Connected Captioning and Generative Networks for Natural Language Based Scene Retrieval

CU Boulder Computer Science Senior Thesis (Spring 2020)

Sousheel Vunnam *, Nisar Ahmed †, Alessandro Roncone *, James Martin *
University of Colorado, Boulder
* Department of Computer Science
† Ann and H.J. Smead Aerospace Engineering Sciences
Boulder, CO, USA
{sovu8155, niah8949, alro6039, james.martin}@colorado.edu

***Abstract*—This paper proposes a system for retrieving scenes from a robot's memory using natural language inquiries through the connected use of captioning and generative text to image systems. People may have differing models of an environment, with varying levels of knowledge, syntax usage, and word choice. Also, in a physical setting, disadvantaged viewpoints and images have limited the application of existing deep learning classification methods. Visual recognition can be grounded to an environment through natural language interactions between a human operator and a robot. Captioning systems can provide detailed descriptions of scenes and AttnGAN can synthesize images from human inquiry, allowing for informed semantic and visual comparisons. Overall, the system does not effectively function in the context that it aims to improve, but gives insight into human robot collaboration problems. The code for this project is available here: https://github.com/sousheel/Scene-Retrieval**

I. Introduction

For physical environments in which a human-assisted robot will act in, it is difficult for the operator and the robot to communicate with each other naturally. There may be different levels of understanding about an environment and a lack of knowledge from the operator about how the robot's language model operates. In previous human-in-the-loop experiments conducted in the COHRINT Lab, an operator conveys information to a robot using a pre-defined word bank [2]. However, it is easier for an operator to communicate with natural language. These natural language statements can then be grounded to the physical environment through visual recognition.

In recent years, research into object recognition has been progressing rapidly due to a combination of improved computational power and curation of massive datasets such as ImageNet. Convolutional neural networks have made category-level object recognition quick and highly accurate. Furthermore, large scale language models have also been advancing for similar reasons. However, in a setting involving robotic navigation through an environment, disadvantaged viewpoints and images have limited the application of these methods. It is also difficult for non-skilled operators to interact with these systems as they may not know which labels or syntax to use.

Visual recognition can be used to ground natural language queries to improve interactions between a human operator and a robot acting in an environment. In reality, one can refer to an object using many names. For example, a "soda" can be referred to as "bottle/can", "coke", or "pop". Furthermore, one may also use different types of phrases, interactions, and references in their language.

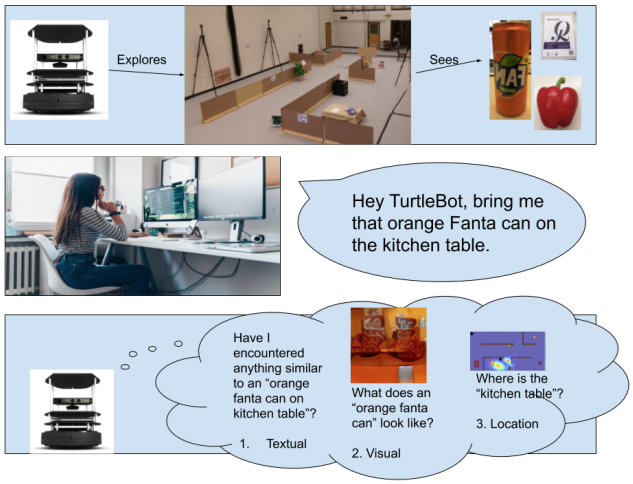

Fig. 1. A robot explores an environment and captions the area. An operator can then request objects which leads to a semantic and visual comparison.

With this research, I aim to ground visual recognition to a specific environment using captioning systems in conjuction with generative text-to-image techniques. It can be seen as a more accessible expansion of object detection in which a user can command an in-home robot to retrieve objects. For instance, "Grab me my mug from the kitchen table".

This research will aim to investigate the use of text to image algorithms to create synthetic images of its environment, allowing for visual predictions of human statements. I hypothesize that it will allow for synthesis of attributes with objects in a scene to create more accurate representations of a human's inquiries, leading to accurate scene retrieval.

II. Related Work

The open vocabulary research problem being investigated in this paper is motivated by the work done in the COHRINT Lab to harness the power of human perception in autonomous systems [2]. Much of the previous work in the lab has been focused on creating internal models of human observations, though the use of perception algorithms to aid these observations has been limited. This research topic has arisen from the symbol grounding problem, which asks how a system can attain meaning from a set of symbols [17]. Due to the vagueness of "meaning", most of the research done in this area has been problem-focused and involves the exploitation of environmental info to attain a useful result. Thus, an interdisciplinary set of techniques is used, especially those from robotics, computer vision, and natural language processing.

A. Visual Language Grounding

In this paper, we will focus on symbol grounding techniques that take advantage of the visual capabilities of a robot acting in a physical environment. One technique involves creating complex relation models of objects in an environment called factor graphs [5]. Much of the research also focuses on grounding a specific type of feature of objects, such as spatial relations [3] and attributes such as color and shape [10]. as well as joint visual-language models that combine language statements and features from a scene [3]. Unfortunately, these approaches are difficult to implement and do not scale well with new advances in individual fields, such as feature extraction and classification or improved language models.

B. Natural Language Processing Techniques

Many systems that aim to encourage human-robot cooperation using natural language rely on literal linguistic features such as part of speech tags, word dependencies, and syntax structures. Other systems use interpreted models to extract object categories, temporal and spatial relations, and task descriptions from human observations [7]. However, these systems are usually hand built to maximize their effectiveness for single use cases. Language models that use deep learning usually tend to be more general and can be applied to multiple use cases. They are also being continuously improved with more data and compute power [21]. In particular, captioning methods are advancing quickly [19]. Captioning methods have previously been applied to robotics through translation of videos into actions [12]. I aim to use captioning to develop a solution to physical symbol grounding by creating interpretable descriptions from a video feed.

C. Generative Adversarial Networks

Text to image generation is a difficult problem being tackled by both natural language and computer vision research communities. Most of the time, it is regarded as the inverse of the captioning problem in which one aims to generate an accurate image given a text description. Generative Adversarial Networks (GANs), in particular, have allowed for an advancement compared to other text to image techniques [9]. GANs have previously been used in many applications such as 3D object construction [18] and image restoration [13]. The first paper to address text to image generation using GANs is Reed [16]. Text captions are converted into images by taking advantage of the usefulness of recurrent neural networks for text embeddings and GANs for synthesizing images. The key difference is the addition of the text embedding as a conditional input to the GAN along with the randomly sampled noise. There is also a key improvement to the loss function which forces the text embeddings to not be factored into the loss until the discriminator cannot determine between real and fake images. Further improvements on this GAN include Stack-GAN which generates images in two stages using two separately trained GANs [23] and AttnGAN. AttnGAN uses an attention mechanism for creating more highly detailed subregions [20], a technique taken directly from captioning [19]. The entire sentence is conditioned with the sampled noise to create a base image, then each word is conditioned for smaller subregions of the image to fine tune details. In this paper, AttnGAN is used as it is the most widely available text to image GAN at the moment. There is not much use of these text to image techniques in real world applications, especially in robotics, so it is useful to investigate the effectiveness of these.

III. Approach

In this section I will formalize the problem being investigated as well as describing the system built to do so. In this experiment, given a set of images and corresponding locations, the goal of the system is to choose the image that is closest to an inquiry. This system involves 3 main phases: an exploration phase, a semantic evaluation, and a visual evaluation.

A. Exploration Phase

The exploration phase signifies the robot's development of knowledge about its environment. In a physical area, the robot wanders around, intermittently capturing and tagging images from a camera feed. This will build up a set of scenes with location tags and corresponding captions. In this experiment, the location data is used simply for navigation.

After a picture is taken, it is fed into an attention based captioning system [19] implemented in PyTorch and trained on the ILSVRC-2012 dataset. This system was used due to a wide availability of resources for assistance in implementation as well as speedy captioning on a less powerful system. The speed of the captioning system did not pose a problem here, but in slower systems that may be put into use in the physical world, it may be necessary to adjust the image capture rate to match the speed of the captioning system. The output of this system gives a one sentence caption that is bundled together with the image and stored.

It was out of the scope to develop optimal exploration tactics and a physical environment for testing, so instead we considered doing a simulation. However, this was not feasible either, due to the lack of expected objects and clutter in many indoor simulation environments. Instead, the Autonomous

Robot Indoor Dataset (ARID) was used [8]. This dataset contains RGB images captured by a robot patrolling a human populated work environment. Each image contains multiple objects and the environment is captured with images from multiple views. Due to the use of a dataset, location data was not implemented. The exploration phase runs on the robot at all times, gaining information about the environment.

B. Semantic Comparison

Once an inquiry is made, then the semantic comparison phase begins. A human inquiry is signified by the request for the system to retrieve a scene from the set acquired by the exploration phase. The inquiry will be a natural language command such as ”can you get me the red mug on the kitchen table?”. In a physical system, this would be implemented by a speech to text program or a chatbox interface, but in this experiment I will ask subjects to give observations about a specific scene from the exploration phase.

The raw text of the inquiry will be input into Word2Vec, a context based similarity metric. A context-based model assumes words that are used in a similar context have similar meanings. Therefore, even if the generated caption and the human inquiry have different wordings, they can still be similar due to the similarity of their underlying context. Furthermore, attributes of an object usually have similar contexts, as it is common to see a description of an object near the naming of one. Word2Vec accomplishes this similarity metric through skip gram negative sampling. A classifier is trained with positive examples of words near the word in a text and negative examples of words farther in distance. The weights of the system then correspond to a multidimensional vector that signifies a space in which similar words lie near it. Since Word2Vec only operates with a word to word comparison and we would like a sentence to sentence comparison, a dot product is taken over the vectors of words. This is a bag of words comparison which may lose info about word order and location.

The similarity metric from the spaCy Python library is used, trained on the Glove [14] word2Vec model. This gives a cosine similarity score between 0 and 1 and it is important to distinguish that this does not signify a probability. The raw text from the inquiry is then compared with all of the captions taken from the exploration phase, leading to a set of closest images. This smaller set of images is necessary, as the visual comparison takes much longer than the semantic comparison. After a set of closest images is taken, these images are sent to the visual comparison stage.

C. Visual Comparison

The raw text from the human inquiry is input into AttnGAN, a generative text to image network. This allows for a synthesis of different attributes and relations of objects to be represented visually. The process of creating an image is fairly slow when run on a CPU, but does not face problems on a computer with a GPU. This produces a representation of the human inquiry in image form, which is then compared to the small set of closest images from the semantic comparison.

Visual similarity between images is a highly studied topic in computer vision, however the use in this context is difficult and largely unsolved. The problem of taking a representation of an image such as a ”notebook” and matching it to a specific image of a green spiral notebook contains many difficulties. I began with using established computer vision techniques such as histogram similarity and sift features. Histogram techniques did not work well with objects of different colors and sizes, so it was largely useless in this context. The implementation of SIFT features slightly improved the system, but still did not work well due to the lack of detail in the generated images. In the end, I used an API from Clarifai to compare images [22]. This forced a lack of transparency within the comparison, so it is unknown how these comparisons work. However, this method was much more accurate than the previous methods.

The most similar image between the generated image and the set of images is chosen as the best image. In a physical environment, the robot will travel to that scene and retrieve the object requested. The actual manipulation and retrieval of an object given a scene with multiple objects is outside the scope of this project.

D. Architecture

The architecture of the scene retrieval system is implemented using Python with the ROS framework. The exploration phase can be concurrently run with the comparison phases. As soon as an inquiry is made, the semantic comparison stage begins. The inquiry is also sent to AttnGAN in order to save time. Once the semantic comparison is done and AttnGAN has generated the image, then the visual comparison stage begins. See Figure 2 for a diagram.

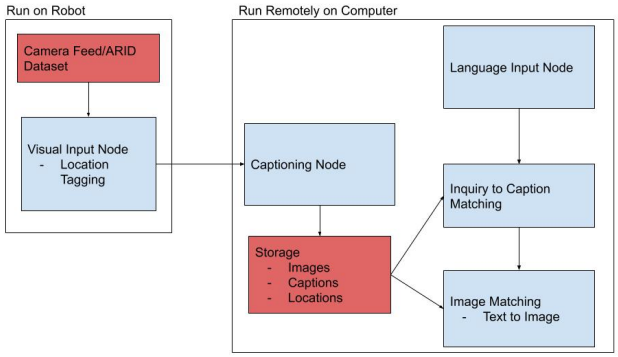

Fig. 2. System Architecture Diagram. The blue squares indicate ROS nodes while the red squares indicate other functions.

IV. Evaluation

Unfortunately, due to the nature of the COVID-19 outbreak, the evaluation of the system has faced difficulties. Due to time constraints, it was not possible to transfer the experiment to an online system such as Amazon Mechanical Turk. Thus, there is a lack of diversity in the test subjects demographics which may be vital to display the effectiveness of the system with a wide range of human operators. Instead, I have devised five personas with differing technical abilities and personal lexicons. This is aimed to test differences in word choice, syntax, familiarity with technical jargon, and varying levels of previous knowledge about an environment. Each person has made an inquiry on a selected set of 100 images. These images are from a pool of images from ARID. To evaluate the system quantitatively, the precision of the outputs will be measured. In this experiment, the output will be considered correct if the correct image is in the top 5 most similar. The semantic and visual comparison stages will be individually tested to see how much they contribute to the systems output. Then the entire system with both stages will be tested. See Fig. 4 for results. Also, see the example sentences used in column 2 which describe the image in Fig. 3.

It seems that as more words are added to the observation, it makes the semantic comparison worse. This can probably be attributed to the bag of words model that this implementation of word2vec uses. The sentences are treated as a unordered collection of tokens, so word dependencies are lost. This may be improved with Doc2Vec [6]. Word2Vec also seems to fail in physical scenarios due to lost information. Words used to convey spatial information have similar contexts, so ”next to the kitchen table” and ”under the kitchen table” are extremely similar. This could be improved with parsing of human requests to single out words with spatial information.

The visual comparison on its own did not seem to work at all. However, with the addition of the semantic comparison, it gives a slight advantage. This is promising, as the order of the two comparison stages works well. It also indicates that as text to image GANs improve in the future, this system will also improve.

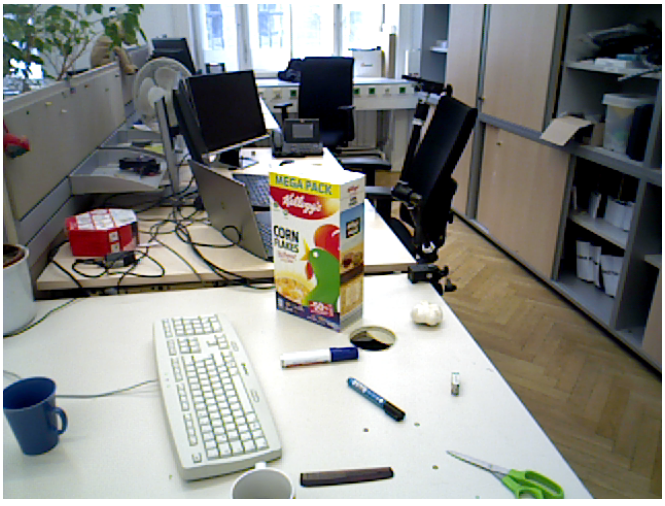

Fig. 3. The image used for example sentences in Fig 4 with generated caption: ”a desk with food and a laptop and a mouse on it.”

V. Conclusions

In this paper, a system for retrieving scenes from a robot’s memory using natural language inquiries is proposed. It aims to take care of attributes of objects, spatial relations, and cluttered scenes through the use of AttnGAN to synthesize images from a human inquiry. The usage of captioning gives a detailed description of a scene, providing more information and speeding the comparison process up. Overall, the system does not work effectively in the context that it aims to improve. This is largely due to the lack of accuracy of the AttnGAN in producing a usable inquiry and the difficulty of visual comparison. However, the approach combines multiple deep learning techniques that will be improved in the future and can lead to more satisfying results. The system also provides human interpretable results that allow for an operator without domain knowledge to gain insight into a robot’s interpretation of it’s environment and human inquiries. These are steps toward more informed human-robot collaboration.

A. Future Improvements and Work

The framework of this system seems to be promising for physical use. However, many improvements can be made. Increased processing throughout the system can lead to advancements in the precision.For example, the images taken during the exploration phase seem to be too general. They contain a multitude of objects which make it difficult for the captioning system to generate info on the entire scene. Increased processing of the images using techniques such as segmentation could be effective. Processing of the human inquiry and generated captions could also be useful to identify key attributes and relations. There are also systems being developed to expand inquiries to predict intent, emotions, and possible actions [1].

Furthermore, the semantic and visual comparisons can be improved, as described in the evaluation section. It is difficult to measure the accuracy of text to image systems quantitatively, but techniques which feed the image back into a classification system are being investigated [4].

Location data was not used in this experiment because of the use of a dataset rather than a physical camera feed. However, this additional info can be used to differentiate rooms and provide navigational details. For example, if one says there is an object in the ”kitchen”, a semantic analysis of the captions can be used to classify which area is most likely to be the kitchen based on the objects seen in it.

Improvements to the captioning systems and text to image GANs are being investigated thoroughly and there are already better systems that have been published. More descriptive captions can be used, giving a wider array of info about a scene, even up to a paragraph [11]. MirrorGAN is able synthesize much more accurate representations of text through

Person	Example Sentence	Precision-5 (Semantic)	Precision-5 (Visual)	Precision-5 (Combined)
Sousheel	"Get me the corn flakes box on my desk."	.41	.11	.43
Office Employee	"Can you please retrieve the box of cereal on my office desk?"	.43	.11	.44
Military Operator	"I am requesting the box of cereal on top of the desk , over."	.37	.08	.39
Grandma	"Honey, can you be a dear and grab me the cereal box over there?"	.35	.07	.36
Yoda	"Give me the cereal box, you will."	.39	.13	.42

Fig. 4. A robot explores an environment and captions the area. An operator can then request objects which leads to a semantic and visual comparison.

continuously readjusting the image after feeding it into a caption generator [15]. However, much of the improvements of these systems will likely be made in time with the curation of larger datasets for the training of captioning systems and GANs.

This system can be applied to research other areas of robotics as well. For example, one could investigate people's trust in deep learning systems being applied in physical space by asking how people view the generated images and captions. It can also lead to insights on how to make a robot's decision processes more interpretable through generative tools which provide interpretable references to a robot's world model.

REFERENCES

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction, 2019.

[2] Luke Burks, Ian Loefgren, Luke Barbier, Jeremy Muesing, Jamison McGinley, Sousheel Vunnam, and Nisar Ahmed. Closed-loop bayesian semantic data fusion for collaborative human-autonomy target search. *CoRR*, abs/1806.00727, 2018.

[3] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1640–1647, 2013.

[4] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. 10 2019.

[5] Thomas Kollar, Stefanie Tellex, Matthew Walter, Albert Huang, Abraham Bachrach, Sachi Hemachandra, Emma Brunskill, Ashis Banerjee, Deb Roy, Seth Teller, and Nicholas Roy. Generalized grounding graphs: A probabilistic framework for understanding grounded commands. 11 2017.

[6] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[7] Rui Liu and Xiaoli Zhang. A review of methodologies for natural-language-facilitated humanrobot cooperation. *International Journal of Advanced Robotic Systems*, 16, 2017.

[8] Mohammad Reza Loghmani, Barbara Caputo, and Markus Vincze. Recognizing objects in-the-wild: Where do we stand? In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[9] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention, 2015.

[10] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning, 2012.

[11] Luke Melas-Kyriazi, Alexander Rush, and George Han. Training for diversity in image paragraph captioning. *EMNLP*, 2018.

[12] Anh Nguyen, Thanh-Toan Do, Ian Reid, Darwin G. Caldwell, and Nikos G. Tsagarakis. V2cnet: A deep learning framework to translate videos to commands for robotic manipulation, 2019.

[13] J. Pan, J. Dong, Y. Liu, J. Zhang, J. Ren, J. Tang, Y. W. Tai, and M. Yang. Physics-based generative adversarial models for image restoration and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[15] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. *CoRR*, abs/1903.05854, 2019.

[16] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *CoRR*, abs/1605.05396, 2016.

[17] Mariarosaria Taddeo and Luciano Floridi. Solving the symbol grounding problem: a critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):419–445, 2005.

[18] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

[19] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[20] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.

[21] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[22] Matt Zeiler. clarifai 2.6.2 documentation.

[23] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016.