SPARTIFICIAL

A Guided Internship project by Spartificial Innovations Private Limited

Snap Narratives - Image Caption Generator

Collaborators:
Aditya R
Ayushi Indu
Shubh Pathak
Urvashi Indu
Vishal Srivastava

Mentor:
Arun M

Guide:
Rohan Shah

Acknowledgements

We would like to take this opportunity to express our heartfelt gratitude to everyone who contributed to the successful completion of this project. Firstly, we would like to thank our project mentor, **Mr. Arun M**, for his invaluable guidance, support, and insights throughout the development process. We would also like to express our sincere appreciation to our Project guide, **Mr. Rohan Shah**, for his tireless efforts in managing and coordinating the project, ensuring that it was completed on time and to the highest standard. Additionally, we extend our thanks to our family members and friends who provided us with unwavering support and encouragement throughout the project. Their constant support and motivation were a driving force behind our success. We are grateful to have had such a dedicated and supportive team, and we look forward to future collaborations.

❖ Introduction

This report details the development and implementation of a cutting-edge Machine Learning based Image Captioning System. An image caption generator is a type of artificial intelligence (AI) model that automatically generates a textual description of an image. The system utilizes advanced technologies such as Image Feature Extraction using EfficientNetB4 and Natural Language Processing with LSTM, and was trained on the Flickr30K dataset. The report will explore the various features of the system, including its innovative approach to image captioning and the technologies that enable it to generate high-quality captions for images.

The image caption generator works by first analyzing the image to identify the objects, people, and other visual features in the picture to generate a textual description of the image that accurately describes the visual content.

Image caption generators have a wide range of applications, from helping people with visual impairments to understand images to improving search engine optimization by generating descriptive image captions. They are also used in fields such as e-commerce, social media, and digital marketing to generate captions that help users better understand and engage with visual content.

❖ Data Preprocessing

- **Image Feature Extraction**: Images are the most essential part of the dataset that we work with in this project. For the system to generate an accurate caption, it should be able to understand the contents of the image and essentially, what is going on in the image. Feature extraction is a part of the dimensionality reduction process, in which an initial set of the raw data is divided and reduced to more manageable groups. For the task of feature extraction, we use transfer learning techniques. Transfer learning involves using a pre-trained convolutional neural network model for image feature extraction and fine-tuning it for a specific task, in our case, image caption generation. This approach can save significant computational resources and improve the accuracy of the model. There exist many architectures in transfer learning that one can use, each varying in the number of layers present in the network. For our project, we worked with four such SOTA models, namely: VGG16, EfficientNet B4, ResNet50 and Inception V3 for determining and selecting the best performing architecture for producing the best results.
- **Caption Processing:** The Flickr dataset along with the images contained corresponding 5 pre-written captions. For caption preprocessing we implemented following steps:
  - Convert sentences into lowercase
  - Remove special characters and numbers present in the text
  - Remove extra spaces
  - Remove single characters
  - Add a starting and an ending tag to the sentences to indicate the beginning and the ending of a sentence

❖ Caption Generation - LSTM

For caption generation we first tokenized our captions and formed a vocabulary consisting of all the cleaned words present in the captions dataset which would be used for generating new captions. Then we created a data generator to get data in batches.Then we split our dataset into training and testing dataset.Then we created an encoder layer which included one input layer, one dropout layer and two dense layers. Following which we created a decoder layer which included one input layer, one embedding layer, one dropout layer and one LSTM layer(Long Short Term Memory model is a type of Recurrent Neural Network that is specifically designed to handle sequential data, such as time series, speech, and text). And at the end we finally compiled our model with 'categorical cross entropy' as our loss function and 'adam' as optimiser.

❖ Results

We developed our project using two different datasets, namely: Flickr8K[b] and Flickr30K[c]. The difference in the two datasets is their respective sizes, which were 8,000 and 30,000 images. Although it is clear that the dataset with larger size would be preferable for developing better performing models, we required a way to measure the performance of a given LSTM model in order to identify the best transfer learning architecture for the project. For this, we used BLEU score as the performance measure. BLEU or Bilingual Evaluation Understudy score is the measure of similarity of two sequences of words or sentences. A perfect match of the two sentences would mean a BLEU score of 1 and a perfect mismatch would result in a BLEU score of 0.

The following table depicts the BLEU scores of the model while using the two different datasets and the four different architectures. In our case, the two sentences would be the one produced by the LSTM model and the one generated by the captions present in the dataset.

Model	Flickr8k				Flickr30k			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
EfficientNet B4	0.590557	0.364770	0.237044	0.146666	0.591614	0.350487	0.219173	0.129545
Resnet50	0.557725	0.332924	0.210045	0.121752	0.571551	0.320197	0.194258	0.109866
VGG 16	0.526105	0.299053	-	-	0.541007	0.296905	0.194258	0.109866
Inception V3	0.432056	0.188727	0.104808	0.053587	0.469754	0.209127	0.112608	0.054758

Since it is clear that using EfficientNet B4 architecture for the flickr30K dataset yielded the best BLEU score, this was used for image feature extraction

❖ Conclusion

In conclusion, for our project *Snap Narratives* we selected EfficientNet B4 + LSTM model architecture and successfully developed a web app using HTML, CSS and JavaScript for frontend and Flask for backend in order to perform Image Captioning for user provided Images. We have also implemented a text to speech feature for the users' accessibility.

❖ References

- ➢ https://www.hackersrealm.net/post/image-caption-generator-using-python
- ➢ Flickr 8k Dataset: https://www.kaggle.com/datasets/adityajn105/flickr8k
- ➢ Flickr 30k Dataset: https://www.kaggle.com/datasets/eeshawn/flickr30k
- ➢ Validation Dataset (Self Compiled): https://www.kaggle.com/datasets/shubhpathak0614/validation-set