Visual Dialog

Xuewen Yang

June 14 2018

Abstract

We introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a question about the image, the agent has to ground the question in image, infer context from history, and answer the question accurately. Visual Dialog is disentangled enough from a specific downstream task so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress. We develop a novel two-person chat datacollection protocol to curate a large-scale Visual Dialog dataset (VisDial). VisDial contains 1 dialog (10 questionanswer pairs) on 140k images from the COCO dataset, with a total of 1.4M dialog question-answer pairs.

1. Introduction

We are witnessing unprecedented advances in computer vision (CV) and artificial intelligence (AI) C from low-level AI tasks such as image classification, scene recognition, object detection to high-level AI tasks such as learning to play Atari video games [4] and Go, answering reading comprehension questions by understanding short stories, and even answering questions about images and videos.

Despite rapid progress at the intersection of vision and language C in particular, in image captioning and visual question answering (VQA) C it is clear that we are far from this grand goal of an AI agent that can see and communicate. In captioning, the human-machine interaction consists of the machine simply talking at the human (Two people are in a wheelchair and one is holding a racket), with no dialog or input from the human. While VQA takes a significant step towards human-machine interaction, it still represents only a single round of a dialog C unlike in human conversations, there is no scope for follow-up questions, no memory in the system of previous questions asked by the user nor consistency with respect to previous answers provided by the system

2. Related Work

2.1. Vision and Language

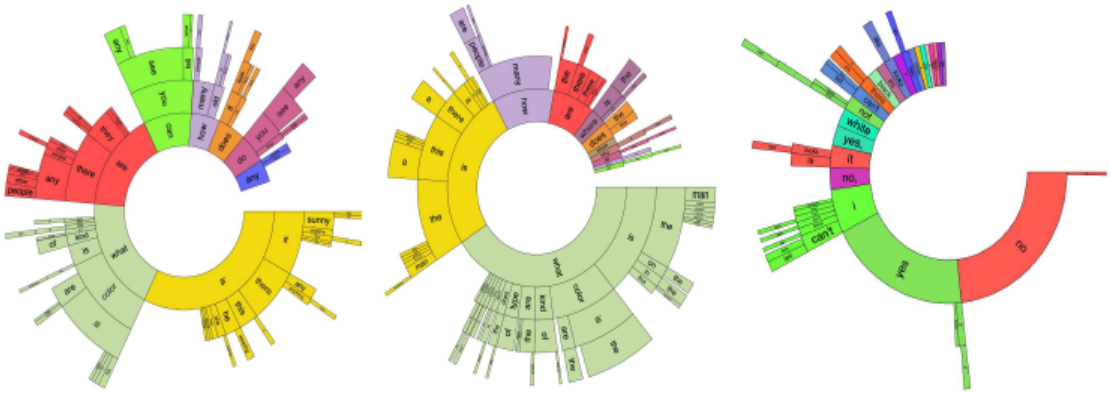
A number of problems at the intersection of vision and language have recently gained prominence C image captioning [2], video/movie description, text-to-image coreference/grounding, visual storytelling, and of course, visual question answering (VQA) [1]. However, all of these involve (at most) a singleshot natural language interaction C there is no dialog. Concurrent with our work, two recent works have also begun studying this problem of visually-grounded dialog.

2.2. Visual Turing Test

Closely related to our work is that of Geman *et al.*, who proposed a fairly restrictive Visual Turing Test C a system that asks templated, binary questions. In comparison, 1) our dataset has free-form, openended natural language questions collected via two subjects chatting on Amazon Mechanical Turk (AMT), resulting in a more realistic and diverse dataset.see figure 1 2) The dataset only contains street scenes, while our dataset has considerably more variety since it uses images from COCO. Moreover, our dataset is two orders of magnitude larger-2591 images in vs 140k images, 10 question-answer pairs per image, total of 1.4M QA pairs.

2.3. Text-based Question Answering

Our work is related to text-based question answering or reading comprehension tasks studied in the NLP community. Some recent large-scale datasets in this domain include the 30M Factoid Question-Answer corpus, 100K Simple Questions dataset, DeepMind Q&A dataset [3], the 20 artificial tasks in the bAbI dataset, and the SQuAD dataset for reading comprehension. VisDial can be viewed as a fusion of reading comprehension and VQA. In VisDial, the machine must comprehend the history of the past dialog and then understand the image to answer the question. By design, the answer to any question in VisDial is not present in the past dialog C if it were, the question would not be asked. The history of the dialog contextualizes the question C the



(a) VisDial Questions (b) VQA Questions (c) VisDial Answers

Figure 1. Distribution of first n-grams for (left to right) VisDial questions, VQA questions and VisDial answers. Word ordering starts towards the center and radiates outwards, and arc length is proportional to number of questions containing the word.

question what else is she holding? requires a machine to comprehend the history to realize who the question is talking about and what has been excluded, and then understand the image to answer the question.

2.4. Conversational Modeling and Chatbots

Visual Dialog is the visual analogue of text-based dialog and conversation modeling. While some of the earliest developed chatbots were rule-based, end-to-end learning based approaches are now being actively explored. A recent large-scale conversation dataset is the Ubuntu Dialogue Corpus, which contains about 500K dialogs extracted from the Ubuntu channel on Internet Relay Chat (IRC). Liu *et al.* perform a study of problems in existing evaluation protocols for free-form dialog. One important difference between free-form textual dialog and Vis- Dial is that in VisDial, the two participants are not symmetric C one person (the questioner) asks questions about an image that they do not see; the other person (the answerer) sees the image and only answers the questions (in otherwise unconstrained text, but no counter-questions allowed). This role assignment gives a sense of purpose to the interaction (why are we talking? To help the questioner build a mental model of the image), and allows objective evaluation of individual responses.

References

[1] A. Agrawal, J. Lu, S. Antol, C. L. Zitnick, C. L. Zitnick, D. Parikh, and D. Batra. Visual question answering. *IJCV*, 2015. 1

[2] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE TPAMI*, 2017. 1

[3] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015. 1

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 2015. 1