Deep Neural Networks

CMSC 422

MARINE CARPUAT
marine@cs.umd.edu

Deep learning slides credit: Vlad Morariu

Training (Deep) Neural Networks

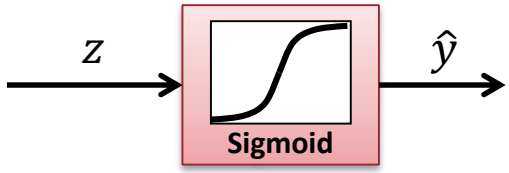
- Computational graphs
- Improvements to gradient descent
  - Stochastic gradient descent
  - Momentum
  - Weight decay

- Vanishing Gradient Problem
- Examples of deep architectures

Vanishing Gradient Problem

In deep networks

- Gradients in the lower layers are typically extremely small
- Optimizing multi-layer neural networks takes huge amount of time



$$\frac{\partial E}{\partial w_{ki}} = \sum_n \frac{\partial z_i^n}{\partial w_{ki}} \frac{d\hat{y}_i^n}{dz_i^n} \frac{\partial E}{\partial \hat{y}_i^n} = \sum_n \frac{\partial z_i^n}{\partial w_{ki}} \frac{d\hat{y}_i^n}{dz_i^n} \sum_j w_{ij} \frac{d\hat{y}_j^n}{dz_j^n} \frac{\partial E}{\partial \hat{y}_j^n}$$

Slide credit: adapted from Bohyung Han

Vanishing Gradient Problem

- Vanishing gradient problem can be mitigated
  - Using other non-linearities
    - E.g., Rectifier: f(x) = max(0,x)
  - Using custom neural network architectures
    - E.g., LSTM

Training (Deep) Neural Networks

- Computational graphs
- Improvements to gradient descent
  - Stochastic gradient descent
  - Momentum
  - Weight decay
- Vanishing Gradient Problem
- Examples of deep architectures

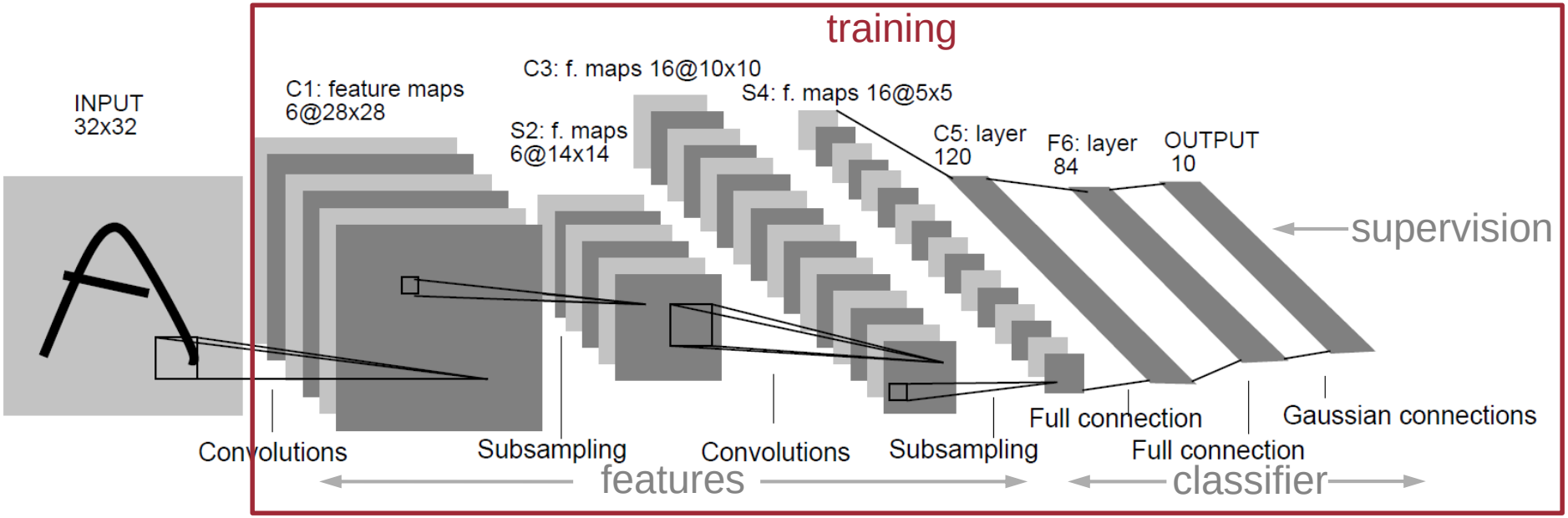


Image credit: LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE, 1998.

An example of deep neural network for computer vision

– learn features and classifiers jointly (“end-to-end” training)

New “winter” and revival in early 2000’s

New “winter” in the early 2000’s due to

- problems with training NNs
- Support Vector Machines (SVMs), Random Forests (RF) – easy to train, nice theory

Revival again by 2011-2012

- Name change (“neural networks” -> “deep learning”)
- + Algorithmic developments
  - unsupervised pre-training
  - ReLU, dropout, layer normalizatoin
- + Big data + GPU computing =
- Large outperformance on many datasets (Vision: ILSVRC’12)

http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning-part-4/

Big Data

- ImageNet Large Scale Visual Recognition Challenge
  - 1000 categories w/ 1000 images per category
  - 1.2 million training images, 50,000 validation, 150,000 testing

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei. **ImageNet Large Scale Visual Recognition Challenge**. *IJCV,* 2015.

AlexNet Architecture

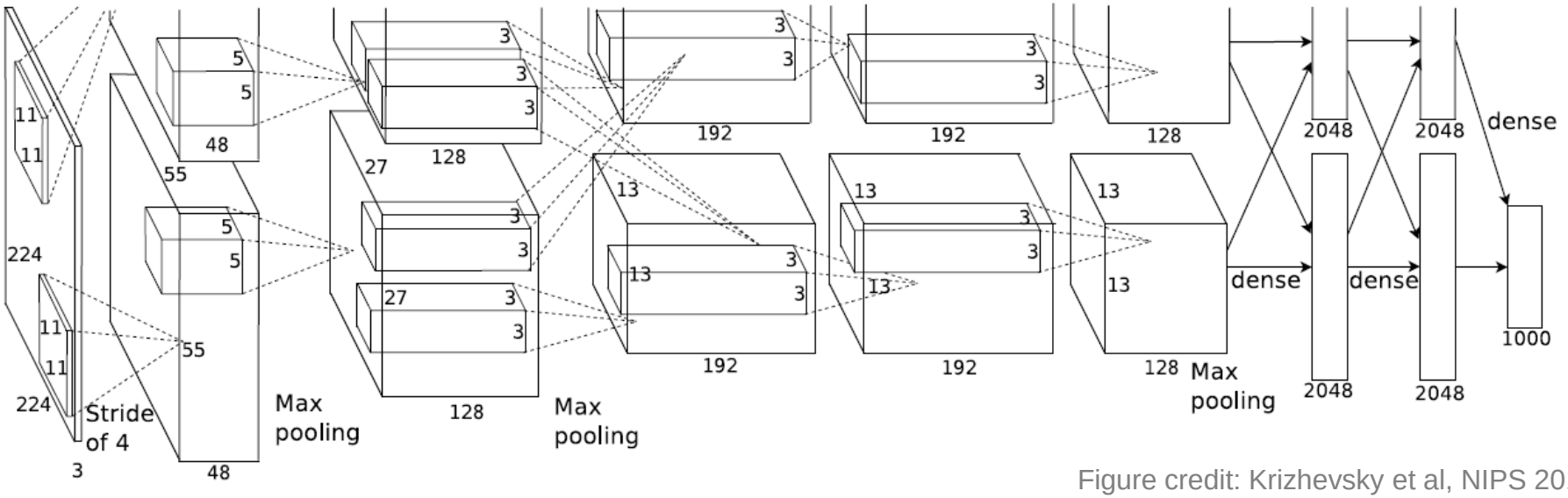


Figure credit: Krizhevsky et al, NIPS 2012.

60 million parameters!

Various tricks

- ReLU nonlinearity
- Dropout – set hidden neuron output to 0 with probability .5
- Training on GPUs
- …

Alex Krizhevsky, Ilya Sutskeyer, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.

GPU Computing

- Big data and big models require lots of computational power

- GPUs
  - thousands of cores for parallel operations
  - multiple GPUs
  - still took about 5-6 days to train AlexNet on two NVIDIA GTX 580 3GB GPUs (much faster today)

Image Classification Performance

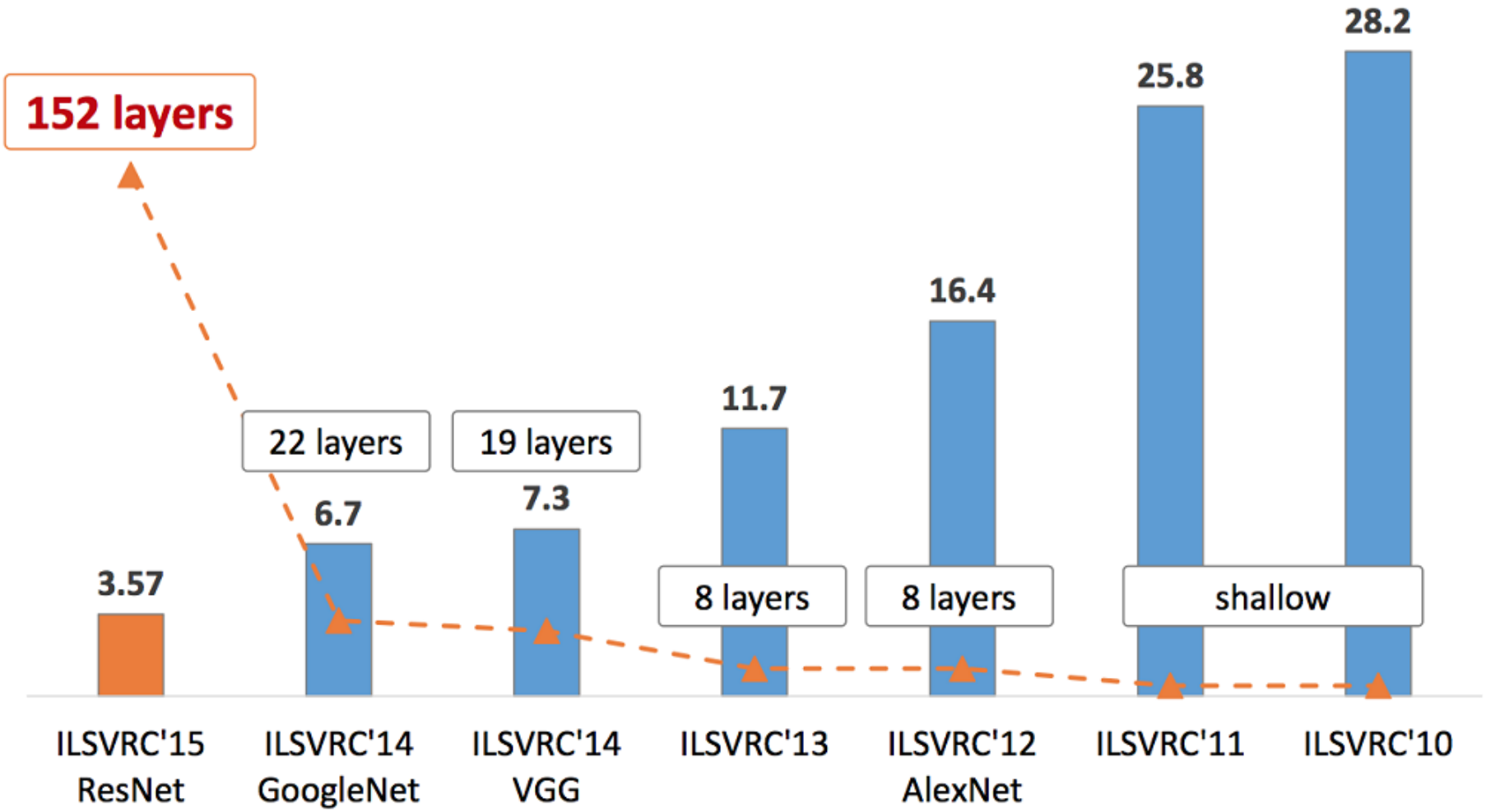
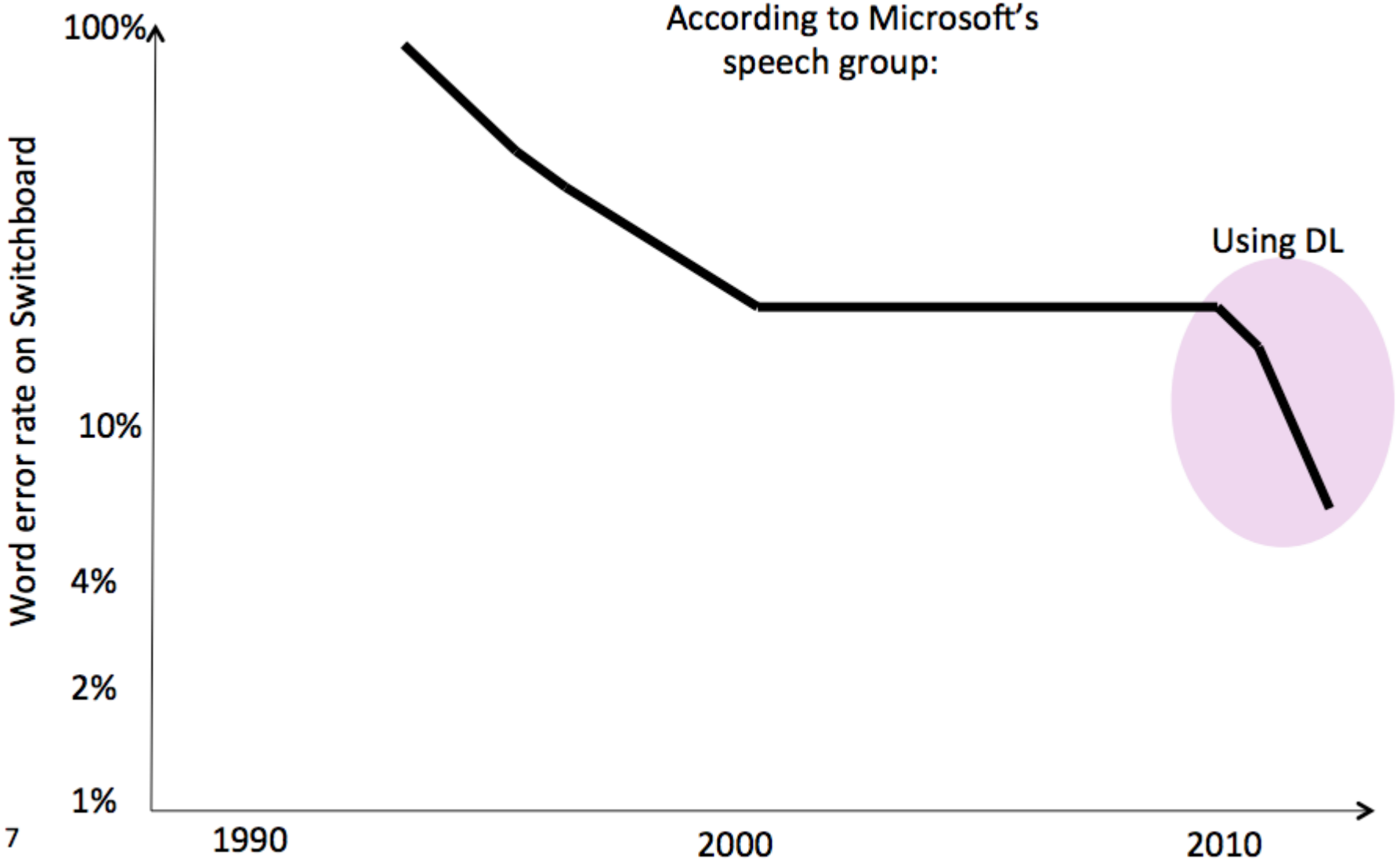


Image Classification Top-5 Errors (%)

Figure from: K. He, X. Zhang, S. Ren, J. Sun. "Deep Residual Learning for Image Recognition". arXiv 2015. (slides)

Slide credit: Bohyung Han

Speech Recognition



Slide credit: Bohyung Han

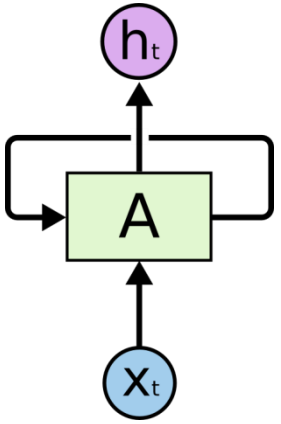
Recurrent Neural Networks for Language Modeling

- Speech recognition is difficult due to ambiguity
  - "how to recognize speech"
  - or "how to wreck a nice beach"?
- Language model gives probability of next word given history
  - P("speech"|"how to recognize")?

Recurrent Neural Networks

Networks with loops

- The output of a layer is used as input for the same (or lower) layer
- Can model dynamics (e.g. in space or time)



Loops are unrolled

- Now a standard feed-forward network with many layers
- Suffers from vanishing gradient problem
- In theory, can learn long term memory, in practice not (Bengio et al, 1994)

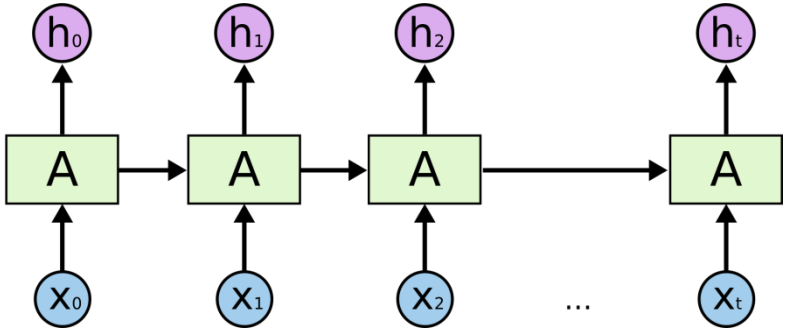
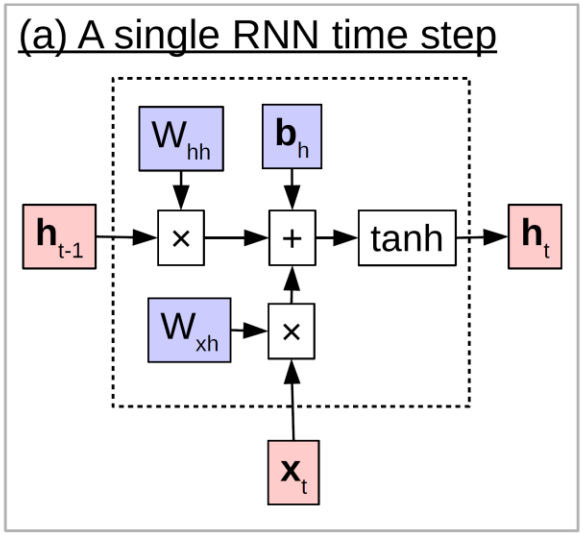


Image credit: Chritopher Olah's blog http://colah.github.io/posts/2015-08-Understanding-LSTMs/

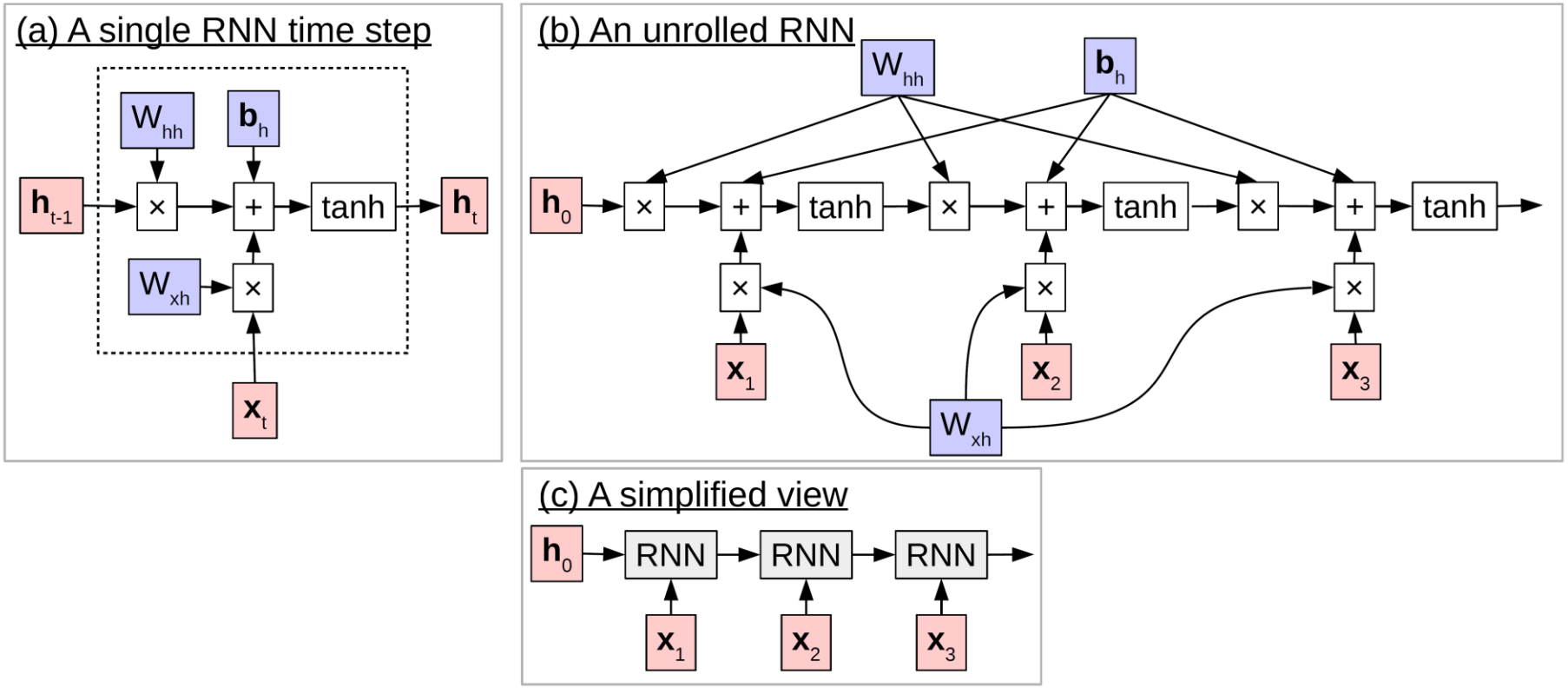
Sepp Hochreiter (1991), Untersuchungen zu dynamischen neuronalen Netzen, Diploma thesis. Institut f. Informatik, Technische Univ. Munich. Advisor: J. Schmidhuber.

Y. Bengio, P. Simard, P. Frasconi. Learning Long-Term Dependencies with Gradient Descent is Difficult. In TNN 1994.

A Recurrent Neural Network Computational Graph



A Recurrent Neural Network Computational Graph



Long Short Term Memory (LSTM)

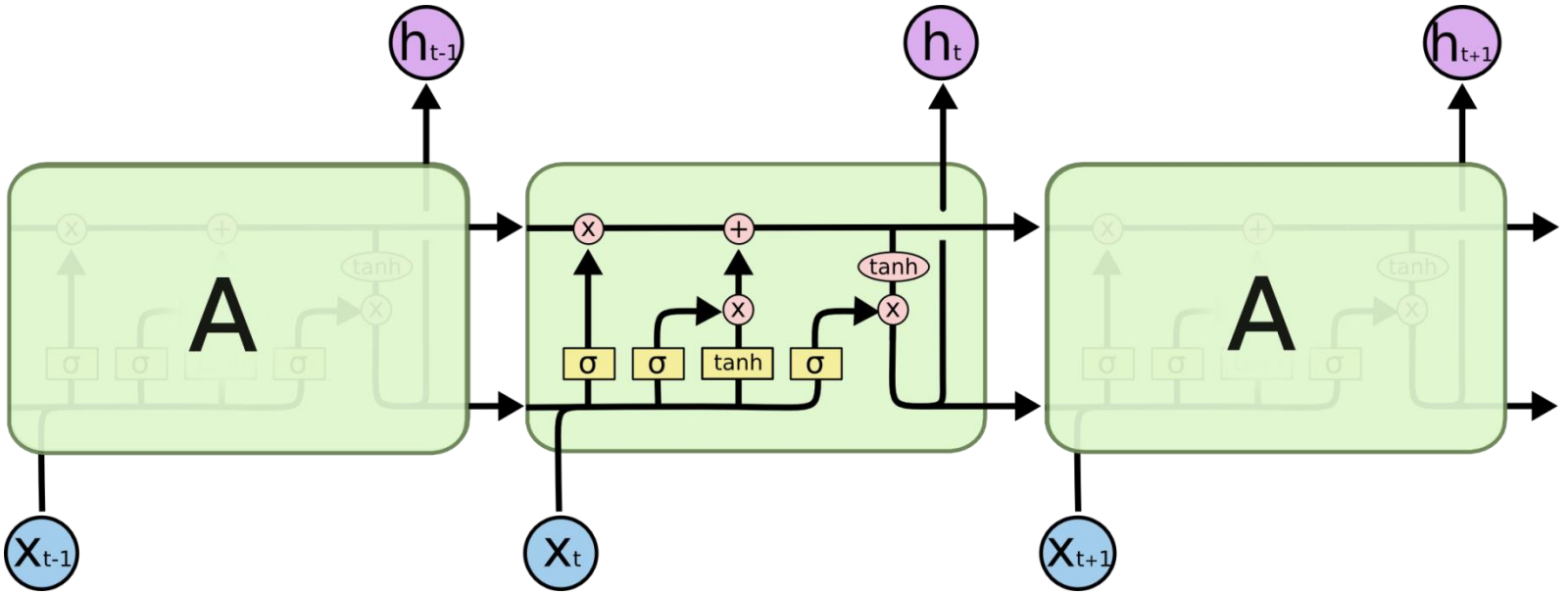


Image credit: Christopher Colah's blog, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- A type of RNN explicitly designed not to have the vanishing or exploding gradient problem
- Models long-term dependencies
- Memory is propagated and accessed by gates
- Used for speech recognition, language modeling …

Hochreiter, Sepp; and Schmidhuber, Jürgen. "Long Short-Term Memory." Neural Computation, 1997.

Long Short Term Memory (LSTM)

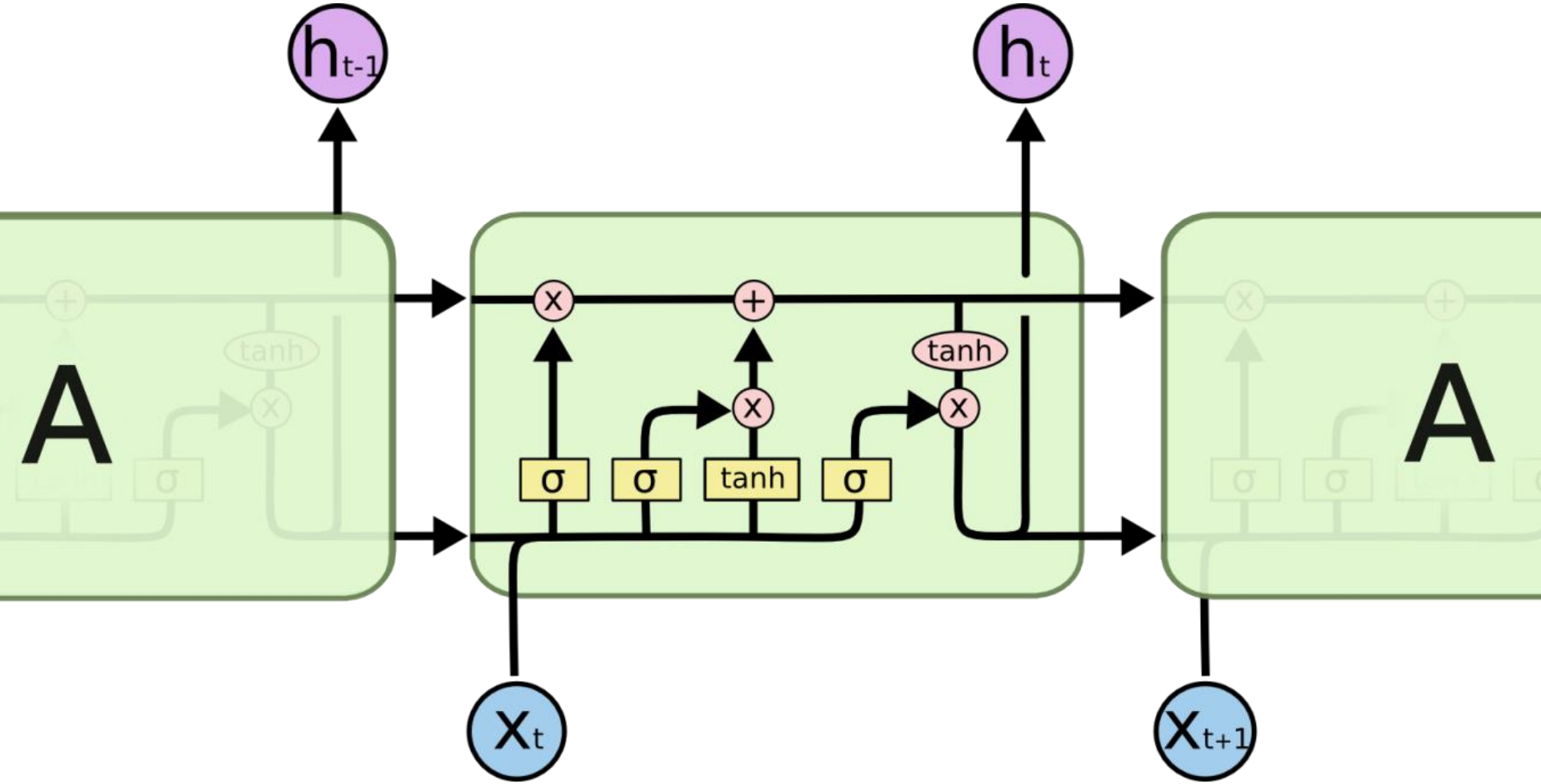


Image credit: Christopher Colah's blog, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

What you should know about deep neural networks

- Why they are difficult to train
  - Initialization
  - Overfitting
  - Vanishing gradient
  - Require large number of training examples

- What can be done about it
  - Improvements to gradient descent
  - Stochastic gradient descent
  - Momentum
  - Weight decay
  - Alternate non-linearities and new architectures

References (& great tutorials) if you want to explore further:
http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning-part-1/
http://cs231n.github.io/neural-networks-1/
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Keeping things in perspective…

In 1958, the New York Times reported the perceptron to be "the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

Project 3

- Due May 10
- PCA, digit classification with neural networks
- 2 important concepts
  - Logistic regression
  - Softmax classifier