Auto-drawer: Generating and Modifying Images Continually by Visual-Relational Knowledge Graph

Wan-Cyuan Fan
National Taiwan University
christine5200312@gmail.com

Abstract

Dialogue-to-image generation is an extended research topic from the text-to-image generation task. In this task, we need a system that generates image iteratively, conditioned on ongoing linguistic input sentences. Such a system must not only know how to draw the image conditioned on given captions but also have to understand and organize the interactions among concepts present in the feedback history. This makes the task significantly difficult than one step Text-to-image generation. Prior work has used text captions in the dialogue and corresponding images to train an end-to-end model. However, there are relatively few datasets with the dialogue-images paired relationship, which makes such a model not easy to be widely used. In this paper, we address the challenge by combining natural language process algorithms and computer vision models. By using the visual-relational knowledge graph as the bridge, we show that this method gets rid of the limitation of a paired dataset, and can effectively generate images based on linguistic input instructions.

1. Introduction

Automatic generation of realistic images from the dialogue between humans and machines has lots of applications, for example in education, in image editing, in animations, or creative arts. However, this topic is a more advanced research field than text-to-image, so there are still many challenges. In 2018, Sharma et al. [12] proposed a non-iterative model called ChatPainter which generates images by adding dialogue data. The author shows that adding a dialogue (obtained from the Visual Dialog (VisDial) [3] dataset.) that further describes the scene leads to a great improvement in the quality of generated images on the MS COCO [8] dataset. Following by El-Nouby *et al*. [4] in 2019, they present a recurrent image generation model that can take both the generated output up to the current step and all past instructions for generation into account. They perform experiments on the Collaborative Drawing [6] (CoDraw) dataset which contains intermediate incremental images for each turn of the dialogue and get a quality result. However, due to the requirement for a dataset with paired data (paired captions and images), it is hard to widely apply the model on different artist scenario.

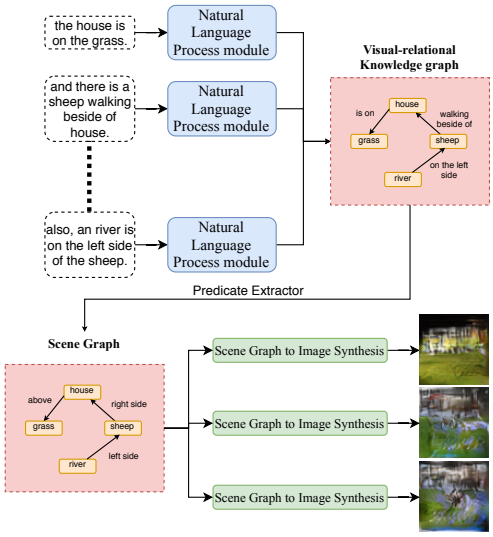

Figure 1. Overview of the Auto-drawer framework.

To avoid such a problem, we split the entire dialogue-to-image task into two parts: (1) semantic understanding and organization of the input dialogue (2) generate corresponding incremental images from the organized dialogue information. Furthermore, the key to the intermediate connection is the Visual-Relational Knowledge Graph which can record the semantic information of the dialogue.

In the first part, we focus on converting input instructions into a visual-relational knowledge graph. J Lafferty *et al*. [7] proposes the conditional random field (CRF) which uses POS tagging to combine observed variables with target variables. Furthermore, by maximizing the likelihood function, the best tagging model is obtained. With the CRF

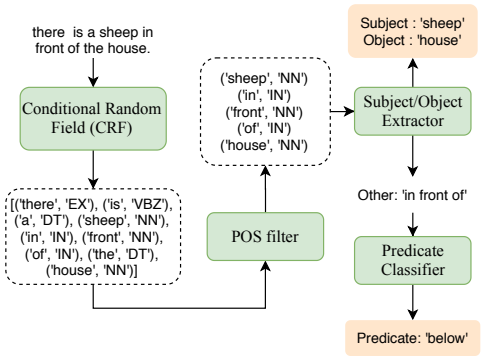

Figure 2. A single input sentence example of the Natural Language Processing Module.

model, we can iteratively input sentences and do the most basic cutting and understanding. In the conversion of predicates and vector space, T Mikolov *et al*.[10] proposes two novel model architectures for computing continuous vector representations of words from very large data sets. With these models, we can get word embedding with semantic meaning, and classify by some classification algorithms.

In the second part, with the help of the knowledge graph, this part becomes simple image synthesis from scene graphs. J Johnson *et al*. [5] presents an end-to-end model that turns an input scene graph into scene layout by using Graph convolutional network (GCN). Then, based on the layout, they generate realistic output images by training the image generation network adversarially against a pair of discriminator networks. Such results have inspired many studies in the future.

In this paper,

- We propose a novel structure to achieve dialogue-to-image tasks. This model does not need the dialogue-images paired data and can widely apply to many datasets.
- We provide an idea using the knowledge graph to connect the field of NLP and image synthesis. This makes it easy to apply to other human interaction algorithms.

2. Our Approach

Our method can be divided into three parts. The first part is Instructions understanding which transforms the input instructions into into a visual-relational knowledge graph. The second part is the scene graph converter which takes the knowledge graph as the input and converts it into a scene graph with predefined relationships based on word embedding classification. The final part generates images from a scene graph by sg2im [5] pre-trained model. An overview of the Auto-drawer framework is shown in Fig. 1.

2.1. Natural Language Processing Module

In the first step, the Auto-drawer takes a series of instructions from designer and iteratively construct a visual-relational knowledge graph. As illustrated in Fig. 2, take a single instruction as an example, this instruction will be sent into conditional random field (CRF) model trained on the treebank dataset provided by NLTK Corpora. The output of the CRF model is a combination of many part-of-speech (POS) tagging. These POS taggings go through a filter to extract important parts of speech, *e.g*. subject and object are nouns. Then, we are using the $Extractor_{SO}$ and $Classifier_{Pred}$ to get the subject ,object and spatial relationship.

Subject and Object Extractor ($Extractor_{SO}$). We find the word whose part of speech is a noun and compare it with the classes in the dataset, and then extract the subject and object.

Predicate Classifier ($Classifier_{Pred}$). Firstly we predefine some spatial relationships such as above, left side and under, etc. Then, we perform query expansion on these predefined relationships on Wordnet. After that, we take out all the spatial relationship words with the same semantic meaning and convert these words into word embeddings using the word2vec [10]. Finally, we apply PCA on these embeddings and use the support vector machine (SVM) to classify these embeddings. The key purpose of this processing is that any relationship word embedding can be classified and get the predefined spatial relationships we want in the inference time.

2.2. Image Generation Module

Once we acquire the scene graph, we feed the scene graph to the generative model (Sg2Im) provided by J Johnson *et al*. [5] and obtain the output images. In this model, they generate the image in three steps: *Graph Convolution Network*, *Layout Prediction Network* and *Cascaded Refinement Network* [2] (CRN). Firstly, they apply the graph convolution network on scene graphs. Then, to move from the graph domain to the image domain, the layout prediction network compute a scene layout by predicting a segmentation mask and bounding box for each object. Finally, the model synthesizes an image based on the given layout by using the Cascaded Refinement Network [2].

3. Experiments

We train the Conditional Random Field and predicate classifier on NLTK treebank dataset [9] and WordNet [11], respectively. Also, we train the Sg2Im model to generate 64 x 64 images on the and COCO-Stuff [1] datasets. In our experiments, we aim to show that our method construct scene graph from iteratively input instructions and generate images based on the scene graph.

Figure 3. Images generated by our method trained on COCO. In each row we start from a simple instruction on the left and progressively add more instructions moving to the right

3.1. Datasets

COCO-Stuff. COCO-Stuff dataset [1], which augments a subset of the COCO dataset [8] with additional stuff categories contains 40K train and 5K val images with bounding boxes and segmentation masks for 80 thing categories (cat, cars, etc.) and 91 stuff categories (sky, grass, etc.). Note that, COCO dataset does not contain intermediate incremental images for each turn of the dialogue. Same as Sg2Im [5] model, we use these bounding boxes annotations to construct synthetic scene graphs based on the coordinates of the boxes of the objects, using six predefined geometric spatial relationships: above, below, left of, right of, inside, and surrounding.

NLTK. NLTK [9] has built-in support for dozens of corpora and trained models. We select Treebank Part of Speech Tagger corpora as the training dataset for condition random field. Treebank Part of Speech Tagger corpora contains 10,156,853 linguistic data with part of speech tagging ('sheep': 'NN', 'of': 'IN' and 'there': 'EX', etc.).

WordNet. WordNet [11] is a large lexical dataset of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet contains 117,000 synsets linked to other synsets by means of a small number of "conceptual relations". This allows us to perform query expansion and train our predicate classifier.

3.2. Results

We now show our experimental results to illustrate the effectiveness of Auto-drawer.

Natural Language Processing Module. We collect Treebank Part of Speech Tagger corpora in NLTK tools as the training data and train our conditional random field model. Then, in order to train our predicate classifier, we use Word2Vec [10] method to convert all the synsets obtained from query expansion on WordNet. After preparing all the word vectors of synsets of predefined spatial relationship, we use support vector machine (SVM) as our classifier and train the classifier on these word vectors. In this way, we can convert input instructions into a scene graph. The results are given in Figure 4.

Image Generation from Scene Graph. After converting the input instructions into the scene graph. We sent the scene graph to the Sg2Im pre-trained on COCO dataset and get the final images.The results are given in Figure 3.

4. Conclusions

In this paper, we design a simple but effective dialogue-to-image generation method. We tackle the issue of datasets limitation of requiring intermediate incremental images for each turn of the dialogue as previous work. By using a visual-relational knowledge graph, we can easily combine the NLP algorithm and computer vision model. This idea can be applied to many tasks such as text-to-image generation and dialogue-to-image generation, etc. By combining the strengths of the NLP and Computer vision model, such tasks can be solved more simply. Furthermore, visual-relational knowledge graphs are not limited to spatial relationships and other visual information such as object attributes and user-defined positions can be added into the graph. This will leave for our future work.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 2, 3

[2] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 2

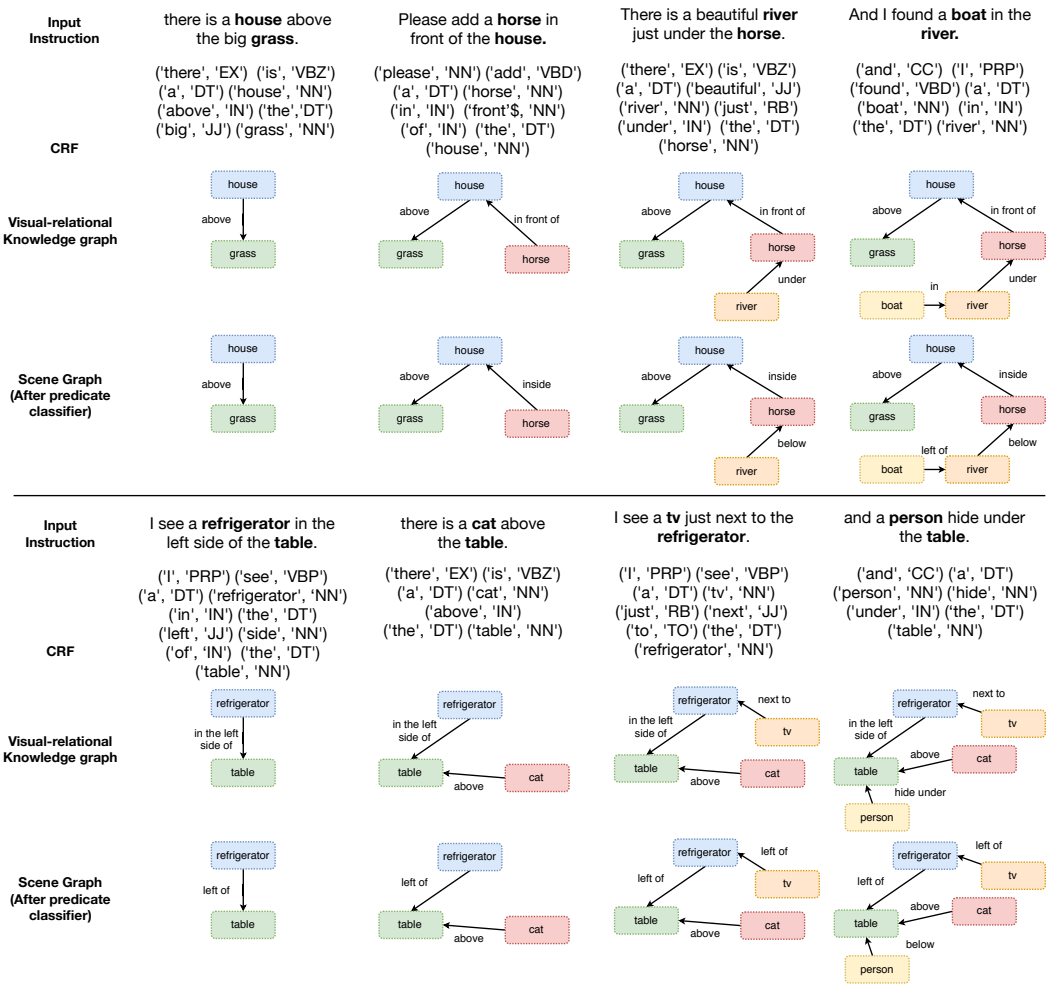

Figure 4. Scene graphs constructed by our natural language processing module.

[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 1

[4] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10304–10312, 2019. 1

[5] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2, 3

[6] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *arXiv preprint arXiv:1712.05558*, 2017. 1

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3

[9] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002. 2, 3

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 3

[11] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998. 2, 3

[12] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216*, 2018. 1