An Empirical Study of Different Deep Learning Models for Image to Text Generation

Mohit Dalal[1] Minakshi Arora[2]
[1]M.Tech. Scholar [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]SKITM, Bahadurgarh, India

Abstract — The technique of creating a written description of an image from its objects and their characteristics is known as image captioning. There are a lot of practical uses for this procedure. One crucial one would be to store a picture's captions so that, in the future, the image may be conveniently accessed based only on this description. The authors have demonstrated the various image captioning techniques in this research. The numerous models for improved performance are also examined, in addition to several approaches for image captioning. Since there are many different kinds of photographs from different domains in real life, this can be utilized to find the right method or strategy for captioning images. Finally, the authors have included evaluation criteria, statistics, and future areas for research.

Keywords: Machine Learning, Deep Learning, Encoder, Decoder

I. INTRODUCTION

The technique of creating a written description of an image from its objects and actions is called "image captioning." For example:

Fig. 1: Image Captioning Real Life Example

There are a lot of practical uses for this procedure. One important one would be to store a picture's captions so that, in the future, the image may be conveniently accessed based only on this description.

II. 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.

III. SELECTIVE MODELS TO IMPROVE PERFORMANCE:

A. *Human Computing:*

While machine learning performs well in specific scenarios, commonsense knowledge is harder for robots to grasp and use. For instance, a machine mistook a surfboard that was vertically positioned against the water and ocean for a human. Humans might easily find these mistakes. These issues are referred to as UUs issues. UUs, or "Unknown Unknown," are an external type of underfitting or a model driven by data bias. They are derived from characteristics that are disproportionately distributed in the training and test sets and can appear anywhere in the data set.

This issue can be resolved by using human computation in the captioning of images. Over time, it has developed into a stand-alone tool for gathering and cleaning data as well as for resolving UU issues [2].

Three phases can be used to explain the human computing approach used in this research to assist locate UUs:

1) Using the initial assessment systems, divide the mass labeled data set into many fragmentations with varying degrees of confidence. Then, attempt to mine as many UUs candidate word pairs in high confidence fragmentations as you can.
2) With a high degree of confidence, extract some correspondingly generated descriptions at random. Then, attempt to evaluate the UUs candidate word pairs in these descriptions using human computation. If any errors are discovered, mark them and note the nature of the errors.
3) Note how frequently UUs candidate word pairings emerge. Give synonym extensions to the UUs candidate word pairs that have the highest frequency of occurrence.

B. *Generation of Ultrasound Image Captioning:*

Effective techniques for a thorough analysis and automatic illness content description in ultrasound picture interpretation are lacking. Based on region recognition, a unique approach of ultrasound image captioning is produced to readily grasp the content of focus areas and determine their position. The technique uses the LSTM to decode the encoding vectors and provide annotation text information to explain the illness content information in ultrasound pictures after concurrently detecting and encoding the focus areas in the images [3].

The experimental findings demonstrate that, in addition to improving 1% the scores of BLEU-1 and BLEU-

2 with fewer parameters and running time as compared to the full-size-image captioning model for ultrasound pictures, the method can accurately identify the focal area's position. The model primarily resolves the issue of organ interference and gathers more precise data regarding the focus region.

C. Relations of Image-Text Semantics:

The goal of image captioning is to accurately describe visual content and translate it into text. However, it usually does not address semantic interpretations or the particular function of an image-text constellation[4]. There aren't many techniques for creating semantic links between images and texts.

D. Using DHEDN to Caption Images:

In order to distinguish between the roles of encoder and decoder, a deep hierarchical structure is investigated in the Deep Hierarchical Encoder-Decoder Network (DHEDN), which is suggested for picture captioning. With the use of deep networks' representational power, this model can effectively combine the high-level semantics of language with vision to produce captions.

In particular, top-level abstraction visual representations are taken into account concurrently, and every degree of abstraction is linked to a single LSTM. Textual inputs are encoded using the bottom-most LSTM. The purpose of the middle layer in an encoder-decoder is to improve the highest-order LSTM's decoding capability [5].

There are various areas to investigate in our ongoing work. First, it makes sense to keep increasing the encoder-decoder's "vertical depth," but the puzzle of how to stack several levels and configure all of those settings still has to be solved. Second, adding additional inputs to the deep hierarchical encoder-decoder, such as characteristics or visual attention. It has been shown that the complementary information improves the performance of the production of image descriptions. It appears quite appealing to try and integrate complementary knowledge into the deep hierarchical encoder-decoder structure.

E. Deep Bidirectional LSTMs and Multi-Task Learning for Image Captioning

Through the integration of two distinct LSTM networks and a deep convolutional neural network (CNN), the model is able to learn long-term visual-language interactions by leveraging context information from the past and future at a high semantic space level. An additional deep multimodal bidirectional model is presented, in which the author finds several methods to learn hierarchical visual-language embeddings by increasing the depth of nonlinearity transition. The author visualizes and qualitatively analyzes the evolution of Bi-LSTM internal states over time to understand how models "translate" image to text. Four benchmark datasets are used to assess the efficacy and generalizability of the suggested models: the Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets [7].

F. Top-Down and Bottom-Up Focus on Image Captioning

The author of this work suggests a bottom-up and top-down combined attention mechanism that allows attention to be determined at the object and other salient picture regions levels. While the top-down technique establishes feature weightings, the bottom-up mechanism (based on Faster R-CNN) suggests image regions, each with a corresponding feature vector. When this method is used to picture captioning, CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5, and 36.9, respectively, set a new state-of-the-art for the task on the MSCOCO test server [10].

G. Unsupervised Image Captioning

Here, the author suggests using an unsupervised method to caption images. Rather than depending on manually annotated image-sentence pairs, the suggested model only needs an image collection, a corpus of sentences, and an already-existing visual concept detector. The captioning model is trained to construct plausible sentences using the sentence corpus. To help the captioning model identify the visual concepts in an image, the visual concept detector's information is condensed into it. To promote semantic consistency in the generated captions, the image and caption are projected onto a shared latent space, allowing them to rebuild each other [12].

H. Creating captions for images with advanced features

In high-level vision tasks, it is hard for the models to choose appropriate subjects in a complicated context and produce acceptable captions. The author put out a novel picture captioning approach based on high-level image attributes, which was inspired by recent publications. To identify focus areas in a picture, the author combines high-level elements like face recognition and motion classification with low-level data like image quality. Experiments on MSCOCO, Flickr 30K, PASCL, and SBU datasets show good performance from this attention model [13].

I. Using two cascading agents for image captioning

Two cascaded agents make up the pipelined image captioning structure that the author suggests. The former, known as the "semantic adaptive agent," creates the input for the decoder by examining data from the ongoing decoding process, and the latter, known as the "caption generating agent," chooses one word from the vocabulary to be the output of the decoder by taking into account both the input and the decoder's current state. The author created a multi-stage training process for the framework of two cascaded agents in order to fully utilize reinforcement learning to train the two agents with distinct aims. Experiments are conducted using the MS COCO dataset, and the findings show that the quantitative and qualitative analysis greatly outperforms baseline methods in a number of evaluation measures [15].

IV. DATASETS:

- MSCOCO
- Flickr30k
- Flickr8k

These datasets are widely used and well-liked for captioning images. The MSCOCO dataset is incredibly huge, with several captions for each image. The primary application of the Visual Genome dataset is region-based image captioning. [1].

A. Evaluation Metrics:

The effectiveness of picture captions is evaluated using a variety of measures.

- ROUGE
- METEOR
- SPICE
- BLEU

V. FUTURISTIC DIRECTIONS:

- New captions for each image can be produced by generation-based methods. Nevertheless, these techniques fall somewhat short in identifying salient features and objects as well as their connections when producing precise and many captions.
- Working with an open-domain dataset will be a fascinating direction for this field of study.
- A significant amount of labeled data is required for training in supervised learning. For this reason, in the future, image captioning will use more unsupervised learning and reinforcement learning techniques.
- Multiple UUs in a single scenario, which we optimize via reinforcement learning and GANs.Automatic understanding of multimodal information is still an unsolved research problem.
- Deepening the encoder-decoder's vertical depth to improve picture captioning.
- To investigate the issue of producing captions in multiple languages.
- Assessments of unsupervised image captioning by humans.
- Using high-level image elements in image captions may draw viewers' attention.
- An improved pipelined model with two cascading 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 2022.

[2] Zhihong Zeng, Xiaowen Li, "Application of human computing in image captioning under deep learning", Springer Nature 2019, May 2021.

[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] Yuting Su, Yuqian Li, 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 LSTMs and Multi-Task Learning", ACM Trans, Multimedia Comput. Commun., Appl. 14, 2s, Article 40, April 2018.

[8] Xiaoxiao Liu, 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, Agrim Gupta, 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] Songtao Ding , 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, Zequn Jie, 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.

