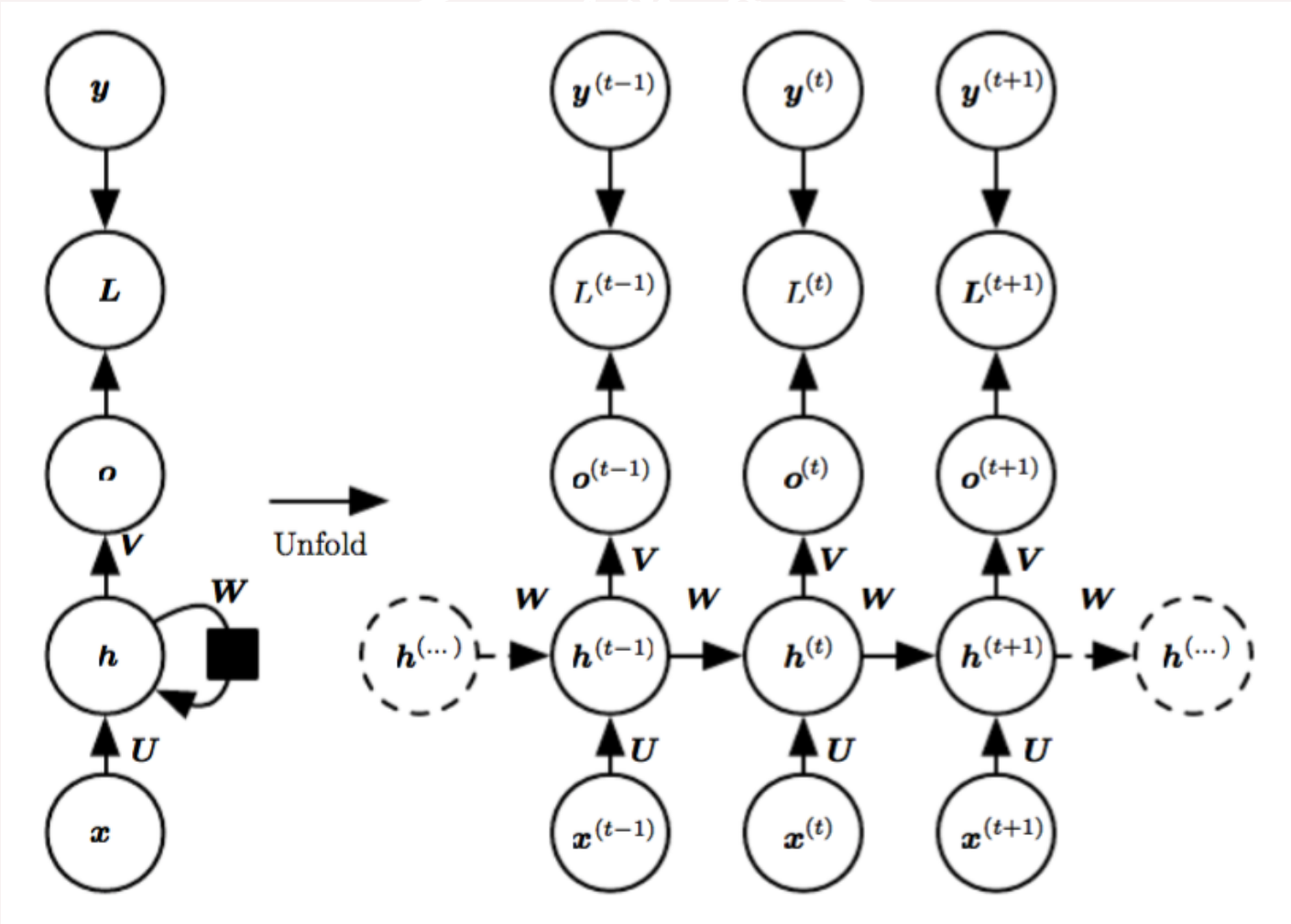
Deep Learning - Study Circle
Sequence Modeling: Recurrent and Recursive Nets

Martin Karlsson

Dept. Automatic Control, Lund University, Lund, Sweden

October 26, 2016

RNN Structure



Warm-up examples

- Recurrent neural network to generate new first names
- I also added two unstructured sequences. Can you see which ones?

Rudi Levette Berice Lussa Hany Mareanne Chrestina Carissy Marylen Hammine Janye Marlise Zstmyoi Jacacrie Hendred Romand Charienna Charisa Allisa Anatha Cathanie Geetra Alexie Lstaogf Jerin Cassen Herbett Cossie Velen Daurenge Robester Shermond

- Compare with auto-generated images. Generally more structured than white noise.
- More "names" can be found here. (Links are clickable.)
- Training names can be found here

Warm-up examples

- Recurrent neural network to generate new first names
- I also added two unstructured sequences. Can you see which ones?

Rudi Levette Berice Lussa Hany Mareanne Chrestina Carissy Marylen Hammine Janye Marlise **Zstmyoi** Jacacrie Hendred Romand Charienna Charisa Allisa Anatha Cathanie Geetra Alexie **Lstaogf** Jerin Cassen Herbett Cossie Velen Daurenge Robester Shermond

- Compare with auto-generated images. Generally more structured than white noise.
- More "names" can be found here. (Links are clickable.)
- Training names can be found here

Warm-up examples

DeepMind's WaveNet

- Generative model for raw audio
- Text-to-speech
- Music generation

Outline

This presentation should serve as an overview.
Details are found under clickable links.

- Introduction
- Unfolding computational graphs
- Recurrent neural networks (RNNs)
  - Design patterns
  - Forward- and back-propagation
  - Bidirectional RNNs
  - Long-short term memory (LSTM)
  - Examples
- Recursive neural networks
- Assignment

Introduction

- Processing sequential data
- Scale well to long sequences
- Commonly process sequences of variable length
- Compare with CNNs; Specialized for grid data, and can sometimes process input of variable size
- Both RNNs and CNNs rely on parameter sharing
- Sources of information:
  - Neural Networks for Machine Learning, by Geoffrey Hinton
  - Ch. 10 in Deep Learning by Bengio *et al.*
  - Supervised Sequence Labelling with Recurrent Neural Networks by Alex Graves

Unfolding computational graphs

Consider the hidden layer h driven by an input x^t

$$h^t = f(h^{t-1}; x^t; \theta) \tag{1}$$

The unfolded recurrence up to t can be represented by g^t:

$$h^t = g^t(x^t, x^{t-1}, x^{t-2}, ..., x^2, x^1; \theta) \tag{2}$$

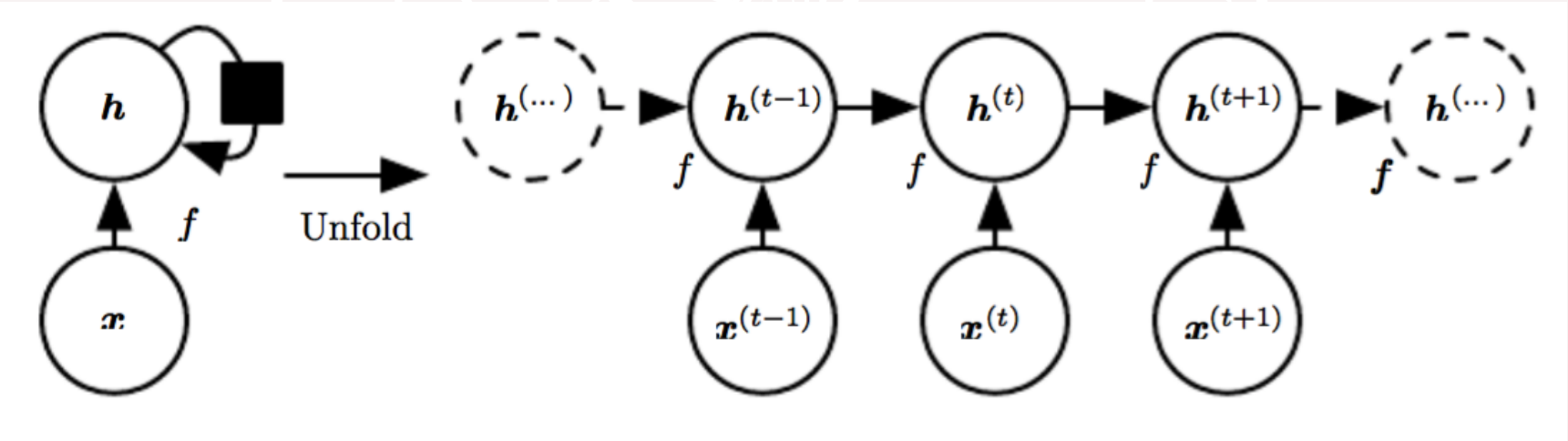
g takes the whole sequence as input. In turn, g can be factorized to repeated application of f.

Major advantages:

- Learned model always same input dimension
- Same transition function f with same parameters θ can be used for every time step (parameter sharing)

Unfolding computational graphs

- Recurrent network (without outputs)
- Black square indicates one time step delay
- The networks below are equivalent



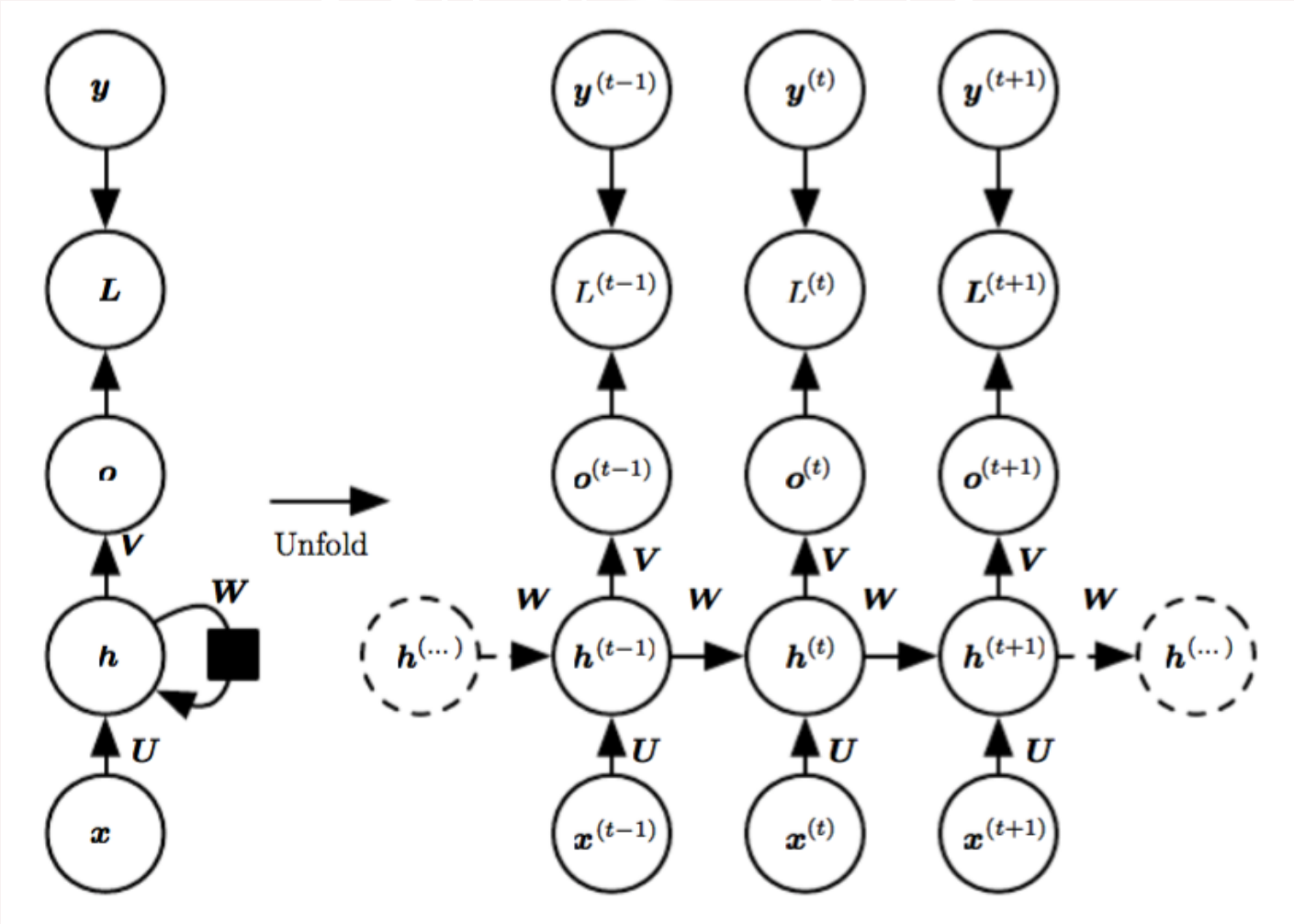
Design patterns

Three typical patterns

1. One output at each time step, recurrence between hidden units
2. One output at each time step, recurrence from output to hidden unit at next time step
3. One output for an entire input sequence, recurrence between hidden units

Design patterns

1. One output at each time step, recurrence between hidden units



Design patterns

1. One output at each time step, recurrence between hidden units

Forward propagation:

$$h^t = \tanh(b + Wh^{t-1} + Ux^t) \tag{3}$$

$$o^t = c + Vh^t \tag{4}$$

$$\hat{y}^t = \text{softmax}(o^t) \tag{5}$$

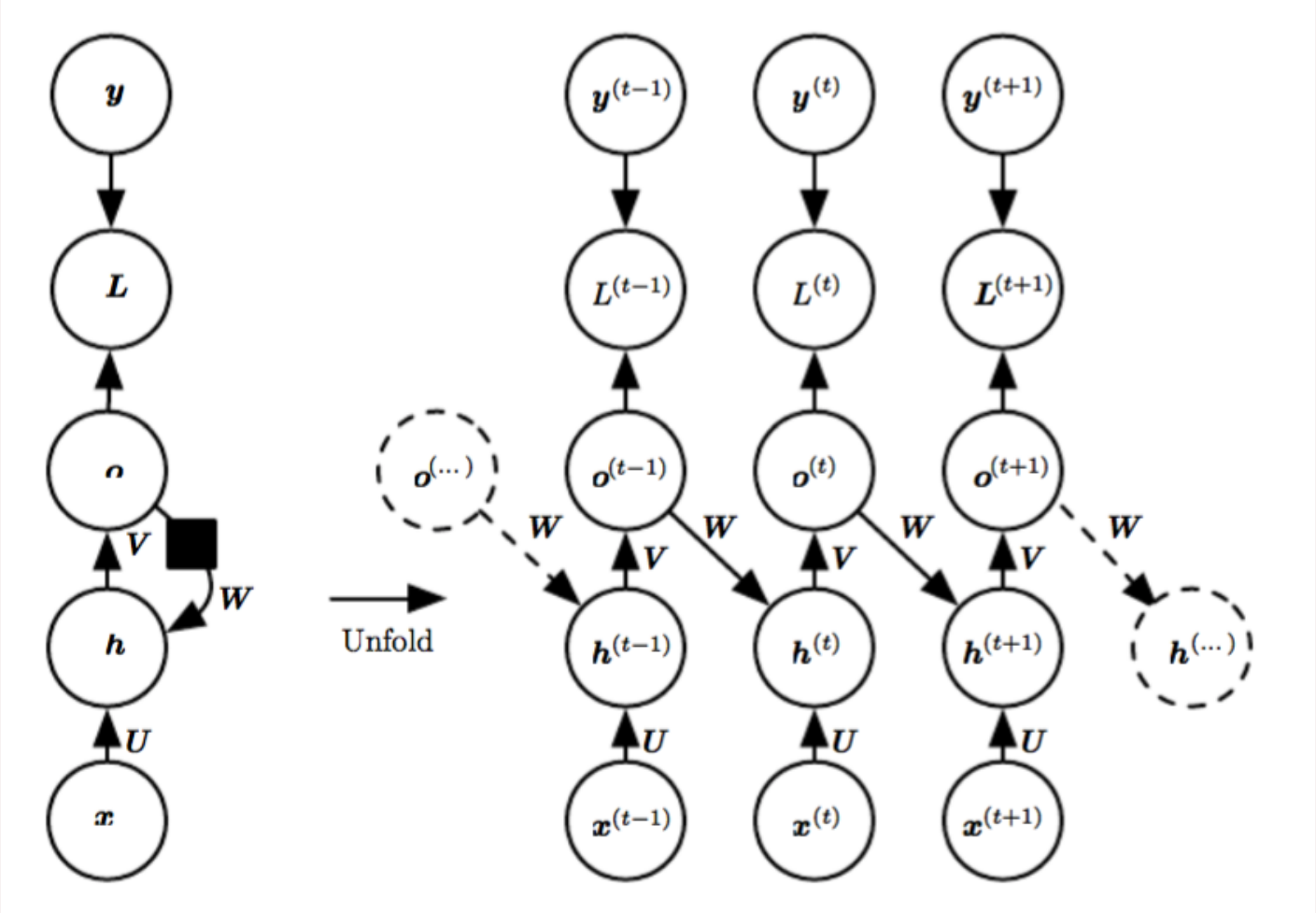
b and c are bias vectors W, U and V are weight matrices

Design patterns

- $\hat{y}$ and $y \Rightarrow$ cost function L
- Minimize L through back-propagation on unrolled graph
- Back-propagation through time (BPTT)
- Mini-batches for training
- Recursively in time?

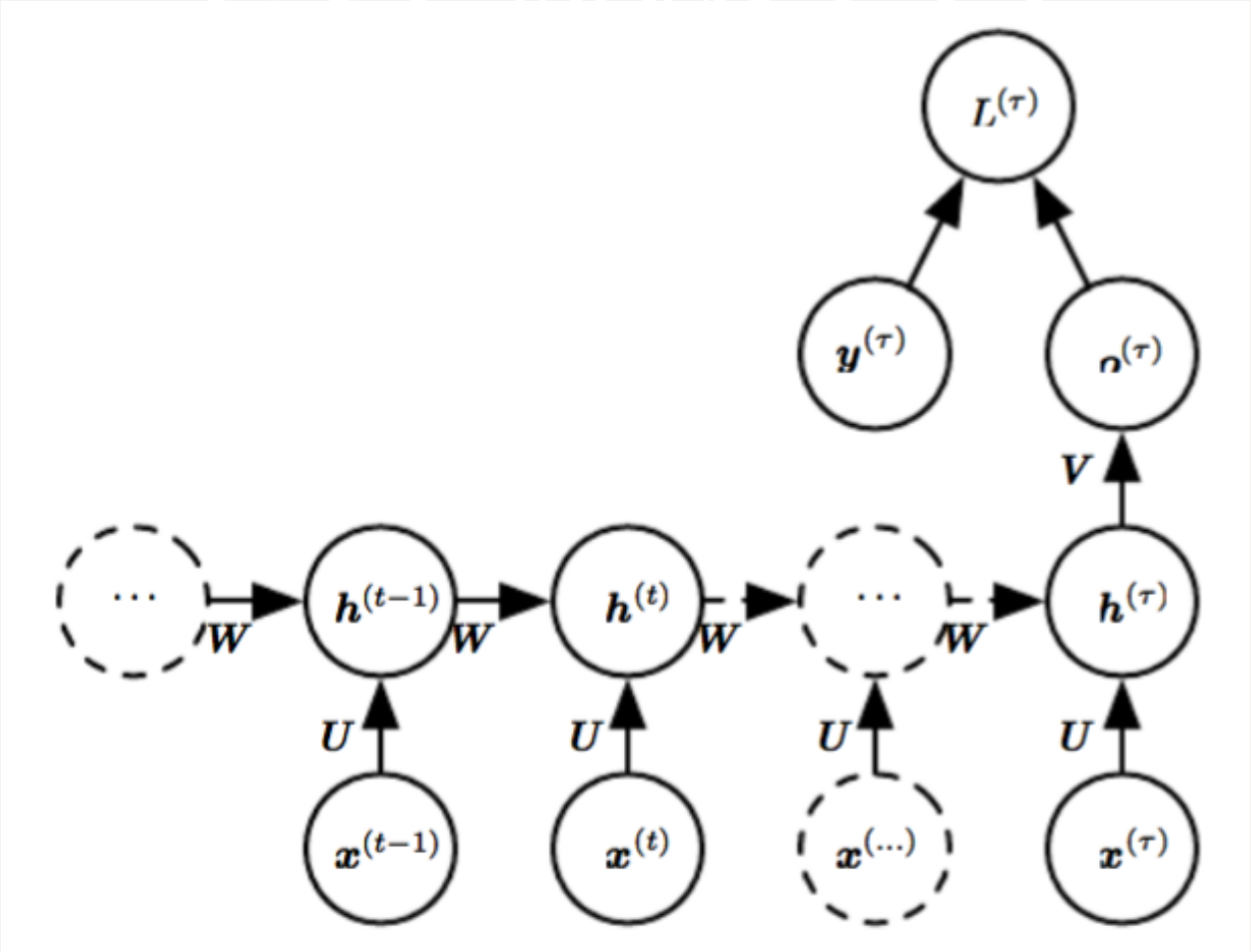
Design patterns

2. One output at each time step, recurrence from output to hidden unit at next time step



Design patterns

3. One output for an entire input sequence, recurrence between hidden units. Sequence classification(?).



Bidirectional RNNs

- We have looked at causal structures
- However, y^t might depend on the whole sequence. For instance
  - Speech recognition
  - Handwriting recognition
  - Bioinformatics
- A bidirectional RNN combines an RNN that moves forward in time, with an RNN that moves backward in time
- Each output unit then depends on both past and future inputs
- This implies a delay (relevant when running in real-time)

Long Short-Term Memory (LSTM)

- Problem with exploding/vanishing gradient
- LSTM very successful in several applications
- Learn long-term dependencies more easily than original RRNs
- One "cell" consists of a state s, a forget gate f, and much more...

Examples

- Examples from The Unreasonable Effectiveness of Recurrent Neural Networks
- Auto-generated "Shakespeare's play"
- letter-by-letter
- 3-layer RNN with 512 hidden per layer was used
- This is non-sense, but follows a certain structure.

PANDARUS: Alas, I think he shall be come approached and the day When little srain would be attain'd into being never fed, And who is but a chain and subjects of his death, I should not sleep.

Second Senator: They are away this miseries, produced upon my soul, Breaking and strongly should be buried, when I perish The earth and thoughts of many states.

DUKE VINCENTIO: Well, your wit is in the care of side and that.

Examples

- "Wikipedia text":

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict.

Recursive neural networks

- Generalizes RNN from a chain to a tree
- Open question how these trees should be structured
- Advantage compared to RNNs: depth reduced from τ to $O(\log(\tau))$, for input of fixed length τ
- Subjective remarks:
  - Otherwise difficult to draw any general conclusion
  - Seem much less ubiquitous than RNNs

Assignment

- Train and evaluate an RNN (or recursive net), on a dataset of your choice.
- For a flying start, you could have a look at this tutorial
- For the Theano expert: Music modeling with RNN-RBM. This would count as both the RBM- and the RNN assignments

Good Luck!