

The Computational Graphs / RNN Language Models

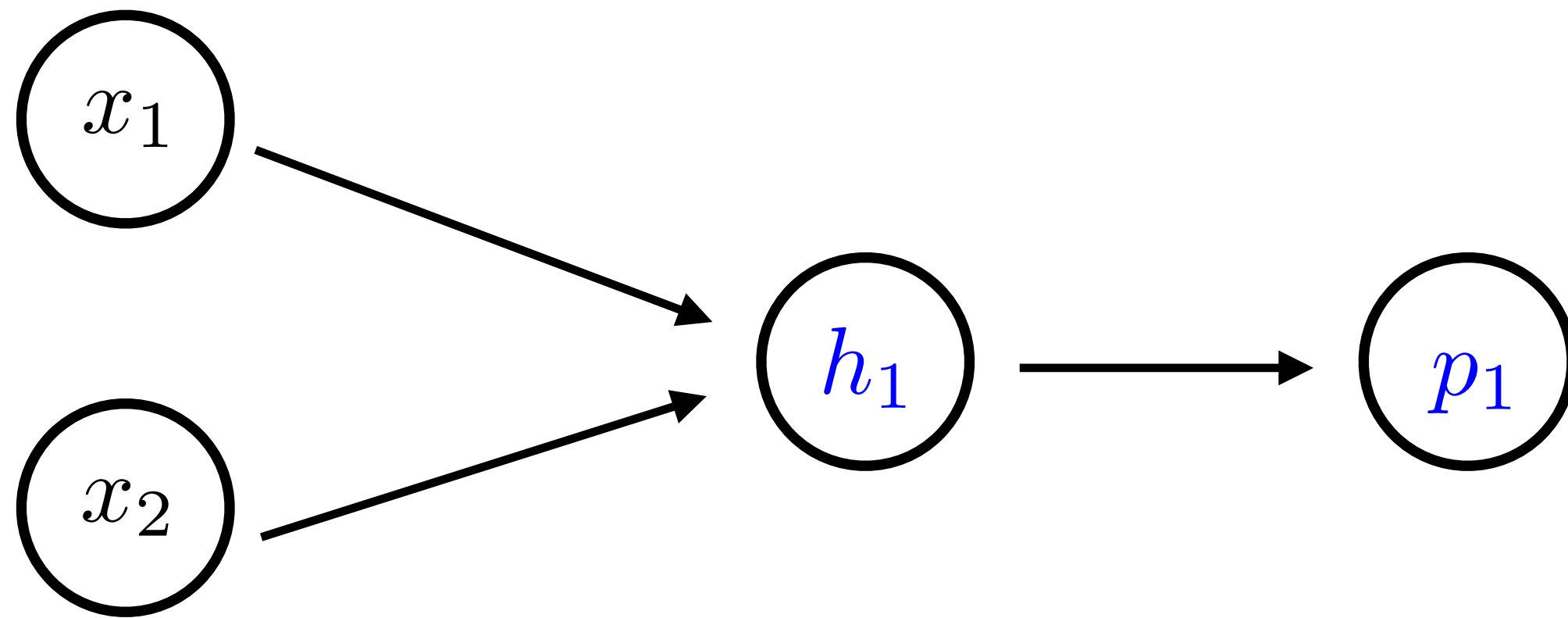
COMP7607—Lecture 2

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Some materials from Stanford University CS224n with special thanks!

Logistic Regression

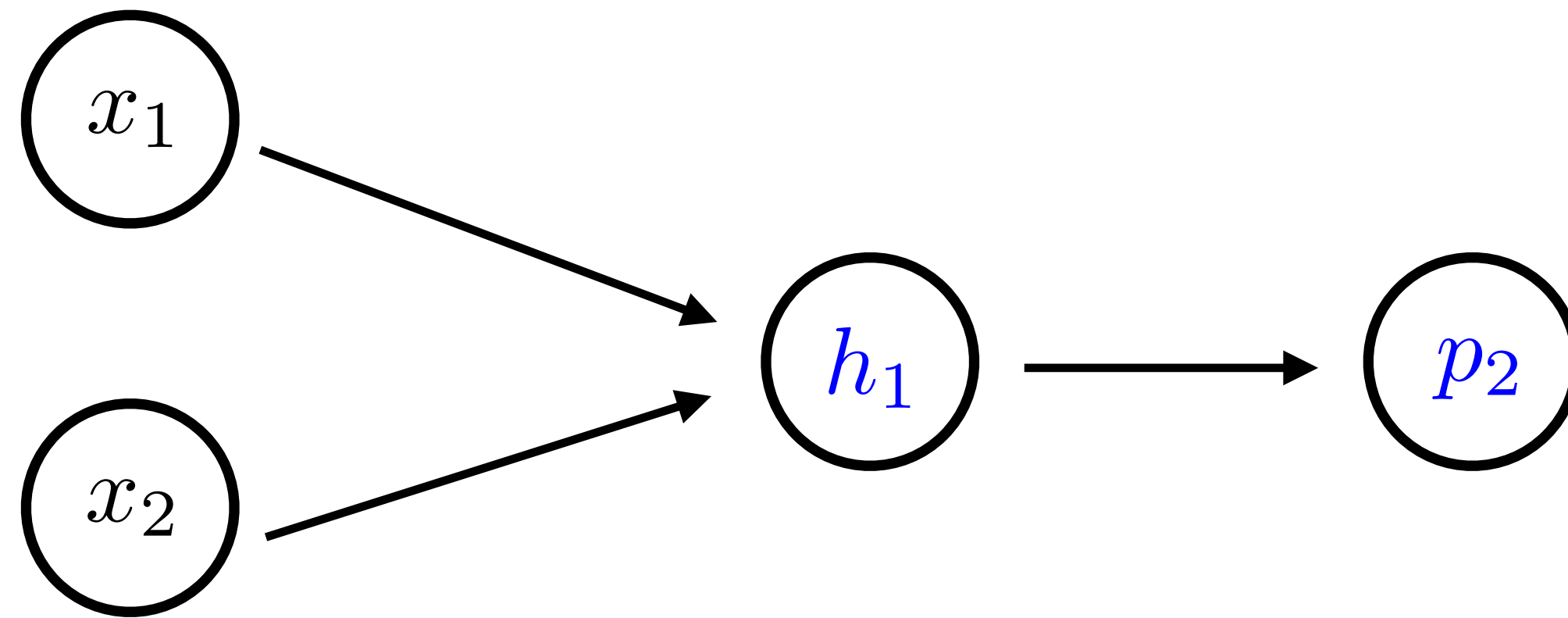


$$p_1 := p(y = 1 \mid x_1, x_2)$$

$$h_1 = w_1 x_1 + w_2 x_2 + b$$

$$p_1 = \frac{1}{1 + \exp(-h_1)}$$

Logistic Regression



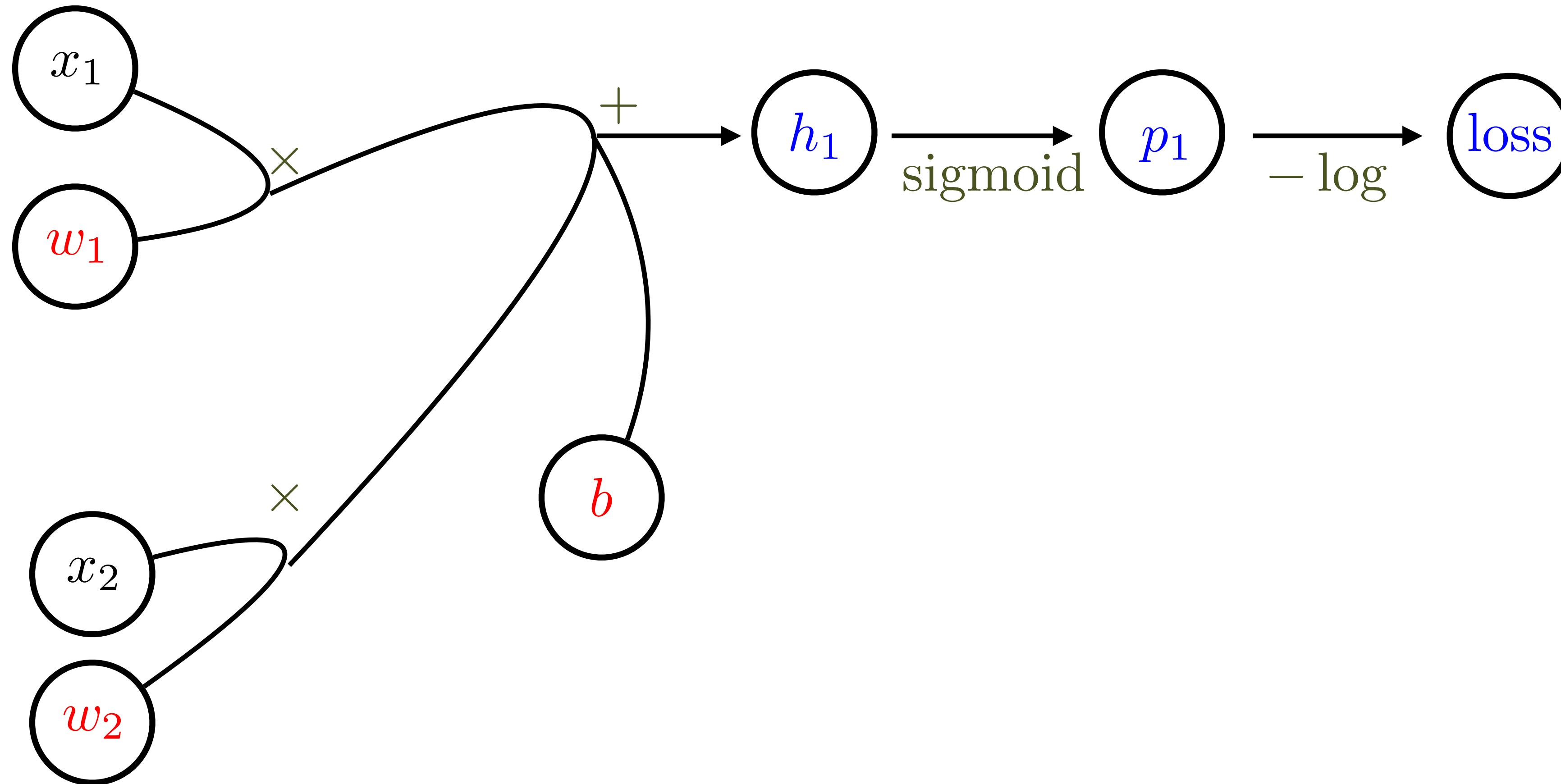
$$p_2 := p(y = 0 \mid x_1, x_2)$$

$$h_1 = w_1 x_1 + w_2 x_2 + b$$

$$p_2 = 1 - \frac{1}{1 + \exp(-h_1)} = \frac{\exp(-h_1)}{1 + \exp(-h_1)}$$

Loss Function

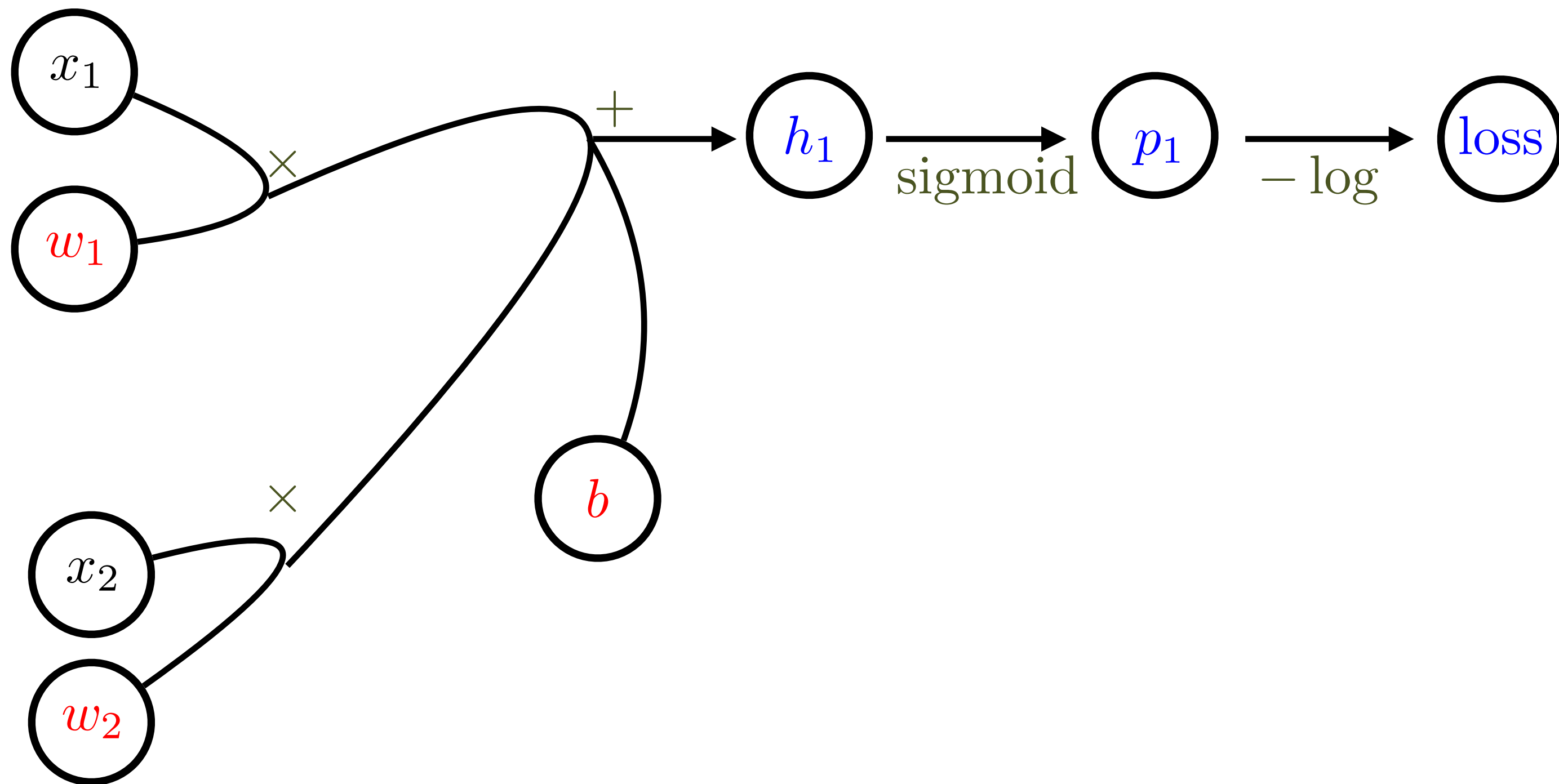
case $y = 1$:



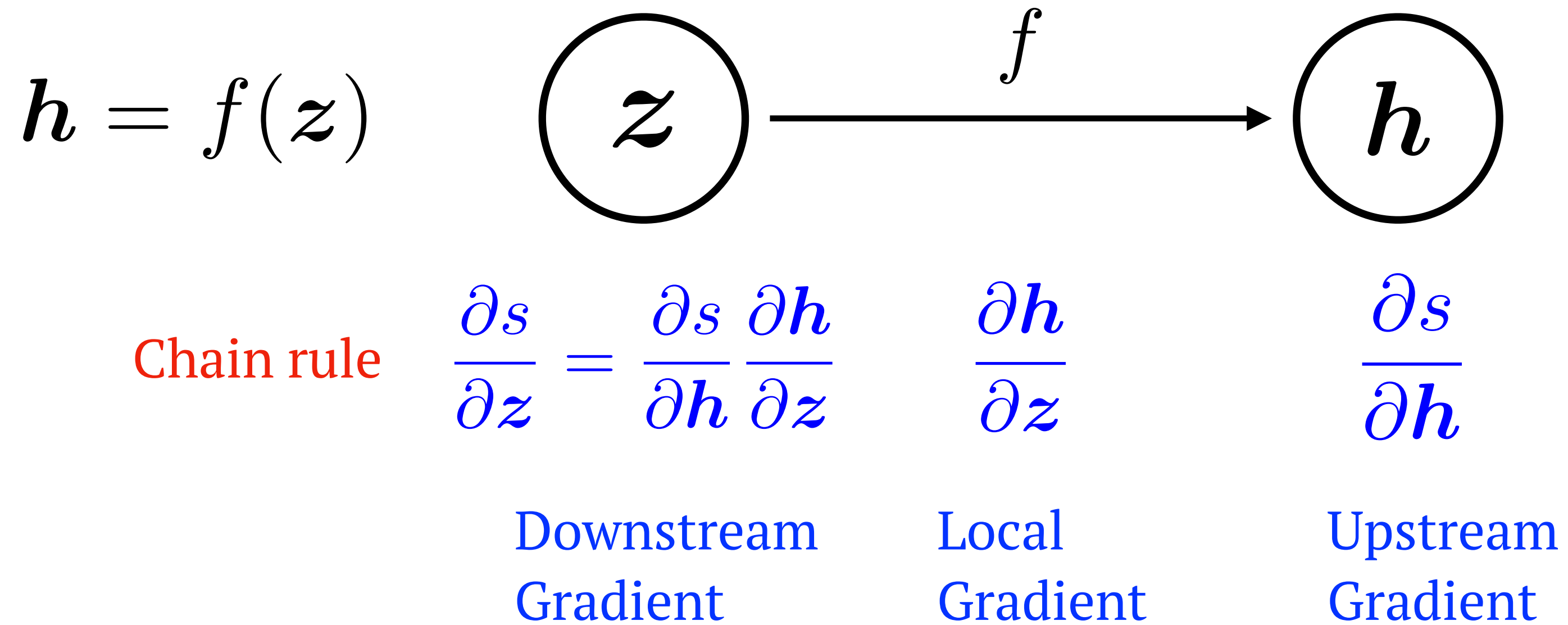
Computational Graphs

Input	x_1	x_2	
Parameter	w_1	w_2	
Expression	h_1	p_1	loss
Operation	\times	$+$	sigmoid $-\log$

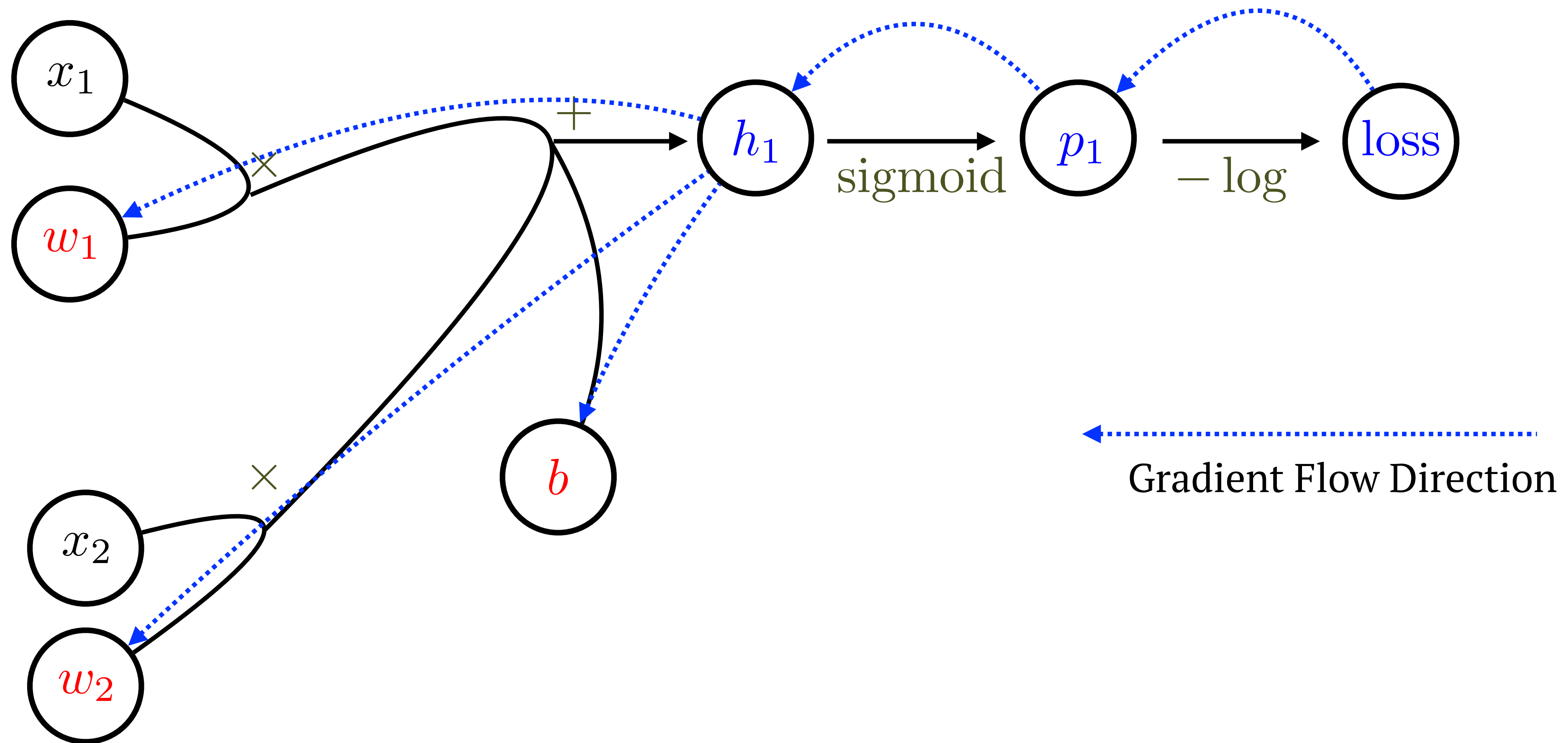
Special



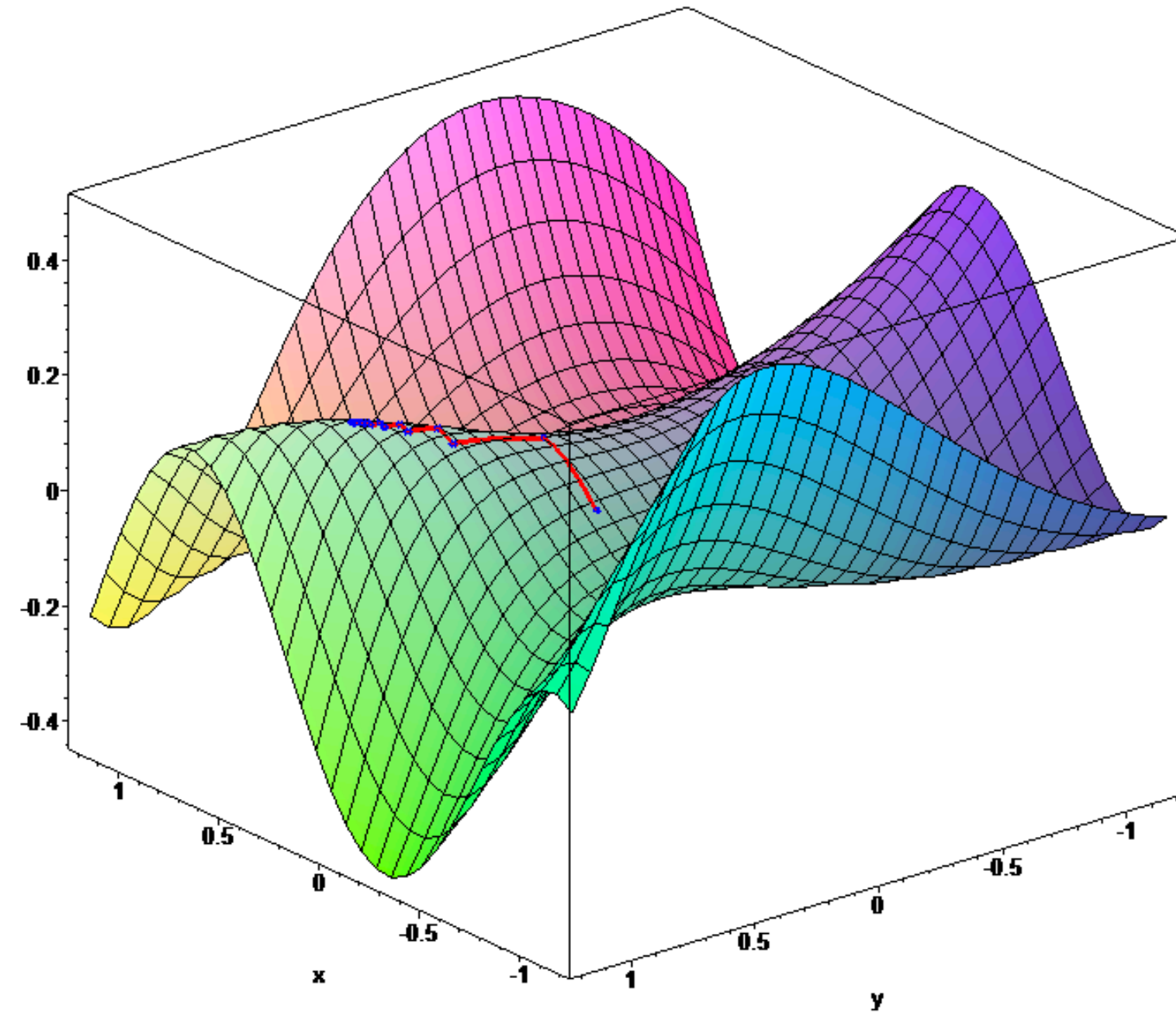
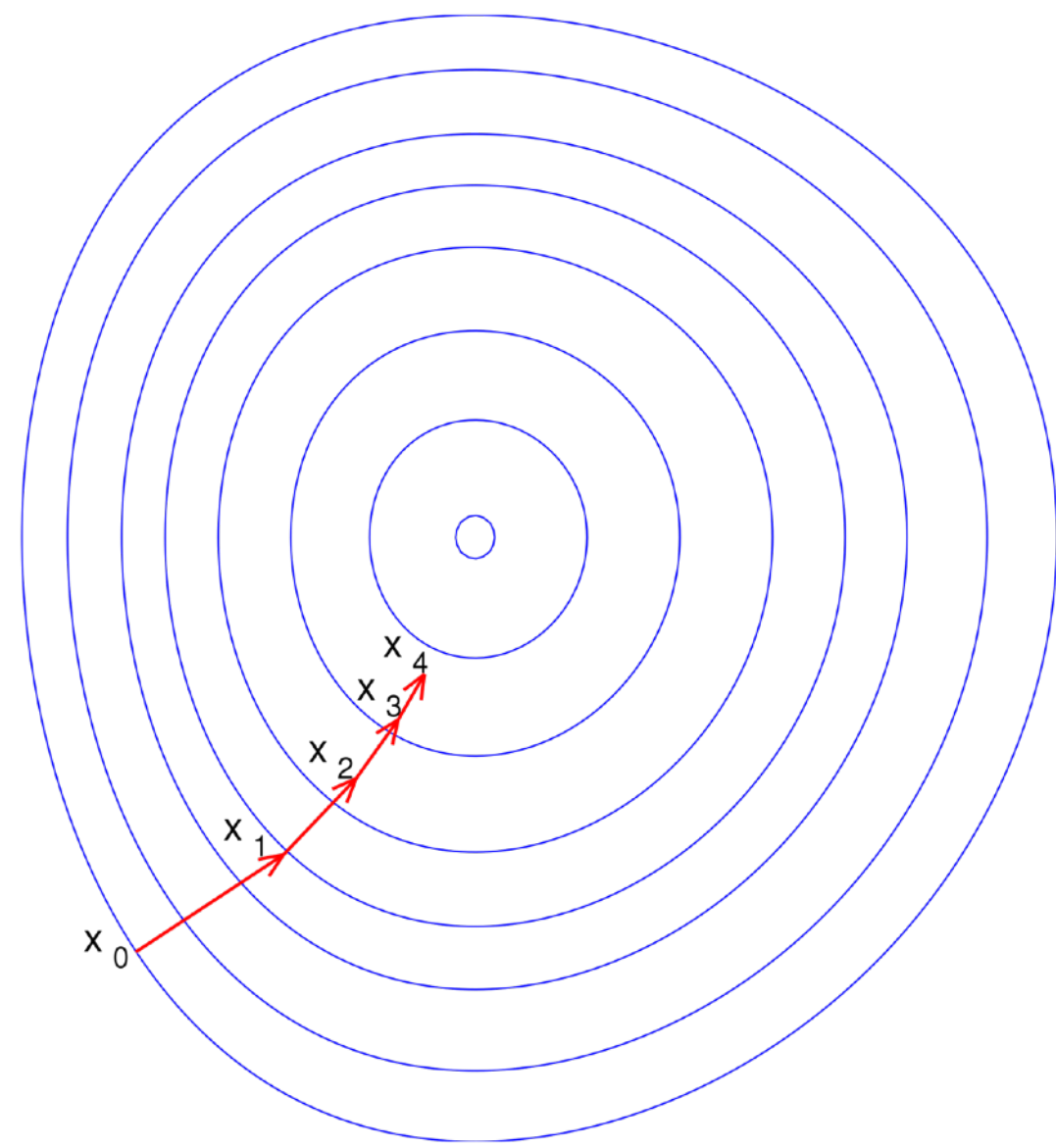
How to minimize? (Automatic Differentiation)



How to minimize? (Automatic Differentiation)



Review: Gradient Descent

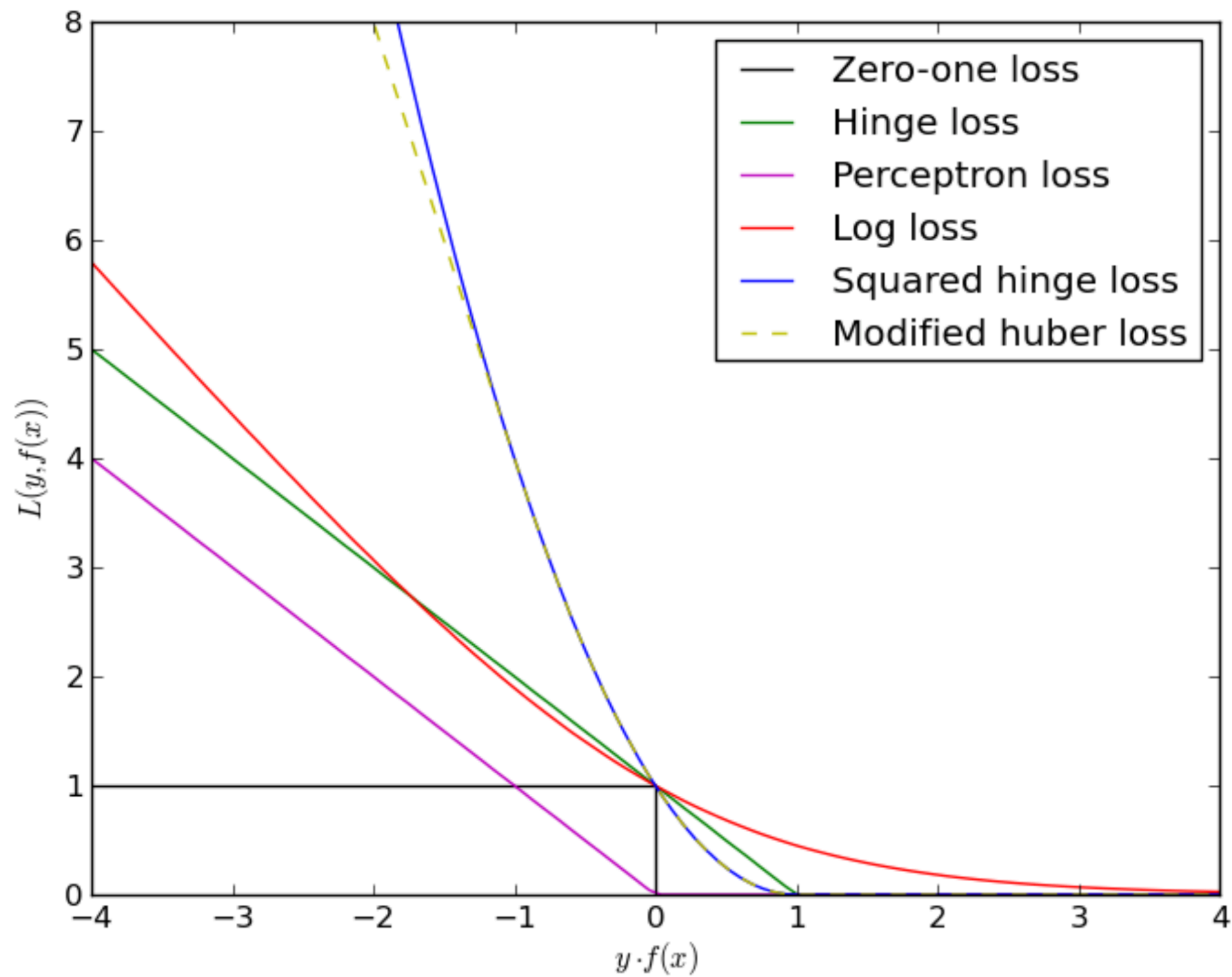


fog in the mountains

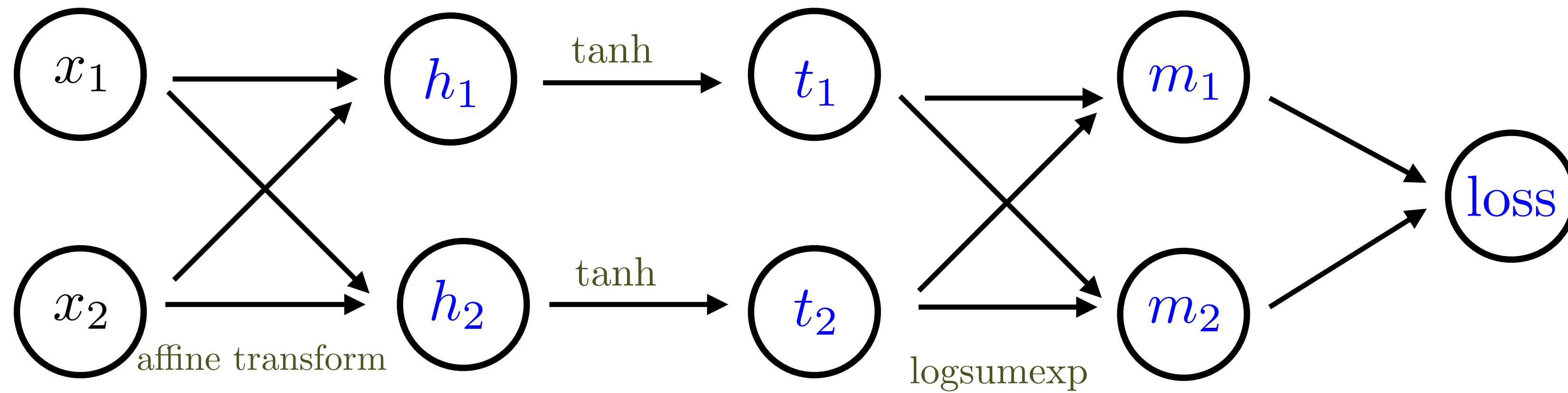
$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} \mathcal{L}(\theta) |_{\theta_t} \quad \text{negative gradient (descent direction)}$$

step size
(learning rate)

Other Loss Function?

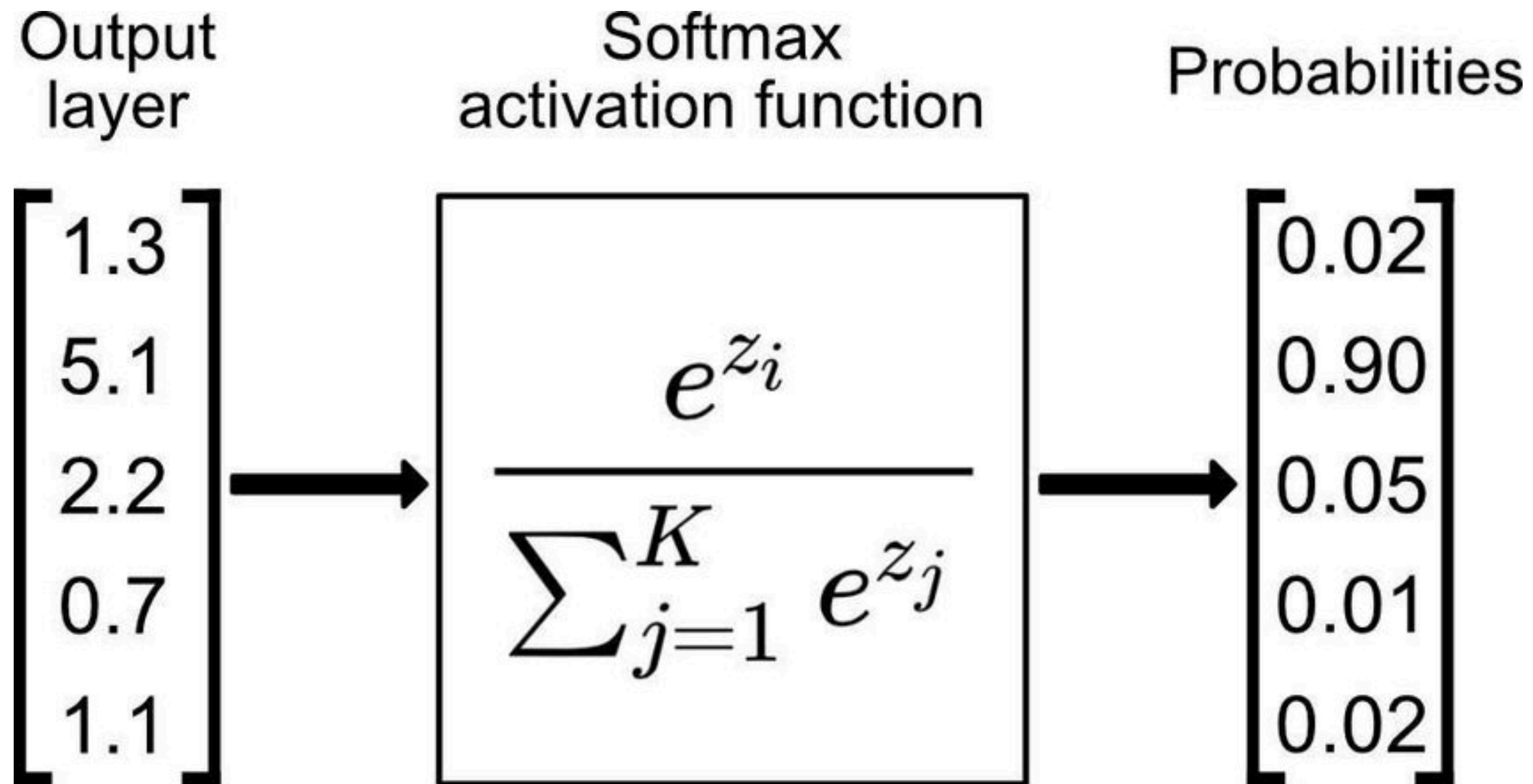


“Deeper” Neural Network

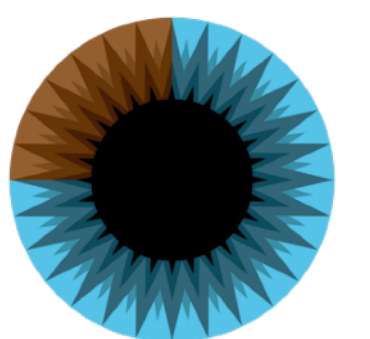
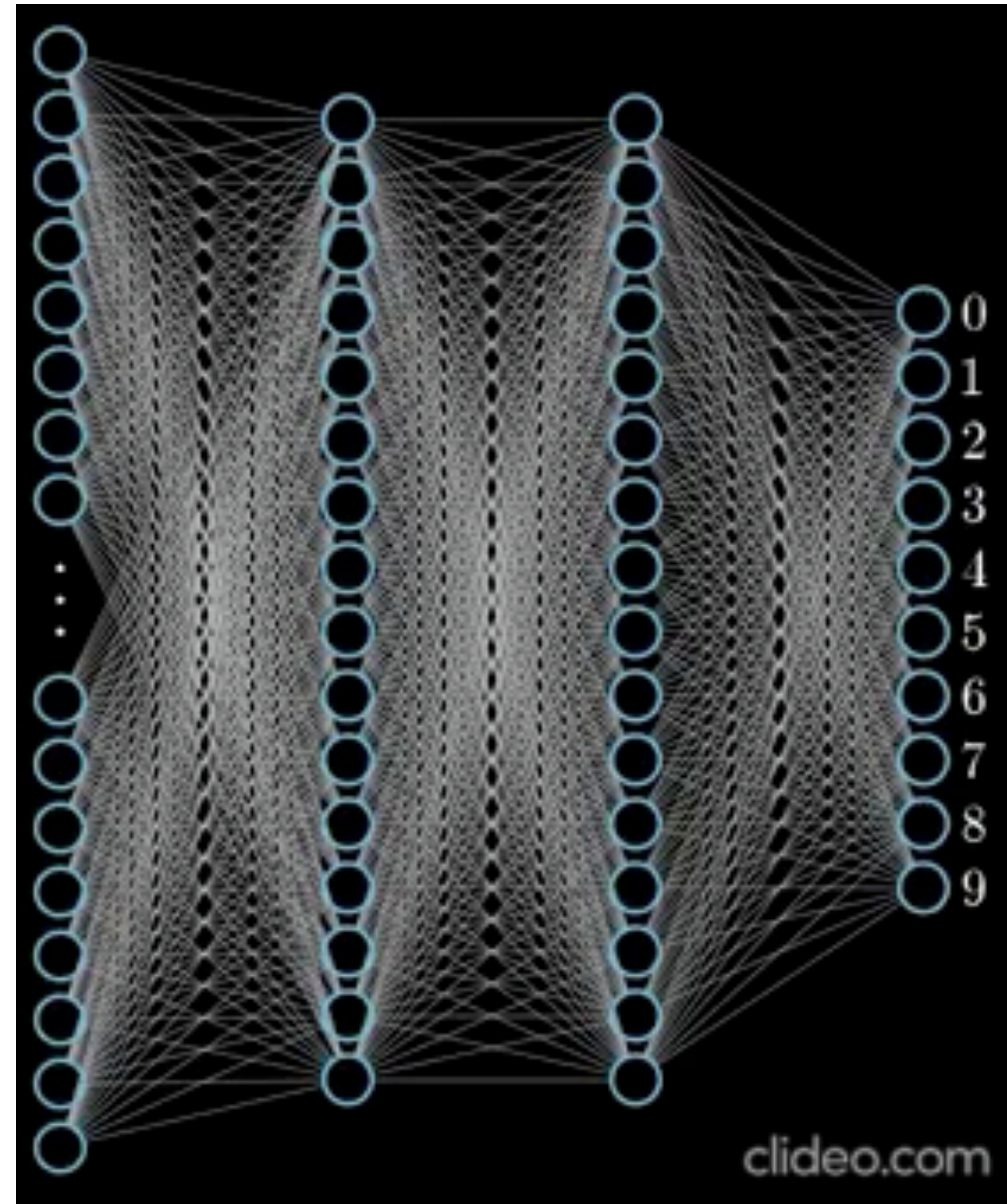


$$m_1 = \log\left(\frac{\exp(t_1)}{\exp(t_1) + \exp(t_2)}\right)$$

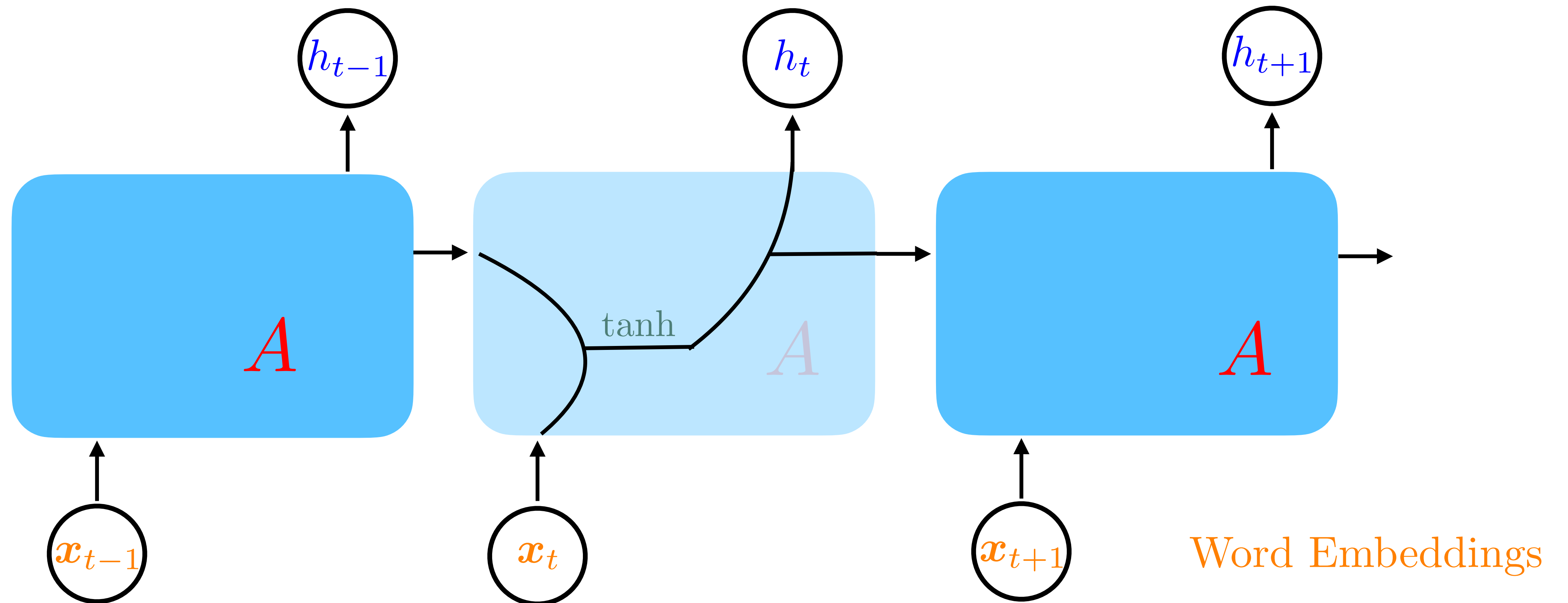
Softmax Function



Neutralize the dice!



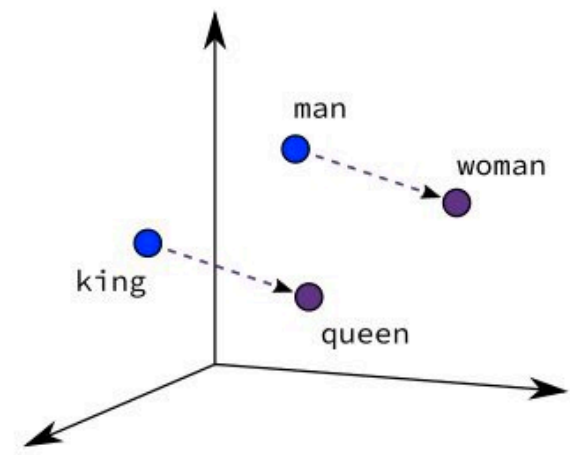
Recurrent Neural Network



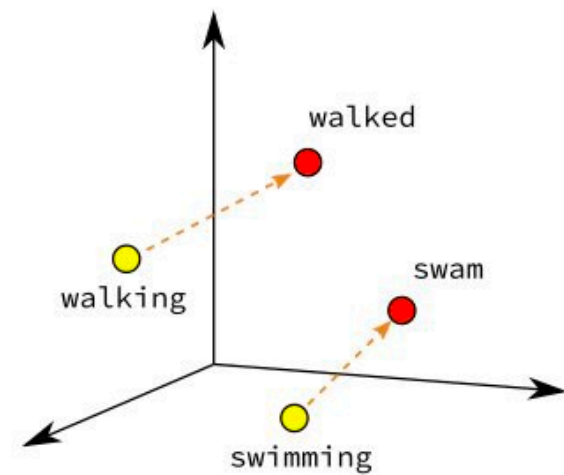
Word Embeddings



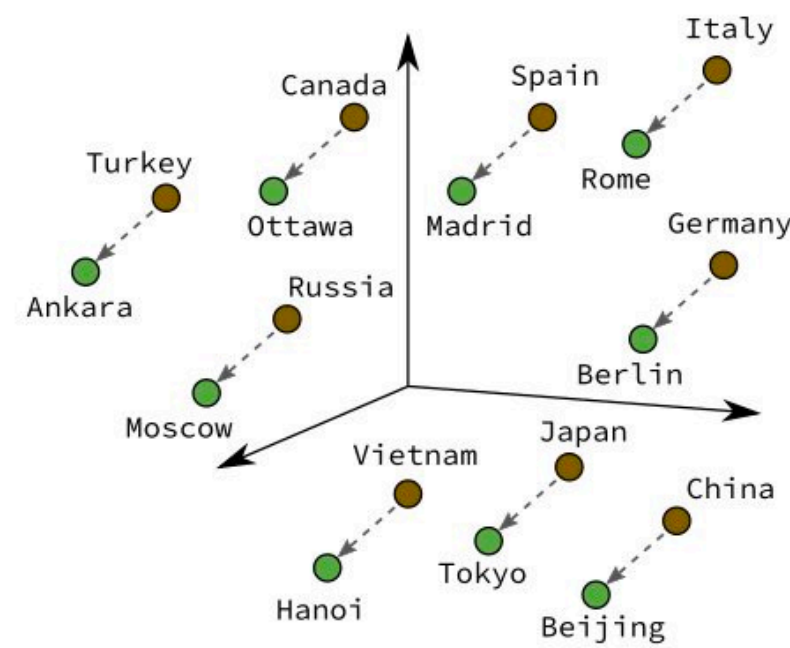
cat



Male-Female



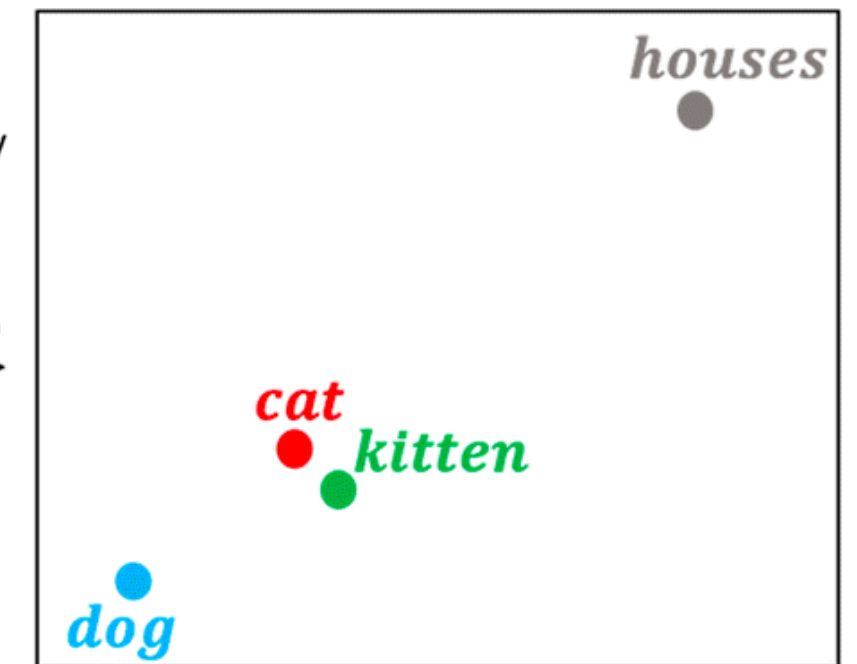
Verb Tense



Country-Capital

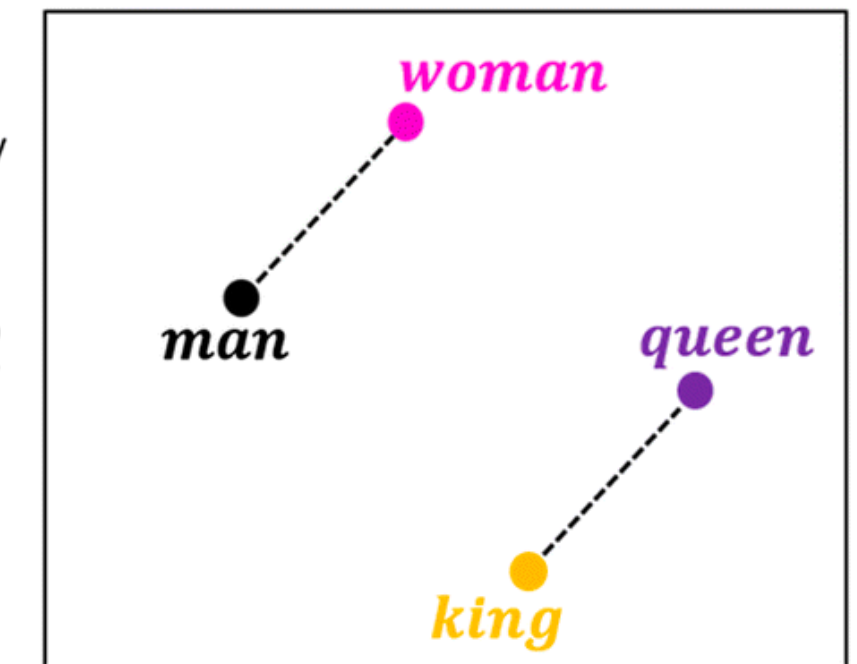
	living being	feline	human	gender	royalty	verb	plural
<i>cat</i>	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i>	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i>	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i>	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Dimensionality reduction of word embeddings from 7D to 2D



<i>man</i>	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i>	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i>	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i>	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Dimensionality reduction of word embeddings from 7D to 2D



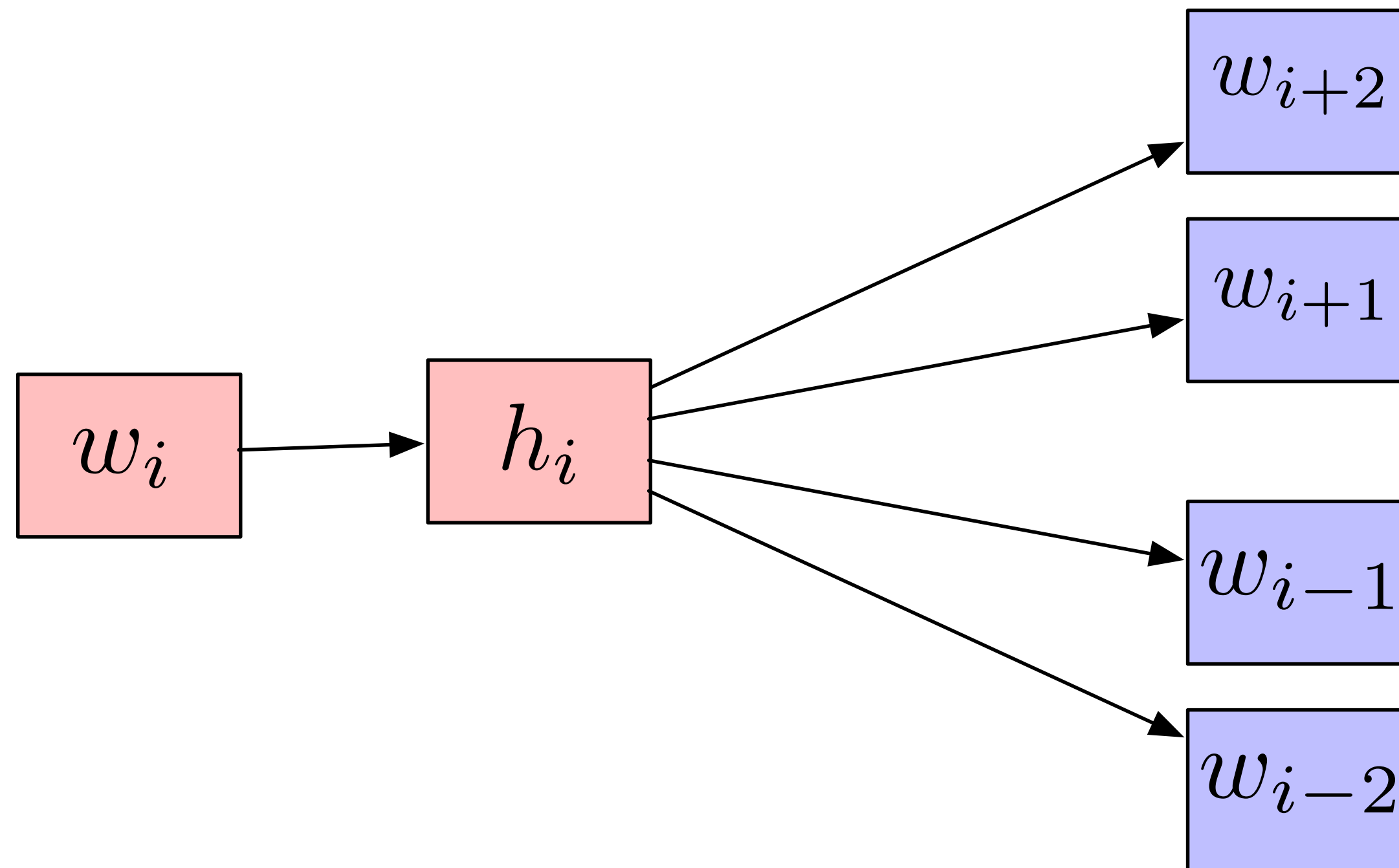
Word

Word embedding

Dimensionality reduction

Visualization of word embeddings in 2D

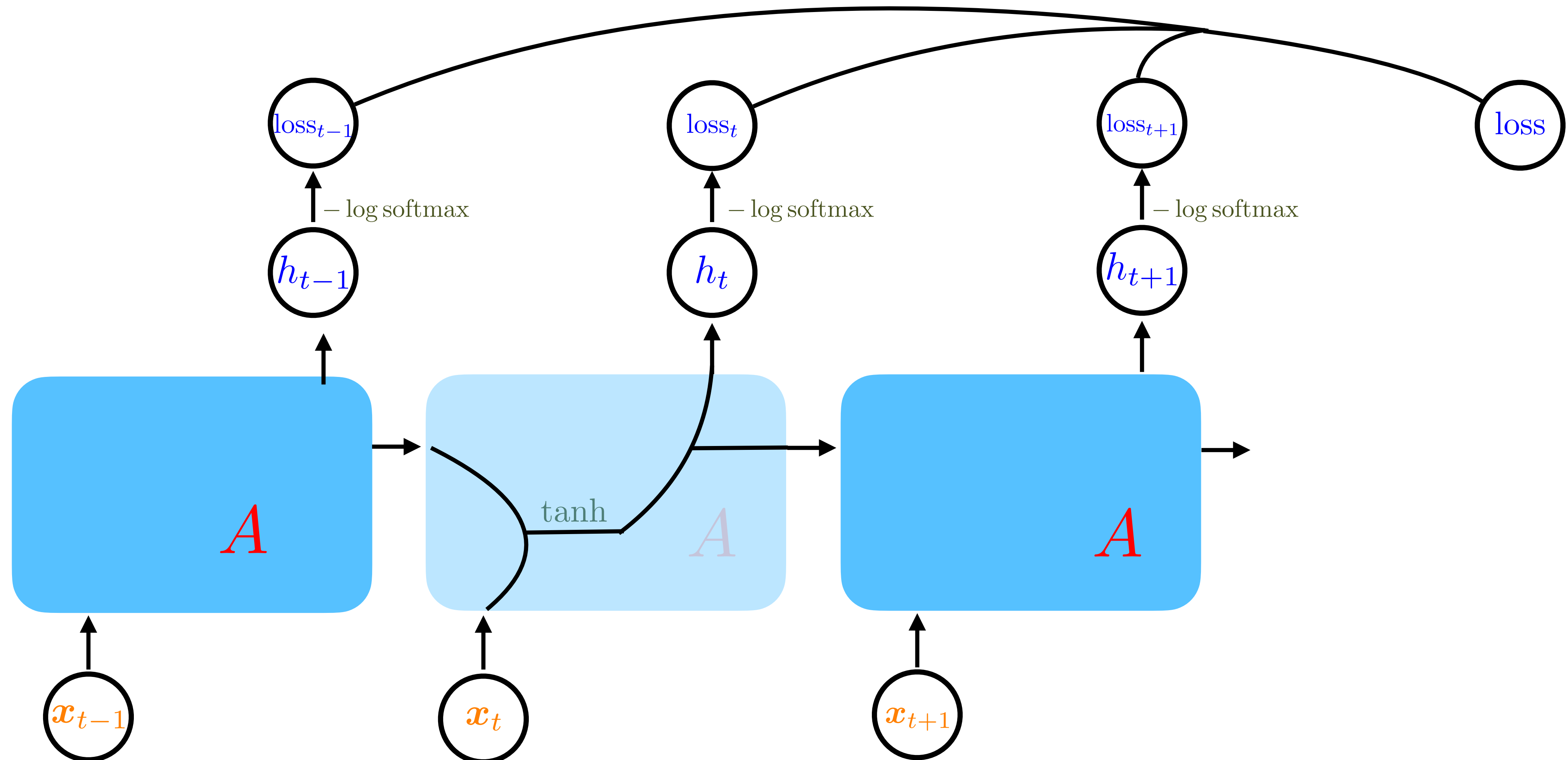
Word Embeddings



1 y the irradiated and refrigerated chicken. Acceptance of radiopasteurization
2 torehouse". Glendora dropped a chicken and a flurry of feathers, and went
3 will specialize in steaks, chops, chicken and prime beef as well as Tom's fa
4 ard as the one concerned with the chicken and the egg. Which came first? Is
5 he millions of buffalo and prairie chicken and the endless seas of grass that
6 "! "Come on, there's some cold chicken and we'll see what else". They wen
7 ves to extend the storage life of chicken at a low cost of about 0.5 cent per
8 CHICKEN CADILLAC# Use one 6-ounce chicken breast for each guest. Salt and pe
9 ion juice, to about half cover the chicken breasts. Bake slowly at least one-
10 d, in butter. Sprinkle over top of chicken breasts. Serve each breast on a th
11 around, they had a hard time". #CHICKEN CADILLAC# Use one 6-ounce chicken
12 successful, and the shelf life of chicken can be extended to a month or more
13 ay from making a cake, building a chicken coop, or producing a book, to found
14 , they decided, but a deck full of chicken coops and pigpens was hardly suita
15 im. "Johnny insisted on cooking a chicken dinner in my honor- he's always bee
16 nutes. Kid Ory, the trombonist chicken farmer, is also one of the solid a
17 y Johnson reaching around the wire chicken fencing, which half covered the tr
18 yes glittering behind dull silver chicken fencing. "That was Tee-wah I was t
19 wine in the pot roast or that the chicken had been marinated in brandy, and
20 yed this same game and called it "Chicken". He could not go through the f
21 f the Mexicans hiding in a little chicken house had passed through his head,
22 I'll never forget him cleaning the chicken in the tub". A story, no doubt
23 . Organ meats such as beef and chicken liver, tongue and heart are planne
24 p. "Miss Sarah, I can't cut up no chicken. Miss Maude say she won't". Aga
25 pot. "What is it"? he asked. "Chicken", Mose said, and theatrically licke
26 im"? Adam shook his head. "Chicken", Mose said. She was a child too m

What is "chicken" ?

Recurrent Neural Network (Language Model)



Flashback: Markov Models in Retrospect

Consider a sequence of random variables X_1, X_2, \dots, X_n , each take any value in \mathcal{V}

The joint probability of a sentence is

$$\begin{aligned} & P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}) \end{aligned}$$



First-order Markov Assumption

$$P(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1})$$

Is it possible to directly model this probability?

Flashback: Supervised Learning

How does this happen?

They all sampled from a distribution!



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

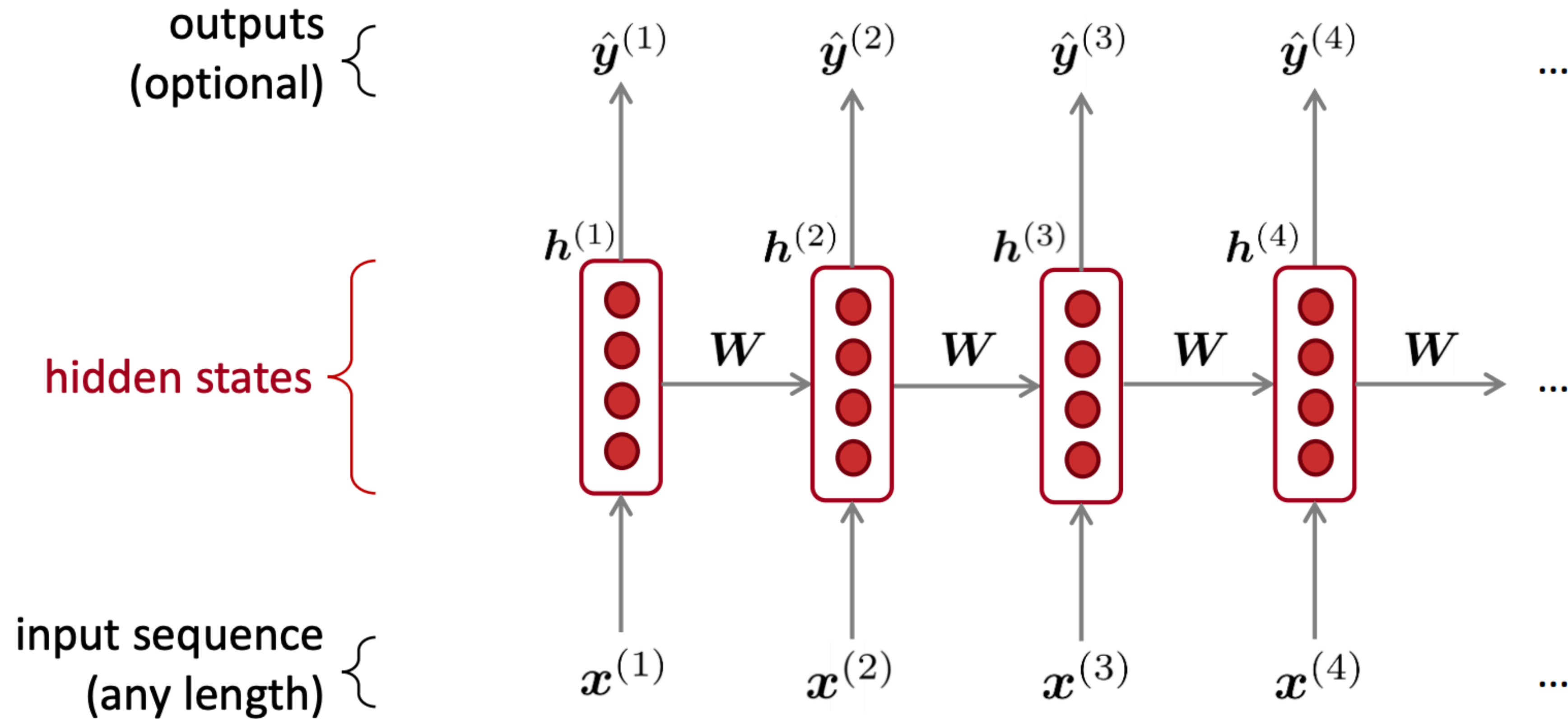
$p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$ Model / Function

$$p(\mathcal{D}_{\text{train}} | \boldsymbol{\theta}) = \prod_{n=1}^{N_{\text{train}}} p(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, \boldsymbol{\theta})$$

Objective / Loss

Learning / Training $\longrightarrow \hat{\boldsymbol{\theta}}$

Recurrent Neural Networks (RNNs)



$$p(x^{(t)} \mid x^{(t-1)}, \dots, x^{(1)})$$

A RNN Language Model

output distribution

$$\hat{y}^{(t)} = \text{softmax}(U\mathbf{h}^{(t)} + \mathbf{b}_2) \in \mathbb{R}^{|V|}$$

hidden states

$$\mathbf{h}^{(t)} = \sigma(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_e \mathbf{e}^{(t)} + \mathbf{b}_1)$$

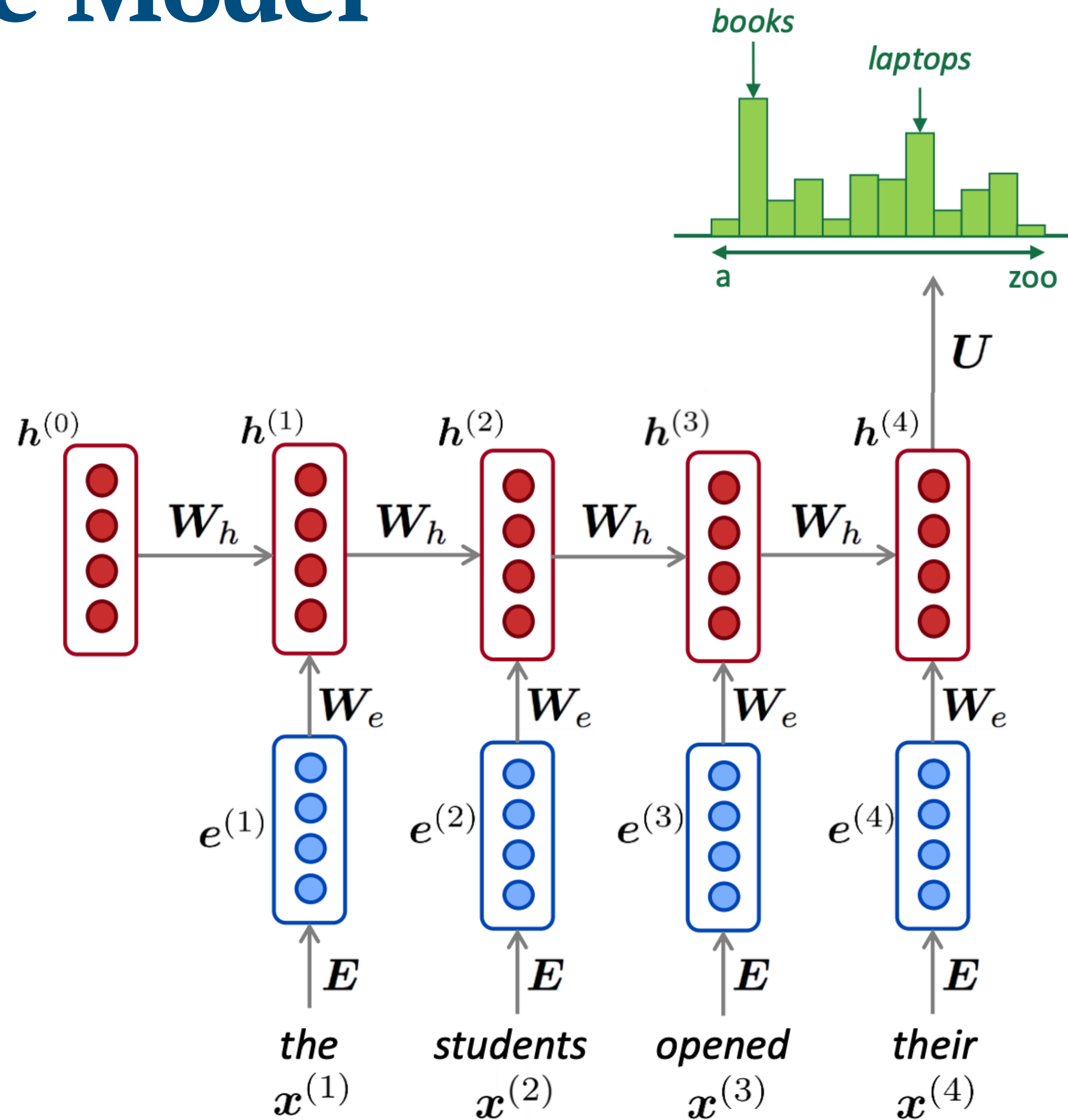
$\mathbf{h}^{(0)}$ is the initial hidden state

word embeddings

$$\mathbf{e}^{(t)} = \mathbf{E} \mathbf{x}^{(t)}$$

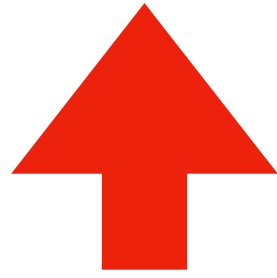
words / one-hot vectors

$$\mathbf{x}^{(t)} \in \mathbb{R}^{|V|}$$



A RNN Language Model

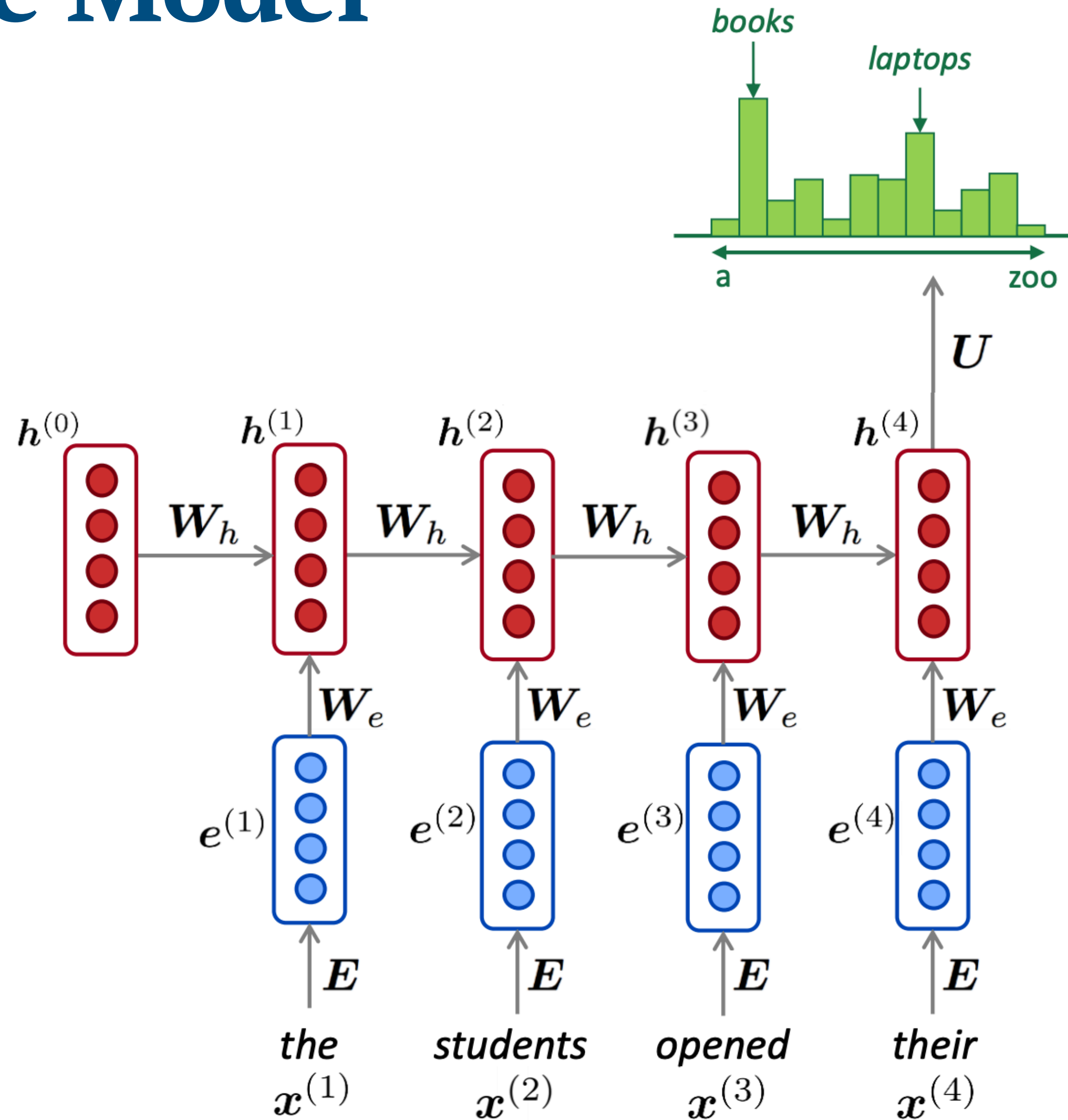
$$p(x^{(t)} \mid x^{(t-1)}, \dots, x^{(1)})$$



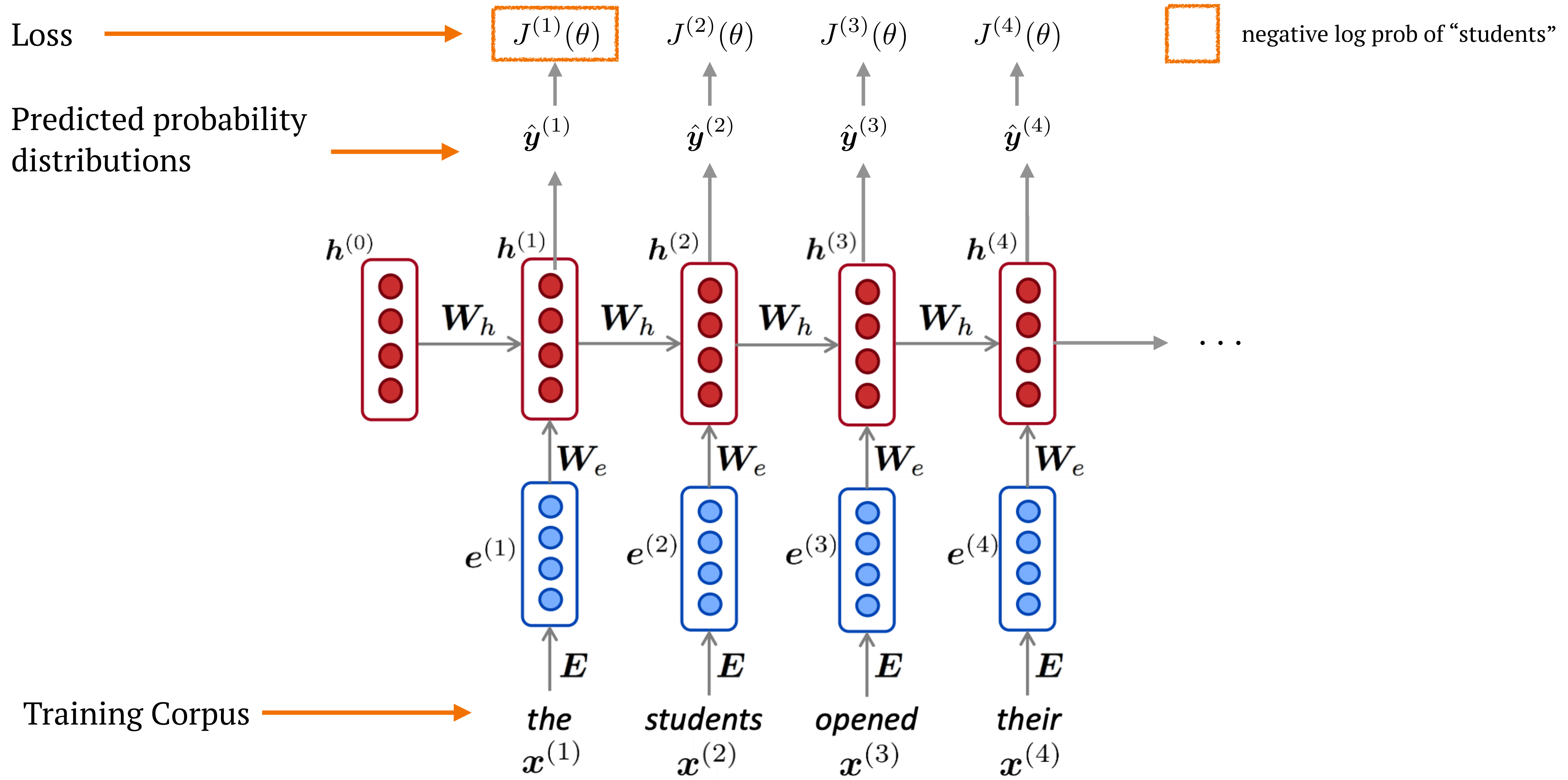
$$\hat{y}^{(t)} = \text{softmax}(U h^{(t)} + b_2) \in \mathbb{R}^{|V|}$$
$$h^{(t)} = \sigma(W_h h^{(t-1)} + W_e e^{(t)} + b_1)$$

$h^{(0)}$ is the initial hidden state

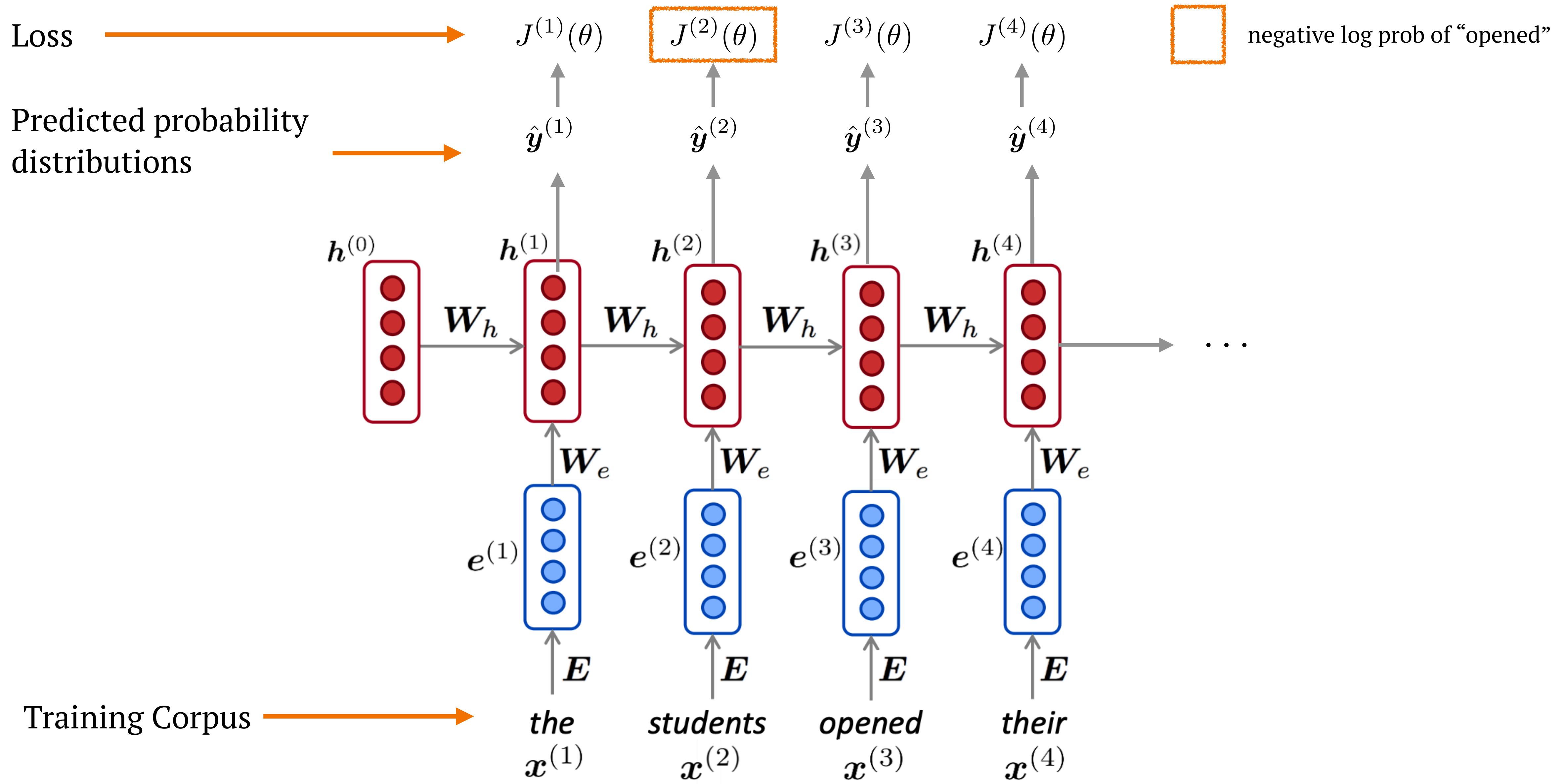
The recurrent function here, takes into consideration **all** the history!



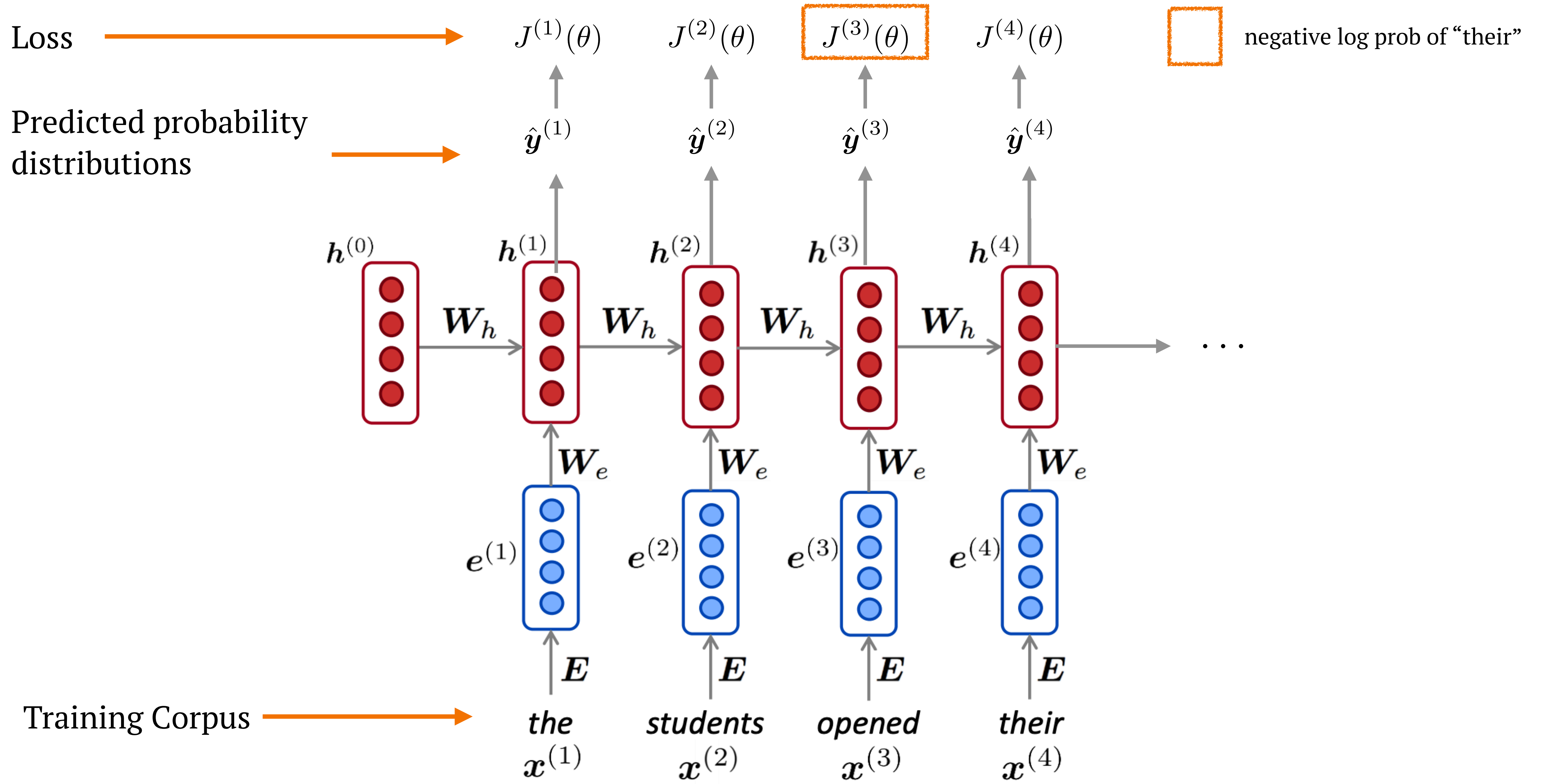
Training a RNN Language Model



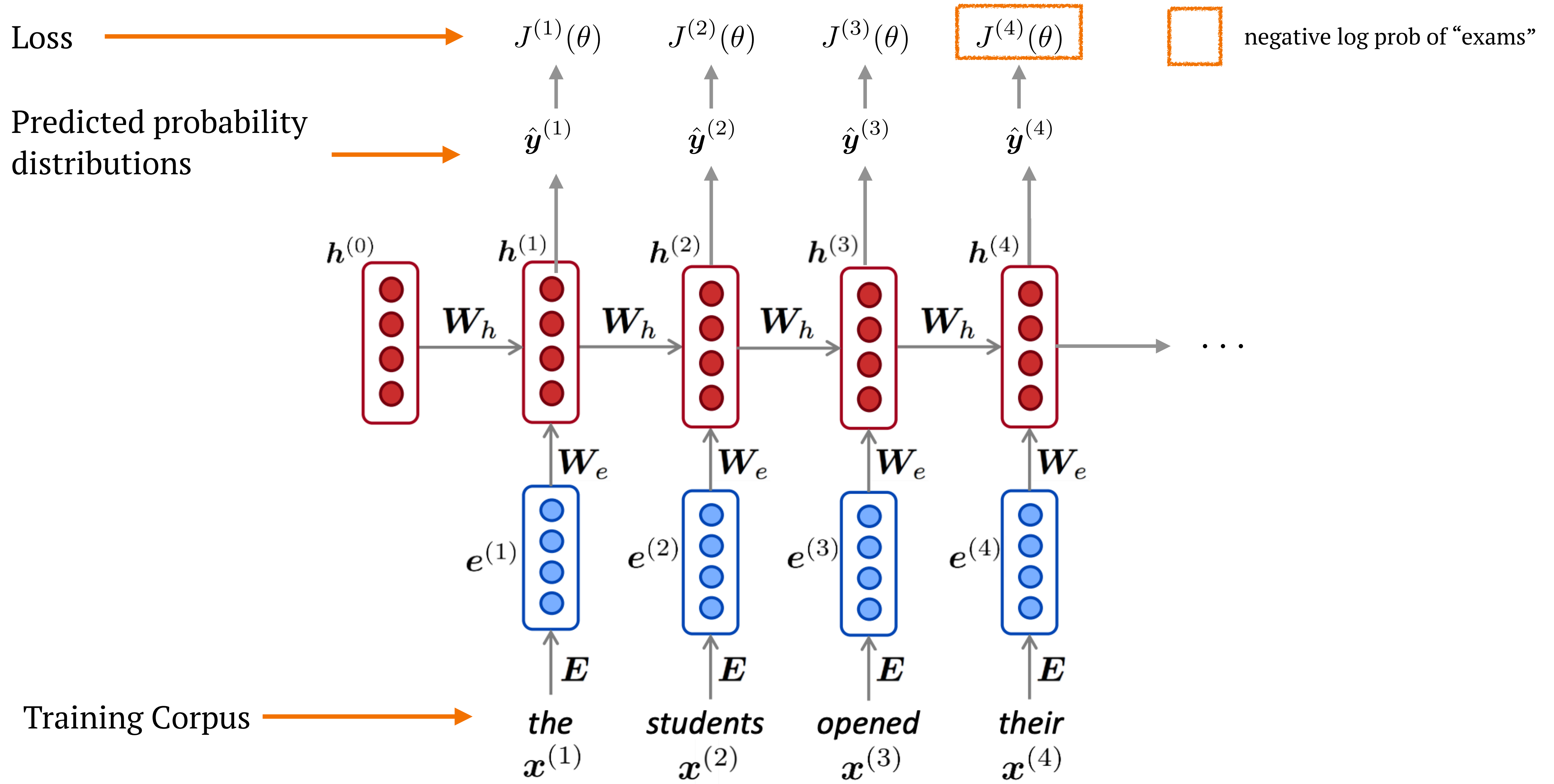
Training a RNN Language Model



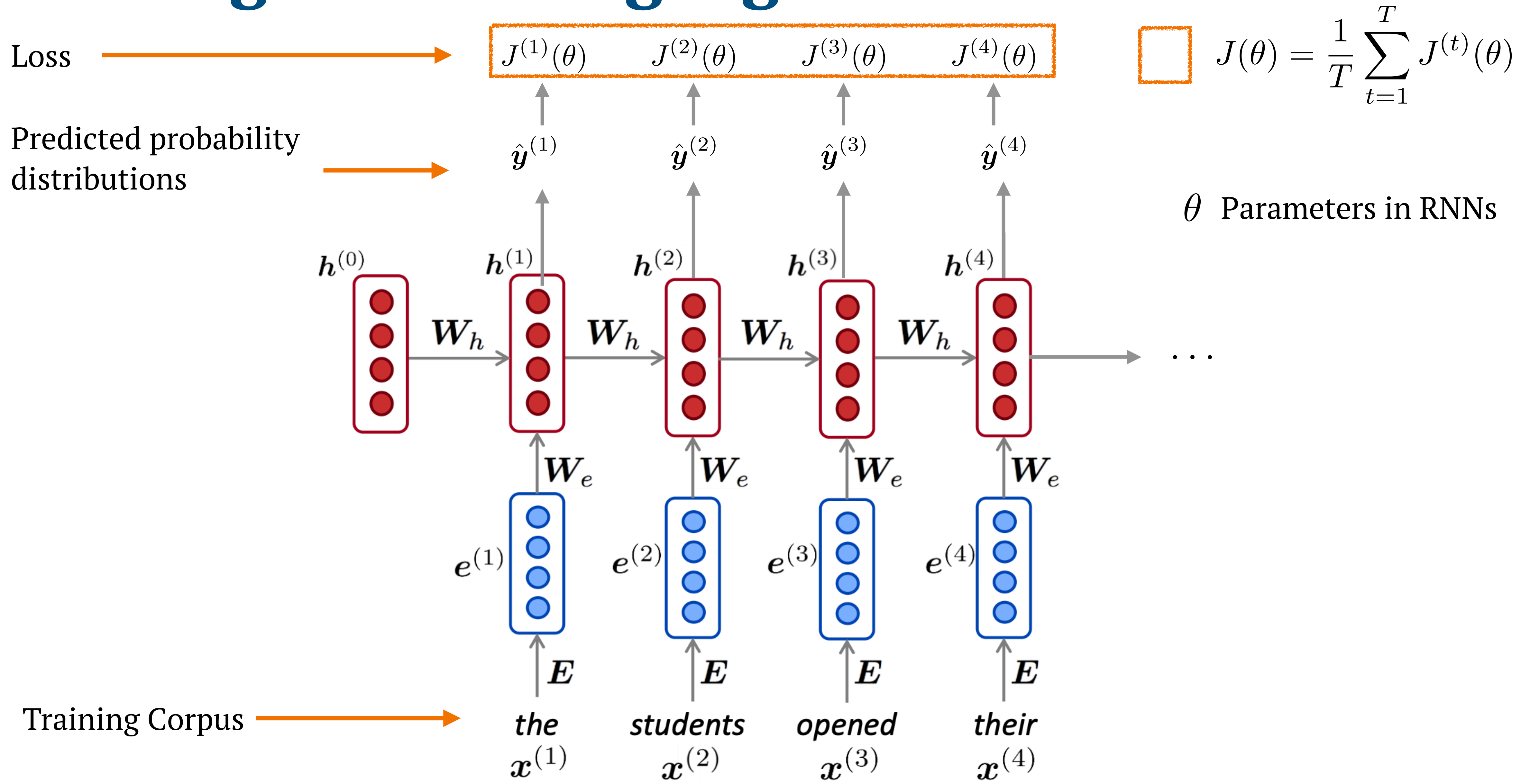
Training a RNN Language Model



Training a RNN Language Model



Training a RNN Language Model



GPT-3 Model

Loss



Predicted probability distributions



$$J^{(1)}(\theta) \quad J^{(2)}(\theta) \quad J^{(3)}(\theta) \quad J^{(4)}(\theta)$$

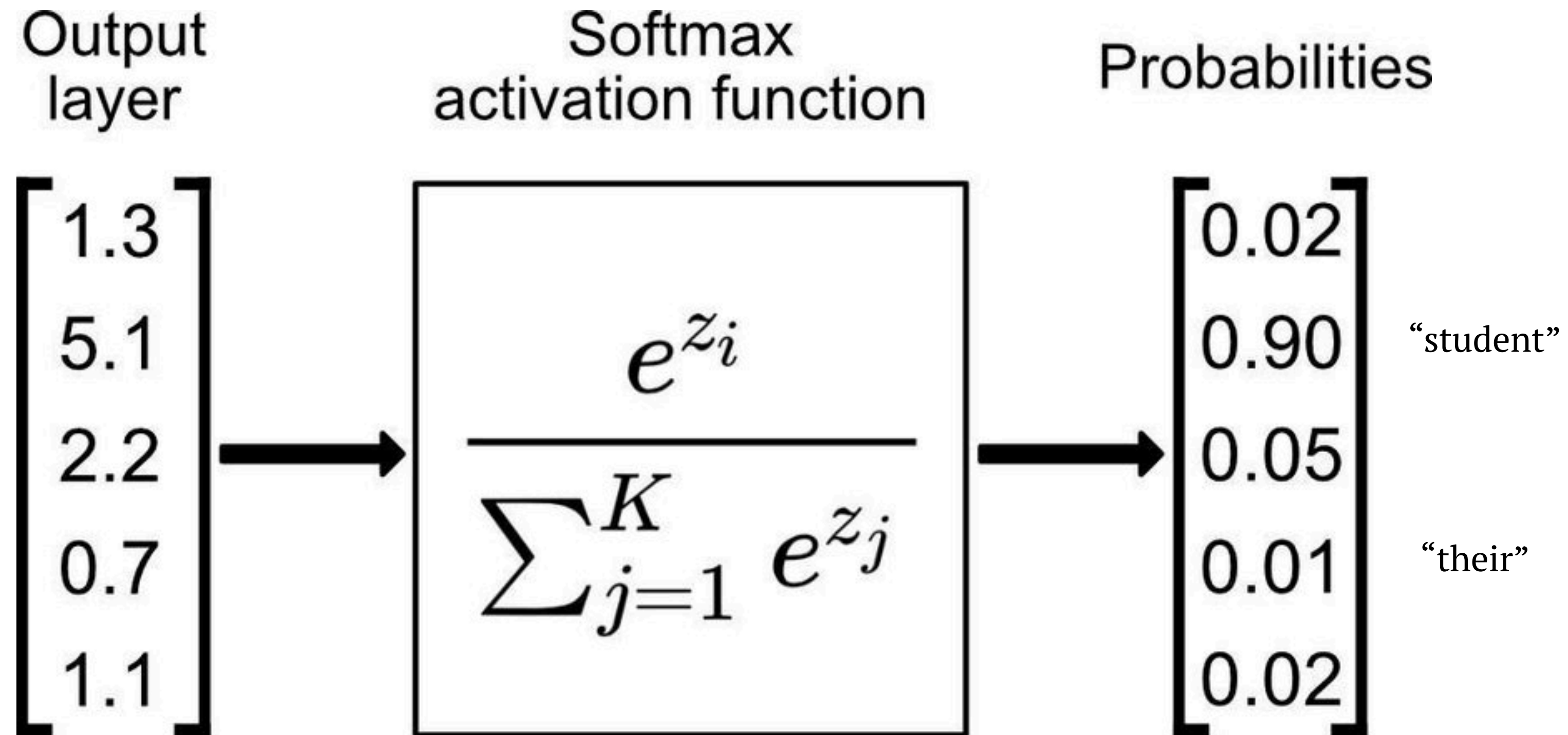
$$\hat{y}^{(1)} \quad \hat{y}^{(2)} \quad \hat{y}^{(3)} \quad \hat{y}^{(4)}$$




$$\square J(\theta) = \frac{1}{T} \sum_{t=1}^T J^{(t)}(\theta)$$

θ Parameters in Transformer

Softmax Function


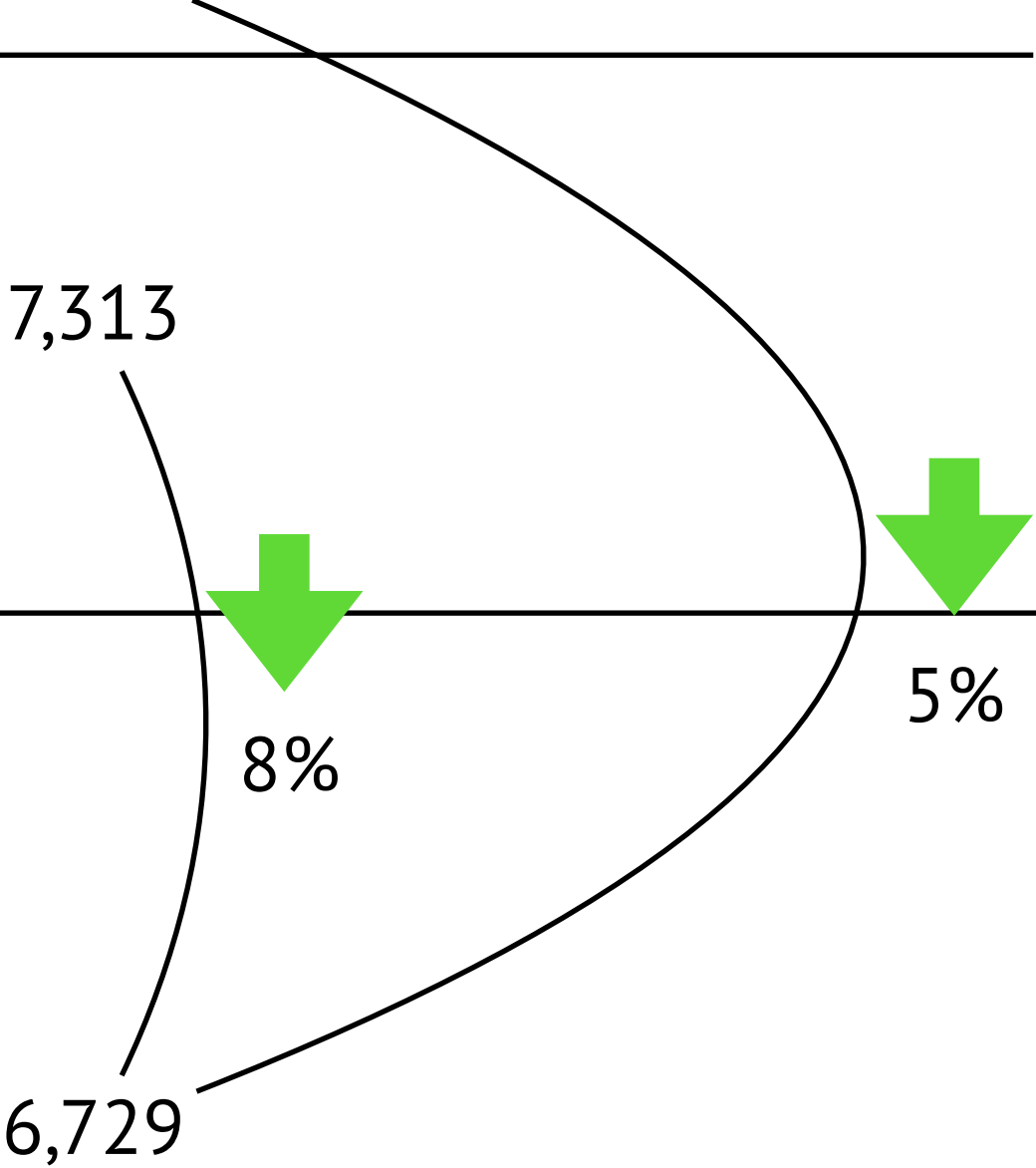


Typical NLP Task: Predicting gross revenues of movies

Models	Mean Absolute Error (\$)
Baseline: Predict median from training data	7,079
Metadata (D): U.S. origin? , log budget, # screens, runtime <i>name, production house, genre(s), scriptwriter(s), director(s), country of origin, primary actors, release date, MPAA rating, and running time</i>	7,313
Text (T): Movie Reviews (from only before the release date) <div style="display: flex; align-items: flex-start; margin-top: 10px;"> <div style="flex: 1;">  </div> <div style="flex: 2;"> <p>70 The New York Times Elvis Mitchell It becomes less crisp on screen than it was on the page, with much of the enjoyable jargon either mumbled confusingly or otherwise thrown away. [11 June 1993, p.C1]</p> <p>67 THE AUSTIN CHRONICLE Marc Savlov I continually found myself longing for the sheer intensity of the director's past glories, like Jaws, or even Duel. Spielberg seems to be trying so very hard for that elusive "Gosh, Wow, Sense of Wonder!" that it all looks strained in spots. Read full review</p> </div> <div style="flex: 1; margin-left: 20px;"> <p><i>Words, bigrams, trigrams, and dependency relations</i></p> </div> </div>	6,729
Metadata (D) + Text (T)	6,725

(Smith, 2010)

Typical NLP Task: Predicting gross revenues of movies

Models	Mean Absolute Error (\$)
Baseline: Predict median from training data	7,079
Metadata (D): U.S. origin? , log budget, # screens, runtime <i>name, production house, genre(s), scriptwriter(s), director(s), country of origin, primary actors, release date, MPAA rating, and running time</i>	7,313
Text (T): Movie Reviews (from only before the release date) <div style="display: flex; align-items: flex-start; margin-top: 10px;"> <div style="flex: 1;">  <p>70 The New York Times Elvis Mitchell It becomes less crisp on screen than it was on the page, with much of the enjoyable jargon either mumbled confusingly or otherwise thrown away. [11 June 1993, p.C1]</p> <hr/> <p>67 THE AUSTIN CHRONICLE Marc Savlov I continually found myself longing for the sheer intensity of the director's past glories, like Jaws, or even Duel. Spielberg seems to be trying so very hard for that elusive "Gosh, Wow, Sense of Wonder!" that it all looks strained in spots. Read full review</p> </div> <div style="flex: 1; padding-left: 20px; color: red;"> <p><i>Words, bigrams, trigrams, and dependency relations</i></p> </div> </div>	6,729 8% 5% 
Metadata (D) + Text (T)	6,725

Models

Linear Regression:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^n (M_i - \mathbf{w}^\top \underbrace{\mathbf{f}_i}_{\text{depends on } D_i, T_i, \text{ or both}})^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2$$

↑
“elastic net” regularization
(Zou and Hastie, 2005; Friedman et al., 2008)

Features

Jurassic Park lacks the emotional unity of Spielberg's classics .

Words:

Jurassic
Park
lacks
the
emotional
unity
of
Spielberg's
classics
.

Bigrams:

Jurassic Park
Park lacks
Lacks the
the emotional
emotional unity
unity of
of Spielberg's
Spielberg's classics
classics .
.<eos>

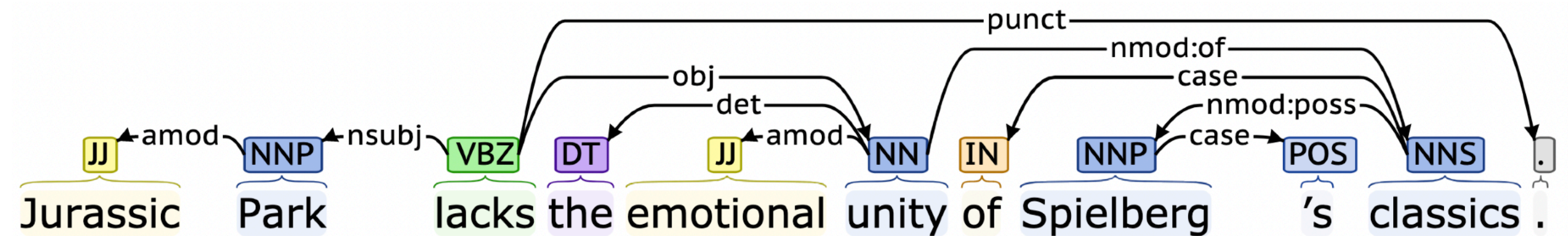
Part-of-speech tags:

JJ
NNP
VBZ
DT
JJ
NN
IN
NNP
POS
NNS
.

Named Entities:

Movie: Jurassic
Park
Person: Spielberg

Dependencies (Syntax Parsing):

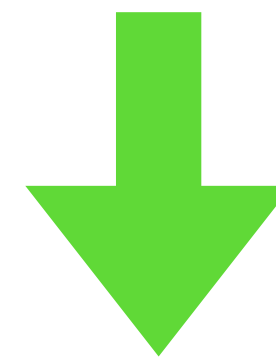


Bag-of-words Models

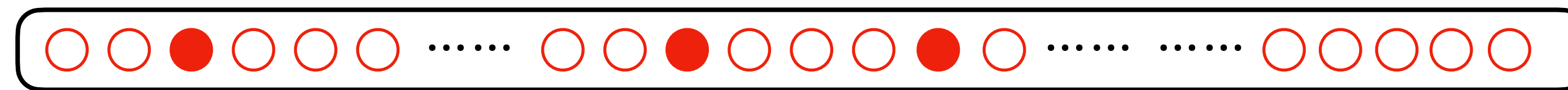
Jurassic Park lacks the emotional unity of Spielberg's classics .

Words:

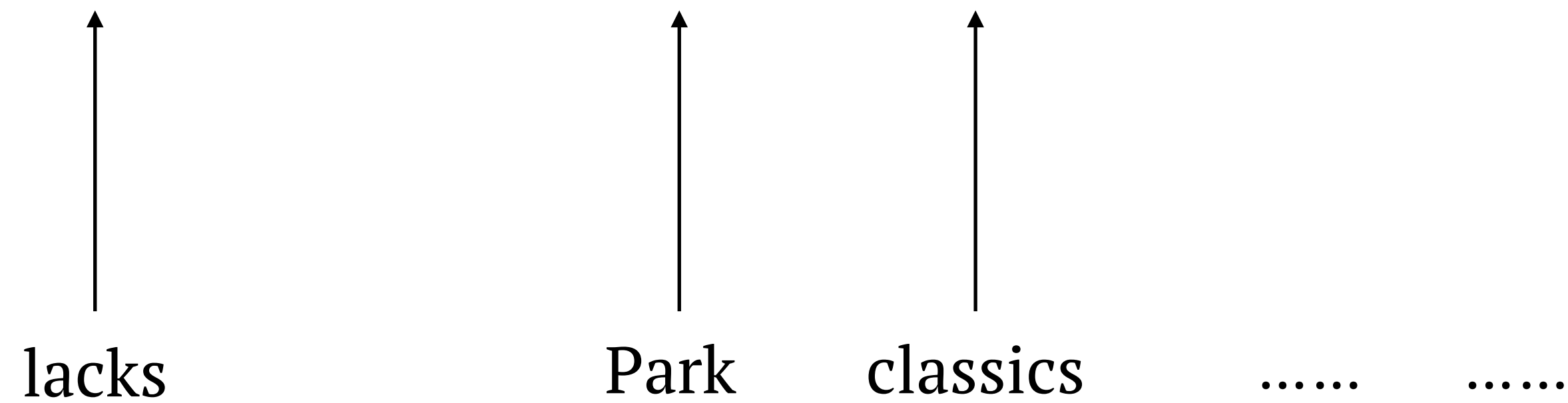
Jurassic
Park
lacks
the
emotional
unity
of
Spielberg's
classics



featurized



Full Vocabulary



Weights Vector
(learned)

.

Natural Language Processing (NLP) Pipeline

General-purpose linguistic modules:

Words Bigrams

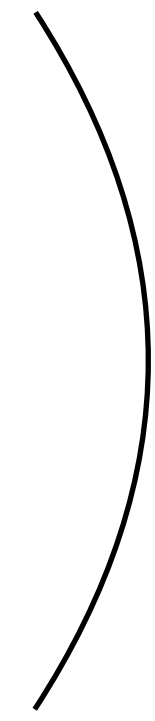


Light preprocessing (mostly rule-based)

Part-of-speech tags: word classes

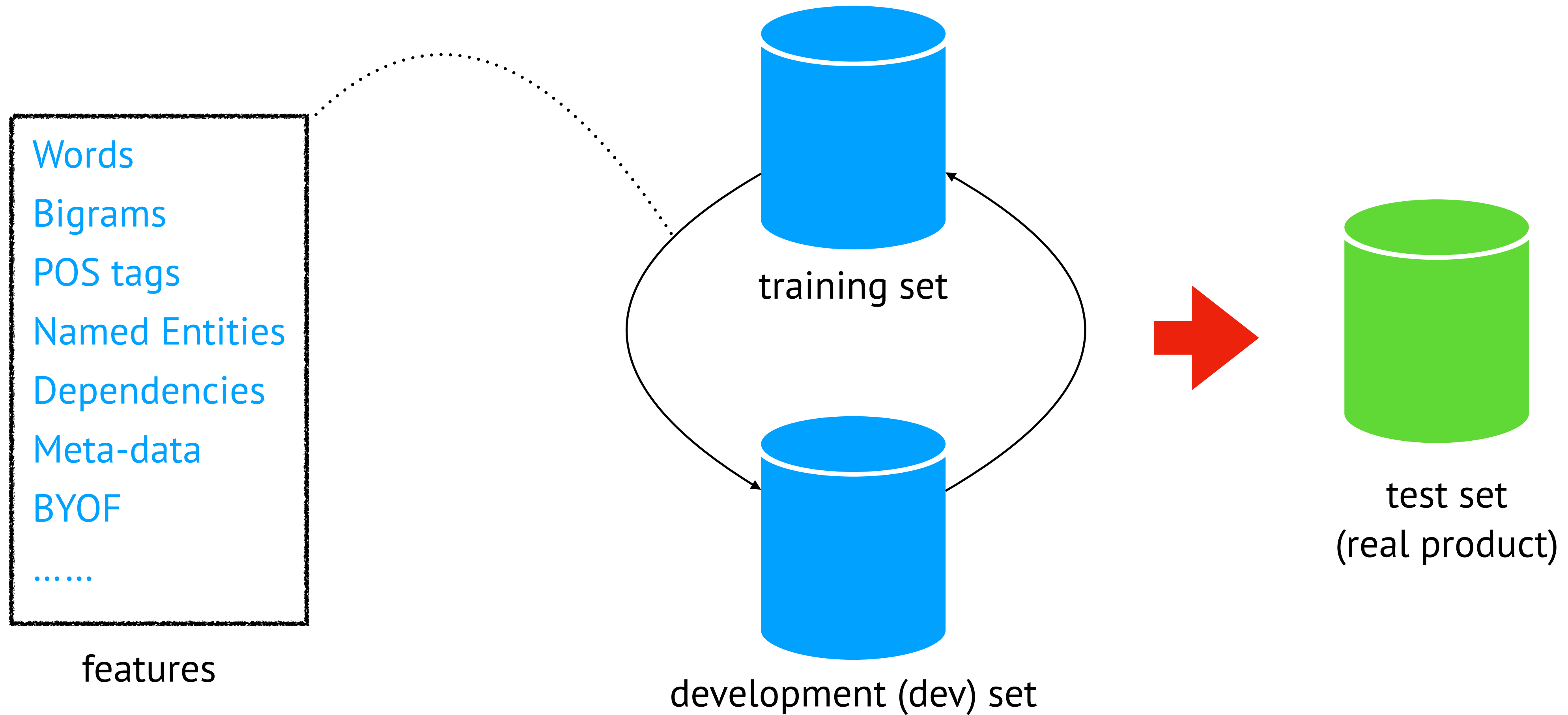
Named Entities: words of interests

Dependencies (Syntax Parsing): Internal structures

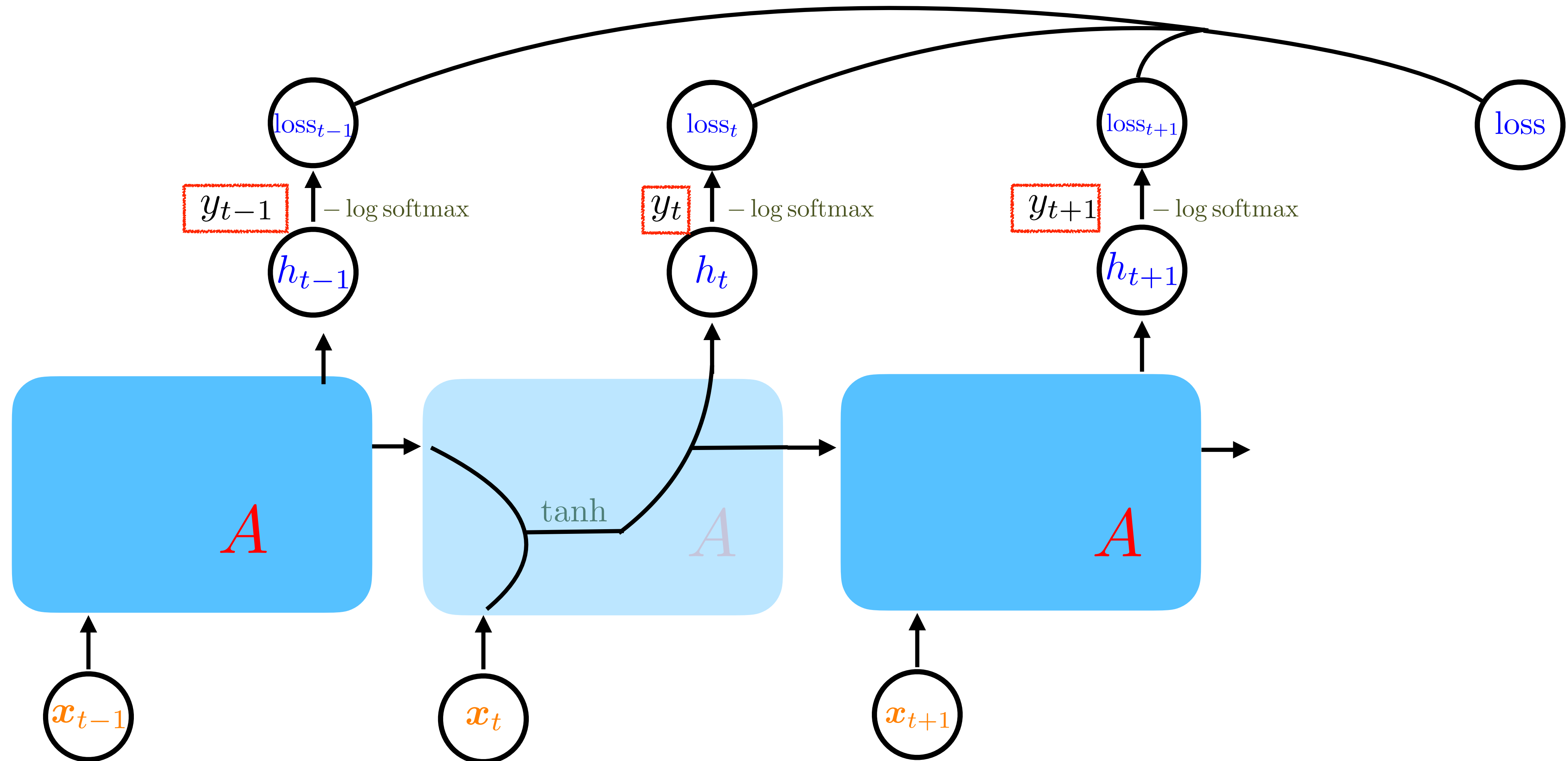


Supervised learning from linguistic data
(CoreNLP pipeline)

Feature Engineering



RNNs for Tagging



Part-of-Speech Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/**N** soared/**V** at/**P** Boeing/**N** Co./**N** ,/, easily/**ADV** topping/**V** forecasts/**N** on/**P** Wall/**N** Street/**N** ,/, as/**P** their/**POSS** CEO/**N** Alan/**N** Mulally/**N** announced/**V** first/**ADJ** quarter/**N** results/**N** ./.

N = Noun

V = Verb

P = Preposition

Adv = Adverb

Adj = Adjective

...

Named Entity Recognition (NER)

INPUT: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT: Profits soared at [Company Boeing Co.], easily topping forecasts on [Location Wall Street], as their CEO [Person Alan Mulally] announced first quarter results.

Named Entity Recognition (NER)

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

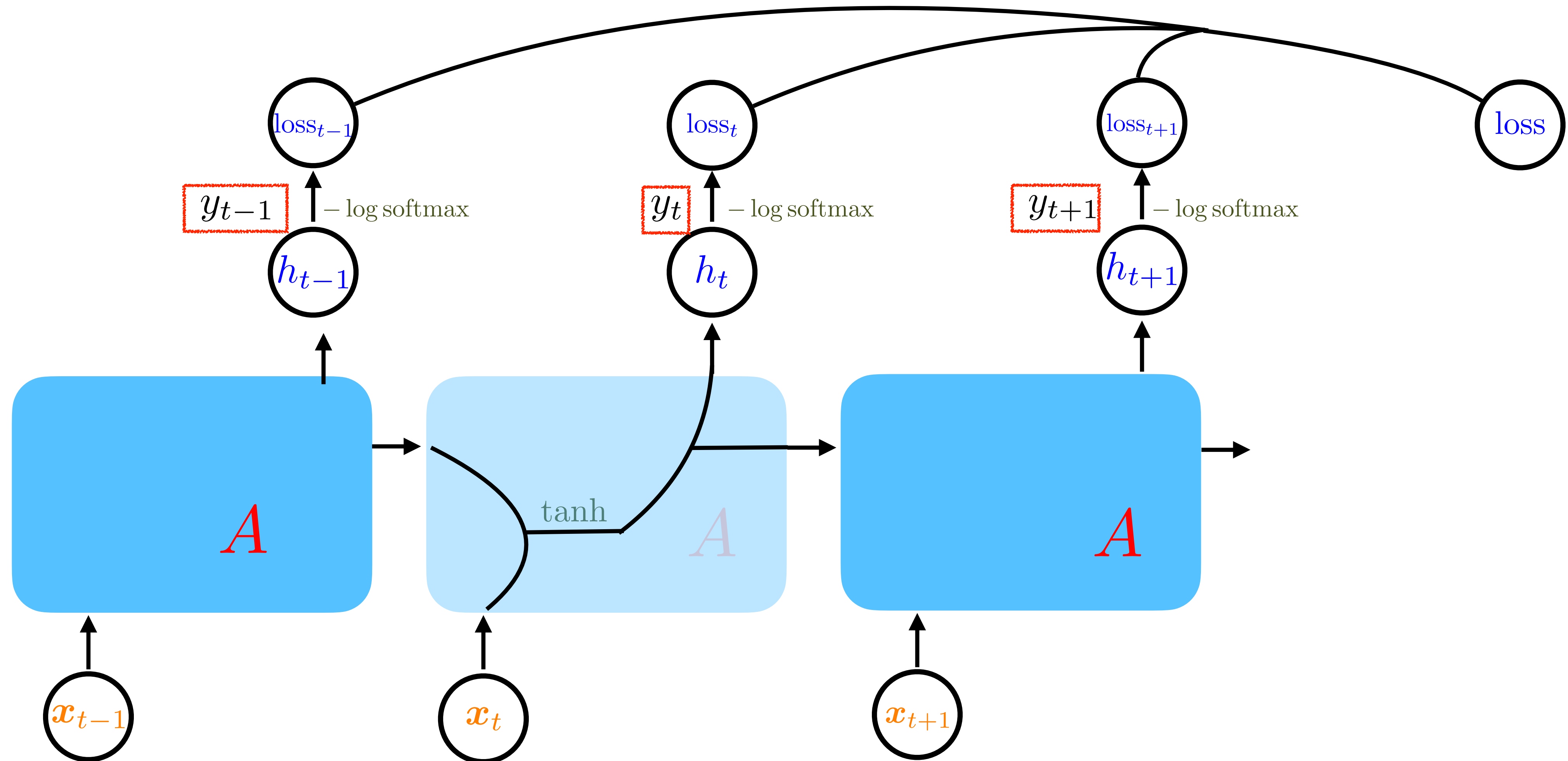
OUTPUT:

Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA
topping/NA forecasts/NA on/NA Wall/SL Street/CL ,/NA as/NA
their/NA CEO/NA Alan/SP Mulally/CP announced/NA first/NA
quarter/NA results/NA ./NA

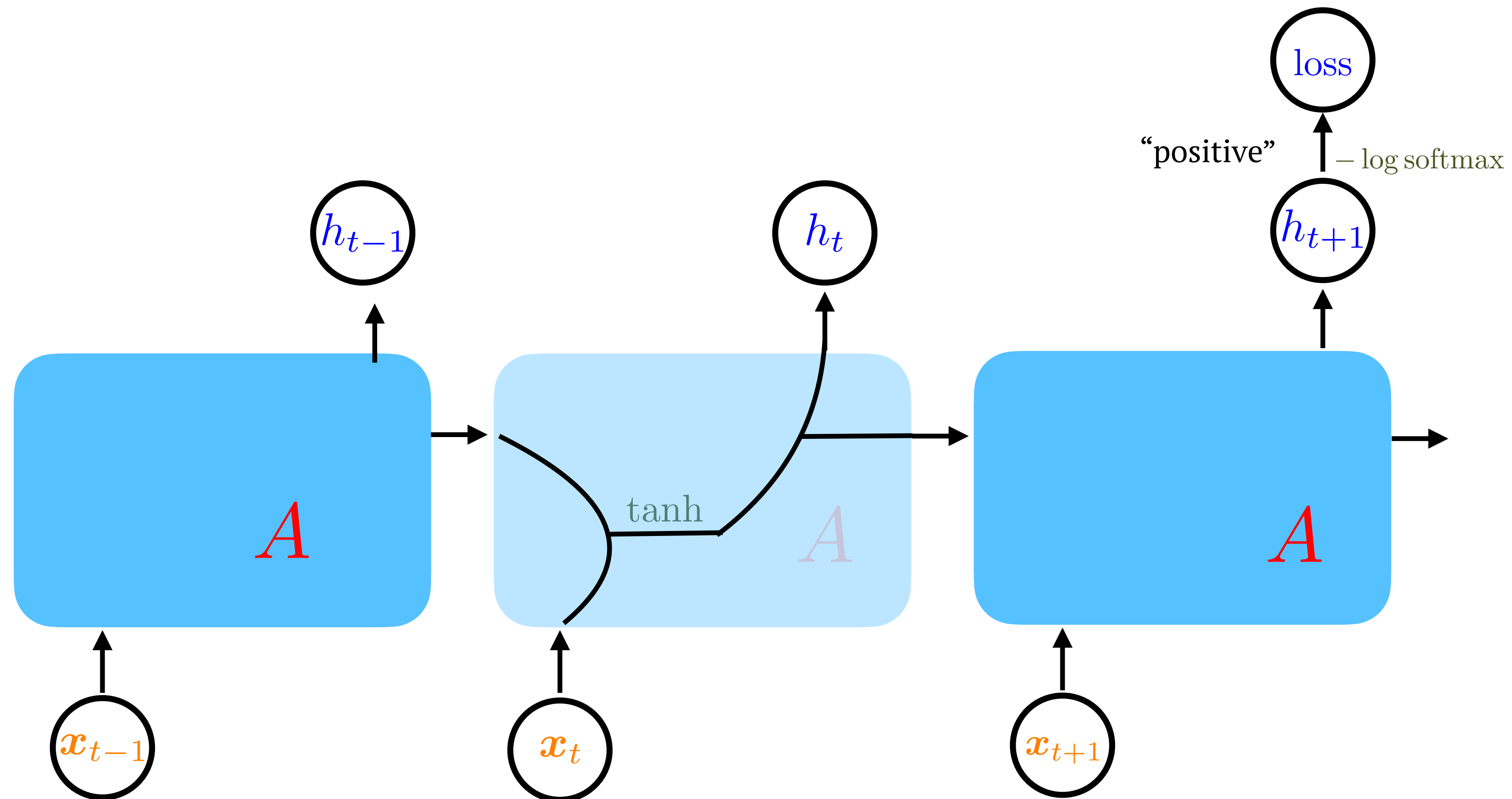
NA = No entity
SC = Start Company
CC = Continue Company
SL = Start Location
CL = Continue Location

....

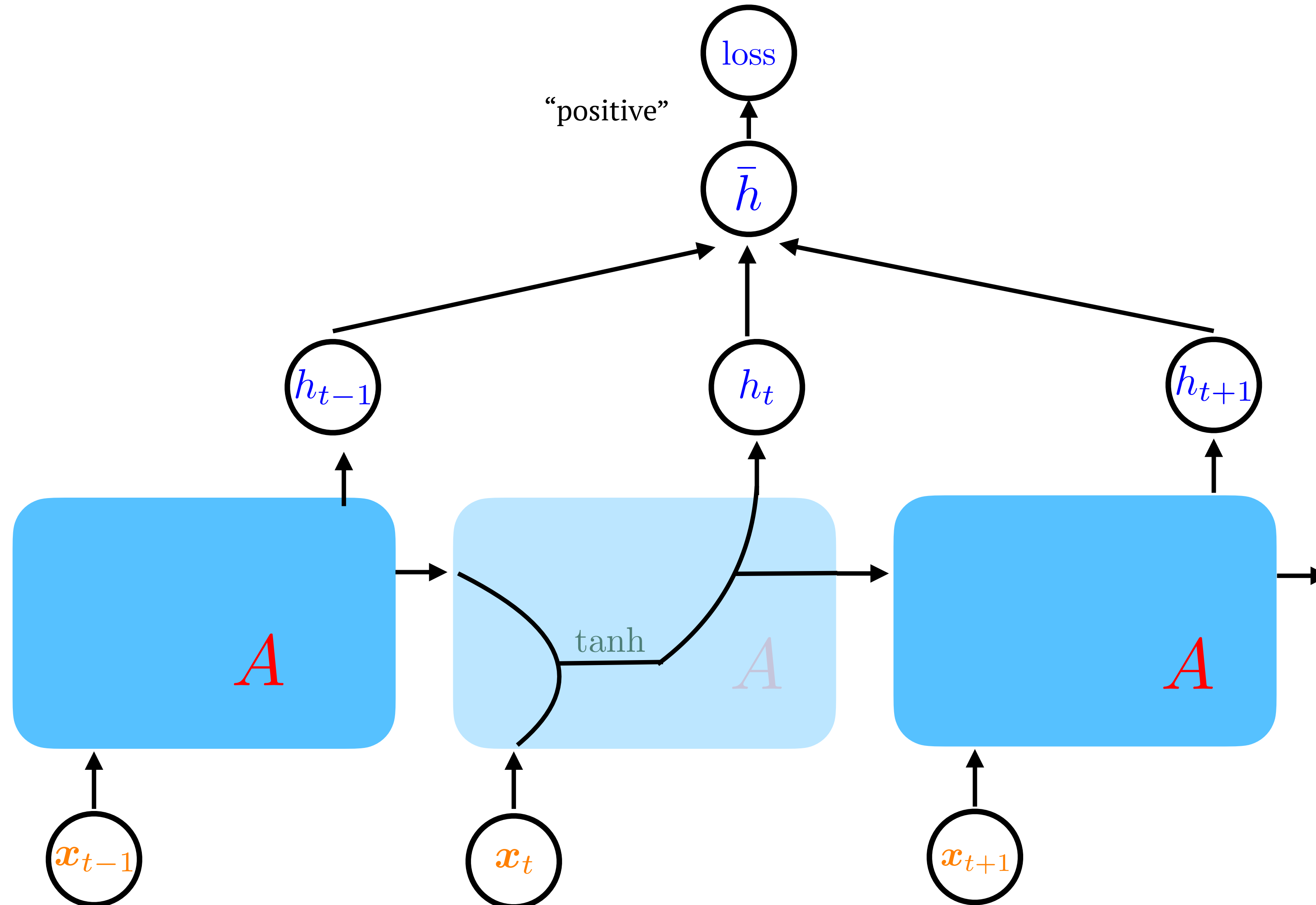
RNNs for Tagging



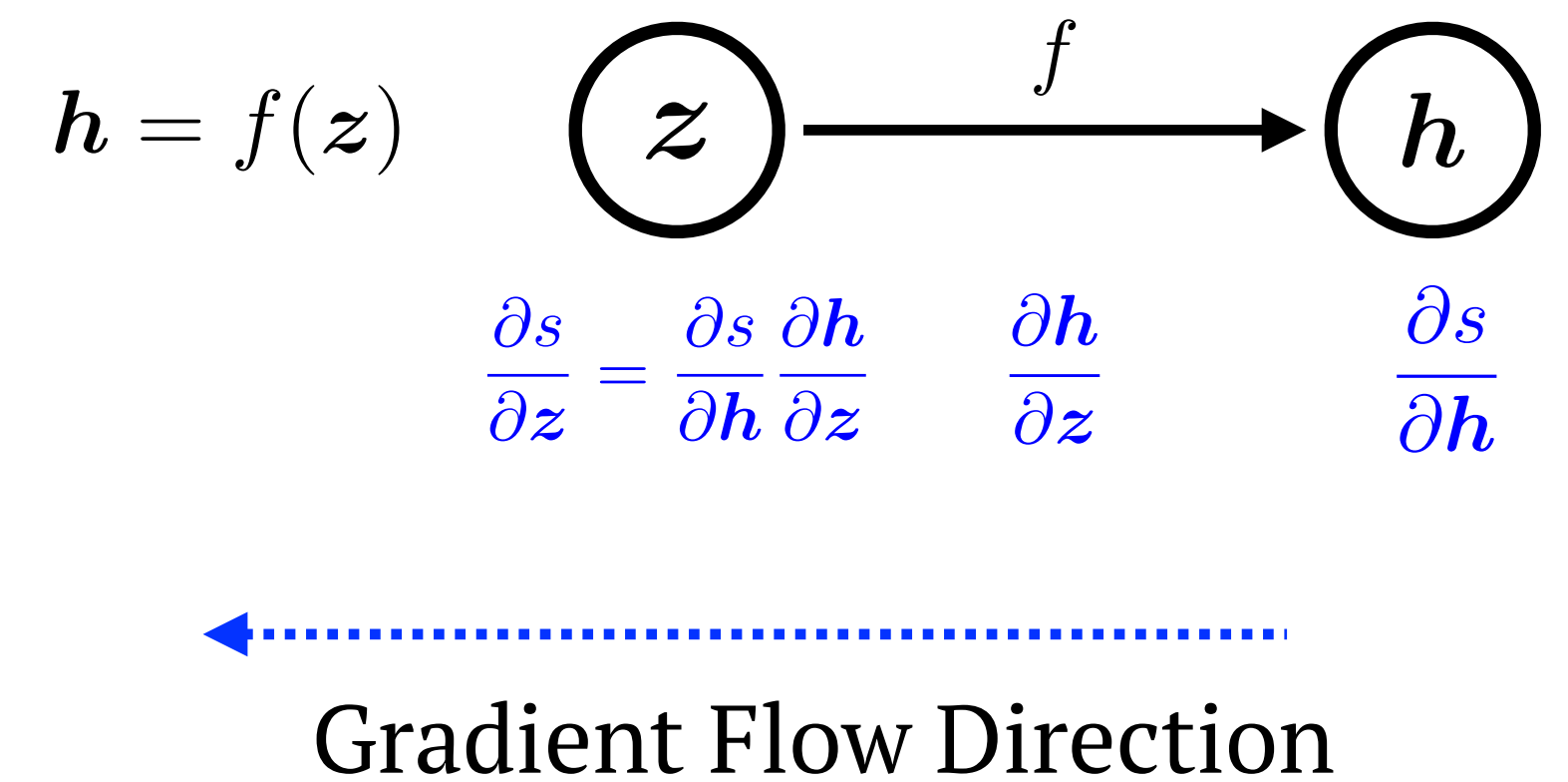
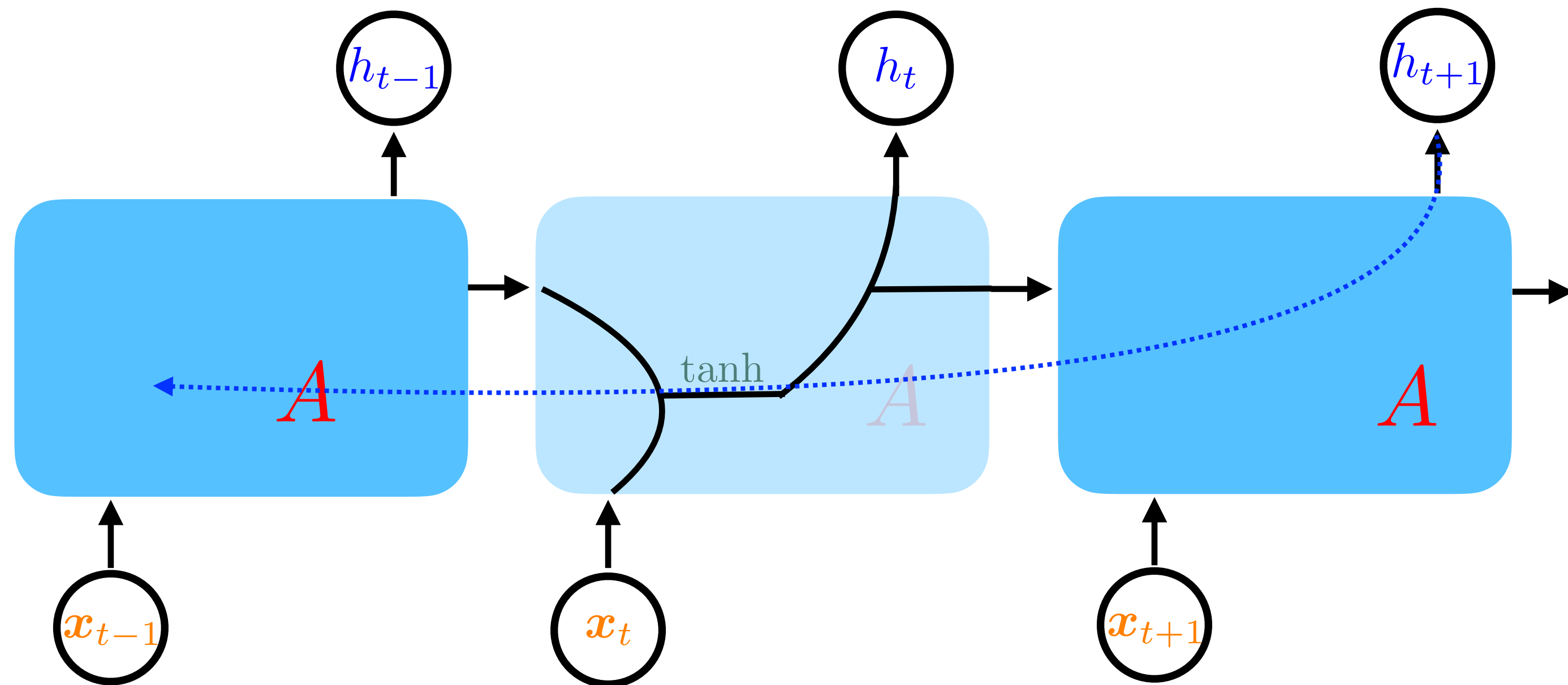
RNNs for Sentence Classification



RNNs for Sentence Classification



Vanishing Gradient in RNNs



In general, the longer the path, the smaller the gradient signal.

Long Short-Term Memory (LSTMs)

