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

Xiao Lin, Devi Parikh
Virginia Tech

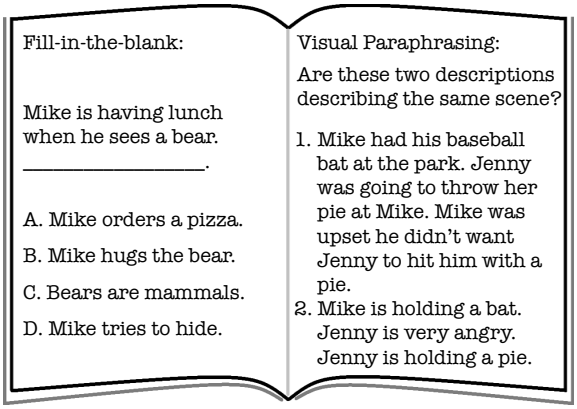

Figure 1: We introduce two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP). While they seem like purely textual tasks, they require some imagination – visual common sense – to answer.

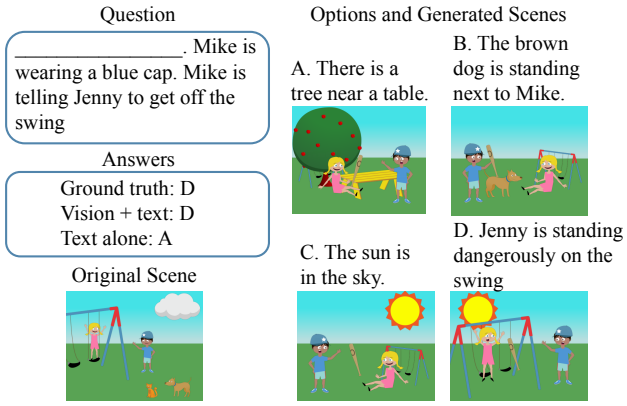

Figure 2: Scenes generated for an example FITB question in the FITB dataset. Text-based approaches only has access to text, while our approach uses both the imagined scene and text to give an answer.

Approach	FITB Accuracy(%)	VP Average Precision(%)
Random	25.00	33.33
Text baseline	44.97	94.15
Text + visual	**48.04**	**95.55**
Human	54.87	94.78

Table 1: FITB and VP performance of different approaches.

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.

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*, *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.

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 are their attributes, how are they interacting, *etc*..). Hence, we imagine our scenes in an abstract representation of our visual world – in particular using clipart [1, 2].

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.

Specifically, given an FITB task with four options, we "imagine" or generate scenes corresponding to each of the four descriptions that can be formed by pairing the input description with each of the four options, using a CRF-based scene generation approach of [2]. We then apply a learnt ranking SVM 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 SVM to classify both descriptions as describing the same scene or not. On both FITB and VP tasks, our approach has access to both text and the imagined scene. Figure 2 shows a qualitative FITB example demosntrating our imagination approach.

We introduce datasets for both tasks based on the Abstract Scenes Dataset, which has 10,020 human-created abstract scenes of a boy and a girl playing in the park. 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. 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 scene descriptions of the same image for each image in the Abstract Scenes Dataset and sampled descriptions of difference scenes for negative pairs. Our FITB dataset contains 8,959 FITB questions (7,198 for training and 1,761 for testing). Our VP dataset contains 30,060 VP questions (20,040 for training and 10,020 for testing.

Results on FITB and VP datasets (Table 1) show that our imagination-based approach that leverages both visual and textual common sense outperforms the text-only baseline on both tasks and brings performance much closer to human performance. 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 demonstrate that a machine that imagines and uses visual common sense performs better at these tasks than a machine that does not.

Our datasets and code are publicly available.

[1] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *ECCV*. 2014.

[2] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.