Urban Sound Challenge
Audio Classification with Deep Learning

From Feature Extraction to Recurrent Neural Networks

Micha Birklbauer, 2020

Urban Sound Challenge – The Dataset

- Audio recordings of urban sounds
- 10 class multiclassification-problem
- 3673 samples with length of up to 4 seconds
- 3637 samples used for training
  - 70% of training samples used for training
  - 30% of training samples used for validation
- 33 samples used for testing (evaluation)
- 3 samples unused because of missing class

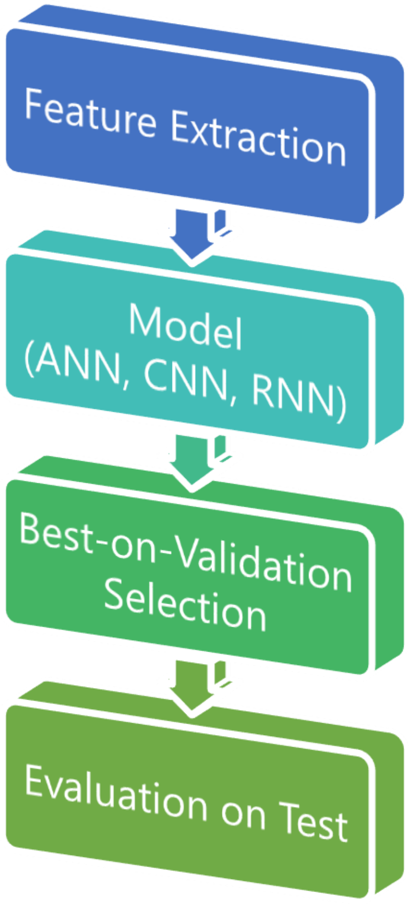
- Baseline:
  - siren: most frequent class – makes up 13.4% of total training samples
  - Therefore the baseline is an accuracy of >13.4%

- Classes (with Training Distribution):
  - air_conditioner: 351 samples
  - car_horn: 249 samples
  - children_playing: 375 samples
  - dog_bark: 433 samples
  - drilling: 468 samples
  - engine_idling: 346 samples
  - gun_shot: 151 samples
  - jackhammer: 377 samples
  - ***siren: 488 samples (13.4% -> baseline)***
  - street_music: 399 samples

Pipeline

- Feature Extraction:
  - Mel-Frequency Cepstrum Coefficients (MFCC)
  - Chromagram
  - Melspectrogram
  - Spectral Contrast
  - Tonnetz
- Training with different Model Architectures:
  - Classical Feed-Forward Neural Networks
  - Convolutional Neural Networks
  - Recurrent Neural Networks
  - Recurrent Neural Networks
    with Gated Recurrent Units (GRU)
  - Recurrent Neural Networks
    with Long-Short Term Memory (LSTM)
- Best-on-Validation Selection:
  - Select model with highest validation accuracy
- Evaluation on Test Data:
  - 33 audio samples
- Model Architectures trained in total: 57

Workflow



Feature Extraction – Mel-Frequency Cepstrum Coefficients (MFCC)

- Definitions:
  - **Power Spectrum:** Describes the distribution of power (=the energy that is transfered) as frequencies.
  - **Mel Scale:** Perceptual scale of pitches judged by listeners to be equal in distance from one another.
  - **Mel-Frequency Cepstrum:** Representation of the short-term power spectrum of a sound, based on the linear cosine transform of the log power spectrum on the mel scale.
- Mel-Frequency Cepstrum Coefficients (MFCC):

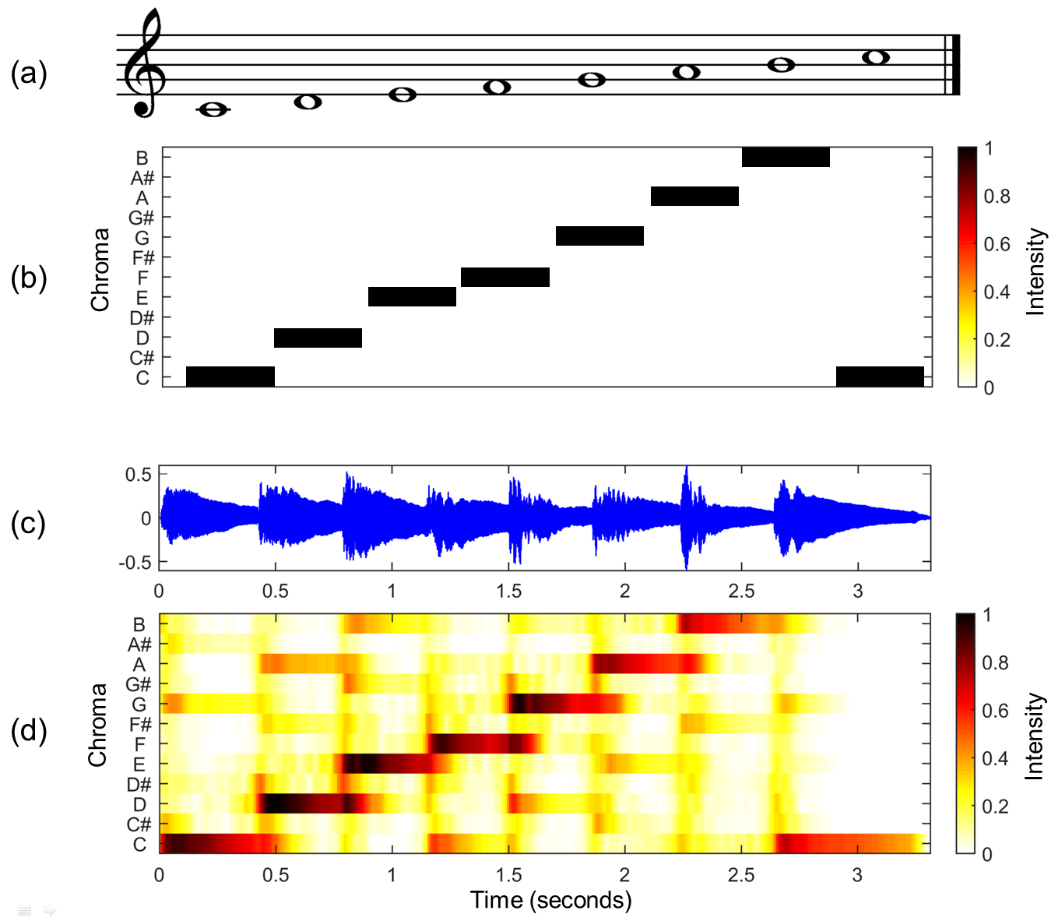
  Coefficients that collectively make up a Mel-Frequency Cepstrum.
- MFCC Usage:
  - Speech Recognition
  - Audio Classification
  - Audio Clustering

- MFCC Extraction:
  - Take the Fourier transform of (a window of) a signal.
  - Map the powers of the spectrum obtained onto the mel scale, using overlapping windows.
  - Take the logarithms of the powers for each of the mel frequencies.
  - Take the discrete cosine transform of the mel log powers.
  - The MFCCs are the amplitudes of the resulting spectrum.

Source: Wikipedia

Feature Extraction – Chromagram

- A chromagram is a categorization for pitches of music or sound into usually 12 different categories. Typically a chromagram captures harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation.
- The picture on the right illustrates:
  - A – the musical score of a C-major scale.
  - B – the chromagram obtained from the score.
  - C – audio recording of the C-major scale played on a piano.
  - D – chromagram obtained from the audio recording.

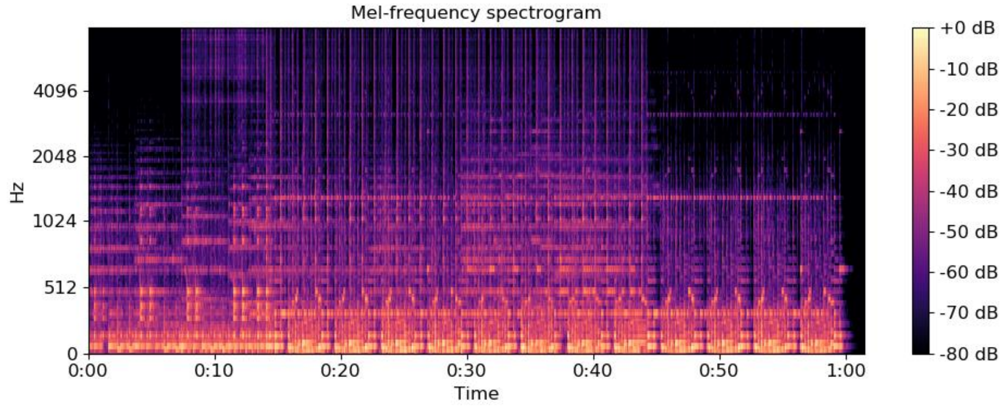


Source: [Wikipedia]()

Feature Extraction – Melspectrogram

- Definitions:
  - **Spectrogram:** Visual representation of the spectrum of frequencies of a signal as it varies with time.
  - **Mel Scale:** Perceptual scale of pitches judged by listeners to be equal in distance from one another.
- Melspectrogram:

  The mel-scaled spectrogram of an audio segment.

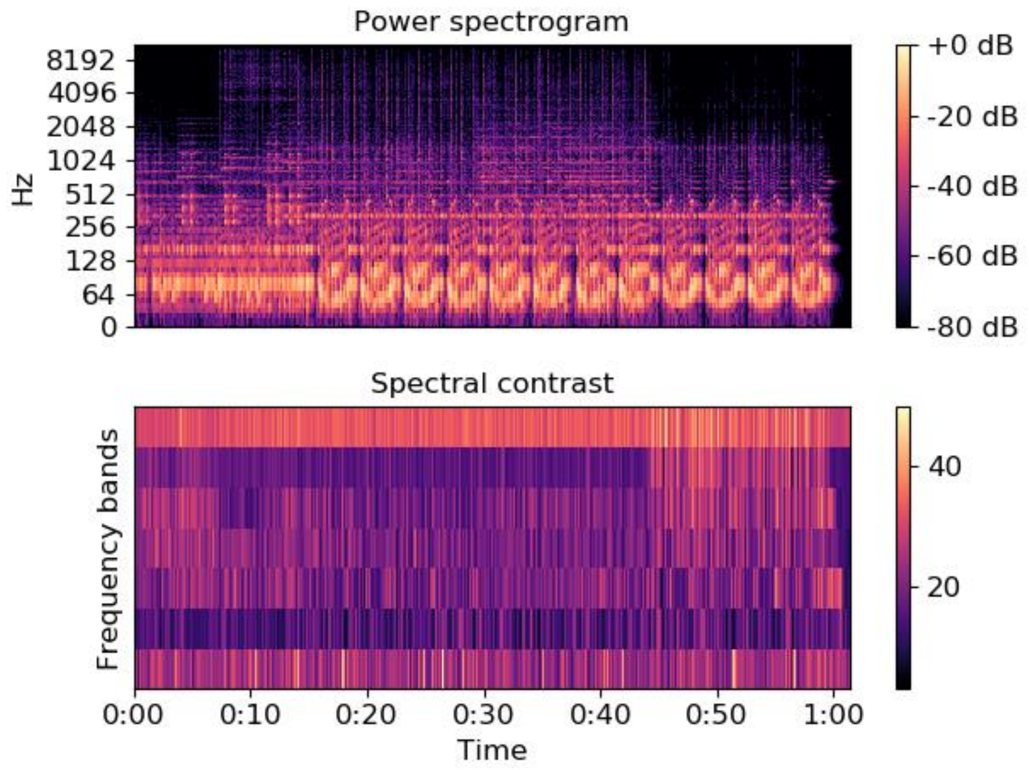


Source: LibROSA

Source: Wikipedia

Feature Extraction – Spectral Contrast

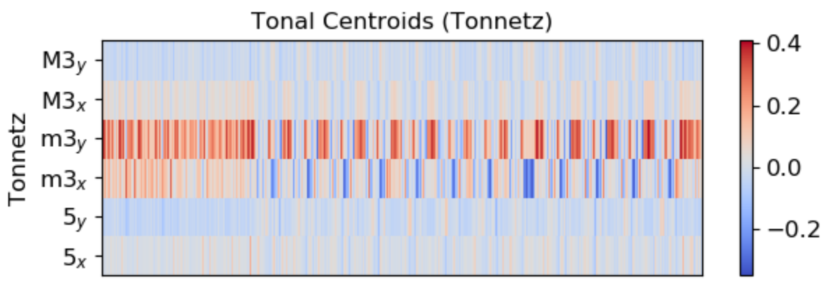
- Each frame of a power spectrogram is divided into bands. For each band, the energy contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy).

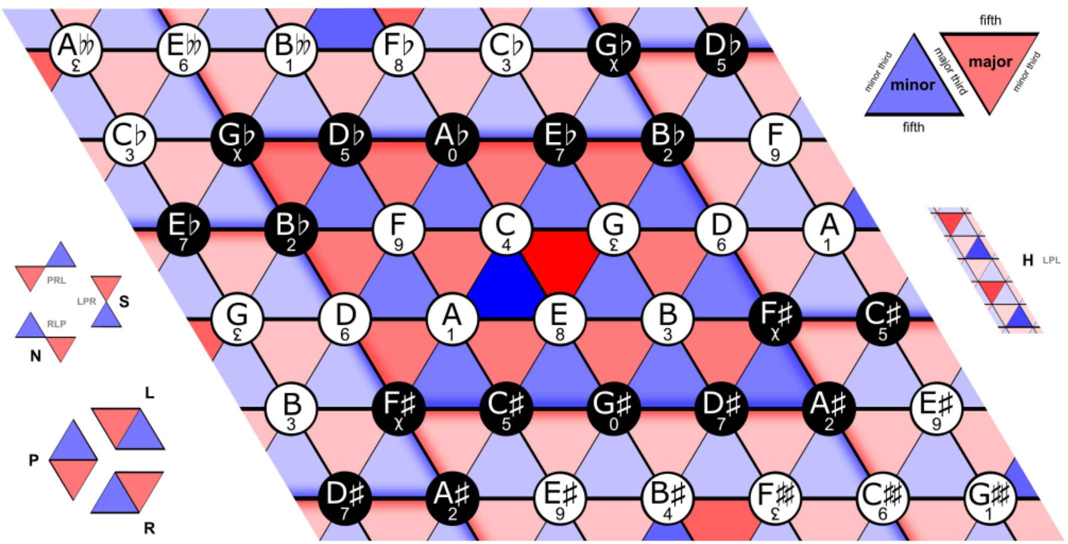


Source: [LibROSA]

Feature Extraction – Tonnetz

- The Tonnetz is a conceptual lattice diagram representing tonal space.
  (left picture)
- In LibROSA represented as the 6 tonal centroid features at each timestep.
  (below picture)





Sources: Wikipedia / LibROSA

Model Architectures

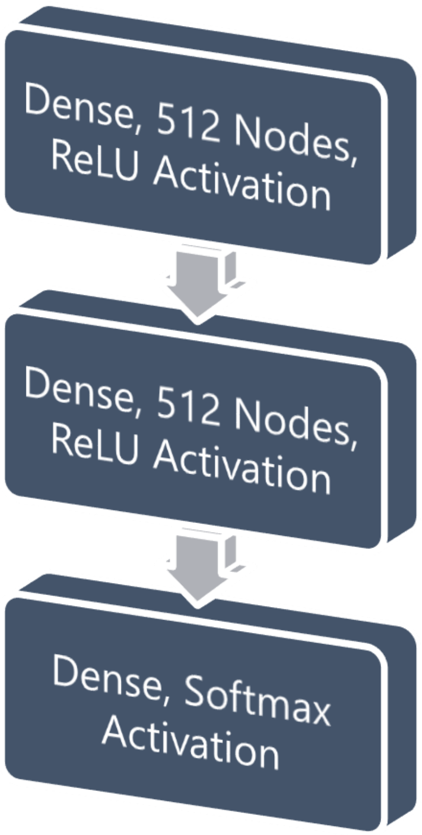
- Feed-Forward Neural Nets
- Convolutional Neural Nets
- Recurrent Neural Nets

Feed-Forward Neural Networks – Architecture I

Description & Results

- Simple 2 layer network with 512 nodes per layer and ReLU activation functions. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.907509***

Architecture

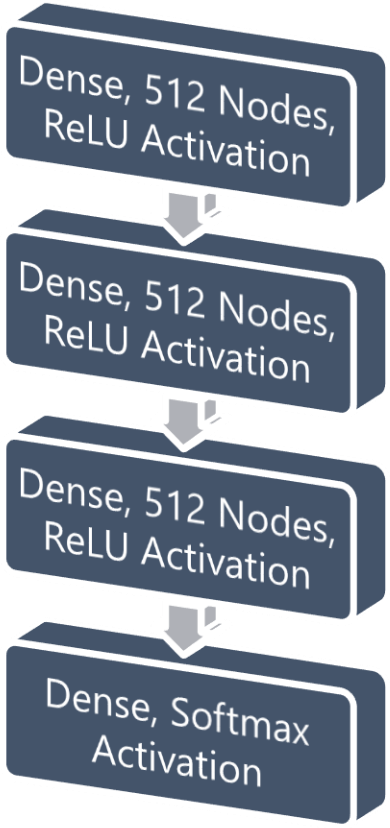


Feed-Forward Neural Networks – Architecture II

Description & Results

- 3 layer network with 512 nodes per layer and ReLU activation functions. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.912088***

Architecture

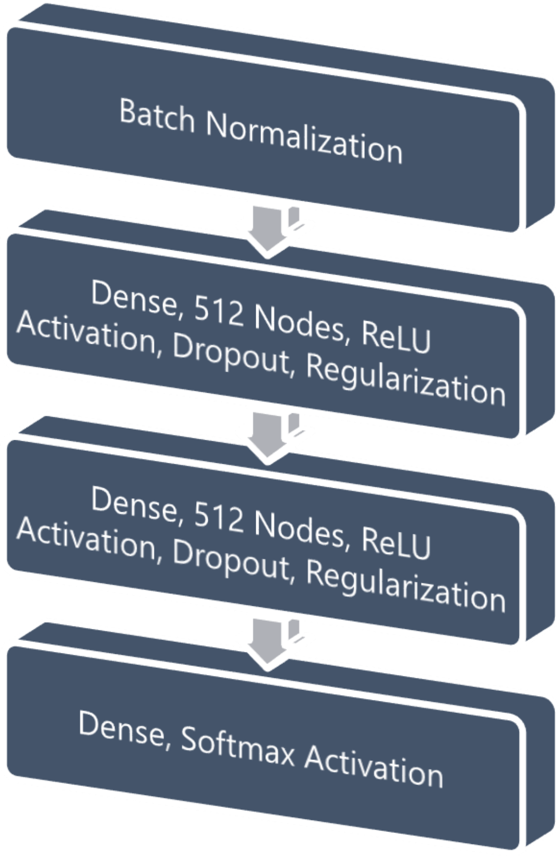


Feed-Forward Neural Networks – Architecture III

Description & Results

- 2 layer network with 512 nodes per layer and ReLU activation functions. Additionally layers have 0.3 dropout and 0.01 L2 kernel- and bias regularization. Inputs are batch normalized. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.946886***

Architecture

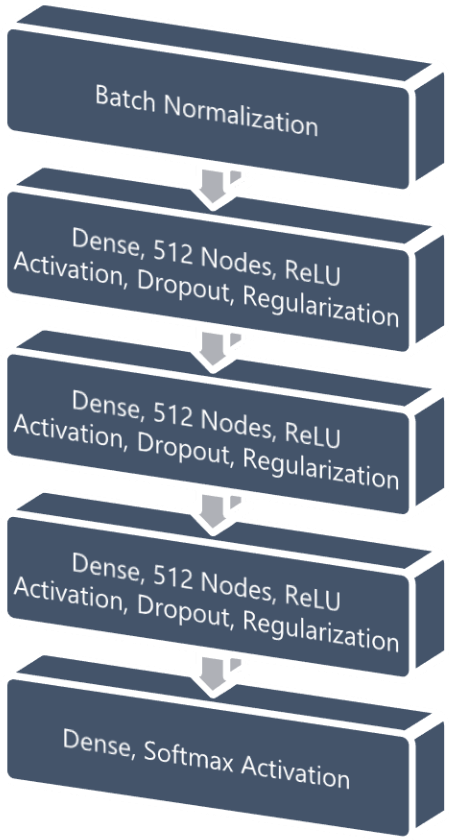


Feed-Forward Neural Networks – Architecture IV

Description & Results

- 3 layer network with 512 nodes per layer and ReLU activation functions. Additionally layers have 0.3 dropout and 0.01 L2 kernel- and bias regularization. Inputs are batch normalized. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.945971***

Architecture

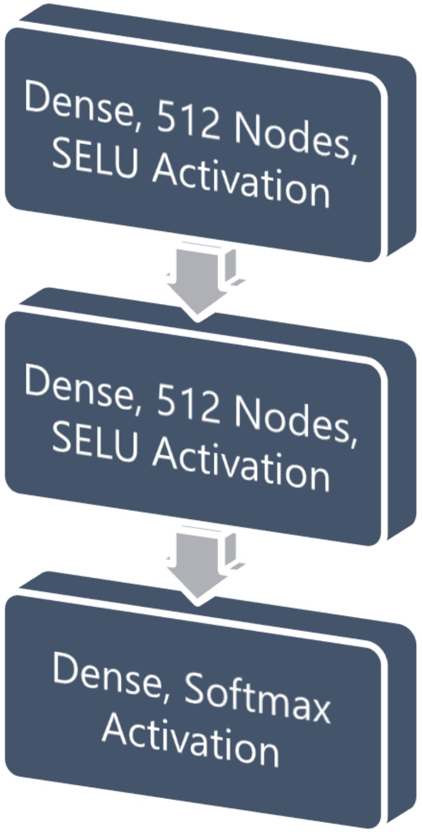


Feed-Forward Neural Networks – Architecture V

Description & Results

- 2 layer network with 512 nodes per layer and SELU activation functions. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.913004***

Architecture

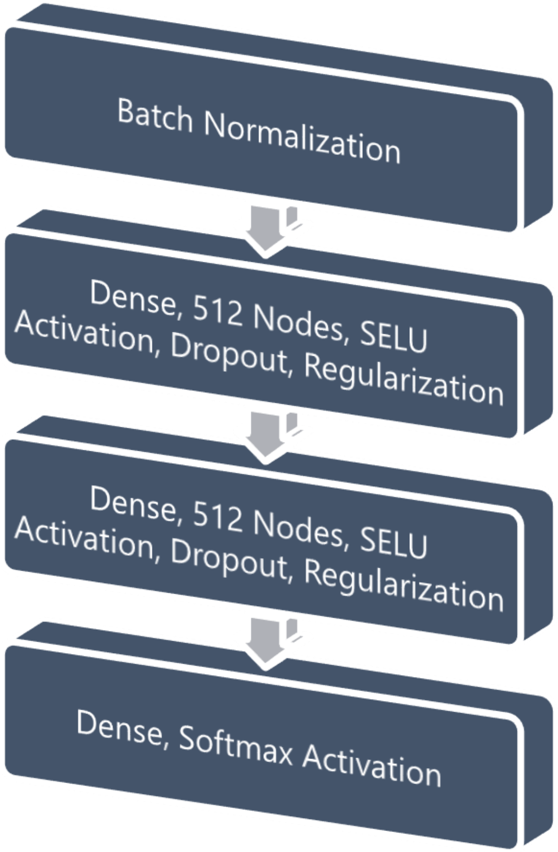


Feed-Forward Neural Networks – Architecture VI

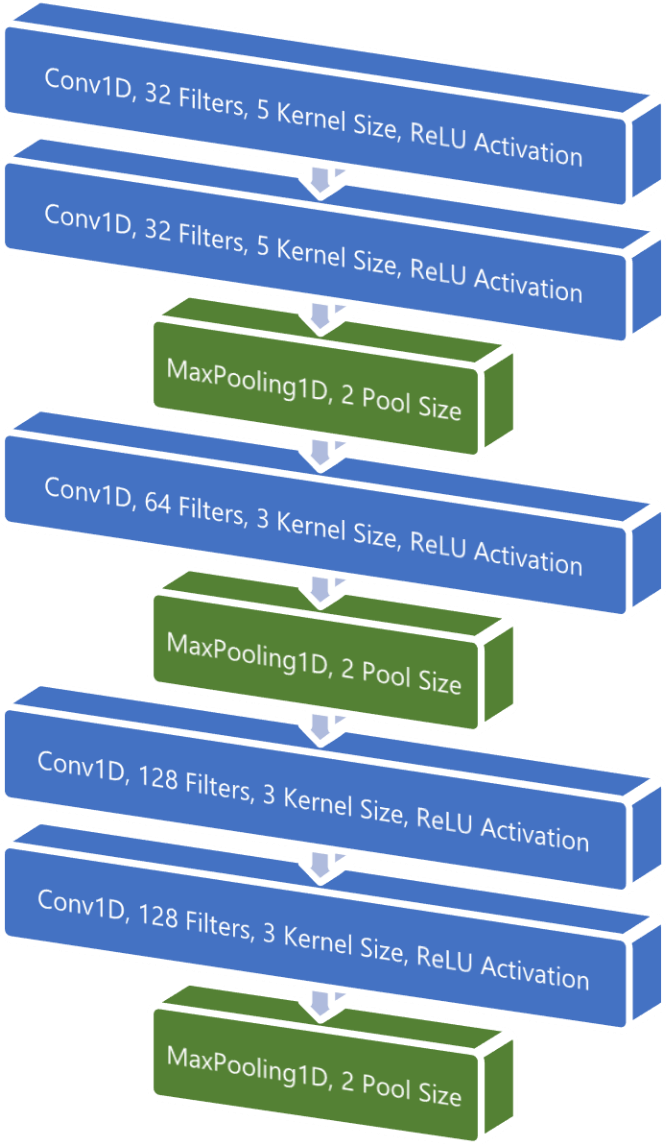
Description & Results

- 2 layer network with 512 nodes per layer and SELU activation functions. Additionally layers have 0.3 dropout and 0.01 L2 kernel- and bias regularization. Inputs are batch normalized. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.874542***

Architecture



Convolutional Neural Networks – Architecture I



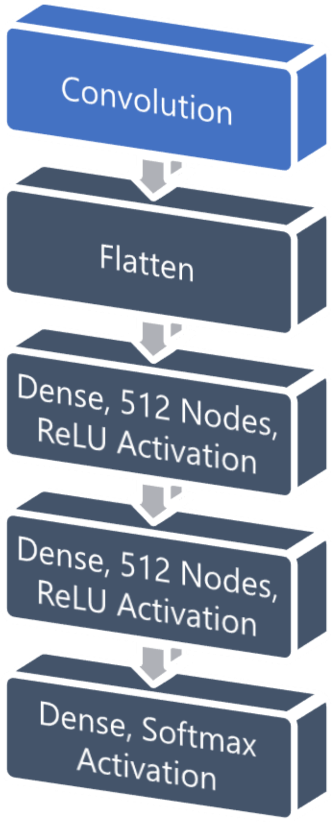
Convolution – Description

- 5 convolutional layers, all with ReLU activation functions.
- Different number of filters and varying kernel size in convolutional layers.
- 3 pooling layers doing max pooling with a pool size of 2.

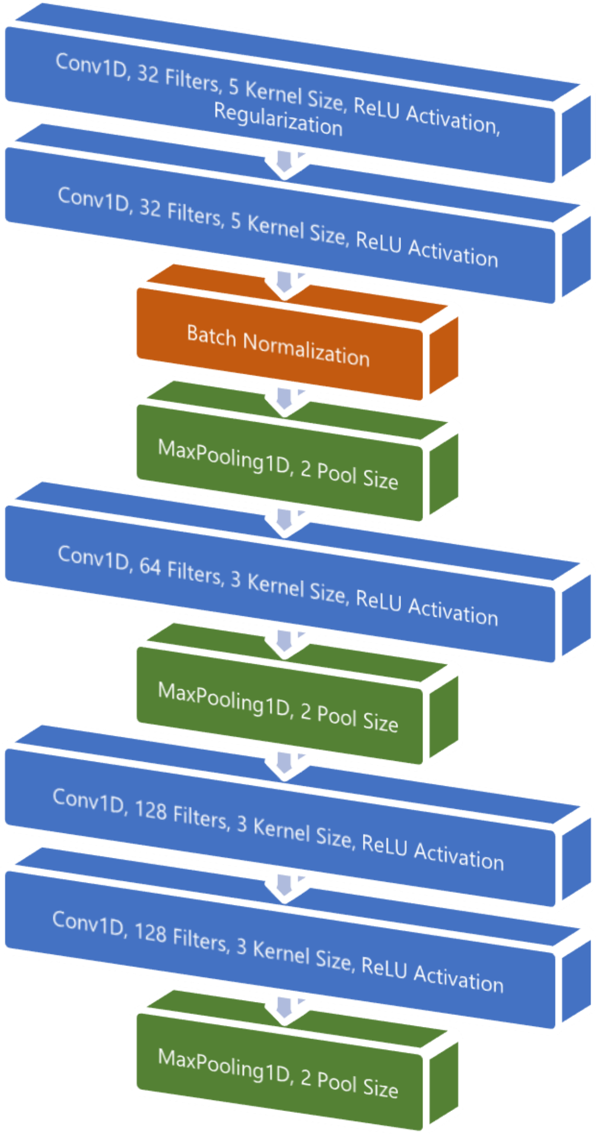
Convolutional Neural Networks – Architecture I

Feed-Forward Description & Results

- 2 layer network with 512 nodes per layer and ReLU activation functions. Convolution input is flattened. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.931319***



Convolutional Neural Networks – Architecture II



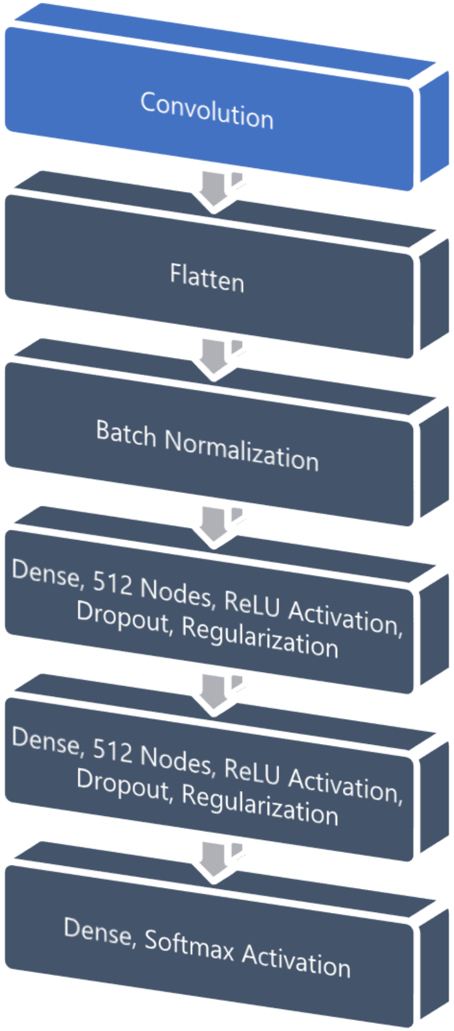
Convolution – Description

- 5 convolutional layers, all with ReLU activation functions.
- Different number of filters and varying kernel size in convolutional layers.
- 0.01 kernel- and bias regularization in the first convolutional layer.
- Batch normalization before first pooling layer.
- 3 pooling layers doing max pooling with a pool size of 2.

Convolutional Neural Networks – Architecture II

Feed-Forward Description & Results

- 2 layer network with 512 nodes per layer and ReLU activation functions. Additionally layers have 0.3 dropout and 0.01 L2 kernel- and bias regularization. Convolution input is batch normalized and flattened. Output layer with softmax activation.
- Best Features for this Architecture: Combination of all
- Best-on-Validation Accuracy: ***0.946886***

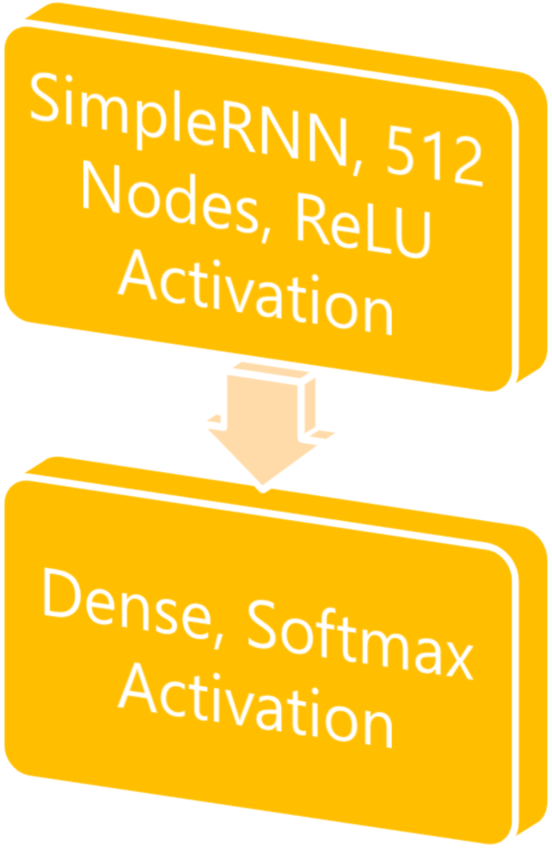


Recurrent Neural Networks – Architecture I

Description & Results

- Simple 1 layer RNN with 512 nodes and ReLU activation function. Ouput layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.901099

Architecture

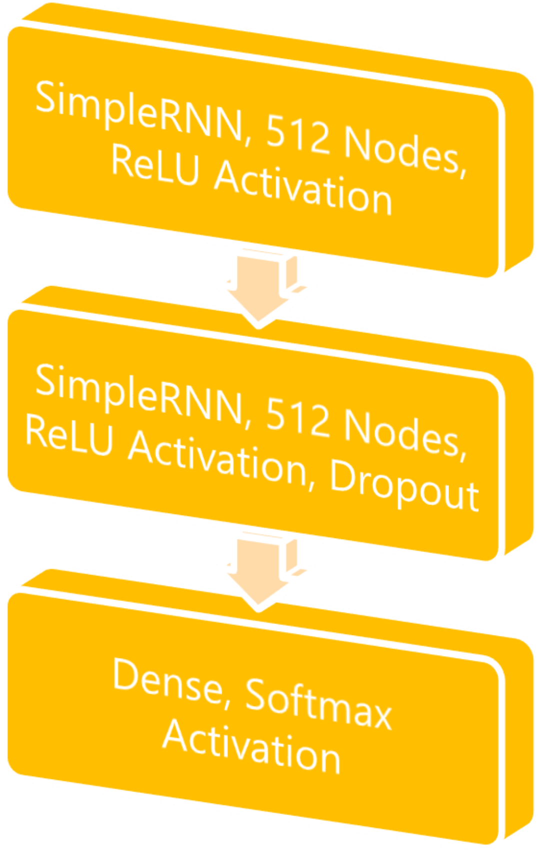


Recurrent Neural Networks – Architecture II

Description & Results

- 2 layer RNN with 512 nodes per layer and ReLU activation functions. The second layer also has 0.2 dropout and recurrent dropout. Ouput layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.880952

Architecture

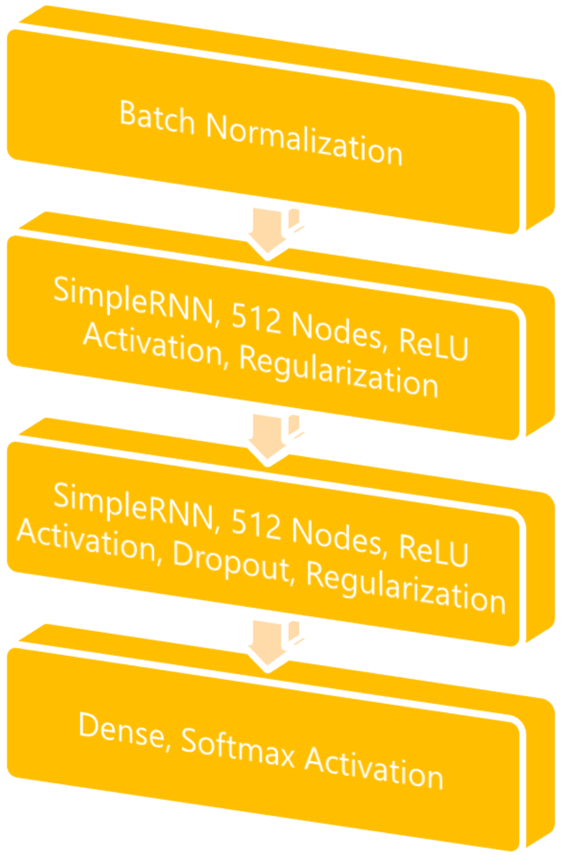


Recurrent Neural Networks – Architecture III

Description & Results

- 2 layer RNN with 512 nodes per layer and ReLU activation functions. The second layer also has 0.2 dropout and recurrent dropout. Additionally layers have 0.01 L2 kernel-, bias- and recurrent regularization. Inputs are batch normalized. Ouput layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs (small samplesize due to bad results)
- Best-on-Validation Accuracy:
  0.326007

Architecture

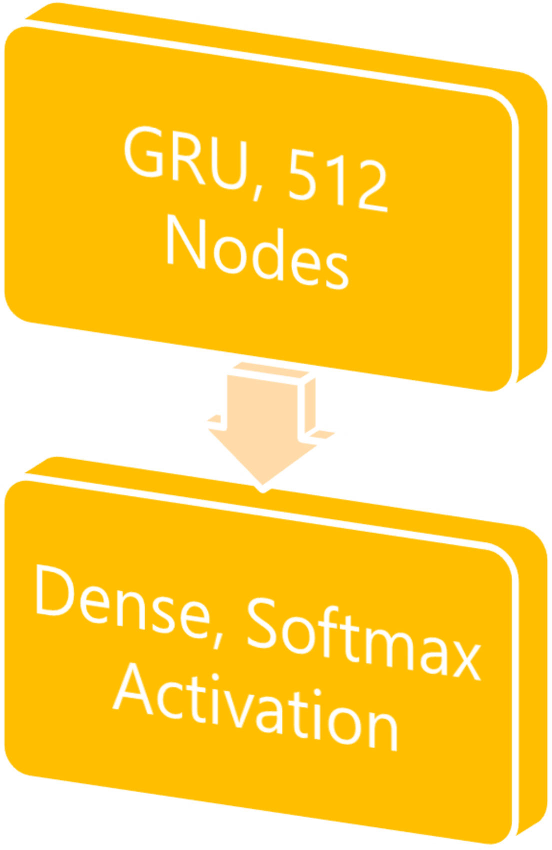


Recurrent Neural Networks – Architecture IV (light)

Description & Results

- Simple 1 layer Gated Recurrent Unit-RNN with 512 nodes. Output layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.578755

Architecture

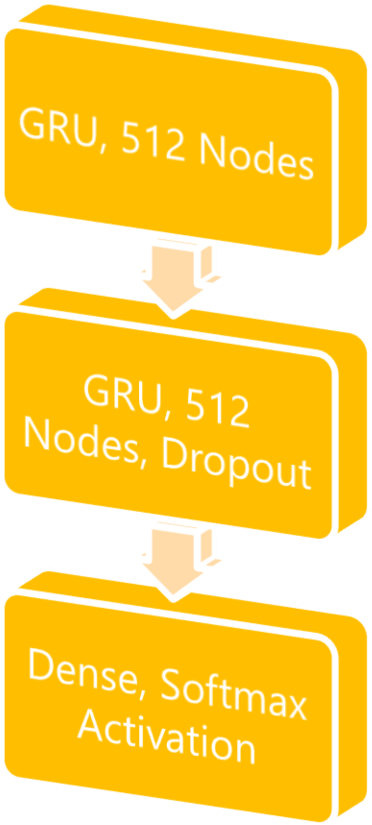


Recurrent Neural Networks – Architecture IV

Description & Results

- 2 layer Gated Recurrent Unit-RNN with 512 nodes per layer. The second layer also has 0.2 dropout and recurrent dropout. Output layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.769231

Architecture

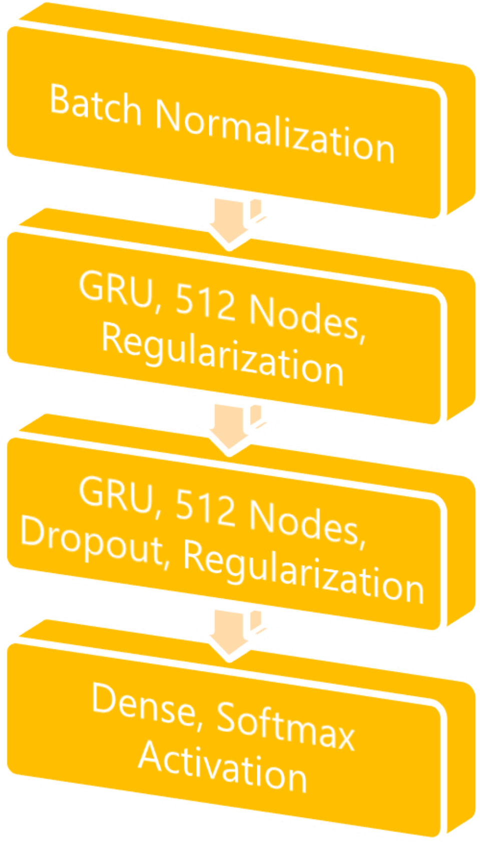


Recurrent Neural Networks – Architecture V

Description & Results

- 2 layer Gated Recurrent Unit-RNN with 512 nodes per layer. The second layer also has 0.2 dropout and recurrent dropout. Additionally layers have 0.01 L2 kernel-, bias- and recurrent regularization. Inputs are batch normalized. Ouput layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs (small samplesize due to bad results)
- Best-on-Validation Accuracy:
  0.13461 (Baseline!)

Architecture

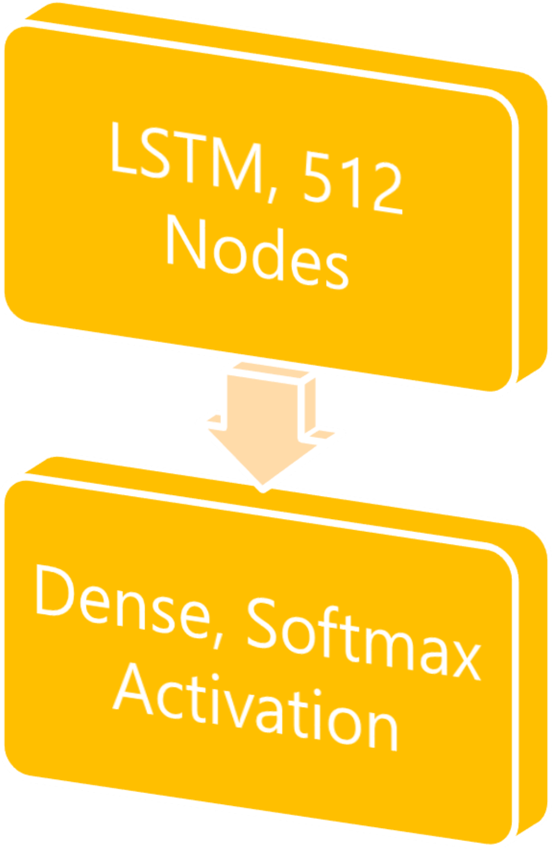


Recurrent Neural Networks – Architecture VI (light)

Description & Results

- Simple 1 layer Long-Short Term Memory-RNN with 512 nodes. Output layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.808608

Architecture

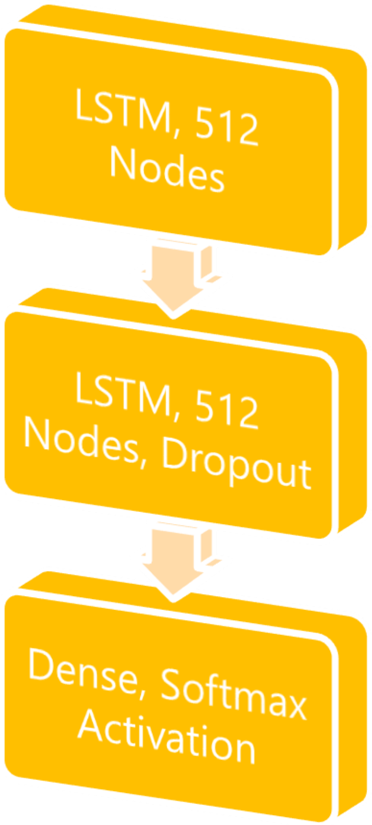


Recurrent Neural Networks – Architecture VI

Description & Results

- 2 layer Long-Short Term Memory-RNN with 512 nodes per layer. The second layer also has 0.2 dropout and recurrent dropout. Output layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs
- Best-on-Validation Accuracy:
  0.801282

Architecture

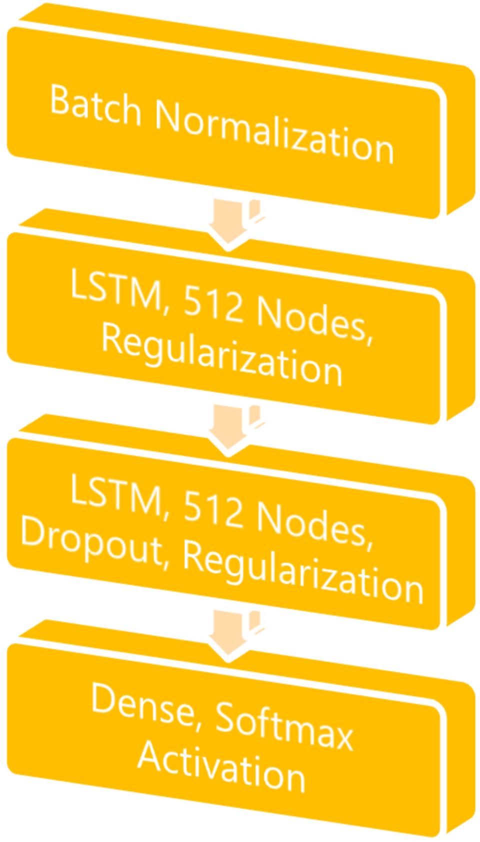


Recurrent Neural Networks – Architecture VII

Description & Results

- 2 layer Long-Short Term Memory-RNN with 512 nodes per layer. The second layer also has 0.2 dropout and recurrent dropout. Additionally layers have 0.01 L2 kernel-, bias- and recurrent regularization. Inputs are batch normalized. Ouput layer with softmax activation.
- Best Features for this Architecture:
  40 MFCCs (small samplesize due to bad results)
- Best-on-Validation Accuracy:
  0.13461 (Baseline!)

Architecture



Results

- Overview
- Model Selection
- Evaluation on Test Data

Results – Complete Overview

Feed-Forward Neural Networks

	Models	Validation_Accuracy
14	mlp_3_all	0.946886
15	mlp_4_all	0.945971
8	mlp_3_mfcc80	0.929487
9	mlp_4_mfcc80	0.928571
3	mlp_4_mfcc40	0.914835
16	mlp_5_all	0.913004
13	mlp_2_all	0.912088
2	mlp_3_mfcc40	0.911172
12	mlp_1_all	0.907509
0	mlp_1_mfcc40	0.905678
7	mlp_2_mfcc80	0.905678
4	mlp_5_mfcc40	0.899267
10	mlp_5_mfcc80	0.892857
6	mlp_1_mfcc80	0.892857
1	mlp_2_mfcc40	0.883700
17	mlp_6_all	0.874542
11	mlp_6_mfcc80	0.814103
5	mlp_6_mfcc40	0.751832

Convolutional Neural Networks

	Models	Validation_Accuracy
9	cnn_2_all	0.946886
5	cnn_2_mfcc80	0.931319
8	cnn_1_all	0.931319
3	cnn_2_mfcc60	0.930403
1	cnn_2_mfcc40	0.923077
4	cnn_1_mfcc80	0.923077
2	cnn_1_mfcc60	0.911172
7	cnn_2_melspec	0.871795
0	cnn_1_mfcc40	0.861722
6	cnn_1_melspec	0.713370
11	cnn_2_chroma	0.652930
10	cnn_1_chroma	0.576007

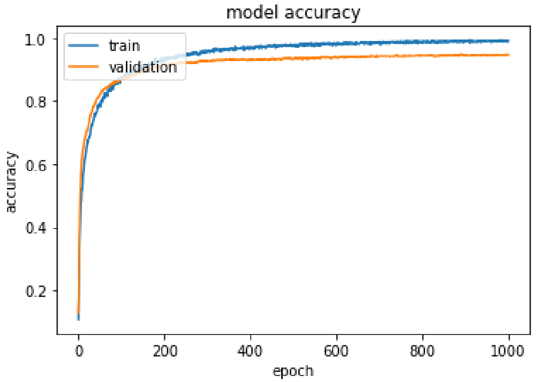
Recurrent Neural Networks

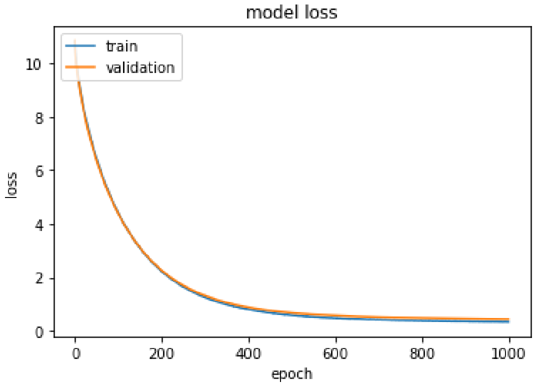
	Model	Validation_Accuracy
0	rnn_1_mfcc40	0.901099
1	rnn_2_mfcc40	0.880952
8	rnn_6light_mfcc40	0.808608
5	rnn_6_mfcc40	0.801282
13	rnn_6light_mfcc80	0.795788
14	rnn_6_mfcc80	0.770147
3	rnn_4_mfcc40	0.769231
9	rnn_1_mfcc80	0.718864
7	rnn_4light_mfcc40	0.713370
25	rnn_6light_melspec	0.618132
10	rnn_2_mfcc80	0.605311
16	rnn_2_chroma	0.581502
12	rnn_4_mfcc80	0.579670
11	rnn_4light_mfcc80	0.578755
15	rnn_1_chroma	0.541209
26	rnn_6_melspec	0.509158
17	rnn_4light_chroma	0.359890
18	rnn_4_chroma	0.355311
19	rnn_6light_chroma	0.348901
20	rnn_6_chroma	0.334249
2	rnn_3_mfcc40	0.326007
22	rnn_2_melspec	0.271978
24	rnn_4_melspec	0.267399
23	rnn_4light_melspec	0.226190
21	rnn_1_melspec	0.187729
6	rnn_7_mfcc40	0.134615
4	rnn_5_mfcc40	0.134615

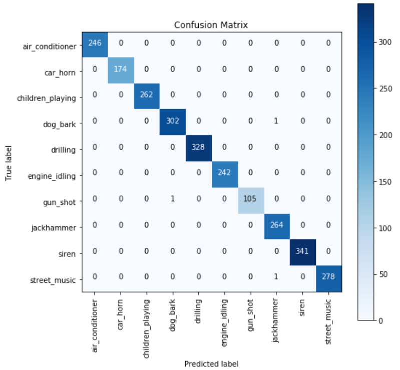
Model Selection – Best-on-Validation

- Two models with ***0.946886 accuracy*** on validation set:
  - Feed-Forward Neural Network – Architecture III
  - Convolutional Neural Network – Architecture II
- **Which one to choose?**
  The Convolutional Neural Network has slightly better training accuracy (about 1% improvement), however going by Occam's razor, the Feed-Forward Neural Network was chosen as it has a simpler network architecture and therefore generalization should be better!
- **To recap – Feed-Forward Neural Network III:**
  2 layer network with 512 nodes per layer and ReLU activation functions. Additionally layers have 0.3 dropout and 0.01 L2 kernel- and bias regularization. Inputs are batch normalized. Output layer with softmax activation.
  Input: All features.

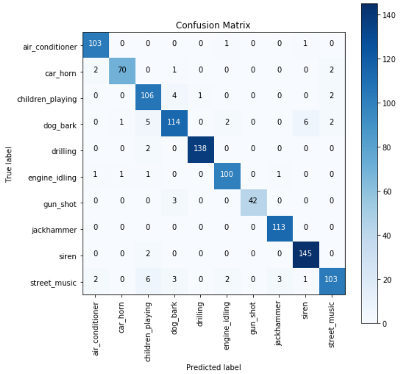
- **Feed-Forward Neural Network III:**
  - Training Accuracy: ***0.998821***
  - Validation Accuracy: ***0.946886***







Training



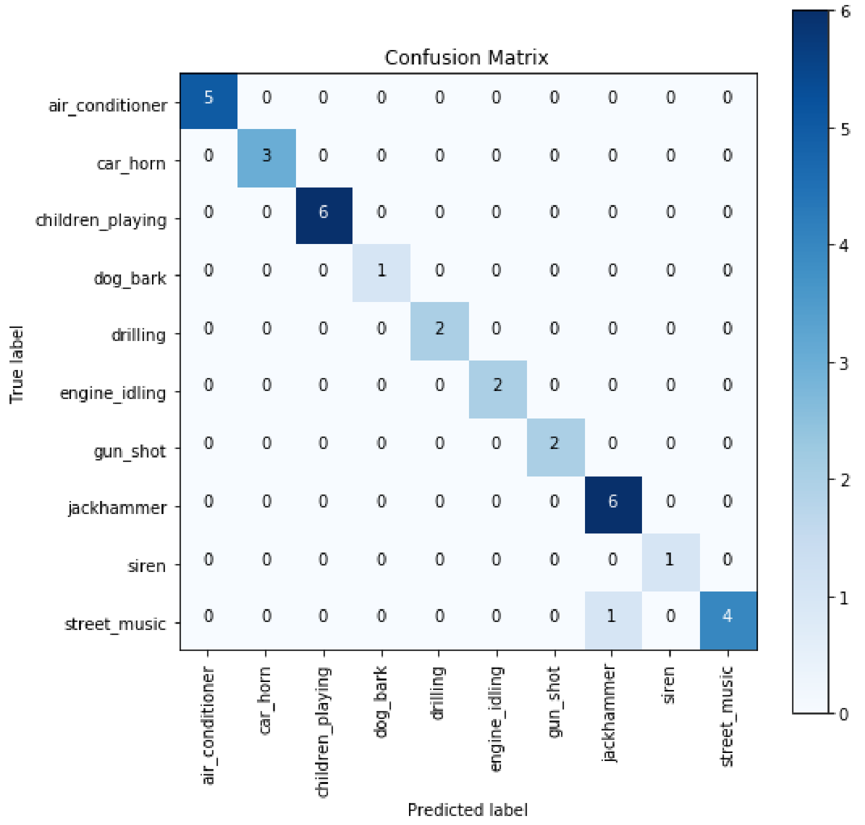
Validation

Model Evaluation – Test Data

Performance

- 33 samples predicted.
- 32 of the predictions are correct.
- Test Accuracy: ***0.969697***

Confusion Matrix



Conclusion

- Discussion
- Take-aways
- Summary

Discussion & Conclusion

Discussion

- Both Feed-Forward and Convolutional architectures performed best when they had all features available (MFCCs, chromagram, melspectrogram, spectral contrast and tonnetz), closely followed by 80 MFCCs.
- Batch normalization, dropout and regularization improved validation performance for both network types.
- Recurrent Neural Network architectures worked best when using only 40 MFCCs as features.
- Batch normalization and/or regularization does not work very well in Recurrent architectures -> after some research it turns out batch normalization is probably the bottleneck, as batch normalization does not consider the recurrent part of the network. Technically it could be done but not with vanilla RNNs. [1][2]

Conclusion & Take-aways

- MFCCs work very well as features for all network architectures.
- One can easily top 90% accuracy already with simple Feed-Forward Neural Networks.
- It is possible to pass the 90% accuracy mark with all network architectures.
- In FFNNs and CNNs batch normalization is a good idea.
- In RNNs batch normalization is a bad idea (if RNNs are not adapted).
- Dropout is a good idea in all network architectures as it reduces overfitting.
- Regularization is a good idea for FFNNs and CNNs.

More?

Read here!

Questions?

micha.birklbauer@gmail.com