

# Project Plan of Generating Comics from Natural Language Description

Bowen Yang, Yuning Wang

September 2019

## 1 Backgrounds

As creation needs a thorough understanding of the subject being created and also the techniques such as designing and drawing [2], it's difficult for people who have a great idea but no previous designing experience to change his/her idea to real artwork. It is a common problem among online writers that they have finished their works but cannot create their own cover images due to unfamiliarity of design or drawing. But for computers, they have both the ability to understand natural language and also the deep knowledge in many fields, generating images based on the natural language description, which is called creation, will be a feasible task for them and help people who have the passion but no ability to create.

There has been an emergent trend in computer vision combining machine learning related to this possible solution, which is constructing scenes from sentence descriptions based on provided references (known as Text-to-Image synthesis) [5]. It requires a natural language process as well as an image process. Additionally, a bridge that synthesis both vision and language modalities are indispensable to the performance. Some recent works have utilized methods concerning Generative Adversarial Networks(GANs) present some admirable upshots

However, current works regarding GANs analyze sentences on a general base, problems might take place when a word-level detail is determined to a picture requirement[5], as a result, it may probably fail to generate a picture fit the description well. Additionally, recent studies majorly focused on a simple object such as birds which do not require the objects' interaction with others or notice to objects' action. As a consequence, complex scenes generation in current works have not been well developed such as in the COCO dataset [6]

Besides, there is still room for further improvement in the image layout control. Most models based on Generative Adversarial Networks (GANs) perform well in one-object-in-the-center problem or single-domain images problem, while can not achieve expected results when the image to be generated contains multiple objects, which have complicated relationships with each other and different locations in the image [7].

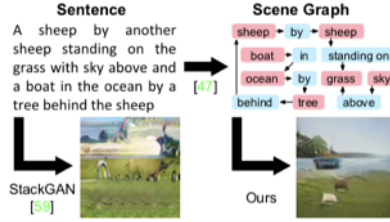


Figure 1. State-of-the-art methods for generating images from sentences, such as StackGAN [59], struggle to faithfully depict complex sentences with many objects. We overcome this limitation by generating images from *scene graphs*, allowing our method to reason explicitly about objects and their relationships.

Figure 1: Scene Graph illustration

## 2 Objectives

In this research, we are devoted to contributing a combined scenery generating model by interpreting the textual description. We majorly focus on comical figures and create a series of interactions and relations with each other. The goals that this research may achieve are modifying poses and facial expressions of figures, identifiable composition among multiple figures and connection to the background environment. We make alterations to the model based on provided dataset COCO, and a self-constructed dataset referring to numerous comic-related materials. On account of our aims to improve the illustration environment who do not gain access to drawing skills as well as the limited time, the reality images generating process will not be considered in the current stage of our research.

Several foreseen difficulties exist from the start of the research. First of all, since the model is worked mainly for comical characters in this stage, a dataset containing comical scenes in COCO format is required to be combined into MS-COCO dataset [6]. This step involves crawling and annotating at least 1k suitable images for training. Secondly, a rich textual semantics sometimes is not easy to be understood by a computer. For example, a computer may get confused if it encounters a description like: “The man is tired” which expects the picture to generate an image of a man with a facial expression of exhaustion

across his face. In this case, the details behind the text may not be obvious and therefore cannot be well presented in the words themselves to the model. Last but not least, the interaction between characters and objects is one of the main concerns in this research. For instance, “the man is embracing a woman” indicates that a man and a woman are embracing, the man’s hands should be on the woman’s waist, and their bodies should have interactions instead of set apart or simply overlapping.

### 3 Methodologies

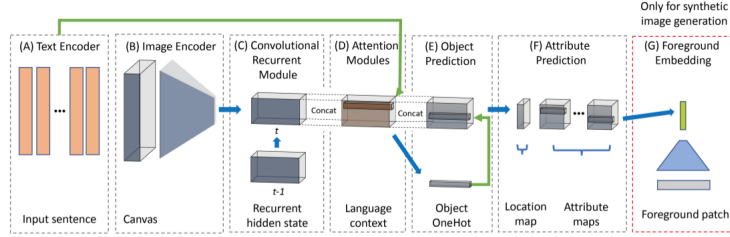


Figure 2: Overview of Text2Scene

Generally, our model obtains a series of structures of textual and vision encoders, a Convolutional Recurrent Module and optimal attention modules, Object Prediction and attribute prediction. Then an object combination aims to synthetic image generation. (Figure 2)

The first step is to build a specific dataset based on COCO data set format. Labelme, an application available on all major OS and coco-annotator, a website application are required to annotate the polygons and convert the annotation files to COCO dataset JSON file.

Some proposed general methods could be applied to achieve the objectives and solve the problems listed

#### - Text2Scene Model [1]

Text2Scene model is not quite similar to what GANs do but their attempts are similar – predicts layouts for picture synthesis. However, a more accurate level which is a pixel-based level output achieved by Text2Scene model without GANs assistance a better result. The advantage of this model is that it is suitable for divergent scenes, such as abstract, reality or composite, merging into the same framework.

This model opts for a Seq to Seq framework [4] who lays focuses on spatial

and sequential designing. This procedure mainly contains several separated models: a text encoder that takes input as a sequence of  $M$  words  $w_i$ , object and attribute decoders foretelling the order of objects  $o_t$  whose locations  $l_t$  and attributes  $R_t^k$  will be identified. To start with an initially empty canvas, this process generates objects within each step allows almost one-to-one mapping with predicted objects.

#### - Scene Graph to Image [2][3]

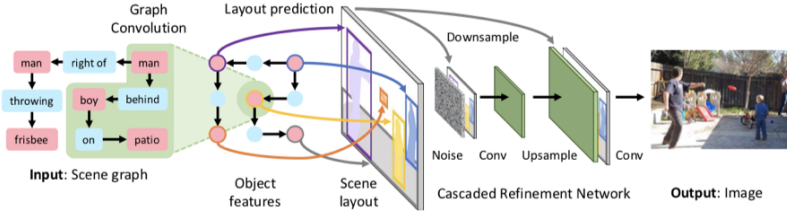


Figure 3: Procedure of Scene Graph

As shown in graph (figure 3), the module takes a scene graph, describing objects and their relationships as input, then process the input using graph convolution network, where the input vectors are passed to functions to generate embedding vectors for each object to predict boxes layout and generate segmentation masks for objects. The model then uses CRN to produce the image. CRN is a cascade of purification modules, where the input is a semantic layout and the output is a color image with the graphic appearance corresponding to the given layout. Each refinement module runs on a specified resolution and contains three feature layers: input, intermediate and output layers. The input to the first module is the semantic layout, while a linear projection needs to be applied to the output of the last module to map to the expected color image. Between successive modules, a resolution is doubled.

## 4 Deliverables

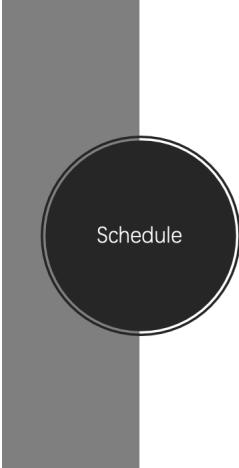
Our major deliverables are

- **The framework we establish to produce compositional picture based on the textual description.**
- **Proofing the above framework is capable to generate various results based on a dataset from COCO, comical materials, as well as proposed semantic layouts.**
- **The outcome and performance to evaluate the above framework,**

involving a comparison among original texts and computer generating semantics analyzed from the generated pictures.

- A website allows a user to type in the textual description and provide computer improvised picture.

## 5 Schedules



Time	Work Accomplished
Sep – Nov 2019 Phase 1	<ul style="list-style-type: none"><li>- Make discussion with teammate and supervisor to set a detailed project plan</li><li>- Research on the topic and implement the model</li><li>- Project implementation</li></ul>
Jan.13-17 2020	First Presentation
Feb 2020 Phase 2	<ul style="list-style-type: none"><li>- Preliminary implementation</li><li>- Detailed interim report</li></ul>
April 2020 Phase 3	<ul style="list-style-type: none"><li>- Runnable implementation</li><li>- Final report</li></ul>
April 20-24 2020	Final Presentation

## References

- [1] Fuwen Tan, Song Feng, Vicente Ordonez. Text2Scene: Generating Compositional Scenes from Textual Descriptions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [2] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- [3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In ICCV, 2017.
- [4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2014.
- [5] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, Jianfeng Gao. Object-driven Text-to-Image Synthesis via Adversarial Training. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.
- [7] Tobias Hinz, Stefan Heinrich, Stefan Wermter. Generating Multiple Objects At Spatially Distinct Locations. In ICLR, 2019.