Image Caption Generator

Navyasri Gangapuram[1], Hemanth Mittapally[2], G.Kadirvelu[3]
[1]*Department of Artificial Intelligence and Machine Learning, Sphoorthy Engineering College, Hyderabad, India*
[2,3]*Assistant professor of Artificial Intelligence and Machine Learning, Sphoorthy Engineering College, Hyderabad, India*

***Abstract*—In the era of digital imagery and social media, the demand for engaging image captions has grown exponentially. This research project explores the fusion of artificial intelligence and natural language processing techniques to automatically generate captivating and contextually relevant captions for images. We present a novel approach that leverages deep learning models and large-scale image-text datasets to train a caption generation system. Our methodology involves pre-training a neural network on a diverse corpus of text data and fine-tuning it on a curated dataset of images and corresponding captions. We evaluate the system's performance using both quantitative metrics, such as BLEU and METEOR, and qualitative assessments by human judges.**

The results demonstrate the effectiveness of our approach in generating captions that not only accurately describe the content of the images but also engage and resonate with human audiences. We discuss the implications of this research for content creators, marketers, and social media platforms, highlighting the potential for enhancing user experiences and automating caption generation at scale. This work represents a significant step towards harnessing AI-driven creativity in the realm of visual content and text generation, with promising applications in social media marketing, content recommendation systems, and accessibility for visually impaired user

Keywords—component, formatting, style, styling, insert.

I. INTRODUCTION

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 mainly focuses on developing semantic image caption generation techniques that leverage image and scene understanding. More particularly, we are interested in addressing image captioning by developing a mixture of object detection and deep learning models. Implementation of 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. Feature extraction is done first and then captions are generated. The flickr_8k dataset is used for training the model. The dataset which we are using contains 8000 images and each image is mapped with various captions. It also produces a semantic evaluation that makes use of sentence encoders to evaluate generated captions.

Deep learning methods became widely used in image captioning since it can generate novel captions by analysing the visual content of the image using an image model and generates the caption using a language model. Generating the novel image captions is a significant approach which solve the problems of using existing captions. 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 long periods of time. Image captioning has various applications such as Image tagging for e-commerce, photo sharing services , virtual assistants and online catalogs.

II. SYSTEM DESIGN

A. System Architecture

System Architecture is the process of designing the architecture, components, and interfaces for a system so that it meets the end-user requirements. The goal of system architecture is to allocate the requirements of a large system to hardware and software components.

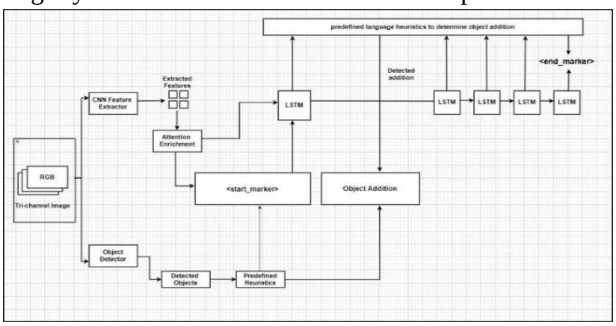

Fig 1. System Architecture

The training happens in a multi-staged manner. Different modules need to be trained for the distributed inference and training purposes in order to compile the final solution. The training can be divided into 3 stages and the specific architectures will be described accordingly.

B. 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 is ImageNet pretrained ResNet50 model. ResNet50 is a Convolutional neural network that is 50 layers deep. We can load a pretrained version of the network trained on more than a million images from the ImageNet datasets. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. Based on this, a pretrained model is satisfactory to provide preliminary results by libraries such as TensorFlow.

C. Object Detection

The second stage of the model requires the development of a detection module. RetinaNet is one of the best one-stage object detection models that has proven to work well with dense and small-scale objects. A RetinaNet model with ResNet50 as its backbone was used as our object detector. This model was trained on the COCO dataset for object detection. The object detector is responsible for detecting objects and detector is responsible for detecting classifying them into various classes.

D. Captioning Language Model:

The third and final stage of the model requires the development of a language model which is responsible for taking the learned image features and to generate a caption from said features. This model needs to identify objects in an image along with language semantics and structures and generate an output in human understandable description of the image content. Due to the variation of architectural nature and performance capabilities of various models, two models were chosen to demonstrate performance capabilities for image captioning. To extract the image features, a Convolutional Neural Network which is an extended version of Recurrent Neural Networks (LSTM) with attention-enrichment is adopted to generate the caption.

III. IMPLEMENTATION

In this section, we detail the implementation of the caption generation model using a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The chosen framework for this implementation is TensorFlow.

Black Dog Runs Through The Water

Fig 2. Image Captioning 1

A. Caption Generation Model Implementation

In this section, we detail the implementation of the caption generation model using a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The chosen framework for this implementation is TensorFlow. The dataset was pre-processed, split into training and testing sets, and used for model training. The training involved fine-tuning the pre-trained ResNet50 and training the LSTM model for caption generation.

Quantitative metrics, including BLEU and METEOR, were used for evaluation, and qualitative assessments were conducted by human judges.

1) Model Architecture

The neural network architecture consists of two main components.

a) Image Feature Extraction (CNNs)

We employed a pre-trained ResNet50 model as a feature extractor to obtain rich image representations.

b) Caption Generation (RNNs)

The caption generation model utilizing Long Short Memory (LSTM) networks for sequence generation. The model was trained on the flickr_8k dataset, where each image is mapped to various captions.

B. Pretrained Feature Extractor Implementation

The pre-trained feature extractor, ResNet50, was employed for image feature extraction. TensorFlow's

pre-trained ResNet50 model was utilized for this purpose. The pre-trained ResNet50 model was fine-tuned on the specific task of image captioning to adapt to the nuances of the dataset. The captioning language model was designed to take learned image features and generate captions. Two models were chosen for demonstration: a Convolutional Neural Network (CNN) and an extended version of Recurrent Neural Networks (LSTM) with attention-enrichment.

C. Object Detection Implementation

The object detection module was implemented using RetinaNet with ResNet50 as its backbone. TensorFlow Object Detection API was utilized for the development of this module. The RetinaNet model was trained on the COCO dataset for object detection. The trained model was then used for detecting and classifying objects in the images. Both models were trained and evaluated, and the final selection was based on their performance capabilities for image captioning.

Man Fishes In The Snow

Fig 3. Image Captioning 2

IV. EVALUATION

A. Datasets

A dataset is a structured collection of data generally associated with a unique body of work. Datasets play a pivotal role in enabling the development and evaluation of robust machine learning models and data-driven solutions. Datasets serve as the bedrock for data analysis and model training.

The Flickr8k dataset is a popular and widely used dataset in the field of computer vision and natural language processing. It is specifically designed for image captioning tasks, where the goal is to generate textual descriptions or captions for images. This dataset covers a wide range of scenes, objects, and concepts, making it diverse and suitable for training models that can generate descriptive captions for various visual content. Flickr8k is used for understanding the visual media that corresponds to a linguistic expression. We implement image captioning by considering detected objects from the image scene and integrate with attention mechanism for caption generation. We also produce a semantic evaluation that makes use of sentence encoders to evaluate generated captions.

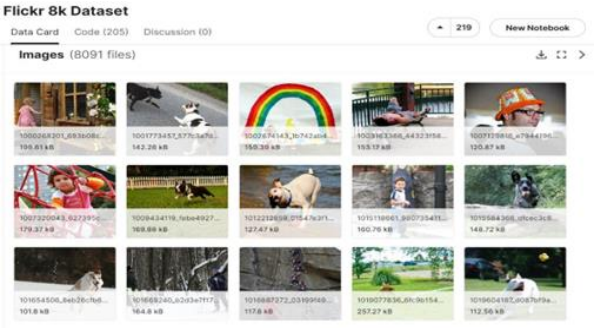

Fig 4. Flickr 8k Dataset

A wide range of training photographs could allow the image captioning version to operate for different types of images making the version more powerful. The Flickr8k datasets helps the model to train effectively without problems. The captions in the Flickr8k dataset are generated by human annotators, providing descriptions that are considered to be semantically and linguistically accurate. The combination of images and textual captions in the dataset also makes it suitable for multimodal research, where the goal is to understand the interactions between visual and textual modalities.

Fig 5. Flickr Dataset 2

B. Evaluation Metrics

An evaluation metric quantifies the performance of a predictive model. This typically involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, then comparing the predictions to the expected values in the holdout dataset. Evaluation metrics play a crucial role in assessing the performance of image captioning models and comparing their results. It is important to note that no single evaluation metric captures all aspects of image caption quality, and the choice of metric depends on the specific requirements and objectives of the task.

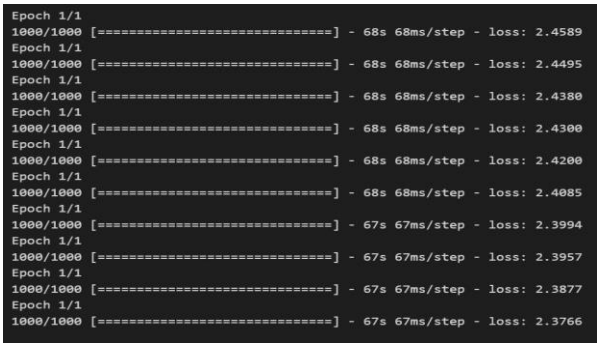

Fig 6. Evaluation Metrics

C. Results

Image captioning is an active area of research, and the quality of the generated captions can vary depending on the model architecture, training data, and other factors. Continued research and development are necessary to enhance the accuracy, coherence, and overall performance of image captioning systems. Our evaluation was upon randomly sampled images and it is observed that the facilitated image captioner gave superior performance on around 64% of the images. The Attention based generators are bit competitive in providing captions for 32% of the images. For only 4% of the images the facilitated generator failed to provide good quality captions.

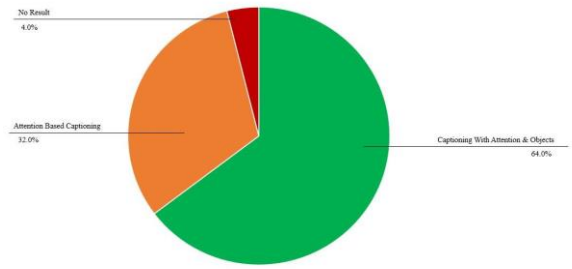

Fig 7. Pie Graph of Results

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. Leveraging techniques such as generative adversarial networks (GANs) or reinforcement learning can help train models with limited labeled data, making image captioning more accessible and scalable. 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. By combining advances in multimodal learning, natural language processing, and context understanding, researchers can push the boundaries of image captioning systems and deliver more immersive and meaningful experiences.

REFERENCES

[1] S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network" in ICET, Antalya, 2017.

[2] O. Vinyals, A. Toshev, S. Bengio and D. Erhan,"A Neural Image Caption Generator" CVPR 2015 Open Access Repository, vol. Xiv, 17 November 2014.

[3] J. Kleenankandy and A. N. K A, “An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies ” Information Processing & Management, vol. 57, 2020.

[4] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, "Neural Image Caption Generation with Weighted Training and Reference" Cognitive Computation, 08 August 2018.

[5] 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.

[6] 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.

[7] 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.