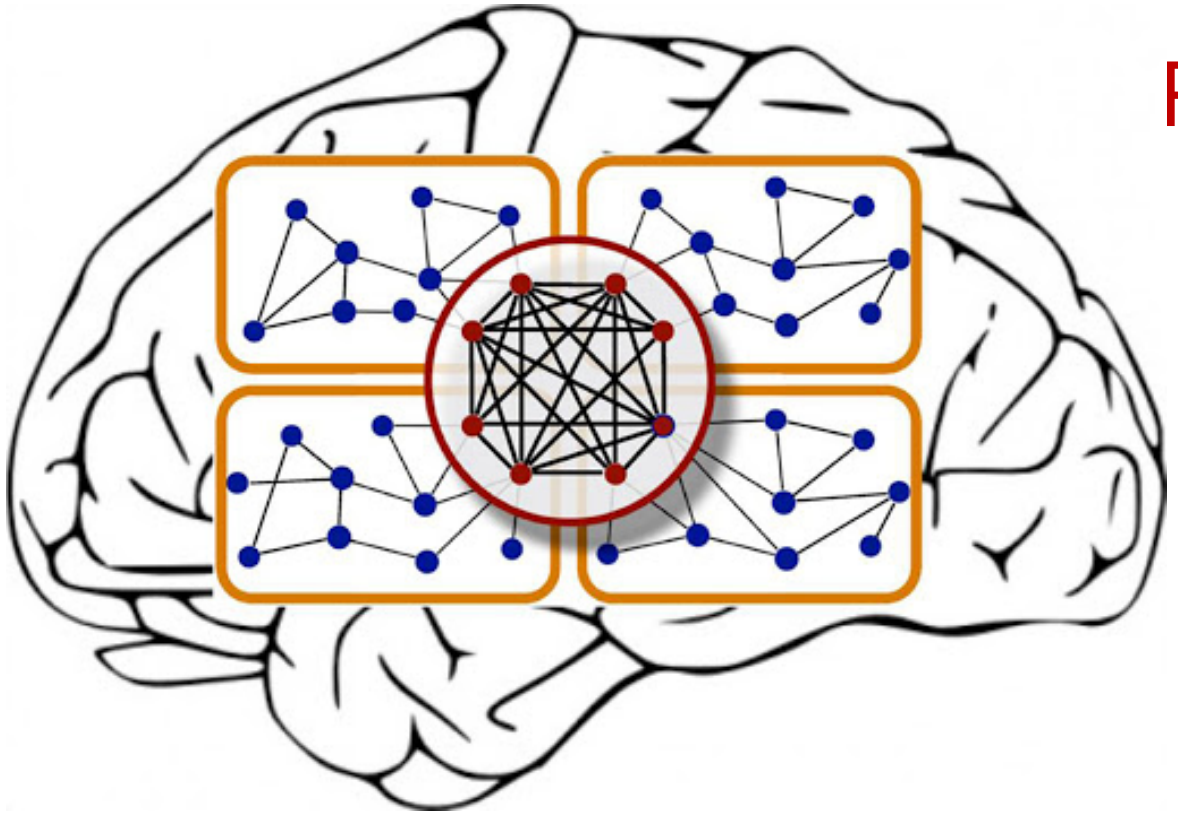


Representation Learning & Transfer Learning

张江

北京师范大学系统科学学院教授

集智俱乐部、集智学园创始人

集智研究中心理事长

What is deep learning?

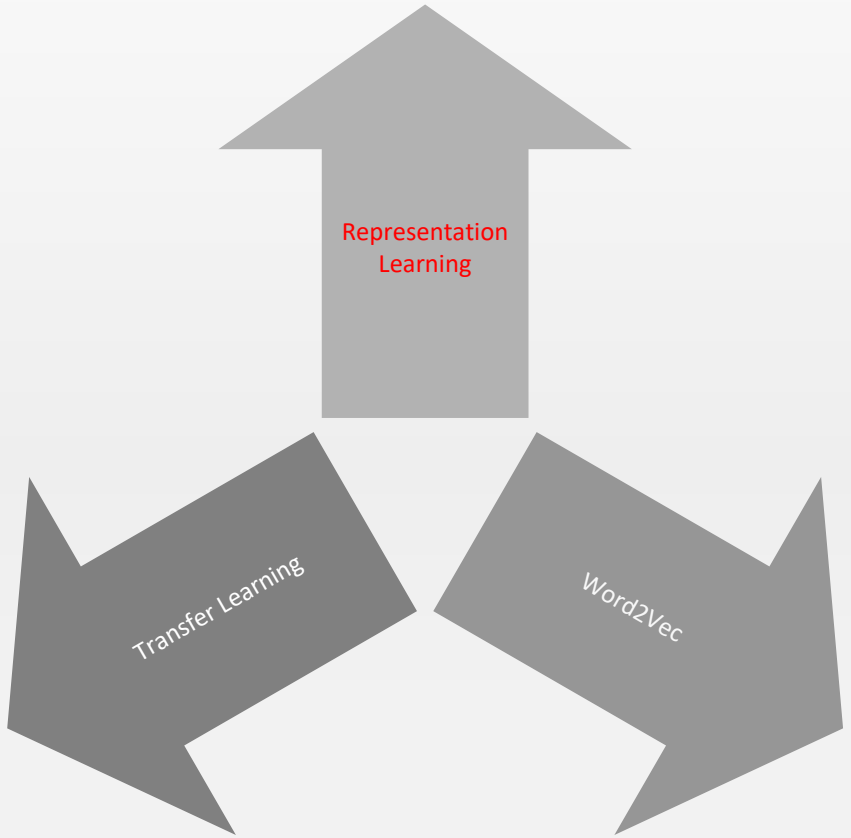
Deep Learning

~~Deep?~~

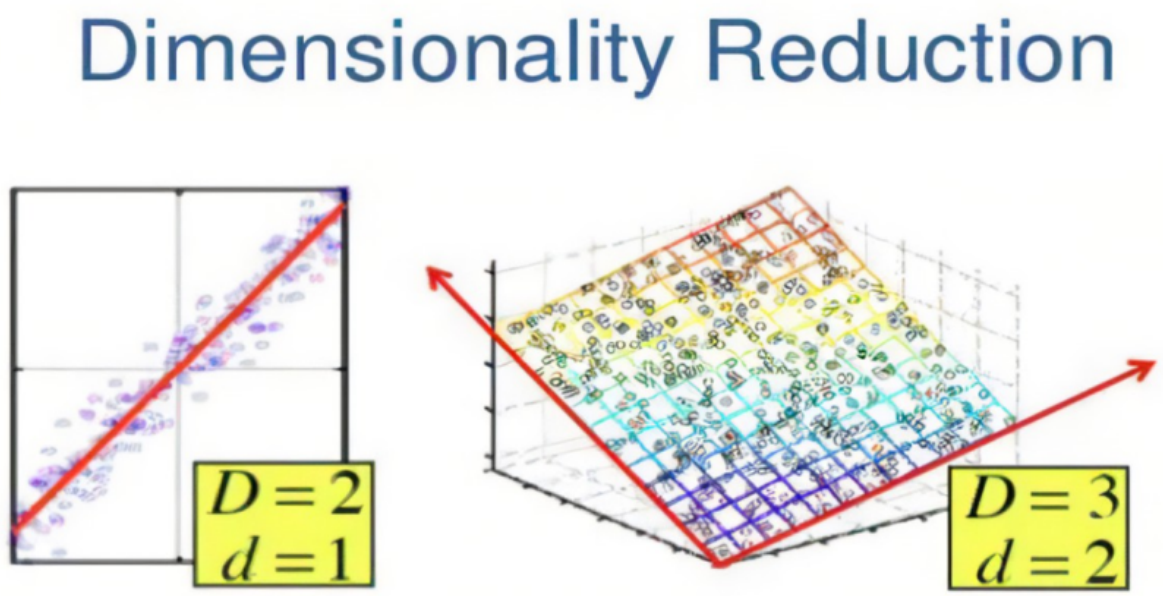
Automatic Differentiation

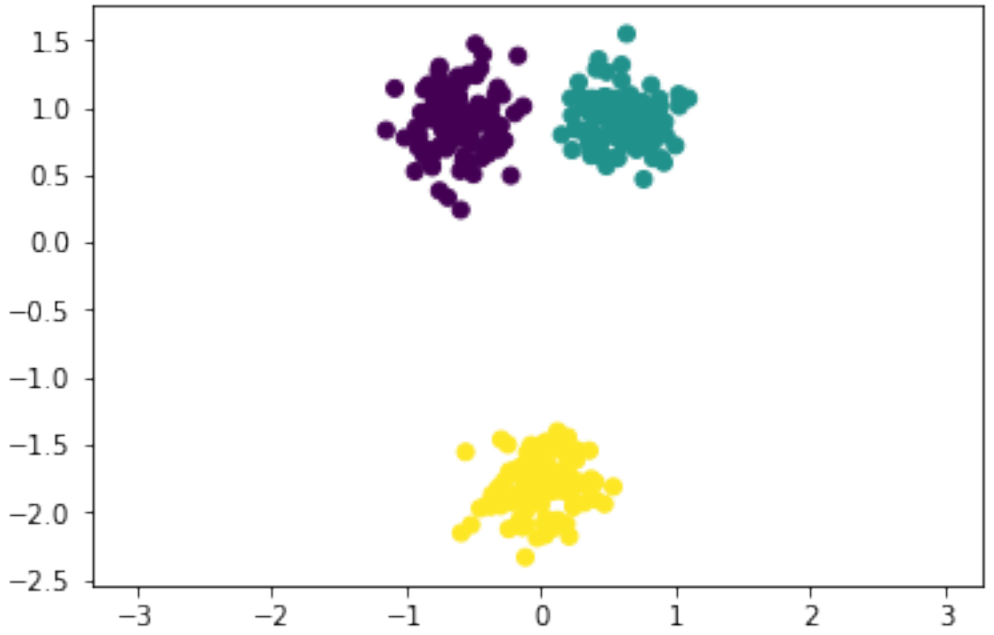
Representation Learning

Outline



Problem I – Dimension Reduction

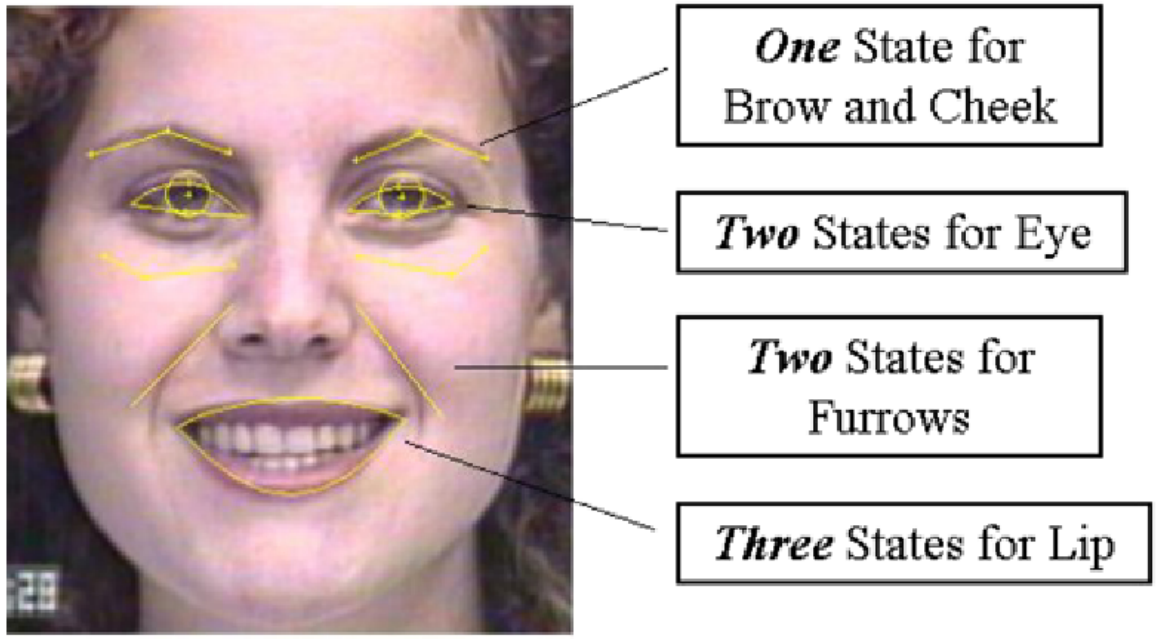




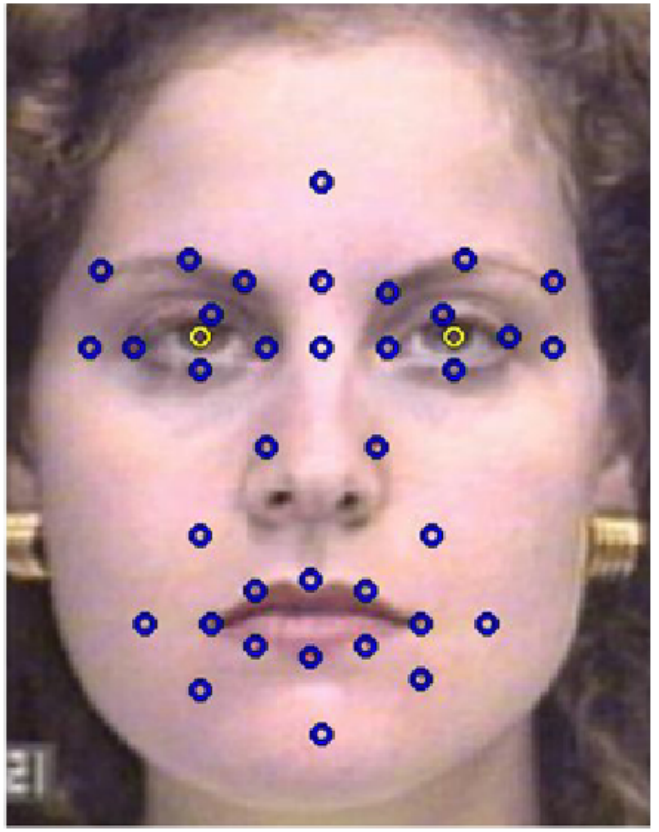
Clustering

Problem II – Feature Extraction

Multi-State Models for Geometric Feature Extraction



(a)

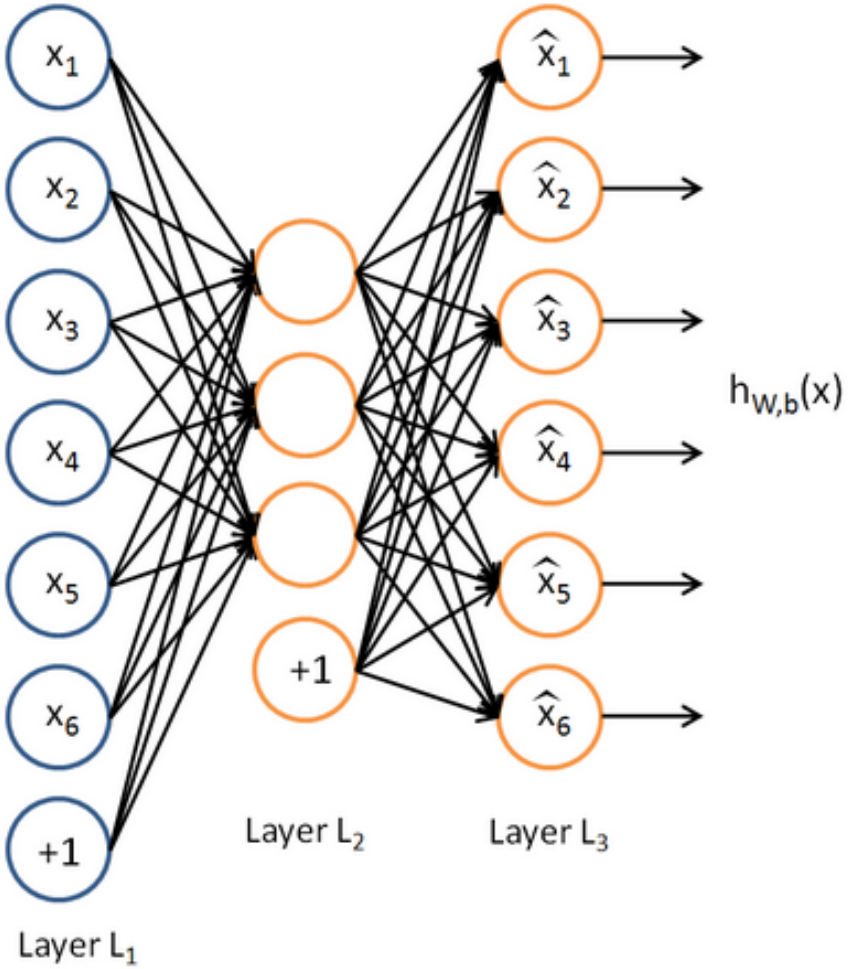


(b)

Problem III – Representation of Complex Objects

Categorical variables are a natural choice for representing discrete structure in the world. However, stochastic neural networks rarely use categorical latent variables due to the inability to backpropagate through samples. In this work, we present an efficient gradient estimator that replaces the non-differentiable sample from a categorical distribution with a differentiable sample from a novel Gumbel-Softmax distribution. This distribution has the essential property that it can be smoothly annealed into a categorical distribution. We show that our Gumbel-Softmax estimator outperforms state-of-the-art gradient estimators on structured output prediction and unsupervised generative modeling tasks with categorical latent variables, and enables large speedups on semi-supervised classification.

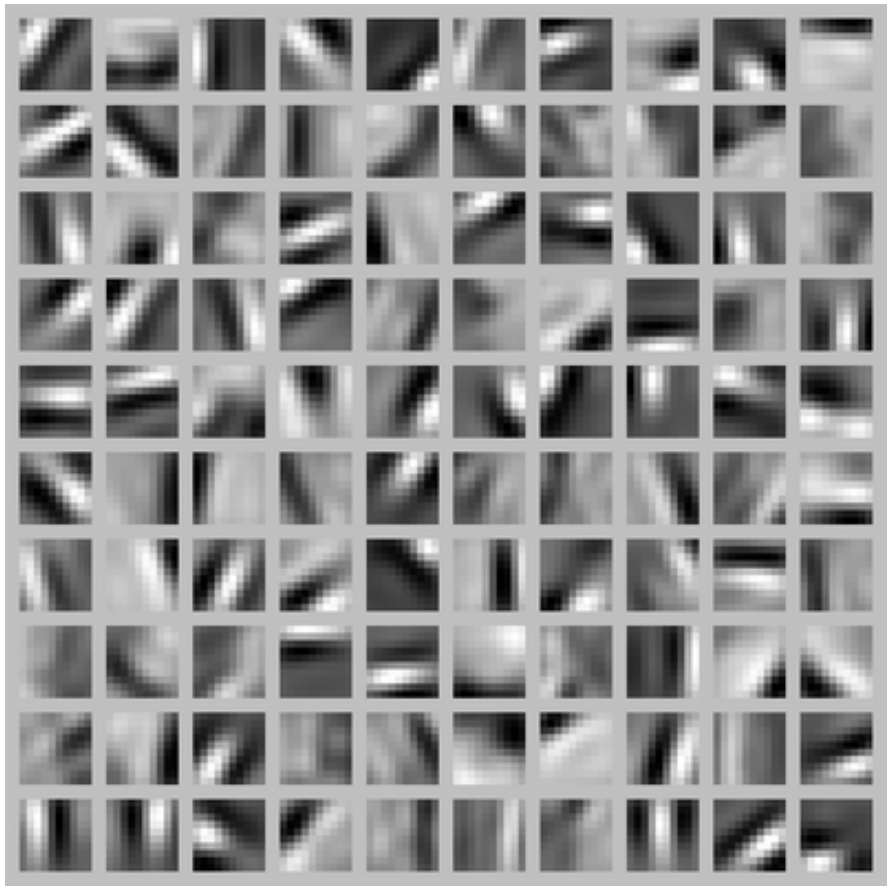
Autoencoder



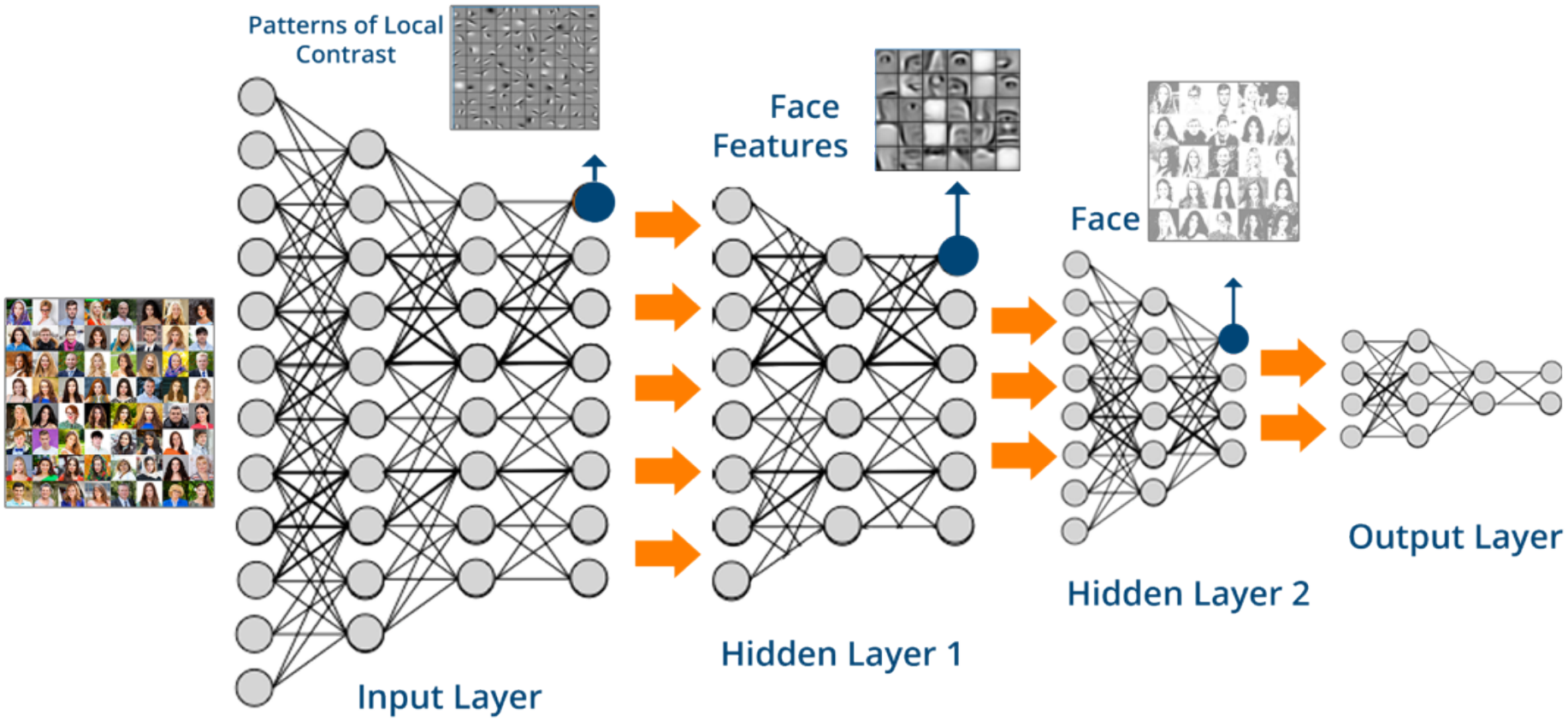
- Unsupervised learning
- Feed-back
  - Reconstruct inpu $h_\theta(x) \approx x$
  - BP Learning
- Constraints

Autoencoder

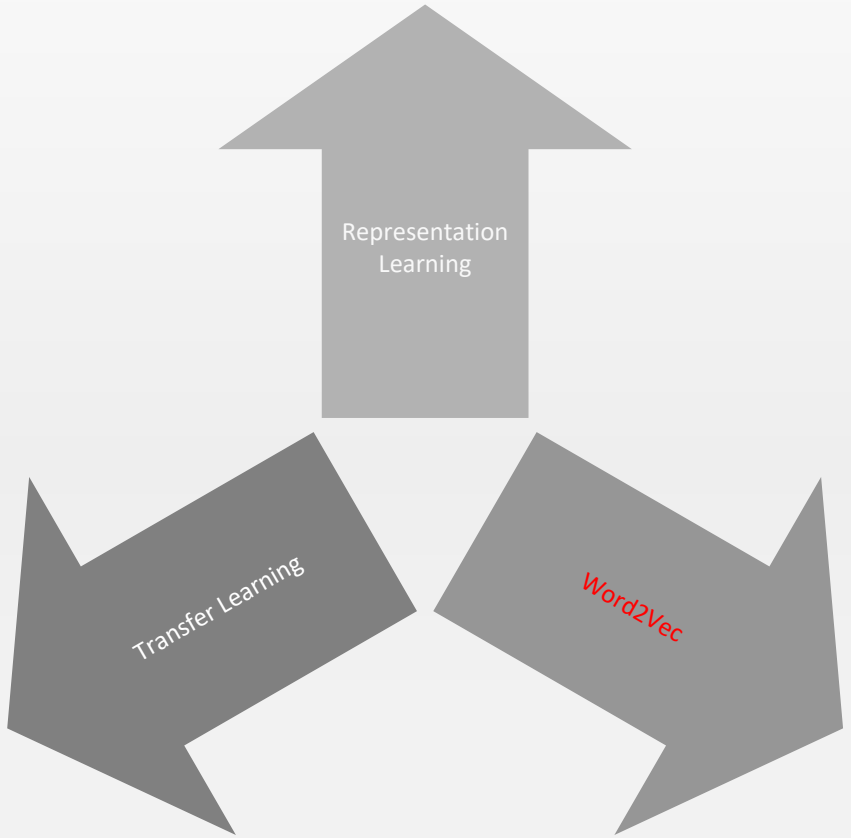
- Constraint activation
- Describe the input data with minimum features.
- When will one unit is activated?
  - when w parallels with xi

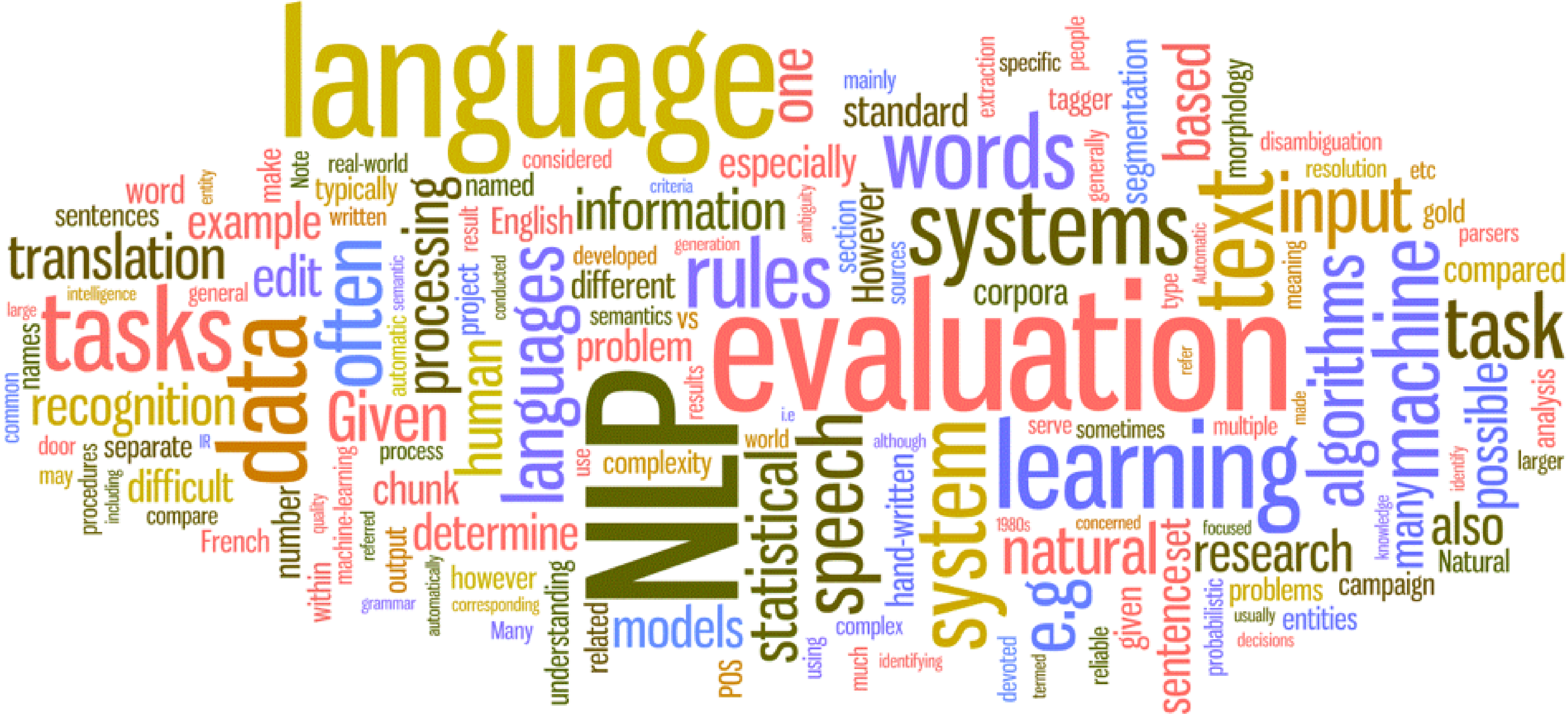


Features of Depth



Outline




language
one
mainly
standard
extraction
specific
people
tagger
segmentation
based
morphology
disambiguation
resolution
etc
words
especially
considered
real-world
Note
make
entity
word
typically
named
criteria
ambiguity
generally
text
input
gold
parsers
compared
sentences
example
written
information
systems
translation
edit
often
processing
result
English
developed
generation
section
However
sources
Automatic
meaning
intelligence
general
semantic
project
conducted
rules
different
corpora
type
algorithms
machine
task
large
tasks
data
automatic
human
languages
semantics
vs
problem
evaluation
names
refer
common
recognition
Given
results
NLP
learning
i.e
possible
analysis
made
serve
sometimes
multiple
door
IR
separate
process
world
although
larger
procedures
use
complexity
statistical
speech
system
may
difficult
including
machine-learning
chunk
identify
knowledge
many
also
quality
hand-written
compare
number
referred
determine
1980s
concerned
focused
French
output
natural
set
research
Natural
within
however
e.g
sentence
problems
campaign
grammar
automatically
corresponding
understanding
related
models
given
probabilistic
usually
entities
Many
complex
POS
using
much
identifying
devoted
termed
reliable
decisions

How to represent words?

Direct Way

- Student
- [‘s’,’t’,’u’,’d’,’e’,’n’,’t’]: [18, 19, 20, 3, 4, 13, 19]

Problems

- Different lengths
- Not similar for similar words in meaning

One-hot coding

单词	序号	编码
A	0	[1, 0, 0, 0, ……, 0]
An	1	[0, 1, 0, 0, ……, 0]
About	2	[0, 0, 1, 0, ……, 0]
……	……	……
Zero	14901	[0, 0, 0, 0, ……, 1]

14902

Disadvantages:

Waste of space

Can not reflect similarity

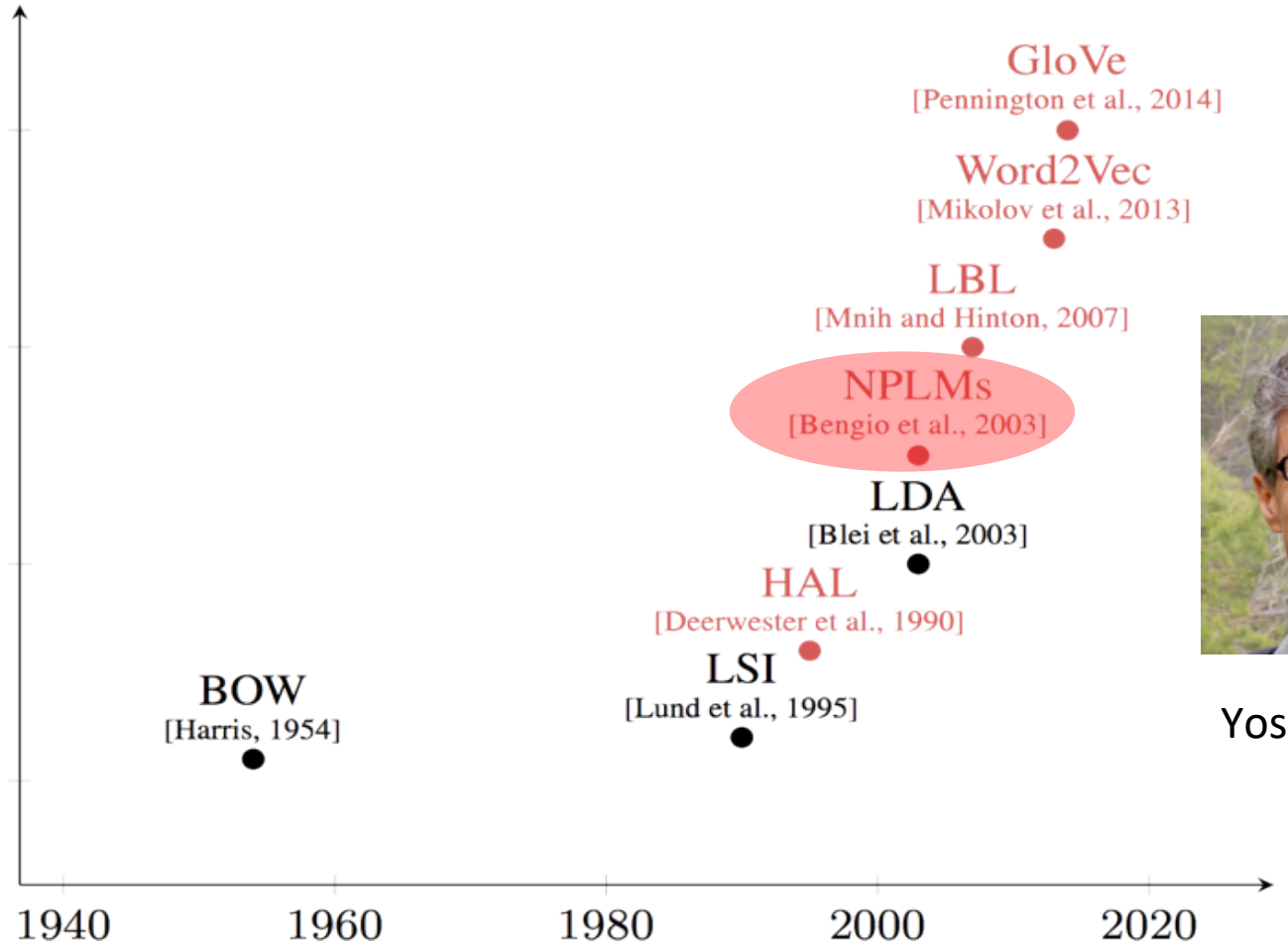
What we want

单词	词向量
star	[0.5, 0.7, -0.3, 0.2]
sun	[0.5, -0.8, 0.6, 0.4]
moon	[0.49, 0.6, -0.3, 0.1]

Features

- Dense vectors
- Similarity

Breakthroughs



Yoshua Bengio

F.Sun et al.: Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations (slides), http://www.bigdatalab.ac.cn/~lanyanyan/slides/2015/ACL2015-sun.pdf

NPLM: An Unsupervised Learning Model

NPLM

红色：(1.0, 2.0, 3.3, 88,…, 0.1)
联合：(0.1, 0.8, 7.1, 9, … , 88)
相比：(2.0, 2.1, 3.3, 4.1,…, 45)
…
理想：(0.1, 0.8, 7.1, 9, … , 88)

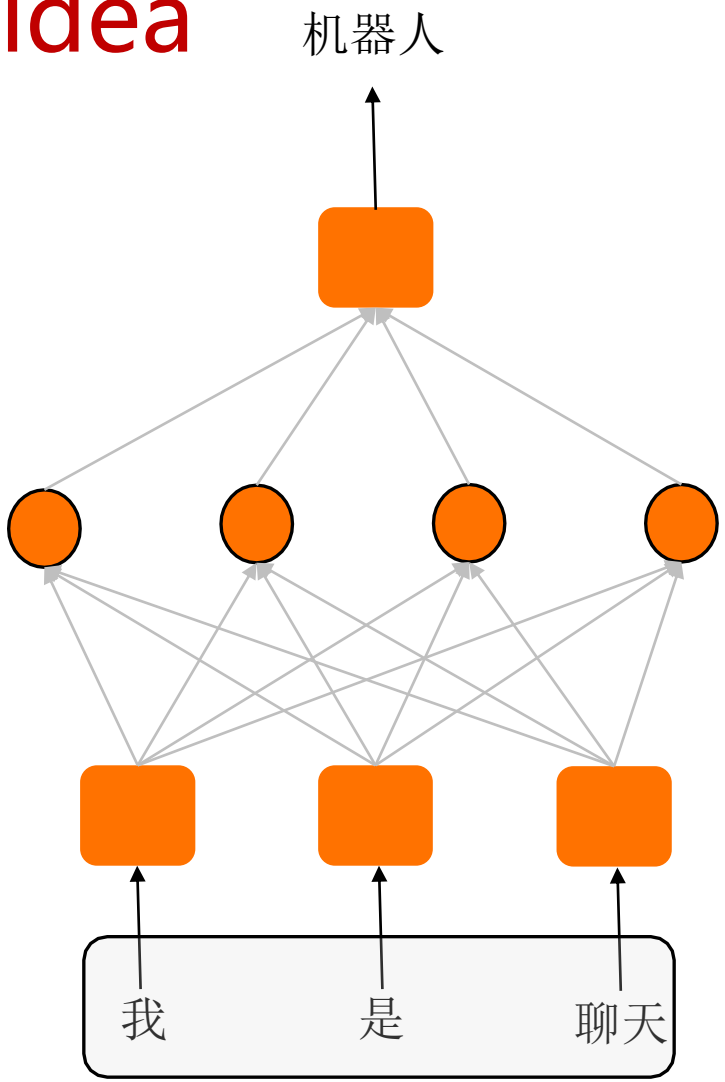
NPLM: An Unsupervised Learning Model

NPLM

语言模型：
根据前N个词预测下一个词

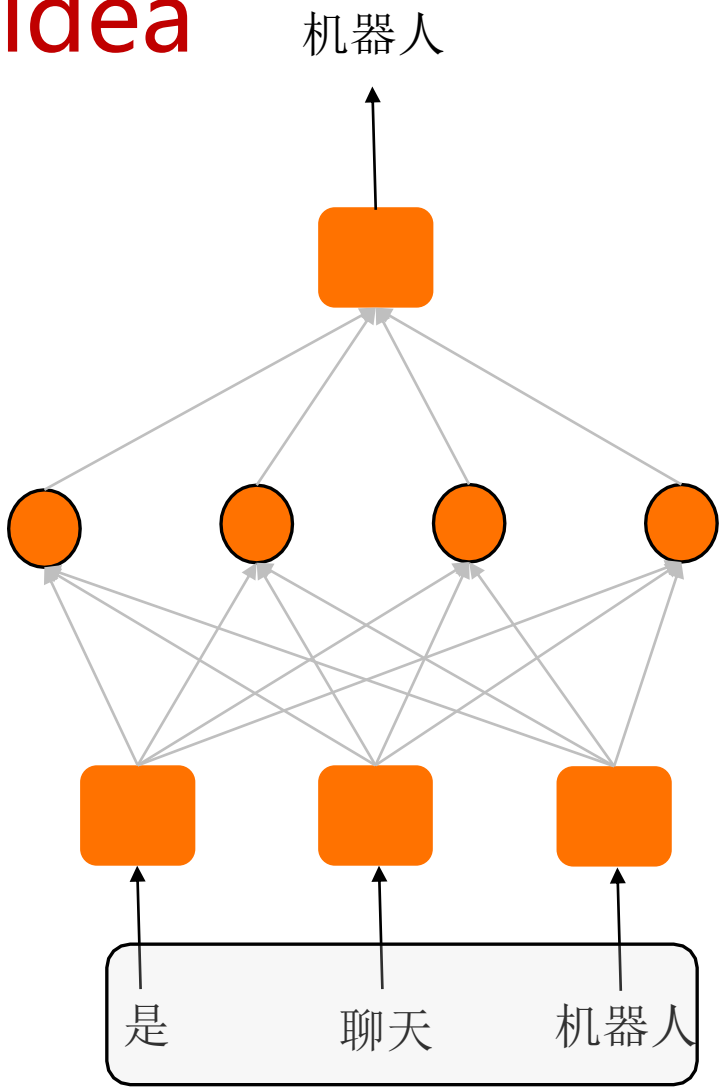
红色：(1.0, 2.0, 3.3, 88,…, 0.1)
联合：(0.1, 0.8, 7.1, 9, … , 88)
相比：(2.0, 2.1, 3.3, 4.1,…, 45)
…

NPLM: Basic idea

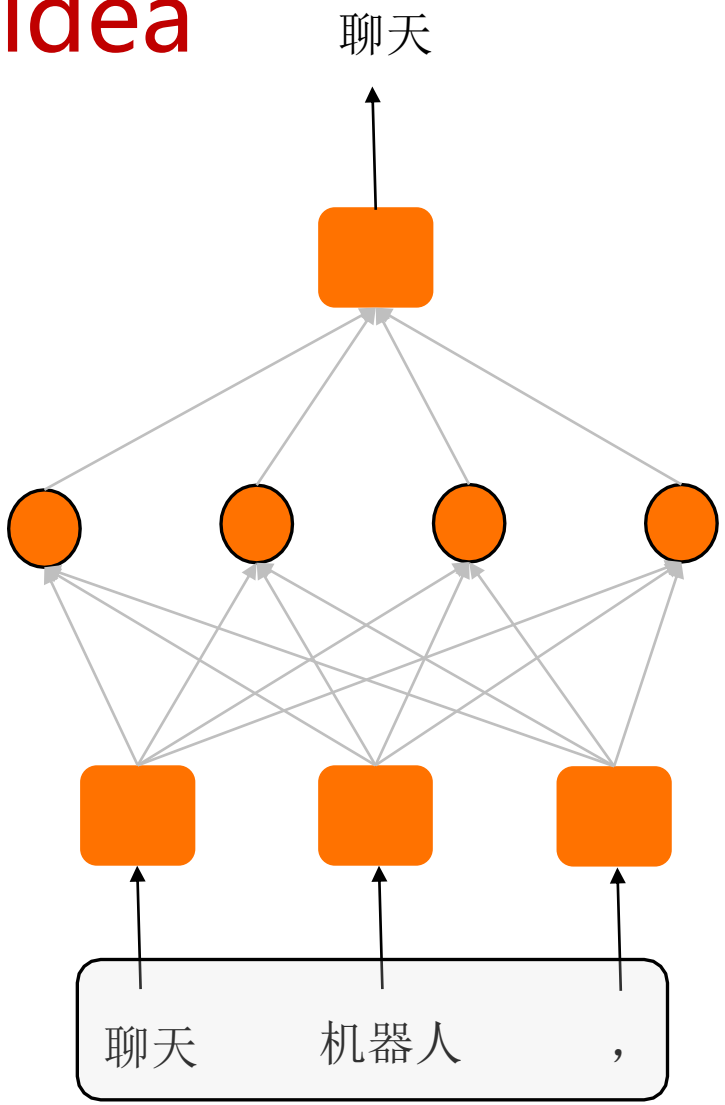


机器人 ， 我 可以 聊天

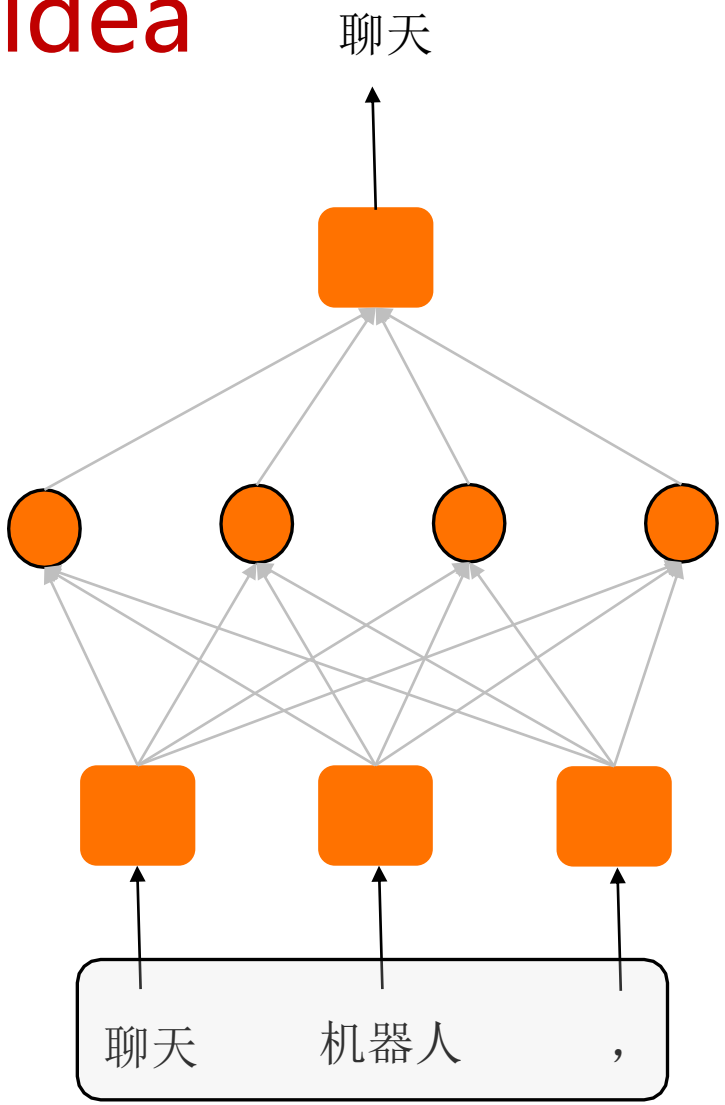
NPLM: Basic idea



NPLM: Basic idea



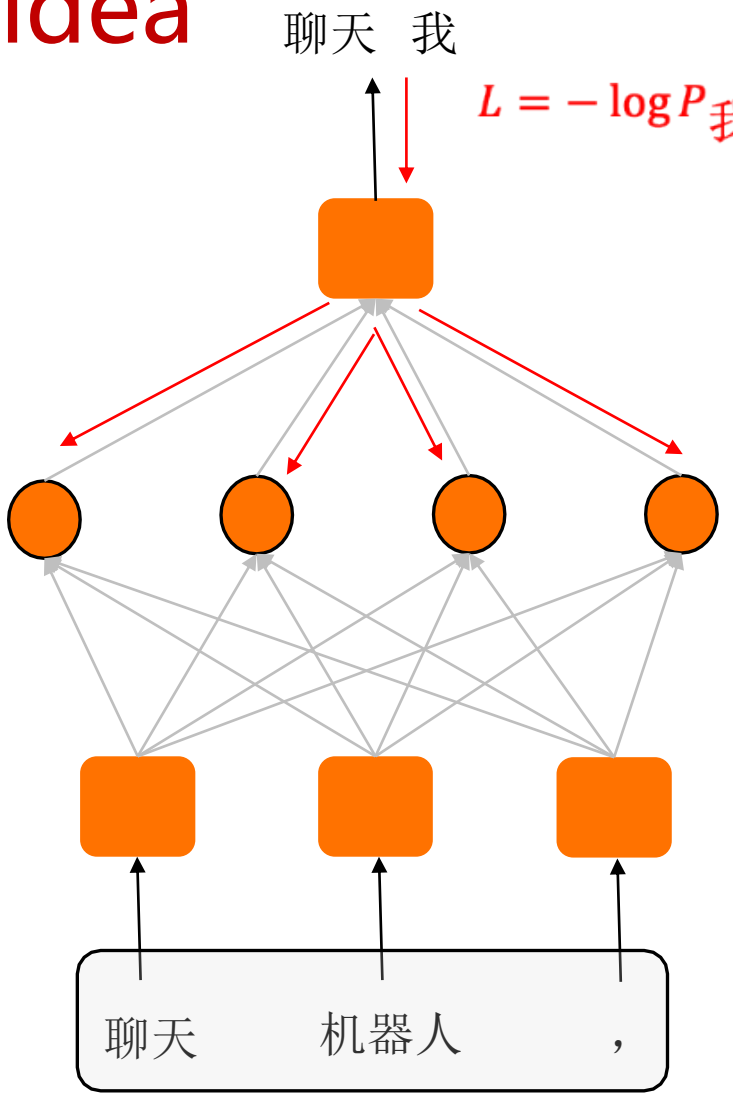
NPLM: Basic idea



N-gram语言模型
N=3

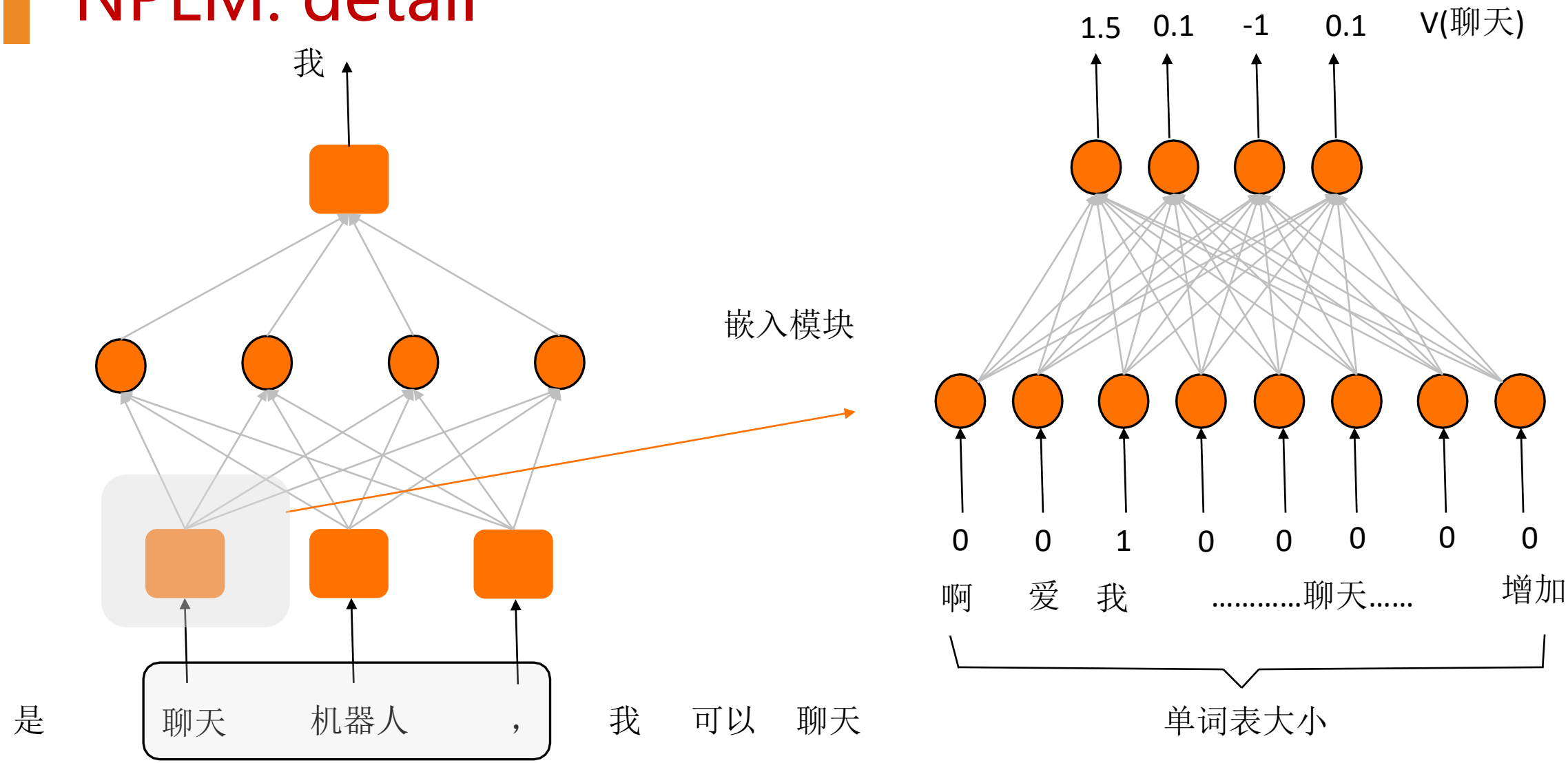
我 是 聊天 机器人 ， 我 可以 聊天

NPLM: Basic idea

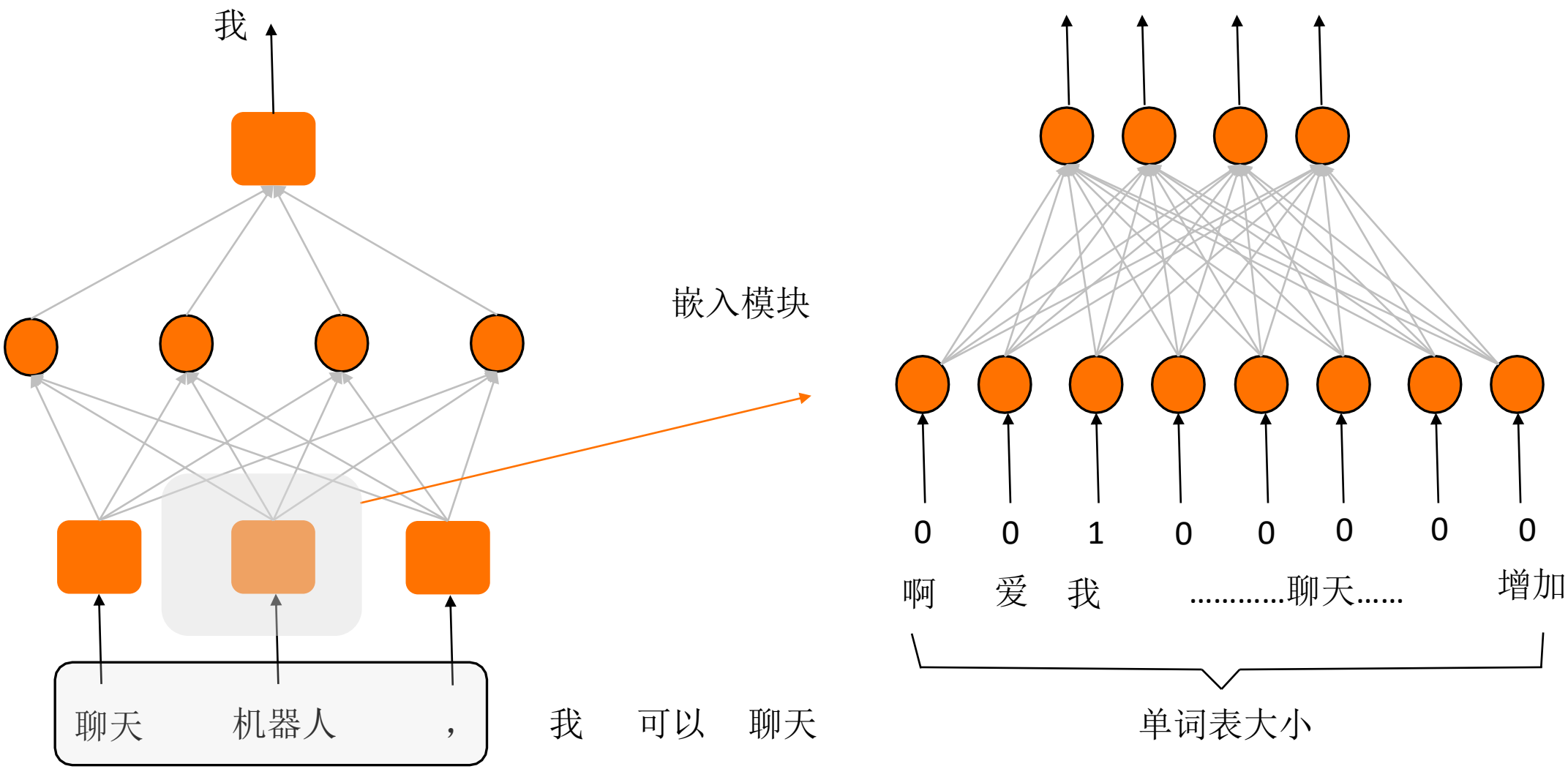


误差反传：BP算法

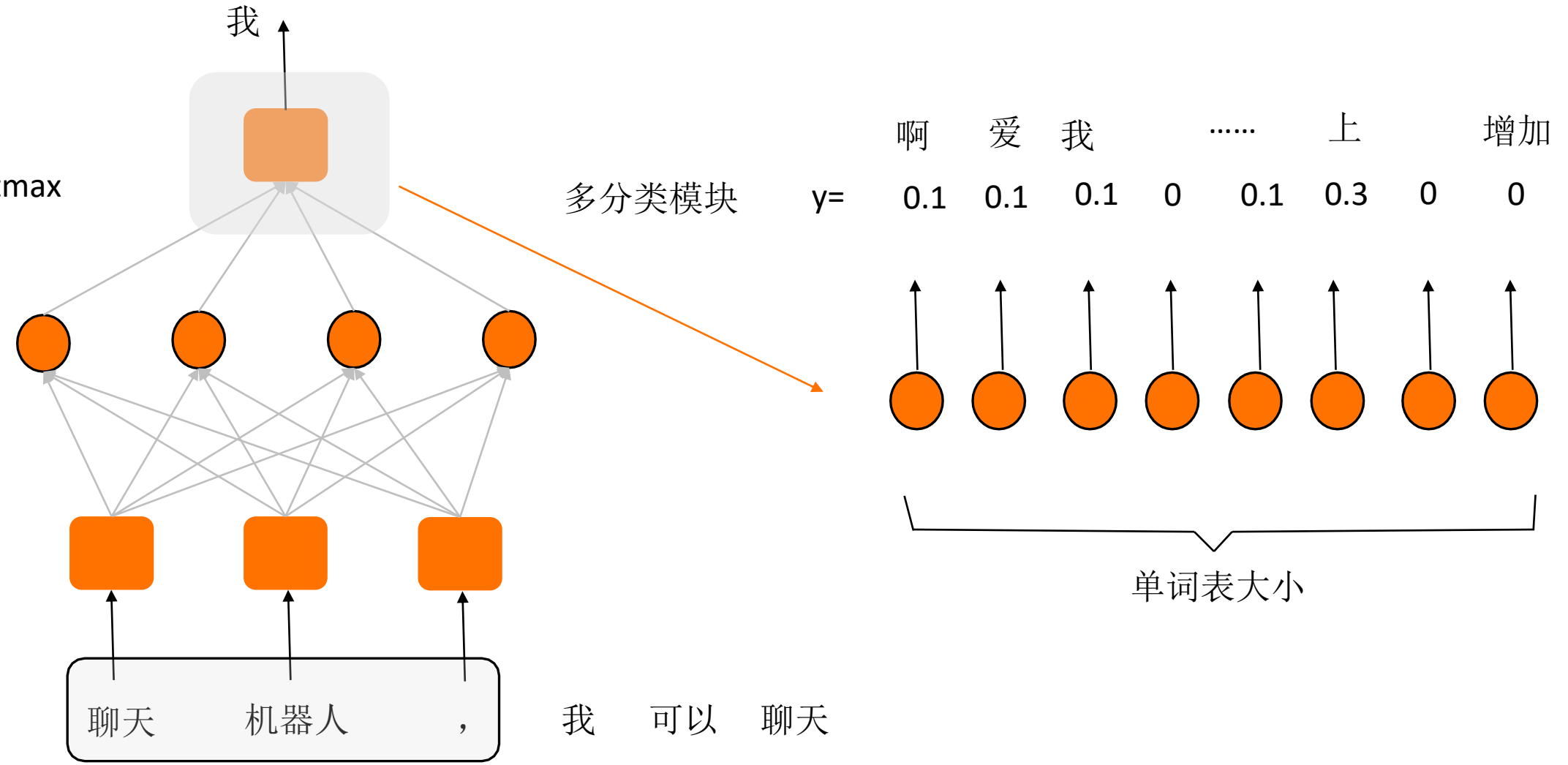
NPLM: detail



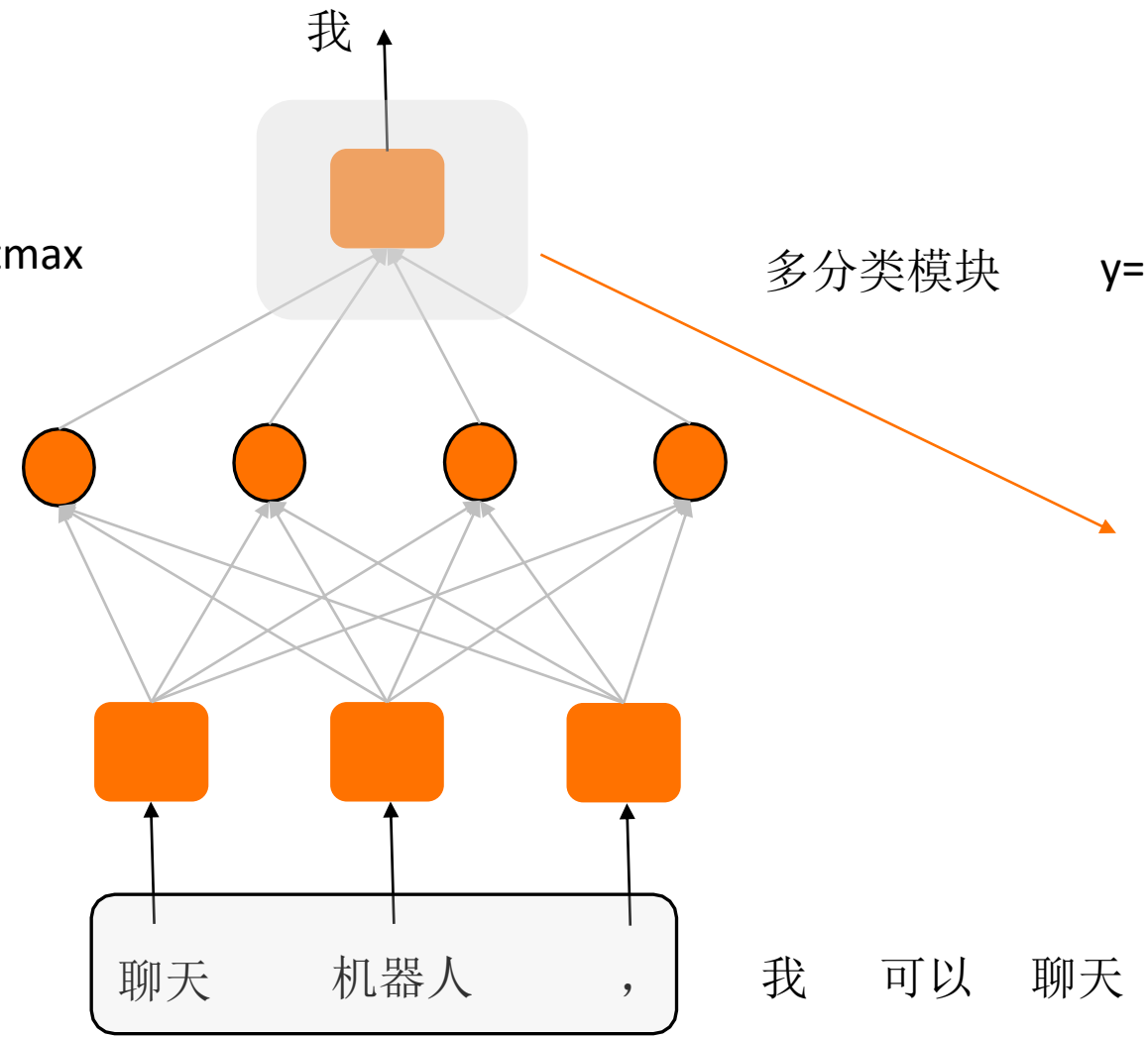
NPLM: detail



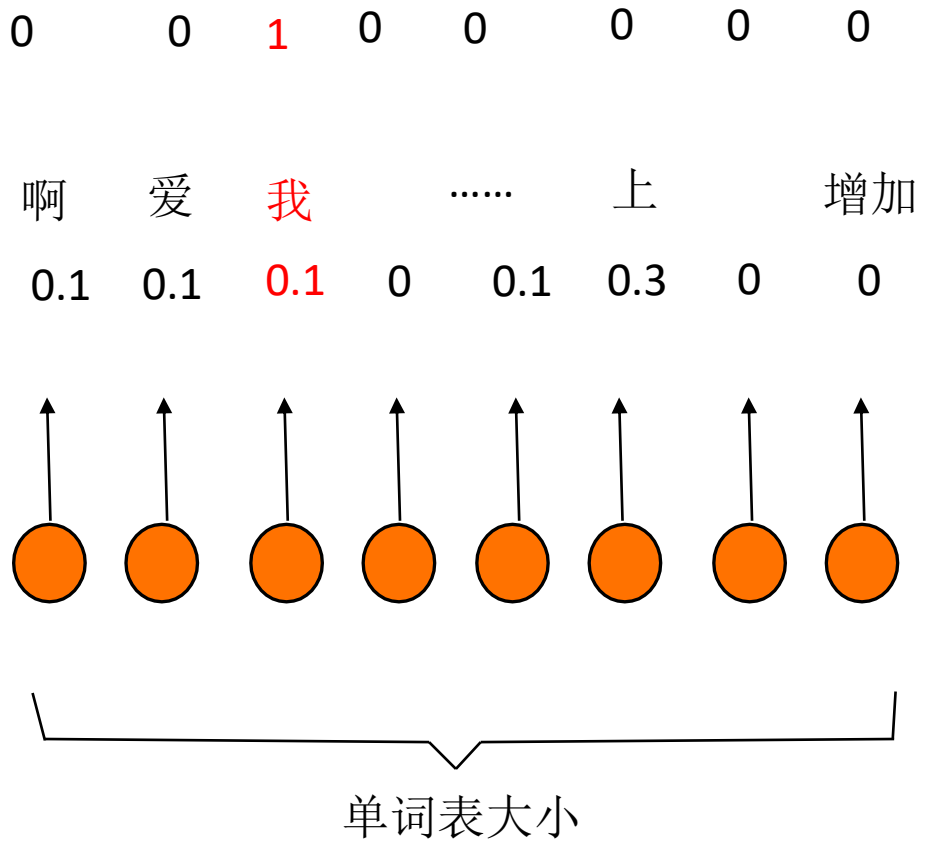
NPLM: detail



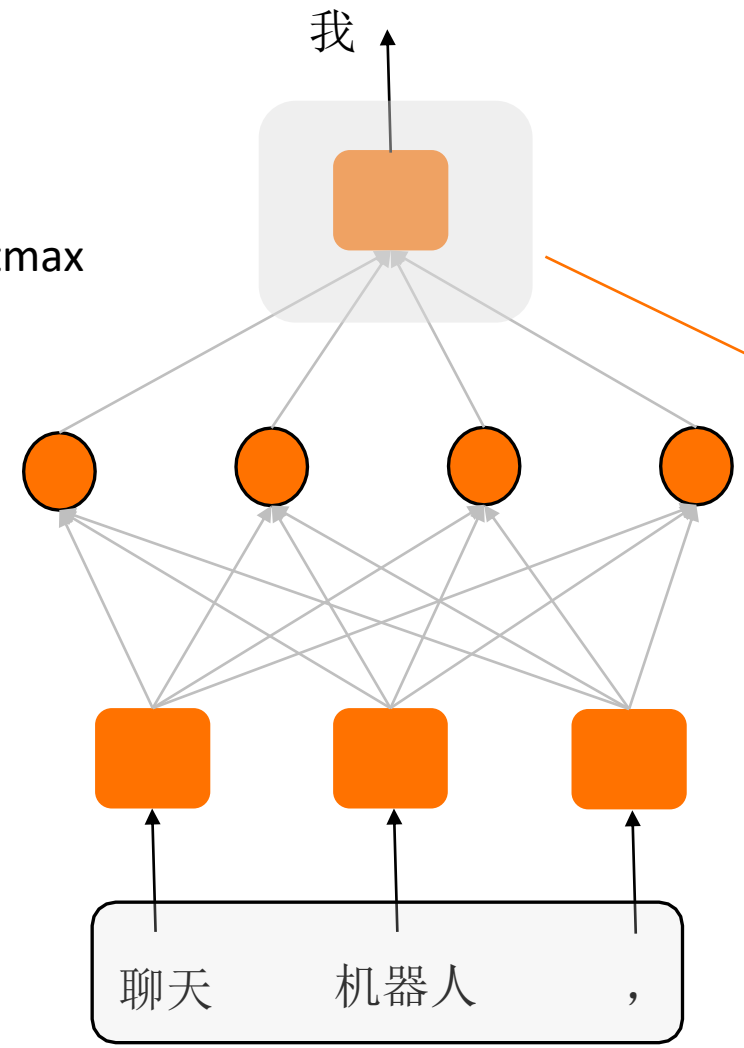
NPLM: detail



多分类模块 y=



NPLM: detail

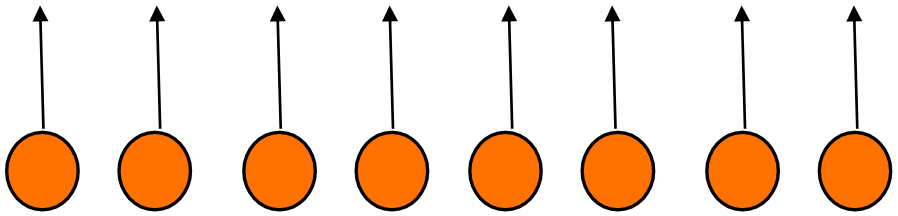


多分类模块

$L = -\log P_{我}$

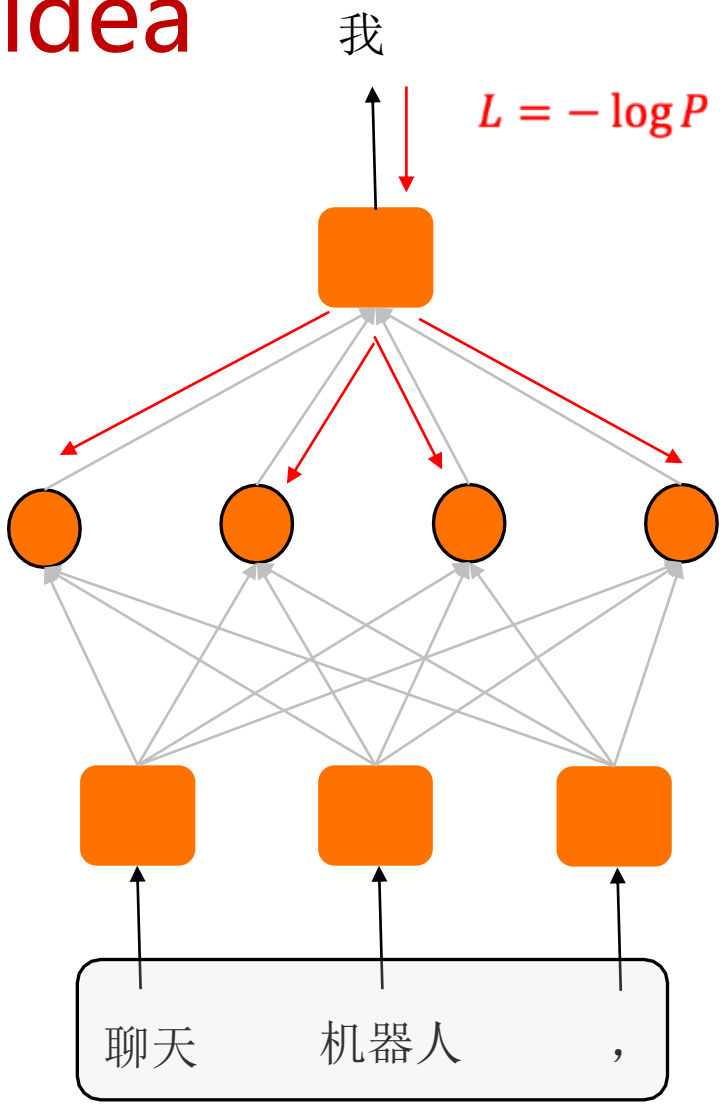
y=

0	0	1	0	0	0	0	0
啊	爱	我	……		上		增加
0.1	0.1	0.1	0	0.1	0.3	0	0



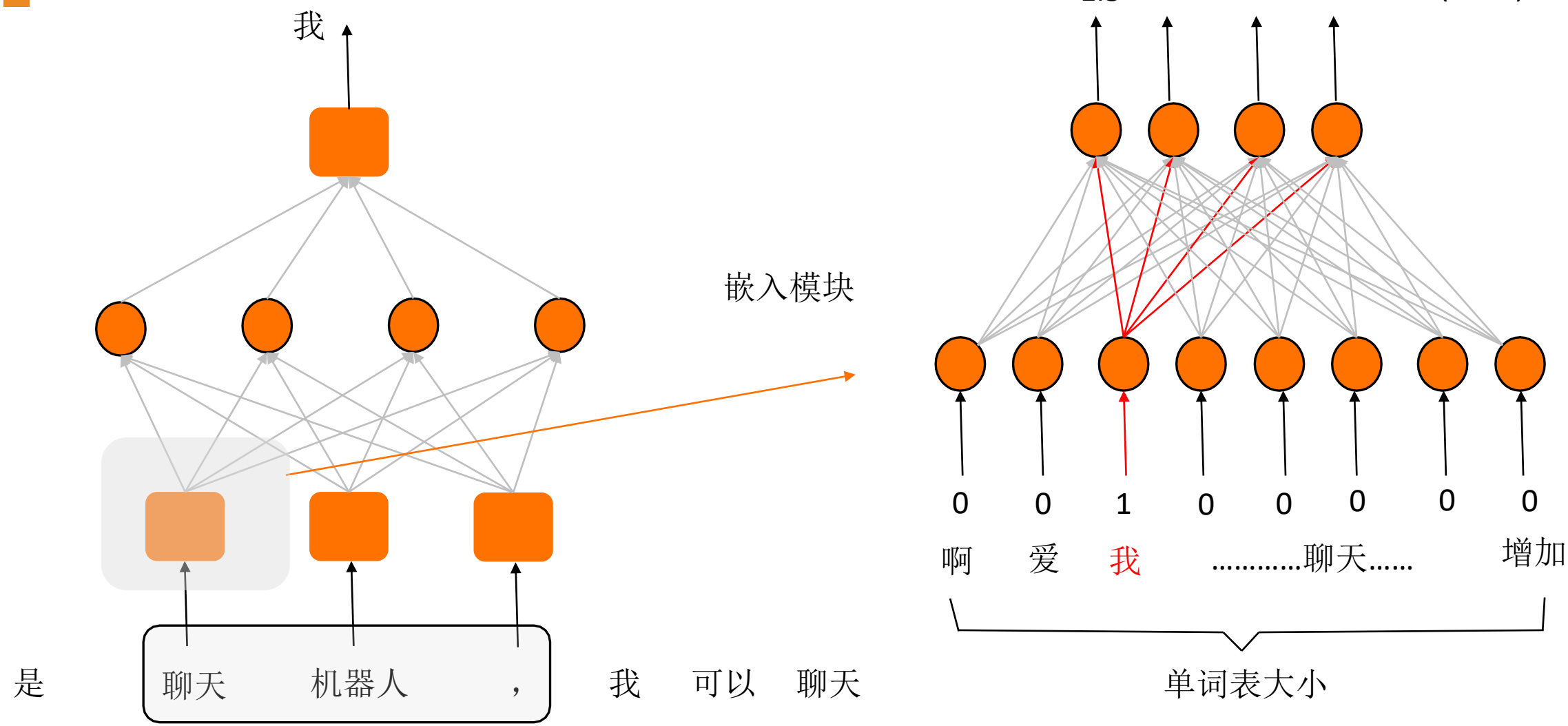
单词表大小

NPLM: Basic idea

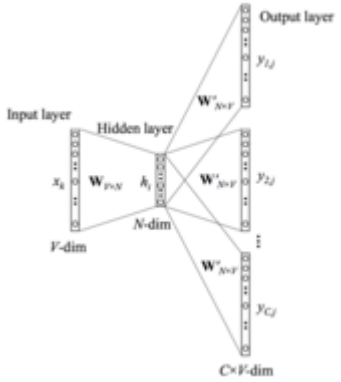


L.backward()

Where is the vector?



Breakthroughs

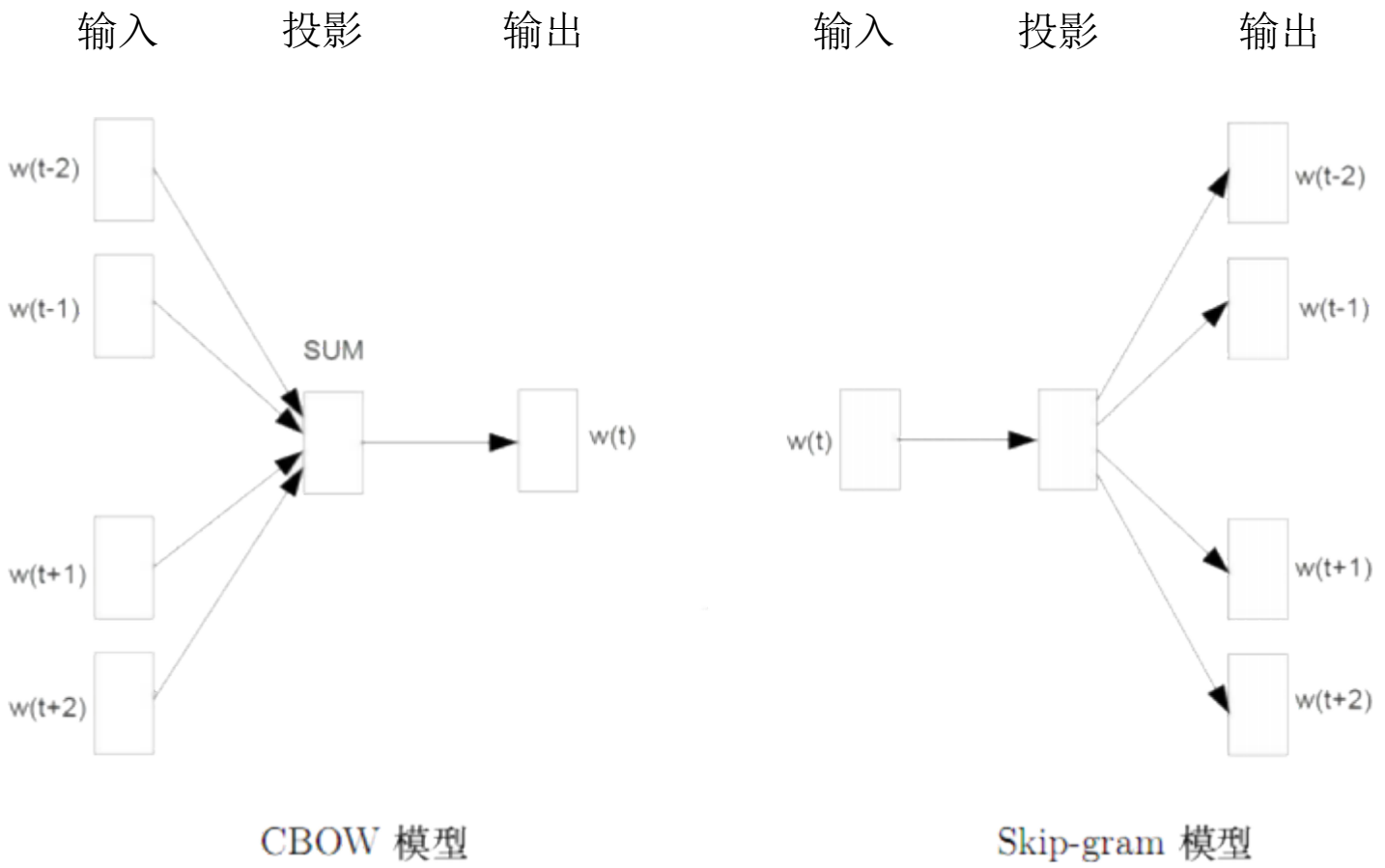


Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Tomas Mikolov

F.Sun et al.: Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations (slides), http://www.bigdatalab.ac.cn/~lanyanyan/slides/2015/ACL2015-sun.pdf

Word2Vec

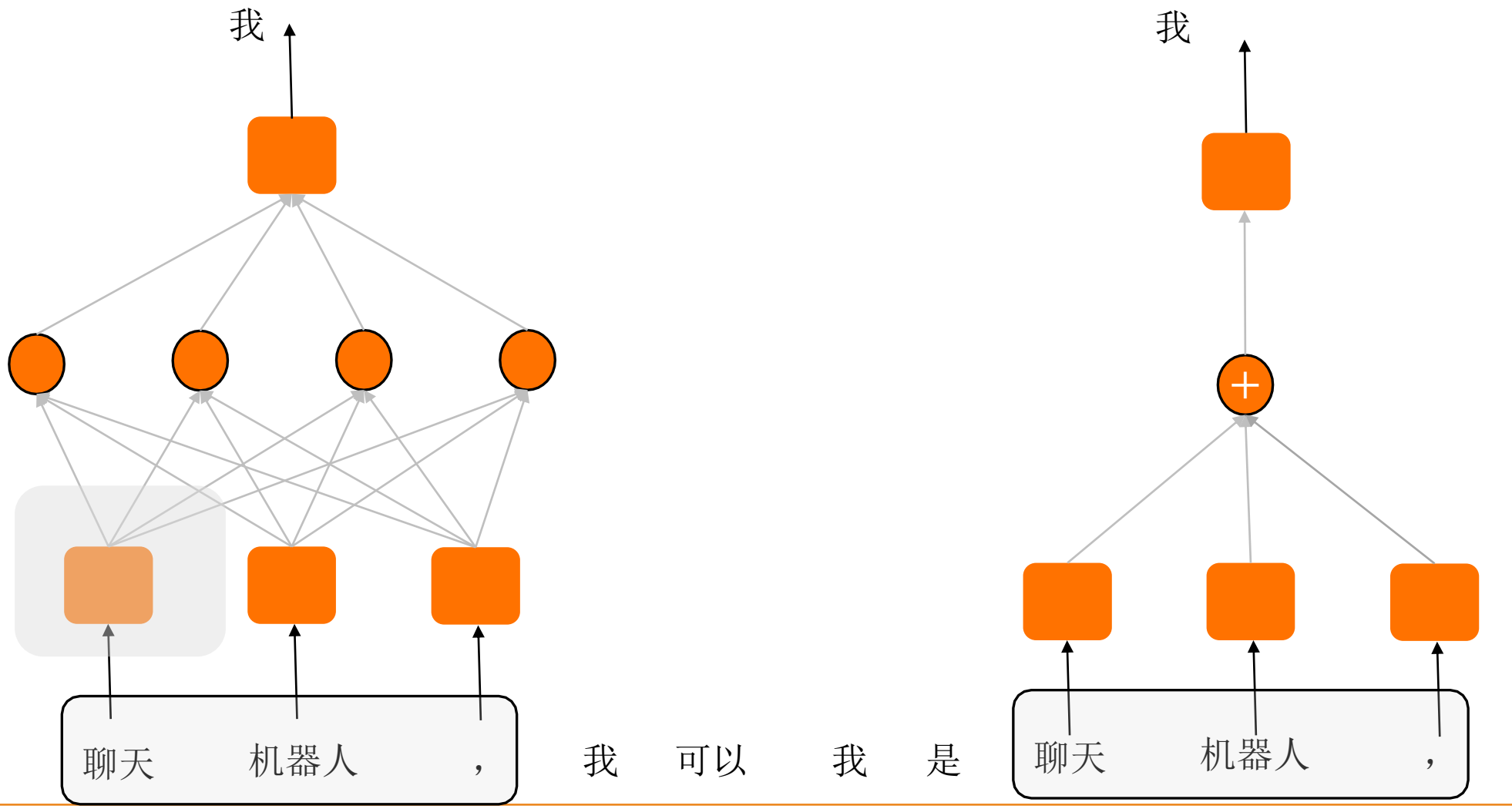


已知上下文预测当前词

已知当前词预测上下文

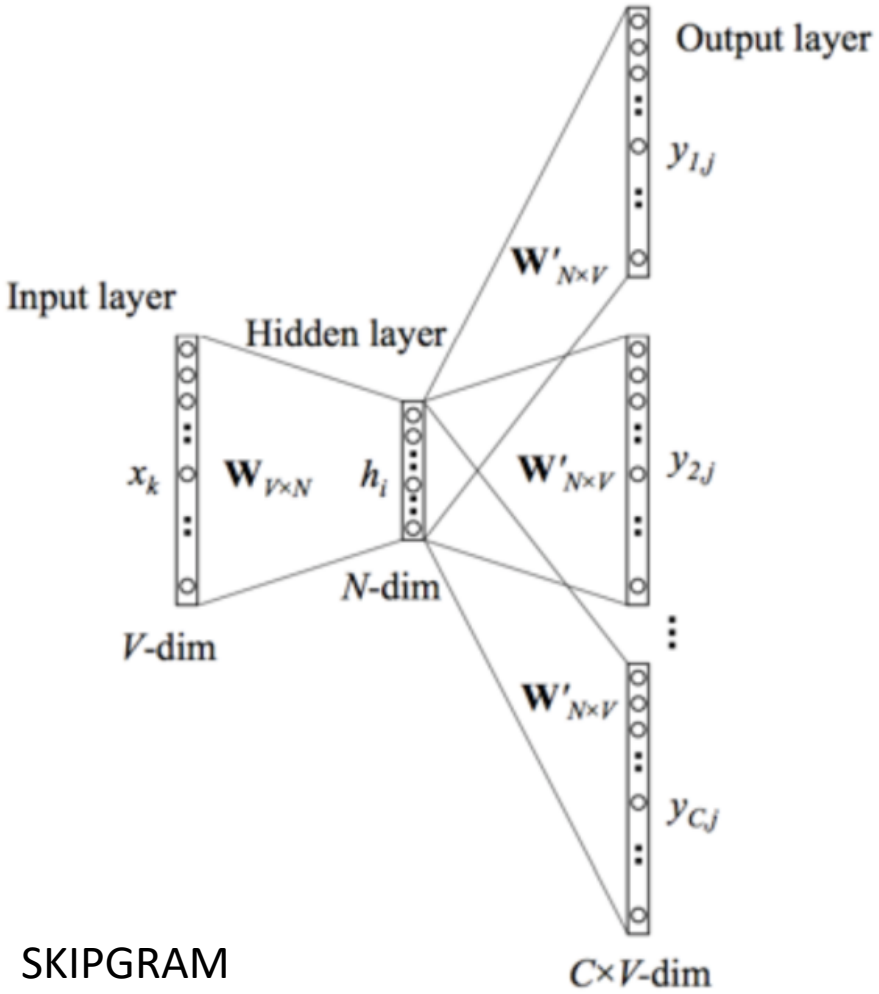
Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

CBOW(Continuous Bag of Words)



Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Skipgram



- 假设输入的句子是：

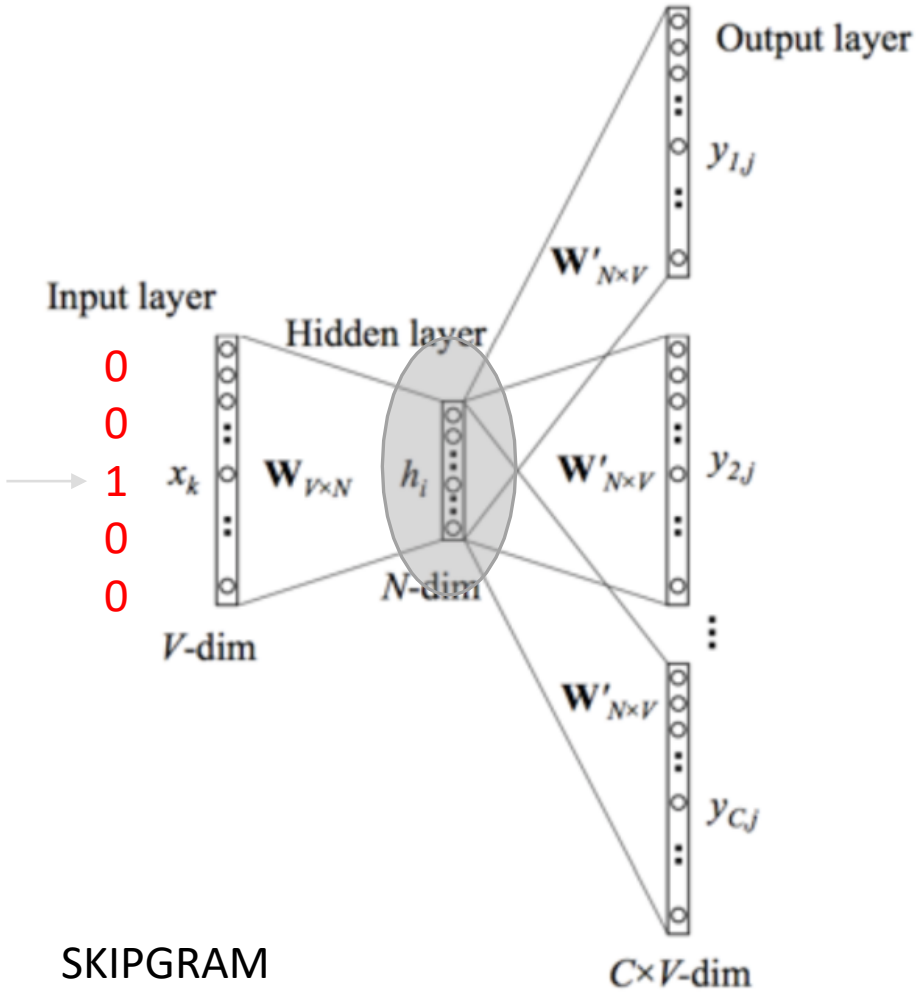
  进入 21世纪 以来，网络 信息 技术 已经 渗透到 经济发展 和 社会生活 的 方方面面

- 假设窗口为3，则得到的分块如下：

  进入_21世纪_以来_网络_信息_技术_已经
  21世纪_以来_网络_信息_技术_已经_渗透到
  以来_网络_信息_技术_已经_渗透到_经济发展
  网络_信息_技术_已经_渗透到_经济发展_和
  信息_技术_已经_渗透到_经济发展_和_社会生活
  技术_已经_渗透到_经济发展_和_社会生活_的
  已经_渗透到_经济发展_和_社会生活_的_方方面面

Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Skipgram



SKIPGRAM

- 考虑第一个序列

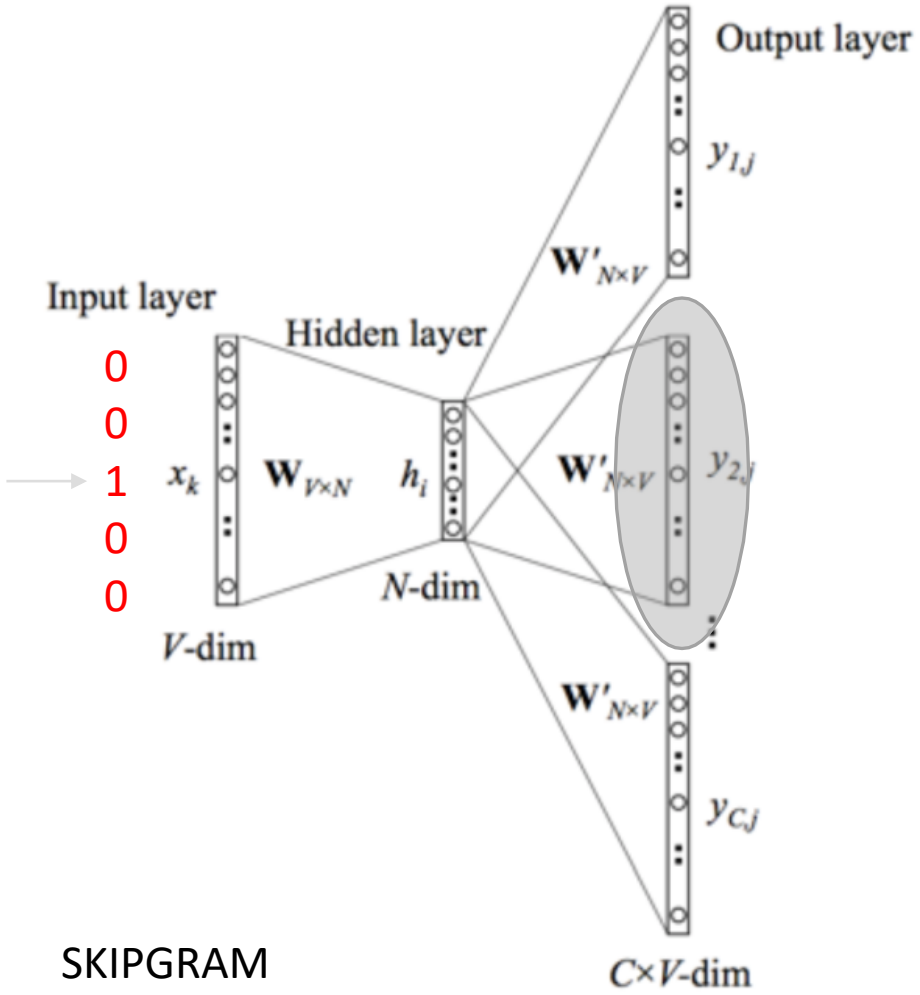
  进入_21世纪_以来_网络_信息_技术_已经

- 网络就是中心词，假设它的one-hot编码是：x_k=00100

$$h_i = W_{V\times N} x_k$$

Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Skipgram



SKIPGRAM

- 考虑第一个序列

  进入_21世纪_以来_网络_信息_技术_已经

- 网络就是中心词，假设它的one-hot编码是：x_k=00100

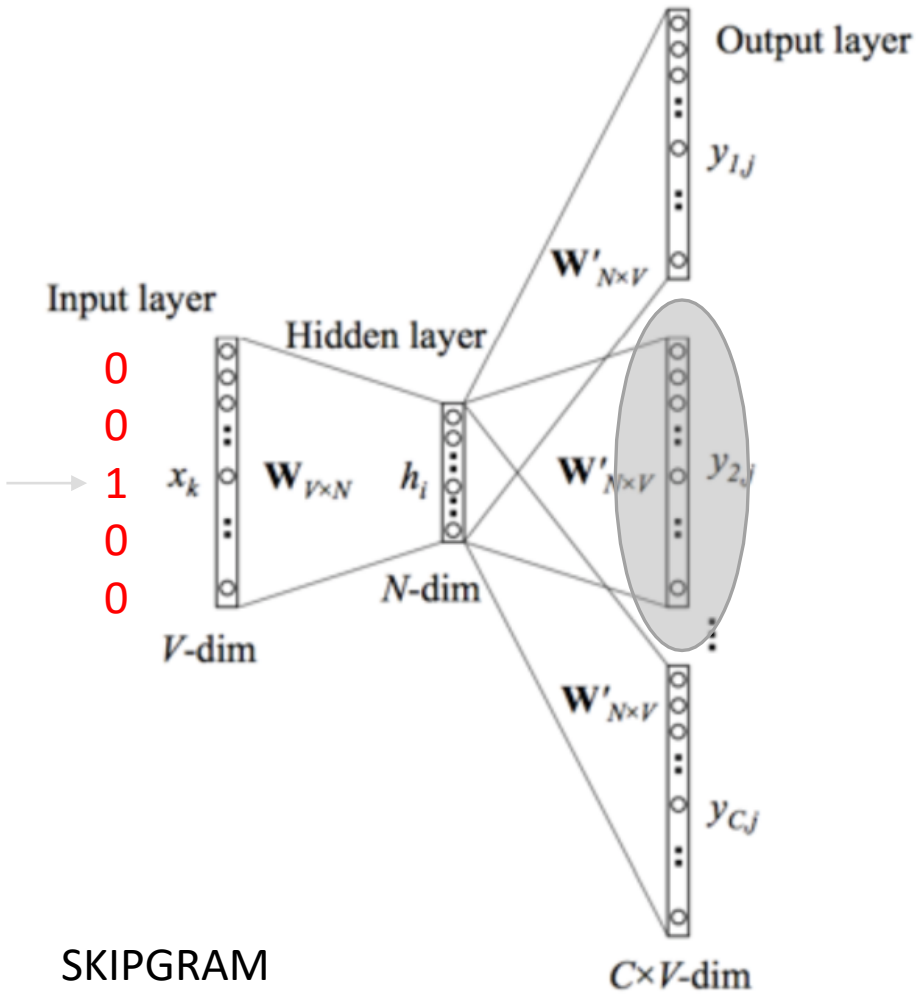
$$h_i = W_{V\times N} x_k$$

$$y_{i,j} = S(W'_{N\times V} h_i) \qquad S(x_i) = \frac{\exp(xi)}{\sum_i \exp(xi)}$$

- $y_{i,j}$表示的是上下文中的第i个单词隶属于单词表中第j个单词的概率

Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Skipgram



SKIPGRAM

- 考虑第一个序列

  进入_21世纪_以来_网络_信息_技术_已经

- 网络就是中心词，假设它的one-hot编码是：x_k=00100

$$h_i = W_{V\times N} x_k$$

$$y_{i,j} = S(W'_{N\times V} h_i) \qquad S(x_i) = \frac{\exp(xi)}{\sum_i \exp(xi)}$$

- $y_{i,j}$表示的是上下文中的第i个单词隶属于单词表中第j个单词的概率
- 给定了上下文，要让这个y最大，从而梯度下降，更新W‘
- W就存储了所有单词的词向量

Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781