L101, lecture 6

Ann Copestake

Computer Laboratory
University of Cambridge

November 2017

Outline

Artificial versus biological NNs

Overview/recap of Restricted Boltzmann Machines

Deep Learning and sequences

RNNs and LSTMs

The plan . . .

- Next lecture: 9am on November 9th
  Compositional semantics and compositional distributional semantics.
- Final lecture: 3pm on November 16th
  Question and answer session on Machine Learning for NLP (with guest):
  ideally questions in advance (email me), ideally before next week's lecture

Outline.

Artificial versus biological NNs

Overview/recap of Restricted Boltzmann Machines

Deep Learning and sequences

RNNs and LSTMs

Artificial vs biological NNs

- ANNs and BNNs both take input from many neurons and carry out simple processing (e.g., summation), then output to many neurons.
- ANNs are still tiny: biggest c160 billion parameters. Human brain has tens of billions of neurons, each with up to 100,000 synapses.
- Brain connections are much slower than ANNs: chemical transmission across synapse. Bigger size and greater parallelism (more than) makes up for this.
- Neurotransmitters are complex and not well understood: biological neurons are only crudely approximated by on/off firing.

Artificial vs biological NNs (continued)

- Brains grow new synapses and lose old ones: individual brains evolve (Hebbian Learning: “Neurons which fire together wire together”).
- Brains are embodied: processing sensory information, controlling muscles. There is no hard division between these parts of the brain and concepts/reasoning (e.g., experiments with *kick* vs *hit*).
- Brains have evolved over (about) 600 million years (more if we include nerve nets, as in jellyfish).
- Brains are expensive (about 20% of a person’s energy), but much more efficient than ANNs.
- and . . .

Brains and syntax

- Exact neurological basis for syntax processing in language is unclear, but assume repurposed neuronal ensembles.
- Limitations: particularly working memory (WM).
- gist representations: store propositions rather than actual linguistic input.
- Processing is incremental, small number of alternatives carried forward, depends strongly on prediction (non-linguistic and linguistic).
- Human languages and human brains co-evolved: so human processing limitations relevant to NLP?

WM constraints

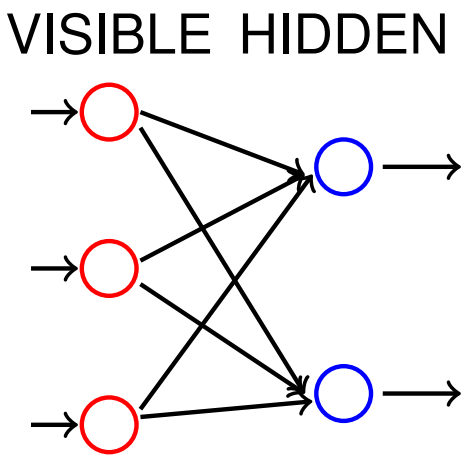
weil ich vielleicht dem Mann den Hund morgen gebe, komme ich
because I possible the-DAT man the-ACC dog tomorrow give come I
'I'm coming because I may give the dog to the man tomorrow'

6 different pieces of information before *gebe* is reached. all separate items in WM???

Outline.

Introduction to RBMs

- Boltzmann machine: arbitrary interconnections between units. Not effectively trainable in general.
- Restricted Boltzmann Machine (RBM): one input and one hidden layer, no intra-layer links.



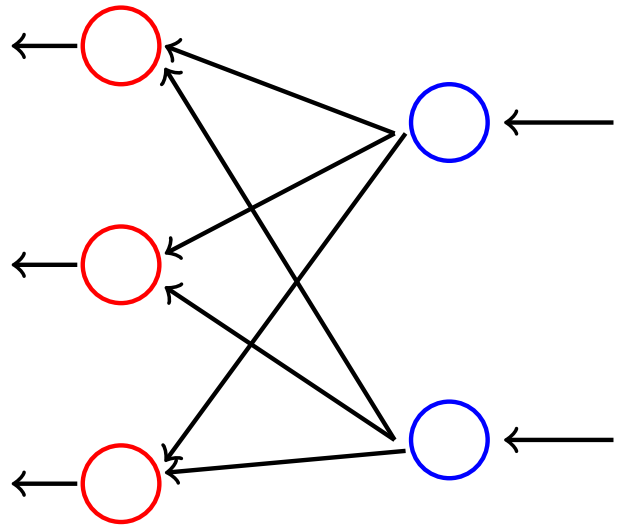
$w_1, \ldots w_6$ b (bias)

Introduction to RBMs

- Hidden layer.
  - One hidden layer can model arbitrary function (see Goldberg 2016), but not necessarily trainable.
- RBM: usually fully connected between two layers but sparse RBMs are possible.
- The layers allow for efficient implementations — weights can be described by a matrix, fast computation.
- Generative probabilistic model: bipartite graph units in hidden layer conditionally independent given input layer and vice versa.
- RBMs allow efficient Gibbs sampling for training (as a step in the overall procedure).
- Goodfellow et al 2016 (`http://www.deeplearningbook.org`)
  Murphy 'Machine Learning: a Probabilistic Perspective'

`https://deeplearning4j.org/restrictedboltzmannmachine`

Training RBMs: reconstruction of input



b' (bias) $w_1, \ldots w_6$

- Forward pass: $P(\text{output}|\text{input}; w)$
- Backprop: $P(\text{input}|\text{output}; w)$
- Overall, joint probability: $P(\text{input}, \text{output})$

Some (hopefully) intuitive explanations of terminology

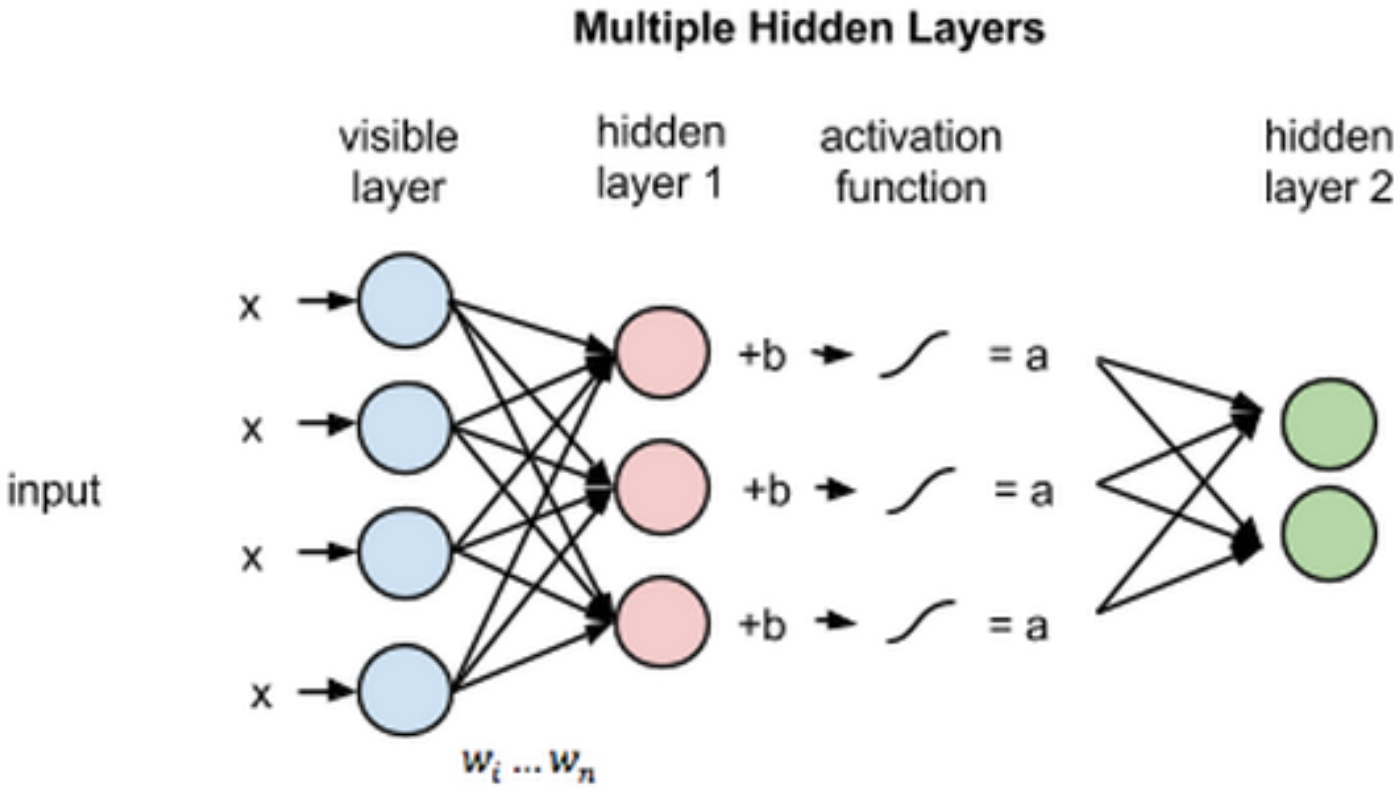
- regularization: methods of choosing the priors to avoid overfitting (less necessary if lots of data). e.g., fitting a smooth curve rather than a wiggly one. dropout often more effective.
- energy function: approximation to probabilities of states (always > 0) in undirected models. Close connection with physics (hence terminology).
- back-propagation aka backprop: information about the cost flowing backward through the network (e.g., computing the gradient).
- stochastic gradient descent: performing learning using the gradient.

Outline.

Deep Learning

- One of the most successful deep learning architectures involves combining RBMs, so the output from one RBM is the input to the next.
- RBMs can be trained separately and then fine-tuned in combination.
- The layers allow for efficient implementations and successive approximations to concepts.
- Unlike LDA (and other similar models), there is no predefined interpretation for the latent variables.
- Different architecture needed for sequences and most language problems (RNN/LSTM).

Combining RBMs: deep learning



https://deeplearning4j.org/restrictedboltzmannmachine

Outline.

Motivation

- Standard NNs cannot handle sequence information well.
- Can pass them sequences encoded as vectors (or CBOW), but input vectors are fixed length.
- Models are needed which are sensitive to sequence input and can output sequences.
- RNN: Recurrent neural network.
- Long short term memory (LSTM): development of RNN, more effective for most language applications.
- More info: `http://neuralnetworksanddeeplearning.com/` (mostly about simpler models and CNNs)
  `https://karpathy.github.io/2015/05/21/rnn-effectiveness/`

  `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

Recurrent Neural Networks

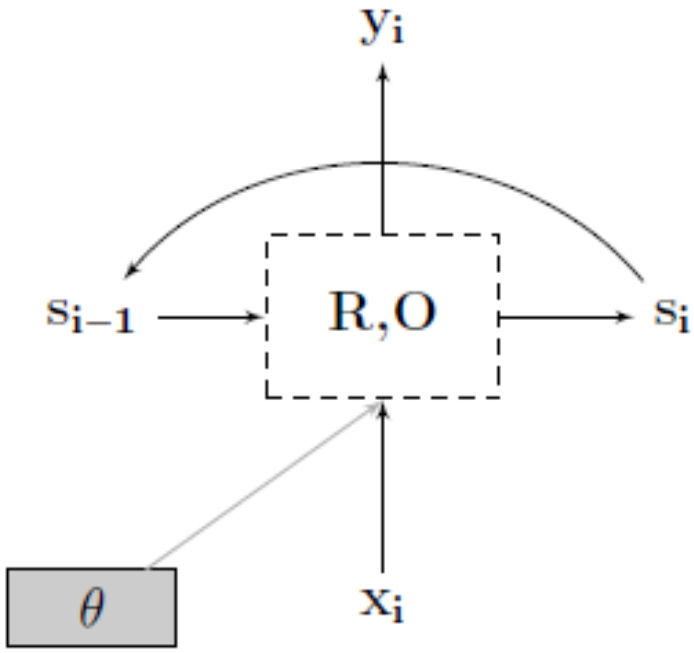


Figure 5: Graphical representation of an RNN (recursive).

From: Goldberg 2016 JAIR `http://www.jair.org/media/4992/live-4992-9623-jair.pdf`

Recurrent Neural Networks

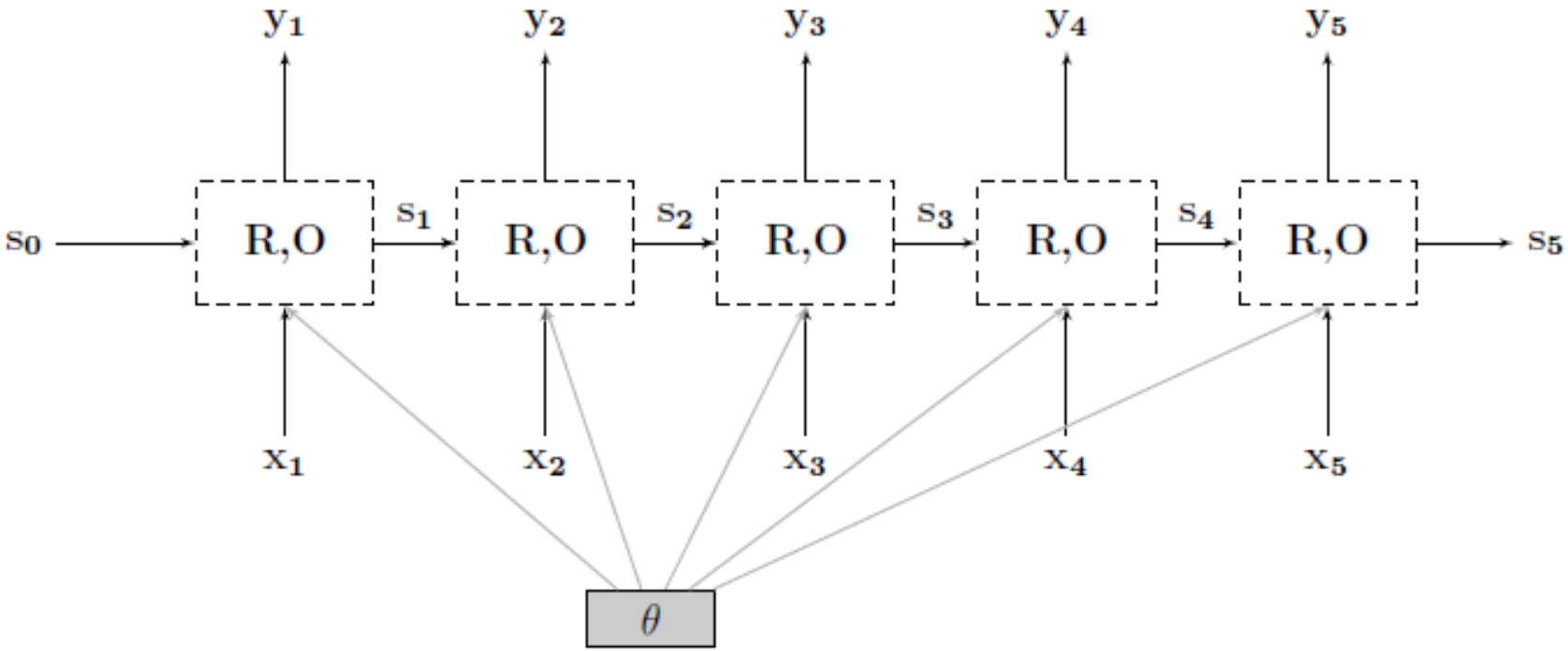


Figure 6: Graphical representation of an RNN (unrolled).

Note: all sequences are finite, parameters are shared.

From: Goldberg 2016 JAIR `http://www.jair.org/media/4992/live-4992-9623-jair.pdf`

Sequences

- Video frame categorization: strict time sequence, one output per input.
- Real-time speech recognition: strict time sequence.
- Neural MT: target not one-to-one with source, order differences: encoder-decoder model.
- Many language tasks: best to operate left-to-right and right-to-left (e.g., bi-LSTM).
- attention: model ‘concentrates’ on part of input relevant at a particular point. Caption generation: treat image data as ordered, align parts of image with parts of caption.

Encoder-decoder

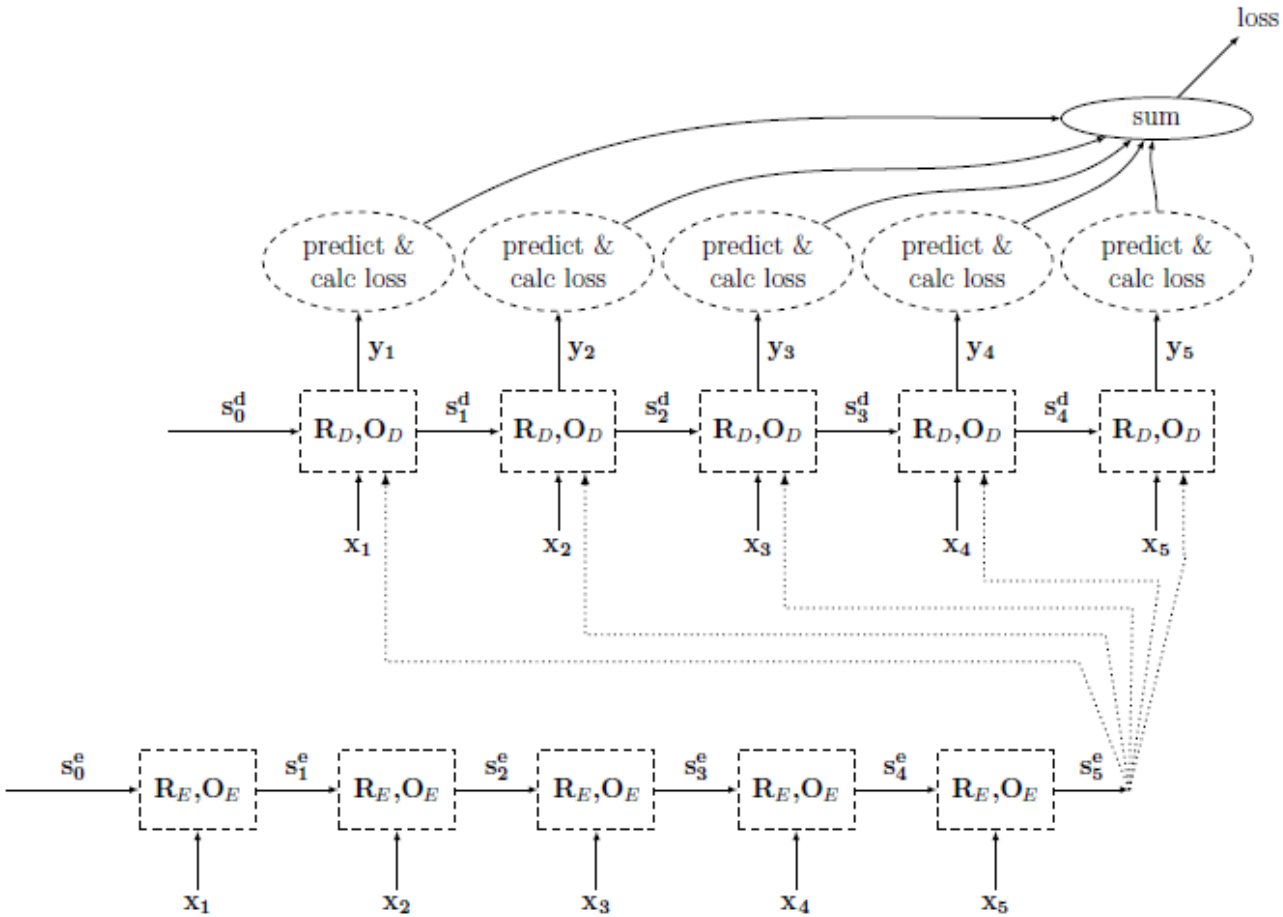


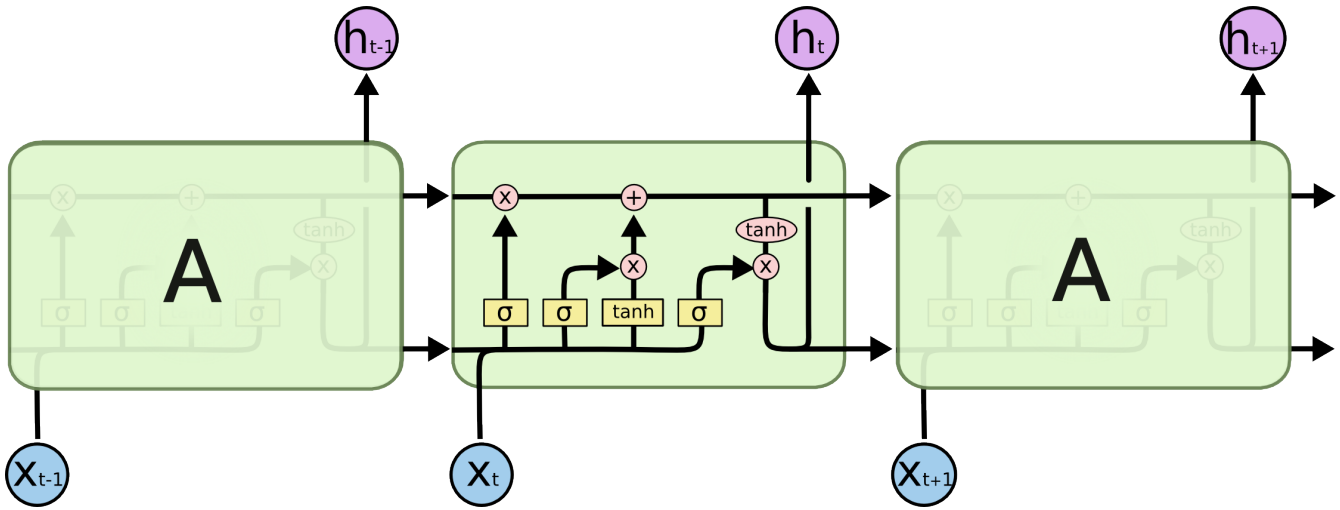
Figure 9: Encoder-Decoder RNN Training Graph.

From: Goldberg 2016 JAIR `http://www.jair.org/media/4992/live-4992-9623-jair.pdf`

Long Short Term Memory Networks

- An RNN has just one layer in its repeating module, suffers from the vanishing gradients problem
- An LSTM has memory cells controlled by gating components:
  - Forget gate layer: look at previous cell state and current input, and decide which information to throw away.
  - Input gate layer: see which information in the current state we want to update.
  - Update layer: propose new values for the cell state.
  - Output layer: Filter cell state and output the filtered result.
- For instance: store number of subject until another subject is seen.

Long Short Term Memory Networks



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTMs vs other models

- LSTMs are now the default for speech recognition.
- Other models have been proposed, but LSTMs still the most effective for language modelling when experiments are done carefully.
  On the State of the Art of Evaluation in Neural Language Models (`https://arxiv.org/abs/1707.05589`)

The plan . . .

- Next lecture: 9am on November 9th
  Compositional semantics and compositional distributional semantics.
- Final lecture: 3pm on November 16th
  Question and answer session on Machine Learning for NLP (with guest):
  ideally questions in advance (email me), ideally before next week's lecture