Improved Music Feature Learning with Deep Neural Networks

Siddharth Sigtia and Simon Dixon

{sss31,simond}@qmul.ac.uk

Centre for Digital Music
Queen Mary University of London

c4dm

Queen Mary
University of London

Motivation

- Try to learn the most optimal features for a particular task and reduce dependency on hand-crafted features.
- How can we learn features for a particular task?:
  Neural nets with several hidden layers (deep neural networks).
- Can we learn features for MIR tasks with neural nets?:
  Lots of recent evidence suggests yes!

c4dm

Queen Mary University of London

Challenges with this approach?

- Optimising networks with several hidden layers is challenging.
- The error surface is highly non-linear w.r.t. parameters and the best we can do is hope to find a useful local minimum.
- The number of hyper-parameters can be quite large if we include momentum, learning rate schedules etc.
- For large networks, Stochastic Gradient Descent (SGD) can take prohibitively long to find useful minima even with unsupervised pre-training.
- In several domains (including music/audio), it is quite important to understand/interpret the learnt features. Something that is not clear with deep neural nets.

c4dm

Queen Mary
University of London

Can we do better?

- The use of neural networks for supervised learning has come full circle in some ways.
- Unsupervised pre-training is not considered to be necessary for finding good solutions.
- Gradient based optimisers starting with random parameter initialisation provide good results.
- Rectified Linear Units (ReLUs), Dropout, Hessian Free (HF) optimisation, Nesterov's Accelerated Gradient have all been applied to problems in various domains.
- The application of these new techniques to learning features for MIR tasks could provide improvements over existing methods.

c4dm

Queen Mary University of London

Problem definition

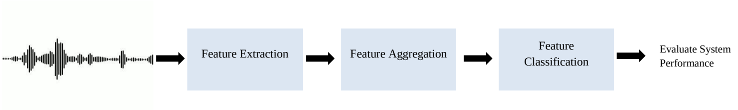


Figure: Pipeline of the genre classification system

- Learn features for a genre classification task using data from the GTZAN dataset.
- Train a classifier on the learned features and evaluate system performance.
- Inspect if features are general by using the same features on the ISMIR2004 genre dataset.

c4dm Queen Mary University of London

Contributions Of The Paper

- Evaluate the use of ReLUs as hidden units.
- Use Dropout for regularisation.
- Use HF Optimisation for training sigmoid nets and compare.

Hypothesis?

- ReLUs+Dropout eliminate the need for pre-training.
- Performance(ReLU+Dropout+SGD) >= Performance(sigmoid nets+SGD)
- More efficient training of sigmoid nets with HF.

c4dm

Queen Mary University of London

Feature Extraction

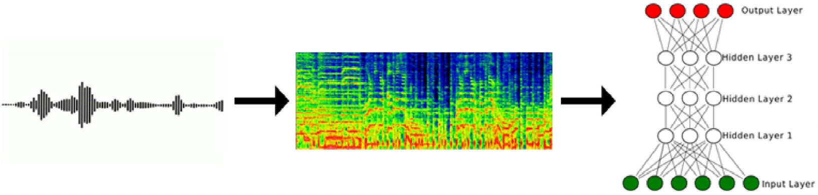
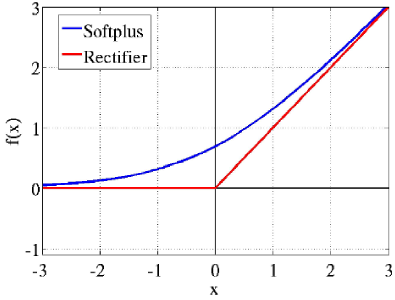
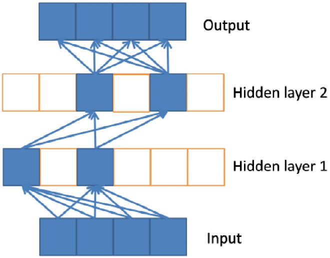


Figure: Feature extraction pipeline

Rectified Linear Units





$$\sum_{i}^{N} \sigma(x - i + 0.5) \approx \log(1 + e^{x})$$

$$\text{NReLU: } f(x) = max(0, x + \mathcal{N}(0, \sigma(x)))$$
$$\text{ReLU: } f(x) = max(0, x)$$

[1] Xavier Glorot, Antoine Bordes and Yoshua Bengio. Deep Sparse Rectifier Networks

Dropout

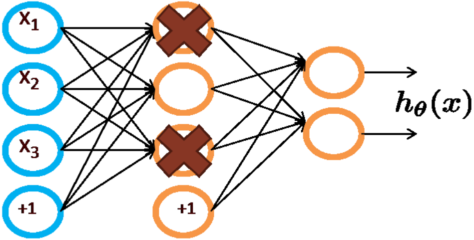


Figure: Dropout: Some of the hidden units are masked during training

Forward Propagation

$$y^l = \frac{1}{1-p} W^l (r^{l-1} * y^{l-1} + b^l)$$

c4dm

Queen Mary University of London

Useful Properties of ReLUs

- No need for supervised pre-training.
- Hard sparsity in hidden layers.
- Gradients flow easily.
- Error surface is less convoluted w.r.t parameters because of the form of the activation function.

c4dm

Queen Mary
University of London

Hessian Free Optimisation

Newton's method: $f(\theta_n + p) \approx f(\theta_n) + \nabla f(\theta_n)^T p + \frac{1}{2} p^T H p$
Newton's update: $\theta_{n+1} = \theta_n - H^{-1} \nabla f(p_n)$
Quasi-newton use an approximation to the Hessian matrix H
Two main insights in HF:

- The products Hp can be easily calculated using finite derivatives.
- The linear CG algorithm can be used to optimise the quadratic objective at each step.

c4dm

Queen Mary
University of London

Datasets

Tzanetakis Dataset:

- 1000 examples, 30 seconds each, 22050 Hz.
- 10 genres.
- 4, 50/25/25 train/valid/test splits.
- Features aggregated over 5 seconds with a 2.5s overlap.

ISMIR 2004 Genre Dataset:

- 1458 examples, truncated to 30 seconds each, downsampled to 22050 Hz.
- 6 genres.
- Original test/train split.
- Features aggregated over 5 seconds with a 2.5s overlap.

Results: Tzanetakis Dataset

	Hidden Units	ReLU+SGD	ReLU+SGD+Dropout	Sigmoid + HF
50	Layer 1	75.0±1.7	76.5±1.5	78.5±2.1
	Layer 2	79.6±2.7	77.0±2.2	80.0±2.6
	Layer 3	81.3±1.8	78.0±1.0	80.8±1.1
	All	81.5±1.9	81.5±1.7	**82.1±1.7**
500	Layer 1	71.8±0.7	75.5±1.1	67.8±1.5
	Layer 2	79.5±1.9	82.5±1.8	74.0±2.6
	Layer 3	83.0±1.2	82.0±1.4	77.1±2.36
	All	82.5±2.3	**83.0±1.1**	76.0±1.0

Table: Genre classification results on the Tzanetakis dataset

c4dm

Queen Mary University of London

Results: ISMIR 2004 Dataset

Hidden Units	Layer	ReLU+SGD	ReLU+SGD+Dropout	Sigmoid + HF
50	1	70.50	68.03	68.72
	2	70.80	66.94	70.23
	3	69.13	68.03	70.50
	All	72.42	69.68	71.20
500	1	68.03	70.09	68.40
	2	71.33	72.01	68.32
	3	71.46	69.41	70.37
	All	72.30	**73.46**	70.23

Table: Genre classification results on the ISMIR 2004 dataset

c4dm

Queen Mary
University of London

Observations

- Best accuracy is achieved with large hidden layers, ReLUs and Dropout.
- Classification accuracy is comparable to current state of the art with hand-crafted features.
- Dropout does not work well for network with small number of hidden units.
- HF achieves comparable results, though the ReLU performs better with large hidden layers.
- The same features when used for the ISMIR dataset outperform MFCCs and PMSC features.

c4dm

Queen Mary
University of London

Conclusions

- ReLUs + SGD + Dropout learn good features are as effective as the state-of-the-art hand-crafted features.
- ReLU network learn good solutions without any pre-training.
- HF is another attractive method for training neural nets.
- The features learnt are general and can be reused for other tasks.

c4dm

[1] Code:http://www.eecs.qmul.ac.uk/s̃ss31/

Queen Mary University of London