Audio Classification

Speech Command Recognition with *torchaudio*

Dataset: SpeechCommands Dataset

No of Commands: **35**

Audio file length: **1** sec each

Sampling frequency: 16kHz

Torchaudio library to convert **audio files(.wav)** to **tensors**

Training**: 105829**

Testing: **11005**

Validation**: 9981**

Excluded while training

Total audio clip : 126815

Speech Command List

'backward', 'bed', 'bird', 'cat', 'dog', 'down', 'eight', 'five', 'follow', 'forward', 'four', 'go',

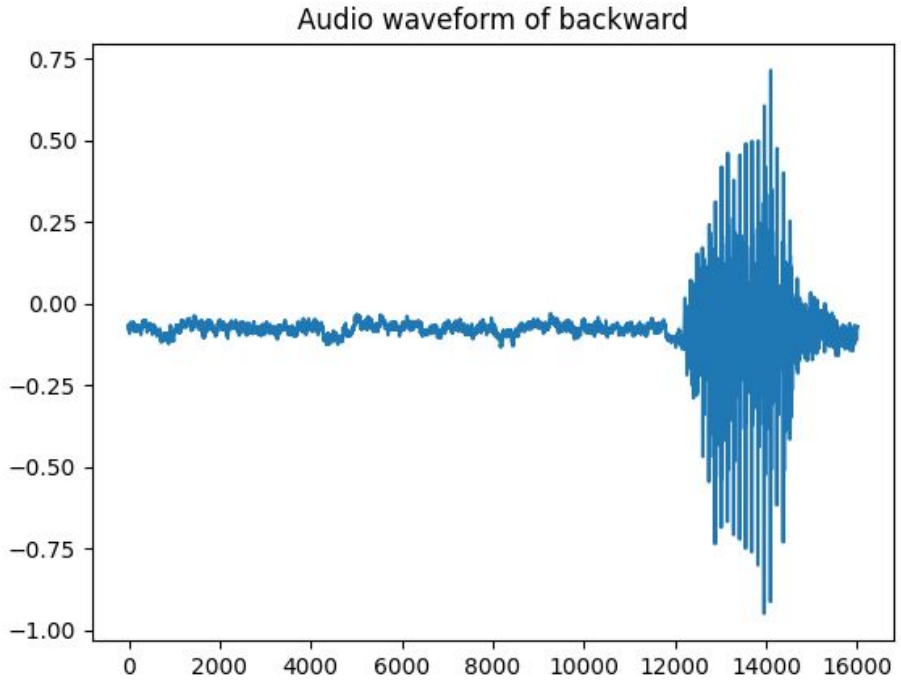
'happy', 'house', 'learn', 'left', 'marvin', 'nine', 'no', 'off', 'on', 'one', 'right', 'seven',

'sheila', 'six', 'stop', 'three', 'tree', 'two', 'up', 'visual', 'wow', 'yes', 'zero'

Sample Audio Clip:

backward

Sample:


Audio waveform of backward
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2000
4000
6000
8000
10000
12000
14000
16000

Network:

```
M5(
  (conv1): Conv1d(1, 32, kernel size=(80,), stride=(16,))
  (bn1): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=True, track running stats=True)
  (pool1): MaxPool1d(kernel size=4, stride=4, padding=0, dilation=1, ceil mode=False)
  (conv2): Conv1d(32, 32, kernel size=(3,), stride=(1,))
  (bn2): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=True, track running stats=True)
  (pool2): MaxPool1d(kernel size=4, stride=4, padding=0, dilation=1, ceil mode=False)
  (conv3): Conv1d(32, 64, kernel size=(3,), stride=(1,))
  (bn3): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track running stats=True)
  (pool3): MaxPool1d(kernel size=4, stride=4, padding=0, dilation=1, ceil mode=False)
  (conv4): Conv1d(64, 64, kernel size=(3,), stride=(1,))
  (bn4): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track running stats=True)
  (pool4): MaxPool1d(kernel size=4, stride=4, padding=0, dilation=1, ceil mode=False)
  (fc1): Linear(in features=64, out features=35, bias=True)
)
Number of parameters: 26915
```

Training Configuration:

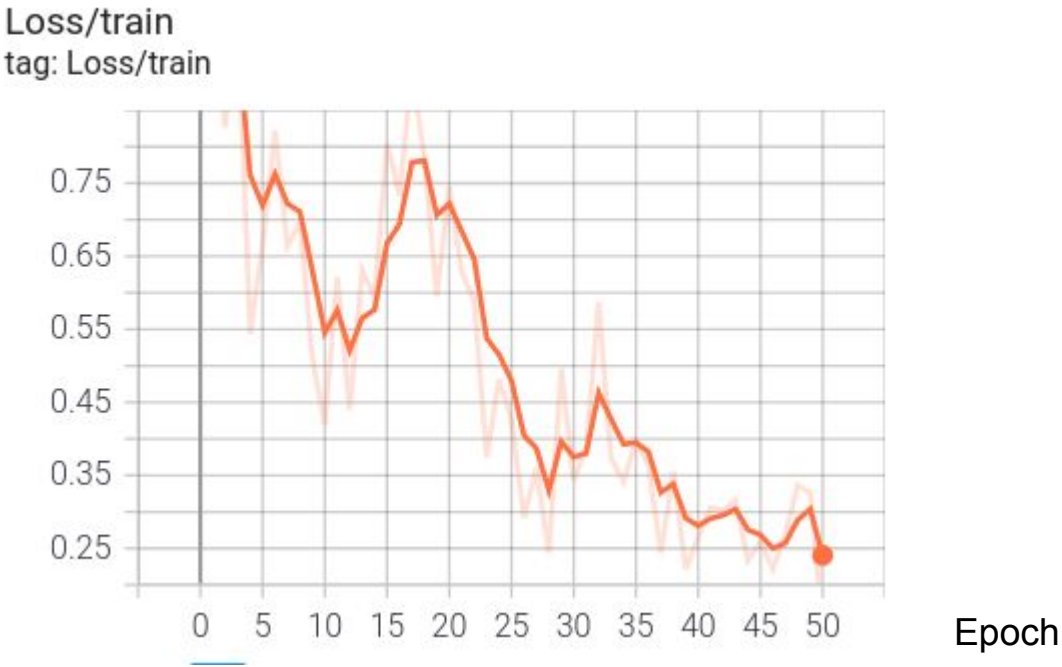
Training_size: **105829**

Lr : **0.01** with scheduler (to decrease by 0.001 after 20 epochs)

Epoch : **50**

Final activation layer: Log_**Softmax**

Loss


Loss/train
tag: Loss/train
0.75
0.65
0.55
0.45
0.35
0.25
0
5
10
15
20
25
30
35
40
45
50
Epoch

Accuracy:

Inference result of Training audio data

Accuracy: 98420/105829 (93%)

Inference result of Testing audio data1

Accuracy: 10289/11005 (93%)

Inference result of Testing audio data2

Accuracy: 9379/9981 (94%)

0 = backward

1 = bed

2= bird

3 = cat

4 = dog

5 = down

6 = eight

7= five

8 = follow

9 = forward

10 = four

11 = go

12 = happy

13 = house

14 = learn

15 = left

16 = marvin

17 = nine

18 = no

19 = off

20 = on

21 = one

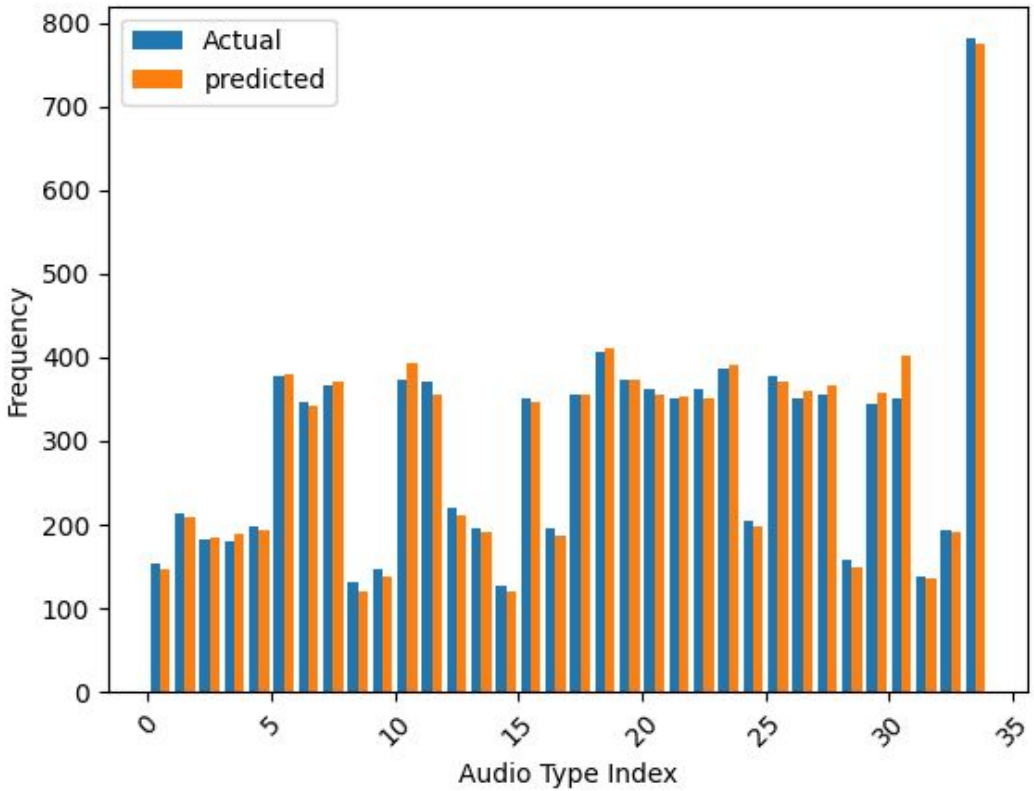
22 = right

23 = seven

24 = sheila

25 = six

26 = stop



27 = three

28 = tree

29 = two

30 = up

31 = visual

32 =wow

33 = yes

34 = zero

Confusion Matrix: