Are deep neural networks really learning relevant features?

Corey Kereliuk
DTU Compute
Technical University of Denmark
cmke@dtu.dk

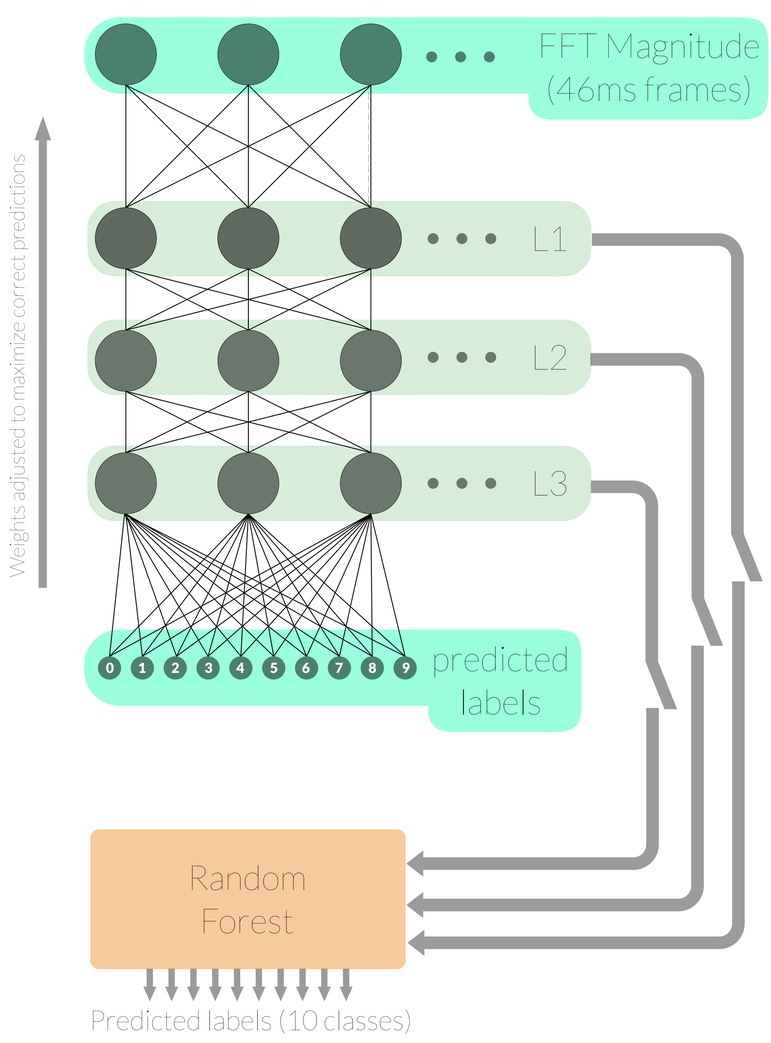
Bob L. Sturm
Audio Analysis Lab
Aalborg University
bst@create.aau.dk

Jan Larsen
DTU Compute
Technical University of Denmark
janla@dtu.dk

ABSTRACT

Deep neural networks (DNNs) are being evaluated more and more for machine learning. This is due to a variety of factors, including improvements to training algorithm efficiency, the capacity of DNNs to implicitly learn features, and their excellent performance in many domains. Two recent works [1,2] apply DNNs to content-based music information retrieval, specifically music genre recognition (MGR). The conclusions in these works are unfortunately drawn from evaluations using the GTZAN dataset, which is now known to contain faults (replicated observations and artists) that have major effects when not taken into consideration [3]. We thus re-examine the conclusions in these works considering these faults, and are led to question the degree to which the learned features are actually an improvement overstandard ``hand-crafted'' features such as Mel-frequency cepstralcoefficients (MFCCs). All our results are reproducible with the open software repository: https://github.com/coreyker/dnn-mgr

SETUP



DNN architecture

- 46 ms FFT magnitude frames (513-d) input
- 50 or 500 rectified linear units (ReLU) per hidden layer
- 3 hidden layers
- Trained using SGD with (and without) 'dropout'

Random forest

- Input L1, L2, L3, or all three layers concatenated
- Trained with (and without) temporal aggregation of frames
- 500 trees in forest

Table of GTZAN classification accuracy from [1]:

		No Aggregation		Aggregation	
Hidden Units	Layer	ReLU+SGD	ReLU+SGD+Dropout	ReLU+SGD	ReLU+SGD+Dropout
50	1	75.0 ± 1.3	75.0 ± 1.4	75.0 ± 1.7	76.5 ± 1.5
	2	75.4 ± 1.2	77.5 ± 2.2	79.6 ± 2.7	77.0 ± 2.2
	3	78.3 ± 1.1	77.0 ± 1.2	81.3 ± 1.8	78.0 ± 1.0
	All	79.0 ± 2.0	78.0 ± 1.6	81.3 ± 1.9	81.5 ± 1.7
500	1	72.7 ± 2.8	73.5 ± 1.9	71.8 ± 0.7	75.5 ± 1.1
	2	78.5 ± 1.9	78.5 ± 2.9	79.5 ± 1.9	82.5 ± 1.8
	3	80.5 ± 1.4	79.5 ± 2.6	83.0 ± 1.2	82.0 ± 1.4
	All	79.0 ± 1.4	80.5 ± 1.8	82.5 ± 2.3	83.0 ± 1.1

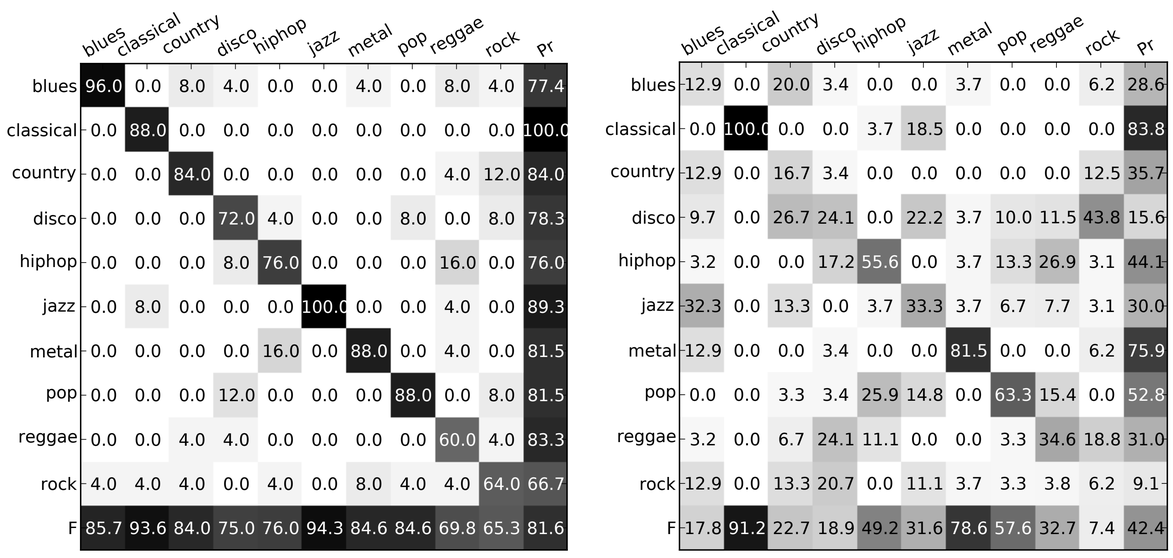
Results from our replication of systems and evaluation in [1]:

		No Aggregation		Aggregation	
Hidden Units	Layer	ReLU+SGD	ReLU+SGD+Dropout	ReLU+SGD	ReLU+SGD+Dropout
50	1	73.20	73.60	74.40	77.60
	2	74.40	73.60	79.20	77.20
	3	76.40	72.80	78.40	76.00
	All	75.20	75.60	81.60	76.80
500	1	67.20	72.80	71.60	76.40
	2	69.60	79.60	74.40	78.80
	3	72.00	81.60	76.40	80.00
	All	71.60	80.40	76.40	82.00

Results with consideration of faults in GTZAN:

		No Aggregation		Aggregation	
Hidden Units	Layer	ReLU+SGD	ReLU+SGD+Dropout	ReLU+SGD	ReLU+SGD+Dropout
50	1	40.34	40.00	39.66	39.66
	2	38.62	39.66	42.07	41.03
	3	38.28	40.00	44.48	38.62
	All	39.31	39.66	42.41	39.31
500	1	42.07	42.07	39.66	40.69
	2	41.72	43.79	41.03	45.52
	3	40.69	44.14	42.76	48.97
	All	40.69	43.79	41.03	47.59

Confusions of two systems (highlighted above) built and tested without (left) and with (right) consideration of GTZAN faults:



Conclusion:

Our evaluation of the DNN systems in [1, 2] produces dramatically lower figures of merit when considering the faults in GTZAN. In other work [3], we find Bayesian classification using standard 'hand-crafted' features (MFCCs) produces a classification accuracy of about 50% in GTZAN when considering its faults. A random forest with MFCCs produces a classification accuracy of 51%. These results thus contradict the conclusion that these DNNs are producing features better or more relevant than ``hand-crafted'' MFCCs for MGR. We are led to the question, are DNNs really learning relevant features?

[1] S. Sigtia and S. Dixon, “Improved music feature learning with deep neural networks,” ICASSP, 2014.

[2] P. Hamel and D. Eck, “Learning features from music audio with deep belief networks.” ISMIR, 2010.

[3] B. L. Sturm, “The state of the art ten years after a state of the art: Future research in music information retrieval,” J. New Music Research, 43(2), pp. 147--172, 2014.