The Computational Graphs / Pre-training and Fine-tuning

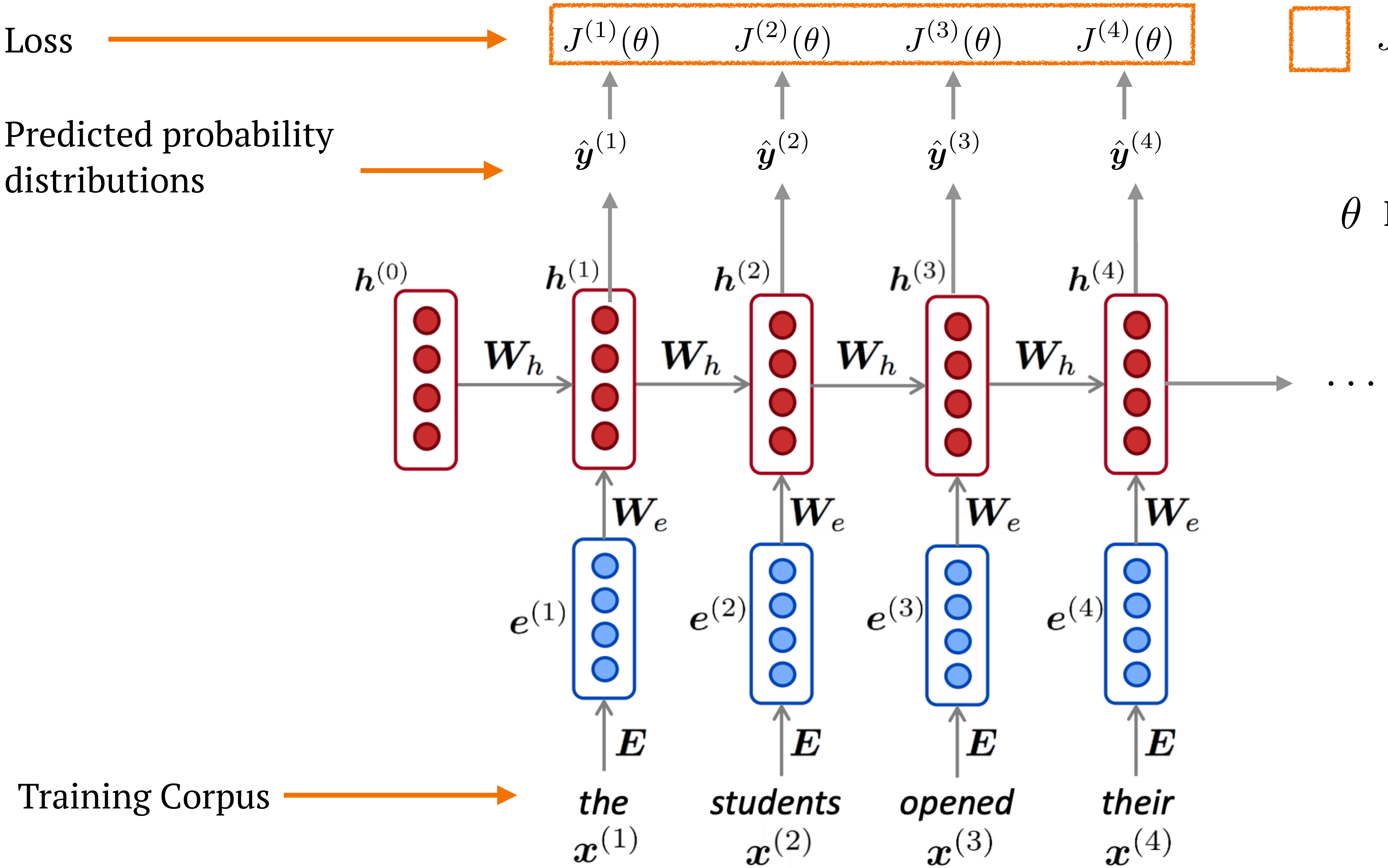
COMP7607— Week 2

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Some materials from Stanford University CS224n with special thanks!

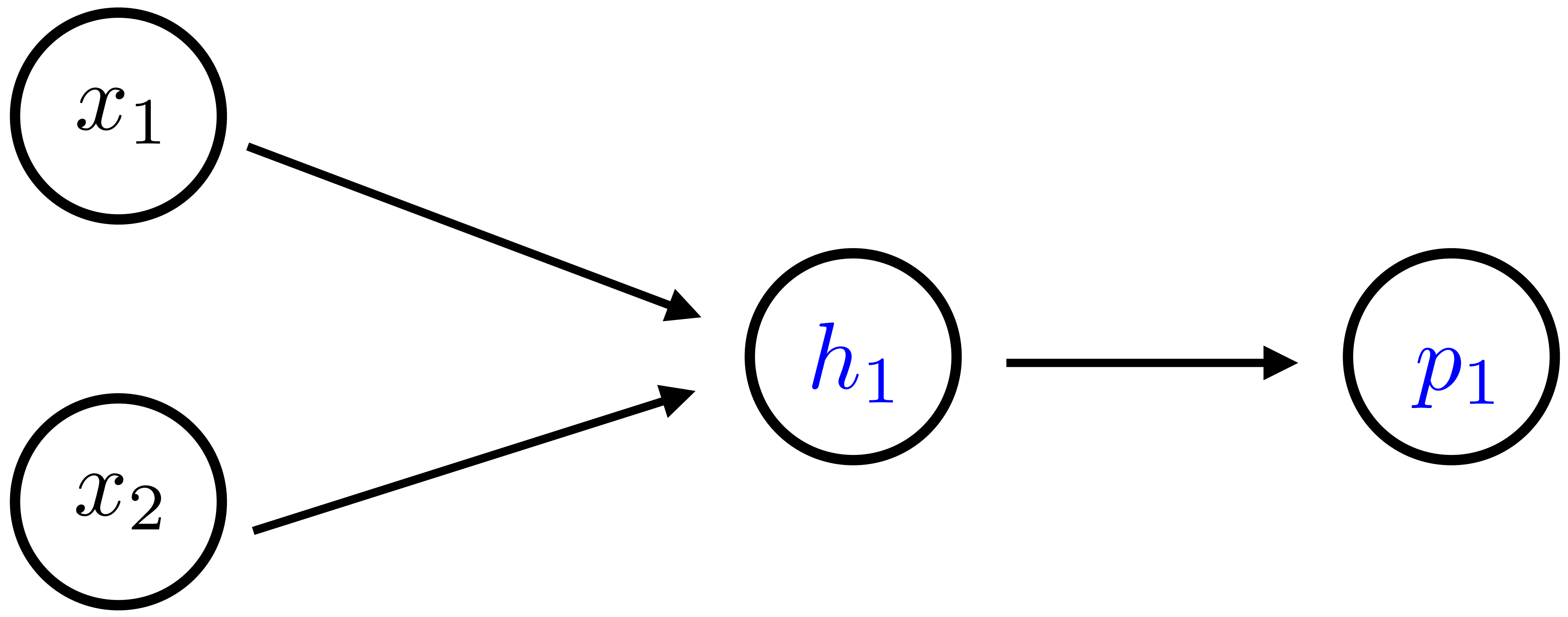
Flashback — Training a RNN Language Model



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

θ Parameters in RNNs

Logistic Regression

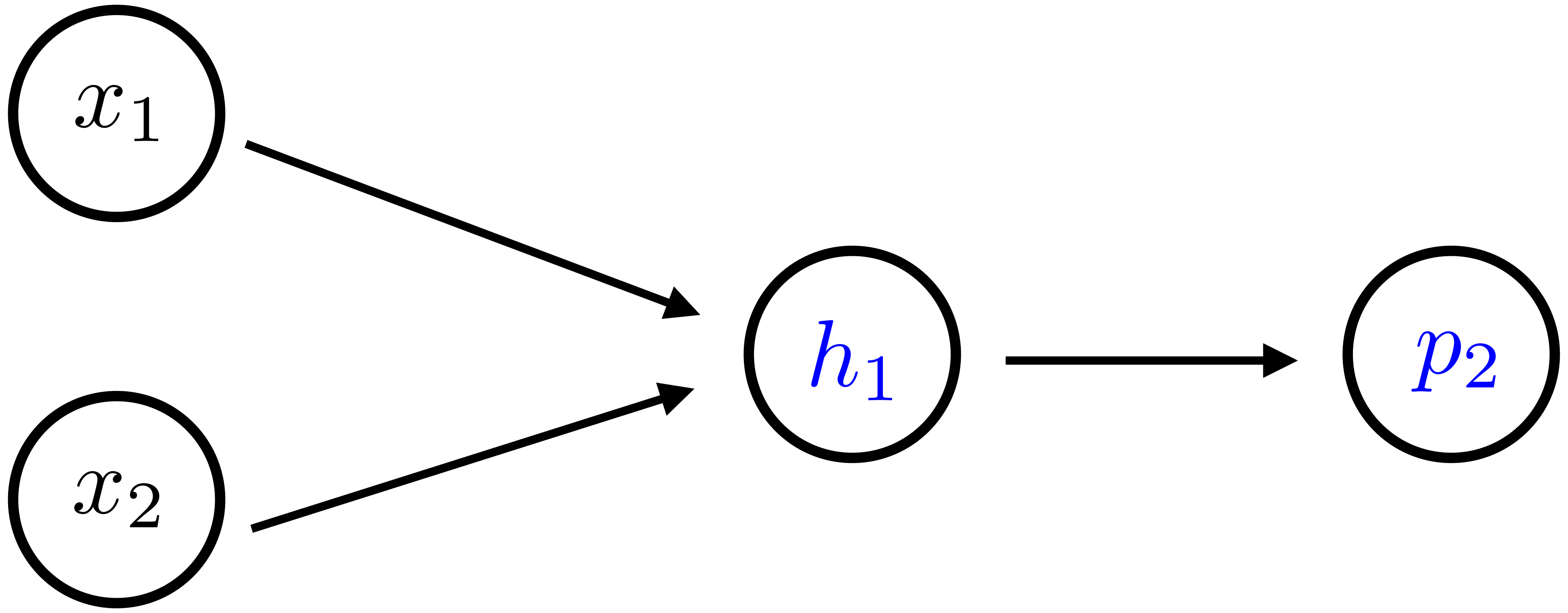


$$p_1 := p(y = 1 \mid x_1, x_2)$$

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

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

Logistic Regression



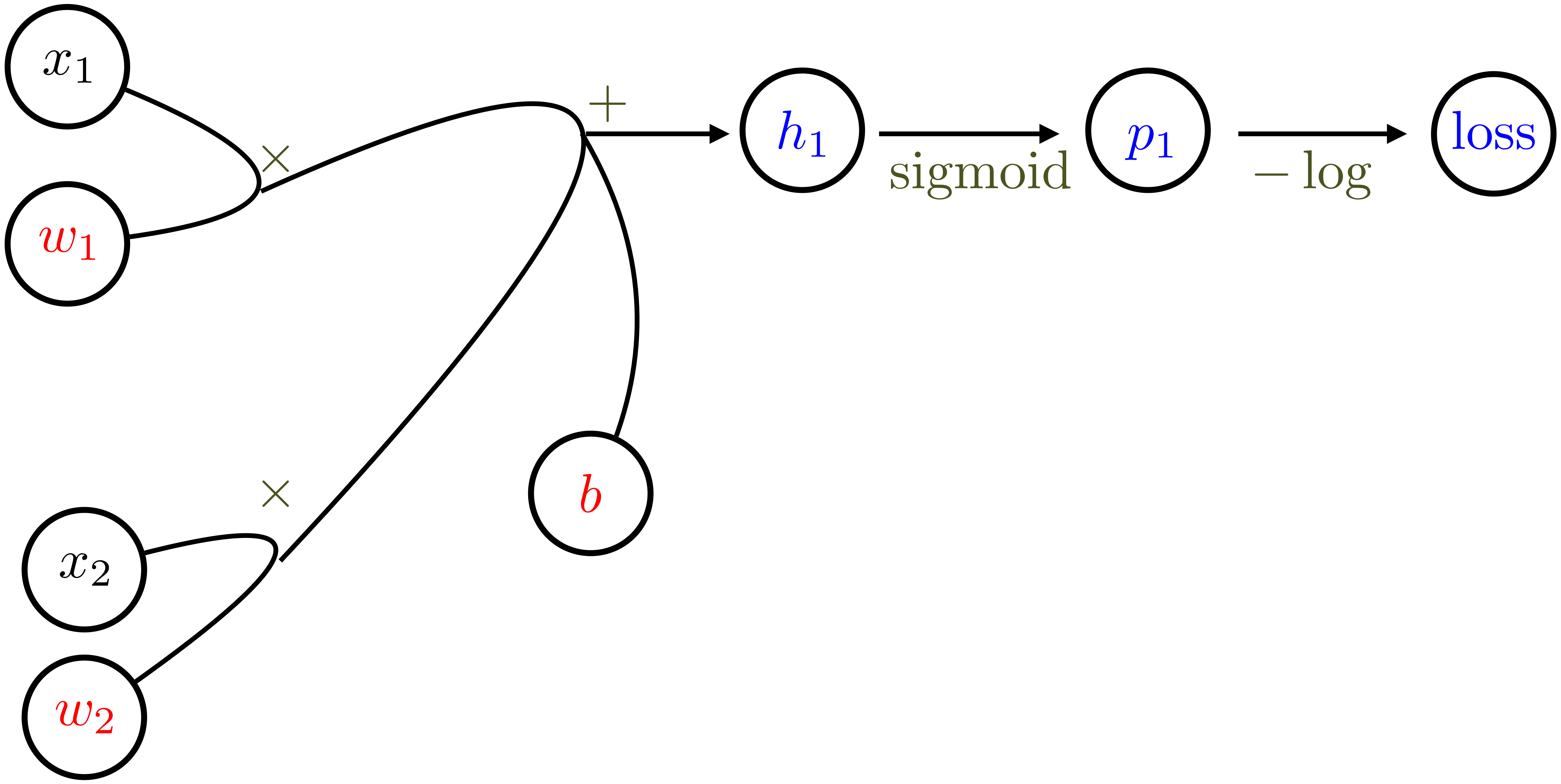
$$p_2 := p(y = 0 \mid x_1, x_2)$$

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

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

Loss Function

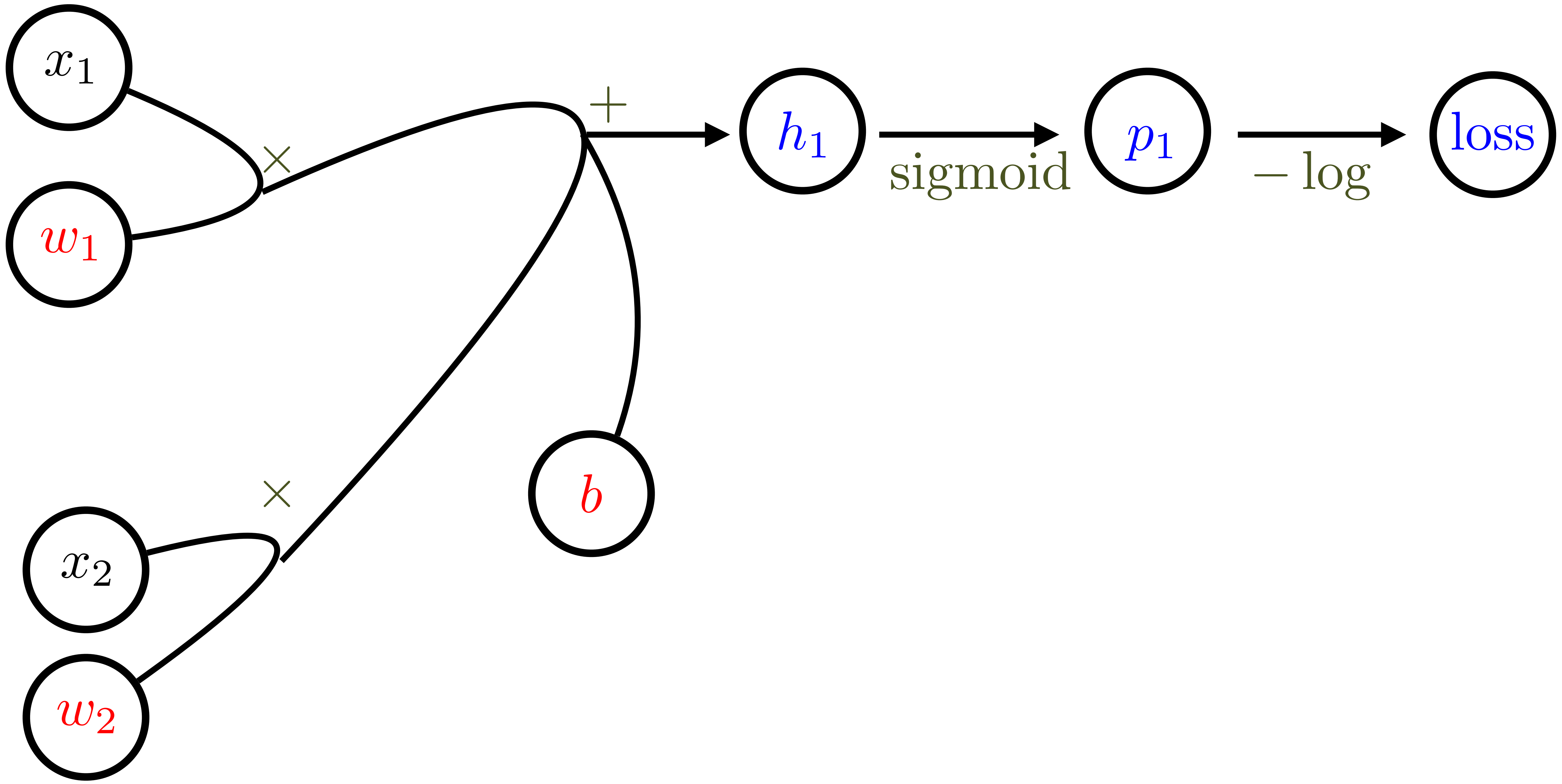
case $y = 1$:



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)

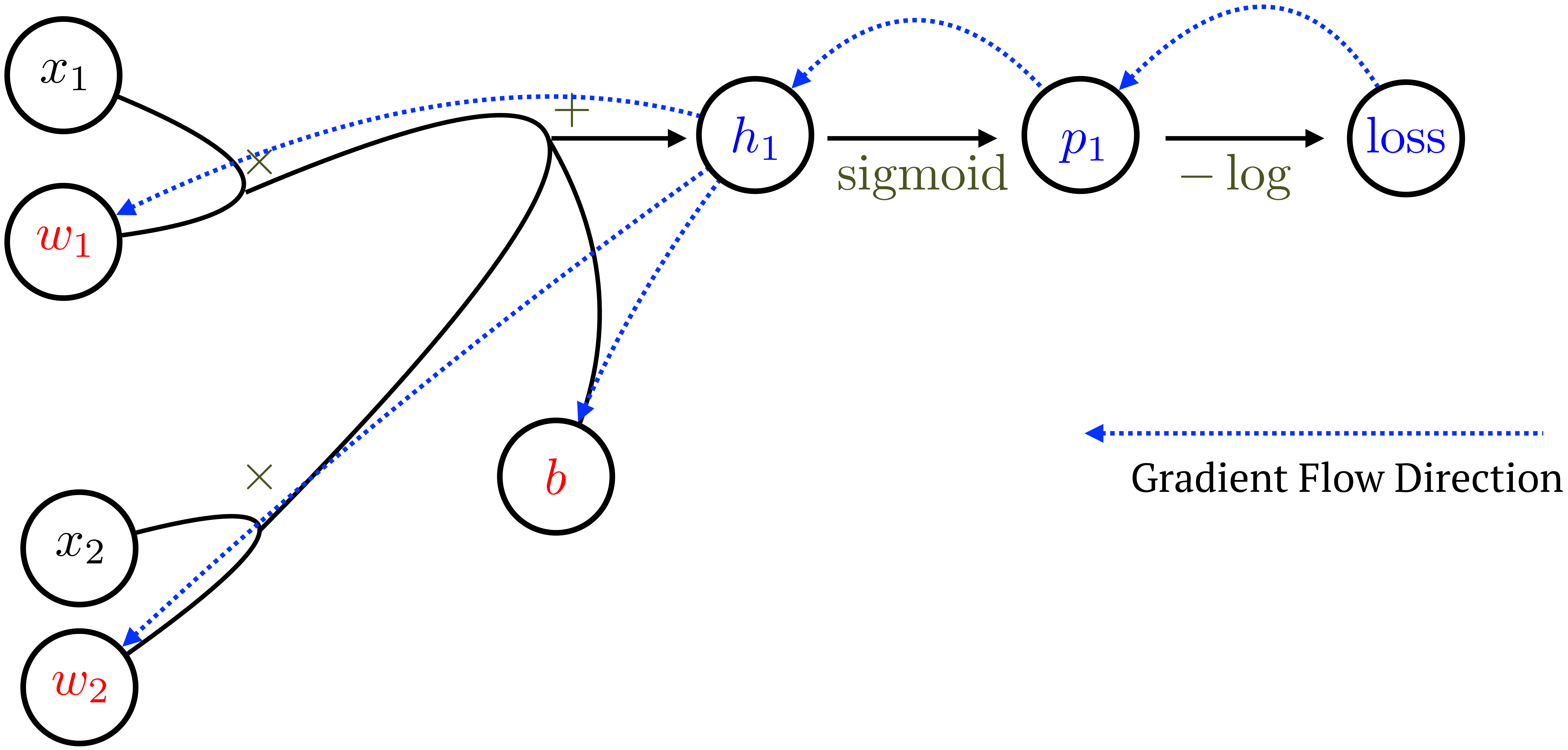
$$\boldsymbol{h} = f(\boldsymbol{z})$$

$$\boldsymbol{z} \xrightarrow{\;f\;} \boldsymbol{h}$$

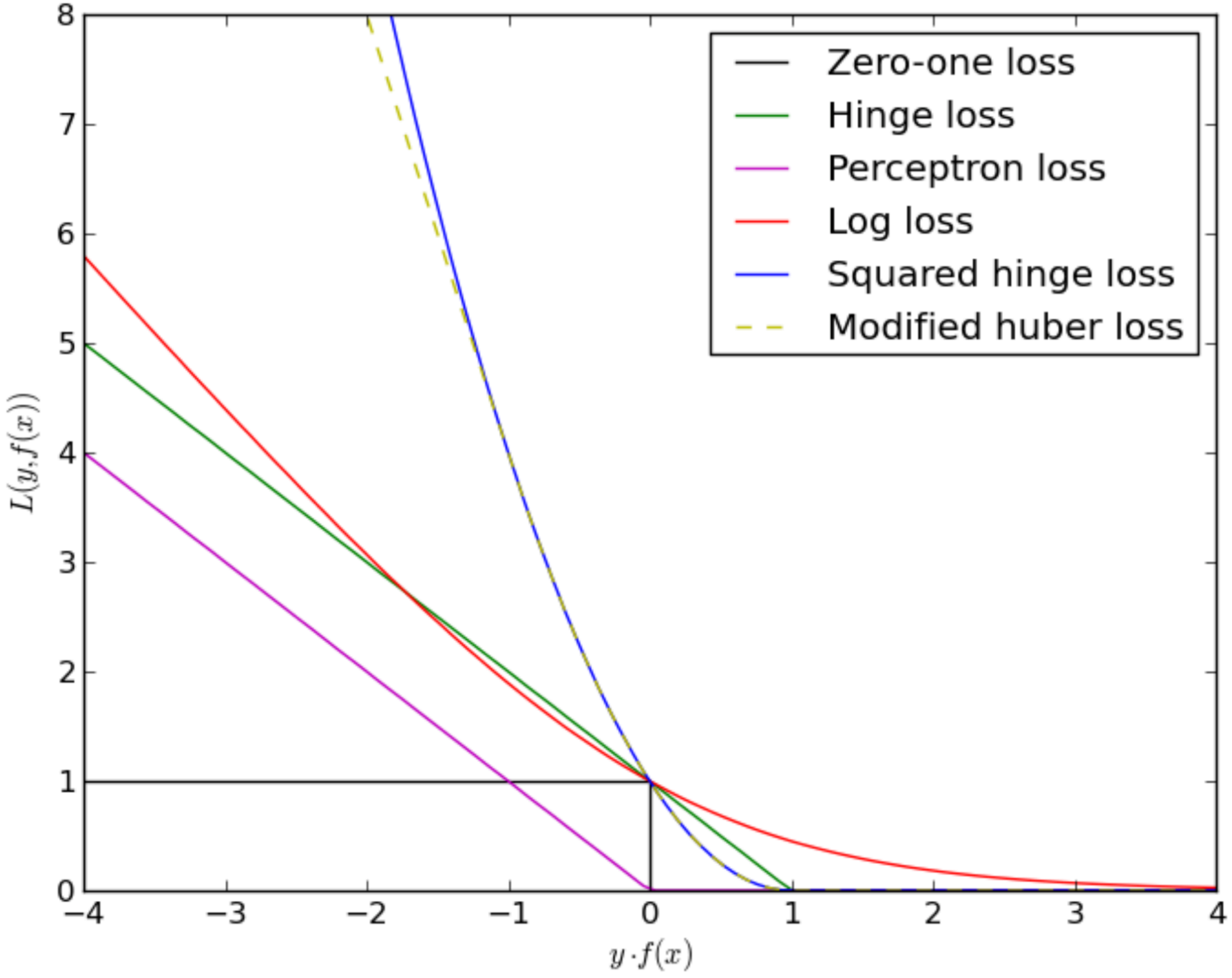
Chain rule $\dfrac{\partial s}{\partial \boldsymbol{z}} = \dfrac{\partial s}{\partial \boldsymbol{h}} \dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}$ $\qquad \dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}$ $\qquad \dfrac{\partial s}{\partial \boldsymbol{h}}$

Downstream Gradient — Local Gradient — Upstream Gradient

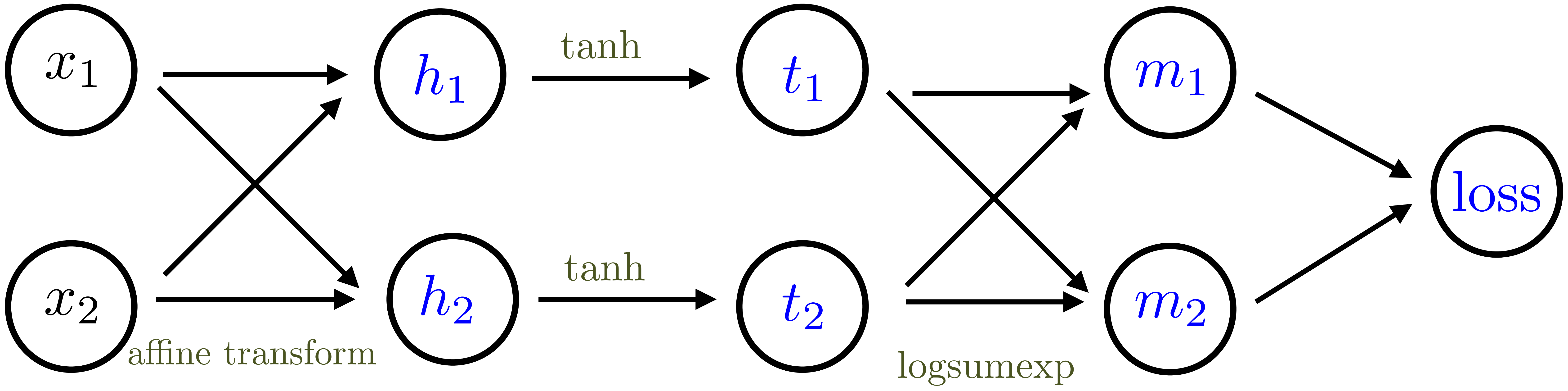
How to minimize? (Automatic Differentiation)



Other Loss Function?

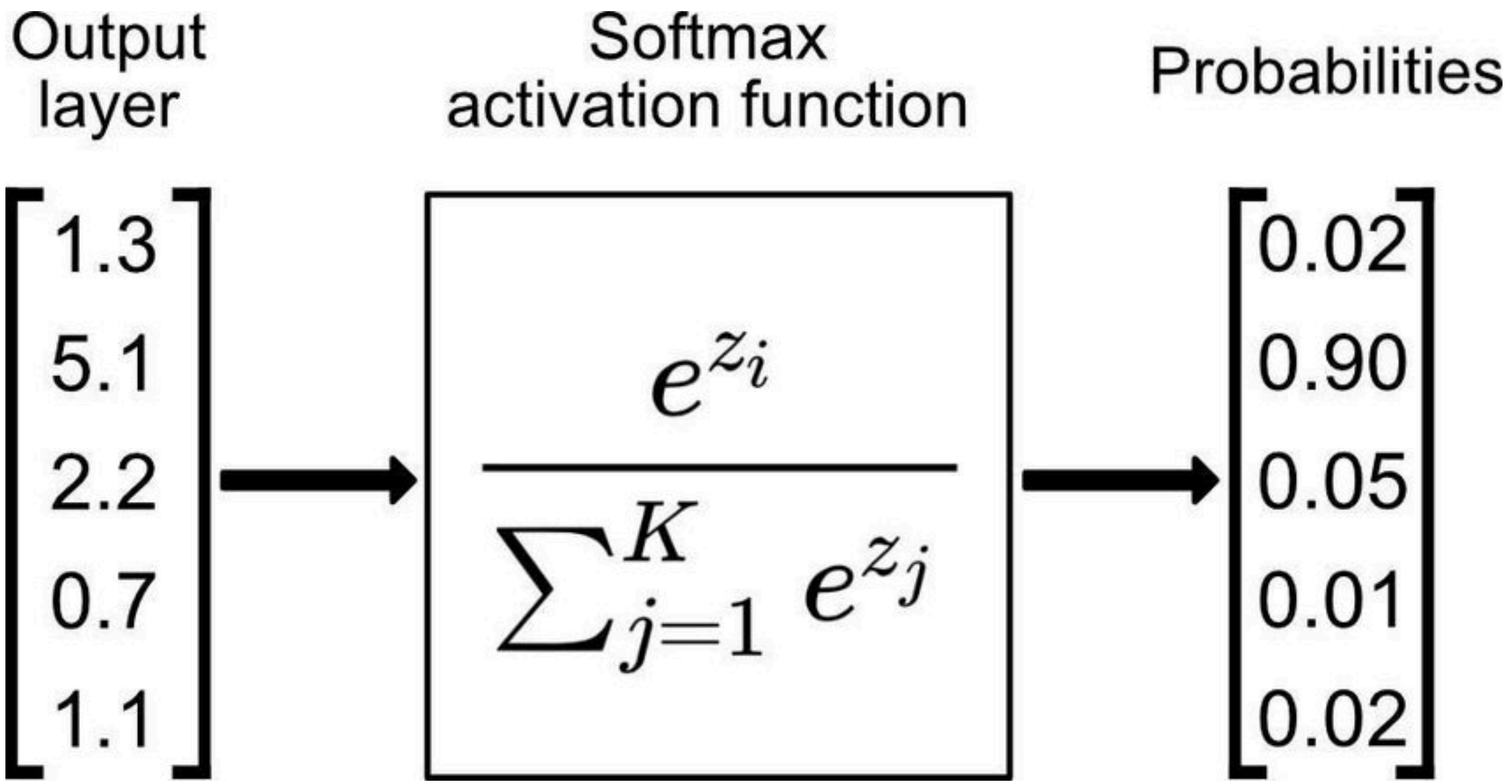


"Deeper" Neural Network

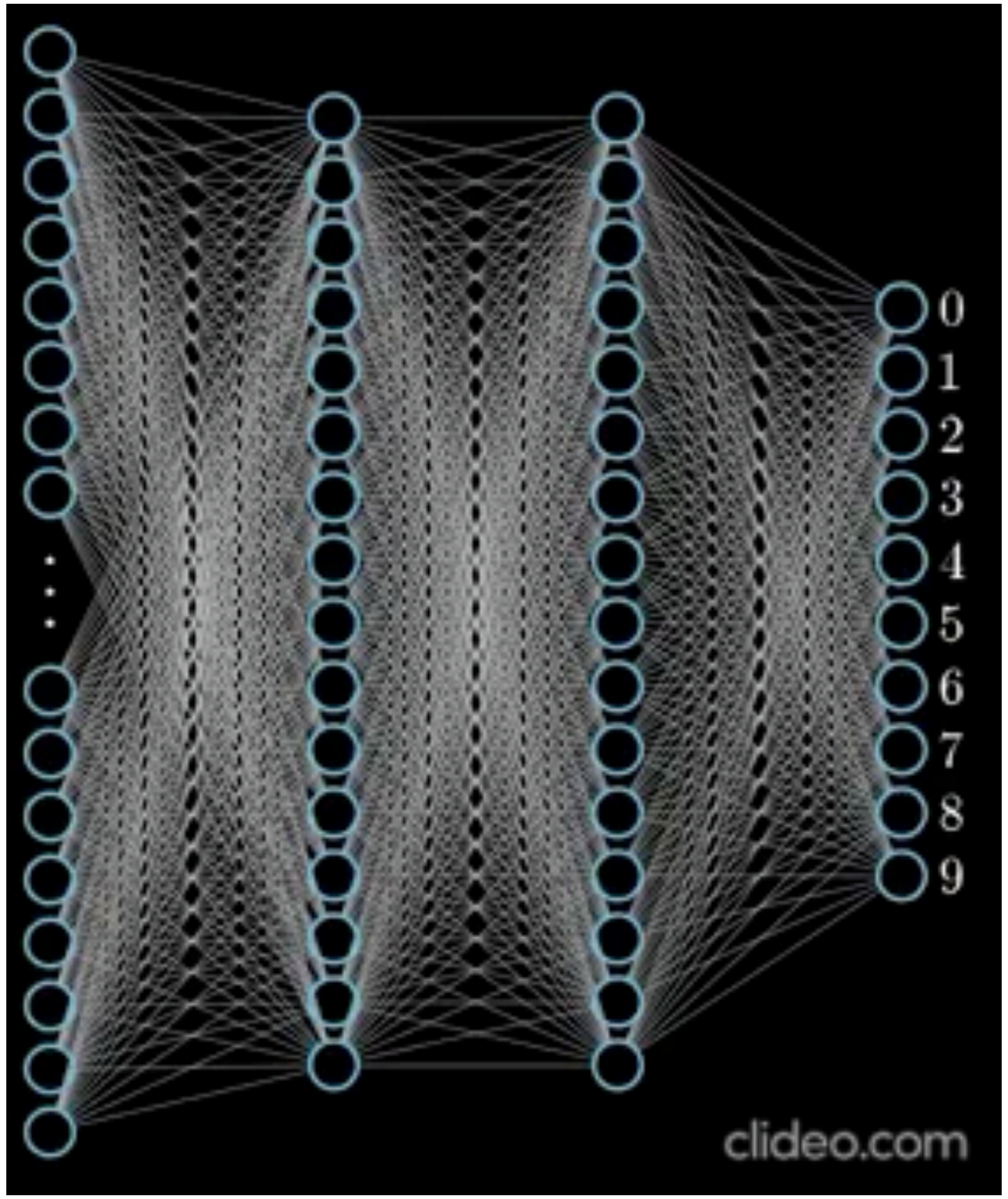


$$m_1 = \log\left(\frac{\exp(t_1)}{\exp(t_1) + \exp(t_2)}\right)$$

Softmax Function

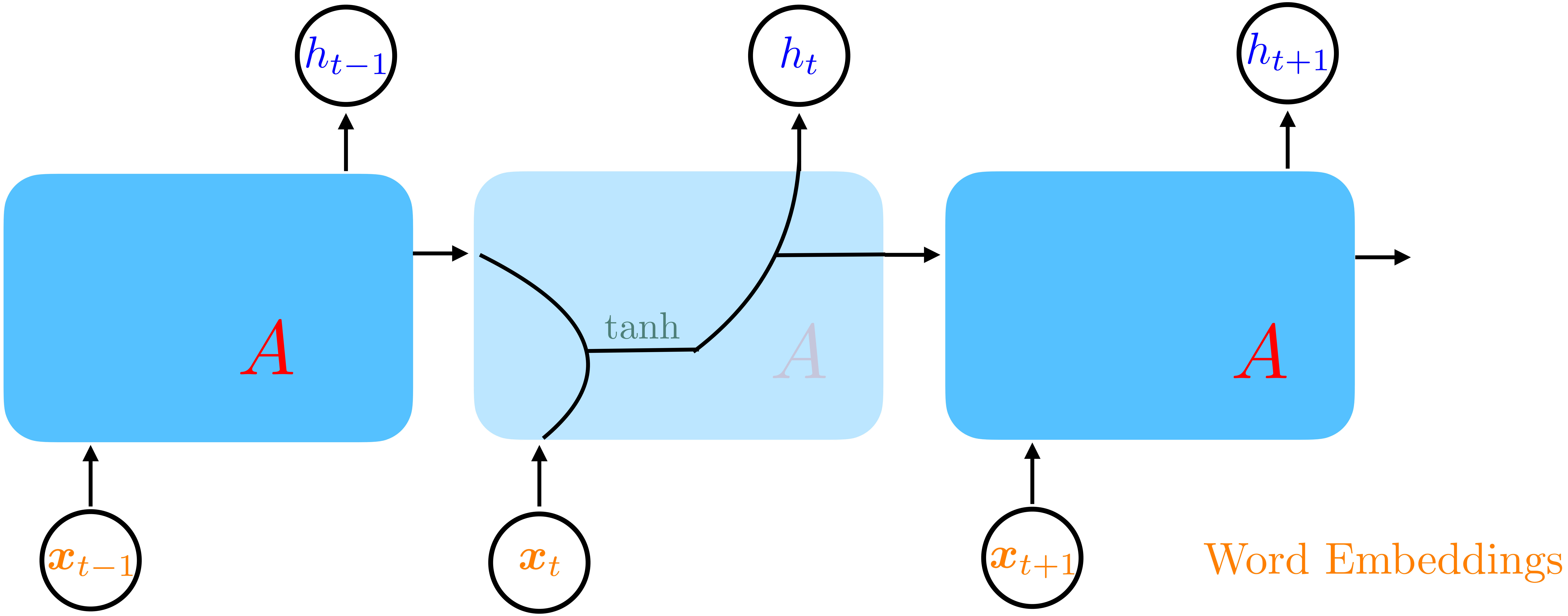


Neuralize the dice!

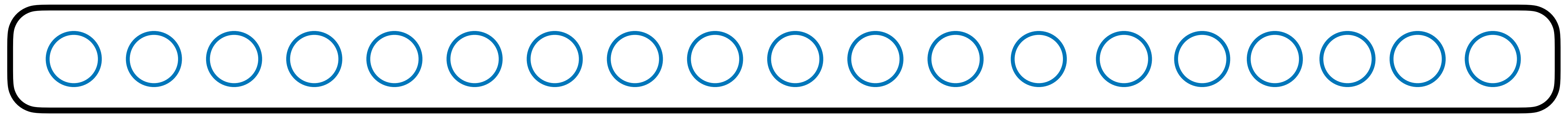


3Blue1Brown

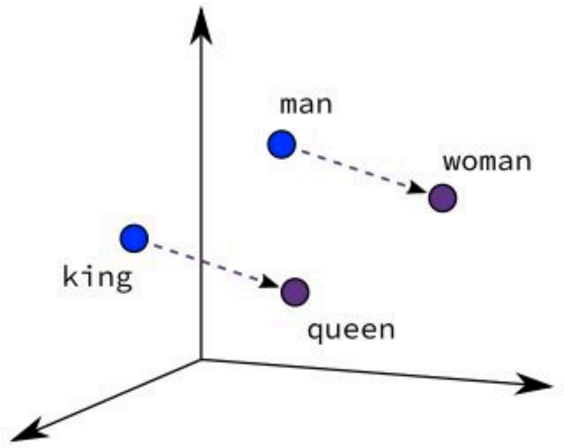
Recurrent Neural Network



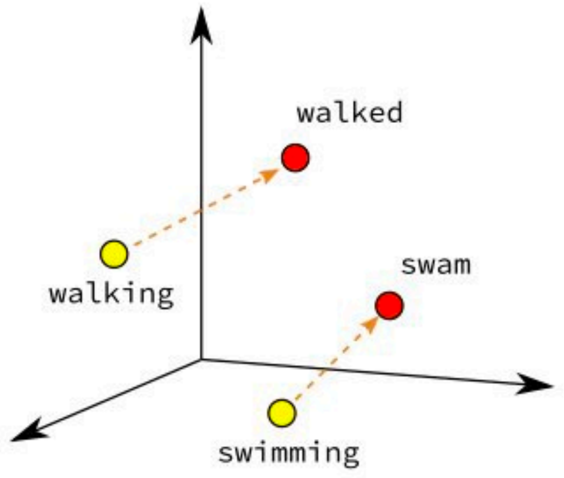
Word Embeddings



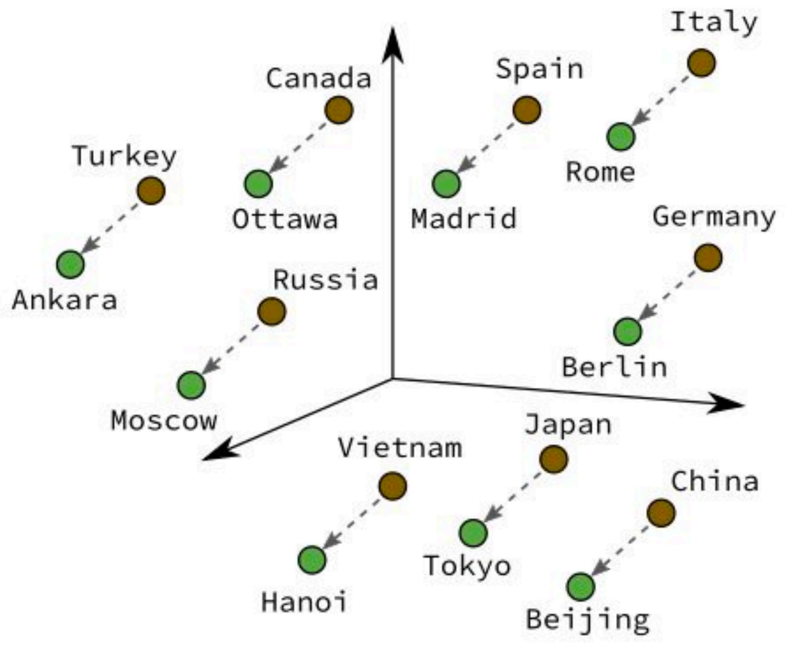
cat



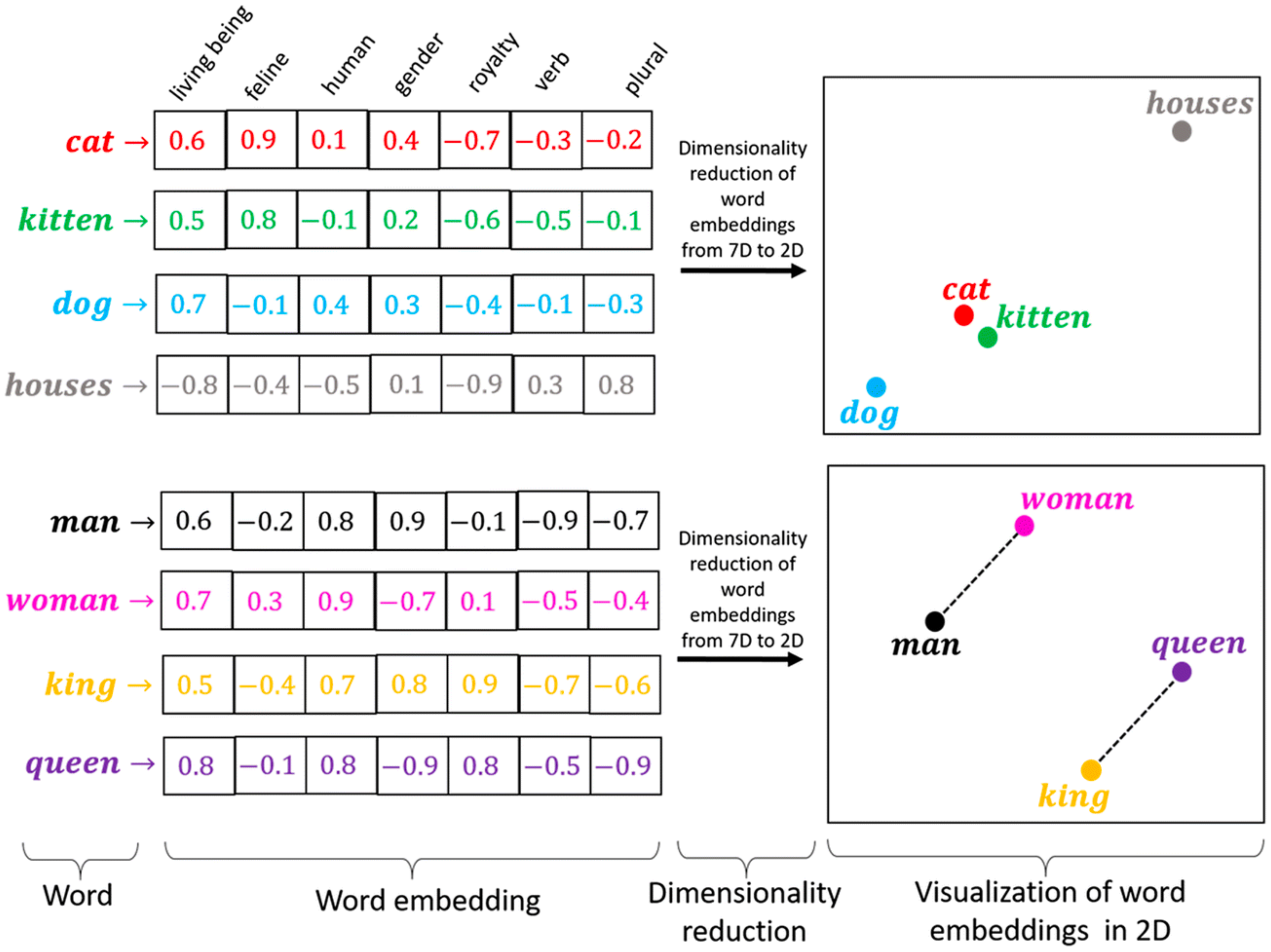
Male-Female



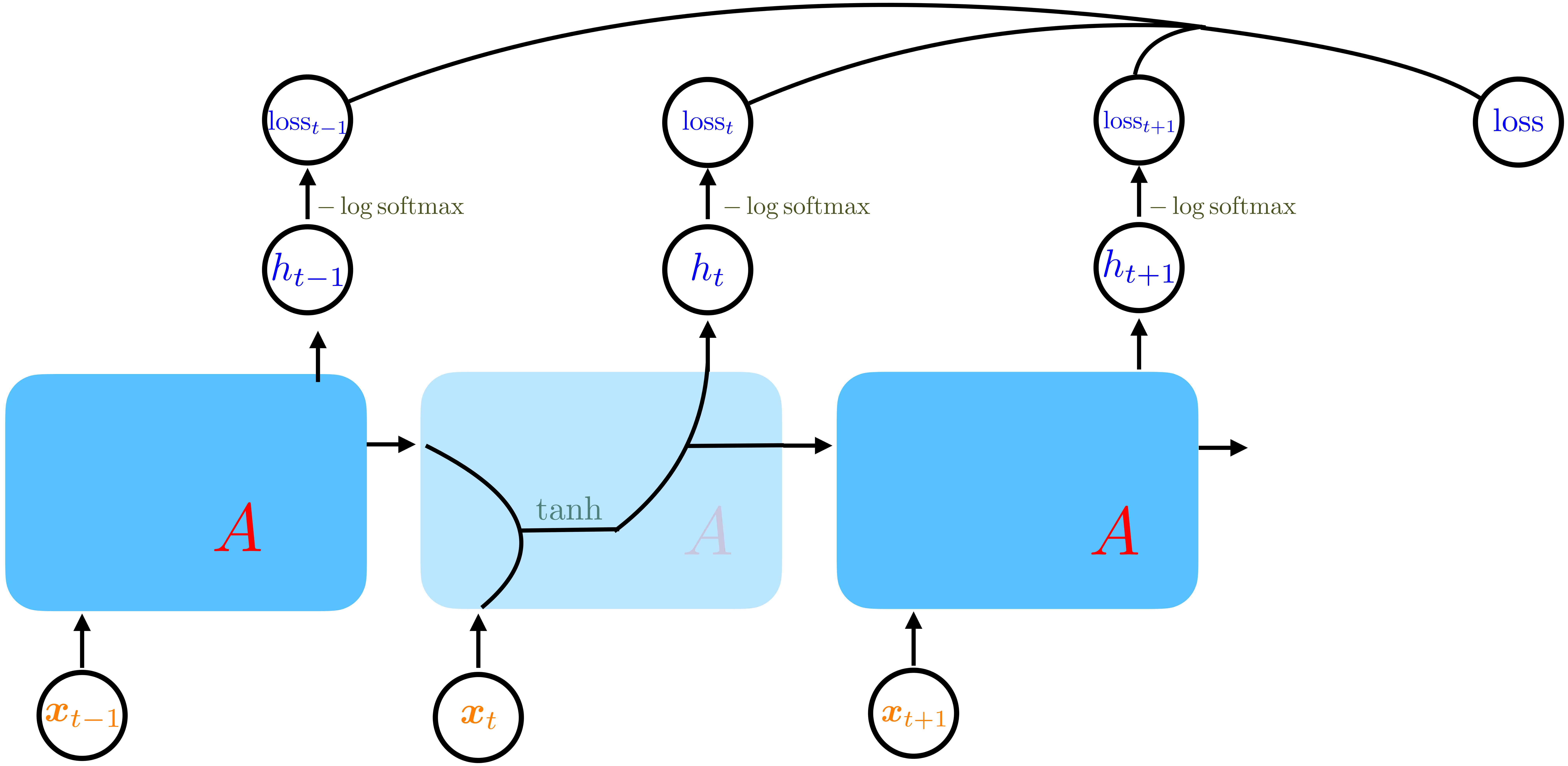
Verb Tense



Country-Capital



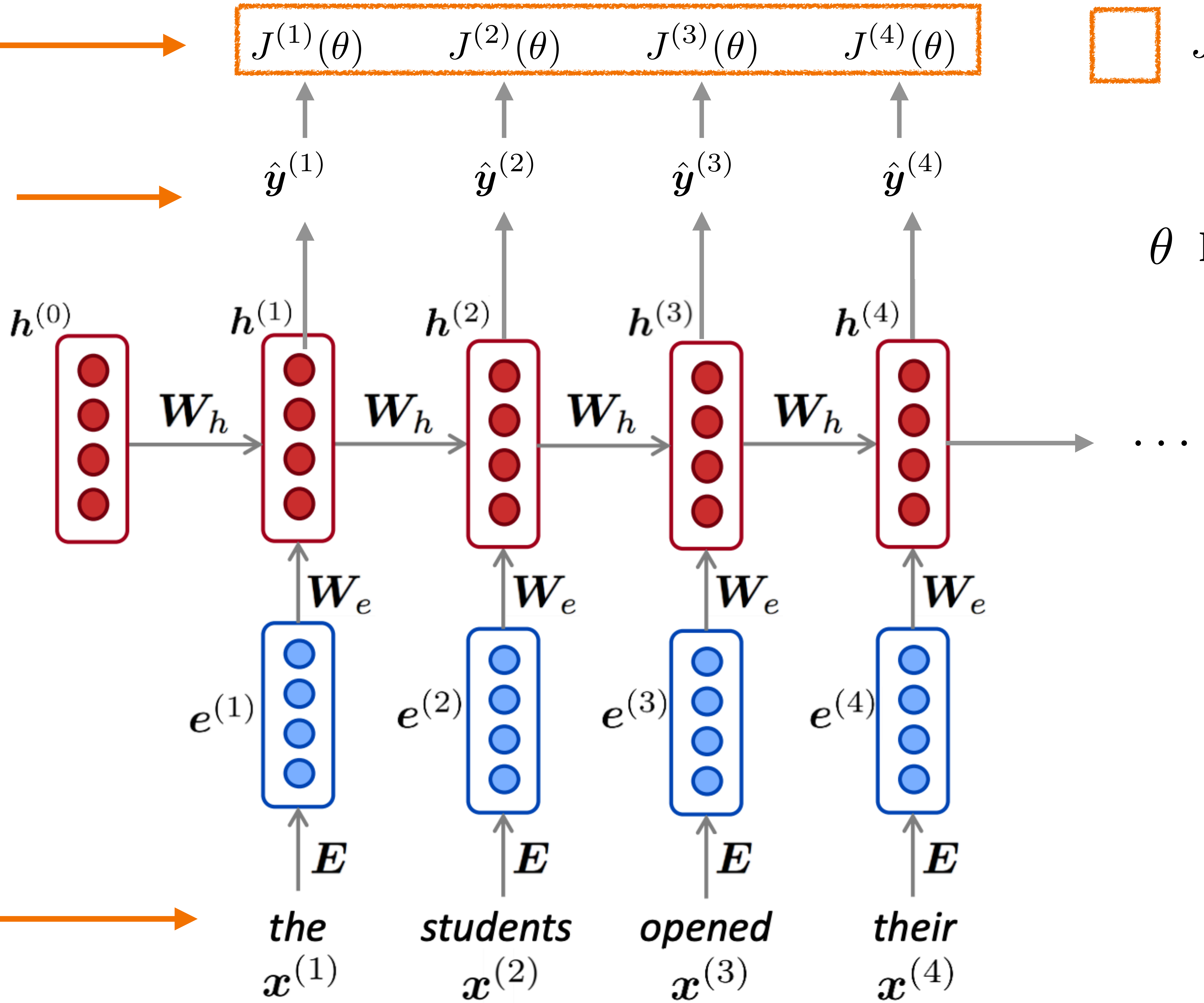
Recurrent Neural Network (Language Model)



Flashback — Training a RNN Language Model

Loss

Predicted probability distributions

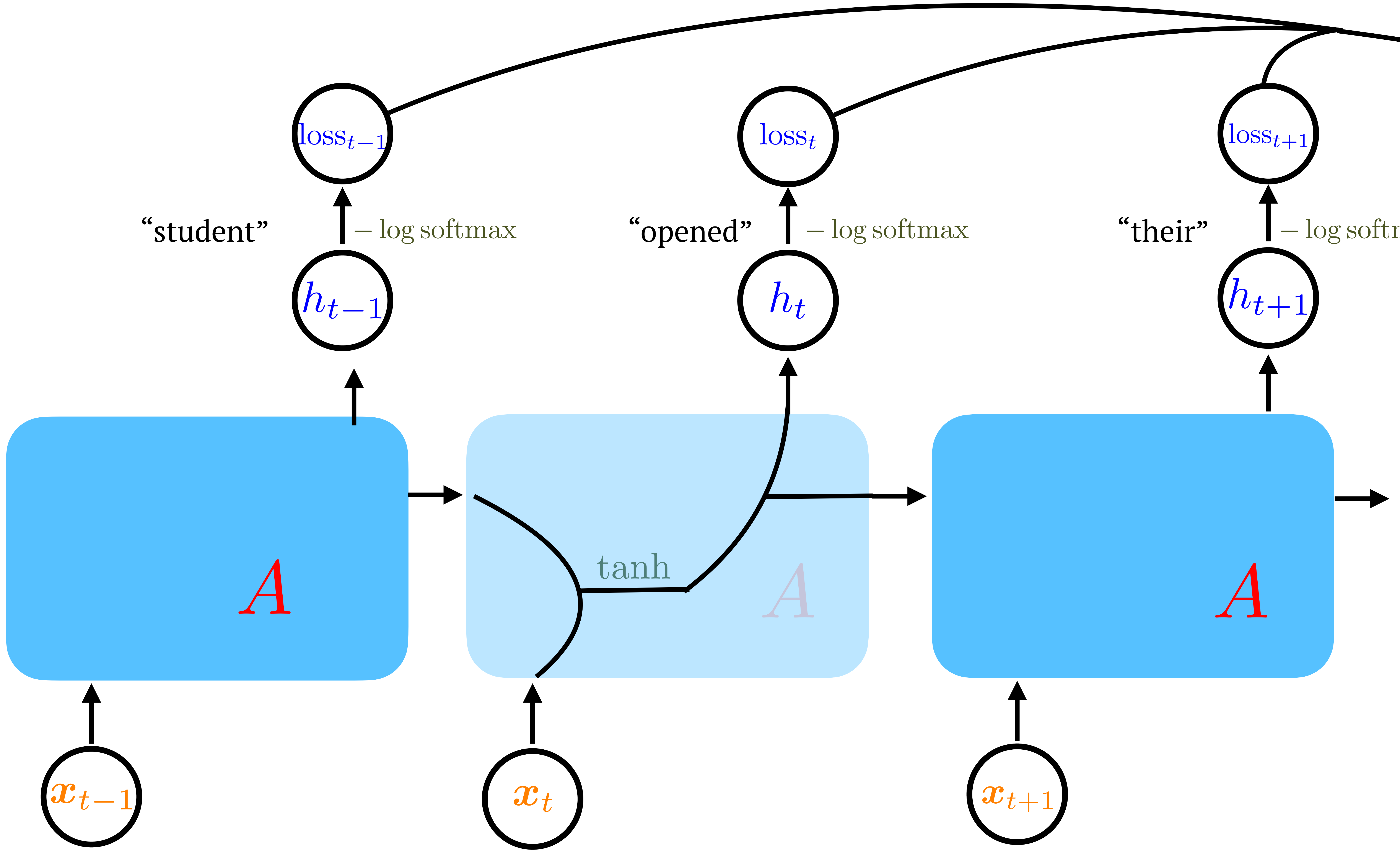


Training Corpus

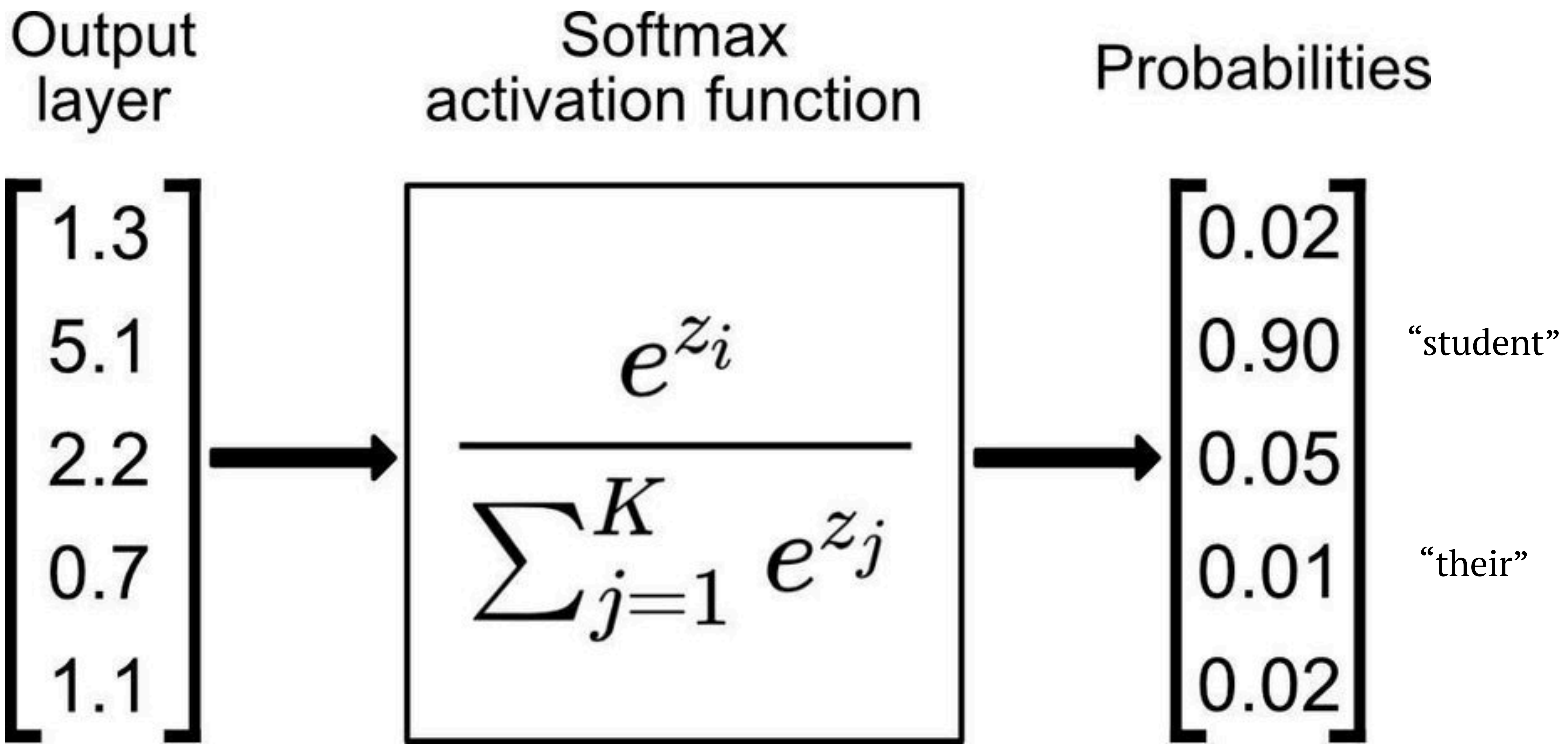
$$J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta)$$

θ Parameters in RNNs

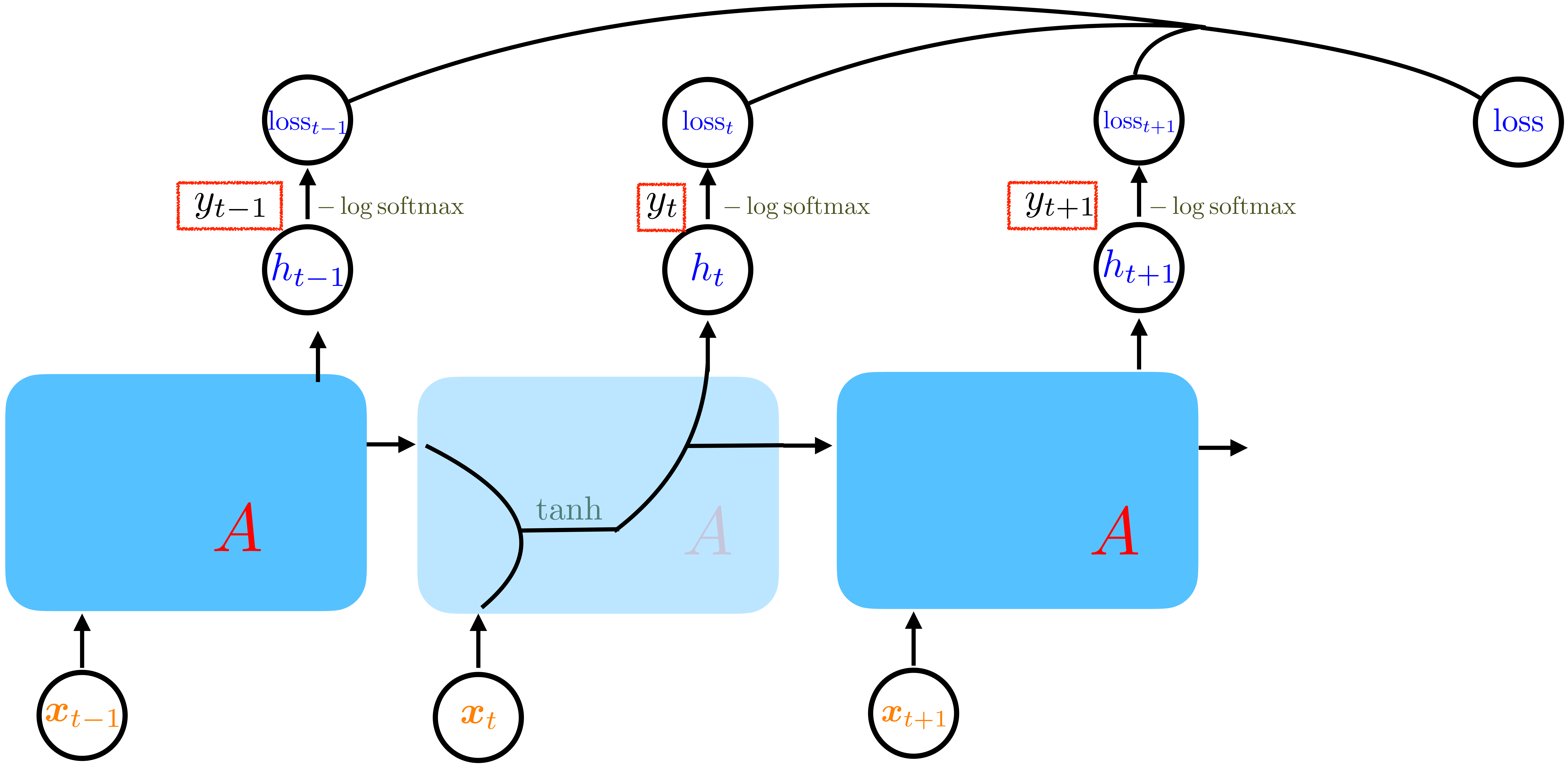
Recurrent Neural Network (Language Model)



Softmax Function



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.

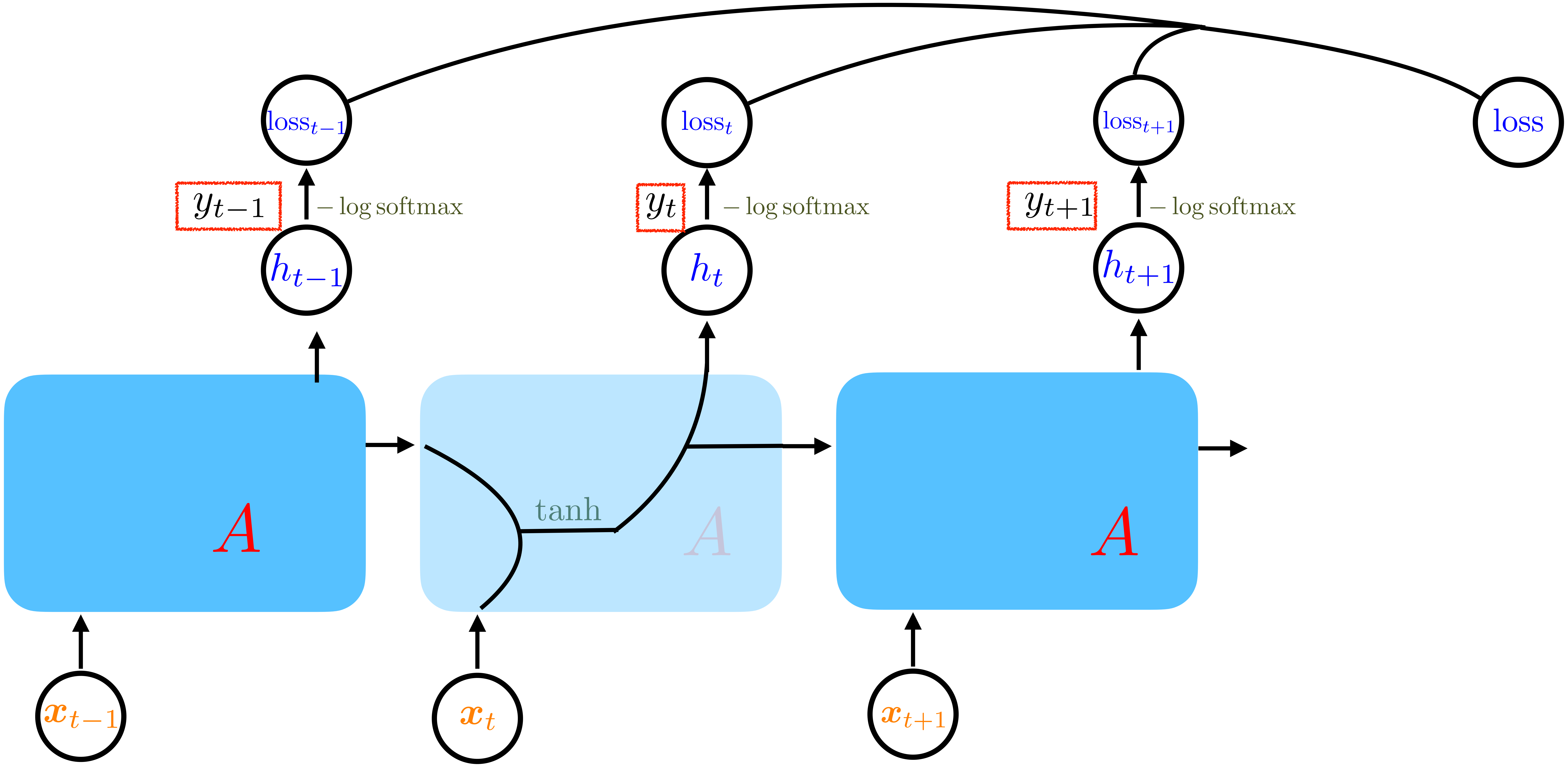
Named Entity Recognition (NER)

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

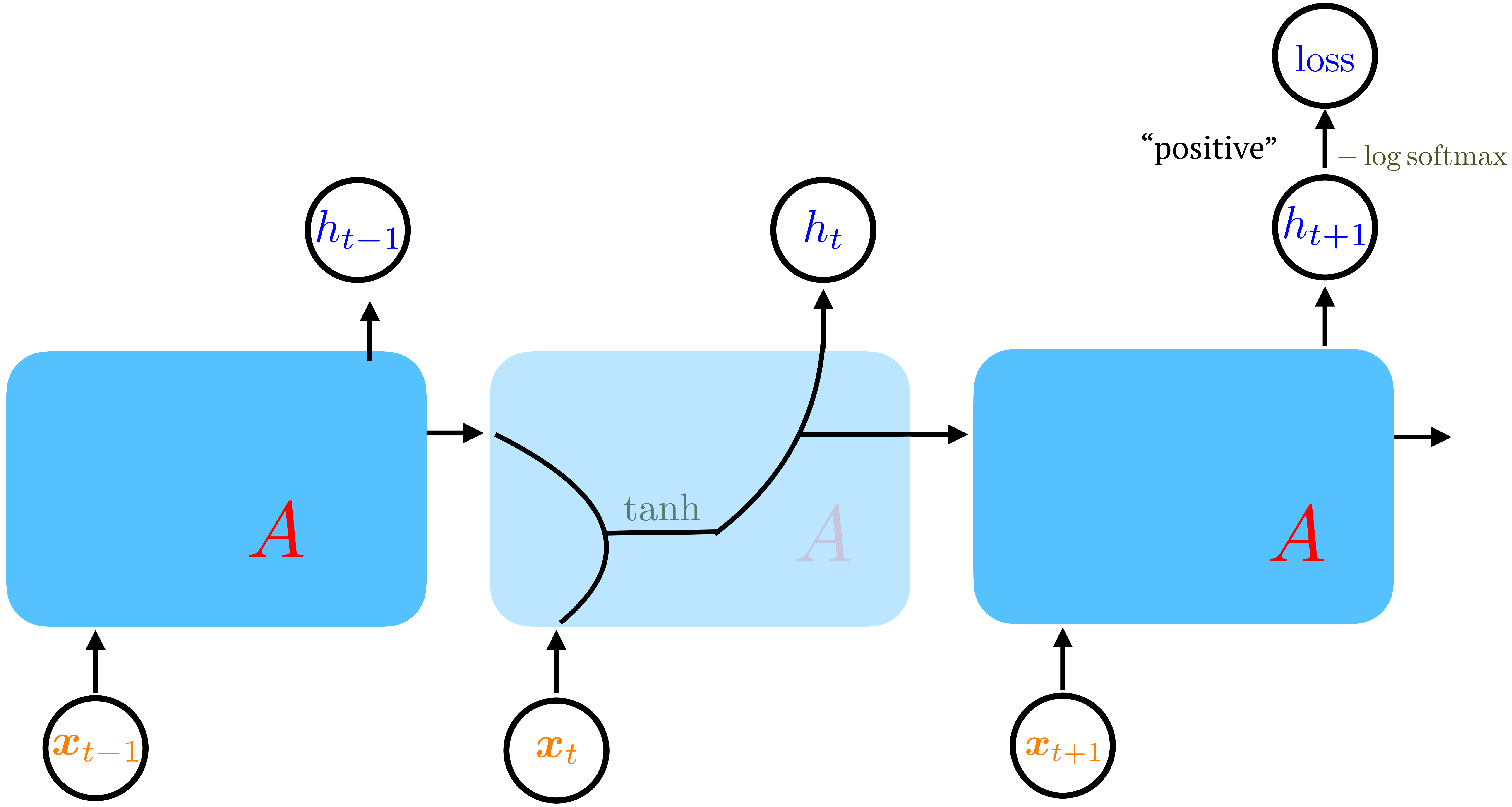
OUTPUT:
Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ,/NA as/NA their/NA CEO/NA Alan/SP Mulally/CP announced/NA first/NA quarter/NA results/NA ./NA

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

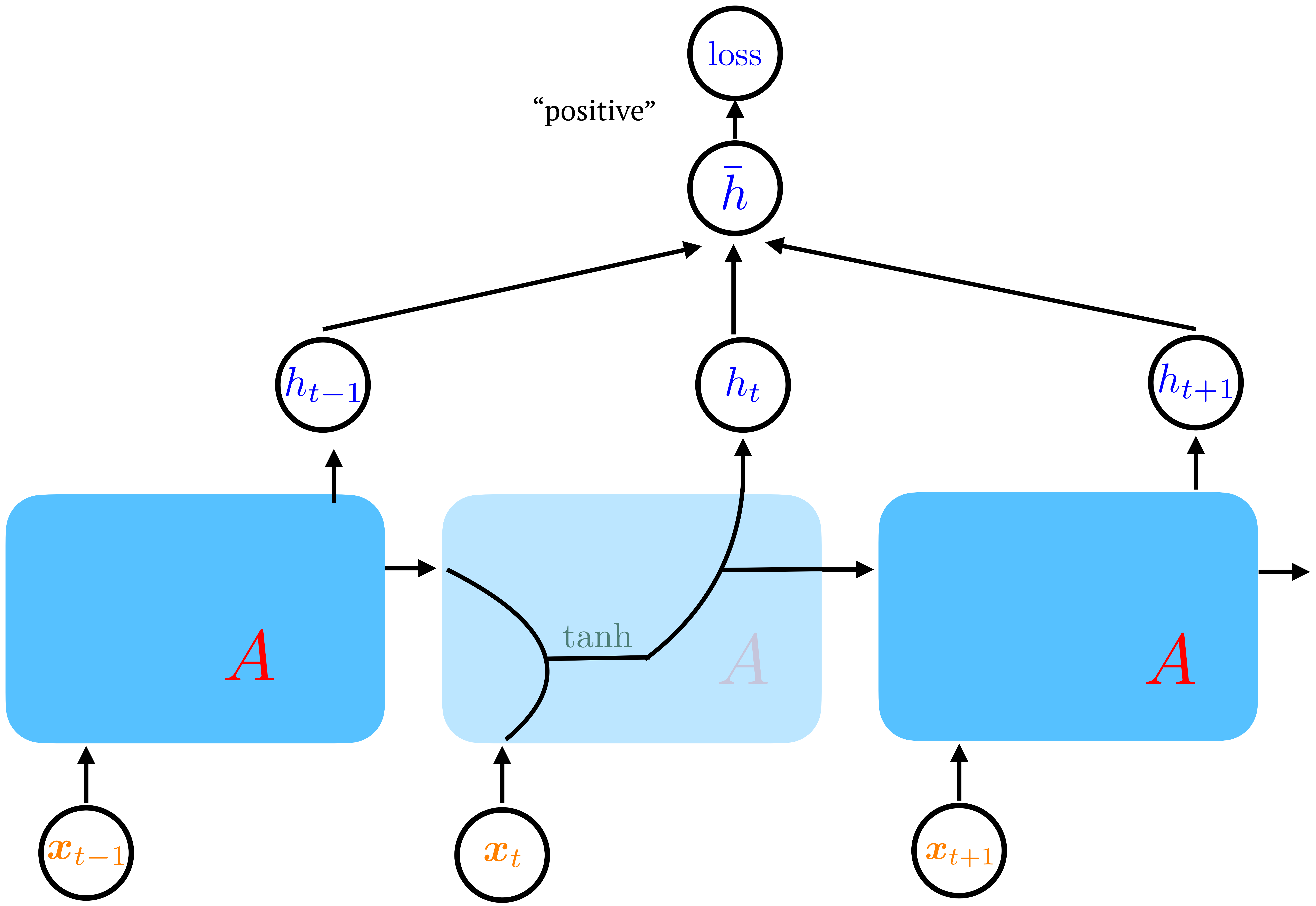
RNNs for Tagging



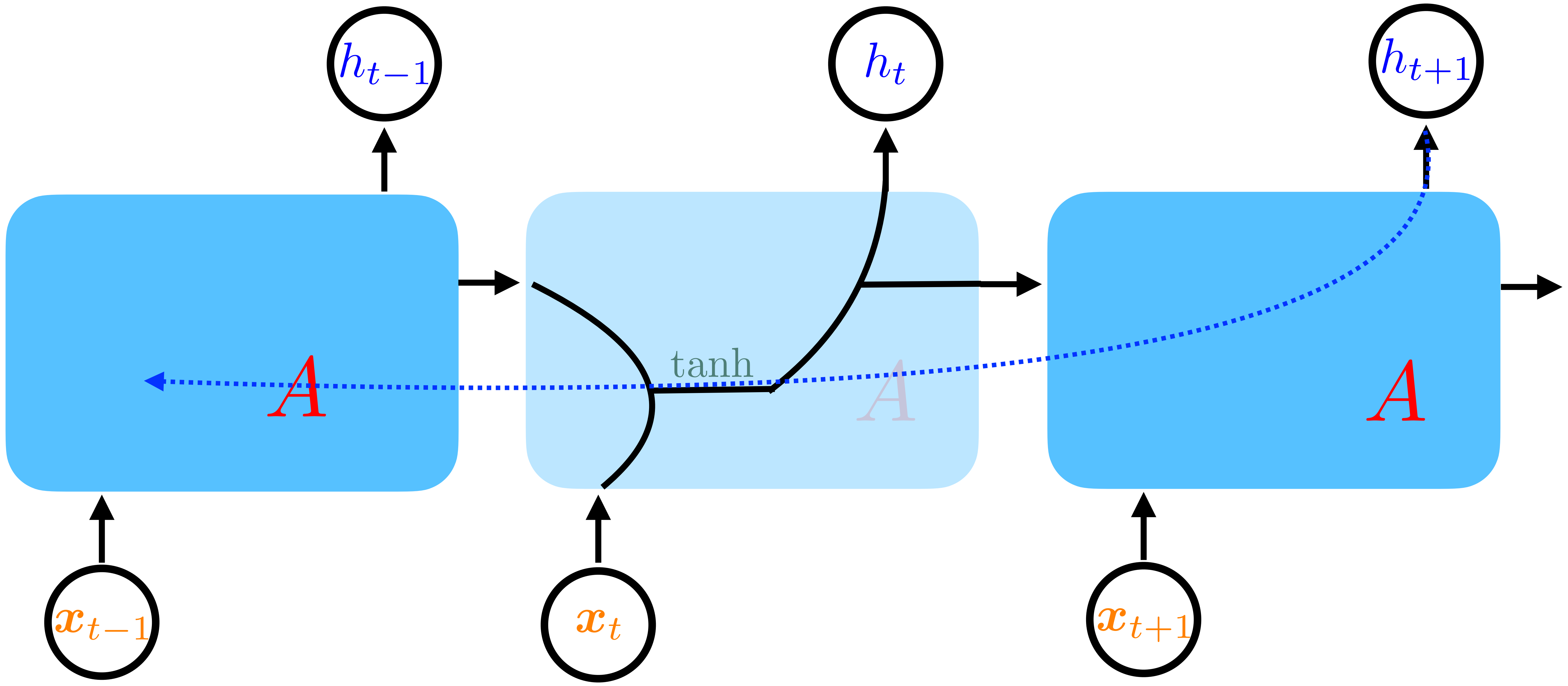
RNNs for Sentence Classification

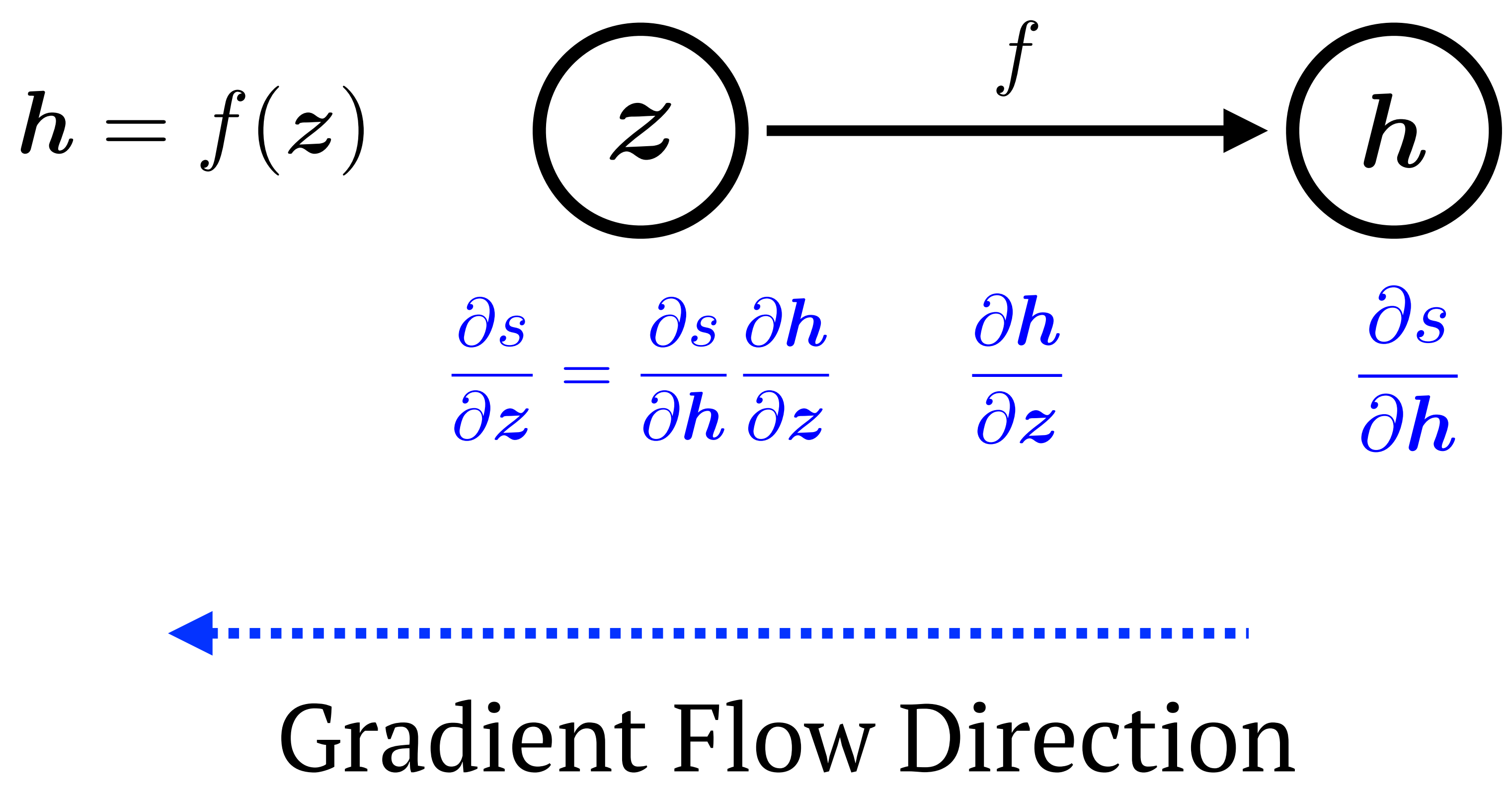


RNNs for Sentence Classification



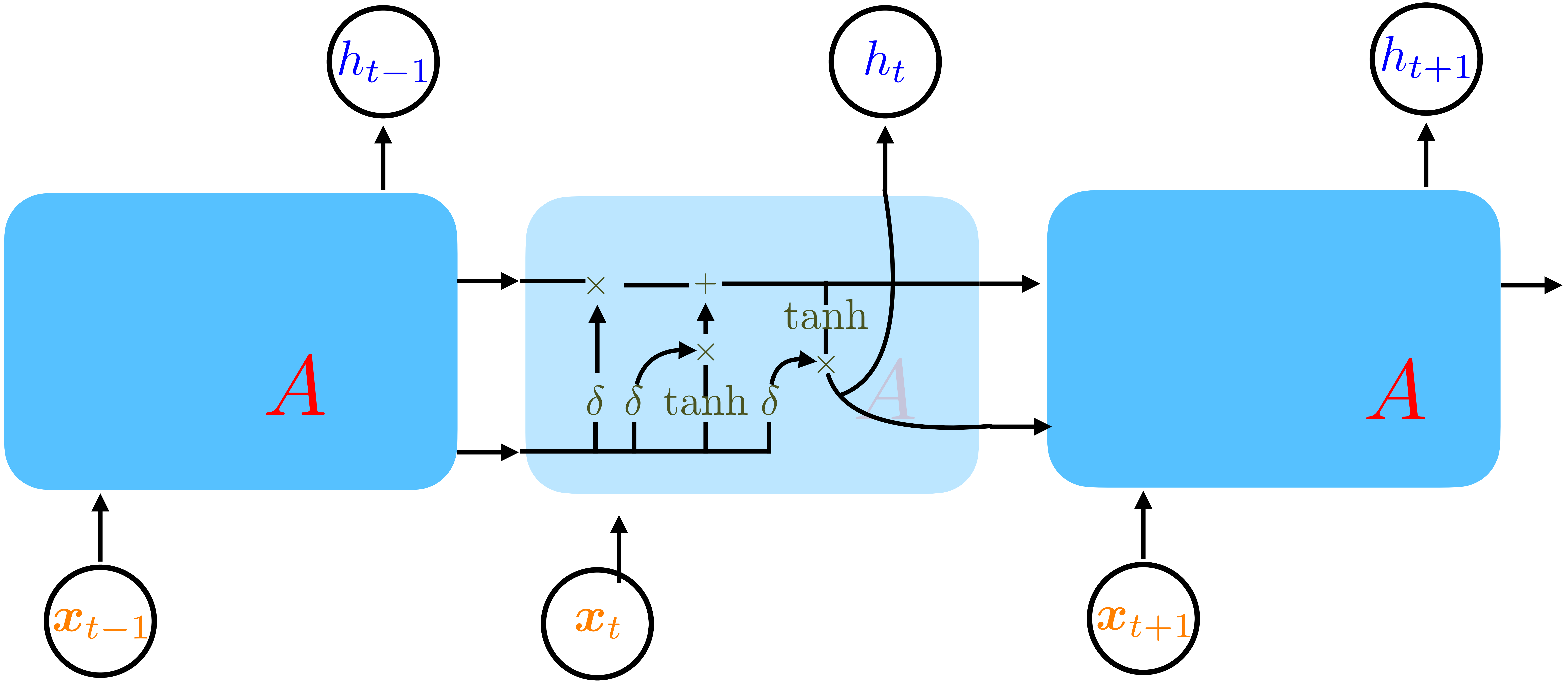
Vanishing Gradient in RNNs





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

Long Short-Term Memory (LSTMs)



http://demo.clab.cs.cmu.edu/cdyer/lstms.pdf

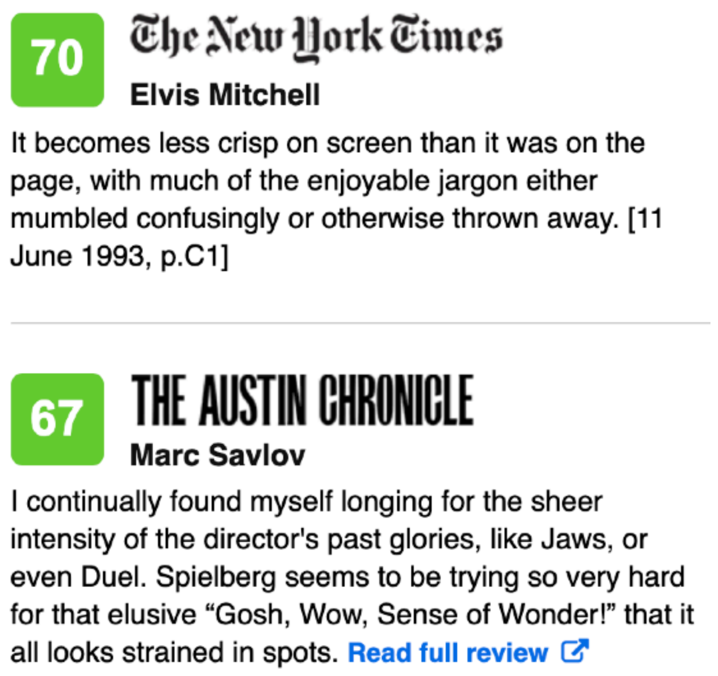
Typical NLP Task: Predicting gross revenues of movies

Models	Mean Absolute Error ($)
Baseline: Predict median from training data	7,079
Metadata (D): U.S. origin? , log budget, # screens, runtime *name, production house, genre(s), scriptwriter(s), director(s), country of origin, primary actors, release date, MPAA rating, and running time*	7,313
Text (T): Movie Reviews (from only before the release date) metacritic 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] 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 *Words, bigrams, trigrams, and dependency relations*	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 *name, production house, genre(s), scriptwriter(s), director(s), country of origin, primary actors, release date, MPAA rating, and running time*	7,313
Text (T): Movie Reviews (from only before the release date) *Words, bigrams, trigrams, and dependency relations*	8% 5% 6,729
Metadata (D) + Text (T)	6,725


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]

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

(Smith, 2010)

Models

Linear Regression:

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

"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:

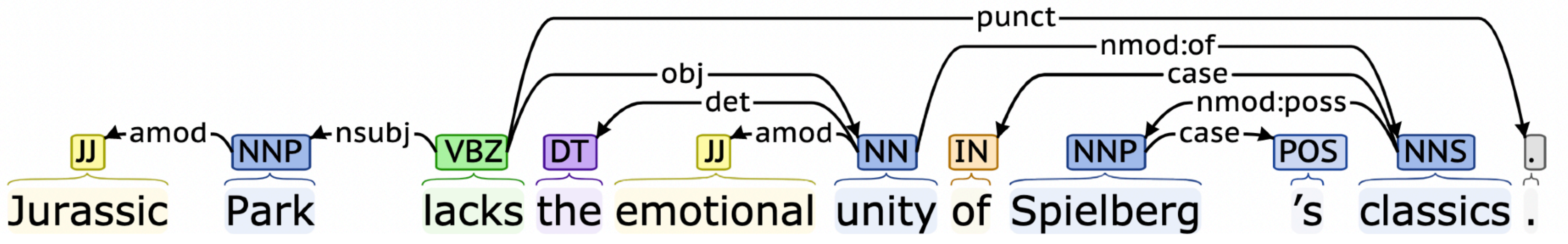
JJ
NNP
VBZ
DT
JJ
NN
IN
NNP
POS
NNS
.

Named Entities:

Movie: Jurassic Park

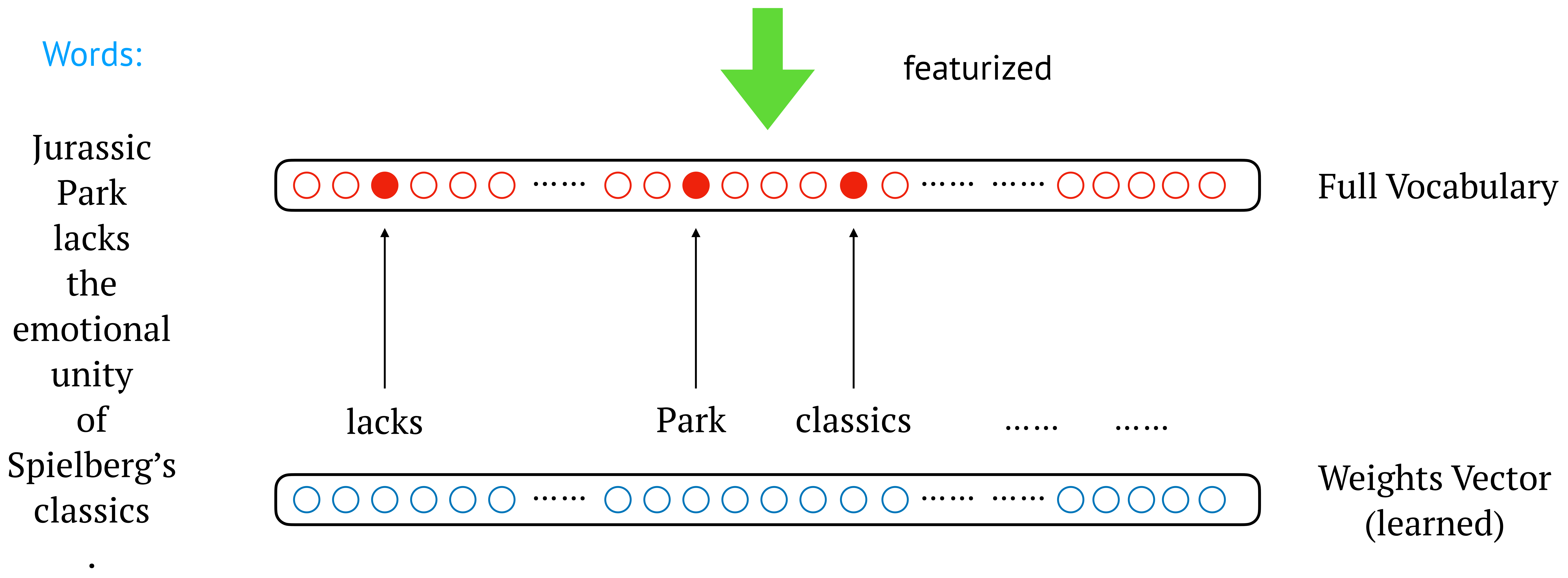
Person: Spielberg

Dependencies (Syntax Parsing):



Bag-of-words Models

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



Natural Language Processing (NLP) Pipeline

General-purpose linguistic modules:

Words Bigrams — Light preprocessing (mostly rule-based)

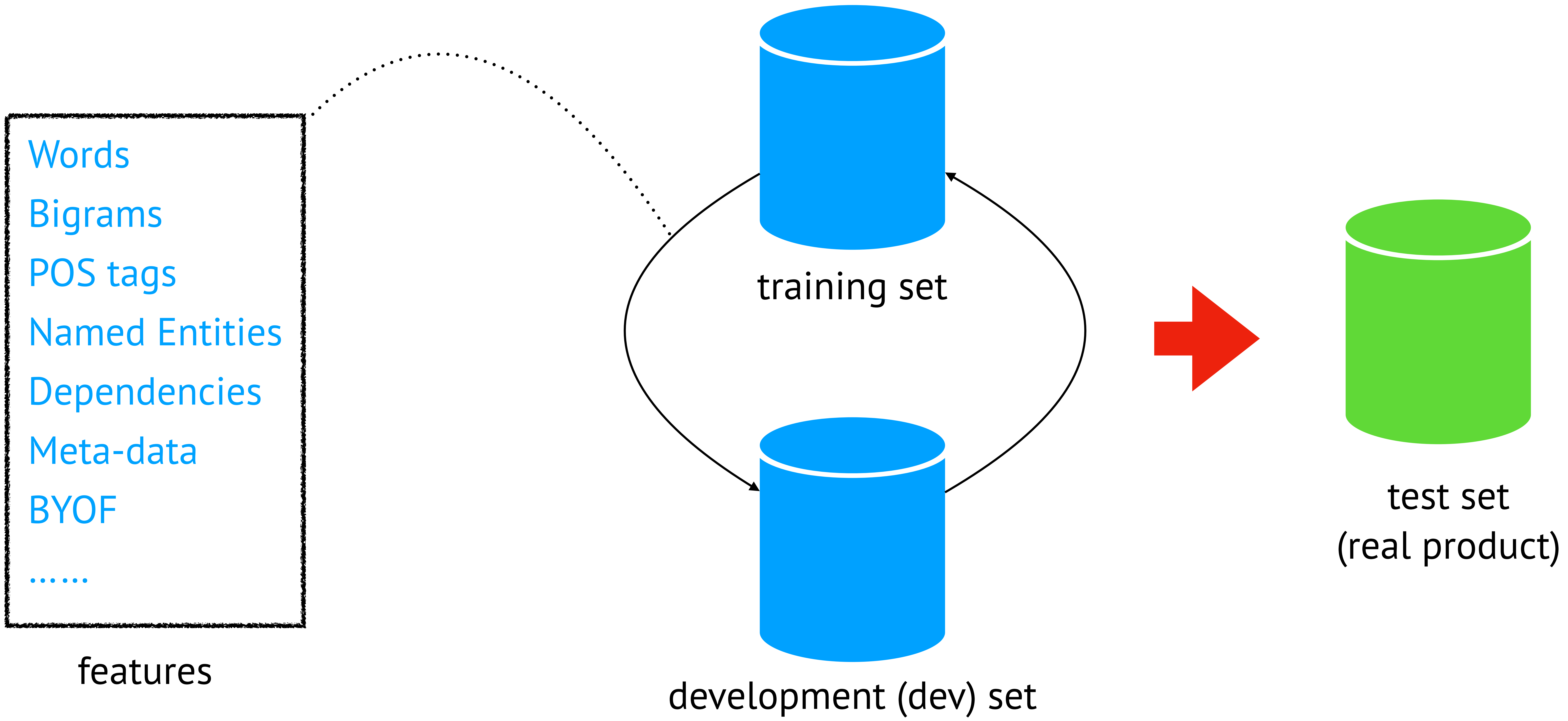
Part-of-speech tags: word classes

Named Entities: words of interests

Dependencies (Syntax Parsing): Internal structures

Supervised learning from linguistic data
(CoreNLP pipeline)

Feature Engineering



Wait, where is deep learning?

Words

Bigrams

POS tags

Named Entities

Dependencies (from a neural parser)

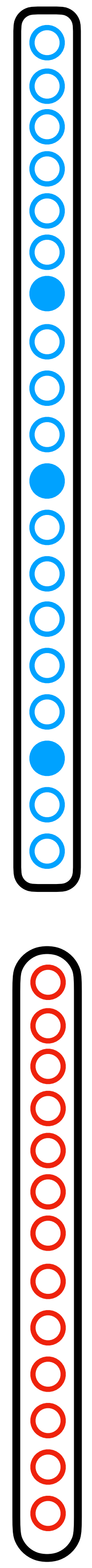
Meta-data

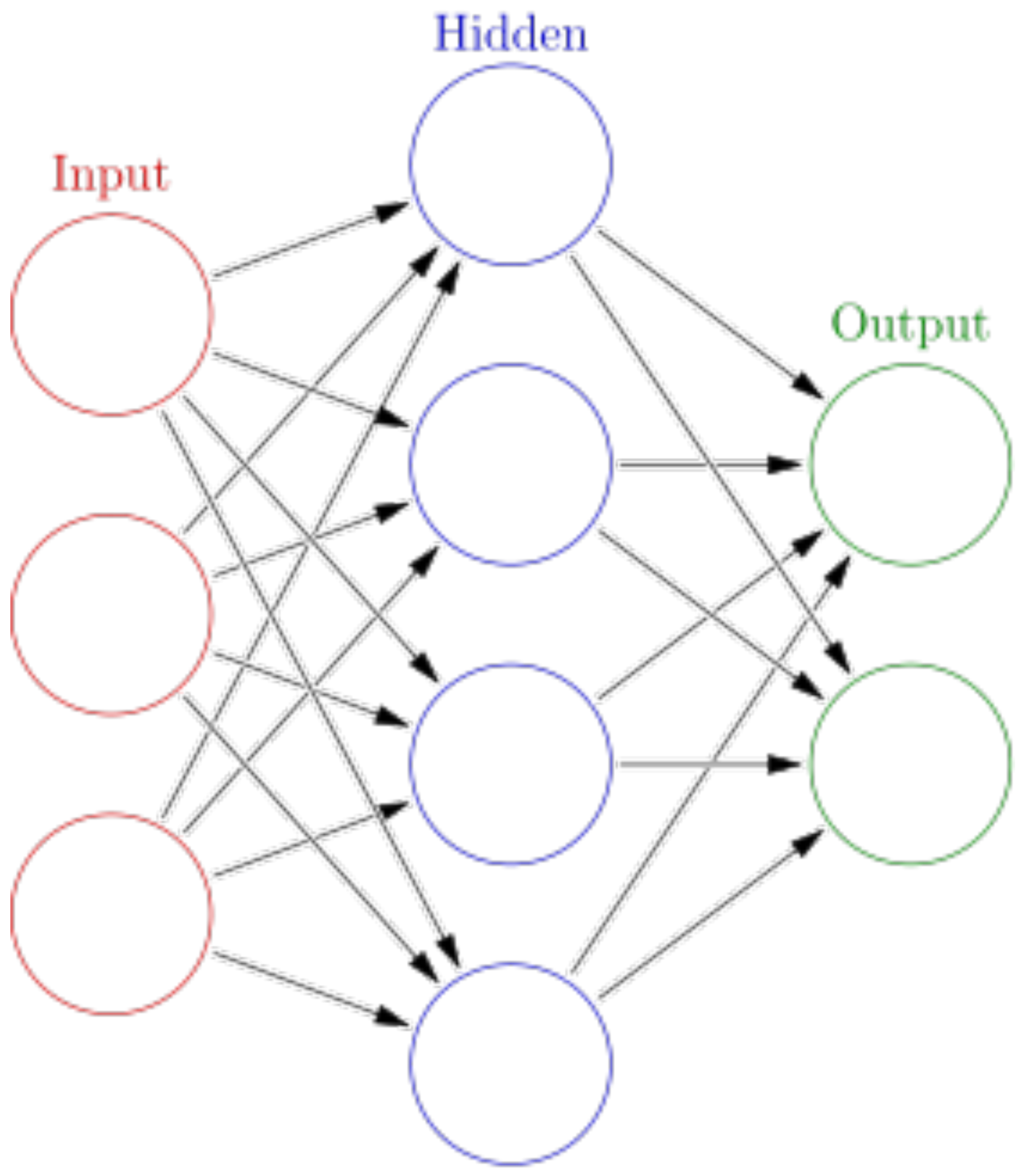
BYOF

Word embeddings

……

features



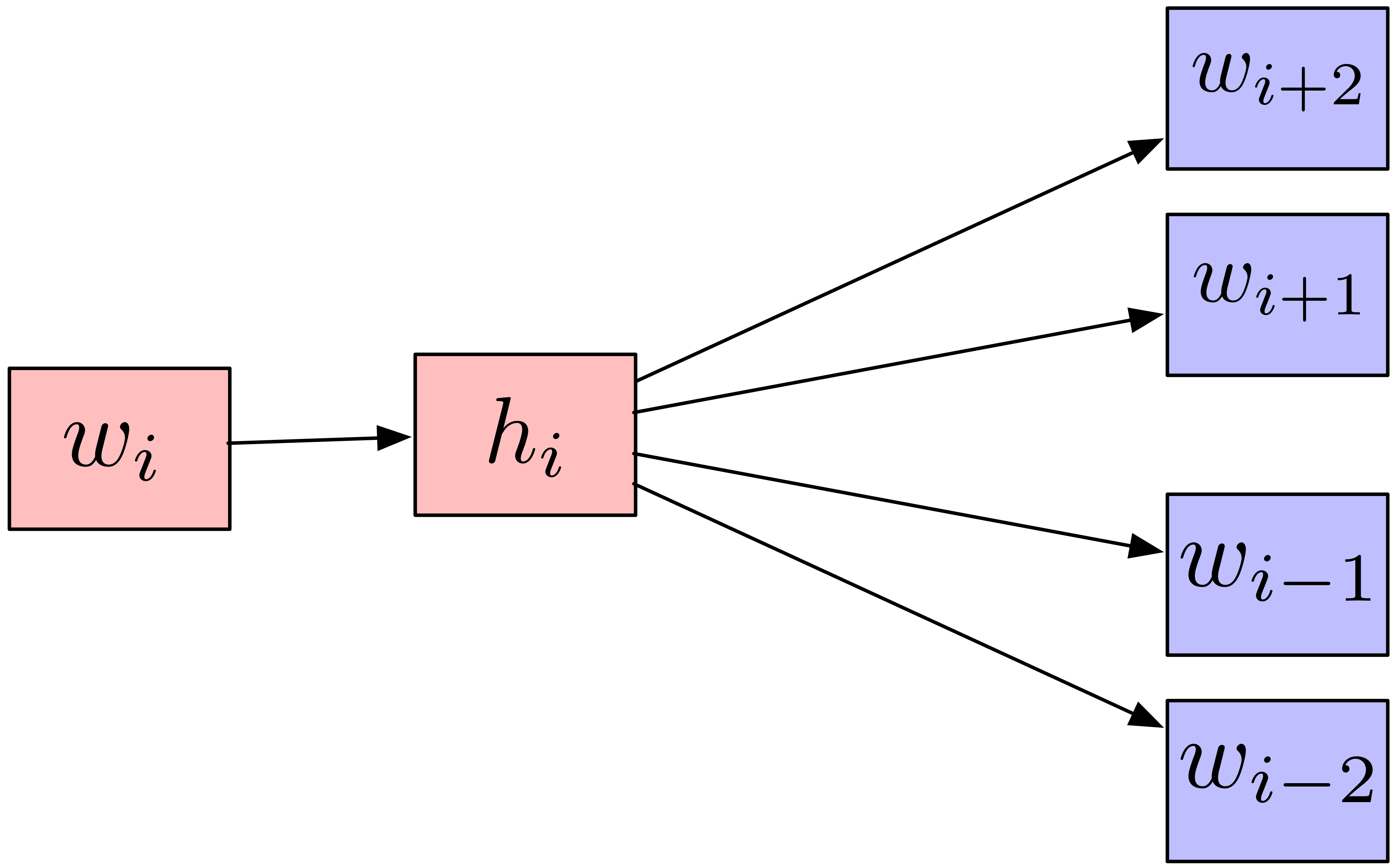


neural networks
(often MLP)

predictions

General-purpose representation learning

Word2vec – pertained word embeddings



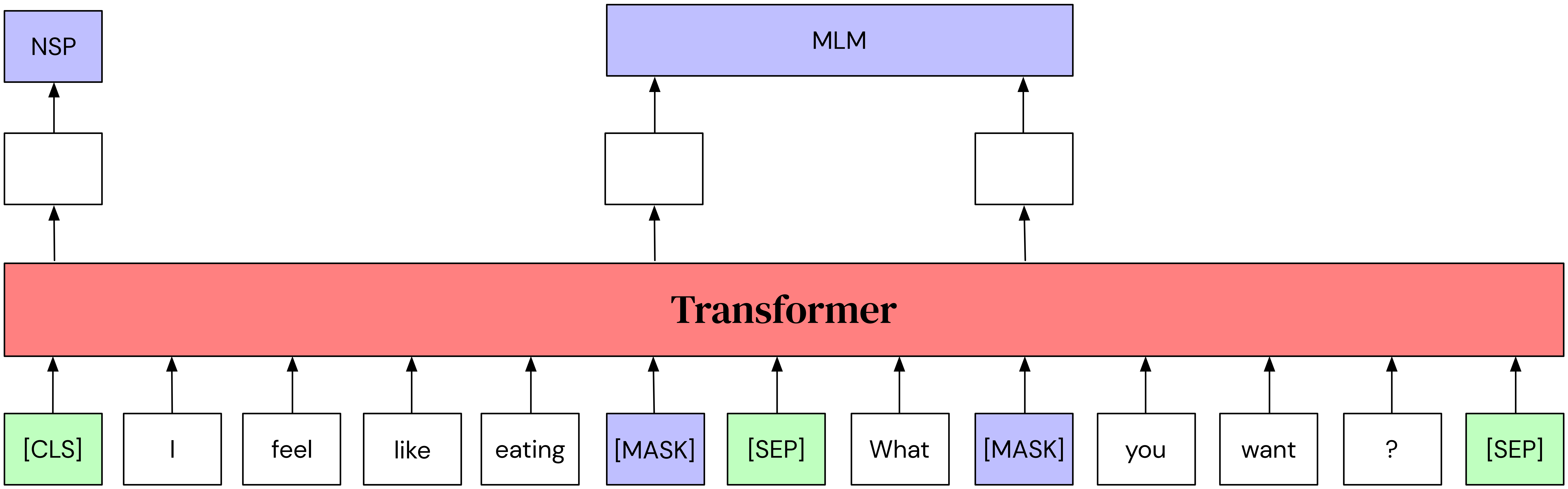

```
1   y the irradiated and refrigerated  chicken. Acceptance of radiopasteurization
2   torehouse".     Glendora dropped a chicken and a flurry of feathers,  and went
3    will specialize in steaks, chops, chicken and prime  beef as well as Tom's fa
4   ard  as the one concerned with the chicken and 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
```

$$\mathbb{E}_{p(x_i, x_j^i)}\left[p(x_j^i \mid x_i)\right]$$

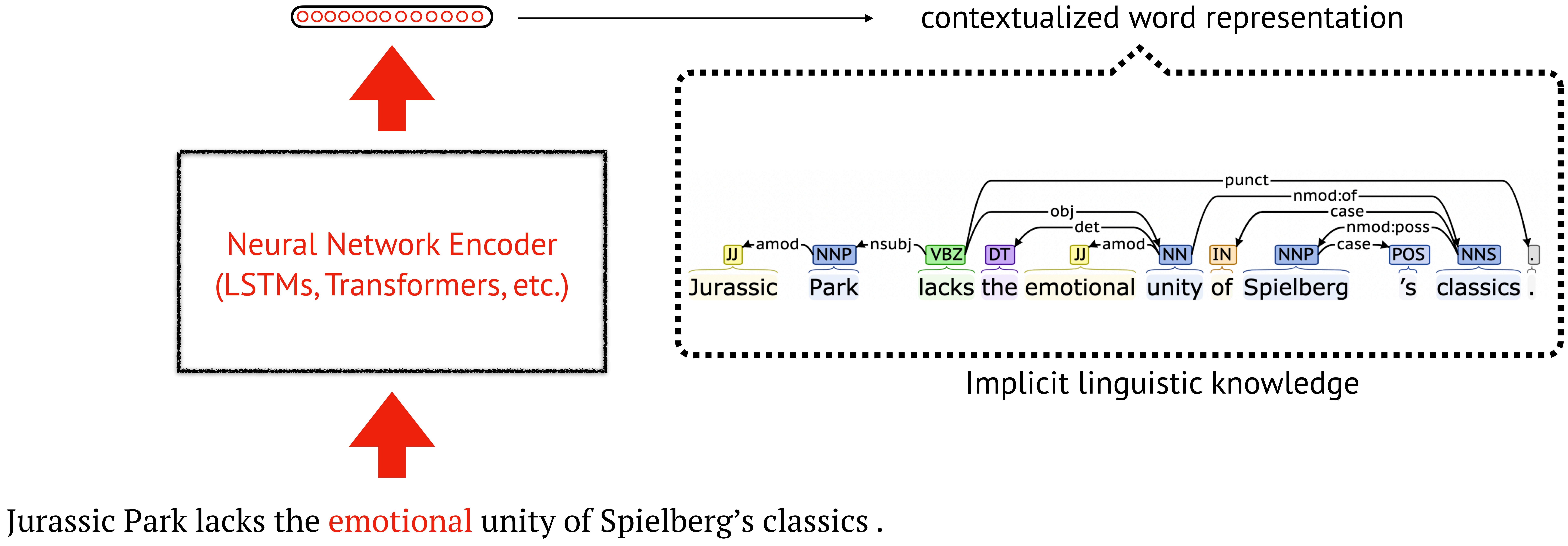
Where is “chicken” ?

Pretraining and Contextualized Word Representations

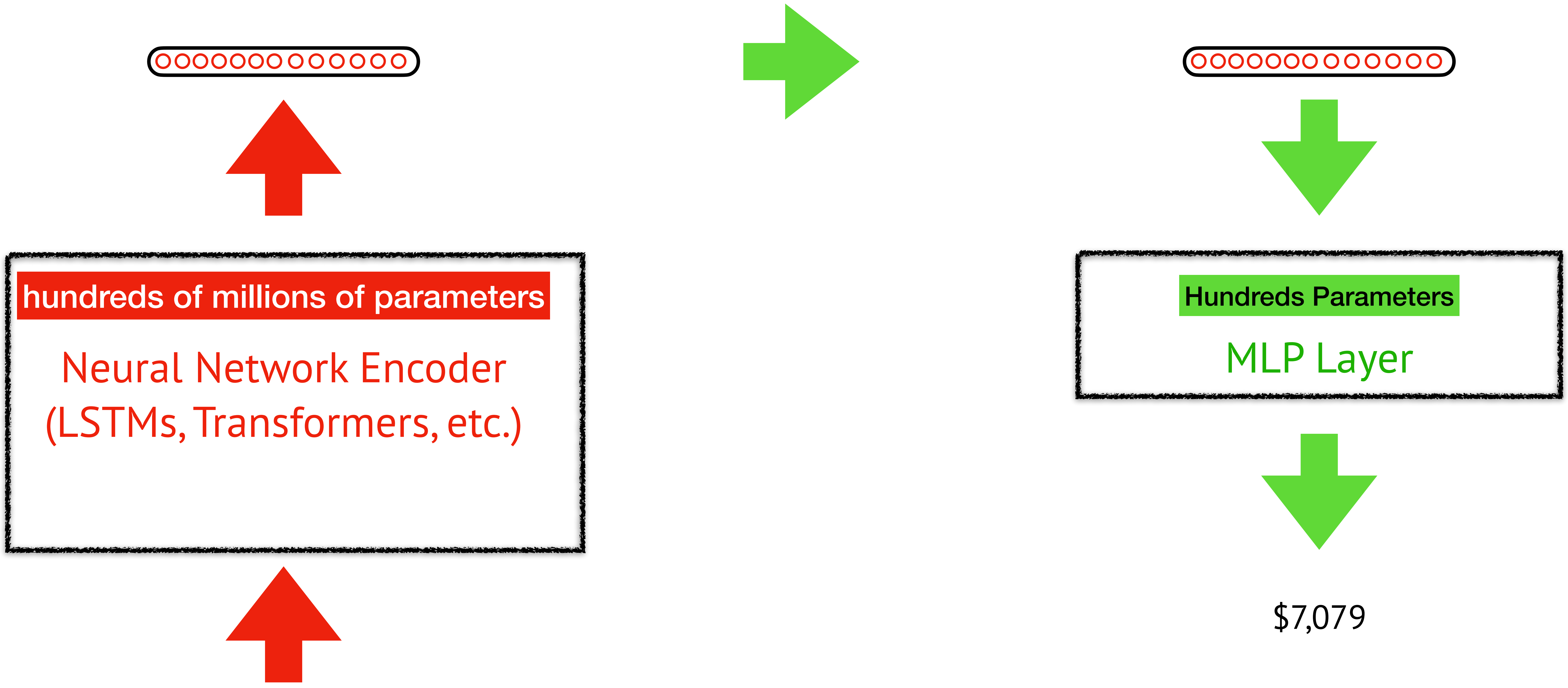
$$\mathbb{E}_{p(x_i, \hat{\boldsymbol{x}}_i)}[p(x_i \mid \hat{\boldsymbol{x}}_i)]$$



Pretraining and Contextualized Word Representations



Pretraining and Fine-tuning



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

This is BERT!

BERT: Bidirectional Encoder Representations from Transformers

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net *Shanghai Jiao Tong U* https://arxiv.org		
4 Aug 04, 2019	XLNet + SG-Net Ve *Shanghai Jiao Tong U* https://arxiv.org		
5 Jul 26, 2019	UPM (sir *Anor*		
6 Mar 20, 2019	BERT + DAE + *Joint Laboratory of HI*		
6 Jul 20, 2019	RoBERTa (*Face*		
7 Sep 12, 2019	RoBERTa+S		
7 Mar 15, 2019	BERT + ConvLSTM + N *Lay*		

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8
3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0
5	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0
6	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4
7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0
8	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9
9	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1
10	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9
11	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1
12	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1
13	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1
14	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1
15	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1
16	廖亿	RGLM-Base (Huawei Noah's Ark Lab)		81.3	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1
17	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1