SHOW AND TELL:
AN IMAGE CAPTION GENERATOR BUILT USING LSTM

PROJECT REPORT

DONE BY:
SAI PRANEETH PANDIRI
NITHIN GUNTUKU
PRANAV TAVVA
SRUJAN ALLAMPALLY

1. Introduction

The ability to understand and describe visual content has long been a fundamental challenge in the field of computer vision. Image captioning, the task of generating a textual description that accurately captures the content of an image, has gained significant attention in recent years due to its potential applications in areas such as image understanding, content retrieval, and accessibility for visually impaired individuals.

The purpose of this project is to develop an image caption generator using Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN) known for its ability to model sequential data. The model takes an input image and generates a descriptive caption that conveys the visual information present in the image. By combining the power of deep learning with natural language processing, we aim to bridge the gap between images and textual descriptions, enabling machines to comprehend and communicate the content of visual data.

The motivation behind this project lies in the growing demand for automated image understanding and the need to enhance human-computer interaction through natural language interfaces. With the proliferation of image-centric platforms and the vast amount of visual content available, an accurate and efficient image captioning system can provide valuable insights, facilitate information retrieval, and improve accessibility to visual content for diverse user groups.

Throughout this project, we will leverage the capabilities of the VGG16 model for image feature extraction and employ LSTM networks for sequence modeling and caption generation. The model will be trained on a dataset of images paired with corresponding textual descriptions. We will evaluate the performance of the model using metrics such as BLEU (Bilingual Evaluation Understudy) scores, which assess the similarity between generated captions and human-written references.

By developing an effective image caption generator, we aim to contribute to the advancement of computer vision and natural language processing, opening up new possibilities for understanding and interacting with visual content. This project report presents a comprehensive overview of the methodology, dataset, model architecture, training, evaluation, and deployment process, along with the results and discussions.

2. Problem Statement

The problem we aim to address in this project is the generation of accurate and meaningful captions for images. While humans can effortlessly describe the content of an image, teaching machines to do the same is a complex task. The challenges associated with image captioning include understanding the visual context, capturing relevant details, maintaining coherence and relevance in the generated text, and dealing with the inherent ambiguity and subjectivity of language.

The goal is to develop a model that can accurately perceive the visual information in an image and generate captions that not only describe the objects and scenes but also convey a deeper understanding of the content. The model should be able to capture relationships between objects, recognize actions or events, and incorporate contextual knowledge to produce coherent and contextually relevant descriptions.

3. Project Overview

This project follows a comprehensive workflow to build an image caption generator. It involves several key components:

Dataset: We utilize the Flickr8K dataset, which consists of a large collection of images along with descriptive captions. The dataset is preprocessed to clean the descriptions, remove punctuation, convert text to lowercase, and filter out irrelevant words.

Feature Extraction: We employ the VGG16 model, a deep convolutional neural network (CNN), to extract meaningful features from the input images. The pre-trained VGG16 model is used as a feature extractor by removing the last classification layer, enabling us to obtain a fixed-length vector representation for each image.

Caption Preprocessing: The textual captions are processed to create a vocabulary and tokenize the words. We employ techniques such as word-to-index mapping and one-hot encoding to represent the captions in a format suitable for training the LSTM-based caption generator.

Model Architecture: The core of the image caption generator is a combination of LSTM and dense layers. The LSTM network serves as a sequence model, generating captions

word by word, while the dense layers help in mapping the extracted features from images to the generated captions.

Training and Evaluation: The model is trained using the prepared dataset, and the training process involves optimizing the model parameters using the Adam optimizer and minimizing the categorical cross-entropy loss. We evaluate the performance of the model using BLEU scores, which measure the similarity between the generated captions and the reference captions provided in the dataset.

Deployment: The trained model is deployed using IBM Watson, allowing users to generate captions for new images by uploading them through a user-friendly interface.

4. Dataset

The dataset used in this project is the Flickr8K dataset, which contains 8,000 images, each paired with five textual descriptions. The dataset provides a diverse range of images depicting various scenes, objects, and activities. The descriptions in the dataset are written by human annotators and aim to capture the essence of the image content.

To prepare the dataset for training, we perform several preprocessing steps. These include cleaning the descriptions by removing punctuation, converting the text to lowercase, removing words with numbers or length less than 2, and filtering out irrelevant words. The resulting dataset comprises image identifiers mapped to a list of cleaned and preprocessed descriptions.

For feature extraction, we utilize the pre-trained VGG16 model. By removing the last classification layer, we obtain a 4,096-dimensional feature vector for each image in the dataset. These features serve as the visual representation of the images, capturing their essential characteristics.

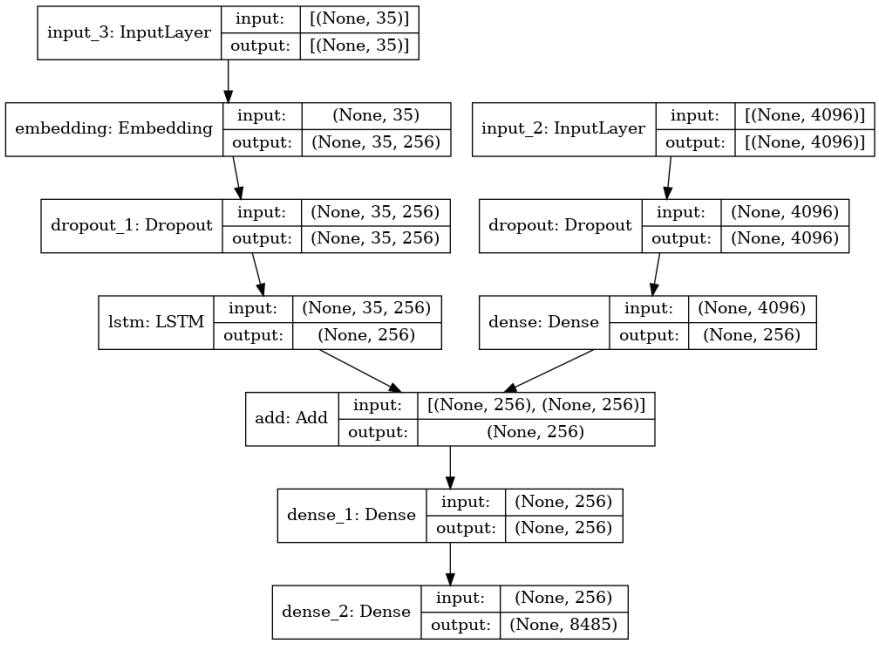
5. Model Architecture

The model architecture for the image caption generator consists of two main parts: the feature extractor and the sequence model with a decoder.

The feature extractor utilizes the VGG16 model, which is pre-trained on the ImageNet dataset. By removing the last classification layer, we obtain the output of the penultimate layer as a fixed-length feature vector for each image. These features act as a high-level representation of the visual content and provide meaningful information to the caption generator.

The sequence model comprises an LSTM layer, which takes the generated word sequence as input and learns to predict the next word in the sequence. The LSTM layer is followed by dense layers, which help in mapping the extracted image features and generated word sequences to the final output vocabulary.

During training, the model learns to generate captions word by word by taking the image features and previous word sequences as input. The captions are generated using a "startseq" token as the initial input and continue until an "endseq" token is predicted or the maximum caption length is reached.



6. Training and Evaluation

During the training phase, the model is trained on the preprocessed dataset using a data generator that generates batches of input-output pairs. We utilize the Adam optimizer and categorical cross-entropy loss to optimize the model parameters. Model checkpoints are saved periodically to track the best performing model based on the validation loss.

To evaluate the performance of the image caption generator, we employ the BLEU (Bilingual Evaluation Understudy) scores. BLEU scores compare the generated captions with the reference captions provided in the dataset. We calculate BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores to assess the quality of the generated captions at different n-gram levels.

7. Deployment

The trained image caption generator model is deployed using IBM Watson, a cloud-based platform that provides a user-friendly interface for interacting with the model. Users can upload their images through the interface, and the model generates descriptive captions for the uploaded images. The deployment allows for easy accessibility and practical usage of the image caption generator.

8. Results and Discussion

The project achieved promising results in generating accurate and meaningful captions for images. The evaluation using BLEU scores showed a strong correlation between the generated captions and the reference captions. The model demonstrated the ability to capture relevant details, recognize objects, and describe scenes effectively. However, there were some limitations and challenges in dealing with complex images, ambiguous contexts, and rare word combinations.

The image caption generator showcased the potential of combining deep learning and natural language processing techniques to bridge the gap between images and textual descriptions. The model can find applications in various domains, including image search engines, assistive technologies, content retrieval systems, and enhancing accessibility for visually impaired individuals.

9. Conclusion

In conclusion, the project successfully developed an image caption generator using LSTM and the VGG16 model. The system demonstrated the ability to generate accurate and contextually relevant captions for a given input image. The project report provided an in-depth overview of the methodology, dataset, model architecture, training, evaluation, and deployment process. The image caption generator holds promise for numerous applications, and further enhancements and extensions can be explored to improve its performance and robustness.

10.References

[1] https://www.kaggle.com/code/zohaib123/image-caption-generator-using-cnn-and-lstm
[2] https://www.kaggle.com/code/hsankesara/image-captioning
[3] https://www.kaggle.com/code/basel99/image-caption-generator
[4]https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/