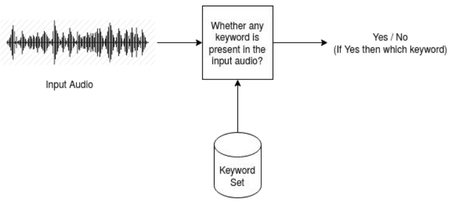
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

Audio Keyword Spotting

OVERVIEW

The goal of this work is to determine when a word of interest is spoken in the audio. We try to adapt the representation for audio from a pre trained Automatic Speech Recognition(ASR) model. The vocabulary set of the keyword is open.
Our key contributions are: (1) improving the performance of learnt representation for audio in a pre trained model while reducing the number of parameters in the model.



VARIOUS METHODS USED TO SPOT KEYWORDS IN AUDIO

1. Convert the speech audio into text using Automatic Speech Recognition models like DeepSpeech2 [1], Wave2letter[2] and check whether the word of interest is present in the text.

Advantage: Can find the word in text using simple string matching.
Disadvantage: Requires a large computing resource to convert the audio into text.

2. Learn a fixed representation for acoustic words by training sequence to sequence autoencoders, and use distance metric to find similarity between two audio signals.

Advantage: Can obtain fixed length representation for variable length audio.
Disadvantage: Preparing a dataset of segmented audio words from continuous speech is itself a separate task.

3. Find a representation using Automatic Speech Recognition Architectures and using it to spot words using distance metrics to compare audio signals at representation level.

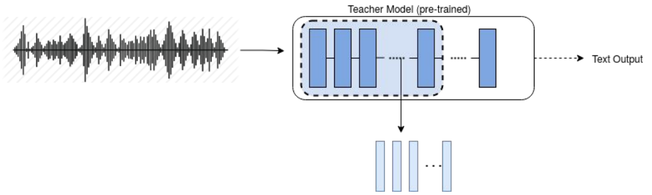
Advantages: More information about the spoken word is captured since it is at higher dimension.
Disadvantage: Difficulty in reproducing the results of ASR performance when trying to reduce the model parameters.

DATASETS USED IN THE WORK

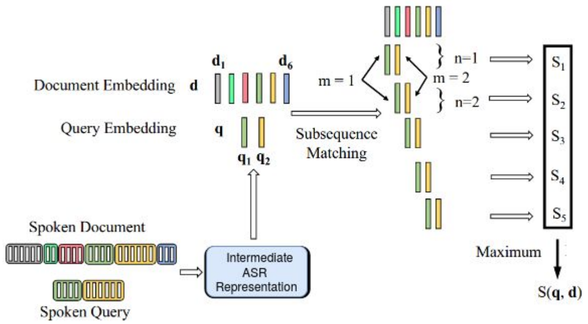
LibriSpeech is a large-scale (about 1000 hours) English speech corpus derived from audio books, sampled at 16kHz. The dataset is divided into clean and other. In the experiment, “train-clean-100”, “train-clean-360”, and “train-other-500” were used in the training phase.
Our experiments were conducted using OpenSeq2Seq, which is a TensorFlow-based toolkit for sequence-to-sequence models, which comes with pre-trained models of DeepSpeech2.

METHODOLOGY

We initially measure the performance of pretrained Automatic Speech Recognition model’s intermediate representation to spot audio. We use pretrained Deepspeech2 model to get the representation.



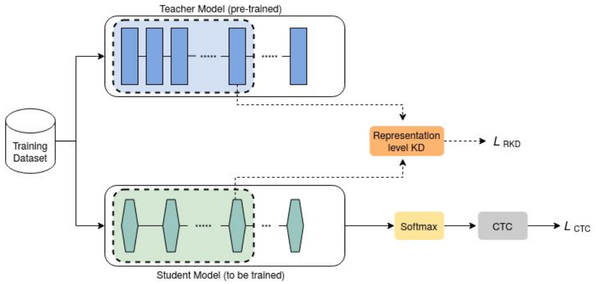
We try to match similarity between two audio representations using cosine similarity. The matching between two audio signals of different length is resolved through sliding window approach.



Directly using intermediate layer representation from pretrained ASR procures using the same number of model parameters as the ASR model. The total model parameters used in DeepSpeech2 Model is 56M. We try to improve the performance and reduce size through knowledge distillation.

DISTILLATION APPROACH

We try to reduce the number of parameters in a ASR model while trying to increase the performance of keyword spotting approach using teacher student distillation setting. We use Mean Squared Error to reduce the distance between the teacher models representation and student models representation.



RESULTS

We measure the performance for segmented audio of query set and retrieval set. One audio for each keyword is present in query set and retrieval set. We convert the audio into the learnt intermediate representation and try to match using cosine similarity. Currently the student model parameters are same as teacher model. The keyword audio is obtained from TIMIT dataset.

Model	Scenario1 (mAP)	Scenario2 (mAP)	WER
DeepSpeech2 Pretrained	0.888	0.822	6.71
DeepSpeech2 Student	0.935	0.827	10.2

Scenario1: The number of characters in the words are more than 4 with all the stop words removed totalling to 571 query-retrieval pair.
Scenario2: The number of characters in the words are more than 3 with all the stop words removed totalling to 730 query-retrieval pair.

REFERENCES

1. Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, JingliangBai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, QiangCheng, Guoliang Chen, et al.DeepSpeech 2: end-to-end speech recognition in english and mandarin. InProc. ICML, pages 173–182,2016.
2. Wang, Yu-Hsuan, Hung-yi Lee, and Lin-shan Lee. "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Authors: Kirandevraj, Dr. Vinod Kumar Kurmi, Prof. Vinay Namboodri, Prof. CV Jawahar

Research Center Name: CVIT