Modeling Sequences Conditioned on Context with RNNs

Sargur Srihari
srihari@buffalo.edu

Topics in Sequence Modeling

- Overview

1. Unfolding Computational Graphs
2. Recurrent Neural Networks
3. Bidirectional RNNs
4. Encoder-Decoder Sequence-to-Sequence Architectures
5. Deep Recurrent Networks
6. Recursive Neural Networks
7. The Challenge of Long-Term Dependencies
8. Echo-State Networks
9. Leaky Units and Other Strategies for Multiple Time Scales
10. LSTM and Other Gated RNNs
11. Optimization for Long-Term Dependencies
12. Explicit Memory

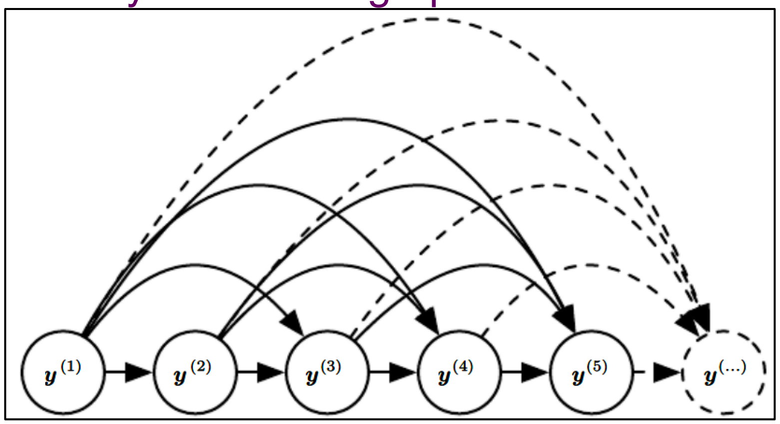
Topics in Recurrent Neural Networks

0. Overview
1. Teacher forcing for output-to-hidden RNNs
2. Computing the gradient in a RNN
3. RNNs as Directed Graphical Models
4. Modeling Sequences Conditioned on Context with RNNs

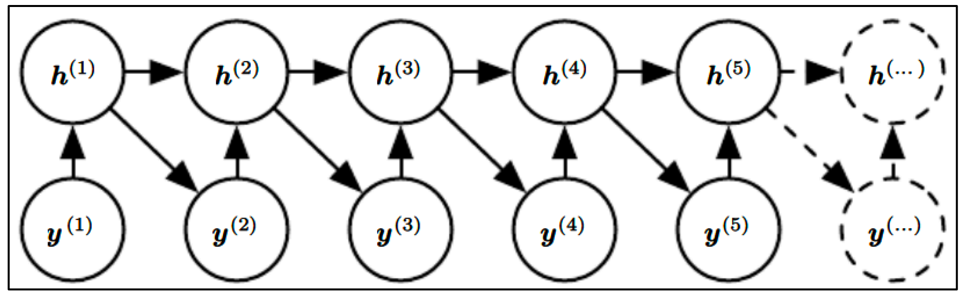
Graphical models of RNNs *without* inputs

- Directed graphical models of RNNs *without* inputs
  - having a set of random variables $\boldsymbol{y}^{(t)}$:

Fully connected graphical model



Efficient parameterization based on $\boldsymbol{h}^{(t)}=f(\boldsymbol{h}^{(t-1)},\boldsymbol{x}^{(t)};\boldsymbol{\theta})$



- RNN graphical models can be extended to the conditional distribution of y *given* the inputs $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$,..$\boldsymbol{x}^{(\tau)}$

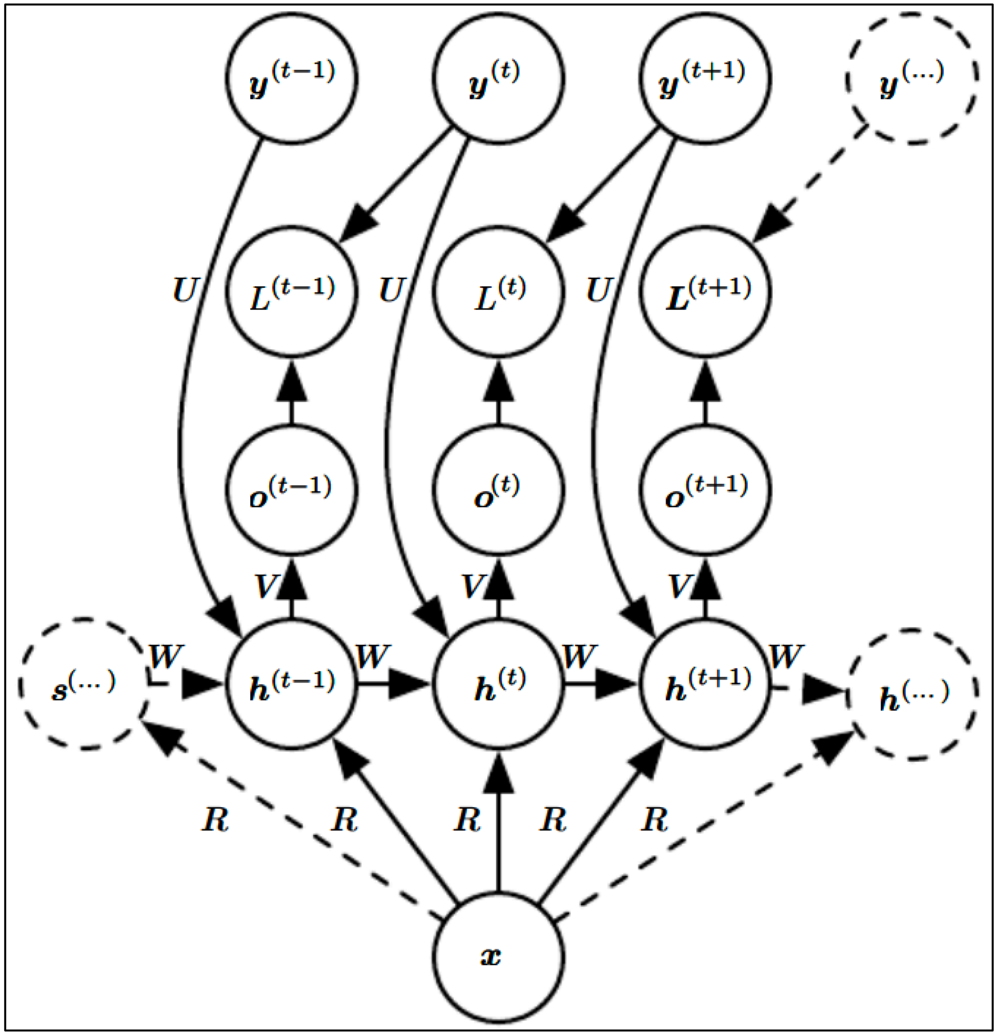
Extending RNNs to represent conditional $P(\boldsymbol{y}|\boldsymbol{x})$

- A model representing a variable $P(\boldsymbol{y}|\boldsymbol{\theta})$ can be reinterpreted as a model representing a conditional distribution $P(\boldsymbol{y}|\boldsymbol{\omega})$ with $\boldsymbol{\omega}=\theta$
- We can extend such a model to represent a distribution $P(\boldsymbol{y}|\boldsymbol{x})$ by using the same $P(\boldsymbol{y}|\boldsymbol{\omega})$ as before but making $\boldsymbol{\omega}$ a function of $\boldsymbol{x}$
- In the case of RNNs this can be achieved in several ways
  - Most common choices are described next

Taking a single vector $\boldsymbol{x}$ as an extra input

- Instead of taking a sequence $\boldsymbol{x}^{(t)},\ t=1,..,\tau$ as input we can take a single vector $\boldsymbol{x}$ as input
- When $\boldsymbol{x}$ is a fixed-size vector we can simply make it an extra input of the RNN that generates the $\boldsymbol{y}$ sequence
- Common ways of providing an extra input to RNN are
    - An extra input at each time step, or
    - As the initial state $\boldsymbol{h}^{(0)}$, or
    - Both
- The first and common approach is illustrated next
    - The interaction between the input $\boldsymbol{x}$ and each hidden unit vector $\boldsymbol{h}^{(t)}$ is parameterized by a newly introduced weight matrix R that was absent from the model with only y values

RNN to map a fixed length vector $\boldsymbol{x}$ over sequences $\mathbf{Y}$



Appropriate for tasks such as image captioning where a single image is input which produces a sequence of words describing the image.

Each element of the observed output $\boldsymbol{y}^{(t)}$ of the observed output sequence serves both as input (for the current time step) and during training as target

RNN to receive a sequence of vectors $\boldsymbol{x}^{(t)}$ as input

- RNN described by $\boldsymbol{a}^{(t)}=\boldsymbol{b}+W\boldsymbol{h}^{(t\text{-}1)}+U\boldsymbol{x}^{(t)}$ corresponds to a conditional distribution $P(\boldsymbol{y}^{(1)},..,\ \boldsymbol{y}^{(\tau)}|\boldsymbol{x}^{(1)},..,\ \boldsymbol{x}^{(\tau)})$
- It makes a conditional independence assumption that this distribution factorizes as

$$\prod_t P(\boldsymbol{y}^{(t)}\,|\boldsymbol{x}^{(1)},..,\boldsymbol{x}^{(t)})$$

- To remove the conditional independence assumption, we can add connections from the output at time t to the hidden unit at time $t+1$ (see next slide)
  - The model can then represent arbitrary probability distributions over the $\boldsymbol{y}$ sequence
- Limitation: both sequences must be of same length
  - Removing this restriction is discussed in Section 10.4

Removing conditional independence assumption

Connections from previous output to current state allow RNN to model arbitrary distribution over sequences of $\boldsymbol{y}$ given sequences of the same length

Compare to model that is only able to represent distributions in which the $\boldsymbol{y}$ values are conditionally independent from each other given $\boldsymbol{x}$ values