FAST NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES

(KARACHI CAMPUS)

Department of Computer Science

Spring 2024

“Image Captioning ”

Deep Learning For Perception Project Report

Section: A

Group Members:

Aliza Hafeez Jahan 20K-0222

Robaisha Masood 20K-0390

Objective

The purpose of this project is to develop an image captioning system that can automatically generate descriptive textual captions for given images. This involves leveraging both computer vision techniques for image understanding and natural language processing (NLP) techniques for text generation. The goal is to create a model that can accurately describe the content of images in natural language.

Problem statement

Image captioning involves converting an image into a descriptive sentence or sequence of words. This project specifically aims to:

- Extract image features using a pre-trained DenseNet201 model.

- Process and preprocess caption text data for training.

- Develop a deep learning model architecture that combines image features with text embeddings using LSTMs.

- Train the model to generate relevant and coherent captions for given images.

- Evaluate the model's performance based on caption generation quality.

Methodology

Importing Libraries

Necessary libraries are imported such as numpy, pandas, TensorFlow, etc. These libraries perform tasks like data preparation, image processing, deep learning model building, and evaluation.

PreProcessing Captions

To preprocess captions a function called **text_preprocessing** converts all characters to lowercase, removes non-alphabetic characters, removes extra whitespaces, and appends start and end tokens to each caption. This function is to prepare data for tokenization.

Tokenization and Encoding

The captions are tokenized into words using Tokenizer from Keras. It broke down the text into words and assigned a unique integer to each word. The maximum length of captions is calculated to prepare for padding sequences later.

Image Feature Extraction

A pre-trained convolutional neural network (DenseNet201) is used in code to extract features from images. The model is instantiated and the last layer is removed to obtain features from the images. Extracted features of all images are stored in a dictionary.

Data Generation

A class of **CustomDataGenerator** is defined to generate batches of data during model training. It takes image features, tokenized captions, and other parameters as input and gives batches of pairs of image captions for training.

Modeling

The architecture of the model is defined using the Functional API of Keras. It has two input layers:

1. For image Features
2. For Tokenized Captions

The image features are passed through a dense layer, reshaped, and concatenated with the embedded captions sequences. For sequence processing the concatenated vector is fed as input into the LSTM layer. In the end, a dropout layer and dense layers process the output before predicting the next words.

There are a total of 10 layers

1. input1: Input layer for image features.
2. input2: Input layer for tokenized captions.
3. Dense: A dense layer for processing image features.
4. Reshape: Reshaping the output of the dense layer.
5. Embedding: Embedding layer for tokenized captions.
6. Concatenate: Concatenation layer to combine image features and embedded captions.
7. LSTM: Long Short-Term Memory (LSTM) layer for sequence processing.
8. Dropout: Dropout layers for regularization.
9. Add: 1 occurrence to add the output of dense and dropout layer.
10. Dense: Dense layer for final prediction.

Model Training

Cross-entropy loss and Adam Optimizer are used to compile the model. For training and validation, Data generators are instantiated. Functions such as ModelCheckpoint, EarlyStopping, and ReduceLROnPlateau are used to adjust the learning rate during training by watching validation loss. The model is trained using the fit method.

Caption Generation

The trained model produces captions using the function predict_caption. It takes the image, tokenizer, maximum length, and image features as input and generates a caption using a greedy search strategy.

Caption Prediction

Randomly some images are selected and their captions are predicted. The predicted captions along with the corresponding images are displayed using the **display_images** function.

Results

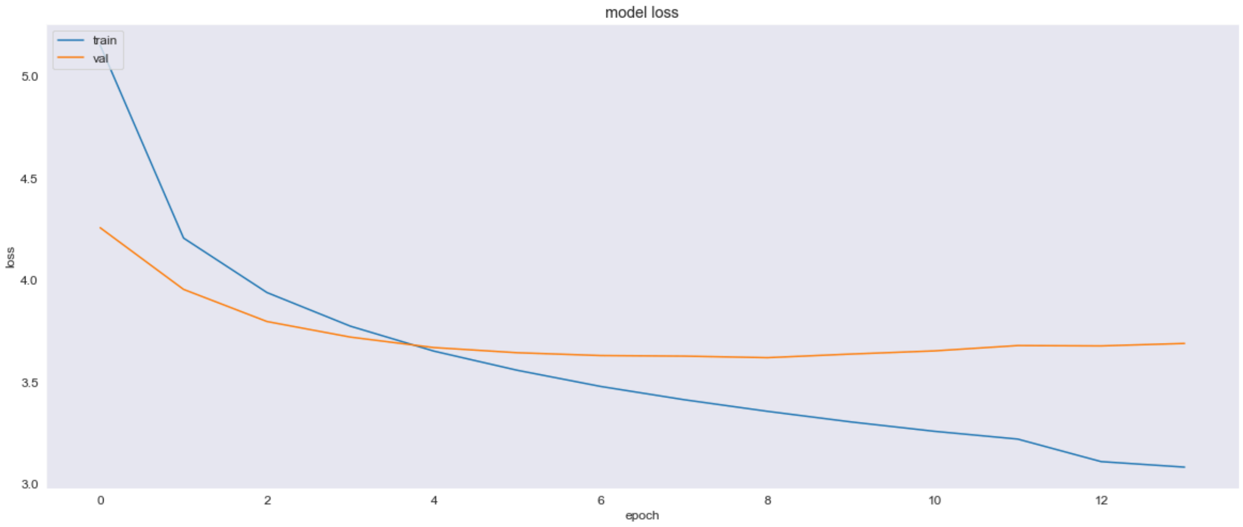
The model achieved promising results in generating captions for images, as demonstrated by the learning curve and sample caption predictions. However, there are areas for improvement:

- The model shows signs of overfitting, likely due to limited data.

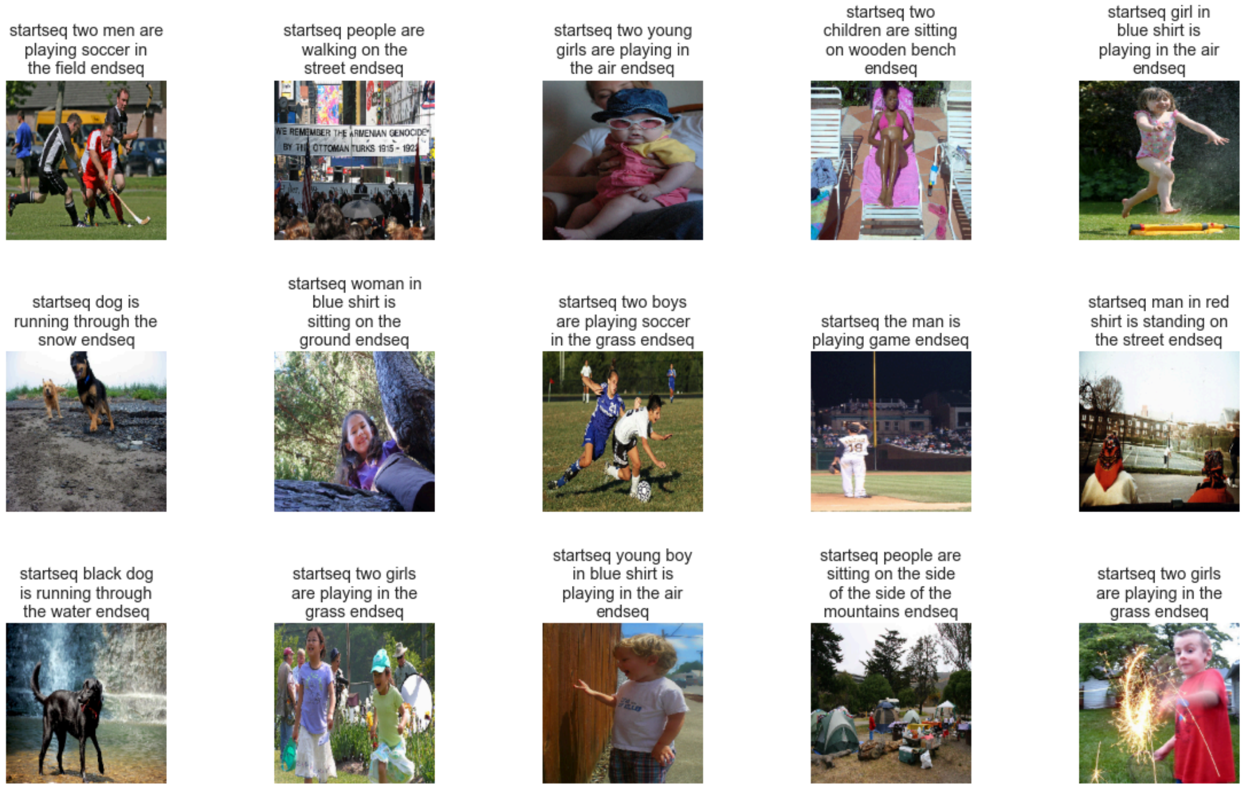
- Caption quality can be further enhanced by training on larger datasets and incorporating attention mechanisms.

-Should be trained for longer with more epochs

Learning Curve:



Sample caption predictions:



startseq two men are playing soccer in the field endseq

startseq people are walking on the street endseq

startseq two young girls are playing in the air endseq

startseq two children are sitting on wooden bench endseq

startseq girl in blue shirt is playing in the air endseq

startseq dog is running through the snow endseq

startseq woman in blue shirt is sitting on the ground endseq

startseq two boys are playing soccer in the grass endseq

startseq the man is playing game endseq

startseq man in red shirt is standing on the street endseq

startseq black dog is running through the water endseq

startseq two girls are playing in the grass endseq

startseq young boy in blue shirt is playing in the air endseq

startseq people are sitting on the side of the side of the mountains endseq

startseq two girls are playing in the grass endseq

Conclusion

In conclusion, this project demonstrates the implementation of an image captioning system using deep learning techniques. While the model performs reasonably well, there are avenues for enhancement, including training on larger datasets and exploring attention-based architectures for improved captioning accuracy. This project serves as a foundational approach to tackling image captioning problems and sets the stage for future advancements in the field.

References

[1]"Features to mimic human image understanding," J. Big Data, vol. 9, p. 20, 2022.

[2]K. Rungta, G. Chau, A. Dewangan, M. Wagner, and J. Huang, "Image captioning using an LSTM network," University of California, San Diego.

[3]O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Google, 2015.