Tutorial

Deep Learning for Music Information Retrieval

September 23, 2018

Tutorial on Github

https://github.com/slychief/ismir2018_tutorial

or

http://tiny.cc/dlismir2018

Download the data sets linked in the README!

(prepared subset of MagnaTagaTune dataset)

Deep Learning for Music Information Retrieval

Presenters:

Thomas Lidy
Head of Machine Learning, Musimap

tom@musimap.com
www.musimap.com

Alexander Schindler
Research Scientist, AIT & TU Wien

Alexander.Schindler@ait.ac.at
http://ifs.tuwien.ac.at/~schindler

Sebastion Böck
Post Doc, OFAI & TU Wien

sebastian.boeck@tuwien.ac.at
http://ifs.tuwien.ac.at/~schindler

Tutorial Outline

I. Convolutional Neural Networks:

- Instrumental vs. Vocal
- Genre Recognition
- Mood Recognition

II. Similarity Retrieval and Representation Learning:

- Similarity Retrieval
- Siamese Networks
- Learning Music Similarity from Tags

III. Onset and Beat Detection with RNNs:

- Recurrent Neural Networks
- Onset and Beat Detection

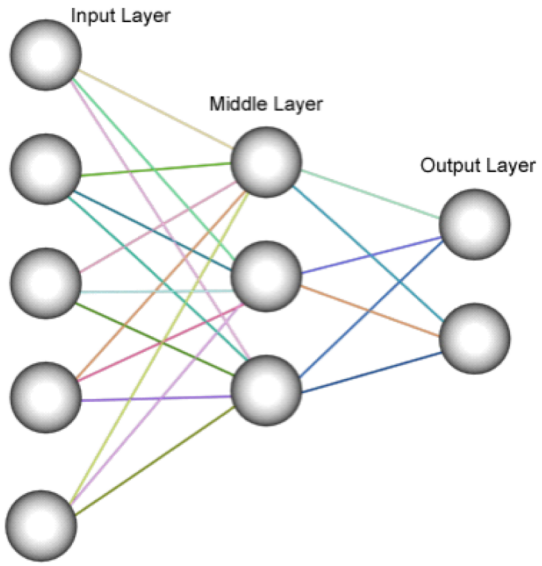
Deep Learning

= (Deep) Artificial Neural Networks (ANNs)

Neural Networks are loosely inspired by biological neurons that are interconnected and communicate with each other

Neural Networks

In reality, a neural network is a **mathematical function**:



- in which the **“neurons”** are **sets of adaptive weights**, i.e. numerical parameters that are **tuned by a learning algorithm**
- which has the capability of approximating non-linear functions of their inputs

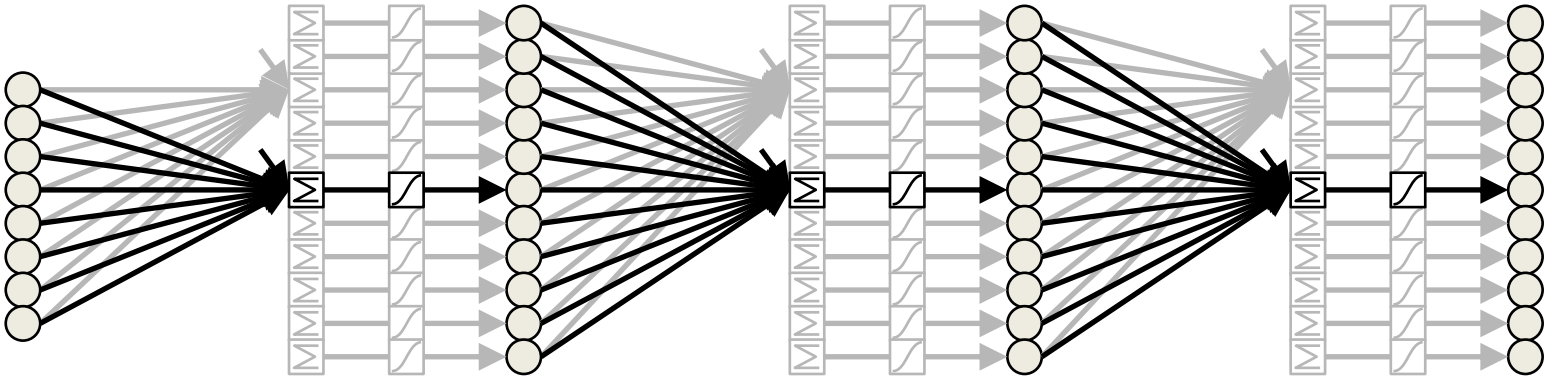
Src: Img: https://devcentral.f5.com/weblogs/images/devcentral_f5_com/weblogs/macvittie/WindowsLiveWriter/Applicationdeliveryneuralnetworksandinte_3132/ann_2.png
Text: https://en.wikipedia.org/wiki/Artificial_neural_network

What are Artificial Neural Networks?

Mathematical expressions, such as:

$$\mathbf{y} = \sigma(\mathbf{b}_3 + \mathbf{W}_3^T\sigma(\mathbf{b}_2 + \mathbf{W}_2^T\sigma(\mathbf{b}_1 + \mathbf{W}_1^T\mathbf{x})))$$

expression can be visualized as a graph:



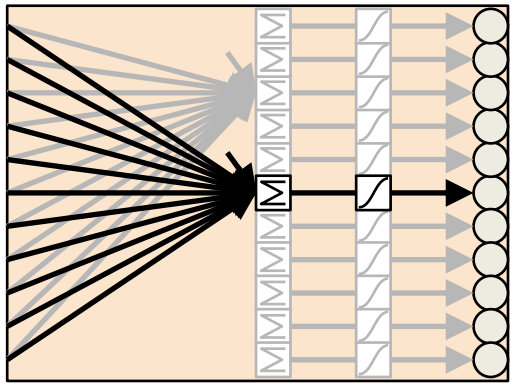
$\mathbf{x}$ $\mathbf{b}_1 + \mathbf{W}_1^T\mathbf{x}$ $\mathbf{h}_1$ $\mathbf{b}_2 + \mathbf{W}_2^T\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{b}_3 + \mathbf{W}_3^T\mathbf{h}_2$ $\mathbf{y}$

What are Artificial Neural Networks?

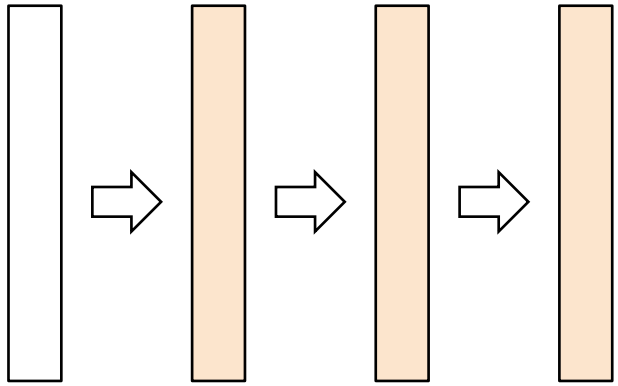
Mathematical expressions, such as:

$$f_{\mathbf{W},\mathbf{b}}(\mathbf{x}) = \sigma(\mathbf{b} + \mathbf{W}^T\mathbf{x}) \qquad \mathbf{y} = (f_{W_3,b_3} \circ f_{W_2,b_2} \circ f_{W_1,b_1})(\mathbf{x})$$

expression can be visualized as a graph – as connected layers:



“dense layer”

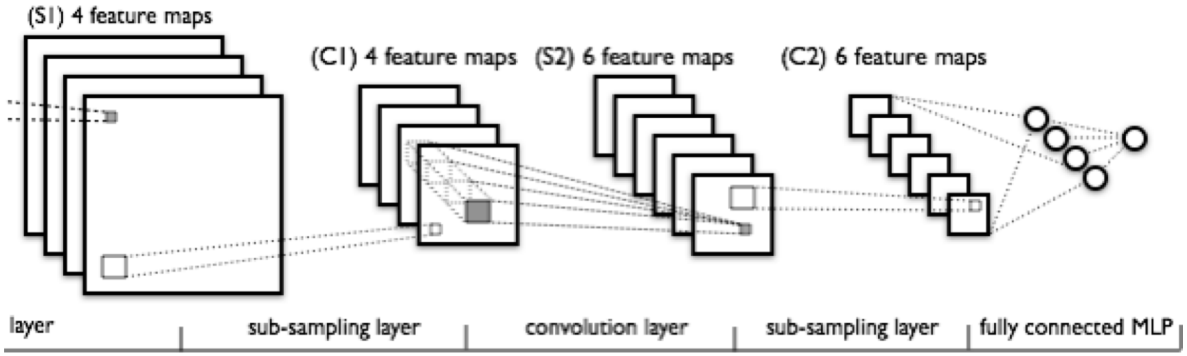


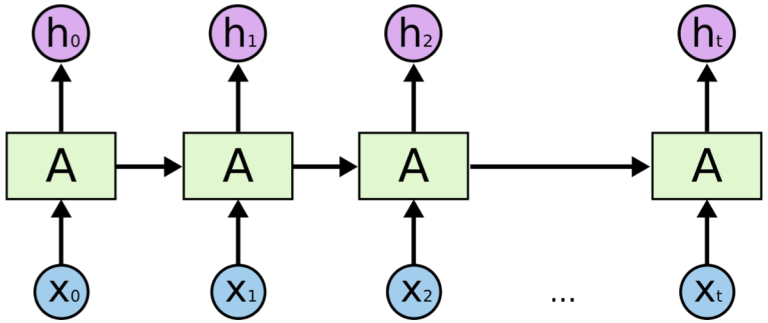
composed of simpler **functions**, commonly termed “**layers**”

Neural Network Architectures

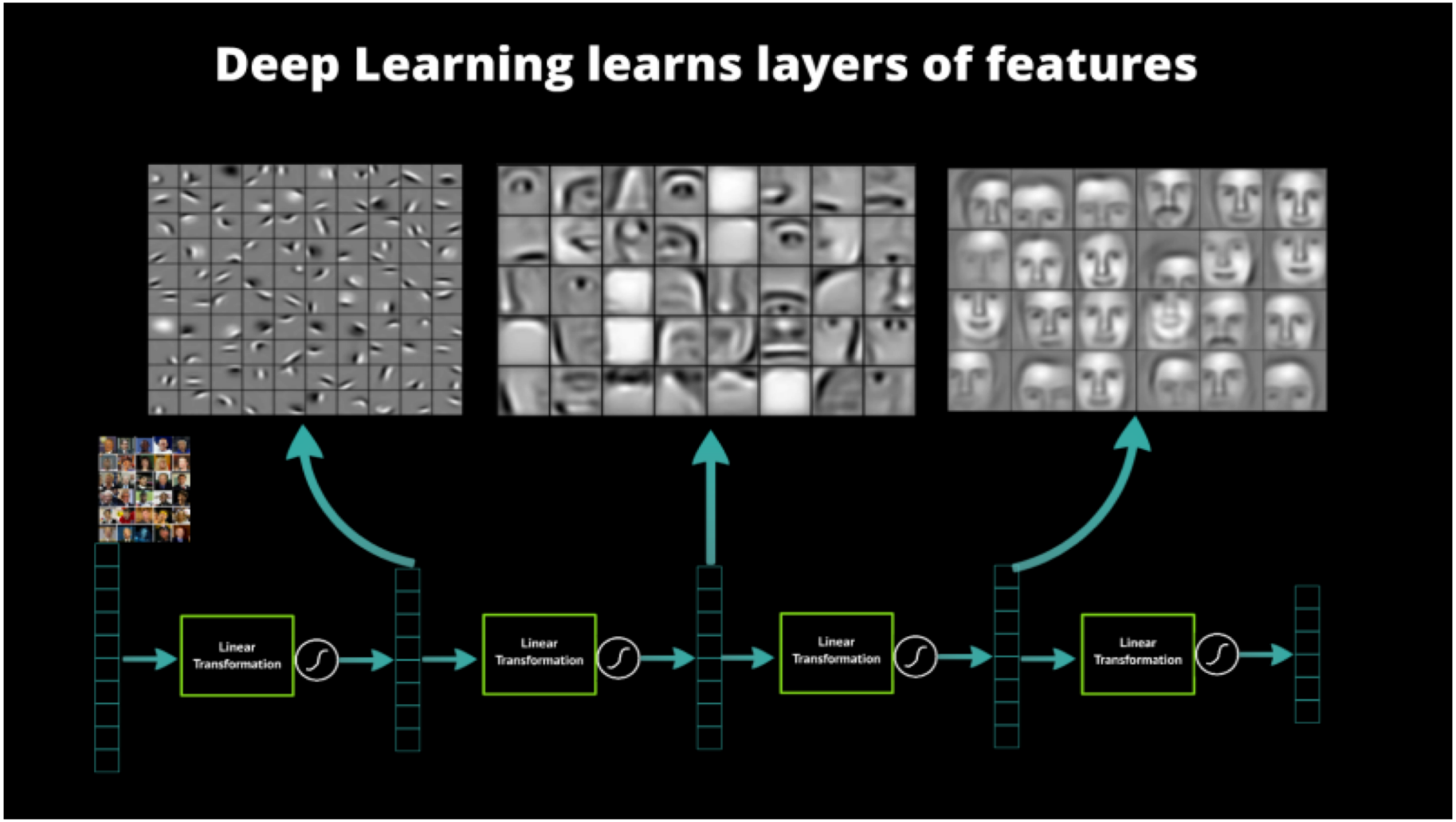
Two main Neural Network types in use today:

- **Convolutional Neural Networks** (ConvNets or CNN)
- **Recurrent Neural Networks** (RNN, LSTM, GRU)



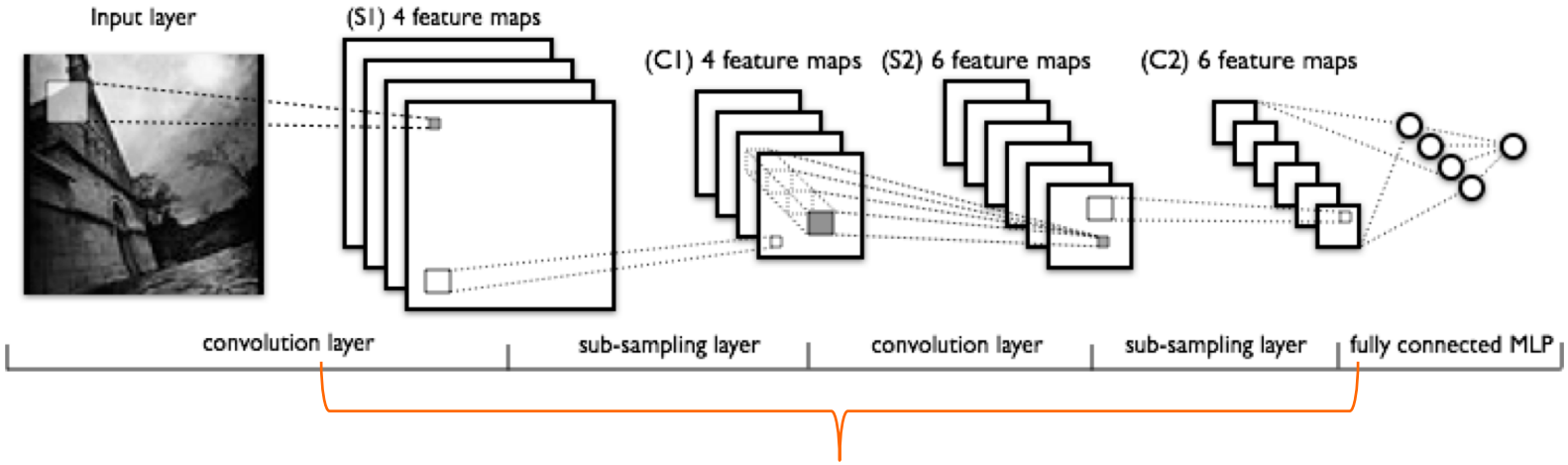


Convolutional Neural Networks (CNN)



Note: the images are conceptual here and do not represent the actual output of the neurons.

Convolutional Neural Network (CNN)



Convolution + Pooling repeat and alternate in deeper networks

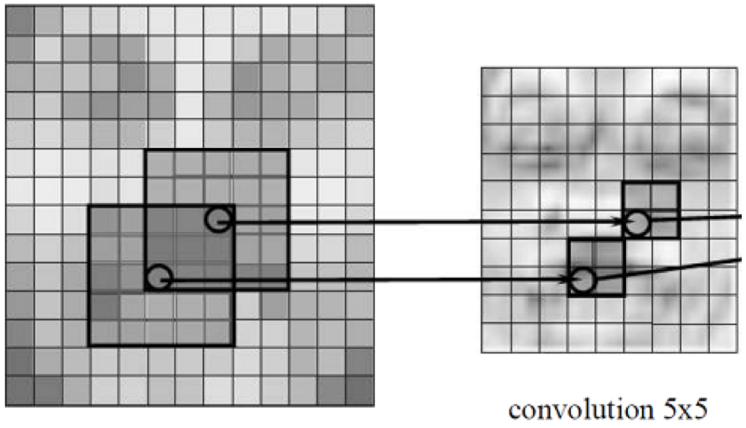
Combines three types of layers:

- **Convolutional layer:** performs 2D convolution of 2D input with multiple learned 2D kernels – learns shapes
- **Subsampling layer:** replaces 2D patches by their maximum ("max-pooling") or average ("average-pooling") – reduces resolution
- **Fully-connected layer:** computes weighted sums of its input with multiple sets of learned coefficients – maps to output

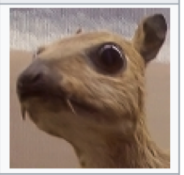
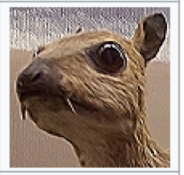
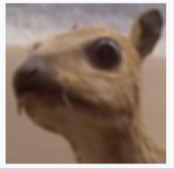
Img: http://deeplearning.net/tutorial/lenet.html

What is a Convolution?

- Apply local filter kernels and slide them over the input
- Instead of using predefined kernels, these kernels are the neurons that are learned!



convolution 5x5

Operation	Kernel	Image result
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	

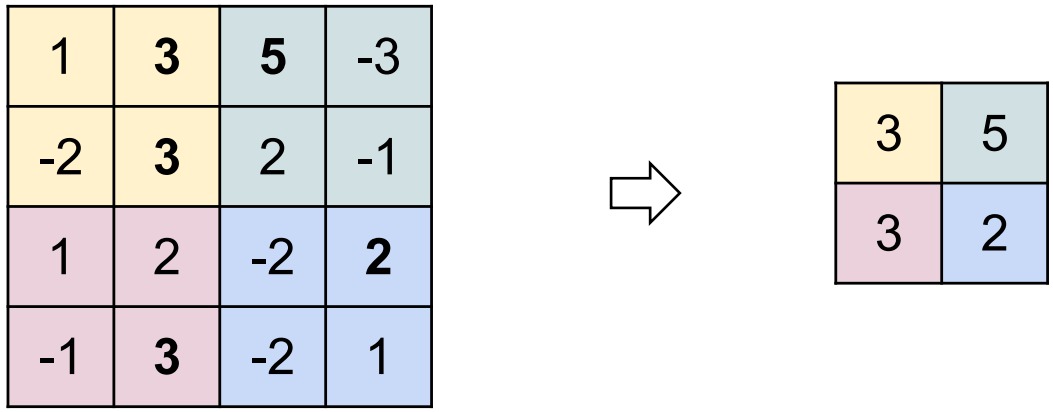
Images: http://sanghyukchun.github.io/75/
https://en.wikipedia.org/wiki/Kernel_(image_processing)

What is Pooling?

Second very important aspect of a CNN:

(also called subsampling or downsampling)

A **pooling layer** reduces the size of feature maps (i.e. output of a CNN layer and thus the input to the next layer)

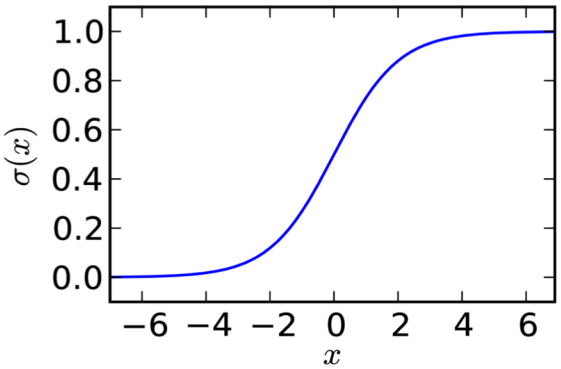


Max pooling: take the max. activation across small regions

(e.g. 2x2, as in the example above)

it can also be considered as an aggregation step

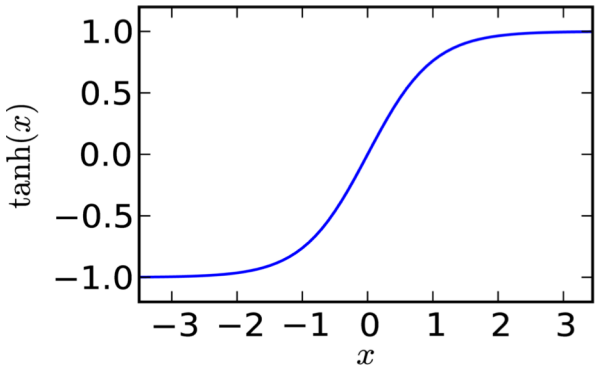
Activation Functions:
Linear Rectifier (ReLU)



Sigmoid

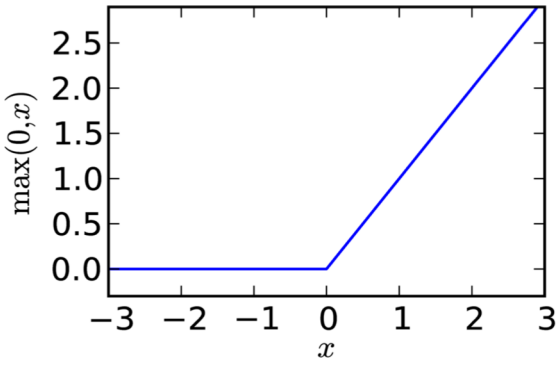
saturates for large inputs
(small slope, weak gradient!)

has nonzero mean
(slows learning)



TanH

saturates for large inputs
(small slope, weak gradient!)



ReLU

has nonzero mean
(slows learning)
has zero gradient for
negative input

Variants of ReLU: Leaky Rectifier (LReLU),
Parametric Rectifier (PReLU), ...

Benefits:
no saturation
low computational costs

ICML 2013: Rectifier Nonlinearities Improve Neural Network Acoustic Models

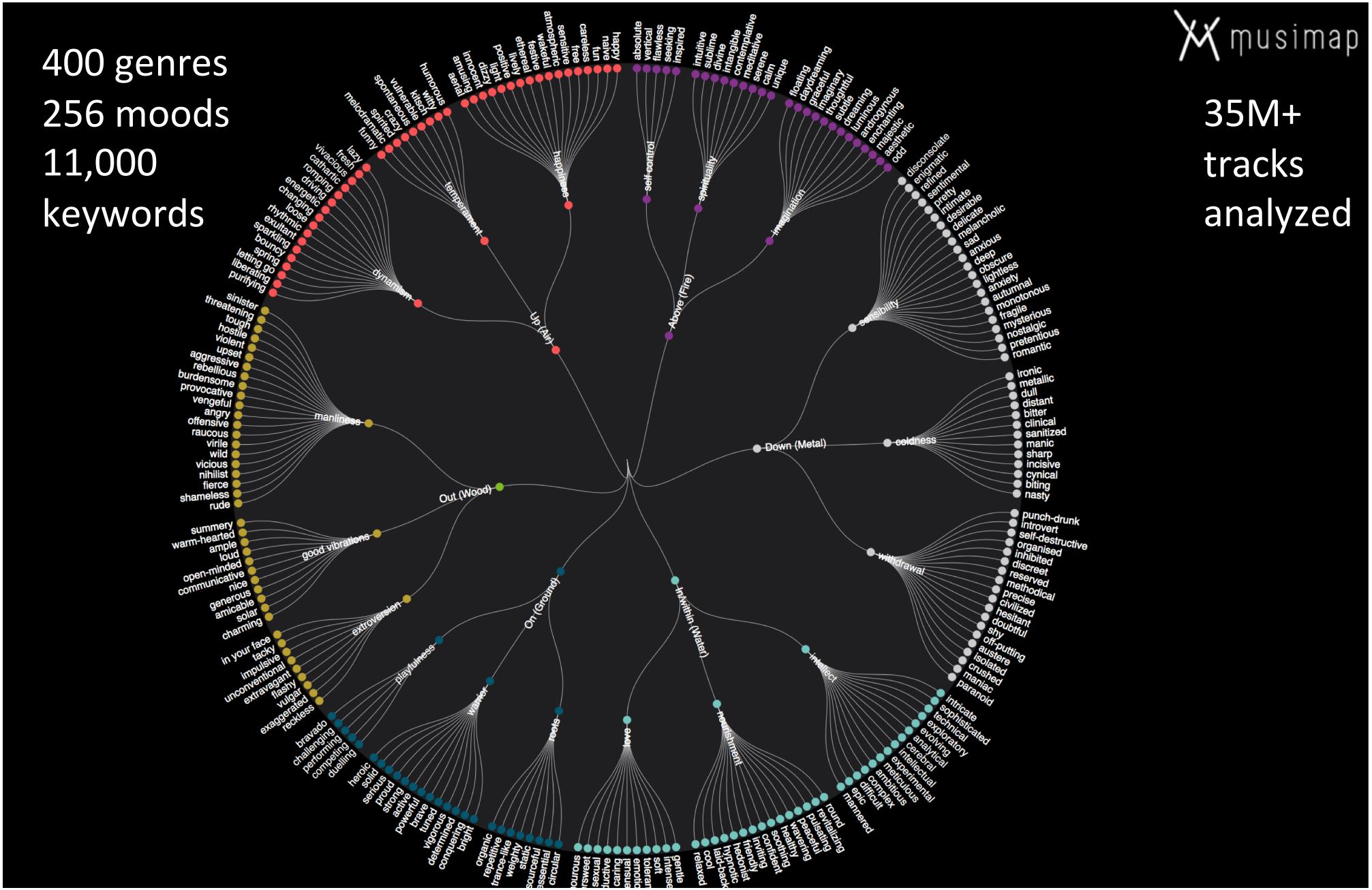
Coding Tutorial:

Jupyter Notebook

Part_1_Convolutional_Neural_Networks.ipynb

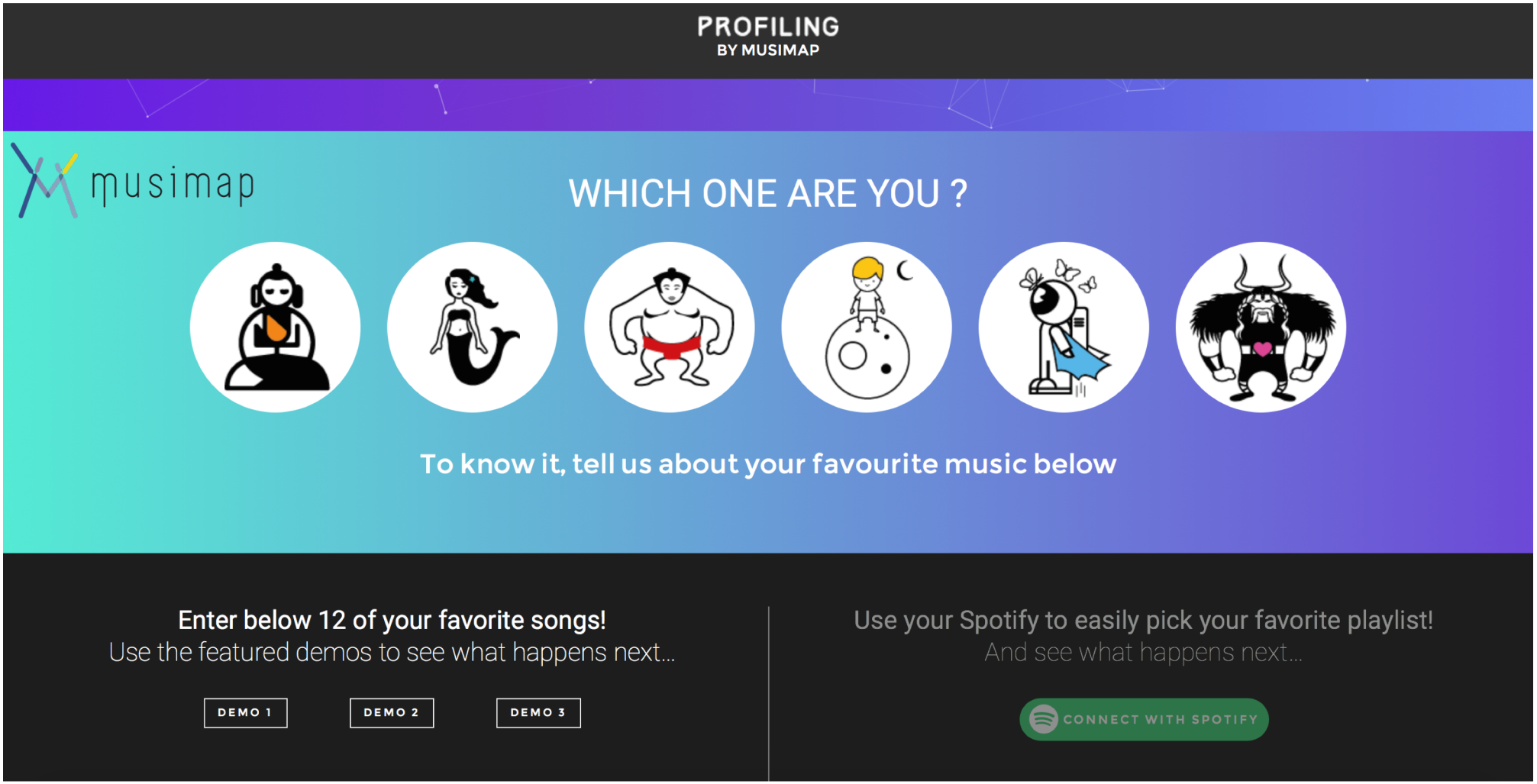
musimap

MOODS, GENRES, SUB-STYLES and more ...



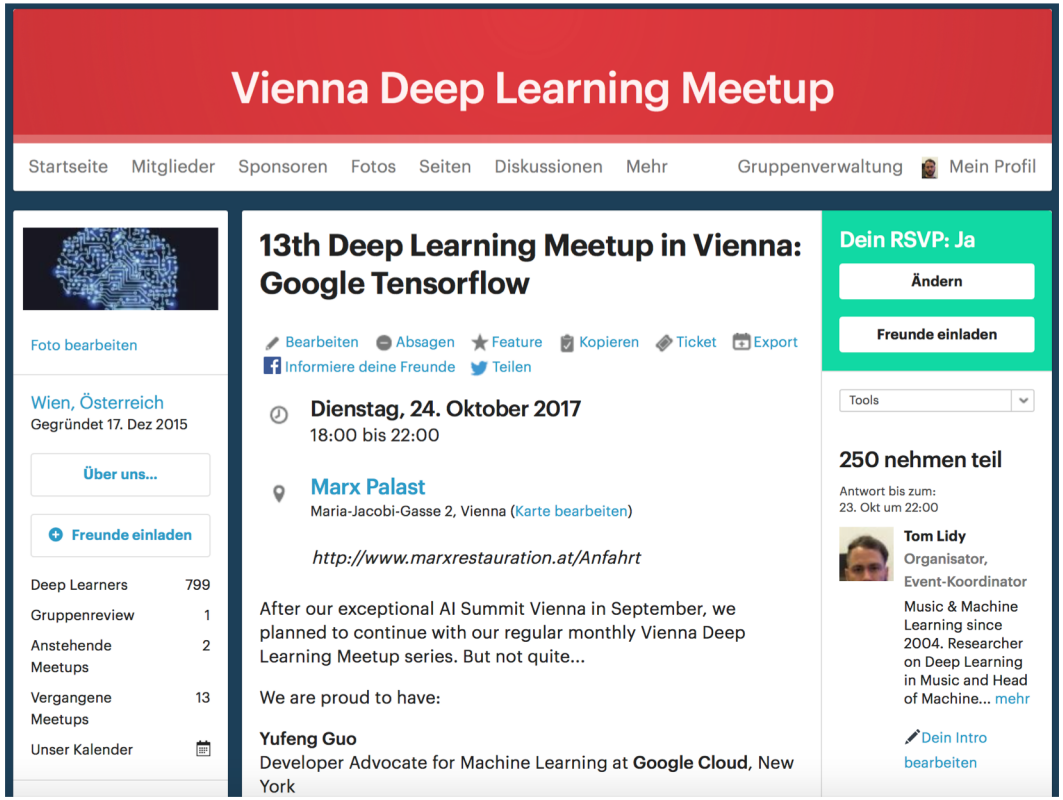
Industry Meetup: Thursday 14:00 – 18:00

Come and find out what person you are given your taste in music :-)



Organizers and Hosts: Vienna Deep Learning Meetup

1400 Members



~200 monthly participants

https://www.meetup.com/Vienna-Deep-Learning-Meetup

Slides from 20 past meetups + further resources:
https://github.com/vdlm/meetups