Discriminative Language Models

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

January 26, 2017

Based on slides from Noah Smith, Richard Socher, and everyone else they copied from.

Language Models

$$P(\text{"I love food"}) = P(\text{"I"} \mid \text{<s>})\, P(\text{"love"} \mid \text{"<s> I"})\, P(\text{"food"} \mid \text{"<s> I love"})$$

Probability of a Sentence

- Is a given sentence something you would expect to see?
- Syntactically (grammar) and Semantically (meaning)

$$P(\text{"food"} \mid \text{"I love"})$$

Probability of the Next Word

- Predict what comes next for a given sequence of words.
- Think of it as V-way classification

Outline

Discriminative Language Models

Feed-forward Neural Networks

Recurrent Neural Networks

Upcoming..

Outline

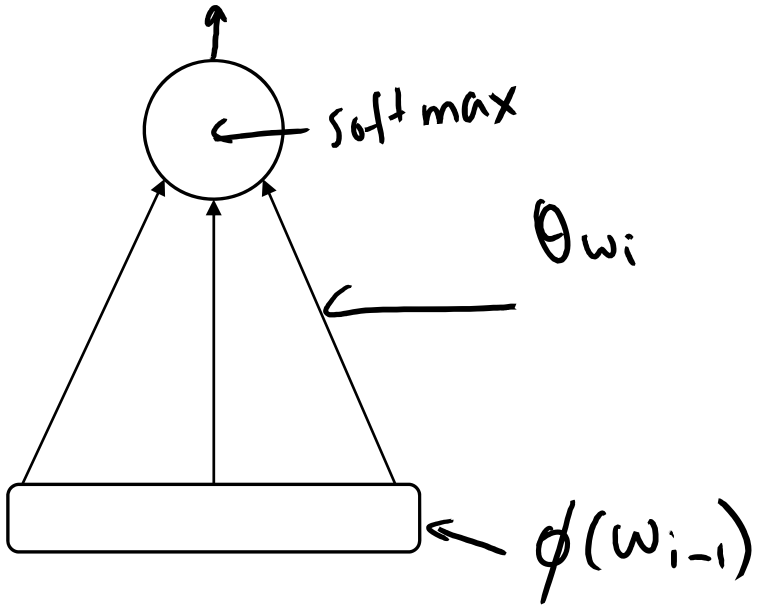
- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks
- Upcoming..

Logistic Regression Model

$p(w_i|w_{i-1})$

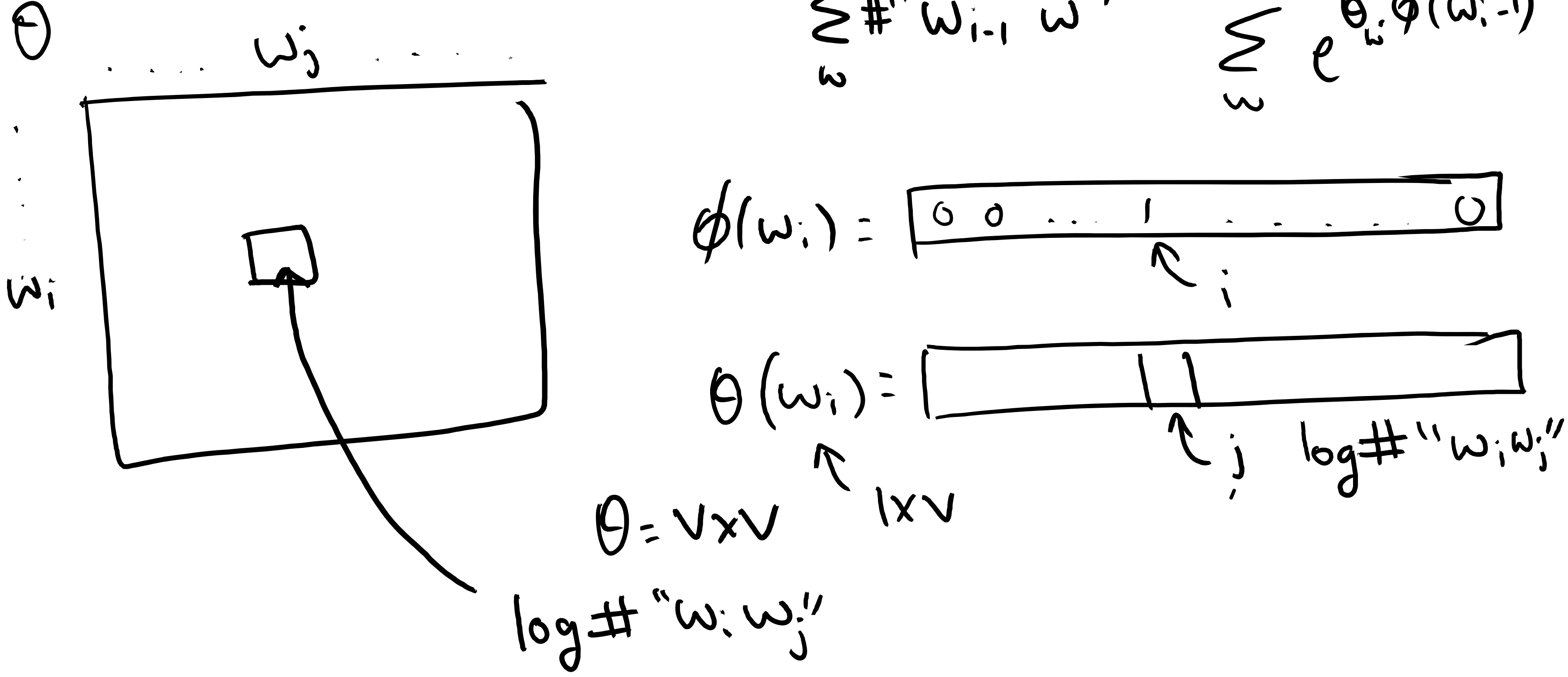
$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-1})$$

$$= \frac{e^{\theta_{w_i} \cdot \phi(w_{i-1})}}{\sum_w e^{\theta_w \cdot \phi(w_{i-1})}}$$

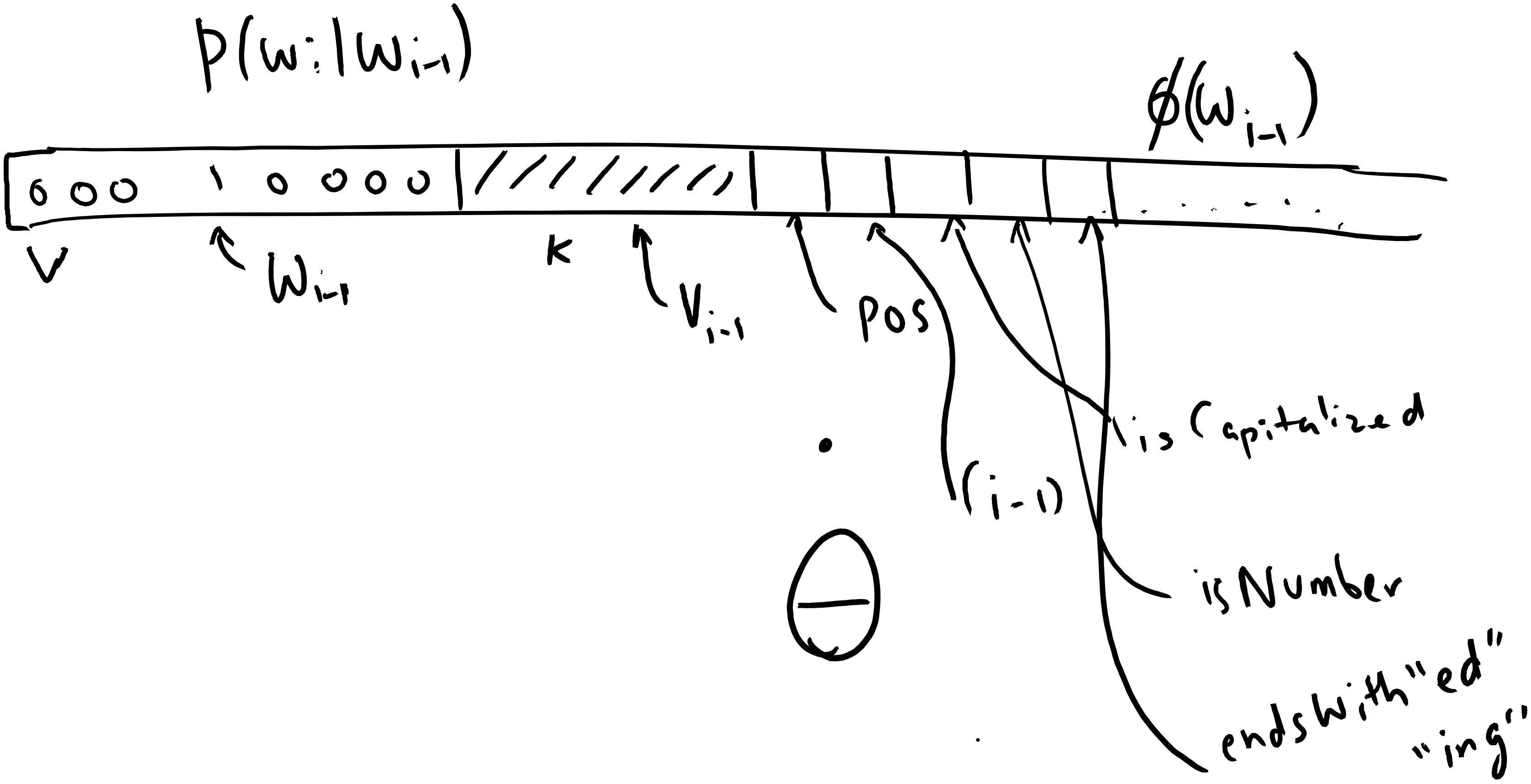


N-Grams as Logistic Reg.

$$p(w_i | w_{i-1}) = \frac{\#\text{"}w_{i-1}\ w_i\text{"}}{\sum_w \#\text{"}w_{i-1}\ w\text{"}} \approx \frac{e^{\theta_{w_i} \cdot \phi(w_{i-1})}}{\sum_w e^{\theta_w \cdot \phi(w_{i-1})}}$$



Other features...



Outline

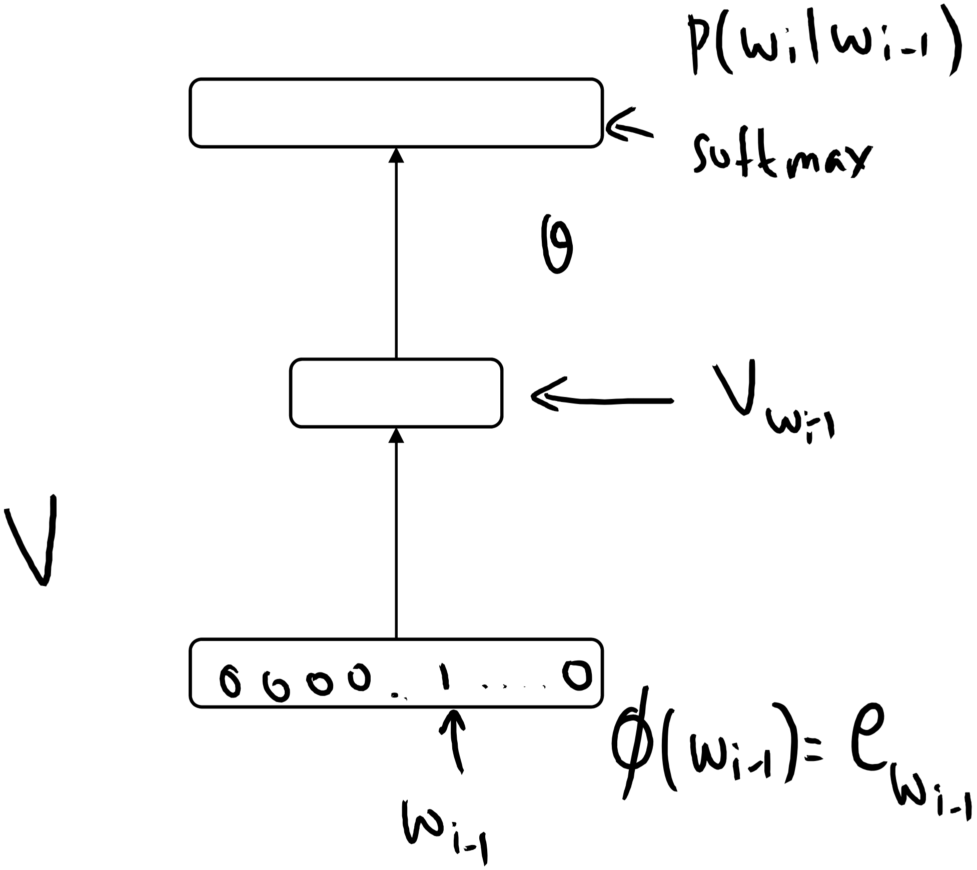
Discriminative Language Models

Feed-forward Neural Networks

Recurrent Neural Networks

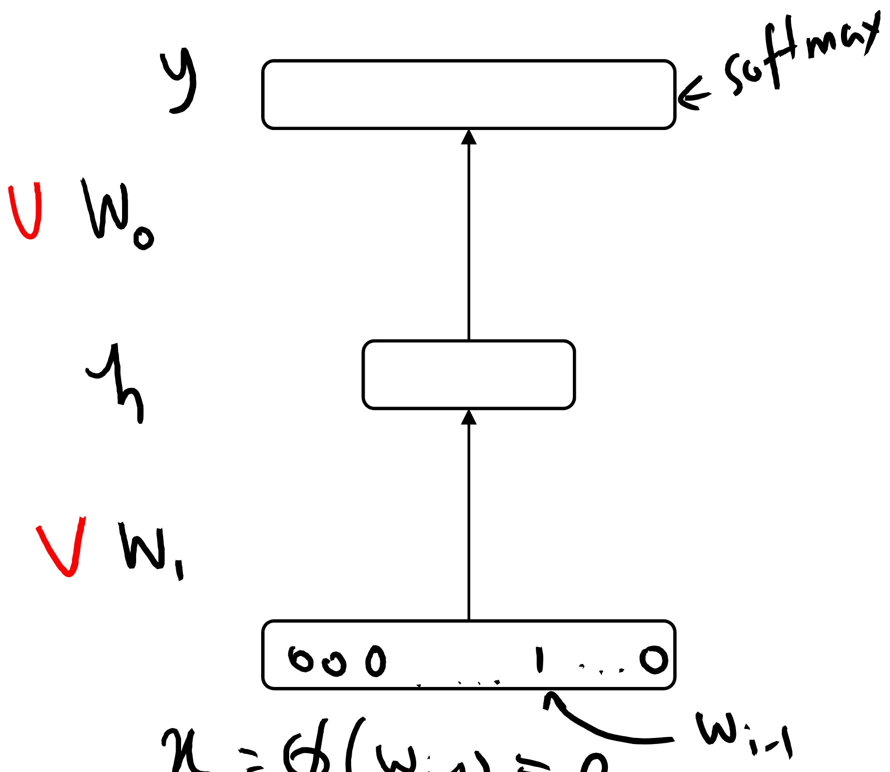
Upcoming..

Logistic Reg. w/ Embeddings



$$y = \text{softmax}(\theta \times V_{w_{i-1}} + b)$$

Neural Networks



$$x = \phi(w_{i-1}) = e_{w_{i-1}}$$

$$y = \mathrm{softmax}(W_o \times h)$$

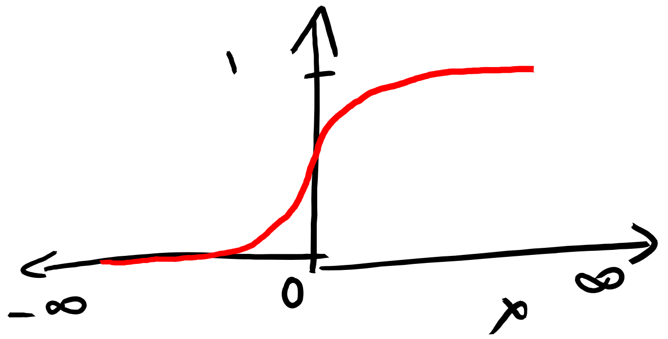
$$h = f(W_1 \times x)$$

↳ sigmoid, σ

Activation Functions, f

sigmoid

$$f(x) = \frac{e^{x}}{1 + e^{x}}$$

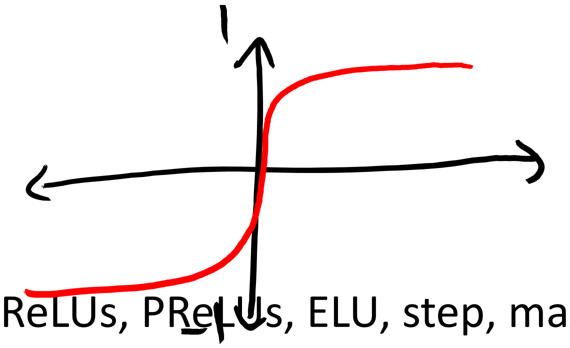


softmax

$$f_i(\vec{x}) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

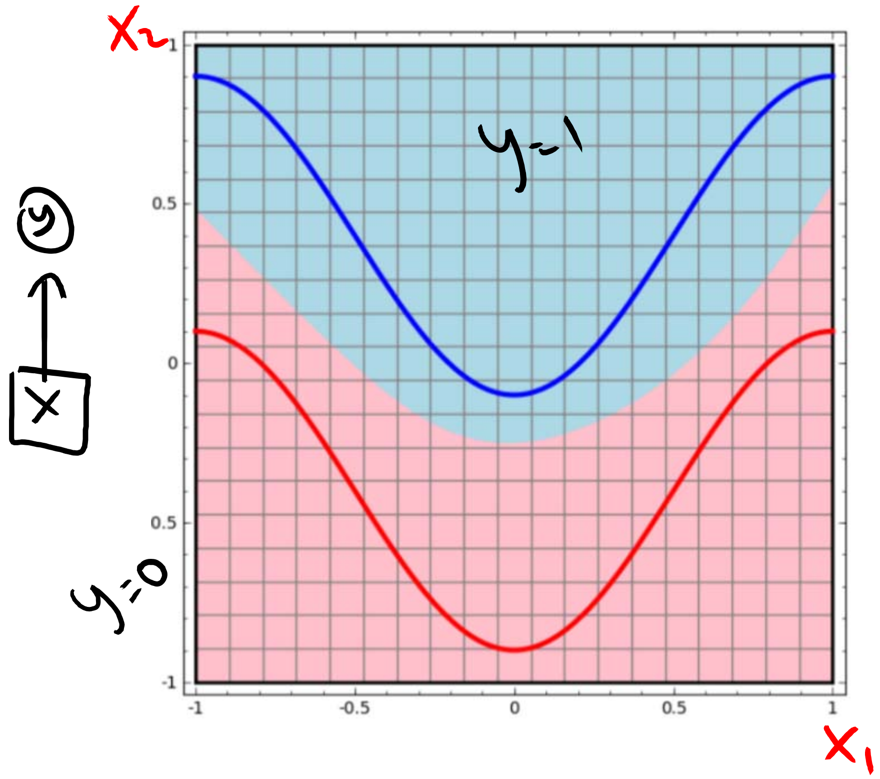
tanh

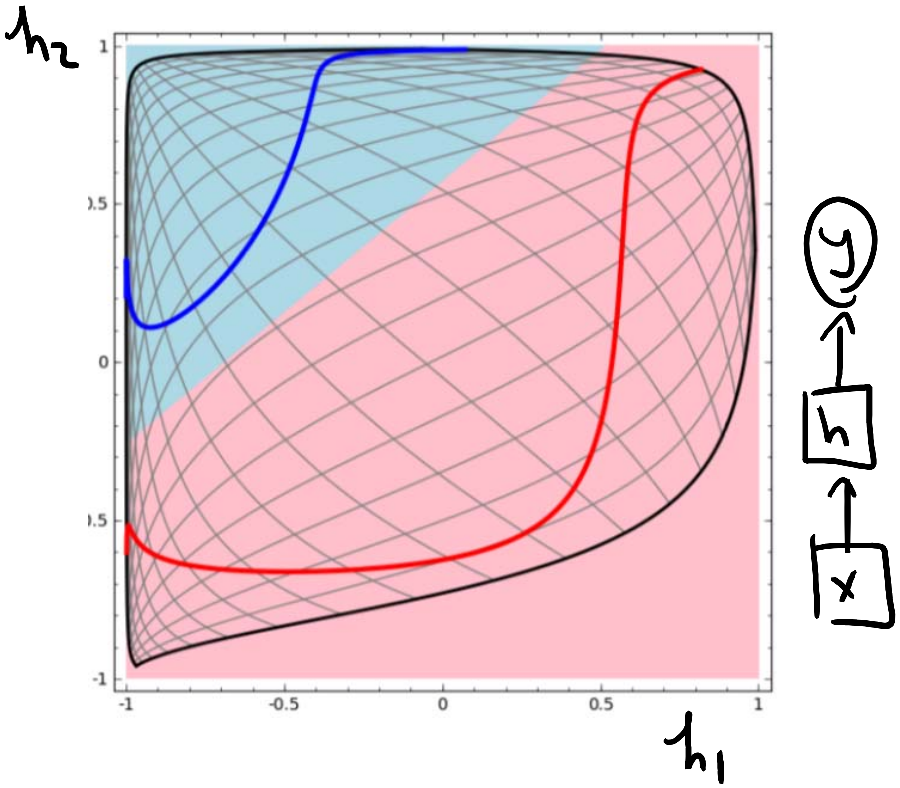
$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{x}}$$



And many others... ReLUs, PReLUs, ELU, step, max, and so on..

Why do they work?





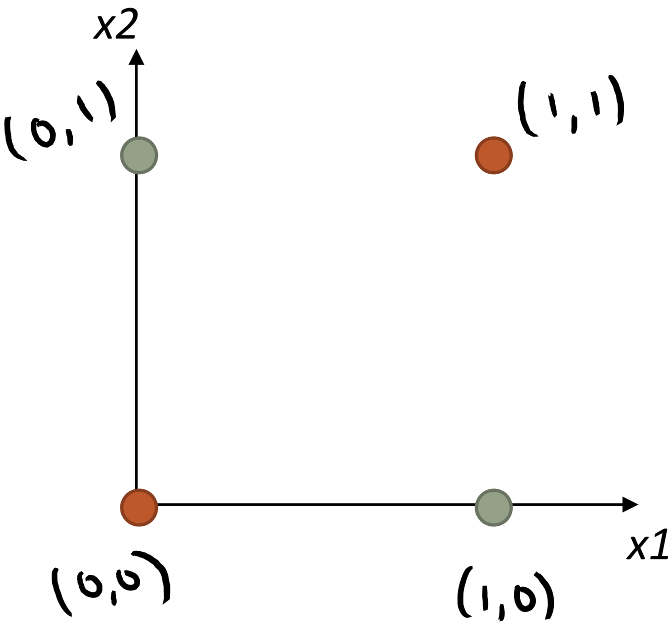
$$h = f(w \times x)$$

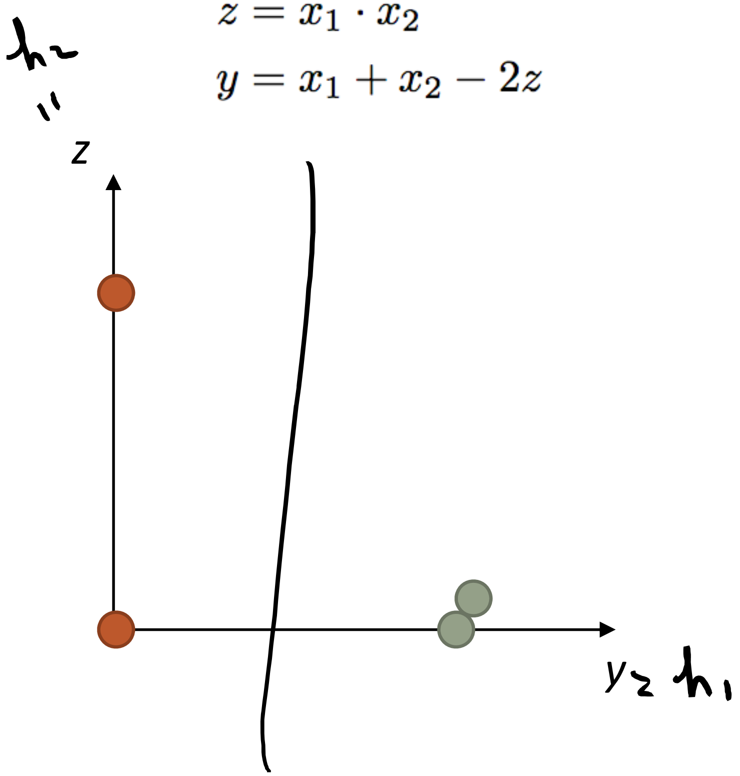
https://colah.github.io

Why do they work?

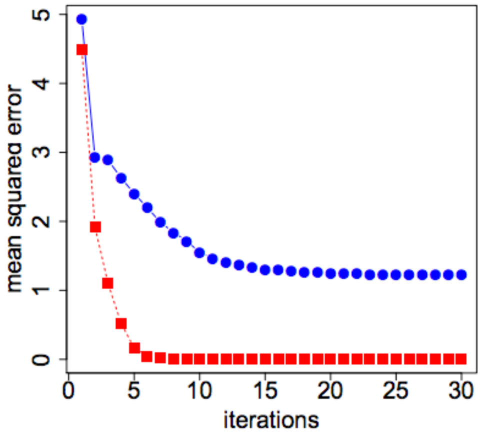
$O = \mathrm{xor}(x_1, x_2)$

$$z = x_1 \cdot x_2$$
$$y = x_1 + x_2 - 2z$$





Simulated Example



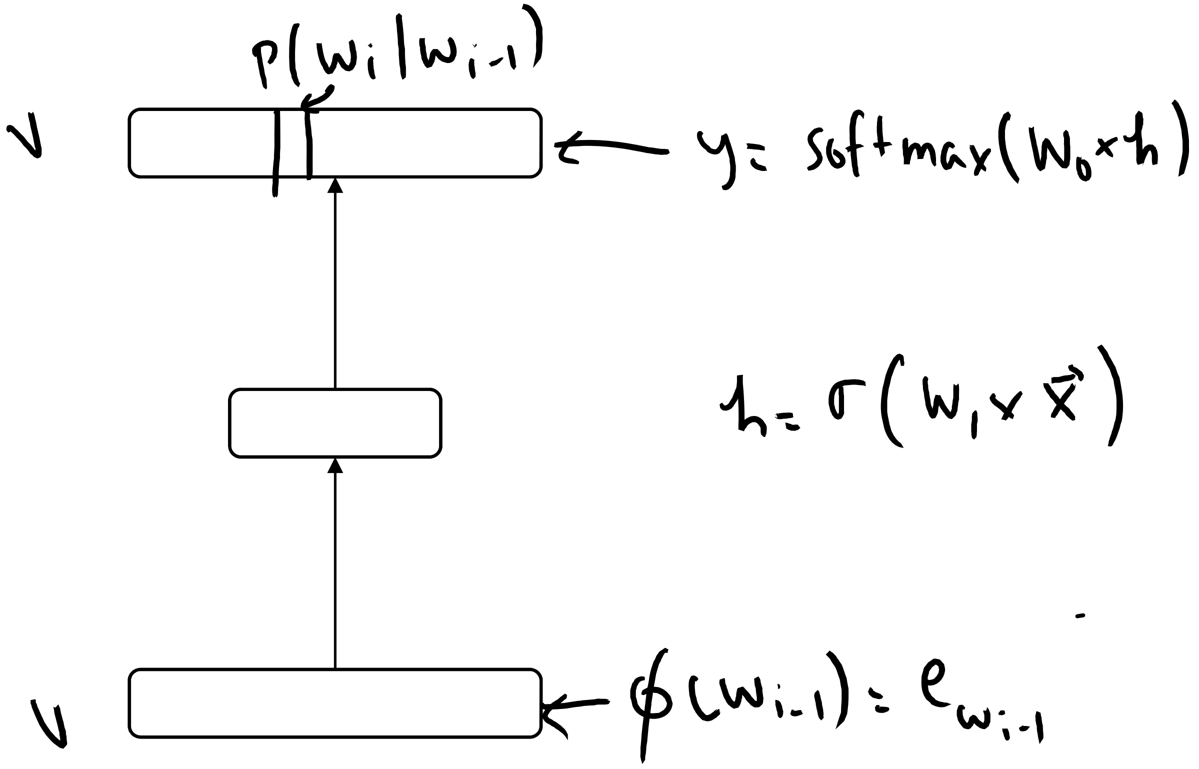
$$\min_{\mathbf{v},a,\mathbf{W},\mathbf{b}} \sum_{x_1\in\{0,1\}} \sum_{x_2\in\{0,1\}} \left(\text{xor}(x_1,x_2) - \underset{3}{\mathbf{v}}^\top \overset{\text{linear}}{\left(\underset{3\times 2}{\mathbf{W}}\underset{2}{\mathbf{x}} + \underset{3}{\mathbf{b}}\right)} + a\right)^2$$

$$\min_{\mathbf{v},a,\mathbf{W},\mathbf{b}} \sum_{x_1\in\{0,1\}} \sum_{x_2\in\{0,1\}} \left(\text{xor}(x_1,x_2) - \underset{3}{\mathbf{v}}^\top \underbrace{\tanh}_{\text{non-linear}}\left(\underset{3\times 2}{\mathbf{W}}\underset{2}{\mathbf{x}} + \underset{3}{\mathbf{b}}\right) + a\right)^2$$

https://github.com/clab/cnn/blob/master/examples/xor.cc

Simple Feedforward NN LM

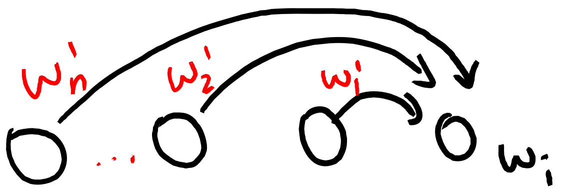
Bigram Model



$y = \text{softmax}(W_0 \times h)$

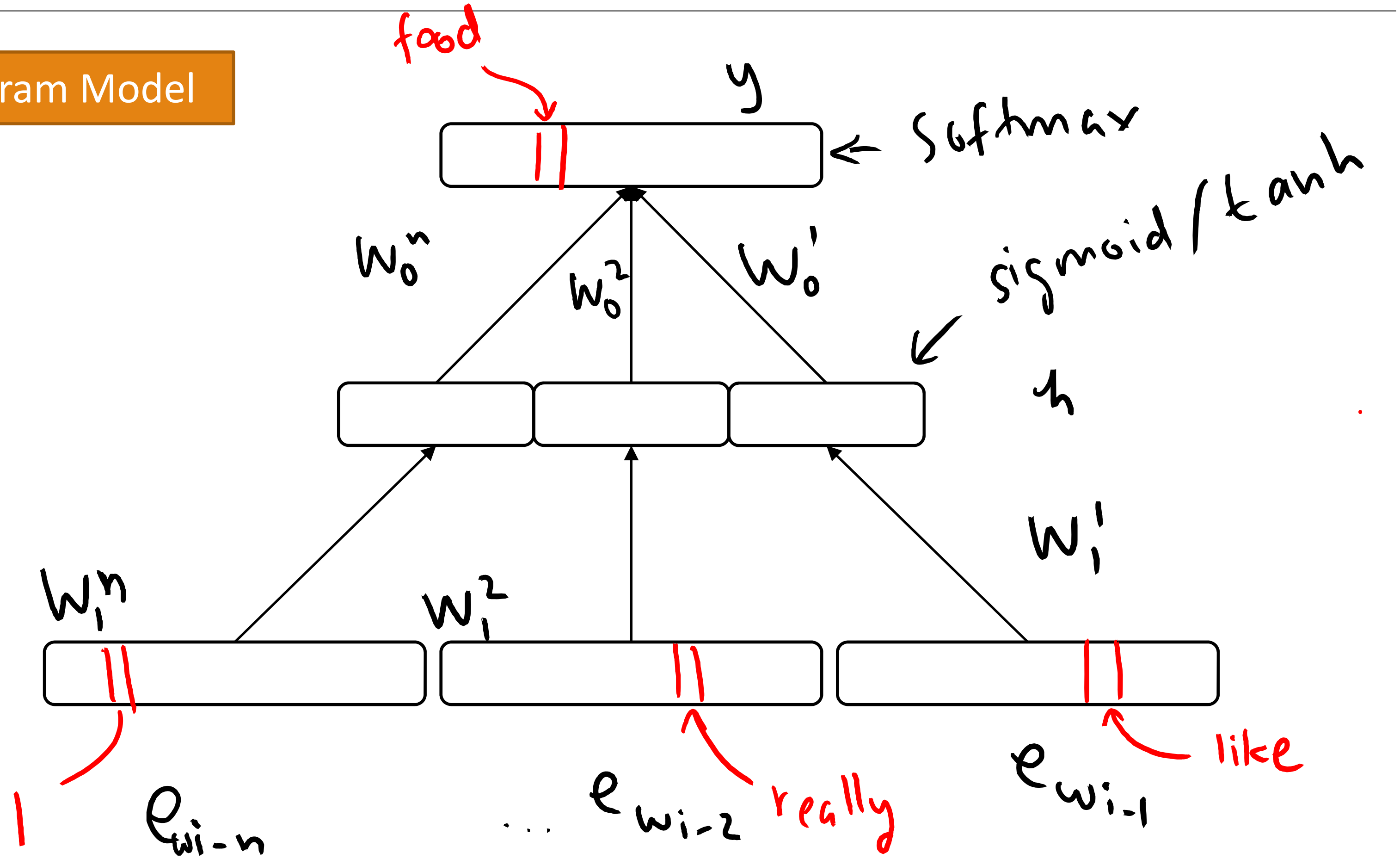
$h = \sigma(W_1 \times \vec{x})$

$\phi(w_{i-1}) = e_{w_{i-1}}$

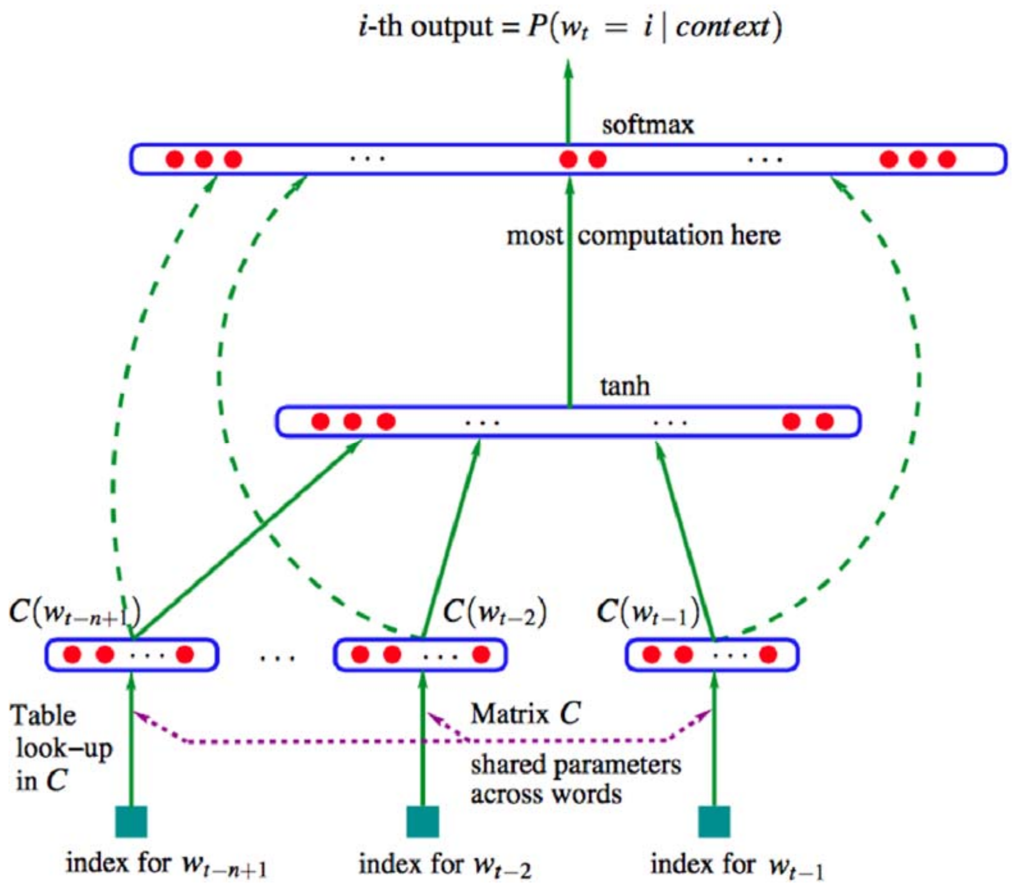


Simple Feedforward NN LM

N-gram Model



Deep Feedforward NN LM



Bengio et al. 2003

Outline

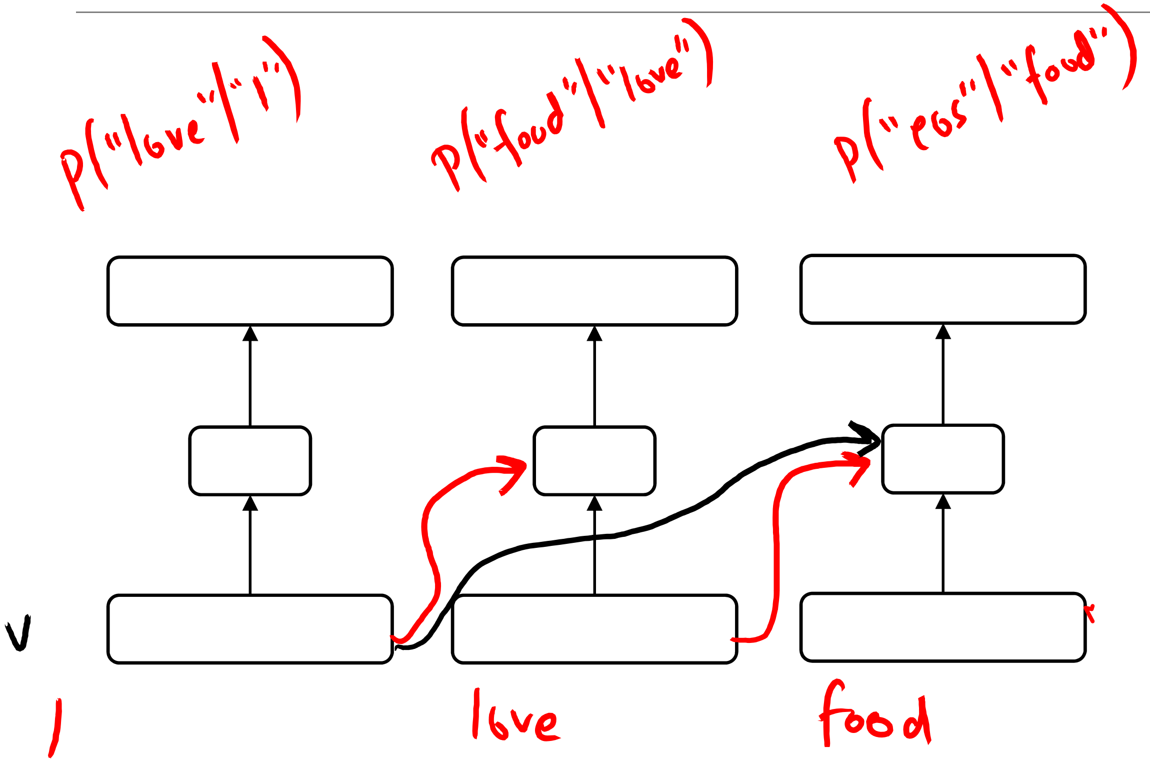
Discriminative Language Models

Feed-forward Neural Networks

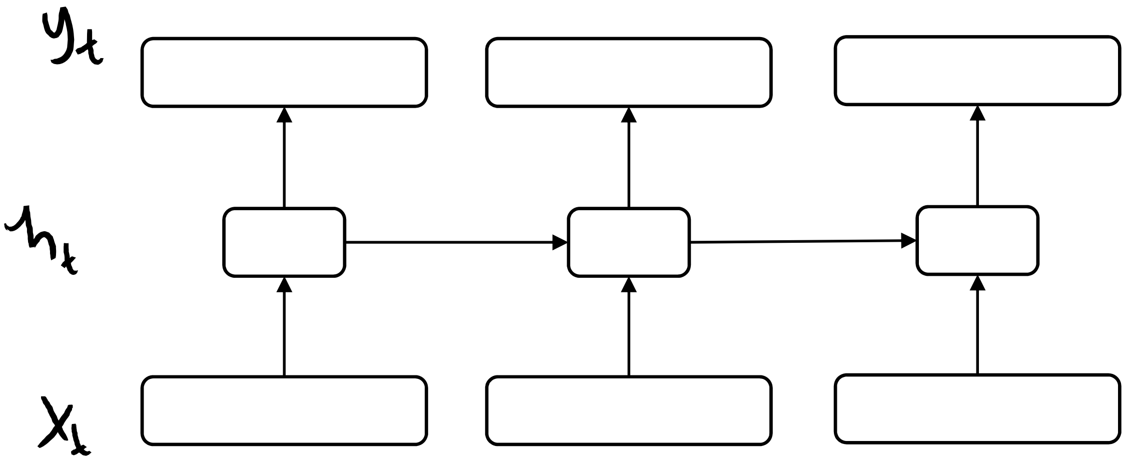
Recurrent Neural Networks

Upcoming..

Sequence View of Simple NNs



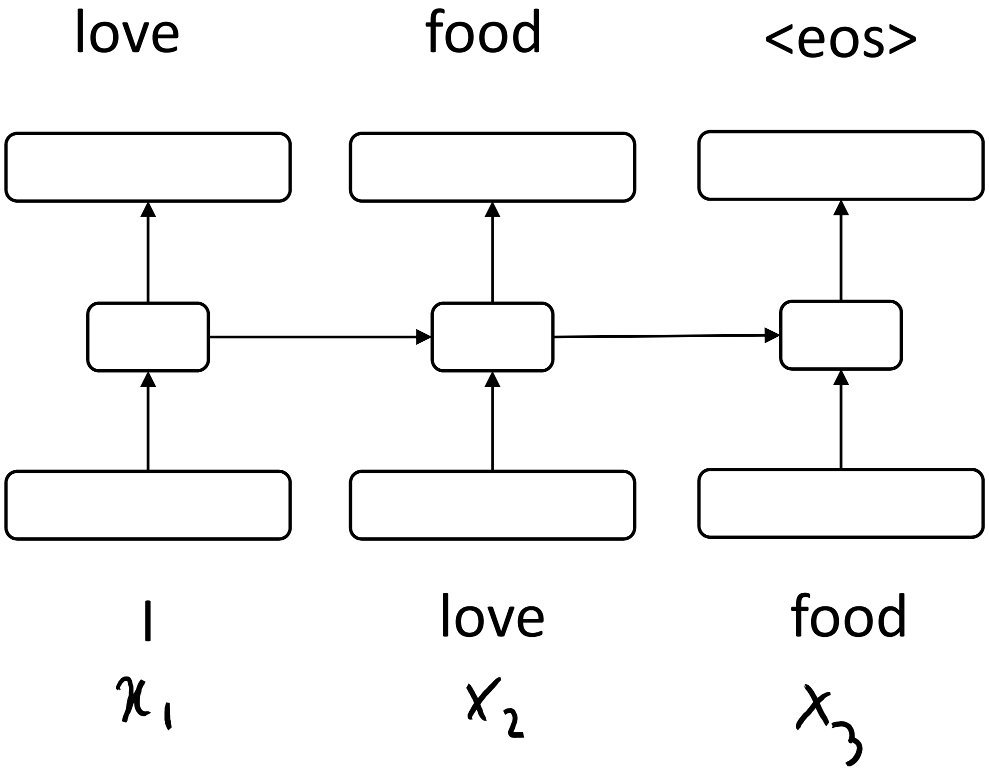
Recurrent Neural Networks



$$y_t = \mathrm{softmax}(W_o^{x} h_t)$$

$$h_t = \tanh(W_1 x + W_1' h_{t-1})$$

Example: “I love food”



love food <eos>

I love food

x_1 x_2 x_3

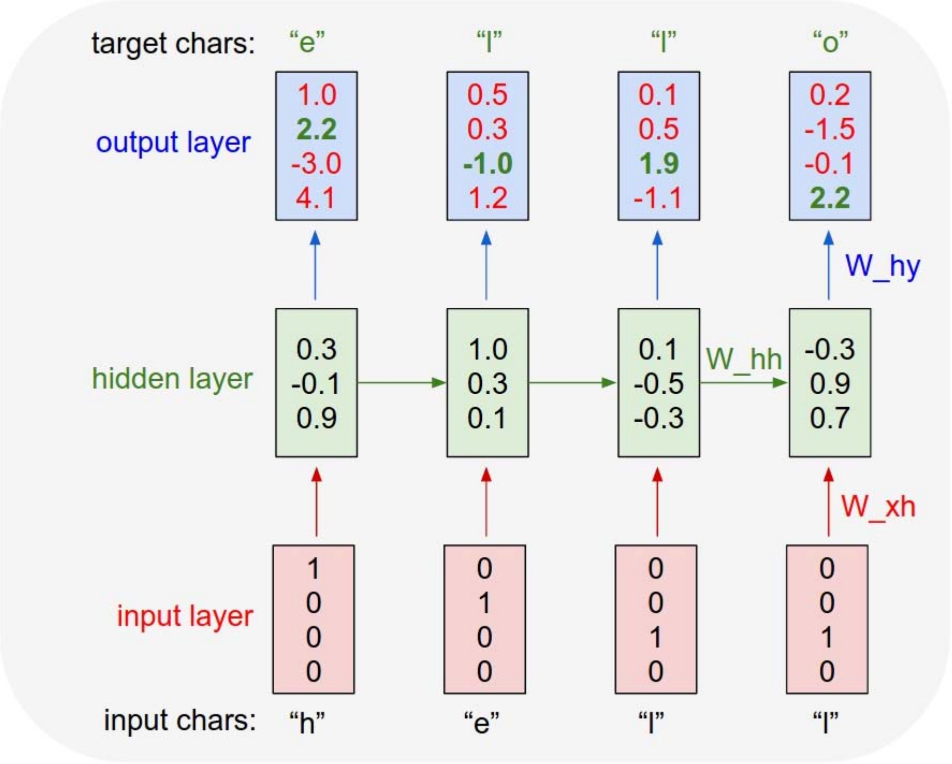
$$y_3 = s.m(W_0 h_3)$$

$$h_3 = \tanh(W_1 x_3 + W_1' h_2)$$

$$h_2 = \tanh(W_1 x_2 + W_1' h_1)$$

$$h_1 = \tanh(W_1 x_1 + W_1' \underbrace{h_0}_{\text{fix}})$$

Power of RNNs: Characters!



target chars:	"e"	"l"	"l"	"o"
output layer	1.0 **2.2** -3.0 4.1	0.5 0.3 **-1.0** 1.2	0.1 0.5 **1.9** -1.1	0.2 -1.5 -0.1 **2.2**
hidden layer	0.3 -0.1 0.9	1.0 0.3 0.1	0.1 -0.5 -0.3	-0.3 0.9 0.7
input layer	1 0 0 0	0 1 0 0	0 0 1 0	0 0 1 0
input chars:	"h"	"e"	"l"	"l"

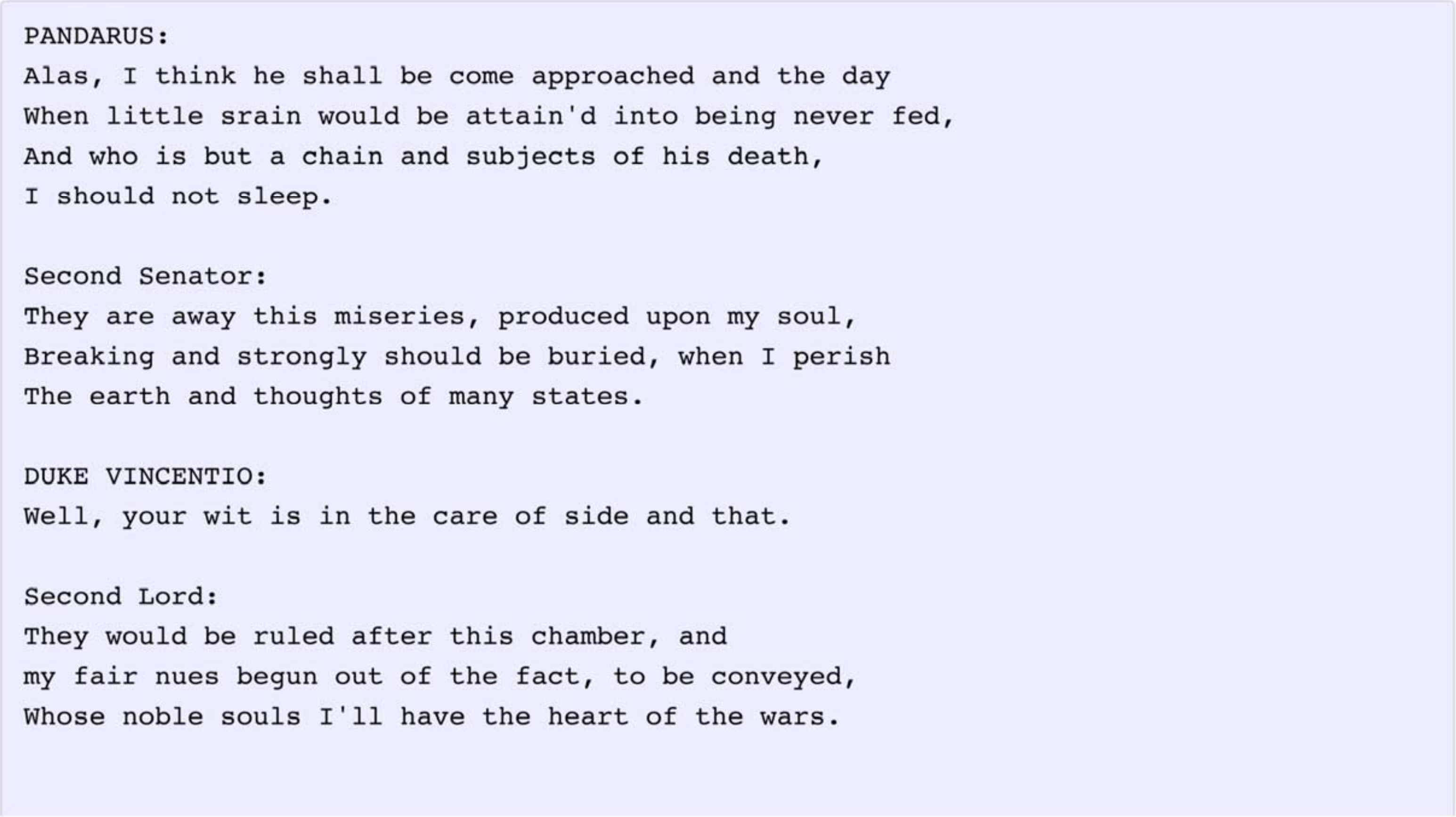
W_hy

W_hh

W_xh

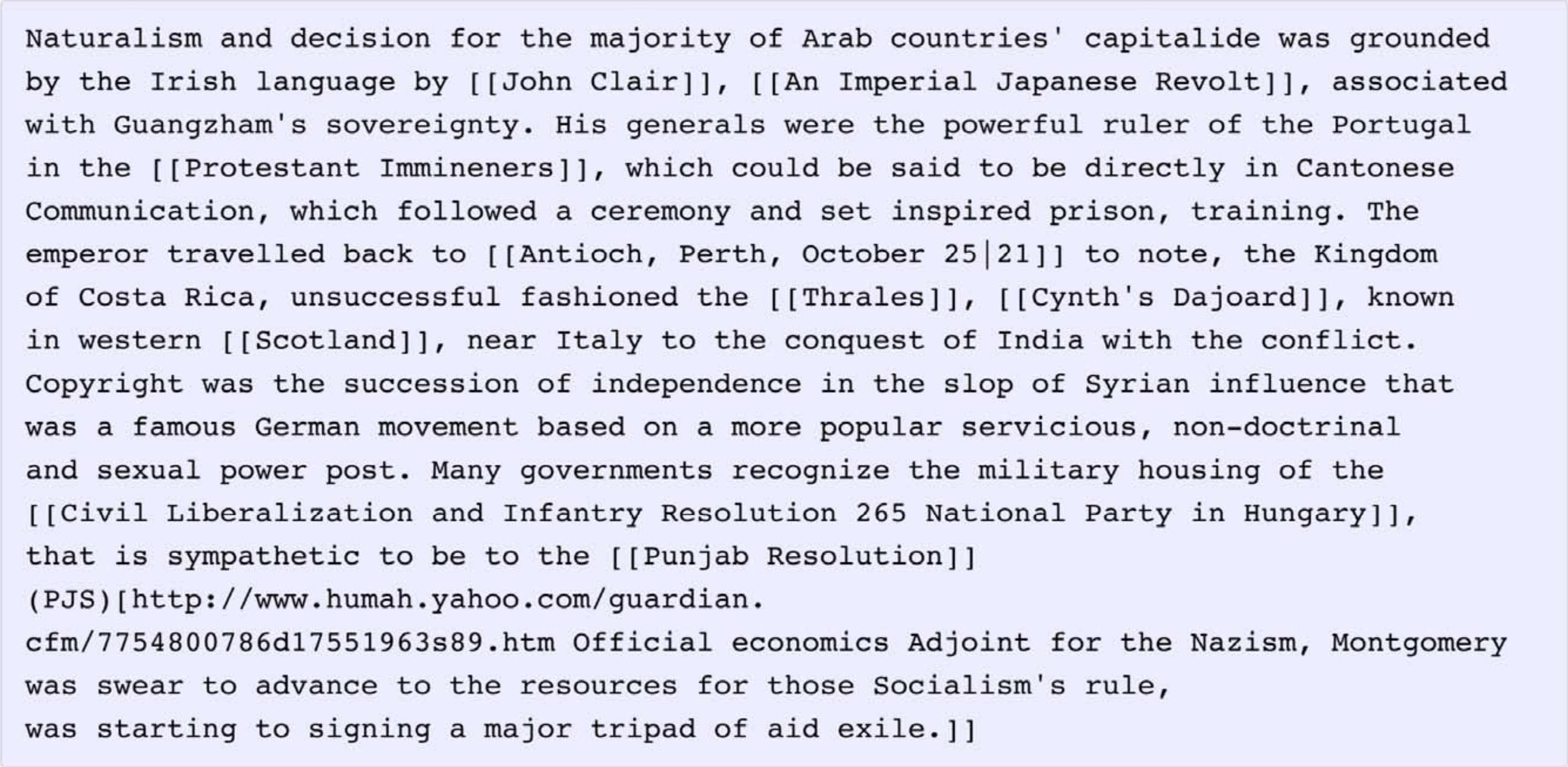
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Char-RNNs: Shakespeare!



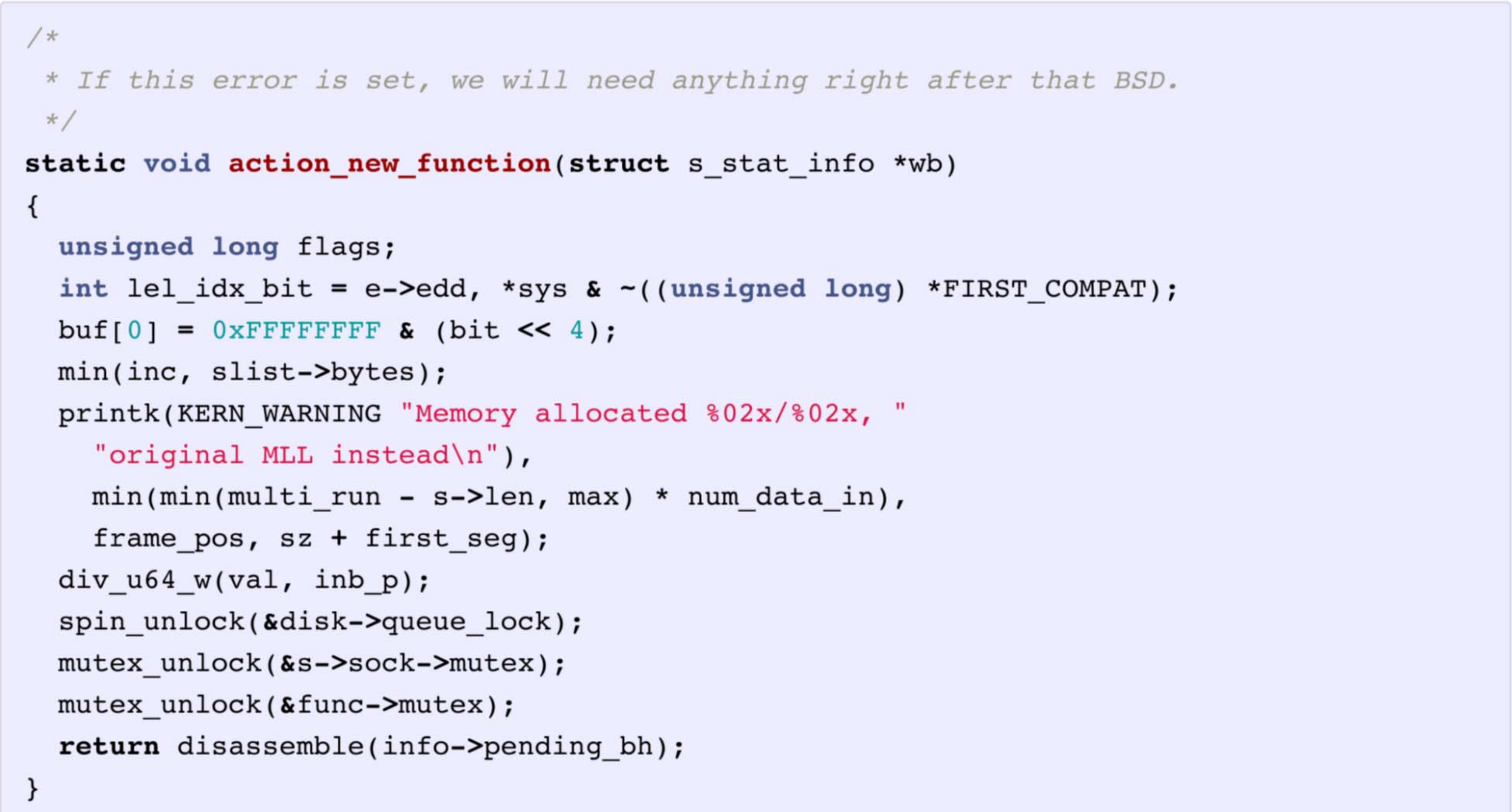
```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

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.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.
```

Char-RNNs: Wikipedia!



```
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.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

Char-RNNs: Linux Code!

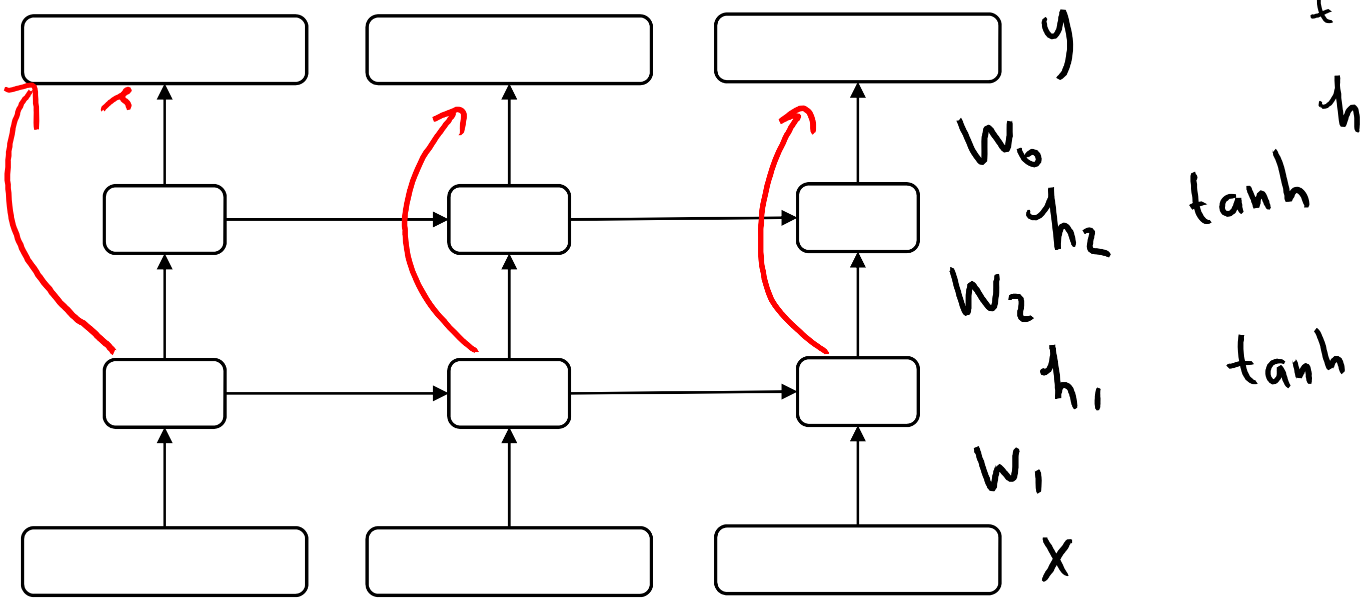


```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

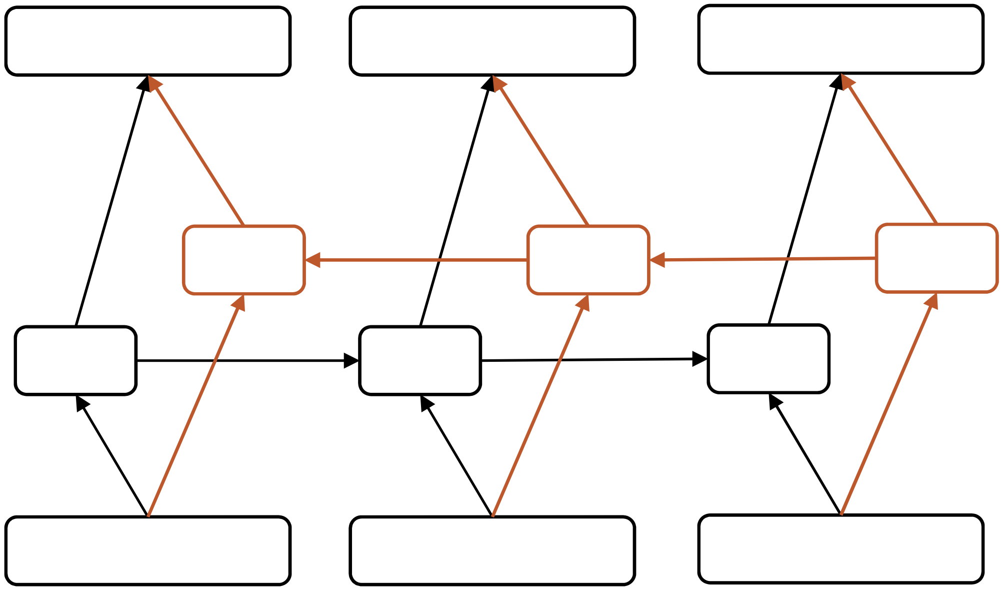
Extension: Stacking

$$y_t = \sigma(W_0 \times h_2^t)$$

$$h_2^t = \sigma(W_2 h_1^t + W_2' h_2^{t'})$$



Extension: Bidirectional RNNs

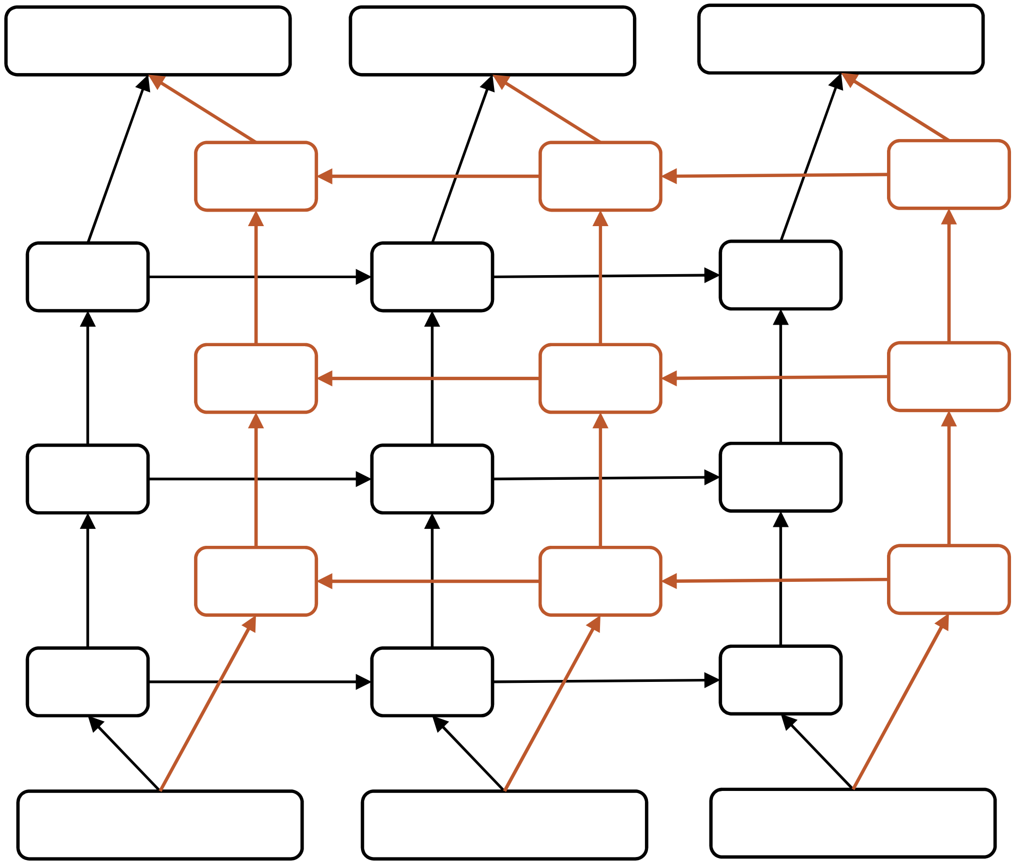


$$y_t = \sigma\left(W_o^f h_t^f + W_o^b h_t^b\right)$$

$$h_t^f \leftarrow h_{t-1}^f$$

$$h_t^b \leftarrow h_{t+1}^b$$

Deep Bidirectional RNNs

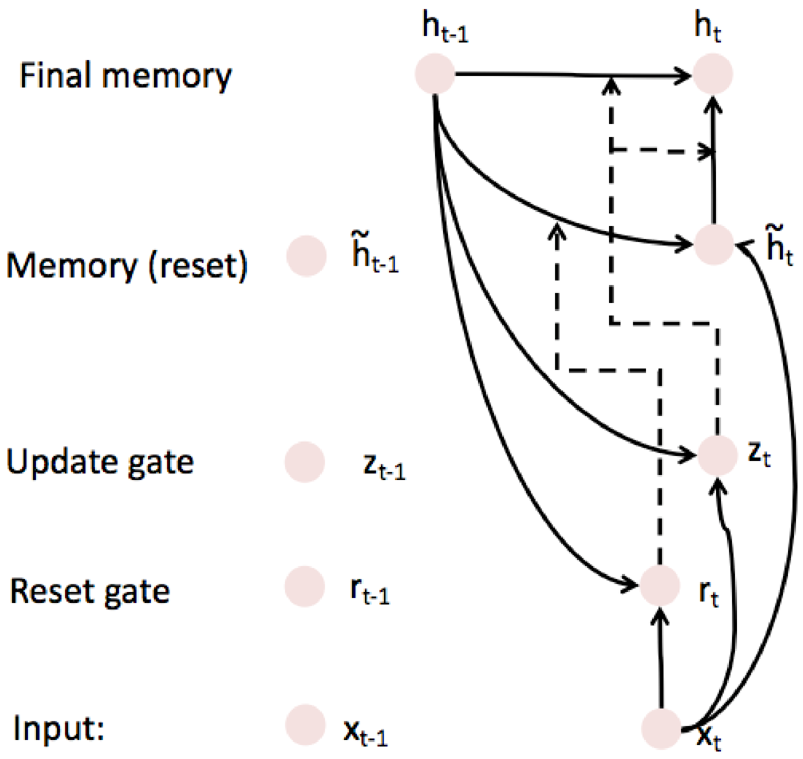


Extension: GRUs

Gated Recurrent Units

Extension: GRUs

Gated Recurrent Units



$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right)$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right)$$

$$\tilde{h}_t = \tanh\left(Wx_t + r_t \circ Uh_{t-1}\right)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t$$

Estimating Parameters

Beyond the scope of the course

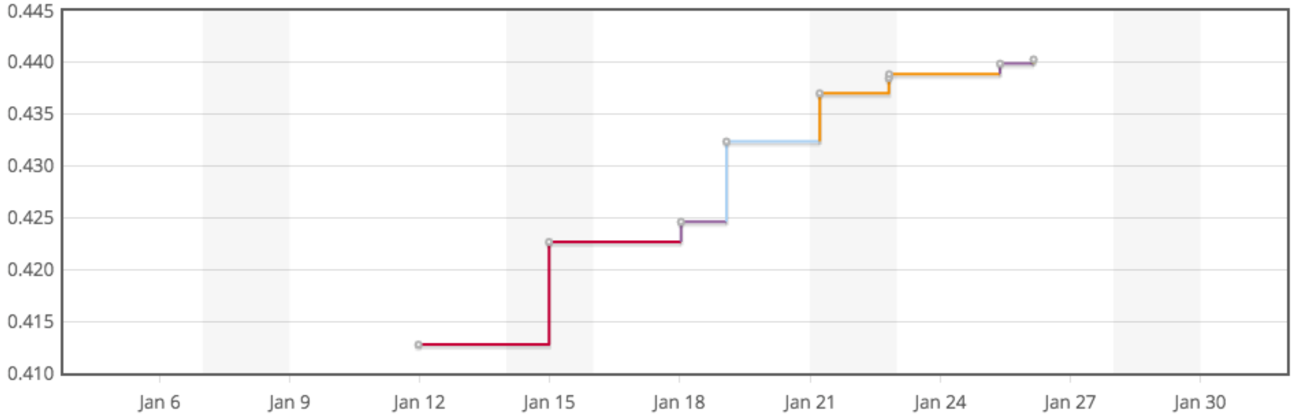
- Lots of tricks, heuristics, “domain knowledge”
- Lot of engineering for efficiency, e.g. GPUs
- New training algorithms being proposed every year
  - sometimes, architecture-specific
- Lots of available tools you can use!
  - Tensorflow, Torch, Keras, MxNET, etc.

Outline

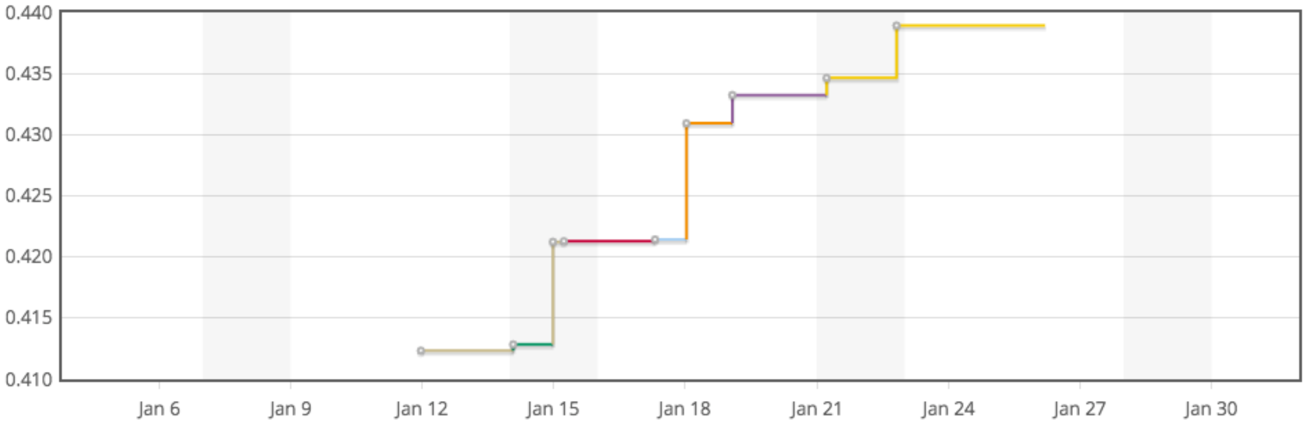
- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks
- Upcoming..

Homework 1 so far…

Public



Private



Ruslan Salakhutdinov

Professor at Carnegie Mellon University
Director of Artificial Intelligence, Apple Inc.

Learning Deep Unsupervised and Multimodal Models

Location: DBH 6011
Time: 11am - 12pm
Date: January 27, 2017

Meeting with PhD students, will post on Piazza

Upcoming...

Homework

- Homework 1 is due tonight: January 26, 2017
- Write-up, data, and code for Homework 2 is up
- Homework 2 is due: February 9, 2017

Project

- Proposal is due: February 7, 2017 (~2 weeks)
- Only 2 pages