Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

"Is It Rectangular?" Using I Spy as an Interactive, Game-Based Approach to Multimodal Robot Learning

†* Natalie Parde, ‡ Michalis Papakostas, ‡ Konstantinos Tsiakas, and † Rodney D. Nielsen
† Department of Computer Science and Engineering, University of North Texas
‡ Department of Computer Science and Engineering, University of Texas Arlington and NCSR Demokritos
*natalie.parde@unt.edu | http://students.cse.unt.edu/~npp0016

Abstract

Training robots about the objects in their environment requires a multimodal correlation of features extracted from visual and linguistic sources. This work abstracts the task of collecting multimodal training data for object and feature learning by encapsulating it in an interactive game, *I Spy*, played between human players and robots. It introduces the concept of the game, briefly describes its methodology, and finally presents an evaluation of the game's performance and its appeal to human players.

Introduction

As interactive robots become increasingly ubiquitous in a diverse array of applications, it is important to develop methods with which they can learn about the world around them. One way to do this is to create mappings between visual and linguistic input in order to build models of new objects and features in their environment. Although this can be done by simply and successively presenting new pairings of images and text descriptions to robots (Holzapfel, Neubig, and Waibel 2008), doing so is not particularly exciting for human participants. This work proposes to embed the tedious process of multimodal data collection in an interactive, vision- and dialogue-based game—specifically, in a robotic variant of the traditional guessing game, *I Spy*.

Background

Previously, the use of an "I Spy" or "20 Questions" format as a means of improving robot vision was explored in (Vogel, Raghunathan, and Jurafsky 2010), which focused on asking "yes" or "no" questions to improve vision-based knowledge of hard-coded object categories and attributes. Moreover, the recent work of (Krause et al. 2014) tackles the concept of "one-shot" learning, where a robot learns to recognize an object via an image and a natural-language object description. Several systems have achieved large-scale success in conceptual learning based on natural-language input: the Never-Ending Language Learner (NELL) (Carlson et al. 2010) continuously extracts information from the web to construct beliefs, and IBM's Watson (Ferrucci et al. 2010) generates hypotheses based on information extracted from natural language and learns based on the successes and failures of these hypotheses. The work in *I Spy* primarily builds upon that of (Vogel, Raghunathan, and Jurafsky 2010), by learning and utilizing features extracted from players' natural-language object descriptions within the context of an interactive, dialogue-based, human-robot game. In doing this, the system builds multimodal feature models comprised of visual feature vectors correlated to a set of keywords drawn from descriptions from players, functioning at its deepest level as a multimodal interactive learning system.

Approach

The *I Spy* game developed in this work is characterized by two phases: (1) an initial learning phase, and (2) a gaming phase. In the initial learning phase, the robot begins with an entirely empty knowledge base. To learn about an object, it captures a series of images of the object from different angles and distances. A human user also describes the object to the robot, and the robot extracts linguistic features (content keywords) from the human's description and associates those features with visual features (texture, shape, and color) extracted from the images of the object.

In the gaming phase, the robot is placed in front of a set of objects. The robot captures one or more images of the gamespace, isolates and segments the individual objects in the gamespace, and extracts each object's visual features. It uses this visual information to return associated learned features, which are then used by the decision-making module to determine which features to query the player about. It automatically generates questions regarding features chosen by the decision-making module (e.g., "Is it rectangular?" or "Does it have a stem?"), and uses the player's answers to update its belief probabilities until it is confident enough to guess the target object.

Methods

I Spy performs a number of tasks drawn from natural language processing (NLP), computer vision, and machine learning, which are summarized here. Briefly, the robot uses NLP to extract keywords from users' object descriptions, correlate co-occurring keywords with one another and with IDs representing the learned objects, and automatically generate questions based on keywords. It uses computer vision

to segment captured photos and extract visual features for each resulting segment using local binary patterns (LBP), histograms of oriented gradients (HOG), and RGB and HSV histograms. It uses Gaussian Mixture Models (GMMs) to associate the visual and linguistic features and train feature and object models. Models are retrained every 5 games.

I Spy's decision-making module considers which feature will best split the existing set of high-probability objects in half. It maintains belief probabilities for each object in play throughout the game, updated according to a player's "yes" or "no" answer to a question regarding a feature, until one of the probabilities reaches a confidence threshold.

Experiment

This work evaluates *I Spy's* gaming performance, as well as its appeal to humans, according to the results of an experiment in which 46 games were played between humans and a NAO robot (http://www.aldebaran.com/en).

Initial Learning Phase

To conduct the initial learning phase, 17 objects of varied size, color, shape, and texture were selected and photographed. Descriptions of each object were then solicited via Amazon Mechanical Turk (https://www.mturk.com/mturk/welcome). The first six descriptions matching eligibility requirements (native English speakers age 18 and over) for each object were kept, resulting in 102 descriptions from 43 Turkers (22 female).

Each object was placed on a white floor with a white background (the gamespace), and a NAO robot was programmed to automatically capture a series of 7 images of the object from different distances and angles. Visual features were extracted from these captured images and associated with the content keywords supplied in Turkers' descriptions of the object to train the feature and object models.

Gaming Phase

To capture images for the games, all 17 objects were placed in the gamespace, and the NAO robot captured photos from 3 predetermined locations. The objects were then rearranged, and this process was repeated 34 times. All resulting photos were segmented and visual feature vectors were extracted for each individual object.

Twenty-three volunteers (8 female, Age Range: 19–56) played two games each, with a different target object each time. All 34 gamespace configurations were played, with 12 played twice. 12 objects were played as the target object three times, and 5 were played twice. Following play, each player was asked to fill out a post-survey containing 6 multiple-choice Likert scale questions (ranked from 1 to 5, with 5 being best) and two free-response questions (one for comments, and one for suggestions).

Results

The robot won 32 of the 46 games played (70%). On average, it asked 13 questions before making a winning guess, and 15 before making a losing guess. Seventeen players responded to the post-game survey, with results presented in Table 1.

Table 1: Player Perceptions of *I Spy*

Question	**1**	**2**	**3**	**4**	**5**	**Avg.**
Naturalness of Questions	1	3	5	6	2	3.3
Personal Enjoyment	1	2	4	4	6	3.7
Platform (NAO) Choice	0	1	3	6	7	4.1
Interaction Level	1	2	3	3	8	3.9
Perceived Intelligence	2	3	5	6	1	3.1
Likability	0	2	0	6	9	4.3

Table 1. Answers to the free-response questions were positive (e.g., "It's awesome," "I loved it!"), and some also lent insight into the variability of answers to *I Spy* questions ("I haven't played I Spy in forever, but I remember not knowing whether to answer yes or no to plenty of things."). Finally, some offered suggestions for future versions of the game.

Discussion

The robot performed well in its games with humans, demonstrating a fairly high guessing accuracy. This accuracy was particularly notable given the quite differing perceptions (and, thus, answers) expressed by players for some of the same target objects. The average rankings for all Likert scale questions in the post-game survey were promising, with average scores for each question above 3 and with the robot's likeability, platform choice, and interaction level scoring particularly high. These results indicate that *I Spy* is a feasible and exciting way to motivate people to engage in training robots about the world around them.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1262860, CNS-1035913, IIS-1409897, and CNS-1338118.

References

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka Jr, E. R.; and Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, 3.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; et al. 2010. Building watson: An overview of the deepqa project. *AI magazine* 31(3):59–79.

Holzapfel, H.; Neubig, D.; and Waibel, A. 2008. A dialogue approach to learning object descriptions and semantic categories. *Robotics and Autonomous Systems* 56(11):1004–1013.

Krause, E.; Zillich, M.; Williams, T.; and Scheutz, M. 2014. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *AAAI Conference on Artificial Intelligence*.

Vogel, A.; Raghunathan, K.; and Jurafsky, D. 2010. Eye spy: Improving vision through dialog. In *AAAI Fall Symposium: Dialog with Robots*.