Analyzing Software using Deep Learning

RNN-based Code Completion and Repair (Part 1)

Prof. Dr. Michael Pradel

Software Lab, University of Stuttgart

Summer 2020

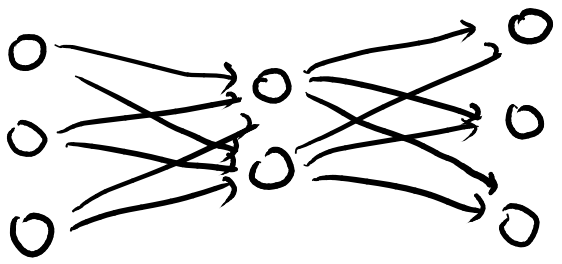
Overview

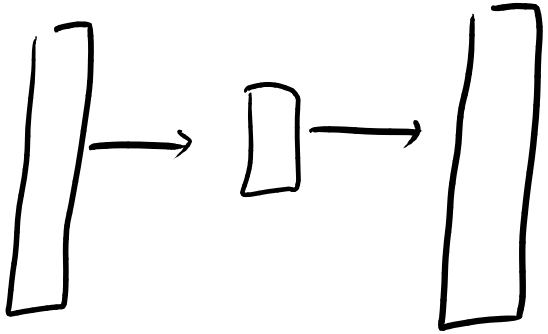
- **Recurrent neural networks (RNNs)** ←
- **Code completion with statistical language models**

  Based on PLDI 2014 paper by Raychev et al.
- **Repair of syntax errors**

  Based on ”Automated correction for syntax errors in programming assignments using recurrent neural networks” by Bhatia & Singh, 2016

From Neurons to Layers





For every neuron:

$$\text{output} = f(w \cdot x + b)$$

x, f, b .. scalars
e.g., in $\mathbb{R}$

w .. vector
e.g., in $\mathbb{R}^n$

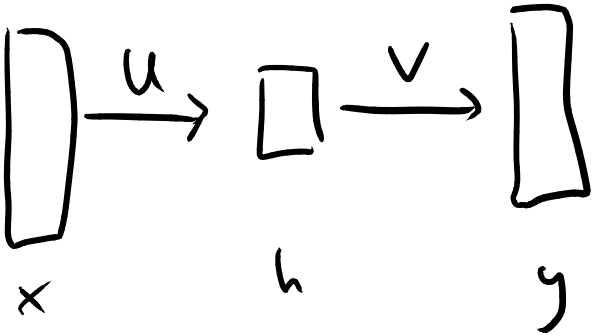
For each layer:

$$\text{output} = f(W \cdot x + b)$$

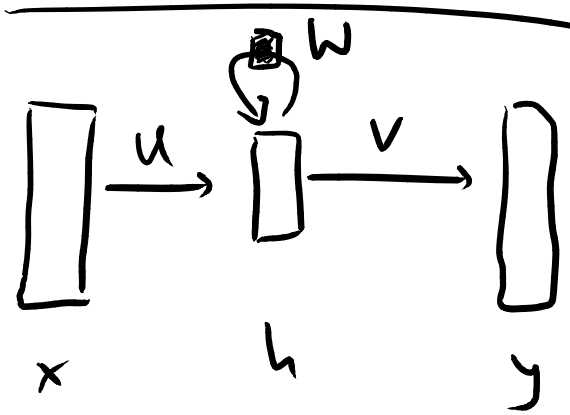
x, f, b .. vectors,
e.g., $\mathbb{R}^n$

W .. matrix,
e.g., in $\mathbb{R}^{m \times n}$

Feedforward networks



Recurrent networks



x, h, y ... input layer, hidden layer, output layer

U, V, W ... weight matrices

→ ... function

→ ... function with <u>delay of single time step</u>

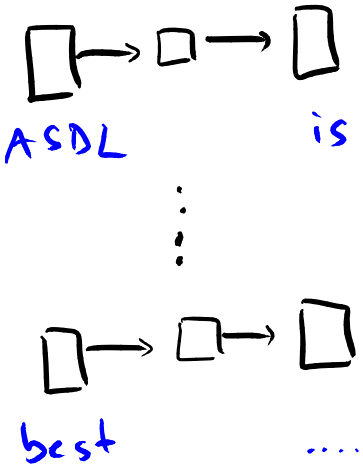
→ useful for representing sequences of inputs & outputs

→ store information about previous inputs

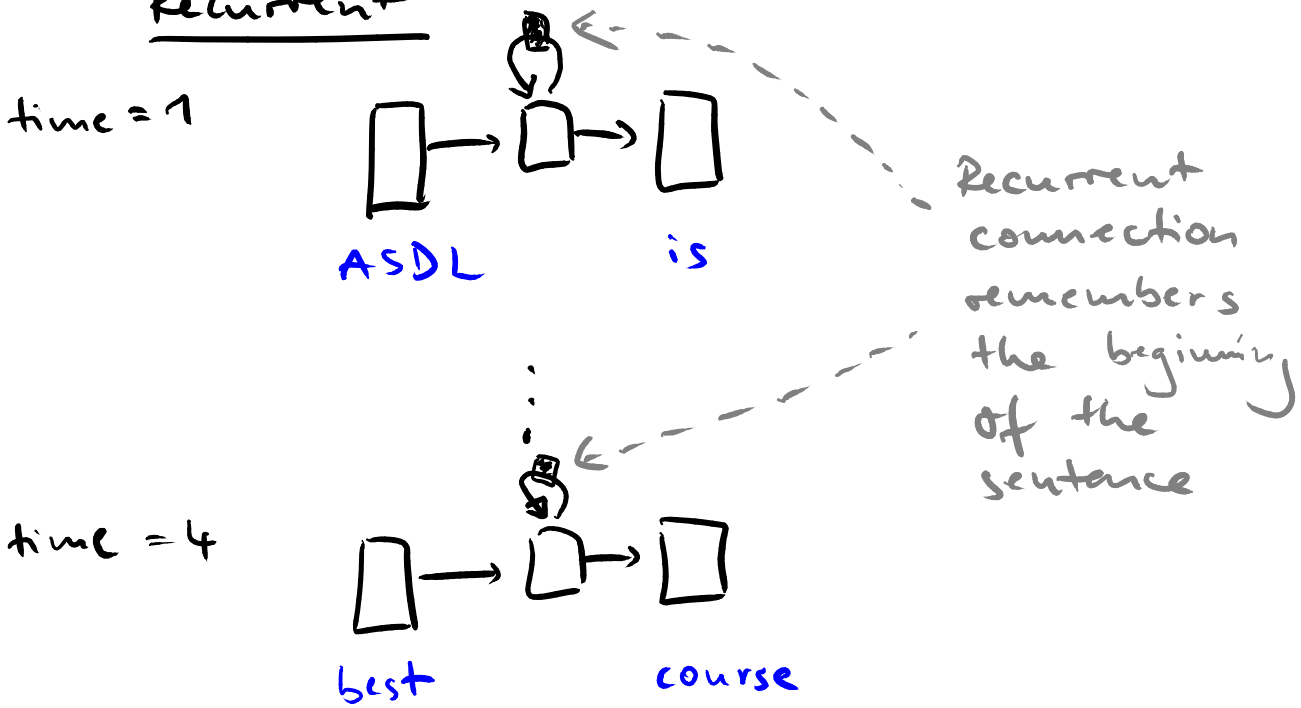
Example: Predict next word in sentence

ASDL is the best ... (course)

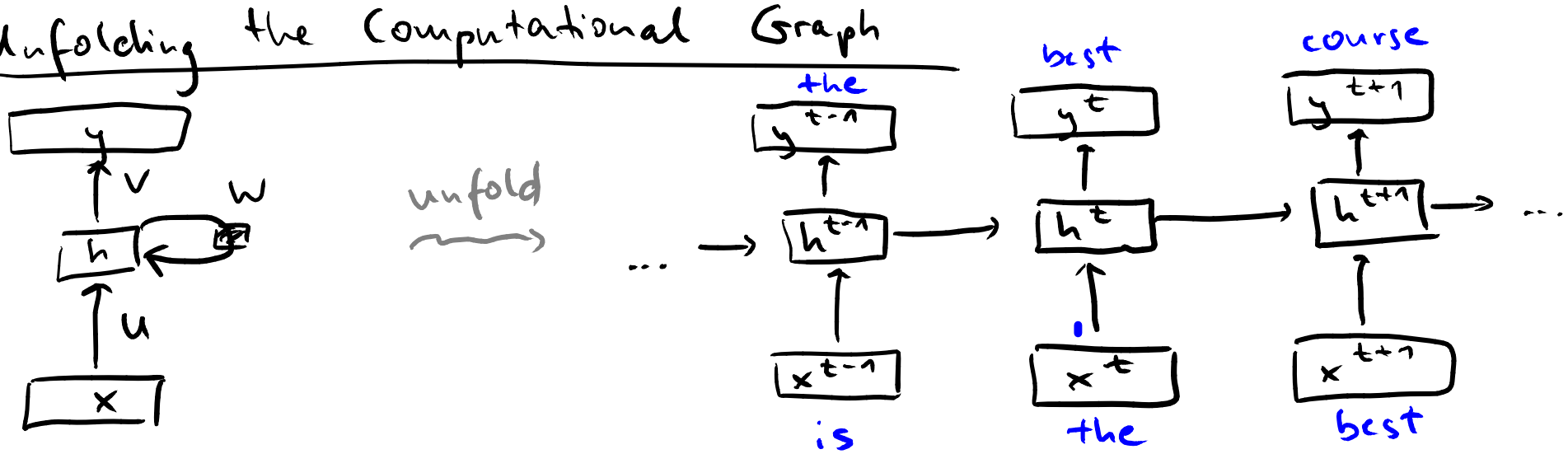
Feedforward



Recurrent

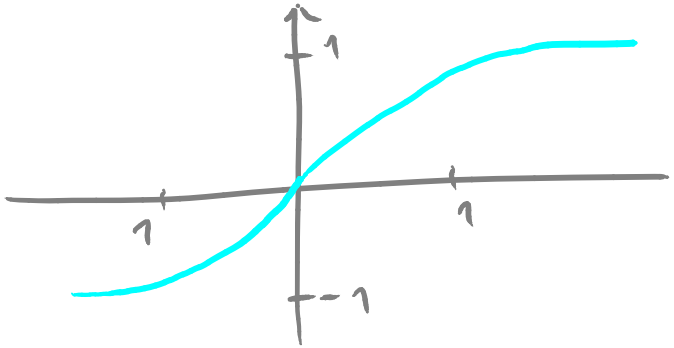


Unfolding the Computational Graph



$h^t = f(h^{t-1}, x^t)$ ------- e.g., $h^t = \tanh(W \cdot h^{t-1} + U \cdot x^t + b)$

$y^t = f(h^t)$ ---------- e.g., $y^t = \text{softmax}(V \cdot h^t + c)$



Softmax Function

- Goal: Interpret output vector as a probability distribution
- "Squashes" vector of k values $\in \mathbb{R}$ into vector of k values $\in [0, 1]$ that sum up to 1
- Definition:

  $\sigma(y)_j = \dfrac{e^{y_j}}{\sum_i^k e^{y_i}}$ for $j = 1, .., k$
- Example:

  $\sigma([1, 2, 3, 4, 1, 2, 3]) =$

  $[0.024, 0.064, 0.175, 0.475, 0.024, 0.064, 0.175]$

Quiz

Which of the following vectors may be the output of the softmax function?

1.) $y = [0.0, 0.0, 0.0, 0.0]$

2.) $y = [0.0, 0.25, 0.25, 0.5]$

3.) $y = [0.0, 1.0, 0.0, 0.0]$

4.) $y = [0.1, 0.1, 0.2, 0.3]$

Quiz

Which of the following vectors may be the output of the softmax function?

1.) $\cancel{y = [0.0, 0.0, 0.0, 0.0]}$ **sum is not 1**

2.) $y = [0.0, 0.25, 0.25, 0.5]$

3.) $y = [0.0, 1.0, 0.0, 0.0]$

4.) $\cancel{y = [0.1, 0.1, 0.2, 0.3]}$ **sum is not 1**

Note: Mathematically, 0 and 1 cannot occur. In practice, they may occur due to rounding of floating point numbers.

Applications of RNNs

Useful for tasks where the input (and maybe also the output) is a sequence

Examples:

- Unsegmented connected handwriting recognition
- Machine translation of natural languages
- Video classification by frames
- Speech recognition
- Sentiment analysis of twitter messages