AIML Project - Image Captioning

170050067 170050068 170050082 170050083

INTRODUCTION:

This project shows how Deep Learning can be used to solve the problem of generating a caption for a given image, hence the name **Image Captioning.**

DATA:

There are many open source datasets available for this problem, like Flickr 8k (containing 8k images), Flickr 30k (containing 30k images), MS COCO (containing 180k images), etc.
We used Flickr 8k data set,6k for training, 1k for validation.One of the files is "Flickr8k.token.txt" which contains name of each image along with its 5 captions we create a dictionary named "descriptions" which contains the name of the image as keys and a list of the 5.

'**startseq**' -> This is a start sequence token which will be added at the start of every caption.

'**endseq**' -> This is an end sequence token which will be added at the end of every caption.

Prepare Photo Data:

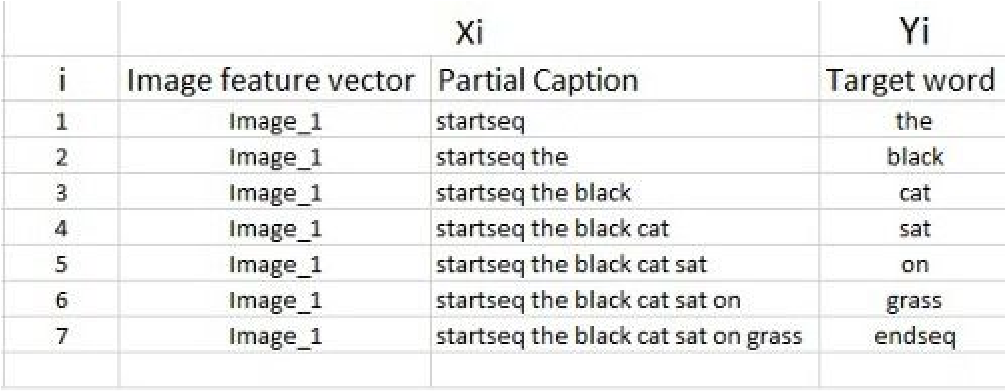
We need to convert every image into a fixed sized vector which can then be fed as input to the neural network. For this purpose, we opt for **transfer learning** by using the InceptionV3 model (Convolutional Neural Network) created by Google Research and VGG16 model.

Hence, we just remove the last softmax layer from the model and extract a 2048 length vector (**bottleneck features**) for every image.(4096 features in case of VGG16)

DATA PREPROCESSING:

We will predict the caption **word by word**. The maximum length of any caption is 34. Total Vocabulary Size 7579.

It's not just the image which goes as input to the system, but also, a partial caption which helps to **predict the next word in the sequence.**



	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

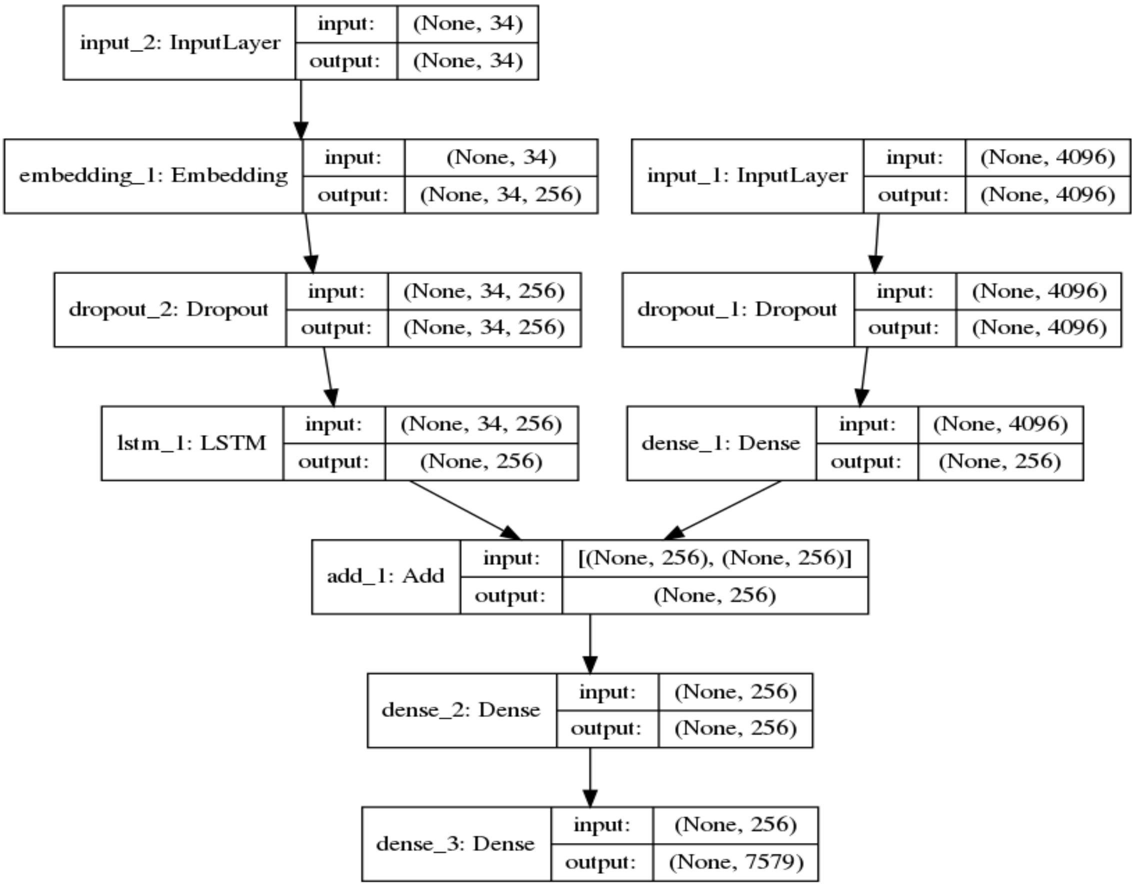
Evaluation Metric:

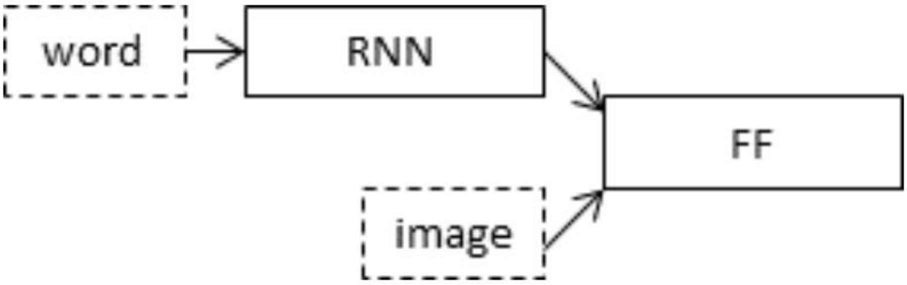
We used BLEU Score to evaluate and measure the performance of the model.
Our approach yields 0.59, compared to human performance around 0.69.

BLEU scores for VGG16 model
BLEU-1: 0.553138
BLEU-2: 0.294281
BLEU-3: 0.194568
BLEU-4: 0.083649

BLEU scores for InceptionV3 model
BLEU-1: 0.541909
BLEU-2: 0.289134
BLEU-3: 0.194673
BLEU-4: 0.087262

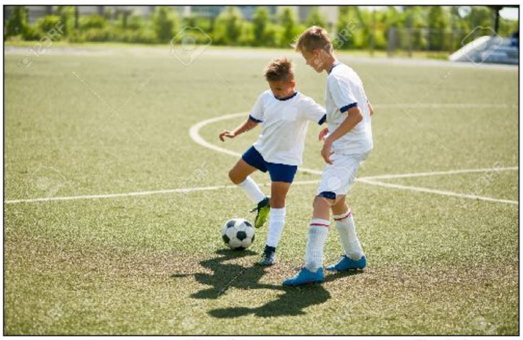
MODEL-ARCHITECTURE:


input_2: InputLayer | input: (None, 34) | output: (None, 34)
embedding_1: Embedding | input: (None, 34) | output: (None, 34, 256)
input_1: InputLayer | input: (None, 4096) | output: (None, 4096)
dropout_2: Dropout | input: (None, 34, 256) | output: (None, 34, 256)
dropout_1: Dropout | input: (None, 4096) | output: (None, 4096)
lstm_1: LSTM | input: (None, 34, 256) | output: (None, 256)
dense_1: Dense | input: (None, 4096) | output: (None, 256)
add_1: Add | input: [(None, 256), (None, 256)] | output: (None, 256)
dense_2: Dense | input: (None, 256) | output: (None, 256)
dense_3: Dense | input: (None, 256) | output: (None, 7579)

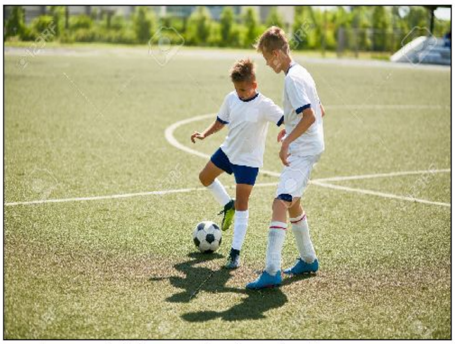


INCEPTION-MODEL BASED

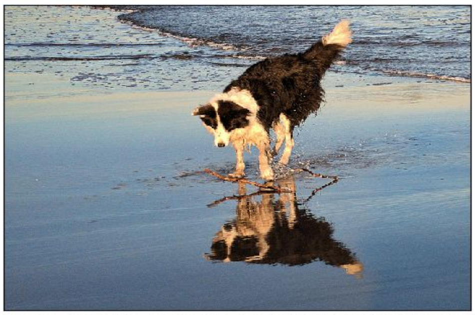
VGG16-MODEL BASED



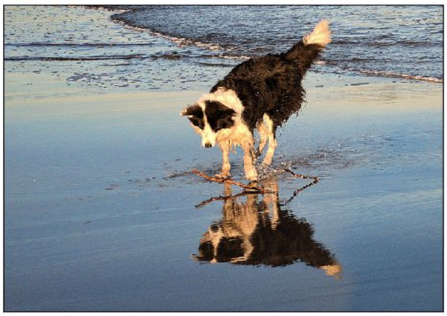
two men playing soccer on field



two children playing in the grass



black dog is running through the water



dog is running through the water

References

1. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf
2. https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8
3. https://www.youtube.com/watch?v=yk6XDFm3J2c