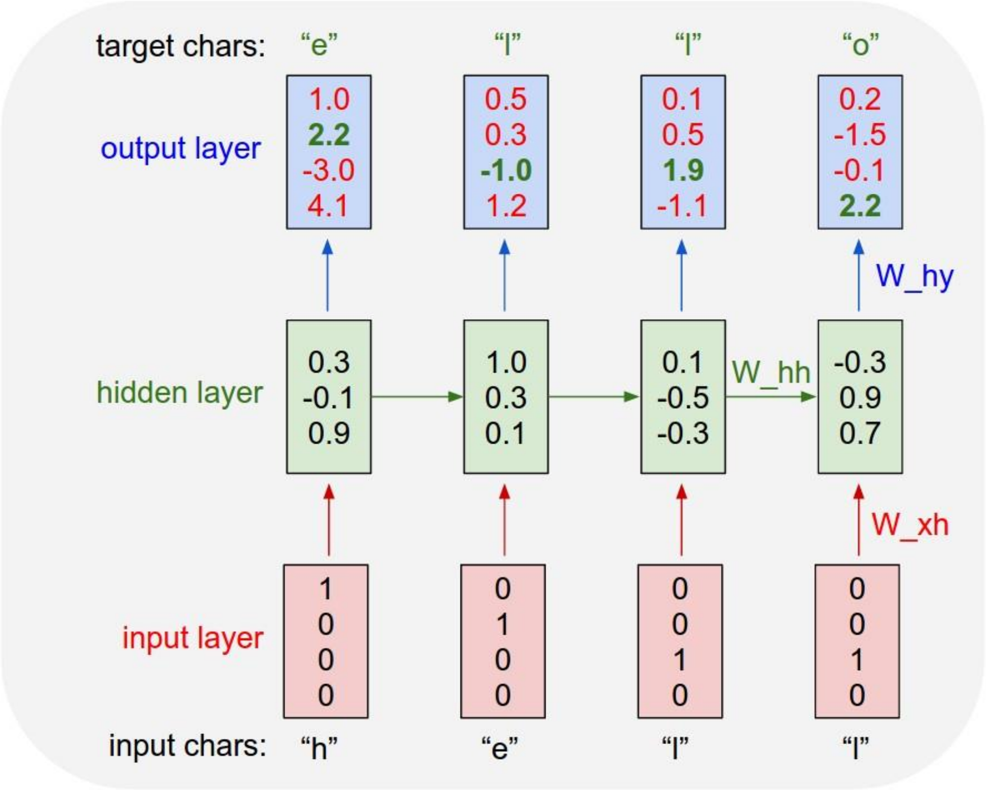
Learning in Recurrent Neural Networks, Part 2



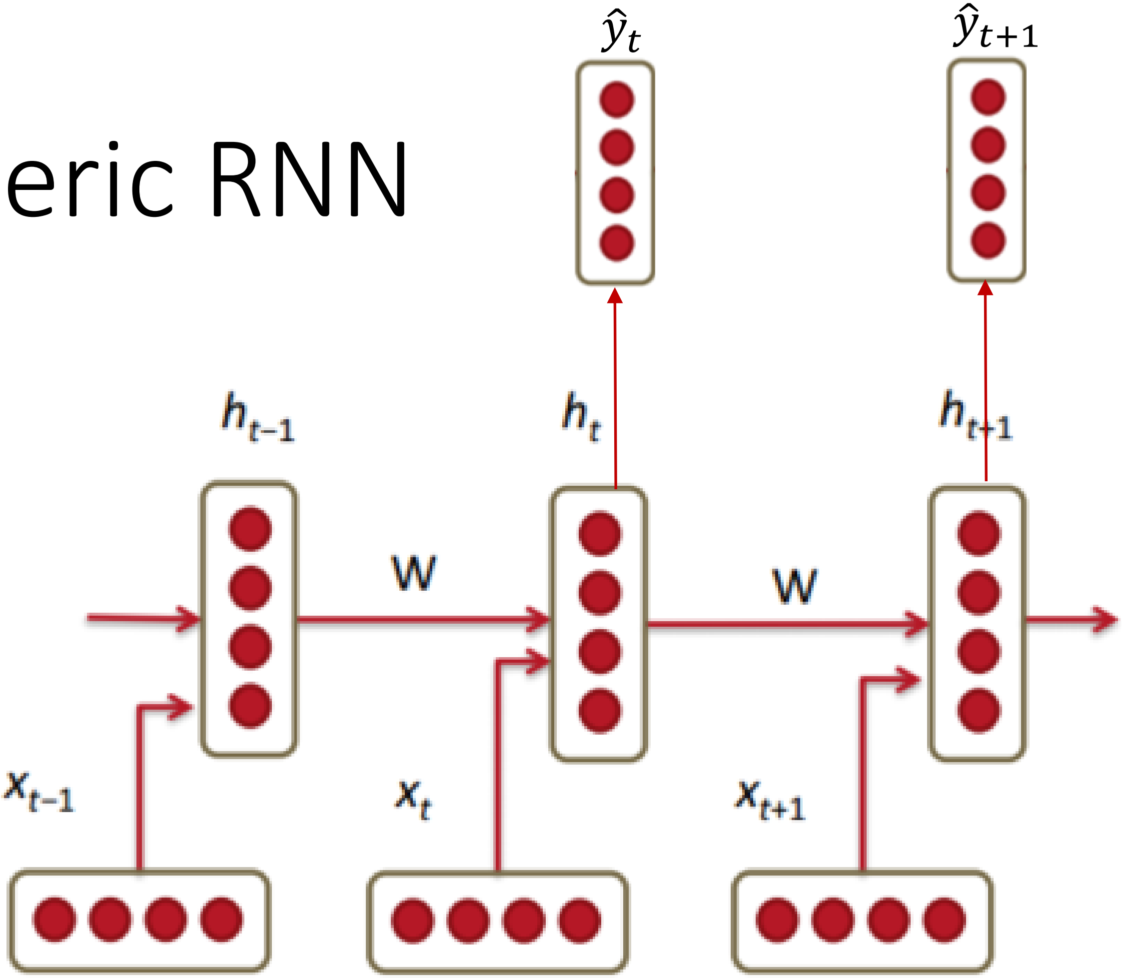
Andrey Karpathy http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Some slides from Richard Socher, Geoffrey Hinton, Andrej Karpathy

CSC321: Intro to Machine Learning and Neural Networks, Winter 2016

Michael Guerzhoy

Generic RNN

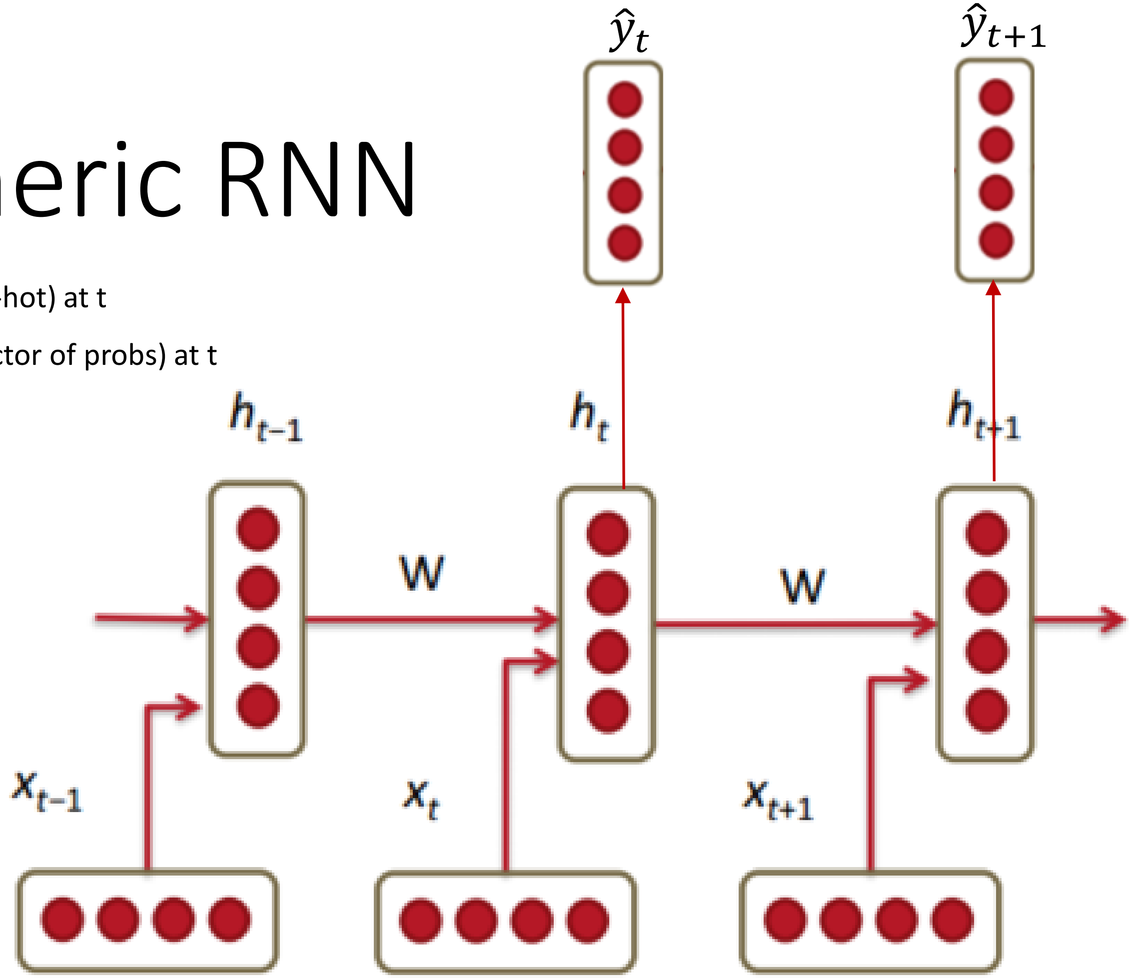


x_t-the t[th] character of the string ("the character at time t")
h_t- the hidden state at time t

Generic RNN

x_t – the input (one-hot) at t

$\hat{y}_t$ - predictions (vector of probs) at t



$$h_t = \sigma\left(W^{(hh)} h_{t-1} + W^{(hx)} x_t\right)$$
$$\hat{y}_t = softmax\left(W^{(S)} h_t\right)$$
$$\hat{P}\left(x_{t+1} = v_j \middle| x_1, x_2, \ldots, x_t\right) = \hat{y}_{t,j}$$

An RNN for strings with valid parentheses

- Valid parens: ()()(()), (()())
- Invalid parens:)(
- Algorithm for recognizing strings w/ valid parents:
  - For all t,
    (#open parens up to t) - (#closed parens up to t) >= 0

(#open parens up to t) - (#closed parens up to t) >= 0

- $x_t^0 = 1$: s[t]==“(“
- $x_t^1 = 1$: s[t]==“)“
- $x_t^2 = 1$: s[t]==“\n“
- h_t: (#open parens) – (#closed parens)
- $h_t = [1 \quad -1]\begin{bmatrix} x_t^0 \\ x_t^1 \end{bmatrix} + [1]h_{t-1}$
- $\hat{y}_t \approx \begin{cases} [.2, .8, 0]^T, \ h_t > 0 \\ [0, 0, 1]^T, \ h_t < 0 \\ [.5, 0, .5]^T, \ h_t \approx 0 \end{cases}$

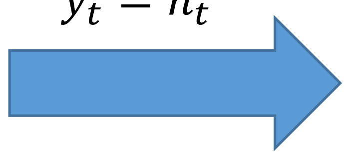
Not doable with a one-layer network, but doable with a two layer network

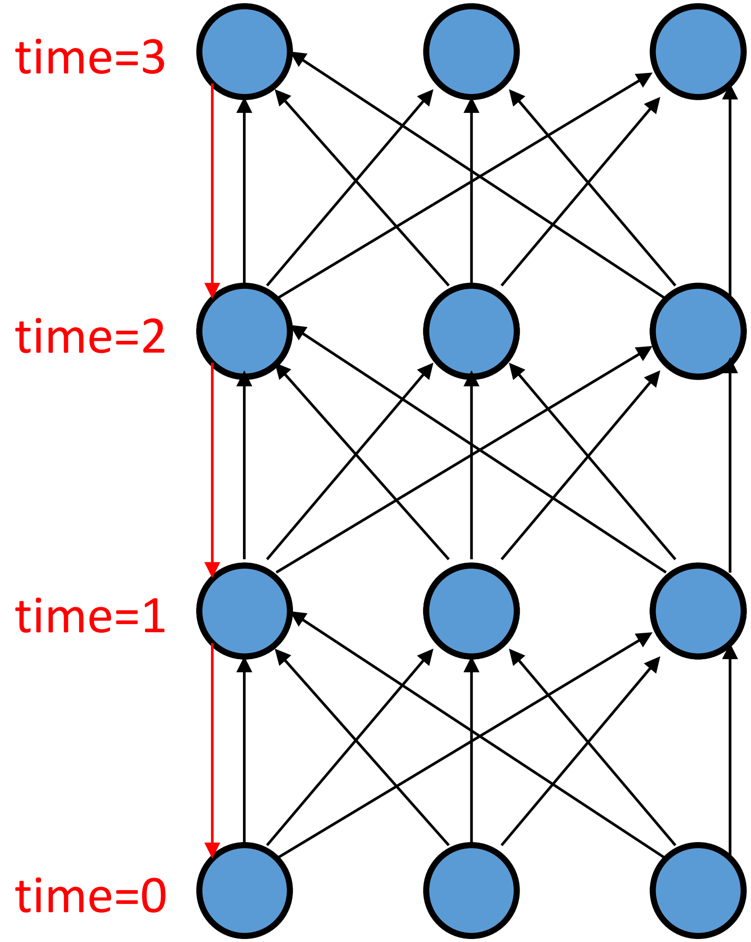
(*Note: really doable with one-layer networks too (approximately), but with a larger h – discussion on the board)*

RNN Gradient

- $\frac{\partial J}{\partial W} = \sum_t \frac{\partial J^{(t)}}{\partial W}$
- $\frac{\partial J^{(t)}}{\partial W} = \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial h_k} \frac{\partial h_k}{\partial W} = \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial \hat{y}_k} \frac{\hat{y}_k}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$

$\hat{y}_t = h_t$





Vanishing Gradient

- $\frac{\partial J^{(t)}}{\partial W}$=$\sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial W} = \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial h_t} \frac{\partial h_t}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial h_k} \frac{\partial h_k}{\partial W}$
- $\frac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}$
- Problem:
  - $\left|\frac{\partial h_j}{\partial h_{j-1}}\right| < 1$ for all j
  - Leads to $\frac{\partial h_t}{\partial h_k}$ being very small