BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

PROJECT REPORT

Course No: EEE 6211

Course Title: Digital Speech Processing

Course Instructor: Dr. Mohammad Ariful Haque

Submitted by:

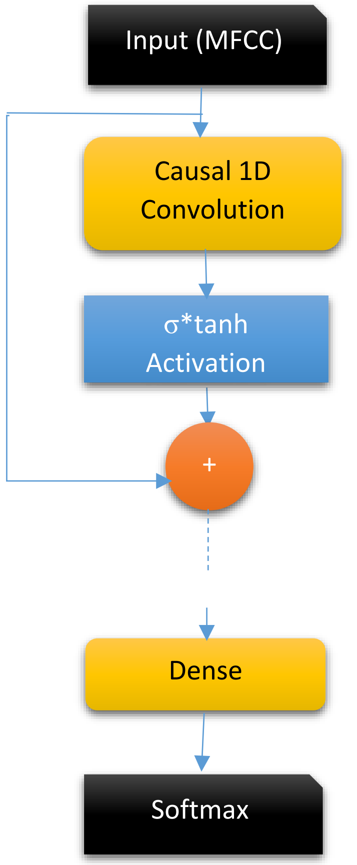
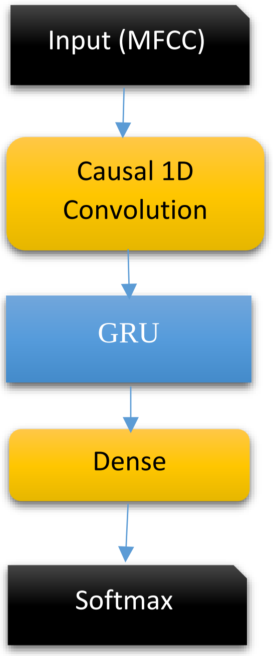
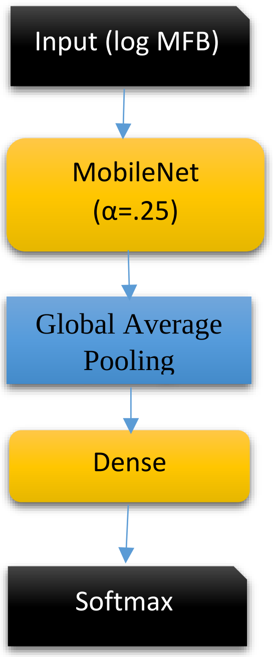
MD. SHAMIM HUSSAIN
Student No.: **0417062229**
Semester: 2nd
E-mail: snirjhar@gmail.com

JAYANTA DEY
Student No.: **0417062214**
Semester: 2nd
E-mail: deyjayanta76@gmail.com

Deep Learning Based Classification:

We considered 3 networks for this classification problem. WaveNet (a 1D convolution network), GRU (a gated recurrent network) and MobileNet (a 2D convolutional network usually used for image classification).

Basic Architecture:

WaveNet	GRU	MobileNet
 Input (MFCC) Causal 1D Convolution σ*tanh Activation + Dense Softmax	 Input (MFCC) Causal 1D Convolution GRU Dense Softmax	 Input (log MFB) MobileNet (α=.25) Global Average Pooling Dense Softmax

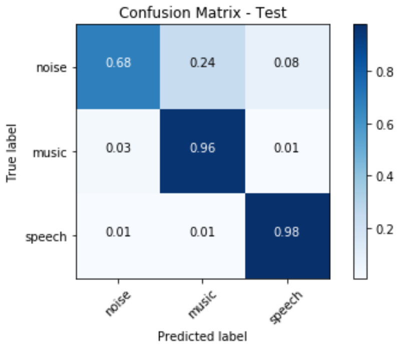
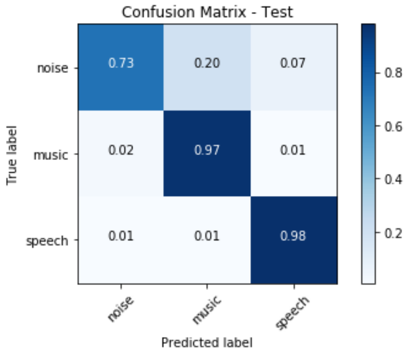
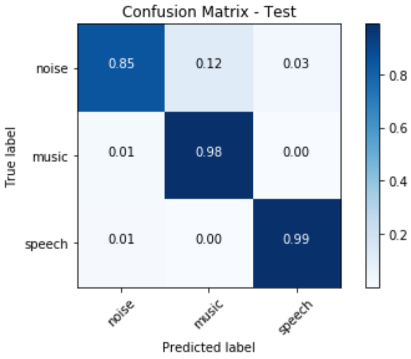
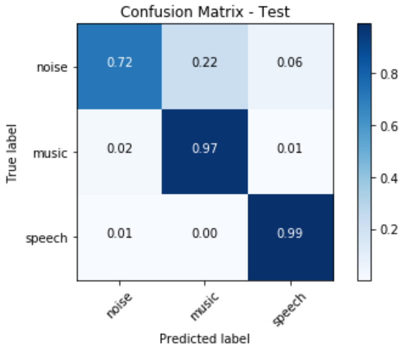
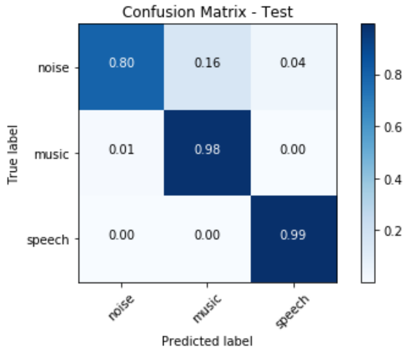
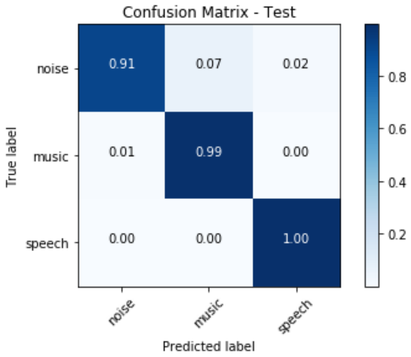
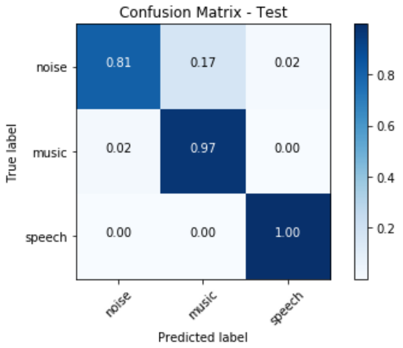
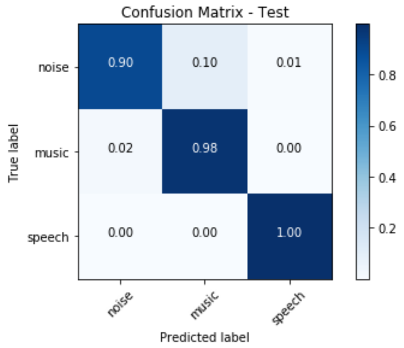
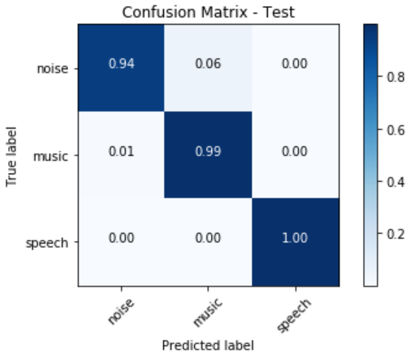
Experimental Method:

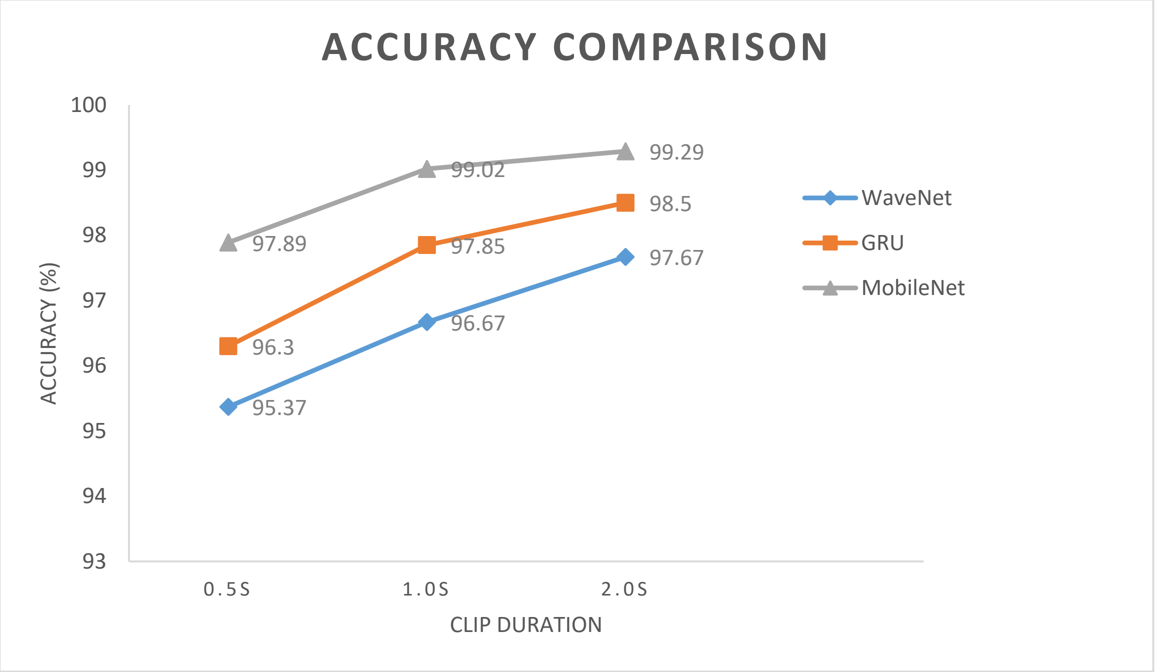
- Of all the files in the dataset, randomly 65 %, 10% and 25% were chosen for training, validation and test correspondingly.
- The files were segmented to form 0.5s, 1s or 2s clips with 50% overlap.
- Very low energy (mostly silent) clips were discarded.
- Segments were framed in 25 ms frames with an overlap of 15 ms (10 ms step).
- 20 MFCC coefficient (from 32 MFB) / 64 log MFB coefficients were extracted from each frame depending on the model to be trained.
- The models were trained with the 2D arrays of MFCC/ MFB features for each clip.
- We initialized MobileNet with ImageNet weights, which resulted in and higher accuracy and faster training.

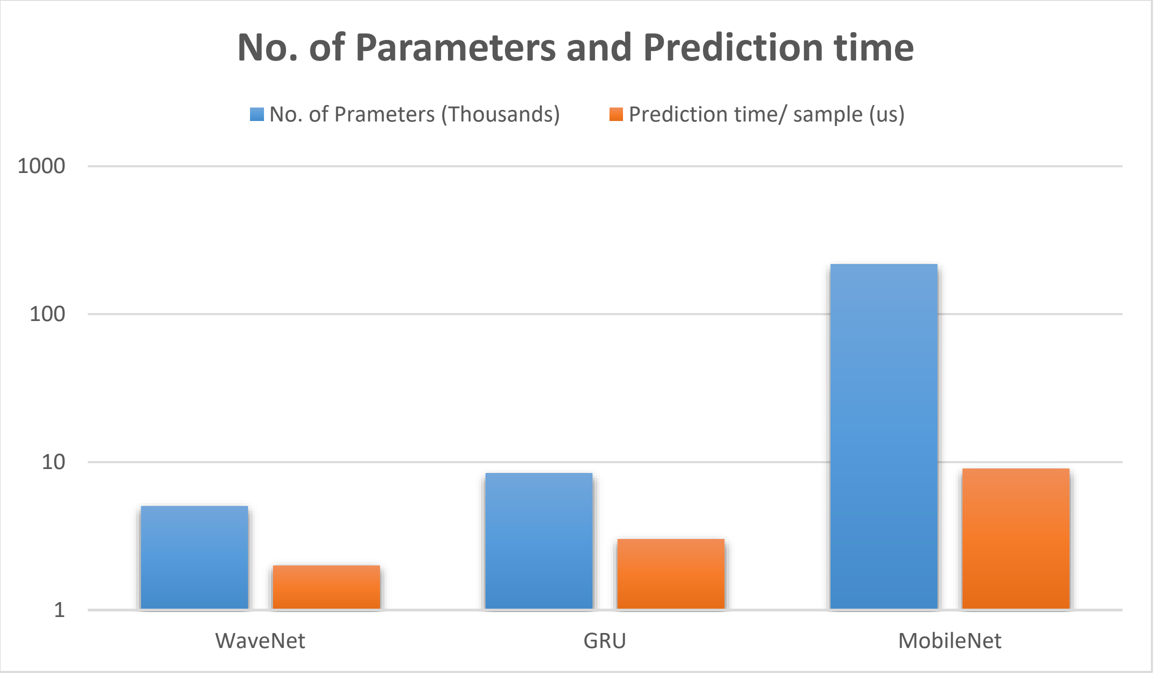
Results: Performance Analysis:

Clip length	0.5 second			1.0 second			2.0 second		
Model	WaveNet	GRU	MobileNet	WaveNet	GRU	MobileNet	WaveNet	GRU	MobileNet
Overall Accuracy	95.37%	96.30%	**97.89%**	96.67%	97.85%	**99.02%**	97.67%	98.50%	**99.29%**
Categorical Cross-entropy (Test)	0.1470	0.1226	**0.0645**	0.1025	0.0760	**0.0567**	0.0825	0.0765	**0.0459**
Speech/Non-speech Classification Accuracy	97.69%	98.27%	**99.09%**	98.80%	99.27%	**99.67%**	99.61%	99.80%	**99.88%**
Noise/Music Classification Accuracy	94.67%	95.52%	**97.25%**	95.20%	96.80%	**98.55%**	95.67%	97.10%	**98.69%**
Prediction-time / Sample (CUDA implemented)	**~ 1.5 us**	~ 2.4 us	~ 7 us	**~ 1.8 us**	~ 2.8 us	~ 8 us	**~ 2 us**	~ 3 us	~ 8.5 us
No. of Parmeters	**5,091**	8,387	217,235	**5,091**	8,387	217,235	**5,091**	8,387	217,235

Normalized Confusion Matrices:

	WaveNet	**GRU**	**MobileNet**
0.5s			
1.0s			
2.0s			


ACCURACY COMPARISON
ACCURACY (%)
100
99
98
97
96
95
94
93
0.5S
1.0S
2.0S
CLIP DURATION
97.89
99.02
99.29
96.3
97.85
98.5
95.37
96.67
97.67
WaveNet
GRU
MobileNet


No. of Parameters and Prediction time
No. of Prameters (Thousands)
Prediction time/ sample (us)
1000
100
10
1
WaveNet
GRU
MobileNet

Advantages and Disadvantages:

WaveNet:

Advantage:

- Very lightweight, only ~5000 parameters.
- Very fast, only ~2 us prediction time/sample.
- Convolutional, easy to implement parallel computation.

Disadvantage:

- Lower accuracy than other networks.

GRU:

Advantage:

- Better performance at relatively low number of parameters, ~8000.
- Faster training (than WaveNet), takes only ~30 epochs to fully train.
- Relatively simple architecture.

Disadvantage:

- Recurrent network, hard to implement parallel computation.
- Slower than convolutional networks (like WaveNet).

MobileNet:

Advantage:

- Excellent accuracy with transfer learning from ImageNet weights.
- Faster training, we trained for only around 10 epochs.
- The same model can be used in a mobile device for image and audio classification (with different weights).

Disadvantage:

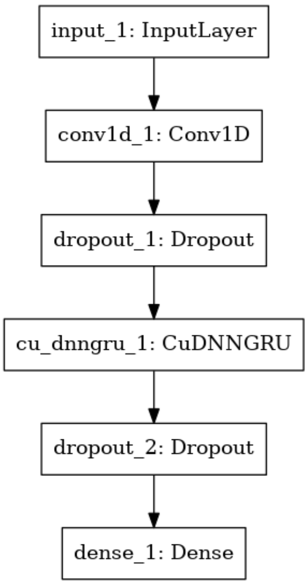
- Relatively bulky, ~200,000 parameters
- Relatively slow due to its big size.
- Non-causal, although causal implementation may be possible.

Future Work:

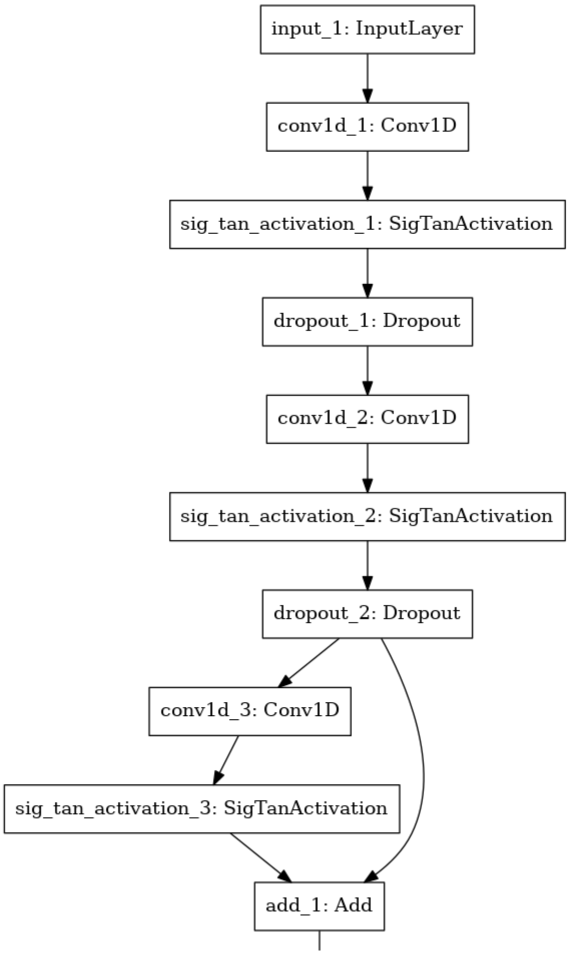
- From the results we observed that it is easy to classify speech and non-speech but it is much harder to classify noise and music, especially for shorter segments. So we could build an ensemble classifier in a boosting configuration to further improve the results.
- From the high performance of MobileNet it is apparent that 2D convolution is very effective for audio classification. So, if we could cut down the size of the model, it may be possible to achieve high performance at faster speed.
- WaveNet can operate on raw audio, sample by sample rather than on derived features like MFCC. But we could not try that due to constraint or computational resources. So, we may try that in the future.
- Transfer learning from ImageNet training shows that it could be useful for audio classification as well. We could, try to implement transfer learning from a big audio dataset for higher performance.

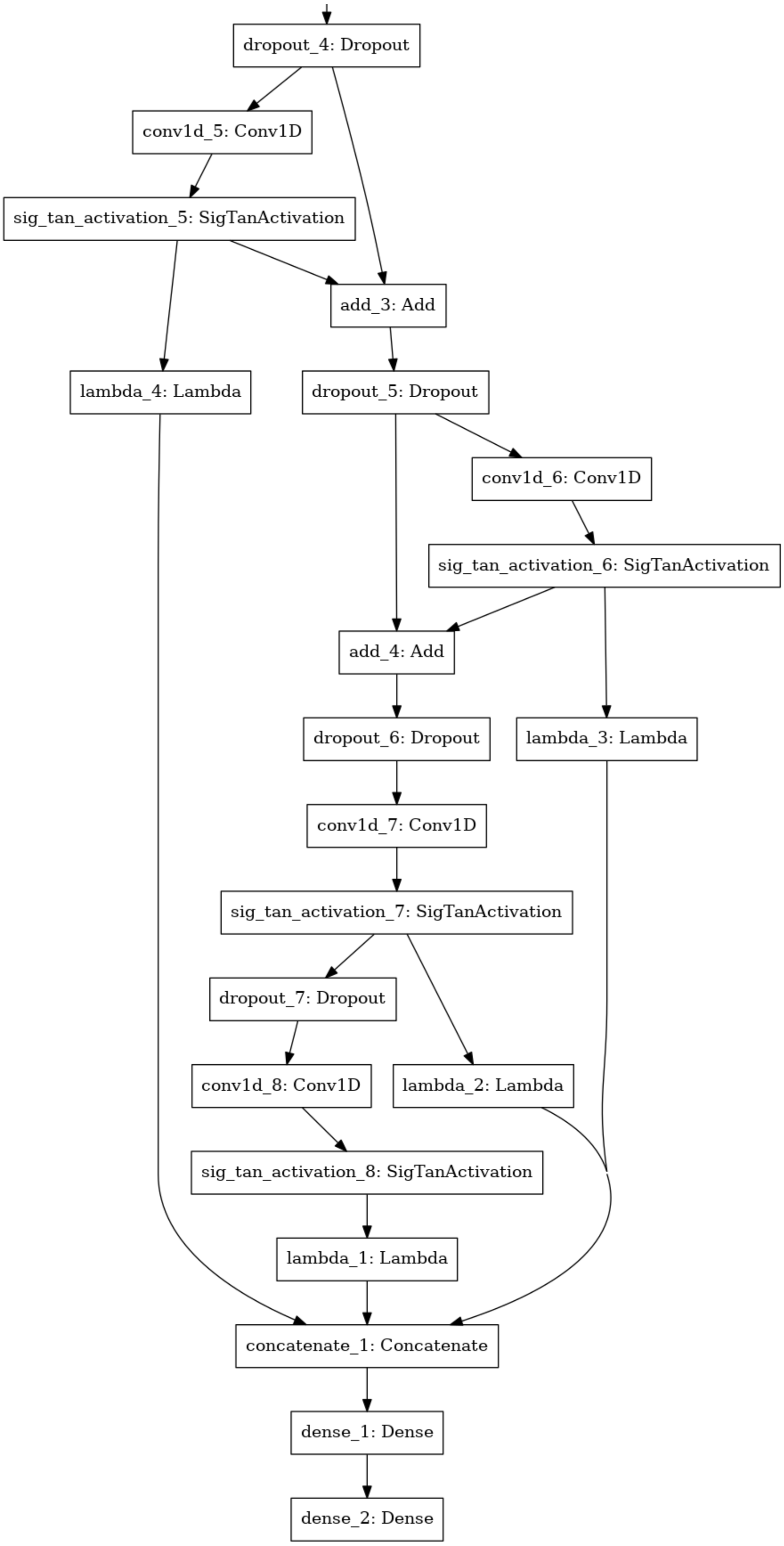
Appendix: Detailed Architectures:

GRU:

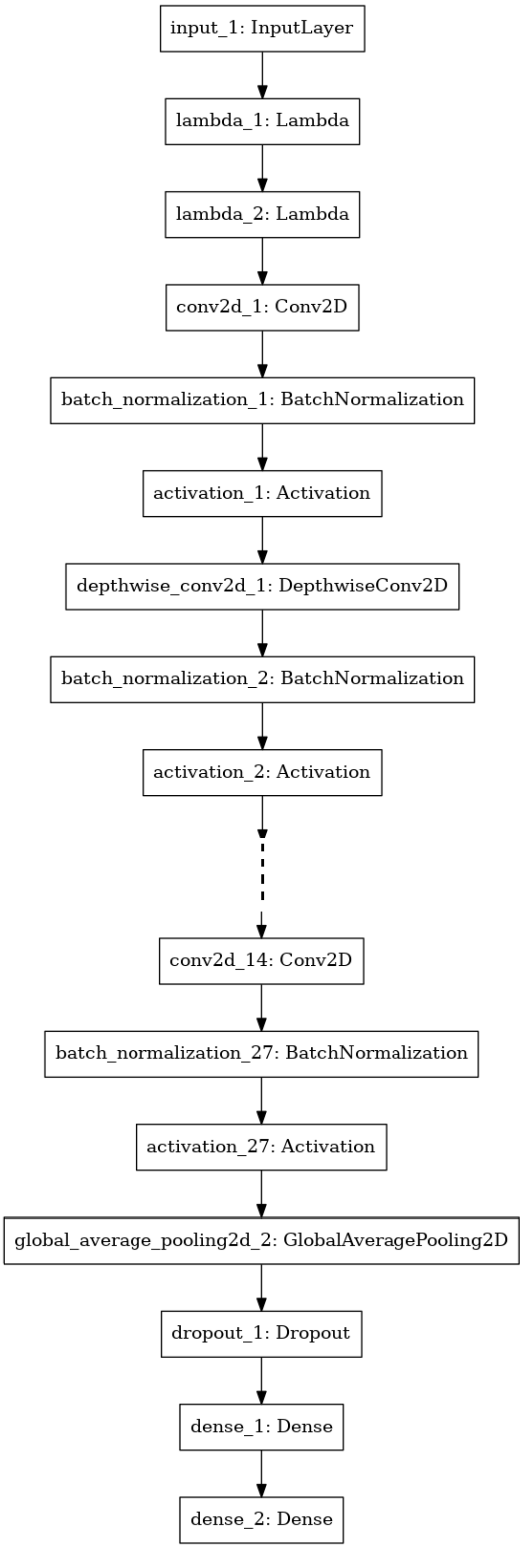


WaveNet:


input_1: InputLayer
conv1d_1: Conv1D
sig_tan_activation_1: SigTanActivation
dropout_1: Dropout
conv1d_2: Conv1D
sig_tan_activation_2: SigTanActivation
dropout_2: Dropout
conv1d_3: Conv1D
sig_tan_activation_3: SigTanActivation
add_1: Add


dropout_4: Dropout
conv1d_5: Conv1D
sig_tan_activation_5: SigTanActivation
add_3: Add
lambda_4: Lambda
dropout_5: Dropout
conv1d_6: Conv1D
sig_tan_activation_6: SigTanActivation
add_4: Add
dropout_6: Dropout
lambda_3: Lambda
conv1d_7: Conv1D
sig_tan_activation_7: SigTanActivation
dropout_7: Dropout
conv1d_8: Conv1D
lambda_2: Lambda
sig_tan_activation_8: SigTanActivation
lambda_1: Lambda
concatenate_1: Concatenate
dense_1: Dense
dense_2: Dense

MobileNet:


input_1: InputLayer
lambda_1: Lambda
lambda_2: Lambda
conv2d_1: Conv2D
batch_normalization_1: BatchNormalization
activation_1: Activation
depthwise_conv2d_1: DepthwiseConv2D
batch_normalization_2: BatchNormalization
activation_2: Activation
conv2d_14: Conv2D
batch_normalization_27: BatchNormalization
activation_27: Activation
global_average_pooling2d_2: GlobalAveragePooling2D
dropout_1: Dropout
dense_1: Dense
dense_2: Dense