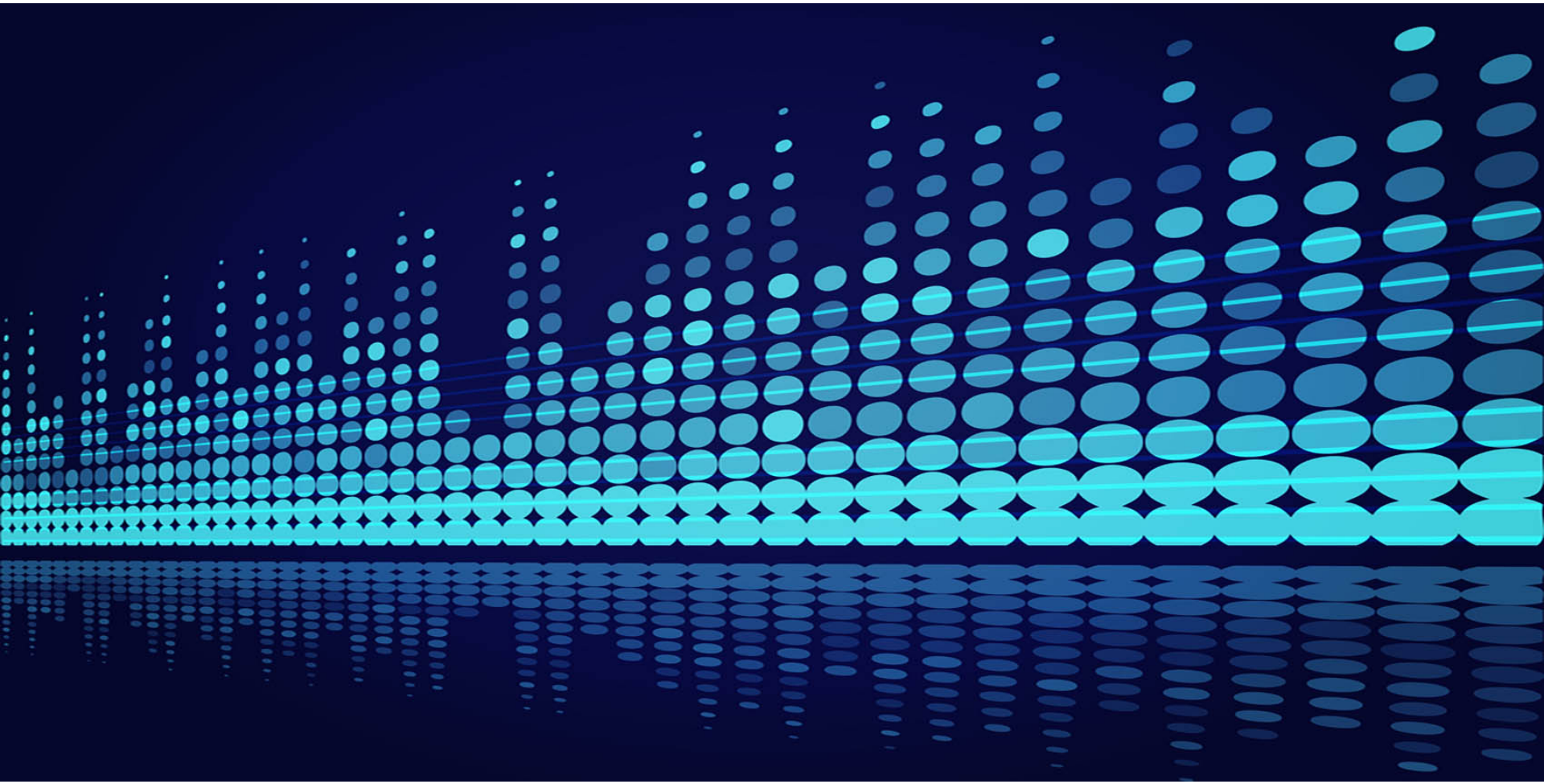
Deep Learning for Music and Audio



Thomas Lidy and Alexander Schindler

21st Vienna Deep Learning Meetup, 15 Oct 2018

Vienna Deep Learning Meetup

Deep Learning for Music and Audio

Thomas Lidy
Head of Machine Learning
Musimap

tom@musimap.com
www.musimap.com

musimap

(TU WIEN TECHNISCHE UNIVERSITÄT WIEN)

Alexander Schindler
Scientist
AIT & TUWien

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

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

Winners of 3 International Benchmarks

with Deep Learning on Audio

Thomas Lidy & Alex Schindler:

Winner IEEE DCASE 2016
Domestic Audio Tagging Contest

Winner MIREX 2015
Music/Speech Categorization

Thomas Lidy & Alex Schindler:

Winner MIREX 2016
Music Genre and Mood Classification

Outline

Intro:

- Images vs. Audio
- Use Cases and Tasks in Audio

Convolutional Neural Networks:

- How CNNs work (Layers, Filters, Pooling)
- Particularities in Music
- Application Domains in Music

Advanced Topics:

- Representation Learning with Siamese Networks

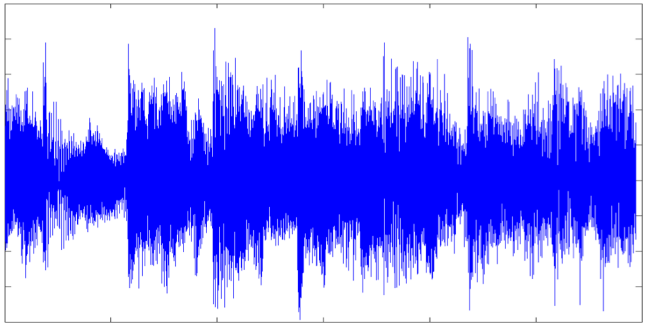
Use Cases and Tasks: Music

- Genre classification
- Mood classification
- Music recommendation
- Artist identification
- Artist similarity
- Cover song detection
- Rhythm and beat detection
- Score following
- Chord detection
- Organization of music
- Audio Fingerprinting
- Audio segmentation
- Instrument detection
- Automatic source separation
- Onset detection
- Optical music recognition
- Melody transcription …..

Use Cases and Tasks: Audio/Speech

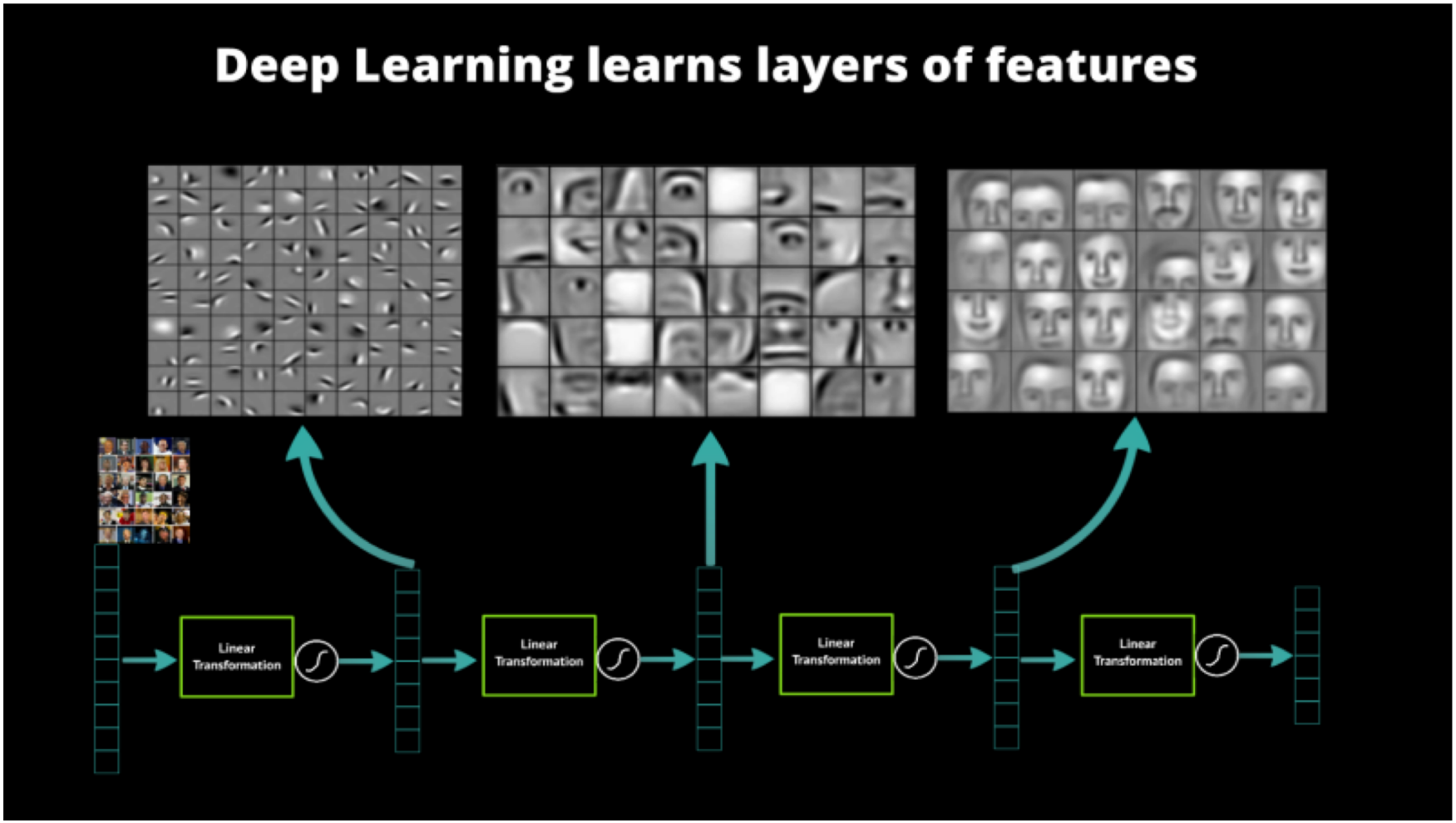
- Text to Speech
- Speech to Text
- Speech to Speech (e.g. Translation)
- Audio FX classification (e.g. instrument library)
- Audio event detection (e.g. gunshot detection, failures of industrial machines …)
- Audio scene recognition
- Audio tagging (urban sounds, domestic, …)
- Species identification (birds, whales, …)
- …

Image vs. Audio



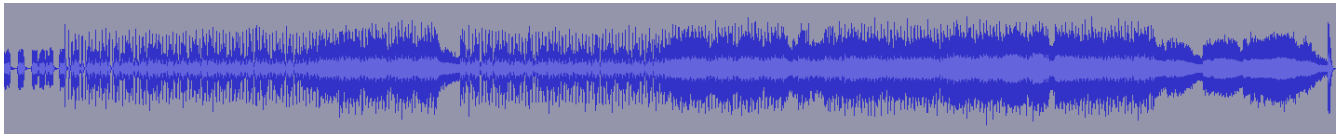
How to analyze audio with Deep Learning?

Convolutional Neural Networks (CNN)
- How to process audio?

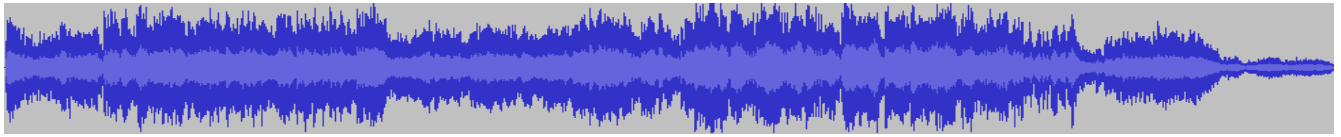


Note: the images are conceptual here and do not represent the actual output of the neurons.

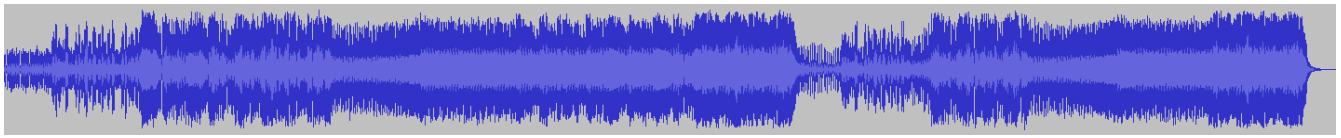
Excercise: What‘s the Genre?



AC-DC – Highway to Hell

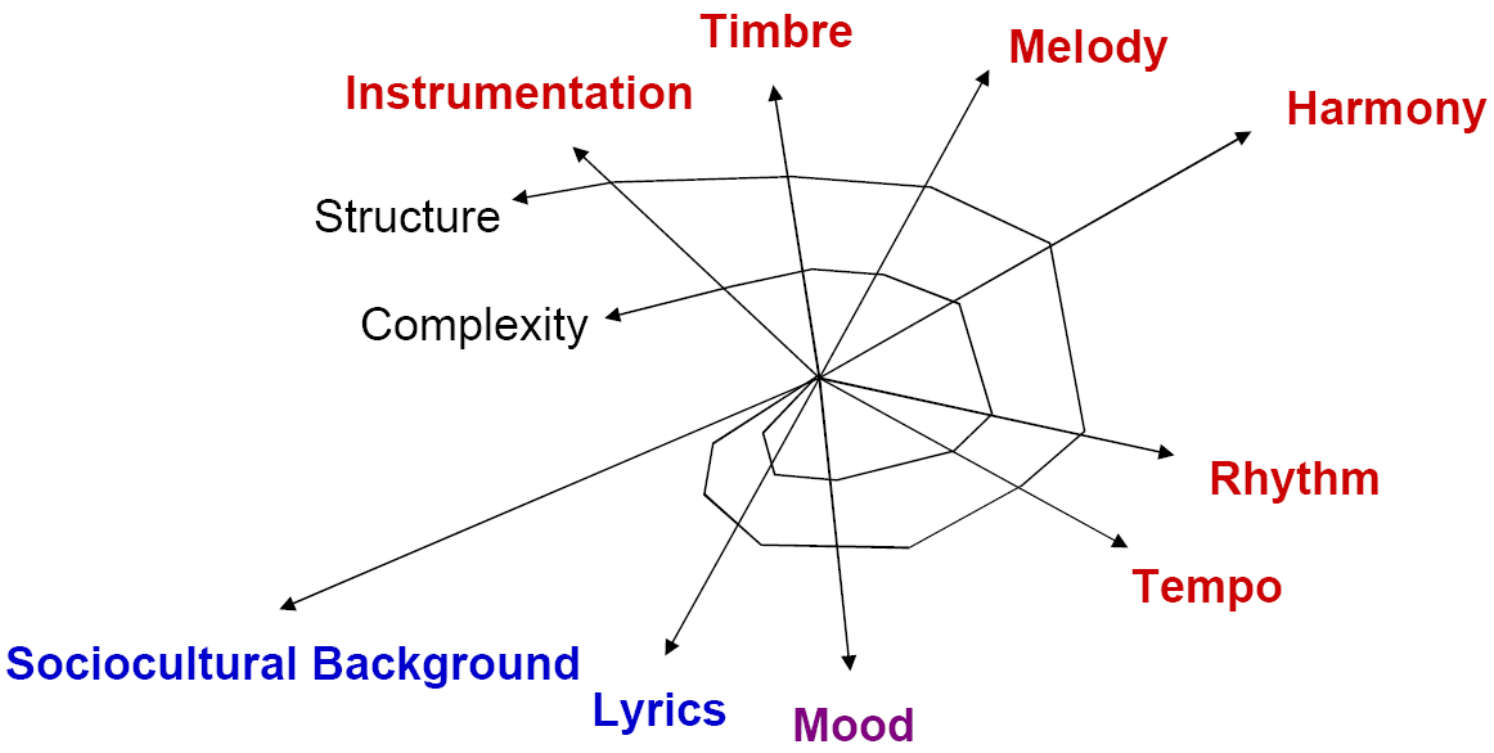


John Williams – Star Wars Main Theme



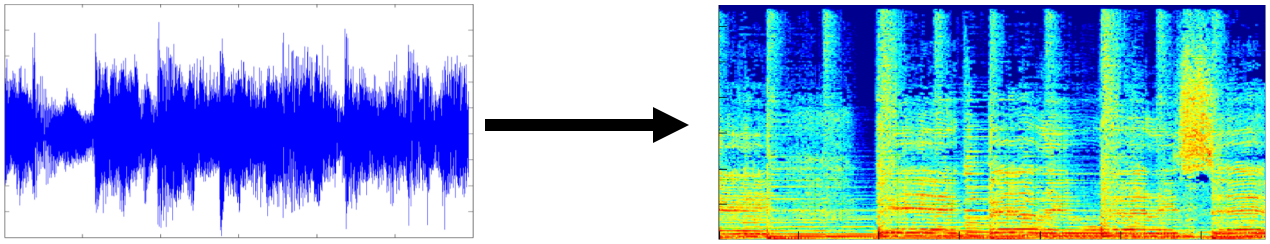
Rihanna feat. Calvin Harris – We Found Love

The Many Dimensions of Music



Signal Processing

Time-Frequency Transformation



Time Domain
(„Wave Form“)

Frequency Domain
(„Spectrum“)

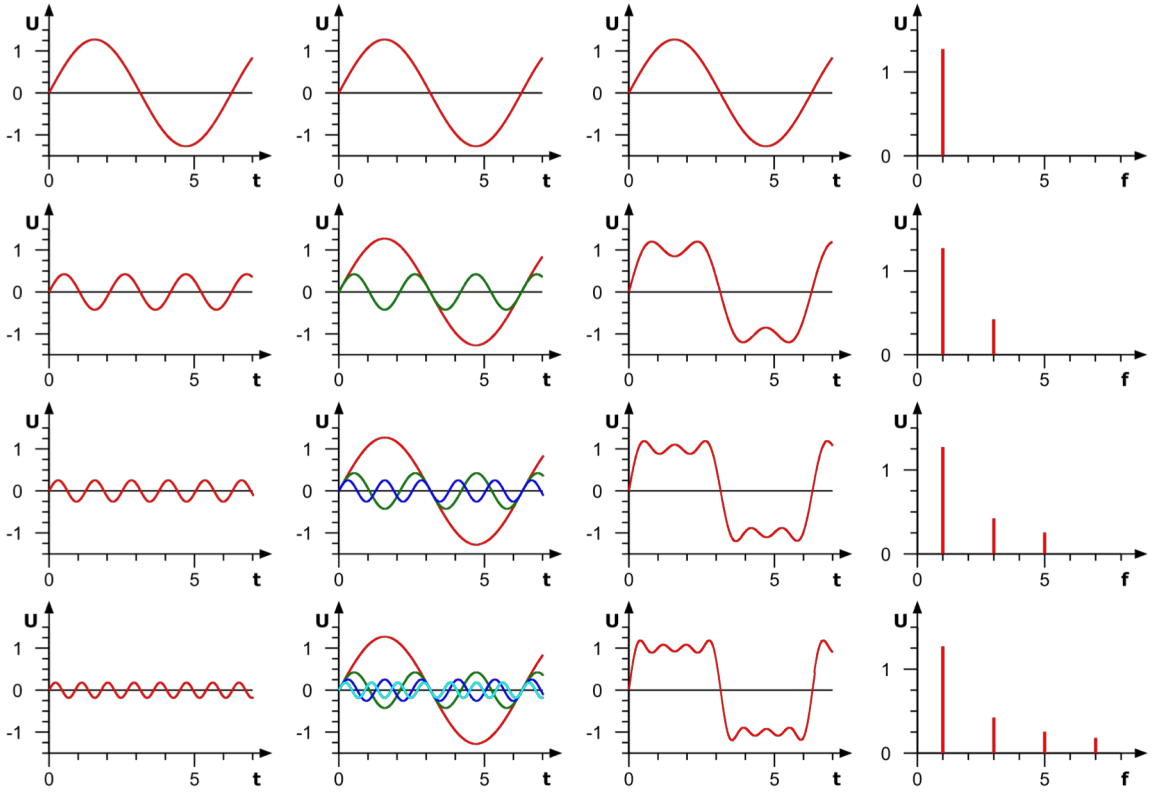
Possibilities:

Fourier Transform (FFT)

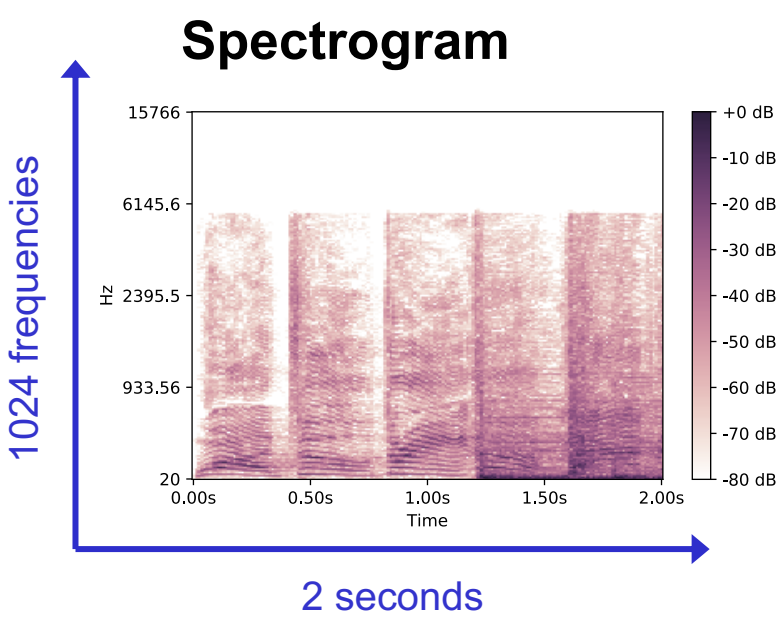
Discrete Cosine Transform (DCT)

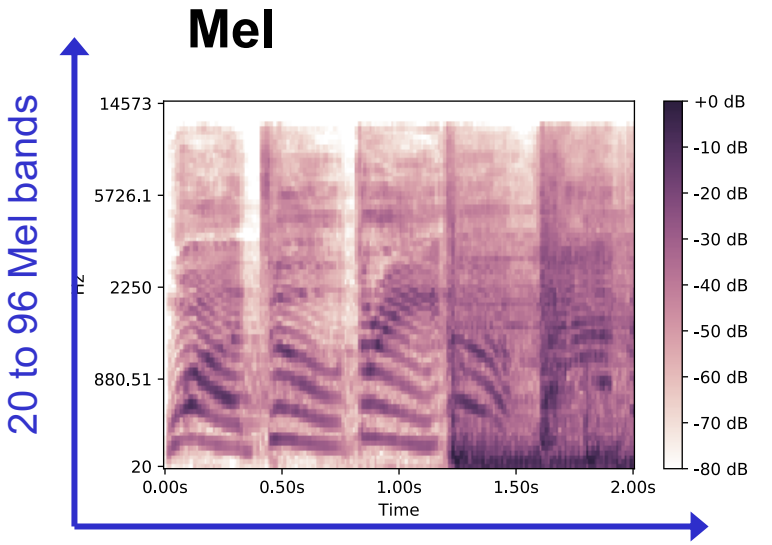
Wavelet Transform

Step 1: Fourier Transform (FFT or STFT)

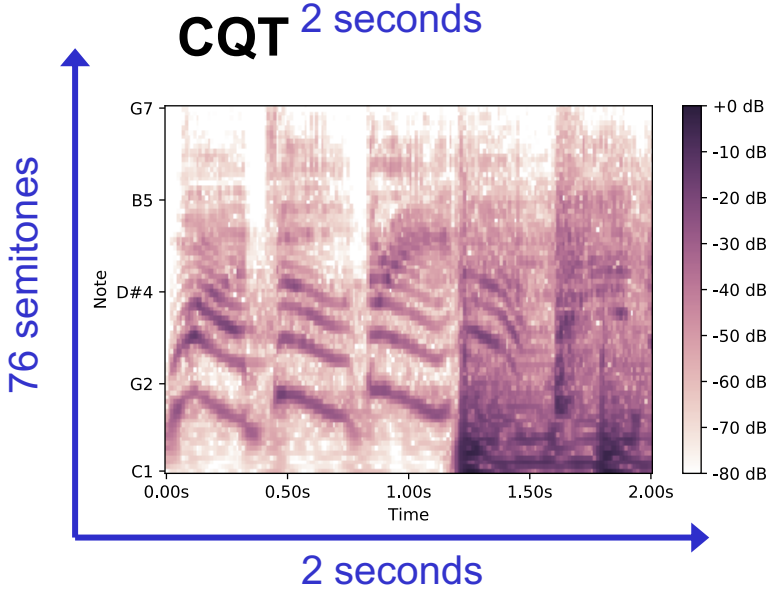


Step 2: Frequency Scale Reduction





or:



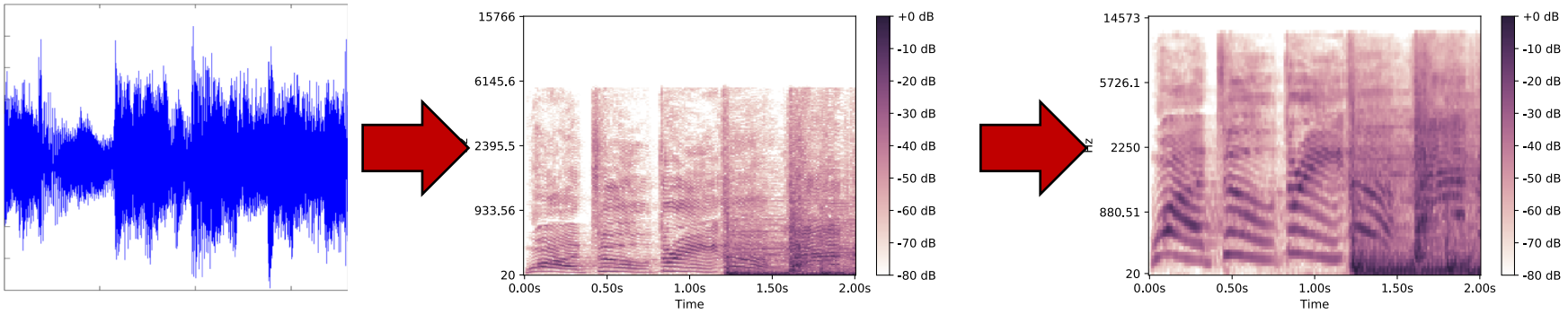
or Bark, or others …

Lidy and Schindler (DCASE 2016)
CQT-BASED CONVOLUTIONAL NEURAL NETWORKS FOR AUDIO SCENE CLASSIFICATION

Audio Preprocessing for Deep Learning

Frequently these 3 steps are done before any DL:

1) **Spectrogram (Fourier Transform)**
2) **Mel scale**
3) **Log transform**



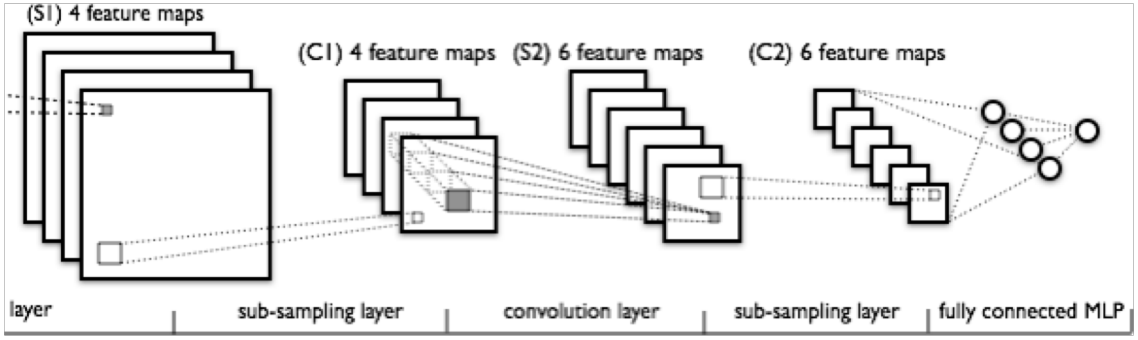
Note: end-to-end learning from Wave is showing more and more successs, but needs massive data.

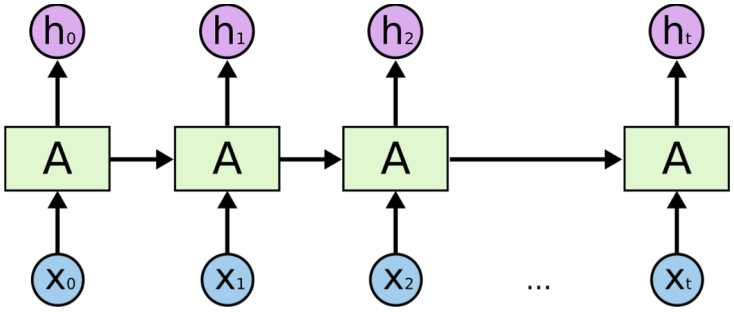
Modern
Neural Network Architectures
for Deep Learning

Neural Network Architectures

Two main Neural Network types in use today:

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





CNNs vs. RNNs

Convolutional Networks:

- input is fixed-sized tensor (e.g. an image or spectrogram)
- produce a fixed-sized vector as output (e.g. probabilities of different classes)

Recurrent Neural Networks:

- operate over **sequences** of vectors or tensors
- Example Applications: Text translation, Text to Speech, Speech to Text, etc.

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Deep Learning Architectures for Music Analysis

Convolutional Networks:

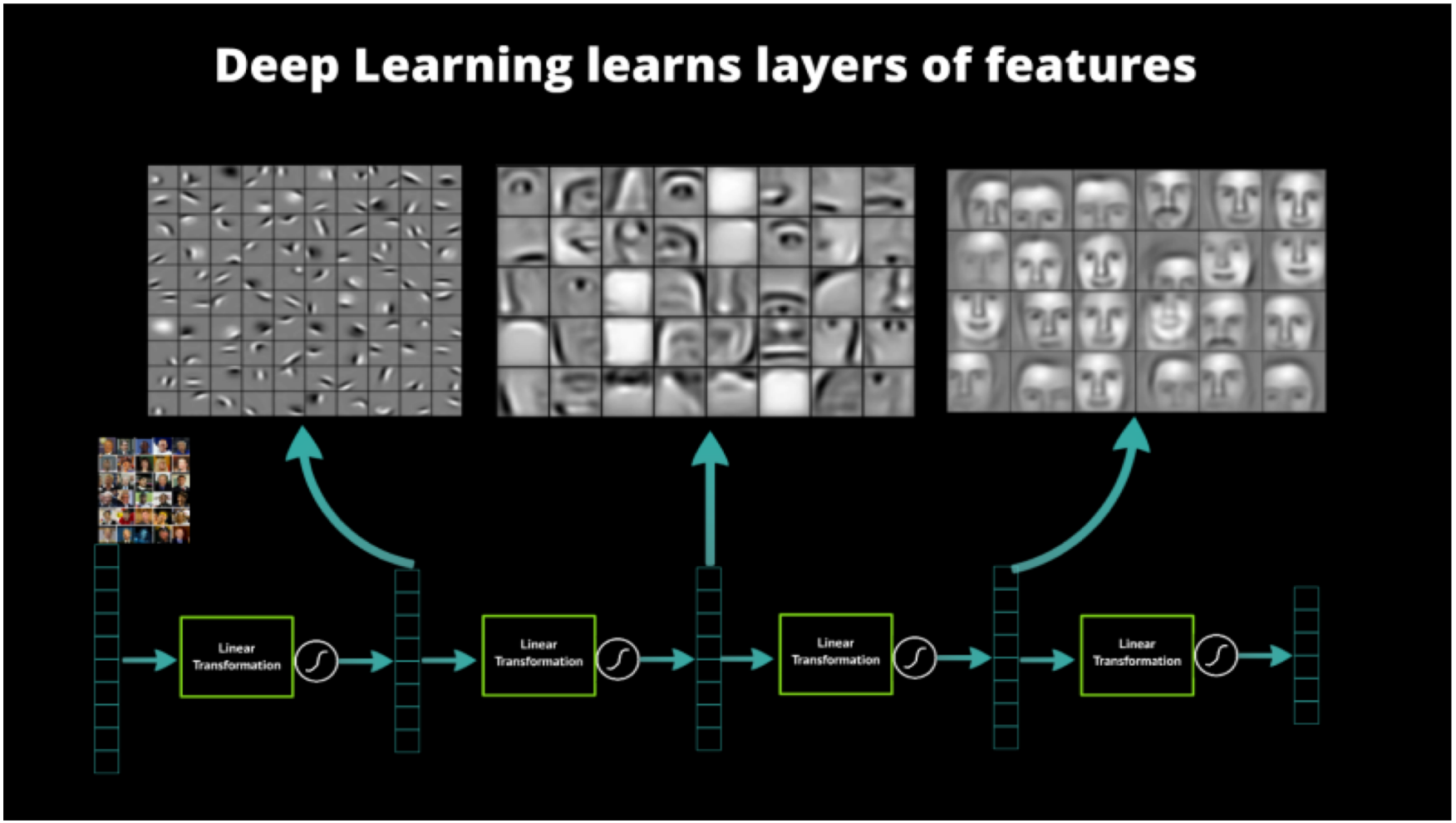
- when analyzing audio spectrogram excerpts
- when time sequence does not necessarily play important role (i.e. processing audio samples as a whole; e.g. for genre, mood recognition, …)

Recurrent Neural Networks:

- when sequence and time series are important (e.g. (genre), melody, beat onset detection)

➔ often a mix of CNNs with RNNs is used

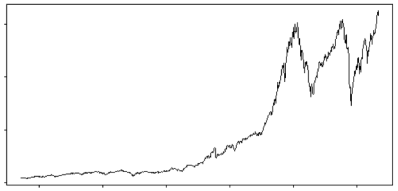
Convolutional Neural Networks (CNN)

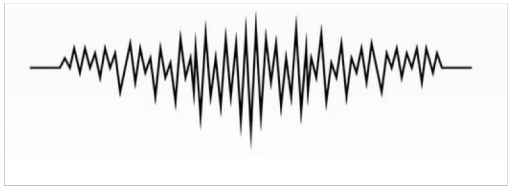


Note: the images are conceptual here and do not represent the actual output of the neurons.

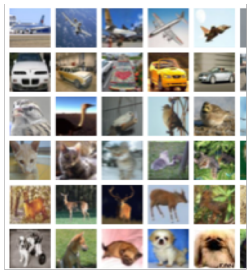
Convolutions in 1D, 2D, 3D

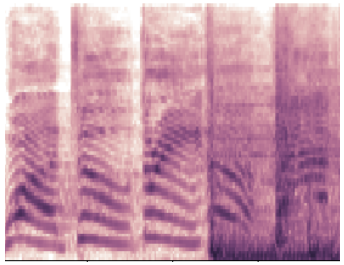
1D: time series,
audio waveforms



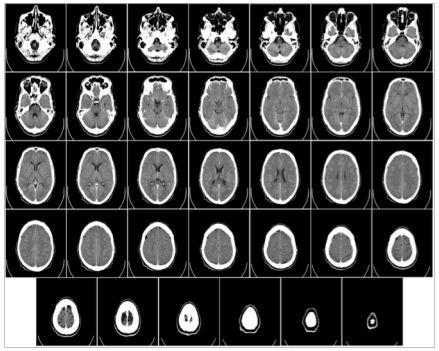


2D: images,
audio spectrograms

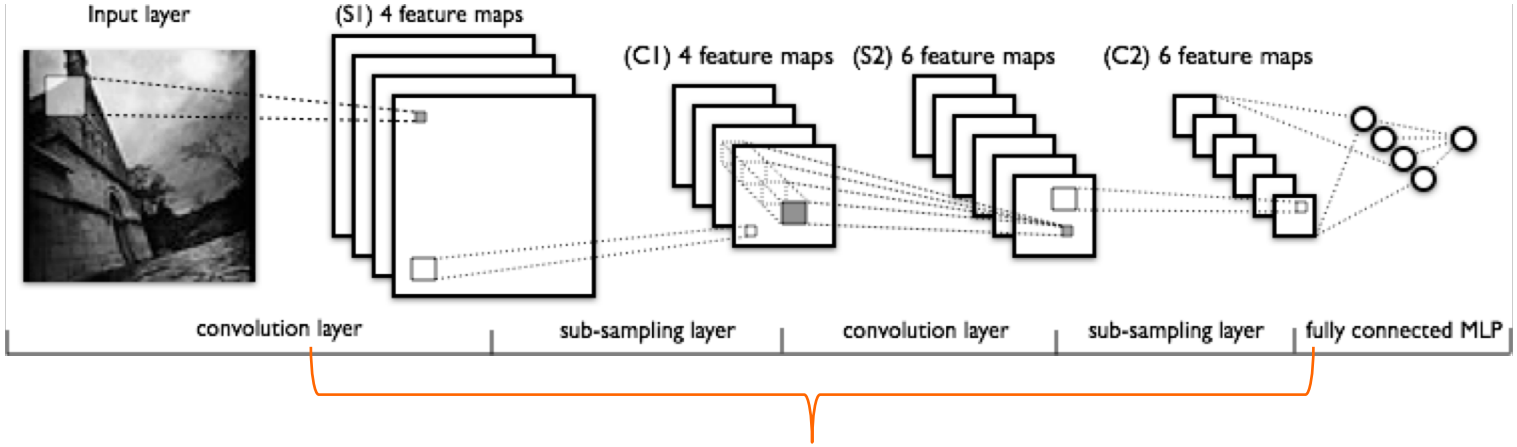




3D: volumetric data,
video



Convolutional Neural Network (CNN)



Convolution + Pooling repeat and alternate in deeper networks

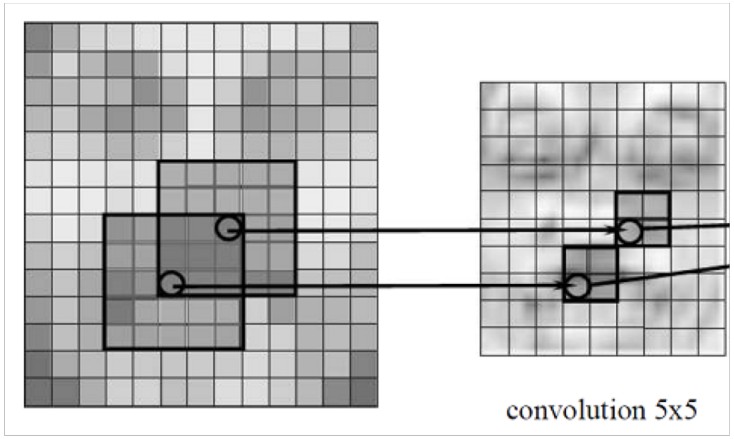
Combines three types of layers:

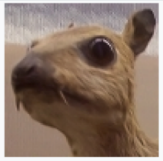
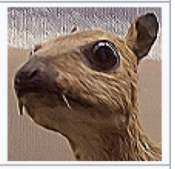
- **Convolutional layer:** performs 2D convolution of 2D input with multiple learned 2D kernels – learns shapes
- **Subsampling layer:** replaces 2D patches by their maximum ("max-pooling") or average ("average-pooling") – reduces resolution
- **Fully-connected layer:** computes weighted sums of its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



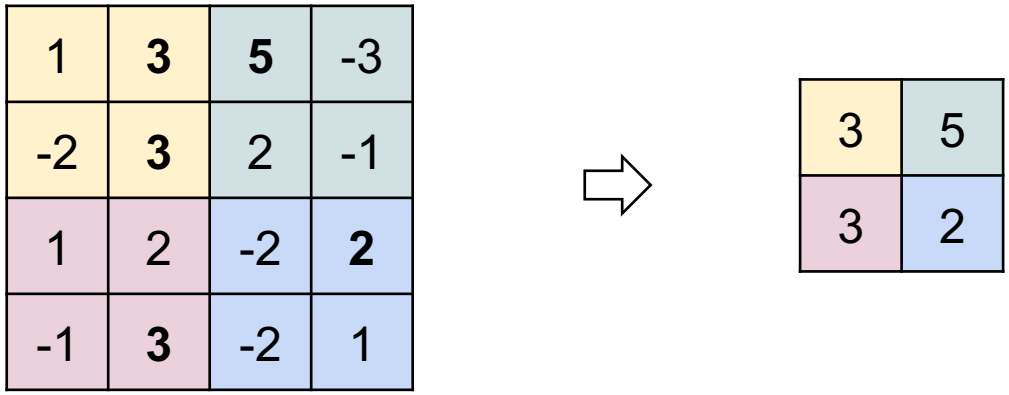
Operation	Kernel	Image result
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
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)

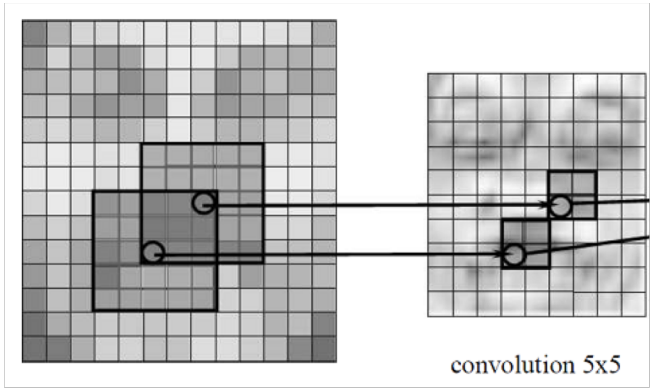
A **pooling layer** reduces the size of feature maps (i.e. output of a CNN layer and thus the input to the next layer)



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

Convolutions: Filter Kernel Sizes



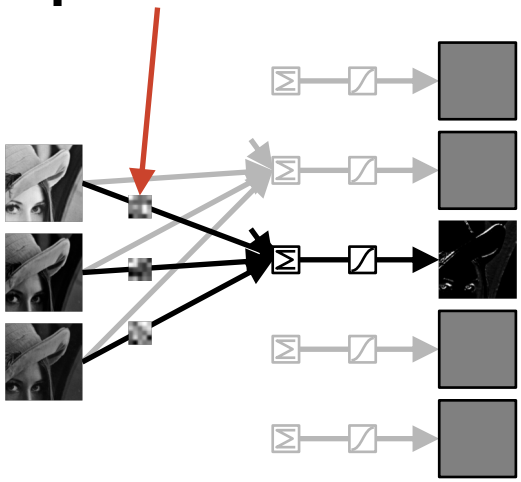
For Image Analysis: usually quadratic shape

3x3 kernel: can only see small structure

7x7 kernel: sees more, but more weights

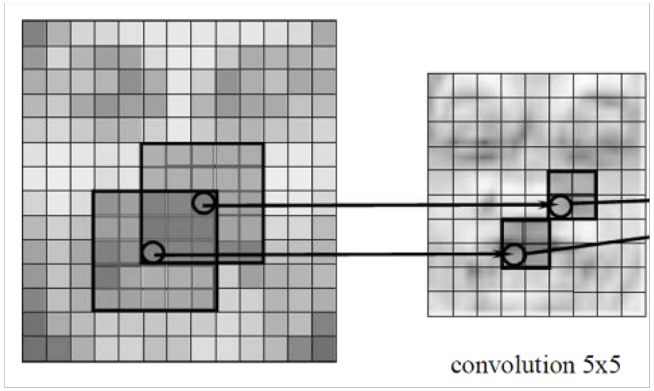
224x224 kernel: possibly sees full input

(equivalent to fully-connected network)



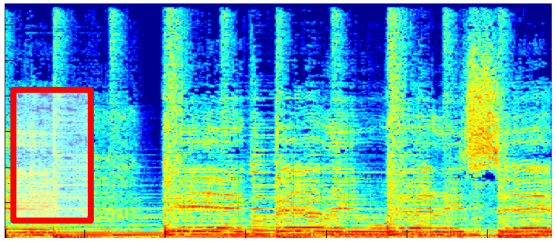
Hard to find sweetspot. But smaller kernels and deeper networks preferable.

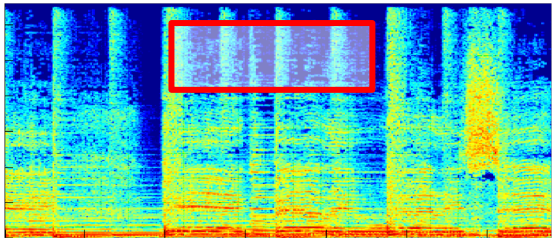
Convolutions: Filter Kernel Sizes



For Audio Analysis: rectangular shapes may help

- **Vertical: captures harmonics/timbre**
- **Horizontal: captures rhythmics**

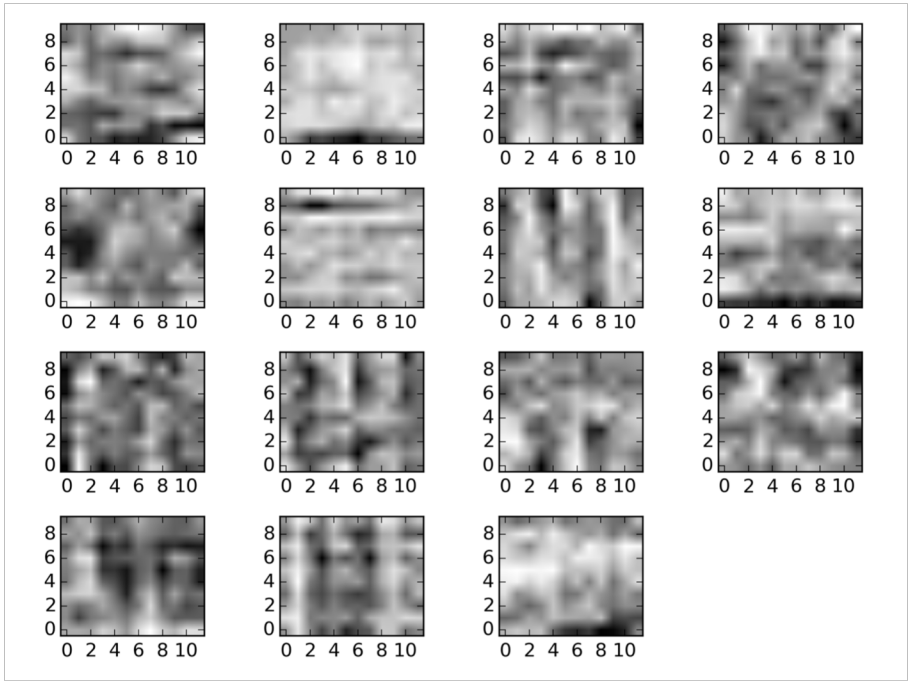




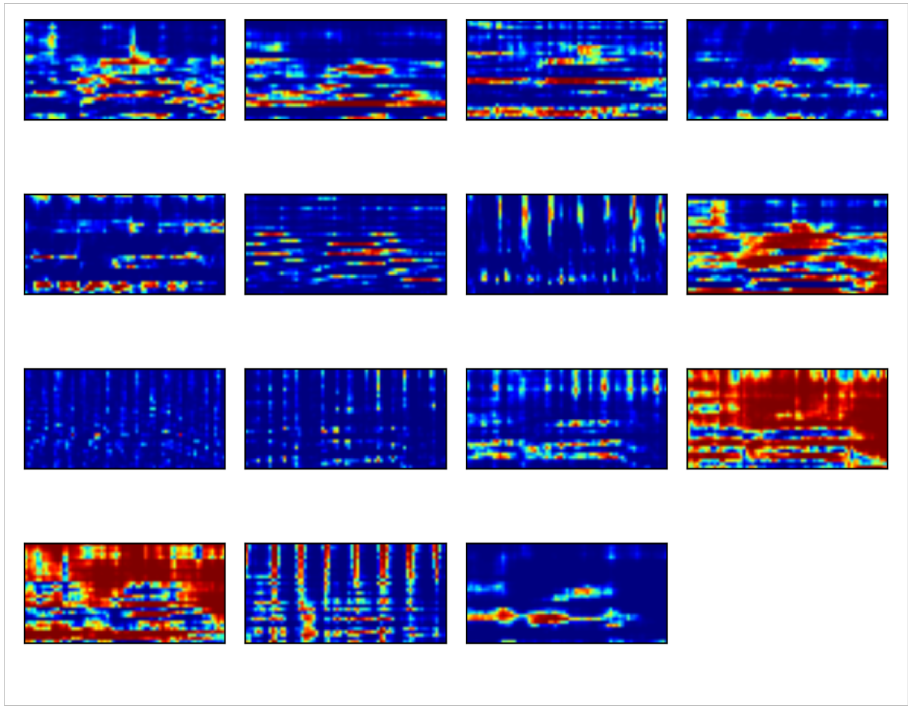
Visualizing CNN Filter Weights and Output

learned for Music/Speech Classification

Learned Filter Weights

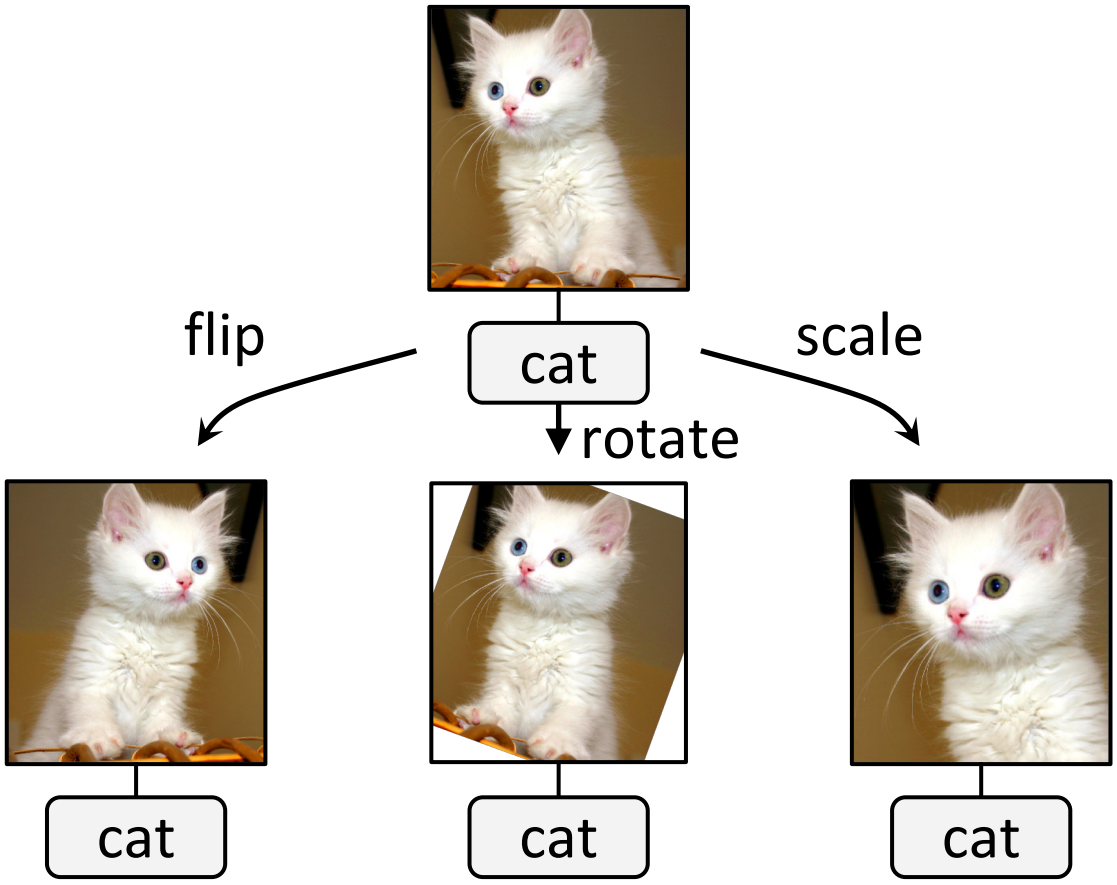


Convolved Spectrograms



Data Augmentation for Images

Transform training data, let classifier learn to ignore irrelevant properties (e.g. orientation, rotation, scale, etc.)



Data Augmentation for Audio

Typical transformations:

- For images: horizontal flip, scale, rotation, color, contrast
- **For audio**:
  - Time stretching (e.g. by factors 0.2x, 0.5x, 1.2x, 1.5x)
  - Pitch shifting (e.g. –5, –2, +2, +5 semitones)
  - Noise addition
  - Loudness, Equalizer, …

Benefits:

- **Regularization**: Less likely to learn training data “by heart”
- **Increases** apparent **training set size**: Can train larger model
- Increases apparent training set size: Can use smaller dataset

Applications

Speech Recognition

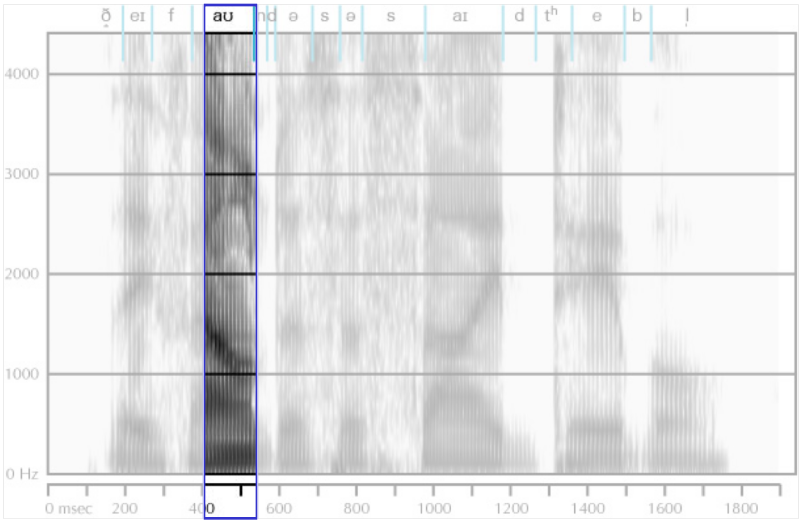
First successful networks in speech recognition:

Input: short spectrogram excerpt

Output: detected phoneme

Method: CNN

Has to be combined with word and language model, requires carefully-aligned training data.



aʊ

Dec 2015: Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, http://arxiv.org/abs/1512.02595

Speech Recognition

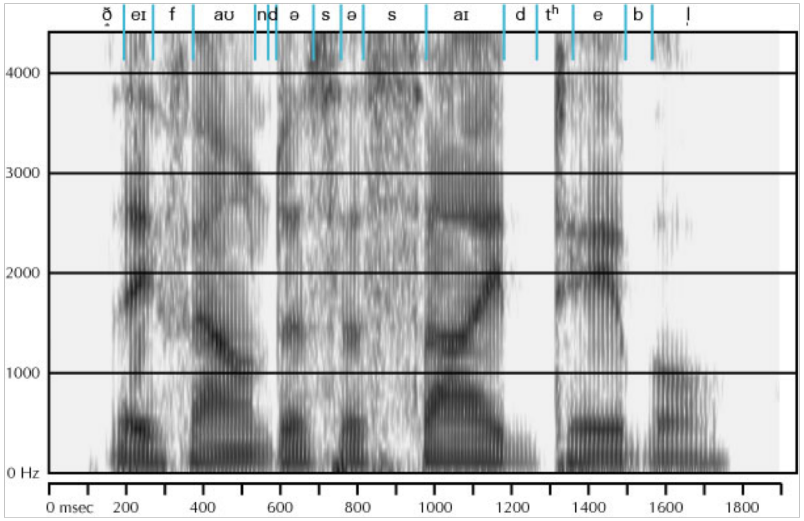
State of the art truly learns **from end to end**:

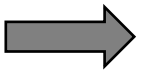
Input: spectrogram of recorded sentence

Output: transcribed sentence as a sequence of characters

Method: CNN for initial processing, RNN for temporal context

Works with coarsely aligned training data, learns word and language model on its own, **learns English and Mandarin with the same architecture**.





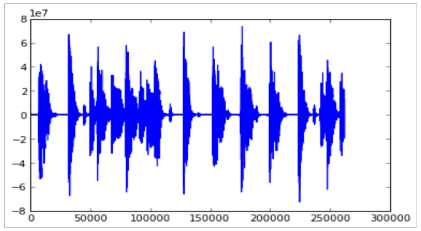
"They found us a side table."

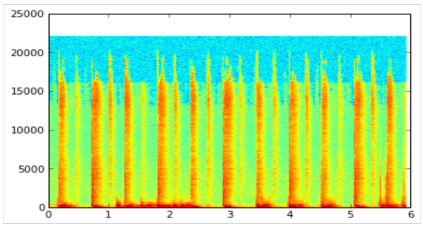
Dec 2015: Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, http://arxiv.org/abs/1512.02595

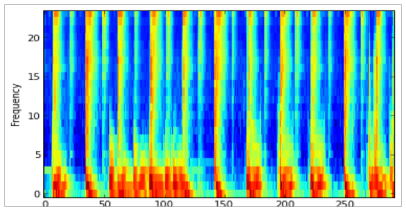
Music / Speech Classification

Distinguish audio excerpts between music and speech

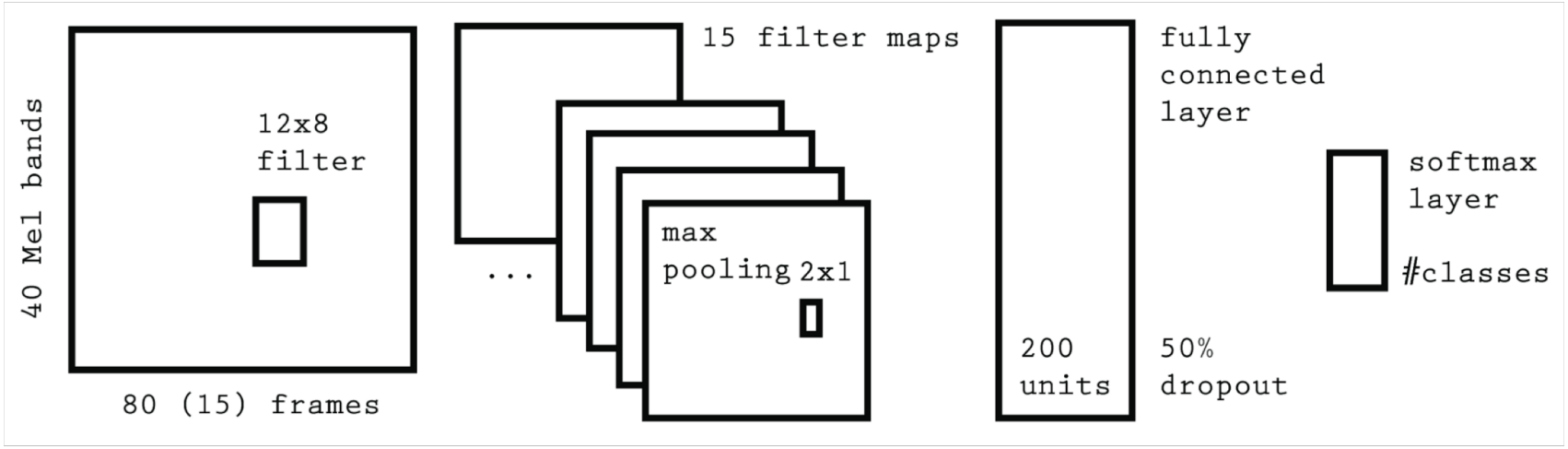
1. Pre-Processing: Waveform → Spectrogram → 40 Mel bands → Log scale







2. CNN with 1 layer, 15 filter maps + 1 full layer (input: 15 40x80 frames per file)



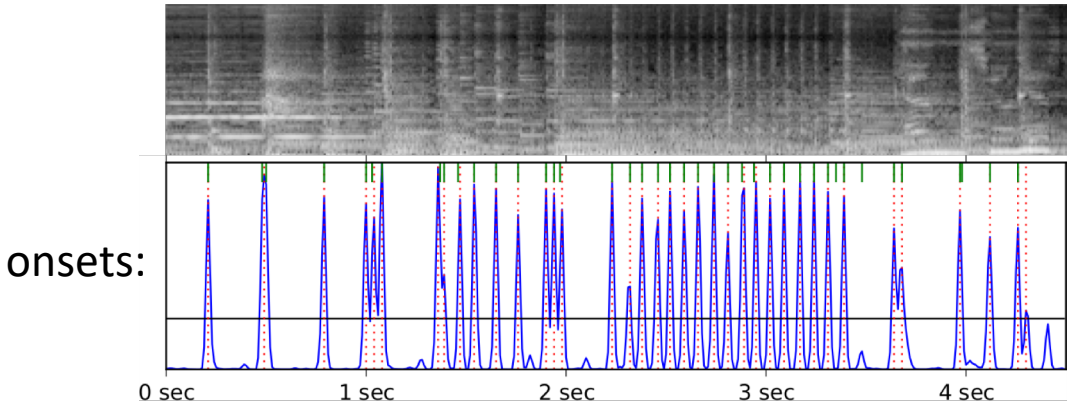
Winning algorithm **MIREX 2015** music/speech classification task **(99.73%) by Thomas Lidy**

Music Analysis: Note Onset Detection

Input: short spectrogram excerpt

Output: whether a note starts at center of excerpt

Combined result: starting positions of all notes played or sung



by Jan Schlüter

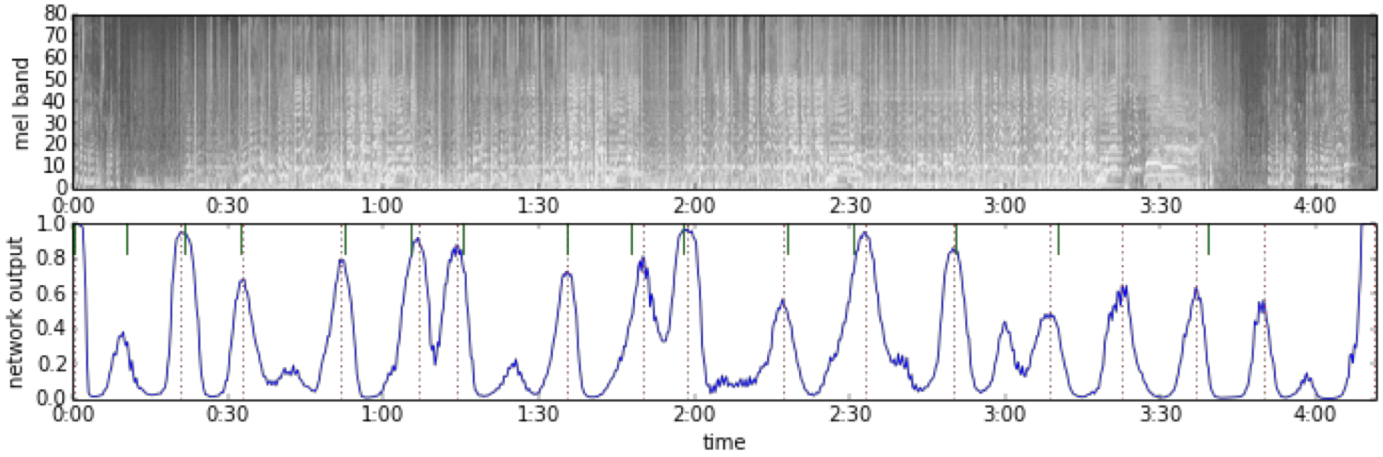
ICASSP 2014: Improved Musical Onset Detection with Convolutional Neural Networks

Music Analysis: Structural Segmentation

Input: long spectrogram excerpt

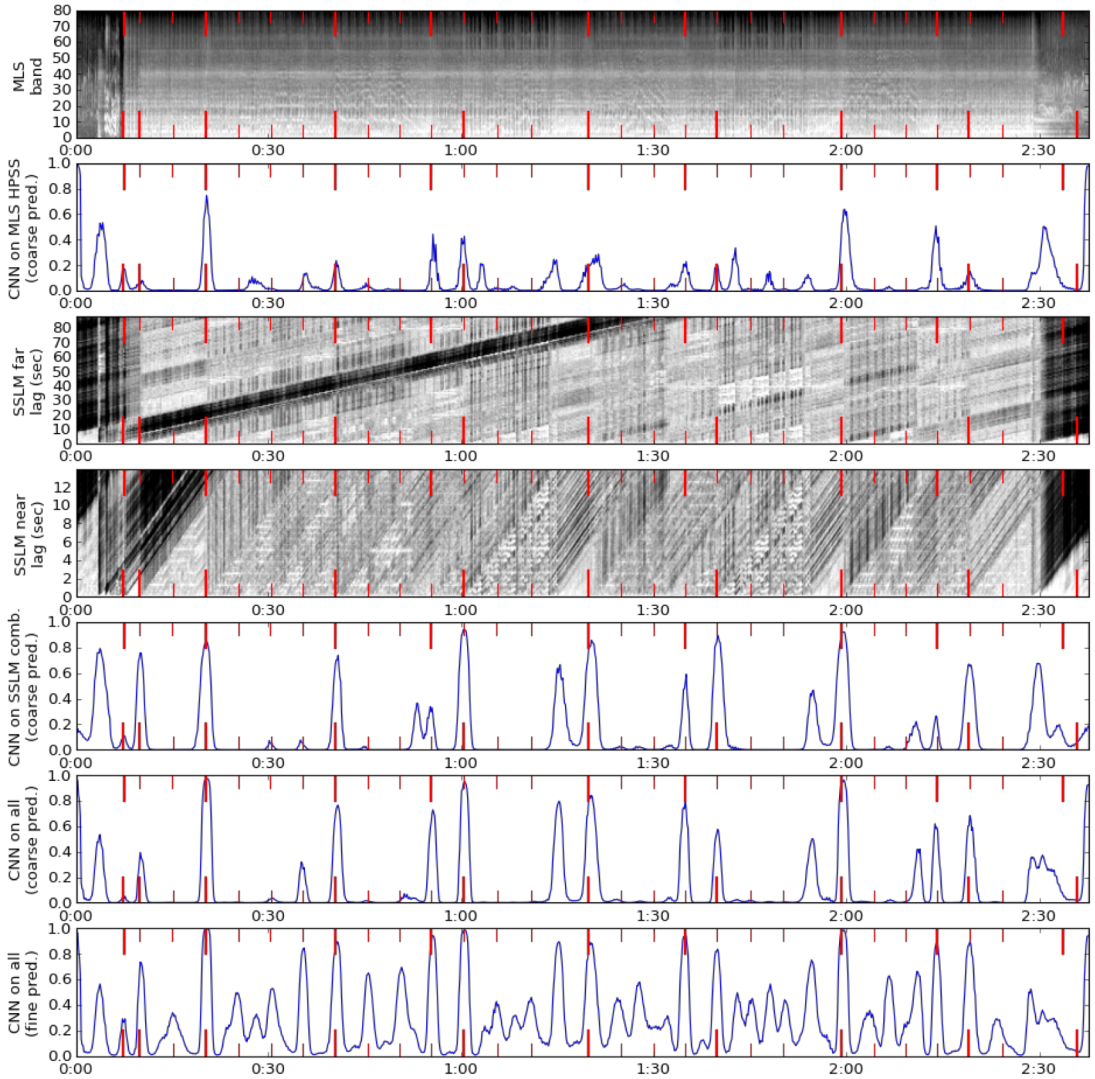
Output: whether music changes at center of excerpt

Combined result: positions of all structural boundaries



by Jan Schlüter

Music Analysis: Structural Segmentation

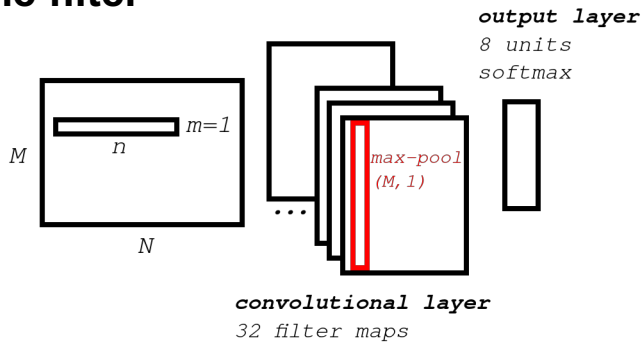


by Jan Schlüter

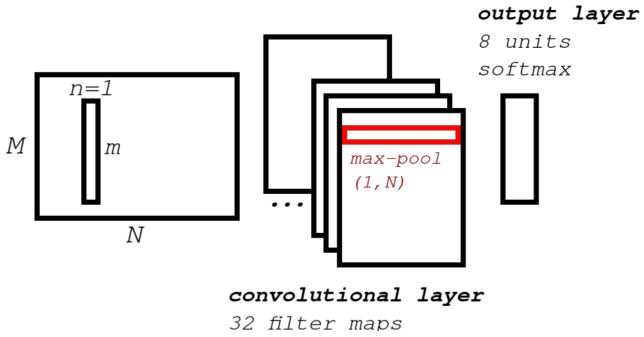
ISMIR 2015: Music Boundary Detection Using Neural Networks on Combined Features and Two-Level Annotations. http://ofai.at/research/impml/projects/audiostreams/ismir2015/

Parallel Networks: Musically Motivated CNN

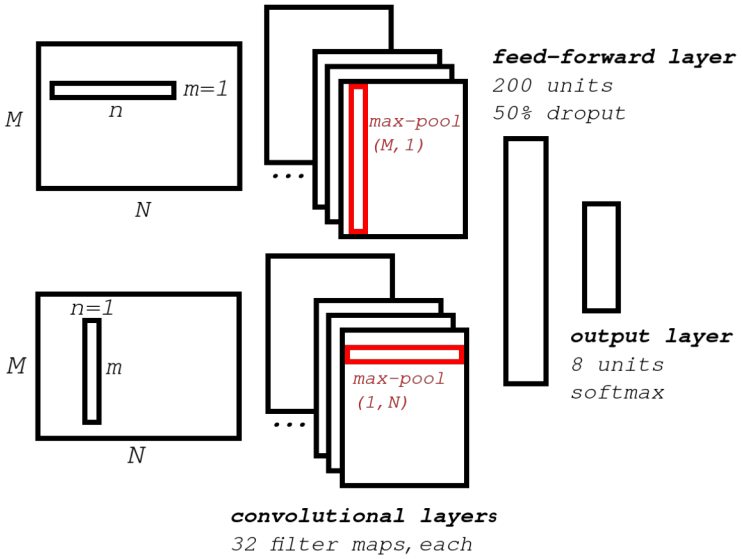
Time filter



Frequency Filter



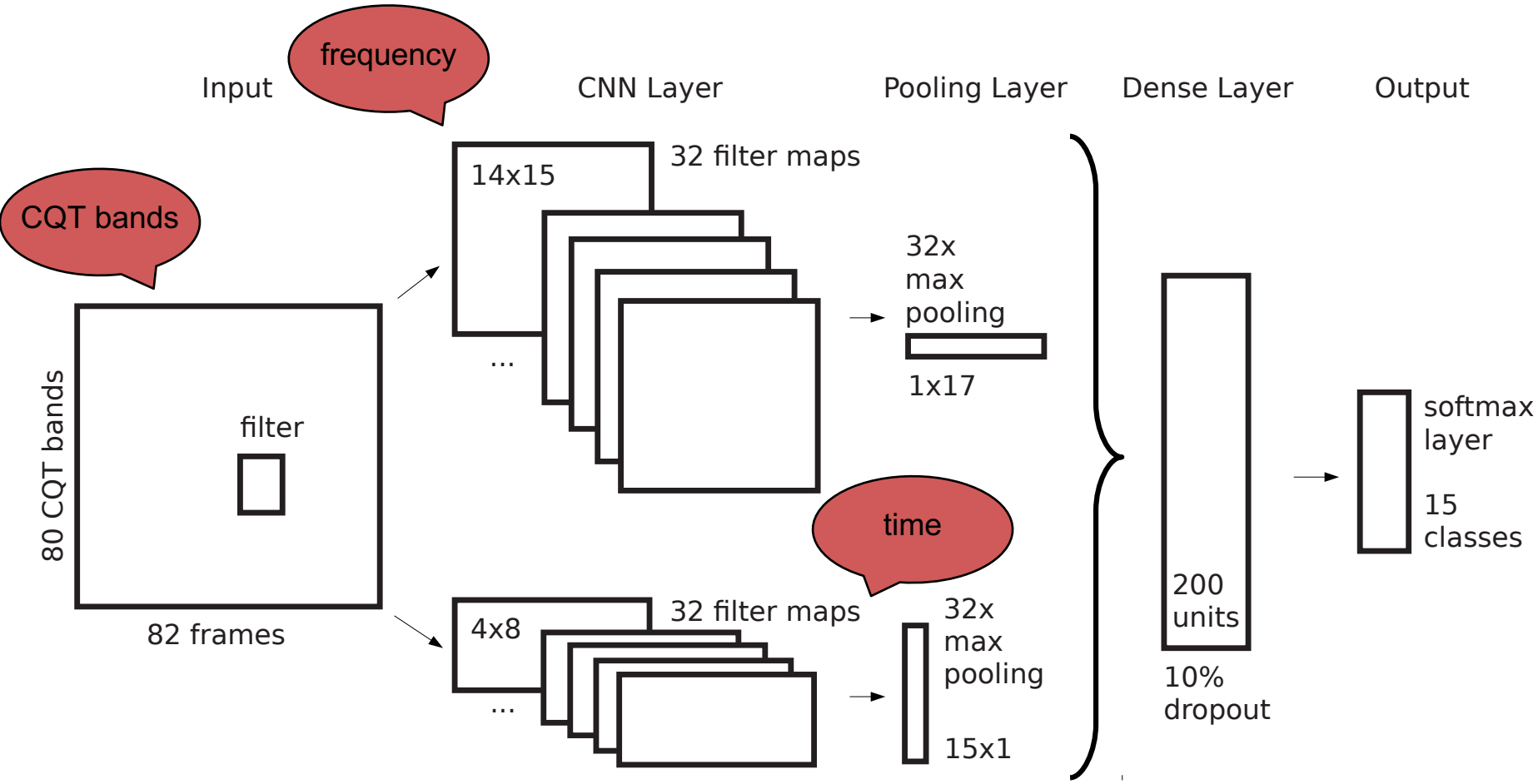
Parallel Model: Time + Frequency Model



Pooling goes in the opposite direction than the filter!

Jordi Pons, T. Lidy (CBMI 2016)

Parallel Network for Audio Tagging



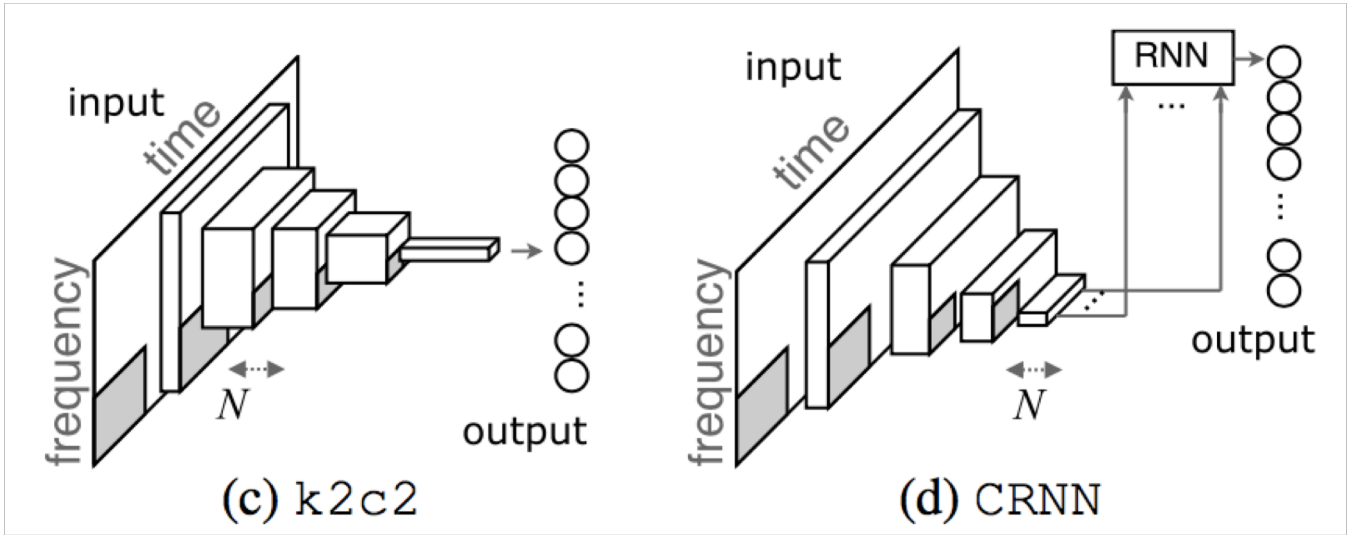
Winning model of DCASE 2016 Domestic Audio Tagging challenge
(child speech, male, female, video-game/TV, percussive sounds, other)

Lidy and Schindler (DCASE 2016)
CQT-BASED CONVOLUTIONAL NEURAL NETWORKS FOR AUDIO SCENE CLASSIFICATION

C-RNNs for Music Classification & Tagging

CNN

C-RNN



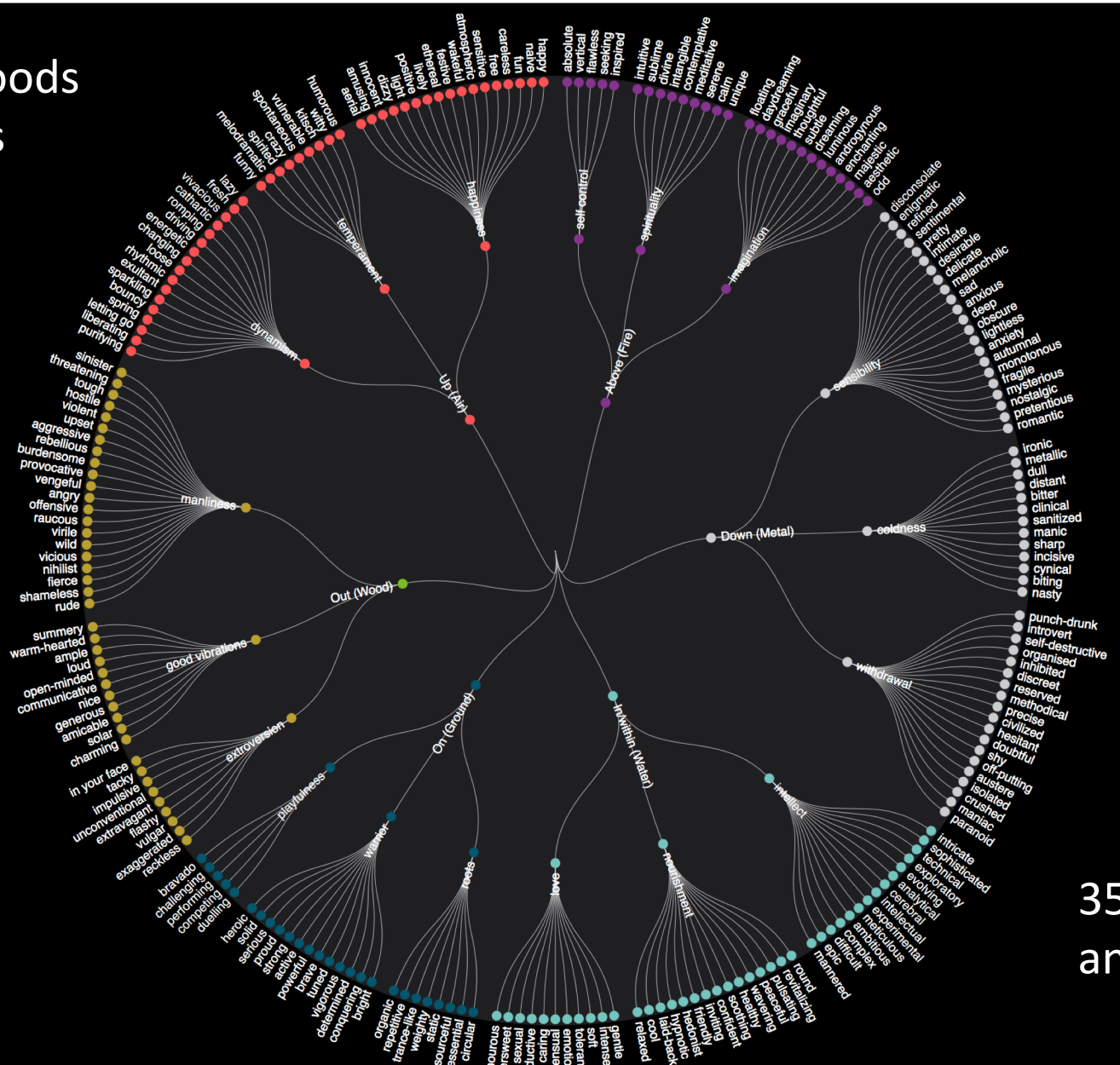
C-RNN (hybrid):

- CNNs for local feature extraction
- + RNNs for temporal summarisation of the extracted „features“

→ strong performance with respect to the number of parameter and training time
→ C-RNN better than CNN with comparable number of parameters

MUSIMAP: MOODS & EMOTIONS

256 moods
3 levels



35M+ tracks
analyzed

musimap **PLAYLISTS** – adapted to the users' tastes

My Ship, Speak low et Mack the Knife de Kurt Weill
The Testament Of Dr Mabuse - Classic Movies
(Das Testament des Dr Mabuse)

Kurt Weill - Wie lange noch

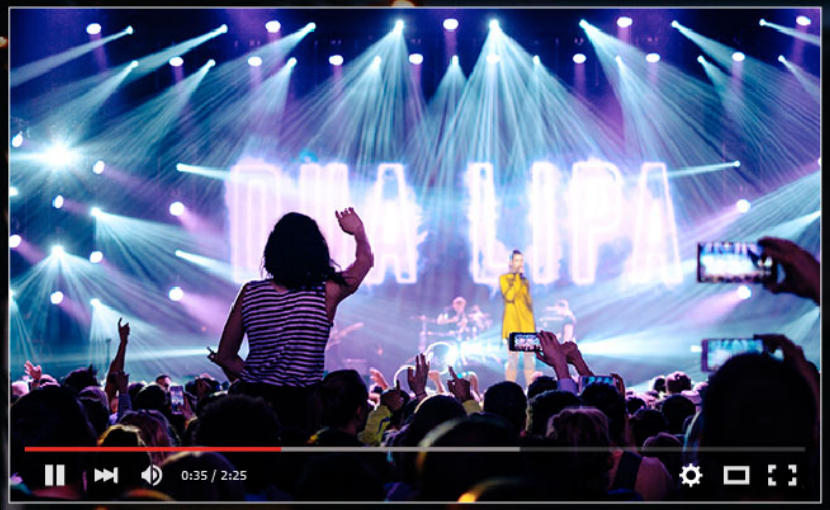
Teresa Stratas - Berlin im Licht-Song

Time Pieces: Jay Electronica
(Directed by Jason Goldwatch)

Lotte Lenya - Moon of Alabama
(Good Quality Audio)

Last Week Tonight with John Oliver:
Voting (HBO)

Mike Posner - I Took A Pill In Ibiza



FRIENDS
AND ACTIVITIES

YOUTUBE MUSIMAP

SLIDE TO SELECT A STYLE

Vocal Jazz

GENERATE

FORGIVENESS

VISUAL:
SUPER HEROES
(DC COMICS
AND
MARVEL COMICS)

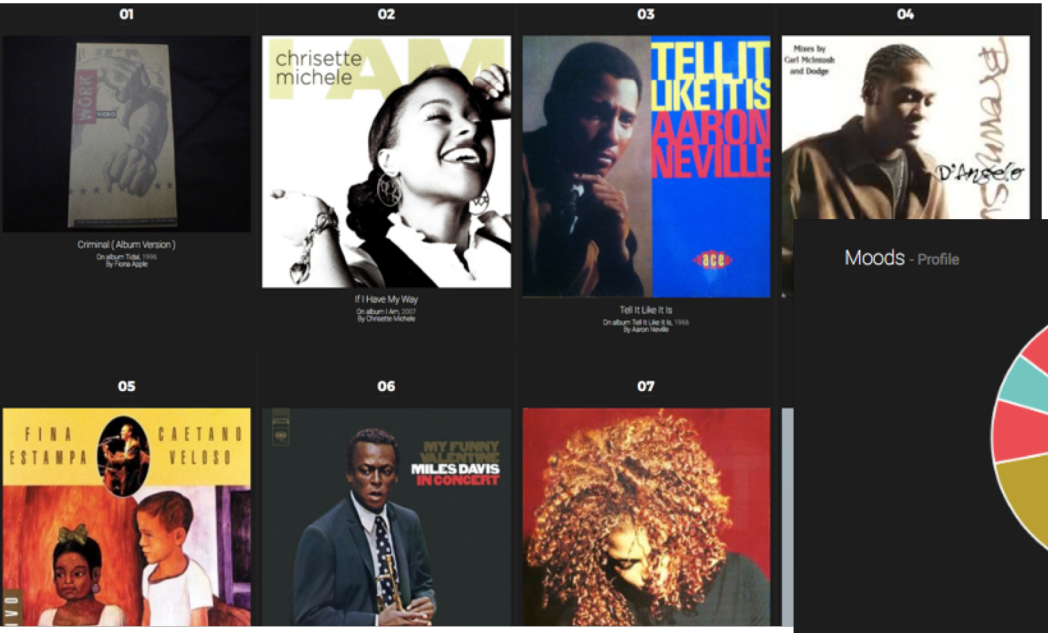
HALLS,
RESTAURANTS,
ELEVATORS
AND
WAITING ROOMS

CARS,
TRANSPORTS
AND
TRAVELLING

ADRENALINE
/
X-TREME SPORT

MUSIME: PSYCH-EMOTIONAL PROFILING

matching emotions in music to people‘s mood & personality

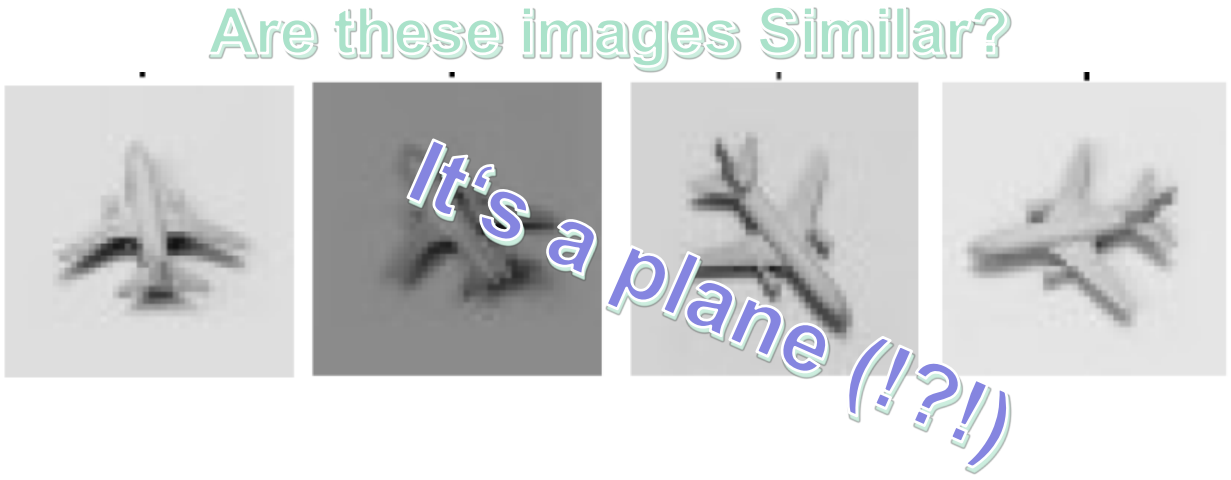


- Targeted Music
- Branding Campaigns
- Stress Reduction, Healthcare

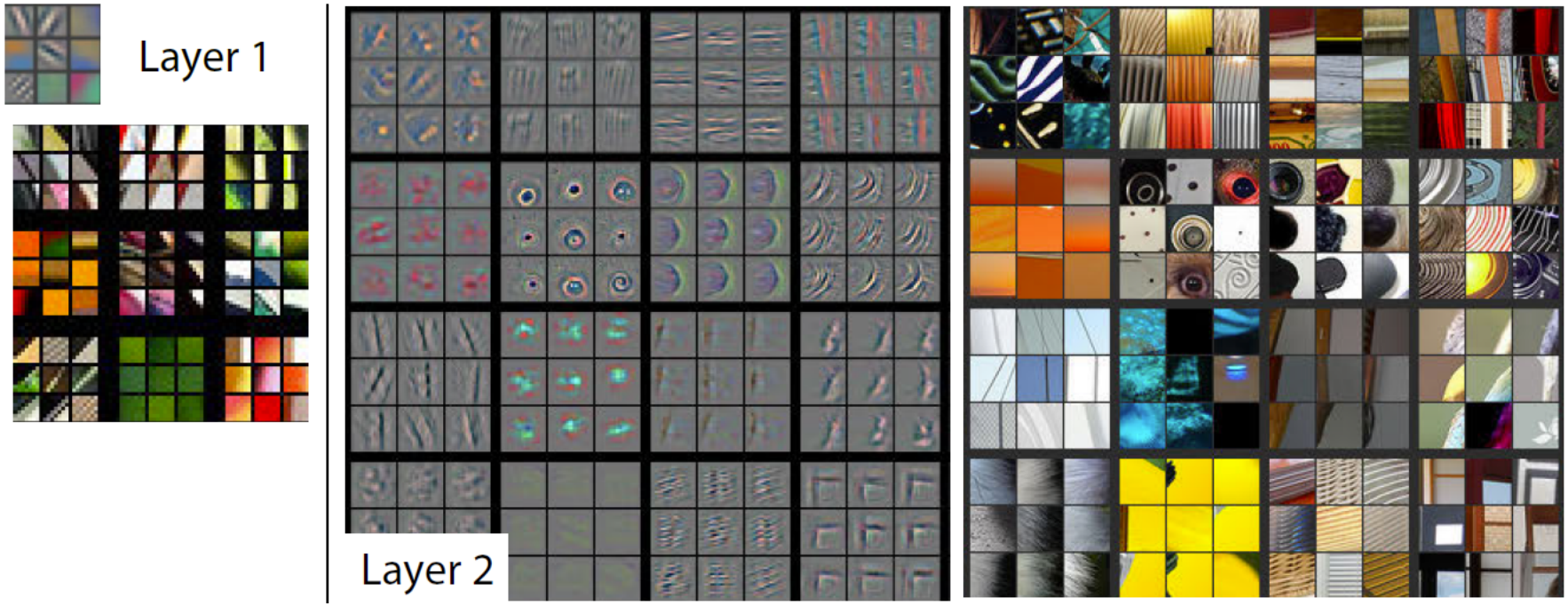
Representation Learning

CNNs trained for Classification

- Cross-Entropy Loss (=identification)
- Filter Kernel learn task-related object filter
- All other information is reduced

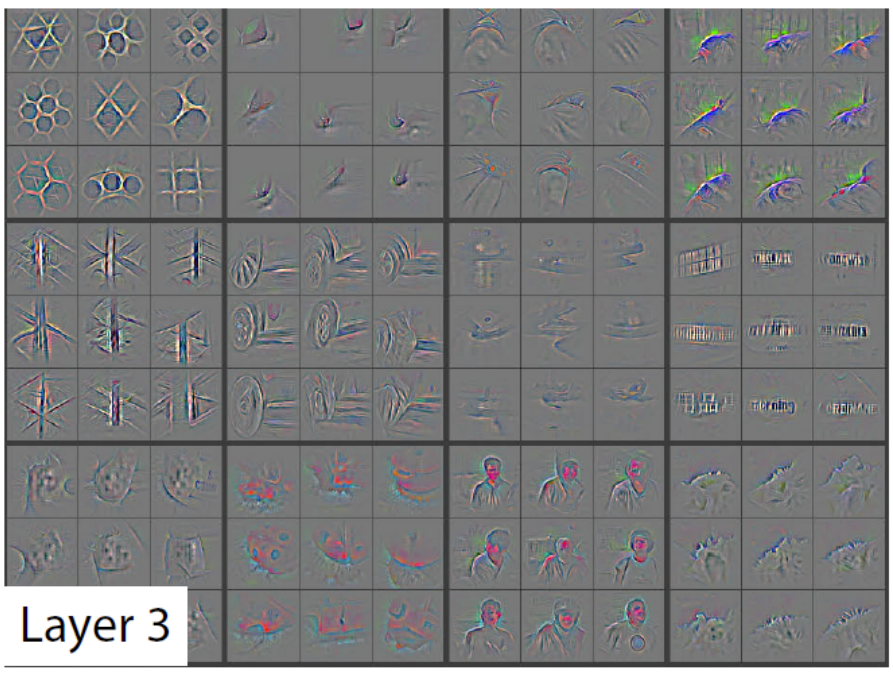


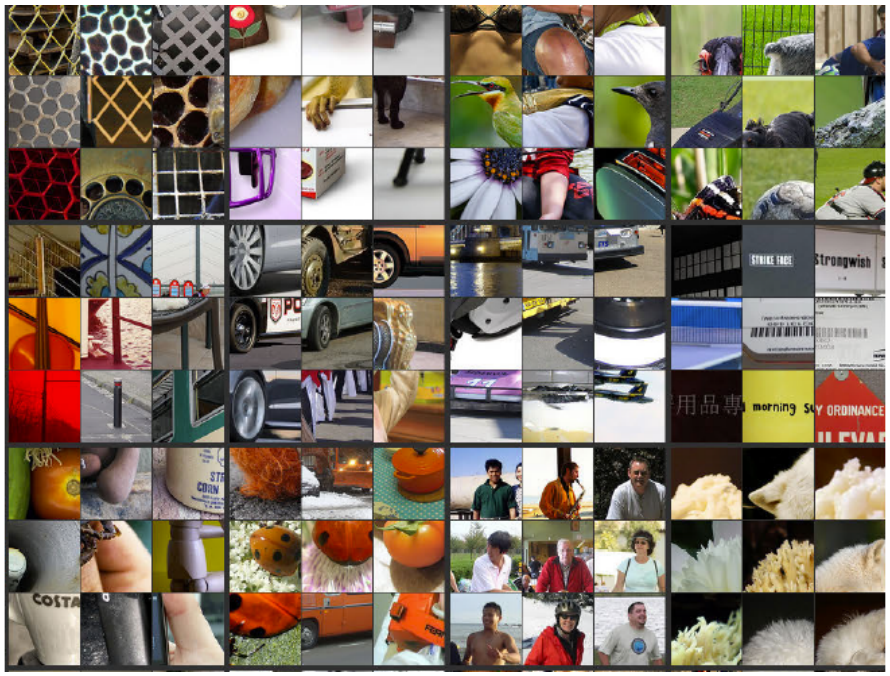
Classification CNN Filter Kernels



Source:
https://arxiv.org/abs/1311.2901

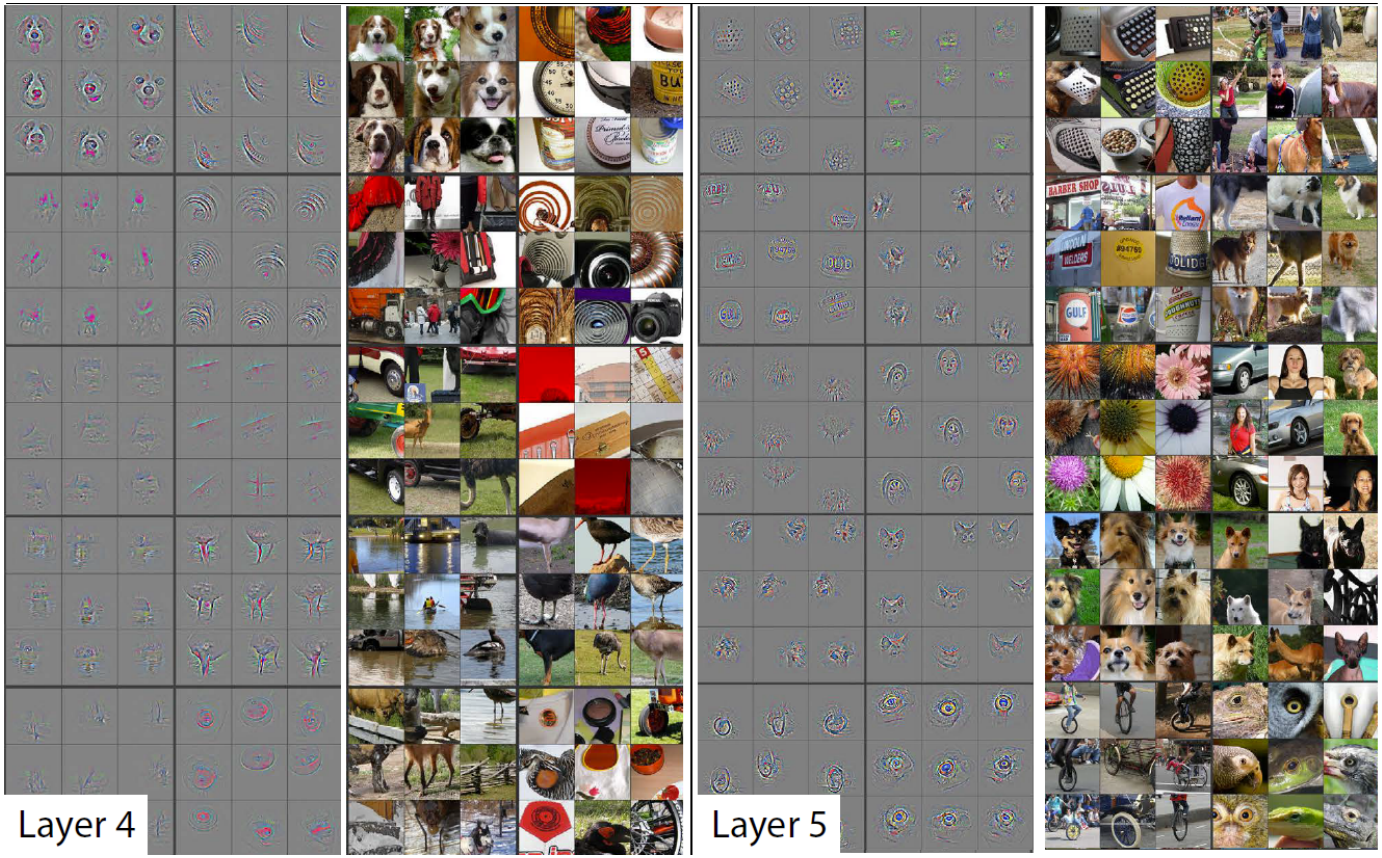
Classification CNN Filter Kernels





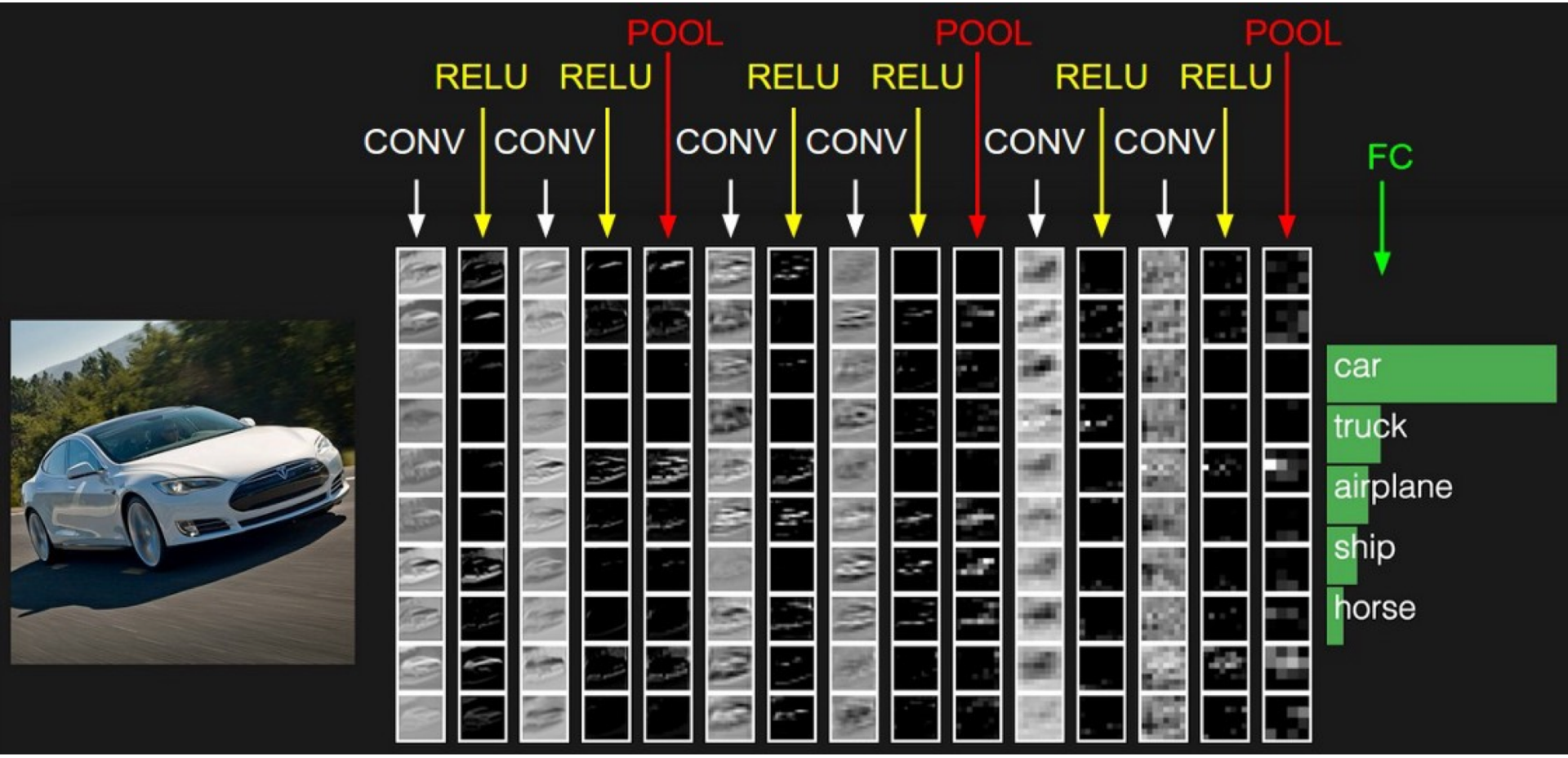
Source:
https://arxiv.org/abs/1311.2901

Classification CNN Filter Kernels



Source:
https://arxiv.org/abs/1311.2901

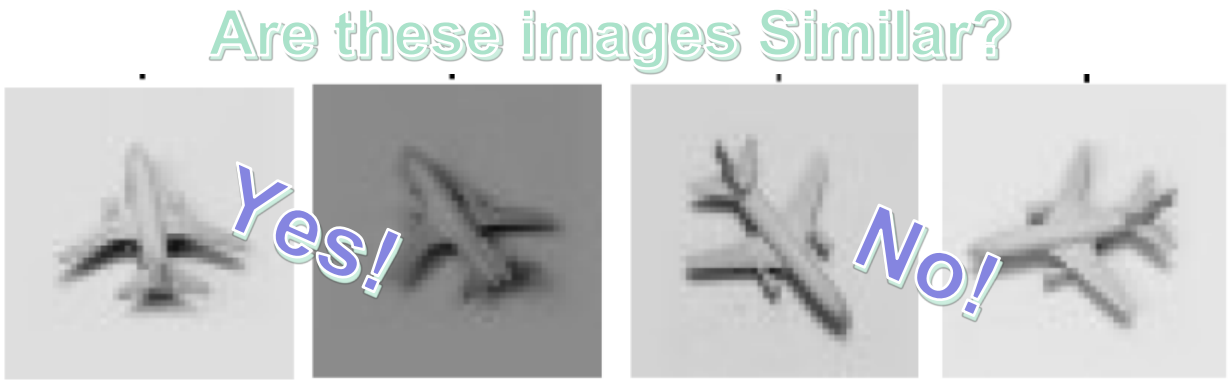
Classification CNN Filter Kernels



Source: https://medium.com/@phidaouss/convolutional-neural-networks-cnn-or-convnets-d7c688b

Representation Learning

- Metric Learning
- Content Representation Learning



Estimating Similarity

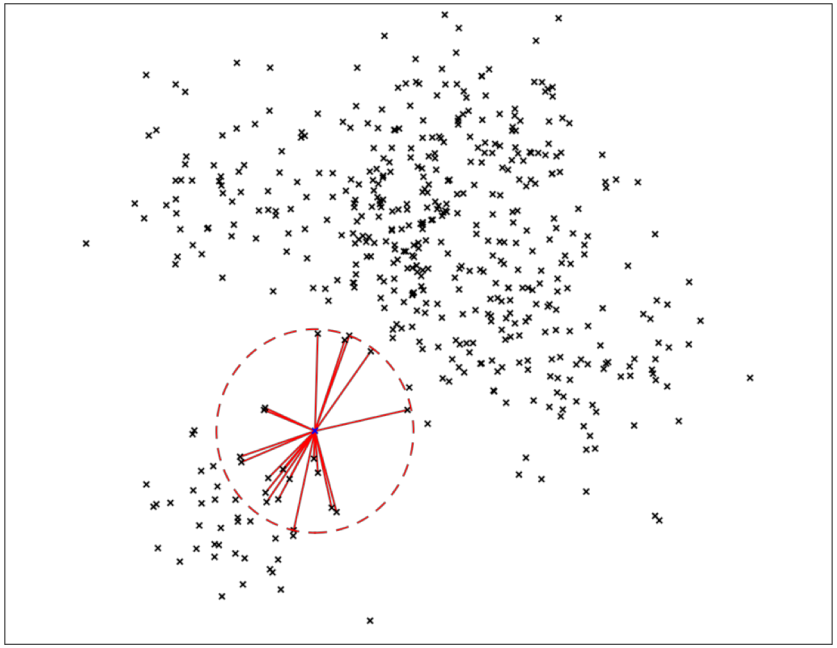
- Feature Vectors (v_1, v_2)
- Similarity Function (f_{sim})

$$sim = f_{sim}(v_1, v_2)$$

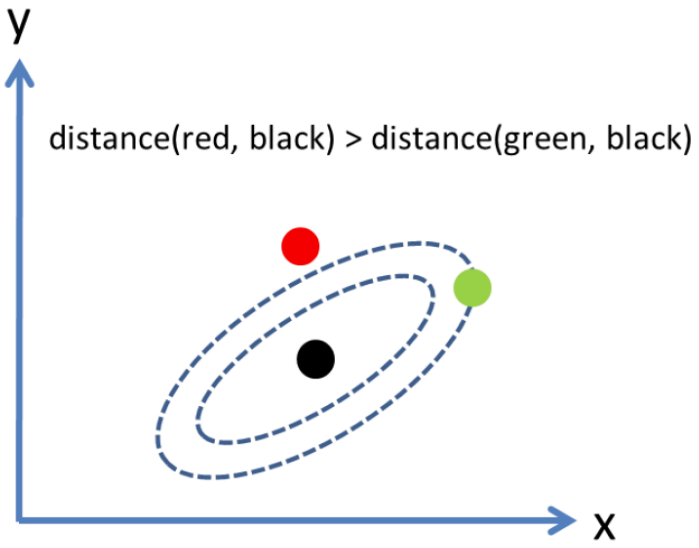
Metric Learning

- Euclidean distance

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$



Source: https://erikbern.com/2015/09/24/nearest-neighbor-methods-vector-models-part-1.html

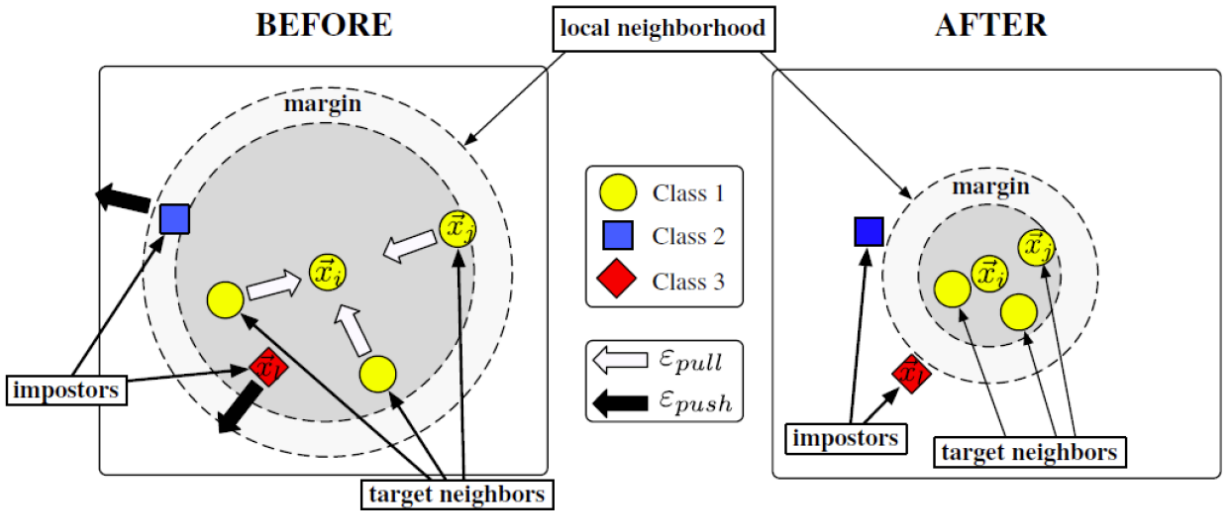


Source: http://horicky.blogspot.jp/2012/08/measuring-similarity-and-distance.html

Metric Learning

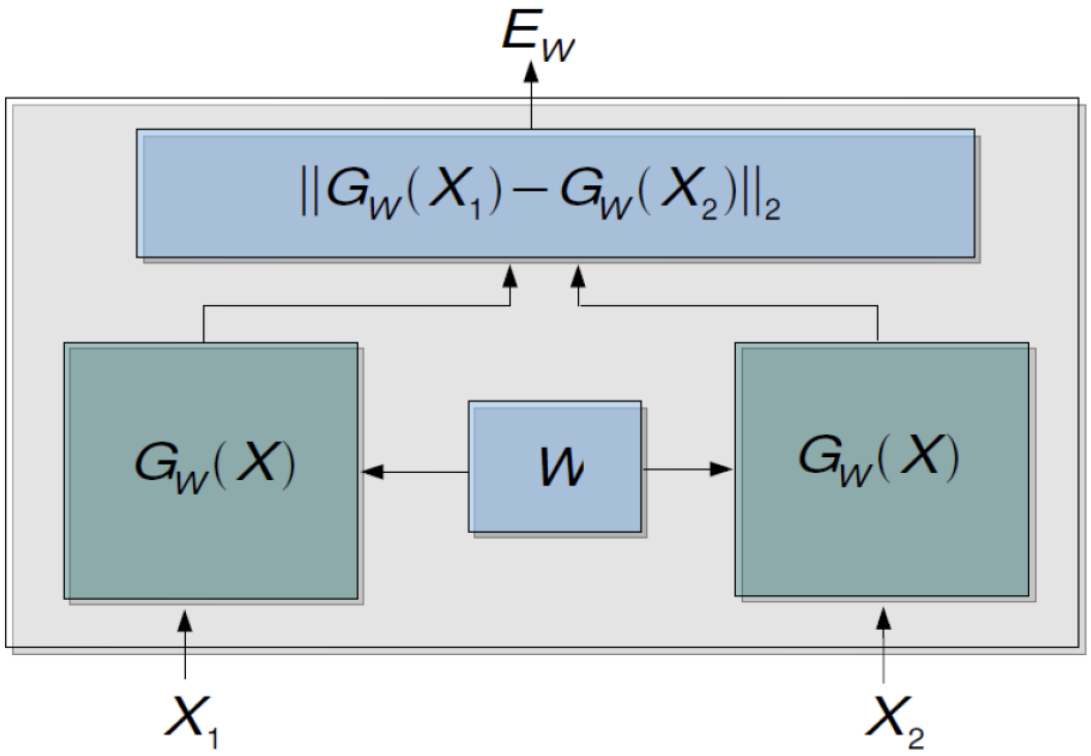
- Large-Margin Nearest Neighbors(LMNN)

$$\min_{A \succeq 0} \sum_{(i,j) \in \mathcal{S}} d_A(\boldsymbol{x}_i, \boldsymbol{x}_j) + \lambda \sum_{(i,j,k) \in \mathcal{R}} [1 + d_A(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_A(\boldsymbol{x}_i, \boldsymbol{x}_k)]_+$$



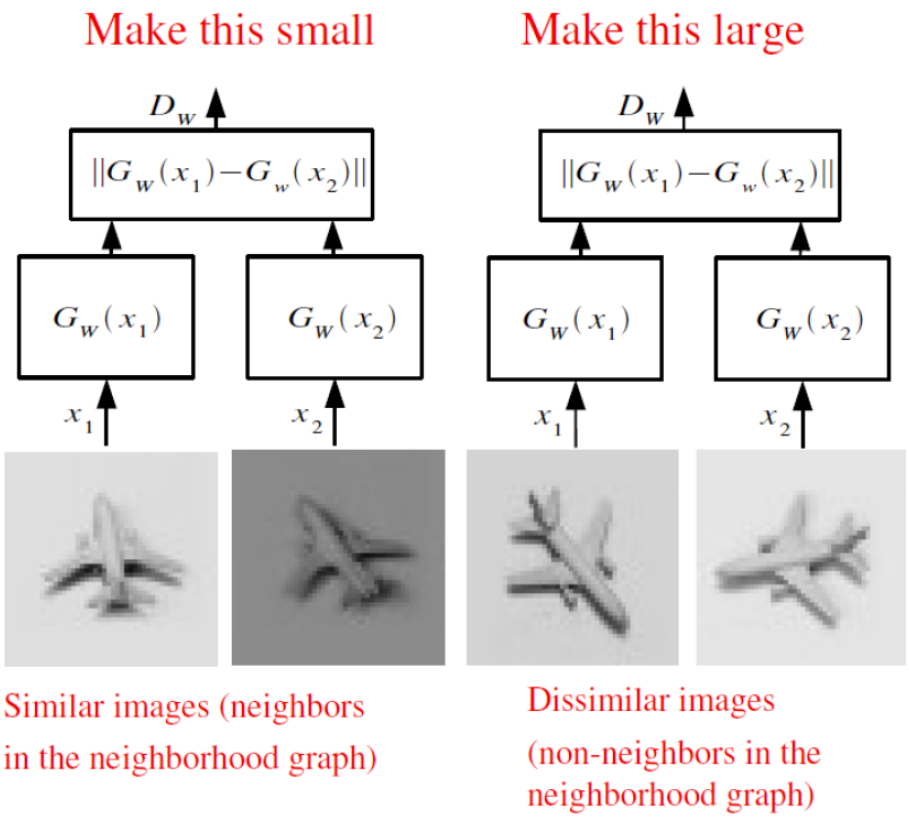
Source: Brian Kulis, Metric Learning: A Survey.web.cse.ohio-state.edu/~kulis/pubs/ftml_metric_learning.pdf

Siamese Architecture



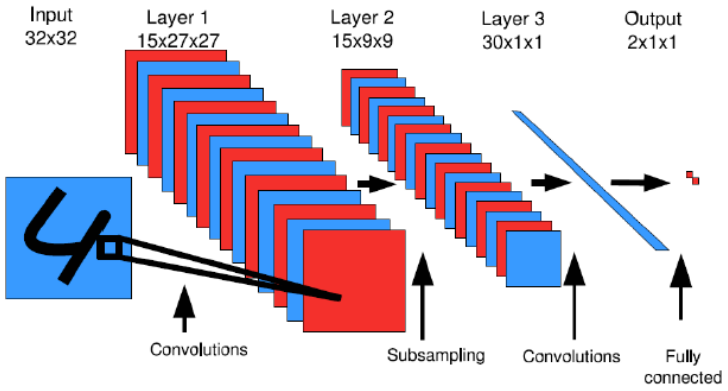
Source: Learning Hierarchies of Invariant Features. Yann LeCun.
helper.ipam.ucla.edu/publications/gss2012/gss2012_10739.pdf

Siamese Archtiecture and loss Function



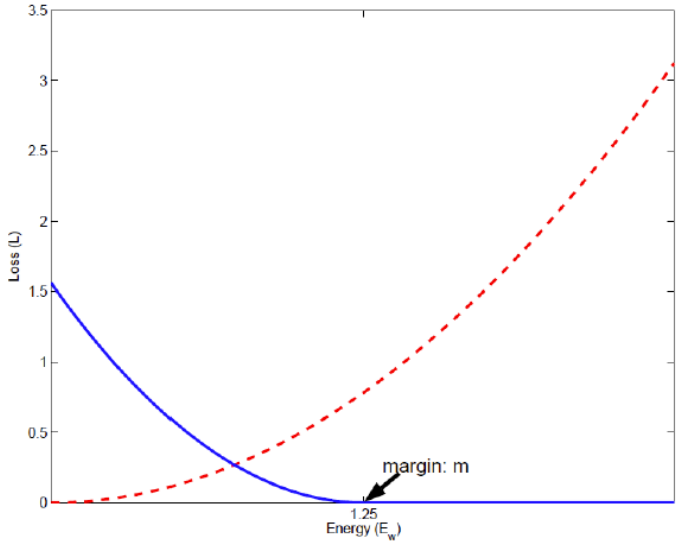
Source: Learning Hierarchies of Invariant Features. Yann LeCun. helper.ipam.ucla.edu/publications/gss2012/gss2012_10739.pdf

Application in Dimensionaltiy reduction



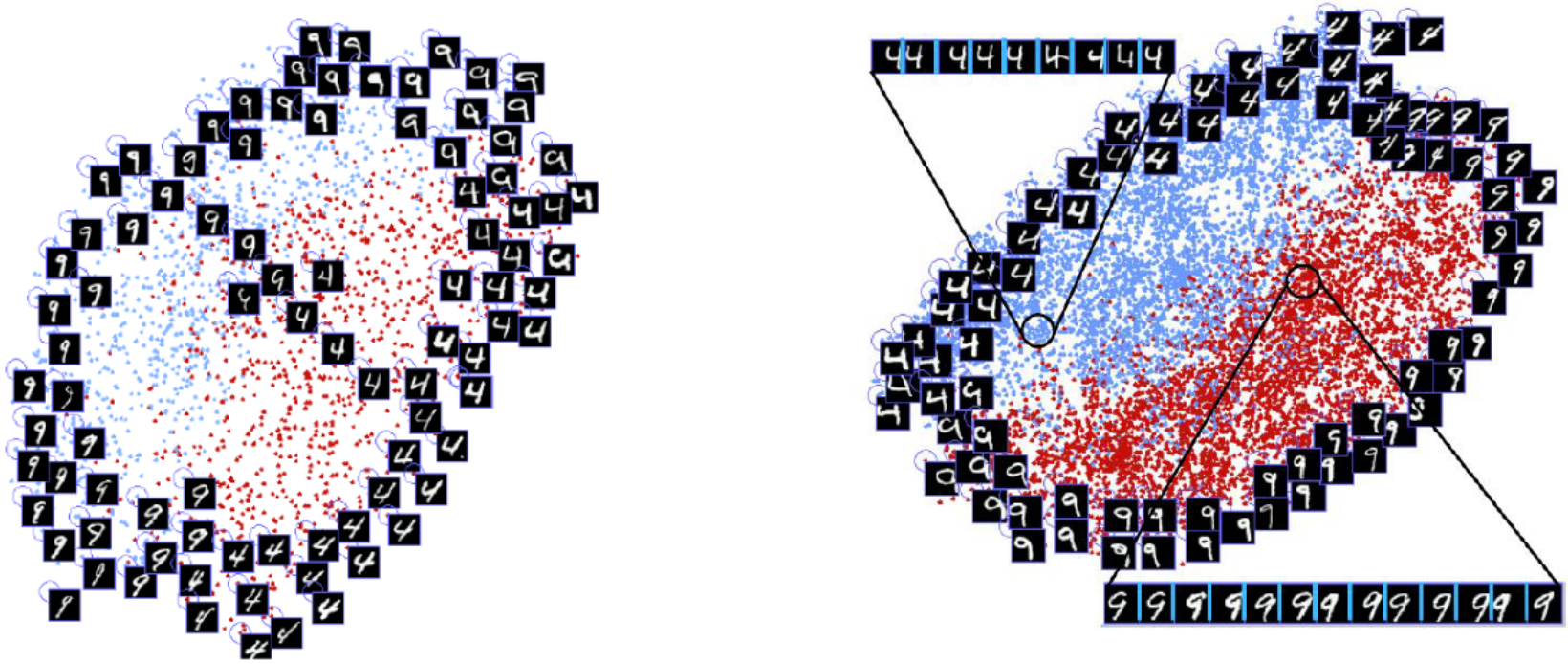
The exact loss function is

$$L(W, Y, \vec{X_1}, \vec{X_2}) =$$
$$(1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$



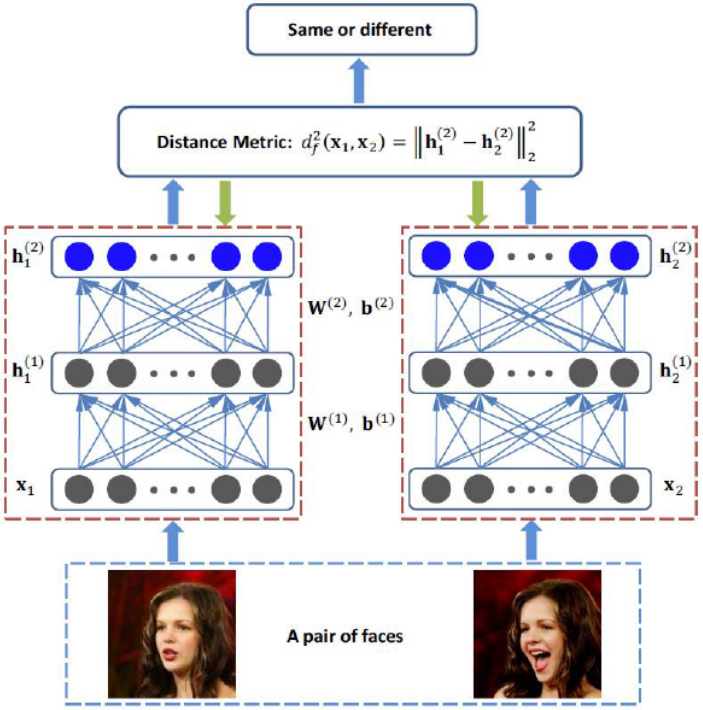
Source: HadsellR, Chopra S, LeCunY. Dimensionality reduction by learning an invariant mapping, CVPR 2006

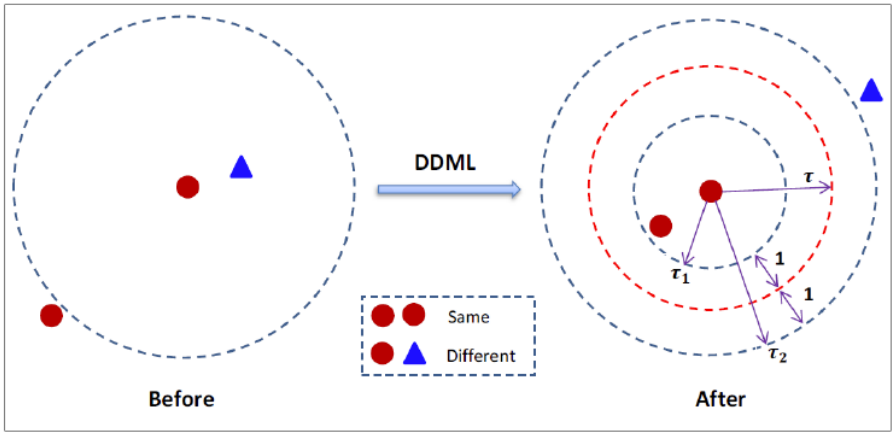
Application in Dimensionaltiy reduction



Source: HadsellR, Chopra S, LeCunY. Dimensionality reduction by learning an invariant mapping, CVPR 2006

Application: Face Verification

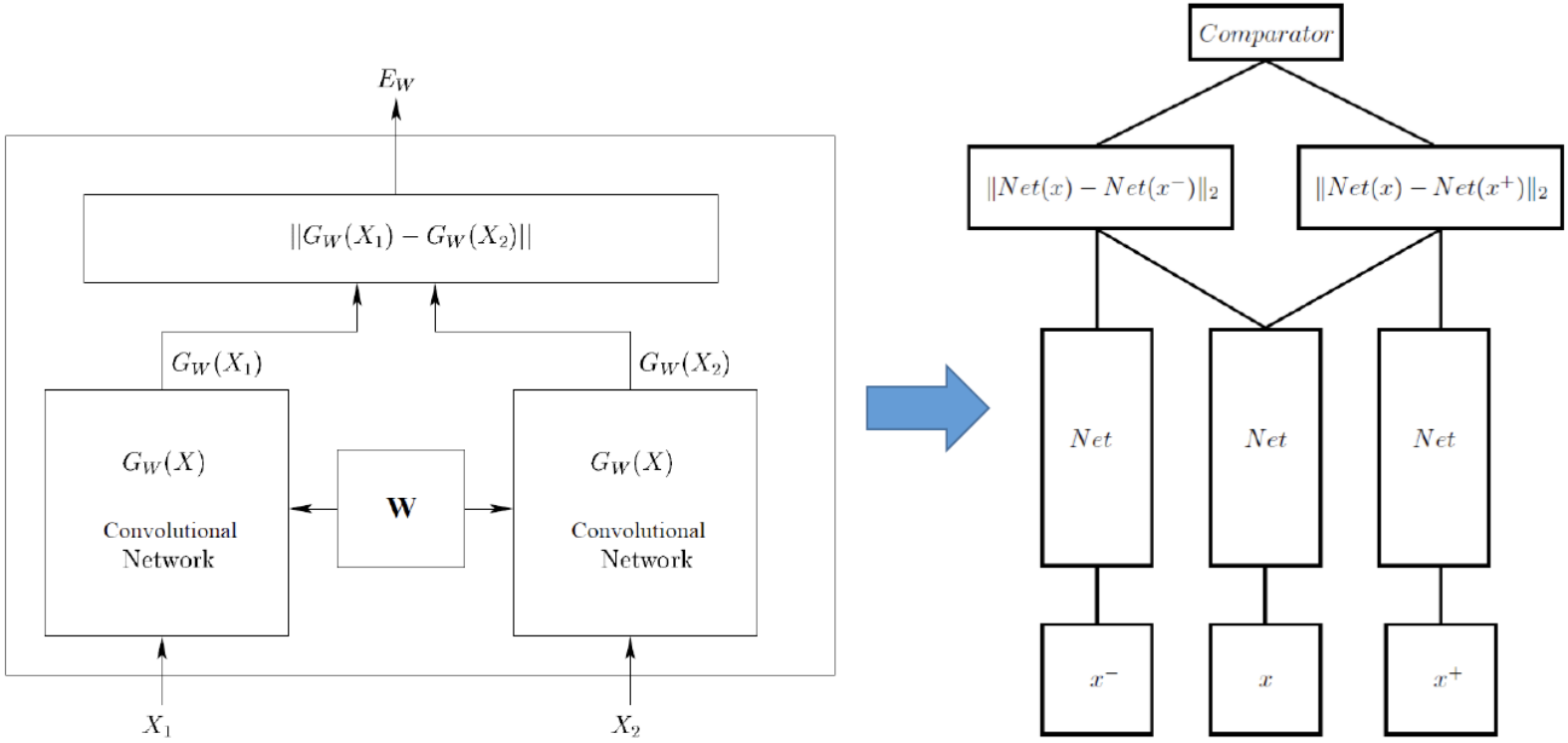




Intuitive illustration of the proposed DDML method

Source: JunlinHu, etc. Discriminative Deep Metric Learning for Face Verification in theWild, CVPR 2014

Triplet Network



Source: EladHoffer, etc. DEEP METRIC LEARNING USING TRIPLET NETWORK. Under review as a conference paper at ICLR 2015 http://arxiv.org/abs/1412.6622

Triplet Network

Sample images from the triplet dataset

Source: Jiang Wang, etc. Learning Fine-grained Image Similarity with Deep Ranking. CVPR 2014

Triplet Network

2D VISUALIZATION OF FEATURES of CIFAR10

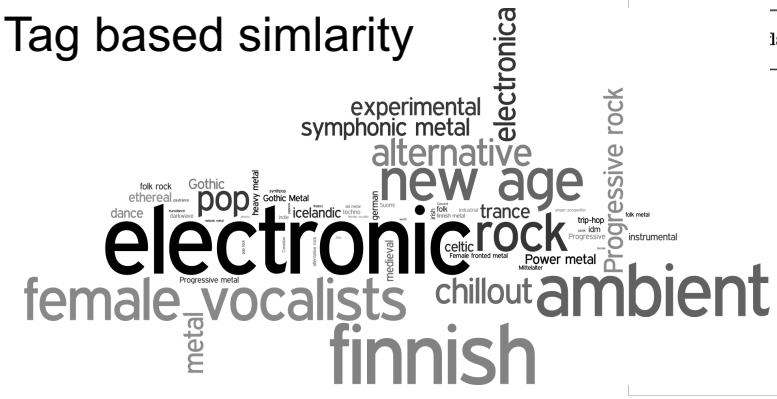
Source: EladHoffer, etc. DEEP METRIC LEARNING USING TRIPLET NETWORK. Under review as a conference paper at ICLR 2015 http://arxiv.org/abs/1412.6622

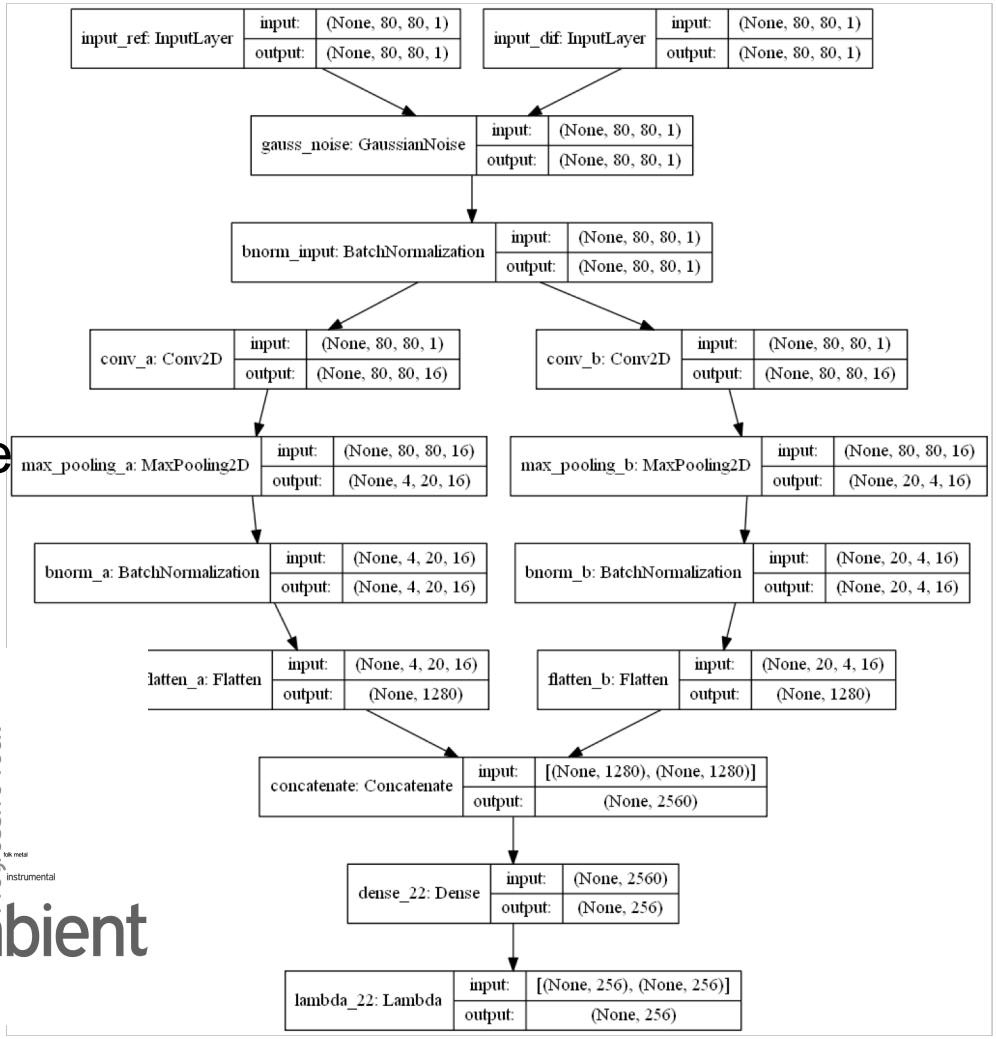
Advantages

- More robust to class imbalance
  - Extreme case: One Shot Learning (from a single sample)
- Learns domain adaptive feature representation
- Loss function is very important

Siamese Networks for Audio/Music

- CNN with Spectrogram input
  - MelSpectrogram, CQT, etc.
- Siamese / Triplet Architecture
- Define Similarity
  - Same genre, mood, theme
  - Tag based simlarity





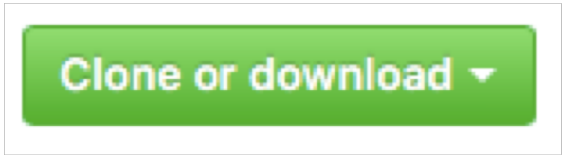
Credits / Contributions

Contributions & sources:

- Jan Schlüter (OFAI Vienna, www.ofai.at/~jan.schlueter)
- Yoshua Bengio
- Andrej Karpathy

… and many others…

Deep Learning for Music



Tutorials on Github

Tutorial 1:

bit.ly/mlmusic18 (or:

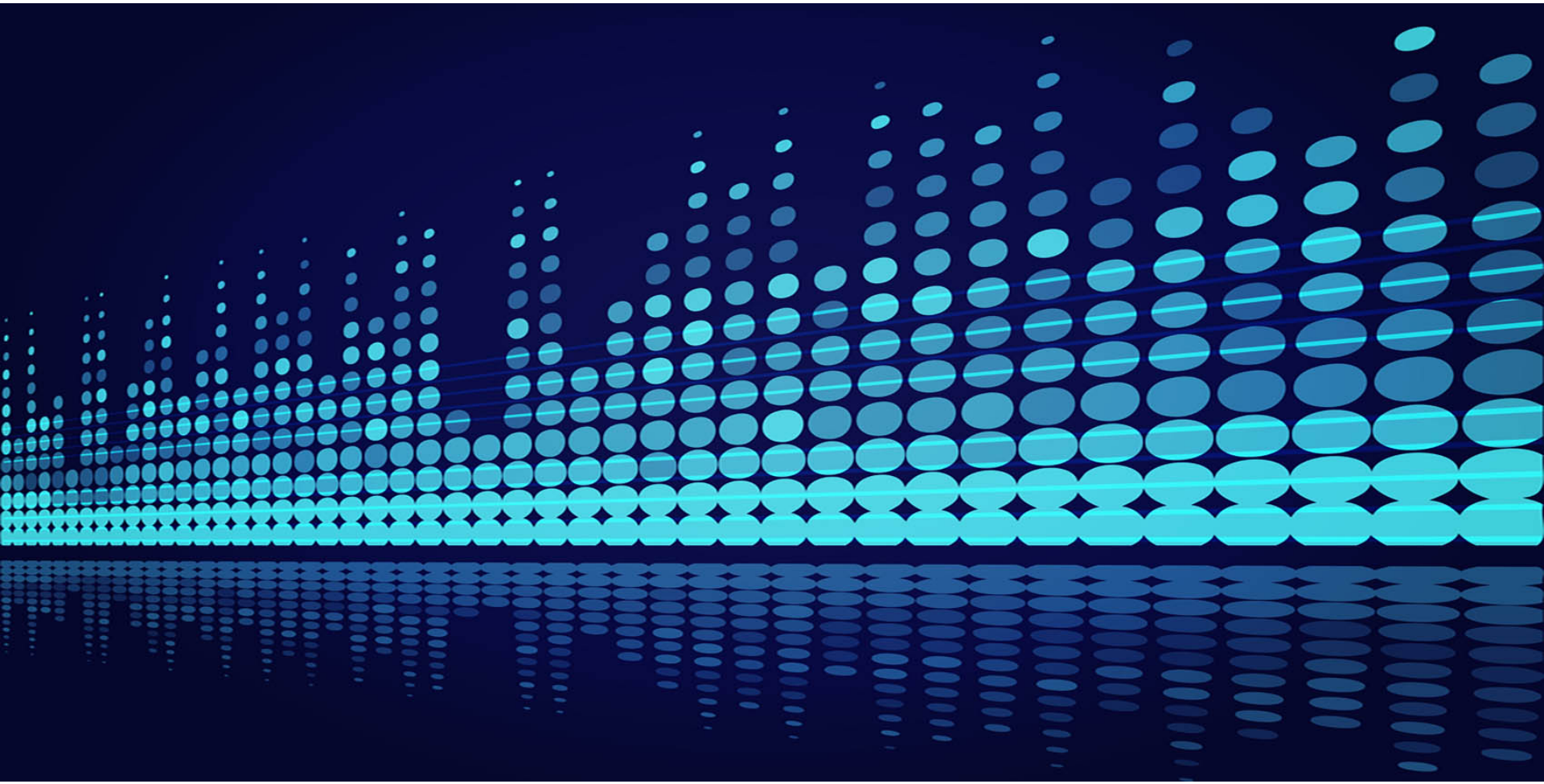
https://github.com/slychief/mlprague2018_tutorial)

Tutorial 2:

http://tiny.cc/dlismir18 (or:

https://github.com/slychief/ismir2018_tutorial)

Deep Learning for Music and Audio



Thomas Lidy and Alexander Schindler

21st Vienna Deep Learning Meetup, 15 Oct 2018

Vienna Deep Learning Meetup