Learning Common Sense Through Visual Abstraction
Supplementary Material

Ramakrishna Vedantam[1*] Xiao Lin[1*] Tanmay Batra[2†] C. Lawrence Zitnick[3] Devi Parikh[1]
[1]Virginia Tech [2]Carnegie Mellon University [3]Microsoft Research
[1]{vrama91,linxiao,parikh}@vt.edu [2]tbatra@andrew.cmu.edu [3]larryz@microsoft.com

We first detail the webpages with qualitative results in Section 1. We then discuss our procedure for extracting commonsense tuples from sentences (Section 2). We give examples of the data we collect about relations via Abstract Scenes which includes the illustrations drawn by users and the tuples provided by them (Section 3). We then show our interface for collecting ground truth on plausibility of TEST/VAL relations (Section 4). These illustrations serve as training data for our vision based similarity model. We show some qualitative examples of cases where vision based similarity helps (Section 5).

1. Webpages

1.1. Webpage with Relation Illustration

A subset of relations along with all corresponding human illustrations collected to form the TRAIN set can be found here.

1.2. Webpage with Visual Model Predictions

The predictions from the classifier trained on visual features, to predict t_P, t_R, and t_S for Figure 5 in the main paper are shown here. These are qualitative visualizations to see which relations are most similar *visually*. We also show similarity between the predictions and the ground truth tuples using our text model based on word2vec.

1.3. Model Scores

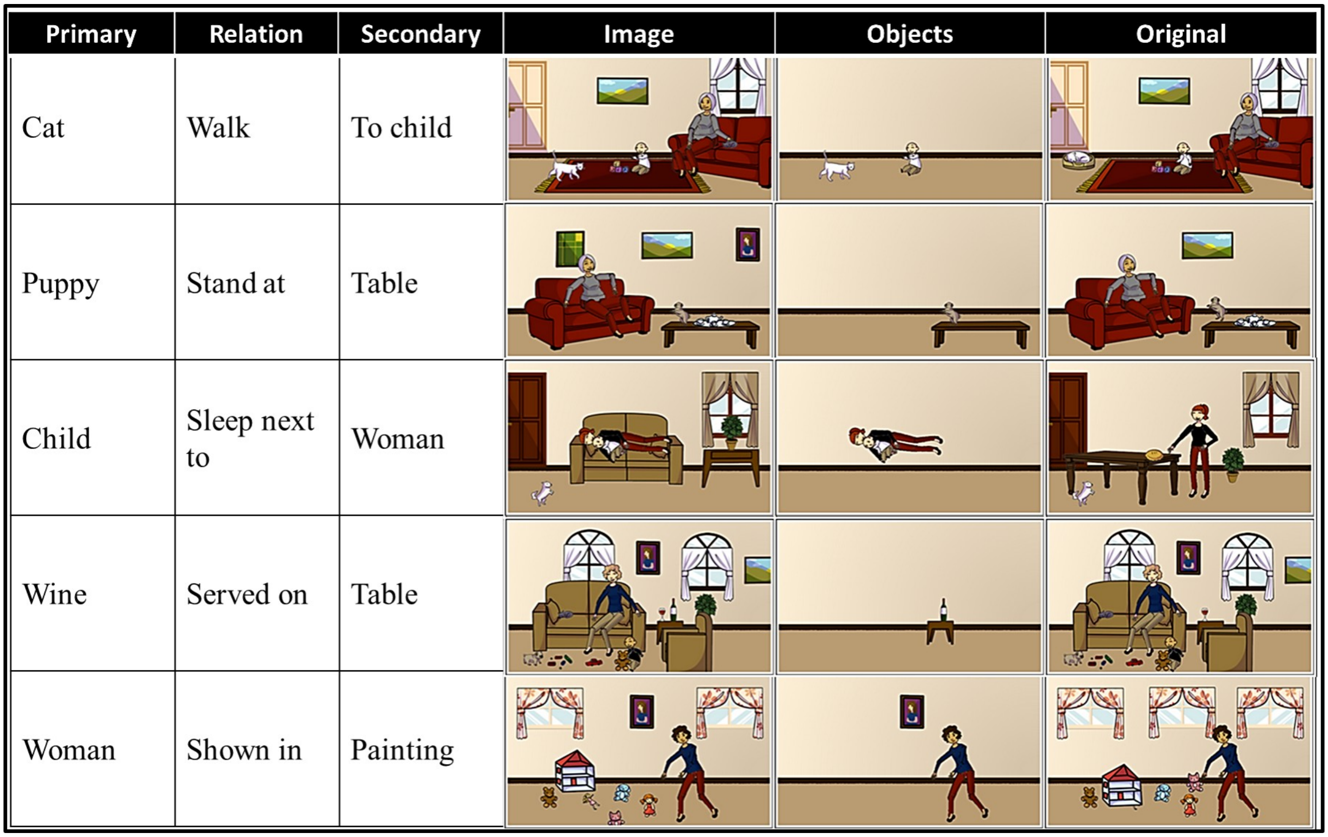
The predictions of the text+vision model, along with text only and vision only models are given, categorized by relation t_R, at the following link. The text tuples and visual illustrations which give most suport to the TEST assertion are also shown.

2. Extracting Tuples from Sentences

As described in Section 3 in the main paper, we build our VAL and TEST sets using the ReVerb information extraction system to extract our commonsense assertions. The ReVerb system segments the image into (typically) three chunks: primary object clause, relation clause and secondary object clause respectively. We do some post-processing to the ReVerb outputs to map them into our final t_P, t_R, and t_S tuples. We describe this post-processing below.

1. Get the Parts Of Speech (POS) tags for each input sentence.
2. Explore minor clauses in sentences by searching for one of the subordinating words (‘because’, ‘although’, ‘unless’, ‘however’, ‘since’) and extracting the shorter (minor) clause. In the minor clause, search for regular expression patterns: “*” is “*” to sample extra sentence chunks.
3. For all relation clauses, remove articles and pronoun instances.
4. For all relation clauses, remove the words “is” and “are”.
5. For all primary and secondary clauses, remove pronouns, articles and adjectives.
6. Split to create new relations for each instance of “and”. For example “Mike and Jenny *play* baseball” is converted to “Mike *play* baseball” and “Jenny *play* baseball”
7. Drop all relation clauses which contain a noun.
8. Perform lemmatization on all relation words. Lemmatization maps verbs to their root forms. Thus “plays” and “playing” are both mapped to “play”.
9. Convert all plural nouns occuring in primary and secondary clauses to singular form. Also remove all instances of words (‘group’, ‘couple’, ‘pair’, ‘bunch’, ‘crowd’, ‘team’, ‘two’, ‘three’, ‘four’, ‘five’).
10. Remove all clauses with empty primary clause, secondary clause or relation clause to get the tuples.

[*] Equal contribution
[†] This work was done as an intern at Virginia Tech



Primary	Relation	Secondary	Image	Objects	Original
Cat	Walk	To child			
Puppy	Stand at	Table			
Child	Sleep next to	Woman			
Wine	Served on	Table			
Woman	Shown in	Painting			

Figure 1: We show the original/ background image (last column) to the worker. The worker then illustrates a scene (column 4) containing the relation (column 2). The worker also selects the objects pariticipating in the relation (column 5) and names them (column 1 and column 3). More examples of the data we collected can be found here

3. Relation and Tuple Illustrations

Figure 1 shows the some sample illustrations created for relations, along with the corresponding tuples (Primary Object, Relation, Secondary Object) phrases. Note that workers are shown a background image, which is a generic abstract scenes image which may or may not contain the relation in question. They then modify the scene to contain the relation and then click on the primary and secondary objects through which they illustrated the relation and name the primary and secondary objects (Section 4.1 in main paper).

4. Human Supervision for Feasibility of Assertions

We describe the interface (Figure 3) we use for collecting ground truth plausibility of tuples or assertions. Workers on Amazon Mechanical Turk are shown a question and asked to rate if the scenario described by the assertion typically happens or not. We also give workers an option to tell us if the scenario described by the assertion makes no sense. We get 10 independent human responses for each such question, as described in Section 3 in the paper.

5. Examples Where Visual Cues Help

We provide qualitative examples where using visual cues with text helps (Figure 2). The figure shows some assertions which are rated by humans as *plausible*. We see that these tuples are rated as more plausible when we take the help of vision. For instance, consider the example boy *have* flower. "having" seems to find support from visual instantiations of images one would describe as "beside" (supporting cliparts row) rather than "have". However, with the visual grounding these lead to a higher score.

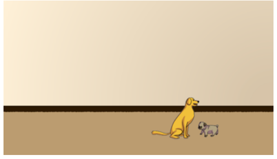
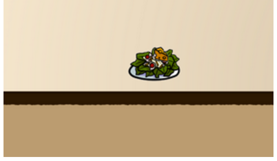
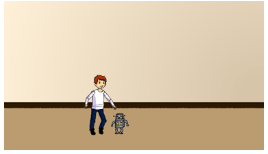
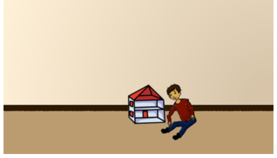
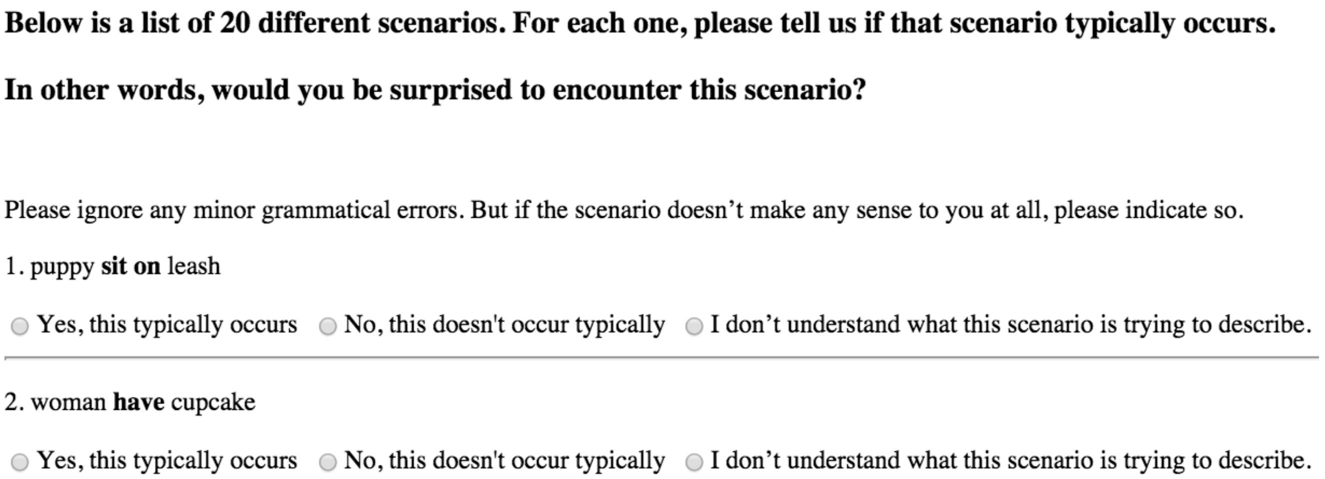
Assertion	Supporting Cliparts		Text Score	Text + Vision Score
dog *“stand with”* blanket			0.29	0.30
plate *“hold”* sandwich			0.009	0.011
boy *“have”* flower			0.01	0.08

Figure 2: We show some plausible assertions which get a higher score using text + vision than using just text, along with the clipart objects which (visually) support the assertions. More examples can be found here



Below is a list of 20 different scenarios. For each one, please tell us if that scenario typically occurs.

In other words, would you be surprised to encounter this scenario?

Please ignore any minor grammatical errors. But if the scenario doesn’t make any sense to you at all, please indicate so.

1. puppy **sit on** leash

○ Yes, this typically occurs ○ No, this doesn't occur typically ○ I don’t understand what this scenario is trying to describe.

2. woman **have** cupcake

○ Yes, this typically occurs ○ No, this doesn't occur typically ○ I don’t understand what this scenario is trying to describe.

Figure 3: Snapshot of the interface used to collect human data about plausibility of assertions