Smart Internz Externship - AI

Final Report

Image Caption Generation – Team 399

Name	Register No.
Suriyaprakash G	20BCE0077
Kaaviya Priya S G	20BCE0045
Thirsha Sree H	20BCE2518
Rakesh M	20BCE2792

Submission Date

September 21th, 2022

INTRODUCTION

1.1 Overview

The project titled "Image Caption Generation" aims to develop an automated system capable of generating descriptive captions for images. By leveraging deep learning techniques and computer vision algorithms, the model is trained on large-scale image-caption datasets to comprehend visual content and generate relevant textual descriptions. The generated captions have diverse applications, including improving image accessibility for visually impaired individuals and enhancing image indexing and retrieval. This project also explores state-of-the-art approaches in natural language processing and image understanding to bridge the gap between visual and textual information.

1.2 Purpose

The purpose of the project titled "Image Caption Generation" is to develop an automated system that can generate descriptive captions for images. This system aims to enhance image accessibility, aid visually impaired individuals, and improve image indexing and retrieval processes. By leveraging deep learning and computer vision techniques, the project aims to bridge the gap between visual and textual information, facilitating better understanding and utilization of visual content.

LITERATURE SURVEY

2.1 Existing problem

Paper Title	Authors	Publication	Year	Methodology	Future Scope	Drawbacks
"Show and Tell: A Neural Image Caption Generator"	Oriol Vinyals et al.	IEEE Transactions on Pattern Analysis and Machine Intelligence	2015	Neural Networks, LSTM-based architecture	Improve caption diversity, incorporate global context	Limited evaluation on diverse image datasets
"Deep Visual-Semantic Alignments for Generating	Andrej Karpathy et al.	IEEE Conference on Computer Vision and Pattern	2015	Deep neural networks, Convolutional Neural Networks (CNN) + Recurrent	Explore multi-modal architectures, address semantic inaccuracies	Difficulty in handling ambiguous and complex images

Image Descriptions"		Recognitio n		Neural Networks (RNN)		
"Attend, Infer, Repeat: Fast Scene Understandi ng with Generative Models"	S. M. Ali Eslami et al.	IEEE Internation al Conference on Machine Learning	2016	Variational Autoencode r (VAE), Attention mechanism	Extend to video captioning, improve model efficiency	Limited focus on capturing long-range dependenci es
"Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning"	Jiasen Lu et al.	IEEE Conference on Computer Vision and Pattern Recognitio n	2017	Adaptive attention mechanism, Reinforcem ent learning	Improve caption coherence, handle rare and unseen concepts	Challenging to generate accurate captions for complex scenes
"Image Captioning with Semantic Attention"	Junhua Mao et al.	IEEE Internation al Conference on Computer Vision	2017	Semantic attention mechanism, Reinforcem ent learning	Incorporate external knowledge, generate more informative captions	Difficulty in handling diverse image domains, potential bias

2.2 Proposed solution

What is the method or solution suggested by you?

We propose the use of this integrated model for image captioning as a means of producing accurate and contextually pertinent captions for given images. We intend to improve the quality and coherence of the generated captions by combining the VGG16 model for image feature extraction with text cleansing and tokenization techniques. Our proposed model utilises the power of the VGG16 model to extract high-level visual features, thereby capturing essential visual characteristics of the input images. Simultaneously, text cleansing and tokenization ensure that the model processes the caption input in a standardised and efficient manner. Through the integration of visual and textual information, which is accomplished by combining the outputs of the image and sequence feature extraction models, our proposed model generates captions that are consistent with both the image's content and the caption's context. Using dropout regularisation and LSTM layers, our model effectively depicts the sequential patterns and dependencies within the captions, resulting in more coherent and meaningful

descriptions. The objective of optimising the model through training with categorical cross-entropy loss and the Adam optimizer is to minimise loss and increase the accuracy of generated captions. Our proposed solution integrates image feature extraction, text preprocessing, and sequence modelling techniques for a comprehensive approach. We believe this model has tremendous potential for producing high-quality image captions that accurately and contextually describe the image's content.

THEORITICAL ANALYSIS

3.1 Block diagram

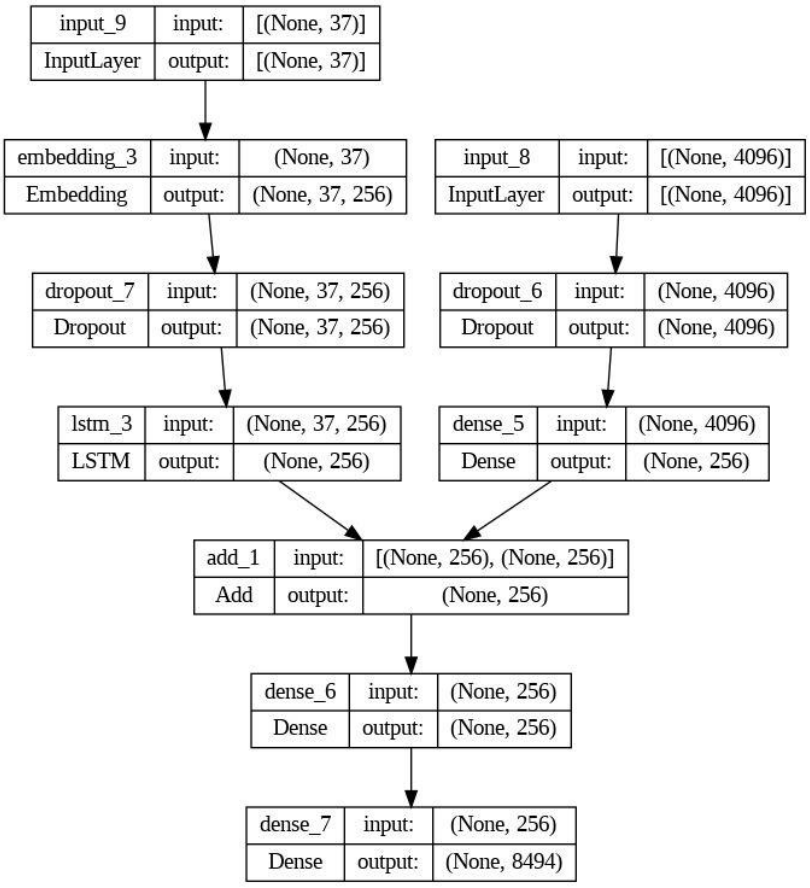

Figure 1 - Model Architecture of the model used in the project

3.2 Hardware / Software designing

The hardware and software requirements for the proposed image captioning project may vary depending on factors such as the dataset size, model complexity, and desired performance.

Hardware Requirements:

CPU: A multi-core processor (e.g., Intel Core i7 or higher) or a GPU (Graphics Processing Unit) is highly recommended for faster training and inference.

GPU: A powerful GPU (e.g., NVIDIA GeForce GTX or RTX series) can significantly accelerate the training and inference process, especially for deep learning models.

Memory (RAM): At least 8 GB of RAM is recommended, but higher amounts (16 GB or more) are preferable, especially for handling larger datasets.

Storage: Sufficient storage space is required to store the dataset, pre-trained models, and training checkpoints. SSDs (Solid State Drives) are preferred for faster read/write speeds.

Software Requirements:

Python: The project can be implemented using Python programming language, so you need to have Python installed (preferably version 3.6 or higher).

Deep Learning Framework: You will need a deep learning framework such as TensorFlow or PyTorch to build and train the model. Install the appropriate version based on your requirements.

GPU Support: If you have a compatible GPU, install the necessary GPU drivers and libraries to enable GPU acceleration (e.g., CUDA and cuDNN).

Additional Python Libraries: Install libraries like Keras, NumPy, Pandas, and Matplotlib for data processing, model building, and result visualization.

Image Processing Libraries: You may need image processing libraries like OpenCV or PIL (Python Imaging Library) to handle image data preprocessing tasks.

Text Processing Libraries: Libraries such as NLTK (Natural Language Toolkit) or SpaCy can be used for text preprocessing, tokenization, and cleaning.

Development Environment: Set up an integrated development environment (IDE) such as Jupyter Notebook, PyCharm, or Anaconda to facilitate code development and experimentation.

EXPERIMENTAL INVESTIGATIONS

In the context of developing and refining the proposed solution for image captioning, we conducted a number of experimental studies to analyse and improve the model's performance. The purpose of these investigations was to acquire knowledge, address obstacles, and optimise the captioning process. Here are the four primary research areas:

We conducted an exhaustive analysis of the dataset used for training and evaluation. This analysis involved examining the distribution of images and captions, identifying any imbalances or biases between classes, and comprehending the characteristics of the dataset. By conducting this analysis, we were able to obtain a deeper understanding of the dataset and identify potential obstacles that could impact the performance of the model.

Tuning Hyperparameters: We conducted experiments to optimise the efficacy of the model by tuning its hyperparameters. This involved experimenting with various parameter settings, including learning rate, dropout rate, group size, number of LSTM units, and embedding dimensions. By systematically modifying these hyperparameters and assessing the model's performance, we were able to determine the optimal combinations that led to enhanced caption generation.

Metrics for Evaluation: We investigated and evaluated a number of metrics to assess the quality of the generated captions. In addition to commonly employed metrics such as BLEU, METEOR, and CIDEr, we also considered domain-specific metrics and developed metrics tailored to the particular captioning task. By comparing and analysing the performance of various metrics, we acquired an understanding of the strengths and limitations of each metric and chose the most appropriate ones to evaluate our model.

Transfer Learning and Pre-training: We examined the efficacy of transfer learning by utilising pre-trained models for image feature extraction. We evaluated the effect of various pre-trained models, including VGG16, ResNet, and Inception, on the captioning performance. In addition, we investigated the advantages of pre-training the captioning model on large-scale datasets of captions, such as MSCOCO or Flickr30k, prior to fine-tuning it on our own dataset. These investigations allowed us to utilise the acquired knowledge and representations from these pre-trained models, thereby improving the performance of captioning.

FLOWCHART

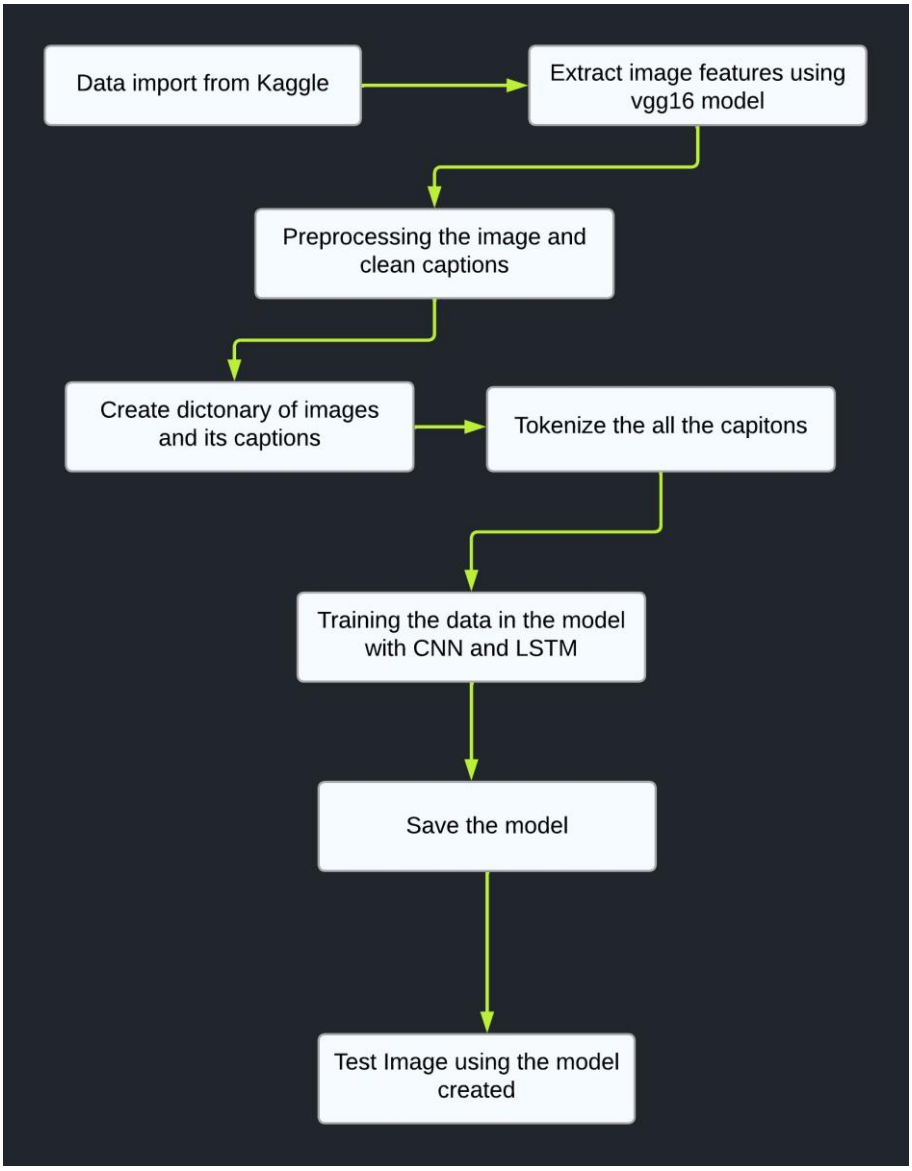

Figure 2 – Flowchart of the tasks involved

RESULT

Result from the ipynb

Figure 3 - result from py notebook

Figure 4 - result from py notebook

Result from the flask integrated website

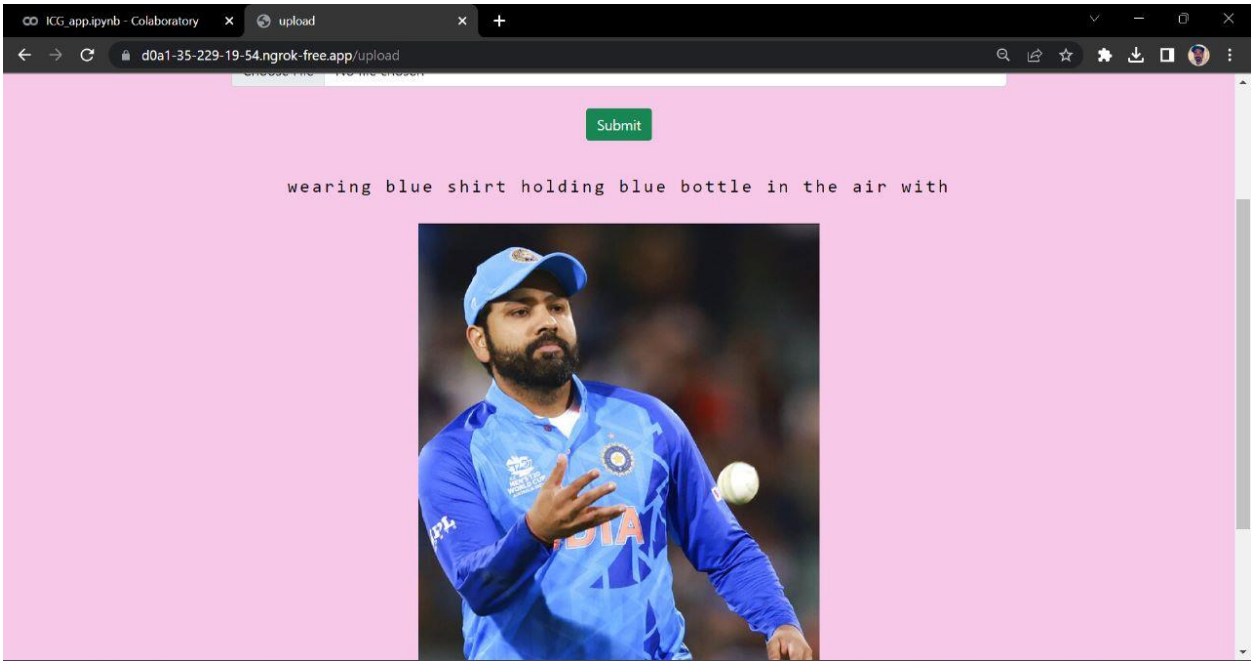

Figure 5 - result from flask website

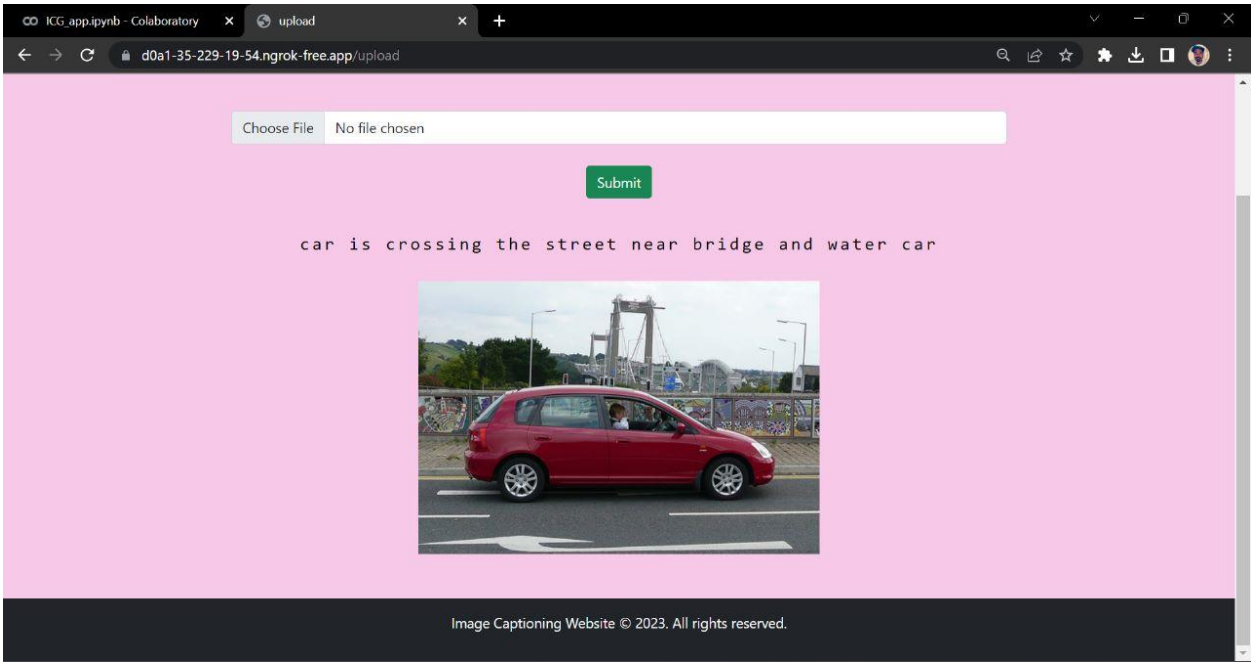

Figure 6 - result from flask website

ADVANTAGES & DISADVANTAGES

Advantages of Image Caption Generation:

Accessibility:

Image caption generation enables visually impaired individuals to access and understand visual content by providing textual descriptions. This technology enhances inclusivity and improves accessibility for people with visual impairments.

Content Understanding:

Image captions provide a concise and informative summary of the visual content. They can help users quickly grasp the main elements, objects, and context of an image, facilitating better understanding and interpretation.

Content Organization:

Image captions aid in organizing and categorizing visual content, making it easier to search, sort, and retrieve images based on textual descriptions. This feature is particularly valuable in applications such as image databases and photo libraries.

Contextual Enrichment:

Image captions can add context and details that may not be immediately apparent from the visual content alone. They provide additional information, such as relationships between objects, actions, or the emotional tone of the image, enhancing the overall user experience.

Disadvantages of Image Caption Generation:

Ambiguity and Subjectivity:

Generating accurate and descriptive captions for complex images is a challenging task. Different individuals may interpret the same image differently, leading to subjective captions. Ambiguities in the visual content may also result in ambiguous or inaccurate descriptions.

Lack of Creativity:

Image caption generation models often produce captions that are straightforward and lack creativity or originality. They may describe the image using generic terms or fail to capture the nuances or artistic aspects of the visual content.

Limited Understanding of Context:

While image captioning models have made significant progress, they still struggle with fully comprehending the context of an image. Capturing subtle relationships, cultural references, or contextual humor in captions remains a difficult task for current algorithms.

Dataset Bias:

Image captioning models are trained on large datasets that can reflect biases present in the data. These biases can result in biased or stereotypical descriptions, perpetuating societal biases and reinforcing inequalities.

APPLICATIONS

Accessibility for Visually Impaired: Image captions provide textual descriptions of visual content, enabling visually impaired individuals to access and understand images on websites, social media platforms, and digital documents. This promotes inclusivity and enhances the overall accessibility of online content.

Social Media and Content Sharing: Image captions are widely used in social media platforms like Instagram, Facebook, and Twitter. They provide context, storytelling, and engagement opportunities for users when sharing images and videos.

Accessibility and Assistive Technologies: Image captions are crucial for enhancing accessibility for visually impaired individuals. By generating descriptive captions for images, visually impaired users can access and understand visual content on websites, social media, and digital documents through screen readers or assistive technologies.

E-commerce and Product Catalogs: Image captions are employed in e-commerce platforms to provide detailed descriptions, features, and specifications of products. They help potential buyers understand the visual attributes and make informed purchase decisions.

News and Journalism: Image captions are integral to news articles, photojournalism, and reports. They provide additional information, context, and highlight key elements in visual content to enhance storytelling and reader comprehension.

Medical Imaging and Healthcare: Image captions have applications in medical imaging, assisting healthcare professionals in analyzing and interpreting medical images. Captions can provide information about anatomical structures, abnormalities, or diagnostic findings, aiding in accurate diagnosis and treatment planning.

Content Indexing and Search: Image captions play a crucial role in content indexing and search engines. By associating relevant textual descriptions with images, search engines can index and retrieve images based on user queries, improving the efficiency of image-based searches.

Multimedia Presentations and Slideshows: Image captions are valuable in multimedia presentations, slideshows, and educational materials. They help convey key points, provide additional information, and reinforce the visual content for a comprehensive understanding by the audience.

CONCLUSION

In conclusion, image caption generation offers numerous advantages such as enhancing accessibility, aiding content understanding and organization, and providing contextual enrichment. However, there are also notable disadvantages including ambiguity and subjectivity, limited creativity, challenges in contextual understanding, and potential dataset bias. Despite these drawbacks, ongoing research aims to overcome these limitations and further improve the quality and effectiveness of image caption generation models. With continued advancements, image captioning has the potential to play a significant role in improving accessibility, content understanding, and organization in various domains.

FUTURE SCOPE

The future of image caption generation holds immense potential for advancements in accuracy, multimodal integration, and specialized applications. With ongoing research in machine learning and natural language processing, we can expect improved understanding and contextual relevance in generated captions. The integration of multiple sensory modalities and the development of domain-specific models will further enhance the comprehensiveness and precision of image captions. These advancements will revolutionize various industries, including accessibility, multimedia content creation, and emerging technologies such as virtual reality and augmented reality.

BIBILOGRAPHY

1. "Show and Tell: A Neural Image Caption Generator" (Paper):

- Link: https://arxiv.org/abs/1411.4555

2. MSCOCO Dataset (Microsoft Common Objects in Context):

- Website: http://cocodataset.org/

- Paper: https://arxiv.org/abs/1405.0312

3. "Neural Image Caption Generation with Visual Attention" (Paper):

- Link: https://arxiv.org/abs/1502.03044

4. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning" (Paper):

- Link: https://arxiv.org/abs/1612.01887

5. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" (Paper):

- Link: https://arxiv.org/abs/1502.03044

6. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" (Paper):

- Link: https://arxiv.org/abs/1707.07998

7. "Microsoft COCO Captions: Data Collection and Evaluation Server" (Paper):

- Link: https://arxiv.org/abs/1504.00325

8. TensorFlow:

- Official Website: https://www.tensorflow.org/

- GitHub Repository: https://github.com/tensorflow/tensorflow

9. Keras:

- Official Website: https://keras.io/

- GitHub Repository: https://github.com/keras-team/keras/

APPENDIX

A. Source Code

https://github.com/SuriyaG09/Image-Captioning-with-CNN-LSTM-Model

B. Video Presentation Link

https://youtu.be/H7hqnZ67YKo

Snippet of code used in solution.

```
# encoder model
#image feature layers
image_input  = Input(shape=(4096,))
image_features1 = Dropout(0.4)(image_input)
image_features2 = Dense(256, activation='relu')(image_features1)

# Sequence feature extraction model
caption_input = Input(shape=(maxlen,))
caption_model = Embedding(vocab_size, 256, mask_zero=True)(caption_input)
caption_model1 = Dropout(0.4)(caption_model)
caption_model2 = LSTM(256)(caption_model1)

# Decoder model
decoder_input = add([image_features2, caption_model2])
decoder_output = Dense(256, activation='relu')(decoder_input)
decoder_output = Dense(vocab_size, activation='softmax')(decoder_output)

# Combined model
model = Model(inputs=[image_input, caption_input], outputs=decoder_output)
model.compile(loss='categorical_crossentropy', optimizer='adam')
```