Auto Caption Generator

Ankit Kumar[1], Kartik Bhat[2], Pratik Chaudhari[3], Varsha Wangikar[4]

[1,2,3]*Excelssior Education Society's, KC College of Engineering and Management Studies and Research, Kopri, Thane (East), Mumbai, Maharashtra, India*

[4]*Assistant Professor, KC College of Engineering and Management Studies and Research, Kopri, Thane (East), Mumbai, Maharashtra, India*

***Abstract*: The world is moving towards digitization, so are the means of communication. Phone calls, emails, text messages etc. have become an integral part of message conveyance in this tech-savvy world. In order to serve the purpose of effective communication between two parties without hindrances, many applications have come to picture, which acts as a mediator and help in effectively carrying messages in form of text, or speech signals over miles of networks.**

***Keywords*: deep-learning, speech-recognition, transcription, LSTM-networks.**

I. INTRODUCTION

One of the key applications of automatic speech recognition is to transcribe speech documents such as talks, presentations, lectures, and broadcast news. Although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve, and reuse speech documents if they are simply recorded as audio signal. Therefore, transcribing speech is expected to become a crucial capability for the coming IT era.

Machine Translation (MT) Systems are used for automated translation of one natural language to another. Machine translation is the research field of Natural Language Processing (NLP) which aims to fill the gap of communication among the different sections of societies. Human translation of any language is time consuming and expensive. By the use of machine translation systems, we can reduce time and cost of human translators. Hindi, the official language of India, is used by about 400 million people. The motivation behind the project English to Hindi machine translation system is that many documents and records like government records, historical, news etc. are written in English, which is not popular among the remote villagers of India. However, the rural villagers, especially in north India, apart from their native language, know Hindi. Hence, there is a need of an automatic translation from English to Hindi.

TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications. TensorFlow offers multiple levels of abstraction so you can choose the right one for your needs. Build and train models by using the high-level Keras API. It is more flexible and eager execution allows for immediate iteration and intuitive debugging. For large ML training tasks, use the Distribution Strategy API for distributed training on different hardware configurations without changing the model definition.

II. METHODOLOGY

A. Proposed Architecture

The core of our system is the neural network trained to ingest MFCC features of speech and generate English text transcriptions. The goal of our network is to convert the features into a sequence of character probabilities for the transcription text.

Our model is composed of 5 layers of hidden units. They use clipped rectified-linear (ReLu) activation function. The fourth layer is a bidirectional LSTM layer. This layer contains two sets of hidden units: a set with forward recurrence and a set with backward recurrence. The fifth (non-recurrent) layer takes both the forward and backward units as input. The output layer is a standard softmax function that yields the predicted character probabilities for each time slice and character in the alphabet.

Once we have computed a prediction, we compute the CTC loss to measure the error in prediction. During training, we can evaluate the gradient with respect to the network outputs given the ground-truth character sequence. From this point, computing the gradient with respect to all of the model parameters may be done via back-propagation through the rest of the network.

B. Audio Extraction

avconv is a very fast video and audio converter that can also grab from a live audio/video source. It can also convert between arbitrary sample rates and resize video on the fly with a high-quality polyphase filter.

avconv reads from an arbitrary number of input "files" (which can be regular files, pipes, network streams, grabbing devices, etc.), specified by the -i option, and writes to an arbitrary number of output "files", which are specified by a plain output filename. Anything found on the command line which cannot be interpreted as an option is considered to be an output filename.

Each input or output file can in principle contain any number of streams of different types (video/audio/subtitle/attachment/data). Allowed number and/or types of streams can be limited by the container format. Selecting, which streams from which inputs go into output, is done either automatically or with the -map option. We use avconv to convert our input video file to audio .wav format

C. *Speech Recognition*

After the completion of audio extraction, the speech recognition part is carried out as shown in. Now, the extracted .wav file is used to generate a .srt file using Speech Recognition in which the audio undergoes three modules, the Front End, Decoder and the Knowledge Base.

First, in the Front End the signal is divided into frames and MFCC (Mel Frequency Cepstrum Coefficient) is calculated. MFCC is most commonly used for feature extraction at front-ends in speech recognition systems. The technique is FFT-based (Fast Fourier Transform), which means that feature vectors are extracted from the frequency spectra of the speech frames. The Mel scale, a non-linear frequency scale is used to make triangular bandpass filters and a series of such filters is called mel frequency filter bank.

The equation (1) given below describes the mathematical relationship between the linear frequency scale and the Mel scale,

freqMel = 2595.0 * (Math.log (1.0 + freq / 700.0) / Math.log (10.0))………(1)

where freqMel is the Mel frequency in mels and freq is the linear frequency in Hz.

The center frequencies of the filter series spread evenly along the Mel frequency scale. The filters are overlapped in such a way that the lower boundary of one filter is situated at the center frequency of the previous filter and the upper boundary is situated at the center frequency of the next filter. The maximum response of a filter, that is, the top vertex of the triangular filter, is located at the filter's center frequency and is normalized to unity.

These features are then further sent to the above described neural network model trained on open source audio speech data as obtained from librispeech. The neural network output is then sent to decoder unit. The decoder unit consists of a CTC Beam search decoder to decode the neural network output text using a language model. The output gives us the text which is used in the further stages of processing.

The entire flow of the proposed system is shown in the following figure

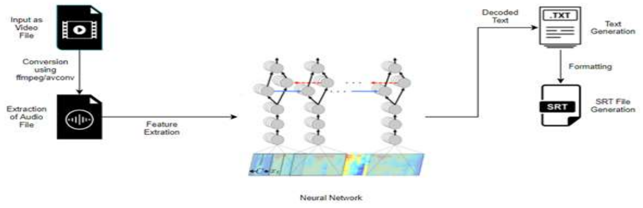

Fig.1: Caption Generating Process

D. *Subtitle Generation*

SubRip (SubRip Text) files are named with the extension .srt, and contain formatted lines of plain text in groups separated by a blank line. This module is expected to get a list of words and their respective speech time-frames from the speech recognition module and then produce a .srt subtitle file. The module formats the text file obtained from speech recognition to match the standard .srt file format. The name of the .srt file will be the same as that of the video file given by the user for convenience and ease of use.

III. CONCLUSION

The proposed system as we described takes video/audio file as input and generates a subrip file for the same. The system first converts the input file into .wav file for further processing. The text is then extracted from the file using speech recognition techniques and at last the subrip file is generated. We believe this approach will yield promising results and the system can be realized soon.

REFERENCES

[1] Jayashree Nair, Amrutha Krishnan K, Deetha R," An Efficient English to Hindi Machine Translation", 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio", 12-19 Sept. 2016

[3] George Frewat, Charbel Baroud and Roy Sammour," Android Voice Recognition Application with Multi Speaker Feature", Proceedings of the 18th Mediterranean Electrotechnical Conference MELECON 2016, Limassol, Cyprus, 18-20 April 2016.

[4] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng, "Deep Speech: Scaling up end-to-end speech recognition", 17-19 Dec 2014

[5] English speech recognition method based on Hidden Markov model, Liaoning Jianzhu Vocational University, 12 Aug. 2016.

[6] Du Guiming , Wang Xia , Wang Guangyan , Zhang Yan1 , Li Dan," Speech Recognition Based on Convolutional Neural Networks", 15 Aug. 2016.

[7] Tianxing He ,Jasha Droppo," Exploiting LSTM Structure in Deep Neural Network for Speech Recognition", 25 March 2016.

AUTHOR BIOGRAPHIES

Ankit Kumar

Born in Patna, Bihar, India on 22/03/1998.

The author is currently pursuing Bachelors of Engineeringin the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.

Kartik Bhat

Born in Mumbai, Maharashtra, India on 09/11/1998.

The author is currently pursuing Bachelors of Engineering in the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.

Pratik Chaudhari

Born in Kalyan, Maharashtra, India on 29/04/1999.

The author is currently pursuing Bachelors of Engineering in the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.