Supplementary Material: Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions

Arijit Ray[1], **Gordon Christie**[1], **Mohit Bansal**[2], **Dhruv Batra**[3,1], **Devi Parikh**[3,1]
[1]Virginia Tech [2]UNC Chapel Hill [3]Georgia Institute of Technology
{ray93,gordonac,dbatra,parikh}@vt.edu
mbansal@cs.unc.edu

Abstract

In this supplementary material, we provide the following:

- Additional details for **RULE-BASED** (the visual vs. non-visual question detection baseline).
- Qualitative results.
- Results with other methods of feature extraction for our question-caption similarity and question-question similarity approaches used for true- vs. false-premise question detection.
- Implementation details for training the models.

1 Rule-based Visual vs. Non-Visual Classification

Section 4.1 in the main document describes **RULE-BASED**, a hand-crafted rule-based approach to detect non-visual questions. Rules were added to make this baseline as strong as possible, where some rules take precedence over others. We list a few examples:

- If there is a plural noun, without a determiner before it, followed by a verb (*e.g., "Do dogs fly?"*), the question is non-visual.
- If there is a determiner followed by a noun (*e.g., "Do dogs fly in this picture?"*), the question is visual.
- If there is a personal or possessive pronoun before a noun (*e.g., "What color is his umbrella?"*), the question is visual.
- We use a list of words that frequently occur in the non-visual questions but infrequently in visual questions. These include words such as: 'God', 'Life', 'meaning', and 'universe'. If any words from this list are present in the question, the question is classified as non-visual.

2 Qualitative Results

Here we provide qualitative results for our visual vs. non-visual question detection experiment, and our true- vs. false-premise question detection experiment.

2.1 Visual vs. Non-visual detection

Here are some examples of non-visual questions correctly detected by **LSTM**:

- "Who is the president of the United States?"
- "If God exists, why is there so much evil in the world?"
- "What is the national anthem of Great Britain?"
- "Is soccer popular in the United States?"

Here are some non-visual questions that **RULE-BASED** failed on, but that were correctly identified as non-visual by **LSTM**:

- "What color is Spock's blood?"
- "Who was the first person to fly across the channel?"

Here are some visual questions correctly classified by **LSTM**, but incorrectly classified by **RULE-BASED**:

- "Where is the body of water?"
- "What color are the glass items?"
- "What is there to sit on?"
- "How many pillows are pictured?"

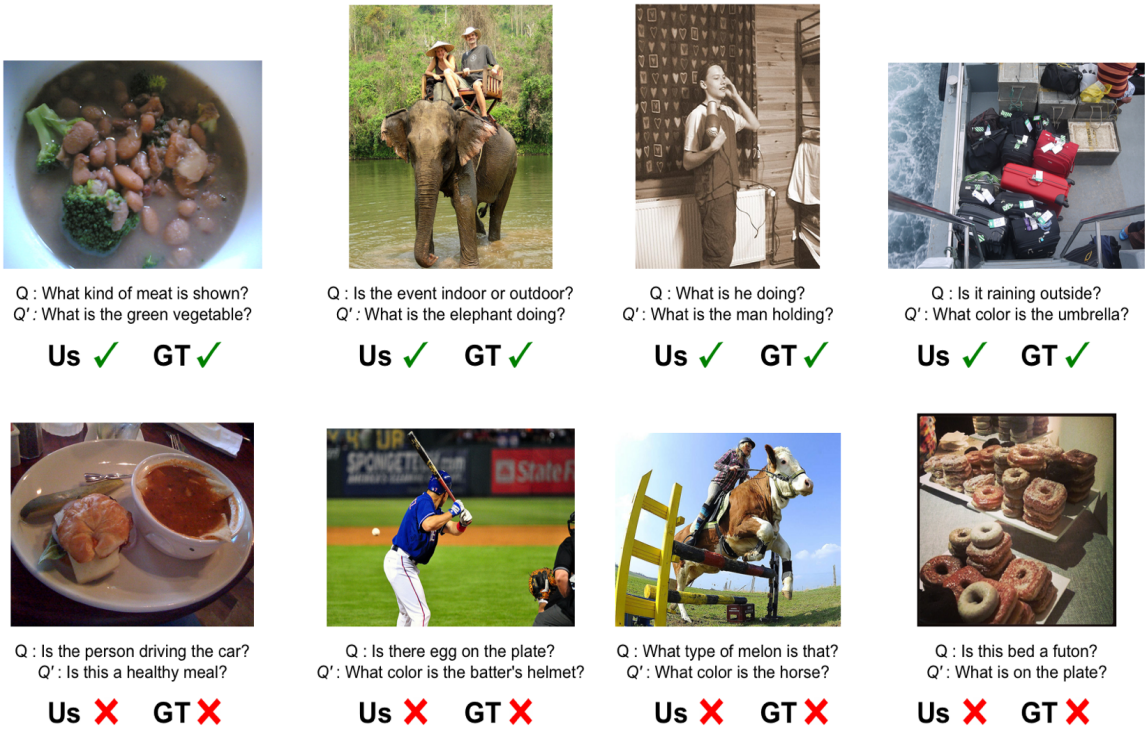

Figure 1: Success Cases: The first row illustrates examples that our model thought were true-premise, and were also labeled so by humans. The second row shows success cases for false-premise detection.

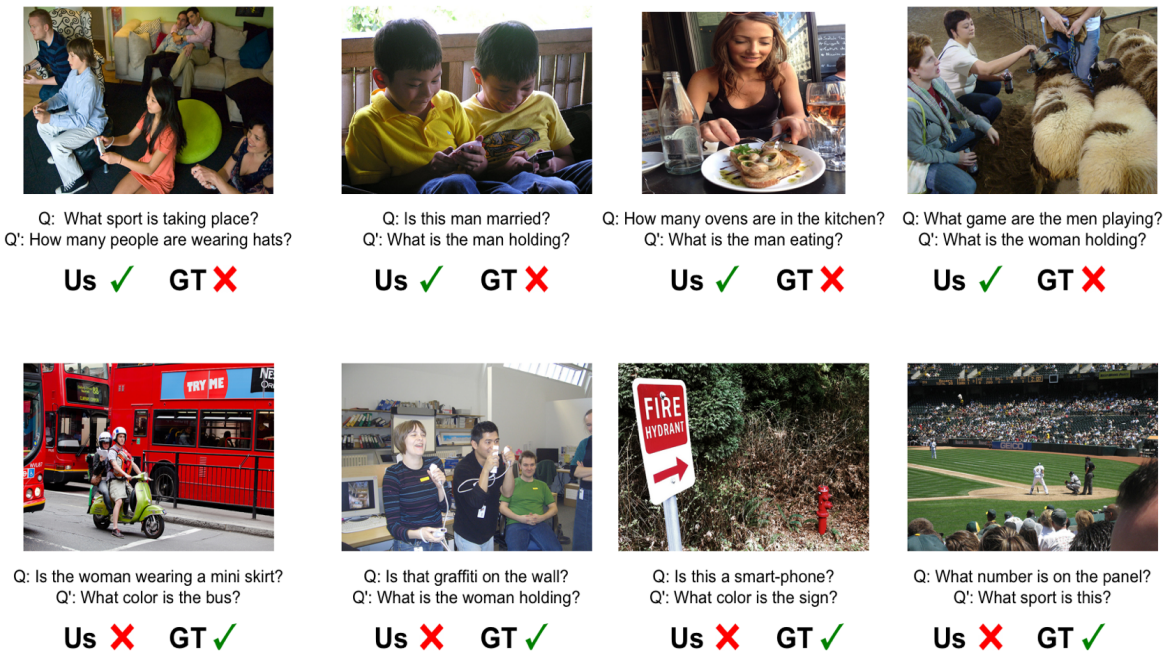

Figure 2: Failure Cases: The first row illustrates examples that our model thought was true-premise, but were actually labeled as false-premise by humans. Vice versa in the second row.

		True-Premise		False-Premise		
		Recall	Precision	Recall	Precision	**Norm Acc.**
ENTROPY		68.07	28.28	51.25	85.05	59.66
Q-GEN SCORE		64.73	25.23	50.09	84.51	57.41
VQA-MLP		57.38	36.13	71.01	85.62	64.19
Q-C SIM	**BOW**	70.48	40.19	69.91	90.46	70.19
	AVG. W2V	69.88	**48.81**	**78.35**	91.24	74.12
	LSTM W2V	72.37	46.08	76.60	91.55	74.48
Q-Q' SIM	**BOW**	68.05	44.00	75.79	90.28	71.92
	AVG. W2V	**74.62**	46.51	74.77	**92.27**	**74.69**
	LSTM W2V	74.25	44.78	74.90	91.93	74.58

Table 1: Results for true- vs. false-premise question detection, which are averaged over 40 random train/test splits.

2.2 True- vs False- Premise Detection

Figures 1 and 2 show success and failure cases for true- vs. false- premise question detection using **Q-Q' SIM**. Note that in the success cases, contextual and semantic similarity was learned even when the words in the question generated by the captioning model (Q') were different from the input question (Q).

3 Performance of Other Features

We explored three choices for feature extraction of the questions and captions:

1. **BOW**. We test a bag-of-words approach with a vocabulary size of 9,952 words to represent questions and captions, where we train an MLP to predict whether the question is relevant or not. The representation is built by setting a value of 1 in the features at the words that are present in either the question or the caption and a 2 when the word is present in both. This means each question-caption pair is represented by a 9,952-dim (vocab length) vector. The MLP used on top of **BOW** is a 5-layer MLP with 30, 20 and 10 hidden units respectively.
2. **AVG. W2V**. We extract word2vec (Mikolov et al., 2013) features for the question and captions' words, compute the average of the features separately for the question and caption and then concatenate them. Similar to **BOW**, we train a 5-layer MLP with 200, 150 and 80 hidden units, respectively.
3. **LSTM W2V**. These are the features we used in the main paper. The LSTM has 40 hidden units using a 4-layer MLP with 40 and 20 hidden units respectively.

Table 1 shows a comparison of the performance in recall, precision and normalized accuracy, where we have averaged over 40 random train/test splits.

4 Training Implementation Details

For training **BOW**, **AVG. W2V**, **LSTM W2V** and **VQA-MLP**, we use the Keras Deep learning Library (Chollet, 2015) for Python. For pre-training the question and caption generation models from scratch, we use the Torch Deep Learning Library (Collobert et al., 2011). We use *rmsprop* as the optimization algorithm (with a learning rate of 0.001) for **LSTM W2V**, and *adadelta* for **BOW** and **AVG. W2V** (initialized with a learning rate of 1). For all our models, we use a gaussian random weights initialization and no momentum.

References

Franois Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

R. Collobert, K. Kavukcuoglu, and C. Farabet. 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.