GA-DSI-123

Developing Image Captioning Model With Deep Learning

Presented By Farhad Farahani

April 17th, 2023

Hello!

Farhad Farahani

Data Scientist working on developing image captioning model in Research Department at Google

Problem Statement

Visual contents are ubiquitous in today's digital world and are used in a wide range of applications including Google apps.

Aid visually impaired individuals in understanding the content and context of the image.

Developing an image captioning model with the ability to generate accurate and relevant captions for images.

Agenda

01 **Data Collection**
Flickr8k with 8,000 images and 40,000 captions

02 **EDA and Preprocessing**
Cleaning captions, length of captions and find the words with the most frequencies

03 **Models**
DenseNet201 and VGG16 with LSTM

04 **Conclusions & Recommendations**

01 Data Collection

Data Collection

- Provided from **Flickr 8k** Dataset from Kaggle.
- **8,000** images are each paired with **five** different captions.
- Clear descriptions of the salient entities and events.
- Were chosen from six different Flickr groups, and tend not to contain any well-known people or locations.

Data Collection

Image

Captions

- A child in a pink dress is climbing up a set of stairs in an entry way .
- A girl going into a wooden building .
- A little girl climbing into a wooden playhouse .
- A little girl climbing the stairs to her playhouse .
- A little girl in a pink dress going into a wooden cabin .

Data Collection

A man in a beret rides a bicycle down the street .

a bike rider jumping into the air over a wooden ramp .

A blond girl standing in a crowd holding a goat on a leash .

Some samples

One child is walking ahead of the other .

Two people running on a beach .

Three dogs in different shades of brown and white biting and licking each other .

02

Exploratory Data Analysis (EDA) and Preprocessing

EDA and Preprocessing

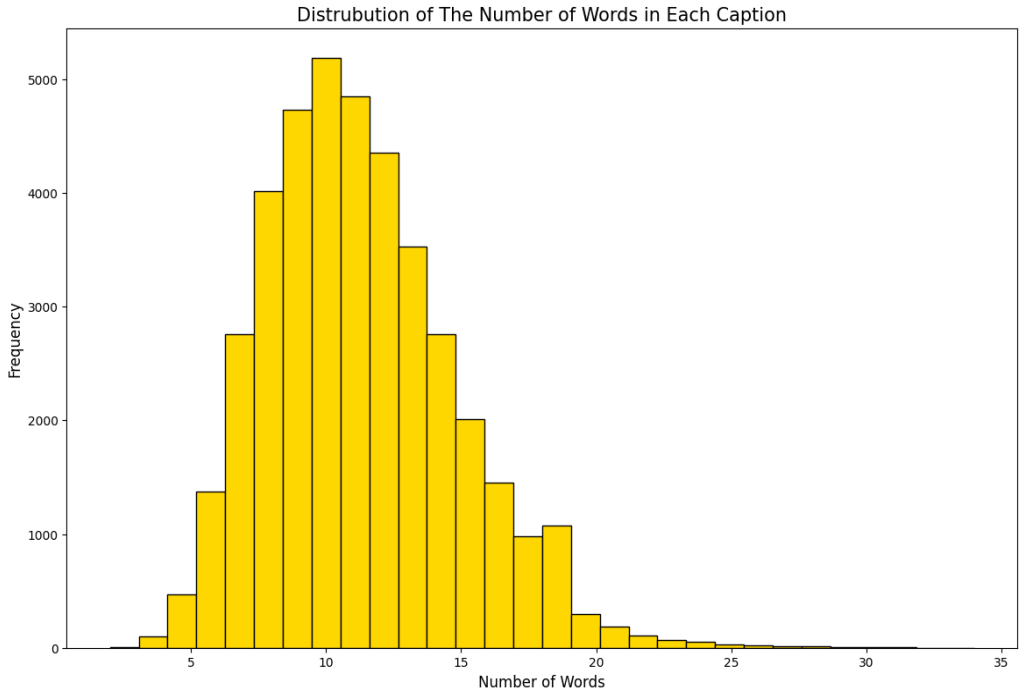
Convert to **lowercase**

Remove **special characters** and **numbers**, **punctuations**, **extra spaces** and **single characters**

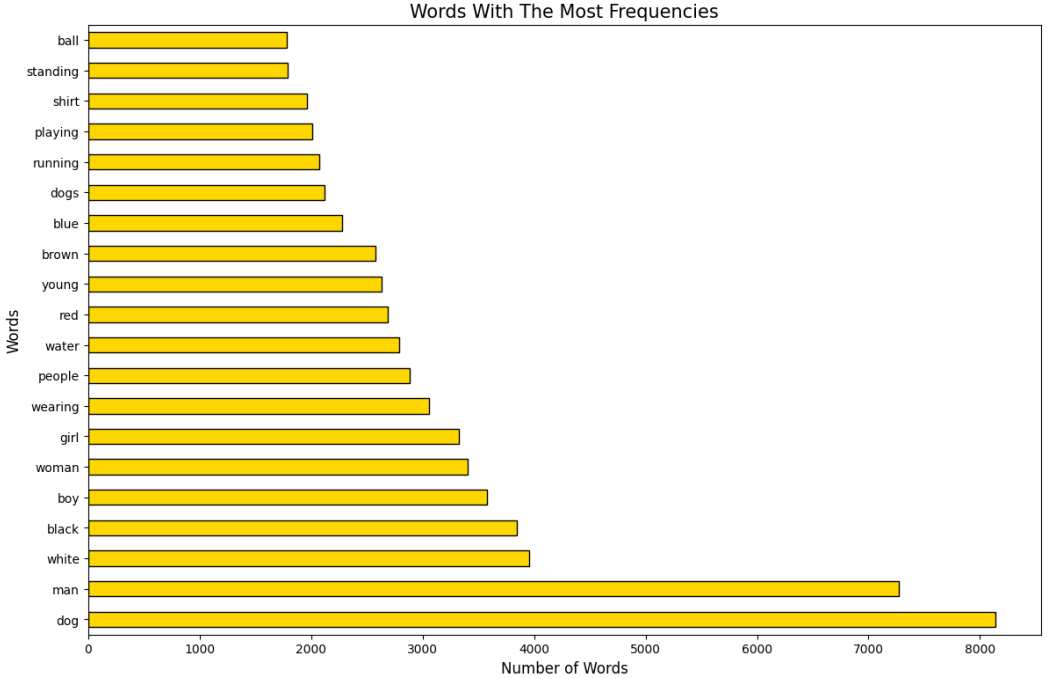
Add **starting** and **ending tags** to the sentences to indicate the **beginning** and the **ending** of a sentence

EDA and Preprocessing

- Length of the generated captions can have a significant impact on the performance and quality of the model.
- Most captions lengths are **10**
- Set the maximum caption length to **25**.



EDA and Preprocessing

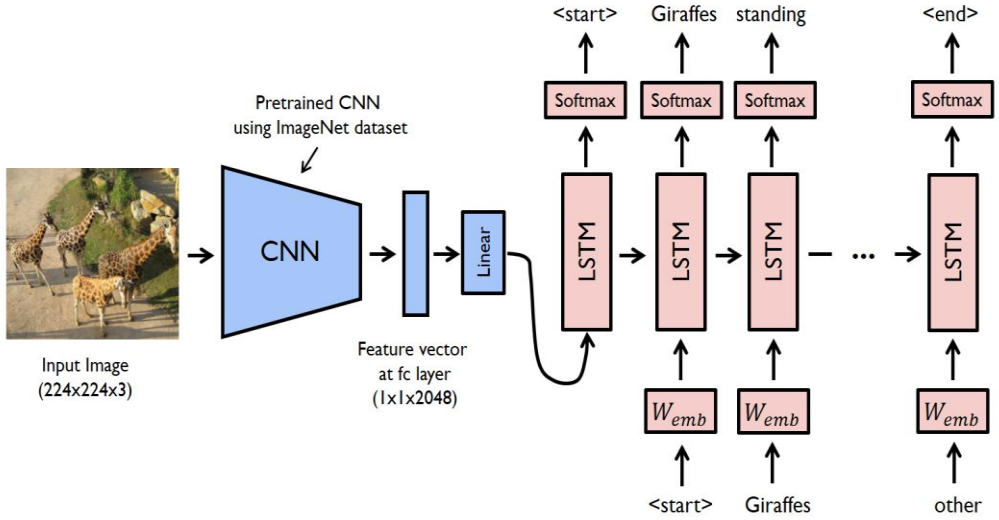


After cleaning, by having this plot we could see that most images are about **people** and **dogs**, explaining what they are **wearing** or **doing** with **color** detail descriptions.

03 Models

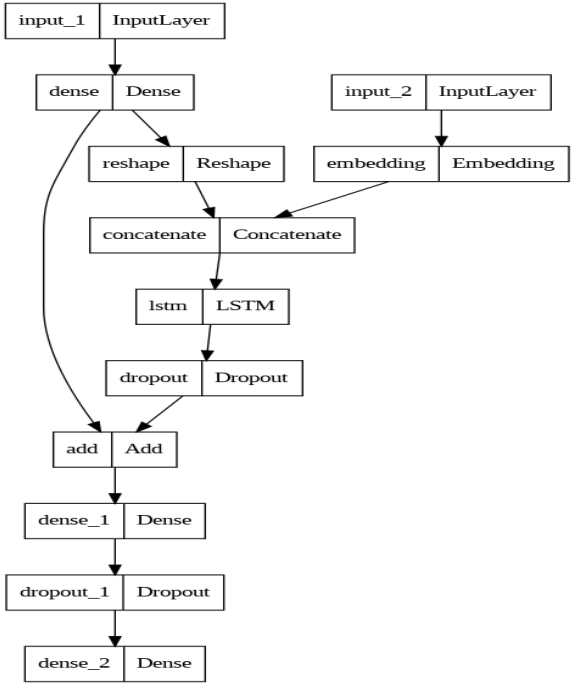
Models

- Utilize pretrained CNN models, **VGG16** and **DenseNet201**, to extract image features.
- Extracted features will be passed through an **LSTM** model to generate captions.
- Utilize **GloVe** embeddings to represent the words in the captions.



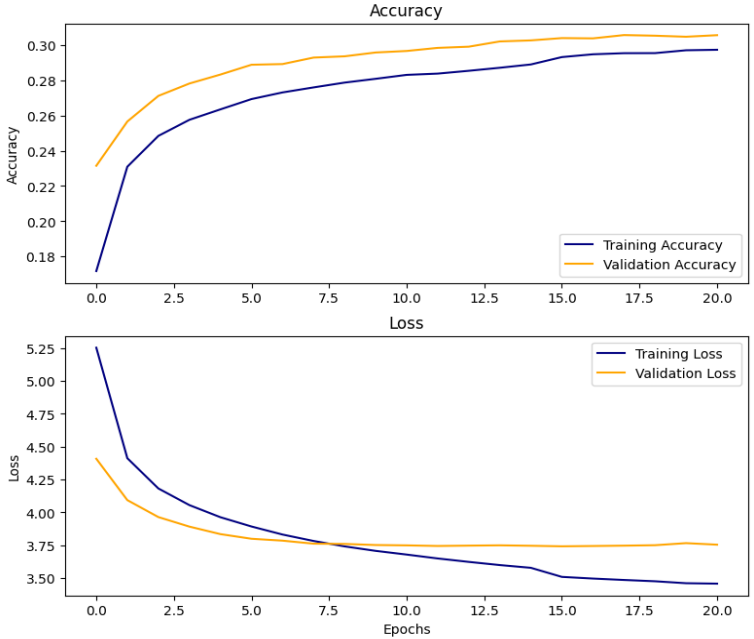
Models

- Use extracted image and caption features.
- Combine features and fed them to LSTM layer, to generate a sequence of words to make the predicted caption.
- Utilize the dropout technique to prevent overfitting.
- Finally, generate a probability distribution over the possible words.

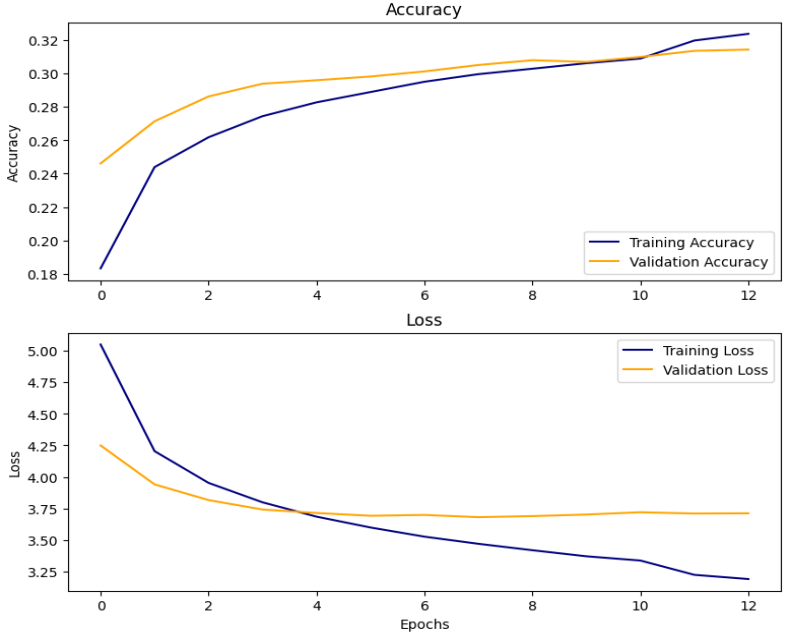


Models

Accuracy and Loss Score With DensNet201-LSTM Model



Accuracy and Loss Score With VGG16-LSTM Model



Models

Model Performance Evaluation

Model	Accuracy Score
Baseline	0.000025
DensNet201-LSTM Model	0.30
VGG16-LSTM Model	0.31

04 Conclusions & Recommendations

Conclusions

Challenge

- Aid visually impaired individuals in understanding the context of the image.
- Adding image caption feature to Google Assistant.

Solution

Developing an image captioning model with the ability to generate accurate and relevant captions for images.

Results

- Successfully developed an image captioning model which is a combination of CNN and LSTM networks.
- Convert the generated captions into audio descriptions.

Recommendations

Larger Datasets

Using larger datasets, such as Flickr30k, MSCOCO and SBU can potentially improve the accuracy of the model.

Pretrained Models

Explore different pretrained models and different word embeddings on the accuracy of the model.

Expand

Consider expanding the application of the model for generating video descriptions or captions for images in a different domain, such as medical imaging.

Thanks!

Do you have any questions?

Momenifarahani.farhad@gmail.com
farhadfarahanii.github.io