Implementation of Speech Recognition on Edge Devices

Shaik Shabana,
SSG OTC,
Intel Corporation
Bangalore,India
shaik.shabana@intel.com

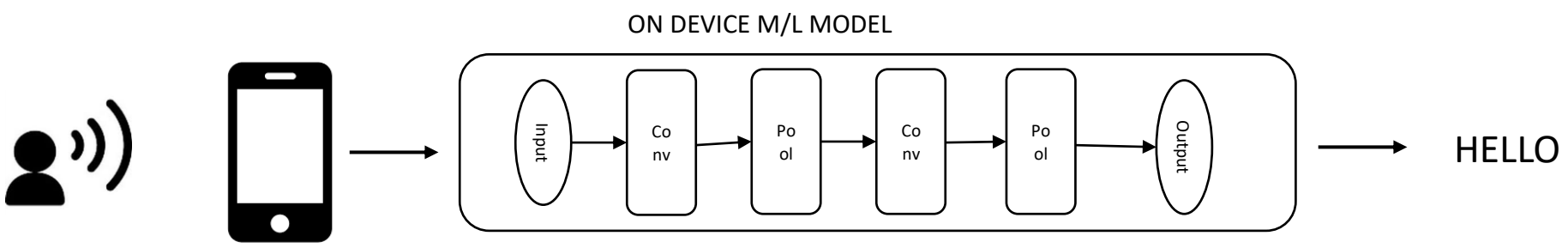
Sujith Thomas,
SSG OTC,
Intel Corporation,
Bangalore, India
sujith.thomas@intel.com

Theme of the Poster:

- This paper explains the challenges, techniques and provides an insight in the implementation of speech recognition models on edge devices.
- It also showcases the process to integrate and enable speech model for on-device inferencing on Android devices.
- This research is aimed at developers who want to deploy and enable speech recognition with ease and at low cost.

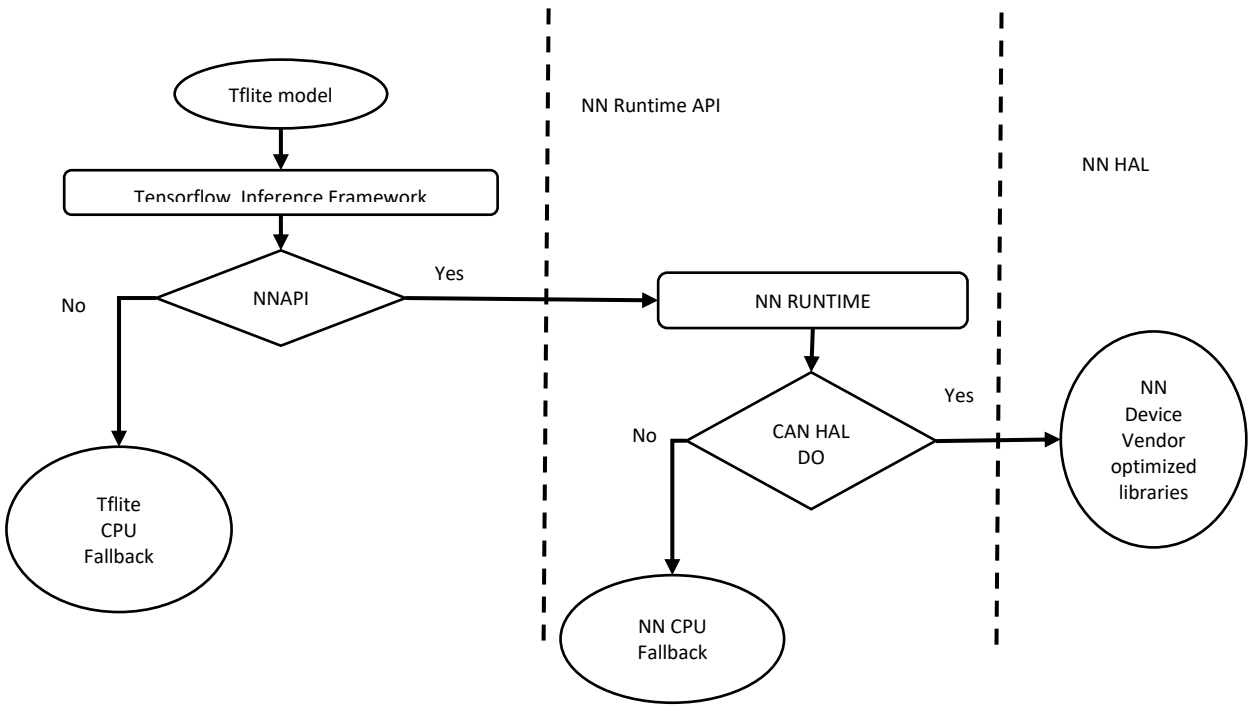
Technology:

- Android has always been a popular option for edge devices because it is an open sourced Linux based software which provides an easy to use front end interface for the users.
- With the advent of neural network API's in Android 8.1 and above, the traction for deep learning based solutions has gained pace.
- The android neural networks API (NNAPI) is an android C API designed for running computationally intensive operations for machine learning on mobile devices. It provides a base layer of functionality for higher-level machine learning frameworks (Tensorflow Lite, Caffe2, or others) that build and train neural networks.



System Methodology:

- Deep learning framework: Choose deep learning frameworks Tensorflow, caffe, torch, theano etc., based on hardware & application
- Selection of model: On-device inferencing, size of model is proportional to the time taken to load model, extensive dataset, Application.
- Wavenet: Popular machine learning algorithm for speech synthesis. Wavenet is a deep generative model of raw audio waveforms.
  - Wavenet are dilated CNN so that there are a number of input layers, which take an audio signal as input and synthesize the output sample by sample.
  - Higher the number of input layers, higher is the receptive field of this network, which is the input for generating the next sample.
- Input Pre-processing techniques: MFCC, Audio Spectrogram.
- Training the model: train on a highly computational CPU, GPU or using dedicated AI hardware. Batch Size, Epochs



Inference Time	SDK	"HOW ARE YOU"	"I AM A BOY"
Device 1	8.1	1.43s	1.55s
Device 2	8.1	0.93s	0.98s
Device 3	8.1	1.46s	1.58s
Device 3	7.0	2.1s	2.3s

Experimental Setup:

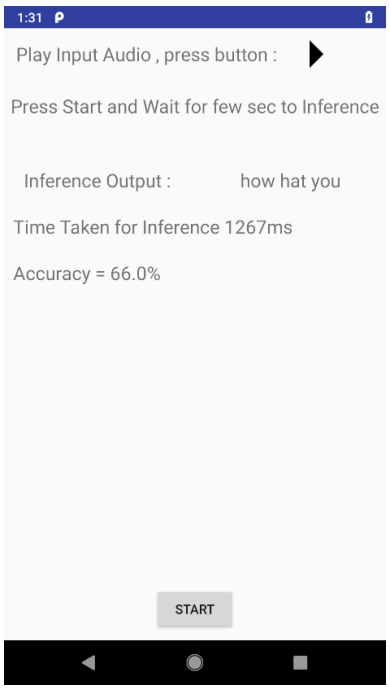
- An application is created as an insight to showcase the true strength of speech recognition and how it can be enhanced with android as an operating system.
- Microphone can be used as an input source. We have recorded an input audio file and used it for inferencing to maintain consistency.
- Wavenet Model in Tensorflow format is used. The sampling rate is set to 16 KHz.
- Inference times are close to 1-1.5 seconds for different market devices, which can be further optimized if the model is converted to TensorflowLite format.

Applications:

- This application can be deployed to identify panic events in automotive to prevent accidents. Model should be trained with words like "SAVE","HELP", "THIEF" etc.
- A person when in trouble first screams or cries. If the model detects such events with good accuracy, the proposed model is going to be a preventive solution for kidnapping cases.

Future Scope of Work:

- The Tensorflow speech model can be converted into Tensorflow Lite format which allows us to accelerate the computation through dedicated AI hardware.
- The model can be trained on a broader dataset which will improve the accuracy of the model.



"Presented at WinTechCon-2018, organized by IEEE CAS Bangalore Chapter, IEEE Bangalore Section, and IEEE WiE Council"