ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 5 | October 2024

www.ijirid.in

Deep Learning Based Image Caption Generation System

Prof. Suraj Dhole[1], Jyotirmaan Singh[2], Yogesh Chandak[3], Yamini Sapkal[4], Shrijay Matode[5]

[1]*Assistant Professor, G. H. Raisoni University, Amravati, (M.S.), India*

[2,3,4,5]*Undergraduate Students, G. H. Raisoni University, Amravati, (M.S.), India*

Abstract: *With the rapid advancements in deep learning, the ability to automatically generate descriptive captions for images has seen significant improvements. This paper presents an image captioning system that leverages the power of Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for sequence generation. The combination of these two architectures enables the system to interpret and describe the content of an image in natural language. The project utilizes the Flickr dataset, which provides a rich source of image-caption pairs, enabling the training of deep learning models for accurate image caption generation. The system first applies CNNs to extract visual features from the image, followed by LSTM networks to generate coherent and contextually relevant textual descriptions. Key tools in the implementation include Keras for model development, NumPy for data manipulation and Jupyter Notebooks for experimentation and analysis. This work provides an overview of the foundational principles behind image captioning, explores common methodologies used in the field, and discusses the challenges and potential of combining computer vision and natural language processing to enhance human-computer interaction.*

Keywords: Deep Learning, Image Captioning, Convolutional Neural Networks, Long Short-Term Memory, LSTM.

I. INTRODUCTION

In our daily lives, we come across countless images from diverse sources such as the internet, news articles, document illustrations, and advertisements. These images typically lack accompanying descriptions, leaving viewers to interpret their content independently. While humans are generally capable of understanding visual information without detailed captions, machines require structured descriptions to interpret and generate meaningful captions automatically. Image captioning plays a crucial role in addressing this challenge, offering benefits like faster, more accurate image searches and enhanced indexing when captions are available for online images.

The primary aim of this project is to develop a system capable of generating contextually relevant captions for a given image. These captions will aim to encapsulate the semantic and contextual information present in the images. Modern approaches rely on deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), or their derivatives, to perform this task. These models adopt an encoder-decoder architecture, where CNNs extract feature representations of the image, and RNNs generate descriptive text based on these features.

Image captioning has a wide range of applications, including e-commerce, biomedical imaging, web search optimization, and military operations. Additionally, platforms like Instagram and Facebook can

DOI 10.5281/zenodo.14252783

leverage automatic image captioning to enhance user engagement by providing descriptive captions for uploaded photos.

II. LITERATURE REVIEW

In the method proposed by Liu et al. [1], two deep learning models are explored for image captioning: CNN-RNN and CNN-CNN. The CNN-RNN model uses CNNs for encoding images into feature vectors, which are then passed to RNNs for decoding into captions, utilizing NLTK libraries. The CNN-CNN model, on the other hand, uses CNNs for both encoding and decoding, with a vocabulary dictionary to map image features to words for caption generation. The CNN-CNN model has faster training times but may suffer from higher loss, whereas the CNN-RNN model, though more time-consuming due to its sequential nature, results in lower loss and more accurate captions.

In the method proposed by Subrata Das, Lalit Jain et al. [2], a CNN-RNN framework is used for military image captioning, employing the Inception model for image encoding and LSTMs to address gradient descent issues during caption generation. Similarly, in the work of G. Geetha et al. [3], a CNN-LSTM model is utilized for image captioning, with CNNs serving as the encoder to extract image features and LSTMs as the decoder to generate descriptive captions. Both models integrate CNNs for feature extraction and LSTMs for language generation, enhancing the quality of the captions produced.

The method [4] incorporates CNN and RNN architectures for image captioning, enhanced with high-level attributes detected from the image. To capture semantic correlations between these attributes, a Multiple Instance Learning (MIL)-based model with Inter-Attributes Correlations (IAC) is proposed, enabling the prediction of attribute probability distributions. These attributes are then integrated into the LSTM to improve caption generation. By leveraging large-scale datasets like YFCC-100M, the model can learn more accurate attributes, leading to better captions. This approach also enables the generation of open-vocabulary and free-form sentences, enhancing caption diversity and relevance.

III. METHODOLOGY

This project requires a dataset comprising both images and their corresponding captions to train the image captioning model effectively.

1. Flickr 8k Dataset:

The Flickr 8k dataset serves as a widely used benchmark for image-to-text description tasks. It includes 8,091 images, each annotated with five captions, providing detailed descriptions of the objects and events depicted in the photographs. These images are sourced from various Flickr groups, encompassing a diverse range of scenes and events while deliberately excluding photos of well-known individuals or locations to maintain generality.

The dataset offers several characteristics that make it suitable for this project:

- **Multiple Captions Per Image:** Mapping multiple captions to a single image helps the model generalize better and reduces the risk of over fitting.

DOI 10.5281/zenodo.14252783

- **Diverse Image Collection:** The variety of training images ensures the model becomes robust and capable of handling a wide array of image types.

2. **Image Data Preparation:**
   Before training a deep learning model, it is essential to process images into a format suitable for feature extraction. This step transforms the raw images into meaningful representations for the model.
   Dataset Splitting: The dataset is divided into three subsets:
   - **Training Set:** 4,855 images
   - **Validation Set:** 1,618 images
   - **Testing Set:** 1,618 images

3. **Feature Extraction:**
   Feature extraction is performed using the Visual Geometry Group (VGG-16) model in conjunction with Convolutional Neural Networks (CNNs). The VGG-16 model, which won the 2015 ImageNet Large Scale Visual Recognition Challenge, is particularly well-suited for this project due to its superior performance in image categorization tasks.
   Key attributes of the VGG-16 model include:
   - **16 Weight Layers:** These deep layers enhance the model's ability to extract intricate features from images.
   - **Simplicity in Architecture:** It uses 3×3 convolutional layers, with max-pooling layers in between, to reduce the image volume size.
   - **Internal Representation:** The classification layer is removed, and the internal representation just before classification is used to produce a 4096-element vector for each image.

   For training, the input images must be resized to 224×224×3 dimensions, aligning with the VGG-16 input requirements. The extracted features form the foundation for building the image captioning model, ensuring effective image identification and caption generation.

4. **Caption Data Preparation:**
   The Flickr 8k dataset provides five captions for each photograph. For simplicity and to streamline the process, only one caption per image was used in this project. During the data preparation phase, each image ID is treated as a dictionary key, with its corresponding caption stored as the associated value.

5. **Data Cleaning:**
   To make the text dataset suitable for machine learning or deep learning models, raw text must be pre-processed into a usable format. This ensures the model can effectively learn from the data.
   The cleaning process involves:
   - **Removing Punctuation:** Eliminates symbols that may disrupt the model's understanding.

DOI 10.5281/zenodo.14252783

- **Eliminating Numbers:** Numbers are excluded to maintain textual consistency.
- **Removing Single-Character Words:** Filters out words that add little to no context.
- **Converting Text to Lowercase:** Standardizes the text for uniformity.

Unlike traditional pre-processing techniques, stop words are retained in the captions. This decision was made to preserve grammatical structure, which is critical for generating accurate and coherent captions. Removing stop words could compromise the readability and grammatical correctness of the generated descriptions, which are essential for the success of this project.

IV. IMPLEMENTATION

1. **Obtain Dataset:**

   For this project, I opted for the Flickr 8k dataset, which contains 8,091 images, each accompanied by five captions. The multiple captions for each image provide diverse perspectives, enhancing the model's generalization ability. The dataset was downloaded directly from the Flickr website.

2. **Load Data:**

   The dataset is organized by separating the file paths for the image files and caption files. The data is then accessed sequentially from the disk to prepare it for further processing.

3. **Prepare Photo Data:**

   To extract meaningful features from images, I utilized the pre-trained VGG16 Convolutional Neural Network (CNN) model. The classification layers of the VGG16 model were removed since the goal is not classification but feature representation. These extracted features were stored in a file named <feature.pkl>. VGG16 requires images to be resized to 224×224 pixels. The images were resized, converted to arrays, and reshaped accordingly. The resulting feature vectors have a size of 4,096 dimensions, which serve as the input representation of the images.

4. **Prepare Text Data:**

   The descriptions of all images were loaded and organized into a dictionary, mapping image names to their respective captions. The text data underwent cleaning, including:

   - Converting all text to lowercase.
   - Removing punctuation.
   - Eliminating single-character words.
   - Removing words containing numbers.

   After cleaning, a vocabulary of unique words from the captions was constructed to standardize the textual data.

5. **Encode Text Data:**

   A tokenizer was employed to map the vocabulary words to integers. This step created an encoded version of the text data. Additionally, the vocabulary size and the maximum length of descriptions (29 words) were computed, as these parameters would be essential during model training.

DOI 10.5281/zenodo.14252783

6. **Generate Output Text Data:**

The model requires both inputs and outputs to be structured for supervised learning. For training, we generated input and output datasets using the image-caption pairs.

- **Input (x1):** A 4,096-dimensional feature vector from the image.
- **Input (x2):** A sequence of words representing the caption.
- **Output (y):** The next word in the caption sequence.

7. **Define Model:**

Using Keras, the model was divided into two components:

- **Sequence Processor:** Handles textual input through an embedding layer, followed by an LSTM layer for sequential data processing.
- **Decoder:** Combines the outputs of the sequence processor and the image features, passing them through a dense layer to make predictions. The final dense layer's nodes correspond to the vocabulary size, enabling word prediction.

8. **Fit Model:**

The model, combining LSTM and VGG16, was trained on the dataset. During training:

- The process was monitored using training and validation loss.
- The model with the lowest validation loss was saved for later use, as training was computationally intensive.
- The training was run for five epochs to balance time and performance.

Visualization of the training and validation losses helped analyze how well the model learned over time.

9. **Generate Captions:**

Finally, captions were generated for images from the test set to evaluate the model's performance. While some captions contained errors, the model accurately captured many key aspects of the images, demonstrating its effectiveness in understanding visual content.

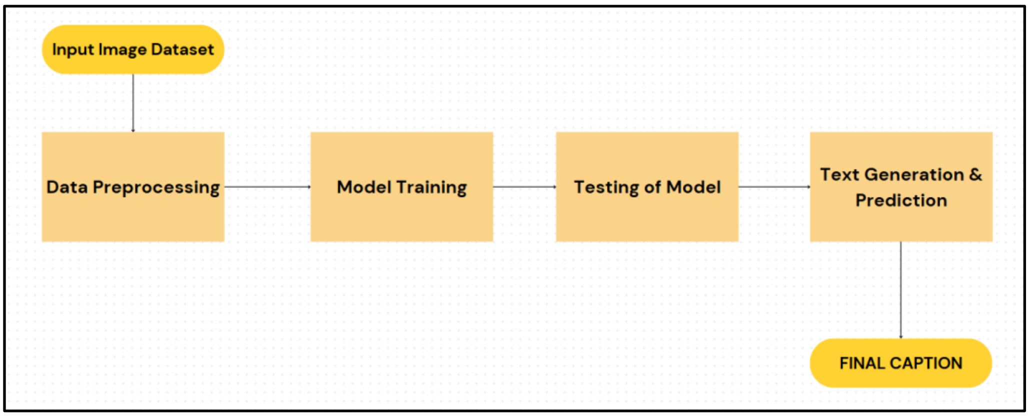
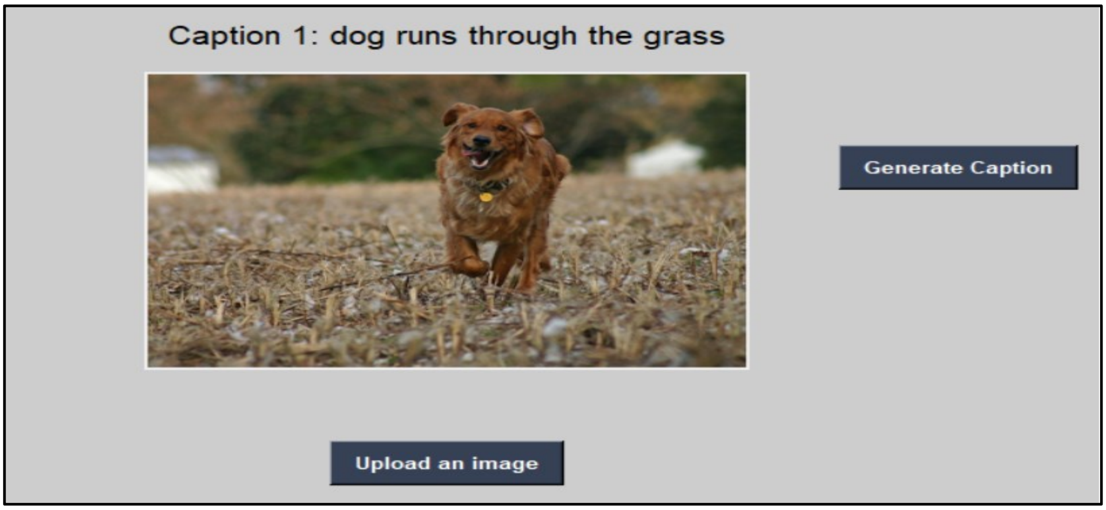


Figure 1: Implementation Diagram

DOI 10.5281/zenodo.14252783

V. RESULT

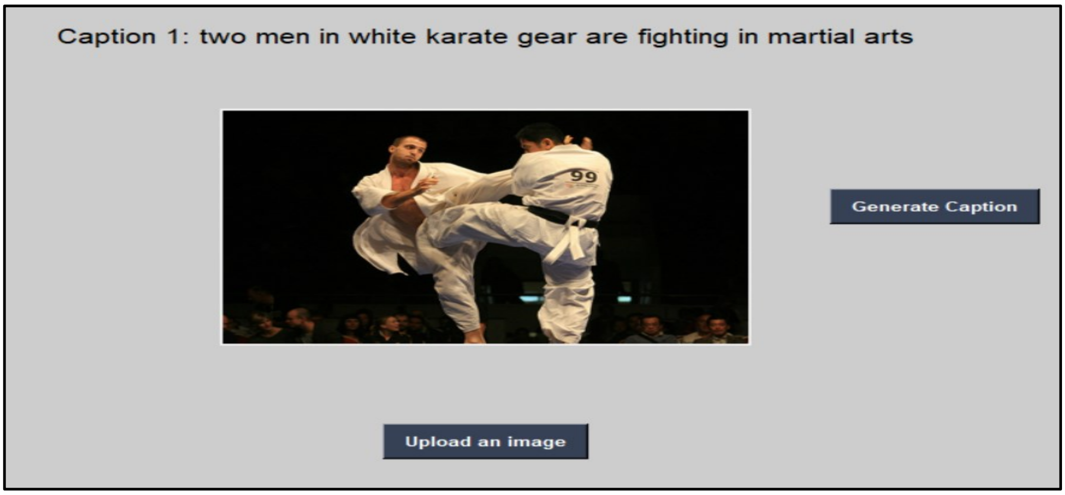
The image caption generator successfully produced a caption for the uploaded image, demonstrating the model's ability to interpret visual content and translate it into a descriptive textual output. In the example shown, the system generated the caption: "**dog runs through the grass**", accurately capturing the primary subject and its activity. The caption aligns well with the image content, highlighting the dog's action in a natural outdoor setting.



Caption 1: dog runs through the grass

Figure 2: Screenshot 1: Generated Caption - 'dog runs through the grass'

The image caption generator produced the caption: "**two men in white karate gear are fighting in martial arts**", which accurately describes the content of the image. The model effectively identified the key elements, including the presence of two individuals, their attire (karate uniforms), and their activity (martial arts combat). This demonstrates the model's capability to not only recognize objects but also understand interactions and contextual details.



Caption 1: two men in white karate gear are fighting in martial arts

Figure 3: Screenshot 2: Generated Caption - 'two men in white karate gear are fighting in martial arts

VI. CONCLUSION

To improve image captioning results, several enhancements can be made, such as using a larger dataset, performing extensive hyperparameter tuning (learning rate, batch size, layers, units, dropout rate, etc.), and incorporating cross-validation to prevent overfitting. Future advancements in image retrieval could leverage contextual image information, enhancing captioning accuracy. The project can be further improved by training with newer image captioning datasets to better identify underrepresented classes. Additionally, combining this approach with other image retrieval methods, like histograms or shape-based techniques, could potentially improve retrieval performance.

REFERENCES

[1] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052

[2] S. Das, L. Jain and A. Das, "Deep Learning for Military Image Captioning," 2018 21st International Conference on Information Fusion (FUSION), 2018, pp. 2165-2171, doi: 10.23919/ICIF.2018.8455321

[3] Geetha, T. Kirthigadevi, G GODWIN Ponsam, T. Karthik, M. Safa, “Image Captioning Using Deep Convolutional Neural Networks (CNNs)” Published under licence by IOP Publishing Ltdin Journal of Physics: Conference Series, Volume 1712, International Conference On Computational Physics in Emerging Technologies (ICCPET) 2020 August 2020, Manglore India, 2015

[4] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016).Image captioning with semantic attention. In Proceedings of the IEEE conference on computervision and pattern recognition (pp. 4651-4659).

DOI 10.5281/zenodo.14252783