Video Captioning Bot

A PROJECT REPORT

submitted in fulfilment of requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

KARTIK KARIRA (18103051)

ROHIT MITTAL (18103081)

under the supervision of

DR. AVTAR SINGH
ASSISTANT PROFESSOR
CSE DEPARTMENT

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DR. B.R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY,
JALANDHAR-144011, PUNJAB (INDIA)

MAY 2022

ACKNOWLEDGEMENT

It is true that hundreds of people work behind the scenes to make a Play a success. We'd want to thank everyone who helped us finish our final project-**Video Captioning Bot** with our sincere gratitude. We ran into a several problems during the project due to a lack of information and competence, but these people helped us overcome these difficulties and develop our idea for shaping sculpture.

We'd like to thank thankful to Professor A.L Sangal, Head, Department of Computer Science & Engineering, for giving his leadership, guidance and all his direct and indirect support, which allowed the entire team to grasp every facet of the project.

We are thankful to the In-charge, Major Project Final Year, for providing us mentor and all other support.

We'd also want to thank Dr Avtar Singh, Assistant Professor, our mentor, who believed in our idea and offered fresh suggestions as needed and also provided his continuous support and monitoring throughout the project.

We are very grateful to get the constant guidance and help from all the Department of Computer Science & Engineering who gave us their valuable time and suggestions whenever we required. Hence were all one of the most important part of our project team. Also, we would like to extend our sincere esteems to all staff in laboratory for their timely support.

Thank you

ABSTRACT

With increase in Internet the videos are becoming an integral part of life and hence understanding it has become very important for purposes like security, social purposes etc. Video captioning provides a very simple and easy way to find the actions in a video and hence summarizing it which can be used for multiple purposes like better searching and indexing. In our work we have introduced a model which first will extract features from the videos using Convolutional Neural Network (CNN) and then using Long Short-term memory (LSTM) try to provide suitable captions to the video clips provided to it.

DECLARATION

*We herewith certify that the work that is being done within the project report entitled "***Video Captioning Bot***" in partial fulfilment of necessities for the award of degree of B.Tech. (Computer Science and Engineering.) submitted to the* ***Department of Computer Science and Engineering*** *of* ***Dr B R Ambedkar National Institute of Technology, Jalandhar****, is an authentic record of our own work carried out during a period from July, 2021 to May, 2022 under the supervision of* ***Dr Avtar Singh, Assistant Professor****. The matter presented in this dissertation has not been submitted by me in any other University/Institute for the award of any degree.*

Kartik Karira	*Rohit Mittal*
(18103051)	*(18103081)*

This is to certify that the above statement made by the candidates is correct and true to the best of my knowledge

Signature of Supervisor
Dr. Avtar Singh
Assistant Professor
(Dept. of CSE)
NIT Jalandhar

Thank you all.

Table Of Contents

FIGURES USED

1. **INTRODUCTION**

1.1. Background

With increase in internet connectivity and use of platforms like YouTube, Instagram reels, the videos has become an integral part of our lives making the task of video captioning very popular in present time and many different ways are being used to get a suitable captions for the video. According to reports about 1 billion videos are watched every day and more than 100 hours of videos uploaded every minute in the world. With such developments the machine learning products for videos have become a high priority requirement of the world.

1.2. Computer Vision

Computer visualization is a machine learning environment dedicated to the interpretation and comprehension of images and videos. It is basically used to provide eyes to the computers i.e., they can have information about how to get the visual information and how to use it for performing different types of actions which humans do every day.

These models are built to get the visual data and then translate them in different types of features and get as much as knowledge possible from them during training which enables them to interpret and make decisions based on the images or videos provided.

Although it seems that image processing and computer view are same but it is not true. Image processing is the process of enhancement and adjustment of images or videos to get better results. It may include blurring, changing brightness, improving, or changing quality of images. While computer view only sees the original images or frames which does not require any special attention.

1.3. Machine Learning

Machine learning is becoming one of the most important and integral part of the world as it provides many advantages like providing customer behaviour trends information, providing companies how to increase their profits. Many tech hubs of the world like Google, Amazon, Facebook, Microsoft have made machine learning their integral part in their every activity in present and have been earning great profits.

1.4. Artificial Neural Networks

Artificial Neural Networks are a very important part of Machine learning especially Deep learning as it tries to mimic the nerve systems of humans providing the machines more power and greater flexibility.

They basically contain several processing layers which try to work in similar way as human nerve systems work and based on inputs try to give outputs which would human brain has provided.

1.5. Convolutional Neural Network

Convolutional neural networks are a one of the great advancements in artificial intelligence networks that use mathematical operations called convolution instead of repeating a normal matrix in at least one of them. They are made specially to process visual data and hence are used for image, video feature extraction and their related other processing

VGG16 is a deep learning model built in convolution neural net (CNN). It is one of the best models for visual information extraction. The great thing about VGG16 is that it even without focusing on large hyper parameter it got an accuracy of 92.6% for ImageNet dataset. Also, the fastest in all the other models competed against it. We are using it by removing the last layer which is just a SoftMax layer used for classification and hence we are able to get the features from images using it.

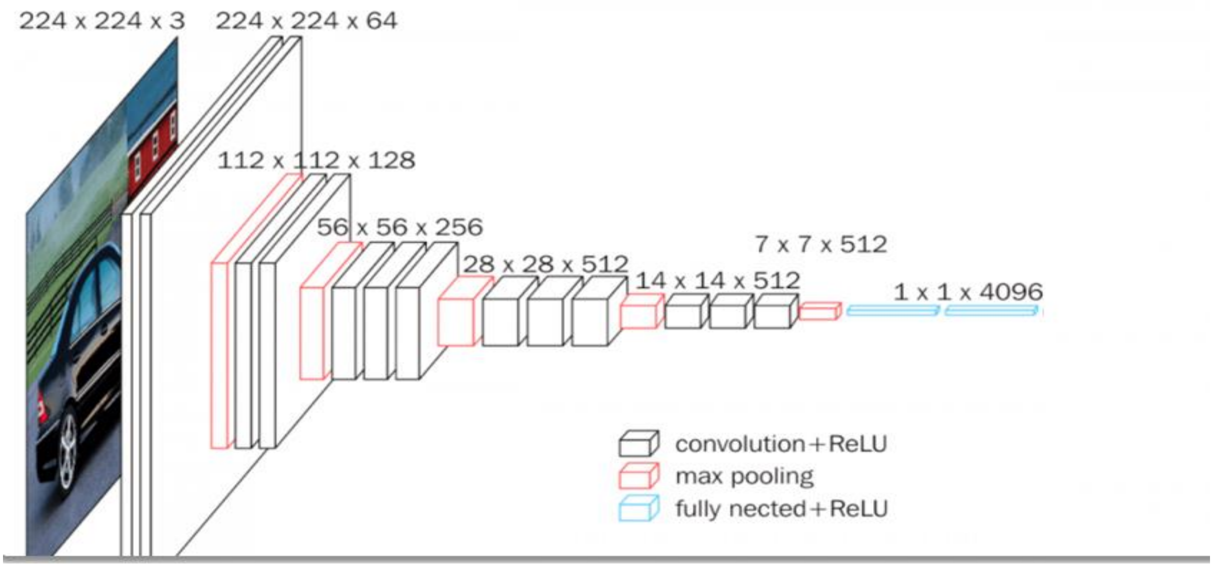


Fig1[3]

1.6. Recurrent Neural Network

Recurrent neural network (RNN) is a special type of arificial neural network in which the output of previous layer is fed as input to get the output for the next layer.

As here the mechanism is allowing us to use the sequential information present to us hence is one of the most important technique for solving problems having sequential relations.

The difference between CNN and RNN is the type of problems that can be solved by both as CNN solves problems having spatial features while RNN does for sequential ones.

But one disadvantage of RNN is the problem of vanishing descent.Hence we can not use it directly for text prediction.

LSTM:

LSTM i.e. Long Short-Term Memory is special type of RNN that with the help of some extra layers i.e.input gate layer, forget gate layer, output gate layer resolves the issue of vanishing descent and hence we can use this in our model for text prediction.

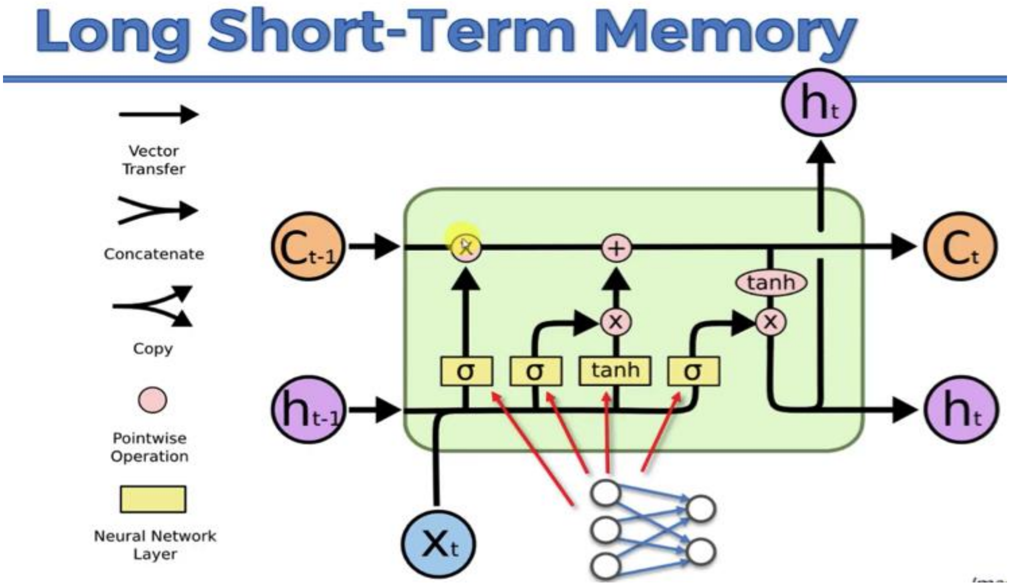


Fig 2[4]

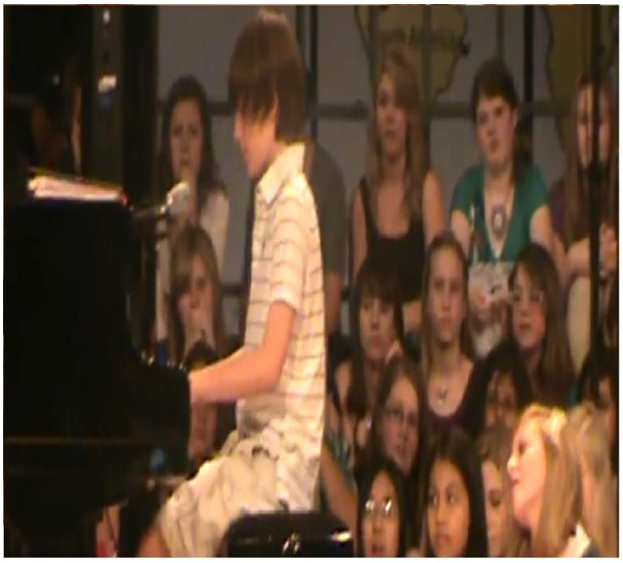
1.7. Data Representation

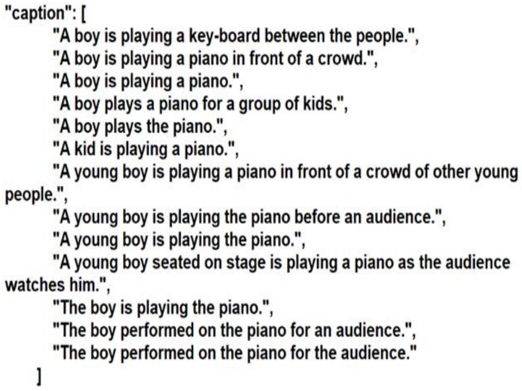
For our work, we are using the MSVD which is Microsoft Research Video Description Corpus and MSR-VTT i.e., Microsoft Research Video to Text dataset created by Microsoft.

MSVD:

The MSVD dataset contains about 1550 videos with on an average having 30 captions for every video. Hence works like a dataset having about 46500 values.

One example can be seen as





```
"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."
    ]
```

Fig 3

2. LITERATURE SURVEY

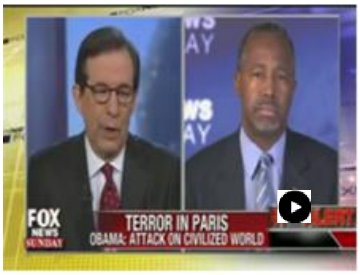


Fig 4

Can we get a suitable caption for the above video?

There may be many captions for the above video like:

- “A news reporter is reporting”
- “One man is talking”,
- “Two men are talking with each other on some topic”

and even more depending upon people to people.

Yes, all these and many other captions are relevant for the video. But the problem is that it is so easy for us as humans that just seeing a video, we can describe it in our language and provide the best caption to it. Even a very small child can do this without any difficulty. But can we have a program which can do it i.e. takes this as an video as input and give us a caption as output which is suitable for the video.

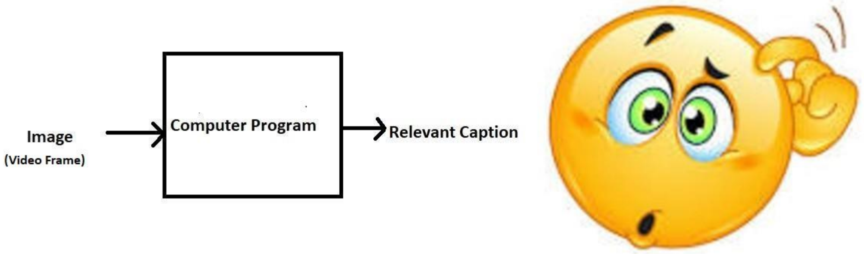


Fig 5

If, we can produce a relevant caption for image, we could also comprehend what is being happening in a small video, by dividing the video in number of frames. The people are working on this from some time one of the popular persons doing it is Andrej Karapathy who is presently working as Director of AI in Tesla

3. PROBLEM STATEMENT AND FEASABILITY

3.1 Problem Statement

For solving a problem, we must first of know what the real-world necessities of it are to be solved. Let's see few of the applications where our work could help in bigger way:

1. **Providing a virtual assistant to the blind:** Using our work and adding the text to voice model we can help the blind know what is happening around them and hence assisting them in their difficult times so they can become independent and can live a better life.
2. **Automatic cars:** One of the biggest challenges in automatic cars is getting inference from visual data. As our model is converting visual data to text hence it can provide a great boost for the automatic or self-driving cars.
3. **Making Search Better:** If there is any way we can get inference from videos i.e., describe them in text the searching could become faster and more accurate at the same time.
4. **Automatic Recommendation Systems:** If we have the captions for every video, we can easily cluster the videos based on this and hence getting better recommendation systems.

3.2. Feasibility- Technical, Non-Technical

Before doing anything, it is very crucial to have knowledge is it feasible to do or not.
The various feasibilities of our project are summed as:

- **TECHNICAL: -**
  - System with high computing and processing power.
  - Camera with good quality precision.
  - As the model will be using cloud the internet connectivity is must.
- **NON- TECHNICAL: -**
  - As development is just based on getting dataset and coding the cost of project is zero but we must deploy it on some server which will have to be paid.
  - As digital world is increasing the video importance is going to increase hence the scope of our project will be increasing in the future.

4. METHODOLOGY

4.1. 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.

4.2. DETAILED SOLUTION

Video captions are a text description of video content production. Compared to picture captions, the location is much more flexible and contains more information than a still image. Therefore, to make a description of the text, video captions need to extract a lot of features, which is much harder than image captions. The most common methods of video caption work are made up of two parts, part of the video element and part of the production of the video description. The standard format of video captions is shown in Figure below.

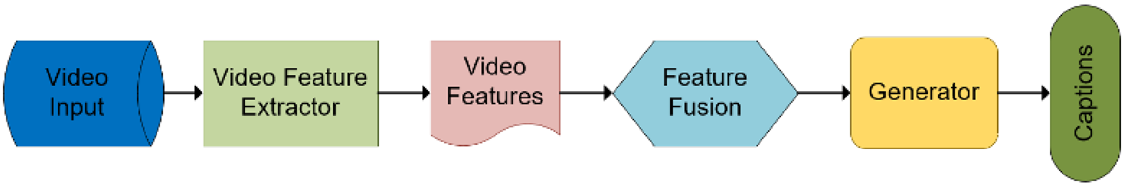


Fig 6 [3]

Video Captioning has two major Parts:

1.Extracting features from video using CNN

2.Generating valid caption for a video frame or image using RNN.

4.3. Model Architecture

Introduction: In the model the main characteristics to be followed are:

a) Authenticity: This means that the caption generated should reflect the actions in the video.

b) Nature: The nature of captions generated should be very close to the descriptions provided by the people and should follow grammar as well.

c) Diversity: The outputs for different videos should be completely different even we could have different captions for the same video as well.

In order to get these features, we propose a model having CNN and LSTM layers . The network structure is shown in Figure below.

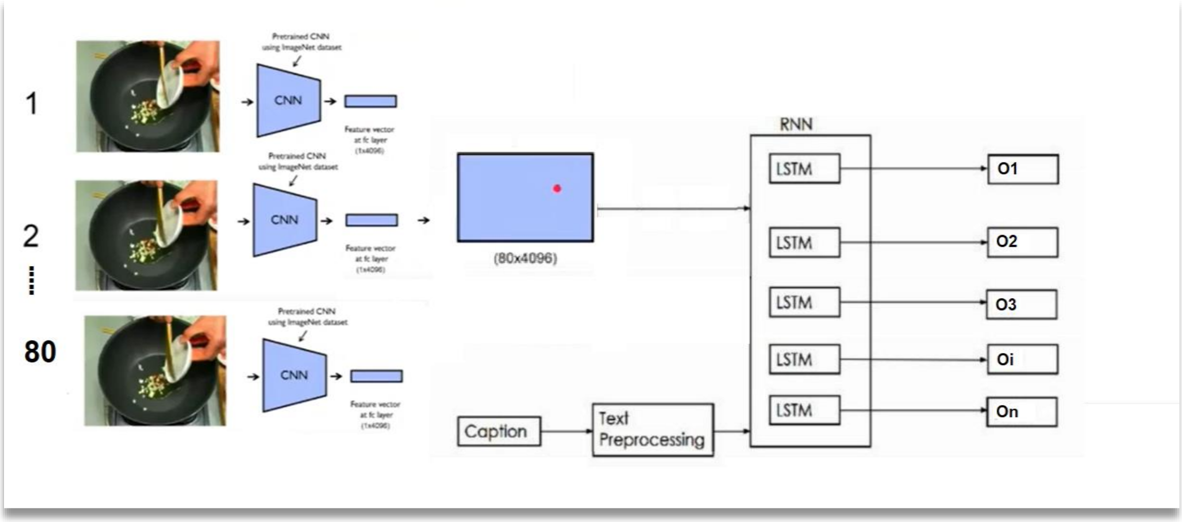

1
2
80
Pretrained CNN using ImageNet dataset
CNN
Feature vector at fc layer (1x4096)
Pretrained CNN using ImageNet dataset
CNN
Feature vector at fc layer (1x4096)
Pretrained CNN using ImageNet dataset
CNN
Feature vector at fc layer (1x4096)
(80x4096)
Caption
Text Preprocessing
RNN
LSTM
LSTM
LSTM
LSTM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

Fig 7

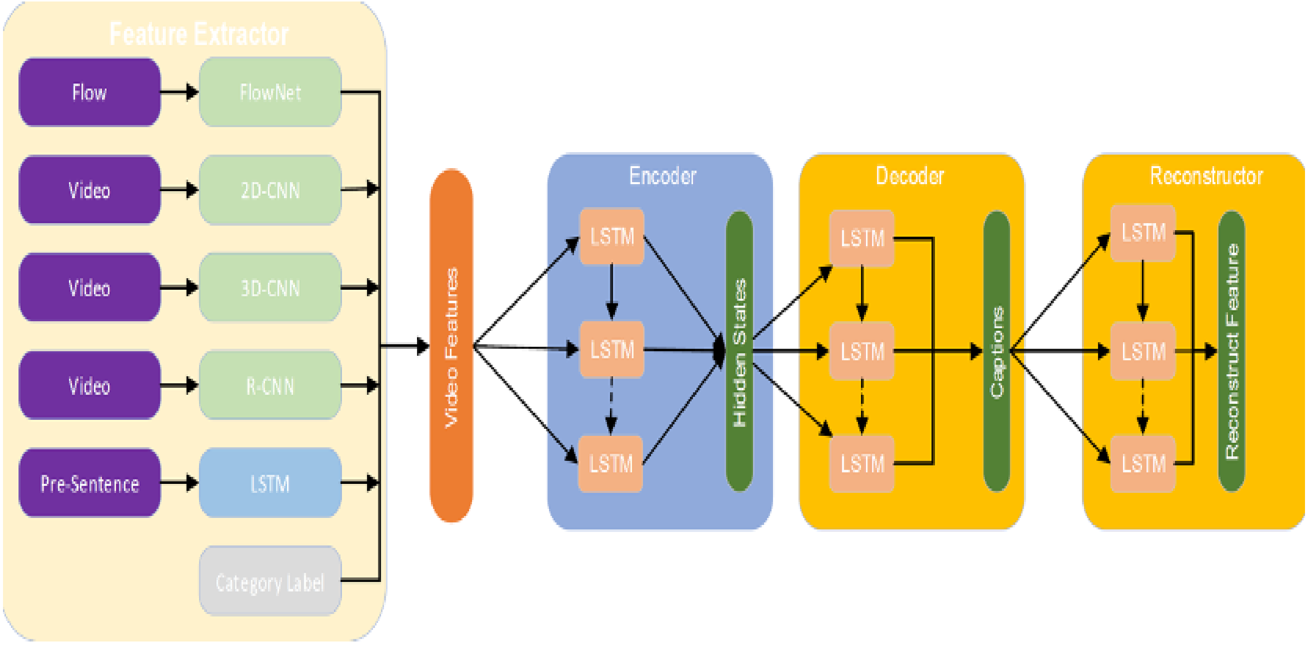


Fig 8 [9]

Architecture: As our work requires text productions so we must use encoder-decoder architecture. Hence for this purpose we are using sequence to sequence structure.

One thing you should know about this structure is the last state of the encoder cell works like it is the original state of cell i.e., it the only cell in encoder. For our model we will use the encoders to take videos features as input and provide us suitable outputs. Hence our encode-decoder model will look like this.

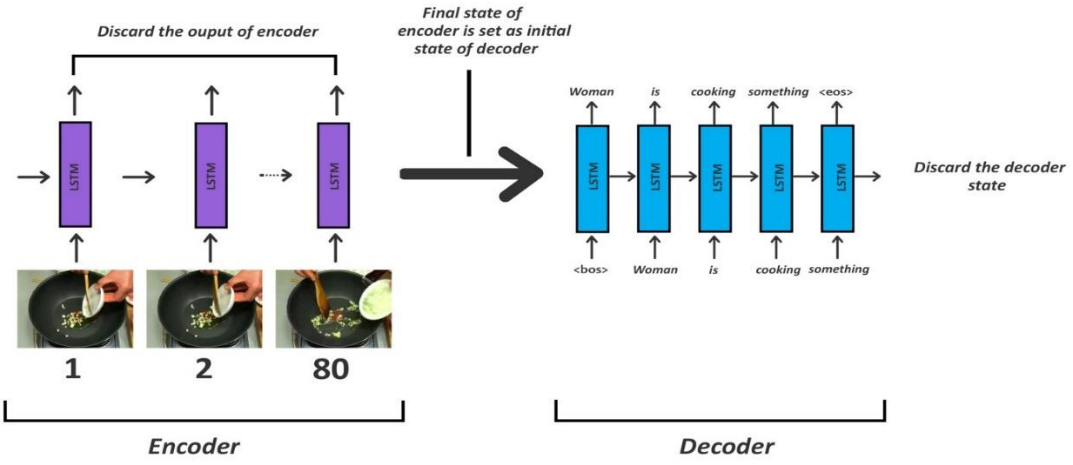

Discard the ouput of encoder
Final state of encoder is set as initial state of decoder
LSTM
1
2
80
Encoder
Woman
is
cooking
something
<eos>
<bos>
Discard the decoder state
Decoder

Fig 9

So now we will see the decoder stage. Here in the first decoder LSTM <BOS> serves as the starting point for a sentence. Each caption from the training data is fed one by one until <EOS>.

So, in the example above, the decoder will start with <BOS> in the first LSTM decoder. Then the next word comes from the real meaning that a woman is fed and then cooked. This ends with a <EOS> token.

As we are extracting 80 frames in a video, we will use 80-bit encoder. Number of features from for every frame gives us 4096 and we have total features to be 80x4096.

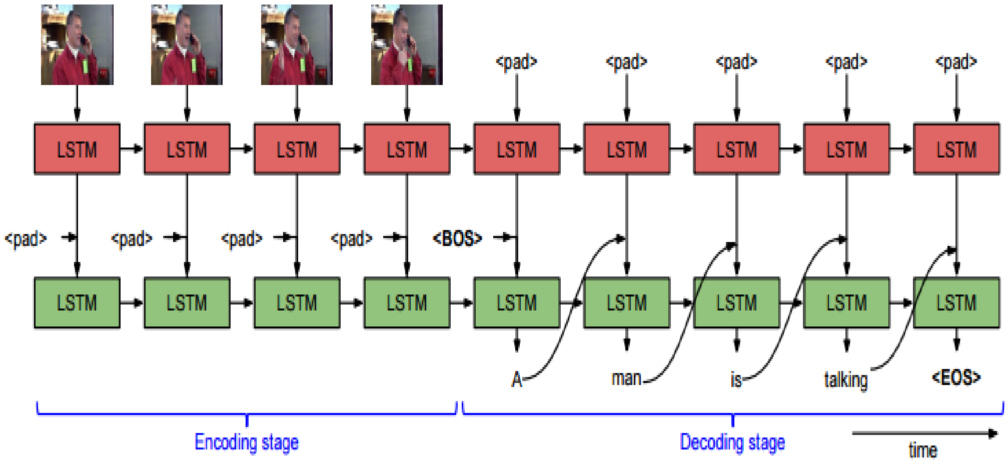


Fig 10

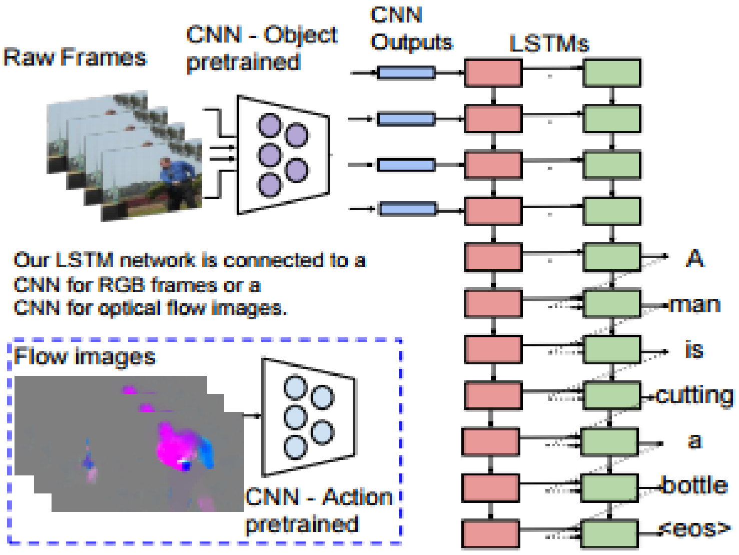


Fig 11 [9]

4.4 Tech Stack Analysis

To get our solution, we must use different Tech stacks. These are chosen on basis of ease of use, time for build and the efficiency of them.

The various technologies used are:

1. **Python**

   Python is a programming language that supports multiple application builds. Engineers consider it a good choice for Artificial Intelligence (AI), Machine Learning, and In-Depth Learning projects.

   It is most used because it provides us with a very large number of libraires, frameworks, easy to use and code.

   Python code is short and readable even for new developers, which benefits machine and in-depth learning projects. Due to its simplicity, the development of Python applications is faster compared to most programming languages. In addition, it allows the engineer to test algorithms without using them.

2. **Tkinter (GUI)**

   Tkinter is a very powerful Python graphical user interface library. Python, in conjunction with Tkinter, gives us very easy and fast GUI applications. We have used it for the GUI part of our project.

3. **Keras**

   Keras is a very important in-depth machine learning API which provides pretrained models to us so that we can focus on the original problems rather than hard coding every algorithm to be used. Also it provides us all the ways, tools required for analysis of models. For our work, we are using Keras for getting pretrained model of VGG16 which is present as an API in Keras.

4. **TensorFlow**

   TensorFlow is one of the most powerful frameworks present in python which was given by Google. Its main purpose is just to create Deep learning models. Hence helping us to focus on our problem and the algorithm to be used and provides us the highest flexibility to declare the number of dimensions to be used, hidden layers and much more.

5. **OpenCV2**

   We all know OpenCV as one of the leading computer libraries out there. Additionally, it also has the functionality of using in-depth reading inference. The best part is that it has very large number of pretrained models which could be used directly in the models and also provides a proper documentation for everything it

5. Result and Discussion

5.1 Performance of our models:

We trained and tested our machine learning models on a dataset of 2000 videos each having 10 captions so total our dataset has 20000 labels We used different combinations of various machine learning and feature extraction algorithms.

We tracked the model's loss on the training set, as well as its performance on the validation set, during training. The validation set is solely used to assess the trained model's generalisation ability, not to train it.

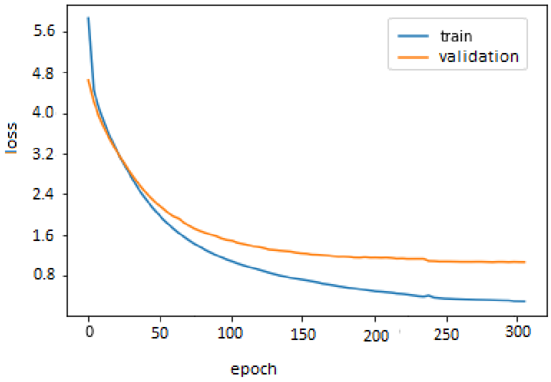
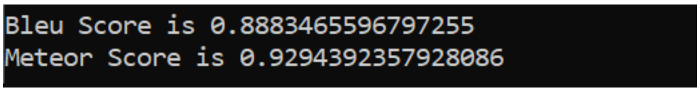


Fig 15

For Quantitative evaluation we have used two types of scores. They are:

BLEU Score: BLEU i.e., BiLingual Evaluation Understudy is one of the most common metrics used for evaluating the translation text. It basically finds the average accuracy for all the unigrams up to length of 4. As it gives output between 0 and 1 we can easily find the similarity of machine generated text to real world text.[9]

METEOR Score: METEOR i.e., Metric for Evaluation of Translation with Explicit Ordering. It is more accurate and a better metric than BLEU score because here harmonic mean is involved of both precision and the recall values.



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

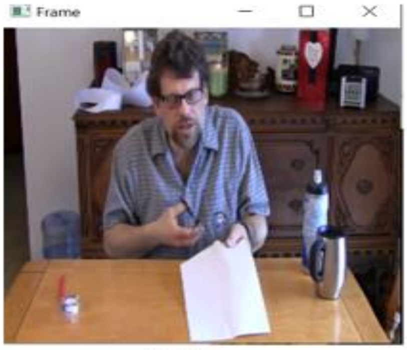
Fig 16

Test Metric	Value
BLEU Score	0.88
METEOR Score	0.92

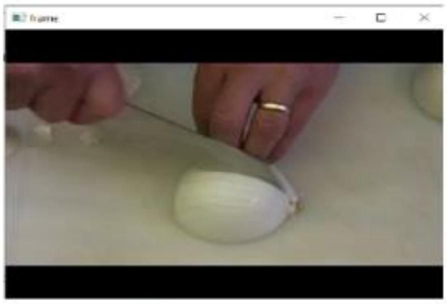
.

5.3 Deployment Status and Testing

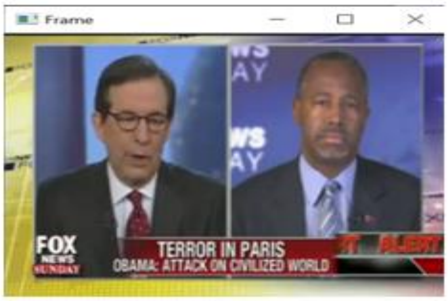
To understand how good our bot is, let’s try to generate captions on videos from the test dataset (of MSVD).



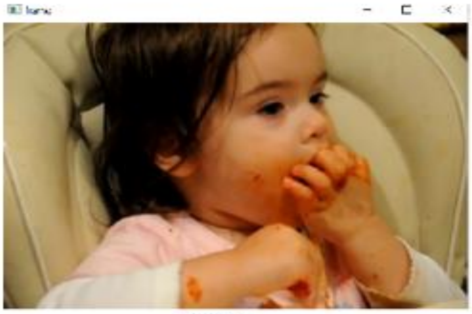
Caption: .
A MAN IS CUTTING A PAPER .



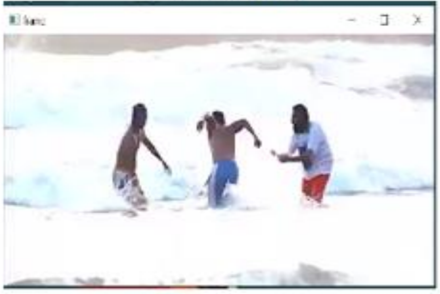
Caption: .
A Man is cutting an onion



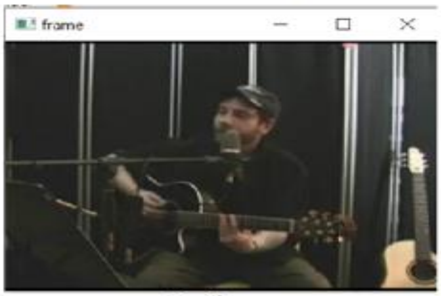
Caption
A man is talking



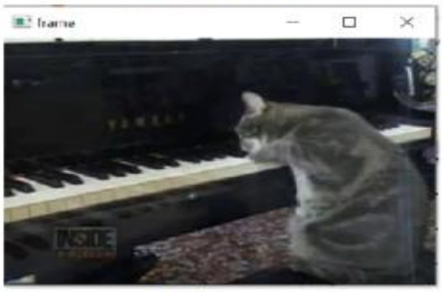
Caption: .
. A baby is eating spaghetti .



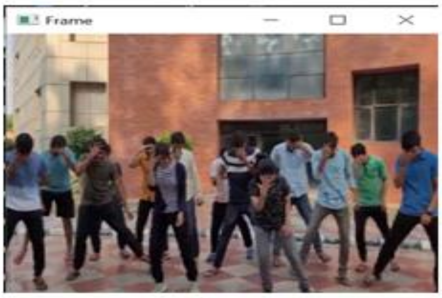
Caption:
Men are playing in the beach



Caption:
a man is playing a guitar



Caption
A cat is playing the piano



Caption: .
A GROUP OF PEOPLE ARE DANCING

Fig 17

6. CONCLUSION AND FUTURE SCOPE

Please refer Google Drive link **here** to access the full code. With increase in internet connectivity the importance of videos is increasing day by day and hence captioning them is becoming more and more important. Video Captioning Bot has come out to be a cost effective, efficient and provides quality captions to the videos at a very faster speed

Future Scope:

There is always a future for a project and hence has some new modifications which can improve the solution. For our project the future scope is as:

- The dataset we are using is small we can work on a larger dataset for better captions.
- The response time are heavily rely on the hardware of the machine, i.e. the processing speed of the processor, the size of the available RAM, and the available features of the webcam, its resolution. Therefore, the program may have better performance when it's running on a decent machine.
- This system can be implemented in many applications that can access government websites whereby video captioning in audio for blind is available or filling out forms online whereby no interpreter may be present to help.

---------------END OF REPORT--------------

References

1. H. Aradhye, G. Toderici, and J. Yagnik. Video2text: Learning to annotate video content. In ICDMW, 2009.
2. X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. CVPR, 2015
3. Sequence to Sequence – Video to Text by Subhashini Venugopalan, Marcus Rohrbach,Jeff Donahue, Raymond Mooney, Trevor Darrell,Kate Saenko
4. M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In EACL, 2014
5. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8), 1997
6. J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification
7. Very Deep Convolutional Networks for Large-Scale Image Recognition by Karen Simonyan, Andrew Zisserman
8. Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images by Srikanth Tammina
9. BLEU: a Method for Automatic Evaluation of Machine Translation by Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu
10. Image Captioning Using R-CNN & LSTM Deep Learning Model by Aditya Kumar Yadav and Prakash.J