Image Caption Generator

Neural Network and Deep Learning Project

MPSTME

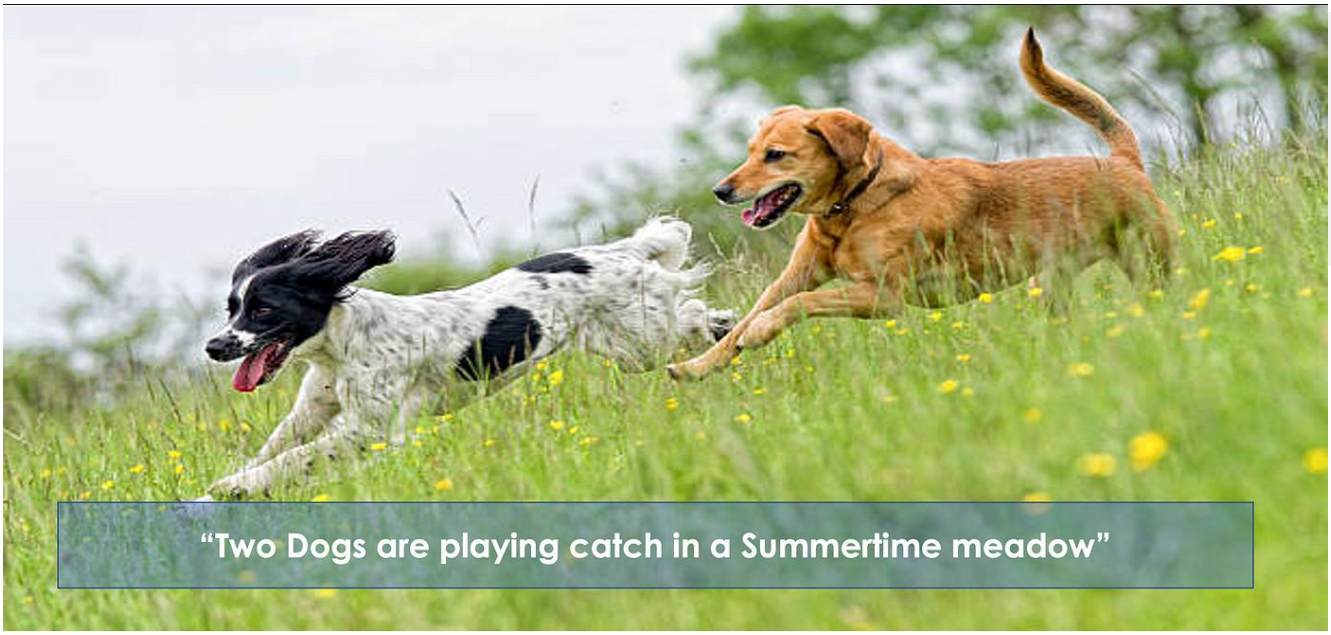
Harsh Raj J011

Pranav Manoj J012

Introduction

About the project

In the age of data-driven artificial intelligence, the ability to understand and describe the content of images is a fundamental challenge with far-reaching implications. The field of computer vision has made significant strides in this direction, and one of the prominent subfields within it is image captioning. Image captioning aims to bridge the gap between visual and textual data by automatically generating descriptive and coherent textual descriptions for images.



The core idea behind image captioning is to create a model that can comprehend the content of an image and express that understanding in natural language. Such a system not only requires an in-depth grasp of the visual world but also a capacity for linguistic creativity. It is this fusion of computer vision and natural language processing (NLP) that makes image captioning a complex and fascinating problem.

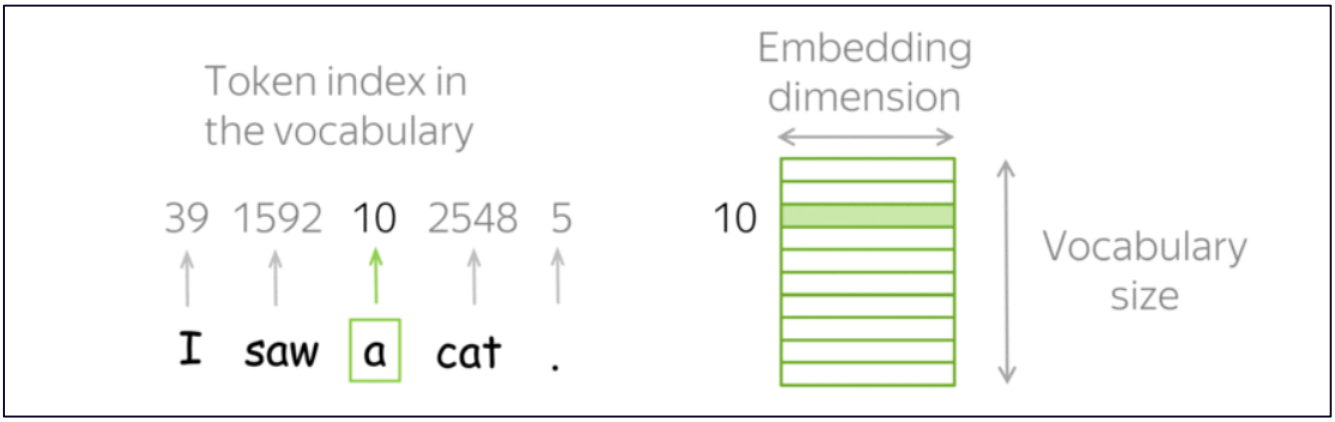
Significance and Objective

The significance in the execution of this project is its potential to improve on the abilities of the AI-powered systems. This contributes to the vast horizon of AI, pertaining to the multi-modal understanding, where images and text converge. The objectives of this project are to come up with a deep learning model that can analyze the images, scan through the contents in the images and accordingly assign an appropriate caption to it. We will also be looking how the current model stands in comparison to other image-captioning models in existence.

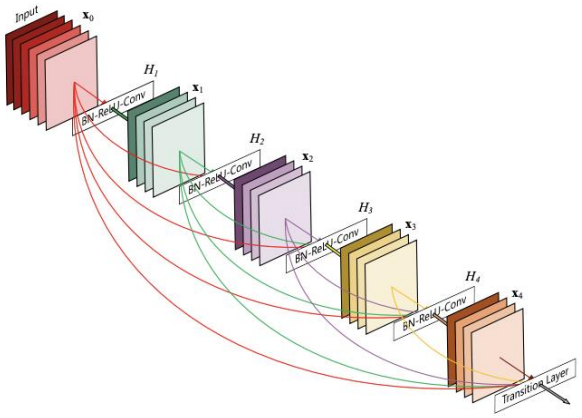
Data Preprocessing

Preparation of the Data

Preprocessing steps both have to be carried out on text (caption) and the images. The text preprocessing steps involve conversion of the characters in the captions to lowercase, removal of special characters, extra spaces, single characters. Additionally, we will be adding a starting and ending tag to the sentences, indicating the starting and ending of the sentences respectively. This is done to suit to the model being applied further on. The next step is to tokenize the words in the sentences and encode them into a one-hot representation. These encodings are further passed on to the embedding layers to generate the respective word embeddings.



Now, the images have to be prepared in the suitable way. Feature extraction is a vital step in any process involving images. This is done so that the complex data gets simplified, dimensionality is reduces, and in the end, improves the performance of the model or algorithm. In this project, we have used DenseNet-201 architecture for feature extraction. This model has been trained on a large dataset, hence applied here to extract meaningful features from images. DenseNet-201 is a version of the DenseNet architecture, which is a deep convolutional neural network (CNN).



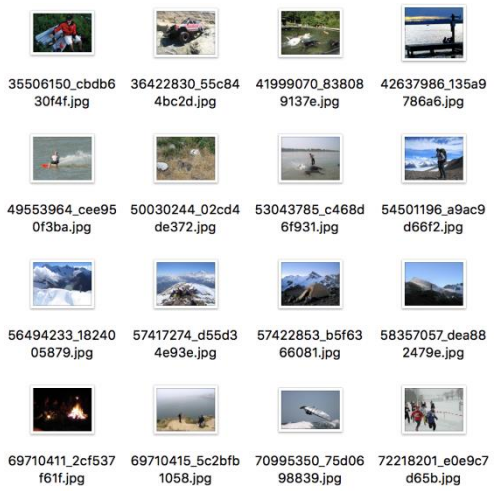
DenseNet is known for its dense connections between layers, where each layer receives input from all previous layers. This architectural characteristic allows for effective feature reuse and can lead to improved feature extraction. We feed an image into this network, we obtain feature representations of the image at various layers of the model. These extracted features represent high-level abstractions of the image, making them suitable for tasks like image captioning.

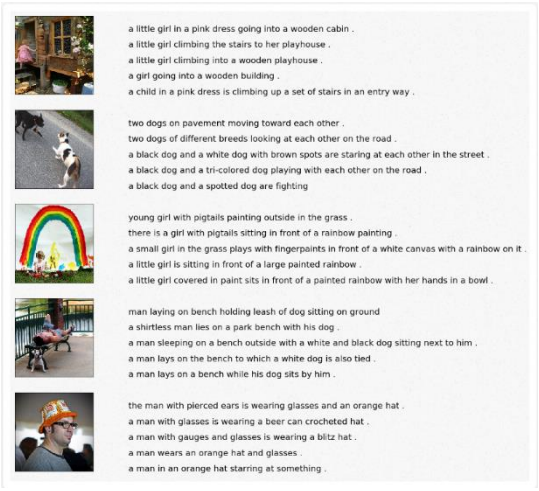
Methodology

Dataset, Model and Architecture

The dataset that has been used in this project is flickr8 dataset. This is a widely used dataset in the fields of Computer Vision and Natural Language Processing. This has been designed for the purpose of image captioning, consisting of both images and respective captions for the images. The dataset consists of approximately 8,000 unique

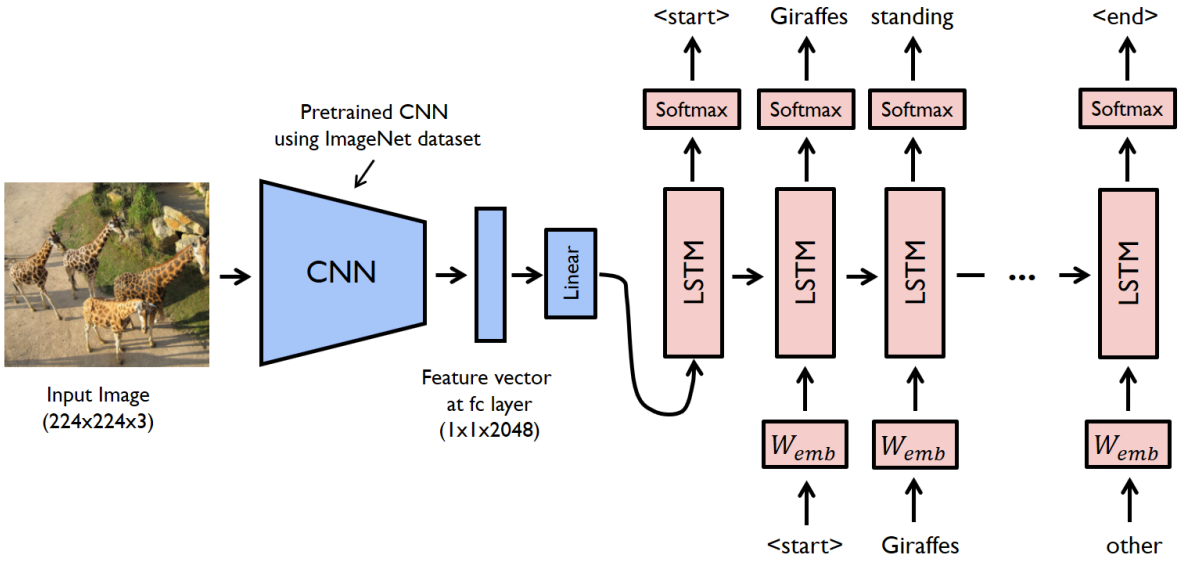
images. Each of the images come with 5 different textual captions. This gives us 40,000 pairs of caption-image set.





The images in this dataset are diverse with a huge variety of actions, scenes and objects, making it highly optimum for training on a huge set of visual concepts. The concepts too are of high quality, every one of them pertaining to the respective images.

Now we have the captions and images ready to be trained. The model that has been used in this project for training is Long Short-Term Memory (LSTM). This is where we enable the model to generate captions for our images on its own.



LSTM is a type of recurrent neural network (RNN), known for its ability to take in sequential data. The embedding representation are concatenated with the starting

words of the sentences. Since we have preprocessed the caption sentences to start with startseq, it will be concatenated with startseq and passed on to the LSTM network. The network then starts generating words after each input. These words in the end form the sentence in the end.

Application and Future work

Future Potential

The future potential for the concept of caption generators for images is certainly very bright. Image caption generators have diverse applications, including enhancing accessibility for the visually impaired, improving content retrieval and search through context-aware captions, automating social media posting, simplifying e-commerce product descriptions, content summarization, aiding robotics and autonomous vehicles, and facilitating education by providing descriptions for visual materials.

Future work in image caption generation will focus on improving language generation quality, deepening multimodal understanding, incorporating common-sense reasoning, providing fine-grained descriptions, expanding to cross-lingual and multilingual capabilities, addressing ethical diversities, enabling real-time and interactive captioning, and enhancing customization and personalization options. The dynamic nature of this field ensures ongoing advancements in image caption generation and its widespread integration into various applications.

Conclusion

In conclusion, our image caption generator project has yielded promising results in the realm of bridging the visual-linguistic divide. As we've observed, the performance of our model can be further enhanced through two key avenues: training on more extensive datasets and incorporating the powerful attention mechanism. The quest for more data not only provides the model with a broader spectrum of visual concepts but also hones its ability to generate richer and more contextually accurate captions. Additionally,

integrating an attention mechanism empowers our model to focus on the most relevant areas of the image when generating text, improving both the quality, relevance and interpretability of the captions.

Appendix

Materials, Repositories referred