Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks

Xiao Lin Devi Parikh
Virginia Tech
{linxiao, parikh}@vt.edu

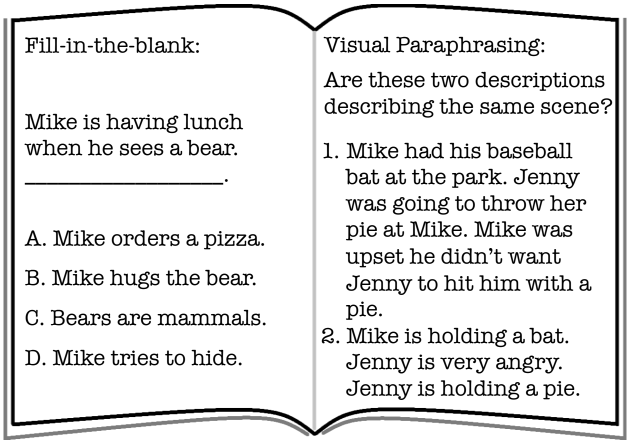
Abstract

Artificial agents today can answer factual questions. But they fall short on questions that require common sense reasoning. Perhaps this is because most existing common sense databases rely on text to learn and represent knowledge. But much of common sense knowledge is unwritten – partly because it tends not to be interesting enough to talk about, and partly because some common sense is unnatural to articulate in text. While unwritten, it is not unseen. In this paper we leverage semantic common sense knowledge learned from images – i.e. visual *common sense – in two textual tasks: fill-in-the-blank and visual paraphrasing. We propose to "imagine" the scene behind the text, and leverage visual cues from the "imagined" scenes in addition to textual cues while answering these questions. We imagine the scenes as a visual abstraction. Our approach outperforms a strong text-only baseline on these tasks. Our proposed tasks can serve as benchmarks to quantitatively evaluate progress in solving tasks that go "beyond recognition". Our code and datasets are publicly available.*

1. Introduction

Today's artificially intelligent agents are good at answering factual questions about our world [10, 16, 46]. For instance, Siri[1], Cortana[2], Google Now[3], Wolfram Alpha[4] *etc*., when asked "How far is the closest McDonald's to me?", can comprehend the question, mine the appropriate database (*e.g.* maps) and respond with a useful answer. While being good at niche applications or answering factual questions, today's AI systems are far from being sapient intelligent entities. Common sense continues to elude them.

Consider a simple fill-in-the-blank task shown in Figure 1 (left). Answering this question requires the common sense that bears are dangerous animals, people like to stay away from and not be noticed by dangerous animals, and hiding is one way of going unnoticed. Similarly, consider the visual paraphrasing question in Figure 1 (right). Answering this question involves common sense that people might throw things when they are angry and in order to throw something, you need to be holding it. Today's systems are unable to answer such questions reliably.



Fill-in-the-blank:

Mike is having lunch when he sees a bear.

__________________.

A. Mike orders a pizza.

B. Mike hugs the bear.

C. Bears are mammals.

D. Mike tries to hide.

Visual Paraphrasing:

Are these two descriptions describing the same scene?

1. Mike had his baseball bat at the park. Jenny was going to throw her pie at Mike. Mike was upset he didn't want Jenny to hit him with a pie.
2. Mike is holding a bat. Jenny is very angry. Jenny is holding a pie.

Figure 1. We introduce two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP). While they seem like purely textual tasks, they require some imagination – visual common sense – to answer.

Perhaps this is not surprising. Most existing common sense knowledge bases rely on knowledge described via text – either mined [6, 26, 31] or manually entered [37, 44, 5, 45]. There are a few short-comings of learning common sense from text. First, it has been shown that people tend not to explicitly talk about common sense knowledge in text [20]. Instead, there is a bias to talk about unusual circumstances, because those are worth talking about. Co-occurrence statistics of visual concepts mined from the web has been shown to not generalize to images [35]. Even when describing images, text is likely to talk about the salient "foreground" objects, activities, *etc*. But common sense reveals itself even in the "background". Second, much of useful common sense knowledge may be hard to describe in text. For instance, the knowledge that "one person is running after another person" implies that the first person is

[1] https://www.apple.com/ios/siri/

[2] http://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana

[3] http://www.google.com/landing/now/

[4] http://www.wolframalpha.com/

facing the second person, the second person is looking in the same direction as the first person, and both people are in running poses, is unnatural (and typically unnecessary) to articulate in text.

Fortunately, much of this common sense knowledge is depicted in our visual world. We call such common sense knowledge that can be learnt from visual data *visual common sense*. By visual common sense we do not mean visual models of commonly occurring interactions between objects [11] or knowledge of visual relationships between objects, parts and attributes [9, 50]. We mean semantic common sense, *e.g*. the knowledge that if one person is running after another person, and the second person turns around, he will see the first person. It can be learnt from visual data but can help in a variety of visual *and* non-visual AI tasks. Such visual common sense is complementary to common sense learnt from non-visual sources.

We argue that the tasks shown in Figure 1 may look like purely text- or language-based tasks on the surface, but they can benefit from visual common sense. In fact, we go further and argue that such tasks can provide exciting new benchmarks to evaluate image understanding "beyond recognition". Effectively learning and applying visual common sense to such tasks involves challenges such as grounding language in vision and learning common sense from visual data – both steps towards deeper image understanding beyond naming objects, attributes, parts, scenes and other image content depicted in the pixels of an image.

In this work we propose two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP) – as seen in Figure 1 – that can benefit from visual common sense. We propose an approach to address these tasks that first "imagines" the scene behind the text. It then reasons about the generated scenes using visual common sense, as well as the text using textual common sense, to identify the most likely solution to the task. In order to leverage visual common sense, this imagined scene need not be photo-realistic. It only needs to encode the semantic features of a scene (which objects are present, where, what their attributes are, how they are interacting, *etc*.). Hence, we imagine our scenes in an abstract representation of our visual world – in particular using clipart [51, 52, 18, 1].

Specifically, given an FITB task with four options, we generate a scene corresponding to each of the four descriptions that can be formed by pairing the input description with each of the four options. We then apply a learnt model that reasons jointly about text and vision to select the most plausible option. Our model essentially uses the generated scene as an intermediate representation to help solve the task. Similarly, for a VP task, we generate a scene for each of the two descriptions, and apply a learnt joint text and vision model to classify both descriptions as describing the same scene or not. We introduce datasets for both tasks. We show that our imagination-based approach that leverages both visual and textual common sense outperforms the text-only baseline on both tasks. Our datasets and code are publicly available.

2. Related Work

Beyond recognition: Higher-level image understanding tasks go beyond recognizing and localizing objects, scenes, attributes and other image content depicted in the pixels of the image. Example tasks include reasoning about *what* people talk about in images [4], understanding the flow of time (*when*) [39], identifying *where* the image is taken [24, 28] and judging the intentions of people in images (*why*) [40]. While going beyond recognition, these tasks are fairly niche. Approaches that automatically produce a textual description of images [22, 14, 29] or synthesize scenes corresponding to input textual descriptions [52] can benefit from reasoning about all these different "W" questions and other high-level information. They are semantically more comprehensive variations of beyond recognition tasks that test high-level image understanding abilities. However, these tasks are difficult to evaluate [29, 13] or often evaluate aspects of the problem that are less relevant to image understanding *e.g*. grammatical correctness of automatically generated descriptions of images. This makes it difficult to use these tasks as benchmarks for evaluating image understanding beyond recognition.

Leveraging visual common sense in our proposed FITB and VP tasks requires qualitatively a similar level of image understanding as in image-to-text and text-to-image tasks. FITB requires reasoning about what else is plausible in a scene given a partial textual description. VP tasks on the other hand require us to reason about how multiple descriptions of the same scene could vary. At the same time, FITB and VP tasks are multiple-choice questions and hence easy to evaluate. This makes them desirable benchmark tasks for evaluating image understanding beyond recognition.

Natural language Q&A: Answering factual queries in natural language is a well studied problem in text retrieval. Given questions like "Through which country does the Yenisei river flow?", the task is to query useful information sources and give a correct answer for example "Mongolia" or "Russia". Many systems such as personal assistant applications on phones and IBM Watson [16] which won the Jeopardy! challenge have achieved commercial success. There are also established challenges on answering factual questions posed by humans [10], natural language knowledge base queries [46] and even university entrance exams [38]. The FITB and VP tasks we study are not about facts, but common sense questions.

[19, 34] have addressed the task of answering questions about visual content. The questions and answers often come from a closed world. [42] introduces self-contained fictional stories and multiple choice reading comprehension

questions that test text meaning understanding. [47] models characters, objects and rooms with simple spatial relationships to answer queries and factual questions after reading a story. Our work can be seen as using the entire scene as the "meaning" of text.

Leveraging common sense: Common sense is an important element in solving many beyond recognition tasks, since beyond recognition tasks tend to require information that is outside the boundaries of the image. It has been shown that learning and using *non-visual* common sense (*i.e.* common sense learnt from non-visual sources) benefits physical reasoning [23, 49], reasoning about intentions [40] and object functionality [50]. One instantiation of visual common sense that has been leveraged in the vision community in the past is the use of contextual reasoning for improved recognition [22, 12, 21, 17, 25, 50]. In this work, we explore the use of visual common sense for seemingly non-visual tasks through "imagination", *i.e.* generating scenes.

Synthetic data: Learning from synthetic data avoids tedious manual labeling of real images. It also provides a platform to study high-level image understanding tasks without having to wait for low-level recognition problems to be solved. Moreover, synthetic data can be collected in large amounts with high density without suffering from a heavy-tailed distribution, allowing us to learn rich models. Previous works have looked at learning recognition models from synthetic data. For instance, computer graphics models were used to synthesize data to learn human pose [43], chair models [2], scene descriptions and generation of 3D scenes [8]. Clipart data has been used to learn models of fine-grained interactions between people [1]. [32] warps images of one category to use them as examples for other categories. [27] uses synthetic images to evaluate low-level image features. Human-created clipart images have been used to learn which semantic features (object presence or co-occurrence, pose, expression, relative location, *etc.*) are relevant to the meaning of a scene [51] and to learn spatio-temporal common sense to model scene dynamics [18]. In this work, we learn our common sense models from human-created clipart scenes and associated descriptions. We also use clipart to "imagine" scenes in order to solve the FITB and VP tasks. Though the abstract scenes [51, 8] are not photo-realistic, they offer a semantically rich world where one can effectively generate scenes and learn semantic variations of sentences and scenes, free from the bottlenecks of (still) imperfect object recognition and detection. Despite being synthetic, it has been shown that semantic concepts learnt from abstract scenes can generalize to real images [1].

3. Dataset

We build our FITB and VP datasets on top of the Abstract Scenes Dataset [51], which has 10,020 human-created abstract scenes of a boy and a girl playing in the park. The dataset contains 58 clipart objects including the boy (Mike), the girl (Jenny), toys, background objects like trees and clouds, animals like dogs and cats, food items like burgers and pizzas, *etc.* A subset of these objects are placed in the scene at a particular location, scale, and orientation (facing left or right). The boy and the girl can have different poses (7) and expressions (5). Each one of the 10,020 scenes has textual descriptions written by two different people. We use this clipart as the representation within which we will "imagine" our scenes. We also use this dataset to learn visual common sense. While more clipart objects, expressions, poses, *etc.* can enable us to learn more comprehensive visual common sense, this dataset has been shown to contain semantically rich information [51, 52], sufficient to begin exploring our proposed tasks. We now describe our approach to creating our FITB and VP datasets.

3.1. Fill-in-the-blank (FITB) Dataset

Every description in the Abstract Scenes Dataset consists of three short sentences, typically describing different aspects of the scene while also forming a coherent description. Since we have two such descriptions for every scene, we arbitrarily place one of the two descriptions (for all scenes) into the source set and the other into the distractor set. For each image, we randomly drop one sentence from its source description to form an FITB question. We group this dropped sentence with 3 random sentences from descriptions of other images in the distractor set. The FITB task is to correctly identify which sentence in the options belongs to the original description in the question.

Removing questions where the NLP parser produced degenerate outputs, our resulting FITB dataset contains 8,959 FITB questions – 7,198 for training and 1,761 for testing. Figure 3 shows one example FITB question from our dataset. The scenes corresponding to the questions in the training set are available for learning visual common sense and text-image correspondence. The scenes corresponding to the test questions are not available at test time.

FITB is a challenging task. Many scenes share the same visual elements such as Mike and Jenny playing football. Sometimes the distractor options may seem just as valid as the ground truth option, even to humans. We conduct studies on human performance on the test set. We had 10 different subjects on Amazon Mechanical Turk (AMT) answer the FITB questions. To mimic the task given to machines, subjects were not shown the corresponding image. We found that the majority vote response (*i.e.* mode of responses) across 10 subjects agreed with the ground truth 52.87% of the time (compared to random guessing at 25%).

Some questions may be generic and ambiguous and can lead to disagreements among the subjects, while other questions have consistent responses across subjects. We find that 41% of the questions in our dataset have 7 or more subjects agreeing on the response. Of these questions, the mode of

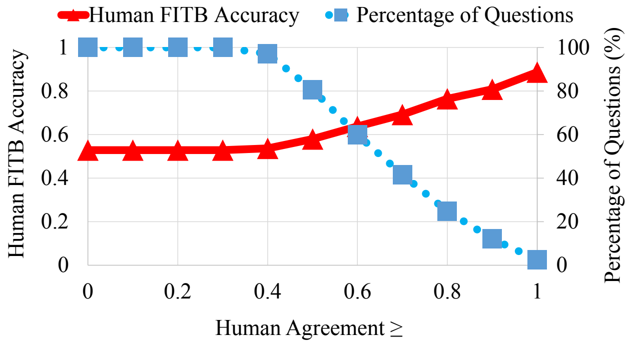


Figure 2. Human performance vs. inter-human agreement on the FITB task. Mode of human responses is more accurate when subjects agree with each other.

the responses across subjects agrees with the ground truth 69% of the time. Interestingly, on the remaining 31% of the questions, 7 out of 10 subjects agree on the *wrong* response. This happens because often the distracting options happen to describe the original image well, or their writing style matches that of the question. In our experiments, we report accuracies relative to the ground truth response, as well as relative to the response that most subjects agree on (the latter might be more relevant from an AI perspective – if the goal is to produce human-like responses).

In Figure 2, we consider different subsets of the dataset formed by only considering questions where a certain minimum proportion of subjects agreed on the response (human agreement). For each subset, we can evaluate the accuracy of the mode response. We also look at what percentage of the dataset falls in each subset. Not surprisingly, human accuracy (mode agreeing with ground truth) correlates well with human agreement (percentage of subjects that agree with mode). Note that even if responses were random, on average 43% of subjects would agree on the mode response.

3.2. Visual Paraphrasing (VP) Dataset

The VP task is to tell if two descriptions are describing the same scene or two different scenes. The correct answer to a pair of descriptions written by two people describing the same scene is "Yes", while to randomly drawn descriptions from two different scenes is "No".

We build our VP dataset using all 10,020 scenes from the Abstract Scenes Dataset, resulting in a dataset with 10,020 positive pairs. We randomly sample 2 ×10,020 pairs as negatives. This leads to a total of 30,060 questions in our dataset. Of these, 24,000 are used for training and the rest 6,060 are used for testing. We choose the negative pairs separately in training and testing sets such that they do not overlap with each other. Figure 4 shows one example VP question from our dataset.

We evaluate human performance on our test set. We had 10 different subjects on AMT solve our tasks. We average their responses (0 for No and 1 for Yes) to obtain a score between 0 and 1 for each question. We can use this score to plot a precision-recall curve. Results show that humans can reliably solve this task with 94.78% average precision (AP), compared to chance at 33%.

FITB and VP tasks are ways to evaluate visual common sense. Some applications of FITB tasks may be automatic story telling and automatic Q&A. Some applications of the VP task may be text-based image retrieval and generating multiple diverse descriptions of the same image.

4. Approach

We first describe the strong baseline approach of using textual features (common sense) to solve the FITB and VP tasks in Section 4.1. We then describe our visual common sense model (Section 4.2.2) and scene generation approach (Section 4.3). Finally in Section 4.4 we describe our approach to using our model to solve the FITB and VP tasks.

4.1. Text Only Model

We first tokenize all words in our dataset and form a vocabulary (1,886 words for the FITB dataset and 2,495 for the VP dataset). We also form a vocabulary of pairs of words by selecting 100 pairs of words which have the highest mutual information in the training data and co-occur more than 100 times.

Both FITB and VP involve reasoning about consistency between two descriptions (question and option for FITB and two input descriptions for VP). Given two descriptions d_1 and d_2, we extract three kinds of textual features from the pair. The first is term frequency, commonly used for text classification and retrieval, which counts how often each word from our vocabulary occurs in (d_1, d_2) (both descriptions concatenated). The second is a 400D word co-occurrence vector indicating for each (of the 100) pair of words whether: (i) the first word occurred in d_1 and the second word occurred in d_2 or (ii) the first word occurred in d_1 and the second word did not occur in d_2 or (iii) the first word did not occur in d_1 and the second word occurred in d_2 or (iv) the first word did not occur in d_1 and the second word did not occur in d_2. The third uses a state-of-the-art deep learning based word embedding representation word2vec [36] trained on questions from our training set to represent each word with a (default) 200D vector. We then average the vector responses of all words in (d_1, d_2). These features capture common sense knowledge about which words are used interchangeably to describe the same thing, which words tend to co-occur in descriptions, *etc.*

Fill-in-the-blank. For N fill-in-the-blank questions and M options per question, we denote the question as $q_i, i \in \{1, \ldots, N\}$ and the options for q_i as $o_{ij}, j \in \{1, \ldots, M\}$. We denote the ground truth option for question q_i as o_i^{gt}, and its index as j_i^{gt}.

The FITB problem is a ranking problem: given q_i, we wish to rank the correct option o_i^{gt} above distractors $o_{ij}, j \neq j_i^{gt}$. For each question-option pair (q_i, o_{ij}), we extract the three kinds of textual features as described above using $d_1 = q_i$ and $d_2 = o_{ij}$. Concatenating these three gives us a 2,486D text feature vector $\phi_{fitb}^{text}(q_i, o_{ij})$. We compute scores $s_{ij} = w^T \phi_{fitb}^{text}(q_i, o_{ij})$ for each option that captures how likely o_{ij} is to be the answer to q_i. We then pick the option with the highest score. We learn w using a ranking SVM [7]:

$$\begin{aligned} \min_{w,\xi\geq 0} \quad & \frac{1}{2}\|w\|^2 + C \sum_{(i,j),j\neq j^{gt}} \xi_{ij}^2 \\ \textit{s.t.} \quad & w^T \phi_{fitb}^{text}(q_i, o_i^{gt}) - w^T \phi_{fitb}^{text}(q_i, o_{ij}) \geq 1 - \xi_{ij}, \\ & \forall (i,j), j \neq j^{gt} \end{aligned} \tag{1}$$

Visual paraphrasing. In visual paraphrasing, for each question i, the goal is to verify if the two given descriptions q_{i1} and q_{i2} describe the same image ($y_i = 1$) or not ($y_i = -1$). We extract all three features described above using $d_1 = q_{i1}$ and $d_2 = q_{i2}$. Let's call this ϕ_{vp1}^{text}. We extract the same features but using $d_1 = q_{i2}$ and $d_2 = q_{i1}$. Let's call this ϕ_{vp2}^{text}. To ensure that the final feature representation is invariant to changing the order of the two descriptions – *i.e.* $\phi_{vp}^{text}(q_{i1}, q_{i2}) = \phi_{vp}^{text}(q_{i2}, q_{i1})$, we use $\phi_{vp}^{text} = [\phi_{vp1}^{text} + \phi_{vp2}^{text}, |\phi_{vp1}^{text} - \phi_{vp2}^{text}|]$ *i.e.* a concatenation of the summation of ϕ_{vp1}^{text} and ϕ_{vp2}^{text} with the absolute difference between the two. This results in a $(2 \times 2,495) + (2 \times 200) + (2 \times 400) =$ 6,190D feature vector ϕ_{vp}^{text} describing (q_{i1}, q_{i2}). We then train a binary linear SVM to verify whether the two descriptions are describing the same image or not.

4.2. Incorporating Visual Common Sense

Our model extends the baseline text-only model (Section 4.1) by using an "imagined" scene as an intermediate representation. "Imagining" a scene involves setting values for all of the variables (*e.g.* presence of objects, their location) that are used to encode scenes. This encoding, along with priors within this abstraction that reason about which scenes are plausible, serve as our representation of visual common sense. This is in contrast with traditional knowledge base representations used to encode common sense via text [50, 40]. Exploring alternative representations of visual common sense is part of future work.

Given a textual description S_i, we generate a scene I_i. We first describe our scoring function that scores the plausibility of the (S_i, I_i) pair. We then (Section 4.3) describe our scene generation approach.

Our scoring function

$$\Omega(I_i, S_i) = \Phi(S_i) + \Phi(I_i) + \Psi(I_i, S_i) \tag{2}$$

captures textual common sense, visual common sense and text-image correspondence. The textual common sense term $\Phi(S_i) = w^T \phi^{text}(S_i)$ only depends on text and is the same as the text-only baseline model (Section 4.1). Of the two new terms, $\Phi(I_i)$ only depends on the scene and captures visual common sense – it evaluates how plausible the scene is (Section 4.2.2). Finally, $\Psi(I_i, S_i)$ depends on both the text description and the scene, and captures how consistent the imagined scene is to the text (Section 4.2.3). We start by describing the representation we use to represent the description and to encode a scene via visual abstractions.

4.2.1 Scene and Description Encoding

The set of clipart in our visual abstraction were described in Section 3. More details can be found in [51]. In the generated scenes, we represent an object O_k using its presence $e_k \in \{0, 1\}$, location x_k, y_k, depth z_k (3 discrete scales), horizontal facing direction or orientation $d_k \in \{-1, 1\}$ (left or right) and attributes f_k (poses and expressions for the boy and girl). The sentence descriptions S_i are represented using a set of predicate tuples T_l extracted using semantic roles analysis [41]. A tuple T_l consists of a primary noun A_l, a relation r_l and an optional secondary noun B_l. For example a tuple can be (Jenny, fly, Kite) or (Mike, be angry, N/A). There are 1,133 nouns and 2,379 relations in our datasets. Each primary noun A_l and secondary noun B_l is mapped to 1 of the 58 clipart objects a_l and b_l respectively which have the highest mutual information with it in training data. We found this to work reliably.

4.2.2 Visual Common Sense

We breakdown and introduce the factors in $\Phi(I_i)$ into per-object (unary) factors $\Phi^u(O_k)$ and between-object (pairwise) factors $\Phi^{pw}(O_{k_1}, O_{k_2})$.

$$\Phi(I_i) = \sum_k \Phi^u(O_k) + \sum_{k_1,k_2} \Phi^{pw}(O_{k_1}, O_{k_2}) \tag{3}$$

Per-object (unary) factors $\Phi^u(O_k)$ capture presence, location, depth, orientation and attributes. This scoring function will be parameterized by w's[5] that are shared across all objects and pairs of objects. Let L be the log probabilities (MLE counts) estimated from training data. For example, $L_e^u(e_k) = \log P(e_k)$, where $P(e_k)$ is the proportion of images in which object O_k exists, and $L_{xyz}^u(x_k, y_k|z_k) = \log P(x_k, y_k|z_k)$, where $P(x_k, y_k|z_k)$ is the proportion of times object O_k is at location (x_k, y_k) given that O_k is at depth z_k.

$$\begin{aligned} \Phi^u(O_k) =& w_e^u L_e^u(e_k) + w_{xyz}^u L_{xyz}^u(x_k, y_k|z_k) + w_z^u L_z^u(z_k) \\ &+ w_d^u L_d^u(d_k) + w_f^u L_f^u(f_k) \end{aligned} \tag{4}$$

Between-object (pairwise) factors $\Phi^{pw}(O_{k_1}, O_{k_2})$ capture co-occurrence of objects and their attributes, as well as relative location, depth and orientation.

[5] Overloaded notation with parameters learnt for the text-only baseline in Section 4.1

$$\Phi^{pw}(O_{k_1},O_{k_2}) = w_e^{pw}L_e^{pw}(e_{k_1},e_{k_2}) + w_{xyd}^{pw}L_{xyd}^{pw}(dx,dy) + w_z^{pw}L_z^{pw}(z_{k_1},z_{k_2}) + w_d^{pw}L_d^{pw}(d_{k_1},d_{k_2}) + w_f^{pw}L_f^{pw}(f_{k_1},f_{k_2}) \tag{5}$$

Here the relative x-location is relative to the orientation of the first object *i.e.* $dx = d_{k_1}(x_{k_1} - x_{k_2})$. Relative y-location is $dy = y_{k_1} - y_{k_2}$. These capture where O_{k_2} is from the perspective of O_{k_1}. The space of (x, y, z) is quite large (typical image size is 500 x 400). So to estimate the probabilities reliably, we model the locations with GMMs. In particular, the factor $L^u_{xyz}(x_k, y_k|z_k)$ is over 27 GMM components and $L^{pw}_{xyd}(dx, dy)$ is over 24 GMM components.

Notice that since the parameters are shared across all objects and pairs of objects, so far we have introduced 5 parameters in Equation 4 and 5 parameters in Equation 5. The corresponding 10 log-likelihood terms can be thought of as features representing visual common sense. The parameters will be learnt to optimize for the FITB (ranking SVM) or VP (binary SVM) tasks similar to the text-only baseline described in Section 4.1.

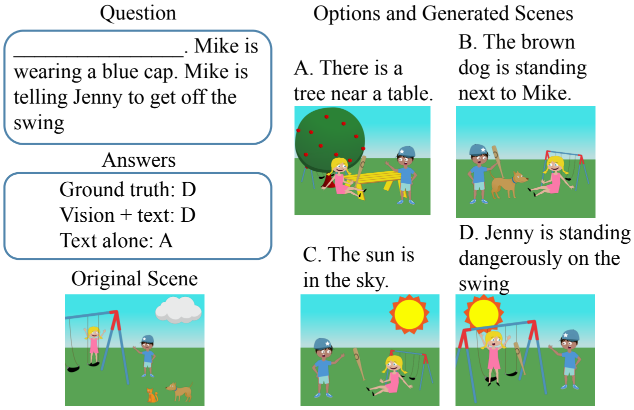
4.2.3 Text-Image Consistency

We now discuss terms in our model that score the consistency between an imagined scene and a textual description. We breakdown and introduce the text-image correspondence factors in $\Psi(I_i, S_i)$ in Equation 2 into per-noun factors $\Psi^{n+}(I_i, T_l)$ and per-relation factors $\Psi^{r+}(I_i, T_l)$ for objects that are mentioned in the description, and default per-object factors $\Psi^{u-}(O_k)$ and default between-object factors $\Psi^{pw-}(O_{k_1}, O_{k_2})$ when the respective objects are not mentioned in the description.

$$\Psi(I_i, S_i) = \sum_l \Psi^{n+}(I_i, T_l) + \sum_l \Psi^{r+}(I_i, T_l) + \sum_{k \notin S_i} \Psi^{u-}(O_k) + \sum_{k_1,k_2 \notin S_i} \Psi^{pw-}(O_{k_1}, O_{k_2}) \tag{6}$$

The per-noun factors $\Psi^{n+}(I_i, T_l)$ capture object presence conditioned on the nouns (both primary and secondary) in the tuple, and object attributes conditioned on the nouns as well as relations in the tuple. For instance, if the tuple T_l is (Jenny, kicks, ball), these terms reason about the likelihood that cliparts corresponding to Jenny and ball exist in the scene, that Jenny shows a kicking pose, *etc*. Again, the likelihood of each concept is scored by its log probability in the training data.

$$\Psi^{n+}(I_i,T_l) = w^{n+}_{abe}\left(L^{n+}_e(e_{a_l}|a_l) + L^{n+}_e(e_{b_l}|b_l)\right) + w^{n+}_{arf}L^{n+}_{arf}(f_{a_l}|a_l, r_l) + w^{n+}_{brf}L^{n+}_{brf}(f_{b_l}|b_l, r_l) \tag{7}$$



______________________. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing

Ground truth: D
Vision + text: D
Text alone: A

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing

Figure 3. Scenes generated for an example FITB question.

The per-relation factors $\Psi^{r+}(I_i, T_l)$ capture relative object location (where is b_l relative to a_l and vice versa), depth and orientation conditioned on the relation. Note that these factors are shared across all objects because "sitting next to" in (Mike, sitting next to, Jenny) and (cat, sitting next to, Jenny) is expected to have similar visual instantiations.

$$\Psi^{r+}(I_i,T_l) = w^{r+}_{rxyd}L^{r+}_{rxyd}(dx, dy|r_l) + w^{r+}_{rxyd'}L^{r+}_{rxyd'}(dx', dy'|r_l) + w^{r+}_{rz}L^{r+}_{rz}(z_{a_l}, z_{b_l}|r_l) + w^{r+}_{rd}L^{r+}_{rd}(d_{a_l}, d_{b_l}|r_l) \tag{8}$$

Here $dx' = d_{b_l}(x_{b_l} - x_{a_l})$ and $dy' = y_{b_l} - y_{a_l}$ captures where the primary object is relative to the secondary object.

The default per-object factors $\Psi^{u-}(O_k)$ and the default between-object factors $\Psi^{pw-}(O_{k_1}, O_{k_2})$ capture default statistics when an object or a pair of objects is not mentioned in the description. $\Psi^{u-}(O_k)$ captures the default presence and attribute whereas $\Psi^{pw-}(O_{k_1}, O_{k_2})$ captures the default relative location, depth and orientation.

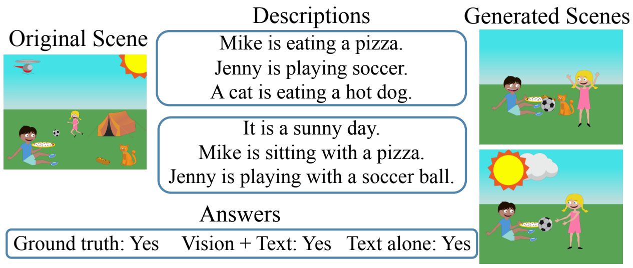
The default factors are object-specific since each object has a different prior depending on its semantic role in scenes. The default factors capture object states conditioned on the object not being mentioned in a description. We use notation D instead of L to stress this point. For example $D^{u-}_e(e_k|S_i) = \log P(e_k|k \notin S_i)$, $D^{pw-}_z(z_{k_1}, z_{k_2}|S_i) = \log P(z_{k_1}, z_{k_2}|k_1, k_2 \notin S_i)$.

$$\Psi^{u-}(O_k) = w^{u-}_{abe}D^{u-}_{abe}(e_k|S_i) + w^{u-}_{abrf}D^{u-}_{abrf}(f_k|S_i)$$
$$\Psi^{pw-}(O_{k_1}, O_{k_2}) = w^{pw-}_{rxyd}D^{pw-}_{rxyd}(dx, dy|S_i) + w^{pw-}_{rz}D^{pw-}_{rz}(z_{k_1}, z_{k_2}|S_i) + w^{pw-}_{rd}D^{pw-}_{rd}(d_{k_1}, d_{k_2}|S_i) \tag{9}$$

We have now introduced an additional 12 w parameters (total 22) that are to be learnt for the FITB and VP tasks. Notice that this is in stark contrast with the thousands of parameters we learn for the text-only baseline (Section 4.1).

4.3. Scene Generation

Given an input description, we extract tuples as described earlier in Section 4.2.1. We then use the approach



Mike is eating a pizza.
Jenny is playing soccer.
A cat is eating a hot dog.

It is a sunny day.
Mike is sitting with a pizza.
Jenny is playing with a soccer ball.

Ground truth: Yes Vision + Text: Yes Text alone: Yes

Figure 4. Scenes generated for an example VP question.

of Zitnick *et al.* [52] trained on our training corpus of clipart images and associated descriptions to generate a scene corresponding to the tuples. Briefly, it sets up a Conditional Random Field (CRF) model with a scoring function very similar to $\Phi(I_i) + \Psi(I_i, S_i)$. It samples scenes from this model using Iterative Conditional Modes with different initializations. Details can be found in [52].

4.4. Answering Questions with Imagined Scenes

Fill-in-the-blank. For FITB, we generate one scene using each question-answer pair $S_{ij} = (q_i, o_{ij})$. Fig. 3 shows qualitative examples of scenes generated for FITB. From the question-answer pair S_{ij} and the generated scenes I_{ij}, we extract features corresponding to our scoring function (Equation 2) and use them to learn the ranking SVM (Equation 1) to answer FITB questions. We choose the ranking SVM C parameter using 5 fold cross validation.

Visual paraphrasing. For VP we generate one scene for each description $S_{i1} = q_{i1}$ and $S_{i2} = q_{i2}$ in the input pair of descriptions. Fig. 4 shows qualitative examples of scenes generated for VP. We capture the difference between the two sentence descriptions by pairing the generated scenes with the *other* description *i.e.* we compute $\Omega(I_{i1}, S_{i2})$ and $\Omega(I_{i2}, S_{i1})$ (Equation 2). We extract features for both combinations, concatenate the addition of the features and the absolute difference of the features to make the mapping symmetric. These features are used to train a binary SVM that determines whether the input pair of descriptions are describing the same scene or not. We choose the SVM C parameter using 5 fold cross validation.

5. Experiments and Results

5.1. Fill-in-the-blank

We present results of our approach on the FITB dataset in Table 1. Our approach of "imagining" and joint visual-text reasoning achieves 48.04% accuracy, significantly outperforming the text-only baseline (44.97%) by 3.07% using only 22 extra feature dimensions (compared to 2,486 dimensions of the baseline). This brings the performance closer to human performance at 52.87%. [6]Leveraging visual common sense does help answering these seemingly purely text-based questions.

Approach	Fill-in-the-blank Accuracy(%)
Random	25.00
Text baseline	44.97
Visual	33.67
Text + visual (presence)	47.02
Text + visual (attribute)	46.39
Text + visual (spatial)	44.80
Text + visual (presence,attribute)	**48.60**
Text + visual (all)	48.04
Human Mode	52.87

Table 1. Fill-in-the-blank performance of different approaches.

By breaking down our 22 parameters (corresponding to visual features) into object presence (w_e^u, w_e^{pw}, w_{abe}^{n+}, w_{abe}^{u-}, 4D), attribute (w_f^u, w_f^{pw}, w_{arf}^{n+}, w_{brf}^{n+}, w_{abrf}^{u-}, 5D) and spatial configuration (w_{xyz}^u, w_z^u, w_d^u, w_{xyd}^{pw}, w_z^{pw}, w_d^{pw}, w_{rxyd}^{r+}, $w_{rxyd'}^{r+}$, w_{rz}^{r+}, w_{rd}^{r+}, w_{rxyd}^{pw-}, w_{rz}^{pw-}, w_{rd}^{pw-}, 13D) categories, we study their individual contribution to FITB performance on top of the text baseline. Object presence contributes the most (47.02%), followed by attribute (46.39%), while spatial information does not help (44.80%). In fact, only using presence and attribute features achieves 48.60%, slightly higher than using all three (including spatial). Visual features alone perform poorly (33.67%), which is expected given the textual nature of the task. But they clearly provide useful complementary information over text. In fact, text-alone (baseline), vision+text (our approach) and humans all seem to make complementary errors. Between text-alone and vision+text, 54.68% of the questions are correctly answered by at least one of them. And between text-alone, vision+text and human, 75.92% of the questions are correctly answered.

Our model is capable of imagining scenes that may contain more objects than the ones mentioned in text. Our model when using only presence does 47.02%, while a visual common sense agnostic model that only infers objects mentioned in the tuples (a_l and b_l) does 46.62%. This further demonstrates the need for visual common sense based imagination, and not treating the text at face value. If the ground truth scenes are available at test time, the performance of our approach reaches 78.04%, while humans are at 94.43%.

In addition to predicting ground truth, we also study how well our approach can mimic human responses. Our approach matches the human majority vote (mode) response 39.35% of the times (text alone: 36.40%). When re-trained using the human mode as the labels, the performance increases to 45.43%. The text-only baseline method does 42.25%. These results suggest that mimicking human is a

[6] Bootstrapping experiments show that the mean bootstrapping (100 rounds) performance of visual+text 46.33% ± 0.14% is statistically significantly better than that of text 43.65% ± 0.15%.

Approach	Visual Paraphrasing Average Precision(%)
Random	33.33
Text baseline	94.15
Visual	91.25
Text + visual (presence)	95.08
Text + visual (attribute)	94.54
Text + visual (spatial)	94.75
Text + visual (presence,attribute)	95.47
Text + visual (all)	**95.55**
Human Average	94.78

Table 2. Visual paraphrasing performance of different approaches.

more challenging task (text-only was at 44.97% when training on and predicting ground truth). Note that visual common sense is also useful when mimicking humans.

We also study how the performance of our approach varies based on the difficulty of the questions. We consider questions to be easy if humans agree on the response. We report performance of the text baseline and our model on subsets of the FITB test set where at least K people agreed with the mode. Fig. 5 shows performance as we vary K. On questions with higher human agreement, the visual approach outperforms the baseline by a larger margin.

Qualitative results can be found in the supplementary material.

5.2. Visual Paraphrasing

We present results of our approach on the VP dataset in Table 2. Our approach of generating and reasoning with scenes does 1.4% better than reasoning only with text[7]. In this task, the performance of the text-based approach is already close to human, while vision pushes it even further to above human performance[8].

Similar to the FITB task, we break down the contribution of visual features into object presence, attribute and spatial configuration categories. Presence shows the most contribution (0.93%). Spatial configuration features also help (by 0.60%) in contrast to FITB. See Table 2.

In VP, a naive scene generation model that only imagines objects that are mentioned in the description does 95.01% which is close to 95.08% where extra objects are inferred. We hypothesize that the VP task is qualitatively different from FITB. In VP, important objects that are relevant to semantic differences between sentences tend to be mentioned in the sentences. What remains is to reason about the attributes and spatial configurations of the objects. In FITB, on the other hand, inferring the unwritten objects is critical to identify the best way to complete the description. The VP task can be made more challenging by sampling pairs of descriptions that describe semantically similar scenes in the Abstract Scenes dataset [51]. These results, along with qualitative examples, can be found in the supplementary material [33].

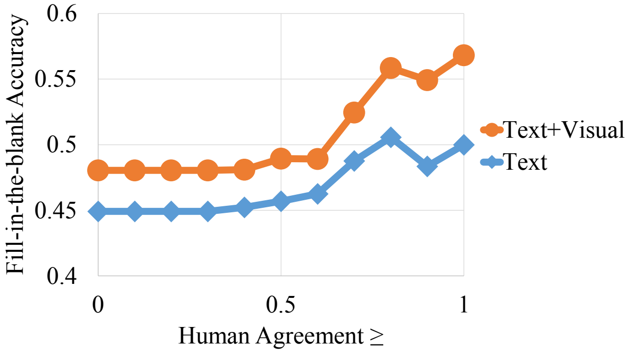


Figure 5. FITB performance on subsets of the test data with varying amounts of human agreement. The margin of improvement of our approach over the baseline increases from 3% on all questions to 6% on questions with high human agreement.

We would like to stress that FITB and VP are purely textual tasks as far as the input modality is concerned. The visual cues that we incorporate are entirely "imagined". Our results clearly demonstrate that a machine that imagines and uses visual common sense performs better at these tasks than a machine that does not.

6. Discussion

Leveraging visual knowledge to solve non-visual tasks may seem counter-intuitive. Indeed, with sufficient training data, one may be able to learn a sufficiently rich text-based model. However in practice, good intermediate representations provide benefits. This is the role that parts and attributes have played in recognition [30, 15, 48]. In this work, the imagined scenes form this intermediate representation that allows us to encode visual common sense.

In this work, we choose clipart scenes as our modality to "imagine" the scene and harness the power of visual common sense. This is analogous to works on physical reasoning that use physics to simulate physical processes [23]. These are both qualitatively different from traditional knowledge bases [9, 50], where relations between instances are explicitly represented and used during inference. Humans cannot always verbalize their reasoning process. Hence, using non-explicit representations of common sense has some appeal. Of course, alternate approaches, including more explicit representations of visual common sense are worth investigating.

Instead of generating one image per text description, one could consider generating multiple diverse images to better capture the underlying distribution [3]. Our approach learns the scene generation model and visual common sense models in two separate stages, but one could envision learning them jointly, *i.e.* learning to infer scenes for the FITB or VP tasks.

[7] Bootstrapping text+visual 95.11% ± 0.02%, text 93.62% ± 0.02%.
[8] Likely due to noise on MTurk.

7. Acknowledgment

This work was supported in part by a Google Faculty Research Award and The Paul G. Allen Family Foundation Allen Distinguished Investigator award to Devi Parikh. We thank Larry Zitnick for helpful discussions and his code.

References

[1] S. Antol, C. L. Zitnick, and D. Parikh. Zero-shot learning via visual abstraction. In *ECCV*. 2014. 2, 3

[2] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 3

[3] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*. 2012. 8

[4] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, 2012. 2

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008. 1

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010. 1

[7] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval*, 13(3):201–215, 2010. 5

[8] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, 2014. 3

[9] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 2, 8

[10] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the trec 2007 question answering track. In *TREC*, volume 7, page 63. Citeseer, 2007. 1, 2

[11] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[12] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 3

[13] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 452–457, 2014. 2

[14] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. 2010. 2

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 8

[16] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010. 1, 2

[17] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*. 2012. 3

[18] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *CVPR*, 2014. 2, 3

[19] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *PNAS*, 112(12):3618–3623, 2015. 2

[20] J. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA, 2013. ACM. 1

[21] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, 2011. 3

[22] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*. 2008. 2, 3

[23] J. Hamrick, P. Battaglia, and J. B. Tenenbaum. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA*, 2011. 3, 8

[24] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 2

[25] V. Hedau, D. Hoiem, and D. Forsyth. Recovering free space of indoor scenes from a single image. In *CVPR*, 2012. 3

[26] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: a spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61, 2013. 1

[27] B. Kaneva, A. Torralba, and W. T. Freeman. Evaluation of image features using a photorealistic virtual world. In *ICCV*, 2011. 3

[28] A. Khosla, B. An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. In *CVPR*, 2014. 2

[29] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 2

[30] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 8

[31] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014. 1

[32] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. In *NIPS*, 2011. 3

[33] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. *arXiv preprint arXiv:1502.06108*, 2015. 8

[34] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, pages 1682–1690, 2014. 2

[35] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 1

[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4

[37] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[38] A. Peñas, Y. Miyao, Á. Rodrigo, E. Hovy, and N. Kando. Overview of clef qa entrance exams task 2014. CLEF, 2014. 2

[39] L. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. Freeman. Seeing the arrow of time. In *CVPR*, 2014. 2

[40] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. *CoRR*, abs/1406.5472, 2014. 2, 3, 5

[41] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W.-t. Yih, L. Vanderwende, and C. Cherry. Msr splat, a language analysis toolkit. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstration Session*, pages 21–24. Association for Computational Linguistics, 2012. 5

[42] M. Richardson, C. J. Burges, and E. Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, pages 193–203, 2013. 2

[43] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 3

[44] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002. 1

[45] R. Speer and C. Havasi. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer, 2013. 1

[46] C. Unger, C. Forascu, V. Lopez, A. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (qald-4). CLEF, 2014. 1, 2

[47] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[48] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. *arXiv preprint arXiv:1311.5591*, 2013. 8

[49] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *CVPR*, 2013. 3

[50] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*. 2014. 2, 3, 5, 8

[51] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2, 3, 5, 8

[52] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, 2013. 2, 3, 7