Automated Image Captioning using CNN & RNN

Anuj Jhunjhunwala | Prof. Radhakrishnan Delhibabu | SCOPE

Introduction

Captioning the image means extracting the meaningful description of an image. The research papers that I have surveyed contains the basis of my project implementation, i.e. use of neural networks. But, it doesn't show any concrete application whereas in my work I have molded my project in such a way that would be beneficial to the blind people to observe their surroundings indirectly.

Motivation

According to the statistics, around 1.5% of India's population comprises of blind people. My project will ultimately help the blind to observe his surroundings independently with use of deep learning and advancements in technology.

SCOPE of the Project

A solution requires both that the content of the image be understood and translated to meaning in the terms of words, and that the words must string together to be comprehensible. It combines both computer vision using deep learning and NLP and marks a true challenging problem in broader AI. Further, the generated text will be converted to speech so that the blind can hear the description of what is there in his surroundings. So, the following objectives are in the scope of this project; **1)** Extract features from the captured image. **2)** Generate a caption for the extracted features. **3)** Convert the caption to speech so that the user can hear.

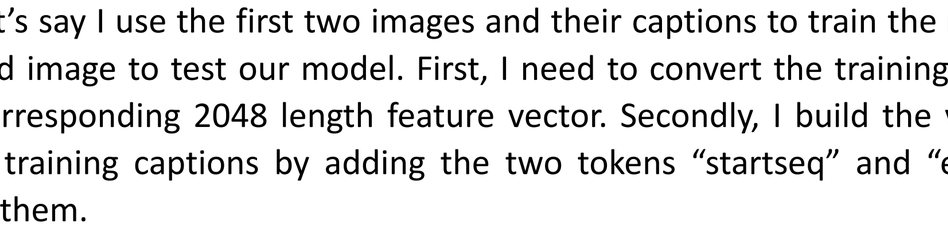
Methodology

There are 6 steps involved in the methodology:

1. Data Cleaning – Involves removing of stopwords and punctuation marks that are not necessary in the computation process.

2. Loading the training set

3. Data Pre-processing on Images and Captions



4. Data preparation using Generator Function

(Train image 1) Caption -> The black cat sat on grass

(Train image 2) Caption -> The white cat is walking on road

(Test image) Caption -> The black cat is walking on grass

Now, let's say I use the first two images and their captions to train the model and the third image to test our model. First, I need to convert the training images to their corresponding 2048 length feature vector. Secondly, I build the vocabulary for the training captions by adding the two tokens "startseq" and "endseq" in both of them.

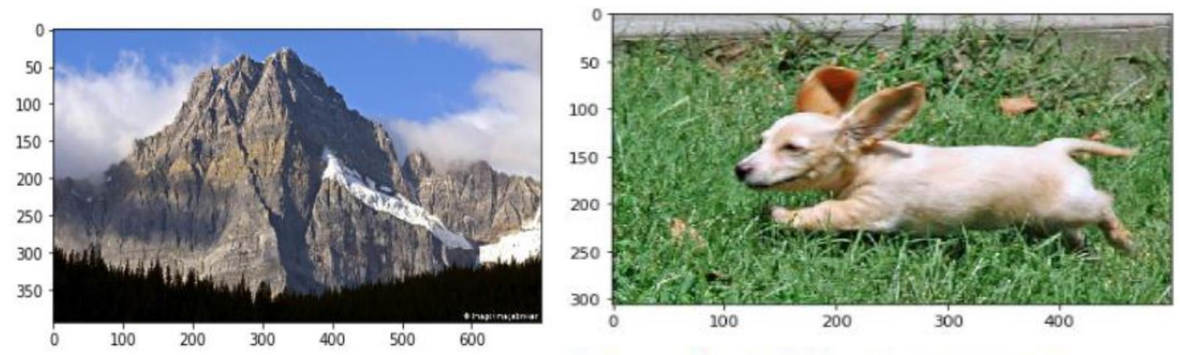
	Xi		Yi
i	Image feature vector	Partial Caption	Target word
1	Image_1	startseq	the
2	Image_1	startseq the	black
3	Image_1	startseq the black	cat
4	Image_1	startseq the black cat	sat
5	Image_1	startseq the black cat sat	on
6	Image_1	startseq the black cat sat on	grass
7	Image_1	startseq the black cat sat on grass	endseq

5. Word Embeddings – These are the texts converted into numbers and there may be different numerical representations of the same text. I have mapped every word (index) to a 200-long vector using a pre-trained GloVe Model.

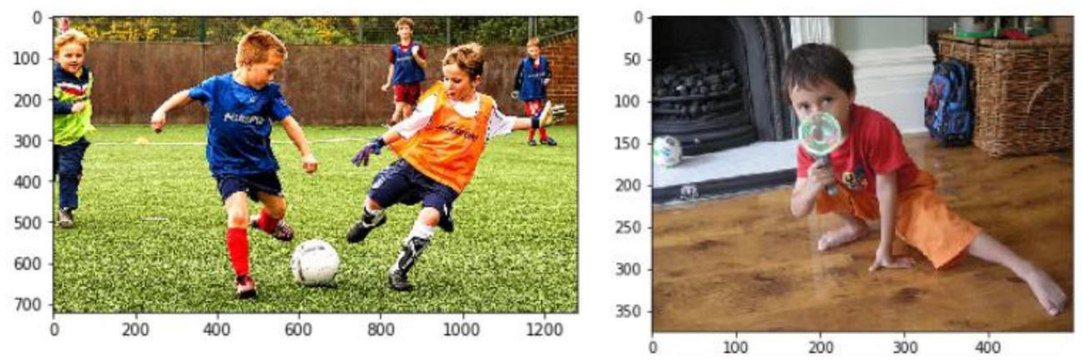
6. Conversion to Speech – The comprehensible caption which is generated is then converted into speech so that the blind person can hear the text. This is done using the gTTS library.

Results

- My neural network system is capable of viewing an image and generating a reasonable representation in English depending on the words in its dictionary generated on the basis of tokens in the captions of train images.
- The model has a convolutional neural network encoder and a LSTM decoder that helps in generation of sentences.
- The purpose of the model is to maximize the likelihood of the sentence given the image.
- There can be many other modifications that can be done to improve the performance of our model such as, Using a **larger** dataset, Doing more **hyper parameter tuning** (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.).



```
(1, 2048)
Caption: man is standing on top of mountain looking at the mountains
```



```
Caption: dog is running through the grass
```

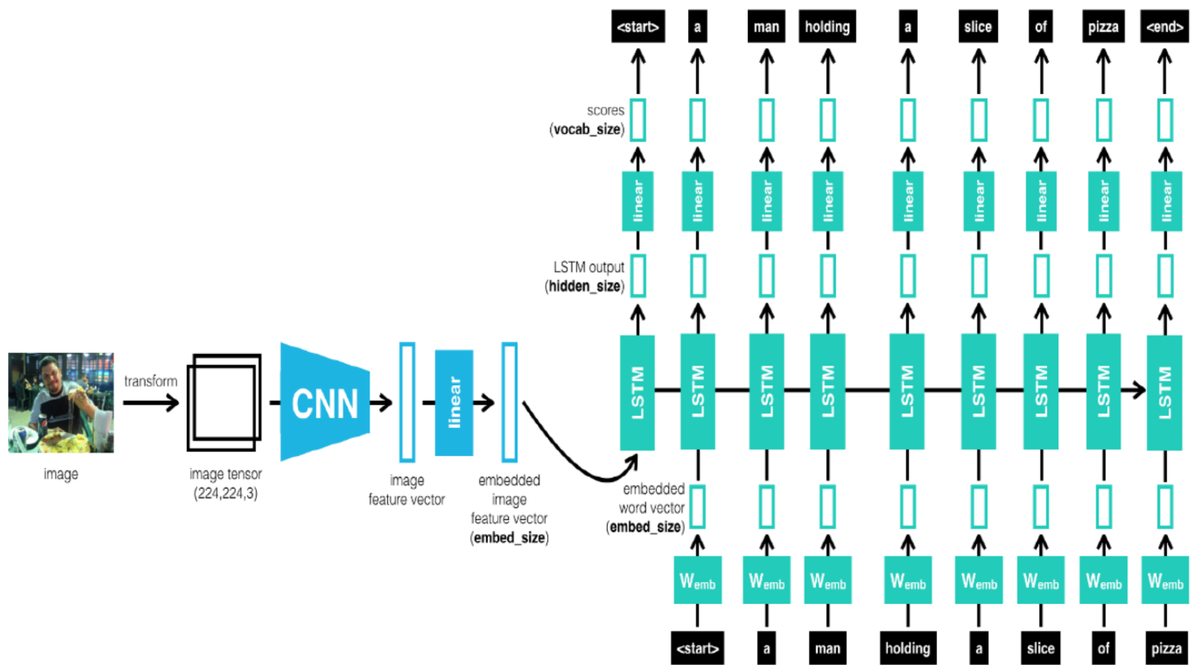


```
(1, 2048)
Caption: two boys play soccer on field
```



```
Caption: young boy is licking water from water fountain
```

Neural Diagram



Conclusion

- From the above set of experimentally obtained results, we can infer the fact that this model provides reasonable accuracy in its output, using the Flickr8k dataset.
- Experimenting the model with Flickr8K dataset shows decent results. The accuracy of the model is estimated to be about 75-80%.
- Working on a bigger dataset such as MS COCO might improve the overall performance of the model and provide higher accuracy in its results.

References

- Wang, E. K., Zhang, X., Wang, F., Wu, T. Y., & Chen, C. M. (2019). Multilayer dense attention model for image caption. *IEEE Access*, *7*, 66358-66368.
- Gupta, N., & Jalal, A. S. (2019). Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. *Neural Computing and Applications*, 1-10.
- Ghosh, A., Dutta, D., & Moitra, T. (2020). A Neural Network Framework to Generate Caption from Images. In *Emerging Technology in Modelling and Graphics* (pp. 171-180). Springer, Singapore.