SPEECH RECOGNITION

evaluating and improving network architectures for speech recognition task

Max Henning Höth

INTRODUCTION

- Speech recognition is an important part of research due to it’s fields of application

- Virtual assistance, Transcription, Customer Service, Automotive, Language learning

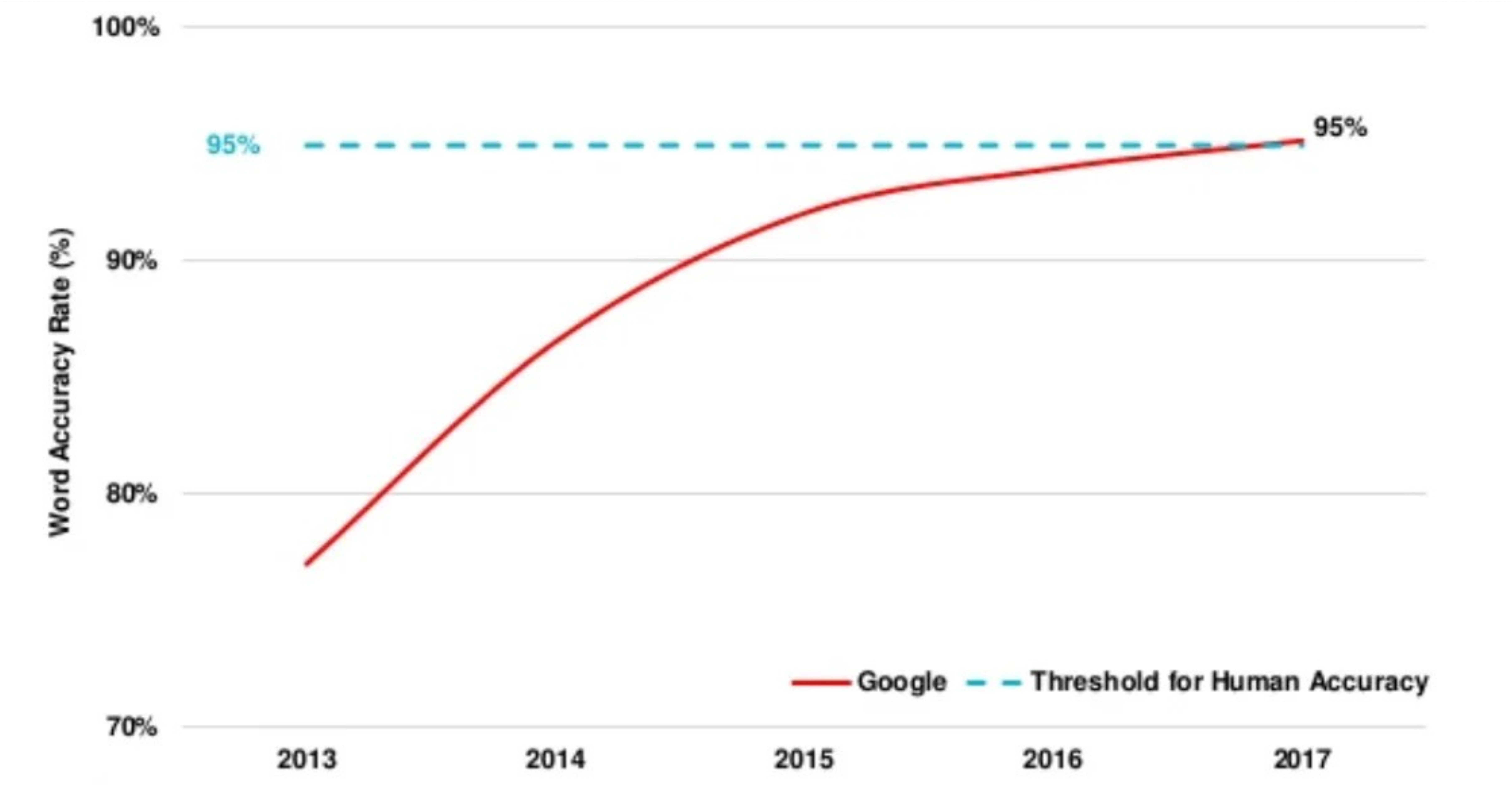
- In recent years the accuracy of such networks rose as high as the one of human perception

INTRODUCTION

- Speech t's fields of appl

- Virtual ve, Langua

- In rece s the one of hum



BACKGROUND

- Speech recognition tries to translate audio signals into words
- A lot of different techniques and architectures are possible
- Many kinds of problems could arise
- Different voice levels, accents, speech rate, background noise

DATASET

- Speech Command dataset from tensorflow provides around 100000 audio files which are 1 second long

- 35 different english words spoken by different people

- Common dataset for train and test networks for speech recognition task

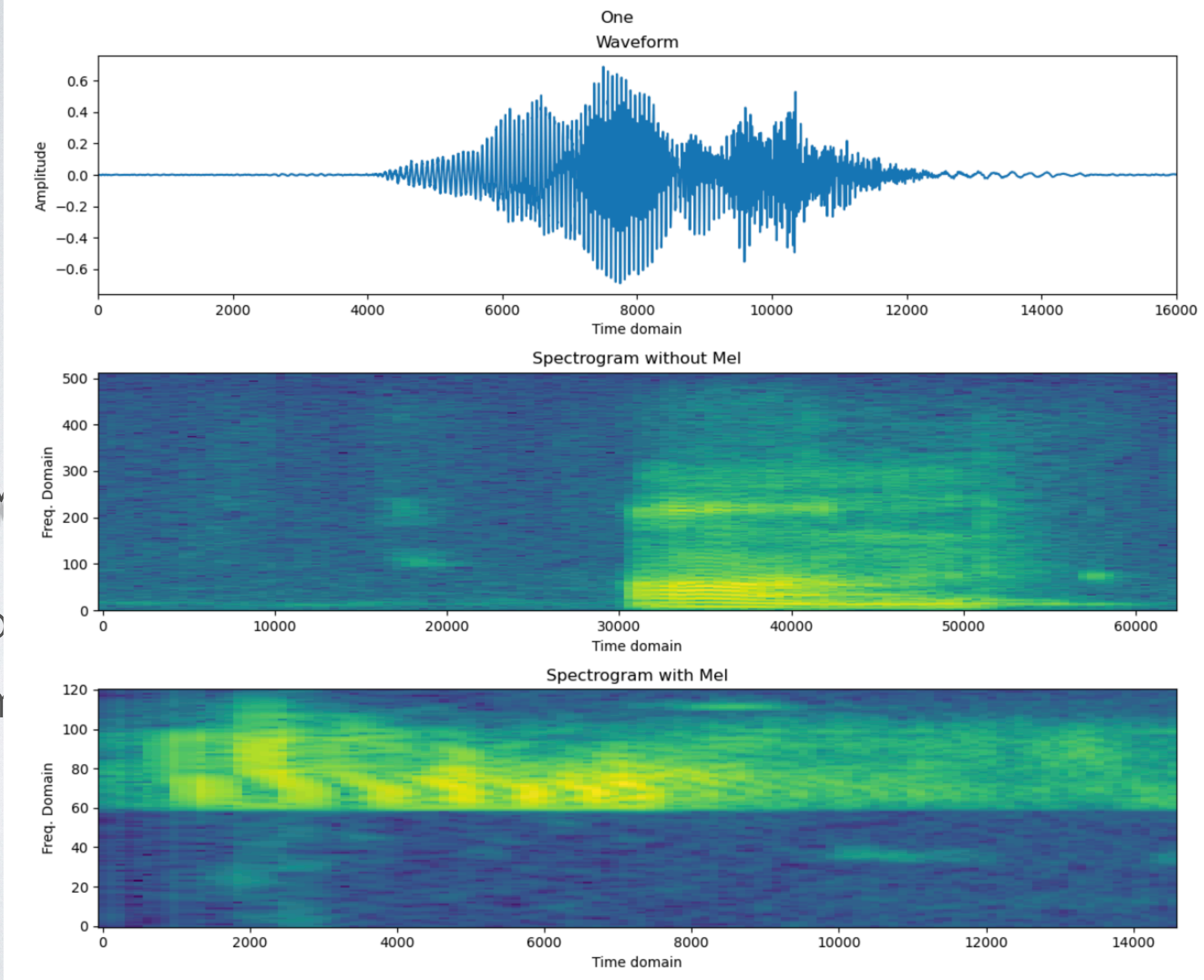
- Easily useable with the tensorflow API

PREPROCESSING

- STFT were applied to each audio file which led to spectrograms
- Each spectrogram was transformed into Mel scale for better performance

3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Mel scale
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
10000
Hertz scale

- STFT w rams
- Each sp
  perform



NETWORK ARCHITECTURES

- 2 x CNN/BatchNorm/MaxPool
- Followed by CNN/LSTM/BiDirectionalLSTM/ATT
- Different number of every layer and combination

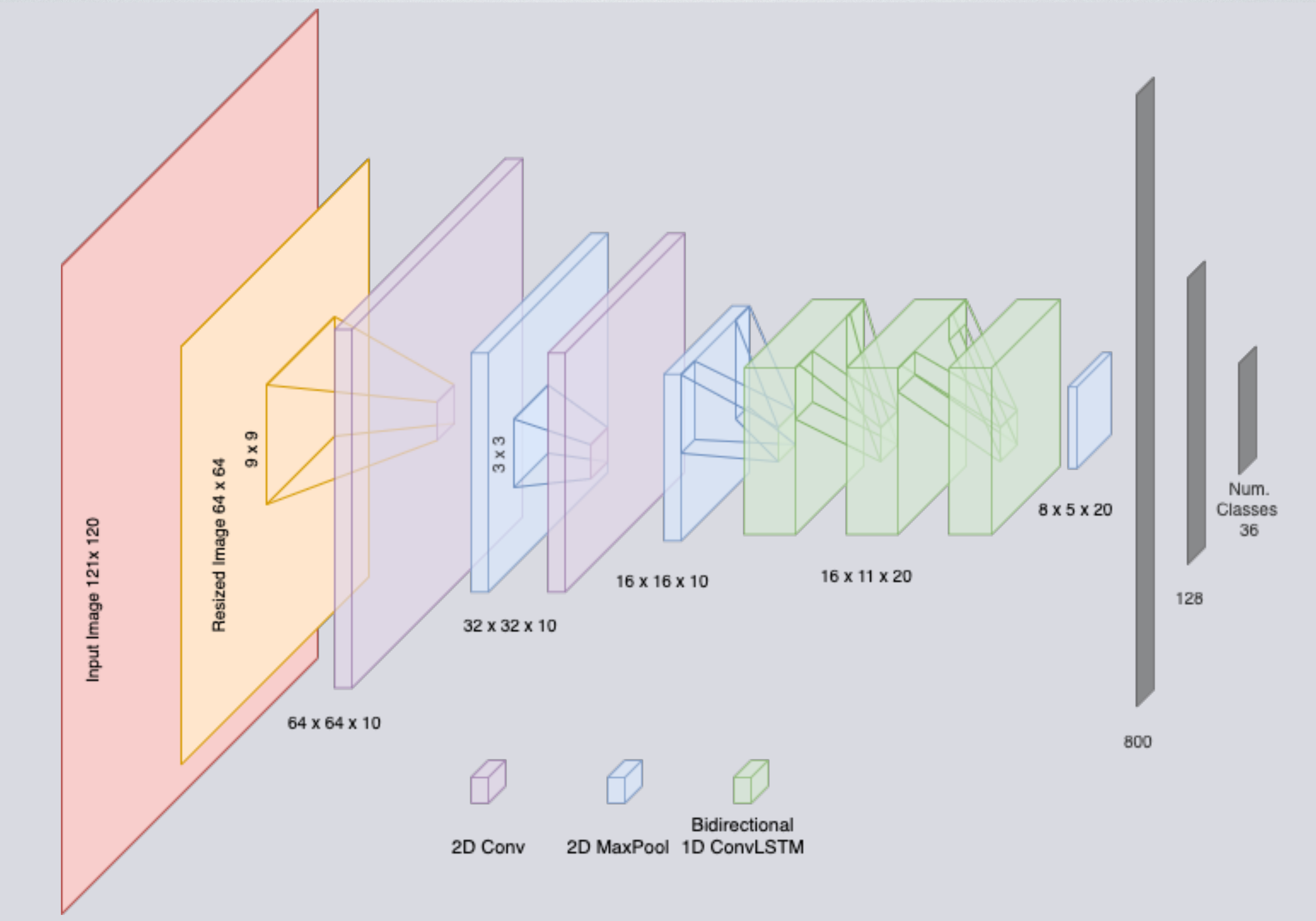
ATTENTION LAYER

- Adding weights to the inputs
- Indicates how important each part of the input is
- Amplifying or suppressing certain parts
- Also increases complexity and computation time

NETWORK ARCHITECTURES

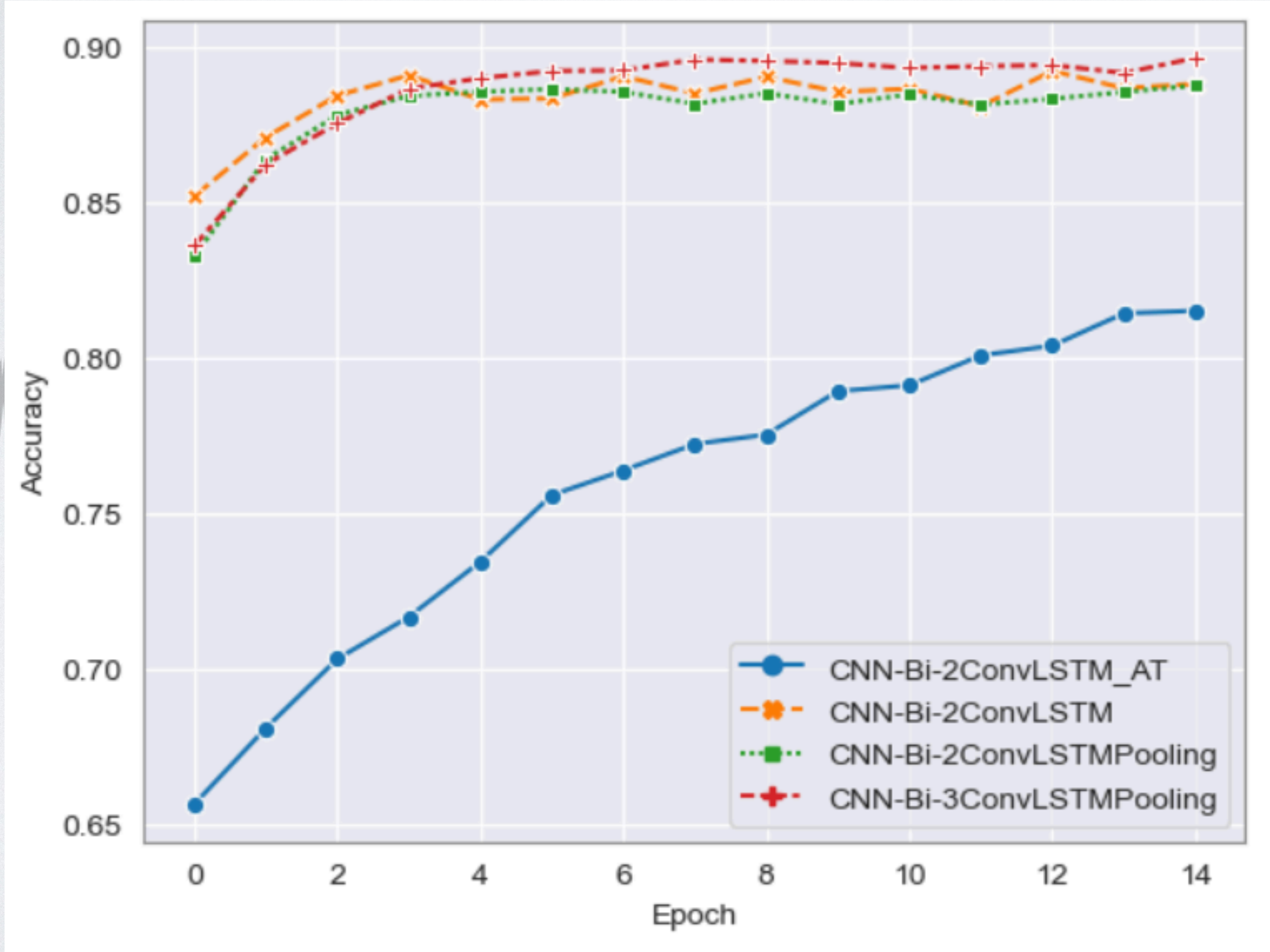
- CNN-Bi-2ConvLSTM_AT
- CNN-Bi-2ConvLSTM
- CNN-Bi-2ConvLSTMPooling
- CNN-Bi-3ConvLSTMPooling

- CN
- CN
- CN
- CN

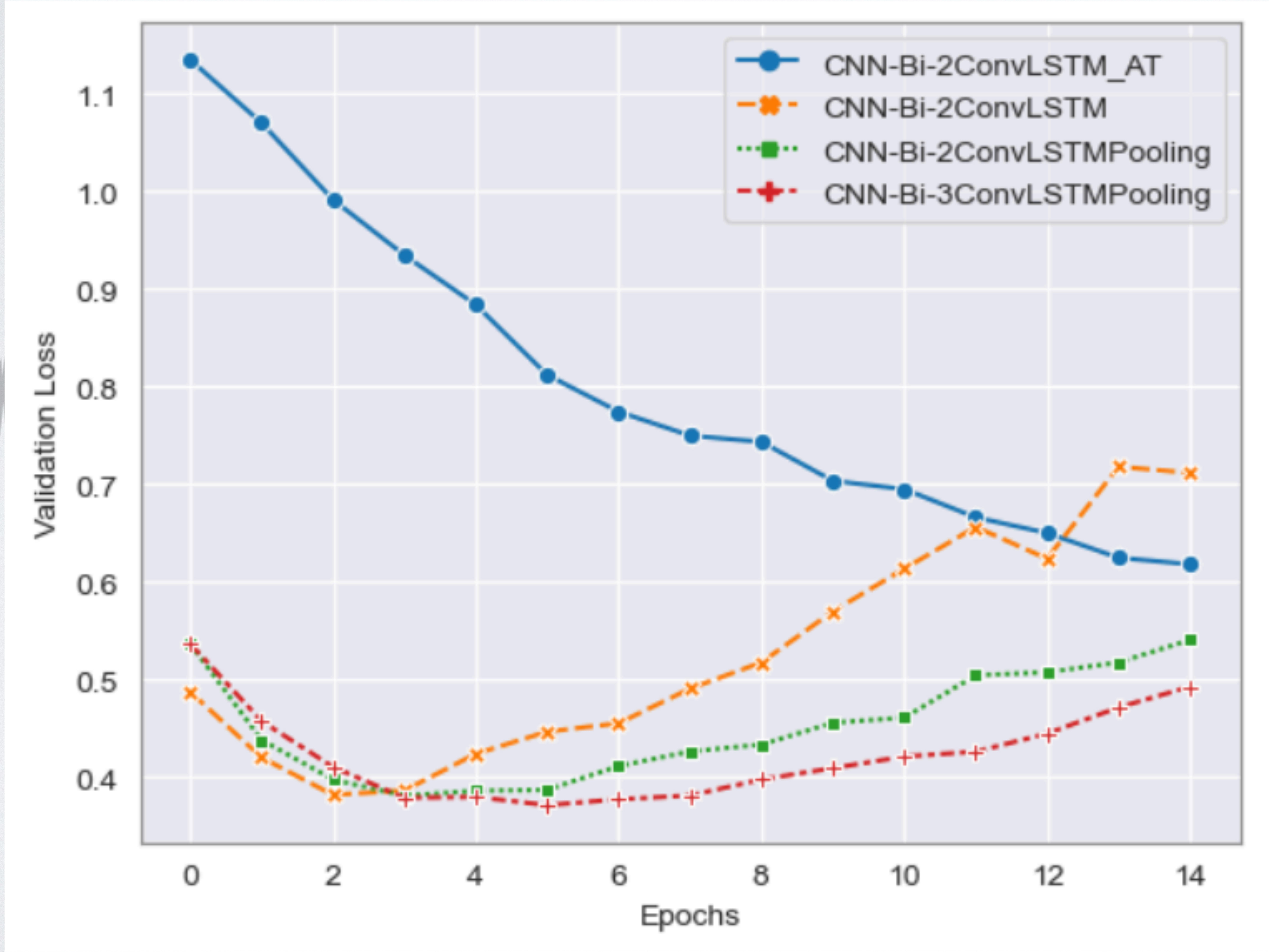

Input Image 121x 120
Resized Image 64 x 64
9 x 9
64 x 64 x 10
3 x 3
32 x 32 x 10
16 x 16 x 10
16 x 11 x 20
8 x 5 x 20
800
128
Num. Classes 36
2D Conv
2D MaxPool
Bidirectional 1D ConvLSTM

EVALUATION

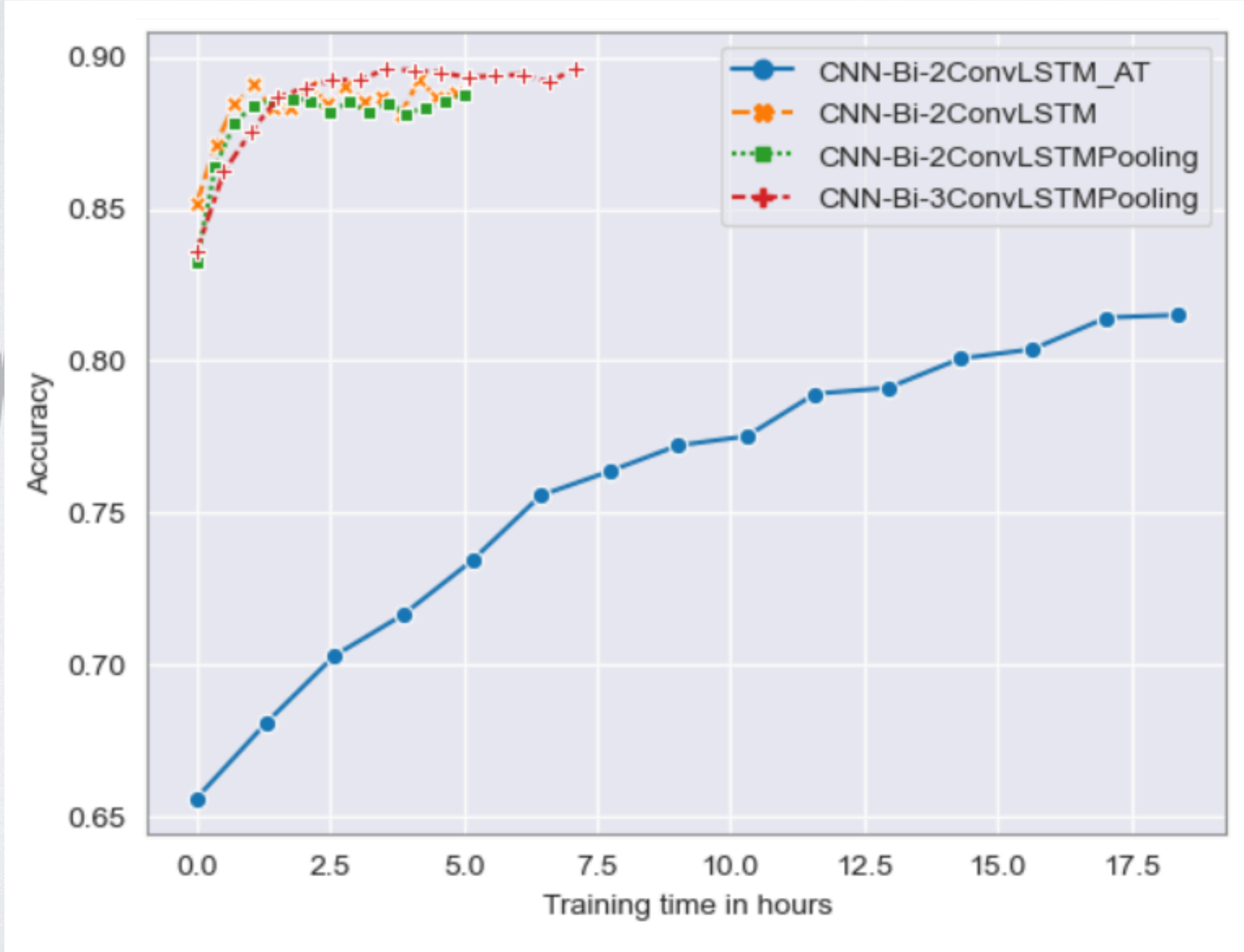
EVALUATION


0.90
0.85
0.80
0.75
0.70
0.65
Accuracy
0
2
4
6
8
10
12
14
Epoch
CNN-Bi-2ConvLSTM_AT
CNN-Bi-2ConvLSTM
CNN-Bi-2ConvLSTMPooling
CNN-Bi-3ConvLSTMPooling

EVALUATION

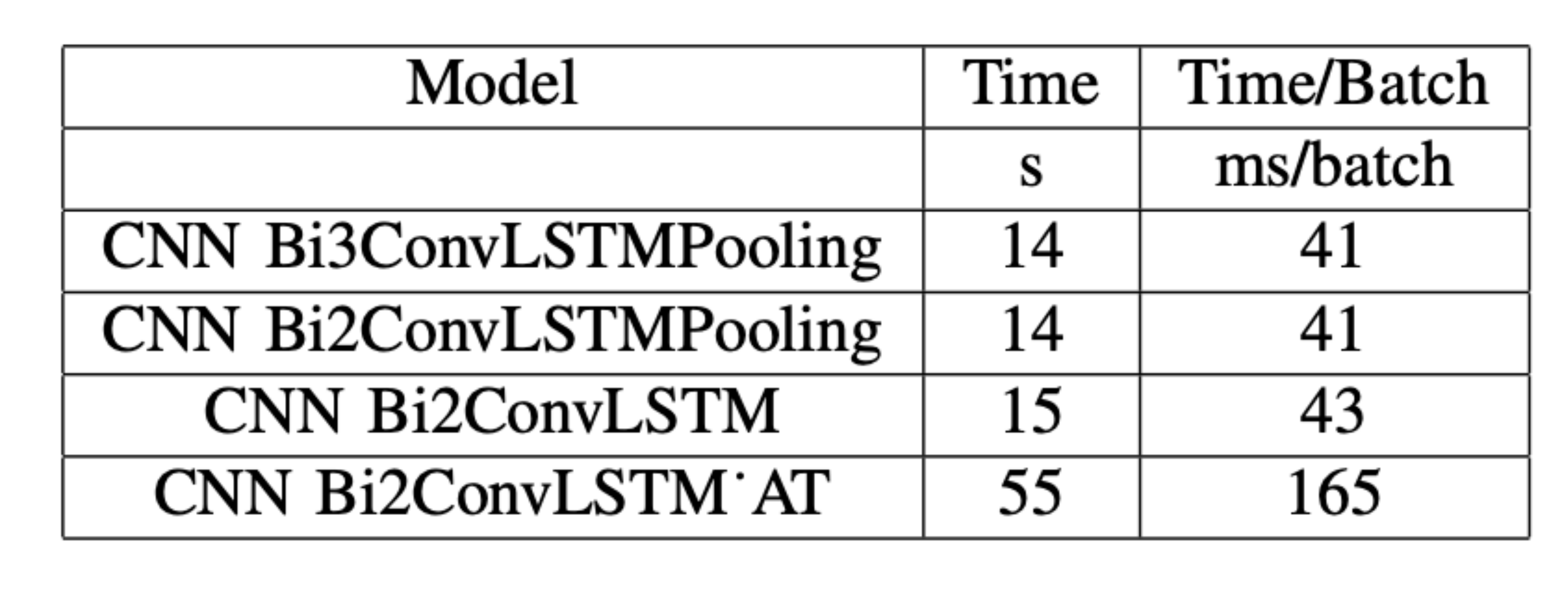

CNN-Bi-2ConvLSTM_AT
CNN-Bi-2ConvLSTM
CNN-Bi-2ConvLSTMPooling
CNN-Bi-3ConvLSTMPooling
Validation Loss
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
12
14
Epochs

EVALUATION

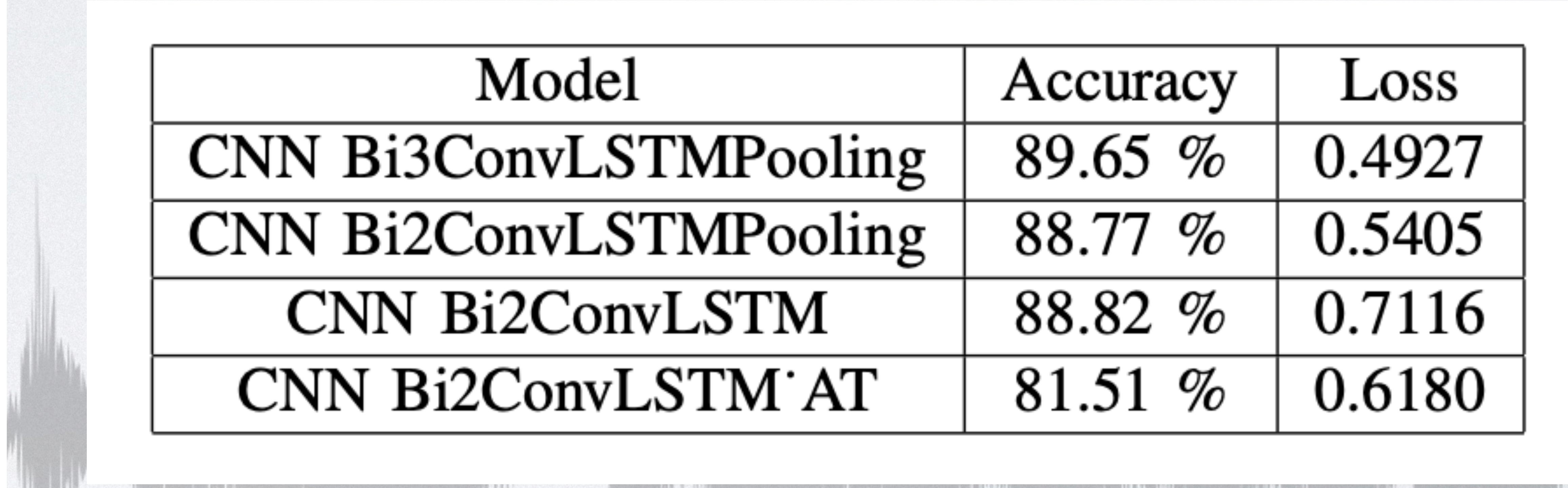


CNN-Bi-2ConvLSTM_AT
CNN-Bi-2ConvLSTM
CNN-Bi-2ConvLSTMPooling
CNN-Bi-3ConvLSTMPooling
Accuracy
0.90
0.85
0.80
0.75
0.70
0.65
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Training time in hours

COMPARISON



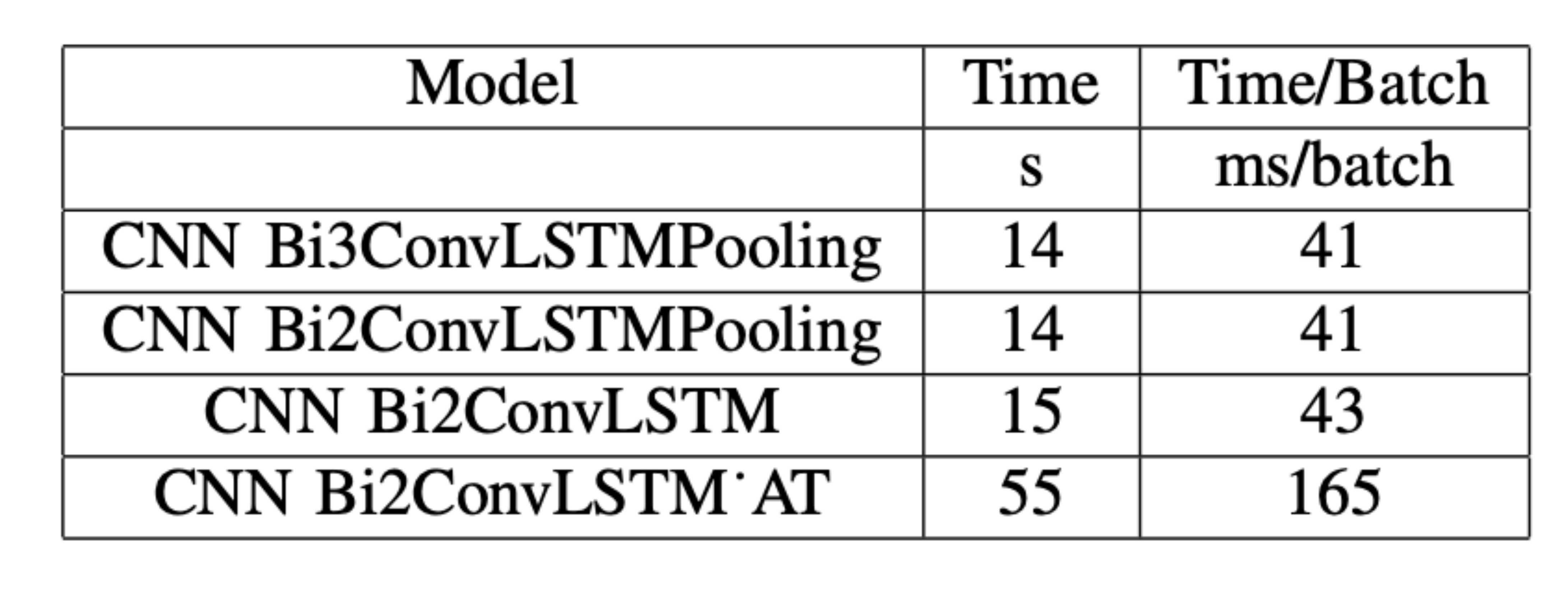
Model	Time	Time/Batch
	s	ms/batch
CNN Bi3ConvLSTMPooling	14	41
CNN Bi2ConvLSTMPooling	14	41
CNN Bi2ConvLSTM	15	43
CNN Bi2ConvLSTM˙AT	55	165



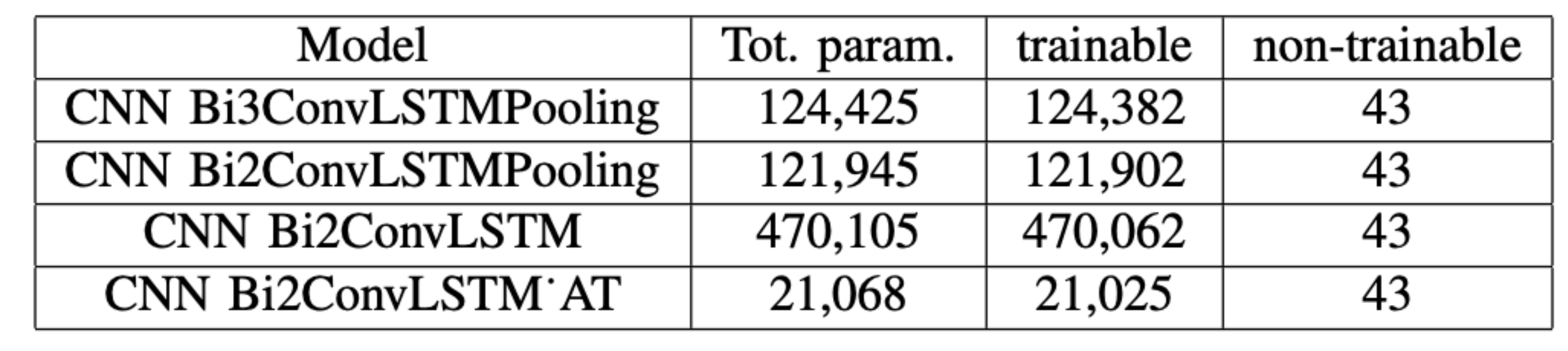
Model	Accuracy	Loss
CNN Bi3ConvLSTMPooling	89.65 %	0.4927
CNN Bi2ConvLSTMPooling	88.77 %	0.5405
CNN Bi2ConvLSTM	88.82 %	0.7116
CNN Bi2ConvLSTM˙AT	81.51 %	0.6180

- Computation time of the attention x4 higher
- After 2.5h of training accuracy:
  - without Att. Layer ~ 88%
  - with Att. Layer ~ 70%

COMPARISON



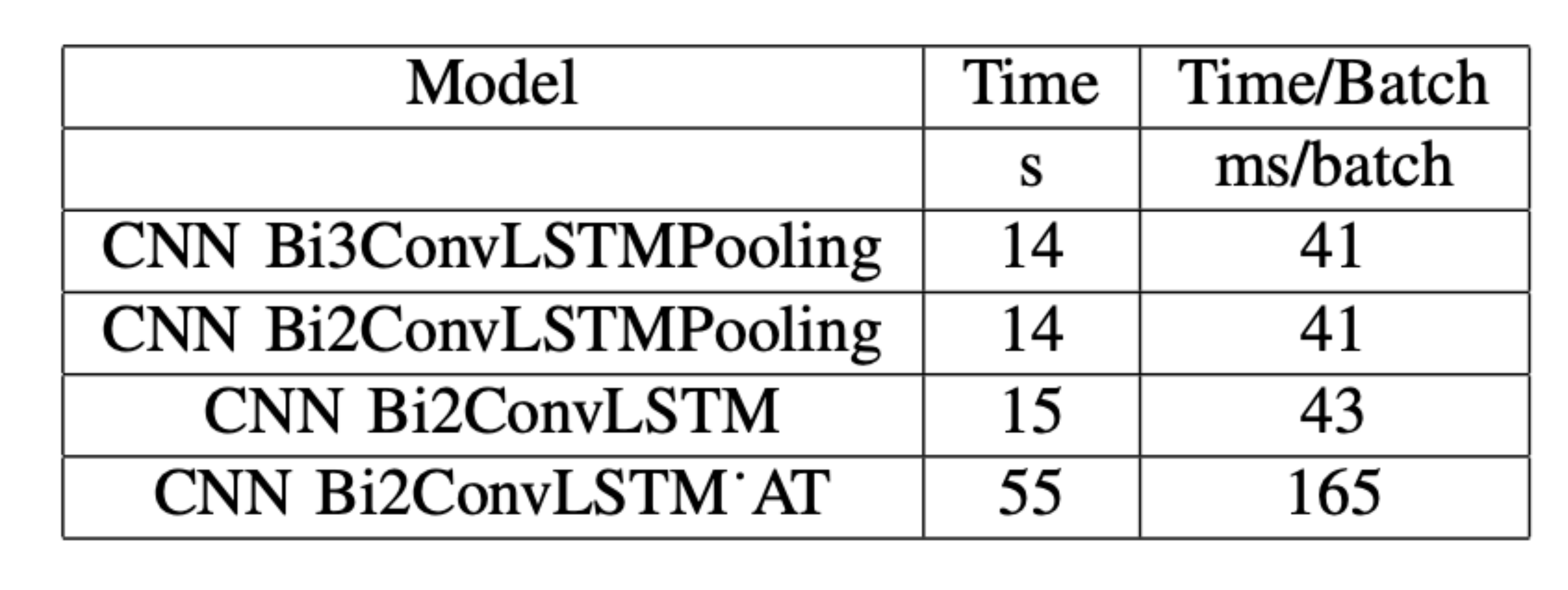
Model	Time	Time/Batch
	s	ms/batch
CNN Bi3ConvLSTMPooling	14	41
CNN Bi2ConvLSTMPooling	14	41
CNN Bi2ConvLSTM	15	43
CNN Bi2ConvLSTM˙AT	55	165



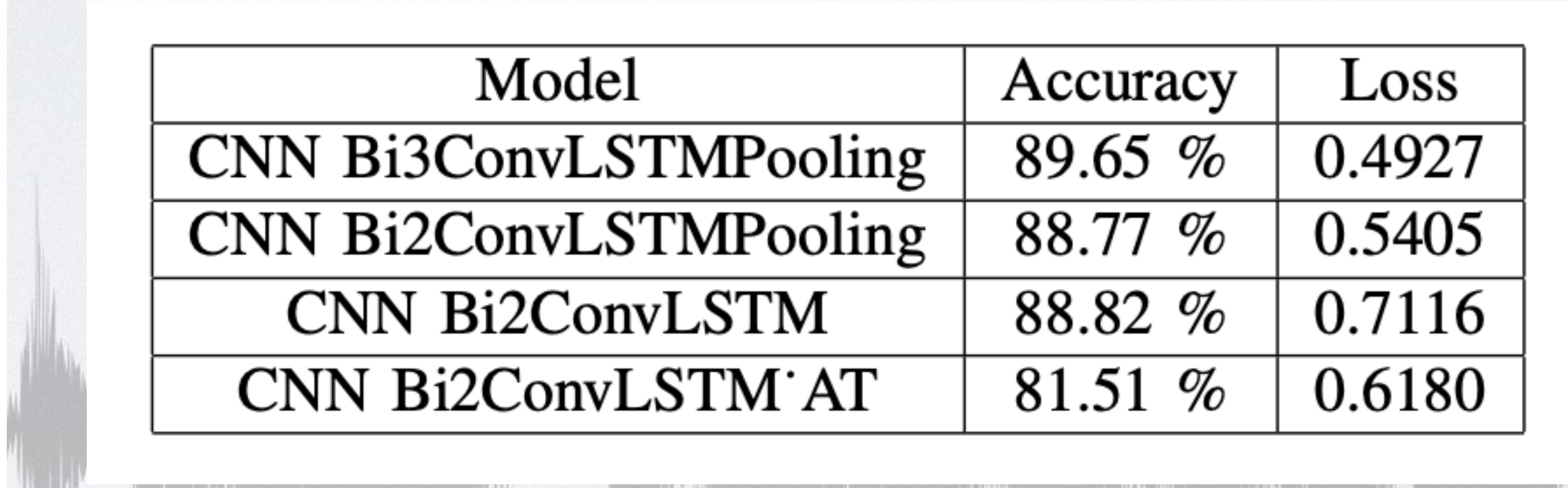
Model	Tot. param.	trainable	non-trainable
CNN Bi3ConvLSTMPooling	124,425	124,382	43
CNN Bi2ConvLSTMPooling	121,945	121,902	43
CNN Bi2ConvLSTM	470,105	470,062	43
CNN Bi2ConvLSTM˙AT	21,068	21,025	43

- Computation time of the attention x4 higher
- After 2.5h of training accuracy:
  - without Att. Layer ~ 88%
  - with Att. Layer ~ 70%

COMPARISON



Model	Time	Time/Batch
	s	ms/batch
CNN Bi3ConvLSTMPooling	14	41
CNN Bi2ConvLSTMPooling	14	41
CNN Bi2ConvLSTM	15	43
CNN Bi2ConvLSTM˙AT	55	165

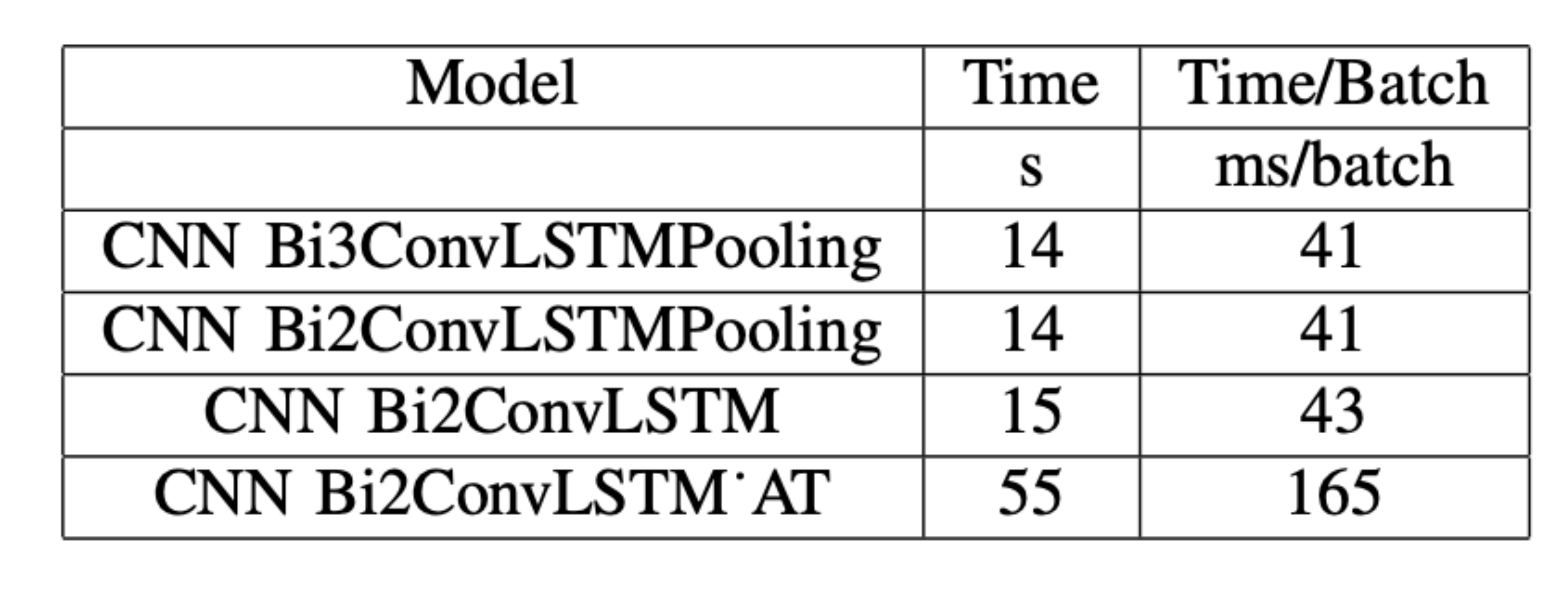


Model	Accuracy	Loss
CNN Bi3ConvLSTMPooling	89.65 %	0.4927
CNN Bi2ConvLSTMPooling	88.77 %	0.5405
CNN Bi2ConvLSTM	88.82 %	0.7116
CNN Bi2ConvLSTM˙AT	81.51 %	0.6180

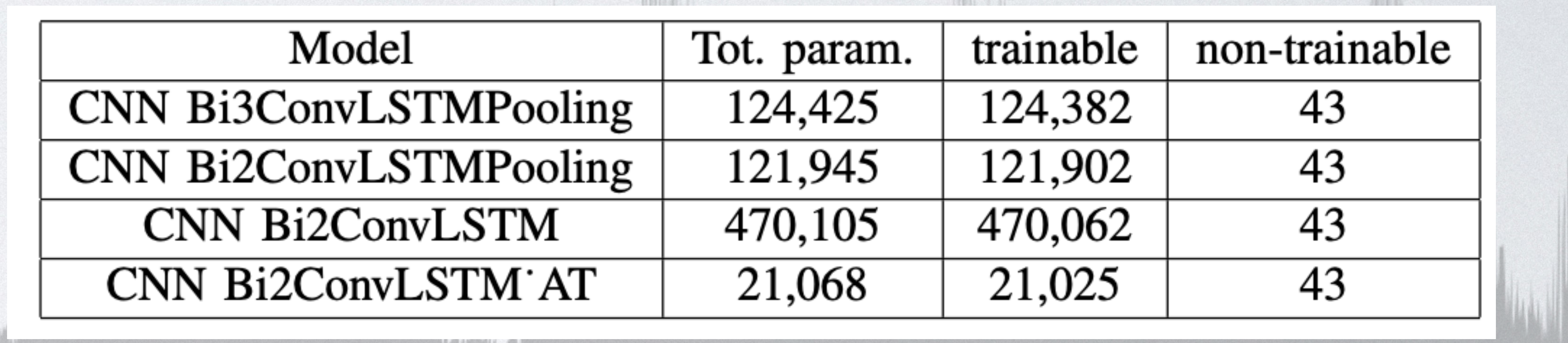
- Still paper of „A neural attention model for speech command recognition“ by Douglas, Sabato, Martin, Bernkopf was able to receive an accuracy of 93.9%

- Trained for 40 epochs and not only 15

COMPARISON



Model	Time	Time/Batch
	s	ms/batch
CNN Bi3ConvLSTMPooling	14	41
CNN Bi2ConvLSTMPooling	14	41
CNN Bi2ConvLSTM	15	43
CNN Bi2ConvLSTM˙AT	55	165



Model	Tot. param.	trainable	non-trainable
CNN Bi3ConvLSTMPooling	124,425	124,382	43
CNN Bi2ConvLSTMPooling	121,945	121,902	43
CNN Bi2ConvLSTM	470,105	470,062	43
CNN Bi2ConvLSTM˙AT	21,068	21,025	43

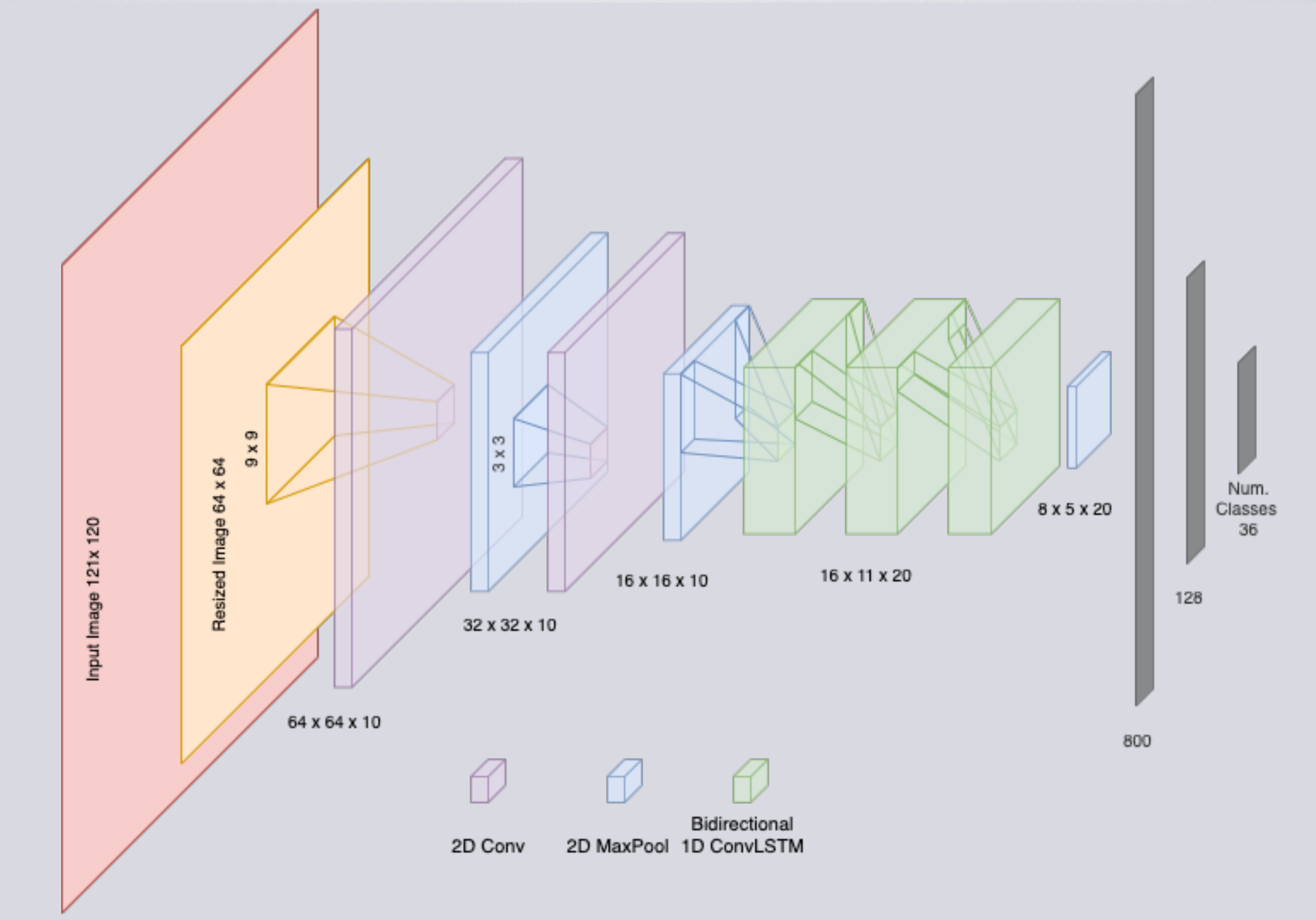
- Still paper of „A neural attention model for speech command recognition“ by Douglas, Sabato, Martin, Bernkopf was able to receive an accuracy of 93.9%

- Trained for 40 epochs and not only 15

CONCLUSION

- For the training steps we used (~40000) models without Att. Layer performed way better
- Accuracy on speech command dataset of 89.65% with only 15 (4) Epochs
- Best model Bidirectional with 3x ConvLSTM + Pooling

- For ver
  per
- Acc 4)
  Epo
- Best



Input Image 121x 120
Resized Image 64 x 64
9 x 9
64 x 64 x 10
3 x 3
32 x 32 x 10
16 x 16 x 10
16 x 11 x 20
8 x 5 x 20
800
128
Num. Classes 36
2D Conv
2D MaxPool
Bidirectional 1D ConvLSTM