Transformers

Marta Szuwarska, Mateusz Nizwantowski

musimy jakis plan zrobic

- data loader (introduction, silence, unknown, mfcc)
- Transformer basic architecture
- Transformer results
- other models listed
- for ever model introduction + results
- for best model plots

musimy jakis plan zrobic

- data loader (introduction, silence, unknown, mfcc)
- struktura plikow, how we handle silence
- mfcc
- dlugo nam sie liczy slabe wyniki

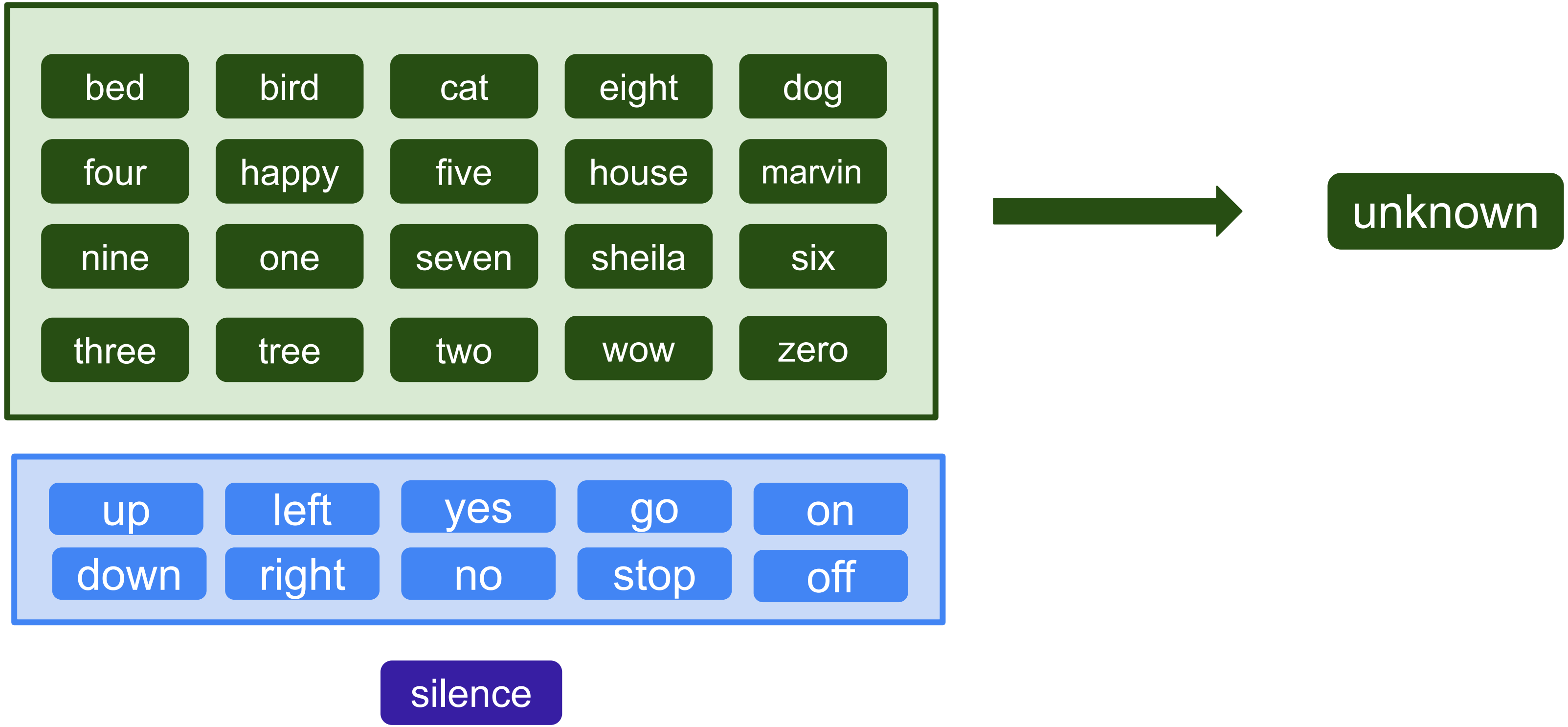
Original classes

bed bird cat eight dog
four happy five house marvin
nine one seven sheila six
three tree two wow zero

up left yes go on
down right no stop off

silence

Original classes



bed bird cat eight dog

four happy five house marvin

nine one seven sheila six

three tree two wow zero

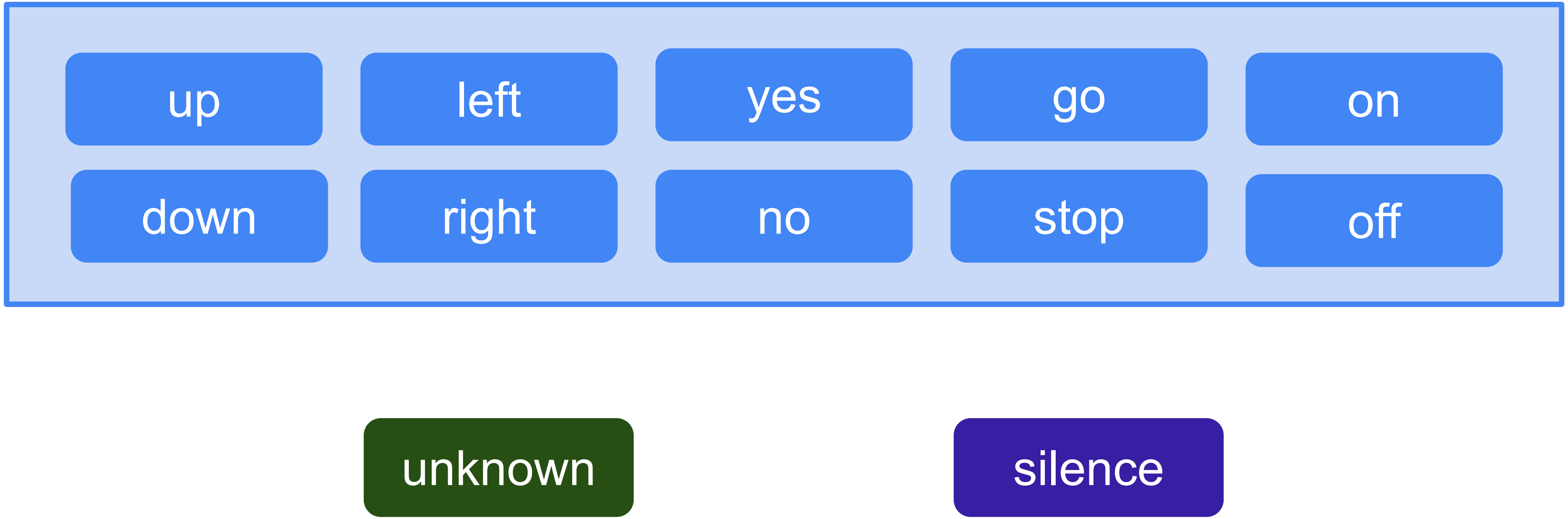
unknown

up left yes go on

down right no stop off

silence

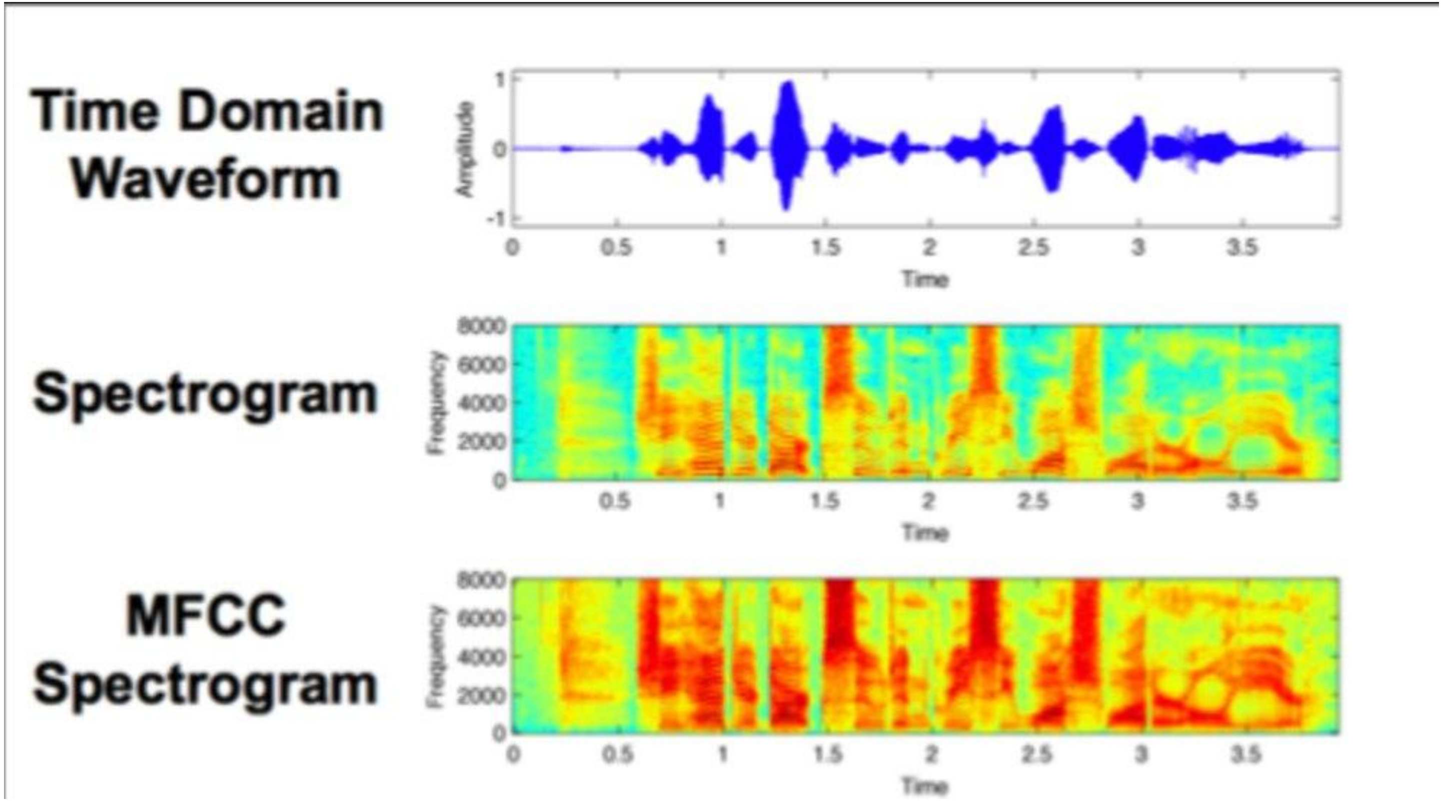
Modified classes



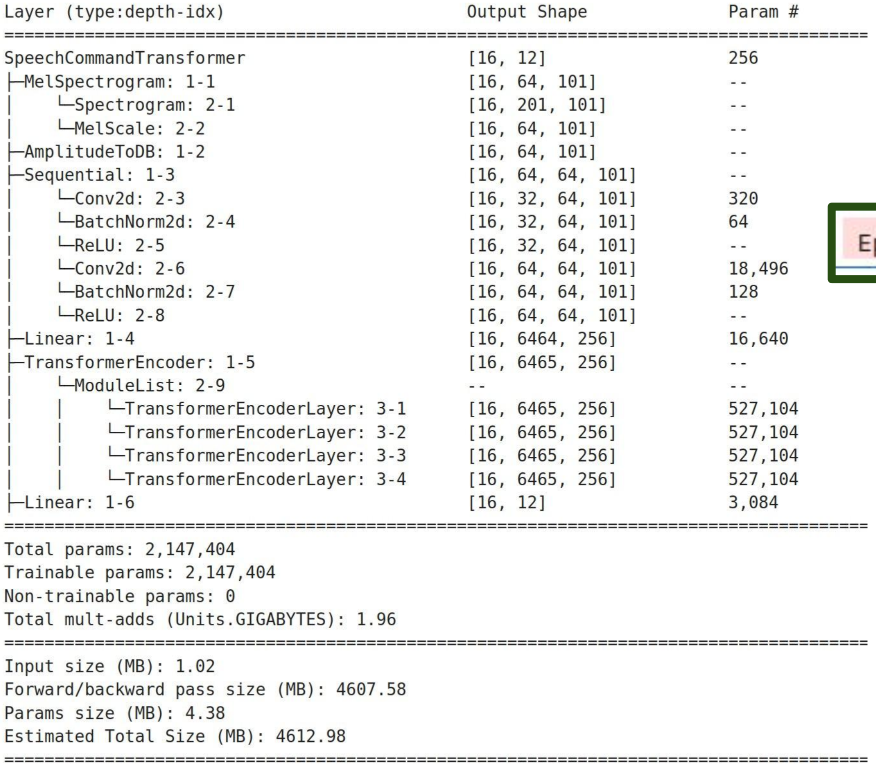
up	left	yes	go	on
down	right	no	stop	off

unknown

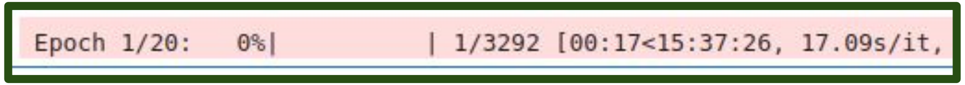
silence


Time Domain Waveform
Amplitude
1
0
-1
0
0.5
1
1.5
2
2.5
3
3.5
Time
Spectrogram
Frequency
8000
6000
4000
2000
0
0.5
1
1.5
2
2.5
3
3.5
Time
MFCC Spectrogram
Frequency
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
Time

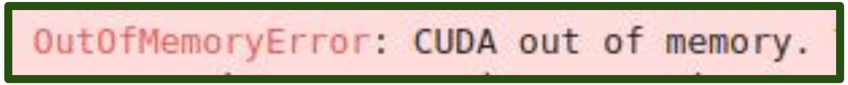
Early transformer architecture



```
Layer (type:depth-idx)                        Output Shape              Param #
=======================================================================================
SpeechCommandTransformer                      [16, 12]                  256
├─MelSpectrogram: 1-1                         [16, 64, 101]             --
│    └─Spectrogram: 2-1                       [16, 201, 101]            --
│    └─MelScale: 2-2                          [16, 64, 101]             --
├─AmplitudeToDB: 1-2                          [16, 64, 101]             --
├─Sequential: 1-3                             [16, 64, 64, 101]         --
│    └─Conv2d: 2-3                            [16, 32, 64, 101]         320
│    └─BatchNorm2d: 2-4                       [16, 32, 64, 101]         64
│    └─ReLU: 2-5                              [16, 32, 64, 101]         --
│    └─Conv2d: 2-6                            [16, 64, 64, 101]         18,496
│    └─BatchNorm2d: 2-7                       [16, 64, 64, 101]         128
│    └─ReLU: 2-8                              [16, 64, 64, 101]         --
├─Linear: 1-4                                 [16, 6464, 256]           16,640
├─TransformerEncoder: 1-5                     [16, 6465, 256]           --
│    └─ModuleList: 2-9                        --                        --
│    │    └─TransformerEncoderLayer: 3-1      [16, 6465, 256]           527,104
│    │    └─TransformerEncoderLayer: 3-2      [16, 6465, 256]           527,104
│    │    └─TransformerEncoderLayer: 3-3      [16, 6465, 256]           527,104
│    │    └─TransformerEncoderLayer: 3-4      [16, 6465, 256]           527,104
├─Linear: 1-6                                 [16, 12]                  3,084
=======================================================================================
Total params: 2,147,404
Trainable params: 2,147,404
Non-trainable params: 0
Total mult-adds (Units.GIGABYTES): 1.96
=======================================================================================
Input size (MB): 1.02
Forward/backward pass size (MB): 4607.58
Params size (MB): 4.38
Estimated Total Size (MB): 4612.98
=======================================================================================
```

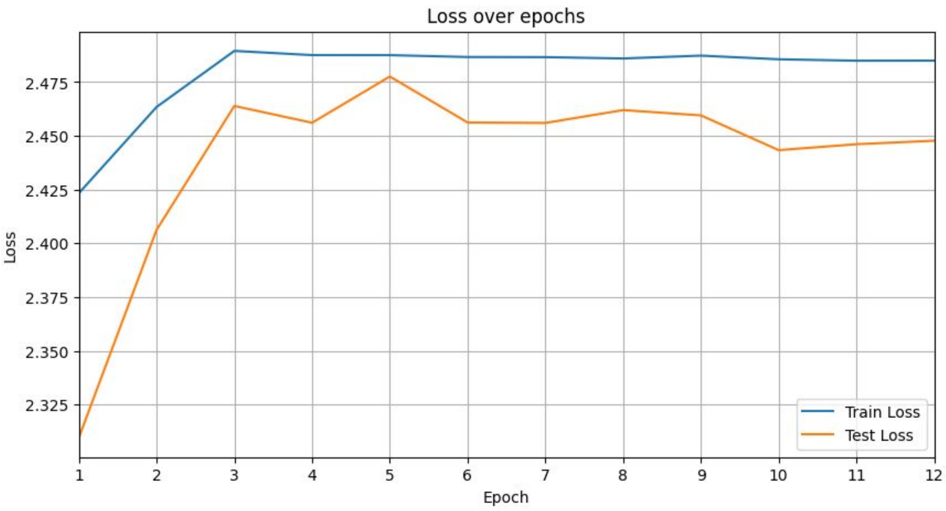


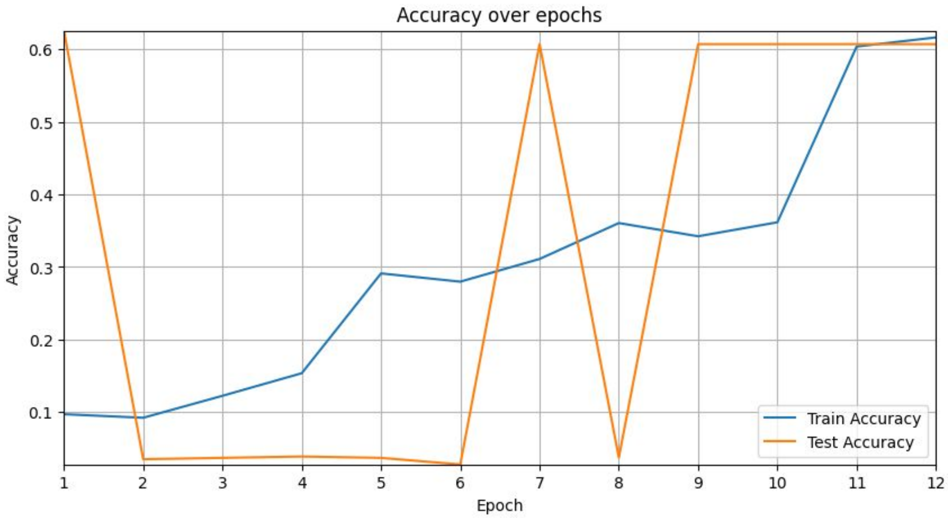
```
Epoch 1/20:   0%|          | 1/3292 [00:17<15:37:26, 17.09s/it,
```



```
OutOfMemoryError: CUDA out of memory.
```

Early transformer results

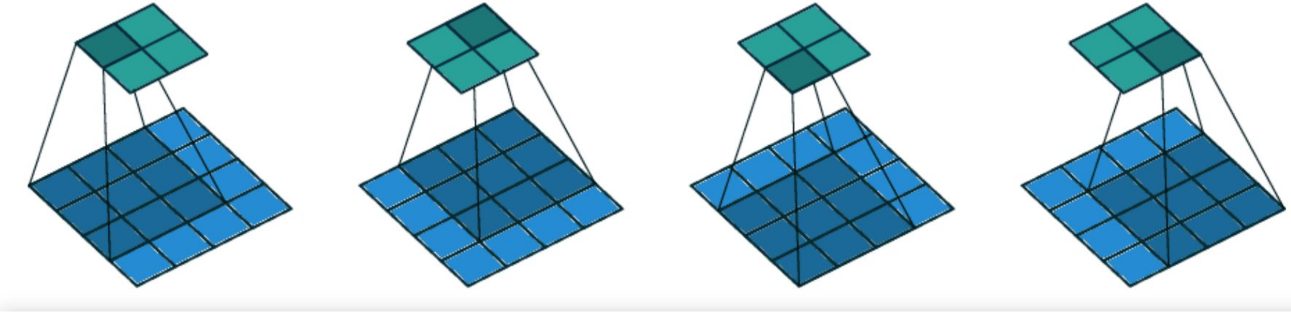




Best test accuracy: 60.71%, corresponding train accuracy: 34.23%

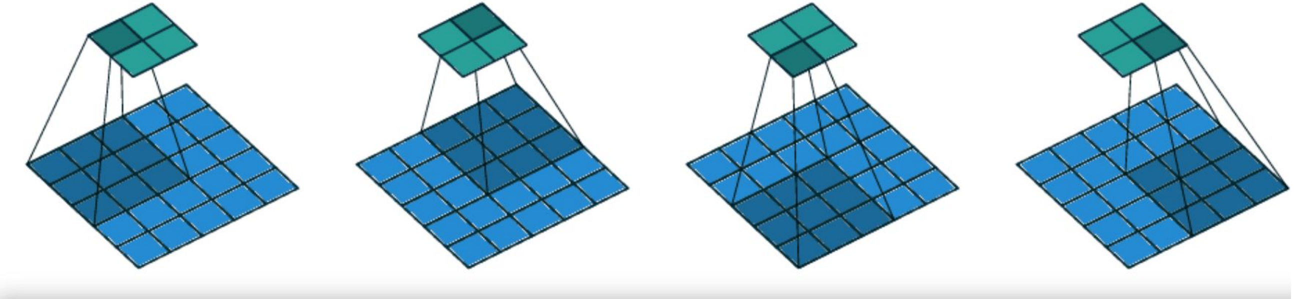
Strided convolution

stride=1



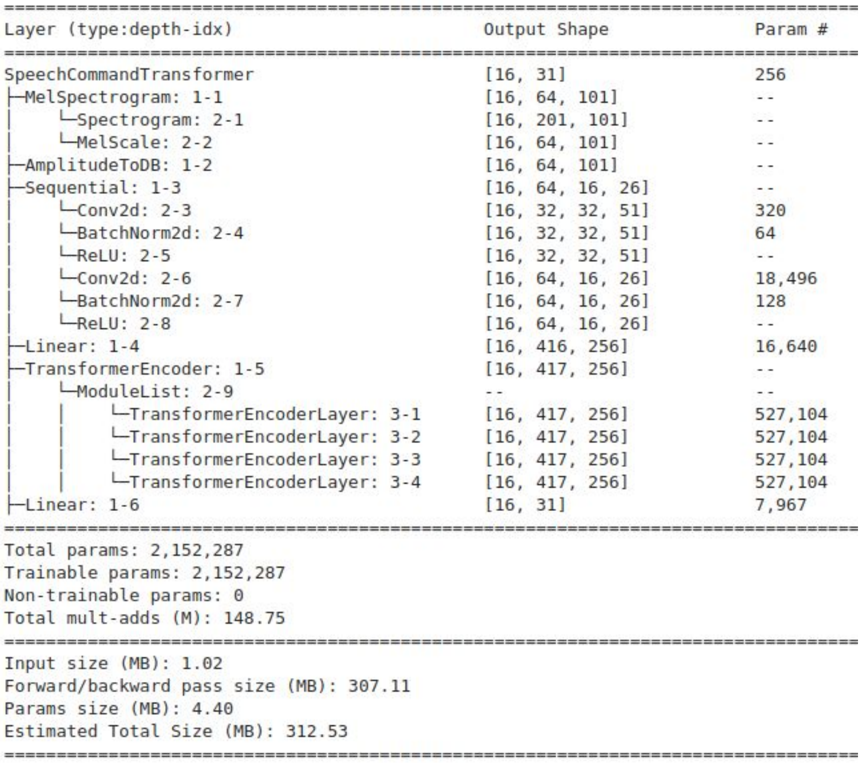
https://www.baeldung.com/wp-content/uploads/sites/4/2023/10/Screenshot-2023-10-10-at-1.11.45-PM.png

stride=2

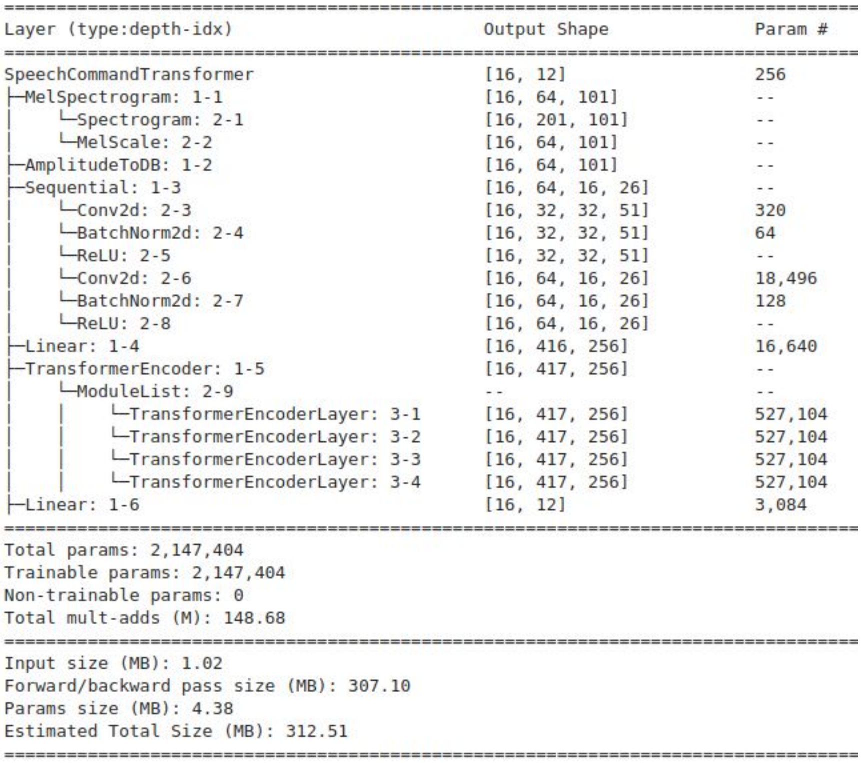


https://www.baeldung.com/wp-content/uploads/sites/4/2023/10/Screenshot-2023-10-10-at-1.11.23-PM.png

Strided transformer architecture - original vs modified



```
===========================================================================================
Layer (type:depth-idx)                        Output Shape              Param #
===========================================================================================
SpeechCommandTransformer                      [16, 31]                  256
├─MelSpectrogram: 1-1                         [16, 64, 101]             --
│    └─Spectrogram: 2-1                       [16, 201, 101]            --
│    └─MelScale: 2-2                          [16, 64, 101]             --
├─AmplitudeToDB: 1-2                          [16, 64, 101]             --
├─Sequential: 1-3                             [16, 64, 16, 26]          --
│    └─Conv2d: 2-3                            [16, 32, 32, 51]          320
│    └─BatchNorm2d: 2-4                       [16, 32, 32, 51]          64
│    └─ReLU: 2-5                              [16, 32, 32, 51]          --
│    └─Conv2d: 2-6                            [16, 64, 16, 26]          18,496
│    └─BatchNorm2d: 2-7                       [16, 64, 16, 26]          128
│    └─ReLU: 2-8                              [16, 64, 16, 26]          --
├─Linear: 1-4                                 [16, 416, 256]            16,640
├─TransformerEncoder: 1-5                     [16, 417, 256]            --
│    └─ModuleList: 2-9                        --                        --
│    │    └─TransformerEncoderLayer: 3-1      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-2      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-3      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-4      [16, 417, 256]            527,104
├─Linear: 1-6                                 [16, 31]                  7,967
===========================================================================================
Total params: 2,152,287
Trainable params: 2,152,287
Non-trainable params: 0
Total mult-adds (M): 148.75
===========================================================================================
Input size (MB): 1.02
Forward/backward pass size (MB): 307.11
Params size (MB): 4.40
Estimated Total Size (MB): 312.53
===========================================================================================
```



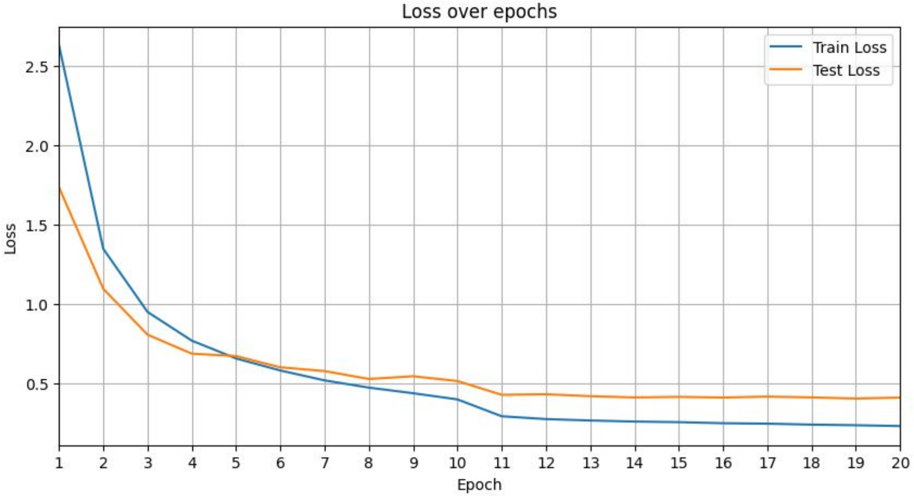
```
===========================================================================================
Layer (type:depth-idx)                        Output Shape              Param #
===========================================================================================
SpeechCommandTransformer                      [16, 12]                  256
├─MelSpectrogram: 1-1                         [16, 64, 101]             --
│    └─Spectrogram: 2-1                       [16, 201, 101]            --
│    └─MelScale: 2-2                          [16, 64, 101]             --
├─AmplitudeToDB: 1-2                          [16, 64, 101]             --
├─Sequential: 1-3                             [16, 64, 16, 26]          --
│    └─Conv2d: 2-3                            [16, 32, 32, 51]          320
│    └─BatchNorm2d: 2-4                       [16, 32, 32, 51]          64
│    └─ReLU: 2-5                              [16, 32, 32, 51]          --
│    └─Conv2d: 2-6                            [16, 64, 16, 26]          18,496
│    └─BatchNorm2d: 2-7                       [16, 64, 16, 26]          128
│    └─ReLU: 2-8                              [16, 64, 16, 26]          --
├─Linear: 1-4                                 [16, 416, 256]            16,640
├─TransformerEncoder: 1-5                     [16, 417, 256]            --
│    └─ModuleList: 2-9                        --                        --
│    │    └─TransformerEncoderLayer: 3-1      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-2      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-3      [16, 417, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-4      [16, 417, 256]            527,104
├─Linear: 1-6                                 [16, 12]                  3,084
===========================================================================================
Total params: 2,147,404
Trainable params: 2,147,404
Non-trainable params: 0
Total mult-adds (M): 148.68
===========================================================================================
Input size (MB): 1.02
Forward/backward pass size (MB): 307.10
Params size (MB): 4.38
Estimated Total Size (MB): 312.51
===========================================================================================
```

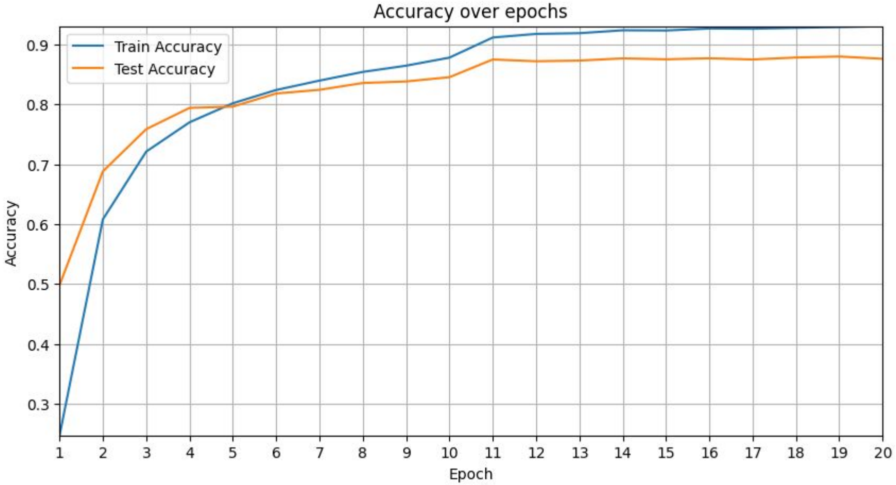
Transformer - original classes

no.	optimizer	learning rate	epochs	train acc	test acc	train time	
1	AdamW	0.000005	10	87.62	83.88	4h 25m	
2	Adam	0.001	9	3.44	3.53	4h 50m	
3	AdamW	0.000005	20	93.04	87.60	14h 17m	🏆

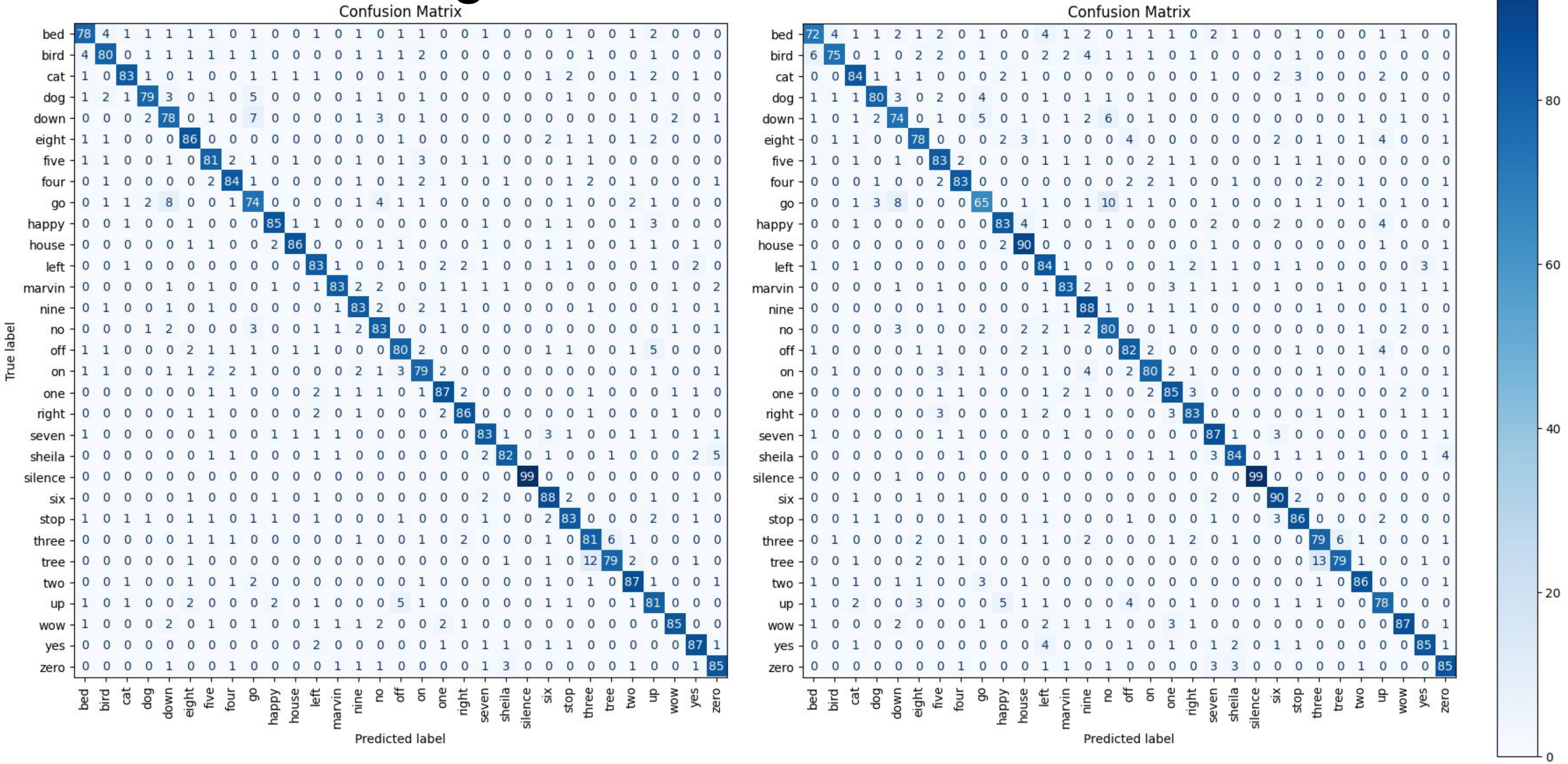
All experiments were conducted with stride = 2, batch size = 16, embedding dimension = 256, StepLR scheduler (step size = 10, gamma = 0.1), no position embedding and weight decay = 0.001 for AdamW optimizer.

Transformer - original classes - best model





Transformer - original classes - best model



	bed	bird	cat	dog	down	eight	five	four	go	happy	house	left	marvin	nine	no	off	on	one	right	seven	sheila	silence	six	stop	three	tree	two	up	wow	yes	zero
bed	78	4	1	1	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1	0	0	0	1	0	0	1	2	0	0	0
bird	4	80	0	1	1	1	1	1	1	0	0	0	0	1	1	1	2	0	0	0	0	0	0	0	1	0	0	1	0	0	0
cat	1	0	83	1	0	1	0	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	2	0	0	1	2	0	1	0
dog	1	2	1	79	3	0	1	0	5	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
down	0	0	0	2	78	0	1	0	7	0	0	0	0	1	3	0	1	0	0	0	0	0	0	0	0	0	1	0	2	0	1
eight	1	1	0	0	0	86	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	1	0	1	2	0	0	0
five	1	1	0	0	1	0	81	2	1	0	1	0	0	1	0	1	3	0	1	1	0	0	0	1	1	0	0	0	0	0	0
four	0	1	0	0	0	0	2	84	1	0	0	0	0	1	0	1	2	1	0	0	1	0	0	1	2	0	1	0	0	0	1
go	0	1	1	2	8	0	0	1	74	0	0	0	0	1	4	1	1	0	0	0	0	0	0	1	0	0	2	1	0	0	0
happy	0	0	1	0	0	1	0	0	0	85	1	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	3	0	0	0
house	0	0	0	0	0	1	1	0	0	2	86	0	0	0	1	1	0	0	0	1	0	0	1	1	0	0	1	1	0	1	0
left	0	0	1	0	0	0	0	0	0	0	0	83	1	0	0	1	0	2	2	1	0	0	1	1	0	0	0	1	0	2	0
marvin	0	0	0	0	1	0	1	0	0	1	0	1	83	2	2	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	2
nine	0	1	0	0	1	0	1	0	0	0	0	0	1	83	2	0	2	1	1	0	0	0	0	0	1	0	0	0	1	0	1
no	0	0	0	1	2	0	0	0	3	0	0	1	1	2	83	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1
off	1	1	0	0	0	2	1	1	1	0	1	1	0	0	0	80	2	0	0	0	0	0	1	1	0	0	1	5	0	0	0
on	1	1	0	0	1	1	2	2	1	0	0	0	0	2	1	3	79	2	0	0	0	0	0	0	0	0	0	1	0	0	1
one	0	0	0	0	0	0	1	1	0	0	0	2	1	1	1	0	1	87	2	0	0	0	0	0	1	0	0	0	1	1	0
right	0	0	0	0	0	1	1	0	0	0	0	2	0	1	0	0	0	2	86	0	0	0	0	0	1	0	0	0	1	0	0
seven	1	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	83	1	0	3	1	0	0	1	1	0	1	1
sheila	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	2	82	0	1	0	0	1	0	0	0	2	5
silence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0
six	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	88	2	0	0	0	1	0	1	0
stop	1	0	1	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	0	1	0	0	2	83	0	0	0	2	0	1	0
three	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	1	0	2	0	0	0	1	0	81	6	1	0	0	0	0
tree	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	12	79	2	0	0	1	0
two	0	0	1	0	0	1	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	87	1	0	0	1
up	1	0	1	0	0	2	0	0	0	2	0	1	0	0	0	5	1	0	0	0	0	0	1	1	0	0	1	81	0	0	0
wow	1	0	0	0	2	0	1	0	1	0	0	1	1	1	2	0	0	2	1	0	0	0	0	0	0	0	0	0	85	0	0
yes	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	0	87	1
zero	0	0	0	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0	0	1	3	0	0	0	0	0	1	0	0	1	85

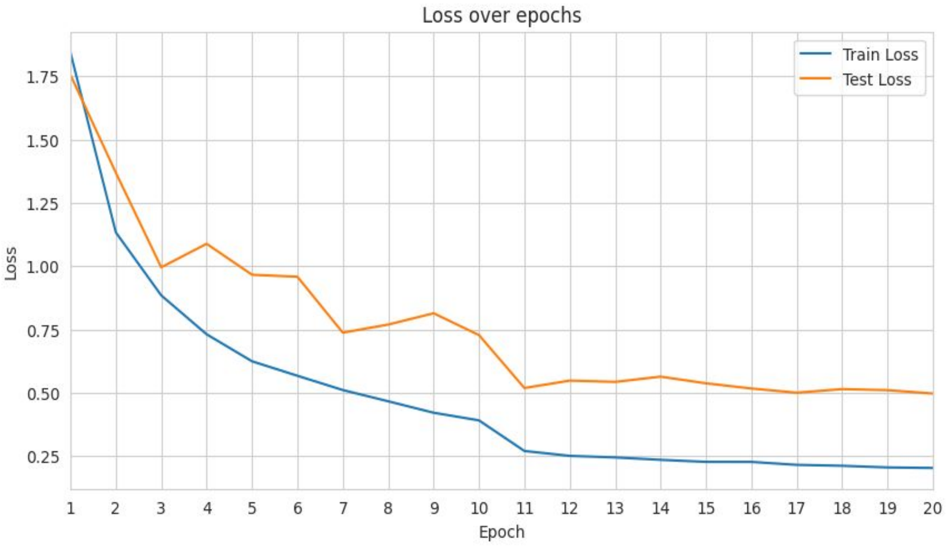
	bed	bird	cat	dog	down	eight	five	four	go	happy	house	left	marvin	nine	no	off	on	one	right	seven	sheila	silence	six	stop	three	tree	two	up	wow	yes	zero
bed	72	4	1	1	2	1	2	0	1	0	0	4	1	2	0	1	1	1	0	2	1	0	0	1	0	0	0	1	1	0	0
bird	6	75	0	1	0	2	2	0	1	0	0	2	2	4	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0
cat	0	0	84	1	1	1	0	0	0	2	1	0	0	0	0	0	0	0	0	1	0	0	2	3	0	0	0	2	0	0	0
dog	1	1	1	80	3	0	2	0	4	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
down	1	0	1	2	74	0	1	0	5	0	1	0	1	2	6	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1
eight	0	1	1	0	0	78	0	0	0	2	3	1	0	0	0	4	0	0	0	0	0	0	2	0	1	0	1	4	0	0	1
five	1	0	1	0	1	0	83	2	0	0	0	1	1	1	0	0	2	1	1	0	0	0	1	1	1	0	0	0	0	0	0
four	0	0	0	1	0	0	2	83	0	0	0	0	0	0	0	2	2	1	0	0	1	0	0	0	2	0	1	0	0	0	1
go	0	0	1	3	8	0	0	0	65	0	1	1	0	1	10	1	1	0	0	1	0	0	0	1	1	0	1	0	1	0	1
happy	0	0	1	0	0	0	0	0	0	83	4	1	0	0	1	0	0	0	0	2	0	0	2	0	0	0	0	4	0	0	0
house	0	0	0	0	0	0	0	0	0	2	90	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
left	1	0	1	0	0	0	0	0	0	0	0	84	1	0	0	0	0	1	2	1	1	0	1	1	0	0	0	0	0	3	1
marvin	0	0	0	0	1	0	1	0	0	0	0	1	83	2	1	0	0	3	1	1	1	0	1	0	0	1	0	0	1	1	1
nine	0	0	0	0	0	0	1	0	0	0	0	1	1	88	1	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0
no	0	0	0	0	3	0	0	0	2	0	2	2	1	2	80	0	0	1	0	0	0	0	0	0	0	0	1	0	2	0	1
off	1	0	0	0	0	1	1	0	0	0	2	1	0	0	0	82	2	0	0	0	0	0	0	1	0	0	1	4	0	0	0
on	0	1	0	0	0	0	3	1	1	0	0	1	0	4	0	2	80	2	1	0	0	0	0	0	1	0	0	1	0	0	1
one	0	0	0	0	0	0	1	1	0	0	0	1	2	1	0	0	2	85	3	0	0	0	0	0	0	0	0	0	2	0	1
right	0	0	0	0	0	0	3	0	0	0	1	2	0	1	0	0	0	3	83	0	0	0	0	0	1	0	1	0	1	1	1
seven	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	87	1	0	3	0	0	0	0	0	0	1	1
sheila	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	3	84	0	1	1	1	0	1	0	0	1	4
silence	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0
six	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	2	0	0	90	2	0	0	0	0	0	0	0
stop	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	3	86	0	0	0	2	0	0	0
three	0	1	0	0	0	2	0	1	0	0	1	1	0	2	0	0	0	1	2	0	1	0	0	0	79	6	1	0	0	0	1
tree	0	0	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	79	1	0	0	1	0
two	1	0	1	0	1	1	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	86	0	0	0	1
up	1	0	2	0	0	3	0	0	0	5	1	1	0	0	0	4	0	0	1	0	0	0	1	1	1	0	0	78	0	0	0
wow	1	0	0	0	2	0	0	0	1	0	0	2	1	1	1	0	0	3	1	0	0	0	0	0	0	0	0	0	87	0	1
yes	0	0	1	0	0	0	0	0	0	0	0	4	0	0	0	1	0	1	0	1	2	0	1	1	0	0	0	0	0	85	1
zero	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	3	3	0	0	0	0	0	1	0	0	0	85

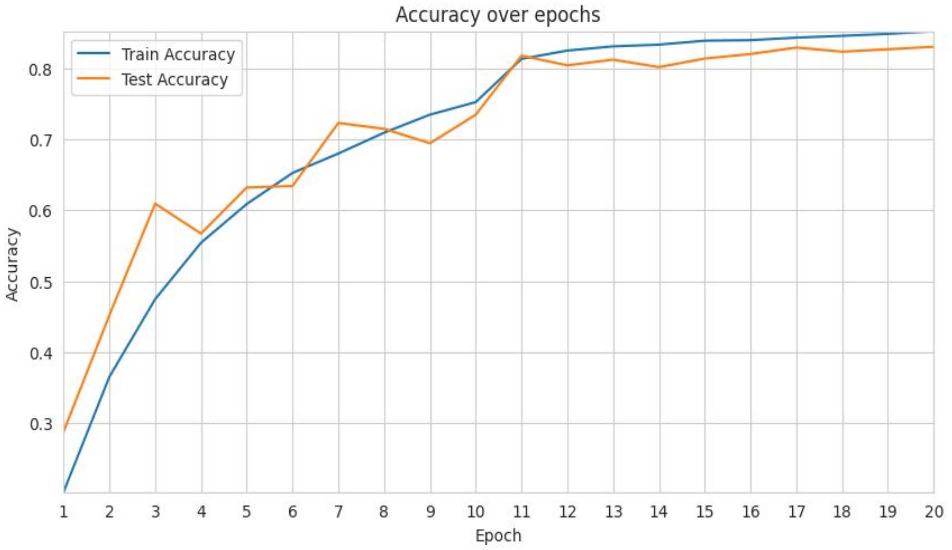
Transformer - modified classes

no.	optimizer	learning rate	epochs	scheduler	pos embed	train acc	test acc	train time
1	AdamW	0.000005	10	StepLR	no	69.88	69.67	4h 6m
2	Adam	0.001	20	StepLR	no	37.33	60.71	1h 30m
3	AdamW	0.000005	20	StepLR	no	85.23	83.04	4h 50m
4	AdamW	0.000005	20	-	no	84.75	77.84	5h 3m
5	AdamW	0.000005	20	StepLR	yes	82.24	80.91	6h 54m
6	AdamW	0.000005	20	CosineAnnealingLR	no	85.83	82.11	14h 16m

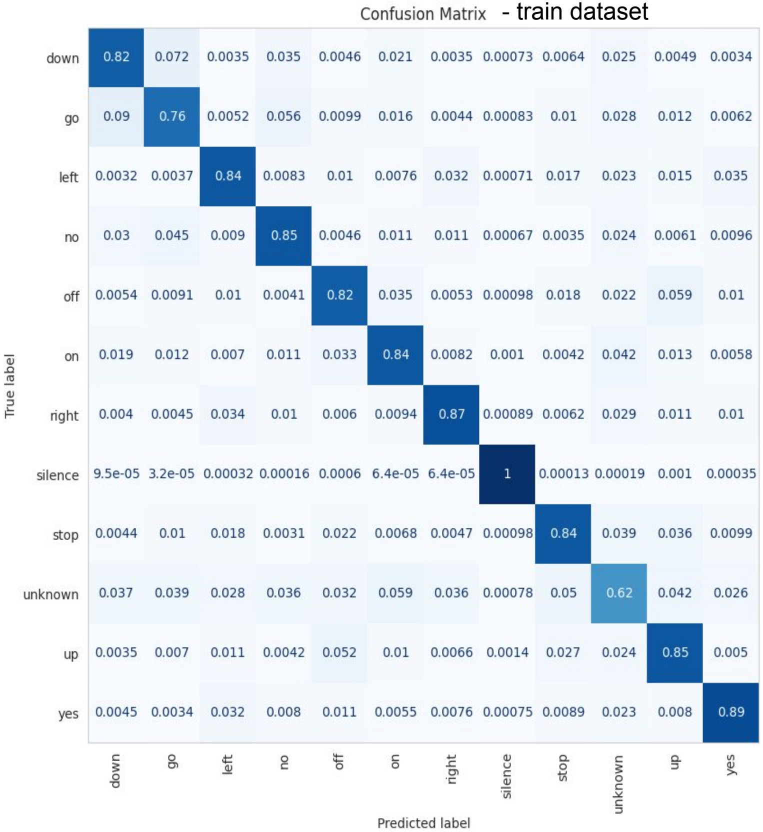
All experiments were conducted with stride = 2, batch size = 16, embedding dimension = 256, 20 epochs and weight decay = 0.001 for AdamW optimizer.

Transformer - modified classes - best model



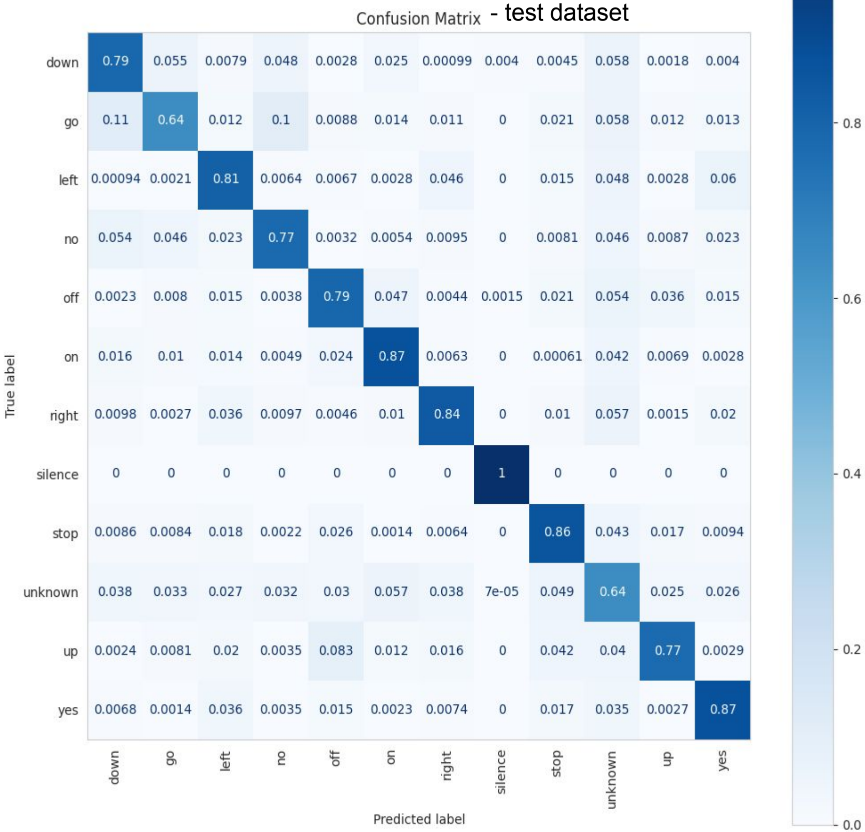


Transformer - modified classes - best model



Confusion Matrix - train dataset

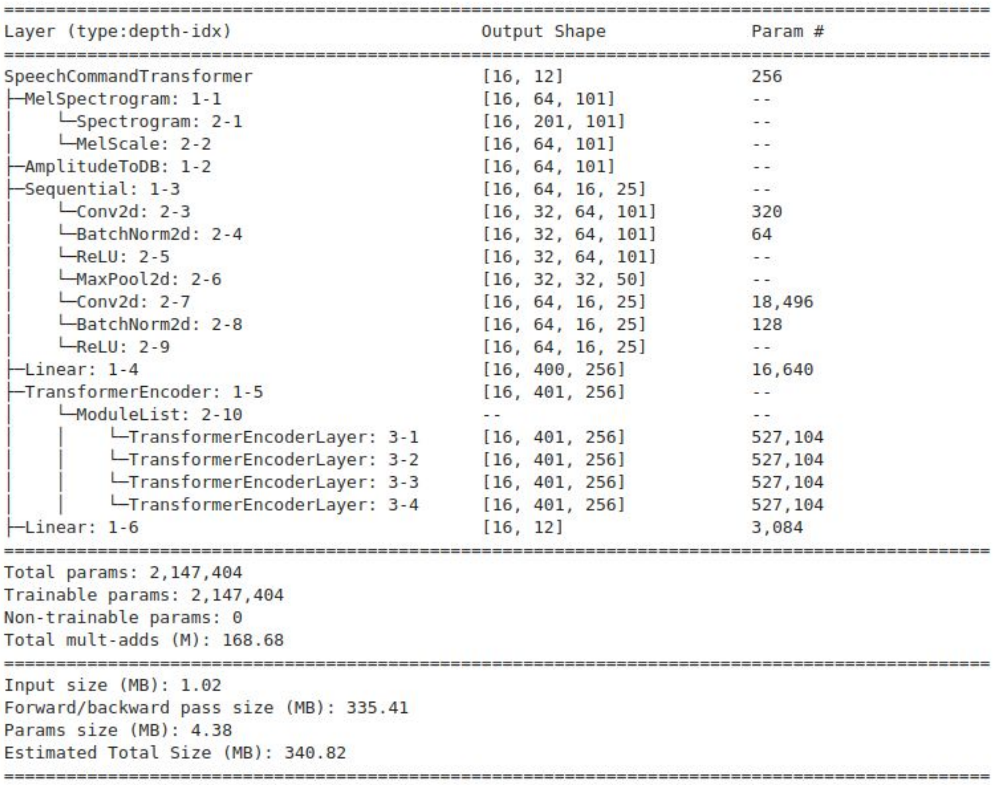
True label \ Predicted label	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.82	0.072	0.0035	0.035	0.0046	0.021	0.0035	0.00073	0.0064	0.025	0.0049	0.0034
go	0.09	0.76	0.0052	0.056	0.0099	0.016	0.0044	0.00083	0.01	0.028	0.012	0.0062
left	0.0032	0.0037	0.84	0.0083	0.01	0.0076	0.032	0.00071	0.017	0.023	0.015	0.035
no	0.03	0.045	0.009	0.85	0.0046	0.011	0.011	0.00067	0.0035	0.024	0.0061	0.0096
off	0.0054	0.0091	0.01	0.0041	0.82	0.035	0.0053	0.00098	0.018	0.022	0.059	0.01
on	0.019	0.012	0.007	0.011	0.033	0.84	0.0082	0.001	0.0042	0.042	0.013	0.0058
right	0.004	0.0045	0.034	0.01	0.006	0.0094	0.87	0.00089	0.0062	0.029	0.011	0.01
silence	9.5e-05	3.2e-05	0.00032	0.00016	0.0006	6.4e-05	6.4e-05	1	0.00013	0.00019	0.001	0.00035
stop	0.0044	0.01	0.018	0.0031	0.022	0.0068	0.0047	0.00098	0.84	0.039	0.036	0.0099
unknown	0.037	0.039	0.028	0.036	0.032	0.059	0.036	0.00078	0.05	0.62	0.042	0.026
up	0.0035	0.007	0.011	0.0042	0.052	0.01	0.0066	0.0014	0.027	0.024	0.85	0.005
yes	0.0045	0.0034	0.032	0.008	0.011	0.0055	0.0076	0.00075	0.0089	0.023	0.008	0.89



Confusion Matrix - test dataset

True label \ Predicted label	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.79	0.055	0.0079	0.048	0.0028	0.025	0.00099	0.004	0.0045	0.058	0.0018	0.004
go	0.11	0.64	0.012	0.1	0.0088	0.014	0.011	0	0.021	0.058	0.012	0.013
left	0.00094	0.0021	0.81	0.0064	0.0067	0.0028	0.046	0	0.015	0.048	0.0028	0.06
no	0.054	0.046	0.023	0.77	0.0032	0.0054	0.0095	0	0.0081	0.046	0.0087	0.023
off	0.0023	0.008	0.015	0.0038	0.79	0.047	0.0044	0.0015	0.021	0.054	0.036	0.015
on	0.016	0.01	0.014	0.0049	0.024	0.87	0.0063	0	0.00061	0.042	0.0069	0.0028
right	0.0098	0.0027	0.036	0.0097	0.0046	0.01	0.84	0	0.01	0.057	0.0015	0.02
silence	0	0	0	0	0	0	0	1	0	0	0	0
stop	0.0086	0.0084	0.018	0.0022	0.026	0.0014	0.0064	0	0.86	0.043	0.017	0.0094
unknown	0.038	0.033	0.027	0.032	0.03	0.057	0.038	7e-05	0.049	0.64	0.025	0.026
up	0.0024	0.0081	0.02	0.0035	0.083	0.012	0.016	0	0.042	0.04	0.77	0.0029
yes	0.0068	0.0014	0.036	0.0035	0.015	0.0023	0.0074	0	0.017	0.035	0.0027	0.87

Transformer - different architecture

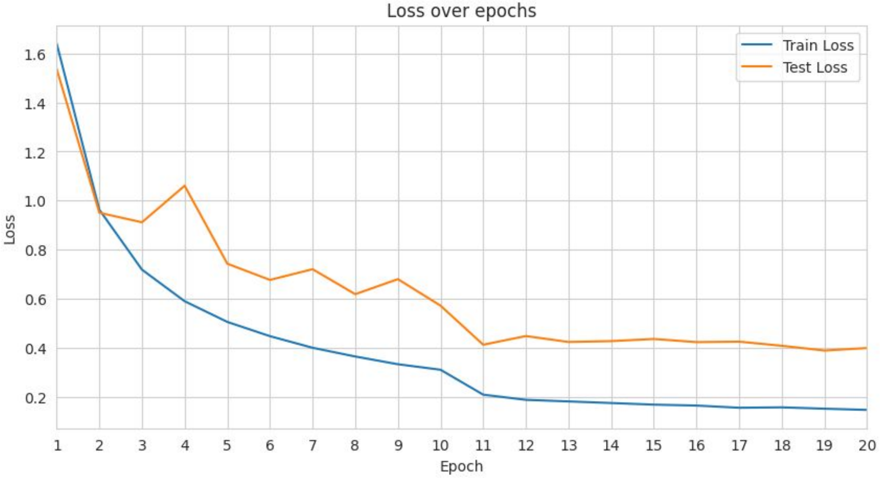


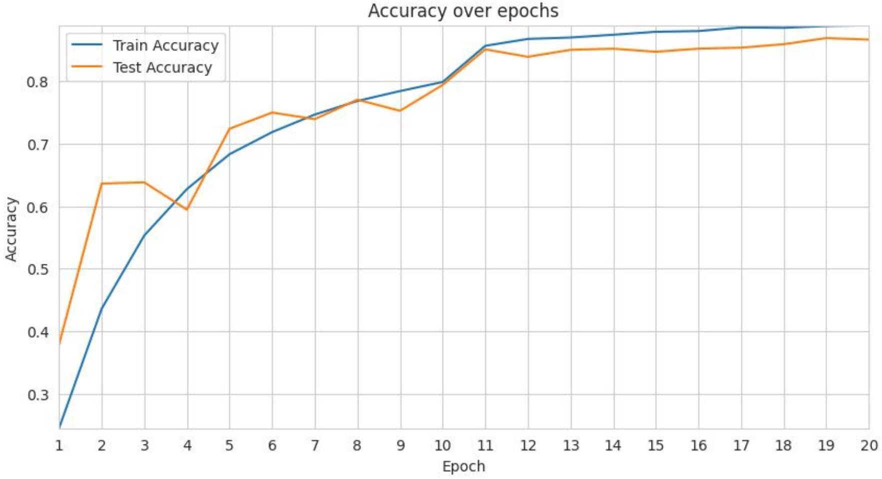
```
===========================================================================================================
Layer (type:depth-idx)                             Output Shape              Param #
===========================================================================================================
SpeechCommandTransformer                           [16, 12]                  256
├─MelSpectrogram: 1-1                              [16, 64, 101]             --
│    └─Spectrogram: 2-1                            [16, 201, 101]            --
│    └─MelScale: 2-2                               [16, 64, 101]             --
├─AmplitudeToDB: 1-2                               [16, 64, 101]             --
├─Sequential: 1-3                                  [16, 64, 16, 25]          --
│    └─Conv2d: 2-3                                 [16, 32, 64, 101]         320
│    └─BatchNorm2d: 2-4                            [16, 32, 64, 101]         64
│    └─ReLU: 2-5                                   [16, 32, 64, 101]         --
│    └─MaxPool2d: 2-6                              [16, 32, 32, 50]          --
│    └─Conv2d: 2-7                                 [16, 64, 16, 25]          18,496
│    └─BatchNorm2d: 2-8                            [16, 64, 16, 25]          128
│    └─ReLU: 2-9                                   [16, 64, 16, 25]          --
├─Linear: 1-4                                      [16, 400, 256]            16,640
├─TransformerEncoder: 1-5                          [16, 401, 256]            --
│    └─ModuleList: 2-10                            --                        --
│    │    └─TransformerEncoderLayer: 3-1           [16, 401, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-2           [16, 401, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-3           [16, 401, 256]            527,104
│    │    └─TransformerEncoderLayer: 3-4           [16, 401, 256]            527,104
├─Linear: 1-6                                      [16, 12]                  3,084
===========================================================================================================
Total params: 2,147,404
Trainable params: 2,147,404
Non-trainable params: 0
Total mult-adds (M): 168.68
===========================================================================================================
Input size (MB): 1.02
Forward/backward pass size (MB): 335.41
Params size (MB): 4.38
Estimated Total Size (MB): 340.82
===========================================================================================================
```

Best test accuracy:
86.67%

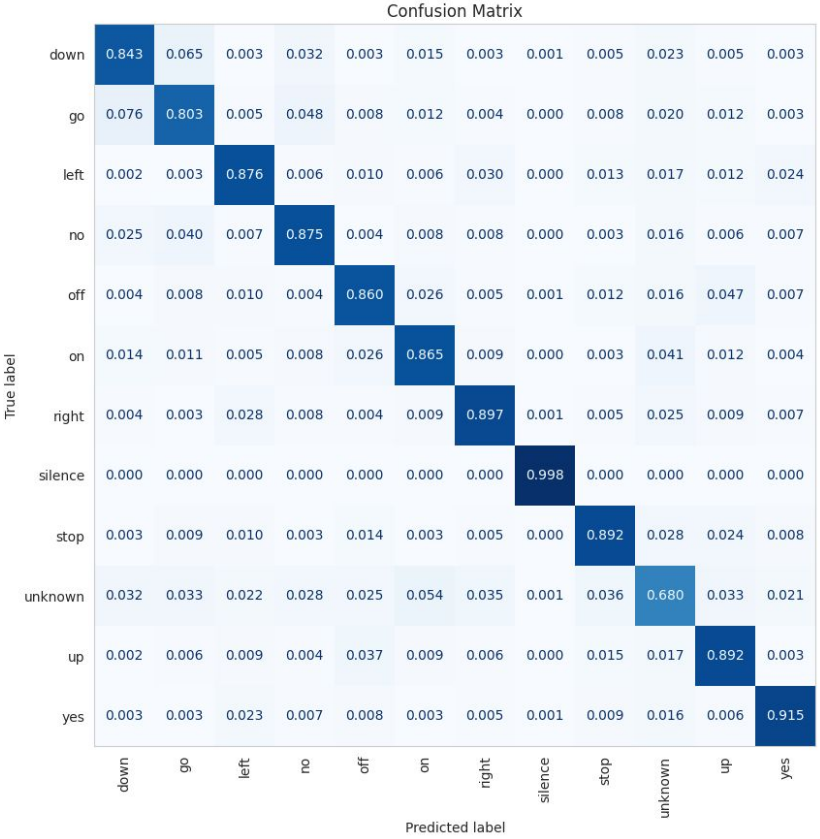
Corresponding train accuracy:
88.98%

Transformer - different architecture - even better results


Loss over epochs

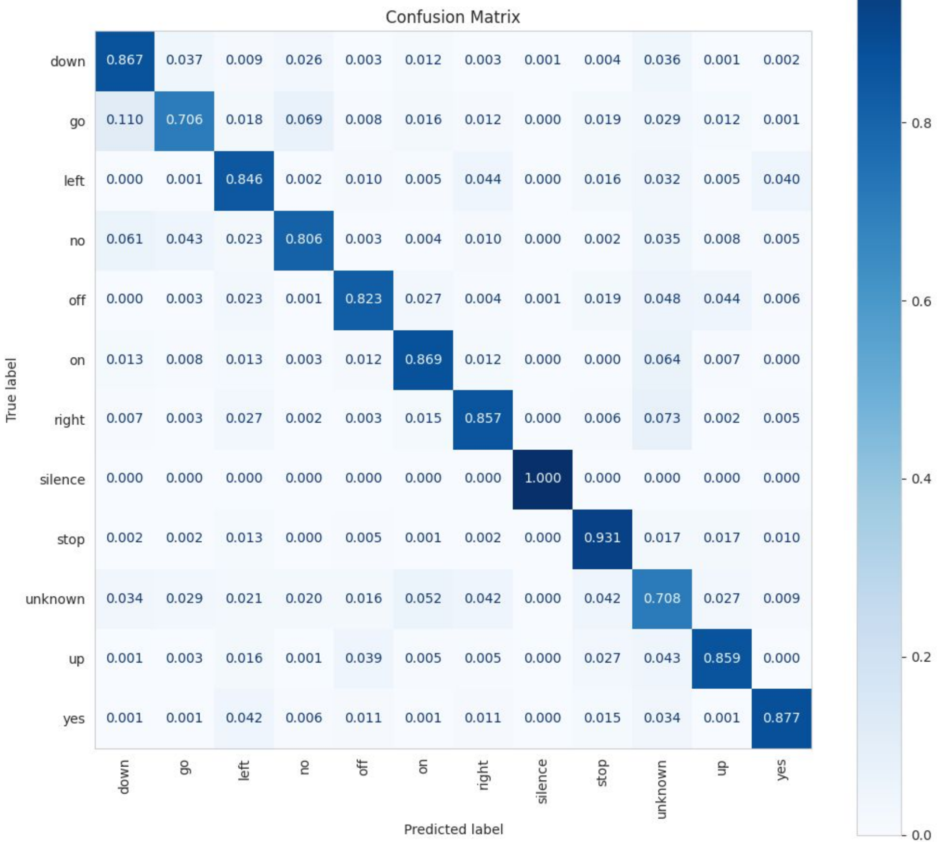

Accuracy over epochs

Transformer - different architecture - even better results



Confusion Matrix

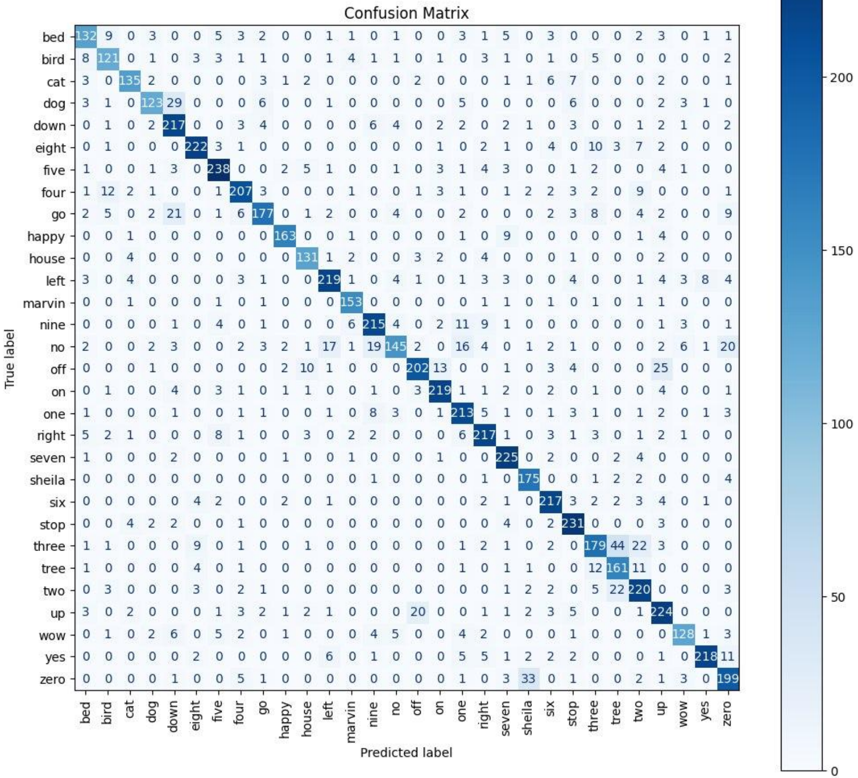
	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.843	0.065	0.003	0.032	0.003	0.015	0.003	0.001	0.005	0.023	0.005	0.003
go	0.076	0.803	0.005	0.048	0.008	0.012	0.004	0.000	0.008	0.020	0.012	0.003
left	0.002	0.003	0.876	0.006	0.010	0.006	0.030	0.000	0.013	0.017	0.012	0.024
no	0.025	0.040	0.007	0.875	0.004	0.008	0.008	0.000	0.003	0.016	0.006	0.007
off	0.004	0.008	0.010	0.004	0.860	0.026	0.005	0.001	0.012	0.016	0.047	0.007
on	0.014	0.011	0.005	0.008	0.026	0.865	0.009	0.000	0.003	0.041	0.012	0.004
right	0.004	0.003	0.028	0.008	0.004	0.009	0.897	0.001	0.005	0.025	0.009	0.007
silence	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.998	0.000	0.000	0.000	0.000
stop	0.003	0.009	0.010	0.003	0.014	0.003	0.005	0.000	0.892	0.028	0.024	0.008
unknown	0.032	0.033	0.022	0.028	0.025	0.054	0.035	0.001	0.036	0.680	0.033	0.021
up	0.002	0.006	0.009	0.004	0.037	0.009	0.006	0.000	0.015	0.017	0.892	0.003
yes	0.003	0.003	0.023	0.007	0.008	0.003	0.005	0.001	0.009	0.016	0.006	0.915



Confusion Matrix

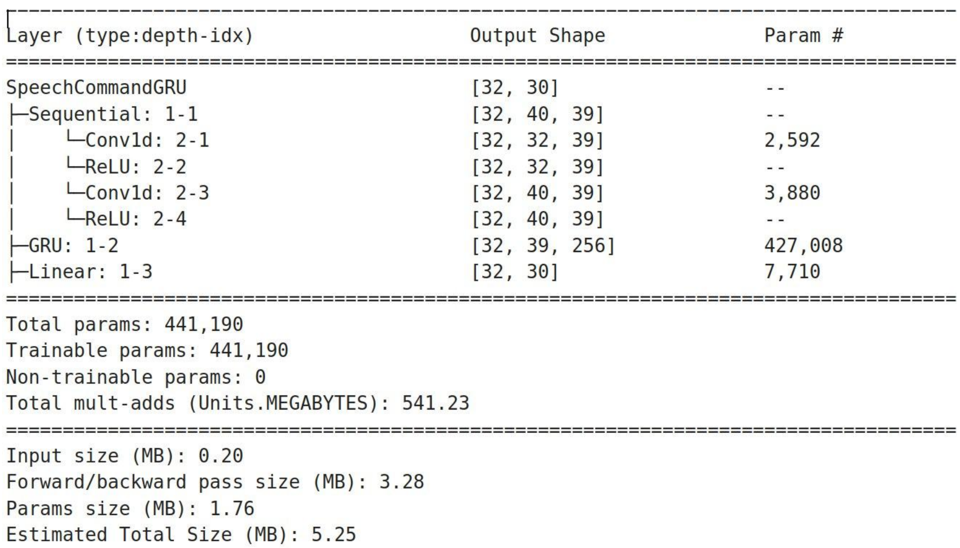
	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.867	0.037	0.009	0.026	0.003	0.012	0.003	0.001	0.004	0.036	0.001	0.002
go	0.110	0.706	0.018	0.069	0.008	0.016	0.012	0.000	0.019	0.029	0.012	0.001
left	0.000	0.001	0.846	0.002	0.010	0.005	0.044	0.000	0.016	0.032	0.005	0.040
no	0.061	0.043	0.023	0.806	0.003	0.004	0.010	0.000	0.002	0.035	0.008	0.005
off	0.000	0.003	0.023	0.001	0.823	0.027	0.004	0.001	0.019	0.048	0.044	0.006
on	0.013	0.008	0.013	0.003	0.012	0.869	0.012	0.000	0.000	0.064	0.007	0.000
right	0.007	0.003	0.027	0.002	0.003	0.015	0.857	0.000	0.006	0.073	0.002	0.005
silence	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
stop	0.002	0.002	0.013	0.000	0.005	0.001	0.002	0.000	0.931	0.017	0.017	0.010
unknown	0.034	0.029	0.021	0.020	0.016	0.052	0.042	0.000	0.042	0.708	0.027	0.009
up	0.001	0.003	0.016	0.001	0.039	0.005	0.005	0.000	0.027	0.043	0.859	0.000
yes	0.001	0.001	0.042	0.006	0.011	0.001	0.011	0.000	0.015	0.034	0.001	0.877

GRU - baseline model - 1D - original classes



Confusion Matrix

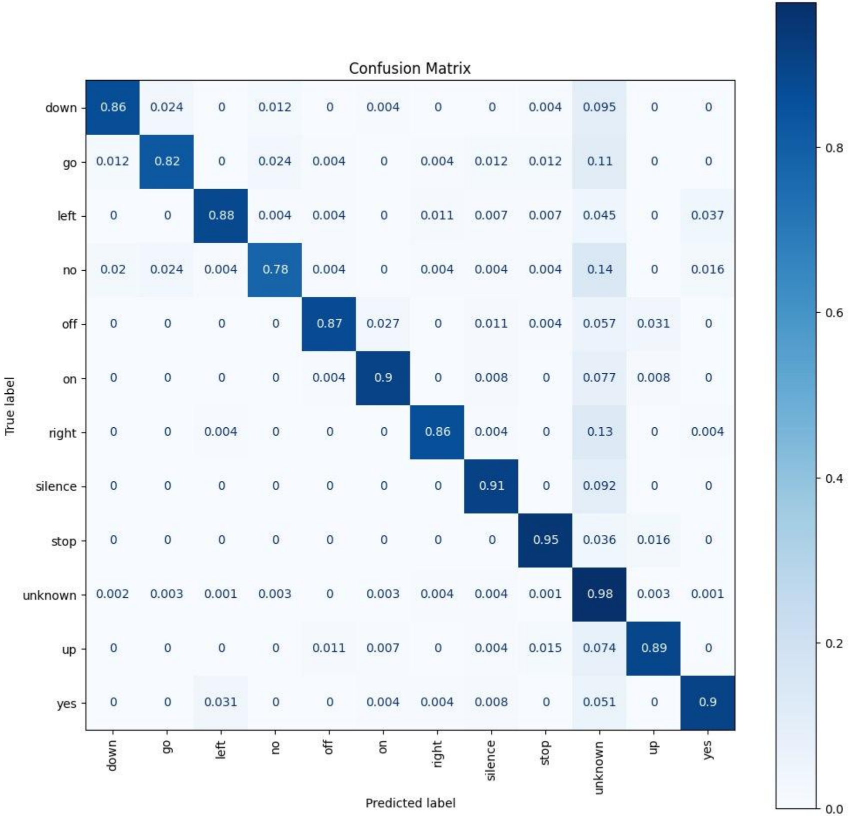
True label \ Predicted label	bed	bird	cat	dog	down	eight	five	four	go	happy	house	left	marvin	nine	no	off	on	one	right	seven	sheila	six	stop	three	tree	two	up	wow	yes	zero
bed	132	9	0	3	0	0	5	3	2	0	0	1	1	0	1	0	0	3	1	5	0	3	0	0	0	2	3	0	1	1
bird	8	121	0	1	0	3	3	1	1	0	0	1	4	1	1	0	1	0	3	1	0	1	0	5	0	0	0	0	0	2
cat	3	0	135	2	0	0	0	0	3	1	2	0	0	0	0	2	0	0	0	1	1	6	7	0	0	0	2	0	0	1
dog	3	1	0	123	29	0	0	0	6	0	0	1	0	0	0	0	0	5	0	0	0	0	6	0	0	0	2	3	1	0
down	0	1	0	2	217	0	0	3	4	0	0	0	0	6	4	0	2	2	0	2	1	0	3	0	0	1	2	1	0	2
eight	0	1	0	0	0	222	3	1	0	0	0	0	0	0	0	0	1	0	2	1	0	4	0	10	3	7	2	0	0	0
five	1	0	0	1	3	0	238	0	0	2	5	1	0	0	1	0	3	1	4	3	0	0	1	2	0	0	4	1	0	0
four	1	12	2	1	0	0	1	207	3	0	0	0	1	0	0	1	3	1	0	1	2	2	3	2	0	9	0	0	0	1
go	2	5	0	2	21	0	1	6	177	0	1	2	0	0	4	0	0	2	0	0	0	2	3	8	0	4	2	0	0	9
happy	0	0	1	0	0	0	0	0	0	163	0	0	1	0	0	0	0	1	0	9	0	0	0	0	0	1	4	0	0	0
house	0	0	4	0	0	0	0	0	0	0	131	1	2	0	0	3	2	0	4	0	0	0	1	0	0	0	2	0	0	0
left	3	0	4	0	0	0	0	3	1	0	0	219	1	0	4	1	0	1	3	3	0	0	4	0	0	1	4	3	8	4
marvin	0	0	1	0	0	0	1	0	1	0	0	0	153	0	0	0	0	0	1	1	0	1	0	1	0	1	1	0	0	0
nine	0	0	0	0	1	0	4	0	1	0	0	0	6	215	4	0	2	11	9	1	0	0	0	0	0	0	1	3	0	1
no	2	0	0	2	3	0	0	2	3	2	1	17	1	19	145	2	0	16	4	0	1	2	1	0	0	0	2	6	1	20
off	0	0	0	1	0	0	0	0	0	2	10	1	0	0	0	202	13	0	0	1	0	3	4	0	0	0	25	0	0	0
on	0	1	0	0	4	0	3	1	0	1	1	0	0	1	0	3	219	1	1	2	0	2	0	1	0	0	4	0	0	1
one	1	0	0	0	1	0	0	1	1	0	0	1	0	8	3	0	1	213	5	1	0	1	3	1	0	1	2	0	1	3
right	5	2	1	0	0	0	8	1	0	0	3	0	2	2	0	0	0	6	217	1	0	3	1	3	0	1	2	1	0	0
seven	1	0	0	0	2	0	0	0	0	1	0	0	1	0	0	0	1	0	0	225	0	2	0	0	2	4	0	0	0	0
sheila	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	175	0	0	1	2	2	0	0	0	4
six	0	0	0	0	0	4	2	0	0	2	0	1	0	0	0	0	0	0	2	1	0	217	3	2	2	3	4	0	1	0
stop	0	0	4	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4	0	2	231	0	0	0	3	0	0	0
three	1	1	0	0	0	9	0	1	0	0	1	0	0	0	0	0	0	1	2	1	0	2	0	179	44	22	3	0	0	0
tree	1	0	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	12	161	11	0	0	0	0
two	0	3	0	0	0	3	0	2	1	0	0	0	0	0	0	0	0	0	0	1	2	2	0	5	22	220	0	0	0	3
up	3	0	2	0	0	0	1	3	2	1	2	1	0	0	0	20	0	0	1	1	2	3	5	0	0	1	224	0	0	0
wow	0	1	0	2	6	0	5	2	0	1	0	0	0	4	5	0	0	4	2	0	0	0	1	0	0	0	0	128	1	3
yes	0	0	0	0	0	2	0	0	0	0	0	6	0	1	0	0	0	5	5	1	2	2	2	0	0	0	1	0	218	11
zero	0	0	0	0	1	0	0	5	1	0	0	0	0	0	0	0	0	1	0	3	33	0	1	0	0	2	1	3	0	199



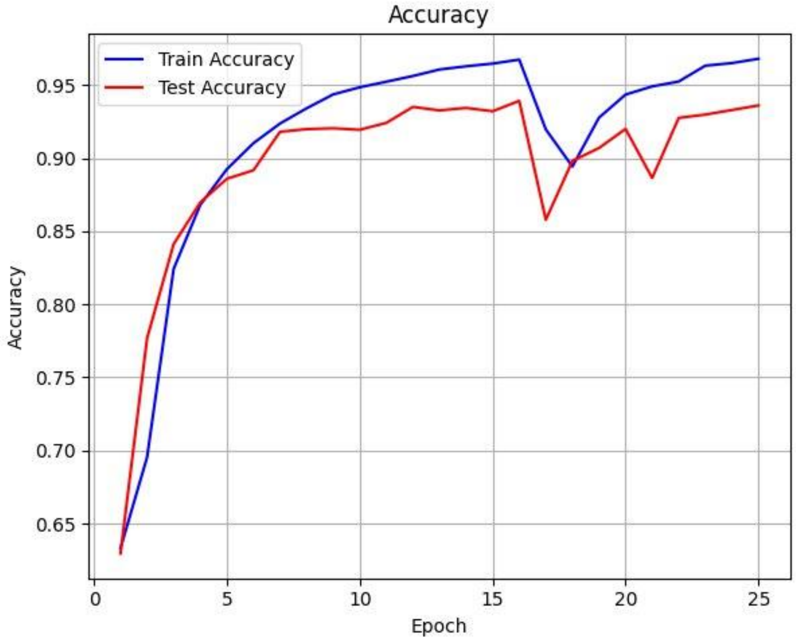
```
Layer (type:depth-idx)                  Output Shape              Param #
==========================================================================
SpeechCommandGRU                        [32, 30]                  --
├─Sequential: 1-1                       [32, 40, 39]              --
│    └─Conv1d: 2-1                      [32, 32, 39]              2,592
│    └─ReLU: 2-2                        [32, 32, 39]              --
│    └─Conv1d: 2-3                      [32, 40, 39]              3,880
│    └─ReLU: 2-4                        [32, 40, 39]              --
├─GRU: 1-2                              [32, 39, 256]             427,008
├─Linear: 1-3                           [32, 30]                  7,710
==========================================================================
Total params: 441,190
Trainable params: 441,190
Non-trainable params: 0
Total mult-adds (Units.MEGABYTES): 541.23
==========================================================================
Input size (MB): 0.20
Forward/backward pass size (MB): 3.28
Params size (MB): 1.76
Estimated Total Size (MB): 5.25
```

Test Accuracy: 82.31%

GRU - baseline model - 1D - modified classes

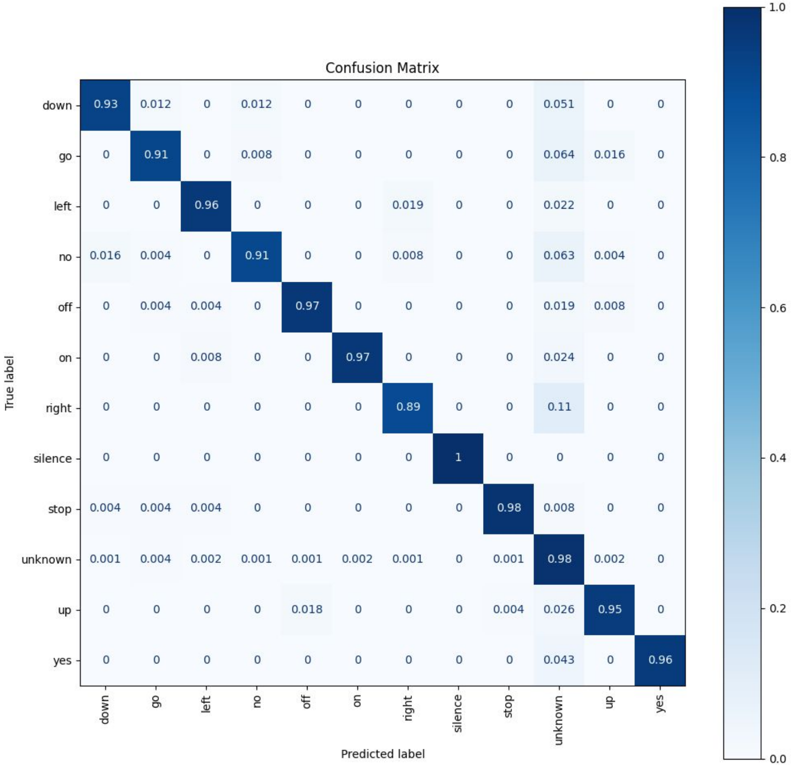


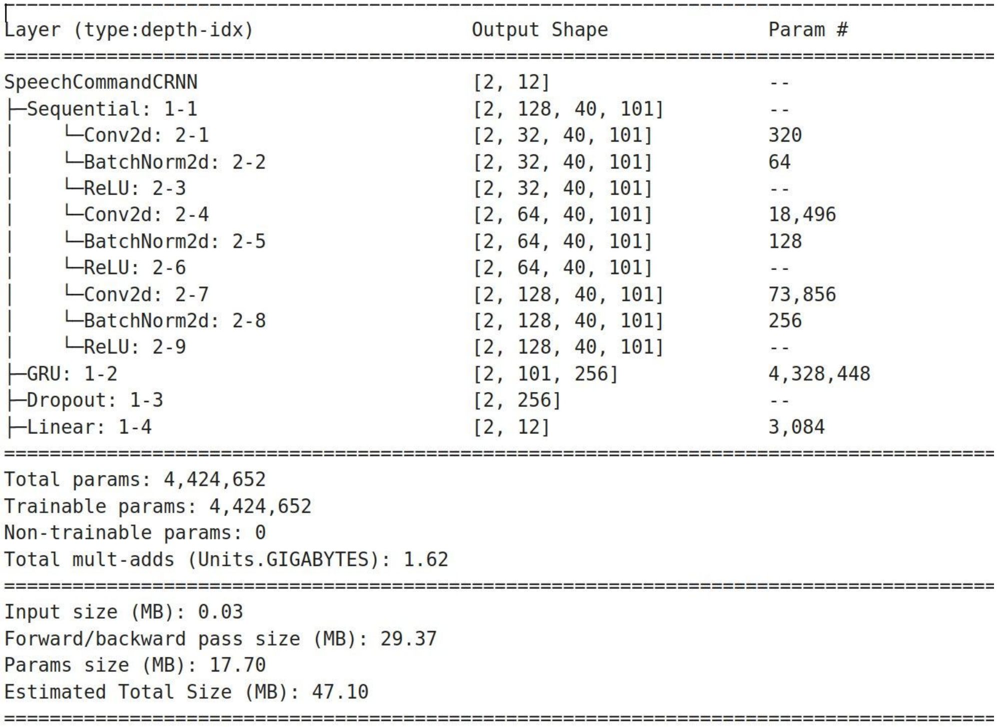
True label \ Predicted label	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.86	0.024	0	0.012	0	0.004	0	0	0.004	0.095	0	0
go	0.012	0.82	0	0.024	0.004	0	0.004	0.012	0.012	0.11	0	0
left	0	0	0.88	0.004	0.004	0	0.011	0.007	0.007	0.045	0	0.037
no	0.02	0.024	0.004	0.78	0.004	0	0.004	0.004	0.004	0.14	0	0.016
off	0	0	0	0	0.87	0.027	0	0.011	0.004	0.057	0.031	0
on	0	0	0	0	0.004	0.9	0	0.008	0	0.077	0.008	0
right	0	0	0.004	0	0	0	0.86	0.004	0	0.13	0	0.004
silence	0	0	0	0	0	0	0	0.91	0	0.092	0	0
stop	0	0	0	0	0	0	0	0	0.95	0.036	0.016	0
unknown	0.002	0.003	0.001	0.003	0	0.003	0.004	0.004	0.001	0.98	0.003	0.001
up	0	0	0	0	0.011	0.007	0	0.004	0.015	0.074	0.89	0
yes	0	0	0.031	0	0	0.004	0.004	0.008	0	0.051	0	0.9



Test Accuracy: 93.6%

GRU - 2D - modified classes

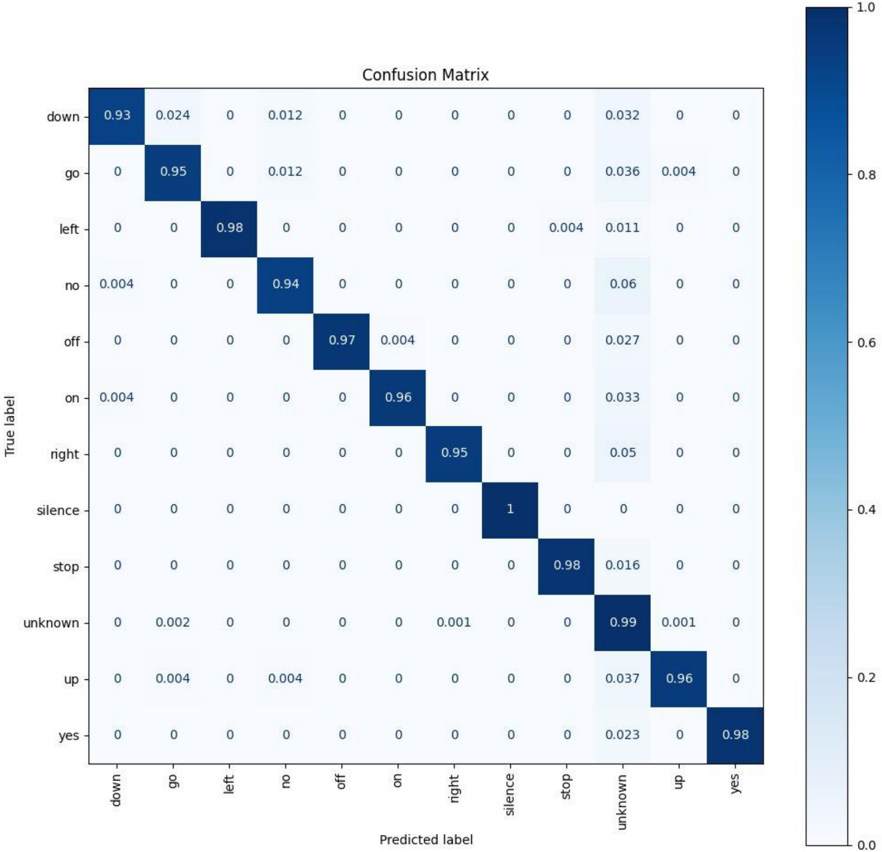




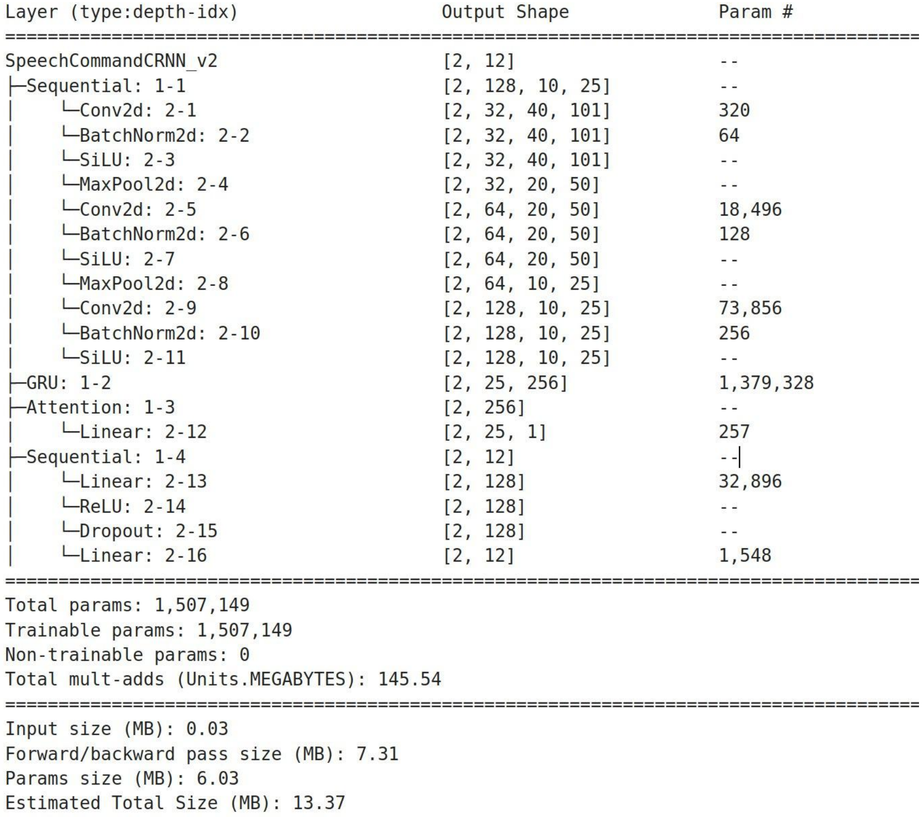
```
--------------------------------------------------------------------------------------
Layer (type:depth-idx)                    Output Shape              Param #
======================================================================================
SpeechCommandCRNN                         [2, 12]                   --
├─Sequential: 1-1                         [2, 128, 40, 101]         --
│    └─Conv2d: 2-1                        [2, 32, 40, 101]          320
│    └─BatchNorm2d: 2-2                   [2, 32, 40, 101]          64
│    └─ReLU: 2-3                          [2, 32, 40, 101]          --
│    └─Conv2d: 2-4                        [2, 64, 40, 101]          18,496
│    └─BatchNorm2d: 2-5                   [2, 64, 40, 101]          128
│    └─ReLU: 2-6                          [2, 64, 40, 101]          --
│    └─Conv2d: 2-7                        [2, 128, 40, 101]         73,856
│    └─BatchNorm2d: 2-8                   [2, 128, 40, 101]         256
│    └─ReLU: 2-9                          [2, 128, 40, 101]         --
├─GRU: 1-2                                [2, 101, 256]             4,328,448
├─Dropout: 1-3                            [2, 256]                  --
├─Linear: 1-4                             [2, 12]                   3,084
======================================================================================
Total params: 4,424,652
Trainable params: 4,424,652
Non-trainable params: 0
Total mult-adds (Units.GIGABYTES): 1.62
======================================================================================
Input size (MB): 0.03
Forward/backward pass size (MB): 29.37
Params size (MB): 17.70
Estimated Total Size (MB): 47.10
======================================================================================
```

Test Accuracy: 96.93%

Gru - modified classes - best model

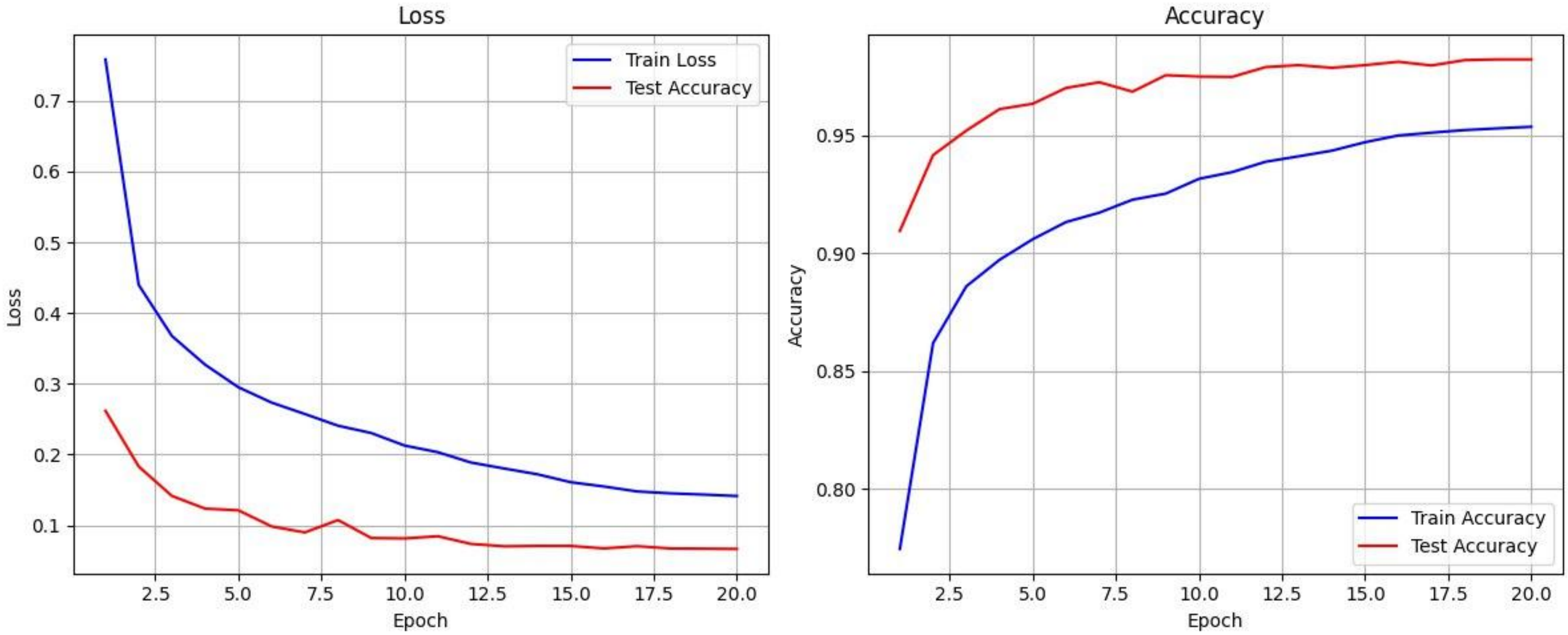


	down	go	left	no	off	on	right	silence	stop	unknown	up	yes
down	0.93	0.024	0	0.012	0	0	0	0	0	0.032	0	0
go	0	0.95	0	0.012	0	0	0	0	0	0.036	0.004	0
left	0	0	0.98	0	0	0	0	0	0.004	0.011	0	0
no	0.004	0	0	0.94	0	0	0	0	0	0.06	0	0
off	0	0	0	0	0.97	0.004	0	0	0	0.027	0	0
on	0.004	0	0	0	0	0.96	0	0	0	0.033	0	0
right	0	0	0	0	0	0	0.95	0	0	0.05	0	0
silence	0	0	0	0	0	0	0	1	0	0	0	0
stop	0	0	0	0	0	0	0	0	0.98	0.016	0	0
unknown	0	0.002	0	0	0	0	0.001	0	0	0.99	0.001	0
up	0	0.004	0	0.004	0	0	0	0	0	0.037	0.96	0
yes	0	0	0	0	0	0	0	0	0	0.023	0	0.98



```
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================================
SpeechCommandCRNN_v2                     [2, 12]                   --
├─Sequential: 1-1                        [2, 128, 10, 25]          --
│    └─Conv2d: 2-1                       [2, 32, 40, 101]          320
│    └─BatchNorm2d: 2-2                  [2, 32, 40, 101]          64
│    └─SiLU: 2-3                         [2, 32, 40, 101]          --
│    └─MaxPool2d: 2-4                    [2, 32, 20, 50]           --
│    └─Conv2d: 2-5                       [2, 64, 20, 50]           18,496
│    └─BatchNorm2d: 2-6                  [2, 64, 20, 50]           128
│    └─SiLU: 2-7                         [2, 64, 20, 50]           --
│    └─MaxPool2d: 2-8                    [2, 64, 10, 25]           --
│    └─Conv2d: 2-9                       [2, 128, 10, 25]          73,856
│    └─BatchNorm2d: 2-10                 [2, 128, 10, 25]          256
│    └─SiLU: 2-11                        [2, 128, 10, 25]          --
├─GRU: 1-2                               [2, 25, 256]              1,379,328
├─Attention: 1-3                         [2, 256]                  --
│    └─Linear: 2-12                      [2, 25, 1]                257
├─Sequential: 1-4                        [2, 12]                   --
│    └─Linear: 2-13                      [2, 128]                  32,896
│    └─ReLU: 2-14                        [2, 128]                  --
│    └─Dropout: 2-15                     [2, 128]                  --
│    └─Linear: 2-16                      [2, 12]                   1,548
==========================================================================================
Total params: 1,507,149
Trainable params: 1,507,149
Non-trainable params: 0
Total mult-adds (Units.MEGABYTES): 145.54
==========================================================================================
Input size (MB): 0.03
Forward/backward pass size (MB): 7.31
Params size (MB): 6.03
Estimated Total Size (MB): 13.37
```

Gru - modified classes - best model



Gru - modified classes - best model - over 10 runs

set	Accuracy		Loss	
	mean	**std**	**mean**	**std**
train	95.42%	0.09%	0.1352	0.0026
test	98.02%	0.08%	0.0721	0.0021
validation	**97.62%**	0.07%	0.0892	0.0046

Train time: under 4 minutes

Thank you for your "attention"

Are there any questions?

Sources

1. Attention is all you need
2. 1b3b transformer series
3. Andrej Karpathy transformers explained
4. Post on medium about speech recognition
5. Pytorch tutorial on speech commands recognition