Advances in Music Information Retrieval using Deep Learning Techniques

- Sid Pramod

Music Information Retrieval (MIR)

- Science of retrieving information from music.
- Includes tasks such as
  - Query by Example, Query by Humming
  - Music Recommendation, Automatic Playlist Generation
  - Genre Classification, Artist Classification, Instrument Classification, Chord recognition
  - Emotion recognition

Solutions

- Commonly use Machine Learning (ML)

Raw Data (Audio Signal)

⇒ Feature Extraction (representation)

⇒ Machine Learning Algorithm (e.g. SVM, kNN, etc.)

Feature Design

- Feature extraction stage usually involves hand-crafted feature design
  - ML Algorithm performance critically depends on feature design
  - Features have to be robust to noise, translations and other variations
  - Hand-crafted features are heuristic based
  - Most common features are Mel-frequency Cepstral Coefficients (MFCCs) and variants

Learning Features/Representations

- Learning representations is far less tedious than engineering features
- Context-dependent feature extraction is made possible
- Not necessarily task-specific, “Transfer learning” allows reusing features across multiple tasks

Learning Features/Representations

- Vector Quantization
  - Use simple feature extraction techniques
  - Perform clustering on all data points
  - Transformed representation is a vector where a single entry is non-zero, whose index corresponds to cluster id.

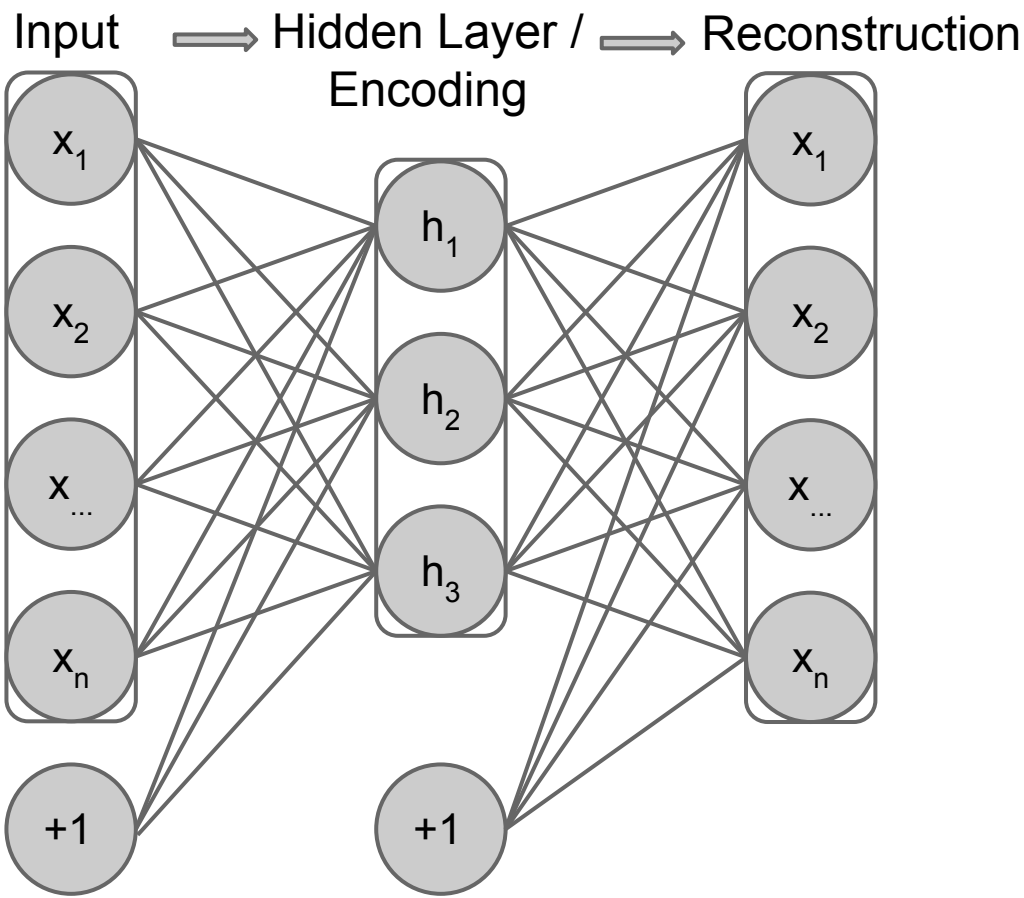
    e.g. $[x_1, x_2, x_3 \ldots x_n] \rightarrow [\ 0, 0, \ldots 1, \ldots 0]$

Learning Features/Representations

- Sparse Coding
  - A class of algorithms that learn to represent each data point as a linear combination of basis vectors (features)
  - Set of basis vectors form the “dictionary”
    e.g. if $[x_1, x_2, x_3 \ldots x_n] = a_1.f_1 + 0.f_2 + 0.f_3 + a_4.f_4 + \ldots$
    then $[x_1, x_2, x_3 \ldots x_n] \rightarrow [a_1, 0, 0, a_4, \ldots]$

Learning Features/Representations

- Autoencoder
  - A feed forward neural network with one hidden layer and same number of output nodes as input
  - Task is to reconstruct the input
  - Hidden layer “learns” a sparse encoding of the data when constraints are placed on hidden layer activations during the training procedure
  - May be combined with supervised criterion


Input
Hidden Layer / Encoding
Reconstruction
x1
x2
x...
xn
+1
h1
h2
h3
+1
x1
x2
x...
xn

Deep Architectures

- A stack of shallow transformations
  - Output of one stage serves as input to next
  - Complex transformation thus modeled as a series of simpler transformations
  - Each transformation encodes some specific variance
- Can be learned
  - Stacked auto-encoders and variants (unsupervised)
  - Deep feedforward neural networks (supervised)

Does it work?

- Humphrey et al. [2012] review initial research in the area
  - All of which achieve state of the art performance
  - Tasks include genre recognition, instrument classification, chord recognition, ...

Does it work?

- Musical Onset Detection using CNNs - Schlüter et al. [2014]

	Precision	Recall	F-score
RNN [10, 5]	0.892	0.855	0.873
CNN [1]	0.905	0.866	0.885
+ Dropout	0.909	0.871	0.890
+ Fuzziness	0.914	0.885	0.899
+ ReLU	**0.917**	**0.889**	**0.903**
SuperFlux [5]	0.883	0.793	0.836

Does it work?

- Deep content-based music recommendation - Oord et al. [2013]

Model	mAP	AUC
MLR	0.01801	0.60608
linear regression	0.02389	0.63518
MLP	0.02536	0.64611
CNN with MSE	0.05016	0.70987
CNN with WPE	0.04323	0.70101

Table 2: Results for all considered models on a subset of the dataset containing only the 9,330 most popular songs, and listening data for 20,000 users.

Model	mAP	AUC
random	0.00015	0.49935
linear regression	0.00101	0.64522
CNN with MSE	0.00672	0.77192
upper bound	0.23278	0.96070

Table 3: Results for linear regression on a bag-of-words representation of the audio signals, and a convolutional neural network trained with the MSE objective, on the full dataset (382,410 songs and 1 million users). Also shown are the scores achieved when the latent factor vectors are randomized, and when they are learned from usage data using WMF (upper bound).

References

Humphrey, Eric J., Juan Pablo Bello, and Yann LeCun. "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." *ISMIR*. 2012.

Schluter, Jan, and Sebastian Bock. "Improved musical onset detection with convolutional neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

Van den Oord, Aaron, Sander Dieleman, and Benjamin Schrauwen. "Deep content-based music recommendation." *Advances in Neural Information Processing Systems*. 2013.

Lee, Honglak, et al. "Unsupervised feature learning for audio classification using convolutional deep belief networks." *Advances in neural information processing systems*. 2009.

Hamel, Philippe, and Douglas Eck. "Learning Features from Music Audio with Deep Belief Networks." *ISMIR*. 2010.

Schluter, Jan, and Christian Osendorfer. "Music similarity estimation with the mean-covariance restricted boltzmann machine." *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE, 2011.