Translation of sign language in videos to English sentences using Deep Learning

Jaspreet Singh, Vivek Mishra, Cherry Khosla

Computer Science and Engineering

Lovely Professional University

Abstract: **Deep learning is doing exceptionally well in the various fields like computer vision, NLP, human computer interaction. There are millions of people who do not have the capacity to understand the world due their inability to listen and speak. With the advent of deep learning, we can make these people hear and speak. In this paper, we have presented a tool, which acts a translator of Indian sign language to English text. The 3-D convolutional neural networks are used to predict the sign language. As a front end, a web application is created which will act as an interface to understand the sign.**

Keywords: Deep Learning Learning, Language translator, CNN.

INTRODUCTION

'SIGN LANGUAGE TRANSLATOR' is an attempt to use deep learning for the favor of deaf community and other people who require sign languages to communicate in various situations. Sign language is completely different from any type of human language. There are more than hundred types of sign languages. Ex. American sign language, German Sign Language etc. A sign language is not a direct translation of any language. For example, translating an American Sign Language does not mean it is a direct translation of English and same is applicable for the other sign languages. These languages consist of all the fundamental characteristics of any spoken language – it possesses its own grammar, order of words, phrases, idioms and rules for pronunciation. Because each type language has its own ways of denoting and identifying different actions, like asking a question rather than making a statement, languages differ in how these actions are done [1] [2]. This paper, presents a translator that translates Indian sign language sentences into English text. Sign language is not only used by dead and hard hearing people, we have another group of people or children who makes use of sign language due to conditions like down syndrome, autism, cerebral palsy, trauma, and brain disorders or speech disorders. Here in this paper, we try to predict the sentences associated with each sign language video and understanding the complexity and difficulties that are needed to be addressed to solve the problem. This task is a really cumbersome task and requires really powerful computing machinery. There are many companies and individuals trying to achieve this feat of sign language translation. But none of them have achieved a satisfactory level of generality and accuracy. When dealing with large amount of unstructured data such as images and audios Deep Learning is a way to go. Deep learning algorithms are some very robust techniques to deal with problems in the different domains like Computer Vision, Speech Recognition, human computer interaction, etc. But as with everything there are trade-offs, when using Deep Learning models, there are trade-offs too. When dealing with Deep Learning models, there are many

parameters involved which in turn makes memory requirement and training time of the models high. But with the advent of powerful computing machinery and cloud computing, it is getting easier.

In any video, individual frames consist of spatial information and sequence of frames consist of temporal information. To capture the spatial information in images, the best way is to use Convolutional Neural Networks. There are mainly two methods applied when extracting temporal information from the frames. The first method is used to extract features from a video that captures temporal information into images. These images are called as *dynamic image* that generates a single image from a video and then apply image classification on that image. The second method is to use 3D-Convolutional Neural Network (3D-CNN) and Long Short-Term Memory (LSTM). Both models are well suited when capturing temporal information in a video. 3D-CNNs are good at capturing spatial information as well as short-term temporal information. As we are trying to predict videos that are less than 10 seconds, we have used 3D-CNNs as our prime method. To make our model accessible to everyone, we have developed a web app. This web app is developed with the help of HTML, CSS and Flask framework.

Related Work

There have been efforts in past to solve this problem. One of the most popular system out there was made by the company Motion Savvy. This was first translator of sign language to voice. The authors have developed an application that make use of motion sensing and pose estimation to predict the sentences. Motion Savvy, was the startup of San Fransisco, is working on different techniques that helps the deaf people to communicate, has started a campaign with a name Indiegogo. This Indiegogo campaign was started to launch their first designed product, known as UNI in the market. The application together Leap Motion technology is used to translate the signs from American Sign Language into audible words. Three different parts of UNI are: a computer, a smart case which is specially-designed for UNI, and a mobile application. The smart case consists of Leap Motion hardware along with couple of cameras which will track the location of both the user's hands and fingers. The app, which is designed for the tablet, is responsible for translating the sign language of hand and finger movements a into an audible speech or text displayed on the screen. Tool called EonTex Conductive Stretchable sensor was developed by [3] which is e-textile technology in which a sensor is in the form of a fabric. The senor reads the analog signals for sensor reading process. Microcontrollers were used to read the analog signals along with sign language translator [7]. In [9], the authors have developed a smart system that served hearing and speech impaired persons. The gloved mapped the hand gestures with the help of bend sensors. contains video of only Indian sign language. We used the Indian Sign language for the task using the spatial information over time to predict sentences. The sign video can be seen as a sequence of frames so we put the array of frames into a 3DCNN where the dimensions are the (frames, width, height, channels).

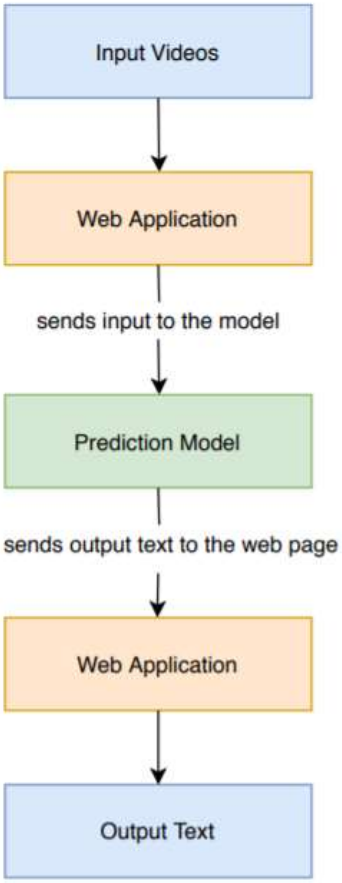

Figure-1 Flow of data

Experimental Plan and Results

The sign language videos we use are taken from YouTube and each video depicts only a single statement. The data we use for this task A web application is developed using flask and python, HTML and CSS in the front end.

The 3D activation maps produced during the convolution of 3D-CNN is really important when analyzing the data where Temporal or Volumetric context is really necessary. This ability to analyze spatial and temporal information leads us to use the 3D-CNNs in our project. Batch Normalization is used so that each layer of a neural network should learn on its own, a little at each step, which is independent of the other layers in the network. To increase efficiency and the stability of model, batch normalization will normalize that is maintain the standard condition the output of the previous activation layer by subtracting the batch mean and dividing by the batch standard deviation.

The model was trained on 66 videos and gives correct predictions for 64 of the videos. So, the accuracy of the model evaluated by the above formula comes out to be 96% and test the result using web application

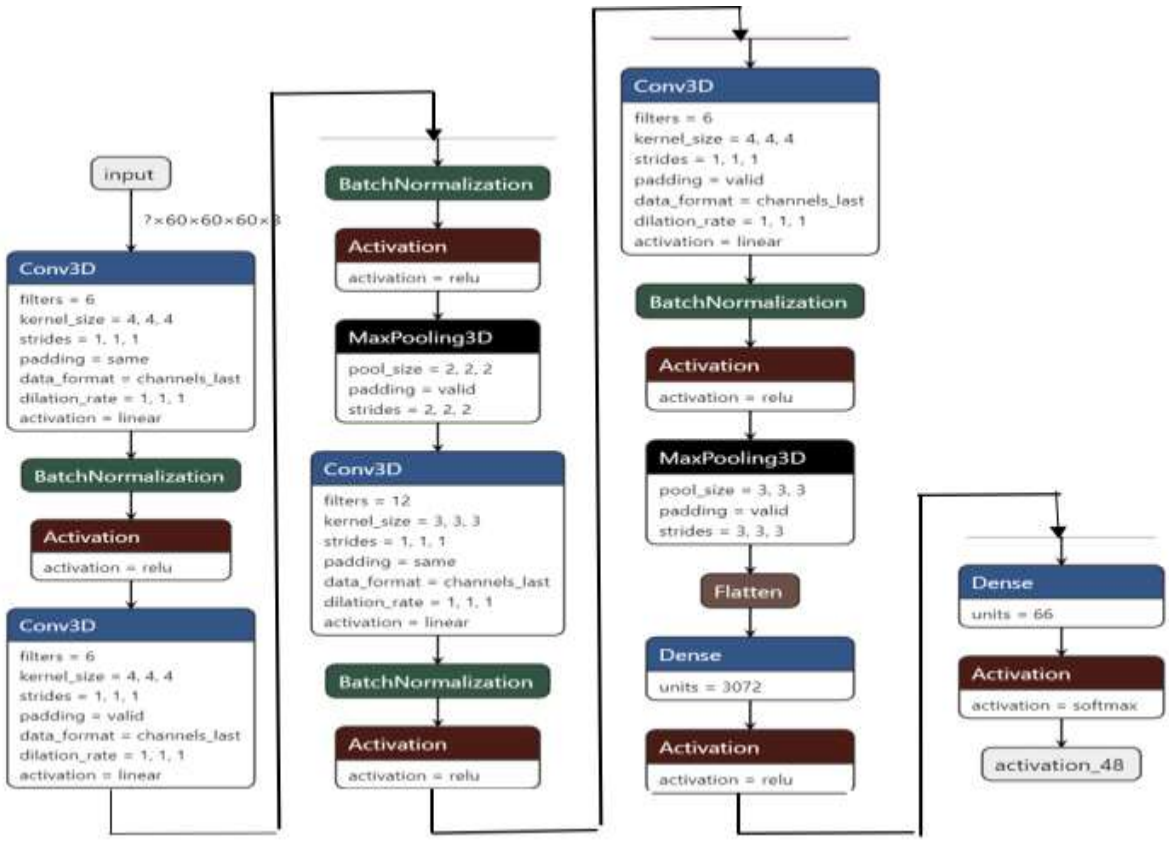

Figure-2 CNN model used

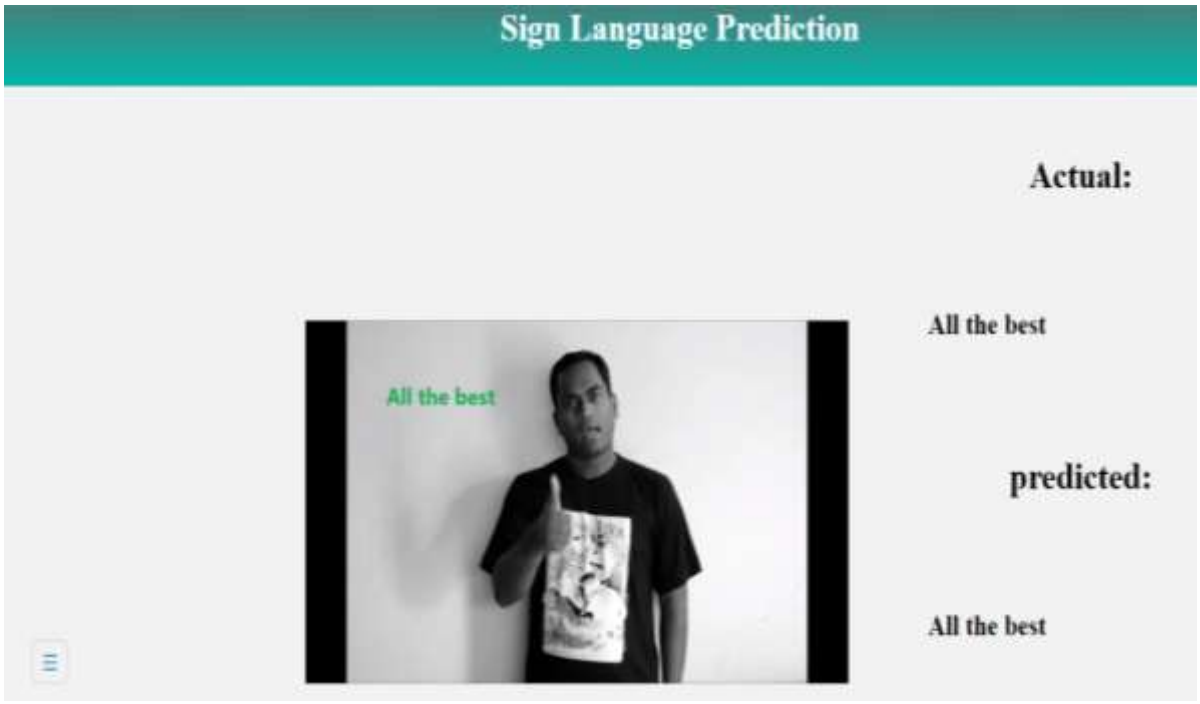

Figure-3 Translation of "All the best" sign

Figure-4 Translation of "any questions" sign

References

1. "Neural Sign Language Translation" by Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, Richard Bowden
2. "Motion Fused Frames: Data Level Fusion Strategy for Hand Gesture Recognition" by Okan K¨op¨ukl¨u, Neslihan K¨ose, Gerhard Rigoll
3. "Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion" by Hao Tang, Hong Liu, Wei Xiao, Nicu Sebe
4. "DenseImage Network: Video Spatial-Temporal Evolution Encoding and Understanding" by Xiaokai Chen, Ke Gao
5. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description" by Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell
6. "LEARNING DEEP AND COMPACT MODELS FOR GESTURE RECOGNITION" by *Koustav Mullick and Anoop M. Namboodiri*
7. Rupasinghe, R.A.D.K., Ailapperuma, D.C.R., De Silva, P.M.B.N.E., Siriwardana, A.K.G. and Sudantha, B.H., 2017. A Portable Tool for Deaf and Hearing Impaired People.
8. Hersh, M., 2013. Deafblind people, communication, independence, and isolation. *Journal of deaf studies and deaf education*, *18*(4), pp.446-463.
9. Chouhan, T., Panse, A., Voona, A.K. and Sameer, S.M., 2014, September. Smart glove with gesture recognition ability for the

hearing and speech impaired. In *2014 IEEE Global Humanitarian Technology Conference-South Asia Satellite (GHTC-SAS)* (pp. 105-110). IEEE..