PREFACE

The present report is the outcome of the internship program with UST Global team. The main objective of the internship was to familiarize with the present state of the art technology in the field of Machine Learning. Tenure of the internship was from June 2020 to July 2020. I learnt lot of things which I was not aware of earlier. All the things I have learned during the internship phase are logged in this report. The report includes the work in the field of Image Captioning, which fascinated me right from the beginning. The work is partitioned into Introduction, Problem definition, Approach and the conclusion is partitioned into Results and Discussion, Learning and Outcome, Summary, Future Scope for better understanding. Though there is a lot of partitioning of the work, Approach alone covers the rough idea of the work. But I really suggest any reader of this report to go through all of it. This internship report attempts to bring under one cover the hardwork and dedication put by me to complete my internship at UST. I have expressed the work in my own simple words hoping that this report will reach every reader. Though I have tried my best to keep the report free from errors, I apologize if any error is found and expect you to understand that it was not deliberate.

V. Sai Vipul Mohan
111801045
Bachelor of Technology
Computer Science and Engineering
IIT Palakkad

ACKNOWLEDGEMENT

I would first like thanks my mentor for the internship, Hanish Sargiya. He was there for me all the time. Especially, as the internship was conducted in online mode, I wouldn't have done any better without his guidelines. Even if it was non-working day, he would happily attend my calls or messages and give me required suggestions.

Next, I would like to thank my Primary Point of Contact with the company, Gopika Bindu. She was so generous that she offered her mobile number and WhatsApp number and even encouraged to call her at any difficult times. She helped me as a friend rather than a member of the company. She helped a lot by conveying all my concerns to the company.

Without support from Ashok G Nair, this internship would be so incomplete. Valuable feedback from his side, during the regular meetings held to evaluate our work every week, kept me driving to put in a lot of effort. The comments by him at the final presentation of the internship will always keep me motivated.

Overall, the team from UST Global did its job in the best way possible. Keeping me backed up always. Without them I would have not completed this internship successfully.

ABOUT INDUSTRY/RESEARCH LAB

The labs at UST Global are really appreciable. But I was unfortunate to not have a chance to work there. The internship took place in online platform due to the unprecendented situations.

ABSTRACT

This report describes the internship I spent at UST Global. The main work is to understand the present state of the art techniques applied in Image Captioning and to tweak some parameters of the state of the art technique with the intention of improving its performance. The more challenging and at the same time the most interesting part of the task is that there is no limit of extending this task. There are a lot of variables which can be adjusted to get better performance or worst performance. Even the slightest mistake would steer off the road that is intended to take. The persent state of the art technique to efficiently caption an image was by Andrej Karpathy. The idea was to predict the next word in the sequence given the image and sequence of words. This idea closely relates to the concept of conditional probability. Initially that state of the art technique is used and a deep learning model is developed and even trained on Flickr 8k dataset. Later the architecture of the deep learning model, transfer learning of the deep learning model and metric of the deep learning model are tweaked to get better performance. There was some improvements made on the course of 8 intermediate models, hours of training for each model, millions of parameters in each model. Considering the technological infrastructure available and the time constraint of 8 weeks, the result of the work is pretty convincing.

A perfect understanding of transfer learning can be drawn from this work. There are multiple areas where transfer learning is applied, to enhance the performance of the model. As the transfer learning requires no additional training time, the efficieny it increased compared with the side effects it has is more. Multiple dropout techniques are also tried in this model intending to improve the performance on the test dataset, by looking at the performance on validation dataset. The key concepts of training a deep learning model are clearly explored. Wandb is explored as the part of getting the overview of the performance of models. Each models stats are observed though Wandb's graphical presentation and the decision is taken on the next step towards betterment of the model. Due to the time constraint, the metrics used were just the value of the loss function, which is not the best metric to evaluate the performance of a image-language model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

GloVe - Global Vectors for Word Representation

LSTM - Long Short Term Memory

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

AI - Artificial Intelligence

BLEU - Bilingual Evaluation Understudy

ROUGE - Recall Oriented Understudy for Gisting Evaluation

CIDEr - Consensus based Image Description

SPICE - Semantic Propositional Image Caption Evaluation

METEOR - Machine Translation Evaluation Metric

ILSVRC -ImageNet Large Scale Visual Recognition Challenge

CCTV - Closed Circuit Television

ARCH - Architecture

PERF - Performance

INTRODUCTION

Deep Learning is a sub field of Machine Learning. The main objective of a machine learning task, is to make predictions given a new data. The machine learning model should be trained on old data earlier to make such predictions. Machine learning is fast growing in the present day A.I. world. Machine learning is quite different from hand coding of the algorithms. There are many tasks identified in the present day scenario, in which very good performance can be achieved by employing machine learning algorithms. Especially, machine learning is advancing to greater lengths in Computer Vision. Computer Vision is to make machines look at images and arrive at conclusions. The challenge here is that machines don't look at images as humans do. They look at them as numbers. Now, machines must be trained to derive conclusions by just looking at those numbers. Due to the work of A.I. scientists like Andrej Karpathy, the computer vision field is advancing a lot. One such advancement is self driving cars. Tesla has now come up with industry leading self driving cars, such a revolutionary advancement was possible only because of computer vision.

Image captioning doesn't completely come under computer vision. It has two parts, image model and language model. The image model comes under computer vision. Image captioning was revolutionized by Andrej Karpathy. By looking at the image, computer describes it. It seems very humanly task, but is very difficult for any machine to achieve. Earlier many of the things were impossible for a machine to do. But in todays world machine does everything. Same is the motivation in the case of image captioning. The idea is very nascent but revolutionary. The motivation to choose such a task in the internship is the very same reason.

PROBLEM DEFINITION

The problem introduces a captioning task, which re-quires a computer vision system to both localize and describe salient regionsin images in natural language. The image captioning task generalizes objectdetection when the descriptions consist of a single word. Given a set of imagesand prior knowledge about the content find the correct semantic label for theentire image(s). In simple terms the problem statement is, To assign a piece of text to an image, describing the image. Input to the model is images and its corresponding captions(in particular five captions describing each image are input). And the output of the model is text describing the corresponding input image.

This task comes into the domain of Image deep learning(Computer Vision) and language deep learning. The state of the art datasetd for this problem statement are the MS COCO and Flickr30K. The dataset used in this work is mentioned below.

DATASET

Flickr 8k dataset provided by kaggle is utilised. The dataset consists of 8,000 high quality images, each image has 5 different captions. The images are not of famous individuals or locations. The images are abstract and depict variety of scenes and backgrounds.

This dataset is split into 6000, 2000, 1000 images each for train dataset, dev dataset and test dataset respectively. The models are trained on 6000 images and tweaked by measuring the performance on dev dataset. Ultimately to perform better on the test dataset. The split is made by the makers of the dataset.

INFRASTRUCTURE

All the deep learning models are trained on Intel Core i5 dual core processor clocking upto 1.8GHz along with Intel HD Graphics 6000 on chip graphics. And a memory stick of 8GB.

APPROACH

The above mentioned problem task is well researched by Andrej Karpathy. A similar approach is followed at the beginning. Later improvisations are made to the model to improve performance on the given dataset.

The dataset is partitioned into train, dev, test sets in the ration of 6:1:1. 6000 images belong to the train set i.e. these are used to train the model. 1000 images belong to the dev set i.e.e these are used to tweak the model to ultimately perform well on the train set which consists of 1000 images. The model majorly works on the principles of conditional probability. In the data preprocessing part, each caption is appended with new words at the beginning of the sentence and at the end of the sentence. This preprocessing part is termed as refining the captions in the work. This refining helps to start the prediction and end the prediction. If the captions are not refined then the predicted captions may be extended so long. Each prediction happens this way, first the startseq word along with the image is given as input to the model, the model predicts the next most probable word given the image and first word. Now, first two words (startseq along with the first predicted word) is passed as input to the model, the model now predicts the next most probable word given the image and first two words. This process is repeated till the endseq word is observed. This is the key concept, `predict` function in the code handles this part of the algorithm. So, from this it is evident that the model just need to predict next word, given an image and sequence of words. Images cannot be directly read as input by the machine, so images need to be converted to numbers. Similar is the case with sequence of words. Transfer learning is employed here to efficiently convert images to n-dimensional linear vectors and convert sequence of words to m-dimensional linear vectors. In case of images conversion, InceptionV3, InceptionResNetV2, ResNet152V2 models along with ImageNet weights are used to extract n-dimensional linear vector. Each of these models are image classification models i.e. they output probability vector correponding to each of the image classes. To get an n-dimensional linear vector which holds the properties of the image, last softmax layer of these models is removed an the output of the neurons in that layer is extracted into a vector. Figure INCEPTIONV3 describes the above process.

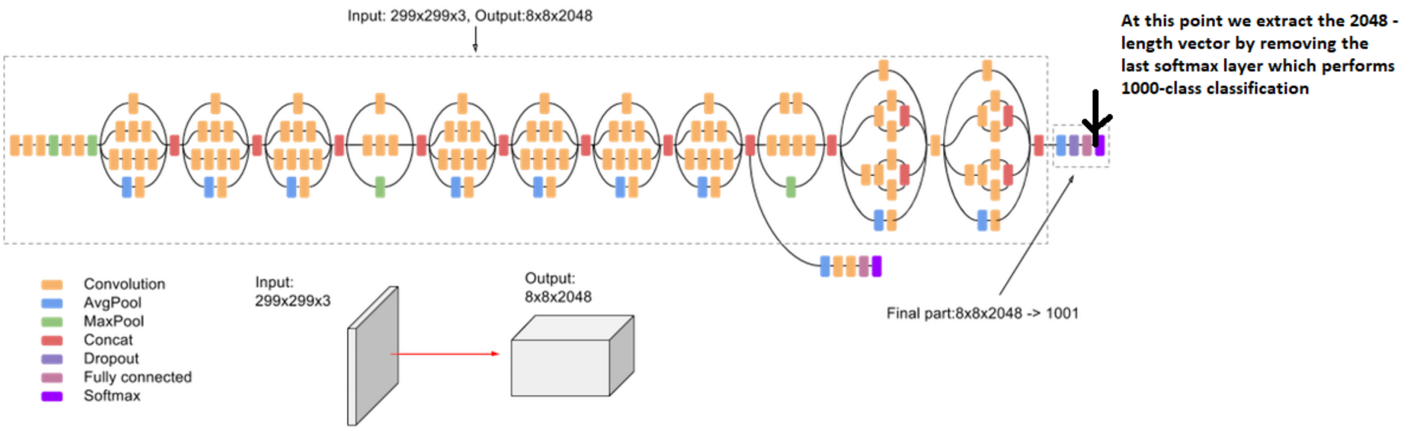


Figure INCEPTIONV3

In case of sequence of words conversion, GloVe is used. GloVe contains vector representation of many words, in particular, GloVe-6B-200d vector representation is used in this work. It contains vector representation of 6 Billion words and each vector representation is 200 dimensional.
After the data gets passed through the transfer learning layers, the data is further passed through custom 10 layer deep learning model. The custom model consists of couple of dropout layers, couple of dense layers, an embedding layer, an LSTM layer. Embedding layer is used as part of the language model transfer learning, to get vector representation of words. Figure MODEL_ARCH_08 displays the layers of the model. There are two input layers in the model. The 2048 neuron input layer is to take the image input which is output from the transfer learning model(InceptionV3, InceptionResNetV2, ResNet152V2) and the 34 neuron input layer is to take the text input which is the words mapped to a number between 1 to 1951(size of the vocabulary). 0 maps to the word startseq and 1952 maps to the word endseq. The output from the input layer is then passed through the embedding layer where each word gets its own representation on the GloVe space. Both the image model and language model are added using the add layer. The last output layer is a softmax layer of size 1952(vocabulary size+1). The argmax of the softmax layer gives the index of the word, thus gives the word itself. Due to the last softmax layer, this models mimics image classification model with 1952 classes. So, Categorical Cross-Entropy loss is used along with Adam optimizer to train the model.

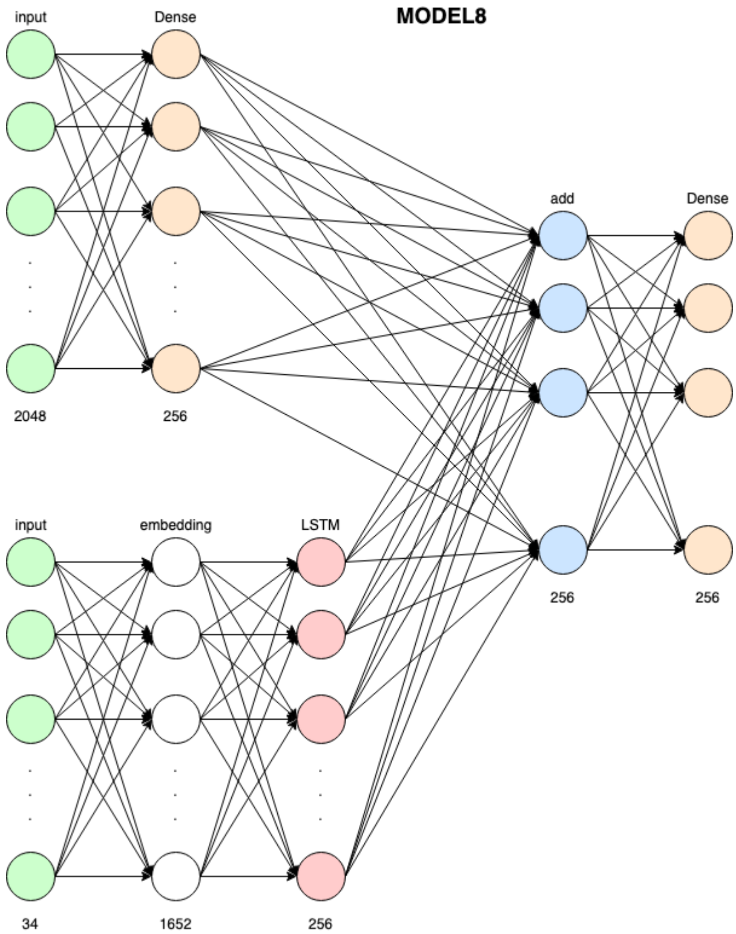


Figure MODEL_ARCH_08

RESULTS AND DISCUSSION

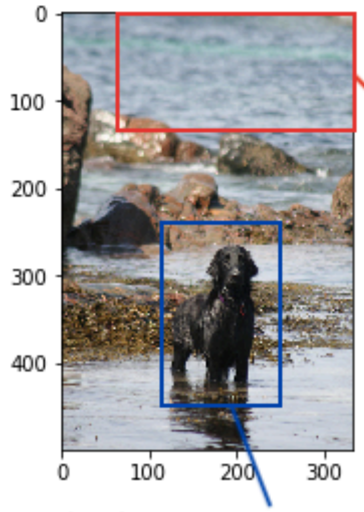
The model is trained on each image with five different captions. In each forward propagation of the deep learning model the next predicted word is found given the image and sequence of words. Now, loss is the deviation of the predicted word from the original word. `categorical_crossentropy` is the loss function used to get the deviation. Later loss is minimized using the `adam` optimiser. The loss of the initial model on the train and dev dataset were correspondingly `2.37(rounded to two decimals)` and `3.48(rounded to two decimals)`. The loss of the final model on the train and dev dataset were correspondingly `2.28(rounded to two decimals)` and `3.89(rounded to two decimals)`. This is not much improvment in the model. At the same time using the loss function as the metrics is not optimal. So, a set of 100 images is considered for testing all the models' performance.

Figure RESULTS_06 is the result of prediction made on first generation model. There are no two dogs in the image as predicted in the caption. Figure RESULTS_07 is the result of prediction made on second generation model. The model understood that there is only one dog in the image. Also, it identified the presence of water in the image. Figure RESULTS_08 is the result of prediction made on the third generation model. The model is capable enough to identify that the dog is taking a leap into the pool to grab a tennis ball. This prediction can be considered to be most nearest to human level performance. Similar trend is observed in other sets of images ranging from figure RESULTS_09, figure RESULTS_10, figure RESULTS_11 and figure RESULTS_12, RESULTS_13. To put the machine a test, an image of my very own laptop is given as input. The model prediction made by the model is not so human level performance, but decent enough. Figure RESULTS_14 demonstrates that. To understand every image result of the model is pretty tiring. So, to analytically compare the models Wandb is emplyed. Results of models ranging from 5 to 8 are comprehensively described by the wandb. And can be found here.

There were some key results observed during the process of improving the efficieny of the model. Such as using highly populated LSTM layer yielded better results. Also, having Guassion dropout rather than usual dropout or alpha dropout resulted increase in performance on the dev dataset. Also, having more denser and more populated neural network resulted in better performance on both train and dev dataset. But the bottle neck here was the size of dataset. Later attempts were made to switch to larger dataset but due to the limited constraints on infrastructure, training on such a larger data seemed impossible. However, the main target was to achieve better performance on the current chosen dataset. Some results of various madels are attached below.

RESULTS

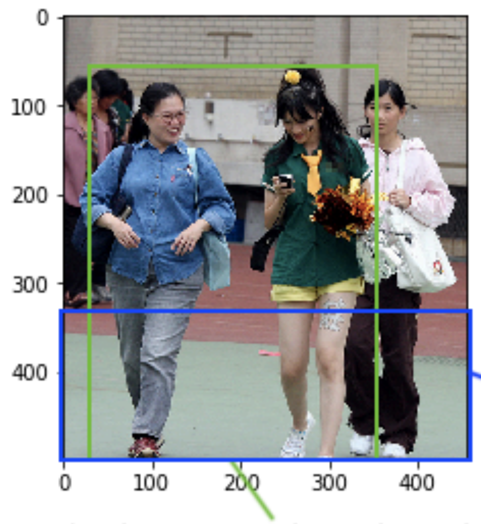
3602838407_bf13e49243



prediction: black dog is running through the water
Ground Truth: black dog in water
Ground Truth: black dog standing in shallow area of water on rocky beach
Ground Truth: dog stands in tide pool
Ground Truth: small black dog in the ocean with some rocks in the background
Ground Truth: small black dog plays in the water

Figure RESULTS_01

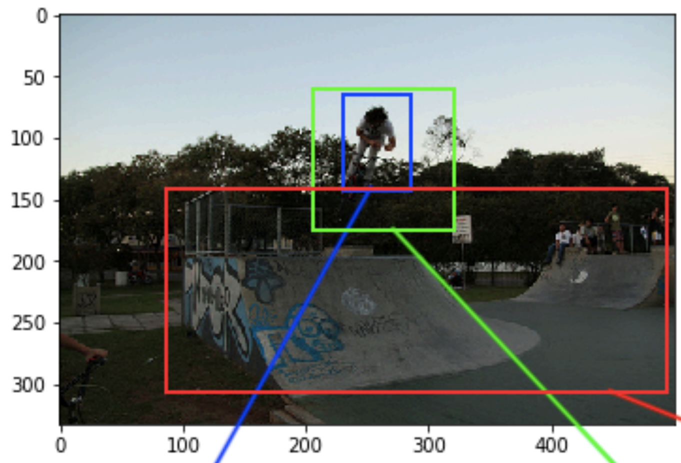
model7
3115901702_f07aa0ef74



prediction: two girls in skirts walking on the sidewalk
Ground Truth: woman walks as she looks at her cellphone and others look on
Ground Truth: several ladies talking together
Ground Truth: three girls walking one is checking her cellphone
Ground Truth: three girls walk together and talk
Ground Truth: three high school girls walk and talk on astroturf

Figure RESULTS_02

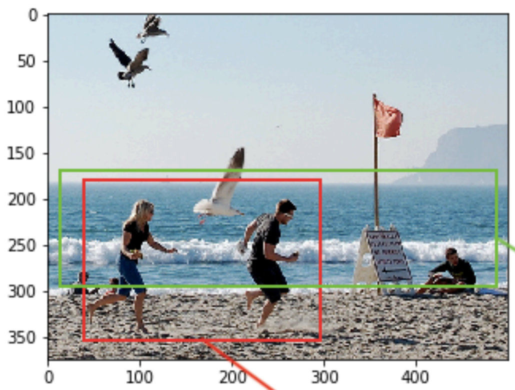
model8
3662963630_8f097e38d4



prediction: man in black shirt is skateboarding on ramp
Ground Truth: biker leaps off halfpipe littered with graffiti as onlookers watch
Ground Truth: bmx rider launches off quarterpipe
Ground Truth: cyclist airborne having jumped his bike off graffiti covered ramp
Ground Truth: kid does trick on bike at skate park
Ground Truth: teenage boy on bicycle is jumping off ramp

Figure RESULTS_03

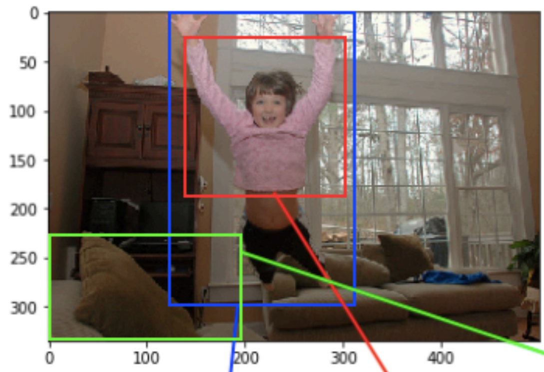
model1
3158327361_6f1a518228



prediction: two people are standing on the edge of body of water and looking at something in the distance
Ground Truth: couple of people running on the shore near the beach
Ground Truth: man and woman hopping acrouss hot sand
Ground Truth: people running away from seagulls on the beach
Ground Truth: two people run in the sand at the beach
Ground Truth: two people run along beach with seagulls

Figure RESULTS_04

model
2210368267_0615754b48



prediction: little girl in pink dress is jumping on bed
Ground Truth: girl in pink shirt jumping from one couch to another in well apportioned living room
Ground Truth: little girl in pink cardigan is jumping from the couch
Ground Truth: little girl in pink shirt has jumped into the air in her house
Ground Truth: little girl jumps through the air to land on the sofa
Ground Truth: girl jumps from couch to couch

Figure RESULTS_05

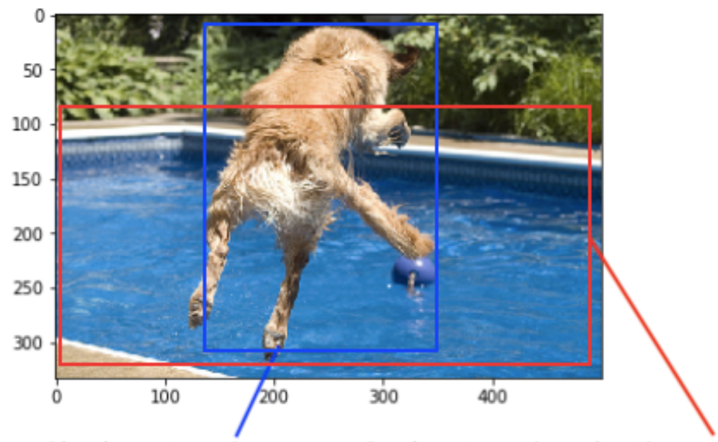


Figure RESULTS_06

prediction: two dogs are playing together in the water

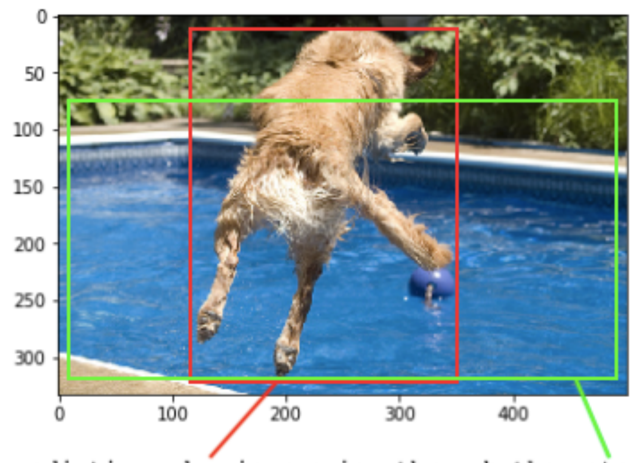


Figure RESULTS_07

prediction: dog is running through the water

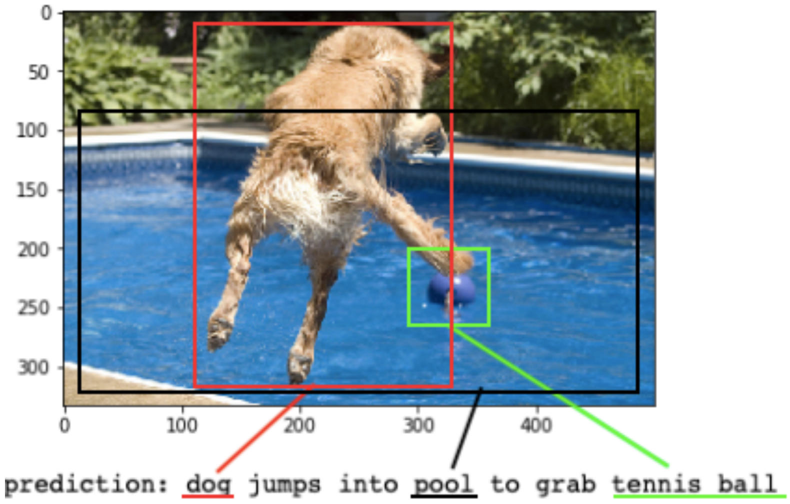


Figure RESULTS_08

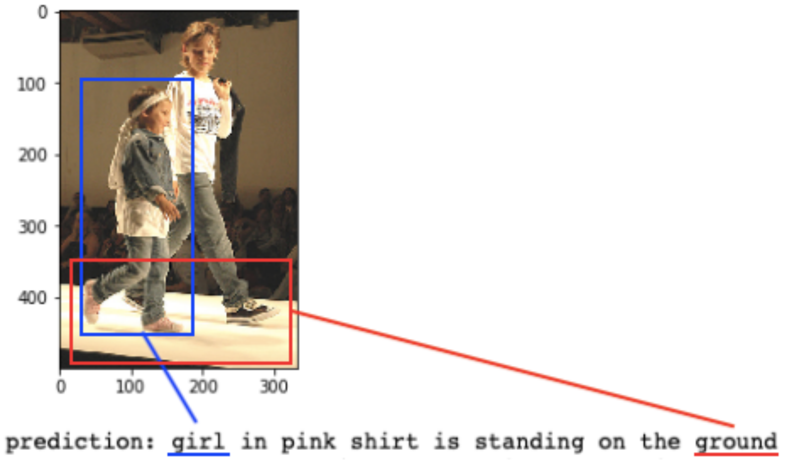


Figure RESULTS_09

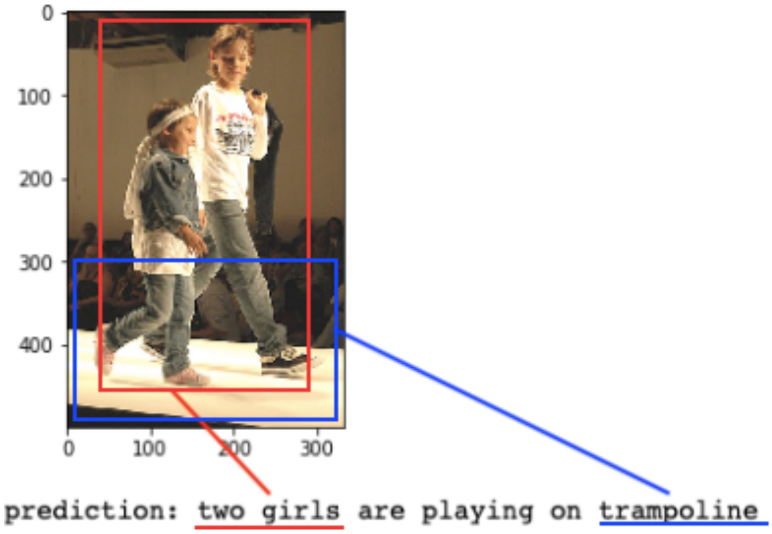


Figure RESULTS_10

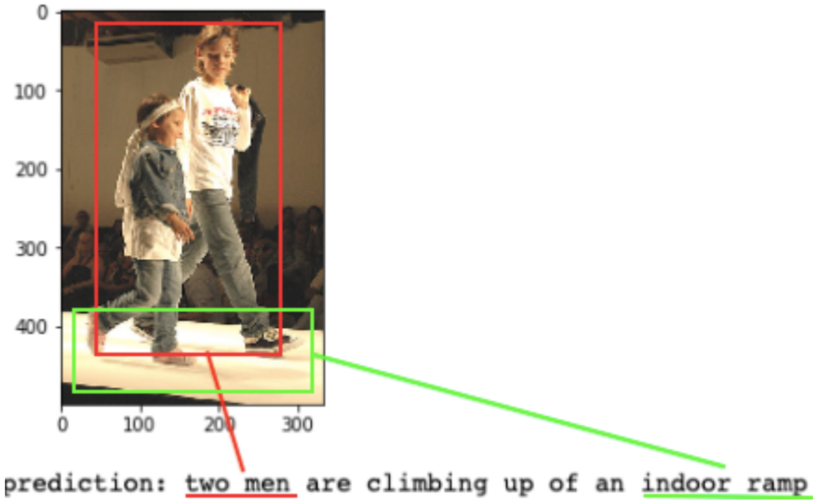


Figure RESULTS_11

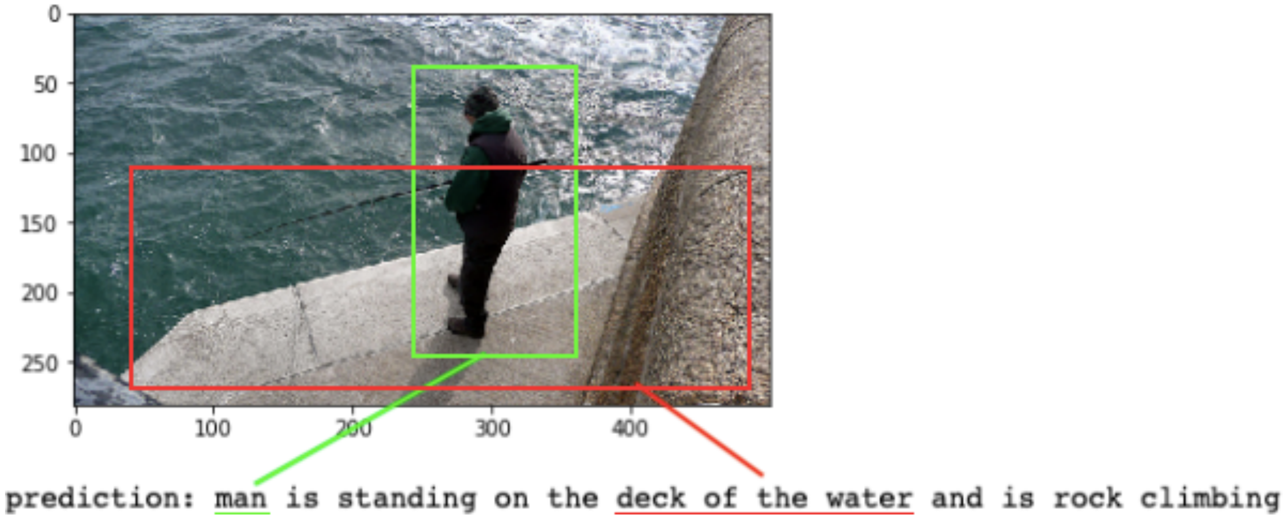


Figure RESULTS_12

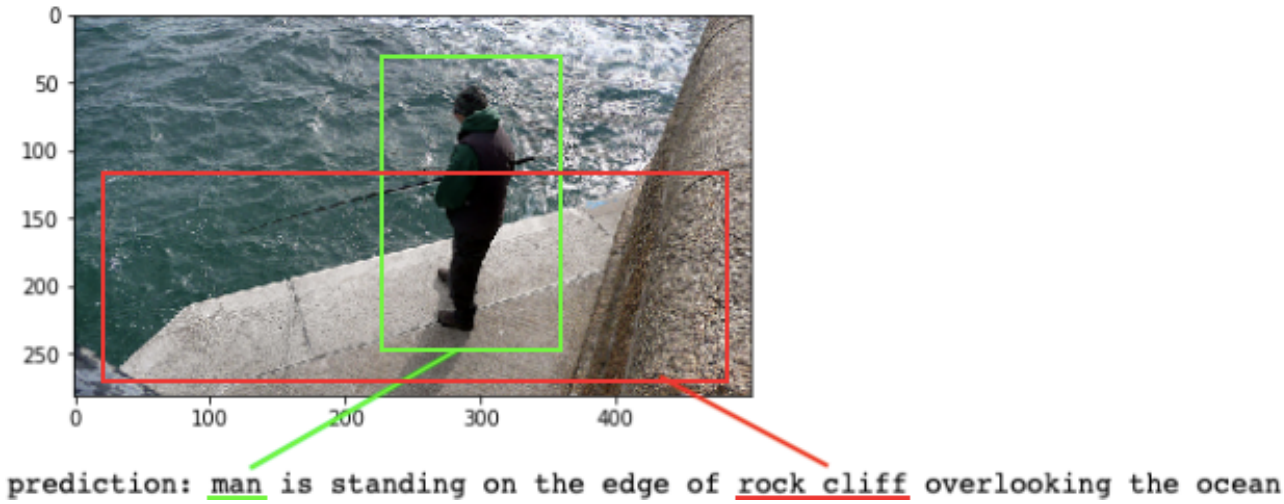


Figure RESULTS_13

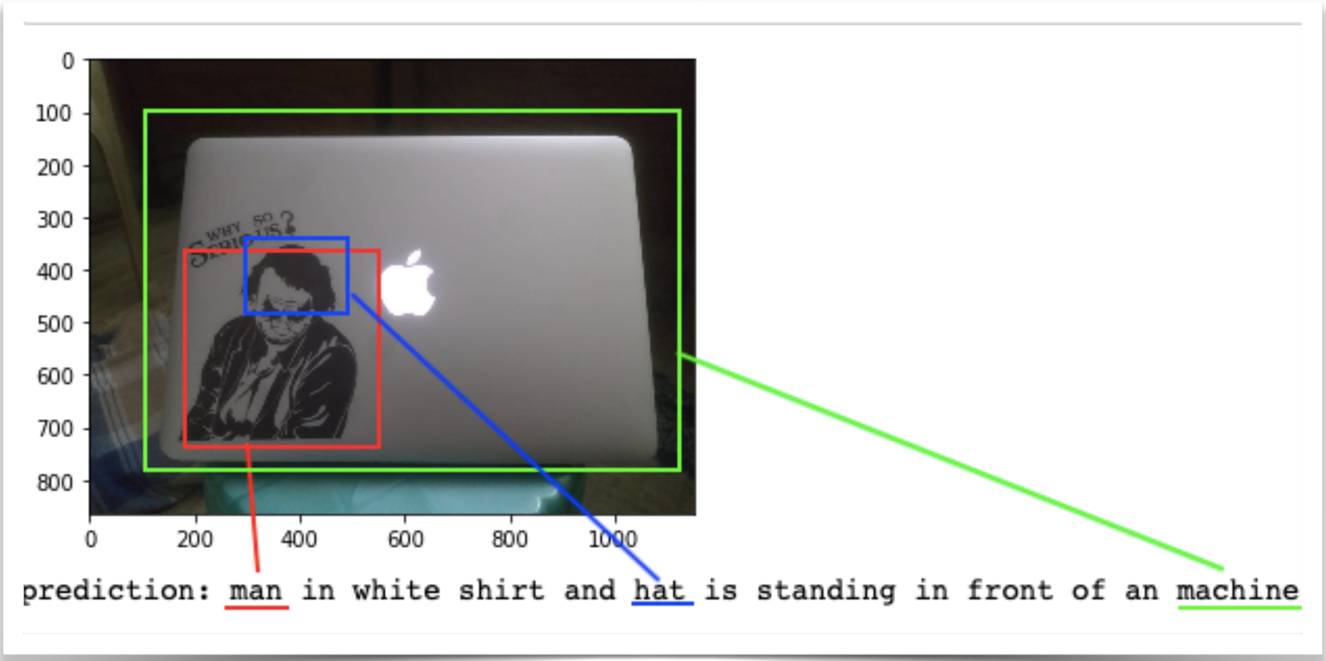


Figure RESULTS_14

LEARNING AND OUTCOME

First there was much of theoretical learning and after that there was practical learning, as I was going on implementing ideas that popped in my head.

The theoretical learning involves understanding of the present state of the art model on image captioning. And implementing it away. A lot of terminology related to deep learning was learnt along the way. Like Convolutional Neural Network and their importance in Computer Vision. Standard deep learning neural network wont be much of our friend when dealing with Computer Vision problem statements. Convolution is a revolutionary idea, instead having populated neurons in each layer, there is convolution matrix, which is convolved with the output from the previous layer to generate output of the current layer. This convolution is element wise multiplication with padding and stride. Another major theoretical learning was as a part of learning transfer learning models used in the current task like InceptionV3, InceptionResNetV2, ResNetV2. They are all presented in the ILSVRC, a prestigious Computer vision competition. They are models which performed exceptionally well on the ImageNet dataset and won prizes. These models have hundreds of layers. Along with these layers, imagenet weights are used as transfer learning to generate feature vectors from images.

Practical learning was from implementing the model and later reiterating on it. Major learning was on how to use the Keras API, to implement various funtionalities and techniques required to solve the image captioning task.

Also, learnt about the dropout in the layers. And its relation to the performance on dev dataset. Mainly dropout is used to avoid overfitting. This model at the beginning didn't have much of an overfitting problem. But later on, as the model went on doing better on the train dataset, overfitting became significant problem. But Keras has nice was of dealing with overfitting by adding additional dropout layers in between the architecture, with out compromising much of the performance on train dataset, the perfromance on dev dataset can be increased using this technique. Various dropouts are available in the Keras API. All the dropout are learnt and understood, the well suited one for this task was Guassian dropout.

Also, learnt about the generator function. Again keras has nice way of dealing with generator functions. If the dataset is large then computing it on any memory device would be difficult, this becomes a bottle neck. So, it is better to load the data on demand. Generator function helps in achieving this. Thus enabling to load much larger datasets, even larger than the memory size.

All the technical learning apart, what I felt was there was much structuring learning too involved due course of this project. Each time come up with some idea intending to improve on the performance and apply that idea. After the implementation is done, training period. Then analysing the improvement in performance by looking at metrics and its results on fixed 100 images variation set. Then adapting new ideas, based on the results. This process is iterated over many times throughout the internship period.

SUMMARY

Deep Learning model was implemented to solve the task of image captioning, many techniques were implemented keeping in mind the efficiency of the model. The summary of the whole internship can be broken down into five segments.

The first segment was more learning segment. There was a lot of reading. My mentor assigned me with quite a lot of reading job. But that reading englightned me in the field of deep learning. Later on in the internship I was never blocked due to not knowing any terminology.

The second segment is about implementing the main model. The main model consisted of implementing the state of the art solution for image captioning. This was not an easy task. Daily I would learn about some part of the implementation and then implement it. Took a good amount of time just for that implementation.

The third segment is iterating on the architecture. Many changes were made to the previously implemented architecture on number of layers, types of layers, neurons in each layers, depth of the neural network and so on.

The fourth segment is iterating on the transfer learning part, to ultimately choose InceptionResNetV2. First the implementation used was basic InceptionV3 model for transfer learning. Later learned that this model is runner up in ILSVRC 2015. So, next adapted to the winner of ILSVRC 2015 model. ResNet152V2 which was a later version of the winner model. Then again tried the combination of both first and second models. InceptionResNetV2, this worked out to be pretty good.

The fifth segment is iterating on the droput layers. The problem that was spotted by now was that the model is overfitting to the train dataset. So, multiple kinds of dropouts are tried to finalise which one works better. The winner was Guassian dropout. It had pretty good results.

Due to the time constraint, later the improvements could not be made. Also the idea of using larger dataset like Fickr30k or MS COCO was dropped due to lack of infrastructure requirements.

FUTURE SCOPE

This work can be aid to the blind. By efficiently captioning the field of view of the blind, it can be helpful for navigating to places.

CCTV cameras are everywhere today, but they require human effort as part of careful monitoring. But there might be times where human can make mistakes and help is needed immediately. Deep learning model created as the part of this internship can be deployed in the CCTV's, it helps to detect any anamoly without human intervention. Anamoly might include burglary, accident, street fights etc. Help can reach immediately as the model triggers the preset alarm.

Robots can make use of this model to interact with humans efficiently. Research to integrate A.I. into robots in currently undergoing. A part of the task of the robot is to interact with humans. If they can caption any scene that is captured by its camera, describing the scene is done. All that the robot can see is the camera feed, continuously captioning the feed can give the robot power to talk to humans in more humanly fashion.

IMPROVEMENTS

Current dataset consists of 8000 images. This can be increased to improve the performance of the model. Deep Learning models are generally data hungry, the more data you feed in the better performance it yields. Due to the limited infrastructure requirements using larger dataset was not possible during the internship phase.

Also, the architecture of the model can be increased currently the model has roughly 174 layers(InceptionResNetV2's 164 layers + model's 10 layers). Using larger image classification model for transfer learning instead of InceptionResNetV2 can improve the performance by some amout. But for that the size of the dataset also needs to be increased. Combined increase of them can improve the performance by some significant amount. More interesting part is the model, it currently has 10 layers. Increasing its architecture by adding suitable layers is proved to improve performance. This was an observation made in the course of the internship.

To get the vector representation of sequence of words, glove-6B-200d dataset is used. Representation of each word in this dataset is 200 dimensional. A better representation of the words would be to use glove-6B-300d dataset, in which, each word is a 300 dimensional vector. Interpreting the words would become better by this improvement.

Metrics is generally very well applicable to any deep learning or infact any machine learning model. But this problem statement's efficiency cannot be very well captured by the metrics. The metric followed in the work is the value of the loss function, which captures the performance of the model only to some extent. Using BLEU, CIDEr, ROUGE, SPICE, METEOR scores as metrics can be helpful to evaluate the models to better extent.

Attention model can also be employed in addition to the current model to yie

WORK

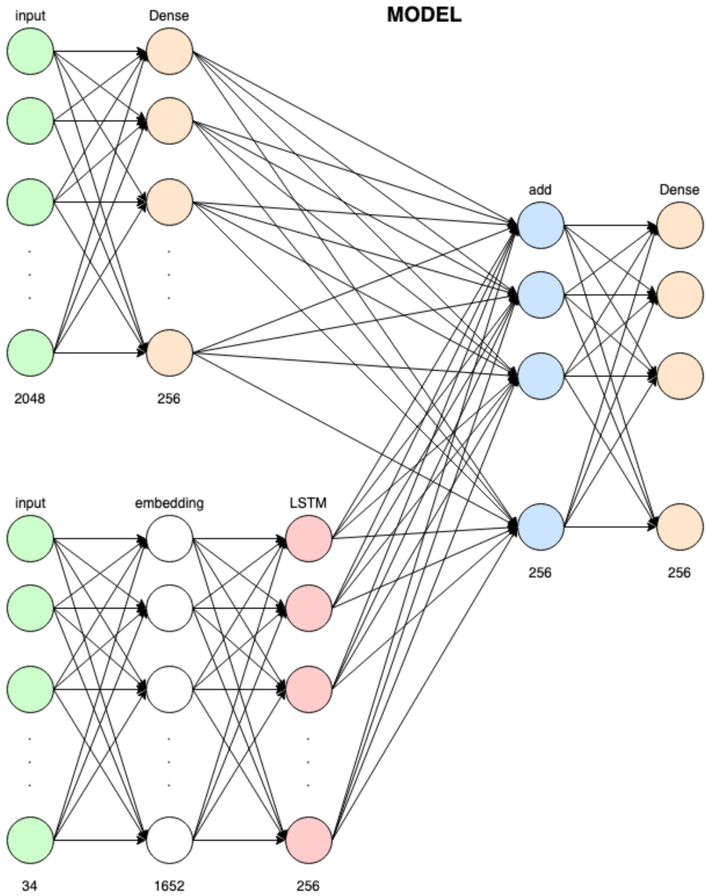


Figure MODEL_ARCH_01

LOSS ON TRAIN DATASET: 2.369915008544922
LOSS ON DEV DATASET: 3.4760239124298096

Figure MODEL_PERF_01

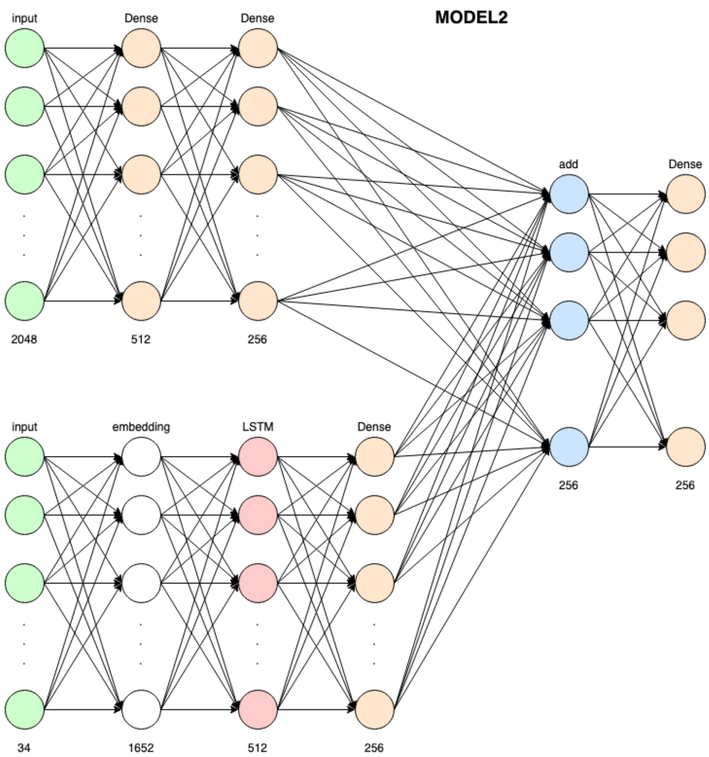


Figure MODEL_ARCH_02

LOSS ON TRAIN DATASET: 2.0370585918426514
LOSS ON DEV DATASET: 3.815887928009033

Figure MODEL_PERF_02

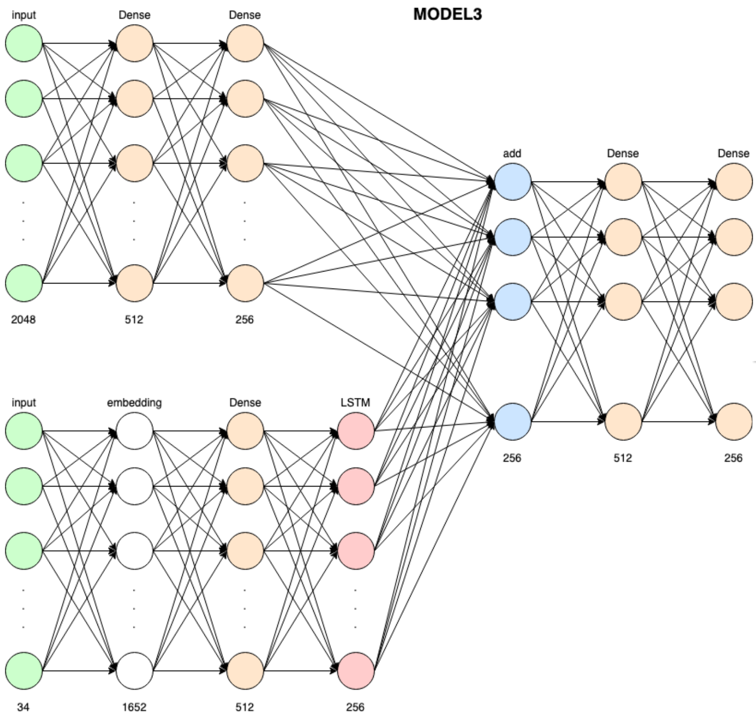


Figure MODEL_ARCH_03

```
LOSS ON TRAIN DATASET: 2.3693161010742188
LOSS ON DEV DATASET: 3.887399435043335
```

Figure MODEL_PERF_03

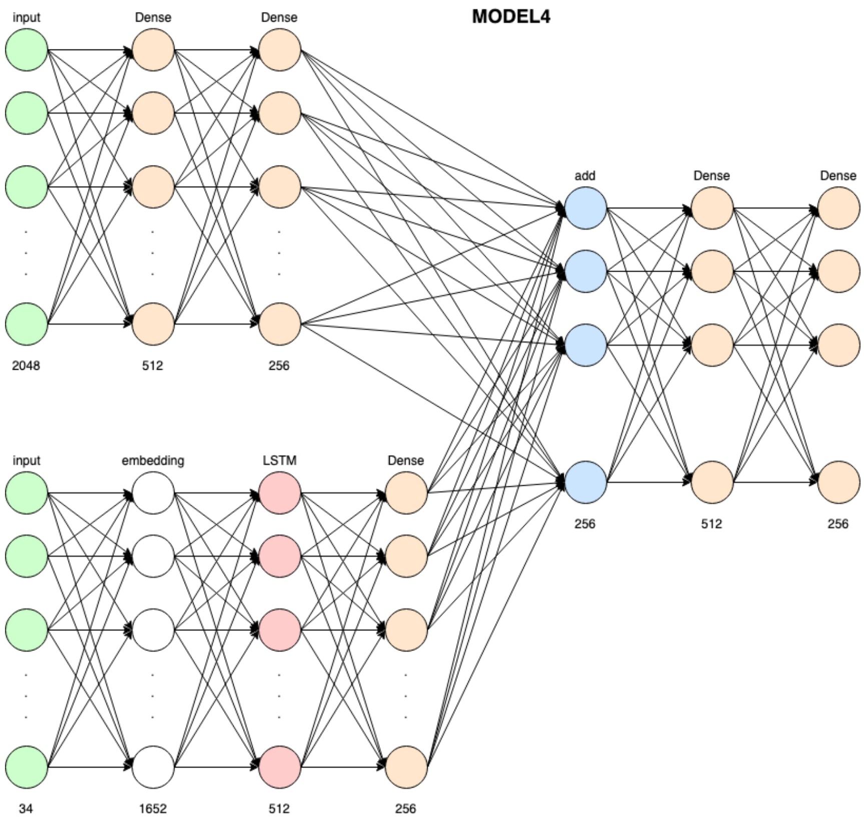


Figure MODEL_ARCH_04

```
LOSS ON TRAIN DATASET: 1.908781886100769
LOSS ON DEV DATASET: 4.075953483581543
```

Figure MODEL_PERF_04

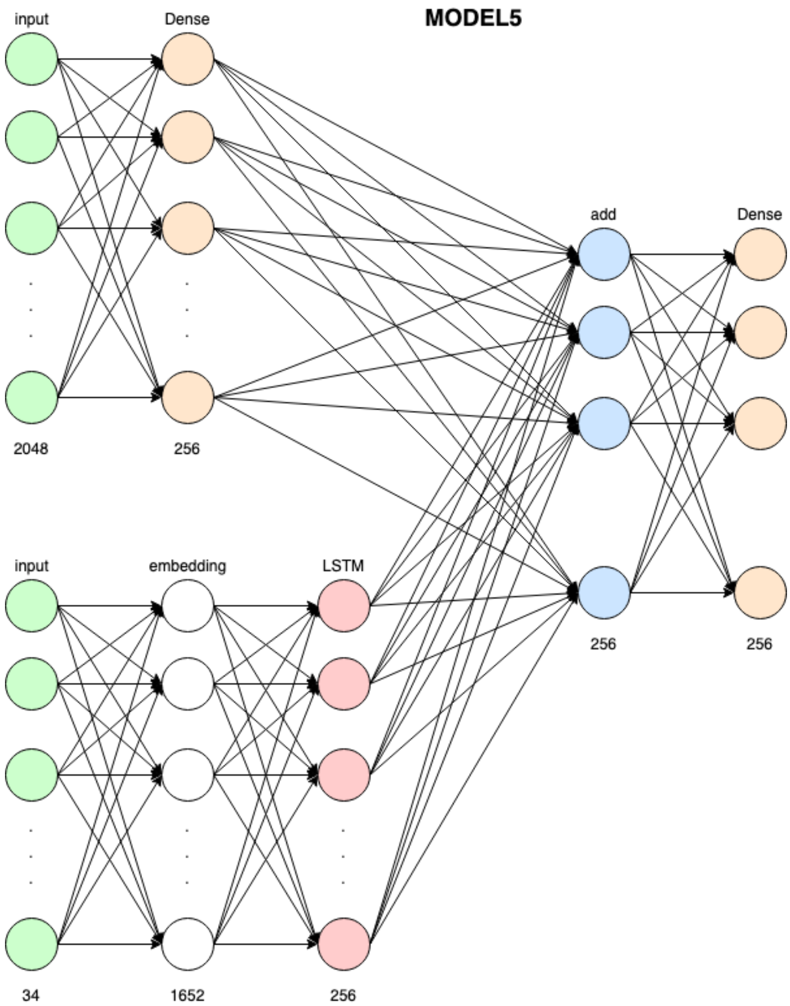


Figure MODEL_ARCH_05

```
LOSS ON TRAIN DATASET: [2.3054563999176025, 0.45
LOSS ON DEV DATASET: [3.6810479164123535, 0.330
```

Figure MODEL_PERF_05

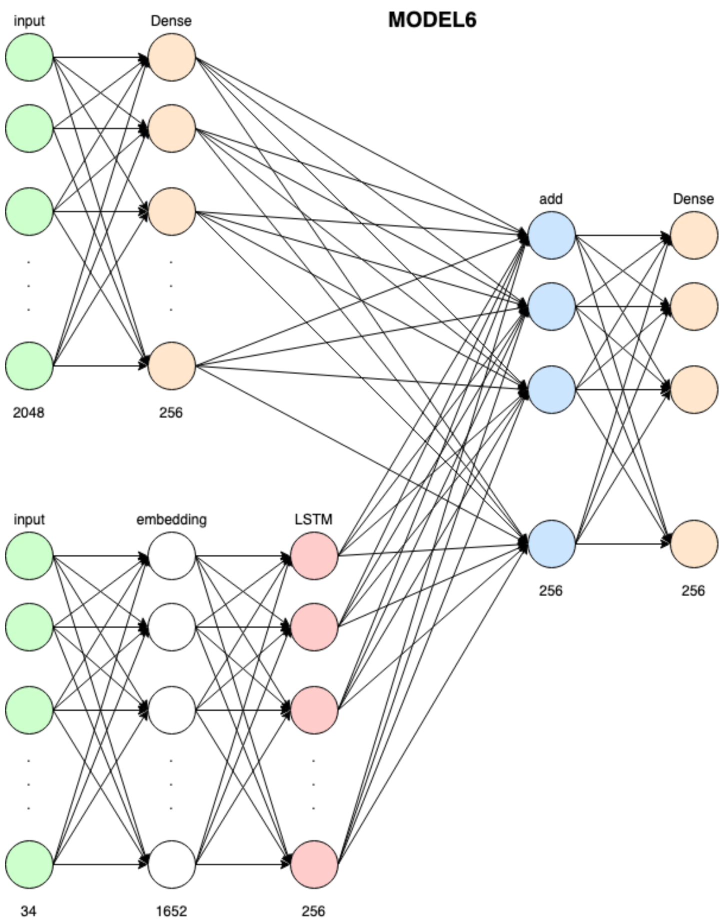


Figure MODEL_ARCH_06

```
LOSS ON TRAIN DATASET: [2.261608600616455, 0.46147
LOSS ON DEV DATASET: [4.086386203765869, 0.3352355
```

Figure MODEL_PERF_06

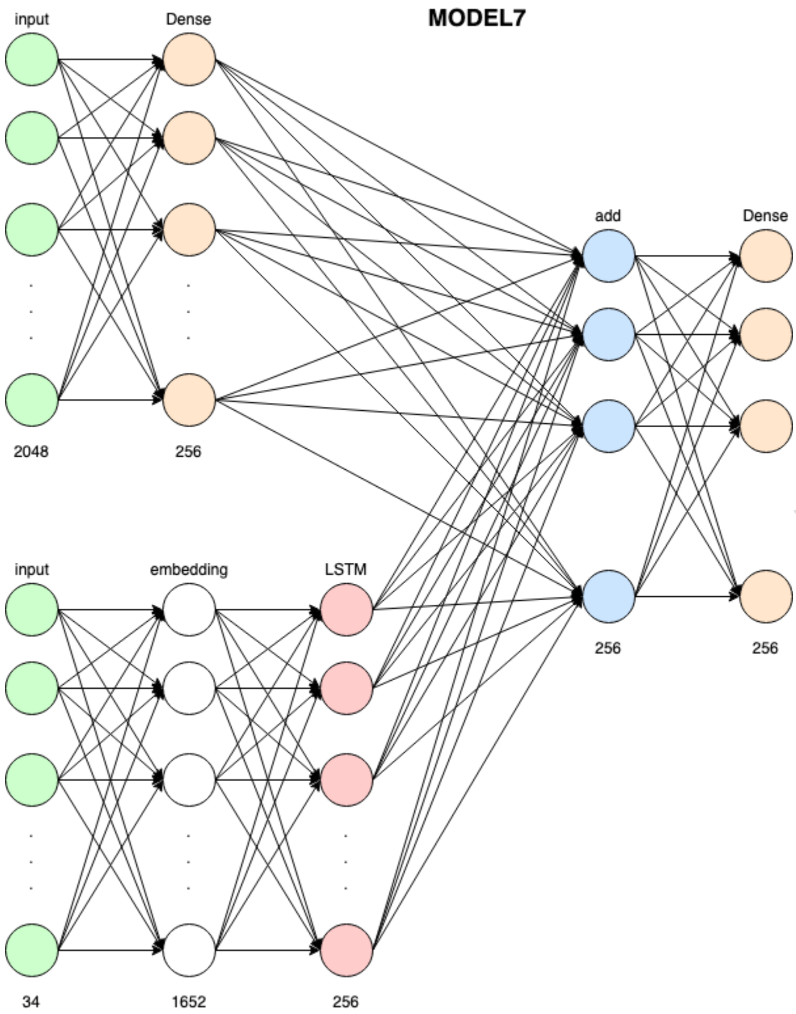


Figure MODEL_ARCH_07

```
LOSS ON TRAIN DATASET: [2.672255039215088, 0.42
LOSS ON DEV DATASET: [4.131709098815918, 0.3226
```

Figure MODEL_PERF_07

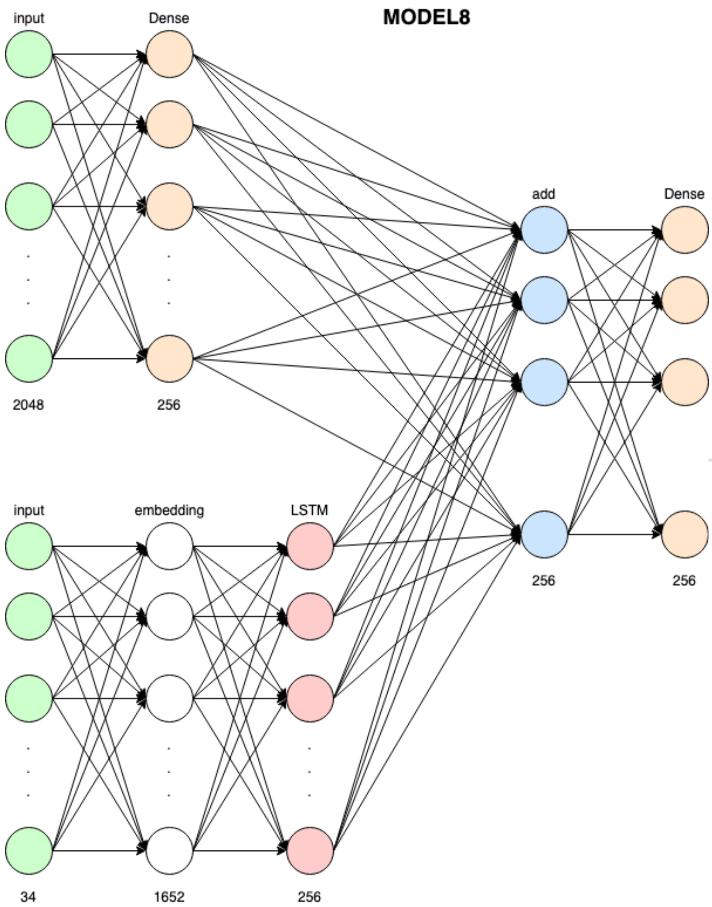


Figure MODEL_ARCH_08

```
LOSS ON TRAIN DATASET: [2.2827000617980957, 0.46
LOSS ON DEV DATASET: [3.8935625553131104, 0.3331
```

Figure MODEL_PERF_08

CERTIFICATE

July 31, 2020

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Sai Vipul Mohan (UID: 153978) from Indian Institute of Technology, Palakkad has completed internship in UST from June 08, 2020 to July 31, 2020.

Sai Vipul Mohan's performance was good during this period and he/she has successfully completed his/her internship training at UST Global on Image Captioning.

UST Global wishes him/her all the success for his/her future endeavours.

For US Technology International Pvt. Ltd

Vinesh George
Group Manager – Human Resources

US Technology International Pvt. Ltd. #721, Nila, Technopark, Trivandrum-695 581, India, T: 91.471. 4040000, F: 91.471.2527276, www.ust-global.com
CIN: U72200KL1999PTC013385