NGUYỄN DUY KHÁNH

TOPIC: VOICE ACTIVITY DETECTION USING 1D TIME-CHANNEL SEPARABLE CONVOLUTION

Mentor: Đỗ Văn Hải

Appreciating Analog

Motivation

RECENTLY, THERE ARE MANY SPEECH-BASED APPLICATIONS ACROSS MOBILE AND WEARABLE DEVICES (E.G., VIRTUAL ASSISTANT). HOWEVER, THEIR MEMORY AND COMPUTING POWER IS LIMITED, WHICH REQUIRE LIGHTWEIGHT MODELS. THEREFORE, I PROPOSE A MODEL WITH JUST 74K PARAMETERS BUT STILL ACHIEVE A GOOD RESULT

1. Dataset AVA-speech

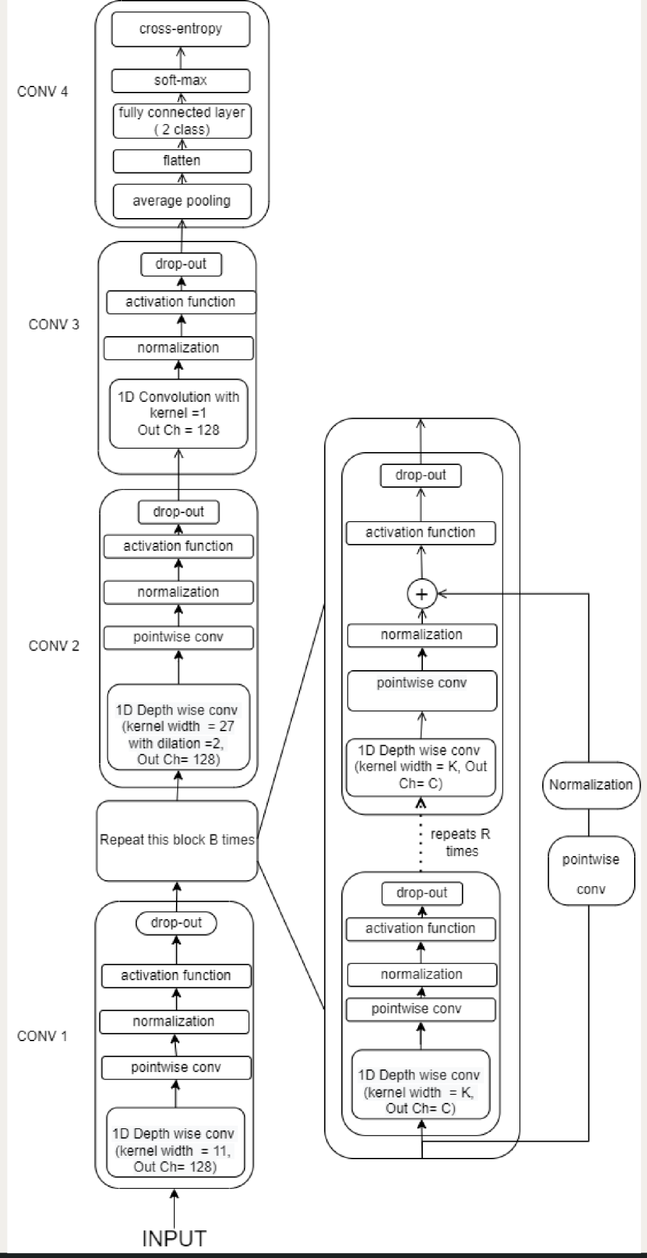
This dataset has 4 label (Clean Speech, Speech with background Music, Speech with background Noise, and No speech), equivalent to ~40K labeled segments spanning 40 hours of data

2. Pre-processing

- Segment to get train, valid, test data (each segment is 0.63s)
• Speech data: 22250 samples
• Background data: 17617 samples
- Pre-processed the audio segments with 64 MFCC features
- Augmentation:
• Add white noise
• Time-shift perputation
• Spectrogram Augmentation

3. Model Architecture

- (B=2 blocks, R=2 sub-blocks per block, C=64 channels) Model includes 1D time-channel separable convolution, batch normalization, ReLU and dropout layers



4. Experiment

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

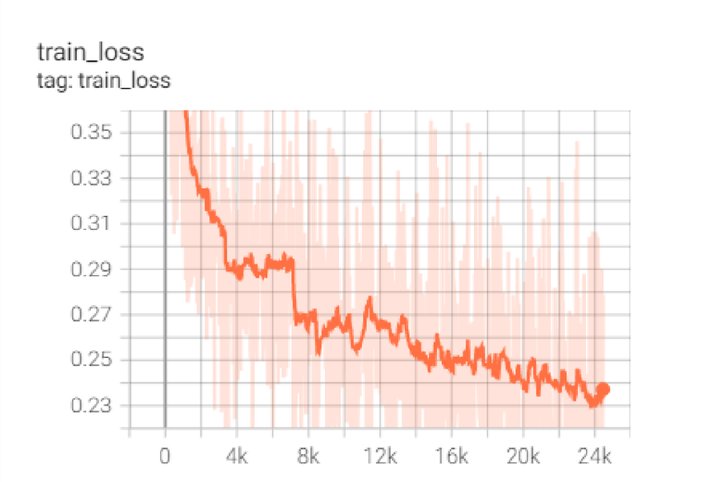
Drop-out rate is set to 0.1 and batch size is set to 256(which is quite stable)

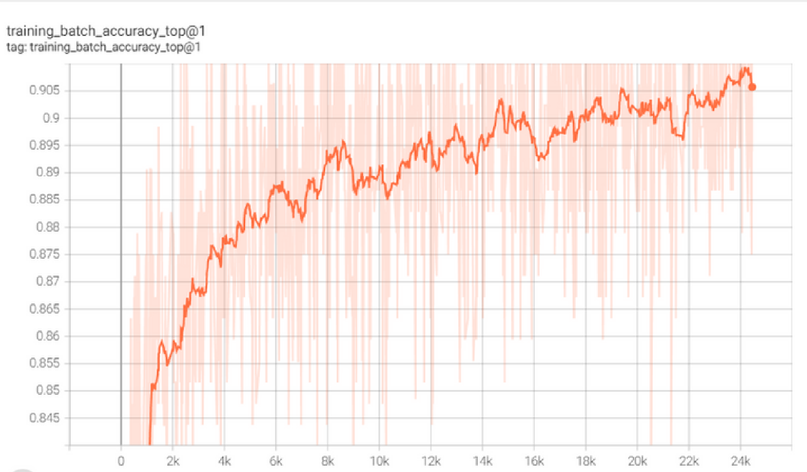
Optimizer: Adam
Learning rate: 0.1

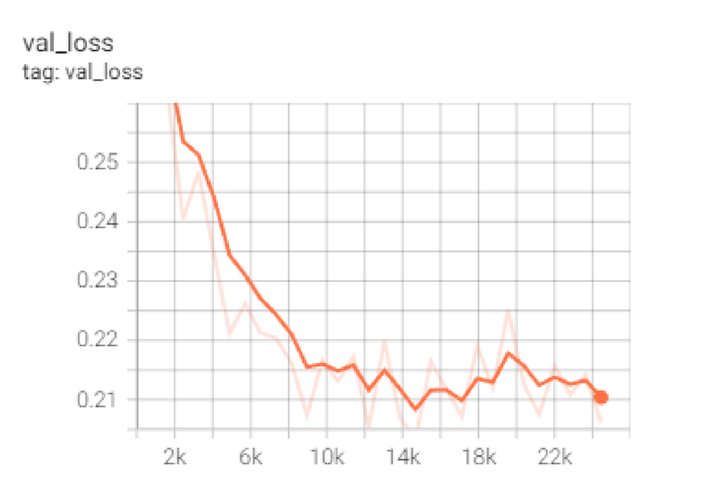
Loss function: cross-entropy

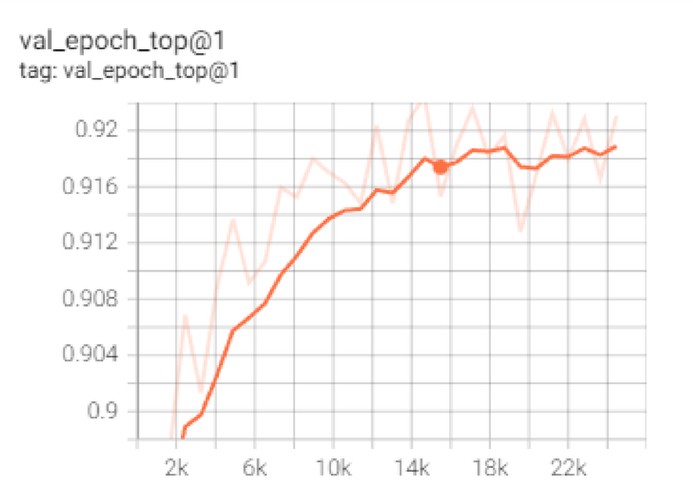
$$L_{CE} = -\sum_{i=1}^{n} q_i \log(p_i)$$

5. Result


train_loss
tag: train_loss
0.35
0.33
0.31
0.29
0.27
0.25
0.23
0
4k
8k
12k
16k
20k
24k


training_batch_accuracy_top@1
tag: training_batch_accuracy_top@1
0.905
0.9
0.895
0.89
0.885
0.88
0.875
0.87
0.865
0.86
0.855
0.85
0.845
0
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
22k
24k


val_loss
tag: val_loss
0.25
0.24
0.23
0.22
0.21
2k
6k
10k
14k
18k
22k


val_epoch_top@1
tag: val_epoch_top@1
0.92
0.916
0.912
0.908
0.904
0.9
2k
6k
10k
14k
18k
22k

Comparision

	All data	clean speech	with music	with noise
Accuracy	93.09%	96.1%	91.27%	92.9%
Precision	96.33%	96.28%	86.3%	91.3%
Recall	91.17%	96.35%	95.18%	93.5%
F1 score	93.67%	96.31%	90.52%	92.39%

	All data	Clean	Music	Noise
CNN-TD	94.5%	98.3%	91.7%	93.9%

5. Limitation

When predicting the audio segment which has human sound (not speech) like laughing, crying, screaming, coughing, the model tend to label it as speech. This may result because they are all human sound like speec

Solution: make a third class, named "non-speech human sound" in order to remove the mistaken of the model

6. Future work

Train system on more class (for example : speech, background, laughter, screaming)