Speech command classification

Team 5

Jeong Jae Yoon, Jin Chang Ho, Lee Dong Ho

Tensorflow Model

https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/audio/simple_audio.ipynb

Our Model

https://colab.research.google.com/drive/1DkhJiNiRlY8czFBjL28K-WxdUI3pxHRp?usp=sharing

1. About Model

Layer (type)	Output Shape	Param #
resizing_2 (Resizing)	(None, 32, 32, 1)	0
normalization_2 (Normalizati	(None, 32, 32, 1)	3
conv2d_4 (Conv2D)	(None, 30, 30, 32)	320
conv2d_5 (Conv2D)	(None, 28, 28, 64)	18496
max_pooling2d_2 (MaxPooling2	(None, 14, 14, 64)	0
dropout_4 (Dropout)	(None, 14, 14, 64)	0
flatten_2 (Flatten)	(None, 12544)	0
dense_4 (Dense)	(None, 128)	1605760
dropout_5 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 8)	1032

Total params: 1,625,611
Trainable params: 1,625,608
Non-trainable params: 3

Tensorflow model: The Tensorflow model is composed of Conv2d, MaxPooling2d, Flatten etc. It used keras API.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 8000, 1)]	0
conv1d (Conv1D)	(None, 7988, 8)	112
max_pooling1d (MaxPooling1D)	(None, 2662, 8)	0
dropout (Dropout)	(None, 2662, 8)	0
conv1d_1 (Conv1D)	(None, 2652, 16)	1424
max_pooling1d_1 (MaxPooling1	(None, 884, 16)	0
dropout_1 (Dropout)	(None, 884, 16)	0
conv1d_2 (Conv1D)	(None, 876, 32)	4640
max_pooling1d_2 (MaxPooling1	(None, 292, 32)	0
dropout_2 (Dropout)	(None, 292, 32)	0
conv1d_3 (Conv1D)	(None, 286, 64)	14400
max_pooling1d_3 (MaxPooling1	(None, 95, 64)	0
dropout_3 (Dropout)	(None, 95, 64)	0
flatten (Flatten)	(None, 6080)	0
dense (Dense)	(None, 256)	1556736
dropout_4 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32896
dropout_5 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 8)	1032

Total params: 1,611,240
Trainable params: 1,611,240
Non-trainable params: 0

The model that we make: Our model is composed of multiple layer, Conv1d, Maxpooling, Dropout, Dense, Flatten etc. We used keras API.
While referring to the tensorflow model, we tried to make a difference.

2. About dataset

We have 8 categories, [yes, no, right, stop, go, up, left, down]. Each category has 1000 wav files and a total of 8000 wav files. 6400 files were used for training, 800 for validation, and 800 for testing.
This data could easily get audio files of words spoken by various people. And to save time, We set the dataset to a smaller size.
Download link:
http://storage.googleapis.com/download.tensorflow.org/data/mini_speech_commands.zip

3. Performed Experiments

We controlled the value of EPOCH, the value of x(just a variable) of Conv2D(x,3), Resizing(x,x) at the tensorflow model to observe how it worked. Also, we changed the number of layer, dropout, and batch size in the model that we make to see the results.

4. Experiment Results

A. Tensorflow Model (x is just a variable)

EPOCH

EPOCH	5	10	15	20	25
Accuracy	80%	84%	86%	85%	87%
			Stop at 5	Stop at 3	Stop at 3

1. EPOCH : the larger EPOCH, the more accurate. However, does not matter of accuracy if it is more than 15.

Conv2D(x, 3)

x's max	32	64	96	128	160
Accuracy	69%	84%	87%	88%	88%

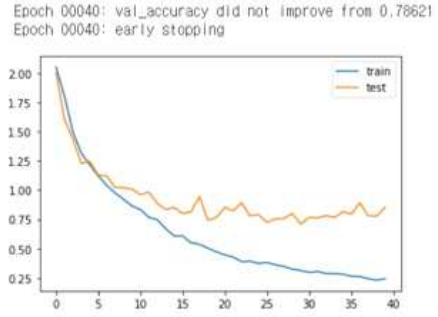
2. Conv2D(x, 3) : the larger x, the more accurate. However, does not matter of accuracy if it is more than 128.

Resizing(x, x)

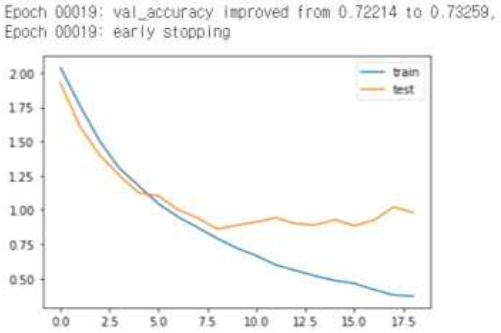
x's max	8	16	32	64	128
Accuracy	54%	75%	84%	85%	79%

3. Resizing(x, x) : As x increases, the accuracy increases rapidly, but the accuracy decreases certain value between 64 and 128.

B. Our Model

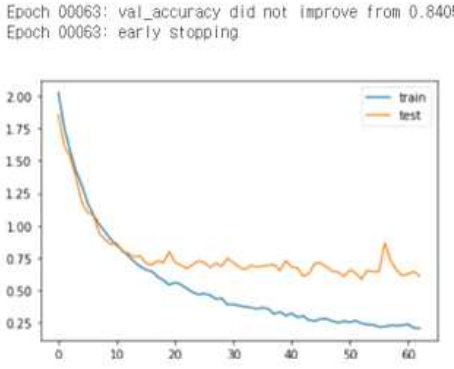


Original model
78.6%

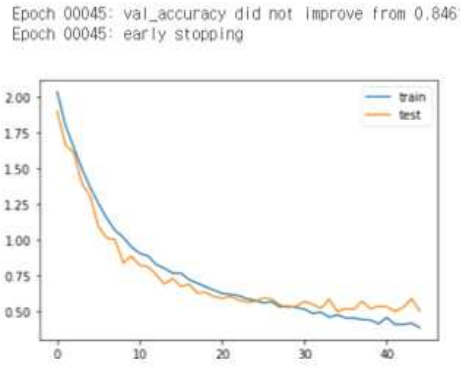


Dropout 0.3 -> 0.2
73.2%

1. Dropout: When we changed dropout value to more or less than 0.3, accuracy decreased.



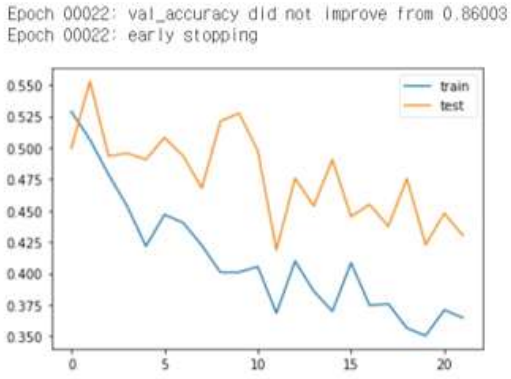
Add 1 layer
84.0%



Add 2 layers
84.6%

2. Layers: Accuracy got higher when we added more layers. However, there was

little increase in accuracy when there were more than seven layers.



Add 2 layer, Batch size 32->16

86.0%

3. Batch size : When we reduced the batch size from 32 to 16, accuracy increased but the loss graph got jagged.