Visual Grounding of Spatial Relationships for Failure Detection

Akanksha Saran
Department of Computer Science
University of Texas at Austin
Austin, Texas 78712
Email: asaran@cs.utexas.edu

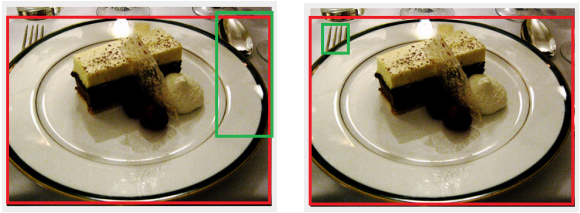
Scott Niekum
Department of Computer Science
University of Texas at Austin
Austin, Texas 78712
Email: sniekum@cs.utexas.edu

***Abstract*—For reliable interaction with everyday objects, personal robots should be capable of visually detecting failures. Understanding spatial relationships between objects (e.g. a screw in a hole) can be useful in determining whether a task was successfully completed or not (e.g. a furniture assembly task). In this work, we show how images, along with natural language descriptions of the spatial relationships between pairs of objects present in those images, can be used in a deep learning framework to ground spatial relationships in the visual space. Thus, failure detection can be performed with natural language, without the extra step of collecting task-specific visual examples of success and failure, allowing robots to verify any spatial concept immediately, just from a description. We use an attention-based deep neural network to ground object names and spatial relationships in natural images, with images and descriptions from the Visual Genome dataset. The network is trained to discriminate correct spatial descriptions pertaining to an image from incorrect ones. We evaluate the performance of our model on a held out test set and obtain an accuracy of 92%.**

I. Introduction

The deployment of personal robots at a large scale will require them to exhibit behaviors that are not only reliable in expectation, but also verifiable: safety-critical tasks must be completed with a high degree of assurance; robots that work with populations that rely on them, such as the disabled or elderly, must be dependable; manufacturing robots that chain together many behaviors, such as in an assembly task, must check their work at each step, or else face multiplicative error rates as the number of steps increase.

A classical approach to detecting task failures perceptually is to gather training data for success and failure conditions and then train a classifier to predict task outcomes in a similar environment at test time. In this work, we attempt to bypass the collection of training images for success and failure detection. Instead, we use natural language descriptions specifying what success and failure should look like in images captured by the robot. This can be especially useful for tasks where training images cannot be collected ahead of time. The natural language descriptions state how any two objects should be spatially located with respect to each other to satisfy a certain subtask completion condition, such as 'screw inside the hole' for a furniture assembly task. Examples of successful task descriptions and corresponding images can be seen in Fig. 1. We can similarly describe conditions which would constitute unsuccessful subtask conditions, such as 'screw outside the hole'. We use such opposing descriptions to learn a model which can distinguish between different prepositions in the visual space. To make a system capable of identifying whether the content of an image matches the spatial relation description for a success condition, we train an attention-based deep neural network using the Visual Genome dataset [8].



(a) Spoon **to the right of** plate (b) Fork **to the left of** plate

Fig. 1. Natural language descriptions annotated for a pair of objects (marked by bounding boxes) in images of the Visual Genome dataset.

Previous approaches have built models of spatial relations by hand-coding their meanings rather than learning these meanings from data [1, 10, 9, 7, 15]. Hand-coding doesn't always work because spatial relationships can appear visually different based on the objects involved. Golland et al. [4] learn a model for spatial relations but assumed perfect visual information consisting of a virtual 3D environment with perfect object segmentation. Also they only refer to objects via object IDs as opposed to the natural language noun reference ('the spoon'), which works for specific references within a small vocabulary. Guadarrama et al. [5] build models of 3D spatial relations learned from crowd-sourcing for robot manipulation responses to human commands. However, their model is trained and tested on clean, uncluttered environments with hand-designed features. They independently ground nouns and 11 common prepositions (referring to spatial relations) in the visual space. In our work, we jointly learn the groundings for objects and 26 spatial relations from a large corpus of 2D images. We work with images in cluttered backgrounds and with occlusions. To the best of our knowledge, we are the first to learn about spatial relations in this manner for failure detection.

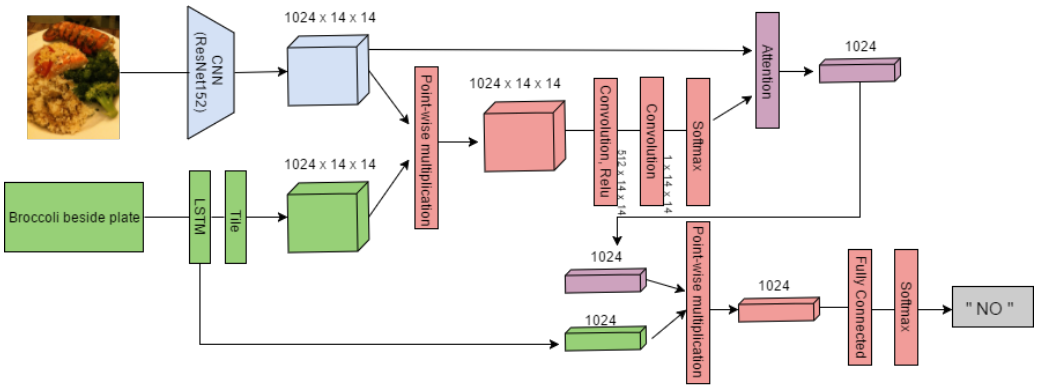


Fig. 2. Architecture of the attention-based network used to learn if a natural language description appropriately describes the relative placement of two objects in an image.

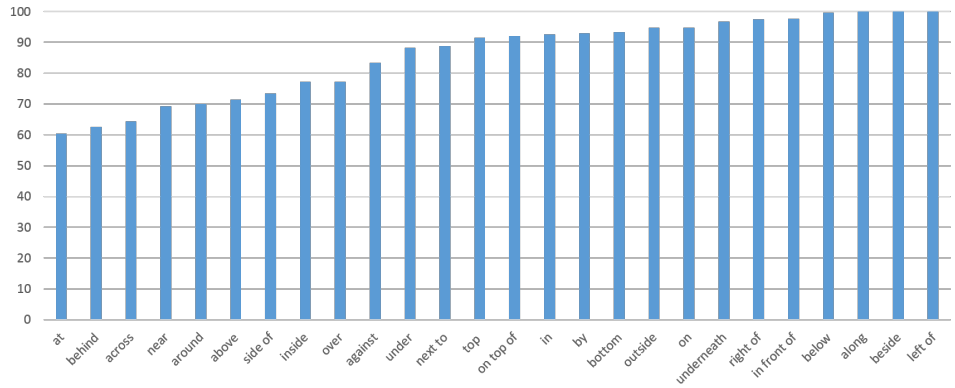


Fig. 3. Percentage accuracy for identifying if descriptions containing specific spatial relations match the respective image content.

II. Approach and Preliminary Results

We train an attention-based deep neural network on a large corpus of natural image data, specifically the Visual Genome Dataset [8]. Attention networks have been used for multi-modal tasks like image captioning [12, 11, 14, 6] and visual question answering (VQA) [3, 2, 13]. Caption generating networks 'attend' to different regions of an image when generating different words of the caption sequentially through a decoder. For VQA, attention networks use the encoded question to 'attend' to different regions of the image to determine the answer.

Inspired by work on attention-based networks in VQA, we use a deep neural network which encodes the natural language description and uses the encoded sentence to learn a soft attention map for the corresponding image. The network eventually performs a two-way classification to determine whether the image content matches the description provided. The architecture is illustrated in detail in Fig. 2. Based on whether the image content matches a success description, the robot can declare success or failure accordingly.

We use the Visual Genome dataset [8] to get annotations for spatial relationships between pairs of objects present in corresponding images. For our experimental results, we use a training set consisting of $66,000$ image-description pairs and a held out test set of $3,200$ image-description pairs. We use a set of 26 common spatial relations from the dataset and generate corresponding negative descriptions with appropriate antonyms. Antonyms for each spatial relationship are chosen manually, as they are not provided with the dataset. The network is trained to identify the descriptions from the original dataset to be in the true class (description matching image content) and the antonym based description to be in the negative class (description not matching the image content). We use the Adam Optimizer with a learning rate of 10^{-4} and a vocabulary size of 4300 words. The network achieves an accuracy of 92% across all prepositions on the held out test set. The accuracy for each preposition is shown in Fig. 3.

III. Discussion

We use an attention based deep learning method to identify appropriateness of a natural language description specifying how two objects spatially relate to one another in the visual space. We show preliminary results for grounding preposition-based spatial relationships between two objects. We will extend this to identifying failures for tasks with natural language descriptions specifying both success and failure conditions. We are working on testing our approach on a new dataset collected by a robot in a home-like environment, to see how well our model generalizes to novel tasks and descriptions. We also plan to work on viewpoint selection for improved discrimination between spatial relations.

ACKNOWLEDGMENTS

This work has taken place in the Personal Autonomous Robotics Lab (PeARL) at The University of Texas at Austin. PeARL research is supported in part by NSF (IIS-1638107, IIS-1617639).

REFERENCES

[1] Wolfram Burgard, Armin B Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. Experiences with an interactive museum tour-guide robot. *Artificial intelligence*, 114(1-2):3–55, 1999.

[2] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[3] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[4] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.

[5] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, et al. Grounding spatial relations for human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[6] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[7] John D Kelleher, Geert-Jan M Kruijff, and Fintan J Costello. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 745–752. Association for Computational Linguistics, 2006.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

[9] Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006.

[10] Marjorie Skubic, Dennis Perzanowski, Samuel Blisard, Alan Schultz, William Adams, Magda Bugajska, and Derek Brock. Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):154–167, 2004.

[11] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[12] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[13] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[14] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.

[15] Hendrik Zender, Geert-Jan M Kruijff, and Ivana Kruijff-Korbayová. Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants. In *IJCAI*, pages 1604–1609, 2009.