** the information you highlighted in the sentence helped you infer its relationship to the image. Formulate **correct English sentences**. Your explanations should be **concise** and **comprehensive** (max 60 words).

Figure 8. Instructions given to workers on Mechanical Turk

Each assignment was completed by three different workers. An example of question is shown in Figure 2 in the core paper.

5.3. Ambiguous Examples from SNLI-VE

Some examples in SNLI-VE were ambiguous and could find correct justifications for incompatible labels, as shown in Figures 9, 10, and 11.

Image: 

Sentence: A woman monitoring a baby as it plays with nearby luggage.

What is the relationship between the image and the sentence?

☒ entailment ☐ neutral ☐ contradiction ☐ ambiguous

Click on words in the sentence that bring information that justifies the relationship between the image and the sentence. With their exact original spelling, use your highlighted words to explain why that relationship was inferred. Formulate concise and comprehensive explanations, written in correct English (max 60 words).

Figure 9. Ambiguous SNLI-VE instance. Some may argue that the woman’s face betrays sadness, but the image is not quite clear. Secondly, even with better resolution, facial expression may not be a strong enough evidence to support the hypothesis about the woman’s emotional state.

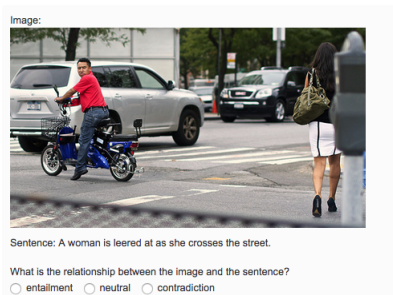


Figure 10. Ambiguous SNLI-VE instance. The lack of consensus is on whether the man is “leering” at the woman. While it is likely the case, this interpretation in favour of entailment is subjective, and a cautious annotator would prefer to label the instance as neutral.


Premise	Hypothesis	Label
	Kindergarten students napping	Neutral

Figure 11. Ambiguous SNLI-VE instance. Some may argue that it is impossible to certify from the image that the children are kindergarten students, and label the instance as neutral. On the other hand, the furniture may be considered as typical of kindergarten, which would be sufficient evidence for entailment.