Deep Learning for Language Understanding: An Introduction

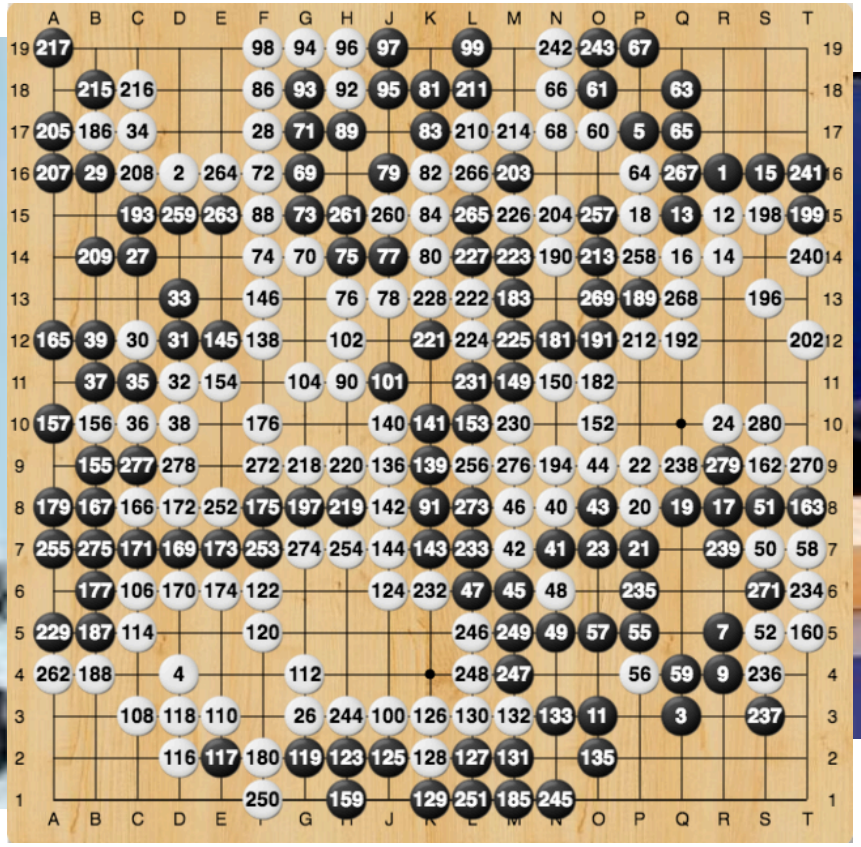
Lei LI

Toutiao Lab

5/29/16

Better than human in GO playing

Via deep reinforcement learning and Monte-Carlo tree search

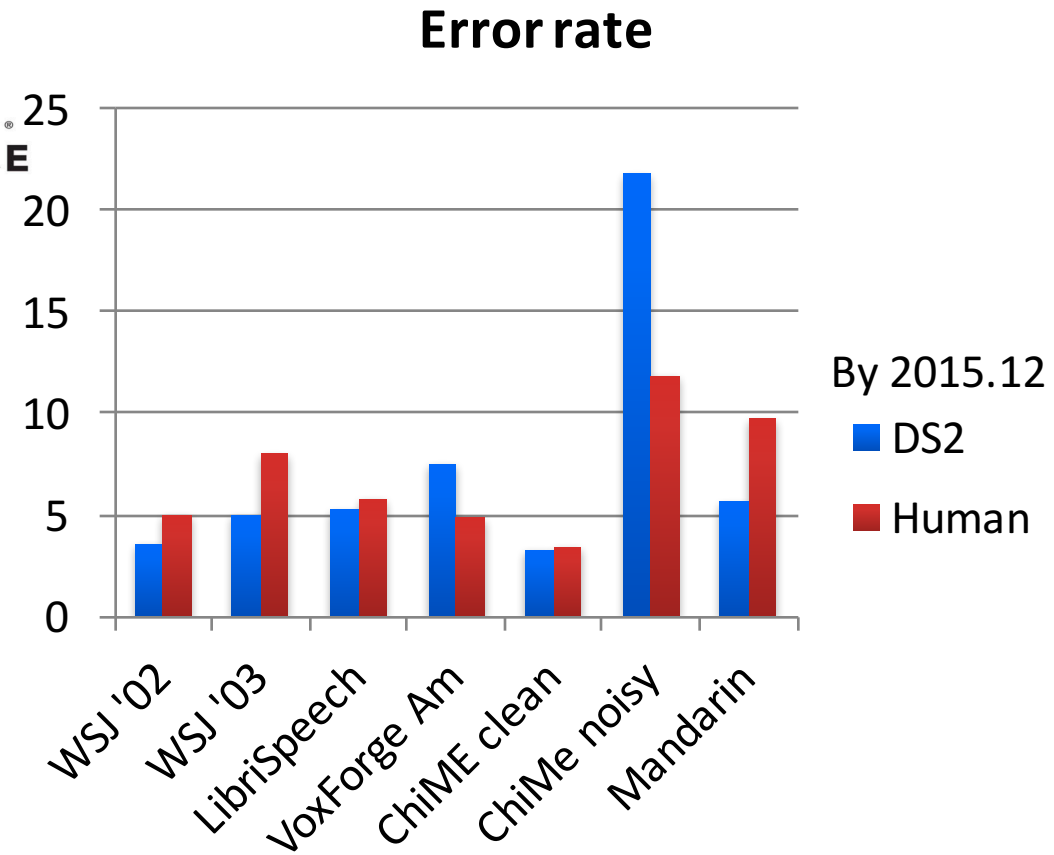


Towards human level in Speech recognition

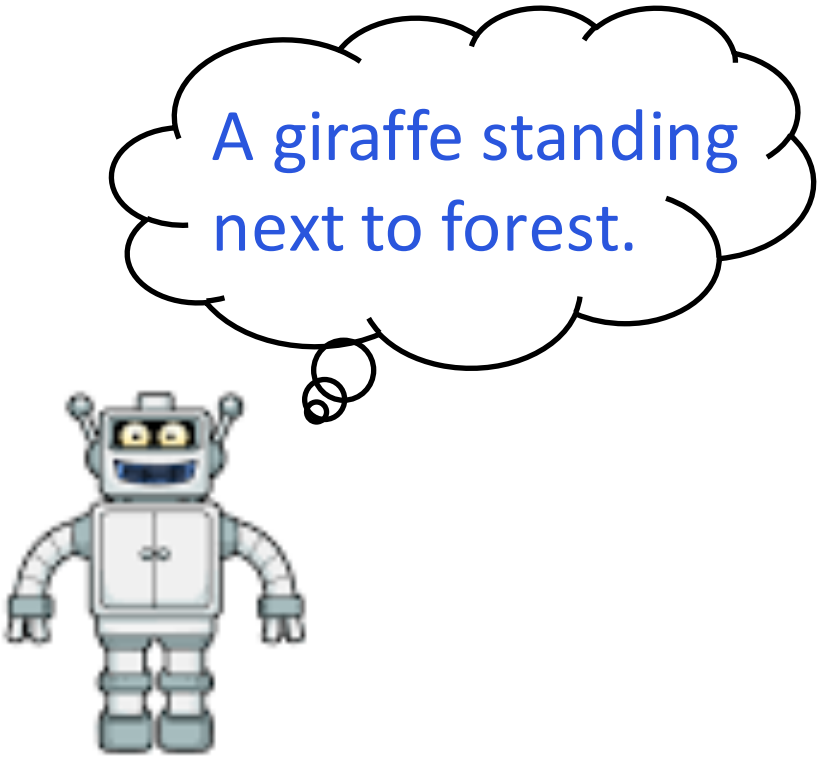
NUANCE

iFLY

Baidu 百度

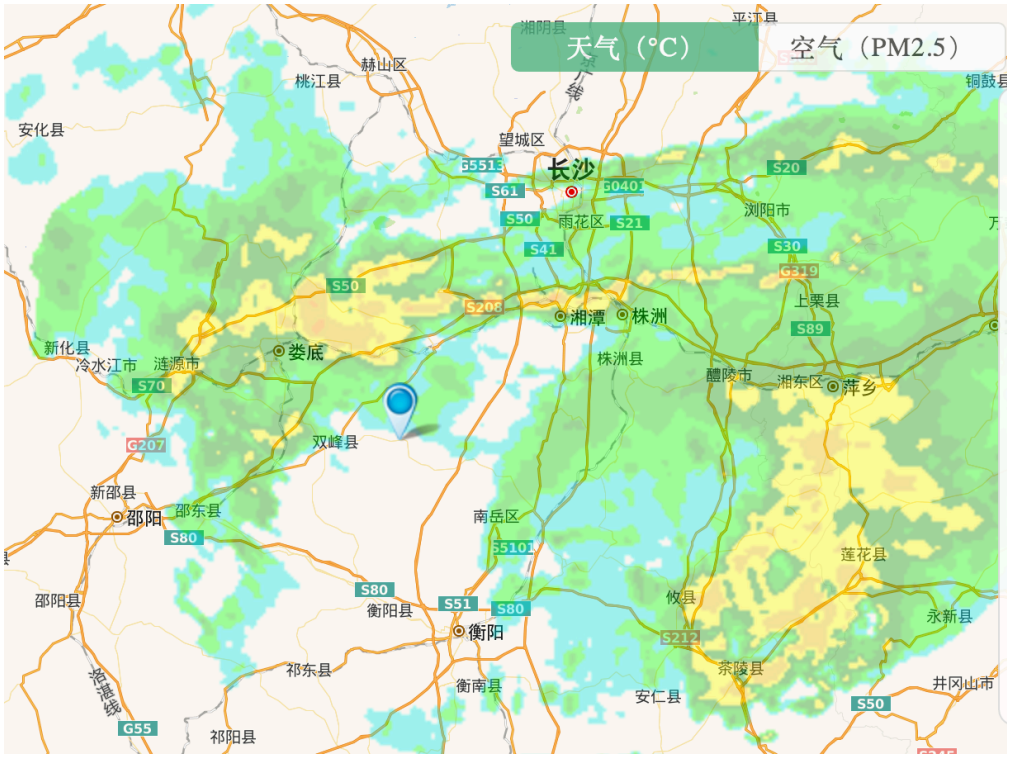


Telling stories in images



Deep Learning for localized minute level weather forecasting

彩云天气

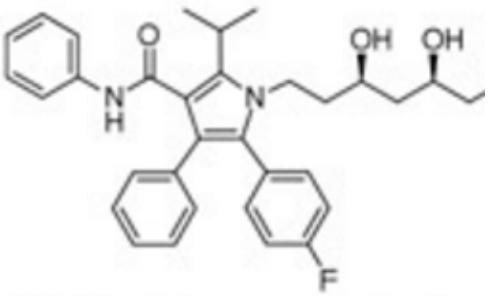


精准位置分钟
级别降雨预报

DL practice at Industry

百度杀毒
Baidu Antivirus

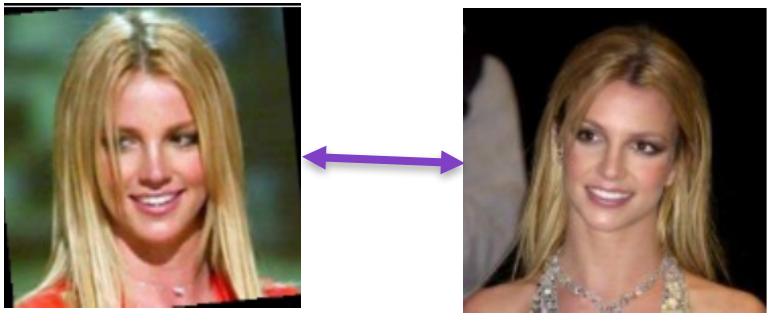
Computational drug discovery with deep learning



Gmail

smart reply

百度医生
导诊科室



music recommendation

SENSETIME

FACE++

DL algorithms work well for

Supervised learning

data

label

X $\longrightarrow$ $f(\cdot)$ $\longrightarrow$ Y

Cat/dog/…

“今天天气不错”

“Today is a nice day”

A giraffe standing next to forest

“打车去故宫”

Deep Learning has advanced Language Understanding Technology

- Neural Language Model
  - Single layer NN for bigram, [Wei Xu and Alex Rudnicky, 2000]
  - Concatenated Word Embedding to predict next word [Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, 2003]
  - RNN Language Model, [Mikolov et al, 2011]
- Basic NLP technology
  - NLP from scratch [Ronan Collerbert, Jason Weston et al 2011]
    - WSJ POS 97.29% acc; CoNLL NER 89.59% F1; CoNLL Chunking 94.32% F1

Language understanding tasks from shallow to deep semantics

- Syntactic parsing
  - Word Segmentation
  - POS tagging
  - Parsing
- Semantic analysis
  - Named entity recognition
  - Sentiment analysis
  - Semantic role labeling
  - Co-reference resolution

- Language generation
- Reading comprehension
- Open-domain question answering
- Machine Translation
- Single/multi-round dialog

Outline

- Part I
  - Artificial Neuron
  - A Single Hidden Layer Neural Network
  - Deep Neural Network
  - Training DNN
- Part II
  - Recurrent Neural Network
  - Adaptive memory and forgetting
  - Deep learning for semantic parsing
  - Single-round dialog
  - Question answering from knowledgebase

Handwriting Recognition

0
1
2
3
4
5
6
7
8
9

Inspired by a biological neuron

Neural networks: massively connected simple units

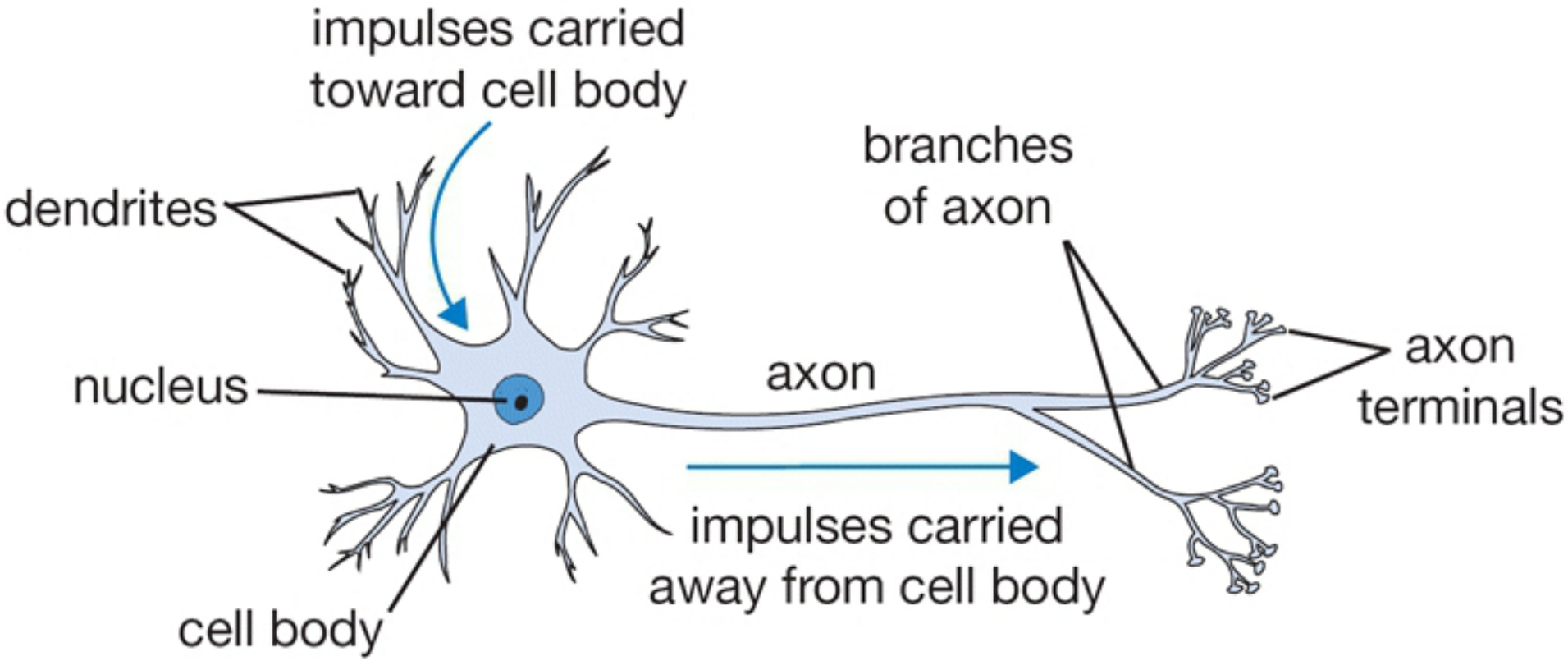
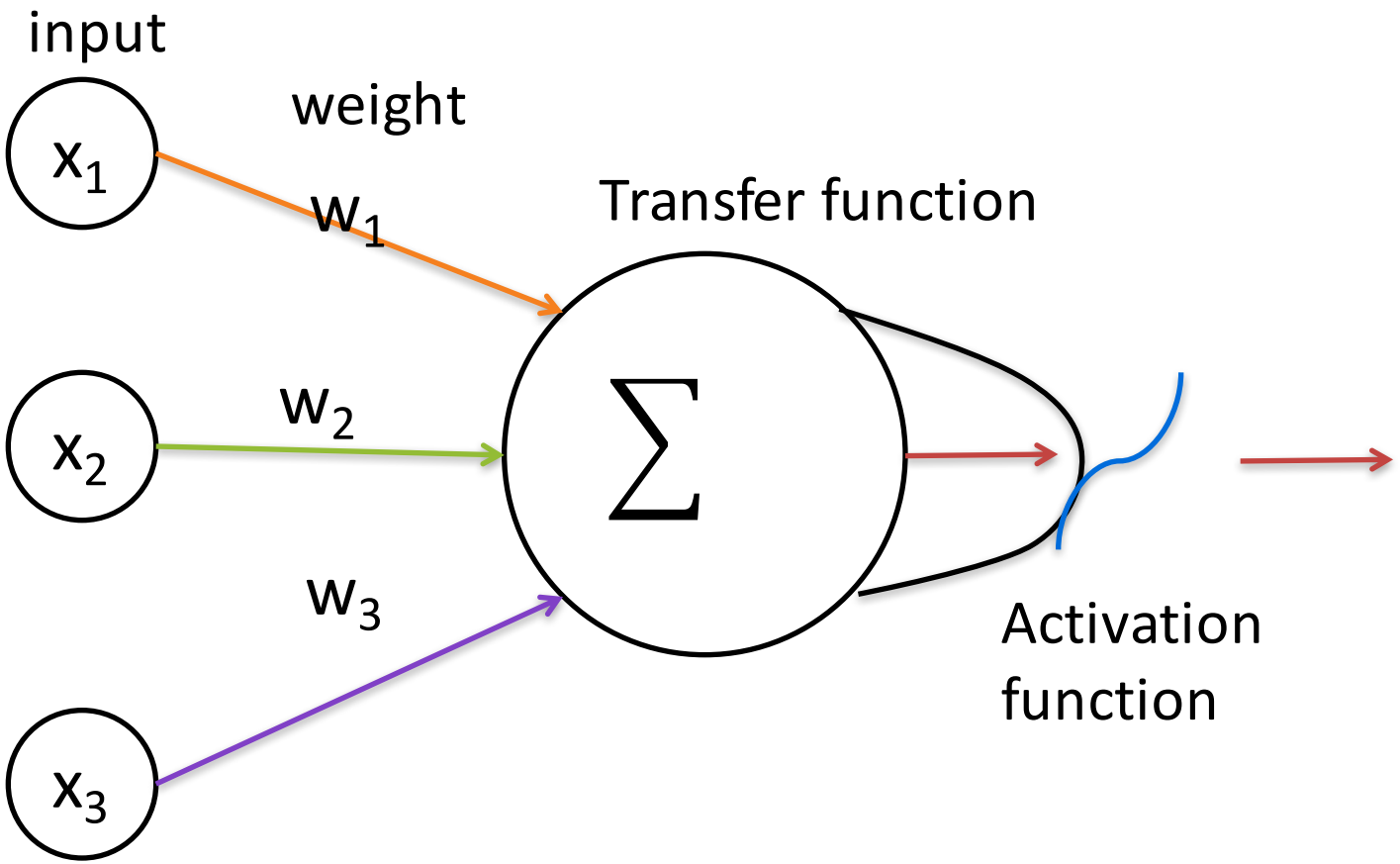


Image credit:
http://cs231n.github.io/neural-networks-1/

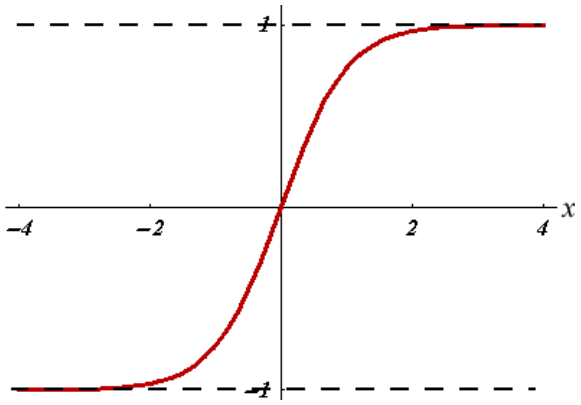
How to model a single artificial neuron?

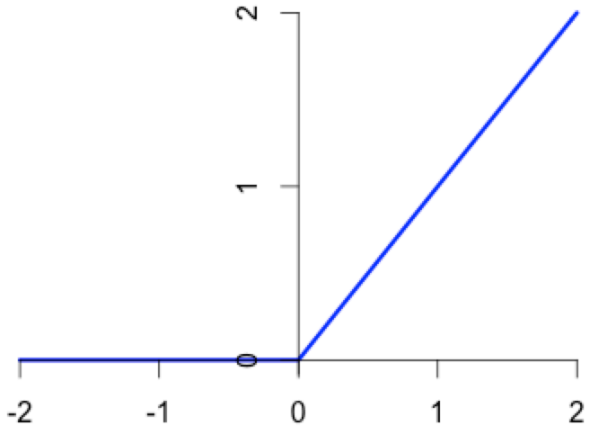


Activation functions

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

$$\text{relu}(x) = \max(0, x)$$

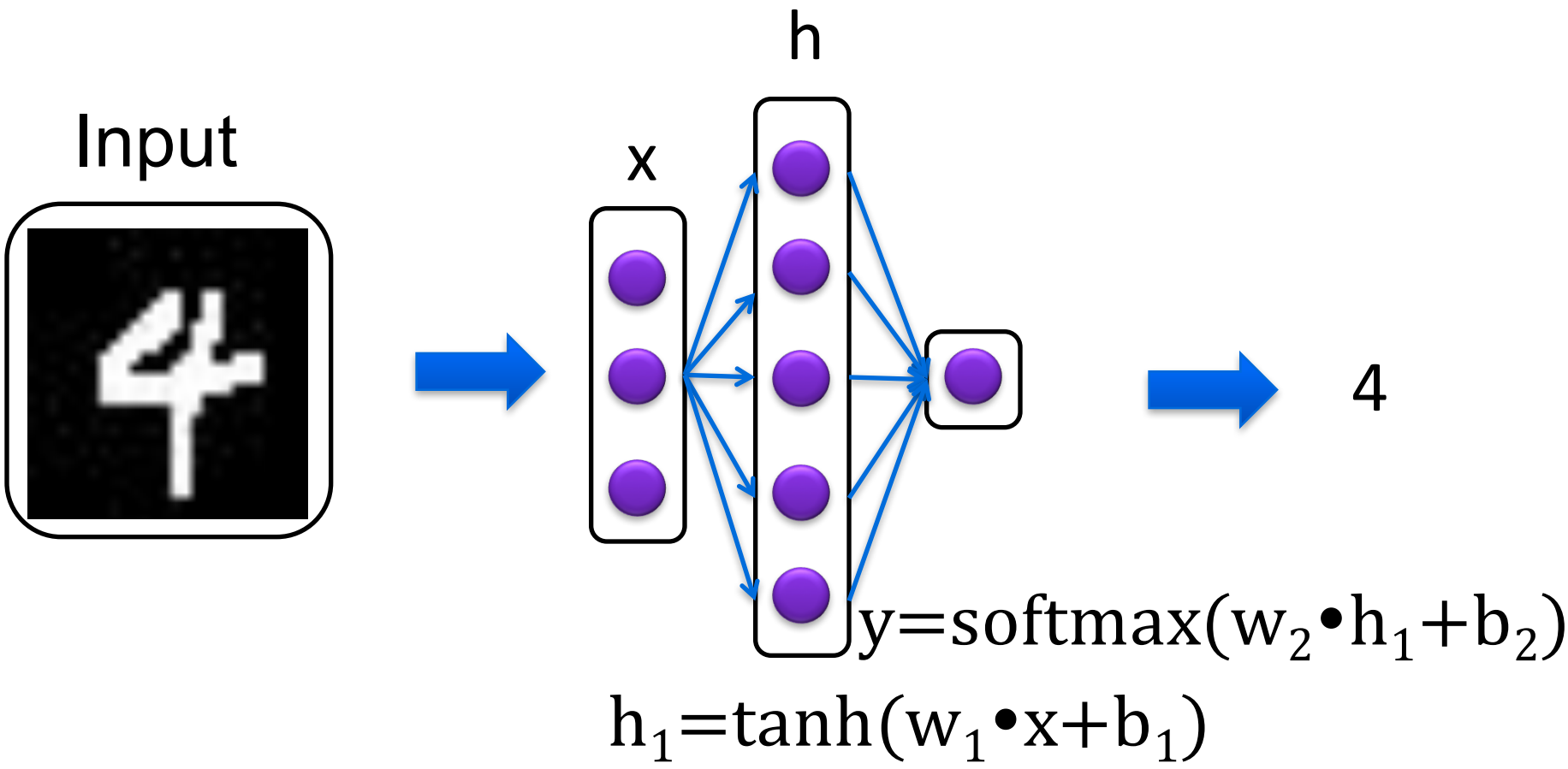




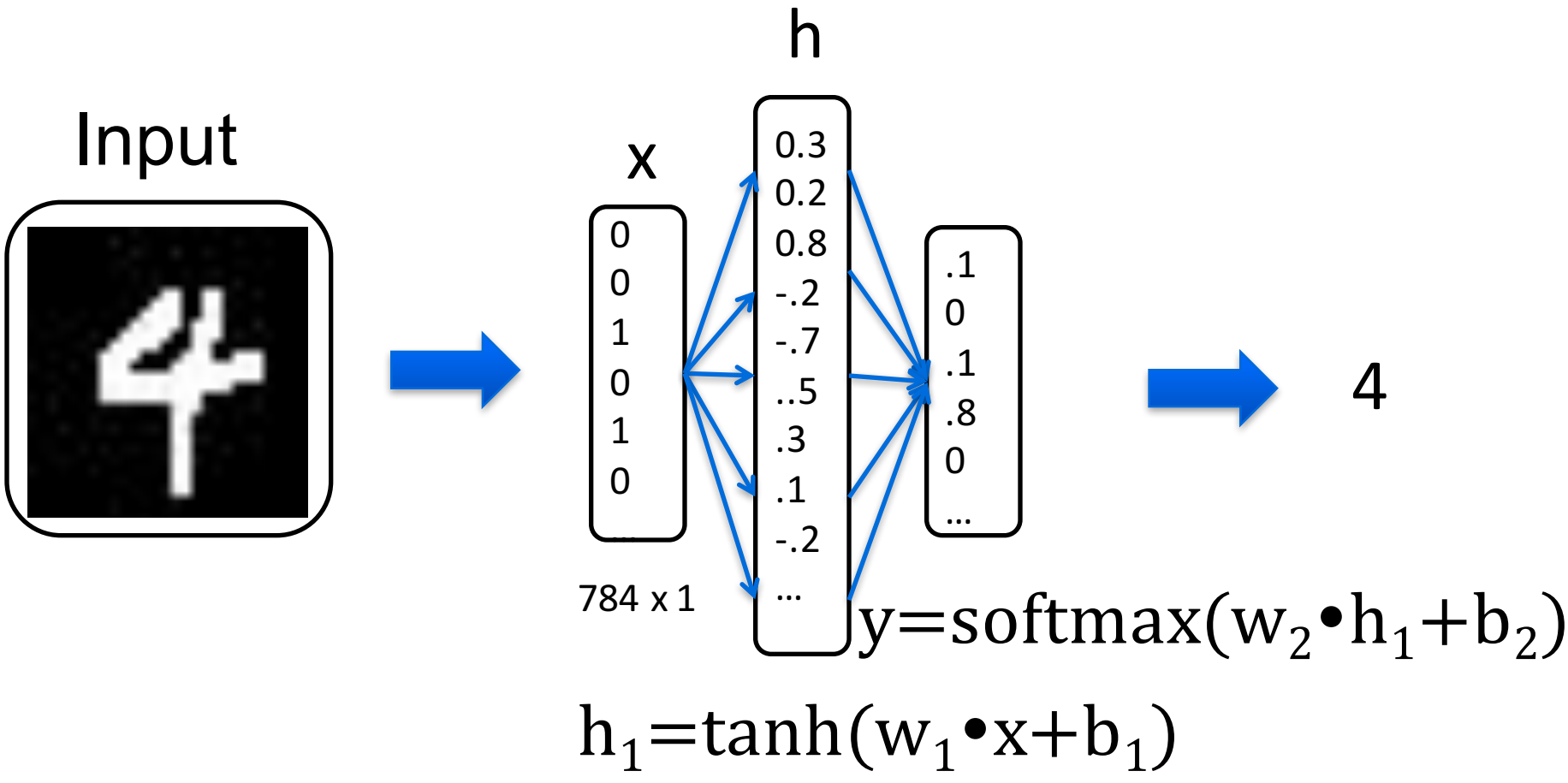
$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum e^{x_i}}$$

Useful for modeling probability (in classification task)

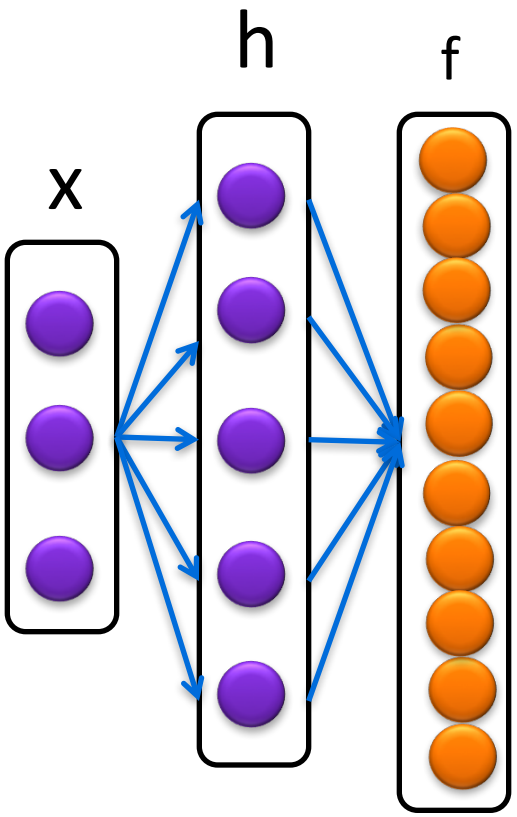
Supervised Learning with Neural Nets



Numerical Example



Objective / Loss: cross-entropy



$$l(f(x_i), y_i) = -\log f(x_i)_{y_i}$$

$f(x_i)$ is a vector (e.g. $\in R^{10}$), representing predicted distribution

y_i is the ground-truth label, can be represented as an one-hot "distribution"
[0,…,0, 1, 0,…,0]

Cross-entropy

$$H(p, q) = -\sum_k p_k \log q_k$$

Cross-entropy

$$H(p,q) = -\sum_k p_k \log q_k$$

Average number of bits needed to represent message in q, while the actual message is distributed in p

OR. roughly

The information gap between p and q + (some const)

Minimizing cross-entropy == diminishing the information gap

$$H\big(y_i, f(x_i)\big) = -\sum_k y_{i,k} \log f(x_i)_k = -\log f(x_i)_{y_i}$$

Ideal case $f(x_i)_{y_i}$ ==> 1.0

Alternative View: Max cond. log-likelihood

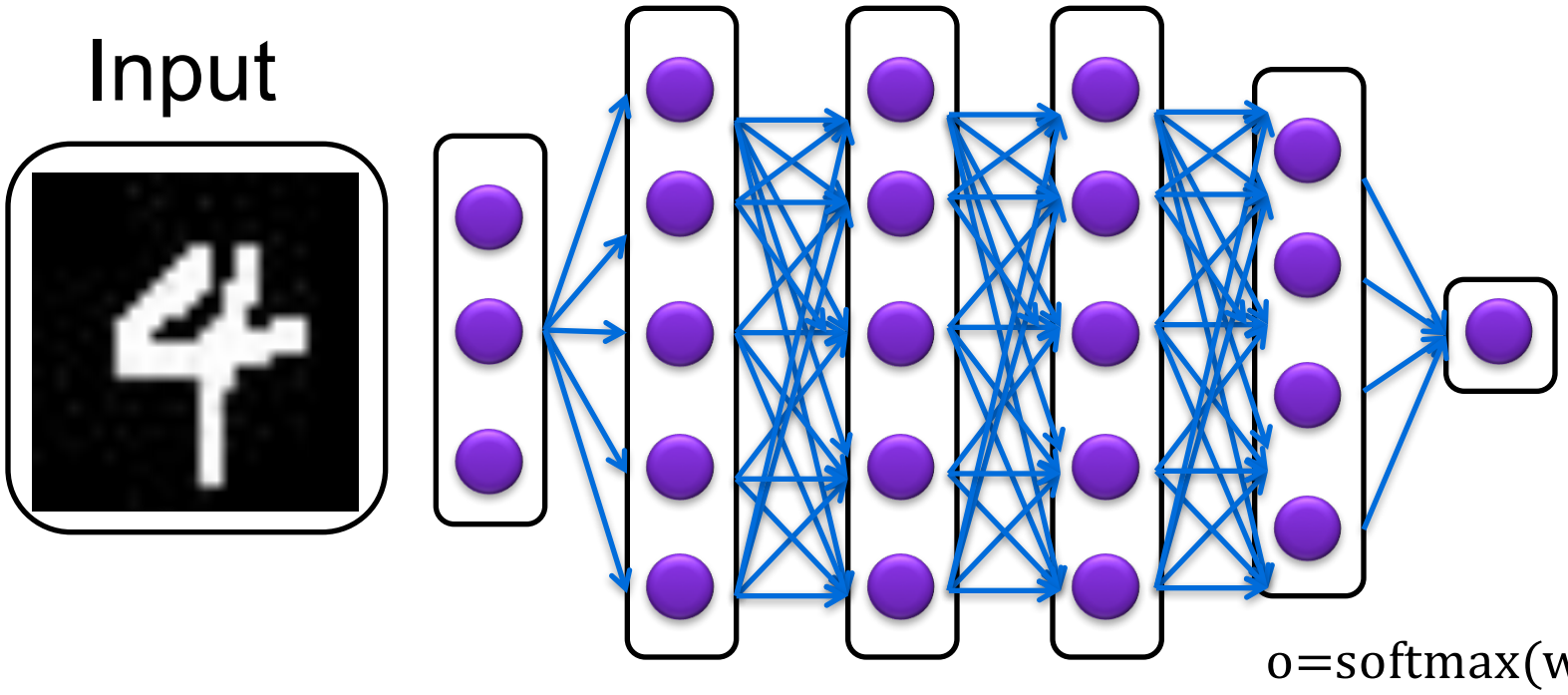
$$\max \log p(y_i|x_i; w) = \sum_k y_{i,k} \log f(x_i)_k$$

Or equivalently

$$\min -\sum_k y_{i,k} \log f(x_i)_k$$

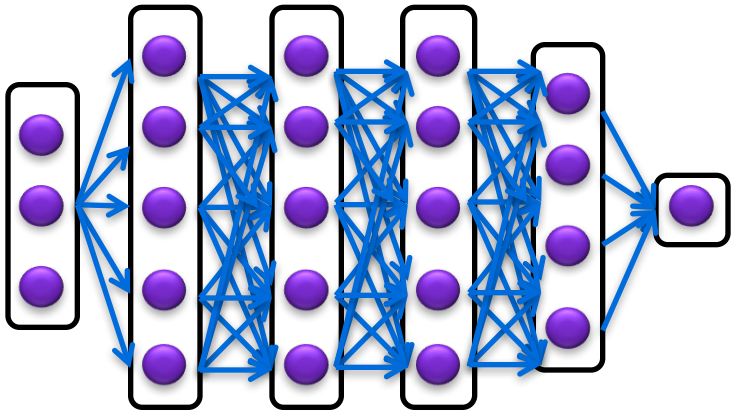
Deep Neural Nets

$$h_1=\sigma_1(w_1 \bullet x+b_1)$$



$$h_2=\sigma_2(w_2 \bullet h_1+b_2)$$

Training DNN

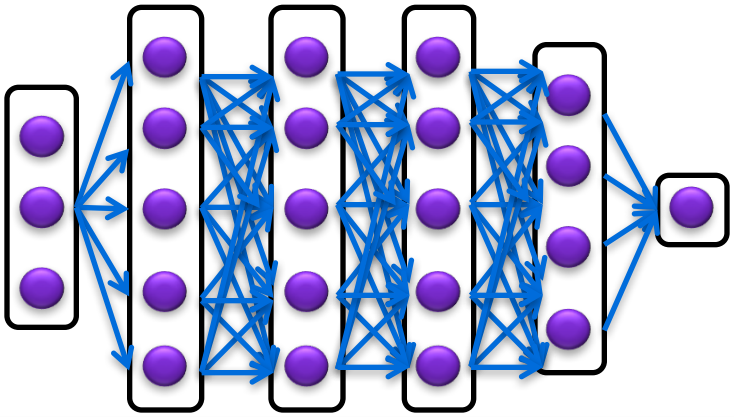


Given: N data points
$(x_1, y_1) \ldots (x_N, y_N)$

Goal: find the best model parameter w, to minimize cost

$$L(w) = \sum_{i=1}^{N} l(f(x_i, w), y_i)$$

Training deep neural nets



To improve efficiency:

Mini-Batch

Compute gradient and update parameters for every batch of k data samples.

Stochastic gradient descent algorithm

for iteration 1 to N (or until convergence)

compute g = $\partial / \partial w_j$

$w = w - a \bullet g$

Step size

gradient

Advanced alg:

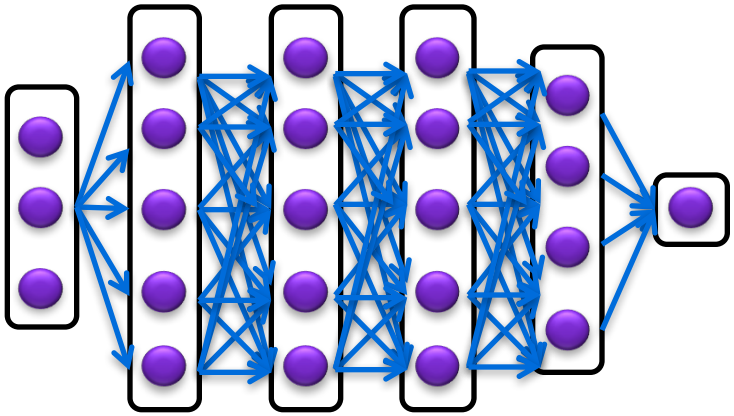
Momentum,

Adagrad,

Adam,

…

Forward and Backward propagation



Chain rule

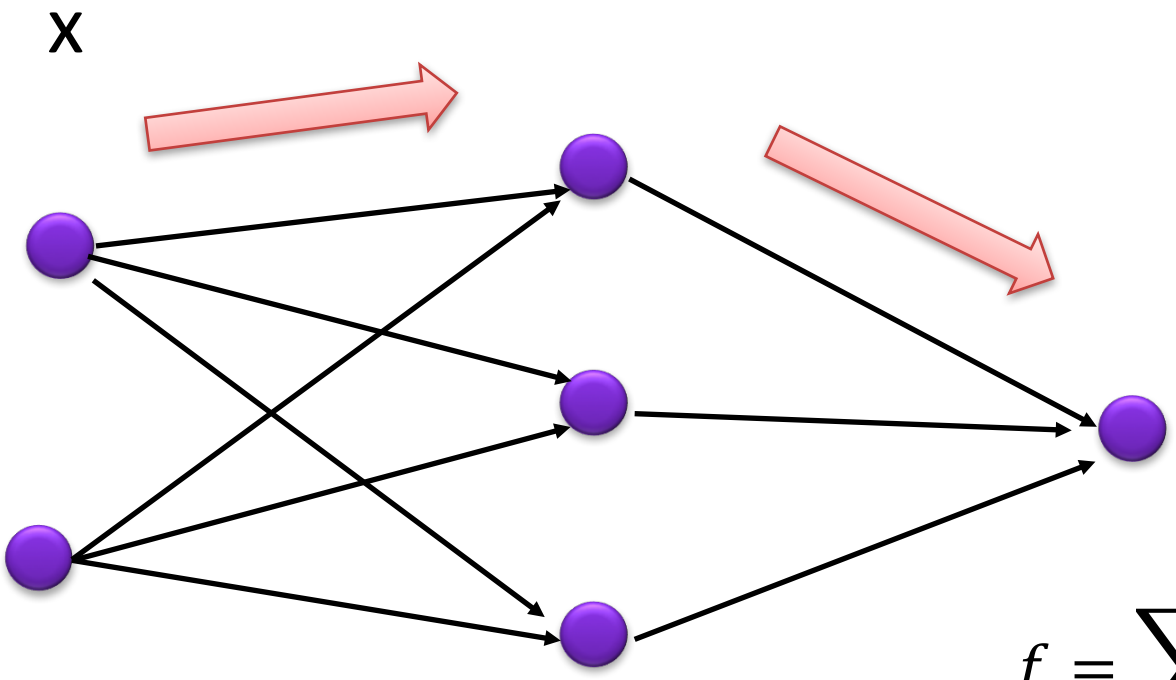
forward pass: computing network prediction

$$h_i = \sigma_i(w_i \bullet h_{i-1})$$

backward prop: computing gradient from layer-wise error

$$\delta_{i-1} = w_i^T \bullet (\delta_i \odot \sigma_i') \qquad \partial / \partial w_j = h_{i-1} \bullet \delta_i^T$$

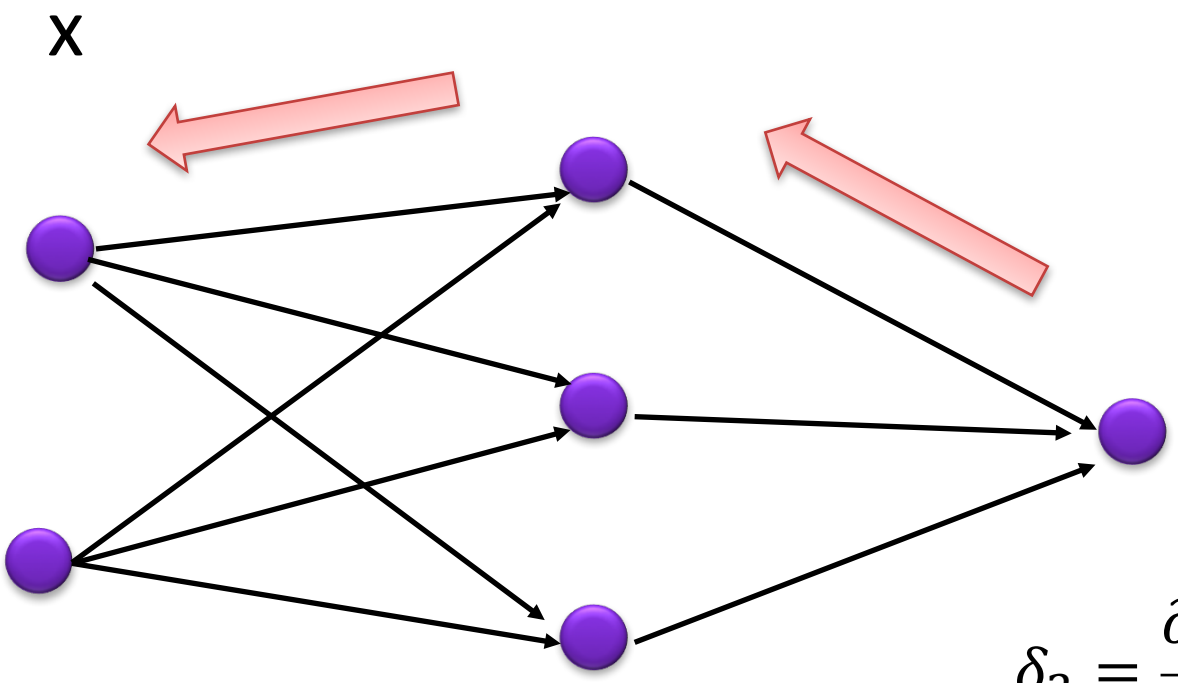
Feedforward computation: computing f(x) given x and parameters w



$$f = \sum h_k$$

$$h = W \cdot x + b$$

Backward propagating error



$$\delta_2 = \frac{\partial l(f, y)}{\partial f}$$

$$\delta_1 = W^T * \delta_2$$

More variation

- Optimization algorithms
  - Momentum
  - Adagrad
  - Adadelta
  - Adam
- Dropout
  - Randomly zeros the output neurons in each layer
- Regularization
  - L1,, L2, to improve generalization

Deep Learning platform

- Tensorflow (Google)
- Torch (NEC, FB)
- Caffe (ucb)
- Theano (U. Montreal)
- MXNet (DMLC, Li Mu et al)

- Provides easy language to construct network
- Rich set of layers, with forward and backward steps
- Library of optimization algorithms
- Many research papers build models based on these

Part II

RECURRENT NEURAL NETWORKS

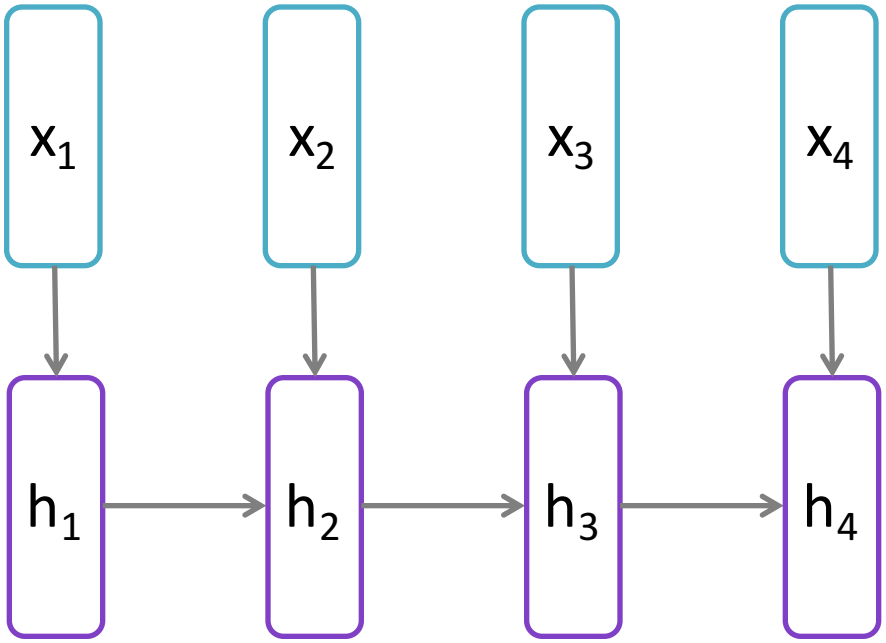
Challenge in processing language

- How to handle variable length of text sequences?
- Solution:
  - Adding Memory to Computation

Recurrent Neural Networks

Basic version: 1 fixed vector memory

- Remember previous state

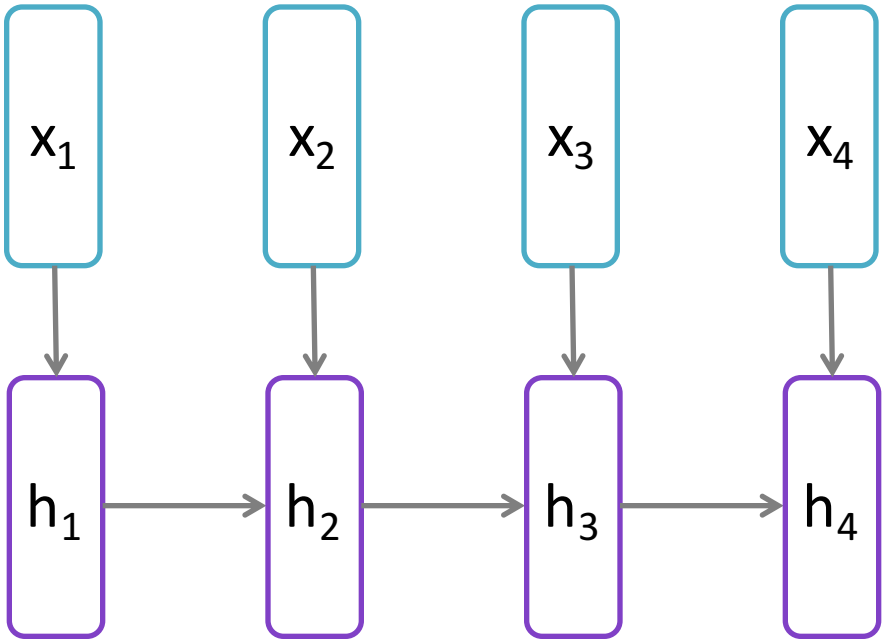


$h_t = f(W \bullet h_{t-1} + U \bullet x_t)$

f = sigmoid, tanh, relu

Recurrent Neural Networks

- Remember previous state

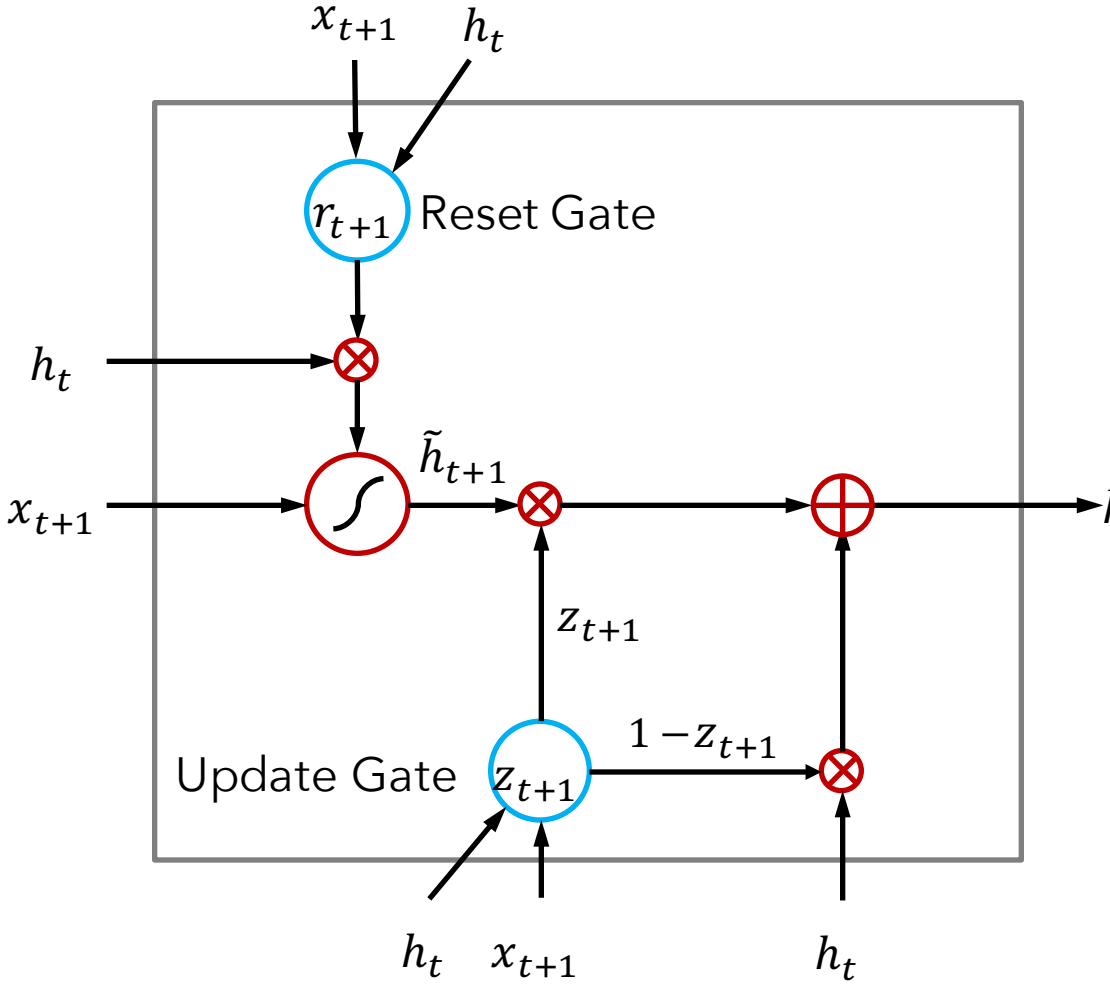


$h_t = f(W \bullet h_{t-1} + U \bullet x_t)$

f = sigmoid, tanh, relu

Gated recurrent unit

Adaptively memorize short and long term information



Input: x_t

Memory: h_t

$$r_{t+1} = \sigma(M_{rx}x_{t+1} + M_{rh}h_t + b_r)$$

$$z_{t+1} = \sigma(M_{zx}x_{t+1} + M_{zh}h_t + b_z)$$

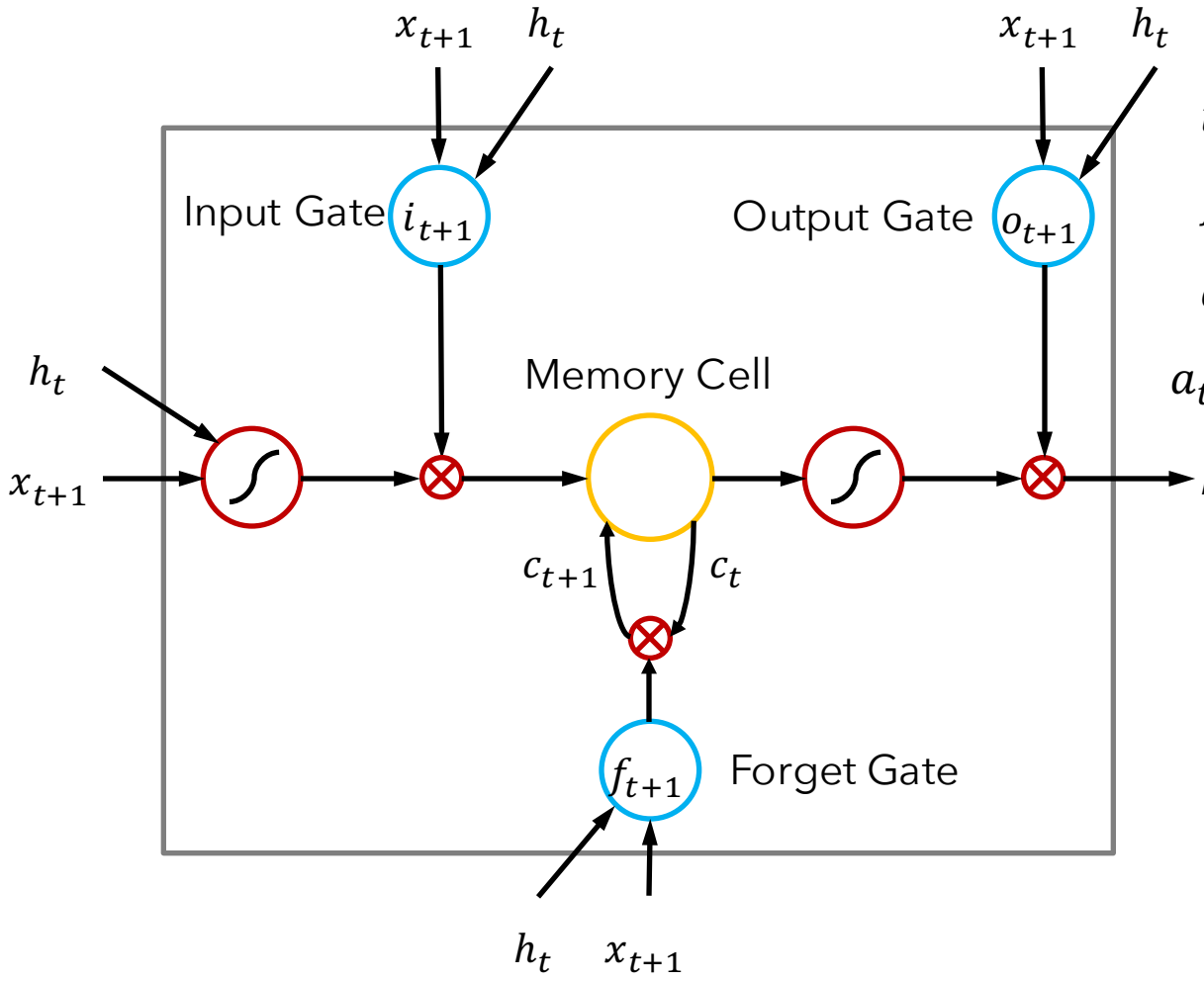
$$\tilde{h}_{t+1} = \tanh(M_{hx}x_{t+1} + M_{hh}(r_{t+1} \otimes h_t) + b_h)$$

$$h_{t+1} = z_{t+1} \otimes \tilde{h}_{t+1} + (\mathbf{1} - \boldsymbol{z_{t+1}}) \otimes \boldsymbol{h_t}$$

[Chung et al 2014]

Long-Short Term Memory (LSTM)

Adaptively memorize short and long term information



$$i_{t+1} = \sigma(M_{ix}x_{t+1} + M_{ih}h_t + b_i)$$

$$f_{t+1} = \sigma(M_{fx}x_{t+1} + M_{fh}h_t + b_f)$$

$$o_{t+1} = \sigma(M_{ox}x_{t+1} + M_{oh}h_t + b_o)$$

$$a_{t+1} = \tanh(M_{cx}x_{t+1} + M_{ch}h_t + b_a)$$

$$c_{t+1} = \boldsymbol{f_{t+1}} \otimes \boldsymbol{c_t} + i_{t+1} \otimes a_{t+1}$$

$$h_{t+1} = o_{t+1} \otimes \tanh(c_{t+1})$$

Deep Learning for (shallow) semantic parsing

Understanding query intention

Wuhan Tech University's nearby handmade noodle house

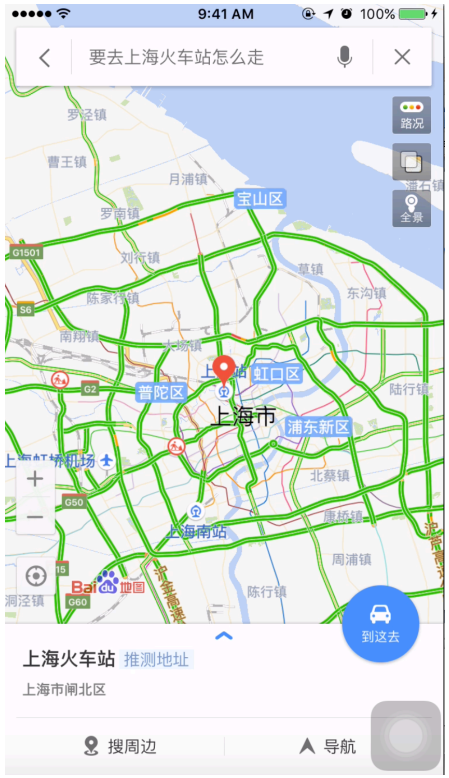
武汉理工大学附近的拉面馆

center keywords

how to go from shanghai to hangzhou

上海到杭州开车怎么走

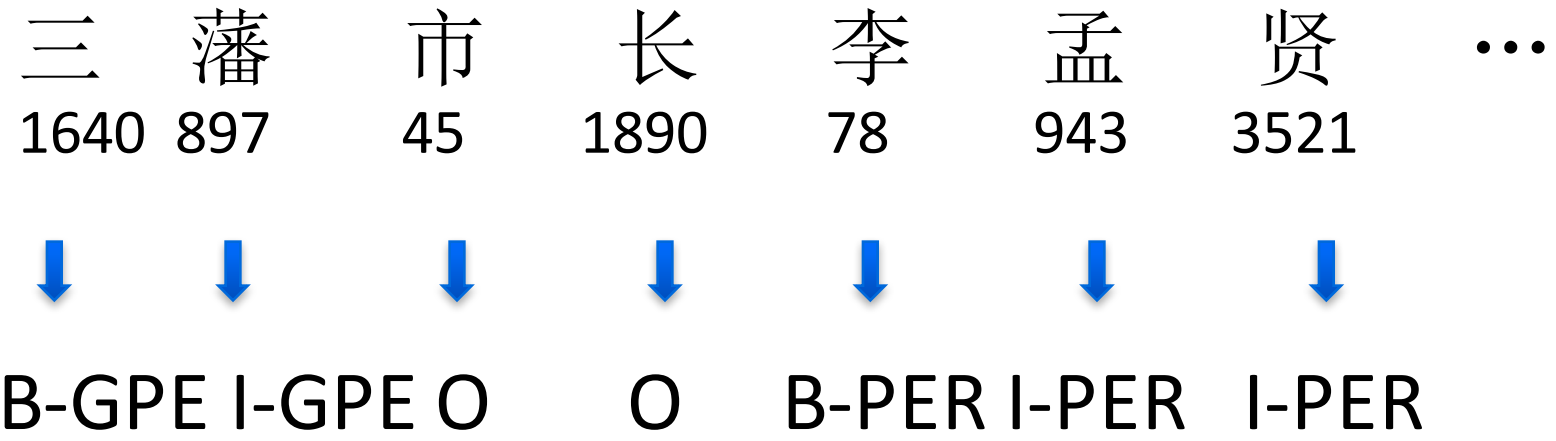
origin destination



Sequence Labelling Task
Named entity recognition

date Location
In **April 1775** fighting broke out between **Massachusetts**
militia units and **British** regulars at **Lexington** and **Concord** .
Geo-Political

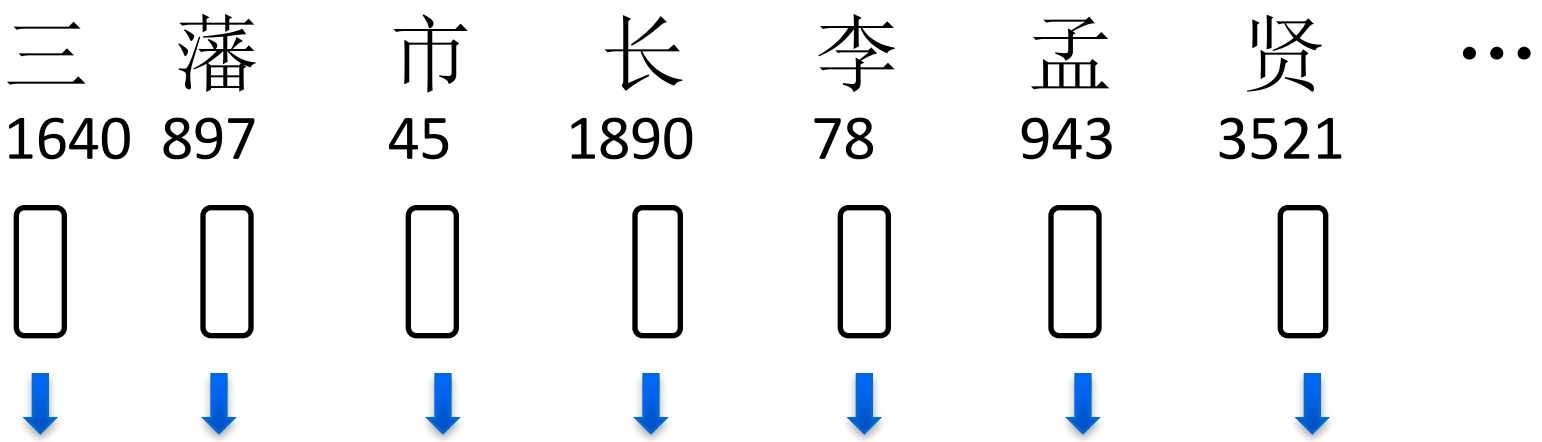
Named entity recognition



Entity chunking scheme: B-I-O Begin of entity chunk, In-middle-of entity chunk, Other (not entity)

Traditional approach

- Conditional random fields with rich expert created features.

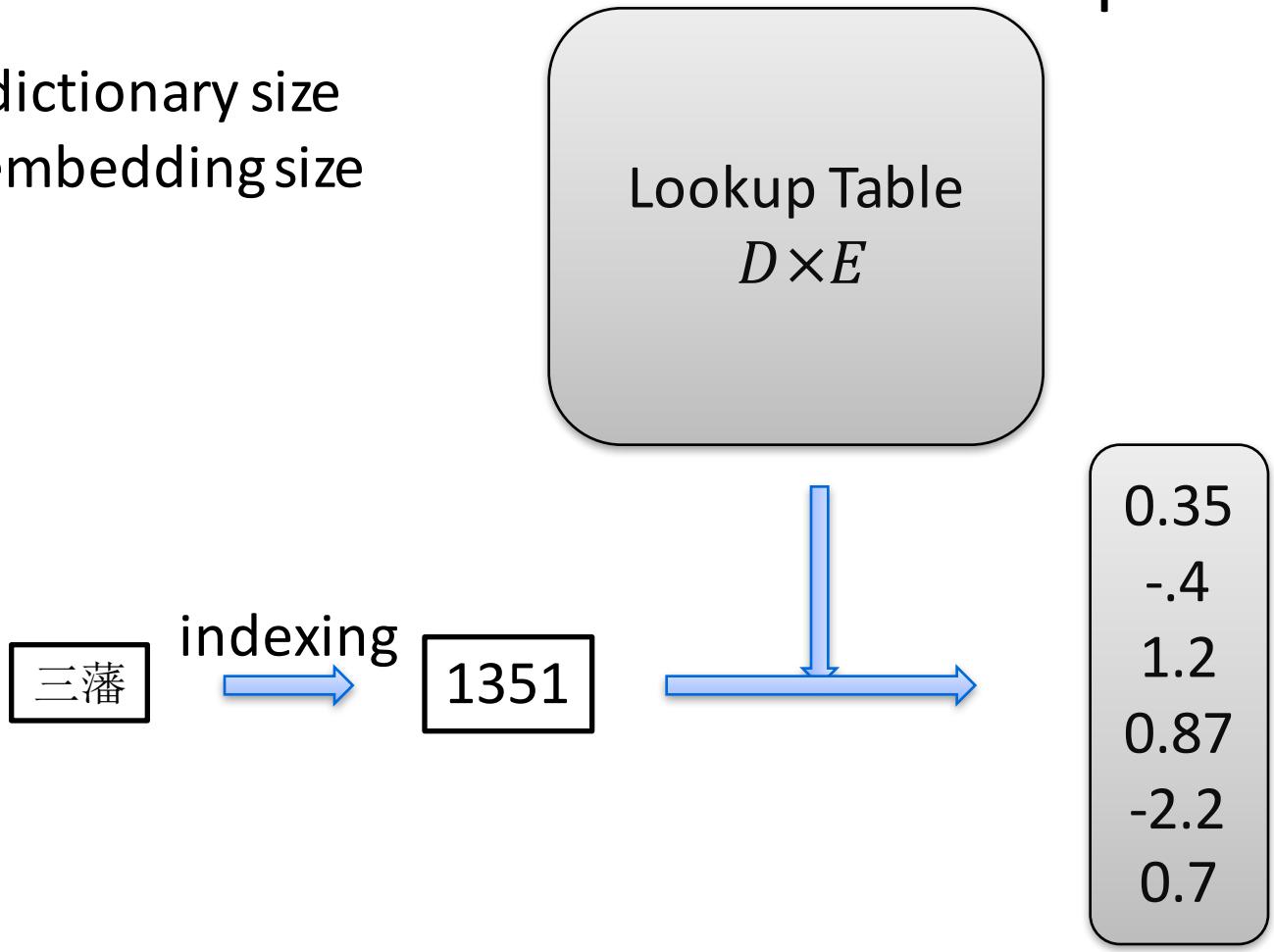


Features: neighboring words,
POS of current word and neighboring words,
Lexical features etc.

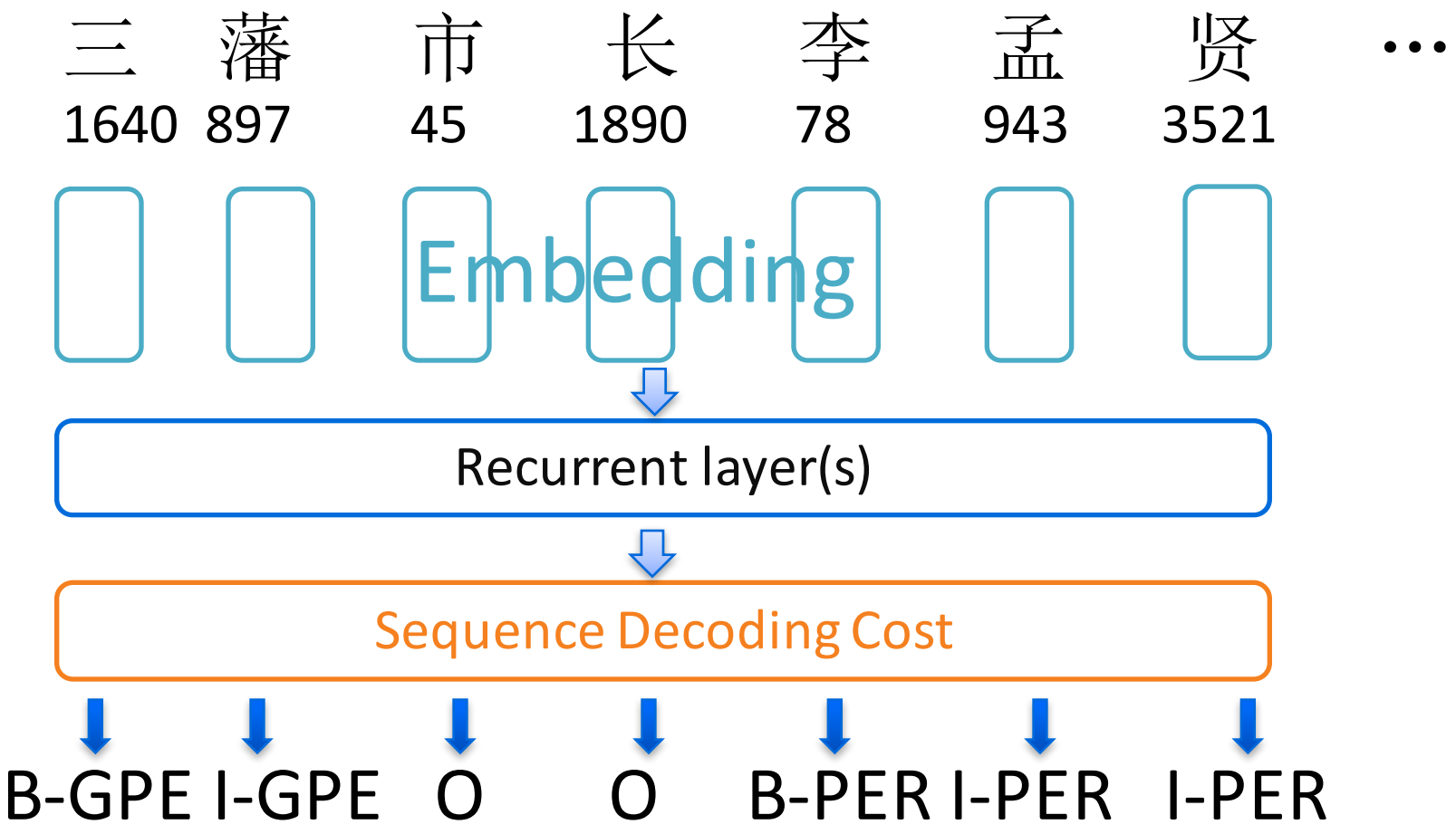
Embedding

- Map every token in the dictionary to a vector.
- The vectors are stored in a lookup-table

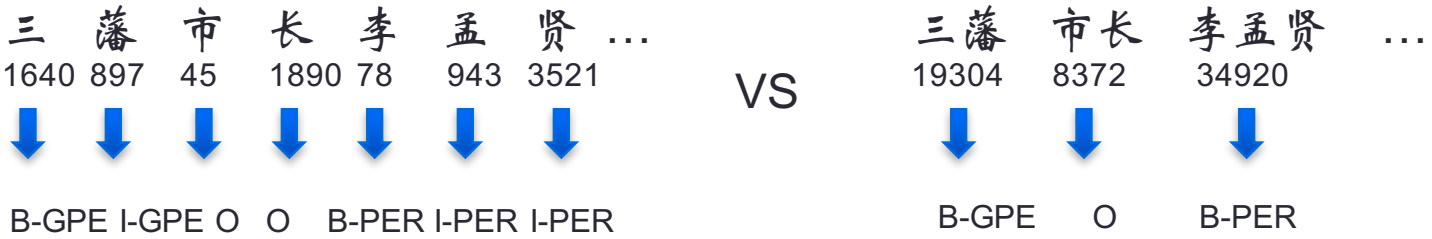
D: dictionary size
E: embedding size



End-to-end training with minimal linguistic features



Chinese: Word or Character?



Vocabulary size : ~4,000 VS ~43,000

Highest Validation Result (18-class)

Basic Model	Precision	Recall	F1
Char-based	77.21	69.05	72.90
Word-based	74.25	61.18	67.08

Char-based is better!

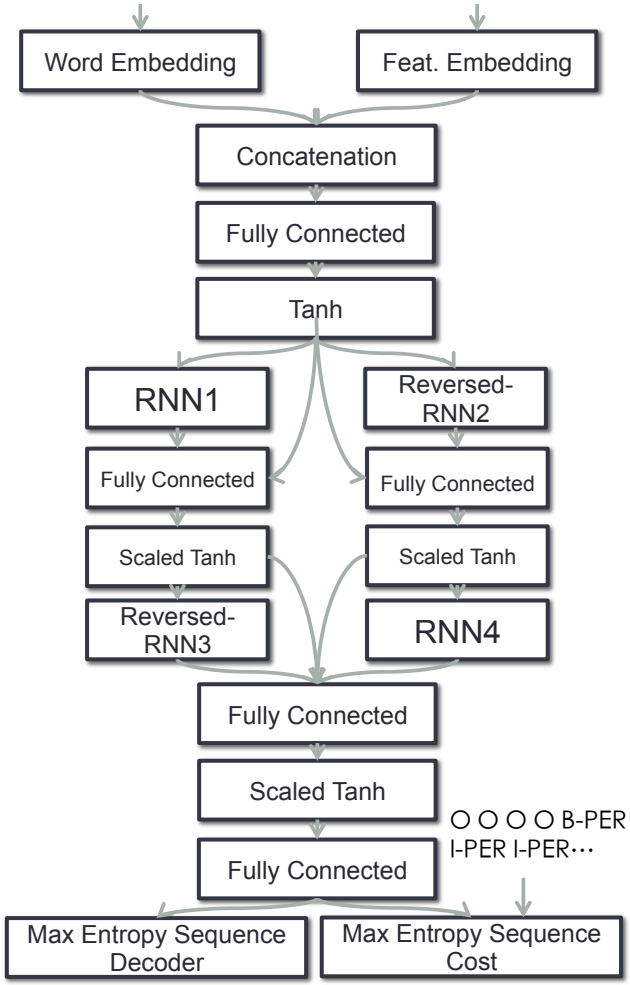
Data & Metric

- Chinese OntoNotes data (release 5.0).
  - ~ 45,000 lines
  - ~ 1,300,000 characters (~ 4000 unique)
  - Training : Validation : Testing = 10 : 1 : 1 (# of chars)
  - 18 Types of Named entities:
    - Cardinal, Date, Event, GPE, FAC, Language, LAW, LOC, Money, NORP, Ordinal, ORG, Percentage, Person, Product, Quantity, Time, Work-of-art
- Standard F1 score
  - Precision: # chunks correct / # chunks produced
  - Recall: # chunks correct / # chunks in the dataset
  - F1 = 2 * P * R / (P + R)

To utilize large unlabeled data

- Pre-trained embedding
- Multi-task training with language modeling

Complete NER Model



Chinese NER

OntoNotes Data 4-class:

Model	P	R	F1
Bi-NER-WA* Wang et al.	84.42	76.34	80.18
RNN-2b with WS ours	84.75	77.85	81.15

* Wang et al used bilingual data

OntoNotes Data 18-class:

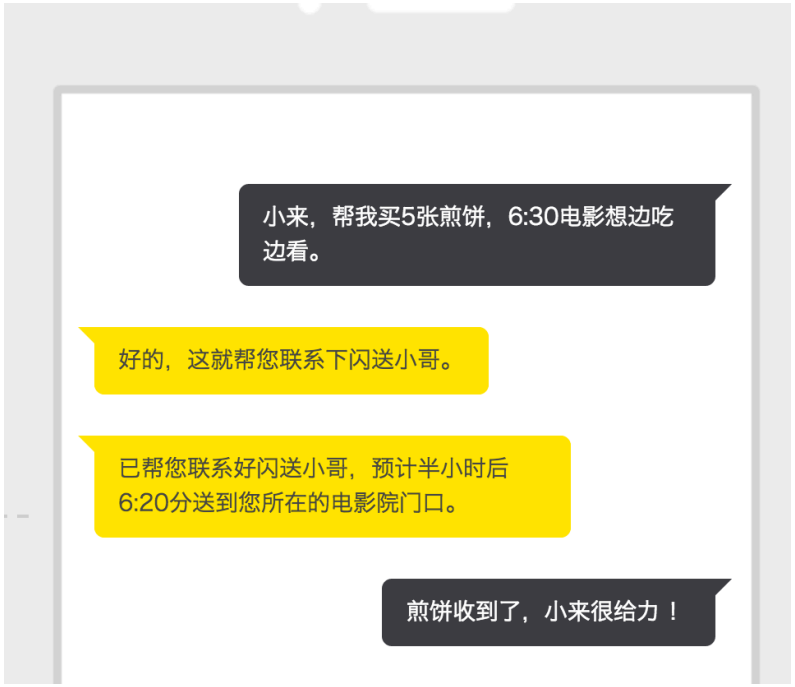
Model	P	R	F1
Sameer Pradhan et al.	78.20	66.45	71.85
RNN-2b with WS ours	78.69	70.54	74.39

[Zefu Lu, Lei Li, Wei Xu, 2015]

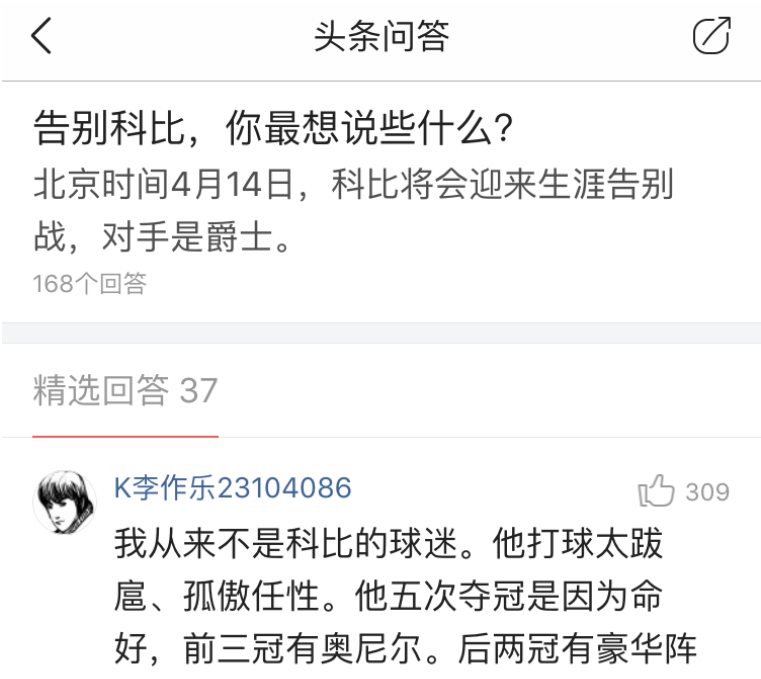
Deep single-round dialog models

Generating Natural Response using deep learning

Personal assistant



Community chatting & QA



Casual chat generated by neural networks

今天午饭好好吃好开心！

So happy to have delicious lunch today!

我也要吃！

I want to eat too!

土豪我们做朋友吧

Let us befriend, rich guy

我不是土豪

I am not rich

你喜欢一见钟情还是日久生情

Would you prefer falling in love at first sight or developing love over time?

一见钟情吧

Falling in love at first sight probably

星球大战好看吗？

Is Star Wars worth watching?

不是很好看

Not very much

[Fei Jiang & Lei Li 2015]

What about longer utterance?

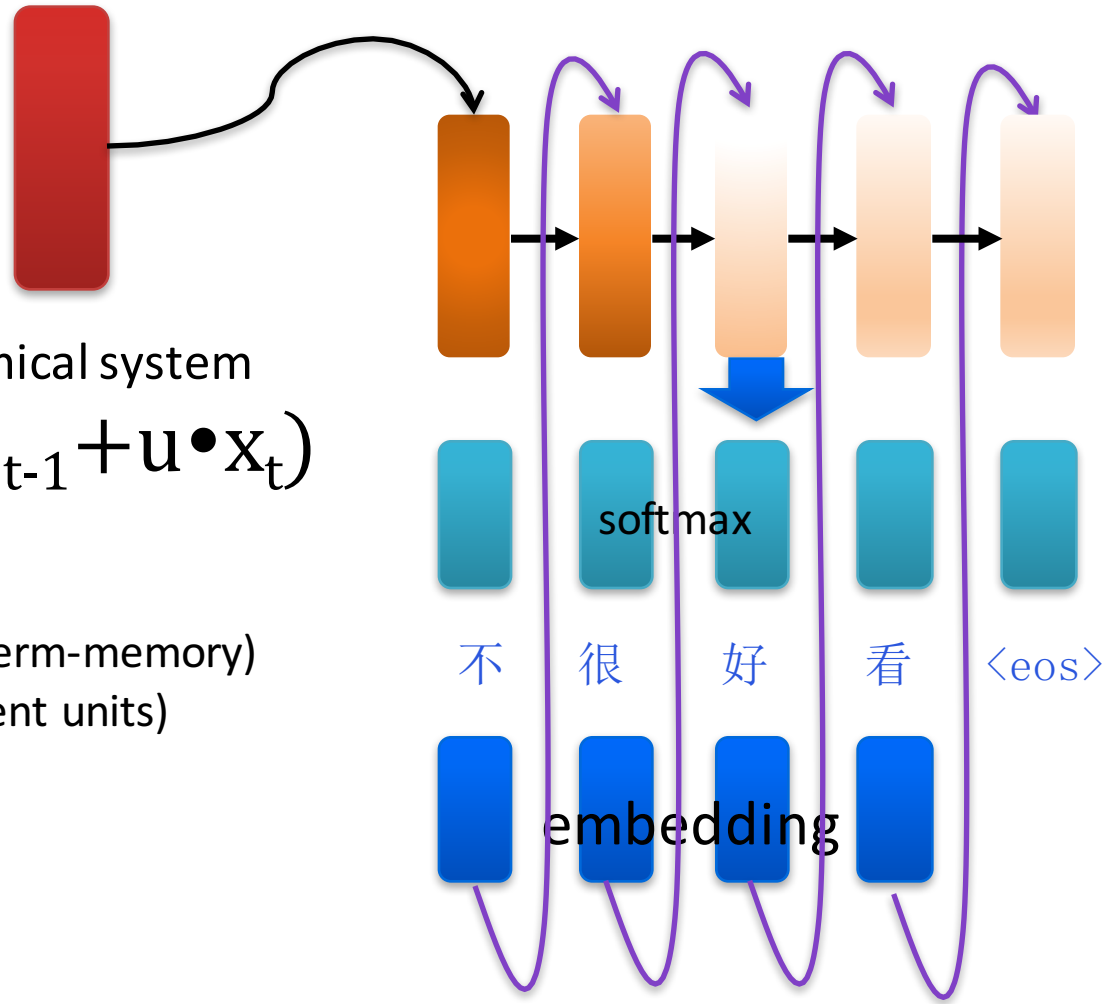
曾经有一段真挚的爱情摆在我眼前，我没有去珍惜，等到失去了才后悔莫及，尘世间最痛苦的事莫过于此，如果上天能给我一个再来一次的机会，我希望能对那个女孩说我爱你，如果非要给这爱加一个期限的话，我希望是一万年。

愿你安好.

I once let the truest love slip away from before my eyes,
Only to find myself regretting when it was too late,
No pain in the world comes near to this,
If only God would give me another chance,
I would say to the girl, I love you!
If there had to be a limit of time,
I pray it's ten thousand years.

Bless you.

RNN for language generation

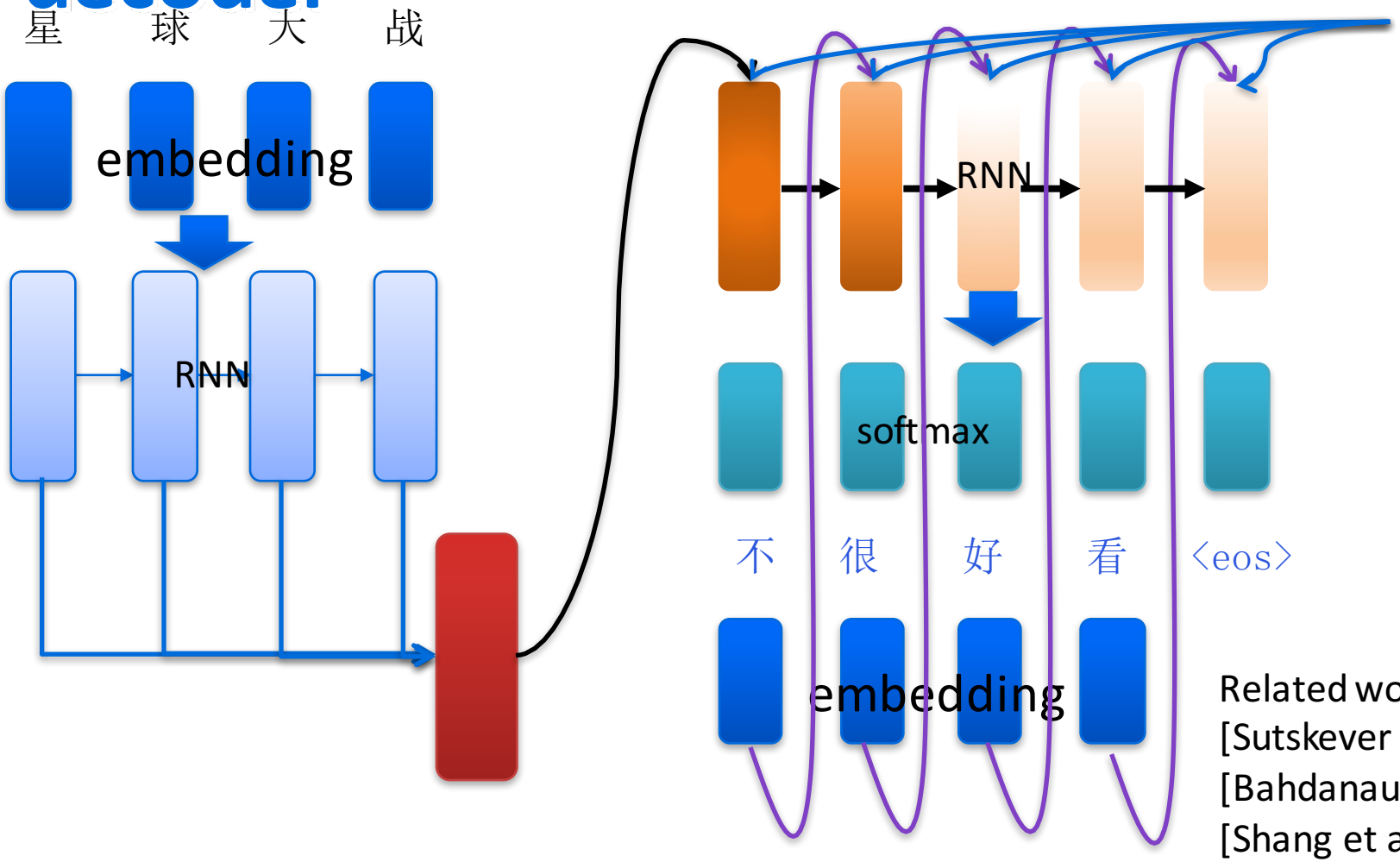


Nonlinear dynamical system

$$h_t = \sigma(w \bullet h_{t-1} + u \bullet x_t)$$

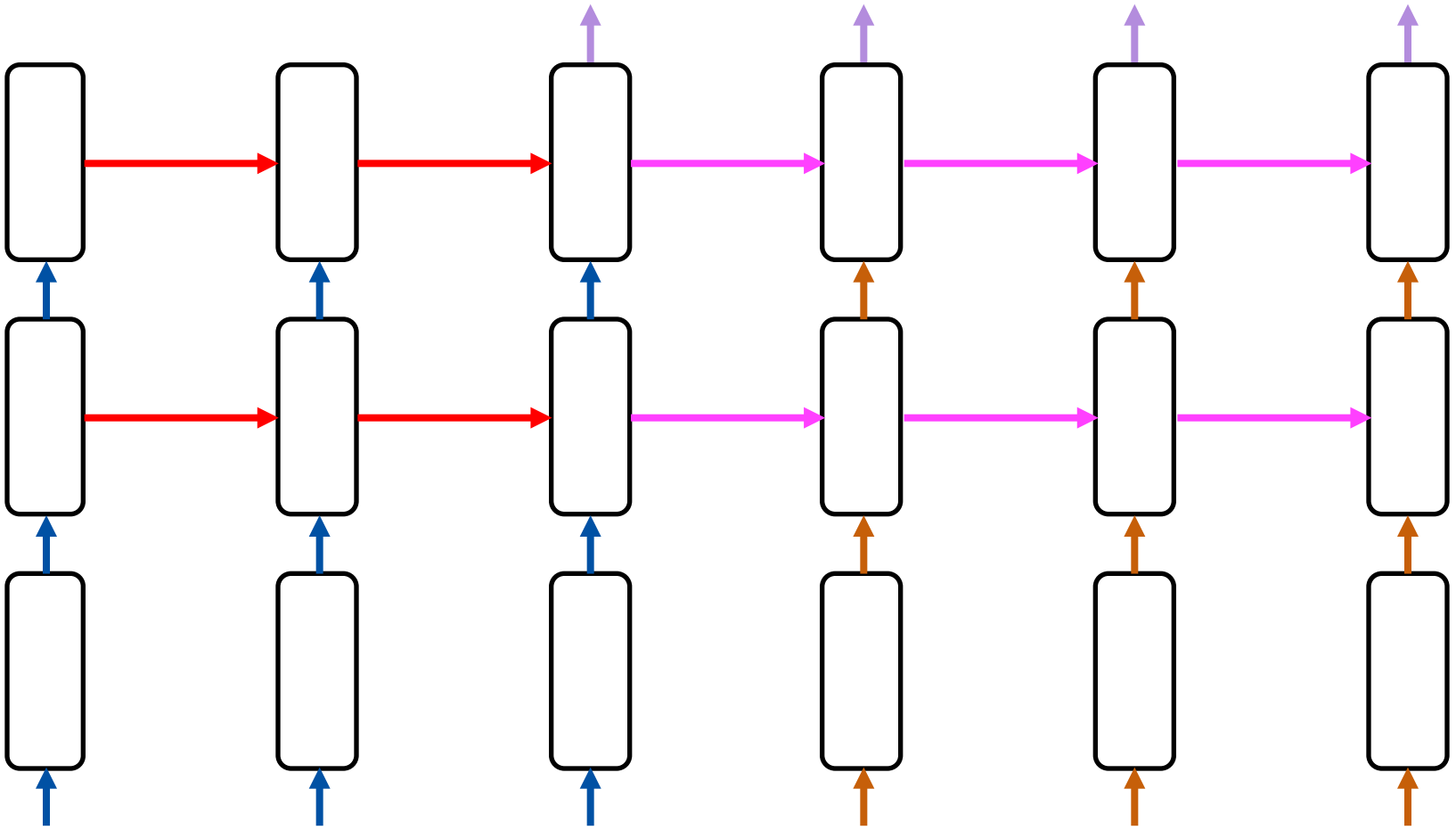
Alternatives:
LSTM (long-short-term-memory)
GRU (gated recurrent units)

Emotional Neural dialog encoder-decoder

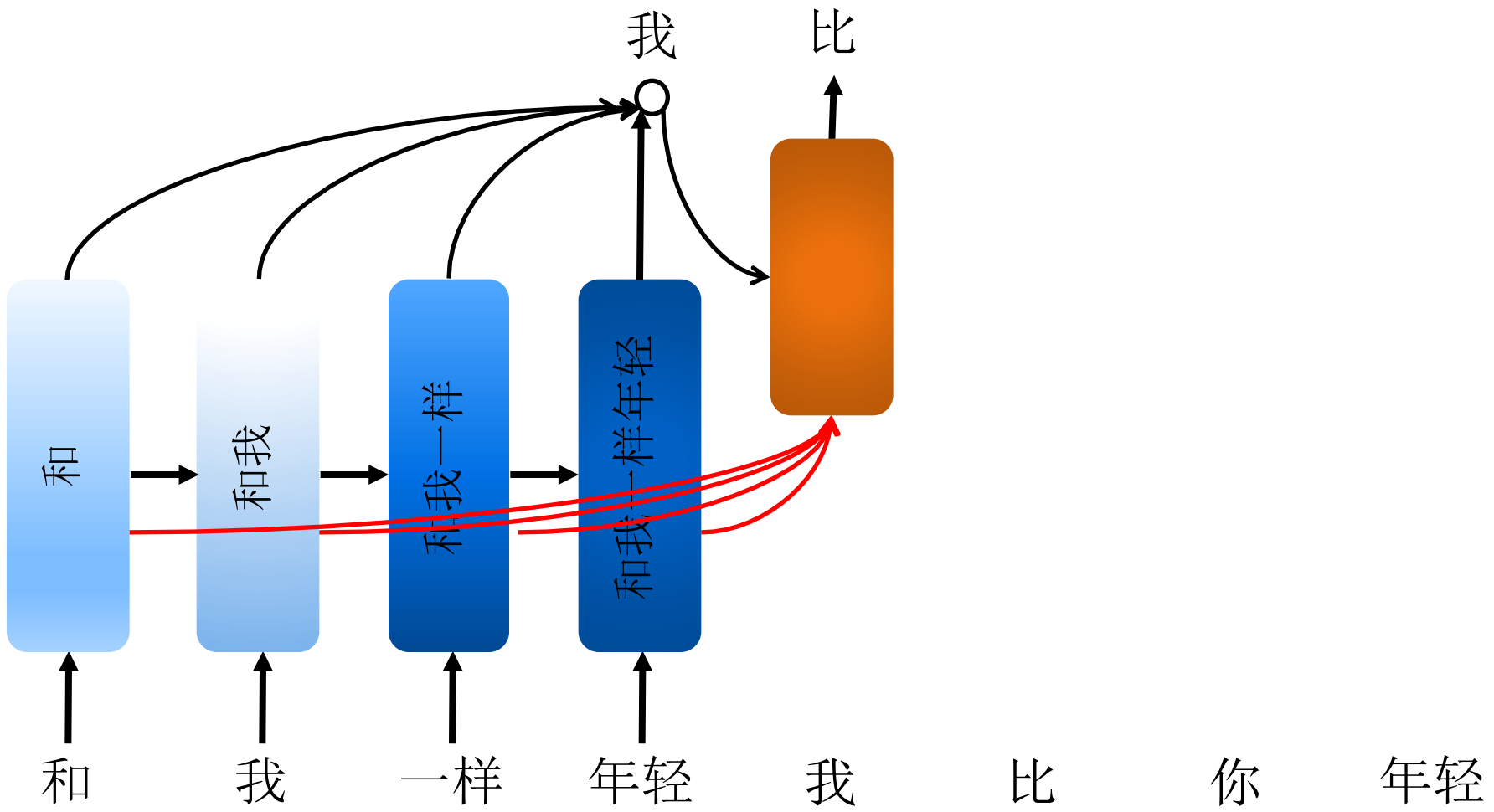


Related work
[Sutskever et al. 2014]
[Bahdanau et al. 2015]
[Shang et al. 2015]
[Vinyals & Le. 2015]

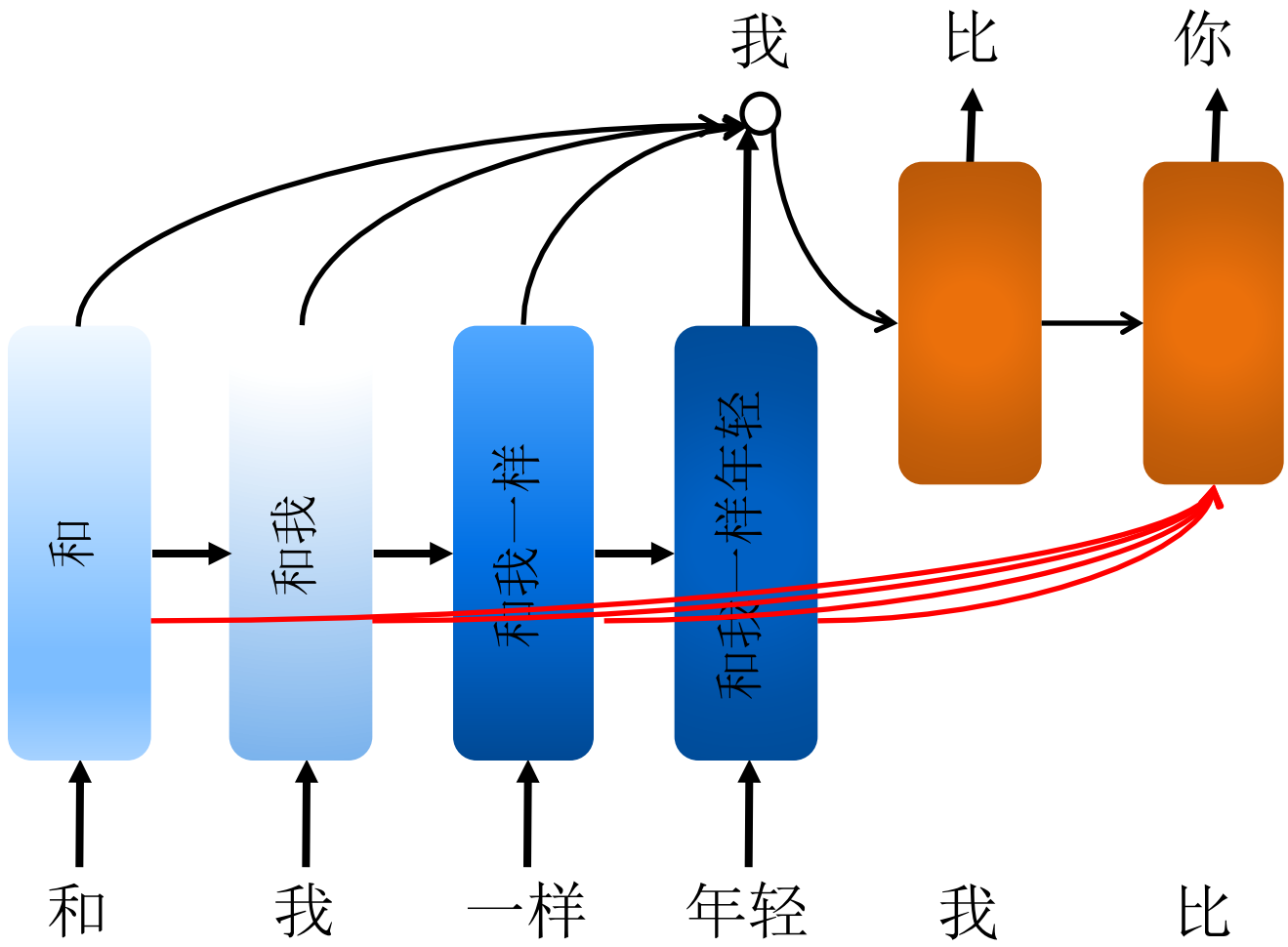
Stacked LSTM for seq-2-seq



Decoding (generation) with attention



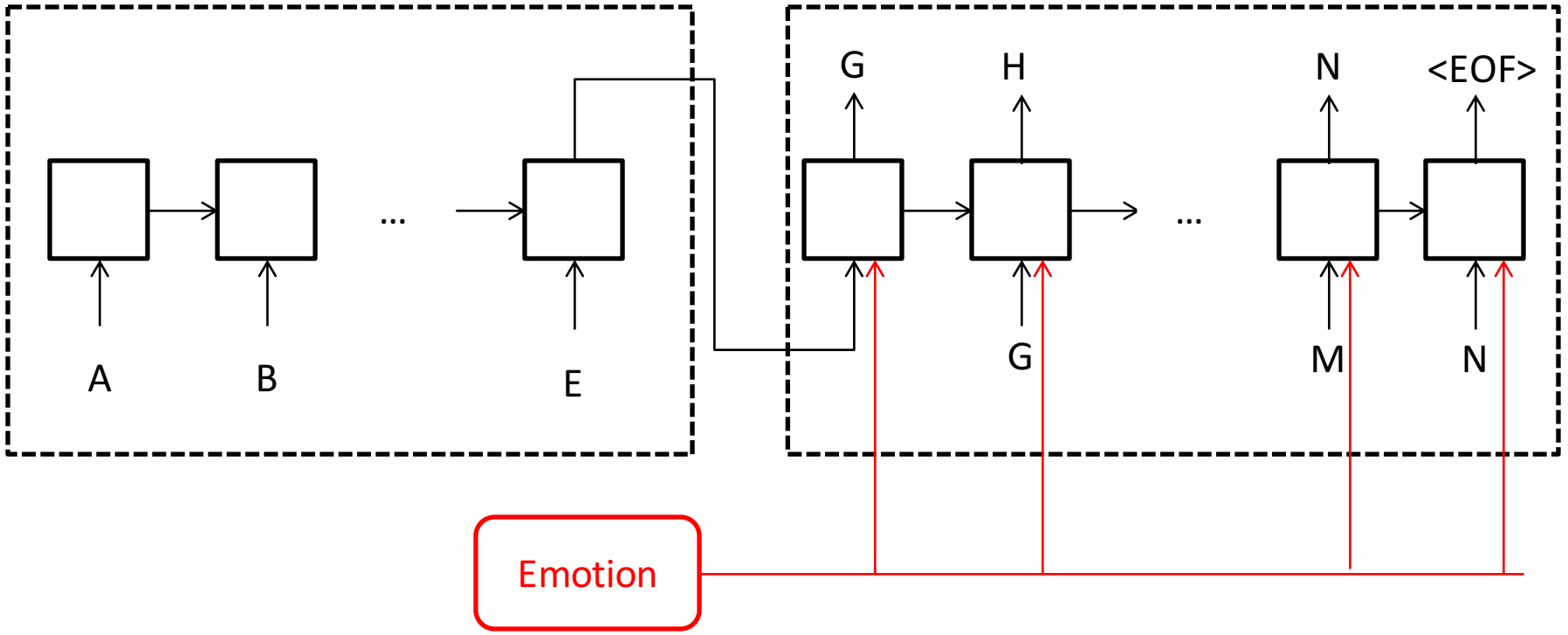
Decoding (generation) with attention



Can machine dialog be emotional?

1 layer lstm with additional targeted emotion labels

Joint w/ Di Zhou

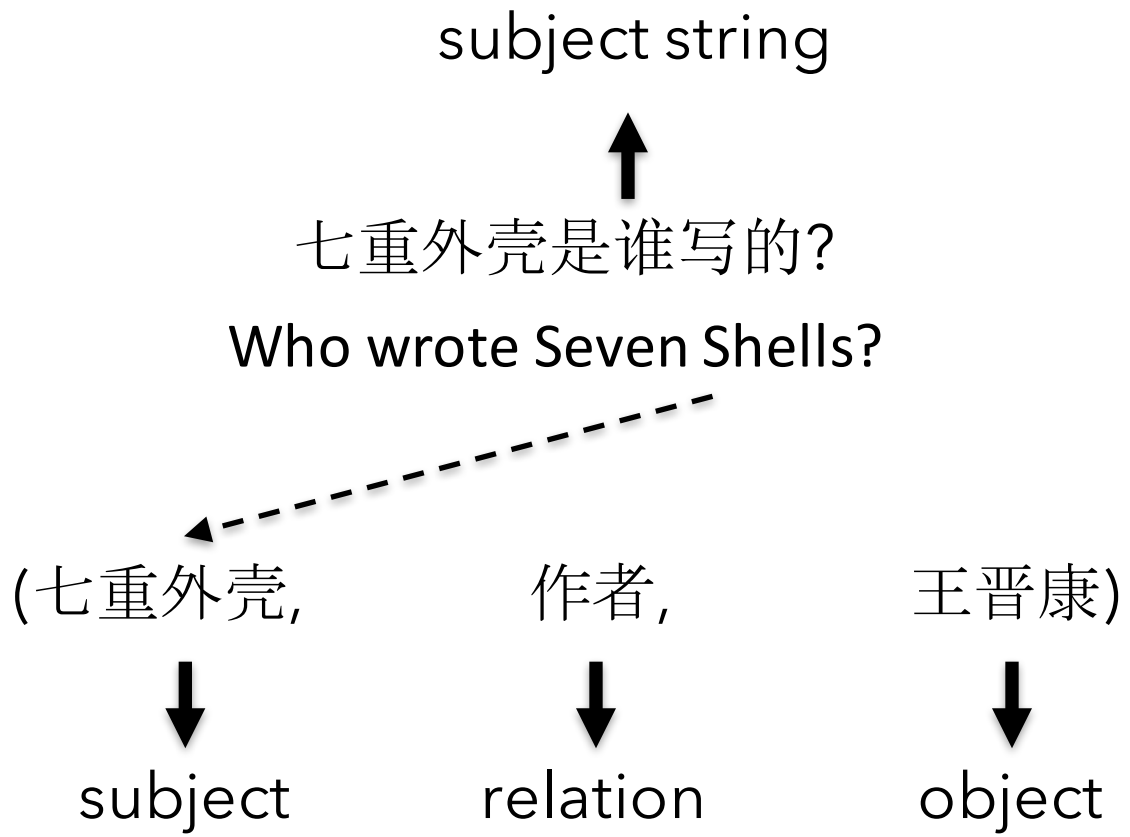


Issue: 爸爸为女儿画卡通眼罩治病(Father help his daughter fight against illness)
R1: 好感动😃
R2: 加油👍

Issue: 南京一宝马闯红灯撞死两人，殃及旁边奔驰(A car in Nanjin killed two people)
R1: 好可怜😭
R2: 尼玛😡

Natural question answering for knowledge base

Answering Knowledge Questions



related work
[Berant 2013]
[Yih 2014]
[Bordes 2015]

Why difficult for a machine?

Language Complexity

- 奥巴马总统在哪儿生的?
- 奥巴马总统出生地在哪里?
- What is the birthplace of Mr. Obama?
- Where was Mr. Obama born?

Entity/Relation Ambiguity

- 麦克乔丹是谁?
- Who is Michael Jordan?

Challenges

1. Insufficient Knowledge Representation

- Where is San Francisco?
- What is Columbus famous for?

→

- **MORE** than **400** entities
- **City, County, Person, Movie,** etc

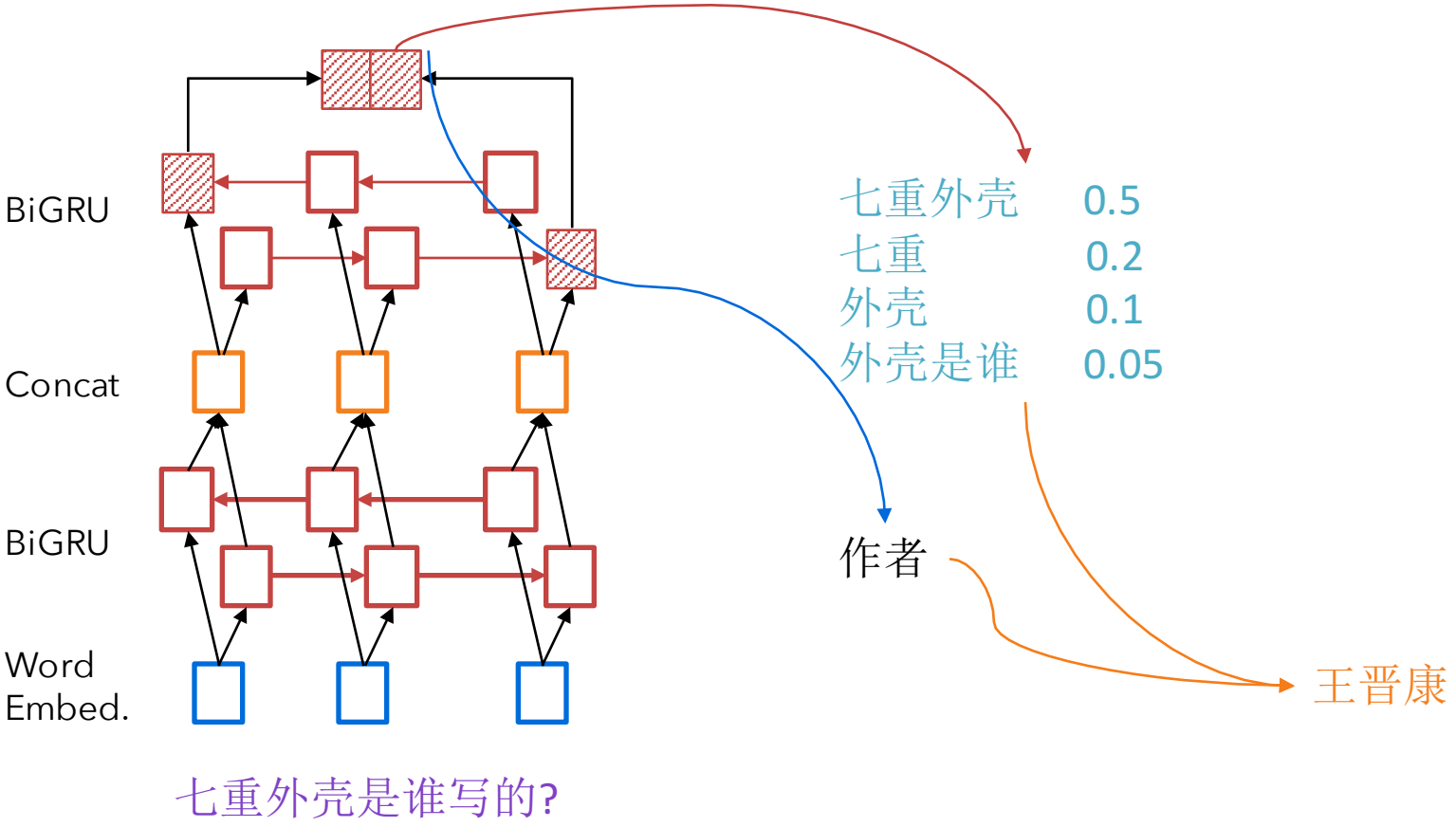
2. Too Much Noise from N-Grams

- What theme is the book the armies of memory?

→

- the book: 73
- theme: 252
- memory: 553
- ……

Stacked bi-directional GRU



[Dai, Li, Xu, 2016]

Answers by our DL system:

哈利波特在哪儿上的学？ Which school did Harry Potter attend?

霍格沃兹魔法学校 Hogwarts School of Witchcraft and Wizardry

格罗格里小学 Gregory Primary school

哈利波特是谁写的？ Who created Harry Potter?

罗琳女士 J.K. Rowling

罗琳的写作风格受谁影响？ Who influenced J.K. Rowling?

乔治艾略特 George Eliot

史蒂文金 Stephen King

史蒂文金写了什么小说？ What books did Stephen King write?

Las cuatro estaciones/different seasons

肖生克的救赎

[Dai, Li, Xu, 2016]

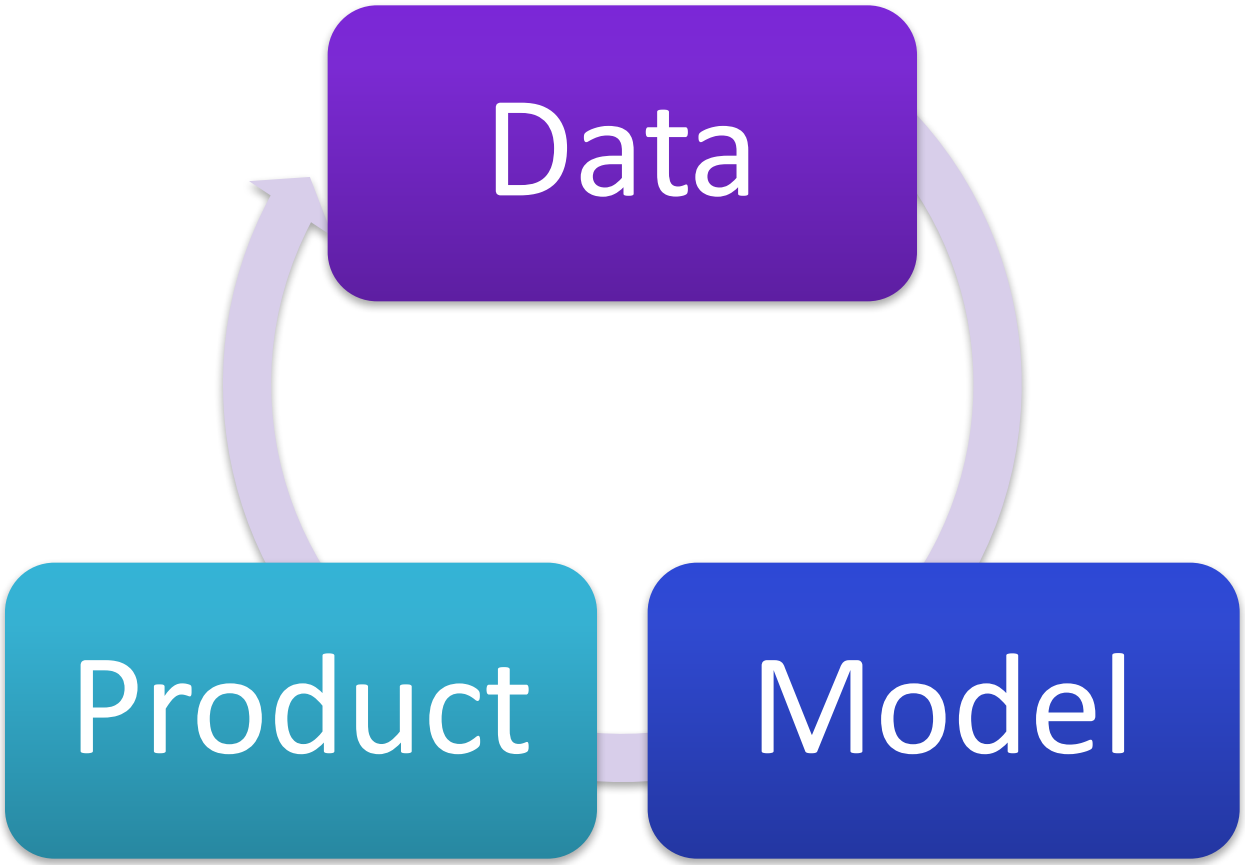
Take-away

- Artificial Neuron
  - Linear summation, Nonlinear activation
- Deep Neural Network
  - Feed forward structure
- Training DNN
  - Stochastic gradient descent
  - Forward and backward propagation

Take-away

- Recurrent Neural Network
  - Suitable for variable length sequence
  - Adaptive memory
- Semantic parsing network
  - Bidirectional RNN + CRF decoding
- Sequence to sequence learning

Success of practical ML depends on iteration speed



Thanks!

Joint work with

Wei Xu (Baidu IDL)
Zihang Dai (CMU): QA
Fei Jiang (Tsinghua): Dialog
Zefu Lu (UIUC): NER
Hieu Pham (Stanford): machine translation
Di Zhou (UT Dallas): emotional dialog

Toutiao Lab is Hiring!

Research Scientist and Software Engineer in
Machine Learning
Natural Language Processing
Computer Vision

http://www.toutiao.com/lab

lab-hr@toutiao.com

Reference

Seq-to-Seq

- Sutskever et al, Sequence to sequence learning with Neural Networks
- Bahdanau et al, Neural machine translation by jointly learning to align and translate
- Shang et al, Neural responding machine for short-text conversation
- Vinyals & Le, A neural conversational model.

Reference

Parsing & Sequence labelling

- Collobert et al, Natural language processing almost from scratch
- Lu et al, Twisted recurrent network for named entity recognition
- Huang et al, Bidirectional LSTM-CRF models for sequence tagging
- Zhou et al, End-to-end learning of semantic role labeling using recurrent neural networks.

Question Answering

- Jonathan Berant, Andrew Chou, Roy Frostig, Percy Liang Semantic Parsing on Freebase from Question-Answer Pairs, EMNLP 2013.
- W. Yih, X. He & C. Meek. Semantic Parsing for Single-Relation Question Answering. In ACL-14.
- W. Yih, M. Chang, X. He & J. Gao. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In ACL-IJCNLP-2015
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, Jason Weston, Large-scale Simple Question Answering with Memory Networks, 2015
- Zihang Dai, Lei Li, Wei Xu, CFO: Conditional Focused Question Answering with Large Knowledge-bases. ACL 2016