International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Facilitated Deep Learning Models for Image Captioning

T. Shravan Kumar
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
shravankumart7@gmail.com

Dr.Subba rao Kolavennu
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
profrao99@gmail.com

Pasham Sai Kishore Reddy
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
saikishorereddypasham10@gmail.com

Manpur ShivaKethan
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
lullumanpur123@gmail.com

Devarapalli Manikanth Reddy
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
dmkr1319@gmail.com

***Abstract*—In recent years, with the rapid development of technology image caption has gradually attracted the attention of many researchers as an interesting and arduous task. Image Captioning is done automatically by generating natural language descriptions according to the content observed in an image which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant. For example, the realization of human-computer interaction. With advanced deep learning techniques, accessibility of big datasets and computer power we can build an efficient model to generate image captions. To extract the image features, a Convolutional Neural Network (CNN) is used and then an extended version of Recurrent Neural Networks (LSTM) with attention-enrichment is adopted to generate the caption. We implement image captioning by considering detected objects from the image scene and then by integrating an attention mechanism for caption generation. This can have multiple advantages from accuracy and semantics perspectives. Our deep learning model generates the relevant natural language caption to the given input image by not only describing a single target object but also detects the multiple target objects in generating image captions. Image captioning helps in deeper understanding of images and monitoring crowd sourced images from social media and other sources.**

Keywords—Deep learning, Natural language processing, Image captioning, Semantics, Convolutional neural network.

I. INTRODUCTION

Image caption generation is a task in which a deep learning model is trained to automatically generate a textual description of the content of an image. The task is challenging because it requires combining two different fields such as computer vision and natural language processing. Computer vision involves the use of algorithms to enable machines to understand and interpret visual information, while natural language processing involves the use of algorithms to enable machines to understand and generate natural language text.

Neural networks that perform image captioning must not only learn the process of extracting features from the image but also encode language heuristics into the generation process that produces the caption. Learning these multimodal features can be a arduous task for a network.

Although many studies have been presented recently to address these challenges but there is still lack of producing highly semantic captions that closely and accurately describe the image content [1]. We propose a deep learning model that facilitates the image captioning process by generating a description based on detected objects and important regions from the image scene. Our approach takes away some computer vision responsibilities from the image captioning model to help the facilitating generation process that results in producing more accurate captions.

II. RELATED WORK

A. Deep Learning Approaches

Deep learning methods became widely used in image captioning since it can generate novel captions by analyzing the visual content of the image using an image model and generates the caption using a language model. Deep learning has made significant progress in many fields including computer vision and natural language processing by allowing models to learn complex patterns in data and perform advanced tasks. For early approaches like template-based approach, Kulkarni et al [2] proposed to extract the image attribute tuples and then generate words by using n-gram based language models to retrieve the final caption. For instance, the survey conducted by Zheng [3] shows the

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

importance of object detection in understanding image content and extracting descriptive semantics of an image. A study carried out in [4] explores the relationship between objects and image captioning. This study promotes the usage of objects in cohesion with captioning models and strengthens our confidence in our proposed model. An Encoder-Decoder architecture is inspired by the machine translation model proposed by Sutskever et al. [5].

B. *Language Models: RNN / LSTM*

Recurrent Neural Networks (RNN) is used in many sequence learning tasks such as machine translation, speech recognition and image captioning tasks [6]. Traditional RNN suffer from vanishing and exploding gradient problem, which means that they cannot predict words in long-range dependencies. Therefore, LSTM an improved version of RNN is adopted in many studies. LSTM has special units in addition to the standard units of RNN that uses a memory cell that maintains information in memory for longer period of time [7]. Vinyals et al. [8] used the LSTM as a decoder for the encoded image to generate the caption. A recent approach proposed an enhanced LSTM architecture that considers the semantic roles of words towards a better sentence modeling [9].

Our deep learning model drives the caption generation by considering the significant objects in the image scene and generates a highly semantic description.

III. Methodology

We propose a facilitated image captioning model that leverages the power of object detectors in cohesion with predefined heuristic, feature extractors and attention enriched language models to generate semantic captions. This end-to-end system can be visualized in Figure 1. The training happens in a multi-staged manner. Different modules need to be trained for distributed inference and training purposes in order to compile the final solution.

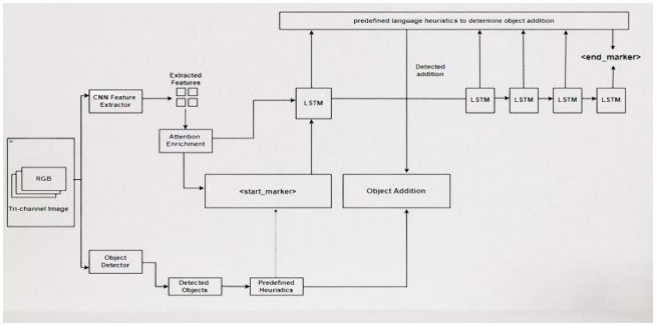

Figure 1: Image Captioning Architecture

A. *Pretrained Feature Extractor*

The first stage of the model is to use a pretrained feature extractor that is able to extract image features from a given image. Our network backbone was chosen to be an ImageNet pretrained ResNet50 model. The ResNet50 is a convolutional neural network architecture that contains 50 layers of convolutional, pooling, and activation functions. It has been trained on more than a million images from the ImageNet dataset which contains images labeled with one of a thousand object categories such as animals, household items, and natural scenes. Libraries such as TensorFlow provide easy-to-use APIs for loading and using pretrained models like ResNet50 making it accessible to researchers and developers with little expertise in deep learning.

B. *Object Detection*

In the second stage of the model, a detection module is developed to identify objects within the input images. One commonly used detection model is RetinaNet which is a one stage object detection model that is known for its effectiveness with small and dense objects. A RetinaNet model with ResNet50 as its backbone was used as the object detector. This means that the ResNet50 network was used to extract features from the input image and these features were passed through the detection module to identify and classify objects. The RetinaNet model was trained on the COCO dataset which is a large-scale dataset for object detection.

C. *Captioning Language Model*

In the final stage of the model, a language model is developed to generate a caption based on the learned image features. The language model needs to identify objects in the image and understand their relationships, as well as incorporate language semantics and structures to produce coherent and grammatically correct sentences. The extended version of LSTM with attention-enrichment is a type of neural network that can process sequential data and incorporate attention mechanisms to focus on relevant parts of the input. This architecture has been shown to be effective in generating high-quality image captions.

IV. Evaluation Datasets And Results

Our evaluation dataset were randomly sampled images from the Flickr8k data set. The Flickr8k dataset is a widely used dataset for the task of image captioning which consists of 8,000 images paired with captions describing the content of the image. The goal of the evaluation was to see the facilitated caption generators performed better than unfacilitated generators. Roughly a total of 350 images were tested on and the results can be seen in Figures 2 and 3. The charts compare the facilitated and unfacilitated variations of caption generator models i.e the simplistic captioning model and the attention based captioning model.

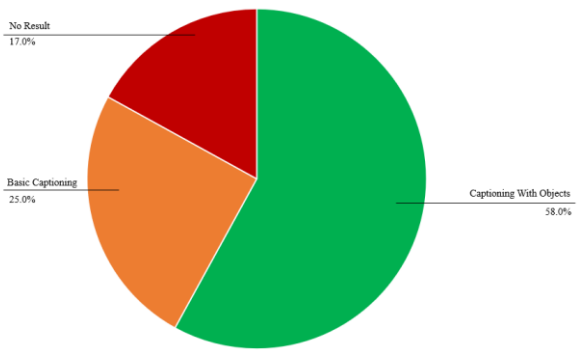

Figure 2: Unfacilitated Image Captioning Model

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

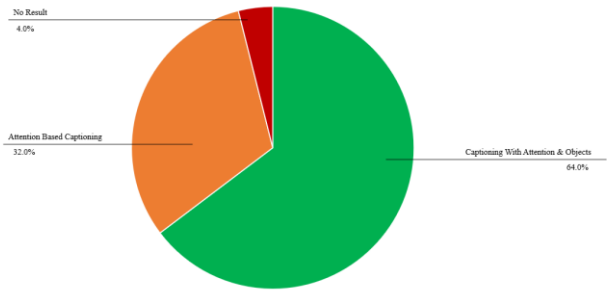

Figure 3: Facilitated Image Captioning Model

It can be seen from Figures 2 and 3 how facilitated caption generators gave superior performance to their unfacilitated caption generators. For the basic generator it could be seen that around 17% of the images in the evaluation set failed to provide good quality captions. It was observed that the facilitated captioner gave superior performance on around 64% of the images, thus it can be seen that the facilitated model was able to provide superior performance than unfacilitated model.

Black Dog Runs Through The Water

Figure 4: A Sample Image Caption Generation Using Facilitated Model

V. Conclusion

We presented the design and implementation of deep learning models for facilitated image captioning by considering object recognition and an enhanced attention mechanism to automatically generate image captions. Results of the different models were evaluated using a semantic similarity analysis between the generated captions and the actual ground truth captions. Our evaluation experiments demonstrates that facilitated image captioning provide superior performance to their unfacilitated models. The future experiments can be run with a various number of changes which include usage of much densely trained detectors, larger captioning datasets and different architectures for generating language models.

References

[1] Y. Wang, J. Xu, Y. Sun, and B. He, "Image captioning based on deep learning methods: A survey," arXiv preprint arXiv:1905.08110, pp. 1-7, 2019.

[2] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891–2903, 2013.

[3] Z-Q. Zhao, P.Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212-3232, 2019.

[4] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in European Conference on Computer Vision. Springer, 2018, pp. 711–727.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2, 2014, pp. 3104-3112.

[6] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, " A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR), vol.51*, no. 6, pp. 1–36, 2019

[7] Q. Wang and A. B. Chan, "CNN+ CNN: Convolutional decoders for image captioning," in 31st IEEE/CVF *Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018, pp. 1–9.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[9] J.Kleenankandy and A. N. K A, "An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies," *Information Processing & Management*, vol. 57, p. 102362, 2020.