UPPSALA UNIVERSITET

A Quick Introduction to Neural MT

Christian Hardmeier

2016-05-16

UPPSALA UNIVERSITET

Why this lecture?

- For about 15 years, the MT world was relatively static.
- State of the art defined by *phrase-based SMT* and *syntax-based SMT*.
- Well-known strengths and weaknesses.
- *Neural MT* is a new, quite different approach to MT that seems to outperform the previous methods.

Deep Learning **Continuous-space NLP**
Neural Networks

UPPSALA UNIVERSITET

Deep Learning

- Machine learning paradigm that gained popularity very recently.
- First breakthroughs in computer vision.
- Multiple layers of prediction:
  "Automated feature engineering"

UPPSALA UNIVERSITET

Deep Learning

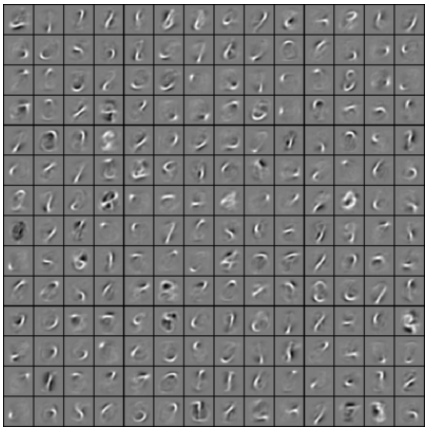


Image source: `http://deeplearning.stanford.edu/wiki/index.php/Exercise%3AVectorization`

UPPSALA UNIVERSITET

Continuous-Space Methods

- NLP traditionally treated words as discrete, incomparable units.
- Continuous-space methods map them into a vector space where you can compute similarities.
- Methods: Word cooccurrence or deep learning.
- With deep learning, we can train word embeddings for specific objectives.

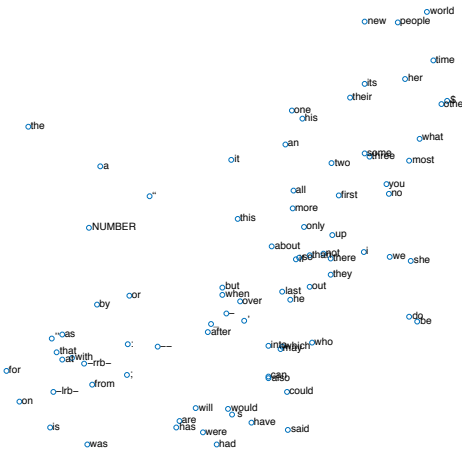
UPPSALA UNIVERSITET

Discrete Words

Lockheed	ICCAT	closed	pride	OHIO
shadowy	Ernest	hangout	solution	homicidal
Pacific	things	far-ranging	enables	Akram
communicates	triangle	taxed	secrets	receipts
taken	Spinelli	dates	Cost	clash
district	relative	visa	captains	abilities
Organization	Austrian	inflows	Loyola	whatever
Primakov	upstaging	guidelines	authors	complaining
oath	marched	soldiers	geology	drifts
seen	provide	adaptation	enterprises	Valdis
un-associated	misguided	non-Serb	writing	doubtless
frankly	anti-Semitism	10-1	operators	Genocide
camouflage	gathered	adopts	bags	shunning
approaching	aspirin	maximum	expenditure	some
footsteps	Dutch	stressed	writers	between
mischief	undertake	attention	degraded	obscene

UPPSALA UNIVERSITET

Word Embeddings (Projected)



(courtesy of Ali Basirat)

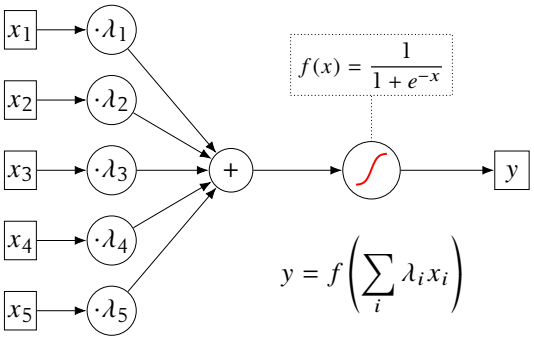
UPPSALA UNIVERSITET

Neural networks

- Neural networks are the machine learning paradigm in which most of this happens.
- Biologically inspired, but doesn't matter very much.
- Very popular in the early 1980's, but the time wasn't ripe.
- The elementary "neuron" is just a nonlinear with some trainable parameters.
- Neurons are combined into a network by function composition.

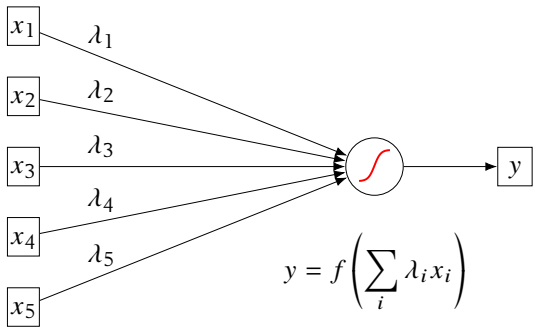
UPPSALA UNIVERSITET

Logistic Regression



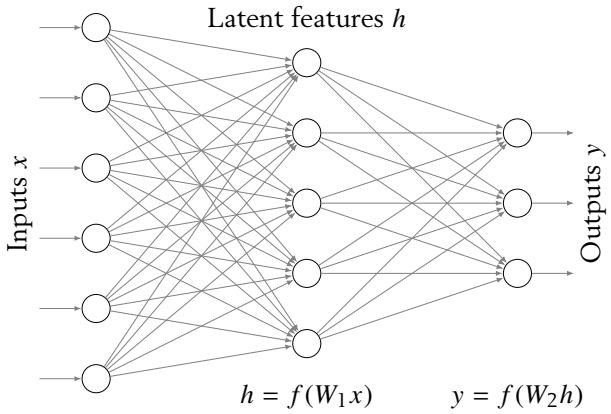
UPPSALA UNIVERSITET

Logistic Regression



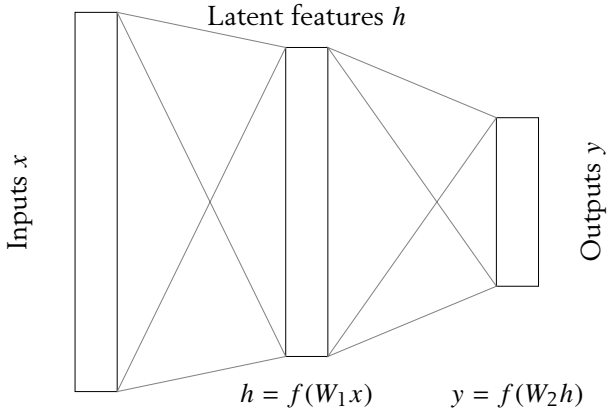
UPPSALA UNIVERSITET

Multiple Decision Steps



UPPSALA UNIVERSITET

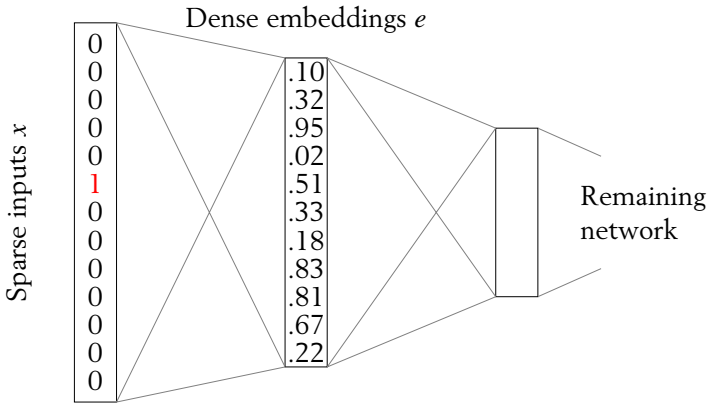
Multiple Decision Steps



Training the Network

- Neural networks are trained by numerically minimising the error of the output for a training set.
- The algorithms used are variants of *gradient descent*.
- The gradients with respect to all weights can be computed efficiently with a dynamic programming algorithm called *back-propagation*.

Word Embeddings in Neural Networks



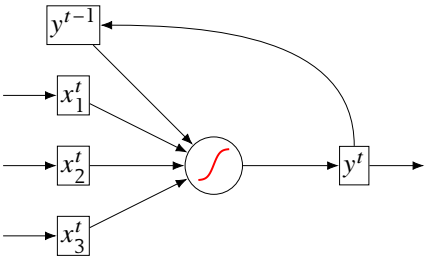
Sequence Length Limits

- A given network takes a fixed number of inputs.
- In MT, we need to process input sentences of arbitrary length and produce output of arbitrary length.
- Input and output length are not necessarily the same.

Input length	Output length	Compression	Network type
fixed	fixed		feed-forward
variable	= input (or fixed)		recurrent
variable	unconstrained	to fixed size	encoder-decoder
variable	unconstrained	no compression	attention-based

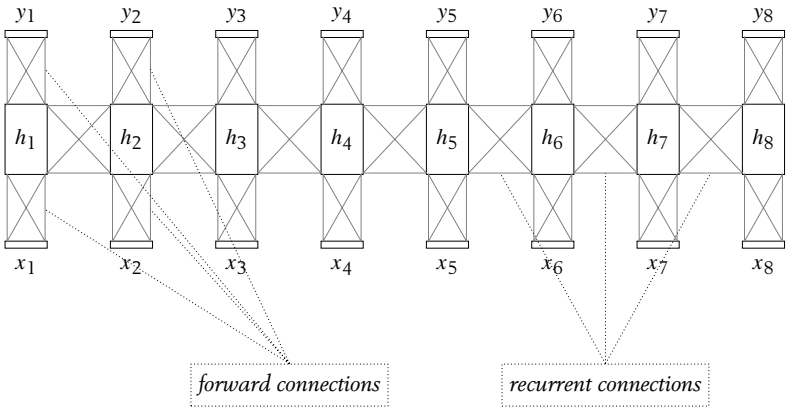
UPPSALA UNIVERSITET

Adding a Time Dimension: Recurrent Nets



UPPSALA UNIVERSITET

Processing Sequences



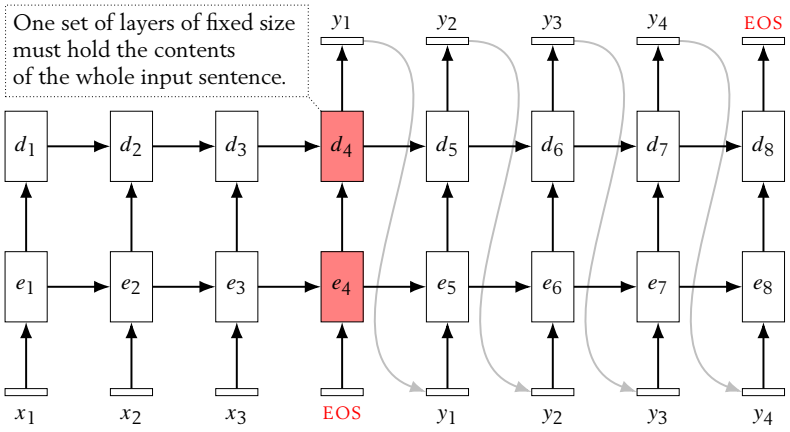
UPPSALA UNIVERSITET

Unequal Sequence Length

- In this architecture, there are equally many inputs x_i as outputs y_i.
- Useful for sequence labelling tasks such as POS tagging.
- In machine translation, the length of the input and output sequences differ.

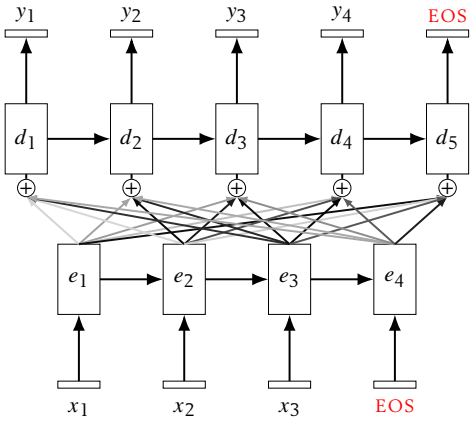
UPPSALA UNIVERSITET

Encoder-Decoder Architecture



UPPSALA UNIVERSITET

Attention Mechanism



UPPSALA UNIVERSITET

Neural MT: Summary

- Very new area: First large-scale systems in 2014.
- Promising results in public evaluations.
- We know little about its strengths and weaknesses yet, but they seem to be very different from earlier approaches.
- I'll tell you more in a few years. . .