NATURAL LANGUAGE PROCESSING – CS6320

Submitted by

KARTHIKA SHALINI POTAIL – ksp230004

REVANTH SAI GOWRISETTY – rsg230005

Abstract
Image captioning is the process of generating textual descriptions for images using artificial intelligence. This project presents an approach that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to generate captions for images from the Flickr8K dataset. By leveraging pretrained CNNs for feature extraction and LSTM-based sequence models for text generation, the system is able to generate human-like captions that describe the content of the images.

What is Image Captioning?

- Image Captioning is the process of generating textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions.
- This task lies at the intersection of computer vision and natural language processing. Most image captioning systems use an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence.

Introduction
The intersection of computer vision and natural language processing has enabled machines to interpret images and describe them in human language. Image captioning is a challenging problem that requires understanding both the visual content and generating coherent and relevant natural language descriptions. This project aims to implement an end-to-end pipeline using CNNs and LSTMs to address this challenge.

Project Goal
To automatically generate accurate and descriptive captions for images by integrating deep learning techniques — CNNs for extracting visual features and LSTMs for generating natural language sentences.

Scope of the Project
The project focuses on the following tasks:
- Loading and preprocessing images from the Flickr8K dataset.

- Extracting meaningful features using a pretrained CNN (VGG16).
- Cleaning and preparing caption data.
- Designing and training an LSTM-based caption generator.
- Generating captions for unseen images.
- Evaluating the generated captions.

Methodology

Dataset

The Flickr8K dataset, containing 8,000 images with five human-annotated captions per image, was used. The captions provide natural language descriptions used as references.

Dataset Overview

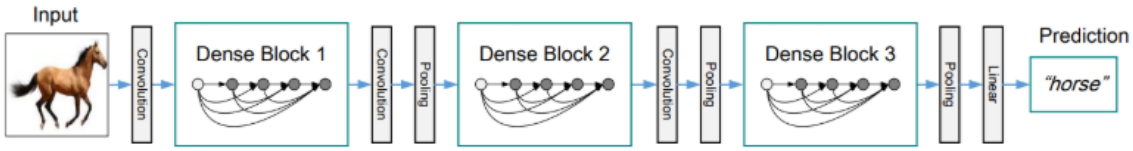
- **Images**:
  The dataset contains exactly **8,000 images** collected from the popular photo-sharing website **Flickr**.
  These images are diverse and reflect everyday scenes involving people and objects in natural environments.
- **Captions**:
  Each image is annotated with **five different textual descriptions**, which were manually written by human annotators.
  These descriptions provide a variety of ways to describe the same visual content and add richness to the dataset.
- **Total Captions**:
  There are a total of **40,000 captions** (8,000 images × 5 captions each).

Feature Extraction (CNN)

The pretrained VGG16 model was employed for feature extraction. The last fully connected layer output (4096-dimensional vector) was used as a compact representation of the image.

Image Feature Extraction

- DenseNet 201 Architecture is used to extract the features from the images
- Any other pretrained architecture can also be used for extracting features from these images
- Since the Global Average Pooling layer is selected as the final layer of the DenseNet201 model for our feature extraction, our image embeddings will be a vector of size 1920



Caption Generation (LSTM)

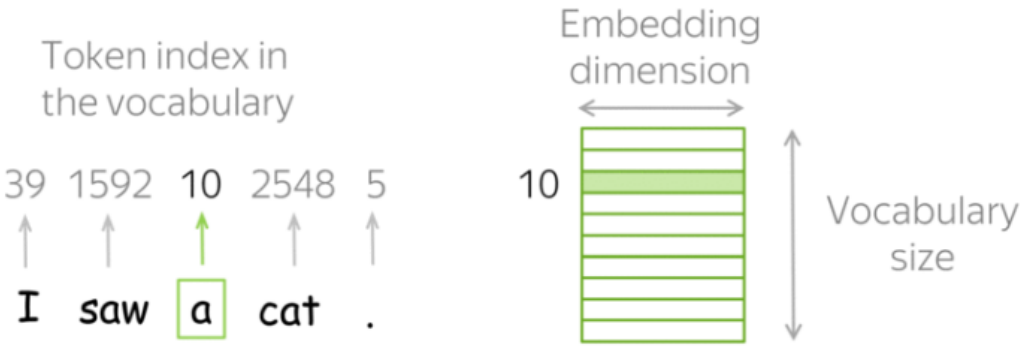
Tokenized captions were converted into padded sequences. The model was built with dense layers, embedding layers, LSTM layers, and a merge layer to combine image and text representations.

Data Generation

- Since Image Caption model training like any other neural network training is a highly resource utillizing process we cannot load the data into the main memory all at once, and hence we need to generate the data in the required format batch wise
- The inputs will be the image embeddings and their corresonding caption text embeddings for the training process
- The text embeddings are passed word by word for the caption generation during inference time

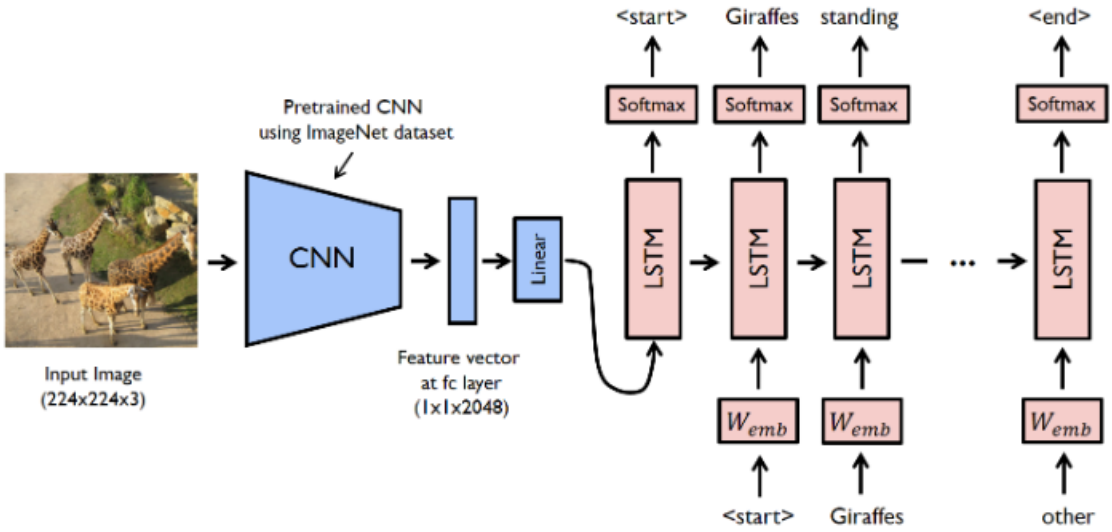
Tokenization and Encoded Representation

- The words in a sentence are separated/tokenized and encoded in a one hot representation
- These encodings are then passed to the embeddings layer to generate word embeddings



Modelling

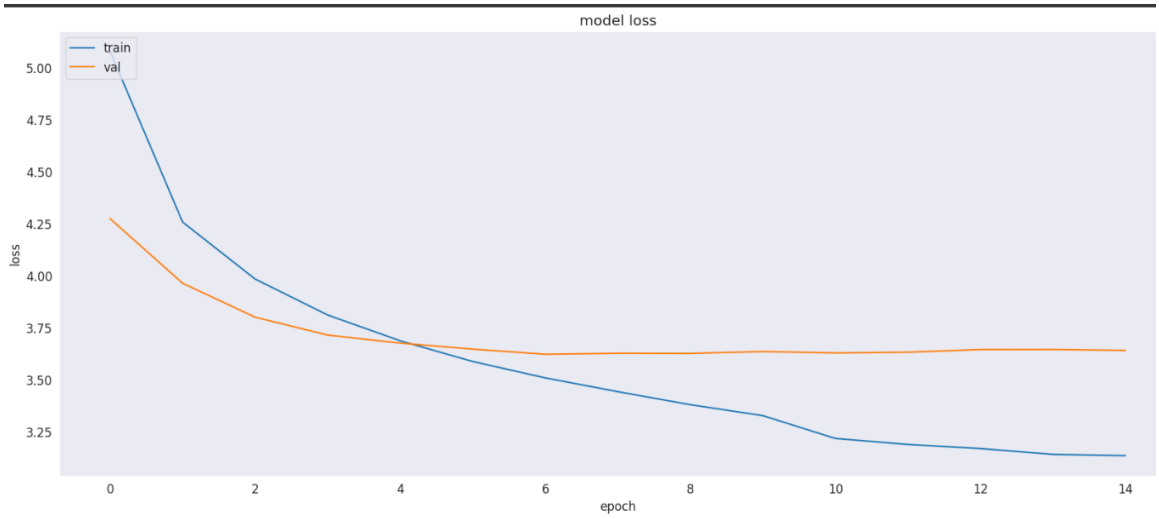
- The image embedding representations are concatenated with the first word of sentence ie. starseq and passed to the LSTM network
- The LSTM network starts generating words after each input thus forming a sentence at the end



Training

The model was trained to predict the next word in a caption sequence given the image and prior words, minimizing categorical cross-entropy loss.

Learning Curve



Problems and Difficulties Faced

- Handling large datasets and pre-processing took a lot of time.
- This model took a lot of computing power and since the GPU power was limited, it took a lot of time to run the epochs.
- Having less GPU power, we have to settle with less number of epochs.

Interesting Aspects

- Combining computer vision and NLP seamlessly.
- Using pretrained CNN for efficient feature extraction.
- Exposure to multimodal AI problems.

Lessons Learned

- Importance of data preprocessing.
- Fine-tuning models for better performance.
- Understanding limitations of LSTM in sequence generation.
- Combining different AI domains effectively.

Future Work

- Replace LSTM with Transformer models for better results.
- Use larger datasets for improved caption diversity.
- Implement attention mechanisms.
- Deploy the model as a web application.
- Explore reinforcement learning for fine-tuning.

Conclusion

The project demonstrates the potential of combining CNN and LSTM for image captioning. Despite challenges, the model achieved reasonable BLEU scores and can be further improved with advanced techniques and larger datasets.