CS280 Final Milestone
Text to Image with Sentiment

Yin Caiyuan Wang Yuzhen

1. Introduction

To truly understand the visual world our models should be able not only to recognize images but also generate them. Aside from imparting deep visual understanding, methods for generating realistic images can also be practically useful. In the near term, automatic image generation can aid the work of artists or graphic designers. One day, we might replace image and video search engines with algorithms that generate customized images and videos in response to the individual tastes of each user.

As a step toward these goals, there has been exciting recent progress on text to image synthesis[1,2,3,5] by combining recurrent neural networks and Generative Adversarial Networks [4] to generate images from natural language descriptions. These methods can give stunning results on limited domains, such as fine-grained descriptions of birds or flowers. And the leading method such as Image Generation from Scene Graphs[6] solved the problem of struggle with complex sentences containing many objects.

However, these methods are processed sentences containing only nouns and prepositions. In realistic, we write sentences with more than nouns and prepositions. Such as:Two man siting on the grass gladly. Methods above can do well to generate image with sentence “Two man siting on the grass” but fail to put up the emotion of “gladly”.

In our project we aim to generate images with sentiment expression. We plan to achieve sentiment expression by adjust the overall color and tone of the images. With this new task comes new challenges. We must do a sentiment analysis for a curtain sentence.

Sentiment analysis is the computational study of people’s opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. The inception and rapid growth of the field coincide with those of the social media on the Web, for example, reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP). It is also widely studied in data mining, Web mining, text mining, and information retrieval. In fact, it has spread from computer science to management sciences and social sciences such as marketing, finance, political science, communications, health science, and even history, due to its importance to business and society as a whole.

By using sentiment analysis, we can classify sentences’ sentiment. We plan to classify them as three categories: positive, neutral and negative. For example, in sentence “Two man sitting on the grass gladly.” we can obviously classify it as positive sentence because the word “gladly”. So, when generating an image, we will choose more warmer tone and more brighter colors as a background of the image. This effort can make the image we generated more humanized.

2. Problem statement

2.1. Dataset

To achieve our goal to generate images from texts with sentiment. We need generate images from texts properly first. And then add the sentiment factor.

A sentence is a linear structure, with one word following another. In this project we aim to generate complex images with many objects and relationships by conditioning our generation on scene graphs, allowing our model to reason explicitly about objects and their relationships. And we use dataset COCO.

2.2. Expected results

In our project, we expected our model can classify the input sentences as positive or negative or neutral. And based this classify, using different color tone to generate scene graph images. For example, three picture Fig.1. , Fig.2. and Fig.3. are stand for image generated from sentence “above the grass is sky with clouds”, but Fig.1. is a very bright picture and Fig.2. is in normal tone, while the

Fig.3. is more darker than the others. Fig.1. is positive, Fig.2. is neutral and Fig.3. is negative.

Fig.1. Positive image generation

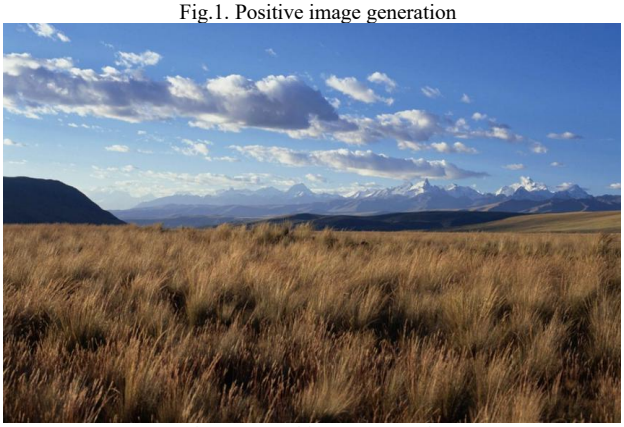

Fig.2. Neutral image generation

Fig.3. Negative image generation

3. Method

3.1 Image generation from scene graphs

Our goal is to develop a model which takes as input a scene graph describing objects and their relationships, and which generates a realistic image corresponding to the graph. The primary challenges are threefold: first, we must develop a method for processing the graph-structured input;second, we must ensure that the generated images respect the objects and relationships specified by the graph; third, we must ensure that the synthesized images are realistic.

We convert scene graphs to images with an image generation network f, shown in Fig.4., which inputs a scene graph G and noise z and outputs an image $\hat{I} = f(G, z)$.

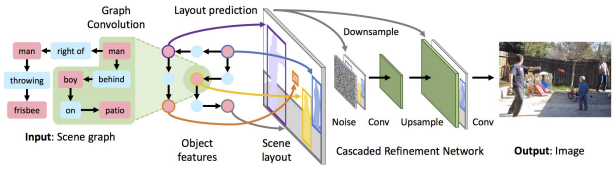

Fig.4. Overview of our image generation network f for generating images from scene graphs.

The scene graph G is processed by a graph convolution network which gives embedding vectors for each object; as shown in Fig.4. and Fig.5., each layer of graph convolution mixes information along edges of the graph.

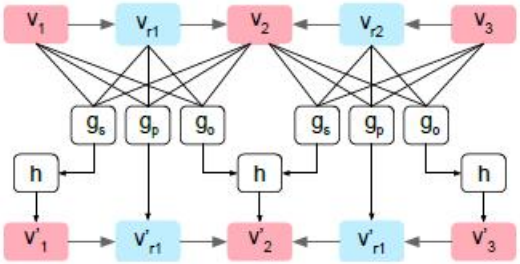

Fig.5. Computational graph illustrating a single graph convolution layer.

We respect the objects and relationships from G by using the object embedding vectors from the graph convolution network to predict bounding boxes and segmentation masks for each object; these are combined to form a scene layout, shown in the center of Fig.4., which acts as an intermediate between the graph and the image domains.

The output image $\hat{I}$ is generated from the layout using a cascaded refinement network (CRN) [7], shown in the right half of Fig.4.; each of its modules processes the layout at increasing spatial scales, eventually generating the image $\hat{I}$.

We generate realistic images by training f adversarially against a pair of discriminator networks D_{img} and D_{obj} which encourage the image $\hat{I}$ to both appear realistic and to contain realistic, recognizable objects.

3.2 Sentiment classify

To achieve the goal of classify sentences by sentiment, we plan use a torchMoji model. As showing in Fig.6., this model contains two layers of LSTM, and attention layer and classifier leave behind.

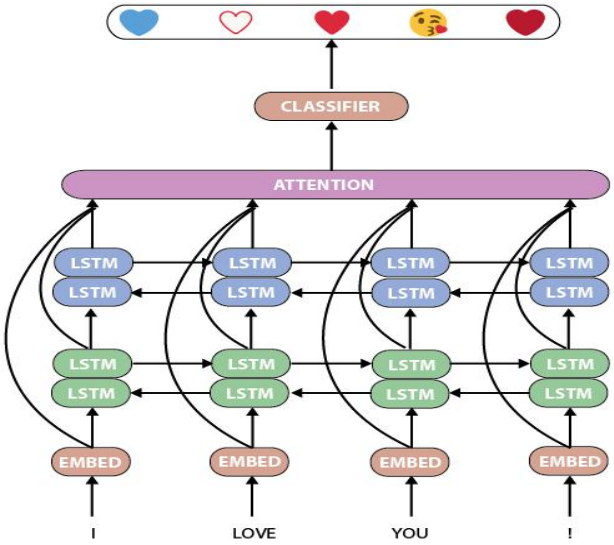

Fig.6. torchMoji Model with two layers of LSTM

4. Preliminary results

We can generate a scene image from curtain sentences now, but still cannot add sentiment expression in our generated images. Following Fig.7. and Fig.8. are two images we generated.


Fig.7. two sheep are eating grass with mountain behind a tree and sky has cloud


Fig.8.8 a boy standing on grass is looking at sky and kite with field under mountain

References

[1] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016. 1,

[2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016. 1, 2

[3] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G´omez, Z. Wang, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. In ICML, 2017. 1, 2

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014. 1, 2, 4

[5] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017. 1, 2, 7, 8

[6] Johnson J , Gupta A , Fei-Fei L . Image Generation from Scene Graphs[J]. 2018.

[7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In ICCV, 2017. 2, 3, 4, 7