Neural Networks and Deep Learning
Course project

Image Caption Generator

Team members:

- Sree Vidya - 180020020
- Sri Varsha Pendem - 180020035
- Divisha Lakshmi - 180020038

Mentor:

Prof. Mahadeva Prasanna

Problem Statement:

To develop a model, which can automatically generate the description of an image with the use of CNN along with LSTM.

Introduction:

Image captioning aims to automatically generate a sentence description for an image.Image captioning can be used in various applications such as :

- Cell Phone application to describe surrounding scenes for blind or visually impaired people.
- We can also use this model to generate video captions for IP cameras which can be used to detect abnormal activities and raise alarm immediately.
- Searching photos within the album as it will be easy searching by keywords.

Dataset:

Our dataset is from COCO where we are getting a large scale of images along with multiple captions corresponding to each image. We had a total of 406095 data points (image, caption pair) with 81219 unique images in total. We are planning to allocate 80% data for training and 20% for testing.

Here are few examples:



Captions :
For image A:
1.Children are playing baseball on a muddy baseball diamond.
2.A child waits for a pitch during a game of baseball.
For image B:
A purple bus and a man dressed as a nun on a tall bicycle.
For image C:
A wine glass sits in front of some plates of food.

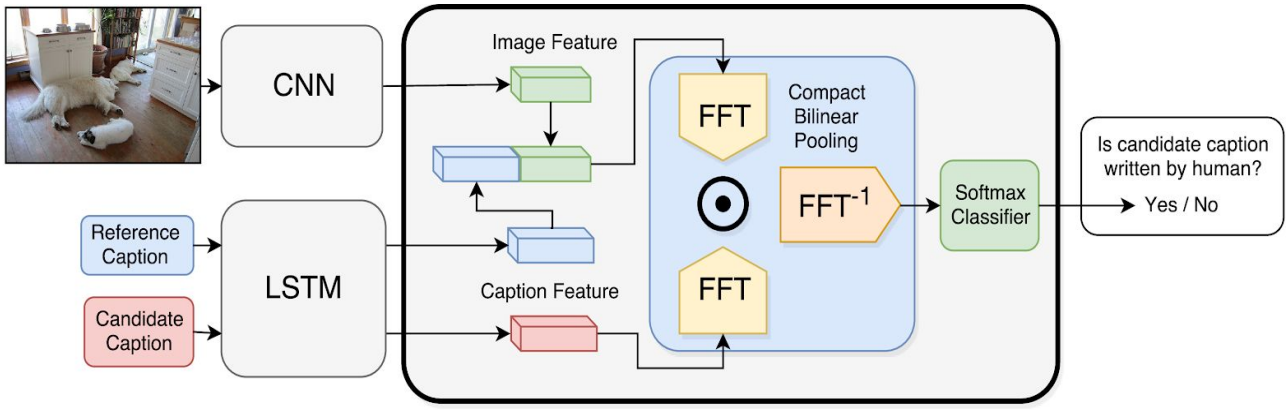
Algorithms:

Convolutional Neural Networks (CNN): Convolutional Neural networks are specialized deep neural networks which processes the data that has input shape like a 2D matrix. CNN works well with images and are easily represented as a 2D matrix. Image classification and identification can be easily done using CNN. It can determine whether an image is a bird, a plane or Superman, etc.
Important features of an image can be extracted by scanning the image from left to right and top to bottom and finally the features are combined together to classify images. It can deal with the images that have been translated, rotated, scaled and changes in perspective.
Object Detection:
Objects are detected from the image with the help of CNN Encoder.

Long Short Term Memory(LSTM):LSTM are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. We can predict what the next words will be based on the previous text. It has shown itself effective from the traditional RNN by overcoming the limitations of RNN. LSTM can carry out relevant information throughout the processing, it discards non-relevant information.

Sentence Generation:

By using LSTM, sentences are generated. Each predicted word is employed to get subsequent words. Using these words, an appropriate sentence is formed with the help of Optimal beam search. Here, Softmax function will be used for prediction of words.



Implementing the Image Caption Generator

1. Data loading and Preprocessing
2. Model building
3. Model training
4. Greedy and Beam Search
5. Evaluation: Metric(BLEU score)

Timeline:

	MODULES	days
1	Convolutional neural networks	By 1st eval
2	Long Short Term Memory(LSTM)	By 2nd eval

References:

- Research paper:
  http://cs230.stanford.edu/projects_spring_2019/reports/18681243.pdf
  Github repository:
  https://github.com/divyanshj16/Image-Captioning/tree/master/datasets
  https://github.com/RoyalSkye/Image-Caption
- Research paper:
  https://www.irjet.net/archives/V7/i4/IRJET-V7I41167.pdf
  Github repository:
  https://github.com/anuragmishracse/caption_generator
- Blog:
  https://www.analyticsvidhya.com/blog/2020/11/create-your-own-image-caption-generator-using-keras/

Future work:

- We will try to improve current model results.
- We will convert the generated caption to speech so that it can be useful for blind people.