Improving Robot Scene Understanding using Referring Expression Comprehension

Xiao Liu
Arizona State Univ.
Tempe, AZ, USA
xliu330@asu.edu

Zijie Wang
Arizona State Univ.
Tempe, AZ, USA
zijiewang@asu.edu

Ziming Dong
Arizona State Univ.
Tempe, AZ, USA
zdong27@asu.edu

Alexandra Nazareno
Arizona State Univ.
Tempe, AZ, USA
annazare@asu.edu

Abstract

This paper explores how to improve robot scene understanding using referring expression comprehension (REC). The synthetic dataset creation portion combined 3 well-crafted datasets: Visual 7W, Visual Madlibs and Visual Genome. Two state-of-the-art models, VL-BERT and ViLBERT, were used for performing validation on the proposed synthetic dataset.

1 Project Description

Human-robot interaction (HRI) applications are enabling progressively more sophisticated, capable technologies to reach large consumer populations. Such systems offer great potential in human-centric applications i.e., elder care, household maintenance, and human-robot collaboration. Natural language processing (NLP) with respect to visual referring expressions can provide great benefit to robot scene understanding. Referring expression comprehension is the task of localizing a target object in an image that is described in a natural language expression or description (Qiao et al., 2020). As shown in Fig. 1, REC can be used to improve the quality of vision-language human-robot interaction by enabling a robot to translate a natural language command or description into some operation on an image or video. We propose our problem as follows: given an image and natural language expression describing a relationship between a subject and object within the image, the system should localize the corresponding items within bounding boxes. Our code is released at [1].

2 Synthetic Dataset Creation

In order to experiment with and enhance current state of the art models, it was necessary to obtain

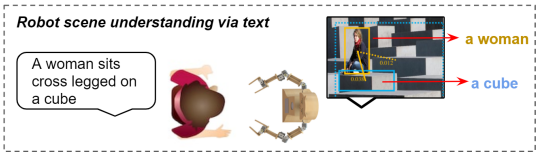


Figure 1: Visual grounding in Human-robot interaction: robot understanding a referring relationship in an image.

"new" data that modeled the task objective, namely a set of images, visual referring expressions describing object relationships within the image, and bounding boxes that encapsulate the described objects. Using a variety of rule-based functions and python packages, 8000 new textual and image samples were synthetically generated from three existing visio-linguistic datasets: Visual 7w, Visual Madlibs, and Visual Genome.

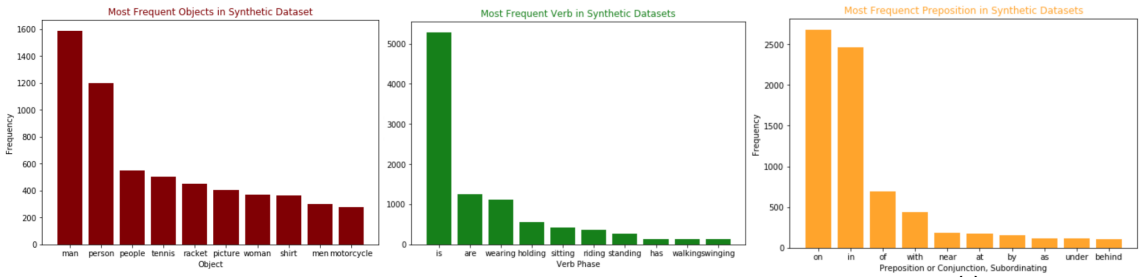
2.1 Visual 7W

The Visual 7W dataset is a visual question answer (VQA) dataset that includes a set of images and question-answer pairs related to them, framed as the seven W questions–*what, where, when, who, why, how*, and *which* (Zhu et al., 2016). The objects being described by the questions and answers are also annotated with bounding boxes. To augment this data to suit the REC task, two scripts were written to combine the "who" and "what" QA pairs into text descriptions of the image scene involving two objects.

2.2 Visual Madlibs

A script was written to access the Visual Madlibs dataset, which consists of annotations created by human-participants in a fill-in-the-blank style task that describe the various attributes, locations, and relationships of persons and objects in a set of images (Yu et al., 2015). The prompts (consisting of a "person", the subject, and an "object")

[1]The code of implementing this project is available at https://github.com/liuxiao1468/Natural-Language-Processing



(a) (b) (c)

Figure 2: (a) Ten most frequent object categories from synthetically-created visual referring expressions; (b) Ten most frequent verbs from synthetically-created visual referring expressions; (c) Ten most frequent prepositions from synthetically-created visual referring expressions.

and human-provided answers were extracted from the pair-relationships annotation set, then concatenated to form a complete referring expression describing a relationship present in the corresponding image. The COCO API was used to retrieve the image metadata and bounding boxes for the subject and object from the MSCOCO image database (https://cocodataset.org/#download). Rules were implemented to further identify and extract attributes for the people and objects, such as gender and appearance, from the annotations and insert them into the expressions. This created final referring expressions that had richer and more precise descriptions for localizing objects.

2.3 Visual Genome

The Visual Genome is a knowledge base part of an ongoing effort to connect structured image concepts to language (Krishna et al., 2016). There are 108077 images, 5.4 million region descriptions, 2.3 million relationships, and 2.8 million attributes in the dataset. Vision and language NLP tasks can implement data augmentation and fine-tuning based on the dataset. To modify this data to suit the desired format, two scripts were written for image and annotation augmentation. The original images were flipped, rotated, and had Additive Gussian Noise added. The augmented annotation was created by replacing verbs with synonymous words. There were three APIs used to complete the augmentation work: imgaug, nlpaug, and Visual Genome API.

3 Synthetic Dataset Analysis

The group's final dataset consisted of 8000 visual referring expressions for 7330 unique images. There were 20418 objects described within those expressions across 2114 categories. The most frequent objects, verbs (relationships), and prepositions are shown in Fig. 2 (a), (b) and (c), respectively. Furthermore, total count, unique, top and frequency of dataset attributes are counted in Tab. 1. The average, shortest, and longest Referring Expression length, as well as the average number of prepositions can be found in Tab. 2. A point of interest to note is the average length of the referring expressions in our proposed dataset is 7.51 words, while in RefCOCO+ the average length is 3.53 (Qiao et al., 2020).

Table 1: Count, Unique, Top, Frequent of dataset attributes.

	Image	Object Label	Subject Label	ReferExpression	Verb	Preposition
Count	8000	8000	8000	8000	12418	7695
Unique	7330	2294	2284	7478	630	67
Top	N/A	man	people	people is present	is	on
Freq	N/A	795	312	17	5291	2682

Table 2: Analysis of Referring Expressions

	Avg Length of Re	Shortest RE	Longest RE	Avg Prep per RE
Number of Words	7.51	3	26	1.91

4 Models

To better understand the state of the art of referring expression comprehension today, we picked the top two performing models, VL-BERT and ViLBERT, from a recent survey by Qiao et al. (2020), and attempted to replicate their respective performances over the RefCOCO+ dataset.

4.1 ViLBERT

ViLBERT is a multi-modal visual and linguistic representation framework that is based on and an extension of the original BERT architecture (Lu et al., 2019). It operates by processing the visual and linguistic features via two separate streams,

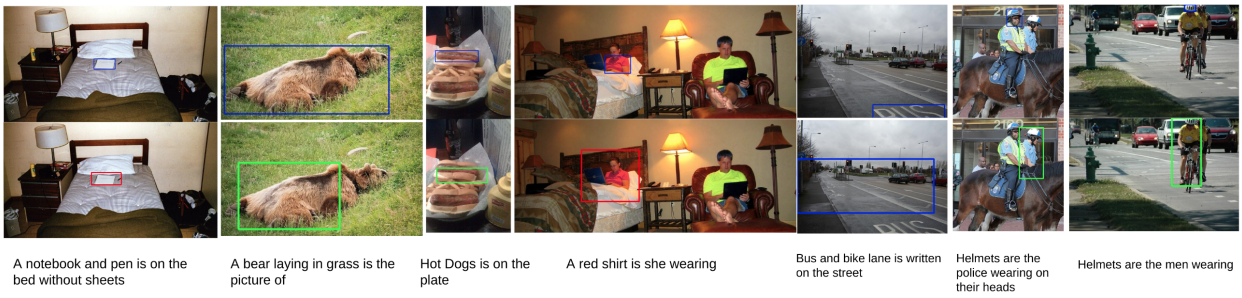


Figure 3: Results from ViLBert model: 1^{st} row shows the ground truth, 2^{nd} row shows the prediction and the referring annotation is the last row.

which interact through co-attentional transformer layers. It was pre-trained on the Conceptual Captions dataset for *masked multi-modal modelling* and *multi-modal alignment prediction* tasks. Results from the team's replicated ViLBERT setup are shown in Tab. 3.

Table 3: Comparison between scores for original ViLBERT model and our group's replicated setup on RefCOCO+ detected region splits.

	val	testA	testB
Lu et al.	72.34	78.52	62.61
Group	72.23	77.51	62.34

4.2 VL-BERT

Similar to ViLBERT, Su et al. (2020) introduced VL-BERT, a multi-modal framework that modifies the original BERT architecture by implementing pieces to encode visual elements, and defines a special element for distinguishing different input formats. It differs from ViLBERT by taking both visual and linguistic features as single-stream input to a stack of bidirectional Transformer encoders, rather than through two separate streams. This allows the related image and text inputs to be processed simultaneously as a combined feature embedding. Having been pre-trained on the visual-linguistic Conceptual Captions dataset, in addition to text-only datasets, it performs especially well with generalization over long, complex expressions, and holds the current top score for Referring Expression Comprehension (Qiao et al., 2020). The VL-BERT system is pre-trained on the *masked language modeling with visual clues* and *masked RoI classification with linguistic clues* tasks.

After setting up the VL-BERT base pre-trained model and fine-tuning it on the RefCOCO+ task, we were able to replicate quite closely the results from the original publication (see Tab. 4).

Table 4: Comparison between scores for original VL-BERT base model and our group's replicated setup on RefCOCO+ dataset splits.

	Detected Regions			GT Bounding Boxes		
	val	testA	testB	val	testA	testB
Su et al.	71.60	77.72	60.99	79.88	82.40	75.01
Group	71.31	77.51	61.03	78.76	81.94	73.98

5 Experiments, Results and Analysis

For experimentation, we evaluated both ViLBERT and VL-BERT's performance on our synthetically-generated dataset. We split our dataset into 80% for training/fine-tuning and 20% for testing. Testing was performed among 20% (1600) images with referring expression annotations. We treat a prediction as positive if t he predicted bounding box has IoU (Intersection over Union) > 0.5 with the ground truth. The fine tuning was performed on ViLBERT model which has 6-Layer co-attention transform modules. For VL-BERT, it is a combination of the BERT and Fast R-CNN model, which works for text and image feature extraction respectively. Some results for ViLBERT and VL-BERT are shown in Tab. 5.

Table 5: ViLBERT and VL-BERT testing results.

Model	Dataset	Test Accuracy
ViLBERT	refcoco+(val)	72.23%
ViLBERT	**proposed dataset**	**36.728%**
VL-BERT	refcoco+(val)	78.76%
VL-BERT	**proposed dataset**	**46.223%**

6 Conclusions

VL-BERT and ViLBERT are both benchmark models for referring expression tasks, but VL-BERT performed better in our case since the system takes visual and linguistic features as single-stream input with performed well with generalization. We

observed that the test accuracy using the proposed dataset remained low because the number of samples we collected is not substantial and the captions we collected were not describing referring relationships with rich adjunct words. However, in the HRI scenario, the locations of predicted bounding boxes are suitable for robot scene understanding.

Acknowledgments

We first would like to thank Dr. Baral for sharing his expert knowledge and experience in Natural Language Processing. We also extend our gratitude to the course teaching assistants, especially Shailaja Sampat, who supported and guided us throughout the steps of this project. Last but not least, we thank the to the School of Computing, Information, and Decision Systems Engineering and the the ASU Fulton Schools of Engineering, for allowing us the opportunity to learn about and take part in this exciting and challenging field.

References

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank image generation and question answering.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images.