ECE 6504: Deep Learning for Perception

Topics:

- Recurrent Neural Networks (RNNs)
- BackProp Through Time (BPTT)
- Vanishing / Exploding Gradients
- [Abhishek:] Lua / Torch Tutorial

Dhruv Batra

Virginia Tech

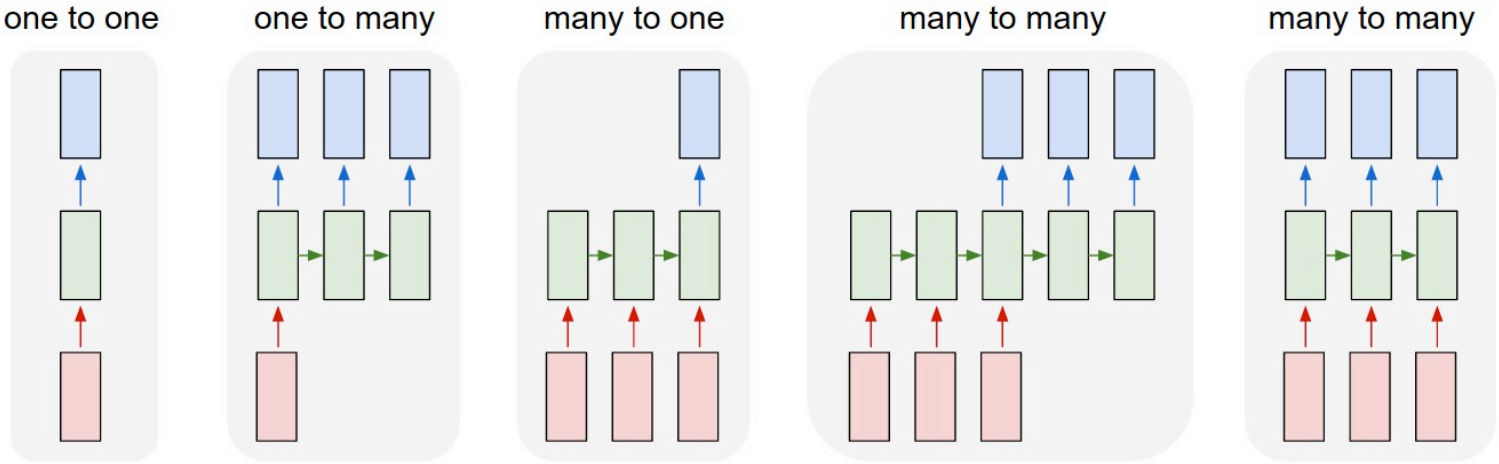
Administrativia

- HW3
  - Out today
  - Due in 2 weeks
  - Please please please please please start early
  - https://computing.ece.vt.edu/~f15ece6504/homework3/

Plan for Today

- Model
  - Recurrent Neural Networks (RNNs)
- Learning
  - BackProp Through Time (BPTT)
  - Vanishing / Exploding Gradients
- [Abhishek:] Lua / Torch Tutorial

New Topic: RNNs



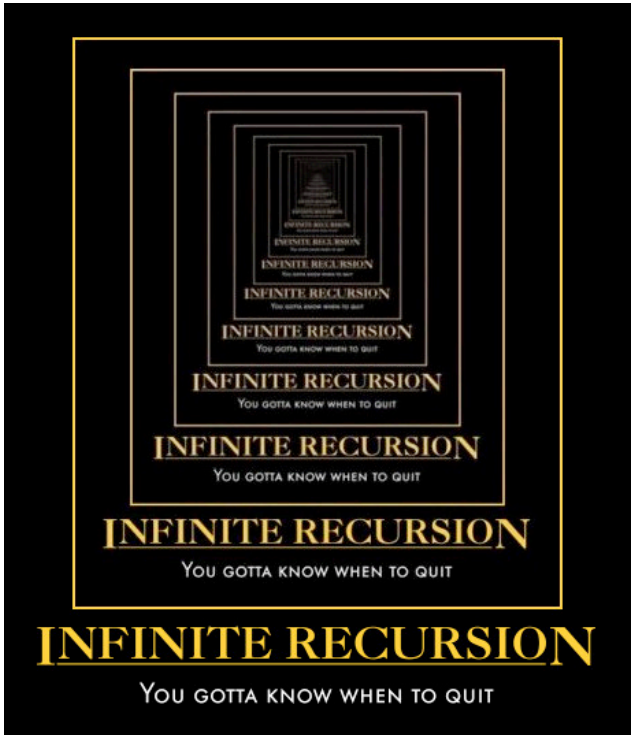


Image Credit: Andrej Karpathy

Synonyms

- Recurrent Neural Networks (RNNs)
- Recursive Neural Networks
  - General familty; think graphs instead of chains
- Types:
  - Long Short Term Memory (LSTMs)
  - Gated Recurrent Units (GRUs)
  - Hopfield network
  - Elman networks
  - …
- Algorithms
  - BackProp Through Time (BPTT)
  - BackProp Through Structure (BPTS)

What's wrong with MLPs?

- Problem 1: Can't model sequences
  - Fixed-sized Inputs & Outputs
  - No temporal structure

- Problem 2: Pure feed-forward processing
  - No "memory", no feedback

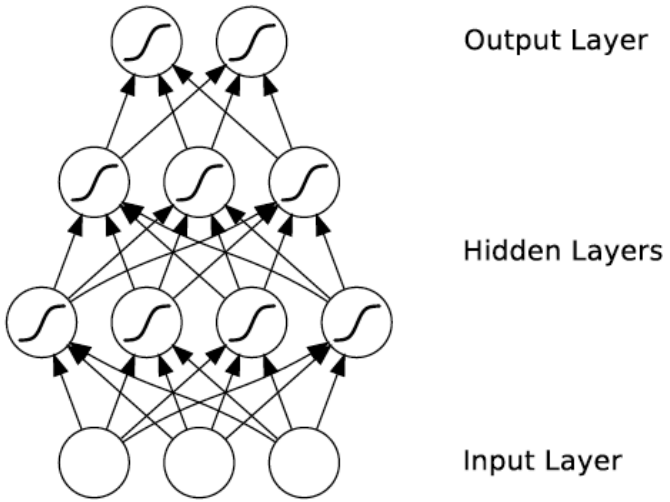
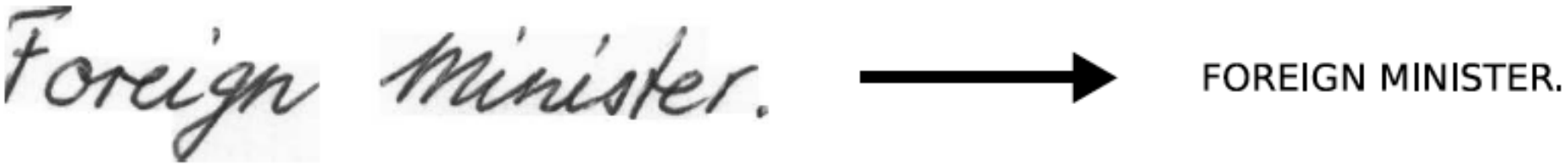


Image Credit: Alex Graves, book

Sequences are everywhere…



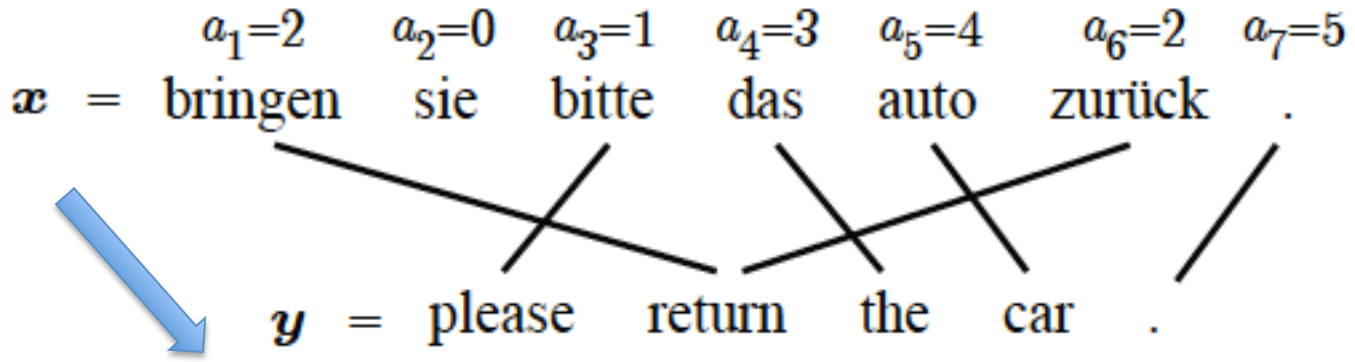
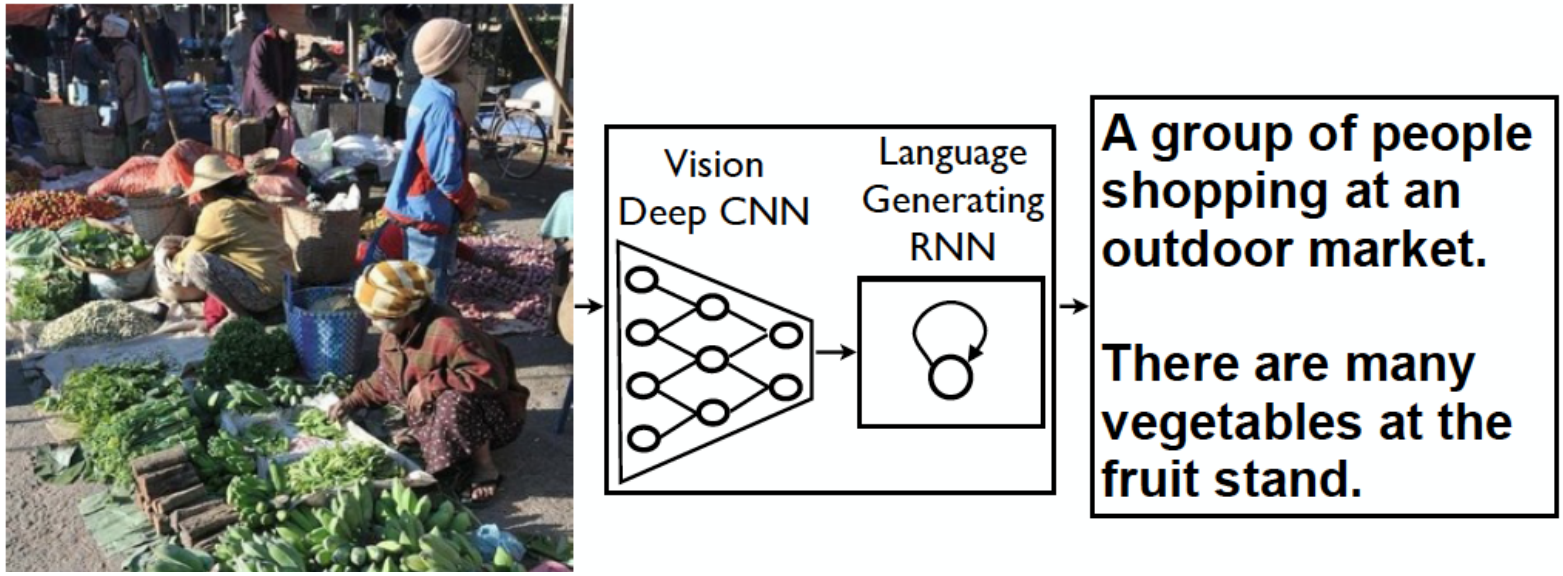
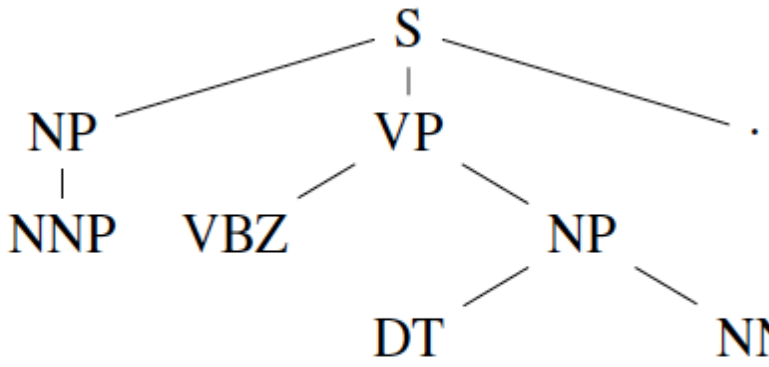


Image Credit: Alex Graves and Kevin Gimpel

Even where you might not expect a sequence…



John has a dog . →

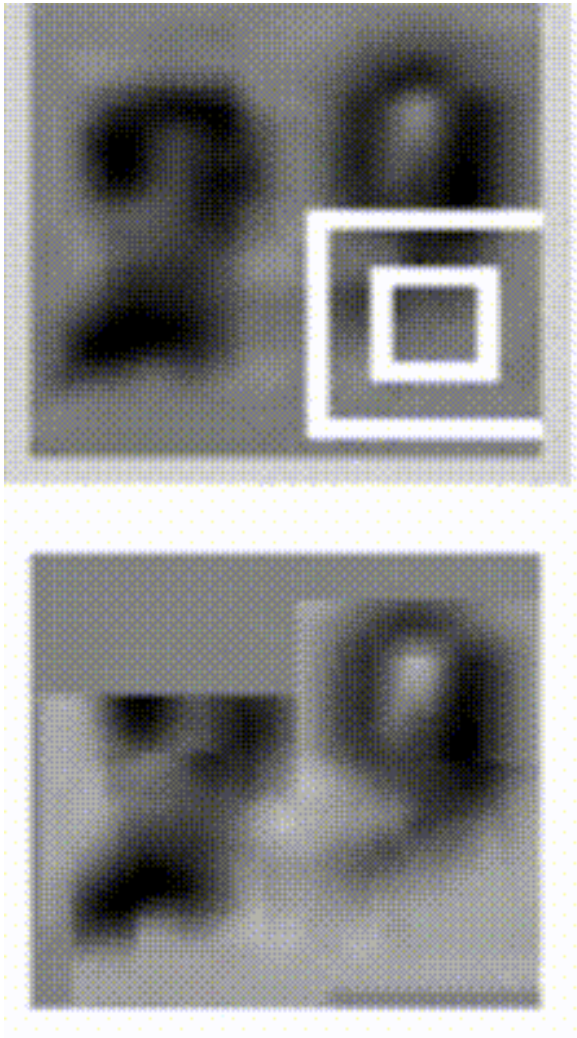


John has a dog . → (S (NP NNP)$_{NP}$ (VP VBZ (NP DT NN)$_{NP}$)$_{VP}$.)$_{S}$

Image Credit: Vinyals et al.

Even where you might not expect a sequence…

- Input ordering = sequence



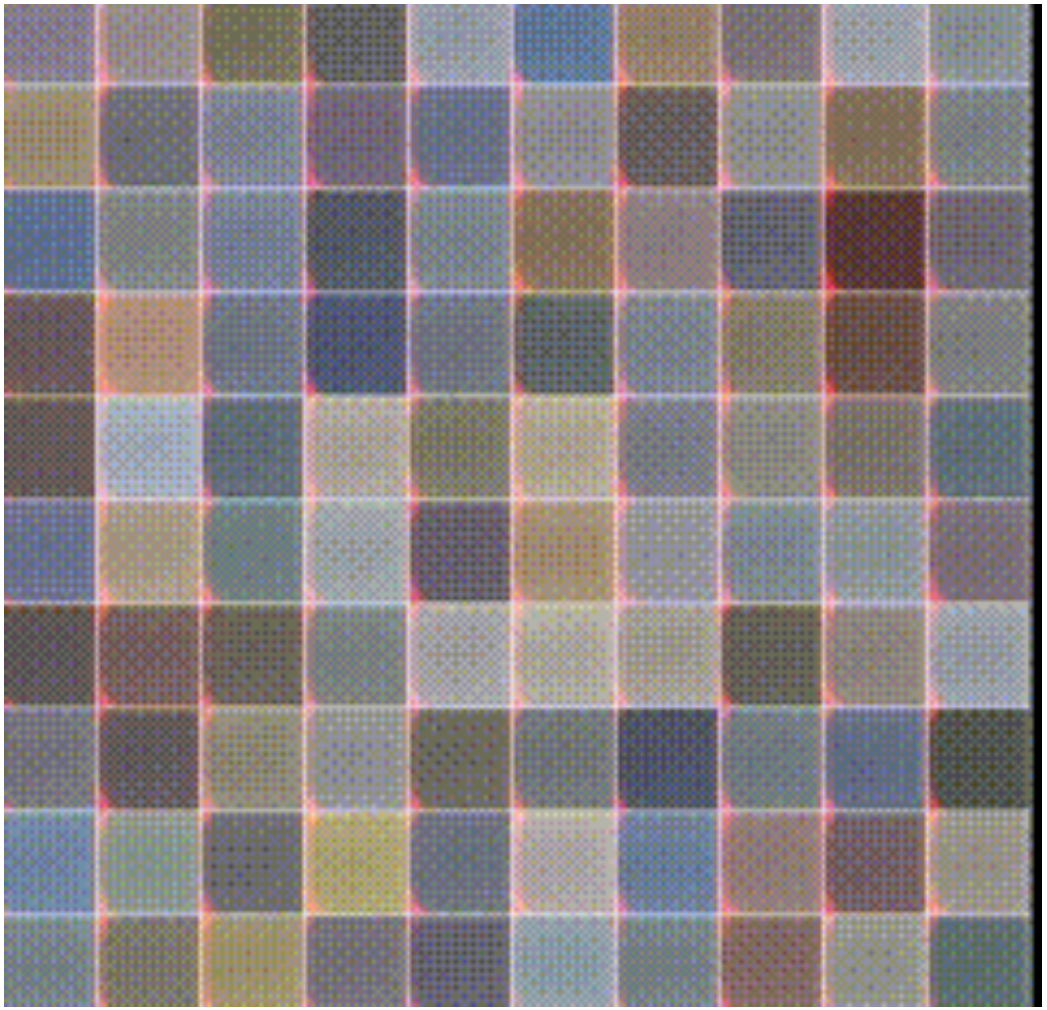


Image Credit: Ba et al.; Gregor et al

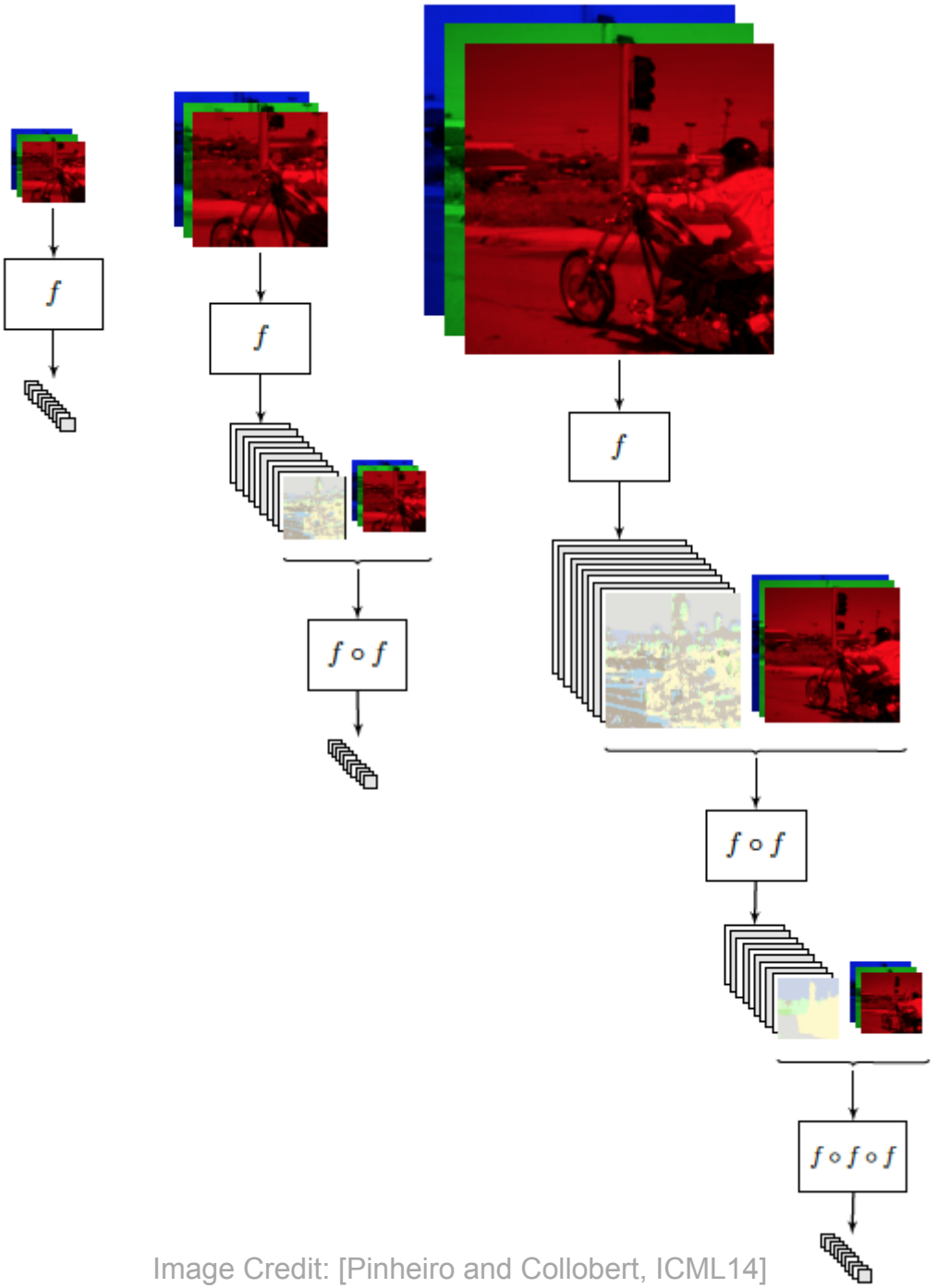


Image Credit: [Pinheiro and Collobert, ICML14]

Why model sequences?

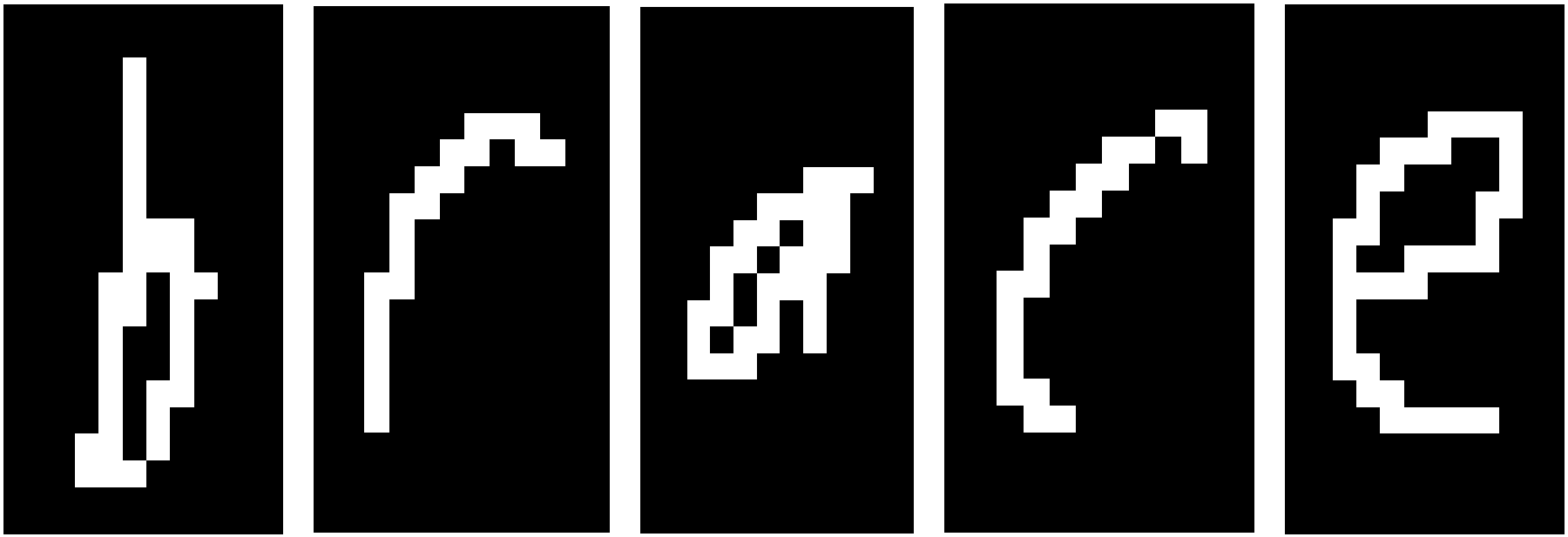


Figure Credit: Carlos Guestrin

Why model sequences?

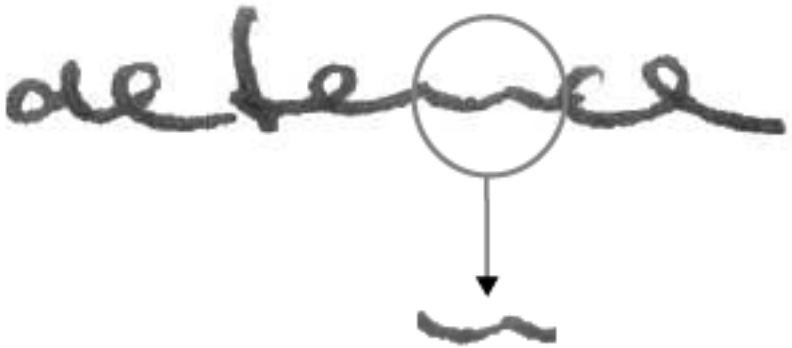
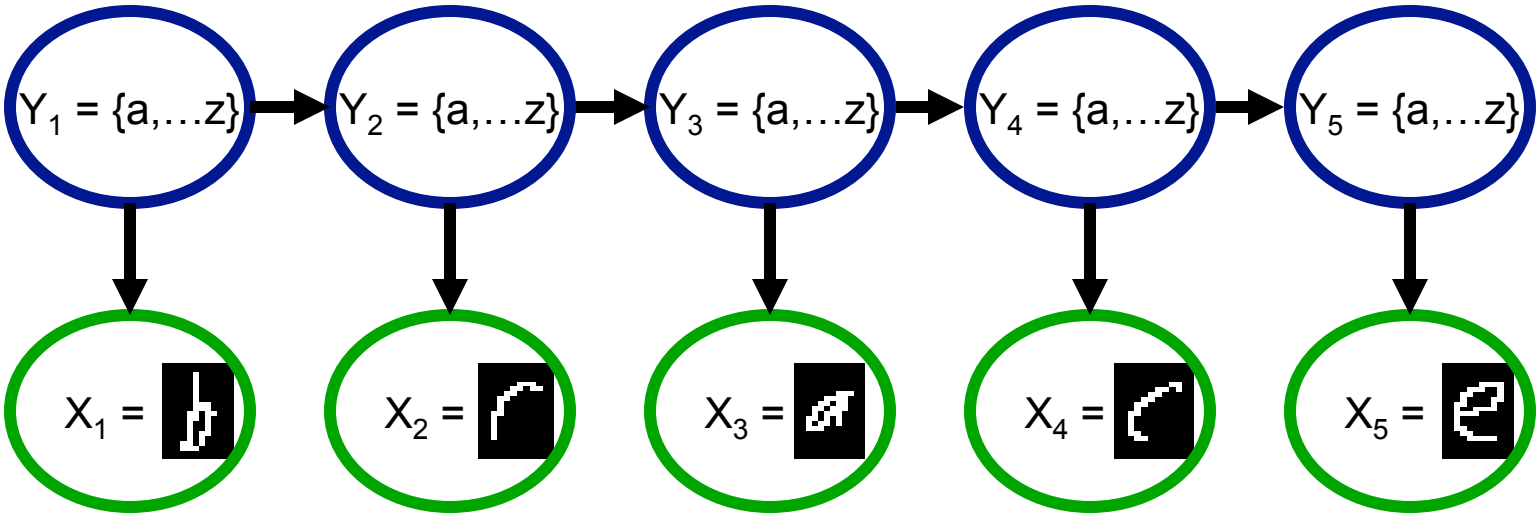


Image Credit: Alex Graves

Name that model



Hidden Markov Model (HMM)

How do we model sequences?

- No input

$$\boldsymbol{s}_t = f_\theta(\boldsymbol{s}_{t-1})$$

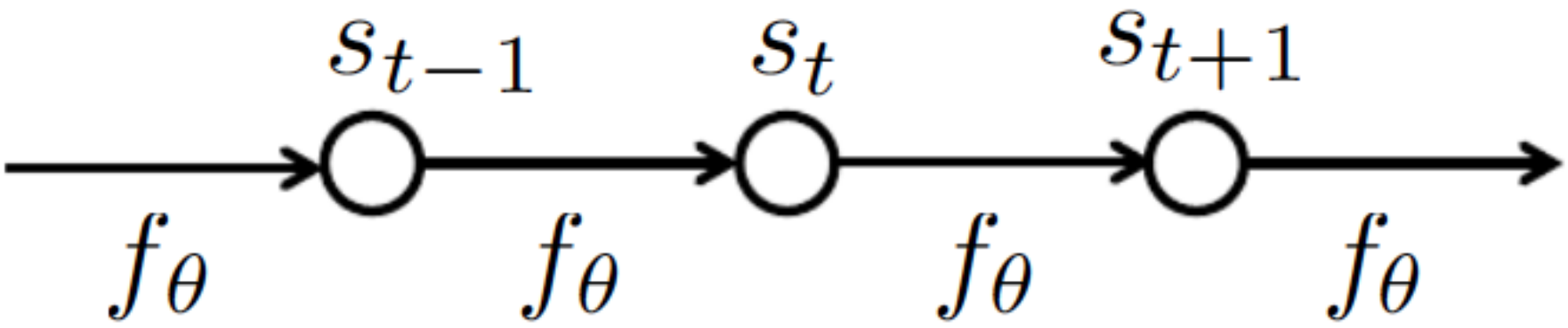


Image Credit: Bengio, Goodfellow, Courville

How do we model sequences?

- With inputs

$$\boldsymbol{s}t = f\theta(\boldsymbol{s}t-1, \boldsymbol{x}t)$$

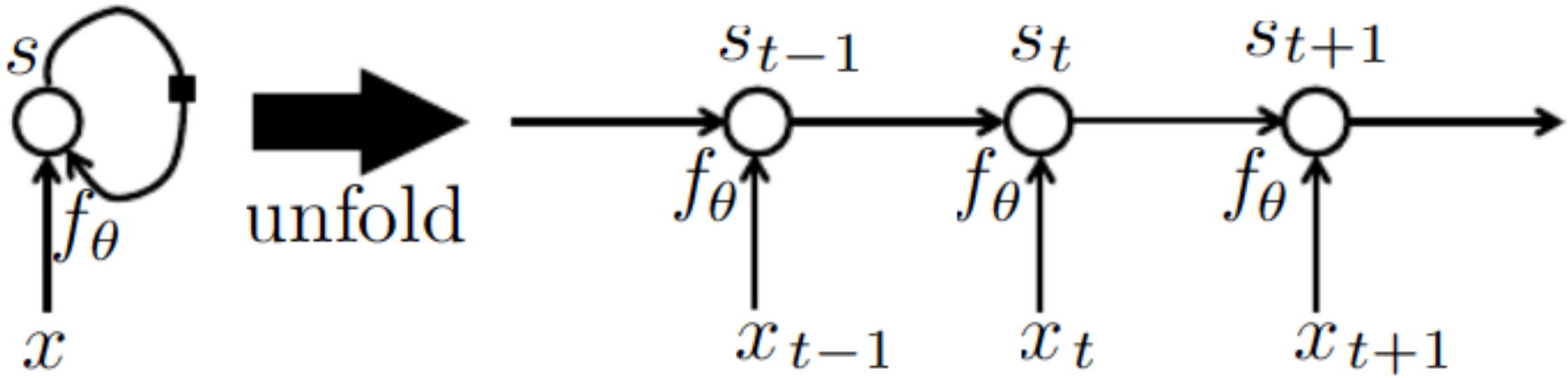


Image Credit: Bengio, Goodfellow, Courville

How do we model sequences?

- With inputs and outputs

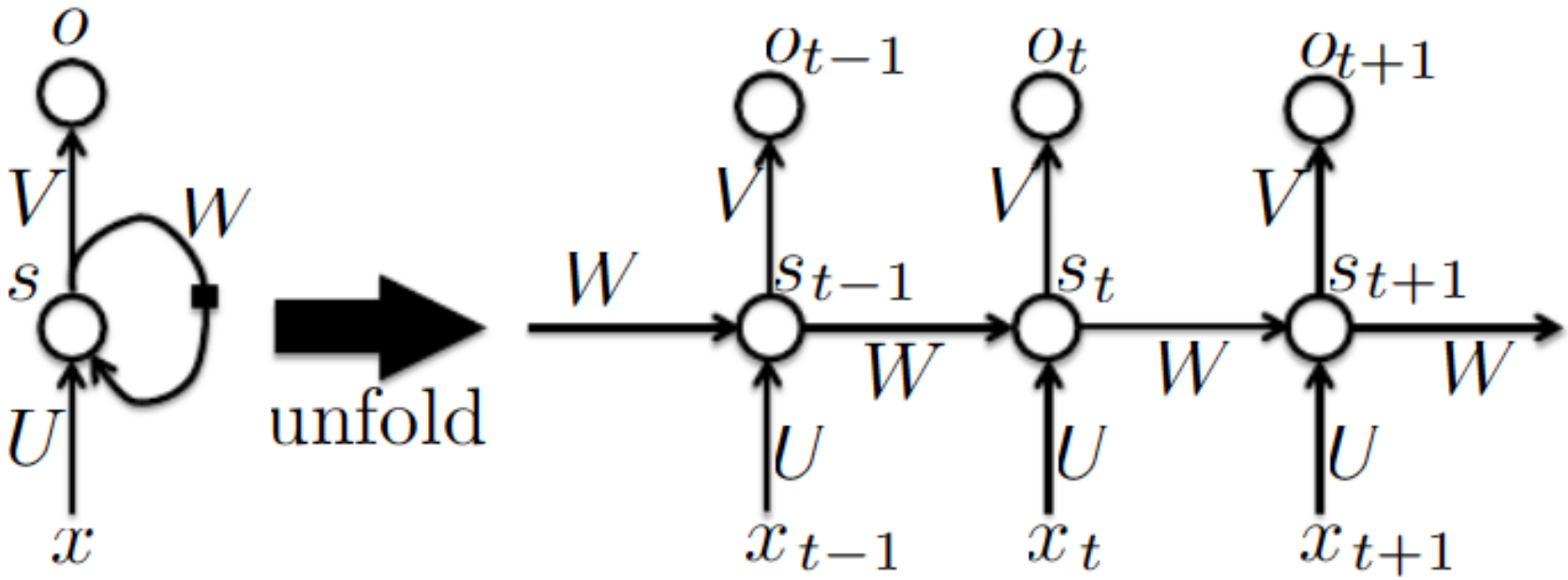
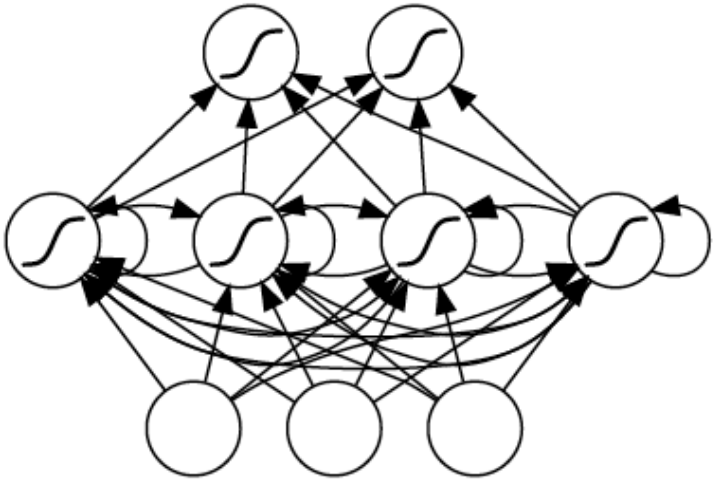


Image Credit: Bengio, Goodfellow, Courville

How do we model sequences?

- With Neural Nets



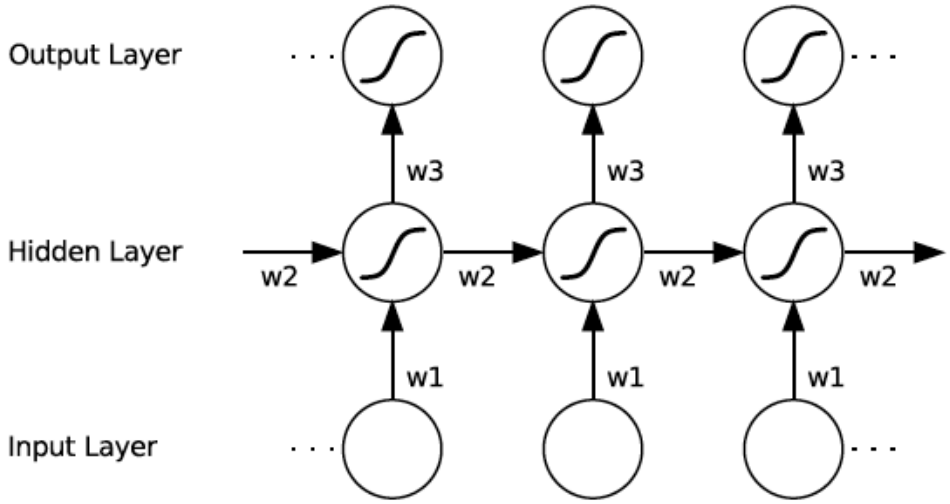
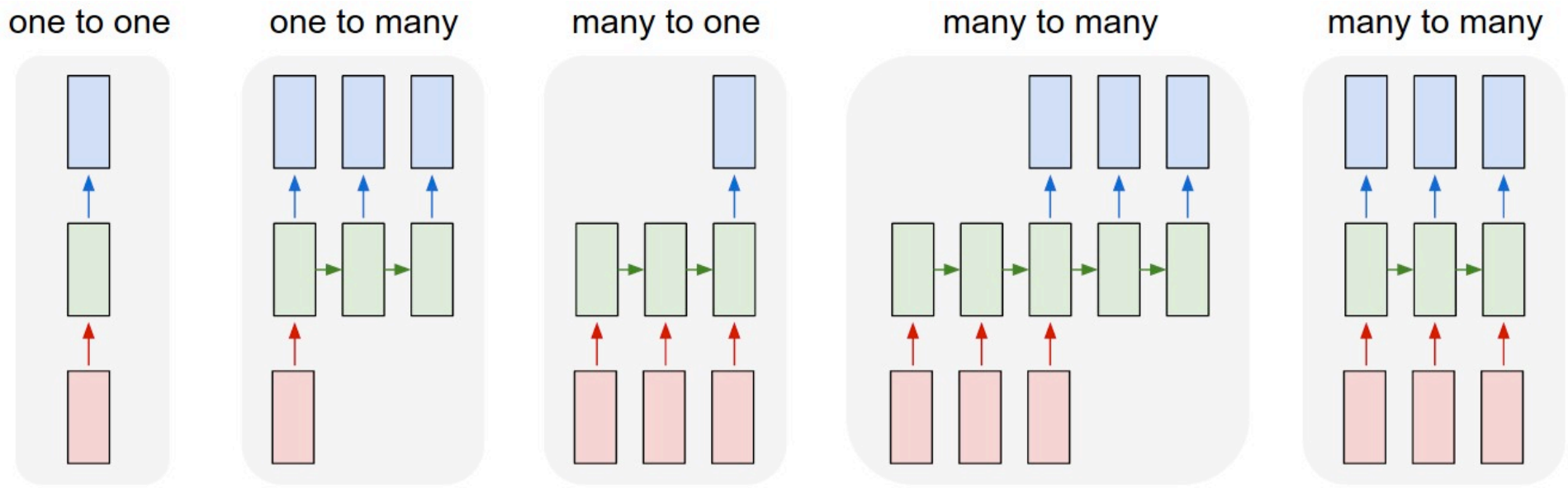


Image Credit: Alex Graves

How do we model sequences?

- It's a spectrum…



Input: No sequence

Output: No sequence

Example: "standard" classification / regression problems

Input: No sequence

Output: Sequence

Example: Im2Caption

Input: Sequence

Output: No sequence

Example: sentence classification, multiple-choice question answering

Input: Sequence

Output: Sequence

Example: machine translation, video captioning, open-ended question answering, video question answering

Image Credit: Andrej Karpathy

Things can get arbitrarily complex

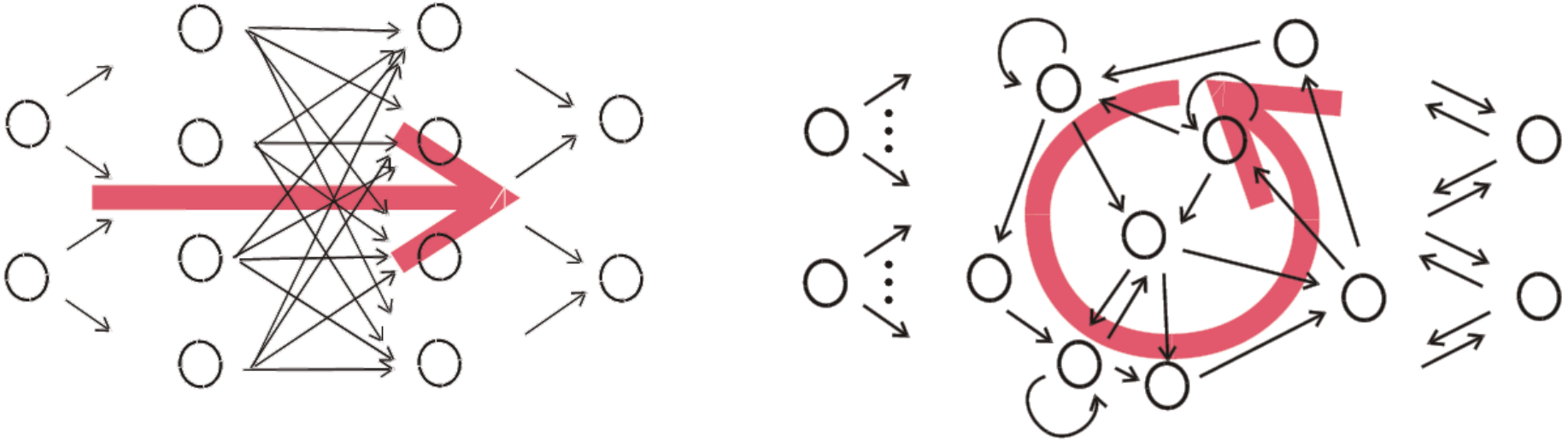


Image Credit: Herbert Jaeger

Key Ideas

- Parameter Sharing + Unrolling
  - Keeps numbers of parameters in check
  - Allows arbitrary sequence lengths!
- “Depth”
  - Measured in the usual sense of layers
  - Not unrolled timesteps
- Learning
  - Is tricky even for “shallow” models due to unrolling

Plan for Today

- Model
  - Recurrent Neural Networks (RNNs)
- Learning
  - BackProp Through Time (BPTT)
  - Vanishing / Exploding Gradients
- [Abhishek:] Lua / Torch Tutorial

BPTT

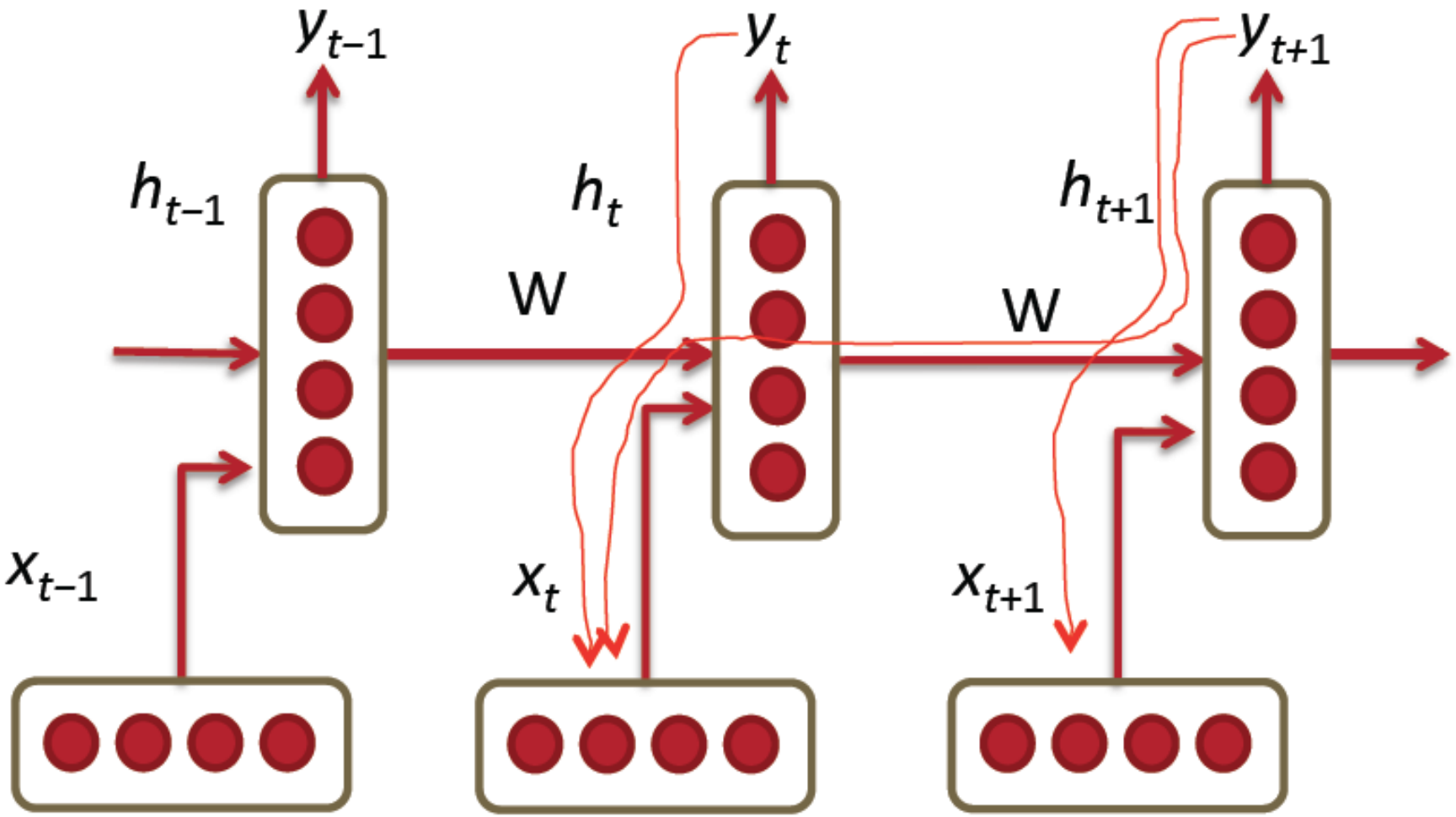
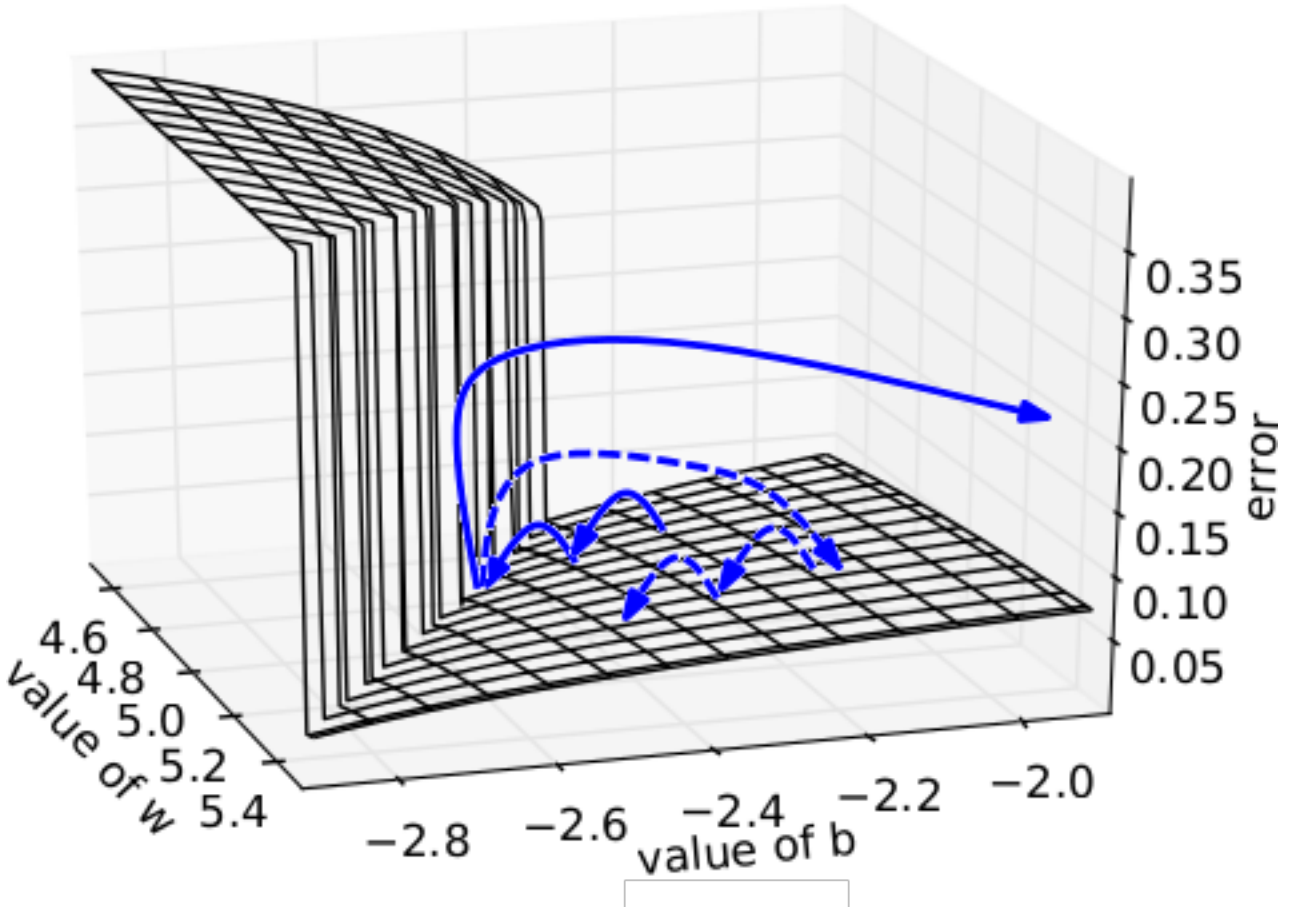


Image Credit: Richard Socher

Illustration [Pascanu et al]

- Intuition
  - Error surface of a single hidden unit RNN; High curvature walls
  - Solid lines: standard gradient descent trajectories
  - Dashed lines: gradient rescaled to fix problem



Fix #1

- Pseudocode

Algorithm 1 Pseudo-code for norm clipping the gradients whenever they explode

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

Image Credit: Richard Socher

Fix #2

- Smart Initialization and ReLus
  - [Socher et al 2013]
  - *A Simple Way to Initialize Recurrent Networks of Rectified Linear* Units, Le et al. 2015