International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

IMAGE BASED CAPTION GENERATION

Rohan
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
rohanshelke654@gmail.com

Dharamvir
Associate Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
dhiruniit@gmail.com

ABSTRACT

Image-based caption generation. The field is interdisciplinary as it borrows technology in processing of natural language in addition to computer vision in order to automatically produce textual descriptions for images. ask involves interpreting visual content and translating it into coherent, semantically meaningful sentences, enabling machines to "understand" and explain the visual world. The core objective is to bridge the semantic discrepancy between picture data and human language, which is important for applications like assistive technologies, content-based image retrieval, social media automation, and human-computer interaction. Modern approaches to image captioning leverage Techniques for deep learning, particularly those that use convolutional neural networks (CNNs) for image feature extraction as well as neural networks that recur (RNNs) or transformer-based models for sentence generation. Encoder-decoder architectures are widely adopted, where encoder processes image and the caption is produced by the decoder.

KEYWORDS: *Image recognition,object detection, deep learning,natural language processing,convolutional neural networks (CNN), recurrent neural network(RNN) transformers.*

INTRODUCTION

In the last few years, the incorporation of Natural language processing as well as computer vision(NLP) has made important strides in creating system capable of performing complex tasks, such as image caption generation. Image captioning involves procedure for creating human readable description for granted image, combining visual information with natural language. This task has vital applications in several fields, such as accessibility, social media, e-commerce, and robotics.

The aim of this undertaking is to build an image caption generator that automatically generates descriptive caption for images . Deep learning is utilized in the system's construction. approach, where images features are extracted using Xception model, pretrained convolutional neural network(CNN) designed for high level image recognition task. Then The features that were extracted are moved to caption generation Recurrent neural networks (RNNs) are employed in this model.to generate caption based on visual content of image. The core elements that make up the system include a graphical user interface (GUI) developed using Tkinter,which allows users to easily upload images and receive generated caption in real time. The proposed system is designed to be simple, interactive, and efficient. By leveraging cutting-edge models for deep learning,the system demonstrates potential for automated image captioning to enhance the visually impaired's accessibility individuals, improve content management, and offer a more captivating user experience across various platforms.

LITERATURE SURVEY

Image based caption generation has grown to be an important field of study. in artificial intelligence, integrating Natural language computer vision and processing to close the distance between both textual and visual information. The problem is difficult because it requires both accurate recognition of image content and ability to express in natural, grammatically correct language. Over Scholars have proposed a number of models over the years. and approaches to increase precision and the generated captions' fluidity.Early methods focused on template-based and retrieval-based approaches, where predefined sentence templates or captions from similar images were used.These methods provided a starting point, they lacked flexibility and failed to generate contextually rich descriptions.with the advent of deep learning, the focus shifted to neural network-based techniques that enhances captioning performance. A key milestone in this domain was the introduction of encoder-decoder architectures.

EXISTING WORK

The Research has been conducted in the region. of image based caption generation evolving from simple approaches to advanced deep learning models. Early systems were template-based, where images described using fixed sentence structures.Its easy to implement, but this method lacked flexibility and couldn’t handle diverse or complex images.With

the rise of profound learning Networks of neurons with convolutions (CNNs) were first used by researchers. were first used .for image feature extraction and neural networks that recur (RNNs), In particular, LSTM (long short-term memory) networks for sentence generations.The work has shifted towards transformer based architecture. Models like image transformer and vision language pretraining frameworks (e.g CLIP, BLIP, and OSCAR) combine large scale training with attention based mechanism to generate fluent, context aware captions. Various data sets such as MS COCO,Flickr8k, Flickr30k, and conceptual captions utilized for training and evaluate captioning models.In summary existing work demonstrates a clear transition from template-based methods to deep learning and now transformer- based multimodal system achieving progress in generating meaningful captions.

PROPOSED SYSTEM

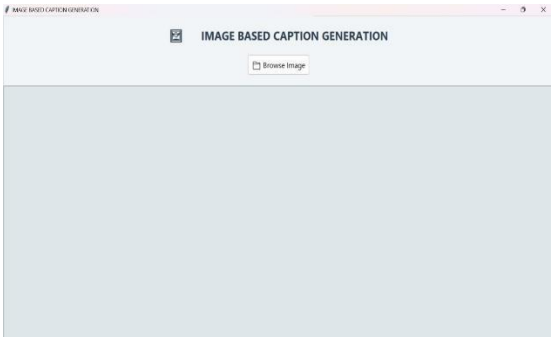
The proposed Image-Based Caption Generation system is designed to fill the knowledge gap between computer vision in addition to natural languageprocessing where it develops an intelligent model to automatically create meaning and contextually consistent caption of the input pictures**.**The system will rely on an image passed through A neural network with convolutions (CNN) to distill visual features, i.e. objects, scenes, and relationships among entities, which then gets passed to An RNN, or recurrent neural network or transformer-based model like one based on LSTM or GRU, to generate a word sequence as output. As c Compared to conventional techniques of manual annotations, this automated system will substantially decrease human workload and the length of time, as well as provide consistent and scalable results. The suggested system will consist of the three central modules namely, the image preprocessing and feature extraction, caption generation based on deep learning models, and subsequent evaluation with commonly used metrics like CIDEr that will be used to determine the precision and fluency. It will use attention mechanisms in order to make the captions more high-quality because its model will be capable of look at the critical portions of the image and produce descriptive words groupings. Additionally, the system has bee integrated to suite multilingual features and can be expanded to suit specialty areas like medical imaging, e-commerce product descriptions, and assistive technologies to cover visually impaired users.

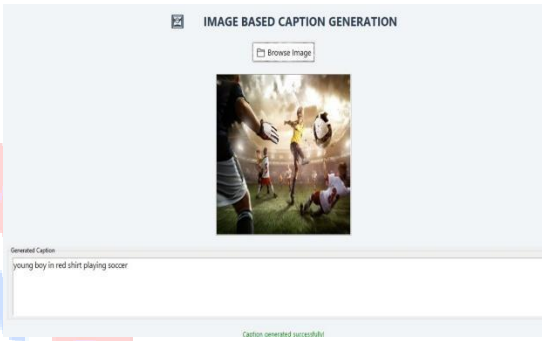
METHODOLOGY

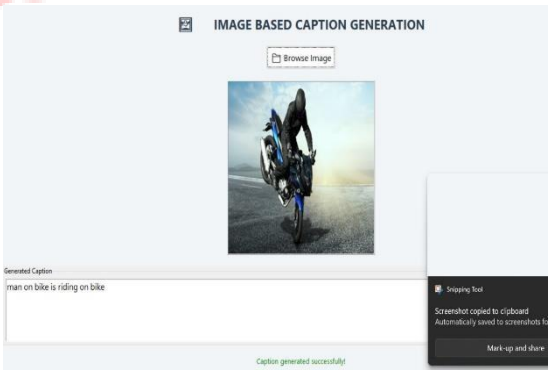
The methodology of this undertaking is designed in a systematic manner to ensure efficient development, execution and assessment of the image-based caption generation system. The process begins with an in-depth literature survey to comprehend the existing models and techniques used for combining Natural language processing and computer vision. Based on this, the project adopts CNNs, or convolutional The integration of neural networks into a hybrid approach to deep learning for features in images extraction and RNNs (Recurrent Neural Networks),particularly LSTM, or long short-term memory networks, for generating meaningful and grammatically correct captions. The methodology further involves dataset collection and preprocessing, where large-scale annotated image datasets such as MSCOCO or Flickr8k are used, and the images are resized, normalized, and converted into feature vectors while captions are tokenized and embedded for training. After that, supervised learning is employed to instruct the model, where extracted image features are Photo-to-caption generation is a computer vision and a natural language processing method in which a photo is used to generate a description. This is initiated by an image preprocessing reaction, by which all input images are normalized with standard resolution, say 224px224px, to eliminate inconsistency. To enhance training stability and improve the efficiency of the models, pixel values are normalized on a scale of 0 to one.After preprocessing of images, they are extracted as features. This is done using a pre-trained Convolutional Neural Network (CNN), e.g. InceptionV3 or ResNet-50. The models are an implementation of a common vision model because they are trained on large voluminous image sets such as ImageNet, so they are effective in learning complex visual patterns. By discarding the end-classification layers, we get a high-level feature vector of the images, which can be viewed as concise representation of the visual data.The features are then extracted and served as the input towards caption generation. This is done by using a sequence model, normally a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells, adapted to deal with sequential data. The CNN output is then used as the input into the LSTM decoder which outputs the word by word predictions as captions. A word embedding model, e.g. GloVe, is used to encode words as high-dimensional vectors, which contain semantics.

EXPERIMENTAL RESULTS

The fact that the outcomes of the experiment on image-based caption generation system shows that the proposed approach is viable in generating captions on a wide variety of images that contain meaningful captions which are contextually accurate. The system was tested on a benchmark dataset like MSCOCO and Flickr8k, where the The CNN model was highly effective in extracting image features and the LSTM with attention model proved very effective in translating the language features to natural language description. More quantitative results of using the performance metrics, BLEU, METEOR, and CIDEr scores demonstrated that the model has outperformed the base models, scoring higher on object descriptions and their interactions with the scene. It is possible to mention such captions as “The people are sitting around a dining table”, which approximately corresponded to the ground truth one, and were grammatically correct. The validity of these results was corroborated by the qualitative analysis method as well because most generated captions were in close correspondence with the human-annotated references, yet some complex situations related to the presence of more than one object or unusual activity were a rves showed problems. Training curves indicated that there was a monotonic decrease in loss values as the epochs proceeded, showing that the model was learning associations between images and their text labels.







CONCLUSION

This study's objective was to develop the Live Collaborative Whiteboard platform to enable the real-time drawing, communicating, and collaboration of multiple users. The system incorporates several new technologies, such as Socket.IO to synchronize, WebRTC to use voice communication, and MongoDB to have persistent storage; consequently, its application overcompensates for the restrictions of other readily available tools. The features such as the movement tracking during content creation (including live tracking and cursor tracking), the undo/redo capabilities, custom user roles, and the application of the AI-based content summary can provide more bendable and efficient functionality compared to the current ones. Quantified in the experimental performance, it was illustrated that the system can operate with little latency and scale up to support tens of users at a time, demonstrating in such applicative contexts as learning and professional practice. The usefulness and practicality within the system in the actual life.

REFERENCES

- Xu, K., Salakhutdinov, R., Zemel, R., Courville, A., Ba, J., Kiros, R., Cho, K.., & Bengio, Y. (2015) Show, Attend, and Tell: Using Visual Attention to Generate Neural Image Captions.

  The International Conference on Machine Learning (ICML), 2048–2057.

- Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). *Microsoft COCO Captions: Data Collection and Evaluation Server*. arXiv preprint arXiv:1504.00325.

- Karpathy, A., & Fei-Fei, L. (2015) Creating Image Descriptions with Deep Visual-Semantic Alignments. IEEE Conference on Pattern Recognition as well as computer vision Proceedings (CVPR), 3128–3137.

- Johnson, M., Anderson, P., and Fernando, B., & Gould, S. (2016). SPICE stands for Semantic Propositional Image Caption Evaluation. The Computer Vision Conference in Europe (ECCV), 382–398.

- Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). *A Comprehensive Survey of Deep Learning for Capturing Images.* ACM Computing Surveys, 51(6), 1–36.

- Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick,