HEAR FROM IMAGES: A STORY GENERATOR USING VISUAL RELATION DETECTION

Kong Lingdong

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Email: KONG0129@e.ntu.edu.sg, **Matri. Number:** G1902089A

ABSTRACT:

Imagine that a machine can automatically write a travel diary for you, all you need to do is feeding in an image stream you took on the journey and wait for a few minutes. This is called visual storytelling. In the computer vision community, generating captions from provided images has been widely studied. However, few works have been done to improve this task from "captioning" to "storytelling" due to the gaps between analyzing a single image to analyzing a stream of images and finding their relationships in between. In this assignment, an idea of designing a visual relation detection-based story generator is proposed. This story generator is desired to have the ability to analyze the images inputted, outline the event fragments, and output a reasonable story.

KEYWORDS: Deep learning, Image captioning, Visual relation detection.

I. IMAGE CAPTIONING

Image captioning is an interesting task in visual understanding, which aims to analyze the visual content of the input image and generate a caption that verbalizes the most salient contents of that image. Different from other computer vision tasks such as object detection and image categorization, image captioning is a multi-modal task that requires the machines to not only capture the salient aspects of an image but also have the ability to describe the image with natural language. The current image captioning methods use a CNN-based encoder to extract visual features and use an RNN-based decoder for caption generation. Some of these models introduce methods such as reinforcement learning and adversarial learning to produce smoother, more expressive, and more reasonable captions.

Let us separate the image captioning task into two components: feature extraction and language modeling. In the feature extraction part, salient features in images inputted are extracted by neural networks, which usually represented by a fixed-length vector. In the language modeling part, when a partial caption has been generated, the language model can predict the probability of the next word of the sequence. A common method is to use an LSTM. Each output time step generates a new word in the sequence. Each generated word is then encoded using a word embedding, such as word2vec, which is then passed as the input to the decoder to generate subsequent words. For image description, a neural network such as a

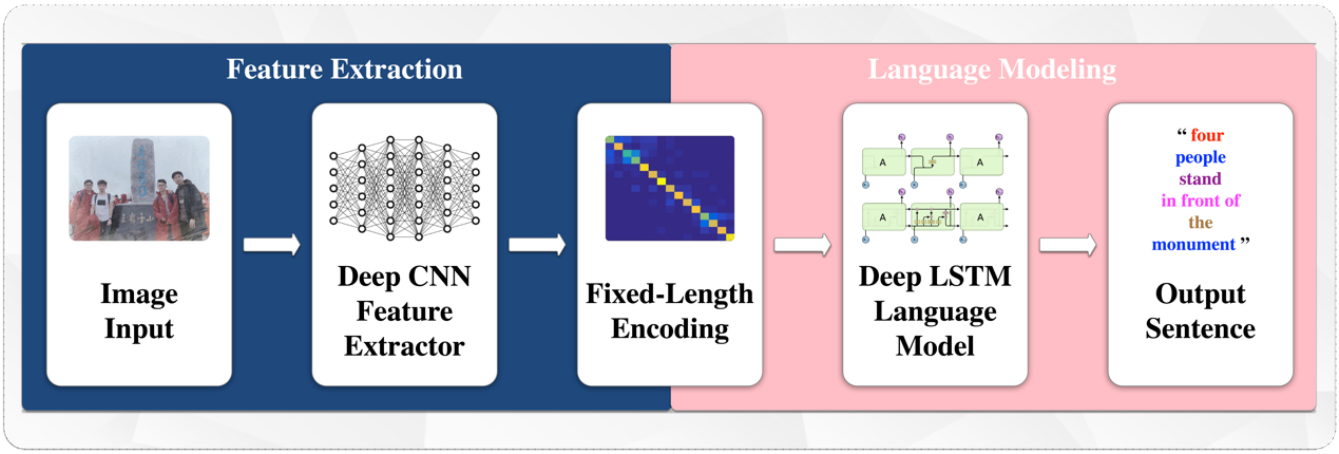

Figure. 1 The Procedures of Image Captioning

language model can predict the word sequence in the description based on the features extracted by the network and construct a description based on the words that have been generated. The image captioning can be achieved by combining the above two parts. Figure. 1 shows the detail procedures.

II. Visual Relation Detection

Although image captioning achieves some good performance, however, it is not intuitive to represent all the visual information of the images with abstract high-level features extracted by deep CNNs, which also impairs the interpretability and reasoning ability of the model. Recall that when we tell a story for an image stream, we will identify the objects in each image, reason about their visual relationships, and then abstract the content into a vision. We then will observe the images in order and infer the relationships between [1].

Different from image captioning, visual relation detection (VRD) aims to combine low-level object detection and the high-level relationship analysis in an image. In [2], researchers proposed a Parallel Pairwise Region-based Fully Convolutional Network to detect "subject-predicate-object" relations in images. Its architecture can be summarized as follows: two parallel modules, i.e., a weakly supervised object detection (WSOD) module, and a weakly supervised predicate prediction (WSPP) module, simultaneously perform pair selection and classification of single regions and region pairs for WSOD and WSPP, while sharing almost all computation over the entire image [2].

An interesting multi-task triple-stream network introduced in [3] gives another solution. The detailed procedures show in Figure. 2. The authors use a localization module based on the region proposal network (RPN) [4] for object detection and a triple-stream LSTM module for captioning. The main procedures can be summarized as follows: first, given an image, use RPN to generate object proposals. Then, the combination layer takes a pair consisting of an

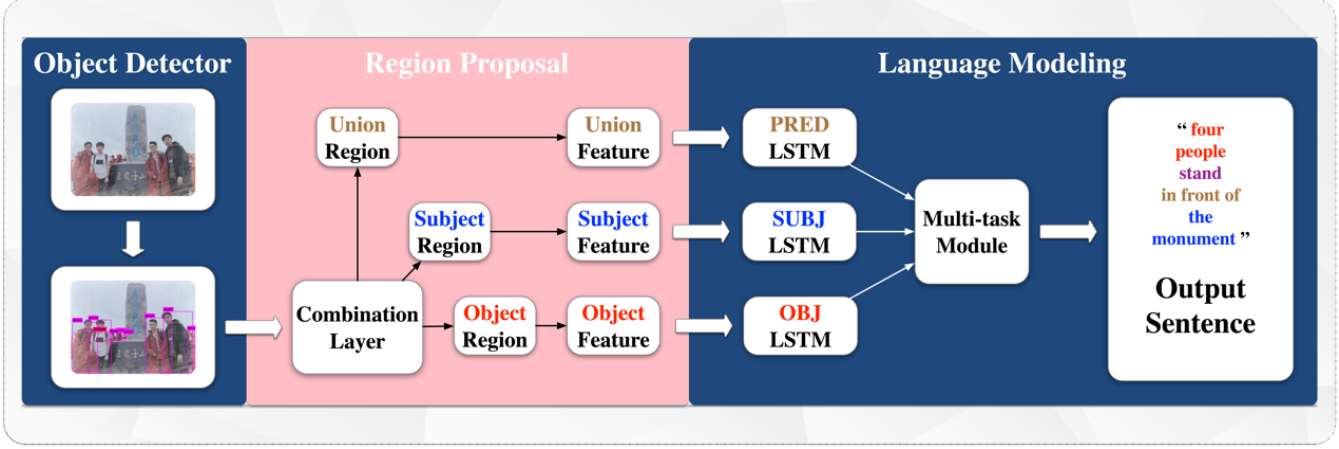

Figure. 2 The Procedures of Relation Detection in Ref. [3]

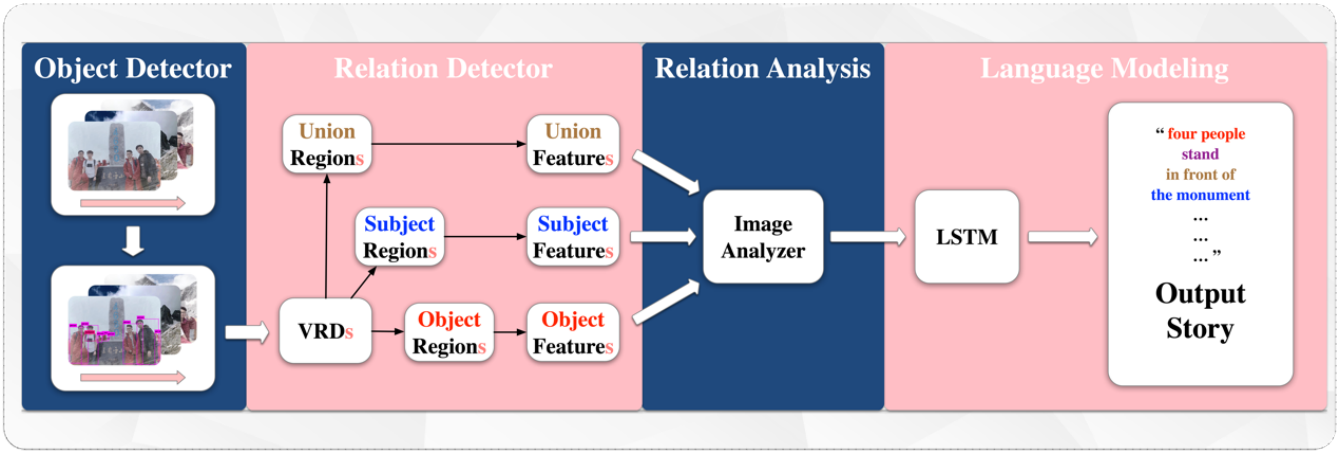

Figure. 3 The Procedures of the Story Generator at Image-Level

object and a subject at a time. To take the surrounding context information into account, use the union region of the subject and object regions. This feature of triplets is fed to the triple-stream LSTMs, where each stream takes its purpose, i.e. union, subject, and object. Given this triplet feature, the triple-stream LSTMs collaboratively generate a caption and the part-of-speech (POS) class tags of each word [4].

III. Method Proposed in This Assignment

Inspired by the aforementioned image captioning model [3] and the VRD models [1, 2, 5], in this assignment, a visual relation analysis-based story generator is designed. The basic building blocks are as follows:

- An **Object Detector module** which is fed with an image stream with a timeline (e.g., a series of photos taken during the tour).

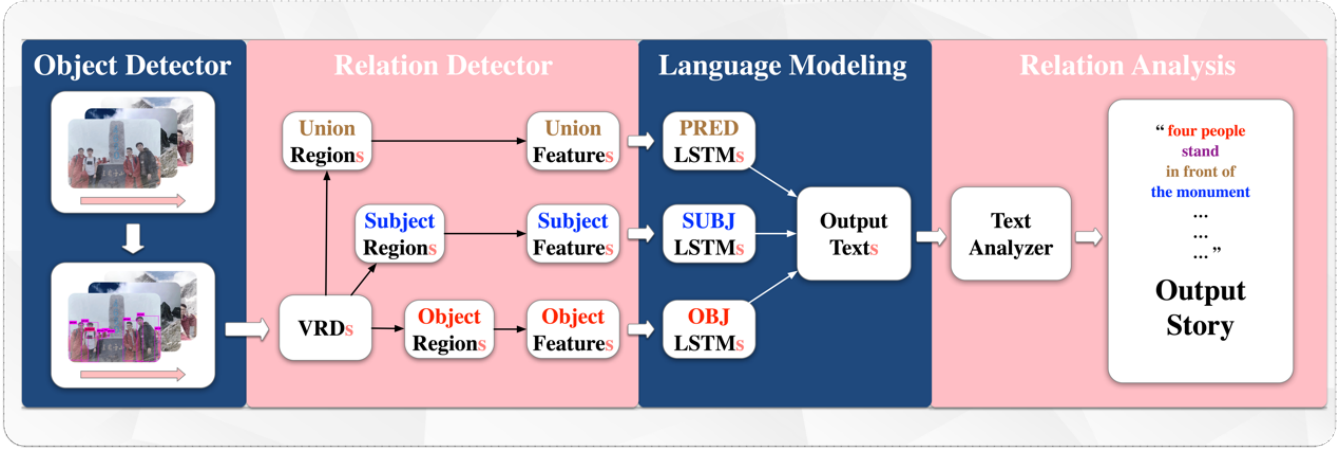

Figure. 4 The Procedures of the Story Generator at Text-Level

- A **Relation Detector module** uses deep CNNs to detect features among images inputted.
- A **Relation Analysis module** to analyze the relations between each image. This could be done at the image-level (see Figure. 3) or at the text-level (see Figure. 4).
- A **Language Modeling module** uses LSTMs to generate text segments. A story is then synthesized based on these segments.

The general principles of the image-level relation analysis are as follows: first of all, with the image stream inputted, we use CNNs to detect the objects in each image and output their features. Second, construct a relationship graph of all the objects detected, particularly focus on the object that occurred in consecutive images. Next, based on this relationship graph constructed, use a parallel model to analyze the relations of the inputted images and output high-level features. Finally, feed an LSTM with high-level features and generate a story.

As for the text-level relation analysis, the general principles are: firstly, extracting the union, subject and object features of the input image stream by CNNs, respectively. We then use a parallel multi-LSTMs network to generate text segments of each image inputted and construct a relationship table correspondingly. Based on this table, we use the text analyzer to analyze the relations among all the objects at the text-level and generate a story directly. To summarize, the key component of this method is the designing of the relation analyzers, we will focus on research in this part in the future.

IV. Conclusion

In this assignment, we design a story generator using relation analysis. This model combines low-level relation detection and high-level relation analysis together and aims to generate an integral story instead of just captions. There are a lot of steps to take before this idea comes true and I will try my best to fulfill the missing parts of this model in the future.

REFERENCES

[1] R. Wang, Z. Wei, P. Li, Q. Zhang, X. Huang. "Storytelling from an Image Stream Using Scene Graphs," *AAAI*, 2020.

[2] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang. "PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN," *ICCV*, 2017.

[3] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon. "Dense Relational Captioning: Triple-Stream Networks for Relationship-based Captioning", *CVPR*, 2019.

[4] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *NIPS*, 2015.

[5] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. "Visual Translation Embedding Network for Visual Relation Detection," *CVPR*, 2017.