Tutorial:
Deep Learning for Music Classification using Keras

Deep Learning for Music Classification using Keras

Alexander Schindler
Scientist
AIT & TUWien

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

Thomas Lidy
Head of Machine Learning
Musimap

tom@musimap.com
www.musimap.com

TUTORIAL AGENDA

AUSTRIAN INSTITUTE OF TECHNOLOGY

I. Deep Learning Basics:

- Audio Processing Basics
- History of Neural Networks
- What is Deep Learning
- Neural Network Concepts
- Coding Examples

II. Convolutional Neural Networks:

- Difference CNN – RNN
- How CNNs work (Layers, Filters, Pooling)
- Application Domains and how to use in Music
- Coding Examples

TUTORIAL AGENDA

AUSTRIAN INSTITUTE OF TECHNOLOGY

III. Instrumental, Genre and Mood Analysis:

- Large-scale Music AI at Musimap
- Instrumental vs. Vocal Detection
- Genre Recognition
- Mood Detection
- Coding Examples

IV. Advanced Deep Learning:

- Similarity Retrieval
- Siamese Networks
- Learning Audio Representation from Tag Similarity
- Coding Examples

TUTORIAL ON GITHUB

AUSTRIAN INSTITUTE OF TECHNOLOGY

https://github.com/slychief/mlprague2018_tutorial

or

bit.ly/mlmusic18

+ scroll to the end of the page to download the data sets!

- **GTZAN Music Speech**
- **MagnaTagaTune**

ALEXANDER SCHINDLER
SHORT BIO

- 2000 – 2004 Software Engineer, Siemens
- 2003 – 2007 Student Assistant, TU-Wien

- since 2010 AIT Austrian Institute of Technology
- 2010 – 2014 High performance image processing Group
- 2014 - Data Science Group „Digital Insights Lab“

- Since 2012 Researcher TU Wien – ML & Music

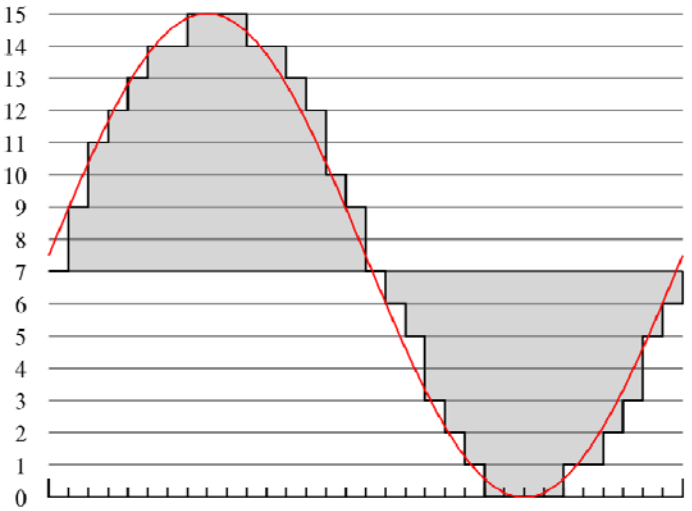
- Since 2017 Guest lecturer, Uni-Wien, FH Burgenland

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AUDIO PROCESSING BASICS

DIGITAL AUDIO

- Digital Audio
  - Sampling Rate: 44,100 Hz
  - 16-bit resolution for each channel
    - 2 channels for stereo
  - 88,200 Integers per second



EXCERCISE: FIND SONGS WITH STRINGS

Song 1:

83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98

Song 2:

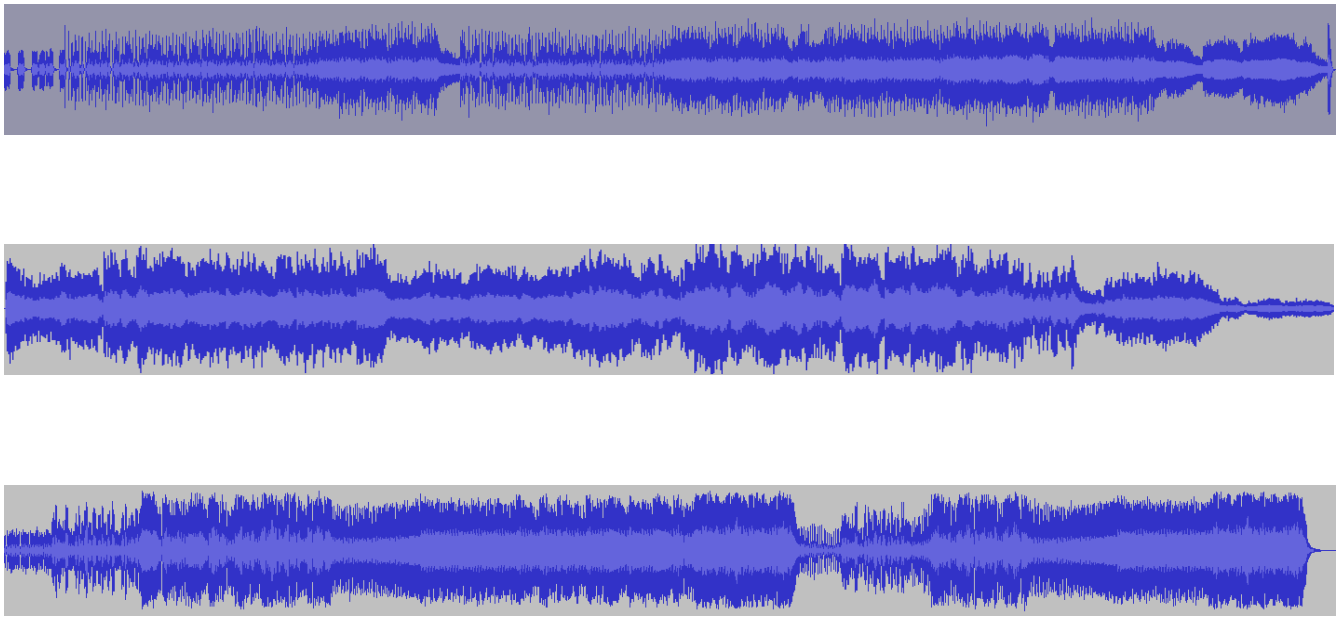
55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7

Song 3:

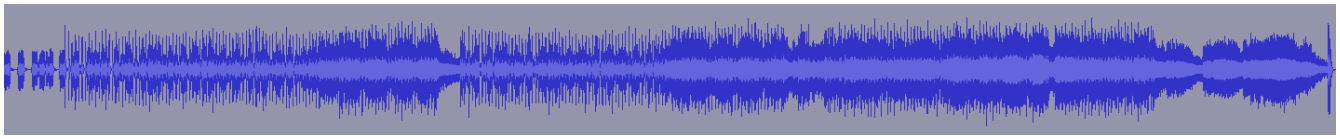
66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2, 60, 9, 23, 43, 21, 98, 34, 29, 62, 26, 72, 38, 98, 55, 96, 11, 49, 83, 58, 11, 11, 9, 60, 96, 25, 39, 42, 87, 90, 12, 26, 99, 69, 10, 56, 64, 41, 47, 61, 6, 40, 94, 23, 43, 52, 31, 77, 32, 57, 40, 89, 91, 28, 38, 96, 3, 90, 43, 18, 25, 16, 79, 97, 83, 64, 46, 70, 63, 34, 38, 39, 7, 66, 89, 95, 9, 47, 11, 59, 9, 17, 46, 92, 27, 58, 87, 46, 39, 100, 10, 2, 5, 53, 73, 56, 43, 46, 47, 67, 2

AUSTRIAN INSTITUTE OF TECHNOLOGY

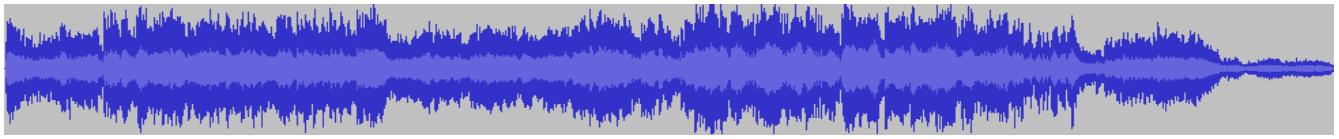
EXCERCISE: SAME GENRE?



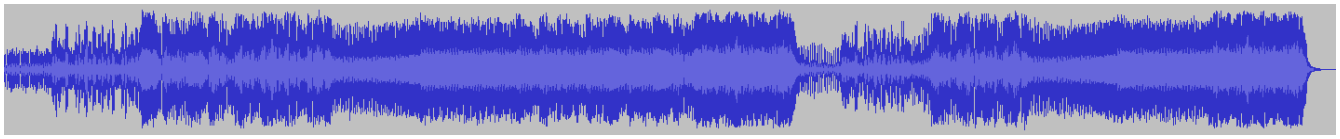
EXCERCISE: IDENTIFY SONGS



AC-DC – Highway to Hell



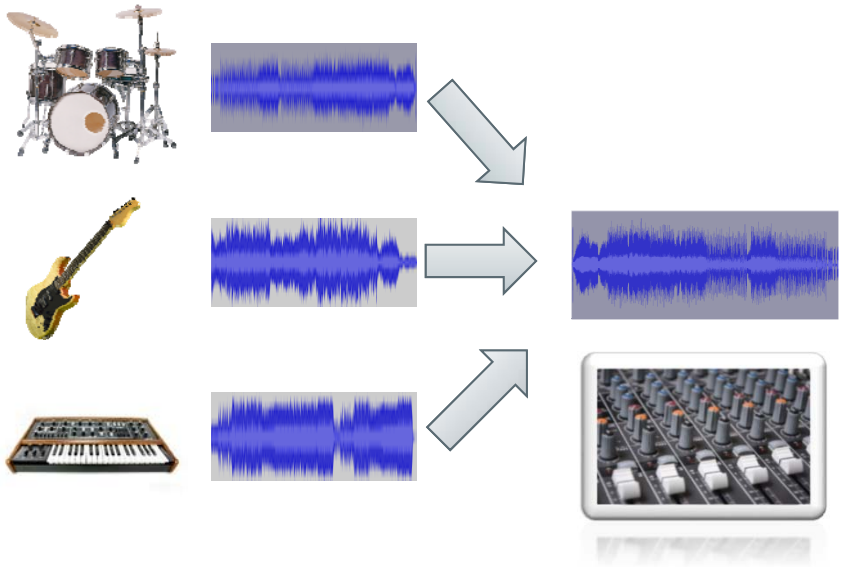
John Williams – Star Wars Main Theme



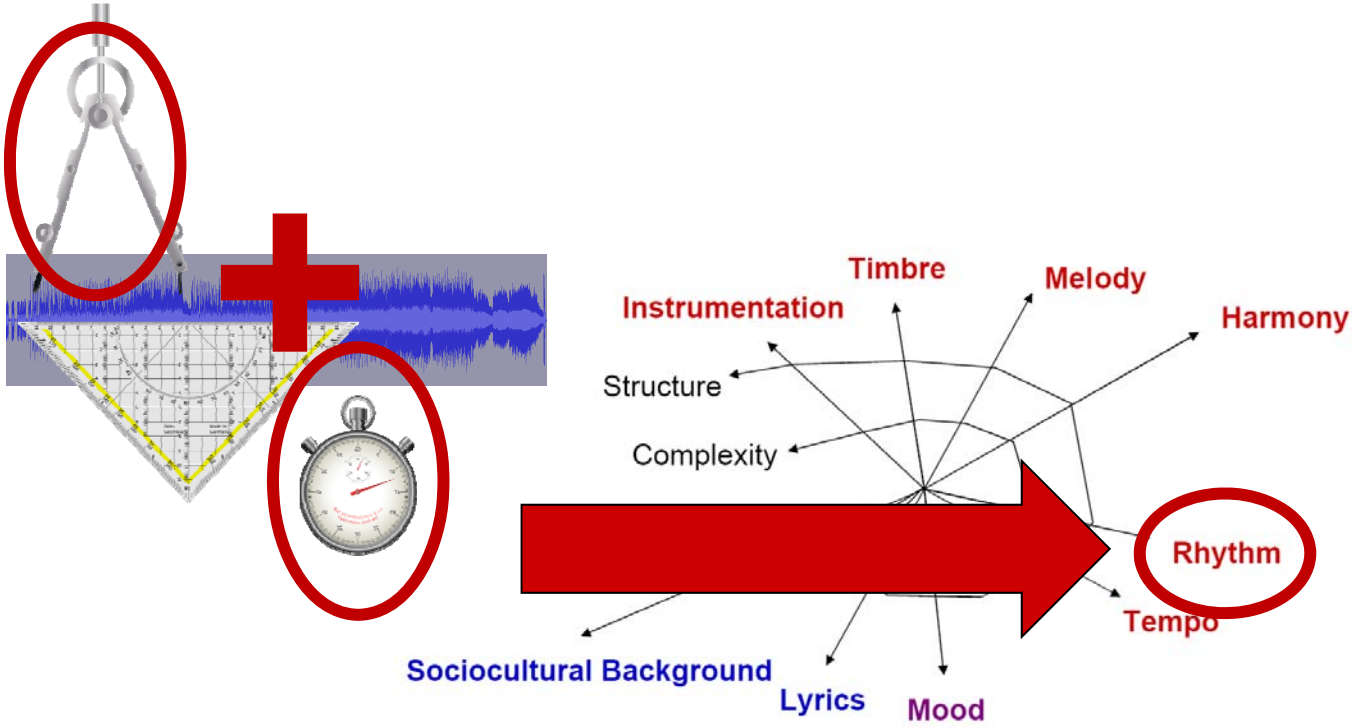
Rihanna feat. Calvin Harris – We Found Love

AUSTRIAN INSTITUTE OF TECHNOLOGY

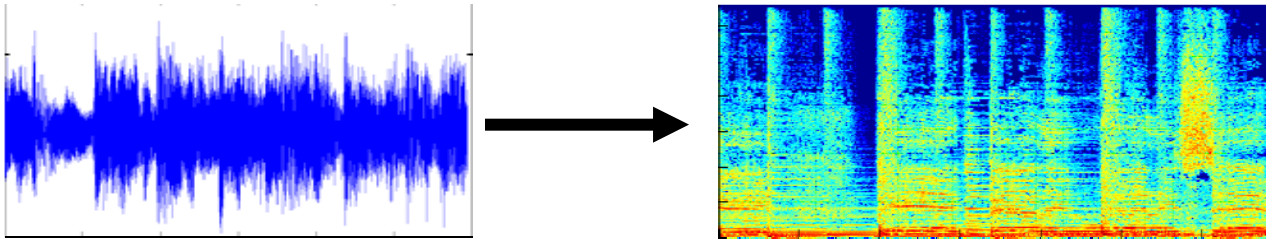
PROBLEM: SOURCE SEPARATION



AUDIO FEATURE EXTRACTION



SIGNAL PROCESSING



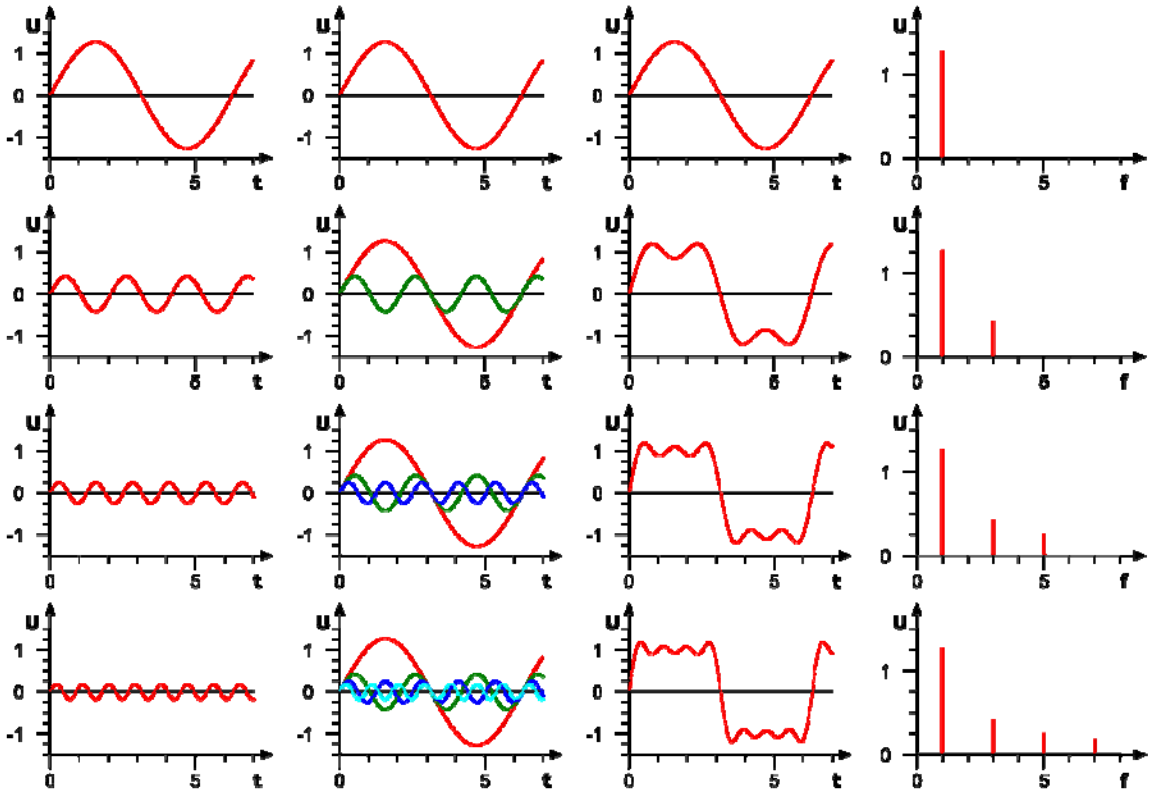
Time Domain
(„Wave Form“)

Frequency Domain
(„Spectrum“)

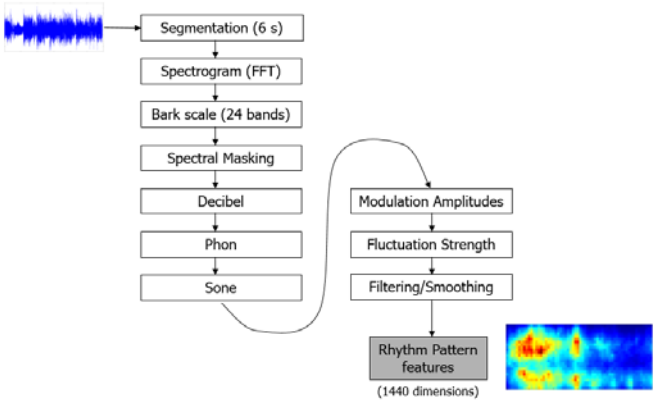
Time-Frequency Transformation

Fourier Transform (FFT)
Discrete Cosine Transform (DCT)
Wavelet Transform

FOURIER TRANSFORM

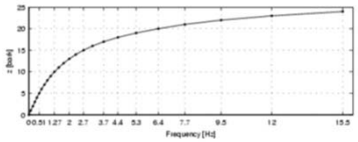


AUDIO/MUSIC FEATURE EXTRACTION



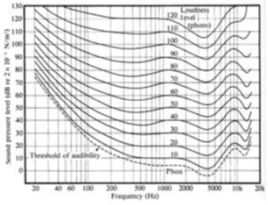
Bark Scale

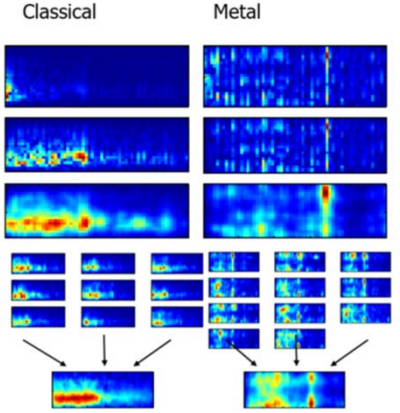
- psychoacoustical scale (related to Mel scale)
- 24 „critical bands" of hearing (non-linear)
- proposed by Eberhard Zwicker in 1961



Equal loudness curves (Phon)

- Relationship between sound pressure level in decibel and hearing sensation is not linear
- Perceived loudness depends on frequency of the tone
- equal loudness contours for 3, 20, 40, 60, 80, 100 phon





modulation amplitude spectrum ("cepstrum")

Fluctuation Strength

Filter (Gradient, Gauss)

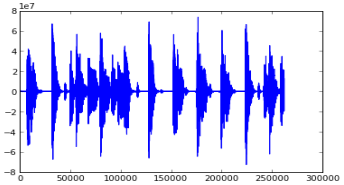
Median

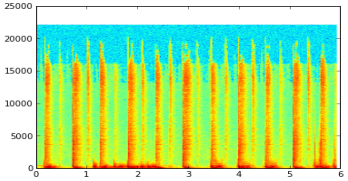
24*60=
1.440-dim feature vec.

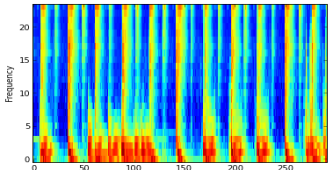
DEEP LEARNING FOR MUSIC IR

AUSTRIAN INSTITUTE OF TECHNOLOGY

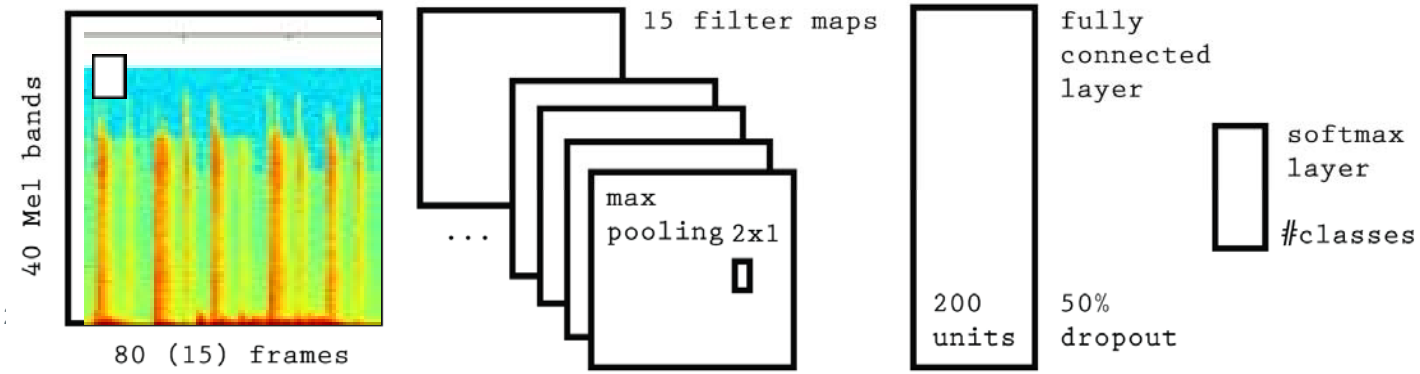
1. Pre-Processing: Waveform → Spectrogram → 40 Mel bands → Log scale







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



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

EXAMPLE 1

Audio Processing in Python

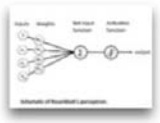
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

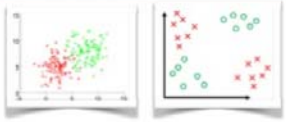
HISTORY OF NEURAL NETWORKS

From the Perceptron to Deep Learning

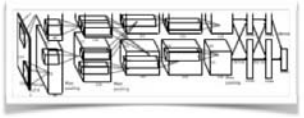
THREE EPOCHS OF NEURAL NETWORKS

	techniques / tricks	hardware	data
1957-69 dawn	perceptron	early mainframes	toy linear, small images, XOR
1986-95 golden age	early NNs	workstations	MNIST
2006- deep learning	deep NNs	GPU, TPU, Intel Xeon Phi	Imagenet





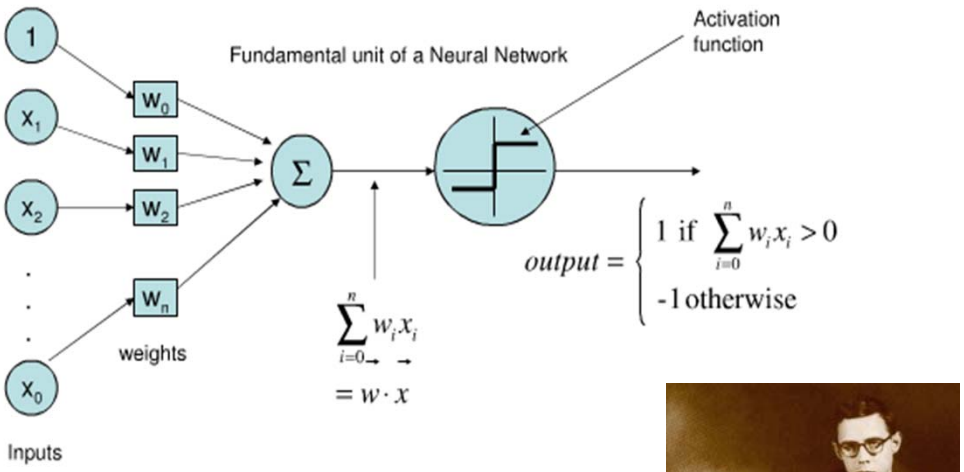


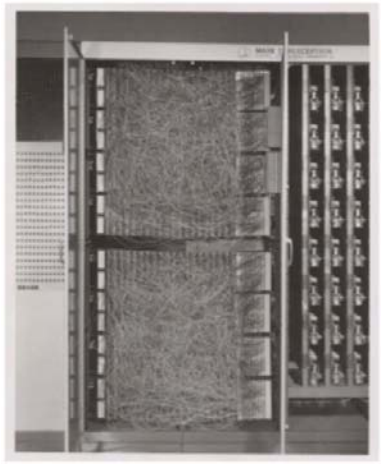




ORIGINS OF NEURAL NETWORKS

1958: Frank Rosenblatt: The Perceptron





Weights were encoded in potentiometers, and weight updates during learning were performed by electric motors.

UPS & DOWNS OF NEURAL NETWORKS

NNs and AI have experienced several hype cycles, followed by disappointment, criticism, and funding cuts:

1950s - 70s: Golden years of AI (funded by DARPA):
solve algebra, play chess & checkers, reasoning, semantic nets
"within ten years a digital computer will be the world's chess champion"

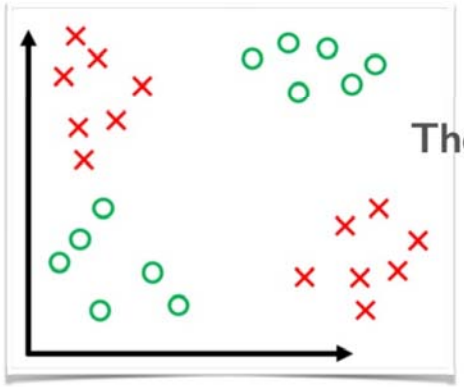
1969: shown that XOR problem cannot be solved by Perceptron
(led to the invention of multi-layer networks later on)

Mid 1970s: Chain reaction that begins with pessimism in the AI community,
followed by pessimism in the press, followed by a severe cutback in funding,
followed by the “end” of serious research (“AI winter”)

FIRST AI-WINTER

- Minsky-Papert (1969)

It is impossible to linearly separate the XOR function



The first winter: 1969 - 86

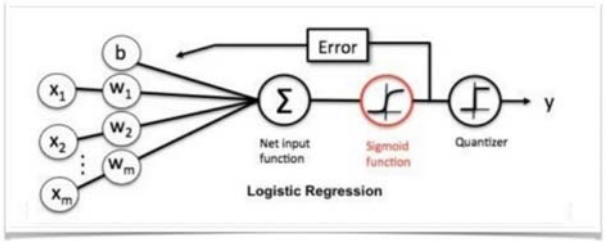
Is it?

UPS & DOWNS OF NEURAL NETWORKS

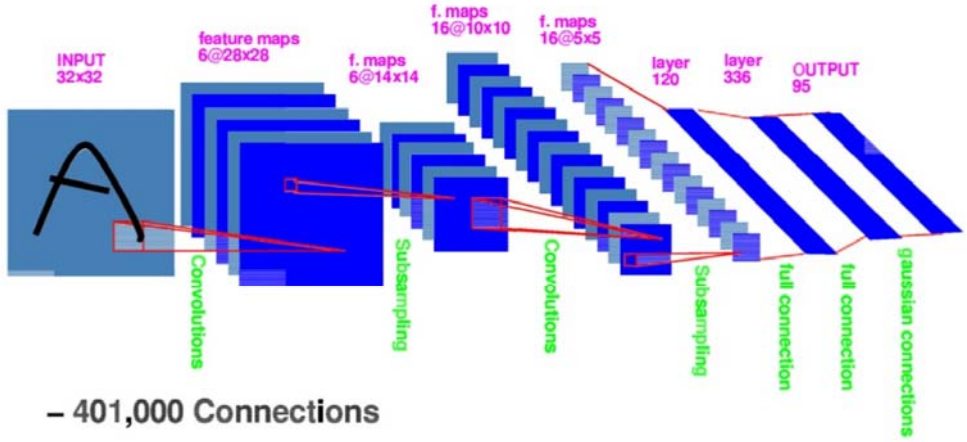
1980s: Governments (starting in Japan) and industry provide AI with billions of dollars. Boom of “expert systems”.

1986: Backpropagation had been invented in the 1970s, but only 1986 it became popular through a famous paper by David Rumelhart, Geoffrey Hinton, and Ronald Williams. It showed that also complex functions became solvable through NNs by using multiple layers.

Late 1980s: Investors - despite actual progress in research - became disillusioned and withdrew funding again.



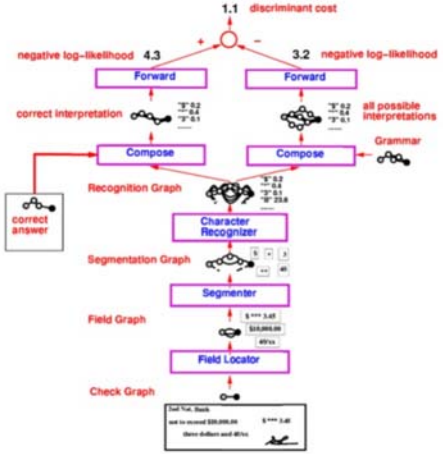
AT&T CHECK READER (LECUN, BENGIO, 1996)



– 401,000 Connections

– 100,000 free parameters

– Trained with 500,000 character samples
(Full ASCII set, machine printed and handwritten)



WHY NO DEEP LEARNING IN THE 1980S?

Neural Networks could not become “deep” yet - because:

- Computers were slow. So the neural networks were tiny and could not achieve (the expected) high performance on real problems.
- Datasets were small. There were no large datasets that had enough information to constrain the numerous parameters of (hypothetical) large neural networks.
- Nobody knew how to train deep nets. Today, object recognition networks have > 25 successive layers of convolutions. In the past, everyone was very sure that such deep nets cannot be trained. Therefore, networks were shallow and did not achieve good results.

SECOND AI-WINTER

1991: Hornik proved 1 hidden layer network can model any continuous function
(universal approximation theorem)

1991/92 Vanishing Gradient: problem in multi-layer networks where training in front layers is slow due to backpropagation diminishing the gradient updates through the layers. Identified by Hochreiter & Schmidhuber who also proposed solutions.

1990s - mid 2000s:
Due to lack of computational power, interest in NNs decreased again and other Machine Learning models, such as Bayesian models, Decision Trees and Support Vector Machines became popular.

RESURRECTION OF DEEP LEARNING IN THE 2000S

2000s: Hinton, Bengio and LeCun (“The fathers of the age of deep learning”) join forces in a project

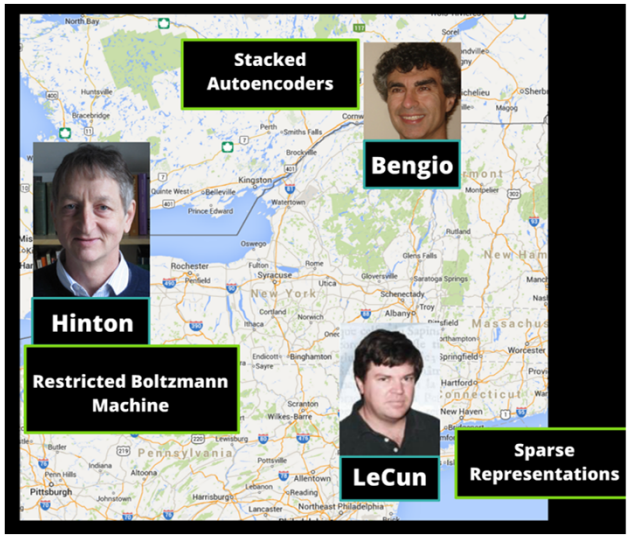
They overcome some problems that caused deep networks not to learn anything at all

2006: Breakthrough with Layer-wise pre-training by unsupervised learning (using RBMs)

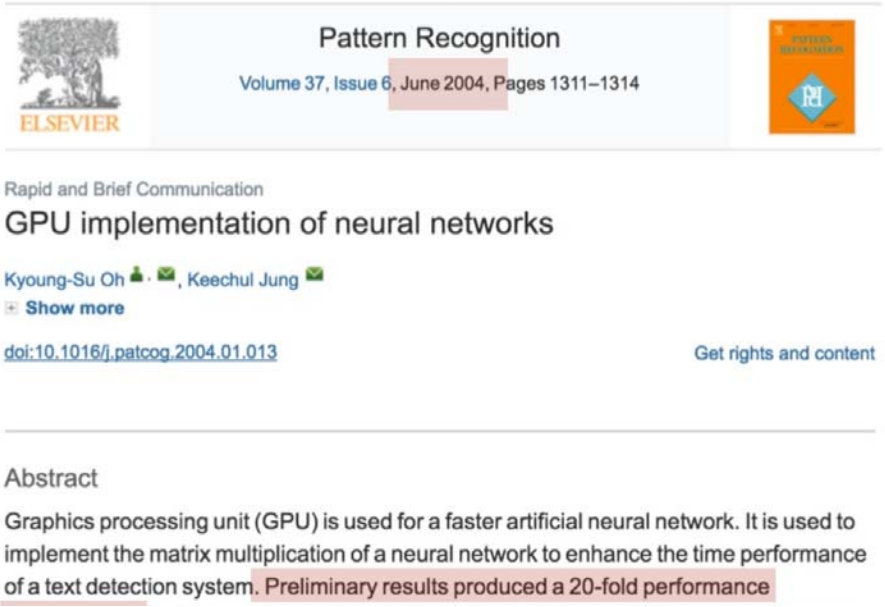
2010s: Important new contributions:

- Simpler initialization (without pre-training)
- Dropout
- Simplier activations: Rectifier Units (ReLUs)
- Batch Normalization

→ not a re-invention of NNs but paved the way for very deep NNs



GPUs (2004)



Pattern Recognition

Volume 37, Issue 6, June 2004, Pages 1311–1314

Rapid and Brief Communication

GPU implementation of neural networks

Kyoung-Su Oh, Keechul Jung

Show more

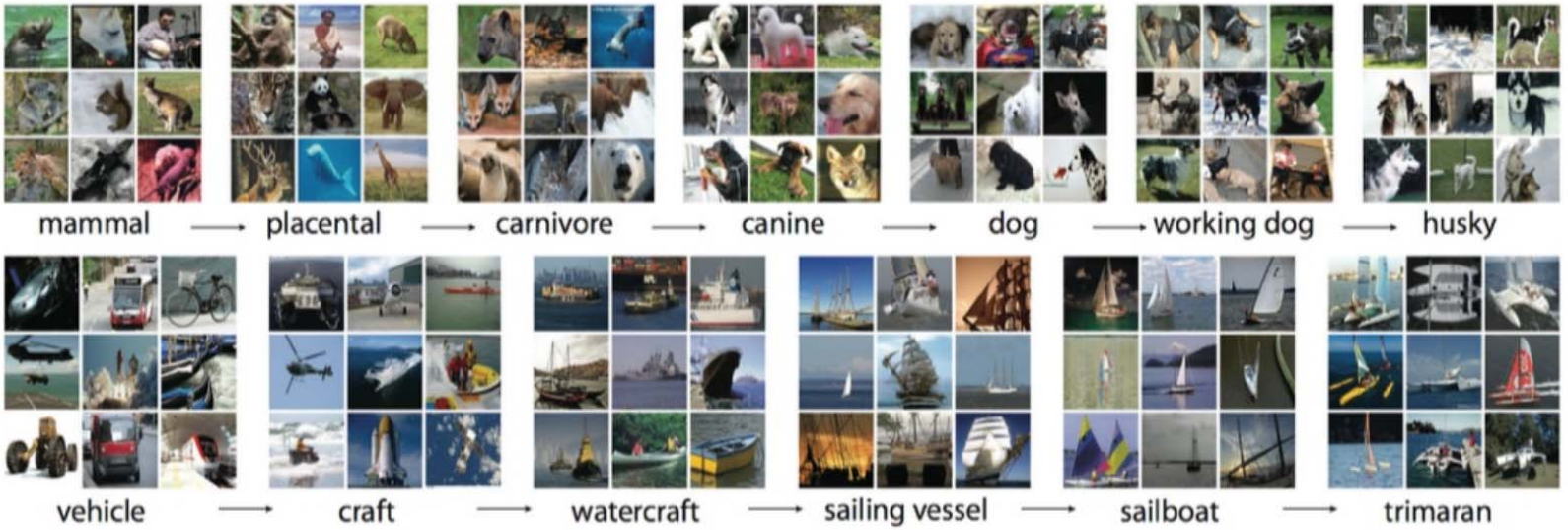
doi:10.1016/j.patcog.2004.01.013

Get rights and content

Abstract

Graphics processing unit (GPU) is used for a faster artificial neural network. It is used to implement the matrix multiplication of a neural network to enhance the time performance of a text detection system. Preliminary results produced a 20-fold performance

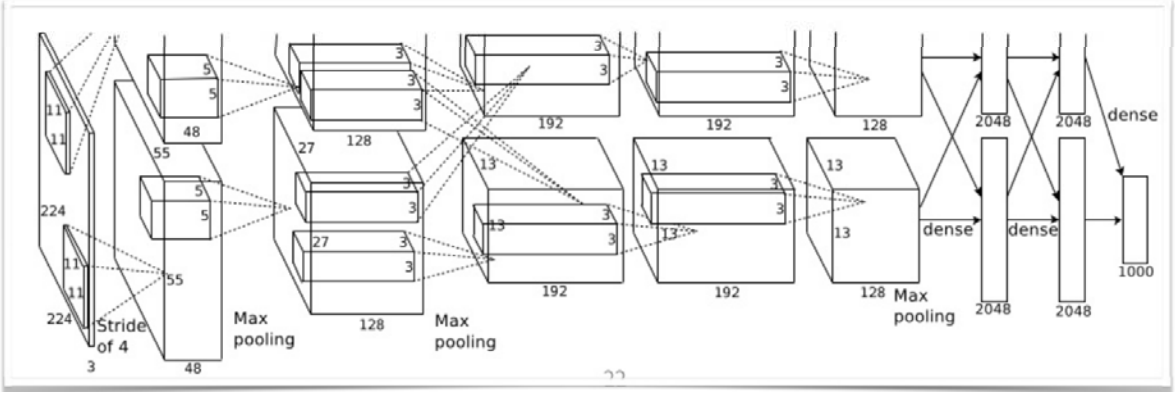
IMAGENET (2009)



ALEXNET (2012)

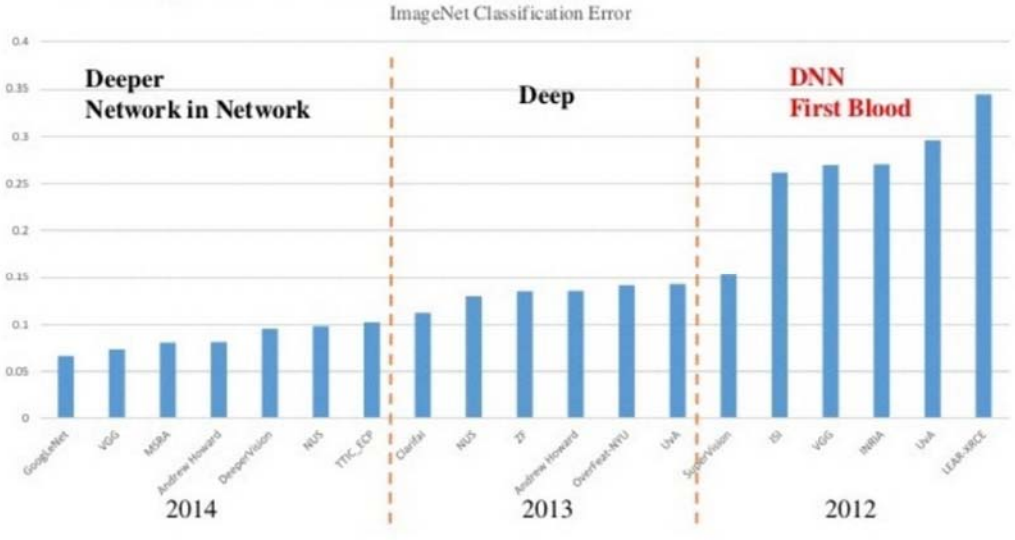
- dropout, ReLU, max-pooling, data augmentation, batch normalizatio automatic differentiation, end-to-end training, lots of layers

- Krizhevsky, Sutskever, Hinton (2012): 1.2M images, 60M paramete 6 days training on two GPUs



DROPPING ERROR RATES SINCE THEN

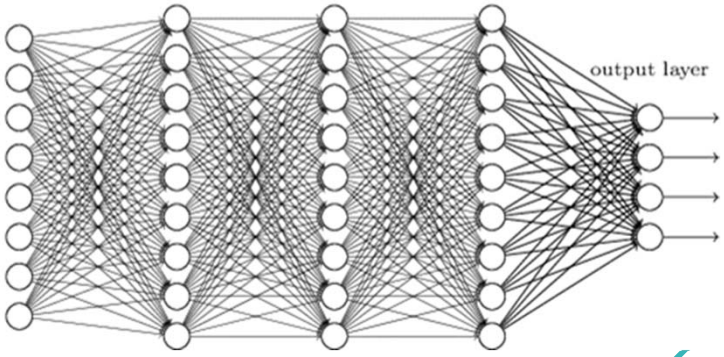
- **1000** categories and **1.2** million training images



Li Fei-Fei: ImageNet Large Scale Visual Recognition Challenge, 2014 http://image-net.org/

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

WHAT IS DEEP LEARNING?



WHAT ARE ARTIFICIAL NEURAL NETWORKS?

a fancy name for particular mathematical expressions, such as:

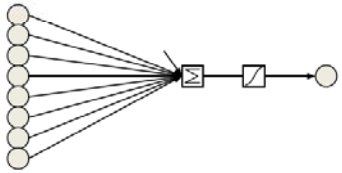
$$y = \sigma(b + \mathbf{w}^T\mathbf{x})$$ (equivalent to logistic regression)

WHAT ARE ARTIFICIAL NEURAL NETWORKS?

a fancy name for particular mathematical expressions, such as:

$y = \sigma(b + \mathbf{w}^T\mathbf{x})$ (equivalent to logistic regression)

expression can be visualized as a graph:

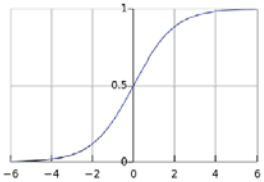


$\mathbf{x}$ $b + \mathbf{w}^T\mathbf{x}$ y

Output value is computed as a **weighted sum of its inputs,**

$b + \mathbf{w}^T\mathbf{x} = b + \Sigma_i w_i x_i$

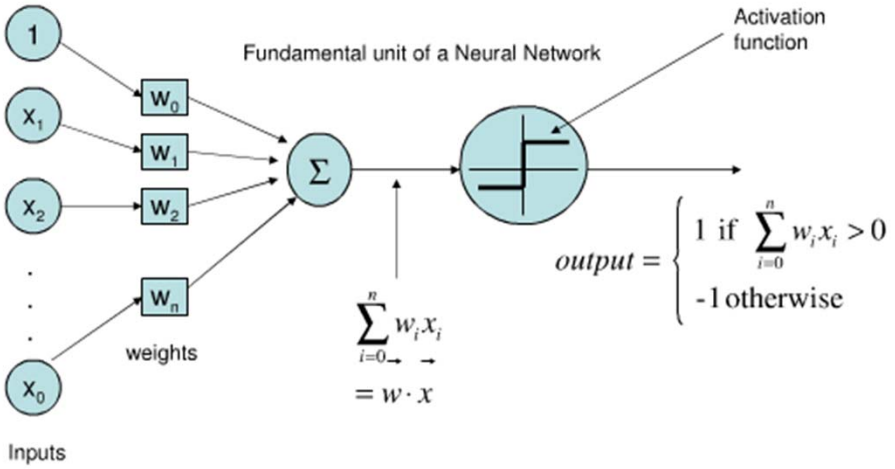
followed by a nonlinear function.



Origins of **Neural Networks**

AUSTRIAN INSTITUTE OF TECHNOLOGY

1958: Rosenblatt: The Perceptron



Linear binary classifier using a step function

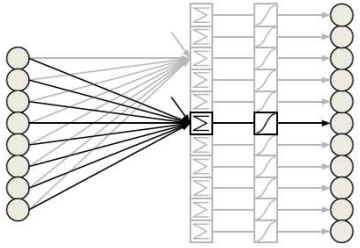
For the first time a NN could solve simple classification problems merely from training data

WHAT ARE ARTIFICIAL NEURAL NETWORKS?

a fancy name for particular mathematical expressions, such as:

$$\mathbf{y} = \sigma(\mathbf{b} + \mathbf{W}^T\mathbf{x}) \qquad \text{(multiple logistic regressions)}$$

expression can be visualized as a graph:

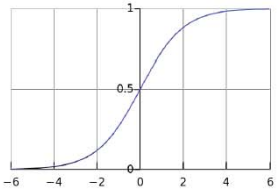


$\mathbf{x}$ $\mathbf{b} + \mathbf{W}^T\mathbf{x}$ $\mathbf{y}$

Output values are computed as **weighted sums of their inputs,**

$$\mathbf{b} + \mathbf{W}^T\mathbf{x} = b_j + \Sigma_i w_{ij} x_i$$

followed by a nonlinear function.

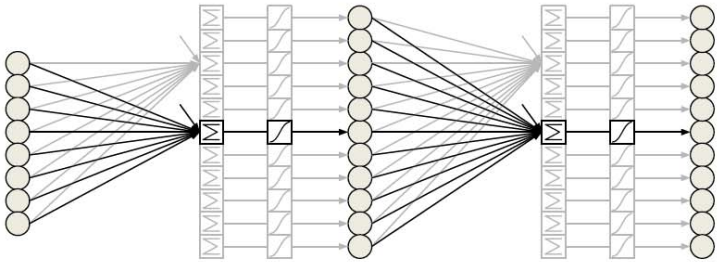


WHAT ARE ARTIFICIAL NEURAL NETWORKS?

a fancy name for particular mathematical expressions, such as:

$$\mathbf{y} = \sigma(\mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\sigma(\mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{x})) \qquad \text{(stacked logistic regressions)}$$

expression can be visualized as a graph:



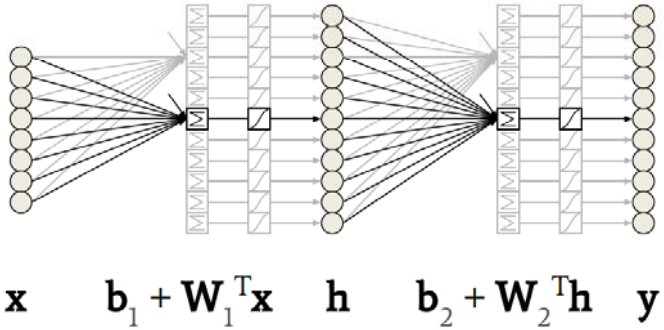
$\mathbf{x}$ $\quad \mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{x}$ $\quad \mathbf{h}$ $\quad \mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\mathbf{h}$ $\quad \mathbf{y}$

WHAT ARE ARTIFICIAL NEURAL NETWORKS?

a fancy name for particular mathematical expressions, such as:

$$\mathbf{y} = \sigma(\mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\sigma(\mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{x})) \quad \text{(stacked logistic regressions)}$$

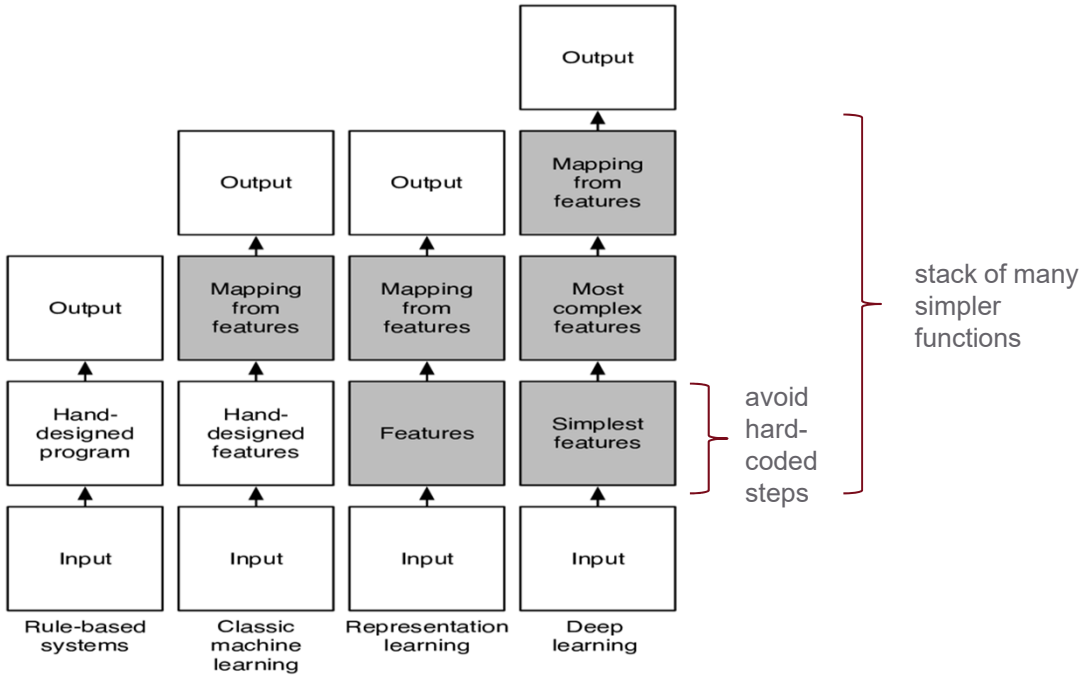
expression can be visualized as a graph:



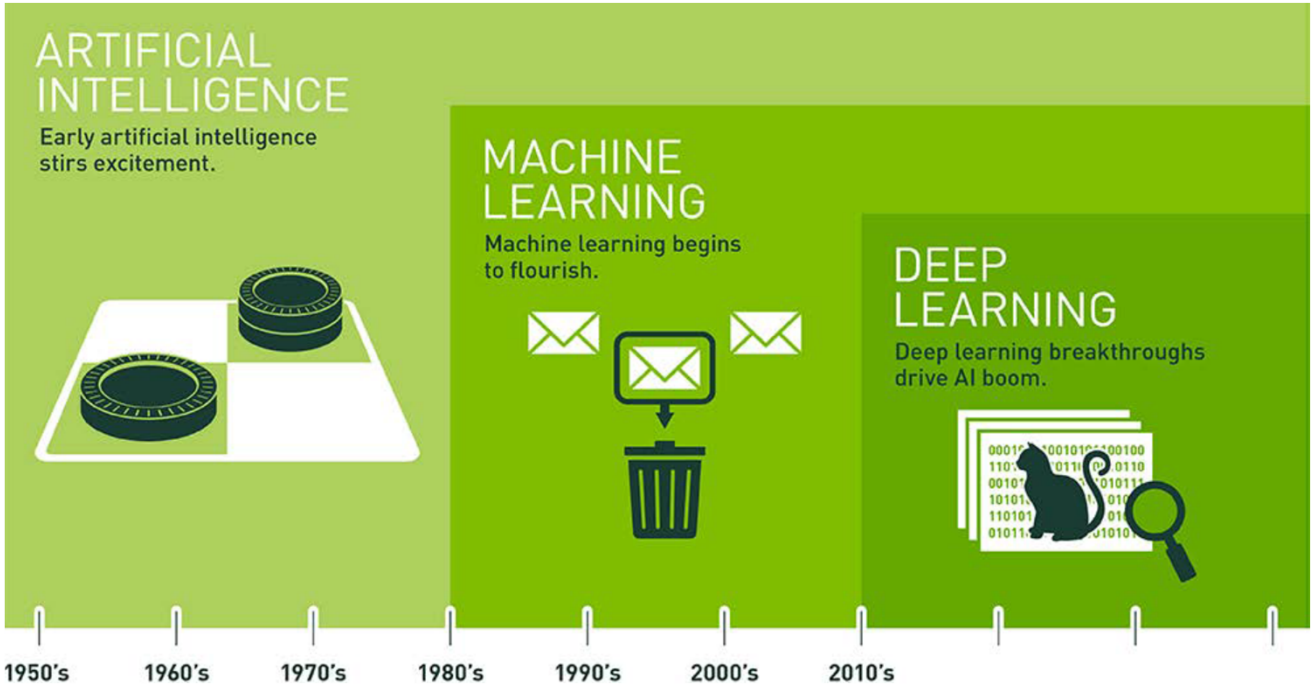
Universal Approximation Theorem:
This can model any continuous function from $\mathbb{R}^n$ to $\mathbb{R}^m$ arbitrarily well (if $\mathbf{h}$ is made large enough).

TERMINOLOGY – DEEP LEARNING

MACHINE LEARNING PARADIGMS



TERMINOLOGY – ARTIFICIAL INTELLIGENCE



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

ADVANTAGES OF DEEP LEARNING

Why it is not a hype

DETACH FROM DOMAIN KNOWLEDGE

- Problems of hand-crafted features
  - Need for expert knowledge
  - Time-consuming and expensive
  - Solution and knowledge does not generalize to other domains

GENERAL MACHINE LEARNING PRINCIPLE

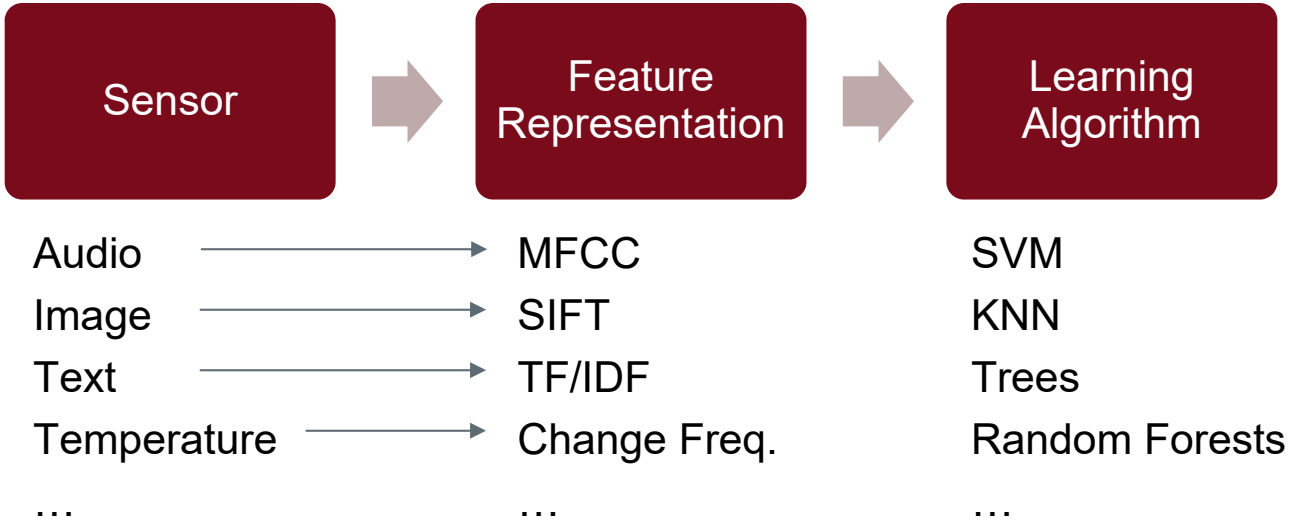
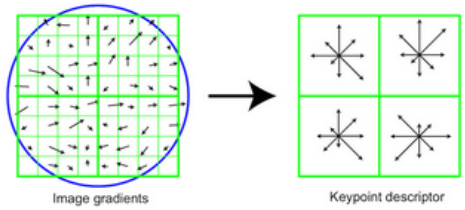
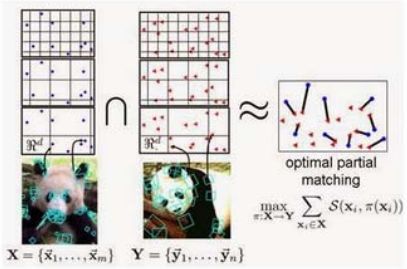


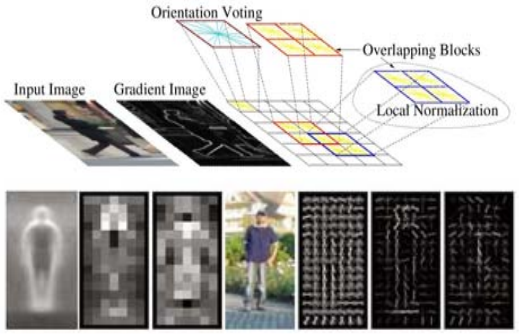
IMAGE PROCESSING



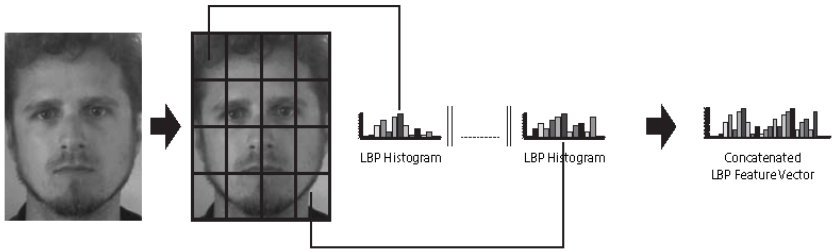
The SIFT recipe: gradient orientations, normalization tricks



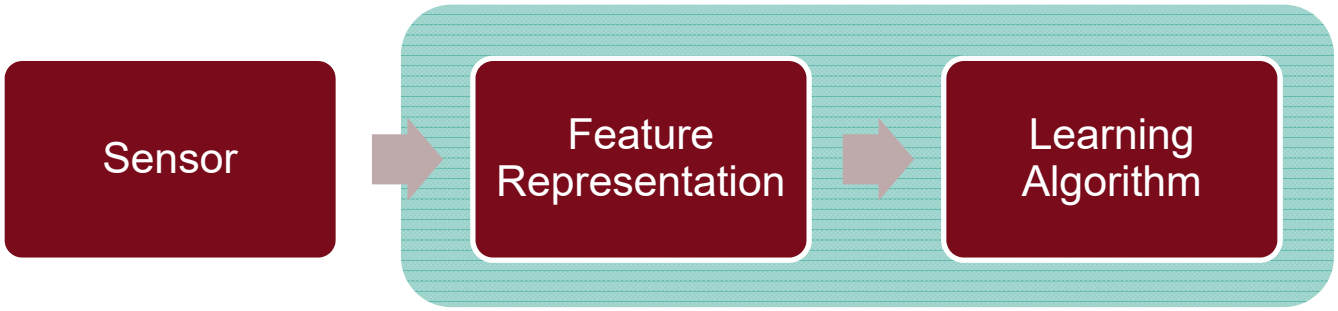
Grauman's Pyramid Match Kernel for Improved Image Matching



Navneet Dalal's HOG Descriptor

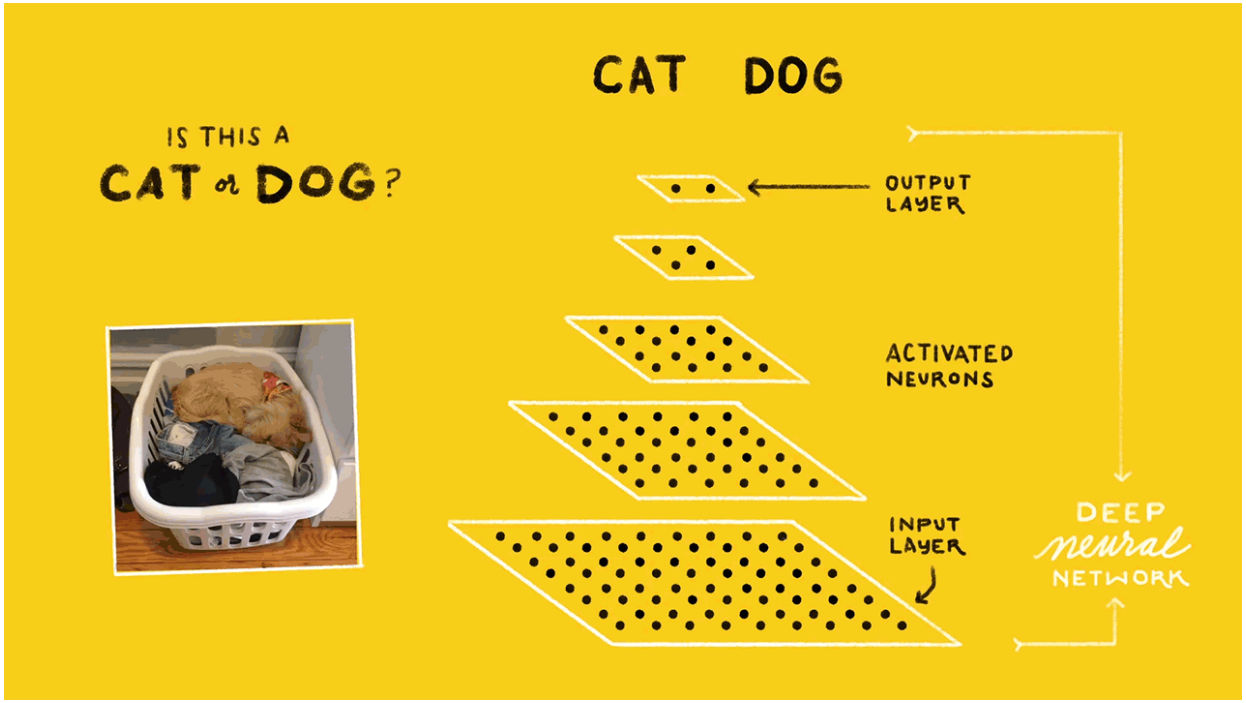


DEEP LEARNING - ADVANTAGE



- Feature representation ➔ learned from data
- Implicit Learning Algorithm
  - Reduces complexity of ML workflow

IMAGE PROCESSING



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

EXAMPLE 2

Defining and training a basic Deep Neural Network

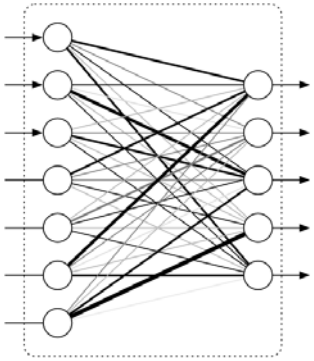
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

NEURAL NETWORKS CONCEPTS

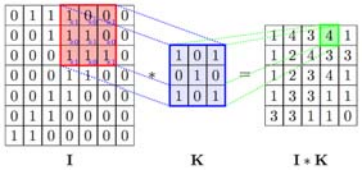
Layers, Activation- and Loss-Functions, Optimization, Overfitting

GENERAL TYPES OF LAYERS

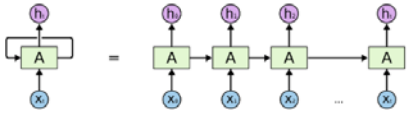
- Fully Connected



- Convolutional

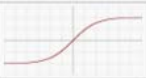
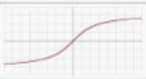


- Recurrent



ACTIVATION FUNCTIONS

- Decides if a neuron should be „fired“
- Adds non-linearity to the model
- Output layers
  - Binary Classificatoin ➔ Sigmoid
  - Multiclass classification ➔Softmax
  - Multilabel classification ➔ Sigmoid

Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$

LOSS FUNCTIONS

- Loss
  - Used to guide training process
- Loss functions
  - Mean squared error $MSE := \frac{1}{n}\sum_{t=1}^{n} e_t^2$
  - Cross Entropy Loss $H_{y'}(y) := -\sum_i y'_i \log(y_i)$
- Task dependancy
  - Multi-class classification ➔ categorical crossentropy
  - Binary/Multi-label classification ➔ binary crossentropy
  - Regression ➔ Mean squared error, L2
  - Siamese Networks ➔ Tripplet Loss, Contrastive Loss

Tutorial:
Deep Learning for Music Classification using Keras

TUTORIAL AGENDA

AUSTRIAN INSTITUTE OF TECHNOLOGY

I. Deep Learning Basics:

- Audio Processing Basics
- History of Neural Networks
- What is Deep Learning
- Neural Network Concepts
- Coding Examples

II. Convolutional Neural Networks:

- Difference CNN – RNN
- How CNNs work (Layers, Filters, Pooling)
- Application Domains and how to use in Music
- Coding Examples