1-20-2019

Visual Entailment: A Novel Task for Fine-Grained Image Understanding

Ning Xie

Farley Lai

Derek Doran

Asim Kadav

Visual Entailment: A Novel Task for Fine-Grained Image Understanding

Ning Xie*
Wright State University
Dayton, OH, U.S.A.
xie.25@wright.edu

Farley Lai
NEC Laboratories America
Princeton, NJ, U.S.A.
farleylai@nec-labs.com

Derek Doran
Wright State University
Dayton, OH, U.S.A.
derek.doran@wright.edu

Asim Kadav
NEC Laboratories America
Princeton, NJ, U.S.A.
asim@nec-labs.com

Abstract

Existing visual reasoning datasets such as Visual Question Answering (VQA), often suffer from biases conditioned on the question, image or answer distributions. The recently proposed CLEVR dataset addresses these limitations and requires fine-grained reasoning but the dataset is synthetic and consists of similar objects and sentence structures across the dataset.

In this paper, we introduce a new inference task, **Visual Entailment** *(VE) - consisting of image-sentence pairs whereby a premise is defined by an image, rather than a natural language sentence as in traditional Textual Entailment tasks. The goal of a trained VE model is to predict whether the image semantically entails the text. To realize this task, we build a dataset SNLI-VE based on the Stanford Natural Language Inference corpus and Flickr30k dataset. We evaluate various existing VQA baselines and build a model called Explainable Visual Entailment (EVE) system to address the VE task. EVE achieves up to 71% accuracy and outperforms several other state-of-the-art VQA based models. Finally, we demonstrate the explainability of EVE through cross-modal attention visualizations. The SNLI-VE dataset is publicly available at* https://github.com/necla-ml/SNLI-VE.

1. Introduction

The pursuit of "visual intelligence" is a long lasting theme of the machine learning community. While the performance of image classification and object detection has significantly improved in the recent years [42, 63, 65, 26], progress in higher-level scene reasoning tasks such as scene understanding is relatively limited [73].

Recently, several datasets, such as VQA-v1.0 [2], VQA-v2.0 [23], CLEVR [32], Visual7w [81], Visual Genome [41], COCO-QA [57], and models [33, 60, 29, 31, 1, 67, 17, 37] have been used to measure the progress in understanding the interaction between vision and language modalities. However, the quality of the widely used VQA-v1.0 dataset [2] suffers from a natural bias [23]. Specifically, there is a long tail distribution of answers and also a question-conditioned bias where, questions may hint at the answers, such that the correct answer may be inferred without even considering the visual information. For instance, of the question "Do you see a . . . ?", the model may bias towards the answer "Yes" since it is correct for 87% of times during training. Besides, many questions in the VQA-v1.0 dataset are simple and straightforward and do not require compositional reasoning from the trained model. VQA-v2.0 [23] has been proposed to reduce the dataset "bias" considerably in VQA-v1.0 by associating each question with relatively balanced different answers. However, the questions are rather straight-forward and require limited fine-grained reasoning.

CLEVR dataset [32], is designed for fine-grained reasoning and consists of compositional questions such as "What size is the cylinder that is left of the brown metal thing that is left of the big sphere?". This kind of questions requires learning fine-grained reasoning based on visual information. However, CLEVR is a synthetic dataset, and visual information and sentence structures are very similar across the dataset. Hence, models that provide good performance on CLEVR dataset may not generalize to real-world settings.

To address the above limitations, we propose a novel inference task, *Visual Entailment* (VE), which requires fine-grained reasoning in real-world settings. The design is de-

*Work performed as a NEC Labs intern

rived from Text Entailment (TE) [12] task. In our VE task, a real world image premise P_{image} and a natural language hypothesis H_{text} are given, and the goal is to determine if H_{text} can be concluded given the information provided by P_{image}. Three labels *entailment*, *neutral* or *contradiction* are assigned based on the relationship conveyed by the (P_{image}, H_{text}).

- *Entailment* holds if there is enough evidence in P_{image} to conclude that H_{text} is true.
- *Contradiction* holds if there is enough evidence in P_{image} to conclude that H_{text} is false.
- Otherwise, the relationship is *neutral*, implying the evidence in P_{image} is insufficient to draw a conclusion about H_{text}.

The main difference between VE and TE task is, the premise in TE in a natural language sentence P_{text}, instead of an image premise P_{image}. Note that the existing of "neutral" makes the VE task more challenging compared to previous "yes-no" VQA tasks, since "neutral" requires the model to conclude the uncertainty between "entailment (yes)" and "contradiction (no)". Figure 1 illustrates a VE example, which is from the SNLI-VE dataset we propose below, that given an image premise, the three different text hypotheses lead to different labels.

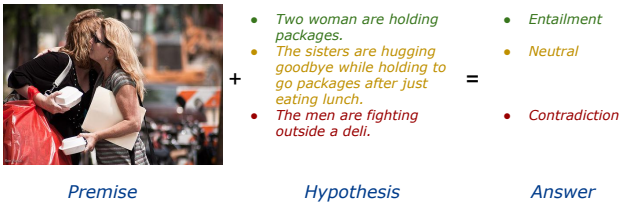

Figure 1. An Example from SNLI-VE dataset

We build the SNLI-VE dataset to illustrate the VE task, based on Stanford Natural Language Inference (SNLI) [4], which is a widely used text-entailment dataset, and Flickr30k [76], which is an image captioning dataset. The combination of SNLI and Flickr30k is straightforward since SNLI is created using Flickr30k. The detailed process of creating the SNLI-VE dataset is discussed in Section 3.2.

We develop an Explainable Visual Entailment (EVE) model to address the VE task. EVE captures the interaction within and between the image premise and the text hypothesis through attention. We evaluate EVE against several other state-of-the-art (SOTA) visual question answering (VQA) baselines and an image captioning based model on the SNLI-VE dataset. The interpretability of EVE is demonstrated using attention visualizations.

In summary, the contributions of our work are:

- We propose a *novel* inference task, Visual Entailment, that requires a systematic cross-modal understanding between vision and a natural language.
- We build a VE dataset, SNLI-VE, consisting of real-world image and natural language sentence pairs for VE tasks. The dataset is publicly available[1].
- We design a VE model, EVE, to solve the VE task with interpretable attention visualizations.
- We evaluate EVE against other SOTA VQA and image captioning based baselines.

2. Related Work

Our work is inspired by previous work on NLI, VQA, image captioning, and interpretable models.

Natural Language Inference. We focus on *textual entailment* as our NLI task [18, 11, 3, 12, 46]. Annotated corpus for TE was limited in size until SNLI [4] was proposed, which is based on the Flickr30k [76] image captions. Since then, several neural-network based methods have been proposed over SNLI that either use sentence encoding models to individually encode hypothesis and premise or attention based models that encode the sentences together and align similar words in hypothesis and premise [8, 50, 62, 59]. Our paper extends the TE task in the visual domain – allowing future work on our SNLI-VE task to build new models on recent progress in SNLI and VQA. Our work is different from the recent work [71] that combines both images and captions as premises.

Visual Question Answering. Recent work on VQA includes datasets [32, 2, 23, 81, 41, 57, 47, 19, 66] and models [33, 60, 29, 31, 1, 67, 17, 37]. The goal of VQA is to answer natural language questions based on the provided visual information. VQA-v2.0 [23] and CLEVR [32] datasets are designed to address bias and reasoning limitations of VQA-v1.0, respectively. Recent work on compositional reasoning systems have achieved nearly 100% results on CLEVR [29] but the SOTA performance on VQA-v2.0 is no more than 75% [15], implying learning multi-modal feature interaction using natural images has room for improvement. There have been a large number of models and approaches to address the VQA task. This includes simple linear models using ranking loss [16, 36], bi-linear pooling methods [45, 20, 55, 17, 37], attention-based methods [1, 52, 64] and reasoning based approaches [54, 27, 33, 38, 29] on CLEVER and VQA-v1.0 datasets.

[1] https://github.com/necla-ml/SNLI-VE

Image Captioning. The problem of image captioning explores the generation of natural language sentences to best depict input image content. A common approach for these tasks is to use temporal models over convolutional features [36, 70, 7]. Recent work has also explored generating richer captions to describe images in a more fine-grained manner [34]. EVE differs from image-captioning since it requires discerning fine-grained information about an image conditioned on the hypothesis into three classes. However, existing image-captioning methods can serve as a baseline, where the output class label is based on a distance measure between the generated caption and the input hypothesis.

Visual Relationship Detection. Relationship detection among image constituents uses separate branches in a ConvNet to model objects, humans, and their interactions [5, 21]. A distinct approach in Santoro et al. [60] treats each of the cells across channels in convolutional feature maps as an object and the relationships are modeled by a pairwise concatenation of the feature representations of individual cells.

Scene graph based relationship modeling, using a structured representation for describing object relationships and their attributes [35, 43, 44, 74] has been extensively studied. Furthermore, pairing different objects in a scene [13, 28, 60, 78] is also common. However, a scene with many objects may have only a few individual interacting objects. Hence, it can be inefficient to model all relationships across all individual object pairs [80], making these methods computationally expensive for complex scene understanding tasks such as VE.

Our model, EVE instead uses self-attention to efficiently learn the relationships between various scene elements and words instead of bi-gram or tri-gram based modeling as used in previous work.

Interpretability. As deep neural networks have become widespread in real-world applications, there has been an increasing focus on interpretability and transparency. Recent work addresses this requirement either through saliency-map visualizations [61, 77, 49], attention mechanism [75, 79, 51, 14], or other analysis [30, 39, 56, 58]. Our work demonstrates interpretability via attention visualizations.

3. Visual Entailment Task

3.1. Formal Definition

We introduce a dataset $\mathcal{D}$ for VE task structured as $\{(i_1, h_1, l_1), (i_1, h_2, l_2) \dots (i_1, h_{m_1}, l_{m_1}), \dots (i_n, h_{m_n}, l_{m_n})\}$, where (i_k, h_s, l_s) is an instance from $\mathcal{D}$, with i_k, h_s, and l_s denoting an image premise, a text hypothesis and a class label, respectively. It is worth noting that each image i_k is used multiple times with different labels given distinct hypotheses $\{h_{m_k}\}$.

Figure 2. **More examples from SNLI-VE dataset**

Three labels e, n, or c are assigned based on the relationship conveyed by (i_k, h_s). Specifically, i) e (entailment) is assigned if $i_k \models h_s$, ii) n (neutral) is assigned if $i_k \not\models h_s \wedge i_k \not\models \neg h_s$, iii) c (contradiction) is assigned if $i_k \models \neg h_s$.

3.2. Visual Entailment Dataset

3.2.1 Dataset criteria

Based on the vision community's experience with SNLI, VQA-v1.0, VQA-v2.0, and CLEVR, there are four *criteria* in developing an effective dataset:

1. *Structured set of real-world images*. The dataset should be based on real-world images and the same image can be paired with different hypotheses to form different labels.

2. *Fine-grained*. The dataset should enforce fine-grained reasoning about subtle changes in hypotheses that could lead to distinct labels.

3. *Sanitization*. No instance overlapping across different dataset partitions. One image can only exist in a single partition.

4. *Account for any bias*. Measure the dataset bias and

	Training	Validation	Testing
#Image	29,783	1,000	1,000
#Entailment	176,932	5,959	5,973
#Neutral	176,045	5,960	5,964
#Contradiction	176,550	5,939	5,964
Vocabulary Size	29,550	6,576	6,592

Table 1. **SNLI-VE dataset**

provide baselines to serve as the performance lower bound for potential future evaluations.

3.2.2 SNLI-VE Construction

We now describe how we construct SNLI-VE, which is a dataset for VE tasks.

We build the dataset SNLI-VE based on two existing datasets, Flickr30k [76] and SNLI [4]. **Flickr30k** is a widely used image captioning dataset containing 31,783 images and 158,915 corresponding captions. The images in Flickr30k consist of everyday activities, events and scenes [76], with 5 captions per image generated via crowdsourcing. **SNLI** is a large annotated TE dataset built upon Flickr30k captions. Each image caption in Flickr30k is used as a text premise in SNLI. The authors of SNLI collect multiple hypotheses in the three classes - *entailment*, *neutral*, and *contradiction* - for a given premise via Amazon Mechanical Turk [68], resulting in about 570K (P_{text}, H_{text}) pairs. Data validation is conducted in SNLI to measure the label agreement. Specifically, each (P_{text}, H_{text}) pair is assigned a *gold label*, indicating the label is agreed by a majority of crowdsourcing workers (at least 3 out of 5). If such a consensus is not reached, the gold label is marked as "-".

Since SNLI was constructed using Flickr30k captions, for each (P_{text}, H_{text}) pair in SNLI, it is feasible to find the corresponding Flickr30k image through the annotations in SNLI. This enables us to create a structured VE dataset based on both. Specifically, for each (P_{text}, H_{text}) pair in SNLI with an agreed gold label, we replace the text premise with its corresponding Flickr30k image, resulting in a (P_{image}, H_{text}) pair in SNLI-VE. Figures 1 and 2 illustrate examples from the SNLI-VE dataset. SNLI-VE naturally meets the aforementioned *criterion 1* and *criterion 2*. Each image in SNLI-VE are real-world ones and is associated with distinct labels given different hypotheses. Furthermore, Flickr30k and SNLI are well-studied datasets, allowing the community to focus on the new task that our paper introduces, rather than spending time familiarizing oneself with the idiosyncrasies of a new dataset.

A sanity check is applied to SNLI-VE dataset partitions in order to guarantee *criterion 3*. We notice the original SNLI dataset partitions does not consider the arrangement of the original caption images. If SNLI-VE directly adopts the original partitions from SNLI, all images in validation or testing partitions also exist in the training partitions, violating *criterion 3*. To amend this, we disjointedly partition SNLI-VE by images following the partition in [22] and make sure instances with different labels are of similar numbers across training, validation, and testing partitions as shown in Table 1.

	SNLI-VE	VQA-v2.0	CLEVR
Partition Size:			
Training	529,527	443,757	699,989
Validation	17,858	214,354	149,991
Testing	17,901	555,187	149,988
Question Length:			
Mean	7.4	6.1	18.4
Median	7.0	6.0	17.0
Mode	6	5	14
Max	56	23	43
Vocabulary Size	32,191	19,174	87

Table 2. **Dataset Comparison Summary**

Regarding *criterion 4*, since SNLI has already been extensively studied, we are aware that there exists a hypothesis-conditioned bias in SNLI as recently reported by Gururangan *et al*. [24]. Though the labels in SNLI-VE are distributed evenly across dataset partitions, SNLI-VE still inevitably suffers from this bias inherently. Therefore, we provide a hypothesis-only baseline in Section 5.1 to serve as a performance lower bound.

3.3. SNLI-VE and VQA Datasets

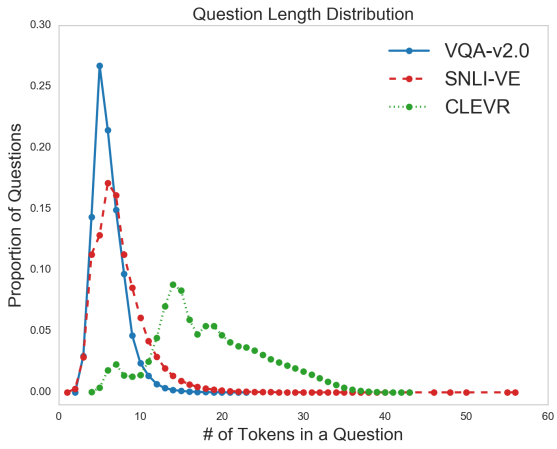

Figure 3. **Question Length Distribution**

We further compare our SNLI-VE dataset with the two widely used VQA datasets, VQA-v2.0 and CLEVR. The

comparison focuses on the *questions* (for SNLI-VE dataset, we consider a hypothesis as a question). Table 2 is a statistical summary about the questions from three datasets. Before generating Table 2, questions are prepossessed by three steps: *i)* split into words, *ii)* lower case all words, *iii)* removing punctuation symbols {‘’“”,.-?!}. Figure 3 depicts a detailed question length distribution.

According to Table 2, among the three datasets, our SNLI-VE dataset, which contains the smallest total number of questions (summing up training, validation and testing partitions), has the largest vocabulary size. The maximum question length in SNLI-VE is 56, which is the largest among these three datasets, and represents real-world descriptions. Both the mean and median lengths are larger than VQA-v2.0 dataset. The question length distribution of SNLI-VE, as shown in Figure 3, is quite heavy-tailed in contrast to the others. These observations indicate that the text in SNLI-VE may be difficult to handle compared to VQA-v2.0 for certain models. As for CLEVR dataset, even though most sentences are much longer than SNLI-VE as shown in Figure 3, the vocabulary size is only 87. We believe this is due to the synthetic nature of CLEVR, which also indicates models that achieve high-accuracy on CLEVR may not be able to generalize to our SNLI-VE dataset.

4. EVE: Explainable Visual Entailment System

The design of our explainable VE architecture, as shown in Figure 4, is based on the Attention Top-Down/Bottom-Up model discussed later in Subsection 5.4, which is the winner of VQA Challenge, 2017. Similar to the Attention Top-Down/Bottom-Up, our EVE architecture is composed of a text and an image branch. The text branch extracts features from the input text hypothesis H_{text} through an RNN. The image branch generates image features from P_{image}. The features produced from the two branches are then fused and projected through fully-connected (FC) layers towards predicting the final conclusion. The image features can be configured to take the feature maps from a pre-trained convolutional neural network (CNN) or ROI-pooled image regions from a region of interest (ROI) proposal network (RPN).

We build two model variants, *EVE-Image* and *EVE-ROI*, for image and ROI features, respectively. EVE-Image incorporates a pre-trained ResNet101 [26], which generates k feature maps of size $d \times d$. For each feature map position, the feature vector across all the k feature maps is considered as an *object*. As a result, there are a total number of $d \times d$ objects of feature size k for an input image. In contrast, the EVE-ROI variant takes ROIs as objects extracted from a pre-trained Mask R-CNN [48].

In order to accurately solve this cross-model VE task, we need: both a mechanism to identify the salient features in images and text inputs and a cross-modal embedding to effectively learn the image-text interactions, which are addressed by employing *self-attention* and *text-image attention* techniques in the EVE model respectively. We next describe the design and implementation of the mechanisms in EVE model.

4.1. Self-Attention

EVE utilizes self-attention [69] in both text and image branches as highlighted with dotted blue frame in Figure 4. Since the hypothesis in SNLI-VE can be relatively long and complex, self-attention helps focus on important keywords in a sentence that relate to each other. The text branch applies self-attention to the projected word embeddings from a multi-layer perceptron (MLP). It is worth noting that although word embeddings, either from GloVe or other existing models, may be fixed, the MLP transformation is able to be trained to generate adaptive projected word embeddings. Similarly, the image branch applies the self-attention to projected image regions either from the aforementioned feature maps or ROIs in expectation of capturing the hidden relations between elements in the same feature space.

Specifically, we use the scaled dot product (SDP) attention in [69] to capture this hidden information:

$$Att_{\text{sdp}} = \text{softmax}(\frac{RQ^T}{\sqrt{d_k}}) \tag{1}$$

$$Q_{Att} = Att_{\text{sdp}} Q \tag{2}$$

where $Q \in \mathbb{R}^{M \times d_k}$ is the *query* feature matrix and $R \in \mathbb{R}^{N \times d_k}$ is the *reference* feature matrix. M and N represent the number of features vectors in matrix Q and R respectively, and d_k denotes the dimension of each feature vector. $Att_{\text{sdp}} \in \mathbb{R}^{N \times M}$ is the resulting attention mask for Q given R. Each element a_{ij} in Att_{sdp} represents how much weight (before scaled by $\frac{1}{\sqrt{d_k}}$ and normalized by softmax) the model should put on each query feature vector $q_{j \in \{1,2,\ldots,M\}} \in \mathbb{R}^{d_k}$ in Q w.r.t. each reference feature vector $r_{i \in \{1,2,\ldots,N\}} \in \mathbb{R}^{d_k}$ in R. The attended query feature matrix $Q_{Att} \in \mathbb{R}^{N \times d_k}$ is the weighted and fused version of the original query feature matrix Q, calculated by the matrix dot product between the attention mask Att_{sdp} and the query feature matrix Q. Note that for the self-attention, the query matrix $Q \in \mathbb{R}^{M \times d_k}$ and the “reference” matrix $R \in \mathbb{R}^{N \times d_k}$ are the same matrix.

4.2. Text-Image Attention

Multi-modal tasks such as phrase grounding [6] demonstrate that high-quality cross-modal feature interactions improve the overall performance. The dotted red frame highlighted area in Figure 4 shows that EVE incorporates the text-image attention to relevant image regions based on the

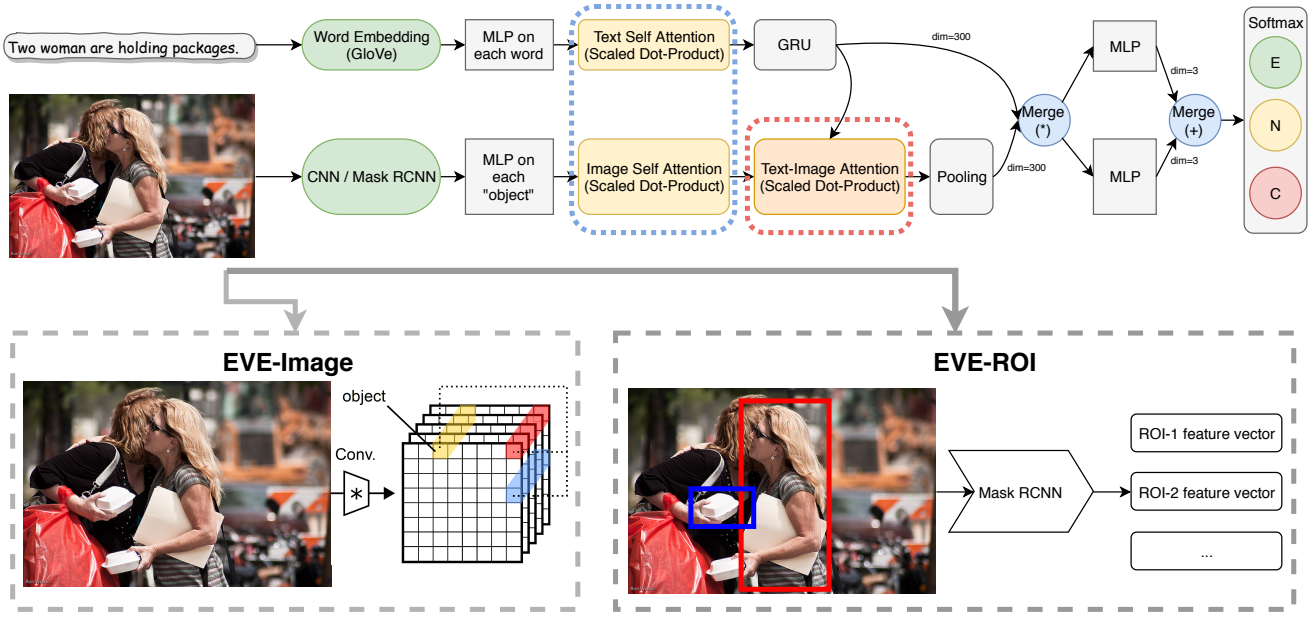

Figure 4. **Our model EVE combines image and ROI information to model fine-grained cross-modal information**

text embedding from the GRU. The feature interaction between the text and image regions are computed using the same SDP technique introduced in Section 4.1, serving as the attention weights. The weighted features of image regions are then fused with the text features for further decision making. Specifically, for the text-image attention, the query matrix $Q \in \mathbb{R}^{M \times d_k}$ is the image features while the "reference" matrix $R \in \mathbb{R}^{N \times d_k}$ is the text features. Note that although Q and R are from different feature spaces, the dimension of each feature vector is projected to be the same d_k in respective branches for ease of the attention calculation.

5. Experiments

In this section, we evaluate EVE as well as several other baseline models on SNLI-VE. Most of the baselines are existing or previous SOTA VQA architectures. The performance results of all models are listed in Table 3.

All models are implemented in PyTorch. We use the pretrained GloVe.6B.300D for word embedding [53], where 6B is the corpus size and 300D is the embedding dimension. Input hypotheses are padded to the maximum sentence length in a batch. Note we do not truncate the sentences because unlike VQA where the beginning of questions typically indicates what is asked about, labels of VE task may depend on keywords or small details at the end of sentences. For example, truncating the hypothesis "The person who is standing next to the tree and wearing a blue shirt is playing _____" inevitably loses the key detail and changes the conclusion. In addition, the maximum sentence length in SNLI is 56, which is much larger than 23 in VQA-v2.0 as shown in Table 2. Always padding to the dataset maximum is not necessarily efficient for training. As a consequence, we opt for padding to the batch-wise maximum sentence length.

Unless explicitly mentioned, all models are trained using a cross-entropy loss function optimized by the Adam optimizer with a batch size of 64. We use an adaptive learning rate scheduler which reduces the learning rate whenever no improvement on the validation dataset for a period of time. The initial learning rate and weight decay are both set to be $1e-4$. The maximum number of training epochs is set to 100. We save a checkpoint whenever the model achieves a higher overall validation accuracy. The final model checkpoint selected for testing is the one with the highest lowest per class accuracy in case the model performance is biased towards particular classes. The batch size is set as 32 for validation and testing. In the following, we discuss the details for each baseline.

5.1. Hypothesis Only

This baseline verifies the existing data bias in the SNLI dataset, as mentioned by Gururangan *et al*. [24] and Vu *et al*. [71], by using hypotheses only without the image premise information.

The model consists of a *text processing component* followed by two FC layers. The text processing component is used to extract the text feature from the given hypothesis. It first generates a sequence of word-embeddings for the given text hypothesis. The embedding sequence is then fed into a GRU [10] to output the text features of dimension 300. The input and output dimensions of the two FC layers are [300, 300] and [300, 3] respectively.

Without any premise information, this baseline is supposed to make a random guess out of the three classes but

Model Name	Val Acc Overall (%)	Val Acc Per Class (%) C	N	E	Test Acc Overall (%)	Test Acc Per Class (%) C	N	E
Hypothesis Only	66.68	67.54	66.90	65.60	66.71	67.60	67.71	64.83
Image Captioning	67.83	66.61	69.23	67.65	67.67	66.25	70.69	66.08
Relational Network	67.56	67.86	67.80	67.02	67.55	67.29	68.86	66.50
Attention Top-Down	70.53	70.23	68.66	72.71	70.30	69.72	69.33	71.86
Attention Bottom-Up	69.34	**71.26**	70.10	66.67	68.90	70.52	**70.96**	65.23
EVE-Image*	**71.56**	71.04	**70.55**	73.10	**71.16**	**71.56**	70.52	71.39
EVE-ROI*	70.81	68.55	68.78	**75.10**	70.47	67.69	69.45	**74.25**

Table 3. **Model Performance on SNLI-VE dataset**

the resulting accuracy is up to 67%, implying the existence of a dataset bias. We do not intend to rewrite the hypotheses in SNLI to reduce the bias but instead, aim at using the premise (image) features to outperform the hypothesis only baseline.

5.2. Image Captioning

Since the original SNLI premises are image captions, a straightforward idea to address VE is to first apply an image caption generator to convert image premises to text premises and then followed by a TE classifier. Particularly, we adopt the PyTorch tutorial implementation [9] as a caption generator. A pre-trained ResNet152 serves as the image encoder while the caption decoder is a long short-term memory (LSTM) network. Once the image caption is generated, the image premise is replaced with the caption and the original VE task is reduced to a TE task. Similar to the Hypothesis-Only baseline, the TE classifier is composed of two text processing components to extract text features from both the premise and hypothesis. The text features are fused and go through two FC layers with input and output dimensions of [600, 300] and [300, 3] for the final prediction.

The resulting performance achieves a slightly higher accuracy of 67.83% and 67.67% on the validation and testing partitions over the Hypothesis-Only baseline, implying that the generated image caption premise does not improve much. We suspect that the generated captions may not cover the necessary information in the image as required by the hypothesis to make the correct conclusion. This is possible in a complex scene where exhaustive enumeration of captions may be needed to cover every detail potentially described by the hypothesis.

5.3. Relational Network

The Relational Network (RN) baseline is based on [60] which is proposed to tackle the CLEVR dataset with high accuracy. There are an image branch and a text branch in the model. The image branch extracts image features in a similar manner as EVE, as described in Section 4, but without self-attention. The text branch generates the hypothesis embedding through an RNN. The highlight of RN is to capture pairwise feature interactions between image regions and the text embedding. Each pair of image region feature and question embedding goes through an MLP. The final classification takes the element-wise sum over the MLP output for each pair as input.

Despite the high accuracy on the synthetic dataset CLEVR, RN only achieves a marginal improvement on SNLI-VE at the accuracy of 67.56% and 67.55% on the validation and testing partitions. This may be attributed to the limited representational power of RN that fails to produce effective cross-modal feature fusion of the natural image premises and the free-form text hypothesis input from SNLI-VE.

5.4. Attention Top-Down and Bottom-Up

We consider the Attention Top-Down and Attention Bottom-Up baselines based on the winner of VQA challenge 2017 [1]. Similar to the RN baseline, there is an image branch and a text branch. The difference between the image branches in Attention Top-Down and Attention Bottom-Up is similar to our EVE. The image features of Attention Top-Down come from the feature maps generated from a pre-trained CNN. As for Attention Bottom-Up, the image features are the top 10 ROIs extracted from a pre-trained Mask-RCNN implementation [25]. No self-attention is applied in both image and text branches. Moreover, the text-image attention is implemented by feeding the concatenation of both image and text features into an FC layer to derive the attention weights rather than using SDP as described in Section 4.1. Then the attended image features and text features are projected separately and fused by dot product. The fused features go through two different MLPs. The element-wise sum of both MLP output serves as the final features for classification.

The SOTA VQA winner model, Attention Top-Down, achieves an accuracy of 70.53% and 70.30% on the validation and testing partitions respectively, implying cross-modal attention is the key to effectively leveraging image premise features. The Attention Bottom-Up model using ROIs also achieves a good accuracy of 69.34% and 68.90% on the validation and testing partitions. The reason why

Attention Bottom-Up performs worse than Attention Top-Down could be possibly due to lack of background information in ROI features and ROI feature quality. It is not guaranteed that those top ROIs cover necessary details described by the hypothesis. However, even with more than 10 ROIs, we observe no significant improvement in performance.

5.5. EVE-Image and EVE-ROI

The details of our EVE architecture have been described in Section 4. EVE-Image achieves the best performance of 71.56% and 71.16% accuracy on the validation and testing partitions respectively. The performance of EVE-ROI is similar, with an accuracy of 70.81% and 70.47%, possibly suffering from similar issues as the Attention Bottom-Up model. However, the improvement is likely due to the introduction of self-attention and text-image attention through SDP that potentially captures the hidden relations in the same feature space and better attended cross-modal feature interaction.

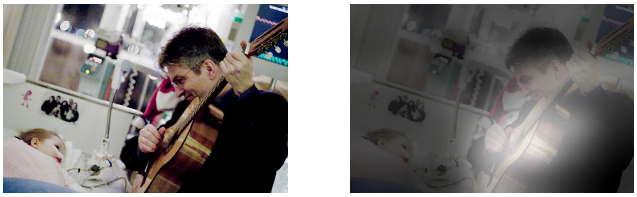

Figure 5. An attention visualization for EVE-Image

Figure 6. **An attention visualization for EVE-ROI**

Attention Visualization. The explainability of EVE is attained using attention visualizations in the areas of interest in the image premise given the hypothesis. Figure 5 and 6 illustrate two visualization examples of the *text-image attention* from EVE-Image and EVE-ROI respectively. The image premise of the EVE-Image example is shown on the left of Figure 5, and the corresponding hypothesis is *"A human playing guitar"*. On the right of Figure 5, our EVE-Image model successfully attends to the guitar area, leading to the correct conclusion: *entailment*. In Figure 6, our EVE-ROI focuses on the children and the sand area in the image premise, leading to the *contradiction* conclusion for the given hypothesis *"Two children are swimming in the ocean."*

5.6. Discussion

In this section, we discuss why existing VQA and CLEVER models have modest performs over SNLI-VE dataset and the possible future directions based on our experience. VQA models are not trained to distinguish fine-grained information. Furthermore, with the same image present across all the three classes in the SNLI-VE dataset, SNLI-VE removes any bias that may originate from just the image premise information and an effectively fused representation is important for high accuracy. Furthermore, models that provide good performance on CLEVR may not work on SNLI-VE since these models have rather simplistic image processing pipelines, often with a couple of convolutional layers that may be sufficient to process synthetic images but works poorly on real images. More importantly, the sentences are not synthetic in the SNLI-VE dataset. As a result, building compositional reasoning modules over SNLI-VE hypotheses is out of reach for existing models.

To effectively address SNLI-VE, we believe three approaches can be beneficial. First, using external knowledge beyond pre-trained models and/or visual entity extraction can be beneficial. If the external knowledge can provide information allowing the model to learn relationships between the entities that may be obvious to humans but difficult or impossible to learn from the dataset (such as "two women in the image are sisters"), it will improve the model performance over SNLI-VE.

Second, it is possible for the hypothesis to contain multiple class labels assigned to its different entities or relationships w.r.t. the premise. However, SNLI-VE lacks annotations for localizing the labels to specific entities in the hypothesis (e.g. as is often provided in synthetic datasets like bABi [72]). Since the hypothesis can be broken down into individual entities and relationships between pairs of entities, providing fine-grained labels for each target in the hypothesis likely facilitates strongly-supervised training.

Finally, a third possible approach is to build effective attention based models as done in TE that encodes the sentences together and align similar words in hypothesis and premise instead of a late-fusion of separately encoded modalities. Hence, the active research on visual grounding can benefit addressing the SNLI-VE task.

6. Conclusion

We introduce a novel task, visual entailment, that requires fine-grained reasoning over the image and text. We build the SNLI-VE dataset for VE using real-world images from Flickr30k as premises, and the corresponding text hypotheses from SNLI. We then develop the EVE architecture to address VE and evaluate against multiple baselines, including existing SOTA VQA based models. We expect more effort to be devoted to generating fine-grained

VE annotations for large image datasets such as the Visual Genome [41] and Open Images Dataset [40] as well as improved models on fine-grained visual reasoning.

Acknowledgments

Ning Xie and Derek Doran were supported by the Ohio Federal Research Network project *Human-Centered Big Data*. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the Ohio Federal Research Network.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018. 1, 2, 7

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2

[3] J. Bos and K. Markert. Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 628–635. Association for Computational Linguistics, 2005. 2

[4] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 2, 4

[5] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. *arXiv preprint arXiv:1702.05448*, 2017. 3

[6] K. Chen, R. Kovvuri, and R. Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[7] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017. 2

[8] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*, 2016. 2

[9] Y. Choi. Image captioning pytorch implementation. https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning. Accessed: 2018-10-30. 7

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 6

[11] C. Condoravdi, D. Crouch, V. De Paiva, R. Stolle, and D. G. Bobrow. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, pages 38–45. Association for Computational Linguistics, 2003. 2

[12] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006. 1, 2

[13] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[14] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 3

[15] EvalAI. VQA challenge leaderboard 2018. https://evalai.cloudcv.org/web/challenges/challenge-page/80/leaderboard/124. Accessed: 2018-11-11. 2

[16] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2

[17] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1, 2

[18] Y. Fyodorov, Y. Winter, and N. Francez. A natural logic inference system. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2)*. Citeseer, 2000. 2

[19] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015. 2

[20] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 2

[21] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*, 2017. 3

[22] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pages 529–545. Springer, 2014. 4

[23] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, page 3, 2017. 1, 2

[24] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018. 4, 6

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 7

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[27] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017. 2

[28] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[29] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 1, 2

[30] H. Jiang, B. Kim, and M. Gupta. To trust or not to trust a classifier. *arXiv preprint arXiv:1805.11783*, 2018. 3

[31] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 1, 2

[32] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017. 1, 2

[33] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017, 2017. 1, 2

[34] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 3

[35] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015. 3

[36] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[37] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018. 1, 2

[38] S. W. Kim, M. Tapaswi, and S. Fidler. Progressive reasoning by module composition. *arXiv preprint arXiv:1806.02453*, 2018. 2

[39] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017. 3

[40] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 8

[41] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1, 2, 8

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[43] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1261–1270, 2017. 3

[44] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[45] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 2

[46] B. MacCartney and C. D. Manning. An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*, pages 140–156. Association for Computational Linguistics, 2009. 2

[47] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014. 2

[48] I. Matterport. Mask rcnn pytorch implementation. https://github.com/multimodallearning/pytorch-mask-rcnn. Accessed: 2018-10-30. 5

[49] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. 3

[50] Y. Nie and M. Bansal. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*, 2017. 2

[51] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016. 3

[52] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*, 2016. 2

[53] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[54] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017. 2

[55] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013. 2

[56] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017. 3

[57] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015. 1, 2

[58] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016. 3

[59] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiskỳ, and P. Blunsom. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*, 2015. 2

[60] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 1, 2, 3, 7

[61] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 3

[62] T. Shen, T. Zhou, G. Long, J. Jiang, S. Wang, and C. Zhang. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*, 2018. 2

[63] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[64] J. Singh, V. Ying, and A. Nutkiewicz. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*, 2018. 2

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[66] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[67] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017. 1, 2

[68] A. M. Turk. Amazon mechanical turk. *Retrieved August*, 17:2012, 2012. 4

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 5

[70] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017. 2

[71] H. T. Vu, C. Greco, A. Erofeeva, S. Jafaritazehjan, G. Linders, M. Tanti, A. Testoni, R. Bernardi, and A. Gatt. Grounded textual entailment. *arXiv preprint arXiv:1806.05645*, 2018. 2, 6

[72] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 8

[73] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017. 1

[74] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 3

[76] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 4

[77] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[78] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[79] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 3

[80] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal. Relationship proposal networks. In *CVPR*, volume 1, page 2, 2017. 3

[81] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 1, 2

Supplementary: Additional Examples

Figure 7 shows random examples from SNLI-VE with predictions from our EVE-Image. Each example consists of an image premise and three selected hypotheses of different labels. Note that for each image premise, the total number of hypotheses are not limited to three.

Cameras are set up on tripods along the side of the road.
-> entailment (pred: entailment)
Cameras are set up to film a high speed police chase.
-> neutral (pred: neutral)
People are holding cameras and taking pictures of people walking inside.
-> contradiction (pred: contradiction)

People in black shirts are having a confrontation.
-> entailment (pred: entailment)
Two men, who are twins, are wearing matching black shirt and are about to fight over a girl.
-> neutral (pred: neutral)
Dogs in black shirts are having a confrontation.
-> contradiction (pred: contradiction)

A police officer dressed in a bright shirt and a black hat wets his lips.
-> entailment (pred: contradiction)
A police officer watches sailors boarding a ship.
-> neutral (pred: neutral)
A police officer driving in a car.
-> contradiction (pred: contradiction)

People sit outside the leaning tower of Piza being photographed.
-> entailment (pred: entailment)
A girl is relaxing in the park.
-> neutral (pred: contradiction)
A women pushed the Leaning Tower of Pisa until it stood straight.
-> contradiction (pred: contradiction)

An energetic boy runs around a group of people.
-> entailment (pred: entailment)
A little boy is bored and decides to run around while watching the school play of Romeo and Juliet.
-> neutral (pred: neutral)
The child was crying.
-> contradiction (pred: contradiction)

There is a person playing sports outdoors.
-> entailment (pred: entailment)
A man bring to the ball back that was thrown out of zone.
-> neutral (pred: entailment)
A man shoots a basketball at a net.
-> contradiction (pred: contradiction)

A group is with a dog outside.
-> entailment (pred: entailment)
Good friends at a park gathering having a picnic.
-> neutral (pred: neutral)
4 women, one child and a black and white dog run outside at a social event.
-> contradiction (pred: neutral)

A woman fell while playing volleyball.
-> entailment (pred: contradiction)
A woman hits the ground while trying to return a spike during a game of volleyball.
-> neutral (pred: neutral)
Girls writing letters.
-> contradiction (pred: contradiction)

A man is watching another man play a game.
-> entailment (pred: entailment)
The game is complicated and needs to be learned by demonstration
-> neutral (pred: neutral)
A young boy is throwing a ball to his dog.
-> contradiction (pred: neutral)

Children playing in a store with floor displays.
-> entailment (pred: entailment)
Two people are checking out the bed to see if they want to buy it.
-> neutral (pred: neutral)
Two children play in a Goodwill, laying under the racks of clothes that line the walls.
-> contradiction (pred: neutral)

A large family poses for a photo.
-> entailment (pred: neutral)
A family member documents a wedding by taking a photo.
-> neutral (pred: neutral)
Some people inside a church at a wedding.
-> contradiction (pred: contradiction)

There is a man wearing a black shaggy hat and a leopard printed sash.
-> entailment (pred: entailment)
a man in fancy attire holds a drum.
-> neutral (pred: entailment)
A man holds a large monkey.
-> contradiction (pred: contradiction)

Group of man playing chess in a pool.
-> entailment (pred: contradiction)
The men are trying to beat the hot, summer heat and still play chess, hence playing chess in the pool.
-> neutral (pred: neutral)
The men are playing checkers in the shade at the park.
-> contradiction (pred: contradiction)

There are people parading around.
-> entailment (pred: entailment)
A group of men is celebrating a team victory by marching down the street waving flags.
-> neutral (pred: neutral)
children are running past a flag.
-> contradiction (pred: entailment)

A few people are getting off a plane.
-> entailment (pred: entailment)
Everybody is boarding in the correct order.
-> neutral (pred: contradiction)
The plane was destroyed.
-> contradiction (pred: contradiction)

People stand on the sidewalk, wearing bright clothing.
-> entailment (pred: entailment)
People stand on a sidewalk near the beach in bright summer clothes.
-> neutral (pred: neutral)
The group of people are inside of a building.
-> contradiction (pred: contradiction)

Figure 7. Random examples from SNLI-VE with prediction results from our best-performed EVE-Image