Empirical Investigation of Various Deep Learning Models For The Production of Image To Text

Ravi Ranjan[1], Parvesh[2]
[1]Dept of Computer Science and Engineering
[2]Assistant Professor, Dept of Computer Science and Engineering
[1, 2] SKITM, Bahadurgarh, India

Abstract- *Image Captioning deals with the process of generating textual description from an image based on the objects and their properties in the image. This process has many potential applications in real life. An important one would be to save the captions of an image so that it can be retrieved easily at a later stage just on the basis of this description. In this paper, we have shown the various methods of image captioning. Along with various methods for image captioning, the various models for better performance are also discussed. Using this, appropriate method or technique for image captioning can be used as we have various types of images in real life from various domains. In last, we have shown datasets, evaluation metrics and futuristic research directions.*

Keywords- Image Captioning, Deep Learning

I. INTRODUCTION

Image Captioning refers to the process of generating textual description from an image based on the objects and actions in the image. For example:

Fig. 1. Image Captioning Real Life Example

This process has many potential applications in real life. A noteworthy one would be to save the captions of an image so that it can be retrieved easily at a later stage just on the basis of this description.

VARIOUS IMAGE CAPTIONING METHODS [1]:

Image captioning methods can be categorized into three main categories:

1. Template Based Image Captioning
2. Retrieval Based Image Captioning
3. Novel Image Caption Generation

Deep Learning Image Captioning methods mostly belong to novel image caption generation method.

Encoder-Decoder Architecture

- Visual Space
- Multimodal Space
- Compositional Architecture
- LSTM language Model Based Architecture
- Supervised Learning Based
- Unsupervised Learning Based etc.

VARIOUS MODELS FOR BETTER PERFORMANCE:

Human Computing:

Although machine learning does well in designated scenes, it's difficult for machines to perform well in commonsense knowledge. For example, a surfboard vertically standing against the backdrop of the sea and ocean was mistaken as a person by a machine. These errors would be easily found by humans. These problems are known as UUs problems. UUs(Unknown Unknown) is originated from features disproportionately distributed in the training set and test set and can be distributed in every corner of the data set, which is an external form of underfitting or model caused by data bias.

Application of human computing in image captioning can solve this problem. It has now gradually evolved an independent tool used in data collection and data cleaning and is used to solve UU problem[2].

The human computing method to help find UUs in this paper can be described in three steps:

1. Divide the mass labeled data set into several fragmentations with different confidence by original evaluation systems, try to mine as many as possible UUs candidate word pairs in high confidence fragmentations.

2. Randomly extract some corresponding generated descriptions with high confidence, try to judge the UUs candidate word pairs in these descriptions with human computing, and if anything wrong is found, label them and record their error types.

3. Record the appearance frequency of UUs candidate word pairs. Select the UUs candidate word pairs with the highest appearance frequency and give them synonym extensions.

Ultrasound Image Caption Generation:

There are lack of effective methods for detailed analysis and automatic description of diseases content information in ultrasound image understanding. In order to find the location of focus areas, and understand the content of focus areas conveniently, a novel method of ultrasound image captioning is generated based on region detection. The method simultaneously detects and encodes the focus areas in ultrasound images, then utilizes the LSTM to decode the encoding vectors and generate annotation text information to describe the diseases content information in ultrasound images[3].

The experimental results show that the method can accurately detect the location of the focus area, and also improves 1% the scores of BLEU-1, BLEU-2 with less parameters and running time, which compared with the full-size-image captioning model for ultrasound images. The model mainly solves the problem of interference between different organs, and obtains more detailed information about the focus area.

Image-Text Semantic Relations:

Image captioning focus on precisely describing visual content and translating it to text, but typically address neither semantic interpretations nor the specific role or purpose of an image-text constellation[4]. Only few methods are available for generating image-text semantic relations.

DHEDN for Image Captioning:

Deep Hierarchical Encoder-Decoder Network (DHEDN) is proposed for image captioning, where a deep hierarchical structure is explored to separate the functions of encoder and decoder. This model is capable of efficiently exerting the representation capacity of deep networks to fuse high level semantics of vision and language in generating captions. Specifically, visual representations in top levels of abstraction are simultaneously considered, and each of these levels is associated to one LSTM. The bottom-most LSTM is applied as the encoder of textual inputs. The application of the middle layer in encoder-decoder is to enhance the decoding ability of top-most LSTM[5].

There are several places of our current work to explore. First, it makes sense to continue to increase the "vertical depth" of encoder-decoder, but how to stack different layers and set those numerous parameters remain difficult to be solved. Second, importing extra inputs like visual attention or attributes into the deep hierarchical encoder-decoder. The complementary knowledge has been demonstrated beneficial for the performance improvement of image descriptions generation. How to integrate complementary knowledge into the deep hierarchical encoder-decoder structure is worth trying and seems very attractive.

Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning

By combining a deep convolutional neural network (CNN) and two separate LSTM networks, model is capable of learning long-term visual-language interactions by making use of history and future context information at high-level semantic space. There is one more deep multimodal bidirectional models, where the author increase the depth of nonlinearity transition in different ways to learn hierarchical visual-language embeddings. To understand how models "translate" image to sentence, author visualize and qualitatively analyze the evolution of Bi-LSTM internal states over time. The effectiveness and generality of proposed models are evaluated on four benchmark datasets: Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets[7].

Bottom-Up and Top-Down Attention for Image Captioning

Here author propose a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. The bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. Applying this approach to image captioning, the

results on the MSCOCO test server establish a new state-of-the-art for the task, achieving CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5 and 36.9, respectively[10].

Unsupervised Image Captioning

Here the author proposes unsupervised image captioning method. Instead of relying on manually labeled image-sentence pairs, proposed model merely requires an image set, a sentence corpus, and an existing visual concept detector. The sentence corpus is used to teach the captioning model how to generate plausible sentences. Meanwhile, the knowledge in the visual concept detector is distilled into the captioning model to guide the model to recognize the visual concepts in an image. In order to further encourage the generated captions to be semantically consistent with the image, the image and caption are projected into a common latent space so that they can reconstruct each other[12].

Image caption generation with high-level image features

It is challenging for the models to select proper subjects in a complex background and generate desired captions in high-level vision tasks. Inspired by recent works, author proposed a novel image captioning model based on high-level image features. Author combine low-level information, such as image quality, with high-level features, such as motion classification and face recognition to detect attention regions of an image. This attention model produces good performance in experiments on MSCOCO, Flickr 30K, PASCL and SBU datasets[13].

Image Captioning with two cascaded agents

The author proposes a pipelined image captioning framework consisting of two cascaded agents. The former is named as “semantic adaptive agent” which generates the input to the decoder by consulting the information from the current decoding process, and the latter as “caption generating agent” which select a single word of the vocabulary as the output of the decoder by taking consideration of the input and the current states of the decoder. For the framework of two cascaded agents, author designed a multi-stage training procedure to train the two agents with different objectives by fully utilizing reinforcement learning. In experiments, quantitative and qualitative analysis on MS COCO dataset is done and results can significantly outperform baseline methods in terms of several evaluation metrics[15].

DATASETS:

- MSCOCO
- Flickr30k
- Flickr8k

These dataset are common and popular datasets used for image captioning. MSCOCO dataset is very large dataset and all the images in these datasets have multiple captions. Visual Genome dataset is mainly used for region based image captioning [1].

EVALUATION METRICS: Different evaluation metrics are used for measuring the performances of image captions.

- ROUGE
- METEOR
- SPICE
- BLEU

FUTURISTIC DIRECTIONS:

- Generation-based methods can generate novel captions for every image. However, these methods fail to detect prominent objects and attributes and their relationships to some extent in generating accurate and multiple captions.
- Working on an open-domain dataset will be an interesting avenue for research in this area.
- Supervised learning needs a large amount of labeled data for training. Therefore, unsupervised learning and reinforcement learning will be more popular in the future in image captioning.
- Multiple UUs in one single scenario, with the help of GANs and reinforcement learning to optimize our models.
- Automatic understanding of multimodal information is still an unsolved research problem.
- Increasing vertical depth of encoder-decoder for better image captioning
- To explore the multilingual caption generation problem.
- Human evaluations for unsupervised image captioning.
- Image captioning with high level image features may lead to visual attention.
- Better pipelined model of two cascaded agents

REFERENCES

[1] MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, and HAMID LAGA, “A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys, Vol. 51, No. 6, Article 118, February 2021.

[2] Zhihong Zeng, Xiaowen Li, "Application of human computing in image captioning under deep learning", Springer Nature 2019, May 2020.
[3] Xianhua Zeng, Li Wen, Banggui Liu, Xiaojun Qi, "Deep Learning for Ultrasound Image Caption Generation based on Object Detection", *Neurocomputing* (2019), doi:https://doi.org/10.1016/j.neucom.2018.11.114, Nov 2018.
[4] Christian Otto, Matthias Springstein, Avishek Anand, Ralph Ewerth, "Understanding, Categorizing and Predicting Semantic Image-Text Relations", ICMR '19, Ottawa, ON, Canada , June 10–13, 2019.
[5] Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan, "Deep Hierarchical Encoder-Decoder Network for Image Captioning", DOI 10.1109/TMM.2019.2915033, IEEE Transactions on Multimedia, 2019.
[6] YutingSu, YuqianLi, Ning Xu, An-An Liu, "Hierarchical Deep Neural Network for Image Captioning", Springer Science+Business Media, LLC, Springer Nature , 2018.
[7] CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL," 40 Image Captioning with Deep Bidirectional LSTMsand Multi-Task Learning", *ACM Trans, MultimediaComput. Commun., Appl.*14, 2s, Article 40, April 2018.
[8] XiaoxiaoLiu, Qingyang Xu, Ning Wang, "A survey on deep neural network-based image captioning", https://doi.org/10.1007/s00371-018-1566-y, Springer Nature, 2020.
[9] Vasiliki Kougia, John Pavlopoulos, Ion Androutsopoulos, "A Survey on Biomedical Image Captioning", arxiv:1905.13302v1, May 2017.
[10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", IEEE Explore, 2017.
[11] Justin Johnson, AgrimGupta, Li Fei-Fei, "Image Generation from Scene Graphs", IEEE Explore, 2020.
[12] Yang Feng, Lin Ma, Wei Liu, Jiebo Luo, "Unsupervised Image Captioning", IEEE Explore, 2019.
[13] SongtaoDing ,Shiru Qu , Yuling Xi , Arun Kumar Sangaiah , Shaohua Wan, " Image caption generation with high-level image features", Pattern Recognition Letters 123 (2019) 89–95, Mar 2019.
[14] Lin Ma, Wenhao Jiang, ZequnJie, Yu-Gang Jiang, and Wei Liu, "Matching Image and Sentence with Multi-faceted Representations", DOI 10.1109/TCSVT.2019.2916167, IEEE Transactions on Circuits and Systems for Video Technology, 2019.
[15] Lun Huang, Wenmin Wang, Gang Wang, "IMAGE CAPTIONING WITH TWO CASCADED AGENTS", ICASSP 2019, IEEE, 978-1-5386-4658-8/18, 2018.