Sequence Modeling: Recurrent and Recursive Networks

Markus Dumke

27th January 2016

Contents

Why RNN's?

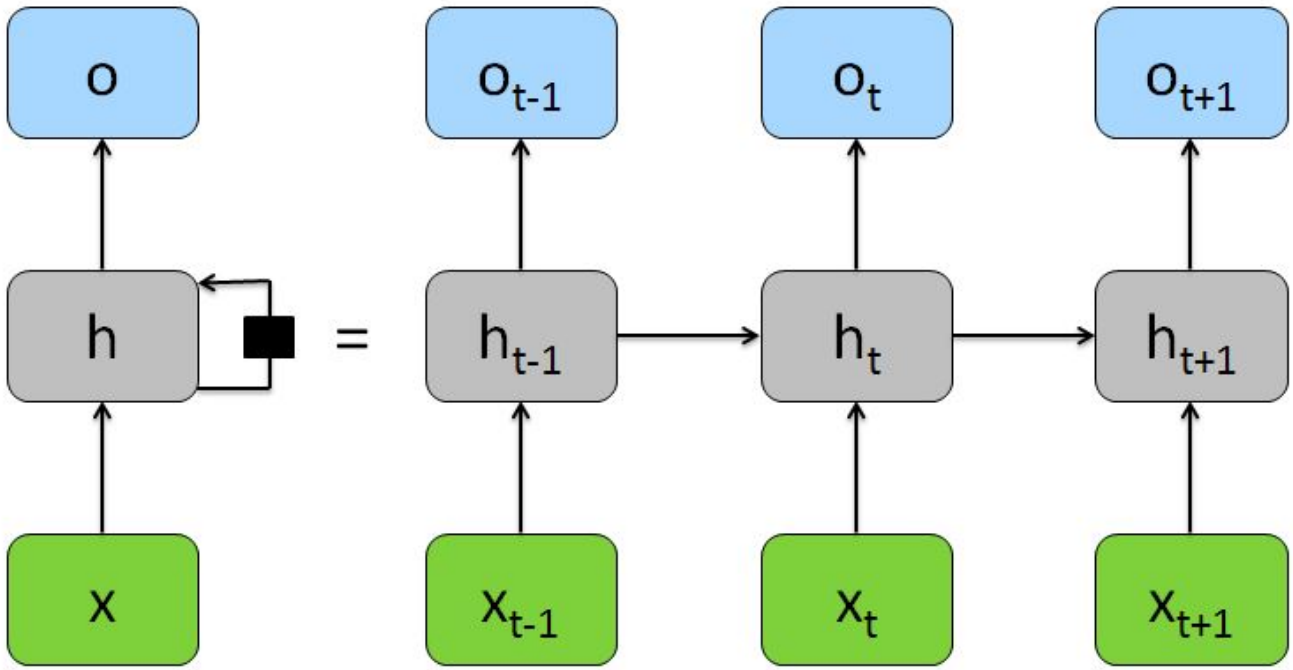
- sequential data
- outputs depend on all previous inputs (no independence)
- long-term dependencies
- memory

Applications

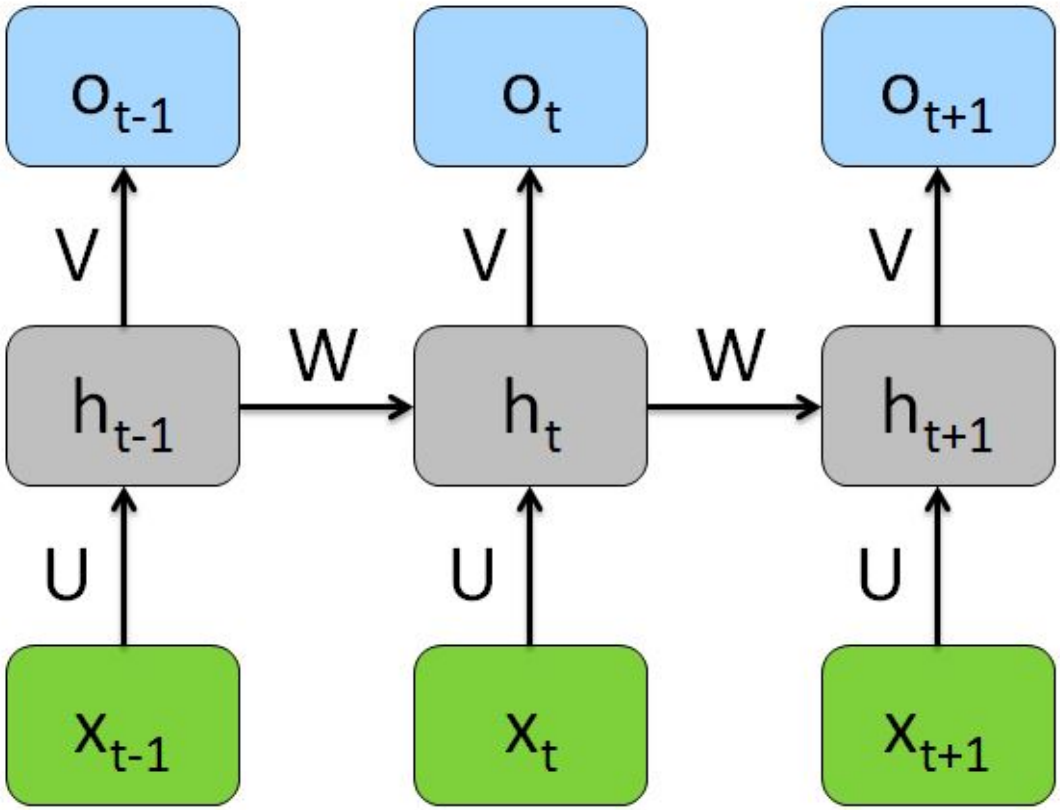
Natural Language Processing

- machine translation
- character- or word-level language model
- text summary or labels
- sentiment analysis
- image captioning
- handwriting recognition and generation
- speech recognition and generation
- time series
- ...

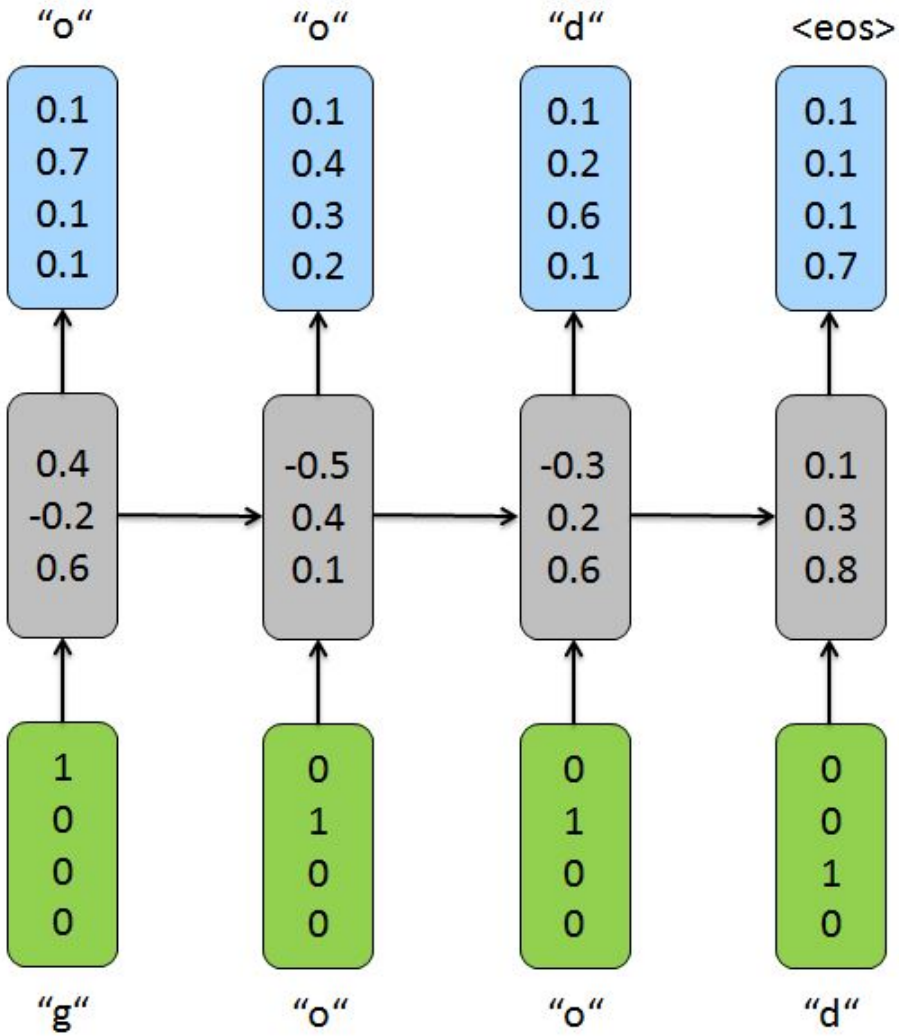
Recurrent Neural Network



Recurrent Neural Network



Recurrent Neural Network



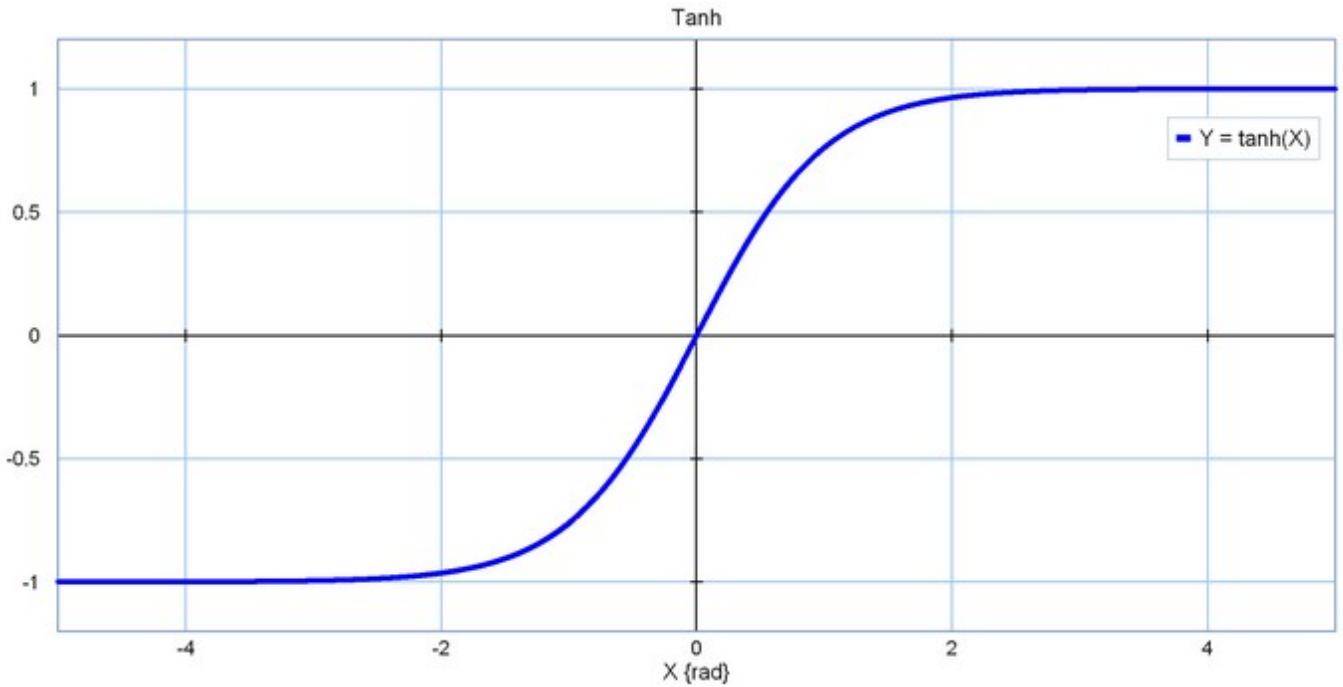
Recurrent Neural Network

for t = 1 to τ:

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)}$$
$$h^{(t)} = f(a^{(t)})$$
$$o^{(t)} = c + Vh^{(t)}$$
$$\hat{y}^{(t)} = softmax(o^{(t)})$$

Which activation function?

$$f(x) = tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

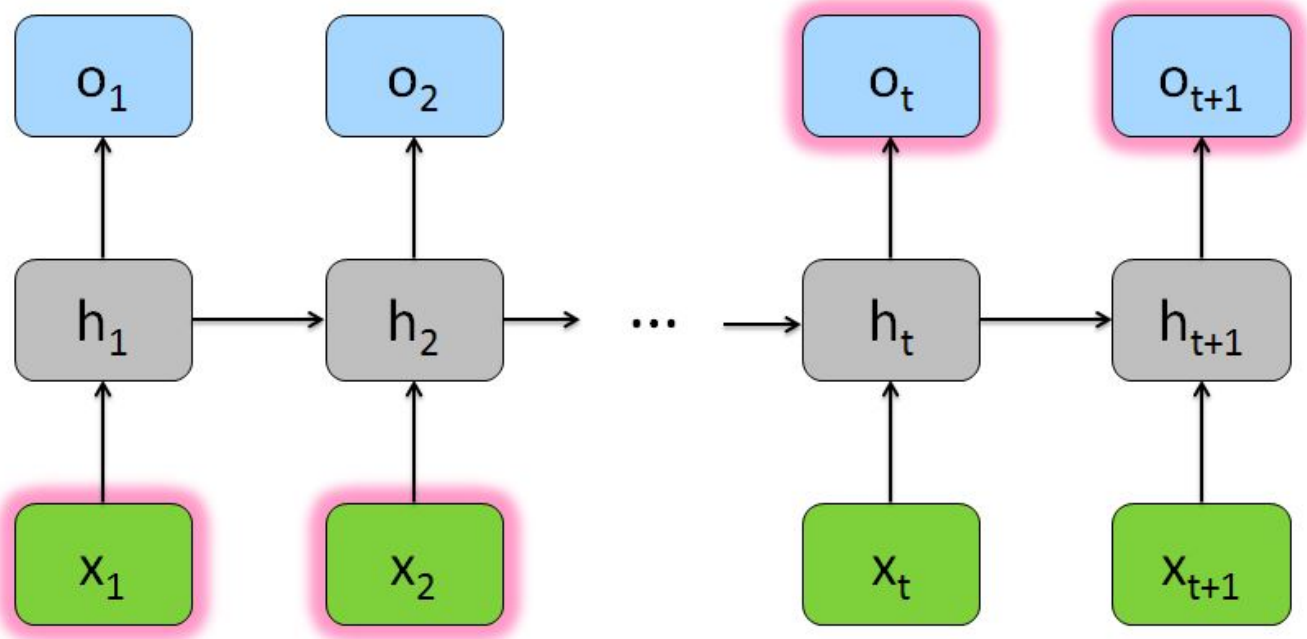


http://www.20sim.com/webhelp/language_reference_functions_tanh.php

Optimization

- Forward Propagation, compute loss
- Backward Propagation through time (BPTT), compute gradients
- Stochastic Gradient Descent (Minibatch)

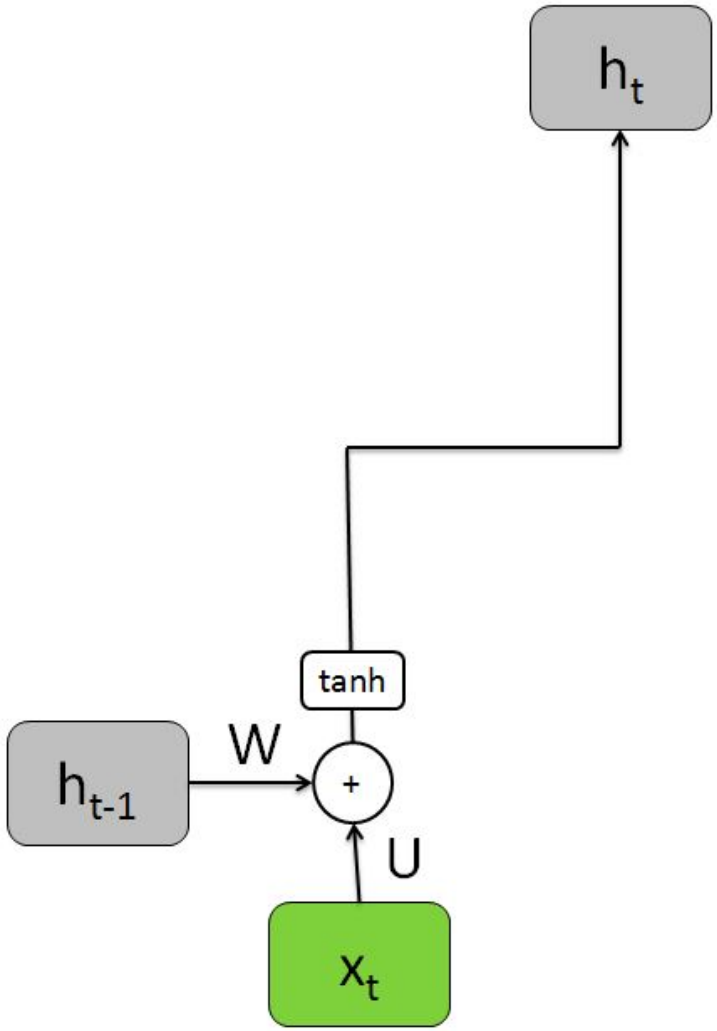
Vanishing (and Exploding) Gradient Problem



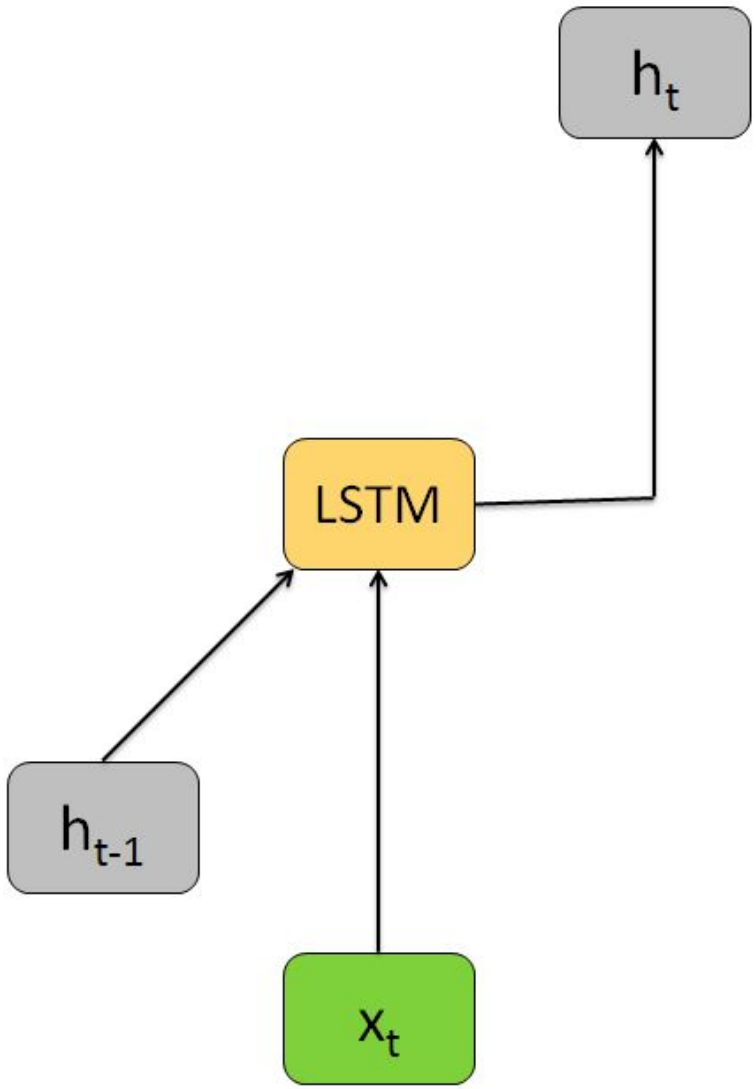
How to deal with vanishing gradients?

- Gradient Clipping
- Regularization
- Leaky Units
- different time scales
- **LSTM**, GRU and variants

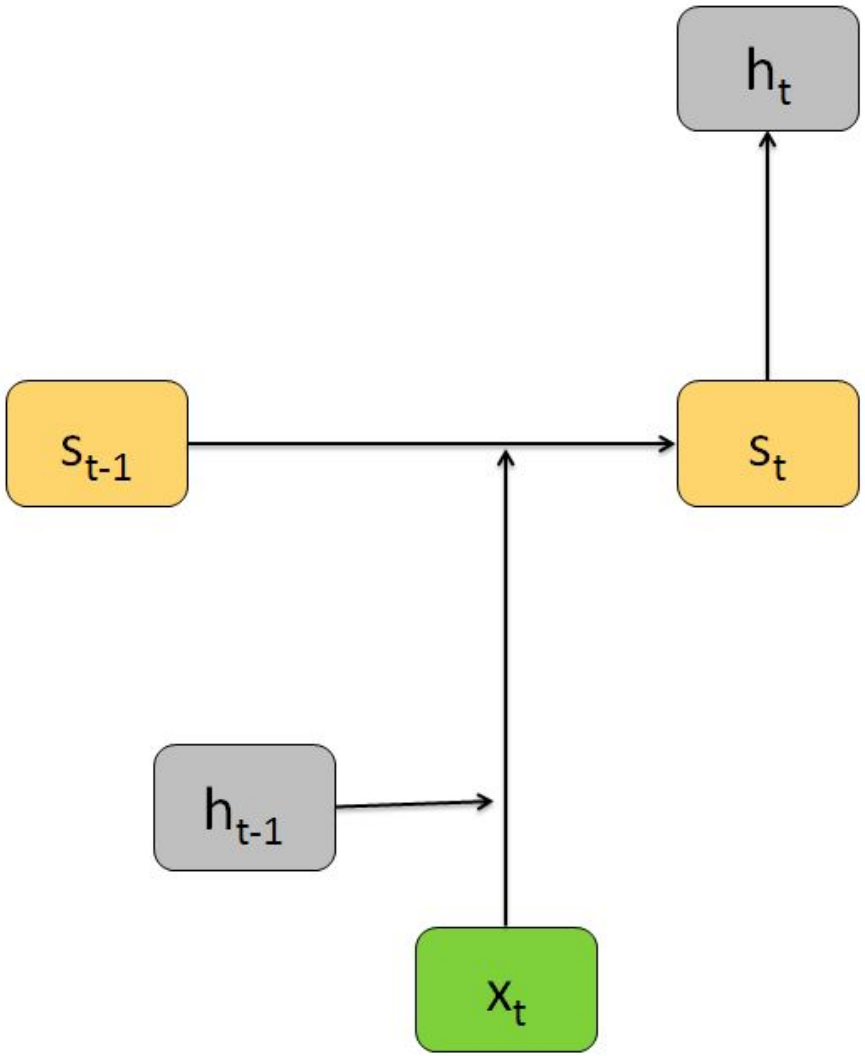
LSTM



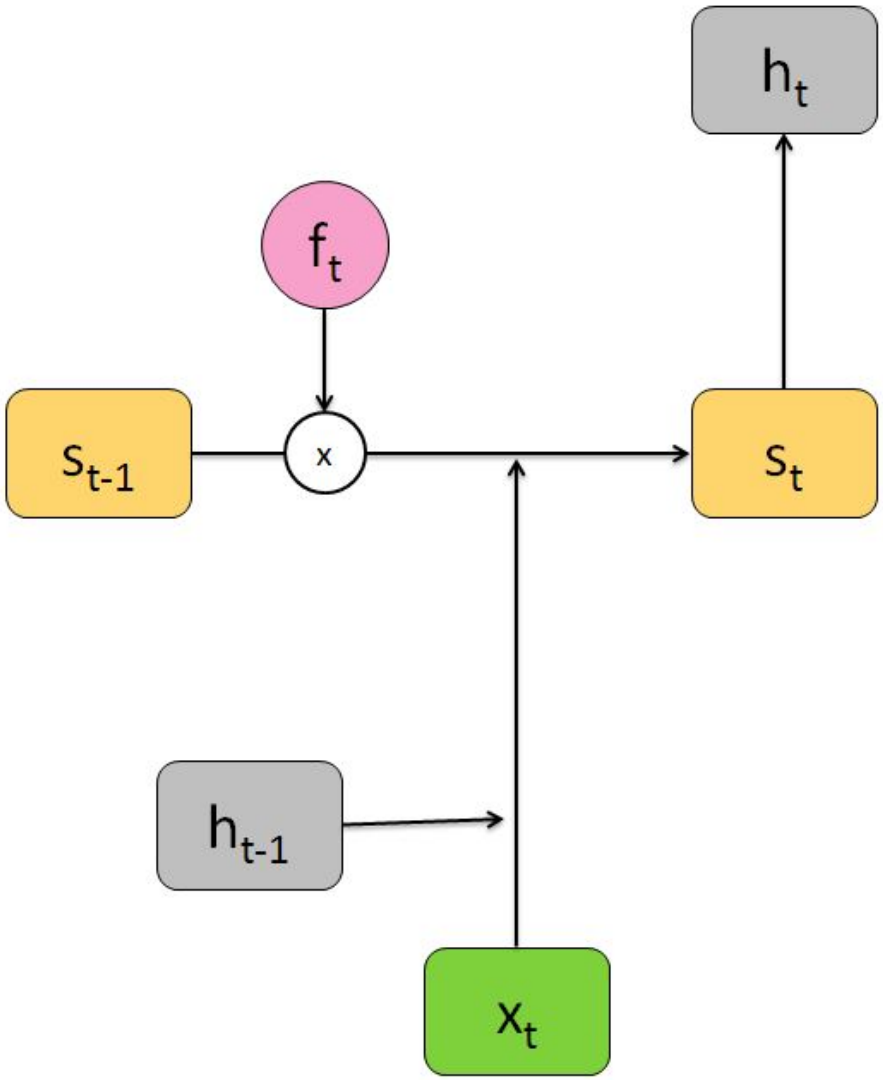
LSTM



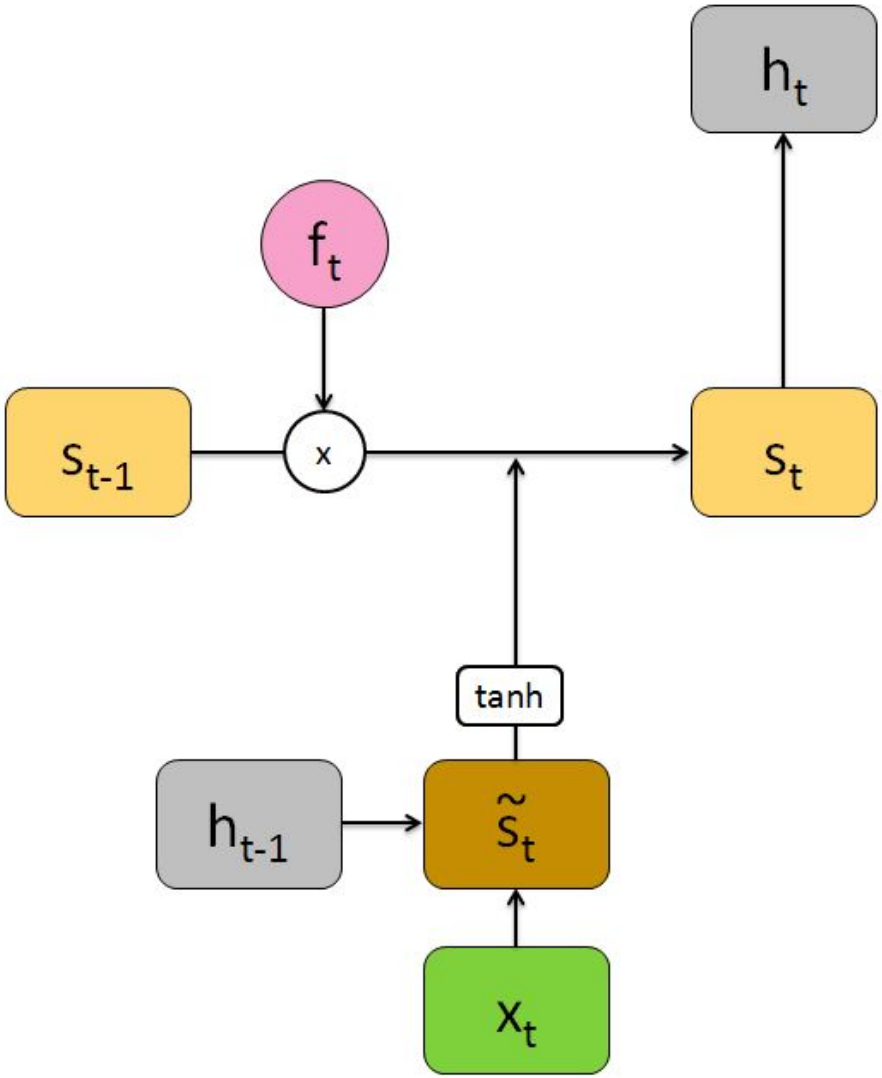
LSTM



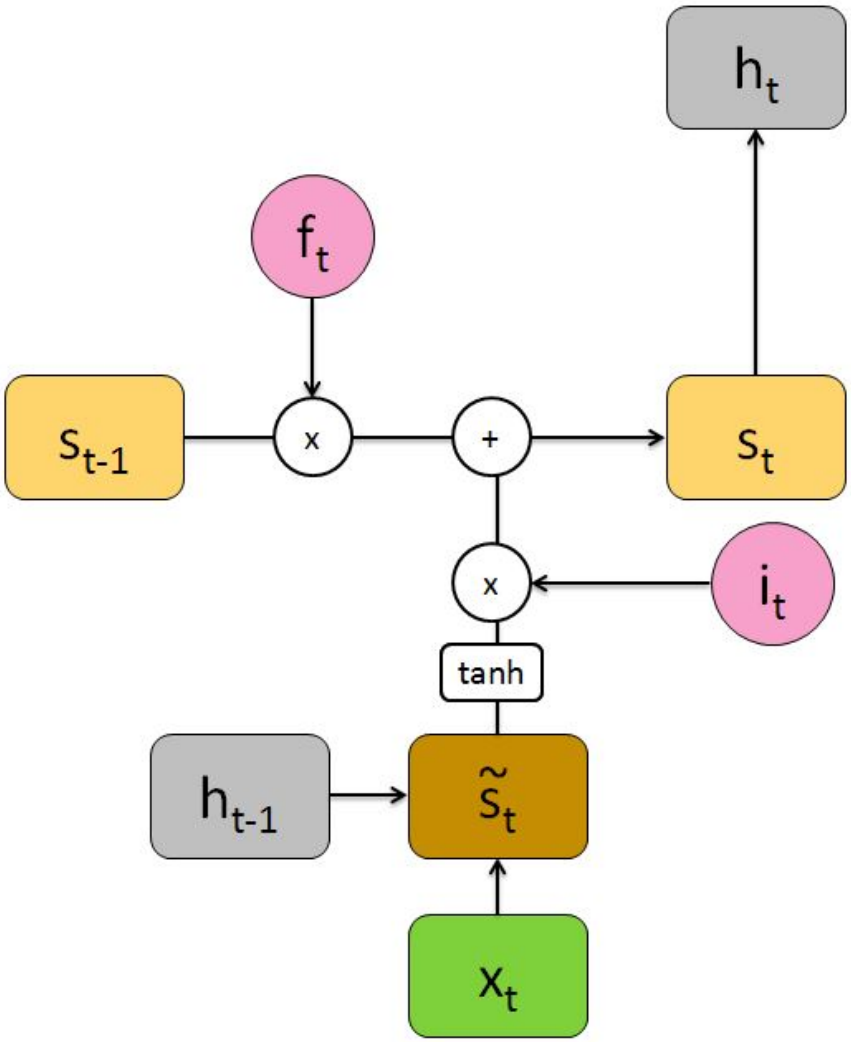
LSTM



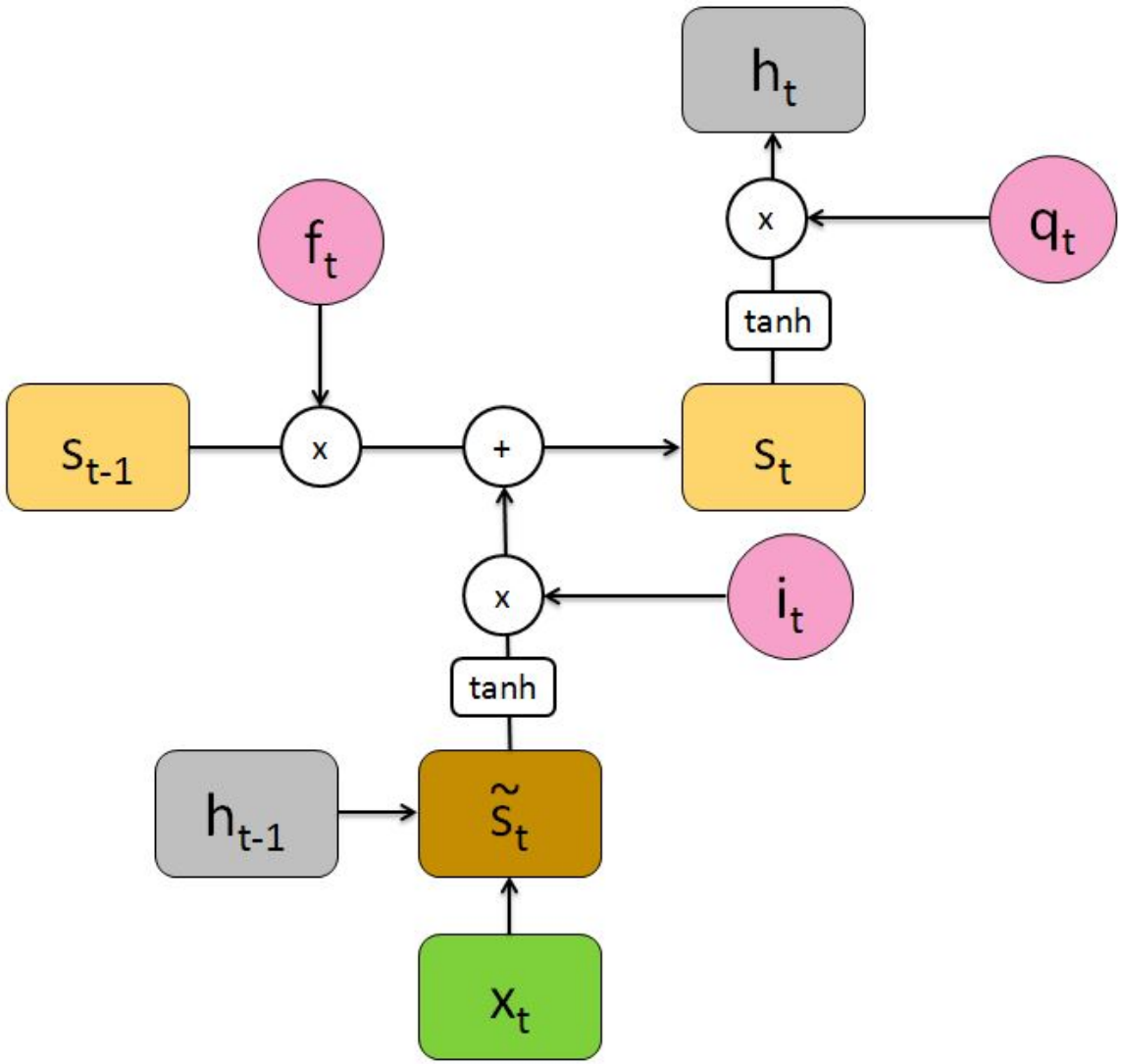
LSTM



LSTM



LSTM



Sampling from an RNN

- sample from conditional distribution at each time step
- how to generate sequence length?
- special end symbol
- Bernoulli random variable
- integer value τ

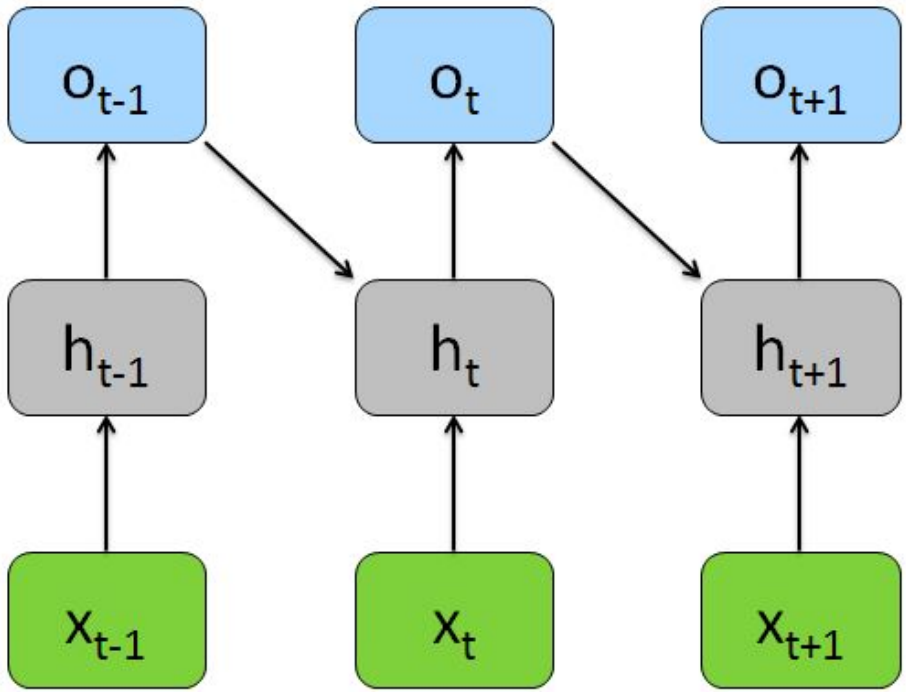
Language Modeling

- Output: Probability distribution over words given previous words

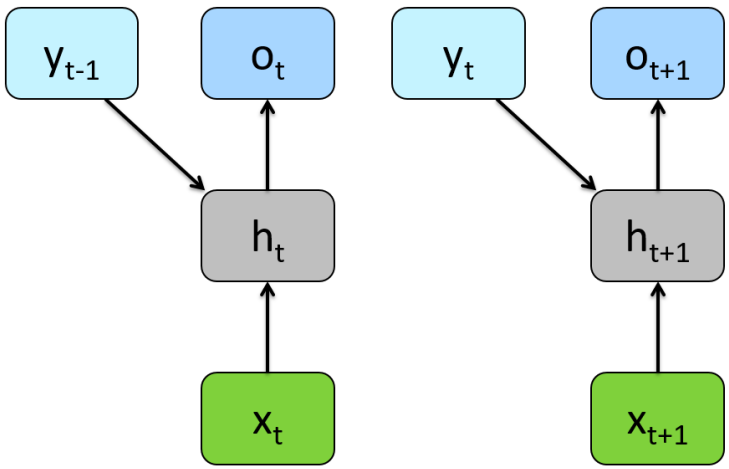
$$\mathrm{P}(\mathrm{y}_1, ..., y_T) = \prod_{i=1}^{T} P(y_i | y_1, ..., y_{i-1})$$

- scoring candidates
- word-level or character-level possible
- Input: word/character encoded as one-hot vector

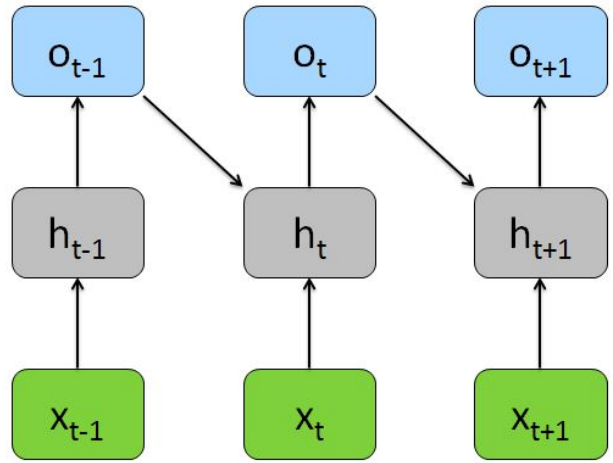
RNN with output recurrence



Teacher Forcing

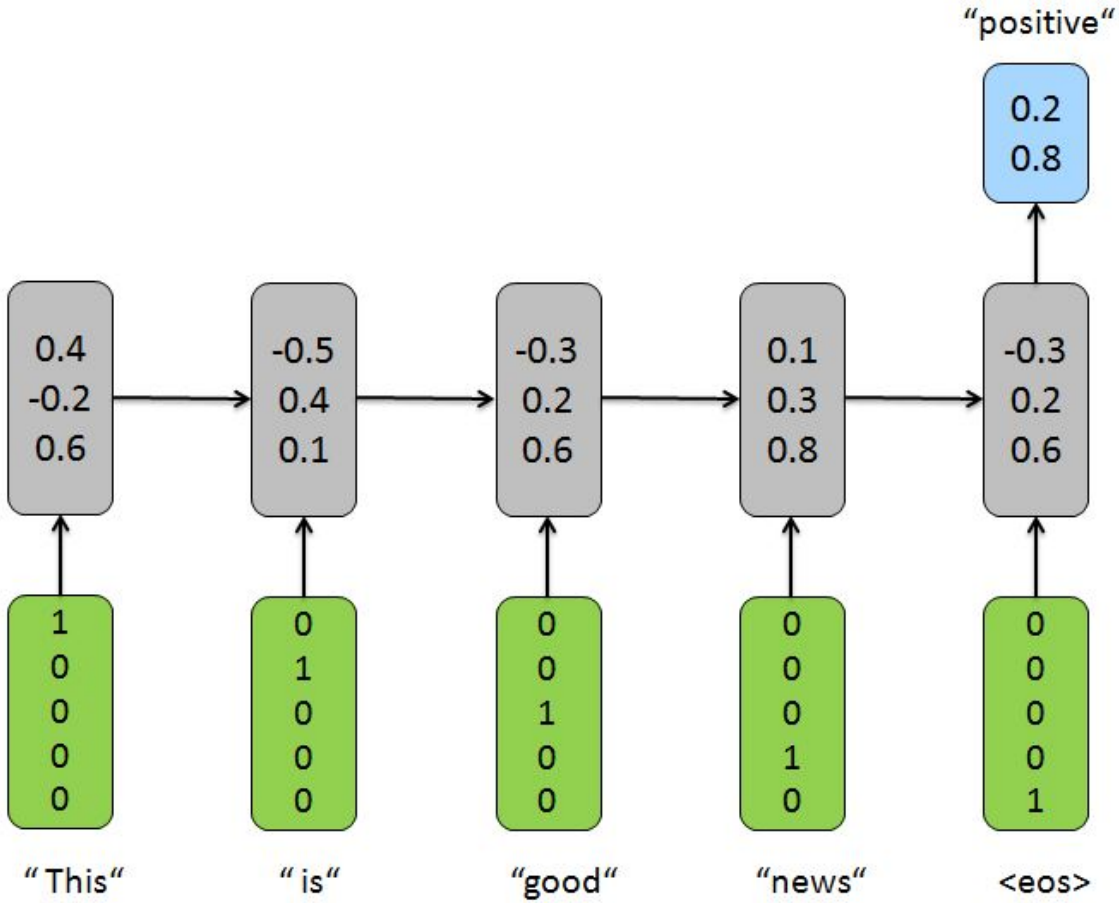


(a) At train time

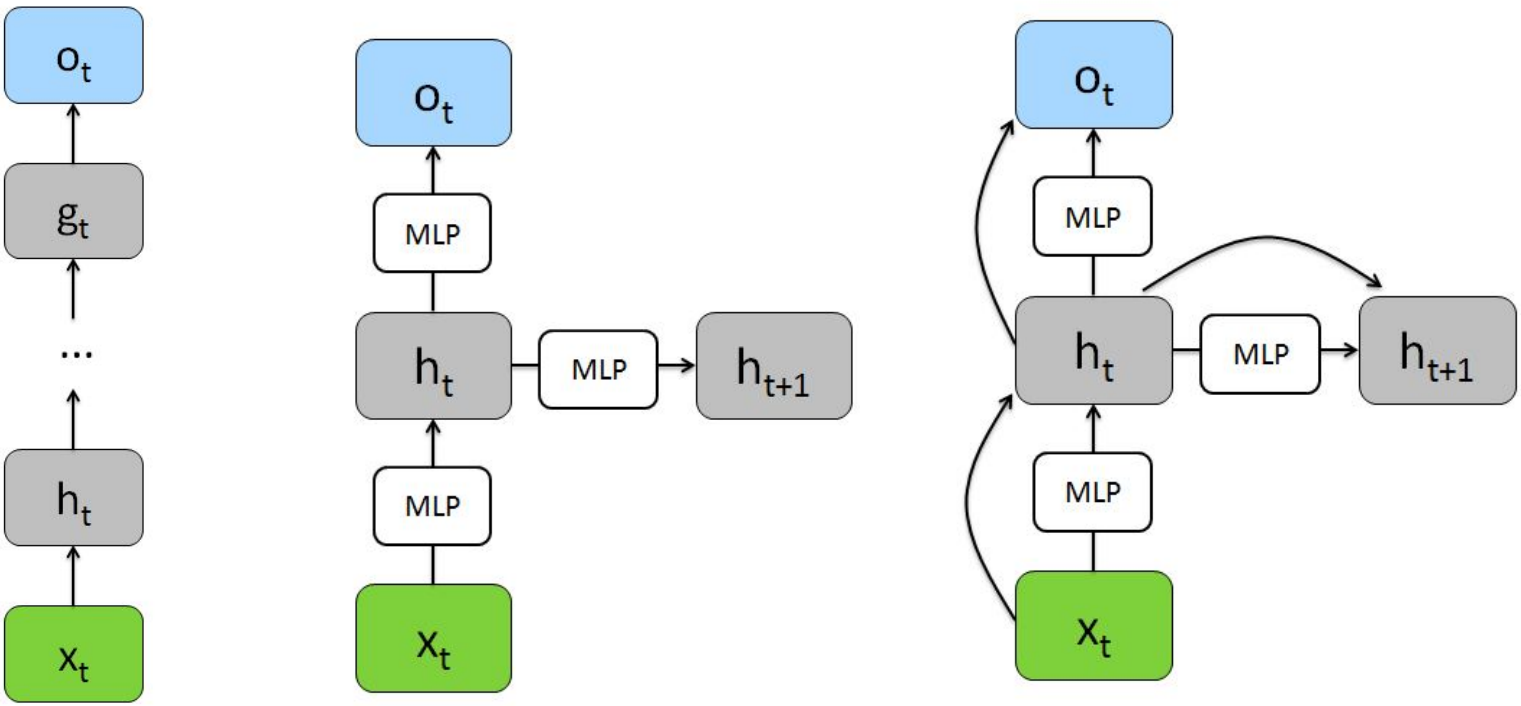


(b) At test time

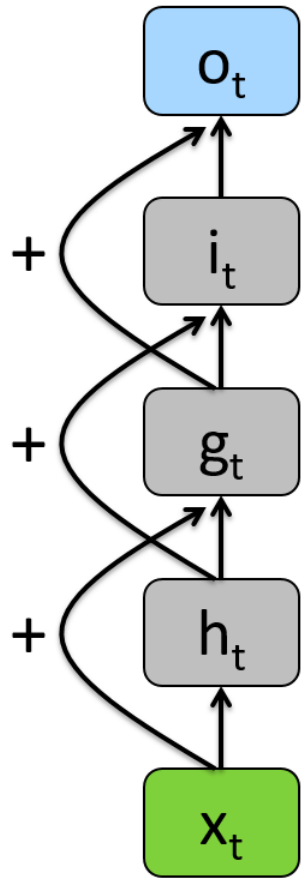
One-output RNN



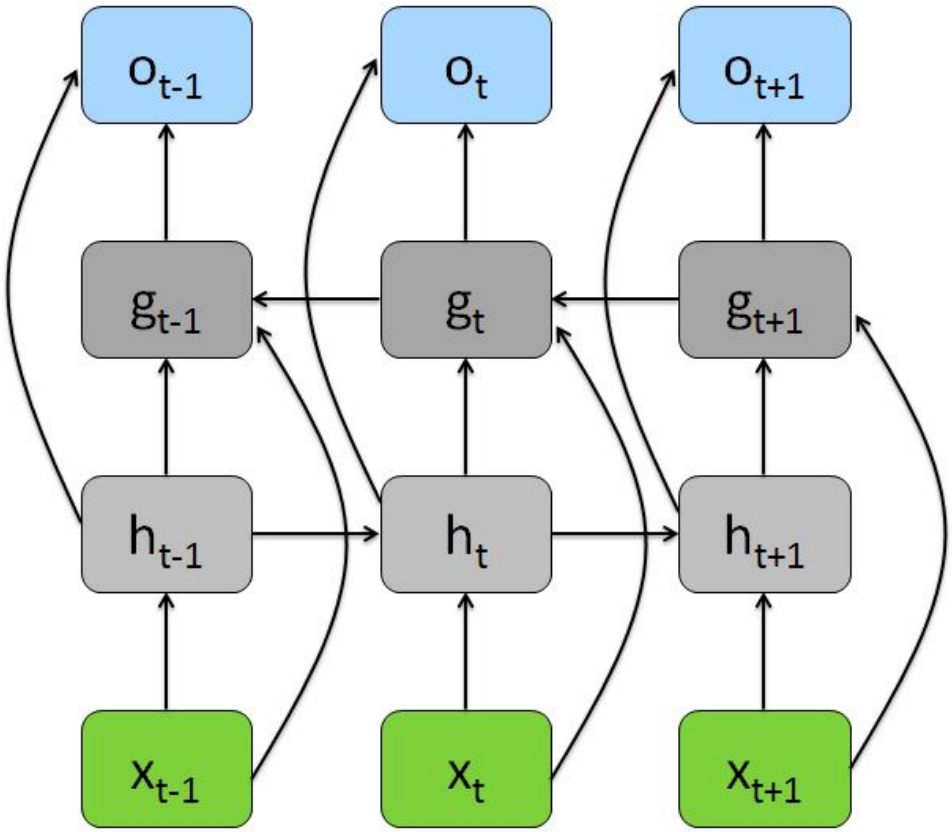
Deep RNNs



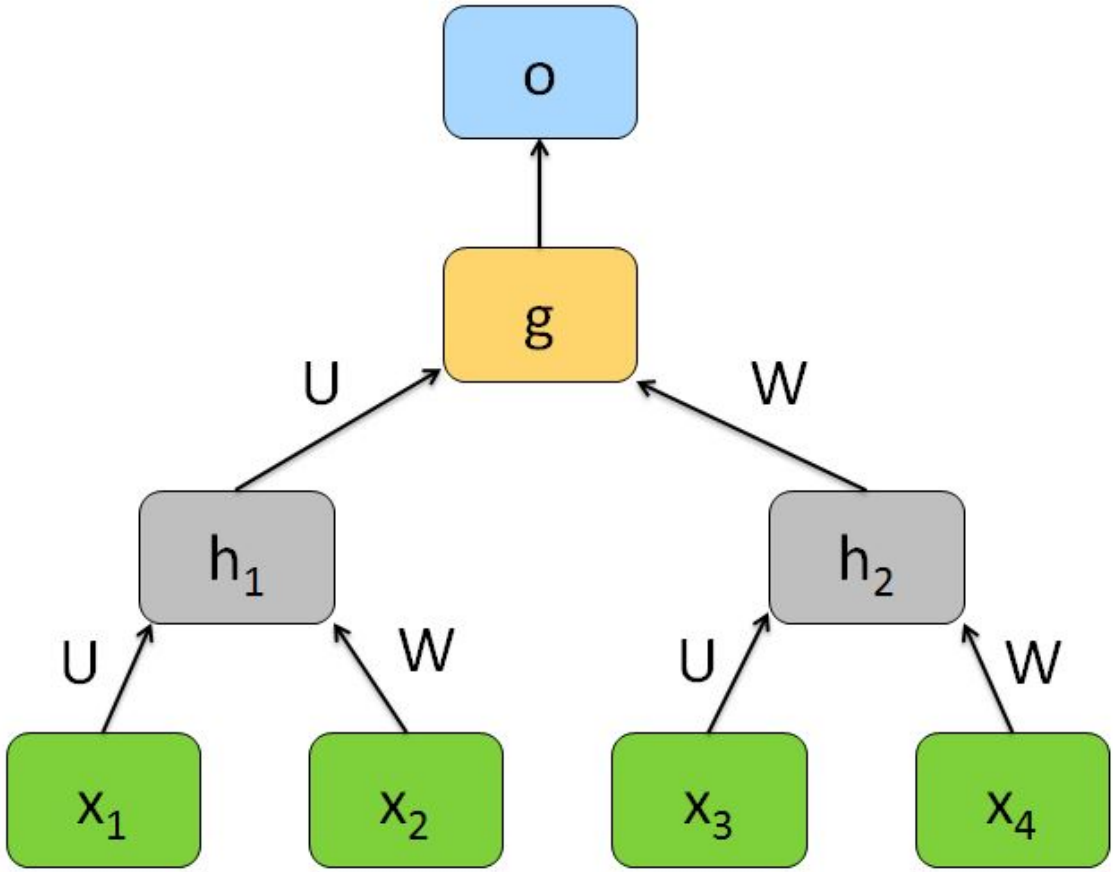
Res-Net



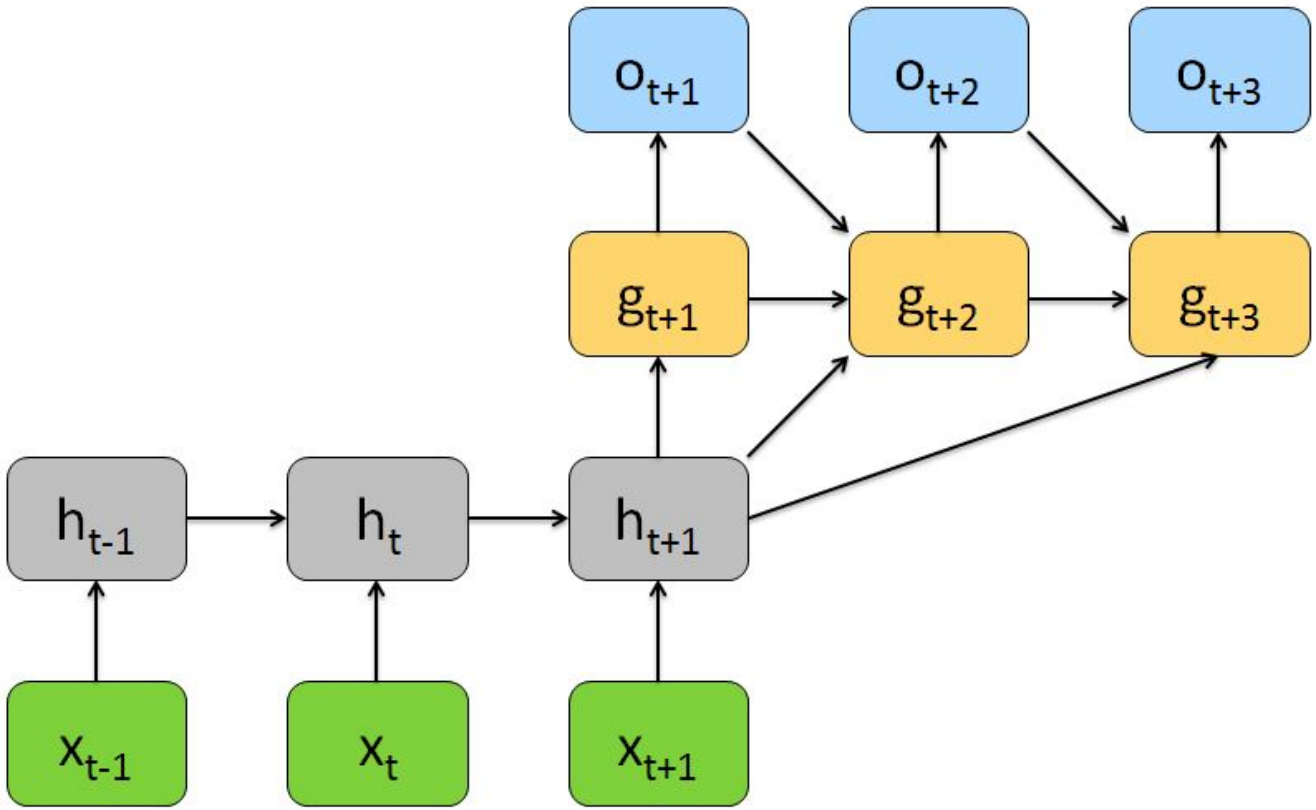
Bidirectional RNN



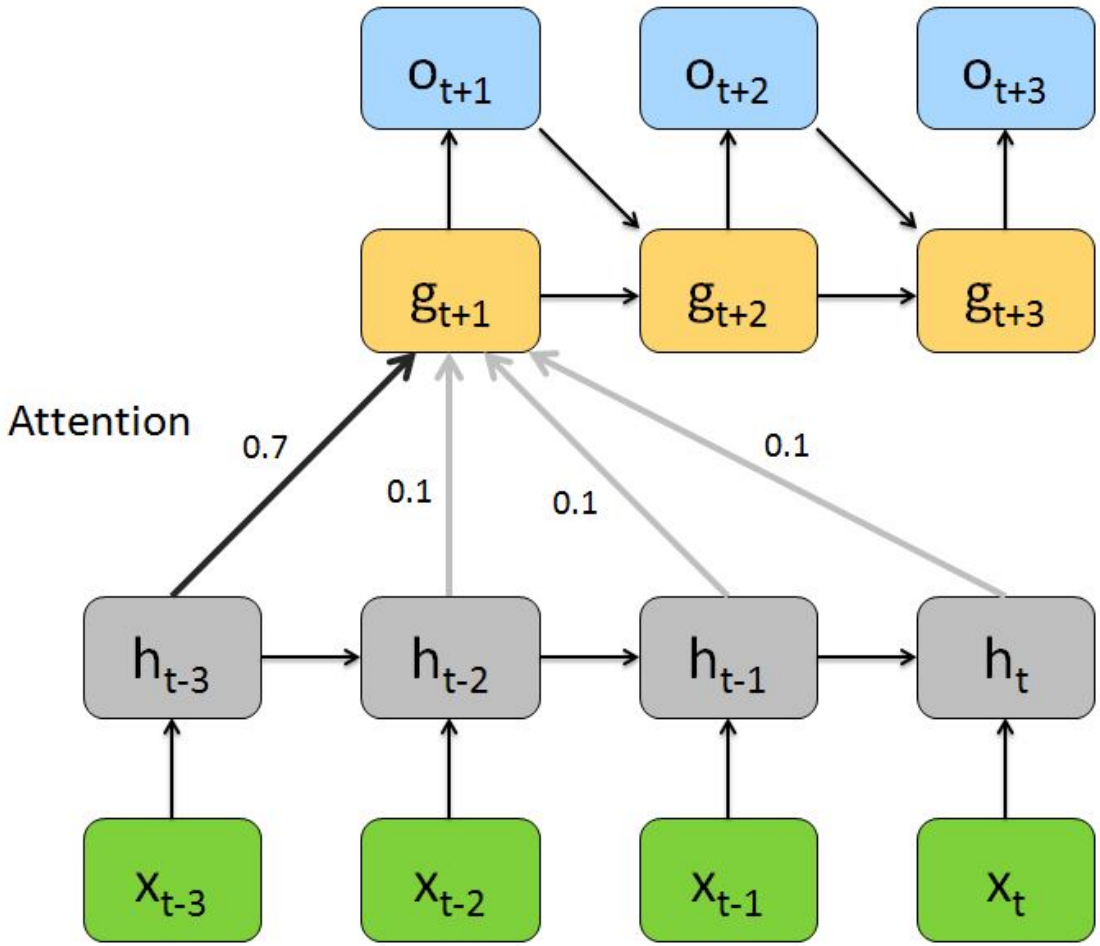
Recursive Neural Network



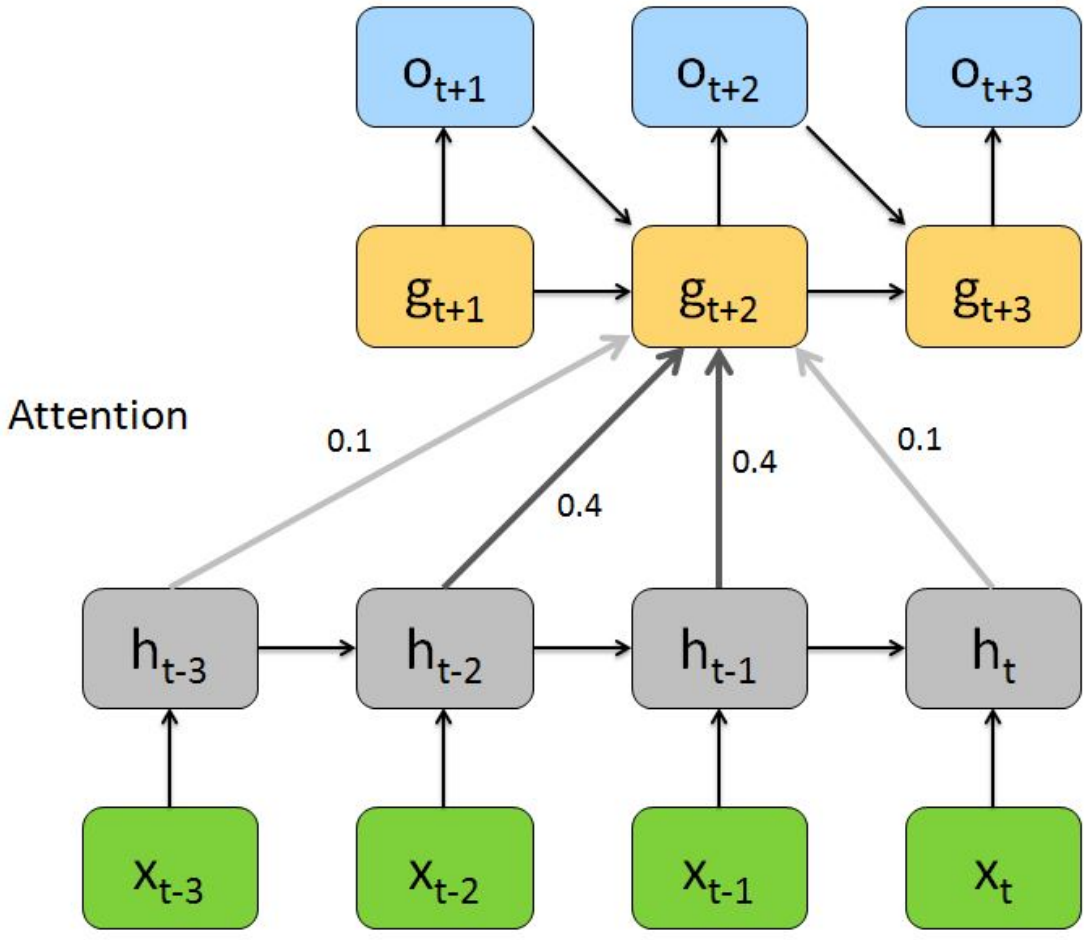
Encoder-Decoder Architecture



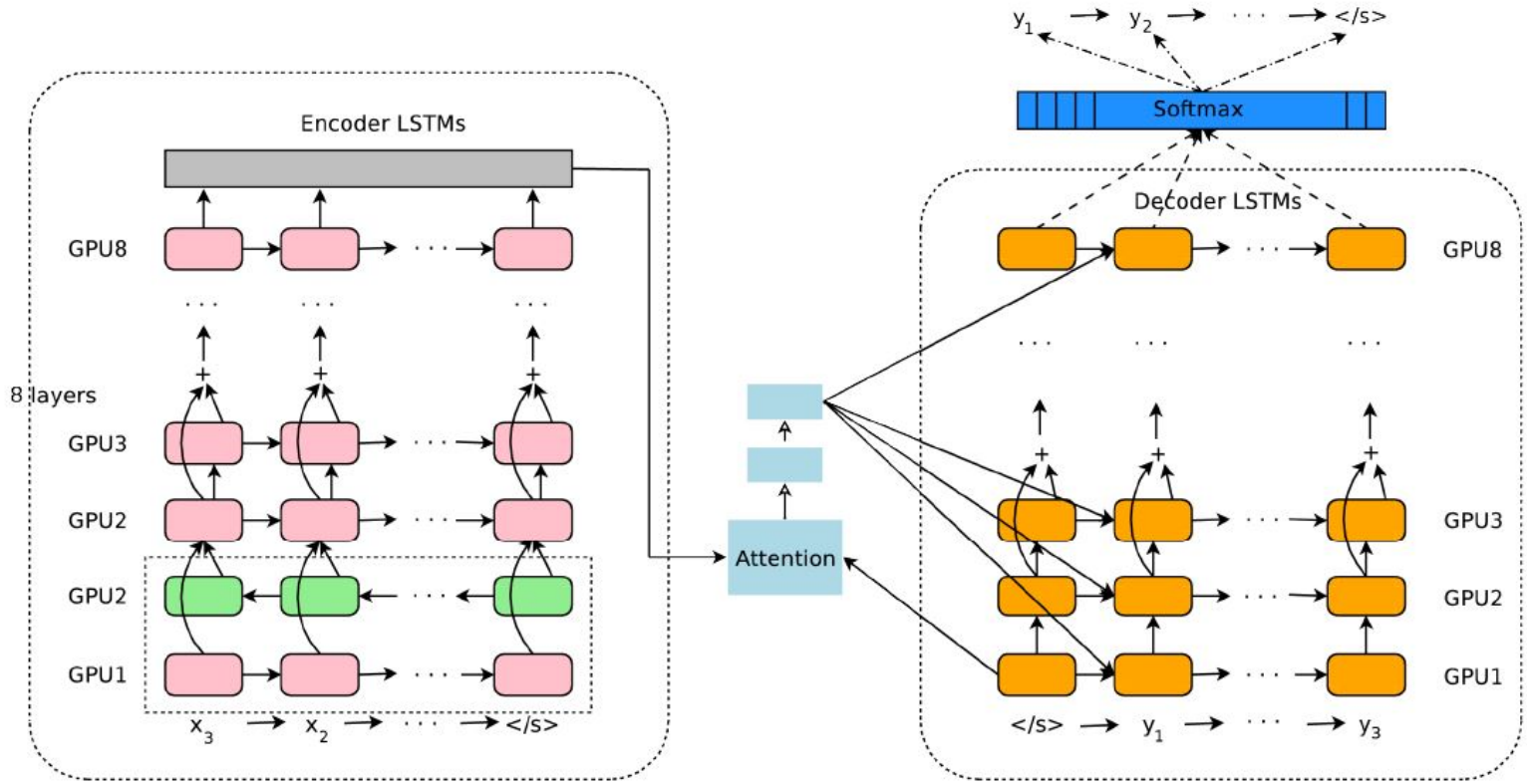
Attention


o_{t+1}
o_{t+2}
o_{t+3}
g_{t+1}
g_{t+2}
g_{t+3}
Attention
0.7
0.1
0.1
0.1
h_{t-3}
h_{t-2}
h_{t-1}
h_t
x_{t-3}
x_{t-2}
x_{t-1}
x_t

Attention

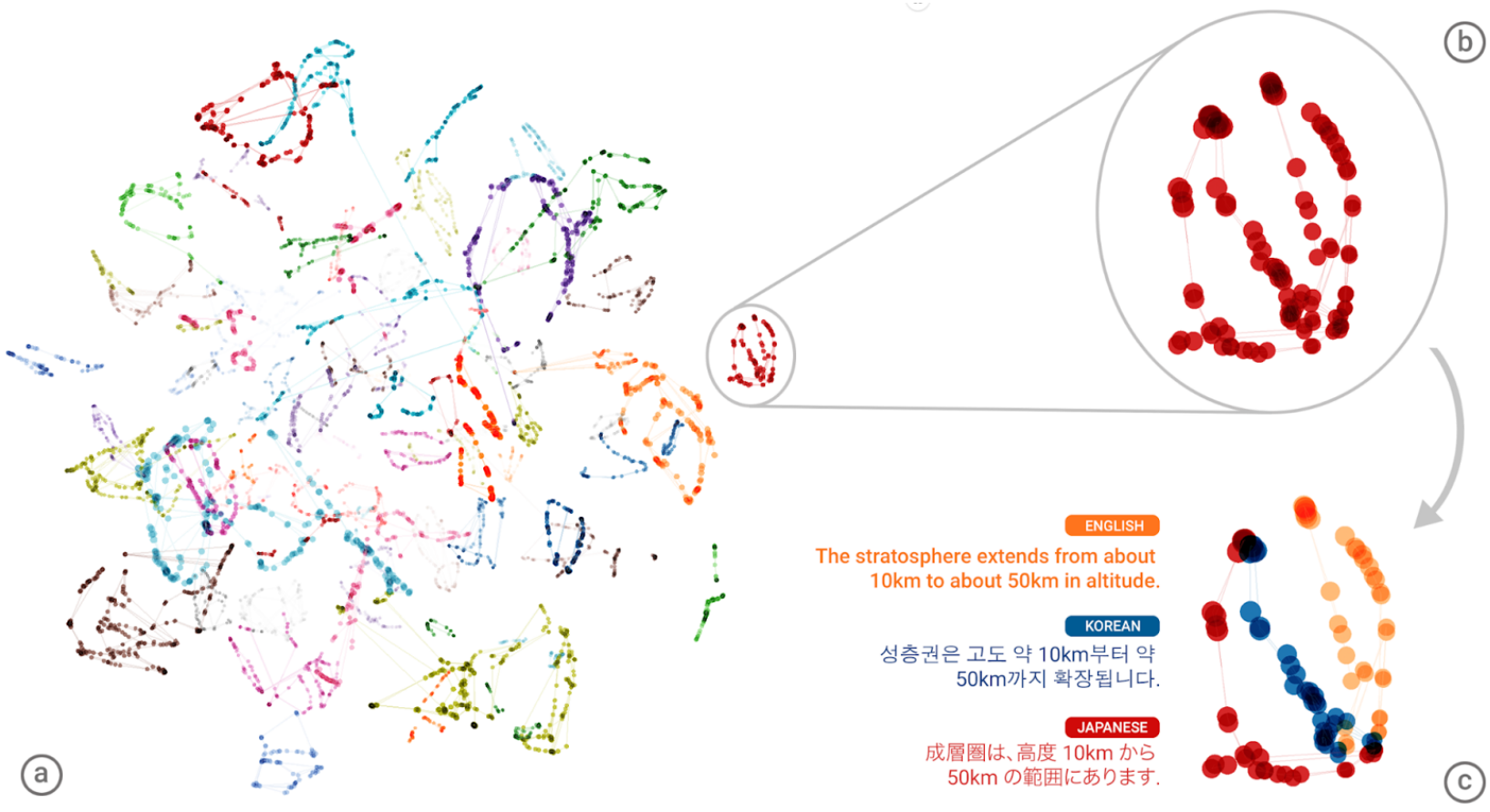

o_{t+1}
o_{t+2}
o_{t+3}
g_{t+1}
g_{t+2}
g_{t+3}
Attention
0.1
0.4
0.4
0.1
h_{t-3}
h_{t-2}
h_{t-1}
h_t
x_{t-3}
x_{t-2}
x_{t-1}
x_t

Google's Neural Machine Translation System



?

Language Embeddings



?

Bibliography