A Framework for Generating Text from Images Using Deep Learning

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 — Image captioning is the process of creating a description for a picture. Identification of the key elements, their characteristics, and their interactions within an image are necessary for image captioning. 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. Our goal in this survey paper is to provide a thorough analysis of the state-of-the-art deep learning-based picture captioning methods. We go into the theoretical underpinnings of the methods to evaluate their effectiveness, drawbacks, and strengths. Additionally, we go over the evaluation measures and datasets that are frequently employed in deep learning-based automatic image captioning.

Keywords: Deep Learning, CNN, RNN

I. INTRODUCTION

Understanding the key elements in an image, their characteristics, and their interactions is necessary for image captioning. 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. Image captioning is crucial for a variety of factors. It can be applied, for instance, to automatically index images. Image indexing has several applications in biomedicine, e-commerce, military, education, digital libraries, and web searches since it is crucial for content-based image retrieval (CBIR). Obtaining image characteristics is essential to understanding an image. The methods employed for this can be generally classified into two groups: (1) conventional methods based on machine learning, and (2) methods based on deep learning. Our goal in this survey paper is to provide a thorough analysis of the state-of-the-art deep learning-based picture captioning methods. We go into the theoretical underpinnings of the methods to evaluate their effectiveness, drawbacks, and strengths. Additionally, we go over the evaluation measures and datasets that are frequently employed in deep learning-based automatic image captioning.

II. IMAGE CAPTIONING METHODS:

Three primary categories can be used to group image captioning techniques:

Three methods of captioning images: Template-based, retrieval-based, and novel image caption generation. Template-based methods use pre-made templates with lots of blank spaces to create captions. These methods fill up the vacant spots in the templates after detecting various objects, characteristics, and actions.

Captions are recovered from a set of preexisting captions in retrieval-based techniques. Using the training dataset, retrieval-based techniques first identify the visually comparable photos and their captions.

The authors refer to these captions as "candidate captions." This captions pool is used to choose the captions for the query image. The majority of unique picture caption creation approaches include Deep Learning picture Captioning techniques.

– Multimodal Space
– Attention-Based Image Captioning
– Semantic Based Image Captioning
– Encoder-Decoder Architecture
– Compositional Architecture
– Dense Captioning

1) For the purpose of creating picture captions, encoder-decoder architecture-based techniques employ a basic CNN and a text generator.

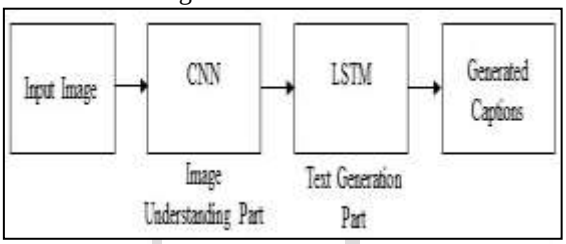

Fig. 1: A simple Encoder-Decoder architecture-based image captioning

2) Architecture-Based Composition for Image captioning. Methods based on compositional architecture and made up of multiple separate, useful building blocks: To extract the semantic concepts from the image, a CNN is first employed. The next step is to create a set of potential captions using a language model. These potential captions are reranked using a deep multimodal similarity algorithm to create the final caption.

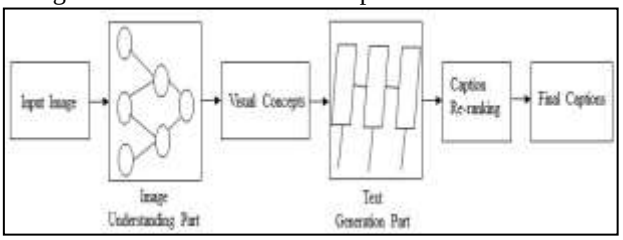

Fig. 2: Compositional network-based captioning.

3) Multimodal space based image captioning: A language encoder portion, a vision part, a multimodal space part, and a language decoder part make up the architecture of a typical multimodal space-based technique. To extract the image features, the vision component employs a deep convolutional neural network as a feature extractor. The language encoder component learns a dense feature embedding for every word by extracting its features. The semantic temporal context is subsequently sent to the recurrent layers. The picture features are mapped into a shared space with the word features in the multimodal space section.

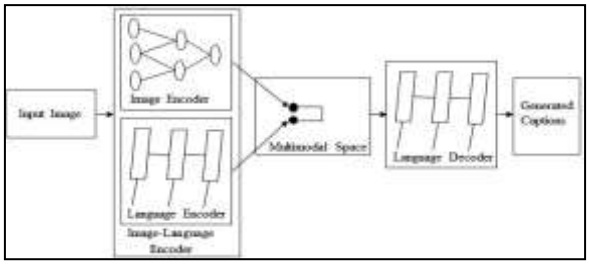

Fig. 3: Multispace architecture-based image captioning

4) Dense captioning Using this technique, all the important areas of an image are located, and after that, descriptions are produced for those areas.

The stages listed below are typical for a method in this category. Region suggestions are produced for each of the specified image's several regions. To acquire the region-based picture attributes, CNN is employed. A language model uses the outputs from previous step to create captions for each region.

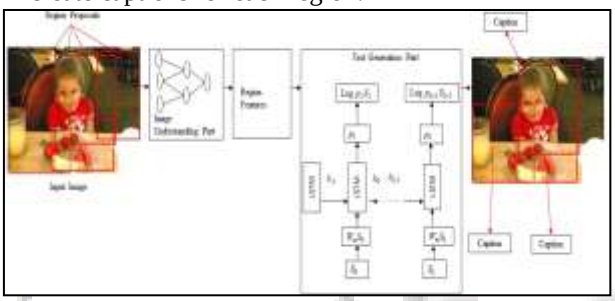

Fig. 4: Dense Captioning Architecture

5) Attention-based image captioning Compared to encoder-decoder architecture-based techniques, these methods perform better since they concentrate on various salient areas of the image.

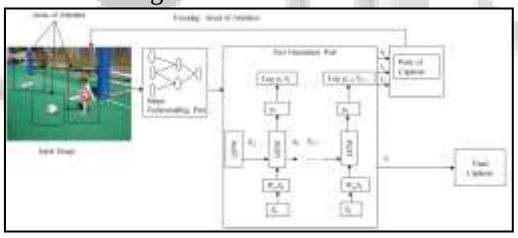

Fig. 5: A typical attention-based image captioning technique.

6) Semantic concept-based image captioning Compared to encoder-decoder architecture-based techniques, these methods perform better since they concentrate on various salient areas of the image.

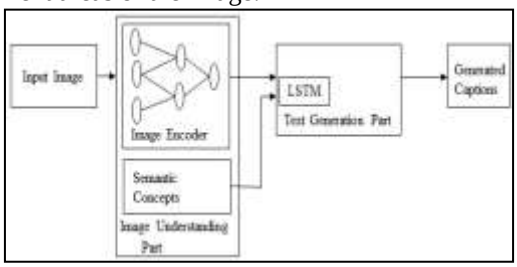

Fig. 6: A semantic concept-based image captioning.

III. DATASETS & EVALUATION METRICS:

A. *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. Flickr8k contains 8K images in dataset and Flickr30K contains 30K images in the dataset. Visual Genome dataset is mainly used for region based image captioning [1].

B. *Evaluation metrics:*

- ROUGE: Text summaries are evaluated using a set of metrics called ROUGE.
- METEOR: The machine translated language is measured using this metric.
- SPICE: A metric called SPICE assesses the semantics of image captioning.
- BLEU: This metric assesses the caliber of text produced by machines.

IV. PERFORMANCE OF DIFFERENT METHODS:

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
MS, SL, WS	0.565	0.386	0.256	0.170	-
VS, SL, WS, EDA	0.647	0.459	0.318	0.216	0.201
VS, SL, WS, EDA, AB	0.670	0.457	0.314	0.213	0.203
VS, SL, WS, EDA, SCB	0.740	0.540	0.380	0.270	-

Table 1: Dataset Flickr8K

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
MS, SL, WS	0.600	0.410	0.280	0.190	-
VS, SL, WS, EDA	0.646	0.466	0.305	0.206	0.179
VS, SL, WS, EDA, AB	0.669	0.439	0.296	0.199	0.184
VS, SL, WS, EDA, SCB	0.730	0.550	0.400	0.280	-

Table 2: Dataset Flickr30K

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
MS, SL, WS	0.670	0.490	0.350	0.250	-
VS, SL, WS, EDA	0.670	0.491	0.358	0.264	0.227
VS, SL, WS, EDA, AB	0.718	0.504	0.357	0.250	0.230
VS, SL, WS, EDA, SCB	0.740	0.560	0.420	0.310	0.260

Table 3: Dataset MSCOCO

It is clear from the above table that the MSCOCO dataset is producing useful results when using different combinations of picture captioning techniques. Nonetheless, distinct methods exist for various categories of image

captioning applications. Therefore, the efficient method for image captioning may be applied based on the needs.

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.
- The automatic comprehension of multimodal information is an open research issue. To improve image captioning, increase the encoder-decoder's vertical depth.

VI. CONCLUSIONS:

In this article, we have reviewed deep-learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagrams of the major groups, and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results was also given. We briefly outlined potential research directions in this area. Although deep-learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that is able to generate high-quality captions for nearly all images is yet to be achieved.

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 2019.
[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 , 2017.
[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 2019.
[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, 2020.
[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, 2020.
[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, 2019.
[16] Alexander G Schwing Jyoti Aneja, Aditya Deshpande. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.
[17] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.
[18] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.
[19] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In Association for Computational Linguistics. 103–111.
[20] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2016. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

