

Report on Speech Command Recognition

1. Logistics

- Start Time: Wed Sep 11, 10:00 AM
- End Time: Wed Sep 11, 04:30 PM

2. Task

1. Paper Summary:

The paper titled ["Deep Speech: Scaling up end-to-end speech recognition"](<https://arxiv.org/abs/1804.03209>) explores deep learning models for speech recognition, particularly focusing on the scalability and end-to-end nature of the models. The authors propose architectures and techniques that improve accuracy and efficiency in recognizing spoken language.

2. Dataset Analysis:

- Dataset: The dataset from the paper was downloaded and analyzed.
- Statistical Analysis: The dataset comprises audio recordings of various spoken commands. Statistical analysis was performed to describe the distribution of command types, sample length, and audio features. This involved code snippets that demonstrate data loading, feature extraction, and visualization.

Dataset Summary:

Total number of audio files: 105829

Number of unique commands: 35

Average duration: 0.98 seconds

Average sample rate: 16000.00 Hz

Average spectral centroid: 1845.07 Hz

Average zero crossing rate: 0.1389

Top 5 most common commands:

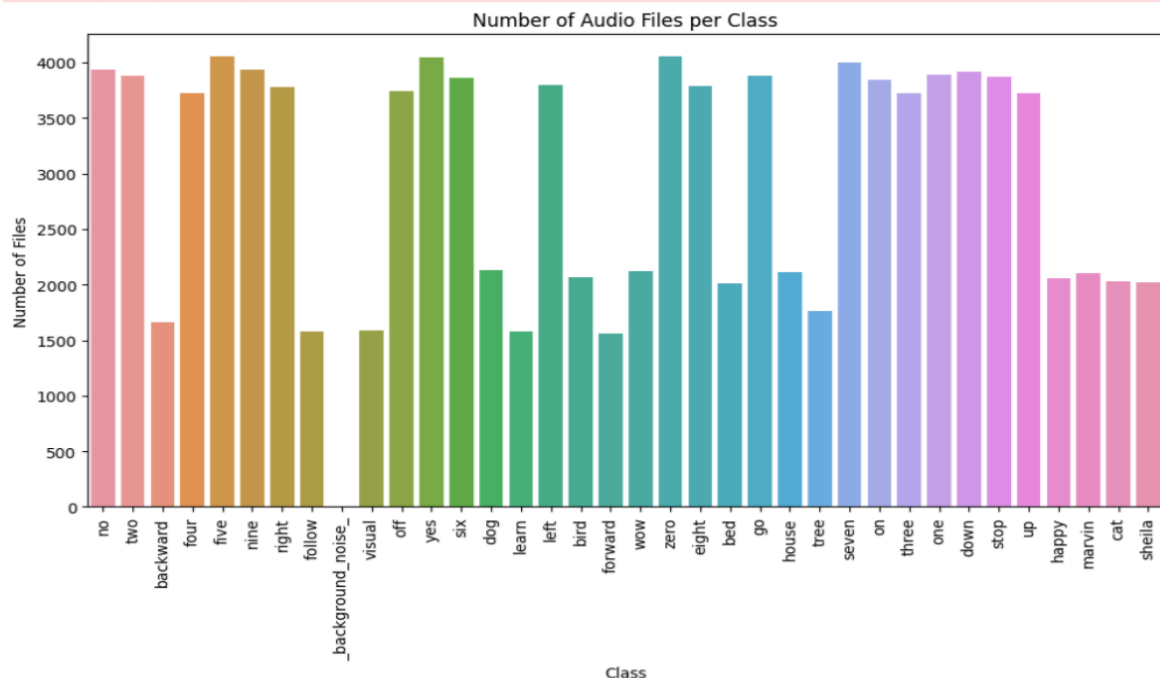
five: 4052

zero: 4052

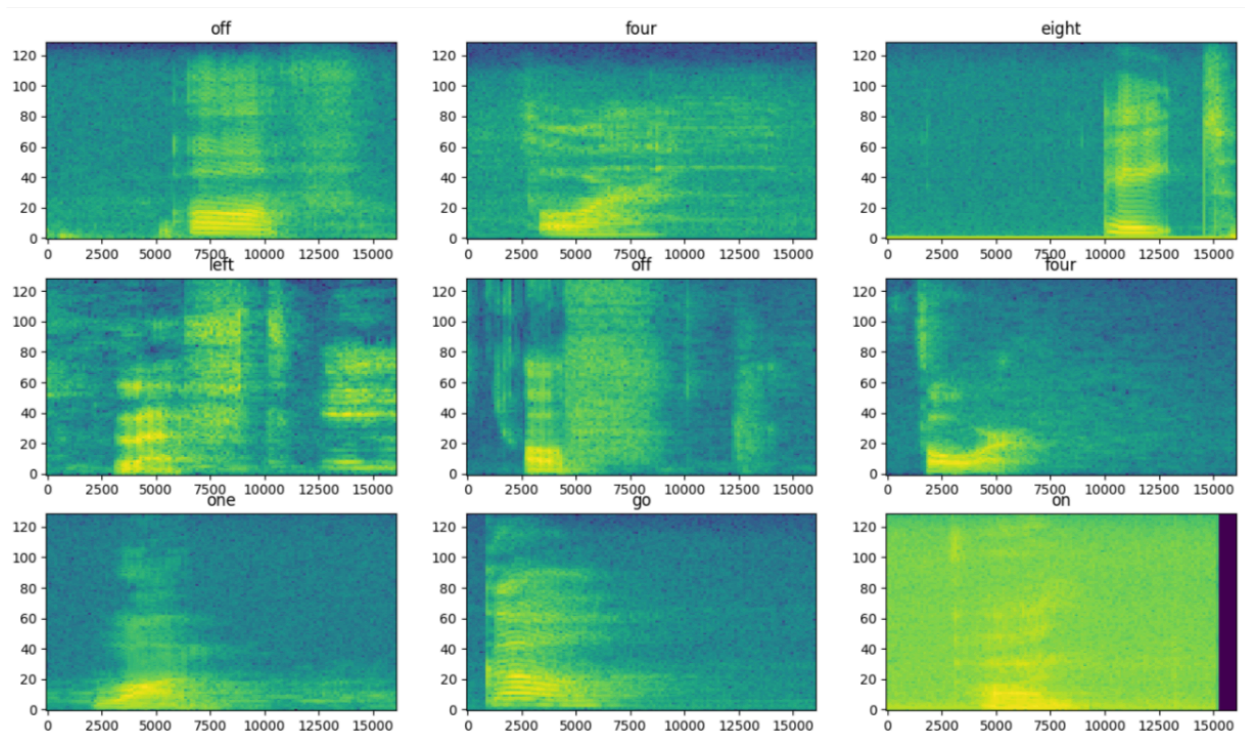
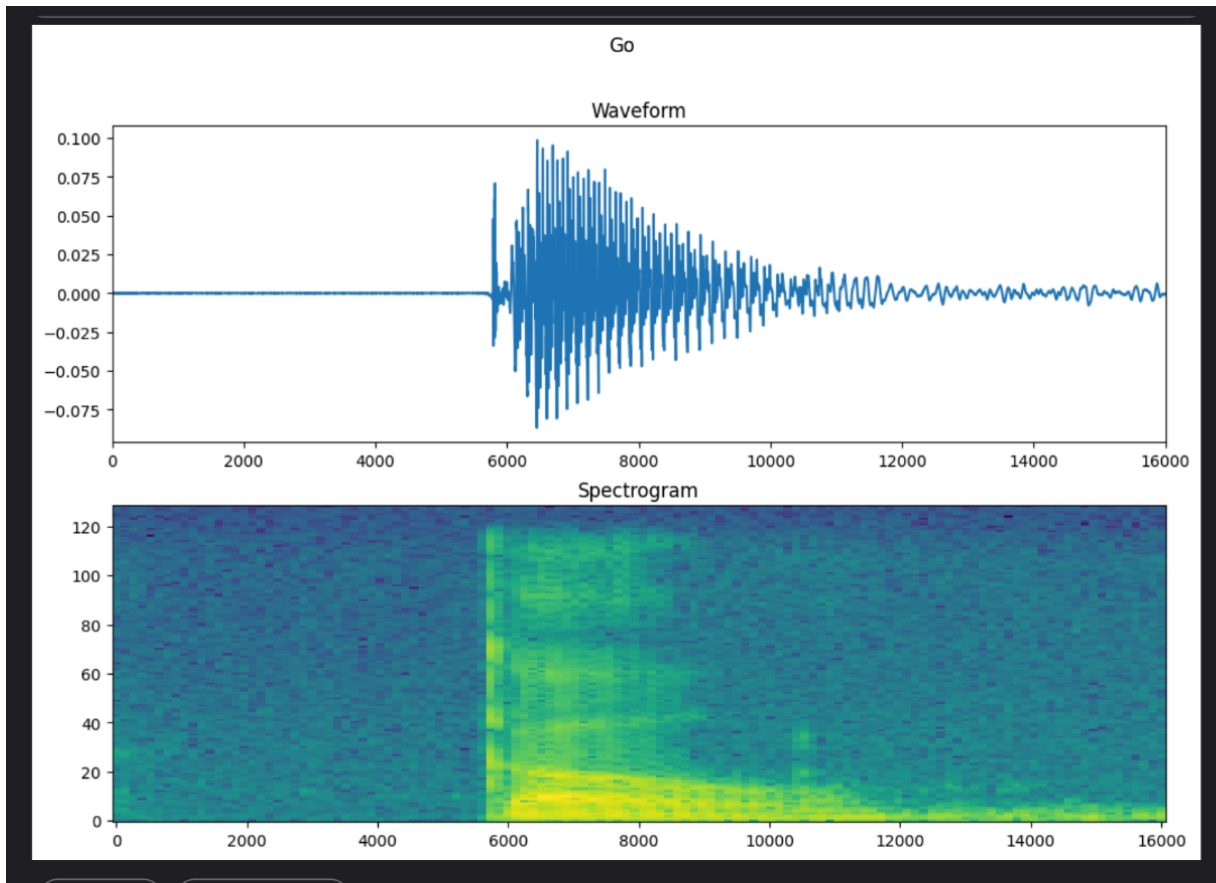
yes: 4044

seven: 3998

no: 3941



Class: no, Average Duration: 0.98 seconds
Class: two, Average Duration: 0.98 seconds
Class: backward, Average Duration: 0.99 seconds
Class: four, Average Duration: 0.98 seconds
Class: five, Average Duration: 0.98 seconds
Class: nine, Average Duration: 0.98 seconds
Class: right, Average Duration: 0.98 seconds
Class: follow, Average Duration: 0.98 seconds
Class: _background_noise_, Average Duration: 66.57 seconds
Class: visual, Average Duration: 0.98 seconds
Class: off, Average Duration: 0.98 seconds
Class: yes, Average Duration: 0.98 seconds
Class: six, Average Duration: 0.99 seconds
Class: dog, Average Duration: 0.97 seconds
Class: learn, Average Duration: 0.97 seconds
Class: left, Average Duration: 0.98 seconds
Class: bird, Average Duration: 0.97 seconds
Class: forward, Average Duration: 0.98 seconds
Class: wow, Average Duration: 0.97 seconds
Class: zero, Average Duration: 0.99 seconds
Class: eight, Average Duration: 0.98 seconds
Class: bed, Average Duration: 0.97 seconds
Class: go, Average Duration: 0.98 seconds
Class: house, Average Duration: 0.97 seconds
Class: tree, Average Duration: 0.97 seconds
Class: seven, Average Duration: 0.98 seconds
Class: on, Average Duration: 0.98 seconds
Class: three, Average Duration: 0.98 seconds
Class: one, Average Duration: 0.98 seconds
Class: down, Average Duration: 0.98 seconds
Class: stop, Average Duration: 0.98 seconds
Class: up, Average Duration: 0.98 seconds
Class: happy, Average Duration: 0.97 seconds
Class: marvin, Average Duration: 0.98 seconds
Class: cat, Average Duration: 0.97 seconds
Class: sheila, Average Duration: 0.98 seconds



3. Classifier Training:

- Model: A classifier was trained to recognize commands in the dataset.

Model: "sequential"

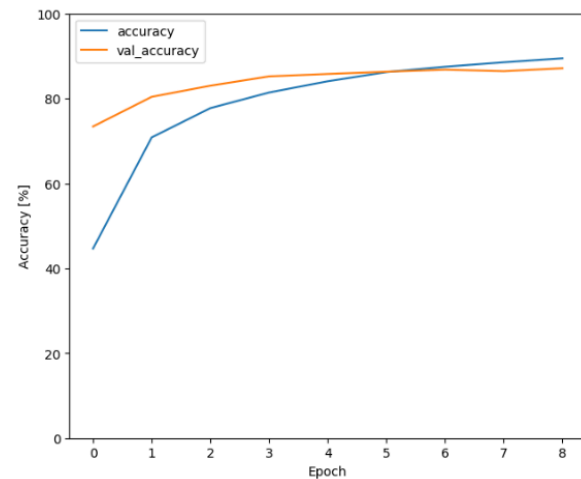
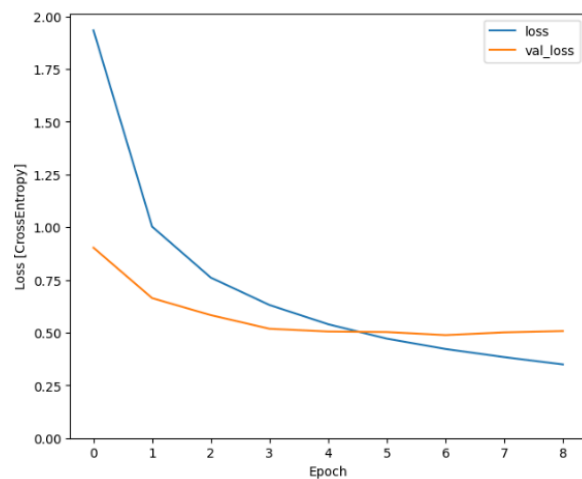
Layer (type)	Output Shape	Param #
resizing (Resizing)	(None, 32, 32, 1)	0
normalization (Normalization)	(None, 32, 32, 1)	3
conv2d (Conv2D)	(None, 30, 30, 16)	160
conv2d_1 (Conv2D)	(None, 28, 28, 32)	4,640
conv2d_2 (Conv2D)	(None, 26, 26, 64)	18,496
conv2d_3 (Conv2D)	(None, 24, 24, 128)	73,856
max_pooling2d (MaxPooling2D)	(None, 12, 12, 128)	0
dropout (Dropout)	(None, 12, 12, 128)	0
flatten (Flatten)	(None, 18432)	0
dense (Dense)	(None, 128)	2,359,424
dense_1 (Dense)	(None, 64)	8,256
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 36)	2,340

Total params: 2,467,175 (9.41 MB)

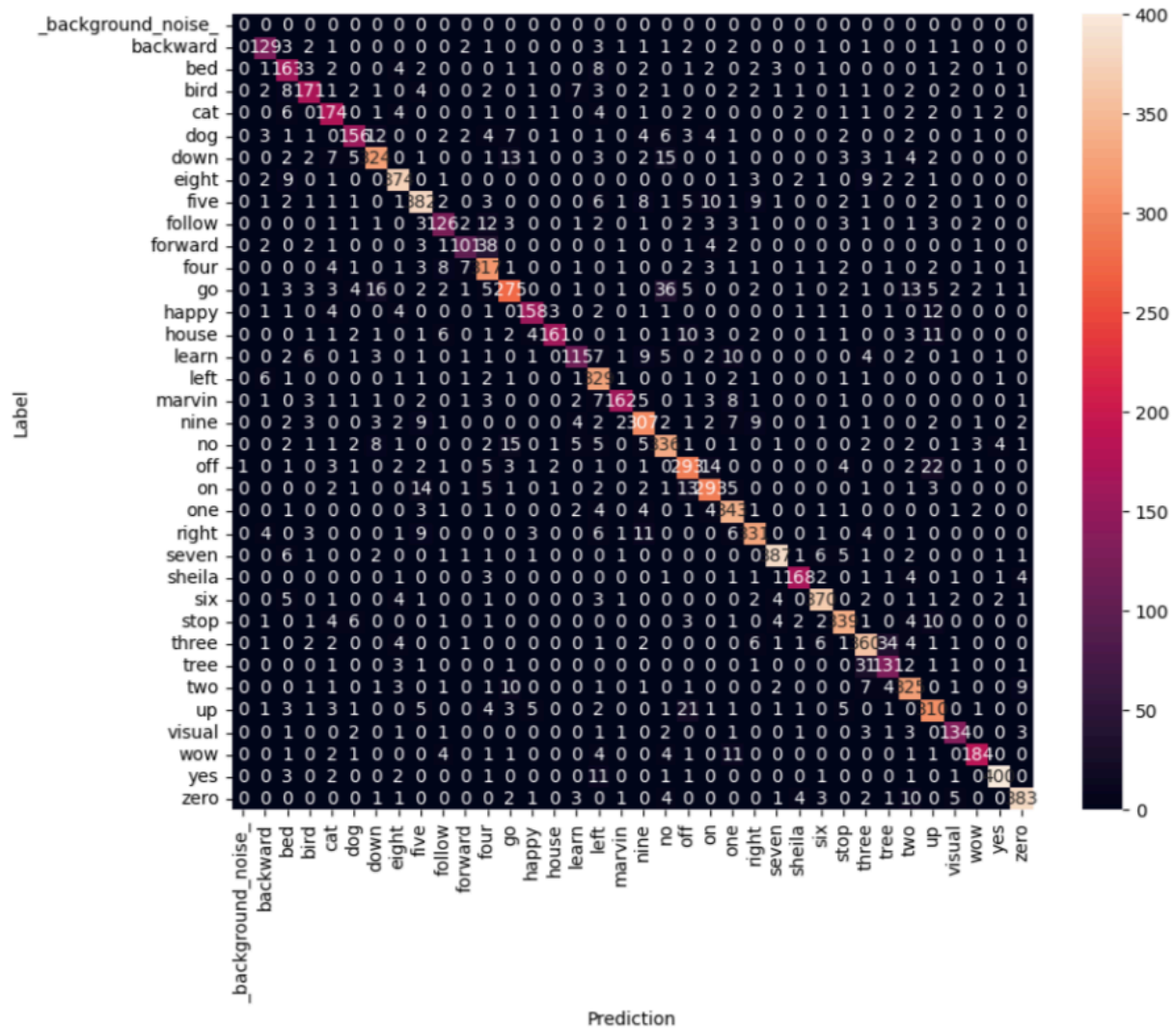
Trainable params: 2,467,172 (9.41 MB)

Non-trainable params: 3 (16.00 B)

Text(0, 0.5, 'Accuracy [%]')



4. Performance Report:



Precision: 0.8712

Recall: 0.8683

F1 Score: 0.8679

	precision	recall	f1-score	support
backward	0.93	0.85	0.89	152
bed	0.87	0.80	0.83	200
bird	0.86	0.79	0.83	218
cat	0.88	0.82	0.85	206
dog	0.83	0.80	0.81	213
down	0.90	0.83	0.86	390
eight	0.93	0.94	0.93	408
five	0.87	0.86	0.87	442
follow	0.83	0.67	0.74	172
forward	0.85	0.69	0.76	157
four	0.81	0.90	0.85	360
go	0.82	0.78	0.80	388
happy	0.94	0.87	0.91	192
house	0.92	0.82	0.87	213
learn	0.80	0.62	0.70	173
left	0.84	0.92	0.88	351
marvin	0.90	0.87	0.89	204
nine	0.81	0.92	0.86	363
no	0.79	0.91	0.85	400
off	0.75	0.89	0.81	360
on	0.79	0.87	0.83	346
one	0.80	0.94	0.87	370
right	0.87	0.91	0.89	381
seven	0.97	0.92	0.95	418
sheila	0.93	0.88	0.91	190
six	0.95	0.93	0.94	401
stop	0.90	0.92	0.91	380
three	0.86	0.89	0.87	428
tree	0.82	0.73	0.77	174
two	0.93	0.86	0.89	369
up	0.83	0.80	0.82	371
visual	0.89	0.90	0.90	155
wow	0.88	0.88	0.88	216
yes	0.95	0.94	0.95	424
zero	0.95	0.92	0.93	422
accuracy			0.87	10607
macro avg	0.87	0.85	0.86	10607
weighted avg	0.87	0.87	0.87	10607