Keywords Spotting in real time

Brief :

Spoken keyword spotting (KWS) deals with the identification of keywords in audio streams in low latency or in real time
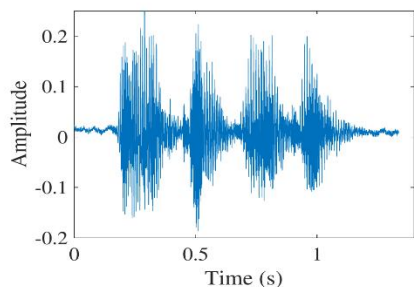
Challenge :

- ➢ Small footprint
- ➢ Low computational complexity
- ➢ Low latency
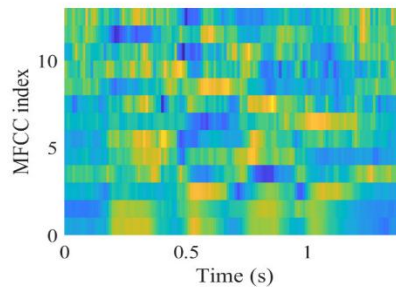- ➢ High accuracy (Large receptive field of audio features)

Team

- Arun Kumar Pradhan
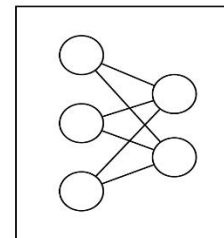- Kathir Ravichandran
- Madhan Gowrinathan

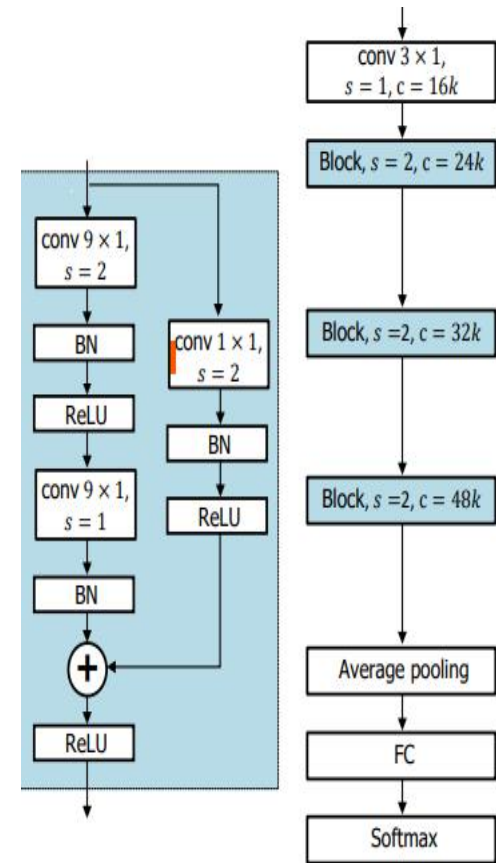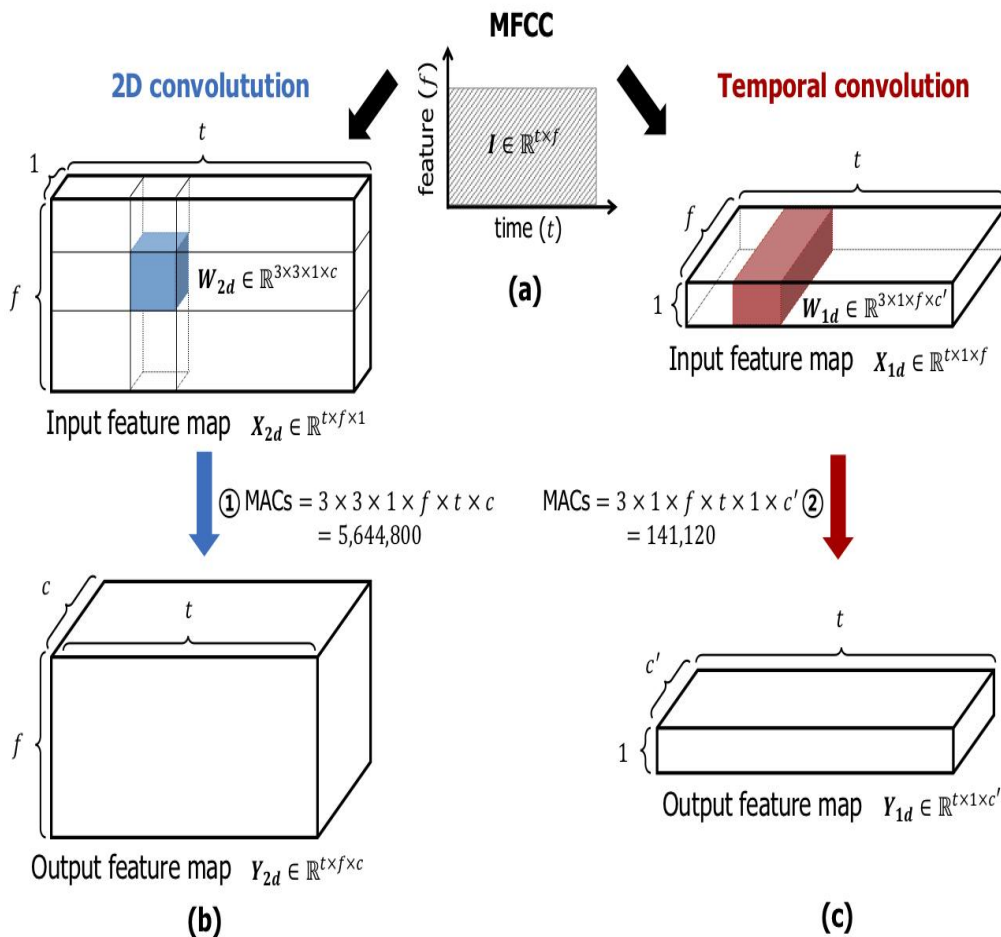Input speech signal

Speech features

Neural network

0.91 "Hey computer"
0.01 "Stop"
0.03 Non-speech
⋮

Model

40 MFCC features (window : 30ms, time shift :10 ms)extracted from Audio stream
Instead of 2D, dilated 1D convolution reduces complexity
RestNet8 with Temporal convolution : TC-RESNET8

2D convolutution

t

$W_{2d} \in \mathbb{R}^{3 \times 3 \times 1 \times c}$

Input feature map $X_{2d} \in \mathbb{R}^{t \times f \times 1}$

MFCC

$I \in \mathbb{R}^{t \times f}$

feature (f)

time (t)

(a)

Temporal convolution

t

$W_{1d} \in \mathbb{R}^{3 \times 1 \times f \times c'}$

Input feature map $X_{1d} \in \mathbb{R}^{t \times 1 \times f}$

① MACs $= 3 \times 3 \times 1 \times f \times t \times c$
$= 5,644,800$

MACs $= 3 \times 1 \times f \times t \times 1 \times c'$ ②
$= 141,120$

c

t

f

Output feature map $Y_{2d} \in \mathbb{R}^{t \times f \times c}$

(b)

c'

t

Output feature map $Y_{1d} \in \mathbb{R}^{t \times 1 \times c'}$

(c)

conv 9×1,
$s = 2$

BN

ReLU

conv 9×1,
$s = 1$

BN

$+$

ReLU

conv 1×1,
$s = 2$

BN

ReLU

conv 3×1,
$s = 1, c = 16k$

Block, $s = 2, c = 24k$

Block, $s = 2, c = 32k$

Block, $s = 2, c = 48k$

Average pooling

FC

Softmax

Tools ,Data Set and Training

Tools : Python 3.6.8 , TensorFlow 1.13.1

Data set :
Google Speech Commands Dataset .The dataset contains 64,727 one-second-long utterance files which are recorded and labeled with one of 30 target categories.

Data split : 80-10-10

Key words Targeted :
we distinguish 12 classes: "yes", "no", "up", "down", "left", "right", "on", "off", "stop", "go", silence, and unknown.

Total training Time: 4 hours and 35 minutes
Training Environment: AWS EC2 m5.2xlarge (32 GiB GiB RAM + 8 vCPU)

Model complexity reduction (Pruning , Quantization) : TensorFlow Lite

Model performance and complexity

Model	Accuracy (%)	Time(ms)	FLOPS	PARAMS
TC-RESNET8	96.1	1.1	3.0M	66K

Vs

Model	Accuracy (%)	Time (ms)	FLOPS	PARAMS
CNN	90.7	32	76.1M	524K

```
[ClassificationReport]
              precision      recall    f1-score      support

    __null__       0.97        1.00        0.99          257
        down       0.97        0.94        0.96          253
          go       0.93        0.96        0.94          251
        left       0.95        0.97        0.96          267
          no       0.97        0.94        0.95          252
         off       0.98        0.96        0.97          262
          on       0.99        0.98        0.98          246
       right       0.97        0.95        0.96          259
        stop       0.98        0.98        0.98          249
     unknown       0.92        0.95        0.94          257
          up       0.95        0.98        0.97          272
         yes       0.98        0.96        0.97          256

avg / total       0.96        0.96        0.96         3081
```
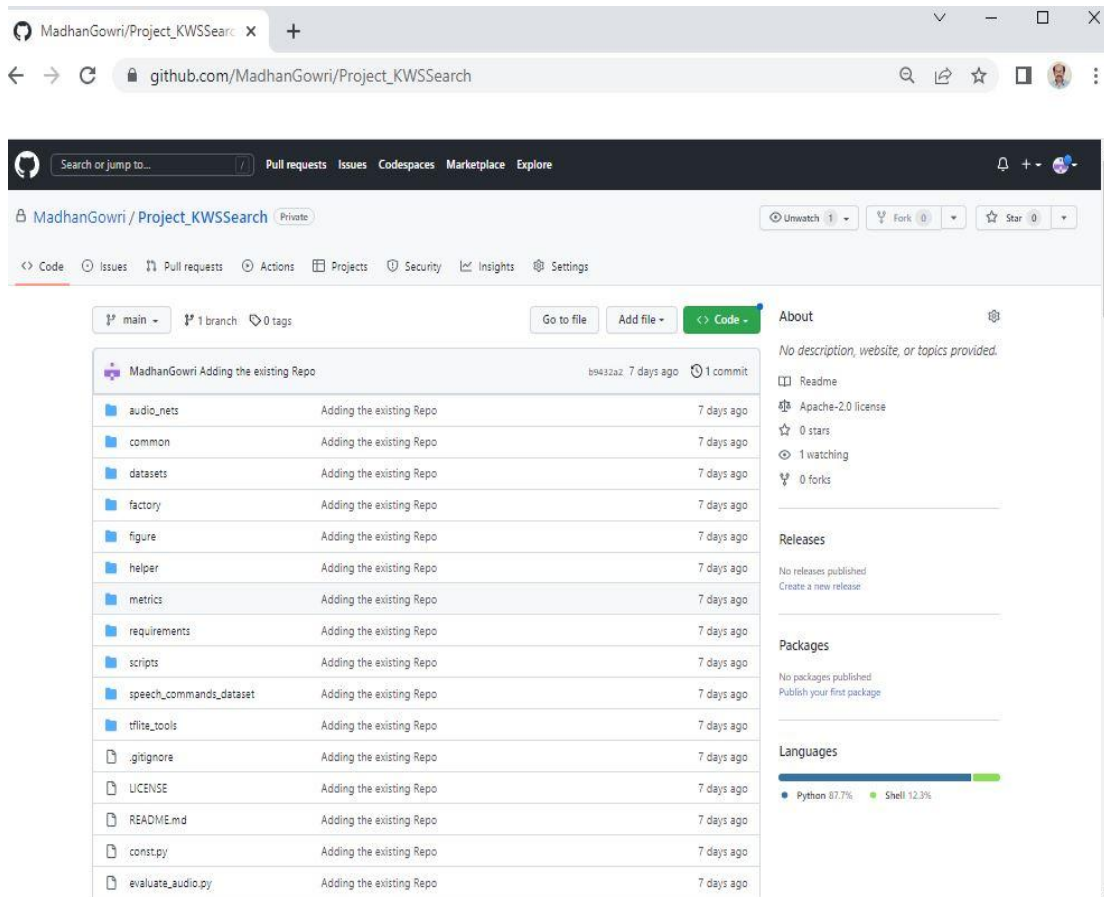
Code repository and Future scope

github.com/MadhanGowri/Project_KWSSearch

Further scope :

➢Model compression and quantization using tflite /Android

➢Model implementation on Android device for Inference

➢Model implementation on low footprint DSP for Inference