Deep Learning by LeCun, Bengio and Hinton

A journey of what, how and why Deep Learning

Presenter: Tai D. Nguyen
Drexel University, Spring 2020
BMES T580

Outline

1. Why Deep Learning?
2. What is Deep Learning?
3. Convolutional Neural Network
4. Distributed Representation
5. Recurrent Neural Network
6. Unsupervised Learning & Future of Deep Learning

Why Deep Learning?

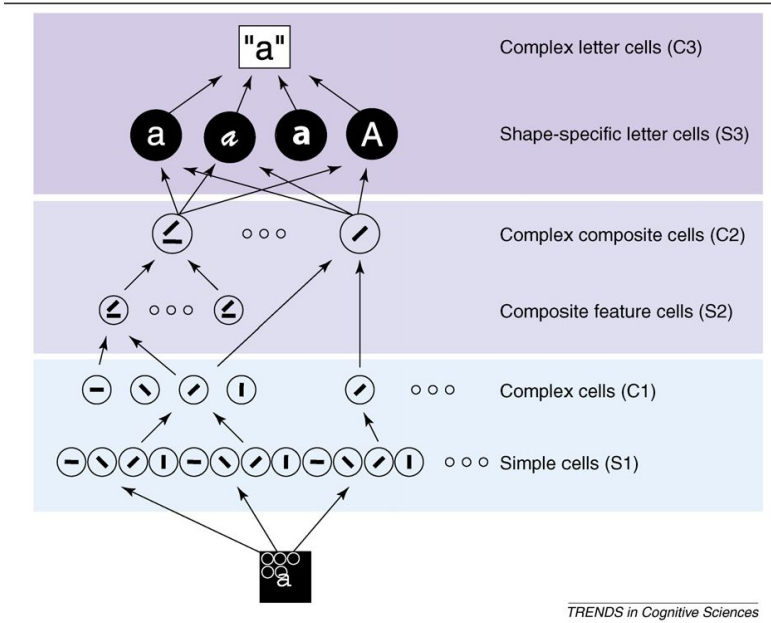
Alan Turing

Alonzo Church

The Universality of Computation

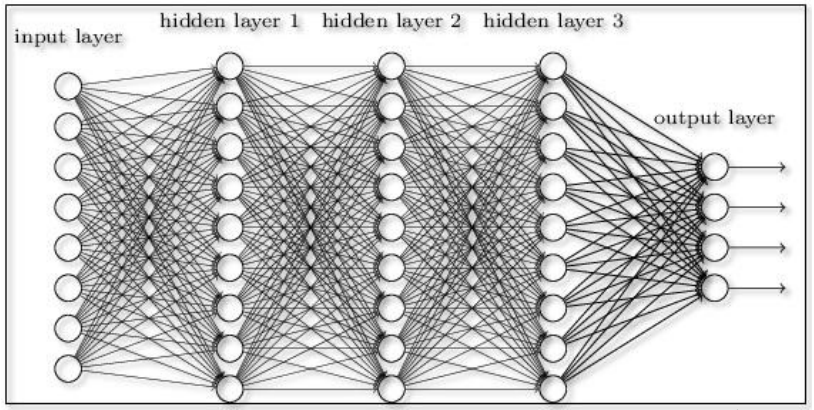
If a problem that can be presented to a Turing machine is not solvable by it, it is also not solvable by any machine; and if any machine is “Turing complete”, then it can compute any algorithm.

Why Deep Learning?



What is Deep Learning?

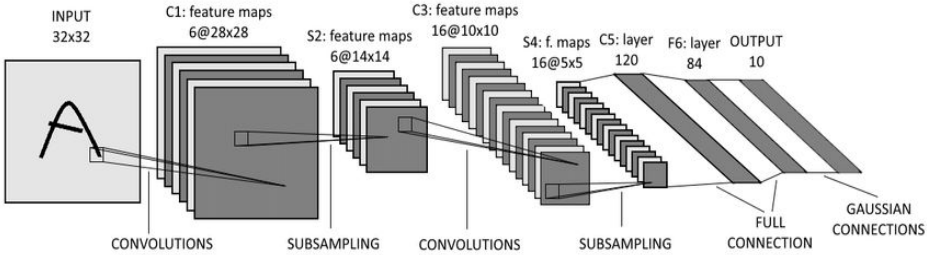
Deep Neural Network (DNN) Model



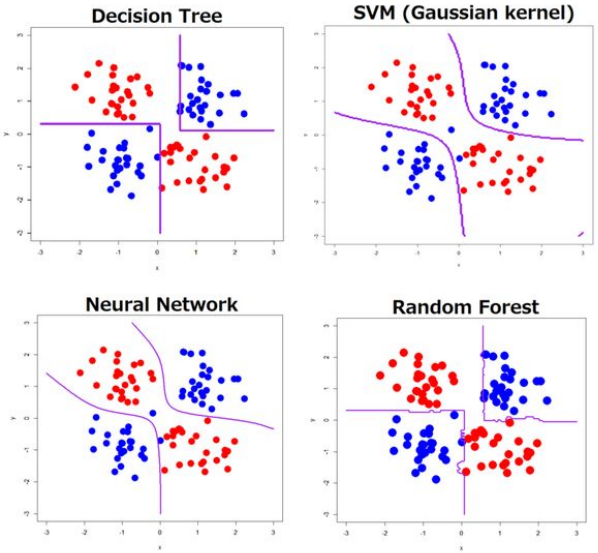
Forward = transformation in the previous layer + nonlinearity + subsampling

Backward = Back-propagation

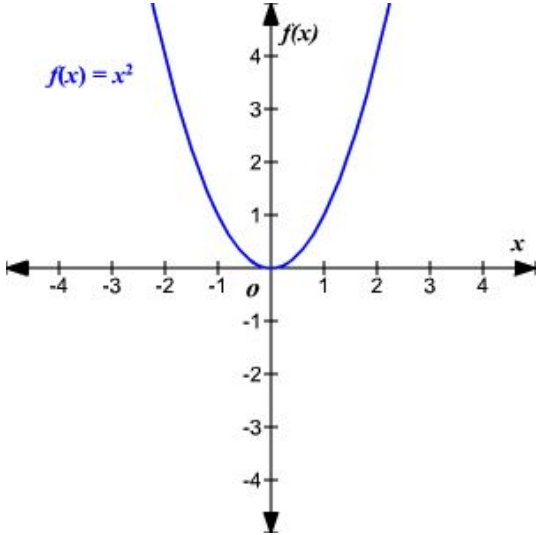
Example: LeNet-5 by LeCun



What is Deep Learning? Objective Function



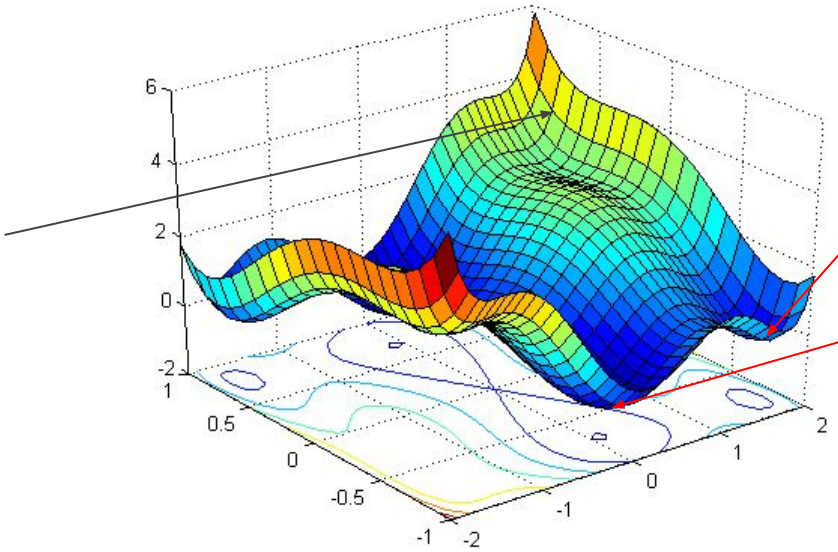
Decision Boundaries drawn by different algorithms



Simple objective function, f=x^2

What is Deep Learning? Stochastic Gradient Descent

Initialized here

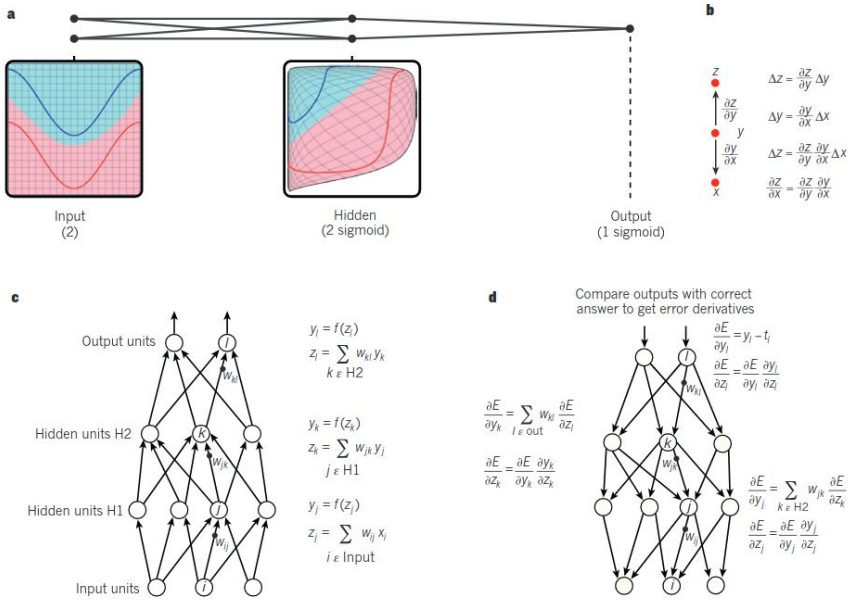


Local minimum here

Global minimum here

A complex objective function. Z-axis is error, X&Y-axis are parameters

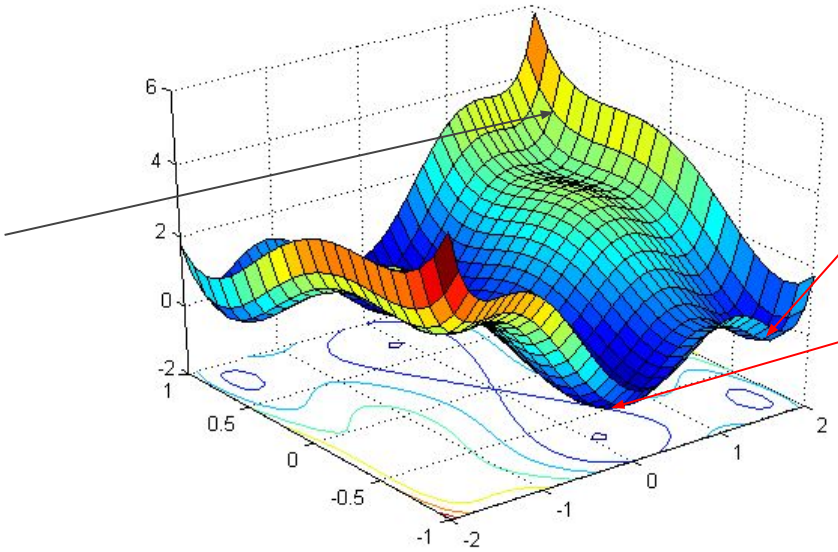
What is Deep Learning? Backpropagation



How backpropagation works mathematically

What is Deep Learning? Stochastic Gradient Descent

Holding a steel ball and drop it here!



Local minimum here

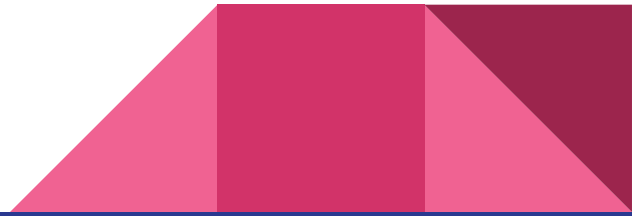
Global minimum here

A complex objective function. Z-axis is error, X&Y-axis are parameters

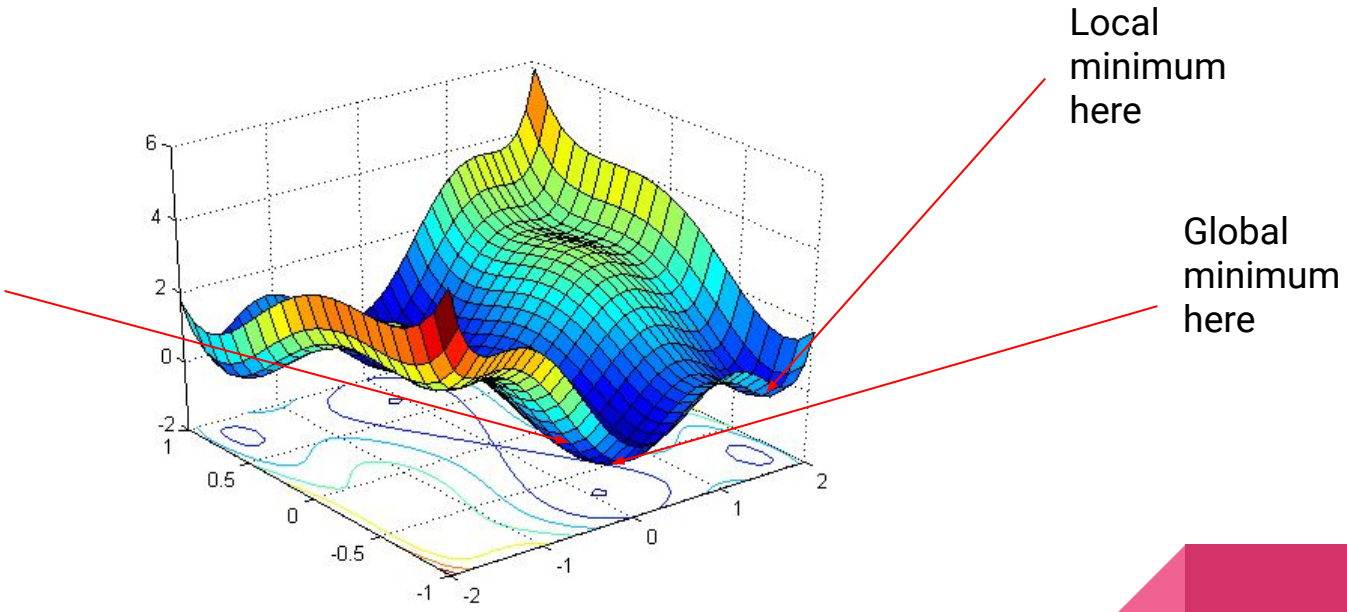
What is Deep Learning? Overfitting & Overgeneralization

Overfitting = Very low error in training but high error in testing

Overgeneralization = High error in training and high error in testing



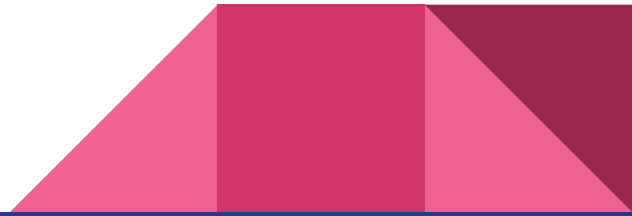
What is Deep Learning? Overfitting & Overgeneralization



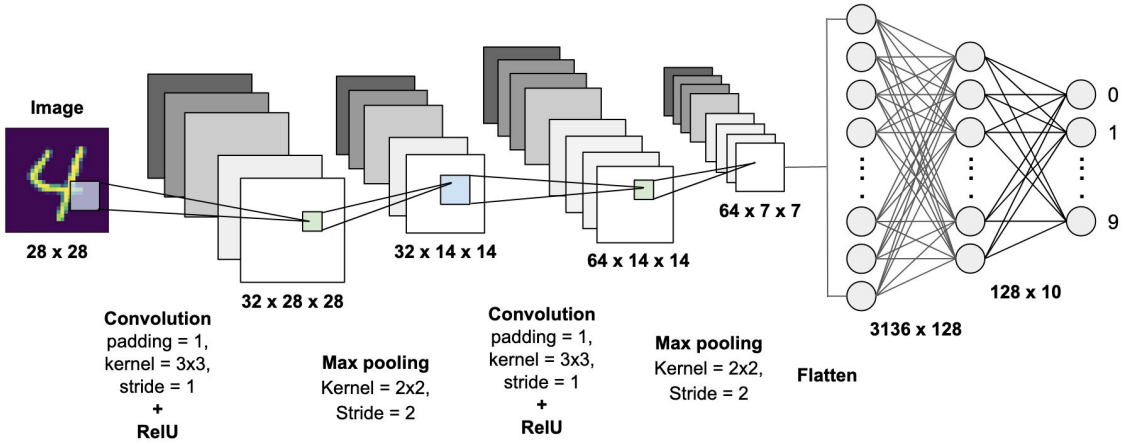
A complex objective function. Z-axis is error, X&Y-axis are parameters

What is Deep Learning? Pre-training

- Let network understand the general shape of the distribution of the input data
- A.K.A knowing the difference between an image of a character vs an image of just noise
- Use the latent parameters from pre-train as initialization
- Backpropagation is used to tune these parameters
- Pre-training is sometime necessary when dataset is limited and/or unknown/noisy labels are present

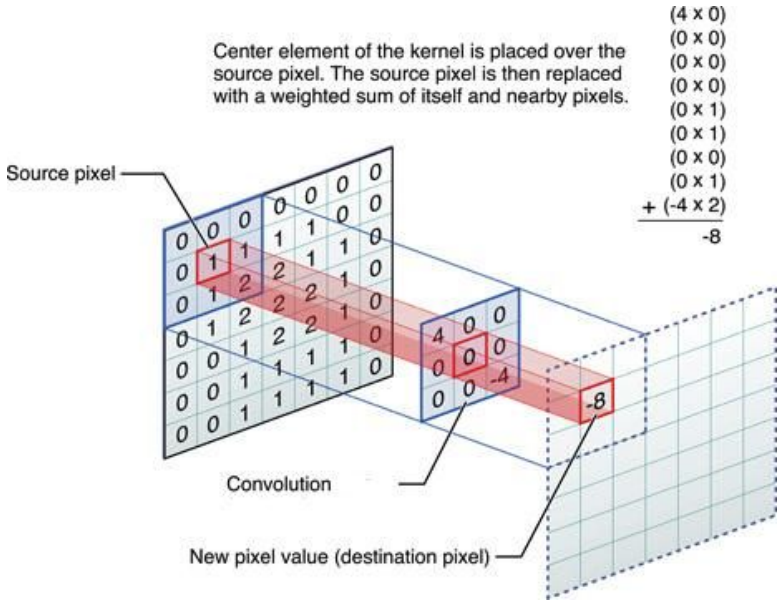


Convolutional Neural Network



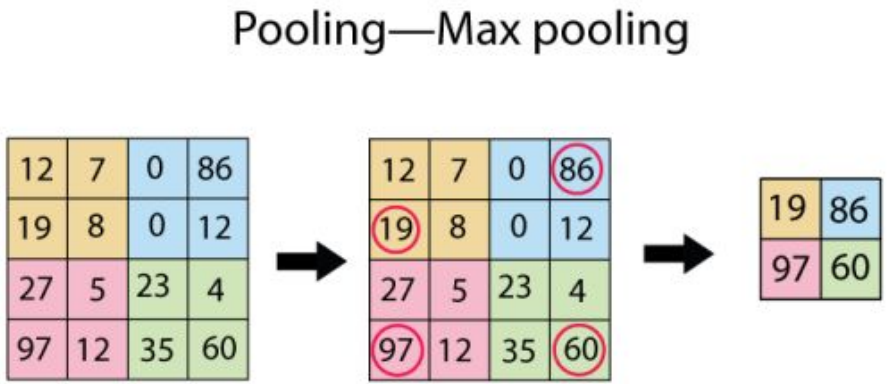
MNIST Handwritten Digits Classification using CNN

Convolutional Neural Network: Convolution



The Convolution operation in CNN is actually the Correlation operation in mathematics

Convolutional Neural Network: Pooling

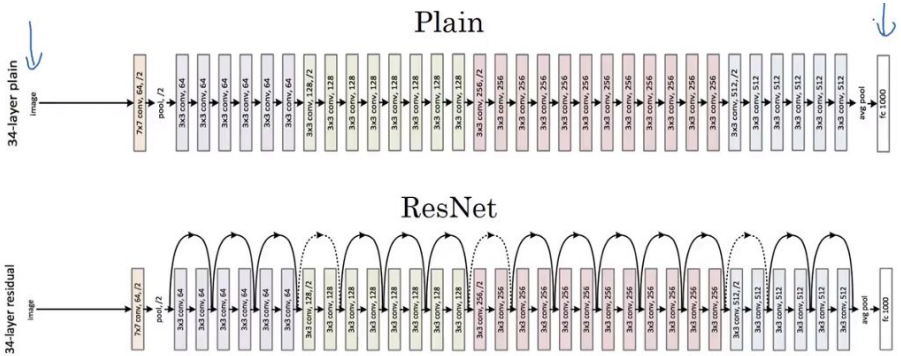


Pooling:

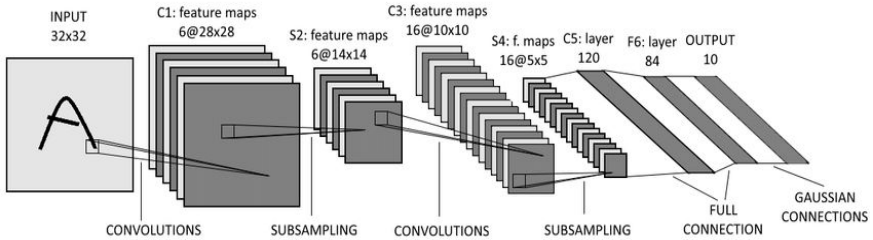
1. Amplify important information and suppressing irrelevant variation
2. Reduce dimensionality of input data
3. Creating an invariance to small shifts and distortions.

Convolutional Neural Network: Examples

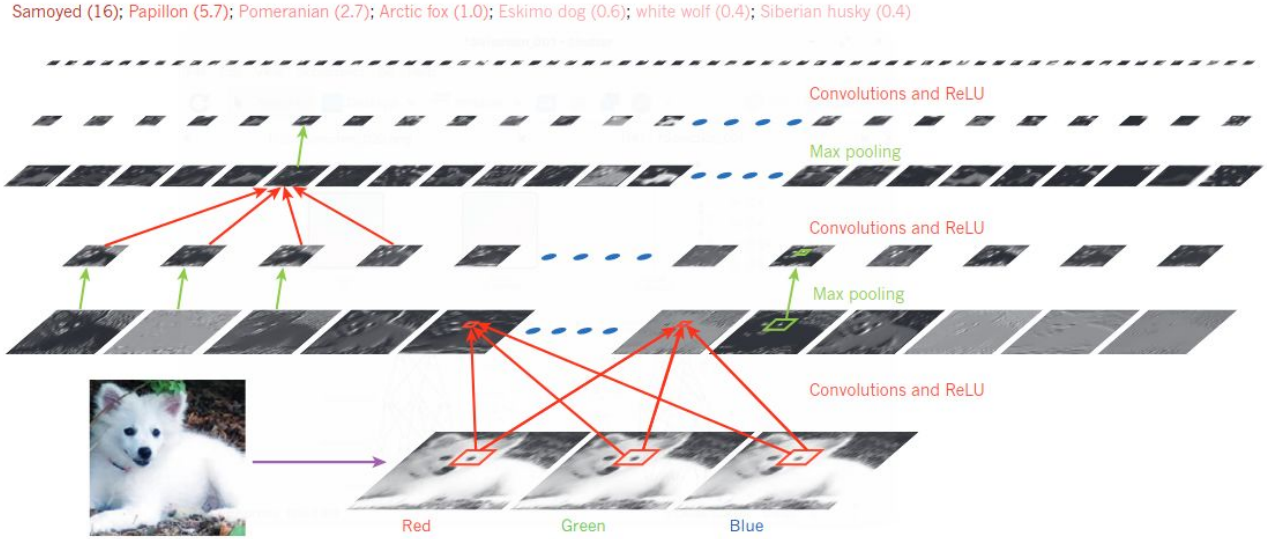
Example: ResNet



Example: LeNet-5 by LeCun



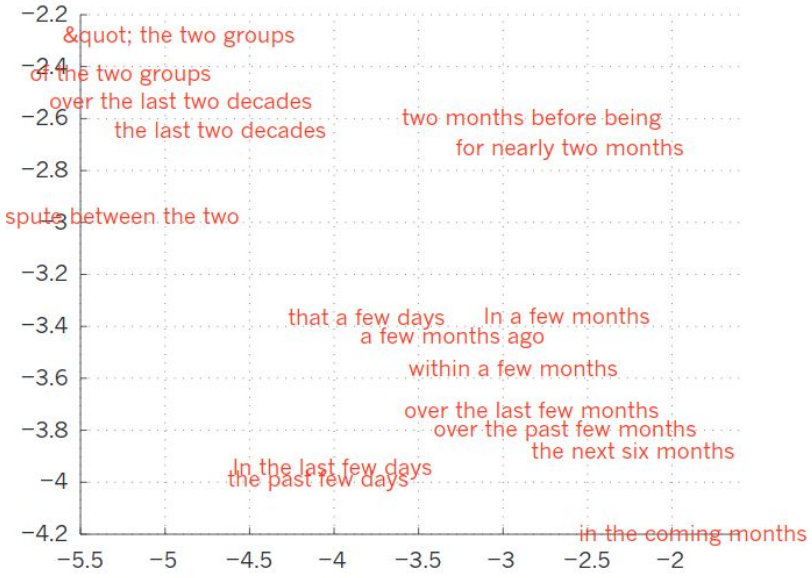
Convolutional Neural Network: Examples



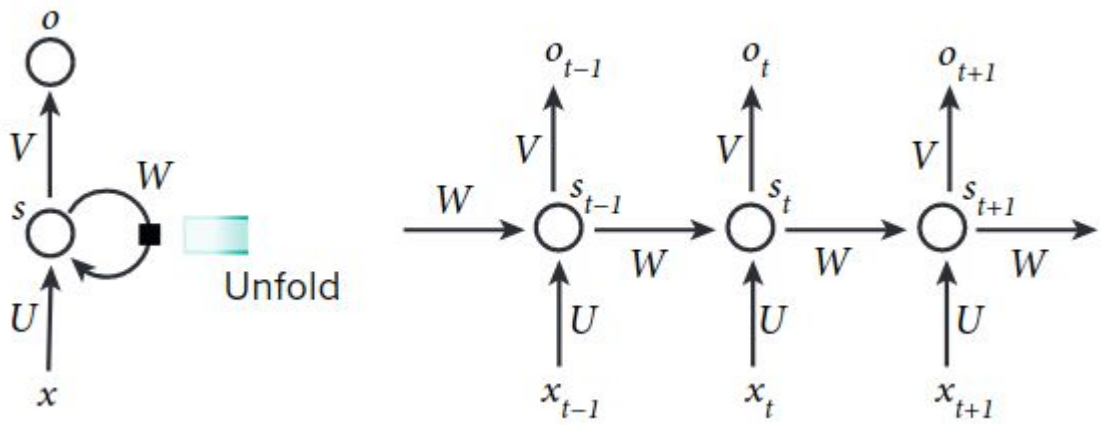
Distributed Representation: Magic of DNN

Word representations learned for modelling language

Representation of phrases learned by an English-to-French encoder–decoder recurrent neural network



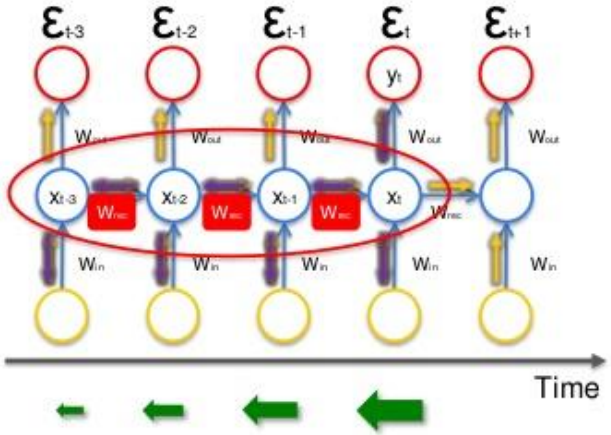
Recurrent Neural Network



An “unfold” recurrent neural network with only 1 unit

Recurrent Neural Network: Backpropagation Issue

The Vanishing Gradient Problem



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

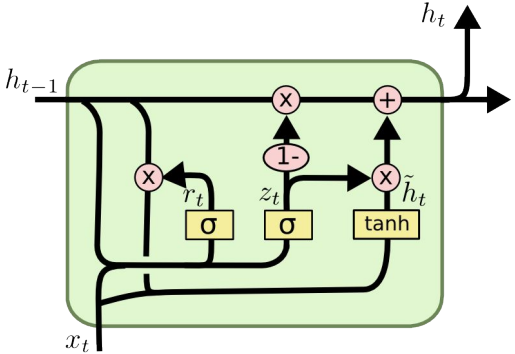
$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

W_{rec} ~ small ⇨ Vanishing

W_{rec} ~ large ⇨ Exploding

Formula Source: Razvan Pascanu et al. (2013)

RNN: Long-Short Term Memory Network



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

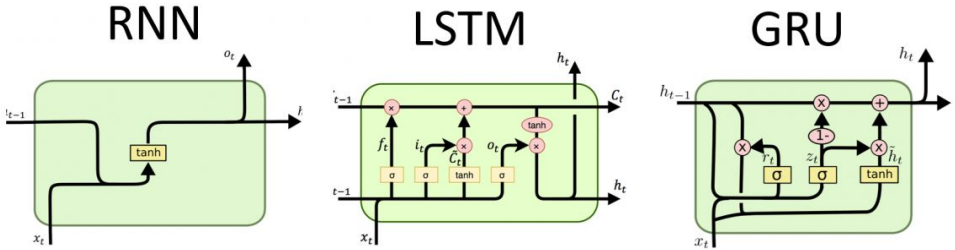
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

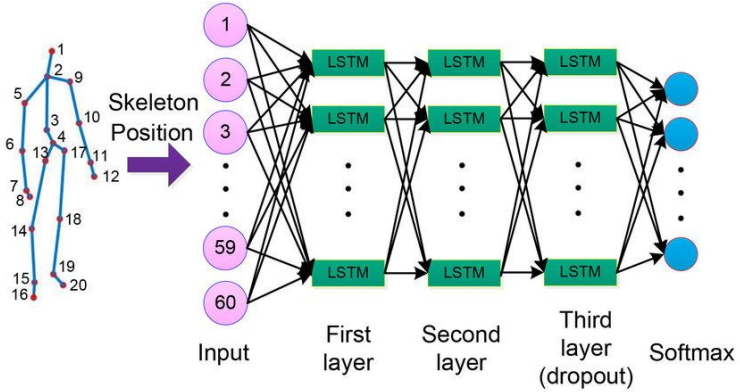
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

An LSTM unit operation

RNN: Long-Short Term Memory Network

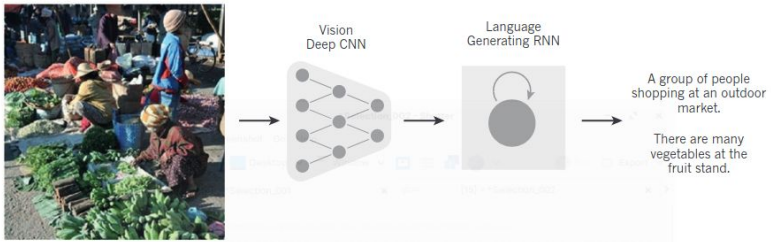


GRU is Gated Recurrent Unit



LSTM network architecture for classification/ prediction of skeleton pose

CNN + RNN: Examples



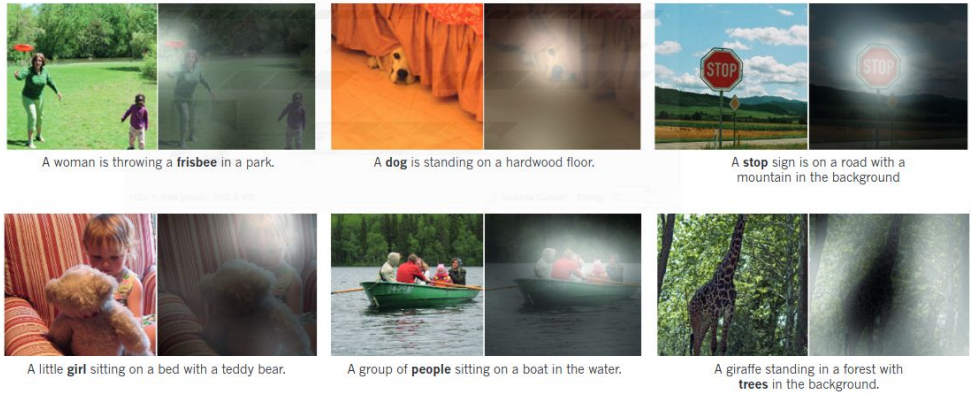


Image Captioning: “**Neural image caption generation with visual attention**” by Xu et. al (2015)

CNN + RNN: Examples

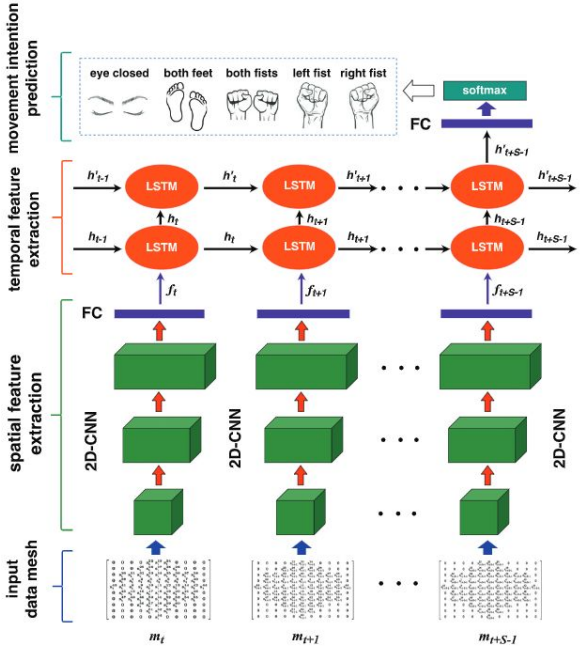


Figure 2: Cascade convolutional recurrent neural network architecture.

Human movement intention prediction: "**EEG-based Intention Recognition from Spatio-Temporal Representations via Cascade and Parallel Convolutional Recurrent Neural Networks**" by Zhang et. al (2018)

Unsupervised Learning & Future of Deep Learning

IBM Deep Blue beating Kasparov - World Chess Champion

OpenAI beating Dota 2 pros

LeCun: "Ultimately, major progress in artificial intelligence will come about through systems that combine representation learning with complex reasoning. Although deep learning and simple reasoning have been used for speech and handwriting recognition for a long time, new paradigms are needed to replace rule-based manipulation of symbolic expressions (Machine Learning) by operations on large vectors (Machine Reasoning)"

Thank you for listening

Question?