Deep Learning

Paper by: LeCun, Bengio, Hinton

Deep Learning

- Methods and practices focusing on machine learning with deep neural network architectures
- learns hierarchical representations through multiple layerwise non-linear units
- NNs can in theory approximate an arbitrarily complex function
- State-of-the-Art in image and speech recognition and machine translation (NLP in general)

Drawbacks of conventional ML-Algorithms

- Conventional algorithms limited in processing raw data
  → careful feature engineering and domain expertise needed
- Linear Classifiers can only categorize regions distinguished by a hyperplane
- Tasks like Image and speech recognition requires the classifier to be insensitive to variation in position, orientation or illumination

Feedforward Neural Networks

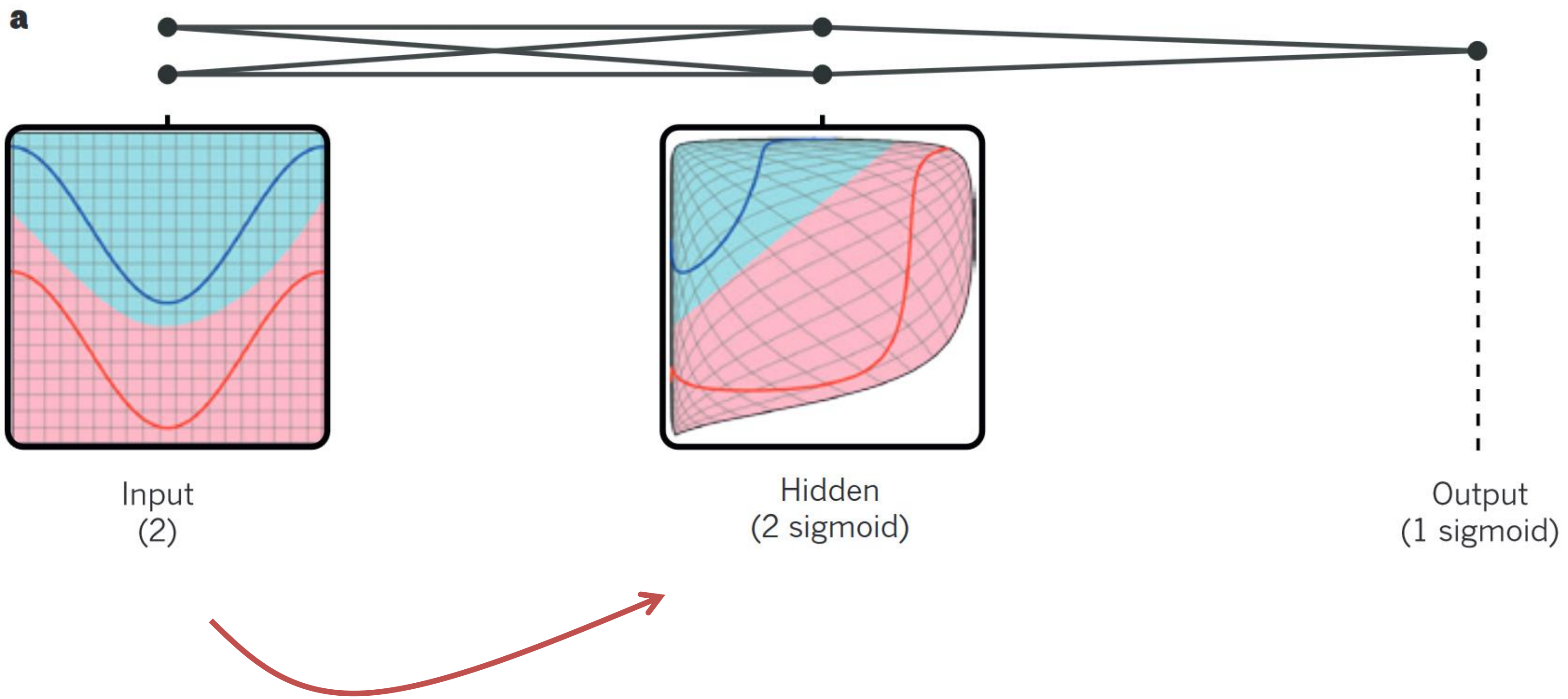
- Mapping of fixed-size Input to fixed-size Output

$$f_\theta : X \rightarrow Y, \qquad where\ X \in \mathbb{R}^m\ and\ Y \in \mathbb{R}^n$$

- multilayered stack of non-linear units
  – Sigmoid, tanh or ReLU activation function

- Universal Approximation of continuous functions
  – Linear Output Layer and min. 1 hidden Layer with non-linear activations and sufficient hidden units
  – Universal approximation theorem (Hornik et al., 1989)

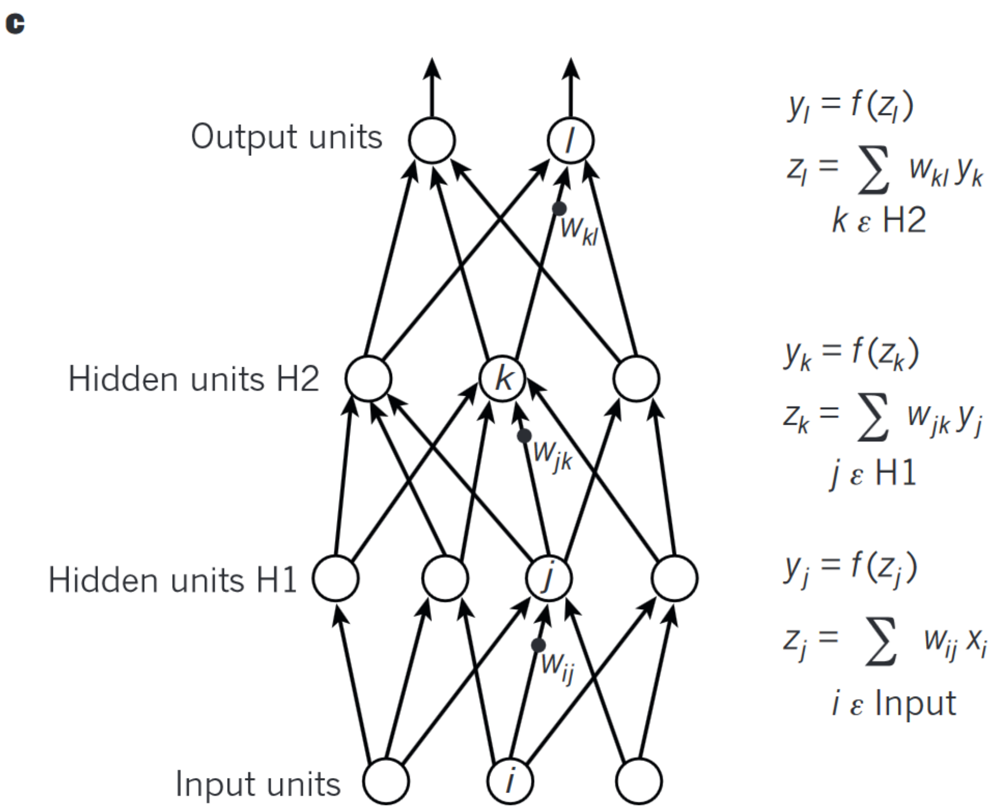
Feedforward Neural Networks

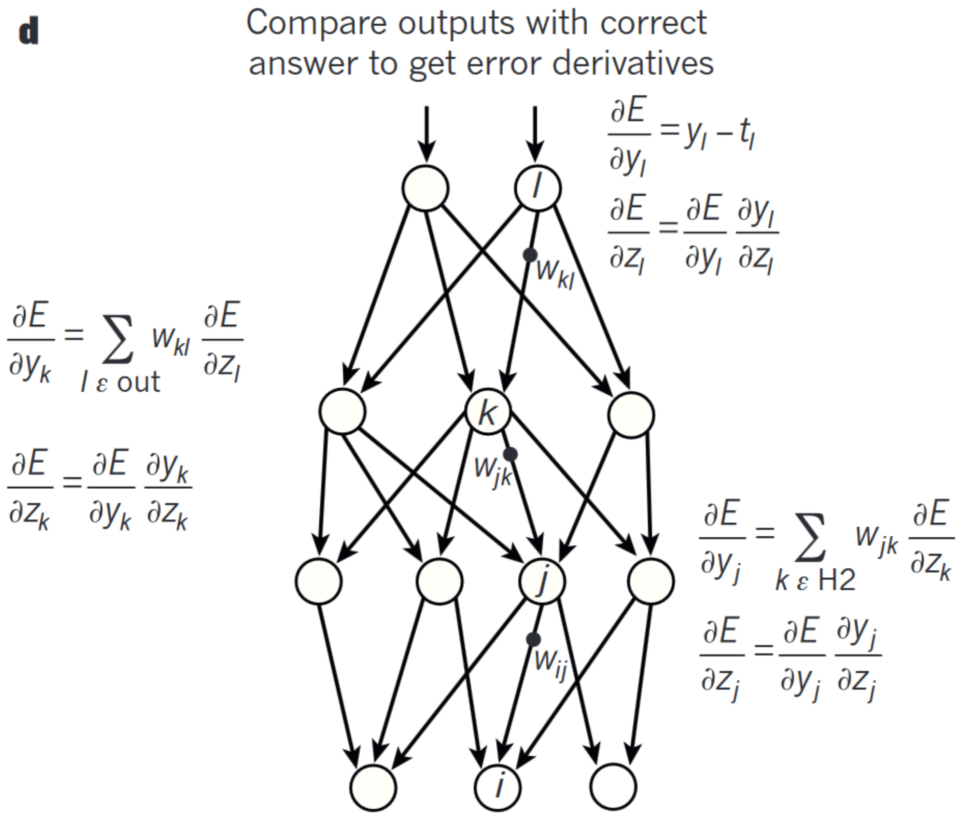


Distortion of input space allows linear separability

(LeCun et al., 2015)

Feedforward Neural Networks





(LeCun et al., 2015)

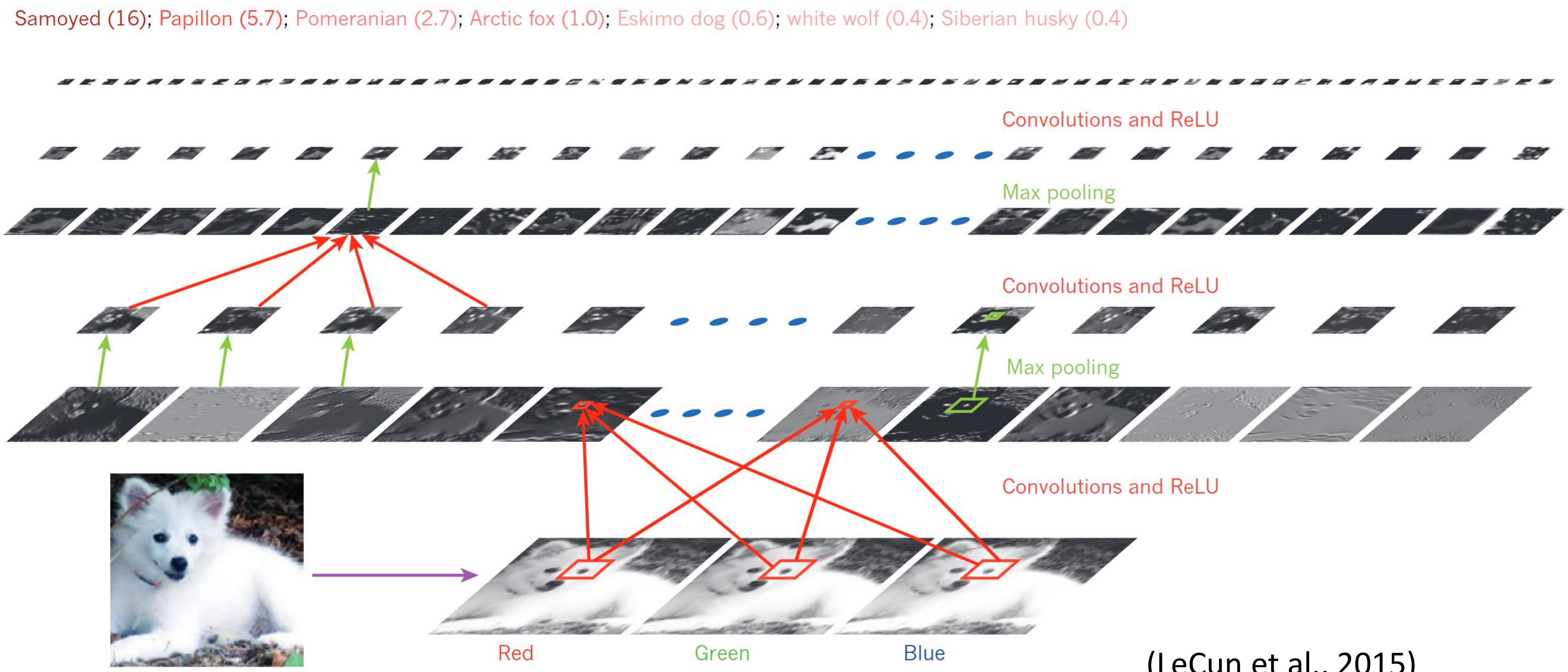
forward computation to calcultate outputs

backwards computation to obtain gradients

Backpropagation

- Neural Architectures trained through Stochastic Gradient Descent (SGD)
- Backpropagation used to compute the Gradients
- SGD previously forsake due to the believe, that optimization would get stuck in poor local minima
- In practice rather large landscape of saddle points with similar quality

Convolutional Neural Networks



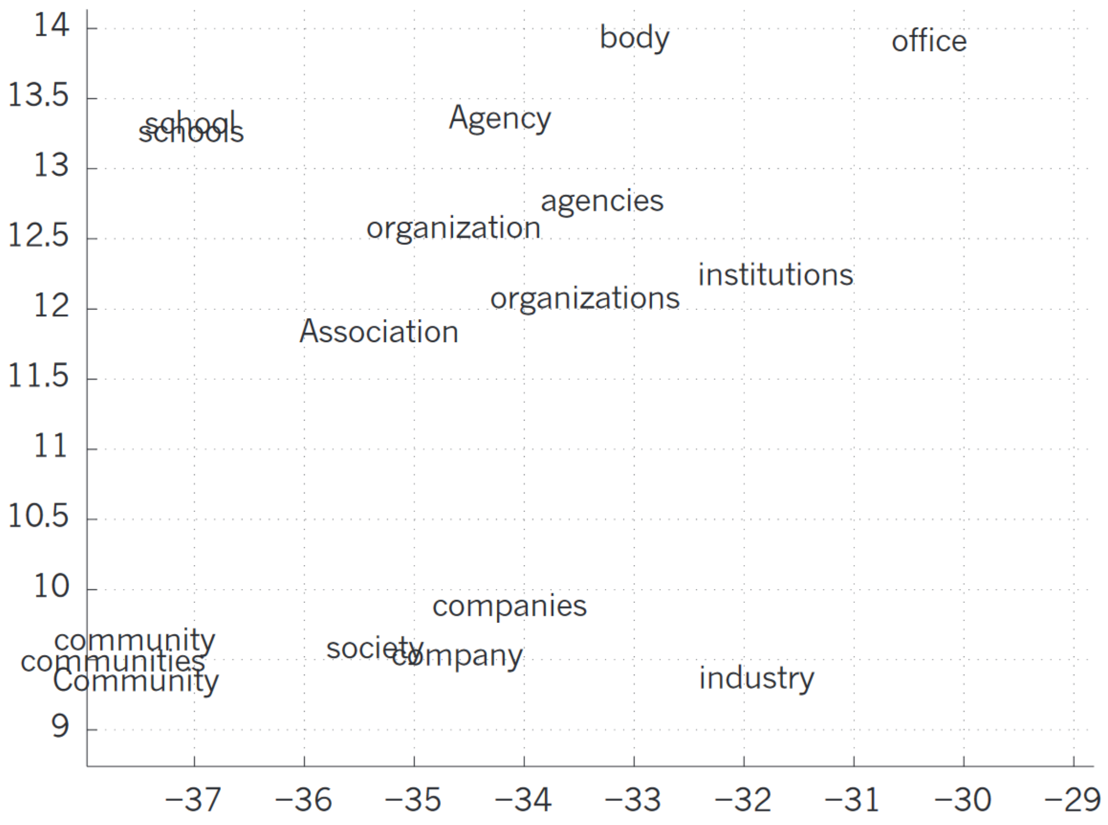
(LeCun et al., 2015)

Convolutional Neural Networks

- Processing multi-dimensional input structures (e.g. images)
- Multiple convolutional kernels extract feature maps
  → detection of distinctive local motives invariant of location
- Pooling merges semantically similar features
  →invariance to small shifts and distortion
- Great success for detection, segmentation or recognition of objects but also in NLP or speech recognition

Distributed representations and Language Processing

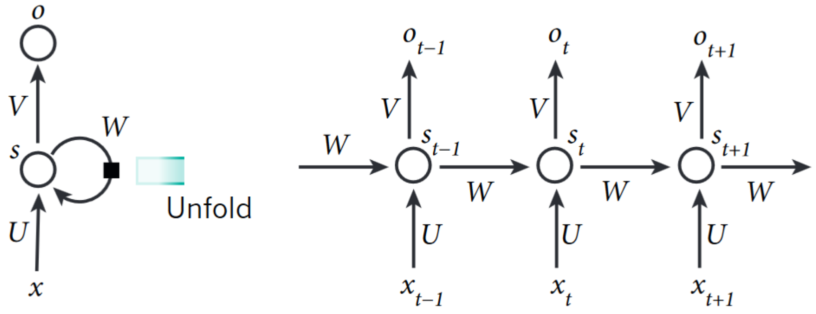
- E.g. vector representation of words in NLP
- captures semantic features not implicitly present in the input
- enables generalization to new combinations



(LeCun et al., 2015)

Recurrent Neural Networks

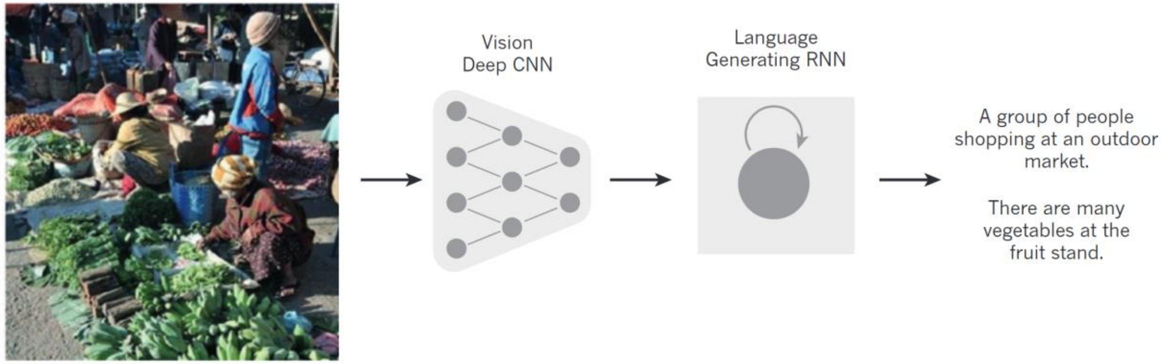
- Computationally more powerful and biologically more plausible
- Continuous internal states

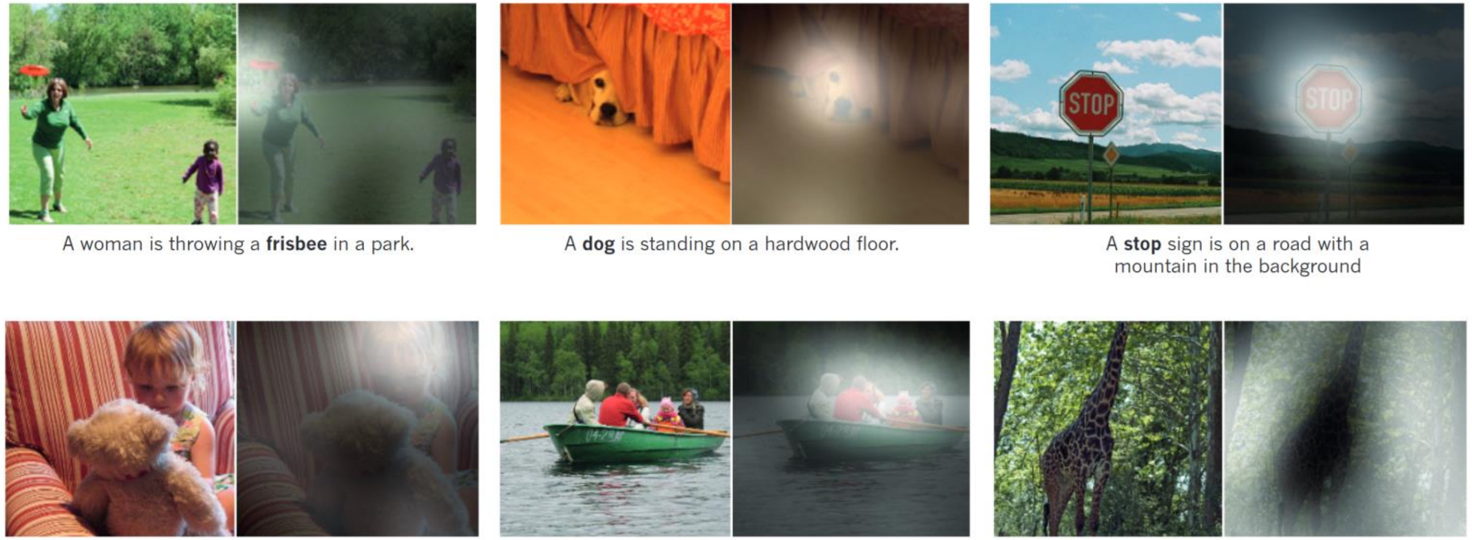


(LeCun et al., 2015)

- Long Short-Term Memory (LSTM) Networks
  - Recognition of temporally extended patterns in noisy sequences
  - Recognition of temporal order of widely separated events in noisy input
  - Extraction of information conveyed by the temporal distance between events

From Image to Text





A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

(LeCun et al., 2015)

Criticism by Jürgen Schmidhuber

- DL was already practically used by Ivankhnenko et al. in 1965
- Paper exclusively promotes CIFAR research as main driver revoking interest in deep neural networks in 2006
- Similar work before 2006 is mostly ignored in the paper
- Earlier deep network approaches have not been commercially successful due to the lack of fast processing resources

Source: people.idsia.ch/~juergen/deep-learning-conspiracy.html

References

- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521.7553 (2015): 436-444.
- Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." *Neural networks* 2.5 (1989): 359-366.