INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, HYDERABAD

Project Report

Image Captioning

Submitted in partial fulfilment of the requirements of the degree of

Bachelor Of Technology in Modern Machine Learning

For the academic year 2024-2025

Submitted By:

Sakshi Warkari	20230015
Himanshi Hatwar	20230008
Prachi Dhore	20230004
Himanshu Katrojwar	20230002
Sakshi Jiwtode	20230006

Under the mentorship of : Amit Pandey Sir

G. H. RAISONI COLLEGE OF ENGINEERING, NAGPUR

Department Of Data Science, Iot , Cyber Security

The goal is to generate descriptive textual representations of images, which requires a deep understanding of the visual content. Traditional methods often relied on template-based approaches or manual feature extraction,which limited their flexibility and effectiveness. To address these limitations, recent advancements have adopted neural network architectures, particularly the CNN-RNN framework, which allows for end-to-end learning from images to captions.

Problem Statement:

With the rapid growth of digital images and multimedia content, there is a pressing need for efficient and accurate methods to generate descriptive textual annotations for images. Traditional manual captioning is time consuming, labor-intensive, and often inconsistent, which makes it impractical for large-scale datasets. The challenge is to develop an intelligent system that can automatically generate meaningful and contextually accurate captions for images. Such a system would greatly enhance accessibility, improve image search engines, and support visually impaired individuals in understanding visual content.The goal of this project is to create a machine learning model that can accurately interpret and describe the contents of images in natural language. The system should be capable of understanding various elements within
an image, such as objects, actions, relationships, and context, and then formulating coherent, human-like captions. The model will be trained on a large dataset of images with corresponding captions and will leverage advanced deep learning techniques, such as convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) or transformers for generating captions.

Real world applications:

Image captioning solves important real-world problems by producing descriptive text for images:

1. Accessibility for Visually Impaired Users:Provides text descriptions for images to improve accessibility for visually impaired people using screen readers.

2. Enhanced Image Search and Organization: Provides more accurate captions for image content, enhancing search and organisation on digital platforms.

3. Efficient Media Management: Automates captioning for large volumes of images, saving time and effort on content-heavy websites.

4. Improved Content Moderation:Creates image descriptions that help identify and filter inappropriate content.

5. Enhanced Ecommerce Experience: Provides detailed product descriptions, allowing customers to make informed purchasing decisions.

6. Autonomous Systems Support:Enables autonomous systems to interpret their surroundings by generating real-time descriptions.

7. Simplified Content Curation: Produces captions for large image collections, assisting with organisation and summarization task

Literature review:

Zhenghang Yuan et.al [1] proposed Multi-level attention module for spatial and scale features extraction and attribute encoder with GCN for learning effective attribute features. The framework enhances remote sensing image captioning using attention and graph convolution. It has achieved superior performance on UCM-captions, Sydney-captions, and RSICD dataset.
Jingqiang Chen et.al [2] proposed Multi-modal attentional mechanism proposed for news image captioning and RNN-based decoder for text and image encoding to generate words.Soheyla Amirian et.al [3] focused on encoder for object and feature extraction in CNN, decoder in neural network for generating natural language sentences and G-MLE and G-GAN strategies for image captioning.

Edy Mulyanto et.al [4] proposed CNN and LSTM models for image caption generation on different languages like English, Chinese, Arabic, and Japanese showing above-average results. It achieved BLEU-1 score of 50.0 and BLEU-3 of 23.9. Shiru Qu et.al [5] proposed LSTM with attention for image captioning achieves state-of-the-art performance. Attention mechanism enhances interpretability and alignment with human intuition. Model combines CNN features with LSTM for semantic sentence generation.

Feng Chen, Songxian Xie et.al [6] addressed Image Captioning challenges in neural network-based models. It uses Template-based methods with attributes, scenes, and objects. Minsi Wang et.al [7] proposed parallel-fusion RNN-LSTM architecture for image caption generation. It uses BLEU, Meteor,and perplexity metrics for sentence generation evaluation.

Xinru Wei et.al [8] focused on image retrieval by dense caption reasoning. It addresses complex image retrieval through structured language descriptions and compares with baseline methods on a large-scale CBIR dataset. Abu Musa Sakib et.al [9] proposed CNN and LSTM layers for image captiongeneration and feature extraction from images using pre trained CNN. Further model used text processing for captions, tokenization, and sequence conversion.

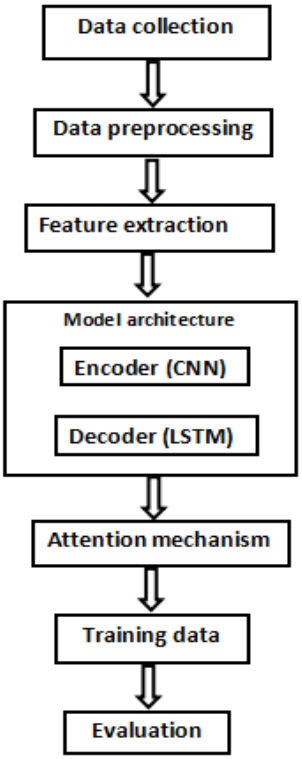
Genc Hoxha et.al [10] proposed CNN-RNN framework for remote sensing image captioning and Reviews existing methods in remote sensing image captioning. Binqiang Wang et.al [11] focused on image captioning using memory cells and topics. It uses Template-based methods

to fill predefined sentence templates with detected objects, Retrieval-based methods treat captioning as an image-sentence retrieval task and RNN based methods view sentence generation as a continuous sequence process

Varsha Kesavan et.al [12] proposed transfer learning approach with pre-trained VGG16 model encoder and attention mechanism, Hard and Soft attention for feature vector representation

Methodology:

This research aims to develop a robust image captioning model that combines convolutional and recurrent neural networks to generate descriptive captions for images. The process involves Data Collection, Data Preprocessing, Feature extraction, model design, and evaluation, as outlined below.



Dig. Flow of model

1. Data Collection

This research leverages the Flickr8k dataset, which contains 8,091 images, each paired with five distinct captions offering diverse descriptions of the image content. This multi-caption structure enhances the dataset's ability to capture various perspectives of a scene, making it an ideal resource for training and evaluating image captioning models. The dataset spans a broad spectrum of subjects, including people, animals, and landscapes, enabling the model to

generalize across different visual elements and contexts. Furthermore, the high-quality captions are consistently written, facilitating the model's learning of coherent, descriptive, and grammatically accurate caption generation.

2. Data Preprocessing

To prepare the dataset for training the image captioning model, several preprocessing steps were applied to the captions to improve data quality and learning efficiency. Each caption was standardized to lowercase and stripped of special characters, reducing vocabulary size and eliminating variations from capitalization and non-alphabetic characters. This cleaned data was then tokenized by appending unique start (startseq) and end (endseq) tokens to each caption, which clearly marks sequence boundaries for the model, enabling it to predict captions word-by-word and recognize when to terminate.

Additionally, a tokenizer was used to build a vocabulary from the training captions, generating a word-to-index mapping that translates each unique word into a numerical index. This mapping converts textual captions into sequences of numbers, facilitating effective processing and interpretation by the model, ultimately enhancing caption generation accuracy.

3. Feature Extraction

To facilitate effective image content analysis and feature extraction for caption generation, this study employs a pre-trained VGG16 model known for its depth and capacity to capture detailed visual information, including textures, shapes, and colors—qualities essential for complex tasks like image captioning. Utilizing weights pre-trained on the extensive ImageNet dataset, the VGG16 model is adapted for feature extraction by removing its final classification layer, enabling it to output a 4096-dimensional feature vector for each image. Each image is resized to 224x224 pixels to align with VGG16's input specifications, followed by standard normalization to ensure uniform input quality. These extracted feature vectors are stored in a dictionary keyed by image IDs, allowing efficient retrieval during model training and ensuring seamless integration with the caption generation process. This structured feature extraction approach provides a robust foundation for generating accurate, descriptive captions.

4. Model Architecture

The proposed image captioning model is structured as an encoder-decoder framework. The encoder is a modified VGG16 model, and the decoder is a bidirectional Long Short-Term Memory (LSTM) network that generates captions sequentially, word-by-word.

Encoder: The VGG16-based encoder extracts deep visual features from each image, outputting a 4096-dimensional feature vector. To prevent overfitting, a dropout layer with a

0.5 rate follows the feature extraction, and a dense layer further reduces the dimensionality to 256, optimizing compatibility with the decoder.

Decoder: The decoder begins with an embedding layer that maps vocabulary words into dense vector representations, facilitating the LSTM's understanding of syntactic and semantic relationships. The embedded words and encoded image features are processed by a bidirectional LSTM, which captures both forward and backward dependencies for improved coherence in generated captions. Finally, a dense layer with a softmax activation outputs a probability distribution over the vocabulary, enabling word-by-word caption prediction. This architecture ensures that the model generates descriptive and contextually accurate captions.

5. Attention Mechanism

The attention mechanism improves captioning accuracy by allowing the model to focus on relevant image regions. It computes attention scores through dot products between encoder outputs and decoder states, normalizes them via softmax, and uses the resulting weights to create a context vector, influencing the generation of each word.

6. Training and Evaluation

Training Details:
The model is trained to minimize categorical cross-entropy loss, using a data generator function to create batches of image-caption pairs. Each input consists of image features and partial captions, with the target being the next word. The model uses batch training (batch size = 32) with dropout (rate = 0.5) and the Adam optimizer for better convergence.

Evaluation Metrics:
Model performance is evaluated using the BLEU score, which compares n-grams between predicted and reference captions. BLEU-1 focuses on unigram matches to assess word-level accuracy, while BLEU-2 evaluates bigram matches to gauge short-phrase coherence. Qualitative analysis manually reviews selected captions for contextual relevance and accuracy.

RESULTS:

The development of the image captioning model yielded significant findings, demonstrating a capability to generate coherent and contextually appropriate captions for images. Leveraging the VGG16 architecture for feature extraction and an LSTM-based model enhanced with an attention mechanism, the following key outcomes were observed:

1. Model Performance Metrics:

Over 50 training epochs, the model exhibited a steady decrease in training loss, indicative of effective learning, culminating in a final loss of 0.4954. The validation loss, reaching 15.6088, reflected the model's generalization potential, albeit with room for improvement. BLEU scores were used to evaluate caption generation proficiency, revealing a **BLEU-1** score of 0.3799, which indicated good precision for single-word matches. The **BLEU-2** score of 0.1670 suggested a fair ability to create relevant two-word sequences. These metrics demonstrated the model's capability to generate captions with reasonable accuracy and highlighted areas for future enhancements.

2. Qualitative Analysis:

The model's predictions generally aligned with human-generated captions, showcasing its capability to understand and describe image content accurately. Examples from test images highlighted its strengths, with most captions reflecting the scene appropriately. However, occasional issues arose with repetitive or overly simplistic descriptions, indicating areas for further refinement. The inclusion of an attention mechanism played a crucial role in improving accuracy by focusing on key image features, thereby enhancing the quality and relevance of the generated captions.

Conclusion:

The integration of pre-trained VGG16 for feature extraction and a bidirectional LSTM model with attention for sequence generation successfully demonstrated the feasibility of automated image captioning. The model's ability to comprehend and narrate scenes represents a significant step toward building systems that bridge visual and textual data. While the model achieved respectable BLEU scores, future work could involve refining the model's architecture, using larger and more diverse datasets, and incorporating more sophisticated attention mechanisms to improve both precision and recall in caption generation.

Overall, this study underscores the potential of deep learning approaches in generating human-like descriptions from images, contributing to advancements in computer vision and natural language processing applications.