Chat-crowd: A Dialog-based Platform for Visual Layout Composition

Paola Cascante-Bonilla¹  Xuwang Yin¹  Vicente Ordonez¹  Song Feng²
¹University of Virginia, ²IBM Thomas J. Watson Research Center.
[pc9za, xy4cm, vicente]@virginia.edu, sfeng@us.ibm.com

Abstract

In this paper we introduce Chat-crowd, an interactive environment for visual layout composition via conversational interactions. Chat-crowd supports multiple agents with two conversational roles: agents who play the role of a designer are in charge of placing objects in an editable canvas according to instructions or commands issued by agents with a director role. The system can be integrated with crowdsourcing platforms for both synchronous and asynchronous data collection and is equipped with comprehensive quality controls on the performance of both types of agents. We expect that this system will be useful to build multimodal goal-oriented dialog tasks that require spatial and geometric reasoning.

1 Introduction

There has been growing interest in building visually grounded dialog systems (Ren et al., 2015; Bisk et al., 2016; Das et al., 2017; Chen et al., 2018). Building interactive agents that can complete goal-oriented tasks in a situated environment using natural language is a challenging problem that requires both robust natural language understanding (NLU) and natural language generation (NLG). Datasets for visually grounded dialog tasks have started to emerge but more general and effective tools for data collection are still missing.

We introduce an interactive data collection and annotation tool¹ for the collaborative tasks of visual layout composition through natural language dialogs (see Figure 1). In this work, we refer to layouts to the spatial distribution of objects in a 2D canvas as well as their attributes such as name, shape, or color. More specifically, Chat-crowd is designed to support a basic model task consisting of dialogs between a director agent that is provided with a visual layout as a reference, and a designer agent that is provided with a canvas where one can add, remove, resize, or relocate visual elements. The two agents can communicate using natural language within the context of various dialog acts (“DA”) as illustrated in Figure 1. The director provides instructions to modify elements in the canvas, and the designer optionally writes clarifying questions or modifies the canvas. While the director can see the progress of the designer as the dialog proceeds, the designer only sees the instructions from the director. The dialog ends when the visual layout composed by the designer matches the given visual layout to the director.

One key feature of our system is that it allows asynchronous conversations, i.e. the director and designer do not need to be online at the same time, or be persistent throughout the task. This means that different users can pick up the task where it was left off in the previous interaction, thus simplifying the overall collection process. Furthermore, it enables a process for data validation and job dis-

¹ chatcrowd.github.io
tribution at a finer level. Additionally, our system optionally employs a bot agent to inject synthetic utterances that trigger diversified or less represented dialog activities. Such injections can also be used for evaluating the responses from human agents since the optimal responses to the synthetic utterances are already given. We show that under this asynchronous mode, crowdsourced contributors are still able to converse based on the dialog history and complete the tasks.

To validate our system, we first apply the system on a task with more controllable visual layouts consisting of visual primitives (e.g., circles, rectangles, triangles); we also test our system on grounding for visual layouts corresponding to objects in real-images from an object recognition dataset (Lin et al., 2014). Separating the pattern recognition task from the visual understanding task through visual layouts allows us to explore richer language for spatial reasoning yet still connects to real-world images.

Our contributions are the following: (1) a new multi-modal dialog simulation system with a focus on spatial reasoning; (2) an asynchronous dialog collection platform that can trigger more diverse dialog activities and evaluate the performance of crowdsourced contributors in ongoing tasks; (3) an analysis of the difficulty and the type of language used by people to accomplish the proposed collaborative task of reconstructing visual layouts from asynchronous dialogs.

2 Visual Layout Dialog Collection

This work aims to demonstrate the usage of Chatcrowd for obtaining dialog data for geometric and spatial reasoning, ranging from abstract to more complex scenes. To this end, we explore two types of visual layouts: layouts in a shape-world with automatically generated simple 2D shape primitives, and layouts of objects from real images.

2D-shape Layouts

We propose a synthetic layout world where objects of different shapes (circle, square, triangle) and colors (blue, red, green) are pinned to a set of $5 \times 5$ grid locations on a canvas. This setup allows us to focus on the language, and the accurate reconstruction of the visual layouts by discarding the additional complexities of real-world scenes. We generate two types of 2D-shape layouts: (1) 2d-shape-random: consisting of scenes with 4 to 6 objects with shapes, color, and locations selected randomly, and 2d-shape-pattern: consisting of objects generated by following a set of customizable production rules that encourage adjacent objects. Figure 3 presents our user interface for the data collection tasks of 2D-shape layouts. For the real-image layouts, the interface includes additional features for resizing, moving, and naming objects.

COCO Layouts

We use as reference and test bed of the object layouts of real-world images from the COCO dataset (Lin et al., 2014). The layout of an image includes objects and their locations. All objects are represented by a set of rectangles (proportional to the size of the corresponding object) and the object class (e.g., people, dog, and surfboard). We also experiment with two types of scenarios: (1) COCO-simple: corresponding to images with simple layouts with 3 to 4 object instances belonging to 3 distinct classes, and (2) COCO-complex: corresponding to images consisting of layouts with 6 to 8 object instances belonging to 6 distinct object classes.

2.1 Crowdsourcing Task Design

In our task, crowd agents interact under two roles: director and designer. In the director mode, agents direct the drawing in the following ways: (1) providing instructions for how layouts should be modified; (2) giving suggestions for correcting or improving the current layout; (3) answering questions from the designer agents. In the designer mode, the agents either follow the instructions to draw on the canvas by specifying the attributes and locations for 2D-shapes/COCO, or ask clarifying questions if needed.

Data Collection

One challenge for such multi-model dialog collection via crowdsourcing is that it could be very complicated and expensive to pair two qualified contributors to converse in real time (Lasecki and Bigham, 2013). Our system is designed to support both synchronous as well as asynchronous interactions. In the asynchronous mode, the agents are asked to review and understand a chat history before taking an action. Thus, we design quizzes to assist agents in learning how to examine the chat history to determine what actions are helpful for reconstructing the layouts.

Quality Control

The most common quality assurance provided by crowdsourcing platforms is to evaluate the performance with gold standard data, which is not applicable in our case. We propose...
Figure 2: Overview of the Chat-crowd System

to verify and ensure the success of the task using the following criteria: a task is considered successful if a layout is reconstructed with high similarity with respect to the reference layout. For 2D shape layouts, a layout similarity is tested by computing an exact match; For COCO layouts the matching is confirmed by the director agents who determines when the task should end.

Output Data The output includes free-form textual utterances from director and designer agents annotated with dialog acts, a sequence of drawings on the canvas and the final layouts of images. In our task setting, the dialog data can be potentially divided into sub-dialogs or atomic dialog interactions accordingly.

3 Experiment Results and Analysis

3.1 System Overview

Figure 2 presents an overview of our system. In synchronous mode, the system allows one director agent and one designer agent to converse with each other in real time to perform a given task. In asynchronous mode, a Job Dispatcher determines a job of a role for next turn to add to the job queue. We implemented an automatic job dispatcher to interface with the crowdsourcing platforms. The agents are provided with the latest canvas and dialog history that is dynamically generated by a Dialog Interface Renderer. Once an input (e.g., dialog act, utterance, or latest canvas) is submitted, Input Validator first examine the content via its sub-module Input Analyzer. It identifies the modification to the previous canvas; object features and locations in the utterance and the dialog acts etc.. Additionally, Synthetic Generator is applied to generate responses: (1) to intrigue more diverse dialog activities such as designer asking clarification questions; (2) to inspect the performance of the contributors, for instance, when a designer submits a canvas given a non-viable instruction by Synthetic Generator.

3.2 Experimental Settings

We post the jobs on the FigureEight crowdsourcing platform. We collected dialogs for 100 2d-shape-random layouts, 100 2d-shape-pattern layouts, and additionally run a pilot study on 10 COCO-simple layouts and 10 COCO-complex layouts, leading to 2,520 individual user interactions for 2d-shapes and 595 for real scene COCO layouts.

3.3 Quantitative and Qualitative Analysis

Director Word Usage Analysis We first analyzed the types of words that people in the director agent role used to provide instructions to the designers. We found that people mention location, color, and shape words in over 90% of the total of instructions and often all three with a slight preference for mentioning shape over color information. For the 2d-shape-pattern task there was slightly lower preference to mention shape and color, than in 2d-shape-random. This is because when each object is placed randomly, then people have to more often refer to each object individually on each round.

Designer Reactions Analyzing the interactions by designer agents we found that about 60% of the times they modify the canvas without necessarily asking clarification questions. Here is a set of example responses: “I did not understand instructions from instructor...”; “please give instruction”, examples of questions are: “where to put circle?”, “do the boxes mean squares”, “where exactly, in the middle, left or right?”, “It is done?”.

Task Duration We additionally analyze the difficulty of each task by the number of rounds that it takes to complete a layout, and the average length (in words) for the instructions issued by the director. Table 1 shows these statistics for our four type of layouts. We found, unsurprisingly, that for 2d-shape-pattern layouts, the average number of rounds is significantly lower than for 2d-shape-random, indicating that the pattern in the distribution of objects in the canvas is indeed being exploited by the agents. Additionally, we can gauge the difference in difficulty between

---

\(^{2}\)We employ NLP tools by spacy.io

\(^{3}\)Formerly known as CrowdFlower.
COCO-complex and COCO-simple, where the number of rounds is more than double even when the average instruction is only two words larger.

**Table 1:** Statistics for task duration for each type of layout based on the number of rounds needed and average number of words in the instructions.

<table>
<thead>
<tr>
<th>Layout-Type</th>
<th>Rounds</th>
<th>Avg-Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>2d-shape-random</td>
<td>7.23</td>
<td>18.0 ± 14.3</td>
</tr>
<tr>
<td>2d-shape-pattern</td>
<td>5.37</td>
<td>19.4 ± 15.8</td>
</tr>
<tr>
<td>COCO-simple</td>
<td>7.6</td>
<td>18.7 ± 13.9</td>
</tr>
<tr>
<td>COCO-complex</td>
<td>22.1</td>
<td>20.9 ± 28.2</td>
</tr>
</tbody>
</table>

**Instruction Efficiency** In terms of single instruction efficiency, we found that for 2d-shape-random layouts, designer agents were able to modify more than three objects per turn only 14% of the times, while this number was 20% for 2d-shape-pattern layouts. Thus, it further confirms that people are effective at using the patterns to optimize language for the task. We also notice how the COCO layouts elicit semantic relations that are not present in the 2D-shape layouts, so while we expect that some of the language from 2D-shape layouts will translate to real-world scenes, such as references to locations, and shapes, the realm of semantic relations might need a separate treatment.

**4 Related Work**

Given some of the limitations of tasks such as human-robot interactions, or visual question answering, there has been recent interest on building more complex multimodal datasets of visually grounded dialogs (Geman et al., 2015; Mostafazadeh et al., 2017; Das et al., 2017; Kim et al., 2017). Our goal oriented task of reconstructing the spatial distribution of objects in a canvas through conversational interactions sets our task apart from these previous works. In our work, we additionally explore re-construction of layouts corresponding to real-world images with a focus on the inclusion of spatial and geometric reasoning and more dynamic dialog activities while completing the tasks.

Another aspect that sets our work apart from previous efforts in this domain is that we leverage asynchronous dialog interactions. However, there have been important previous works studying this type of dialogs in the more general setting (e.g. Blaylock et al. (2002); Joty et al. (2011); Tavafi et al. (2013)). We similarly show that our proposed visually grounded task is feasible under asynchronous dialogs.

Finally, object layouts, and visual grounding on geometric primitives has generally been of interest to study the compositionality of language. The work of Mitchell et al. (2013); FitzGerald et al. (2013) used synthetic object layouts and simple scenes to study referring expressions, while Yin and Ordonez (2017) used layouts from real images for image captioning. Andreas et al. (2016), and Johnson et al. (2017) introduced synthetic abstract scene datasets to test visual question answering. Our work is instead focused on visually grounded dialogs for spatial reasoning.

**5 Conclusions**

We developed Chat-crowd, a framework and associated platform to collect dialogs for goal-oriented tasks involving visual reasoning. Our platform incorporates mechanisms to encourage diverse dialog activities and provides a new way of evaluating the performance of crowdsourcing agents during the task. Our system demonstrated the feasibility of a director-designer agent interaction to reconstruct input visual layouts based only on asynchronous dialog interactions.
References


