ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 1, January 2024

Image Caption Generation using Deep Learning

[1] S. T. Patil, [2] Atharva Narote, [3] Nilesh Binnar, [4] Chaitanya Phad, [5] Payal Rathod

[1] Professor, Vishwakarma Institute of Technology, Pune, Maharashtra, India
[2] [3] [4] [5] Students, Vishwakarma Institute of Technology, Pune, Maharashtra, India
Corresponding Author Email: [1] patil.st@vit.edu, [2] atharv.narote21@vit.edu, [3] binner.nilesh@vit.edu,
[4] chaitaitanya.phad21@vit.edu, [5] payal.rathod21@vit.edu

Abstract— ***Image captioning, is a process of generating descriptions in natural language for images, has garnered substantial interest in the field of natural language processing and computer vision. This literature review examines VGG (Visual Geometry Group) and recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, to address the image captioning task. The integration of pre-trained CNNs, renowned for their prowess in extracting hierarchical features from images, with LSTM networks, capable of modelling sequential data and generating coherent textual descriptions, forms the crux of numerous state-of-the-art image captioning systems. In the following study, we investigate the application of the BLEU (Bilingual Evaluation Understudy) metric as a means to quantitatively assess the quality of captions generated by neural network-based image captioning models.***

Keywords: LSTM, VGG, RNN, Image, Caption Generation, BLEU.

I. INTRODUCTION

People communicate through language whether written or spoken. They frequently describe their visual environment using words like these. For those with physical disabilities, images and signs provide an additional means of understanding and communication. Although it is an extremely tough and complex task to automatically generate suitable phrases from photographs, it can greatly benefit visually impaired persons by helping them understand the descriptions of images on the internet. "Visualising a picture in the mind" is a common way to describe a good image. The analysis of current natural visual descriptions will be the first step towards accomplishing the challenging objectives of human recognition. Compared to picture classification and object identification, automatically creating captions and explaining the image is a significantly more difficult operation. An image's description needs to cover not just what is seen in the picture, but also how the items relate to each other and the behaviours and traits they display.

The connection between visual significance and descriptive elements extends into the realm of natural language processing (NLP), transitioning to the text summarization challenge. Text summarization aims to either select or create a concise summary for a given document. In the context of image captioning, the objective is to generate a descriptive caption that effectively encapsulates various features depicted in the image.

Generating captions for images poses a complex artificial intelligence challenge, requiring the synthesis of a textual description that accurately reflects the content of a given photograph. To achieve this, a combination of methodologies from computer vision and natural language processing is employed. The work involves transforming the visual information captured in an image into a coherent sequence of words, necessitating a cohesive integration of techniques from these two interdisciplinary fields. Novel solutions to this challenge have been developed recently using deep learning techniques. Cutting edge solutions for caption creation tasks have been shown using deep learning techniques. Remarkably, these approaches eliminate the need for intricate data preprocessing or a series of intricately crafted models in a pipeline. Instead, a singular end-to-end model can be formulated to forecast a caption based on a given photograph. The notable aspect of these methodologies lies in the simplicity of their implementation, obviating the necessity for complex preparatory steps or a sequence of specialized models.

The synthesis of deep learning architectures, leveraging convolutional neural networks (CNNs) like VGG (Visual Geometry Group) for image feature extraction and recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) for sequential text generation, has emerged as a pivotal paradigm for generating descriptive captions that encapsulate the content of visual data. This literature review aims to explore the multifaceted landscape of image captioning methodologies harnessing the capabilities of deep neural networks. Central to this exploration is the utilization of VGG models, renowned for their efficacy in extracting hierarchical visual features from images, serving as a foundation for encoding the salient information embedded within diverse visual scenes. The integration of VGG-based feature representations with LSTM networks forms a symbiotic relationship, enabling the generation of semantically meaningful and contextually relevant textual descriptions that correspond to the visual input.

II. LITURETURE SURVEY

In recent research, [1] innovative approaches to image captioning have been explored, combining Convolutional

Neural Networks (CNNs) for image encoding and Long Short-Term Memory (LSTM) networks for decoding. These models showcase promising methodologies for caption generation, utilizing substantial datasets and assessing accuracy through metrics like BLEU. Another study [2] focuses on enhancing image captioning accuracy by combining CNNs for image feature extraction and an extended LSTM with an attention mechanism for caption generation. This integration of object detection and attention proves effective in semantic caption generation, demonstrated through evaluations on COCO and Flickr30k datasets. In a different avenue [3], experiments leverage CNNs as encoders and Recurrent Neural Networks (RNNs) as decoders, showcasing the model's ability to generate cohesive captions. The incorporation of beam search further improves caption quality, as indicated by enhanced BLEU scores.

A unique approach [4] eliminates unnecessary motion features for improved image captioning across multiple datasets. This research underscores the significance of motion-CNN combined with object detection. Image captioning [5] is positioned as a crucial task in the realm of Artificial Intelligence (AI) and Cognitive Internet of Things (CIoT). The fusion of computer vision and NLP techniques is highlighted for automatic text generation from images, showcasing the evolution of deep learning-based methods. The [6] evolution of image captioning architectures is evident in the transition from conventional CNN-RNN models to more efficient transformer models, addressing sequential dependency issues through attention mechanisms. The study focuses on Hindi image captioning, presenting translated data from MSCOCO. An innovative approach [7] introduces an auxiliary classifier within a residual recurrent neural network for image captioning. This method, employing multi-layer LSTM with residual connections and an auxiliary classifier, outperforms existing approaches in generating accurate and contextually rich image captions.

A novel approach [8] named Hyperparameter Tuned Deep Learning for Automated Image Captioning (HPTDL-AIC) incorporates an encoder-decoder framework, showcasing superior performance through hyperparameter tuning. This suggests the potential of hybrid deep learning models in language modeling pursuits. Generative [9] Adversarial Networks (GANs) and CNNs are employed within an encoder-decoder architecture for image caption generation. The system's effectiveness is evaluated against existing models, achieving high accuracy rates for various datasets. A deep learning-based approach [10] employs Bidirectional LSTM (BiLSTM) for information retrieval, text summarization, and image captioning. Evaluation using GigaWord and DUC corpora demonstrates superior precision, recall, and F-scores compared to existing methods. A project [11] utilizing Generative Adversarial Network (GAN) principles introduces synthetic images for training a captioning model with an attention mechanism. The model's performance is evaluated through qualitative and quantitative analyses. Visual features [12] extracted from regions of interest (RoI) and the entire image enhance caption accuracy. This method, incorporating canonical correlation analysis, outperforms baseline gLSTM algorithms on the Flickr8k dataset. An attention-based [13] architecture using both Xception (CNN) and YOLOv4 (object detection) features, along with a novel "importance factor" for object features, shows significant improvement in CIDEr scores. This emphasizes the value of combining object detection features for better image captioning quality.

Research [14] investigates the impact of hyperparameter configurations on an encoder-decoder visual attention architecture for image captioning. Different convolutional architectures are compared, highlighting the superior performance of ResNext-101 and the efficiency of MobileNetV3. An examination [15] delves into the phenomenon of bias propagation within image captioning, particularly within the COCO dataset. This study explores racial and intersectional biases, emphasizing the growing impact of such biases in modern captioning systems. To address the [16] scarcity of Bengali image captioning studies, a strategy for generating Bengali captions from images is proposed. The model, evaluated on Flickr8k and BanglaLekha datasets, outperforms others, showcasing potential advancements in Bengali image captioning.

III. RESEARCH METHODOLOGY

As shown in Fig. 1. The process begins with the input of images, which undergo preprocessing to enhance their quality and usability. Simultaneously, a database of images is connected to a training set, preparing a foundation for the subsequent steps. Both the preprocessed images and the training set pass through a feature extractor, specifically VGG16, to extract meaningful features. These features, along with textual data prepared through a tokenizer generated during the preprocessing phase, are fed into a Long Short-Term Memory (LSTM) network.

The LSTM network, designed for sequence processing, collaborates with the feature extractor in a joint fashion. This integrated information then flows into a decoder, culminating in the generation of descriptive captions for the input images. This approach combines advanced image feature extraction with natural language processing, allowing the model to learn and generate coherent and contextually relevant captions for the given images.

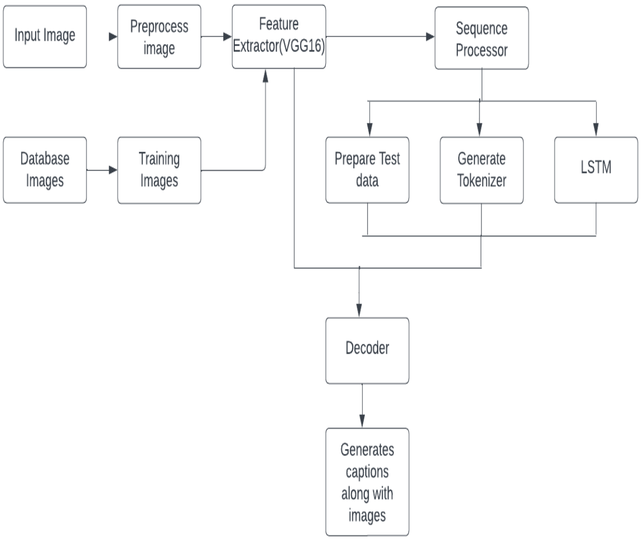


Fig. 1. Architecture of proposed system

A. Data Collection

The Flickr8k dataset holds significant prominence in both computer vision and natural language processing domains. Comprising images sourced from the Flickr website, each image in the dataset is accompanied by a set of five captions that elucidate the visual content captured in the photograph. The dataset is commonly employed for tasks related to image captioning, where the aim is to develop models that can automatically generate descriptive captions for images.

The dataset contains 8,000 images in total each accompanied by five distinct captions offering precise descriptions of prominent entities and events within the images, making it a valuable resource for training and evaluating algorithms that combine visual and textual understanding. Researchers and practitioners often use this dataset to benchmark the performance of image captioning models and explore the intersection of natural language and computer vision understanding.

B. Data Preprocessing

The dataset comprises diverse textual descriptions for each image, necessitating an initial cleaning process for the textual content. The first step involves loading the document containing all descriptions, wherein each photograph is linked to a distinct identifier found in both the filename and the accompanying description text file. Subsequently, a function named `load_descriptions()` is introduced to create a dictionary mapping photo identifiers to their respective descriptions. These descriptions undergo a cleaning process facilitated by the `clean_descriptions()` function, which involves converting all words to lowercase, removing punctuation, eliminating words with one character or less, and excluding words containing numbers. This text cleaning operation aims to streamline the vocabulary, promoting a more concise and efficient representation for subsequent model training.

Additionally, redundant tokens such as 's' or 'a' appended at the end of words were eliminated, fostering a refined and cohesive description corpus. The resultant cleaned descriptions, expressed as coherent strings of words, were meticulously structured and tailored to serve as an integral component for subsequent stages of model training and language modelling. The vocabulary, extracted from the cleaned descriptions, stands as a testament to the rich linguistic diversity encapsulated within the dataset, laying the groundwork for effective language comprehension and generation in image captioning models.

Following the cleaning process, the size of the vocabulary is assessed by transforming the cleaned descriptions into a set, providing insights into the dataset's linguistic richness. The final step involves saving the cleaned dataset, represented by the dictionary of descriptions and image identifiers, to a new file called 'descriptions.txt' using the `save_descriptions()` function. This systematic approach ensures that the dataset is preprocessed and ready for further analysis or utilization in natural language processing tasks.

C. Feature Extraction

VGG16 model has been used for the crucial task of feature extraction from the images. The VGG16, renowned for its depth and performance, is a convolutional neural network architecture that has proven effective in image classification tasks. Specifically, pre-trained weights of the VGG16 model are leveraged, which has been trained on large-scale image datasets. By using this established model, its ability to automatically learn and extract high-level features from input images is harnessed.

VGG16 model removes its classification layer, the penultimate layer's output, holding rich and abstract representations of the image content, is obtained. The preprocessed images are then fed into the model, allowing VGG16 to extract high-level features, encapsulating semantic and spatial information essential for understanding the visual content. As a result, a dictionary structure is created, associating each image's unique identifier with its extracted features. At last, the features are stored in ‘features.pkl’. Input layer taken is of (224,244,3) and we get Dense layer of (4096) neurons.

The utilization of the VGG16 model in workflow not only streamlines the feature extraction process but also ensures that model benefits from the wealth of knowledge encapsulated in the pre-trained VGG16 weights. This approach not only enhances the efficiency of our system but also allows to focus on the intricate task of natural language processing and caption generation, building upon the robust foundation provided by the VGG16 model's image feature representations.

D. LSTM

Subsequent to the feature extraction using the VGG16 model, the neural network architecture of LSTM has been used to unravel the sequential nuances of language within the textual descriptions. LSTMs are well-suited for this task due to their ability to learn and capture long-range dependencies in sequences, making them particularly effective for generating coherent and contextually rich captions. By feeding the pre-processed textual data into the LSTM, the model is enabled to learn the intricate patterns and relationships present in the descriptions, fostering a nuanced understanding of the linguistic context.

The LSTM's recurrent nature facilitates the learning of sequential dependencies, allowing our model to discern and predict the next word in a caption based on the preceding words. This dynamic capability empowers the caption generation process, enabling the model to produce contextually relevant and coherent captions that align with the visual information extracted by the VGG16 model. The synergy between VGG16's image feature extraction and LSTM's sequential learning culminates in a comprehensive framework that seamlessly integrates visual and textual information for the generation of meaningful image captions.

E. Decoder

In the Decoder model, the merging of vectors from the two input models, namely the feature extractor (which processes visual information) and the LSTM (which processes sequential textual information), is accomplished through an additional operation. This operation combines the rich visual features extracted from the images with the context captured by the LSTM from the textual data, creating a fused representation that captures both visual and semantic information.

Subsequently, this merged representation is fed into a Dense layer with 256 neurons. Dense layers are fully connected layers, and the 256 neurons in this layer serve as intermediate nodes to further process and transform the merged information. The Dense layer contributes to the model's ability to learn complex relationships and patterns in the data.

Finally, the output from the Dense layer is connected to another Dense layer responsible for making softmax predictions. The softmax function is applied to produce a probability distribution over the entire output vocabulary, representing the likelihood of each word in the vocabulary being the next word in the generated sequence. This final layer enables the model to generate a well-calibrated probability distribution, assisting in predicting the most probable next word in the sequence, given the merged features from both the visual and textual modalities. The use of softmax ensures that the model's output is a valid probability distribution, facilitating effective language generation.

F. Model Evaluation

This neural network architecture demonstrates a multimodal approach for processing text and image data in the context of a sequence-to-sequence model. The model employs two distinct input streams: one for textual data represented by sequences of length 34 (input_2), and another for image data (input_1) consisting of a high-dimensional vector of length 4096.

The text data undergoes an embedding layer (embedding_1) that transforms the input sequences into fixed-size vectors of length 256, preserving semantic relationships. Concurrently, the image data passes through a dropout layer (dropout_1) for regularization.

Subsequently, a series of dense layers (dense_1, dense_2, dense_3) and an LSTM layer (lstm_1) process the transformed text and image representations. The LSTM layer extracts sequential patterns from the text data, while the dense layers capture higher-level abstractions from both the image and text information.

An 'add' layer (add_1) integrates the learned features from the dense and LSTM pathways, aiding in information fusion. Finally, the network outputs a prediction via a dense layer (dense_3) with 7579 units, possibly used for classification or regression tasks.

IV. MATH

A. Model Evaluation

The softmax activation function is commonly used in neural networks for multi-class classification problems. It transforms a vector of real numbers into a probability distribution, where each element of the vector represents the likelihood of a particular class. The softmax function is particularly useful in the output layer of a neural network when you want to generate probabilities for multiple classes.

$$Softmax(y)_i = \frac{a^{(y)_i}}{\sum_{j=1}^{N} a^{y_i}}$$

Here:

y is the input vector (before applying the softmax function).

a is the base of the natural logarithm (Euler's number).

N is the number of elements in the input vector.

i represents each individual element in the vector.

B. Relu Activation Function

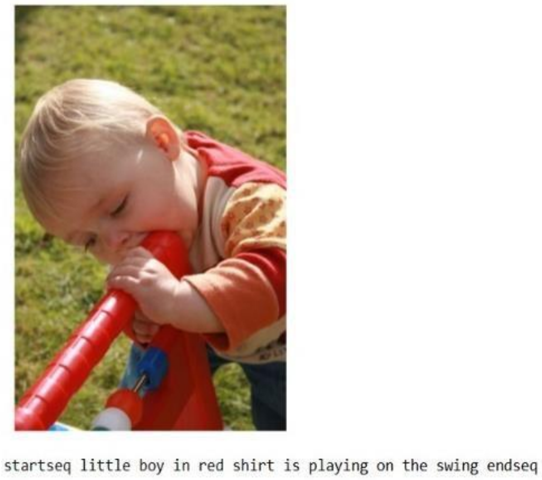
$$f(x) = max(0, x)$$

The rectified linear activation function, often known as ReLU, is a non-linear or piecewise linear function that, if the input is positive, outputs the input directly; if not, it outputs zero.

V. RESULTS

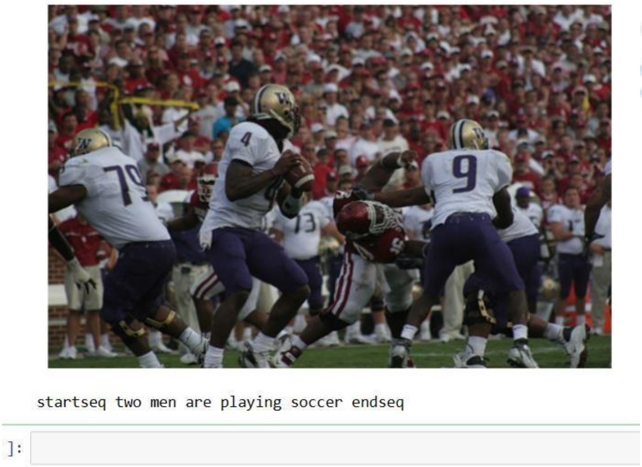
The BLEU evaluation results provide valuable insights into the performance of our caption generation model on the given dataset. With a dataset size of 6000, the training set comprises 6000 descriptions, contributing to a vocabulary size of 7579. This expansive vocabulary reflects the linguistic diversity captured during the model's training phase, indicating its exposure to a broad range of descriptive language elements.

In fig. 2 and 3, visual representations of the model's performance are presented through accompanying images, showcasing the generated captions alongside the corresponding photos. It is observed that the captions exhibit a moderate level of accuracy, with approximately 50% alignment with the content of the images. While not achieving perfect precision, these visualizations provide valuable insights into the model's capabilities and areas for potential improvement. The inclusion of these images enhances the transparency of the research findings, offering a tangible demonstration of the model's output and contributing to a comprehensive understanding of its performance.



```
startseq little boy in red shirt is playing on the swing endseq
```

Fig. 2. Image1 and generated caption



```
startseq two men are playing soccer endseq
```

```
]:
```

Fig. 3. Image 2and generated caption.

VI. FUTURE SCOPE

Most important scope will be running the program with zero loss of model because then only correct captioning can be done.For that good processors TPUs,GPUS will be needed.Also implementing Reinforcement learning will increase ability of less loss and high BLEU score.We may think of Real time image and video captioning.

VII. CONCLUSION

Employing a dataset of 6000 images with corresponding descriptions for training and 2000 images for testing, has yielded promising results as evidenced by the BLUE evaluation metrics. The model, trained on a vocabulary of 7579 words, demonstrates a commendable ability to generate captions that closely align with reference descriptions. The precision scores, ranging from BLEU-1 to BLEU-4, highlight the model's proficiency in capturing both unigram and bigram language patterns, contributing to linguistically coherent and contextually relevant captions.

While the project has achieved notable success, there exists potential for further refinement. The observed decrease in precision for higher order n-grams suggests an avenue for improvement, possibly through the exploration of more sophisticated language modeling techniques or the incorporation of additional context-aware features. Additionally, qualitative assessments and user feedback could complement the quantitative metrics, providing a more comprehensive evaluation of the model's performance in conveying the nuances of image content.

In essence, our caption generation project represents a significant stride in leveraging machine learning techniques for the synthesis of descriptive and meaningful textual content from visual stimuli. As we move forward, continuous iteration and fine-tuning will be essential to enhance the model's linguistic diversity and capture even more intricate semantic relationships, ultimately advancing the state-of-the-art method in image captioning applications.

REFERENCES

[1] Publication, IJRASET. “Image Captioning Using CNN and LSTM.” IJRASET, 2021. doi:10.22214/ijraset.2021.37846.

[2] Azhar, Imtinan, Imad Afyouni, and Ashraf Elnagar. "Facilitated deep learning models for image captioning." In 2021 55th Annual Conference on Information Sciences and Systems (CISS), pp. 1-6. IEEE, 2021.

[3] Arnav, Arnav, Hankyu Jang and Pulkit Maloo. “Image Captioning Using Deep Learning.” (2018).

[4] Iwamura, K.; Louhi Kasahara, J.Y.; Moro, A.; Yamashita, A.; Asama, H. Image Captioning Using Motion-CNN with Object Detection. Sensors 2021, 21, 1270. https://doi.org/10.3390/s21041270

[5] Jaiswal, Tarun. "Image captioning through cognitive IOT and machine-learning approaches." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.9 (2021): 333-351.

[6] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, Amit Kumar Singh, “Image captioning in Hindi language using transformer networks”, ELSEVIER,2021

[7] Çaylı, Özkan, Volkan Kılıç, Aytuğ Onan, and Wenwu Wang. "Auxiliary classifier based residual rnn for image captioning." In 2022 30th European Signal Processing Conference

(EUSIPCO), pp. 1126-1130. IEEE, 2022.

[8] Omri, M.; Abdel-Khalek, S. Khalil, E.M.; Bouslimi, J.; Joshi, G.P. Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning. Mathematics 2022,10, 288. https://doi.org/10.3390/math10030288

[9] Sargar, O. and Kinger, S., 2021, March. Image captioning methods and metrics. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 522-526). IEEE.

[10] P. MAHALAKSHMI AND N. SABIYATH FATIMA, “Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques”, IEEE,2022

[11] MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, HAMID LAGA, AND MOHAMMED BENNAMOUN,”Text to Image Synthesis for Improved Image Captioning”,IEEE,2021

[12] Jing Zhang, Kangkang Li, Zhenkun Wang, Xianwen Zhao, Zhe Wang,”Visual enhanced gLSTM for image captioning”,ELSEVIER,2021.

[13] Muhammad Abdelhadie Al-Malla, Assef Jafar and Nada Ghneim, “Image captioning model using attention and object features to mimic human image understanding”,Journal of Big Data,2018

[14] Castro, Roberto, et al. "Deep learning approaches based on transformer architectures for image captioning tasks." IEEE Access 10 (2022): 33679-33694.

[15] Zhao, Dora, Angelina Wang, and Olga Russakovsky. "Understanding and evaluating racial biases in image captioning." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14830-14840. 2021.

[16] Humaira, Mayeesha, Paul Shimul, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Faisal Muhammad Shah. "A hybridized deep learning method for Bengali image captioning." International Journal of Advanced Computer Science and Applications 12, no. 2 (2021).

IFERP connecting engineers... developing research