CSPCI-400 Project

Video Captioning Bot

SUBMITTED BY:

Kartik Karira(18103051)

Rohit Mittal (18103081)

MENTORED BY:

Dr. Avtar Singh

(Assistant Professor)

CSE Department

Department of Computer Science and Engineering

Dr B R Ambedkar National Institute of Technology Jalandhar

Table of Contents

OVERVIEW

Developing a Machine Learning based approach using OpenCV and python libraries which could provide suitable captions for actions happening in video/gif given as input.

Problem Statement

Dealing with sight loss, already, is a challenge in itself.Individuals with vision impairment are also more likely to experience restrictions in their independence, mobility as well as an increased risk of falls, fractures, injuries.

So if we could have a system or a assistant which will pronounce everything what's happening in the environment So that the risk of falls, fractures, injuries is minimized.

Proposed Solution

We propose a system having Deep Learning model using which blinds can know what they is happening in their surroundings.In this we will record videos of about 2-3 seconds and then provide suitable captions to them using our model and then pronounce it to the person using our system. In this way the blind person will be able to know everything happening around him.

Tech Stack

Dataset Description

For the purpose of this study, we are using the MSVD(Microsoft Research Video Description Corpus) and MSR-VTT(Microsoft Research Video to Text) dataset created by Microsoft.

MSVD:

This dataset has about 1550 videos in which 1450 videos are used for training and validation and remaining 100 for the testing purpose.

MSR-VTT:

MSR-VTT is a large-scale dataset for the open domain video captioning, which consists of 10,000 video clips from 20 categories, and each video clip is annotated with 20 English sentences.

"caption": [
"A boy is playing a key-board between the people.",
"A boy is playing a piano in front of a crowd.",
"A boy is playing a piano.",
"A boy plays a piano for a group of kids.",
"A boy plays the piano.",
"A kid is playing a piano.",
"A young boy is playing a piano in front of a crowd of other young people.",
"A young boy is playing the piano before an audience.",
"A young boy is playing the piano.",
"A young boy seated on stage is playing a piano as the audience watches him.",
"The boy is playing the piano.",
"The boy performed on the piano for an audience.",
"The boy performed on the piano for the audience."
]

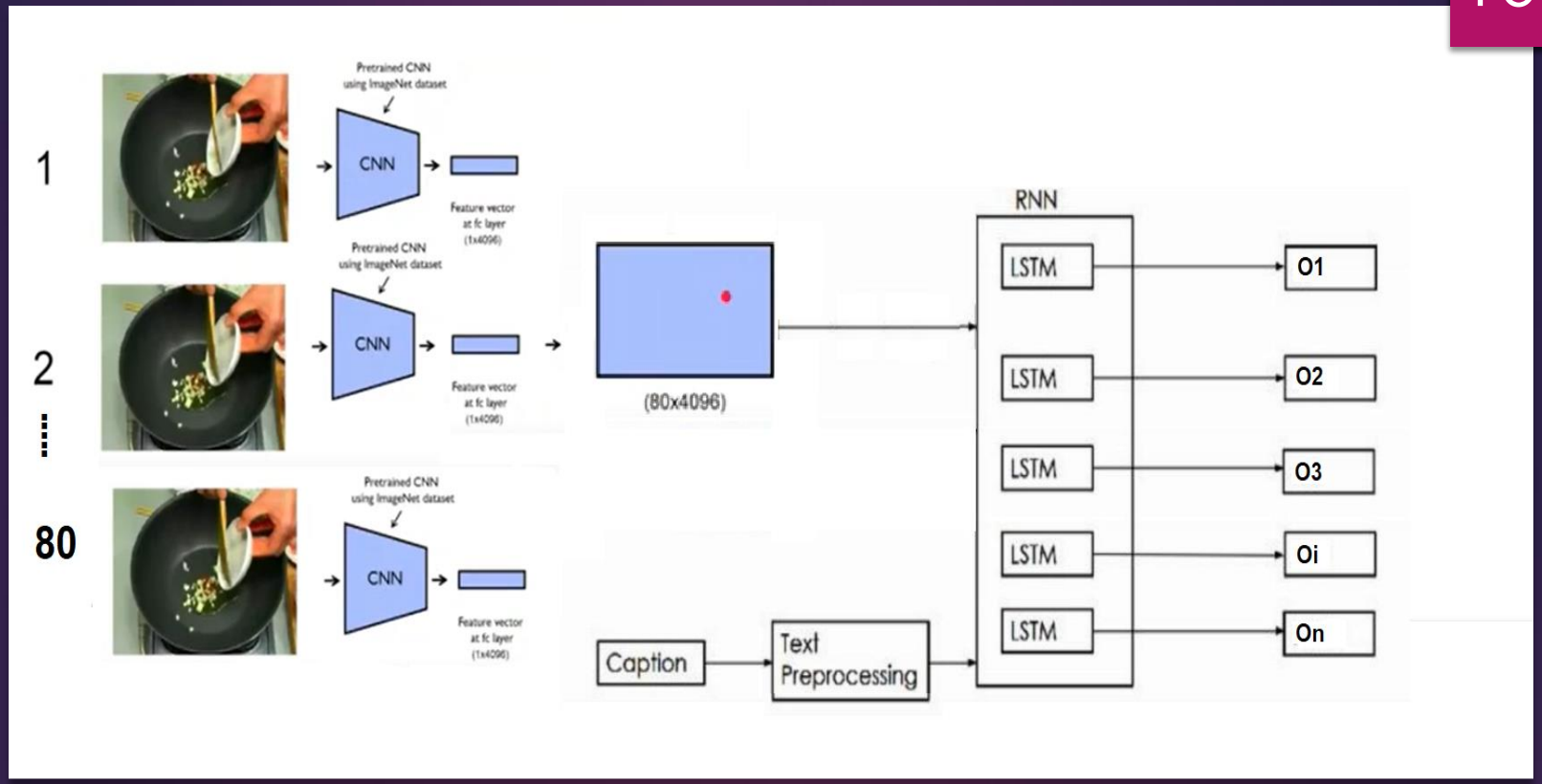
Example

Machine Learning Model: Architecture

Here we are using Multi-Model architecture composed of:

- RNN(Recurrent Neural Network)
- CNN(Convolutional Neural Network)

where CNN will be used for the feature extraction from the video and then the result is fed into a RNN for predicting a suitable caption for it.


1
2
80
Pretrained CNN using ImageNet dataset
CNN
Feature vector at fc layer (1x4096)
(80x4096)
RNN
LSTM
LSTM
LSTM
LSTM
LSTM
O1
O2
O3
Oi
On
Caption
Text Preprocessing

Flow

Concepts

Convolution Neural Network used in VGG16.

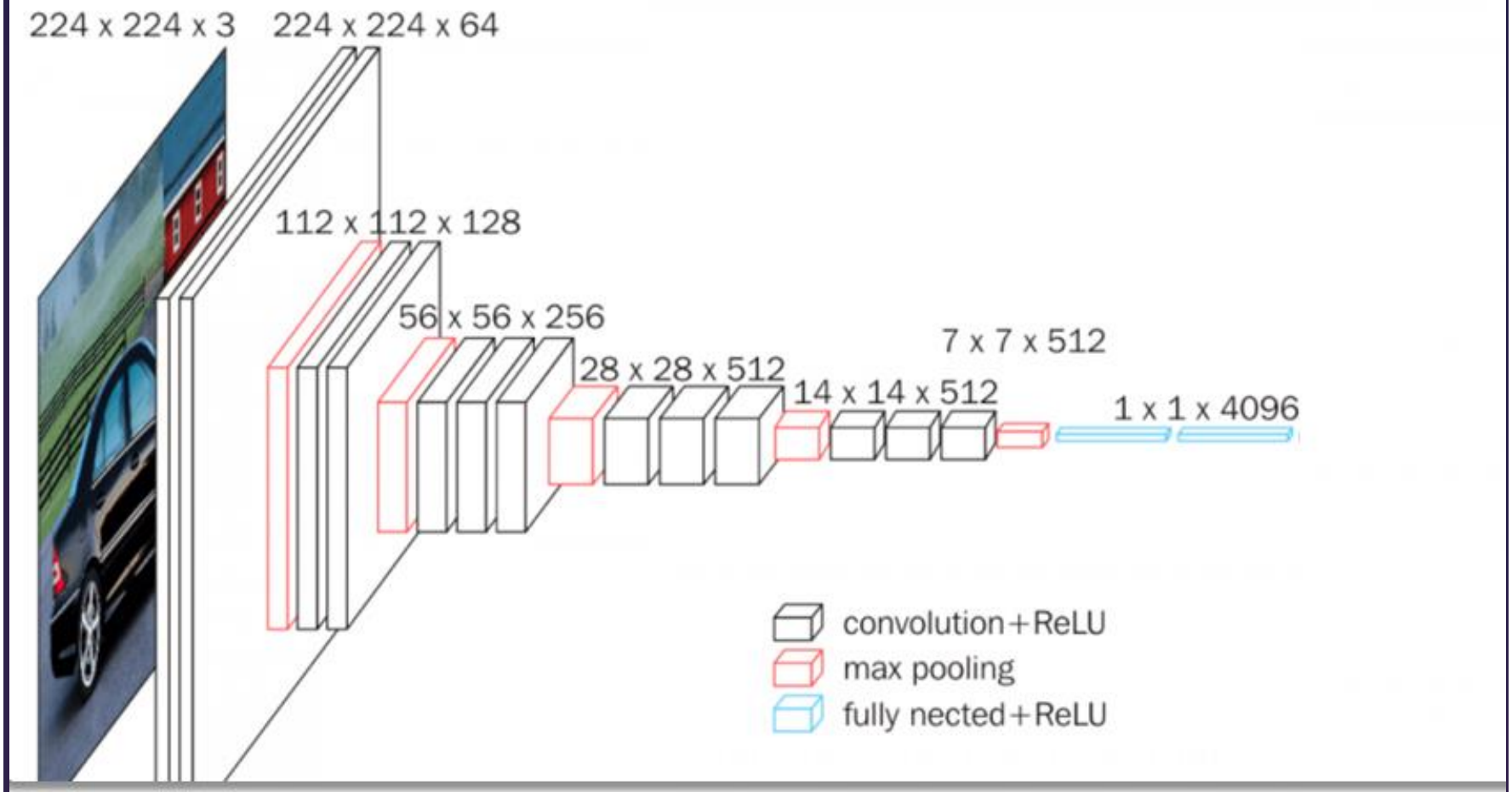
Long Short-Term Memory Recurrent Neural Network.

Text-Processing.

Bleu Score For Textual Evaluation.

CONVOLUTIONAL NEURAL NETWORK (CNN)

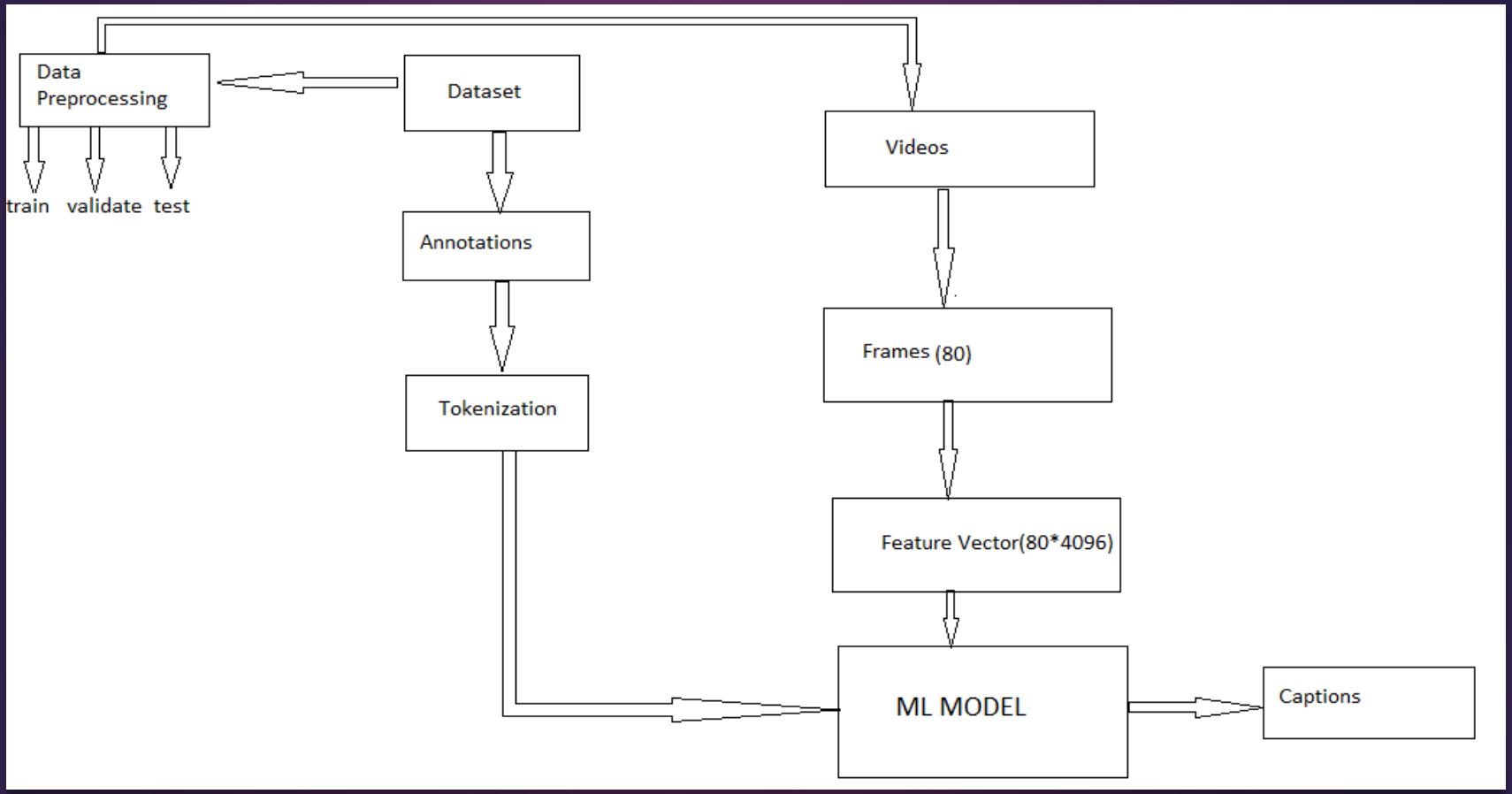
- CNN Are Very Similar To Ordinary Neural Networks.
- Convnet Architectures Make The Explicit Assumption That The Inputs Are Images, Which Allows Us To Encode Certain Properties Into The Architecture.
- It Has Filters, Pooling Layers Because Of The Assumption That Input Is Always An Image.
- For Our Project We Are Using Vgg16 For Extracting The Features From Each Frame And Then Combining All The Frames Result For Our Work.



VGG16(VISUAL GEOMETRY GROUP)

RNN AND LSTM

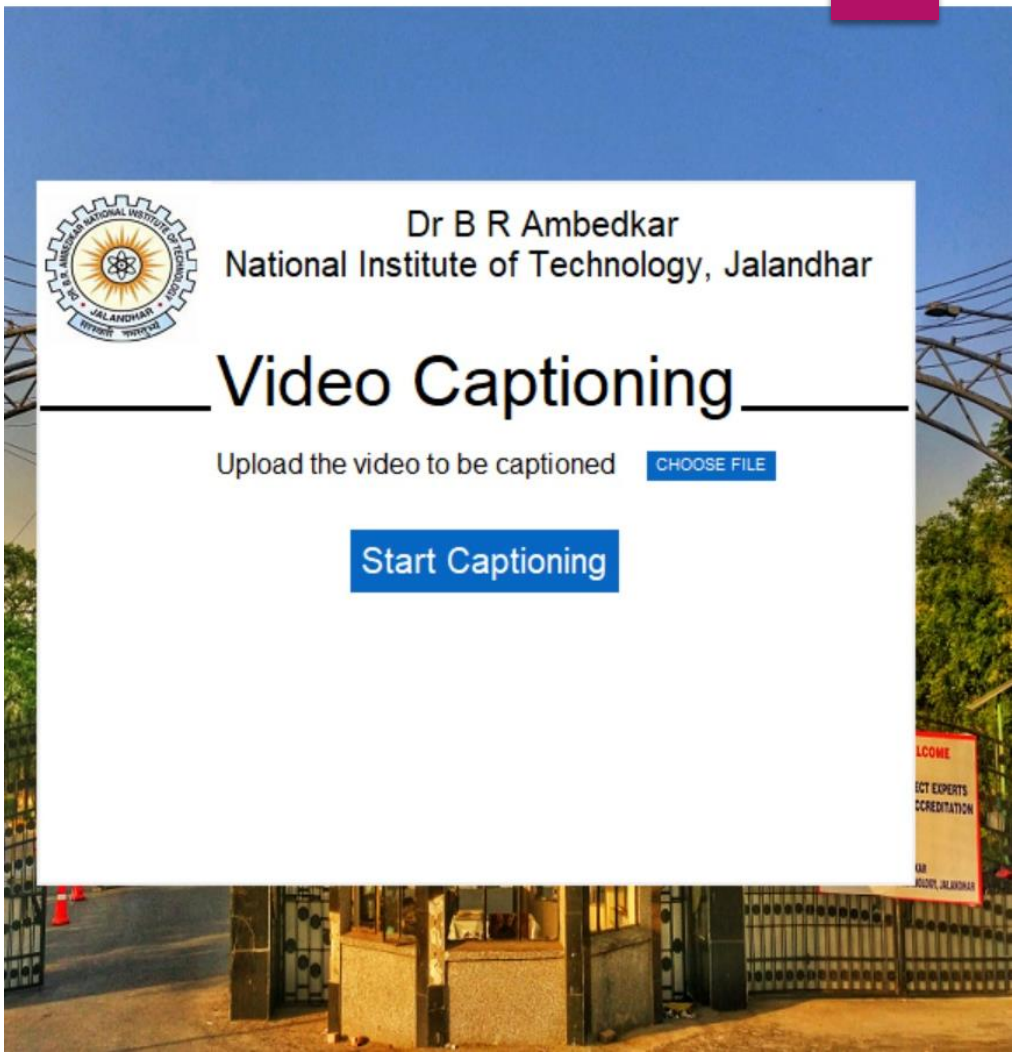
- Recurrent Neural Networks Are The State Of The Art Algorithm For Sequential Data And Among Others Used By Apples Siri And Googles Voice Search.
- It Has An Internal Memory Which Makes It Perfectly Suited For Machine Learning Problems That Involve Sequential Data Because It Can Remembers It's Input.
- For Our Project We Use LSTM As We Require A Sequence Of Words As An Output


Data Preprocessing
Dataset
Videos
train
validate
test
Annotations
Frames (80)
Tokenization
Feature Vector(80*4096)
ML MODEL
Captions

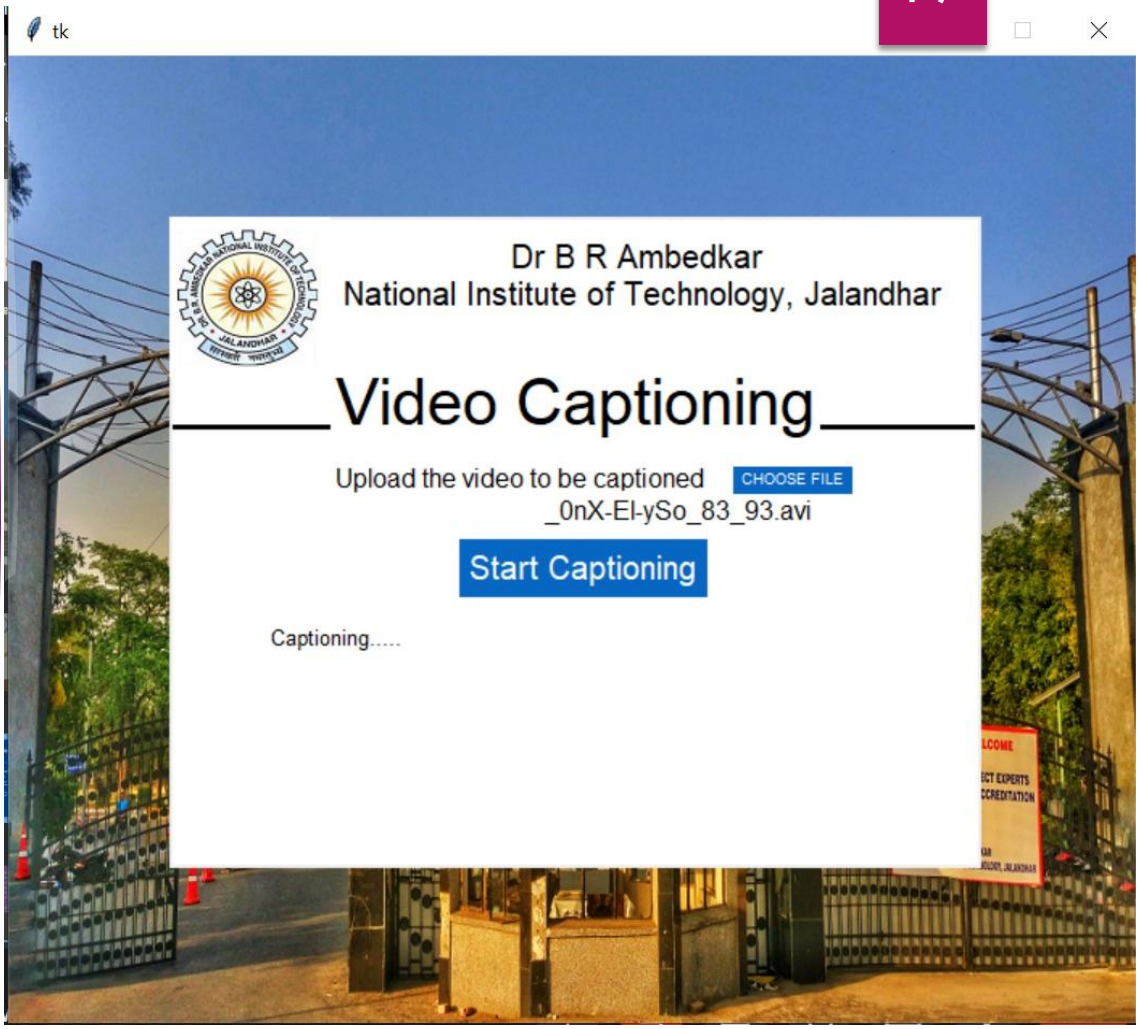
Working

User Interface

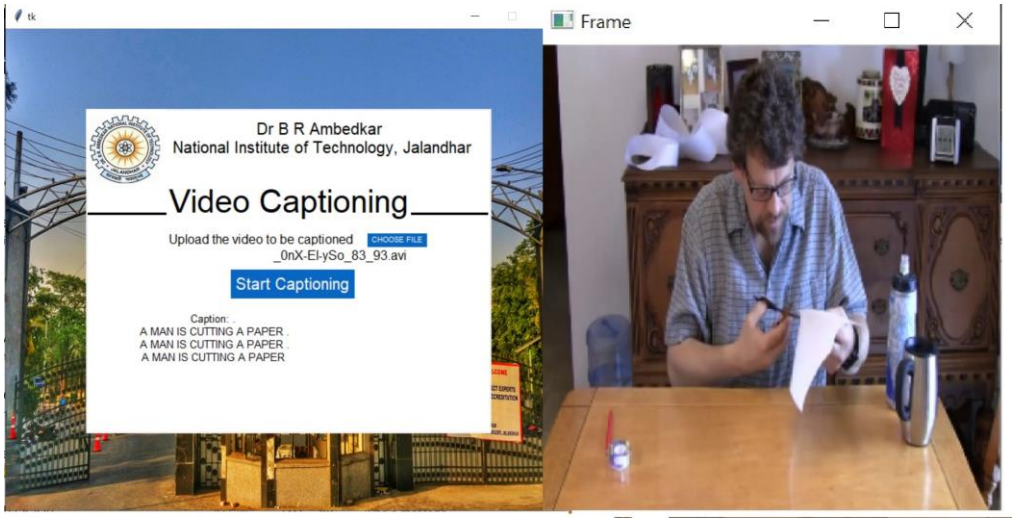
- On running the front.py the GUI below will pop up on screen.



- Now click on choose file and select the video to be captioned and click on Start Captioning.

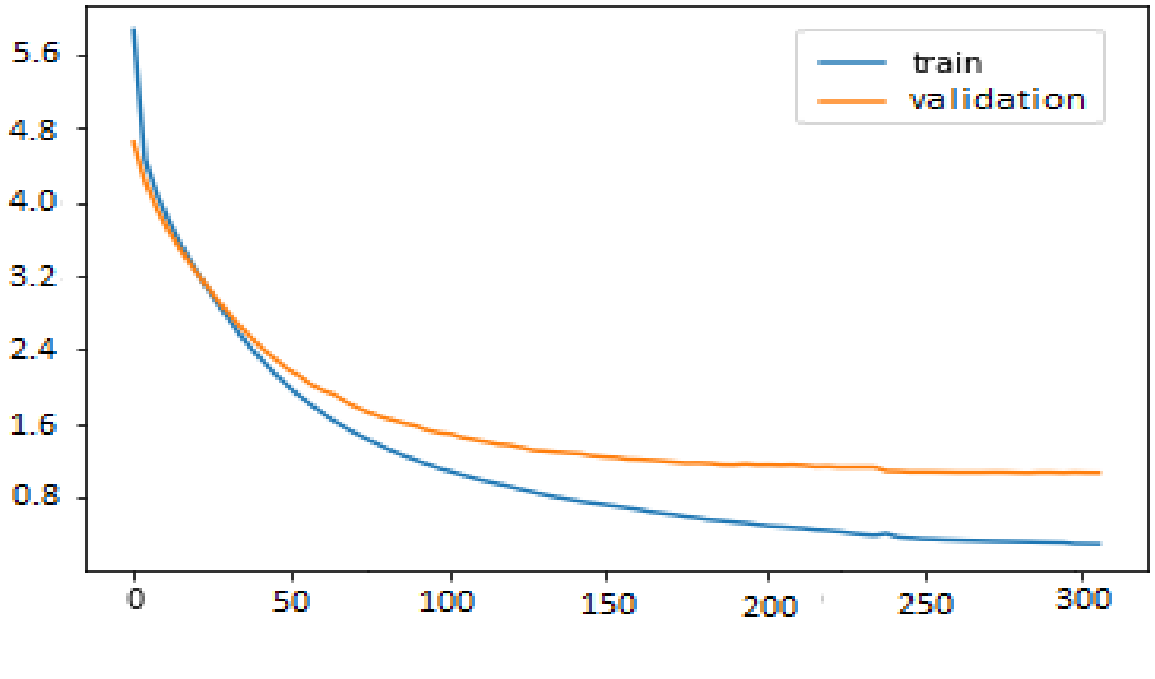


- The output will show the captions generated by the model and will run the video in background and also the caption will also be spoken out



Results

- We tracked the model's loss on the training set, as well as its performance on the validation set, during training.



Continued....

- For quantitative evaluation we have calculated BLEU and METEOR score which comes out to be:

```
Bleu Score is 0.8883465596797255
Meteor Score is 0.9294392357928086
```

Testing Results

- We will look at some of the result captioning

Caption: .
A GROUP OF PEOPLE ARE DANCING

Caption: .
. A baby is eating spaghetti .

Caption
A cat is playing the piano


Thank
you