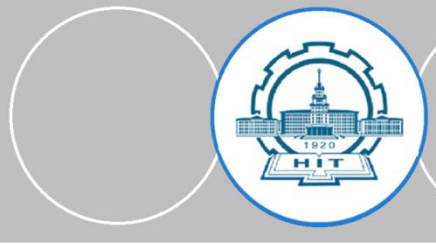


词向量简介

姓 名 陈华杰

导 师 叶允明 教授

目录

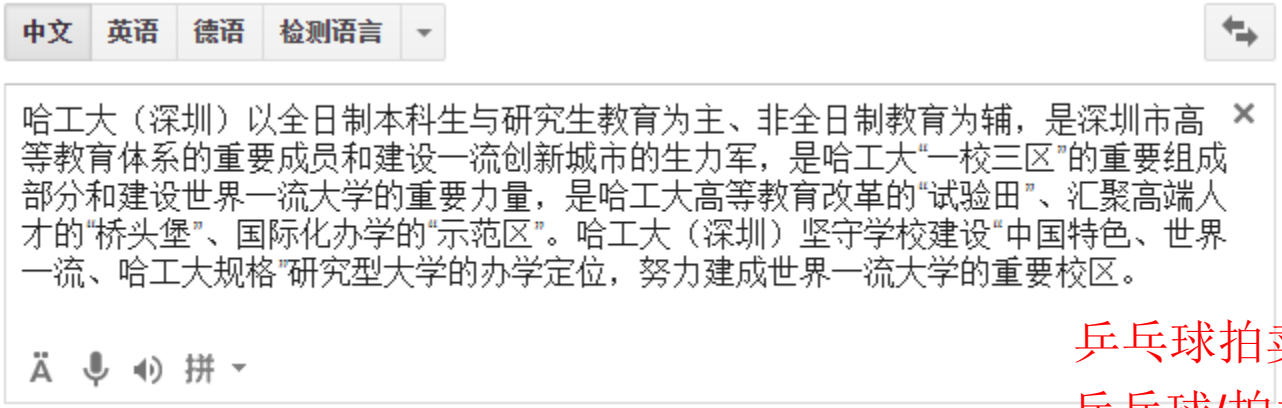
1 引言——自然语言处理

自然语言处理技术的产生可以追溯到20世纪50年代，他是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。随着计算机技术的发展，自然预处理逐渐发展成一门相对独立的学科。

相关的研究问题：

- 分词与词性标注

1 引言——自然语言处理



乒乓球拍卖完了。

乒乓球/拍卖/完了。

乒乓/球拍/卖完/了。

词性分析:

哈工大	（	深圳	）	以	全日制	本科生	与	研究生	教育	为主	、	
非全日制	教育	为辅	，	是	深圳市	高等教育	体系	的	重要	成员	和	
建设	一流	创新	城市	的	生力军	，	是	哈工大	“	一	校	三
区	”	的	重要	组成部分	和	建设	世界	一流	大学	的	重要	力量
，	是	哈工大	高等教育	改革	的	“	试验田	”	、	汇聚	高端	
人才	的	“	桥头堡	”	、	国际化	办学	的	“	示范区	”	。
哈工大	（	深圳	）	坚守	学校	建设	“	中国	特色	、	世界	一流
、	哈工大	规格	”	研究型	大学	的	办学	定位	，	努力	建成	
世界	一流	大学	的	重要	校区	。						

词性类别图示:

组织	标点符号	地名
介词	名词	连词
动词	区别词	助词
形容词	数词	

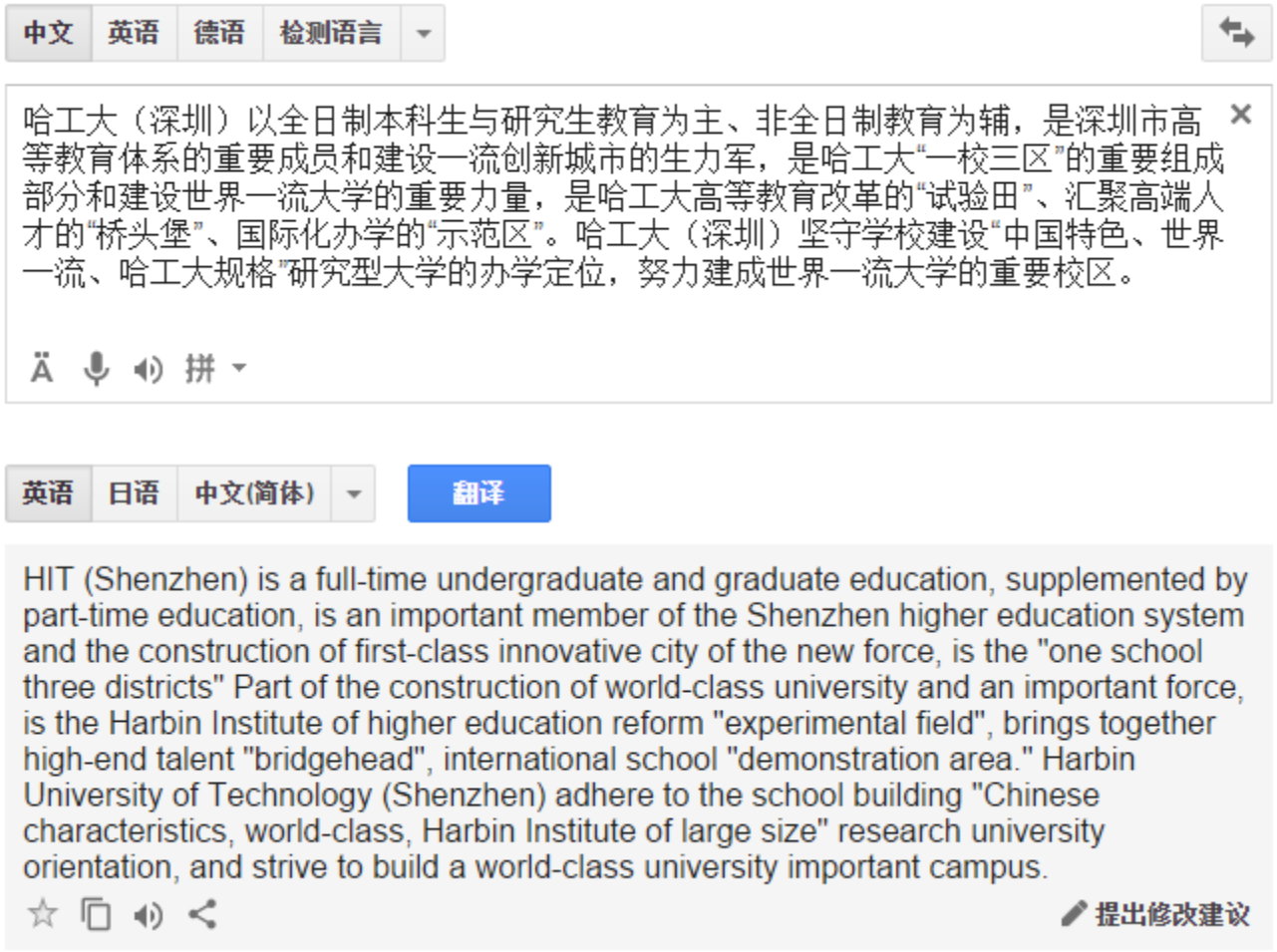
1 引言——自然语言处理

自然语言处理技术的产生可以追溯到20世纪50年代，他是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。随着计算机技术的发展，自然预处理逐渐发展成一门相对独立的学科。

相关的研究问题：

- 分词与词性标注
- 机器翻译

1 引言——自然语言处理


中文
英语
德语
检测语言
哈工大（深圳）以全日制本科生与研究生教育为主、非全日制教育为辅，是深圳市高等教育体系的重要成员和建设一流创新城市的生力军，是哈工大“一校三区”的重要组成部分和建设世界一流大学的重要力量，是哈工大高等教育改革的“试验田”、汇聚高端人才的“桥头堡”、国际化办学的“示范区”。哈工大（深圳）坚守学校建设“中国特色、世界一流、哈工大规格”研究型大学的办学定位，努力建成世界一流大学的重要校区。
拼
英语
日语
中文(简体)
翻译
HIT (Shenzhen) is a full-time undergraduate and graduate education, supplemented by part-time education, is an important member of the Shenzhen higher education system and the construction of first-class innovative city of the new force, is the "one school three districts" Part of the construction of world-class university and an important force, is the Harbin Institute of higher education reform "experimental field", brings together high-end talent "bridgehead", international school "demonstration area." Harbin University of Technology (Shenzhen) adhere to the school building "Chinese characteristics, world-class, Harbin Institute of large size" research university orientation, and strive to build a world-class university important campus.
提出修改建议

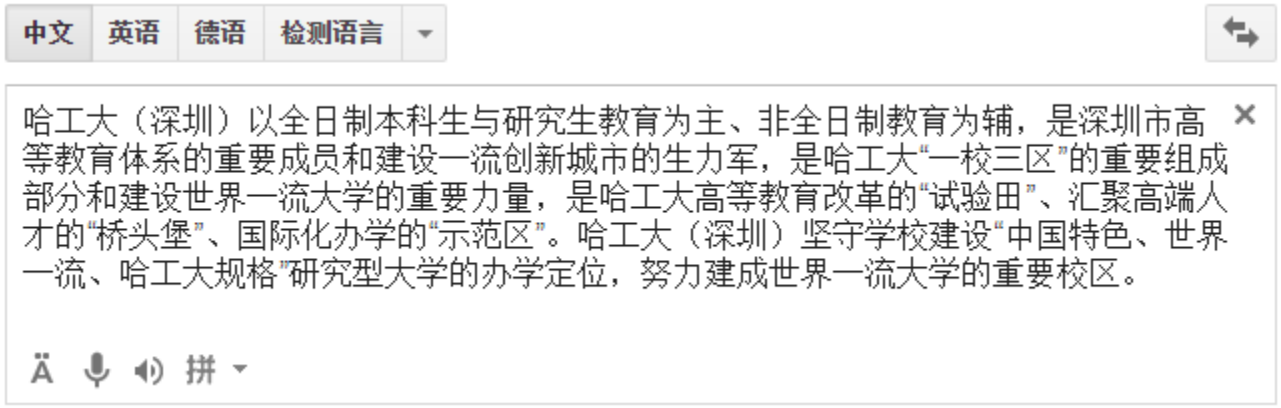
1 引言——自然语言处理

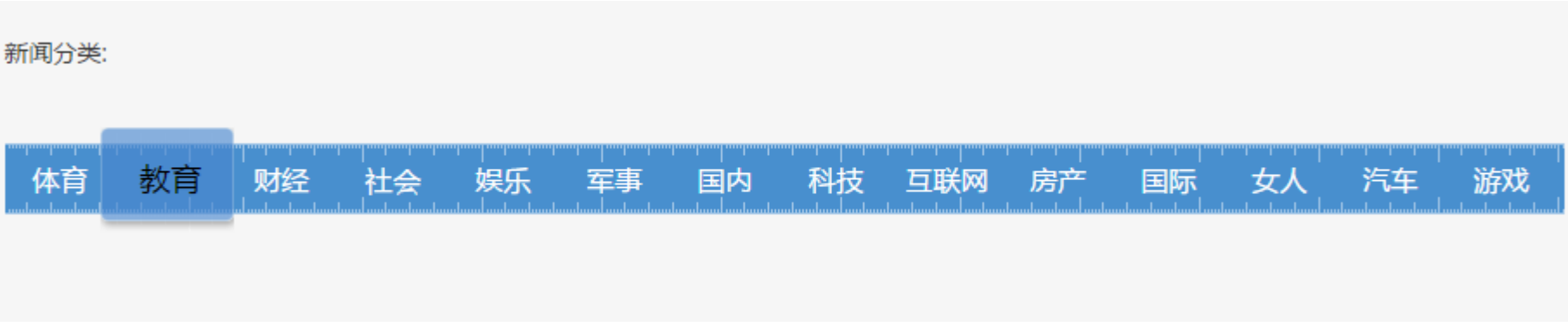
自然语言处理技术的产生可以追溯到20世纪50年代，他是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。随着计算机技术的发展，自然预处理逐渐发展成一门相对独立的学科。

相关的研究问题：

- 分词与词性标注
- 机器翻译
- 文本分类

1 引言——自然语言处理


中文
英语
德语
检测语言
哈工大（深圳）以全日制本科生与研究生教育为主、非全日制教育为辅，是深圳市高等教育体系的重要成员和建设一流创新城市的生力军，是哈工大“一校三区”的重要组成部分和建设世界一流大学的重要力量，是哈工大高等教育改革的“试验田”、汇聚高端人才的“桥头堡”、国际化办学的“示范区”。哈工大（深圳）坚守学校建设“中国特色、世界一流、哈工大规格”研究型大学的办学定位，努力建成世界一流大学的重要校区。
拼


新闻分类:
体育
教育
财经
社会
娱乐
军事
国内
科技
互联网
房产
国际
女人
汽车
游戏

1 引言——自然语言处理

自然语言处理技术的产生可以追溯到20世纪50年代，他是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。随着计算机技术的发展，自然预处理逐渐发展成一门相对独立的学科。

相关的研究问题：

- 分词与词性标注
- 机器翻译
- 文本分类
- 自动问答

1 引言——自然语言处理


智能机器人
卖到假货怎么办
16-10-29 17:16:26
亲，淘宝网对于假货的处理非常严格，会配合消费者妥善处理，一旦核实凭证有效，淘宝将支持退款处理，同时对于卖家售假行为，淘宝也将严肃给予处罚。给您带来不便，请您谅解！
若您收到的商品怀疑是假货，请按照以下步骤操作：
1.若交易还未确认收货，您可以登录【我的淘宝】—【已买到的宝贝】找到对应订单，在订单后面点击【退款】（如何申请退款？），同时上传商品是假货的相关凭证。
2.若订单已经交易成功且时间在15天内，您可以登录【我的淘宝】—【已买到的宝贝】找到对应订单，在订单后面点击【申请售后】（如何申请售后？），同时上传商品是假货的相关凭证。
3. 如果是交易成功15天后，如果您可以提供有效凭证，请您在电脑上尝试在页面点击【申请售后】-【点此查看】，在跳转的新页面点击橙红色按钮【已过维权期处理】（入口开放时间是：9:00—21:00），上传相关假货凭证；如果不能申请售后，建议可以咨询客服哦。
以上内容对您有帮助吗？ 有 没有
16-10-29 17:16:26
请用一句话简要的描述你的问题，比如"如何开店"
发送
保存聊天记录
满意度评价
还可以输入30字

1 引言——自然语言处理

自然语言处理技术的产生可以追溯到20世纪50年代，他是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。随着计算机技术的发展，自然预处理逐渐发展成一门相对独立的学科。

相关的研究问题：

- 分词与词性标注
- 机器翻译
- 文本分类
- 自动问答
- ……

1 引言——自然语言处理

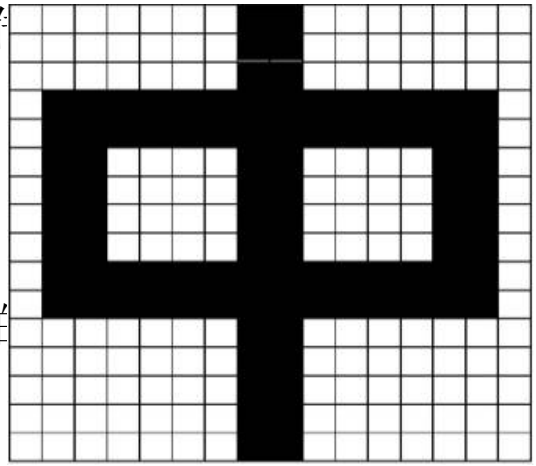
汉字在计算机中的表示：

- 输入码：输入码就是使用英文键盘输入汉字时的编码。

如“保”字，用全拼，输入码为码为“BAO”，用区位码，输入码为“1703”，用五笔字型则为“WKS”。

- 汉字内码：指计算机内部存储，处理加工和传
  1符号组成的代码。

常见的有ASCII码，GB2312，UTF-8



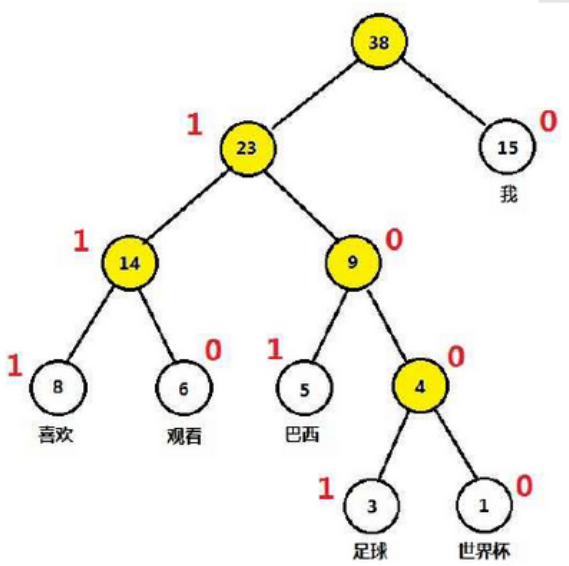
- 汉字字模码：汉字字模就是用0、1表示汉字的
  *n列的正方形内

1 引言——自然语言处理

常用的词表示方法：

- 哈夫曼码

汉字	频数	哈夫曼码
我	15	0
喜欢	8	111
观看	6	110
巴西	5	101
足球	3	1001
世界杯	1	1000



例　句：我 喜欢 观看 巴西 足球 世界杯
传输码：0　111　110　101　1001 1000

1 引言——自然语言处理

常用的词表示方法：

- One-hot Representation

“话筒”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 …]

“麦克”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 …]

每个词都是茫茫 0 海中的一个 1。

存在的问题：

任意两个词之间都是孤立的。光从这两个向量中看不出两个词是否有关系。

- 词向量表示法：用一个相对较短的向量表示

“话筒”表示为 [0.51, 0.72, 0.15, 0.24, 0.89]

“麦克”表示为[0.50, 0.73, 0.15, 0.20, 0.95]

词向量的相关研究

2 基础篇——词向量的相关研究

在实际应用中，我们经常需要解决这样一类问题：如何计算一个句子的概率？

$$\mathrm{W} = w_1^T = (w_1, w_2, w_3, \cdots, w_T)$$

(1) 统计语言模型

$$\begin{aligned}\mathrm{P}(\mathrm{W}) = \mathrm{P}(w_1^T) &= \mathrm{P}(w_1, w_2, w_3, \cdots, w_T) \\ &= \mathrm{P}(w_T | w_1^{T-1})\mathrm{P}(w_1^{T-1}) \\ &= \mathrm{P}(w_T | w_1^{T-1})\mathrm{P}(w_{T-1} | w_1^{T-2})\mathrm{P}(w_1^{T-2}) \\ &= \cdots\cdots \\ &= \mathrm{P}(w_T | w_1^{T-1})\mathrm{P}(w_{T-1} | w_1^{T-2}) \cdots \mathrm{P}(w_2 | w_1)\mathrm{P}(w_1)\end{aligned}$$

条件公式：

P(AB)=P(A|B)P(B)

复杂度：$\mathrm{T}N^T$

2 基础篇——词向量的相关研究

(2) n-gram模型

条件公式：

P(A|B)=P(AB)/P(B)

$$P(w_k|w_1^{k-1}) = \frac{P(w_1^k)}{P(w_1^{k-1})}$$

当语料库足够大的时候：

$$P(w_k|w_1^{k-1}) = \frac{count(w_1^k)}{count(w_1^{k-1})}$$

仅考虑一个词的出现于前面n-1个词有个：

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-n+1}^{k-1})$$

$$= \frac{P(w_{k-n+1}^k)}{P(w_{k-n+1}^{k-1})}$$

$$= \frac{count(w_{k-n+1}^k)}{count(w_{k-n+1}^{k-1})}$$

n的大小	参数的个数
1（unigram）	2×10^5
2（bigram）	4×10^{10}
3（trigram）	8×10^{15}
4（4-gram）	16×10^{20}

*假设有200000个词

2 基础篇——词向量的相关研究

问1：n如何选取？

理论上n越大越好。特别是如今存储容量和计算性能的提升使得我们可以支持n更大的运算。但是在实际应用中随着n增大，实验效果反而下降了。例如，当n从1变到2、从2变到3时，效果提升显著。但是但n从3变到4时，效果提升就不那么显著。当n继续增大时，效果反而下降了。

- 可区别性：当n越大时，可区别性越好；
- 可靠性：当n越大时，可靠性越差。

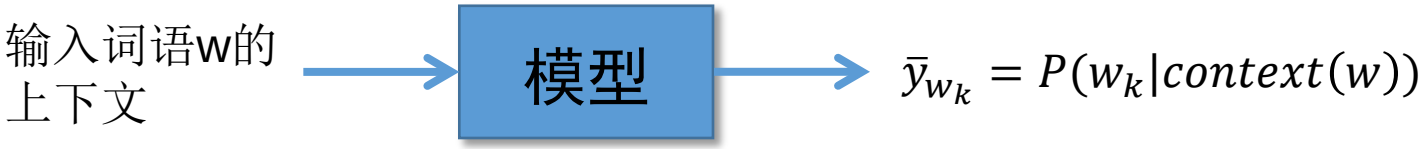
问2：$\mathrm{count}(w_1^k)=0$是否代表$\mathrm{P}(w_k|w_1^{k-1})=0$？

$\mathrm{count}(w_1^k)=\mathrm{count}(w_1^{k-1})$是否代表$\mathrm{P}(w_k|w_1^{k-1})=1$？

显然不能

2 基础篇——词向量的相关研究

(3) 神经网络语言模型



➢ 目标函数：

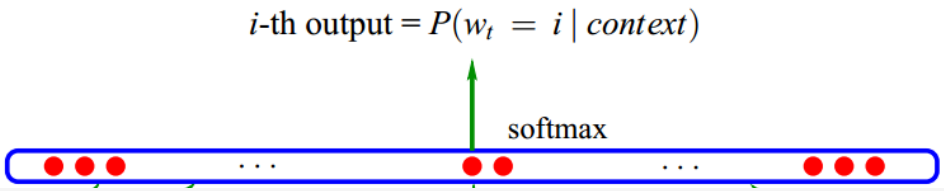
$$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$$

- k=1,2,3…N
- $\bar{y}_{w_k}$表示预测的概率
- y_{w_k}表示实际的概率

➢ 把$P(w_k|context(w))$视为关于w_k和$context(w)$的函数：

$$P(w_k|context(w)) = F(w_k, context(w), \theta)$$

2 基础篇——词向量的相关研究



A Neural Probabilistic Language Model

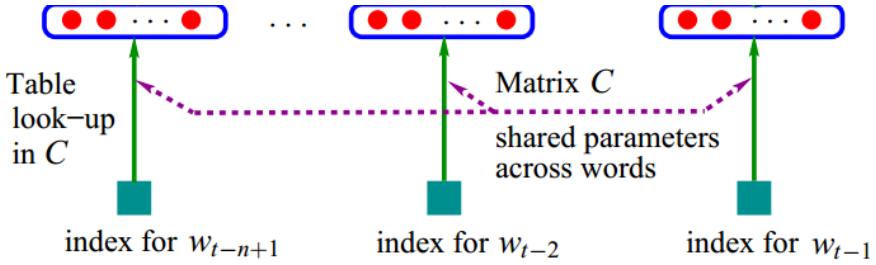
A neural probabilistic language model

[PDF] jmlr.org

Y Bengio, R Ducharme, P Vincent, C Jauvin - journal of machine learning ..., 2003 - jmlr.org

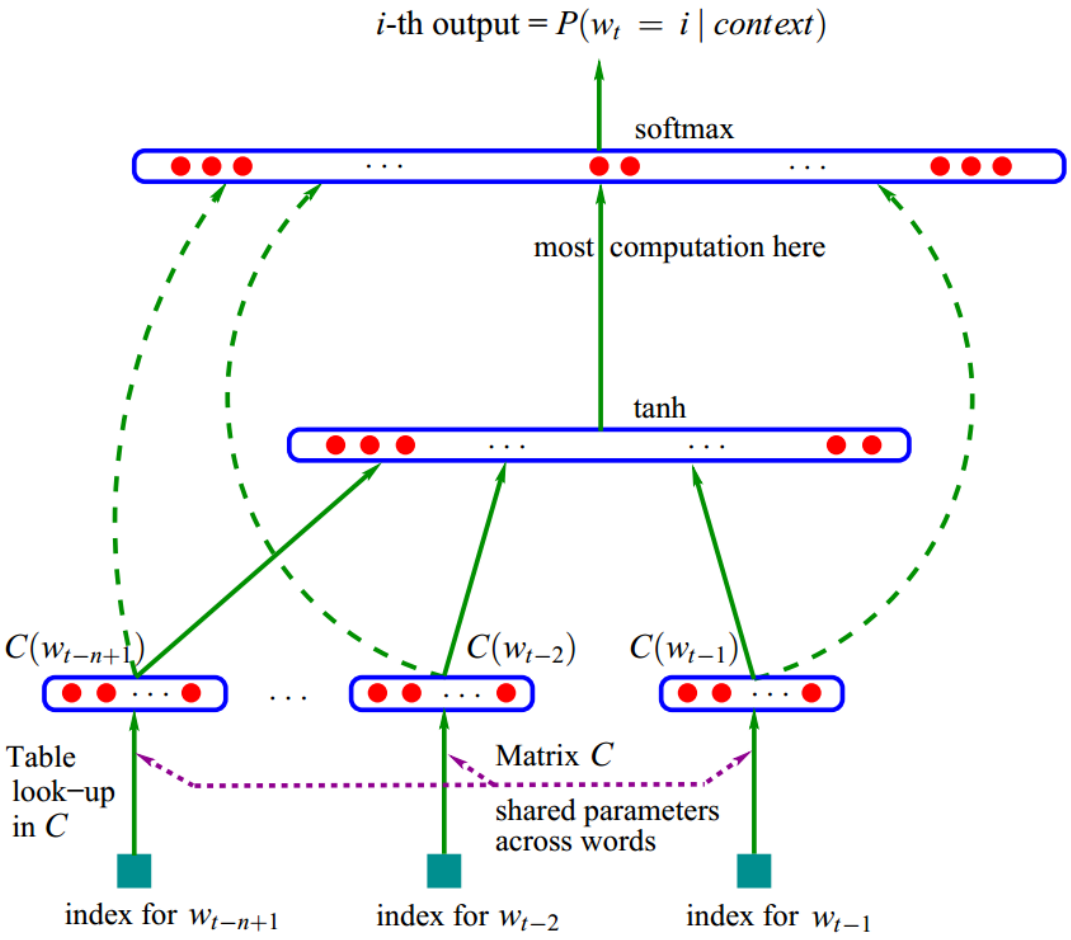
Abstract A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. This is intrinsically difficult because of the curse of dimensionality: a word sequence on which the model will be tested is likely to be different ...

被引用次数：1812 相关文章 所有 54 个版本 引用 保存



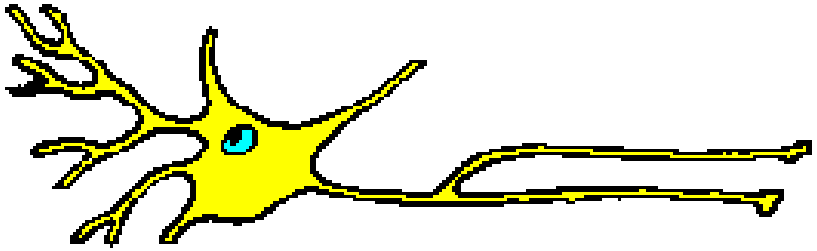
Yoshua Bengio等人在2003年发表了《A Neural Probabilistic Language Model》

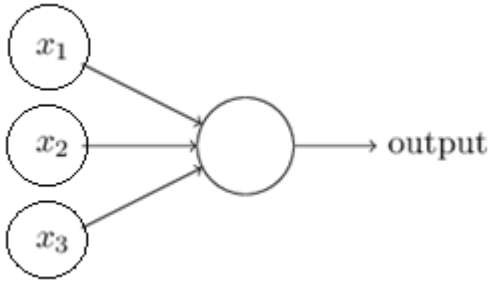
2 基础篇——词向量的相关研究

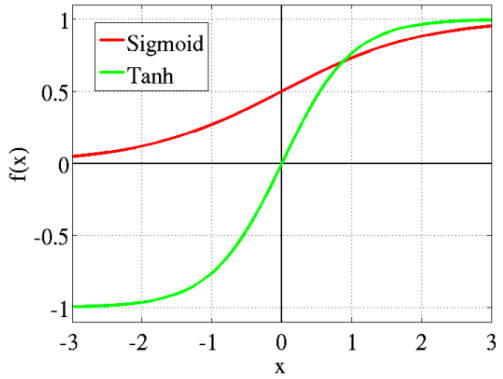


Yoshua Bengio等人在2003年发表了《A Neural Probabilistic Language Model》

2 基础篇——词向量的相关研究







$$a = \begin{cases} 0, & if \ \sum_i w_i x_i + b > threshold \\ 1, & if \ \sum_i w_i x_i + b \leq threshold \end{cases}$$

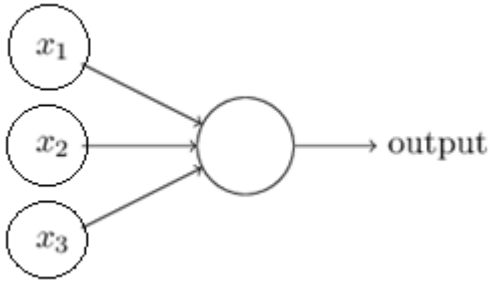
$$z = \sum_i w_i x_i + b$$

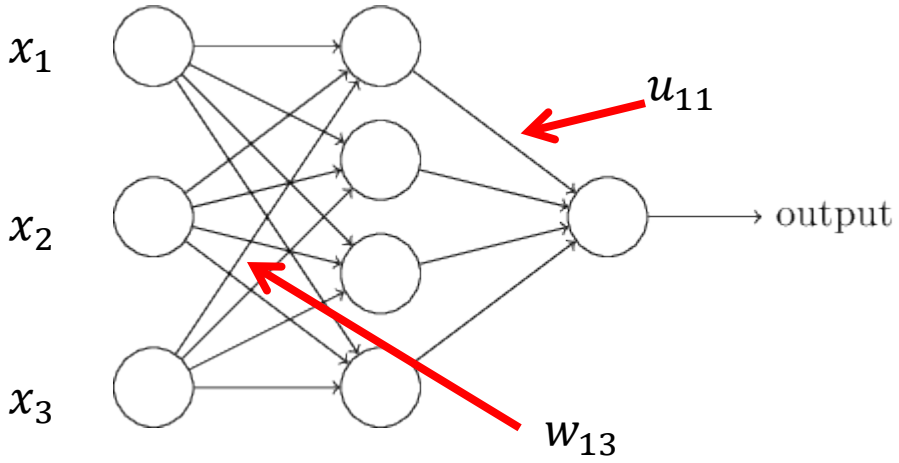
$$a = f(z)$$

$$a = sigmod(z) = \frac{1}{1+e^{-z}}$$

$$a = tanh(\sum_i w_i x_i + b) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

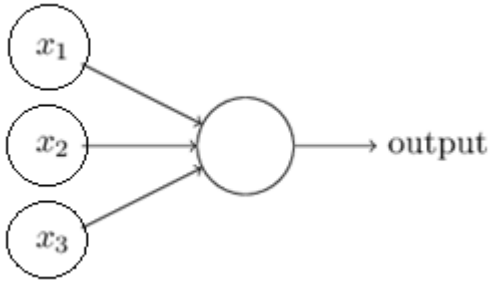
2 基础篇——词向量的相关研究

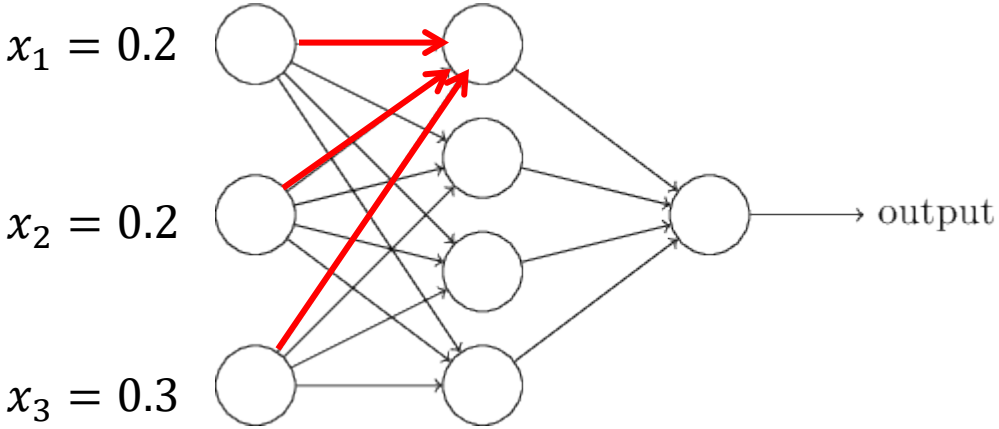




$z_j = \sum_i w_{ji} x_i$ $a_j = \tanh(z_j)$

基础篇——词向量的相关研究



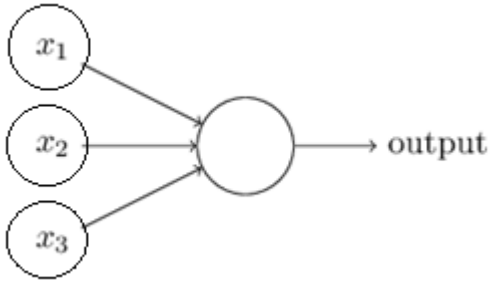


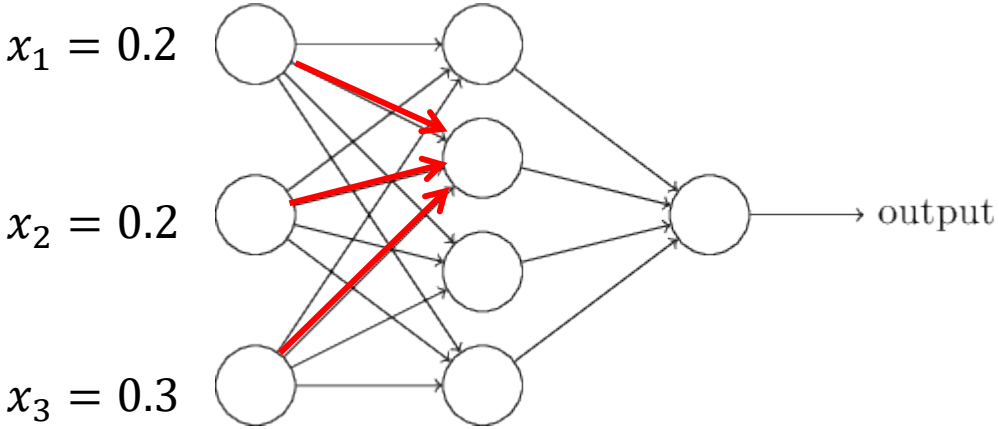
$z_j = \sum_i w_{ji} x_i$ $\quad$ $a_j = \tanh(z_j)$

$w_{11} = 0.2$	$w_{21} = 0.3$	$w_{31} = 0.1$	$p_1 = 0.1$	$z_1 = 0.25$	$a_1 = 0.24$
$w_{12} = 0.3$	$w_{22} = 0.4$	$w_{32} = 0.2$	$p_2 = 0.2$		
$w_{13} = 0.7$	$w_{23} = 0.4$	$w_{33} = 0.6$	$p_3 = 0.3$		
$w_{14} = 0.1$	$w_{24} = 0.6$	$w_{34} = 0.8$	$p_4 = 0.4$		

正向传播算法

基础篇——词向量的相关研究



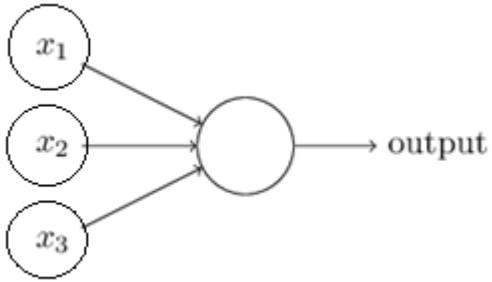


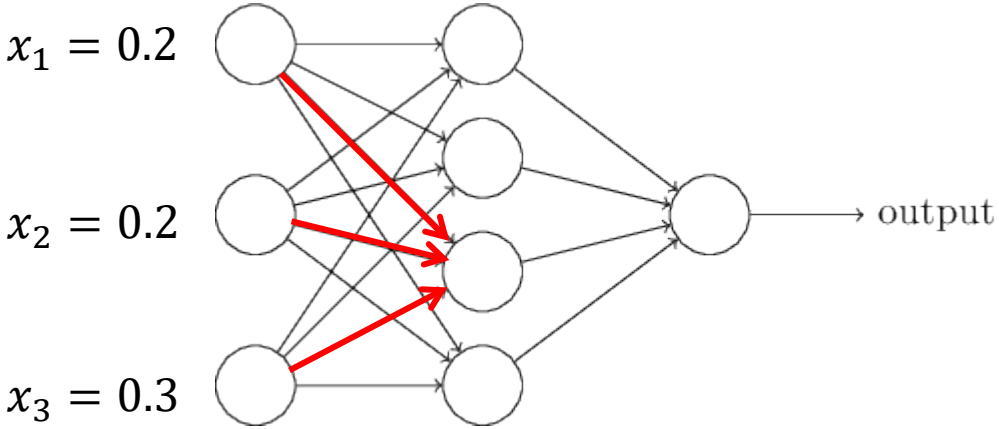
$z_j = \sum_i w_{ji} x_i$ $\quad\quad$ $a_j = \tanh(z_j)$

$w_{11} = 0.2$	$w_{21} = 0.3$	$w_{31} = 0.1$	$p_1 = 0.1$	$z_1 = 0.25$	$a_1 = 0.24$
$w_{12} = 0.3$	$w_{22} = 0.4$	$w_{32} = 0.2$	$p_2 = 0.2$	$z_2 = 0.2$	$a_2 = 0.19$
$w_{13} = 0.7$	$w_{23} = 0.4$	$w_{33} = 0.6$	$p_3 = 0.3$		
$w_{14} = 0.1$	$w_{24} = 0.6$	$w_{34} = 0.8$	$p_4 = 0.4$		

正向传播算法

基础篇——词向量的相关研究



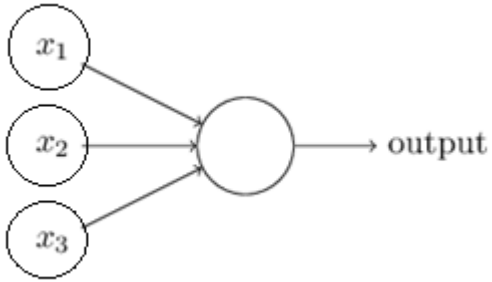


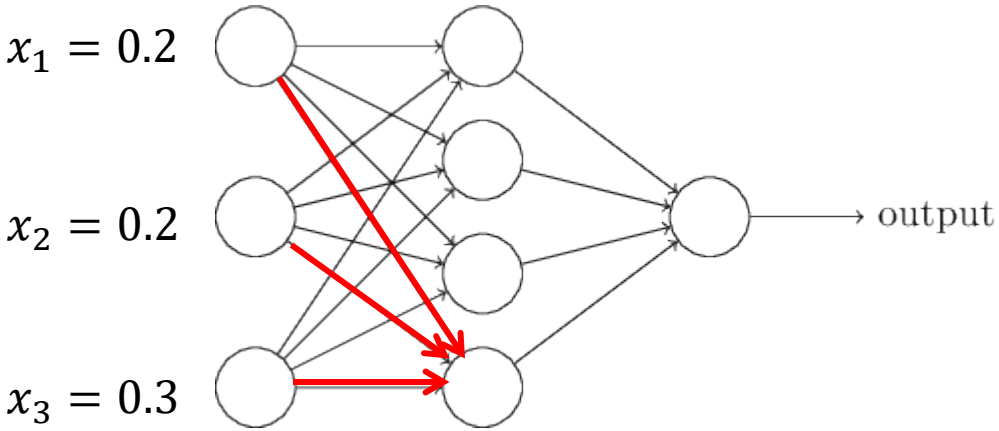
$z_j = \sum_i w_{ji} x_i$ $\quad\quad$ $a_j = \tanh(z_j)$

$w_{11} = 0.2$	$w_{21} = 0.3$	$w_{31} = 0.1$	$p_1 = 0.1$	$z_1 = 0.25$	$a_1 = 0.24$
$w_{12} = 0.3$	$w_{22} = 0.4$	$w_{32} = 0.2$	$p_2 = 0.2$	$z_2 = 0.2$	$a_2 = 0.19$
$w_{13} = 0.7$	$w_{23} = 0.4$	$w_{33} = 0.6$	$p_3 = 0.3$	$z_3 = 0.4$	$a_3 = 0.37$
$w_{14} = 0.1$	$w_{24} = 0.6$	$w_{34} = 0.8$	$p_4 = 0.4$		

正向传播算法

基础篇——词向量的相关研究



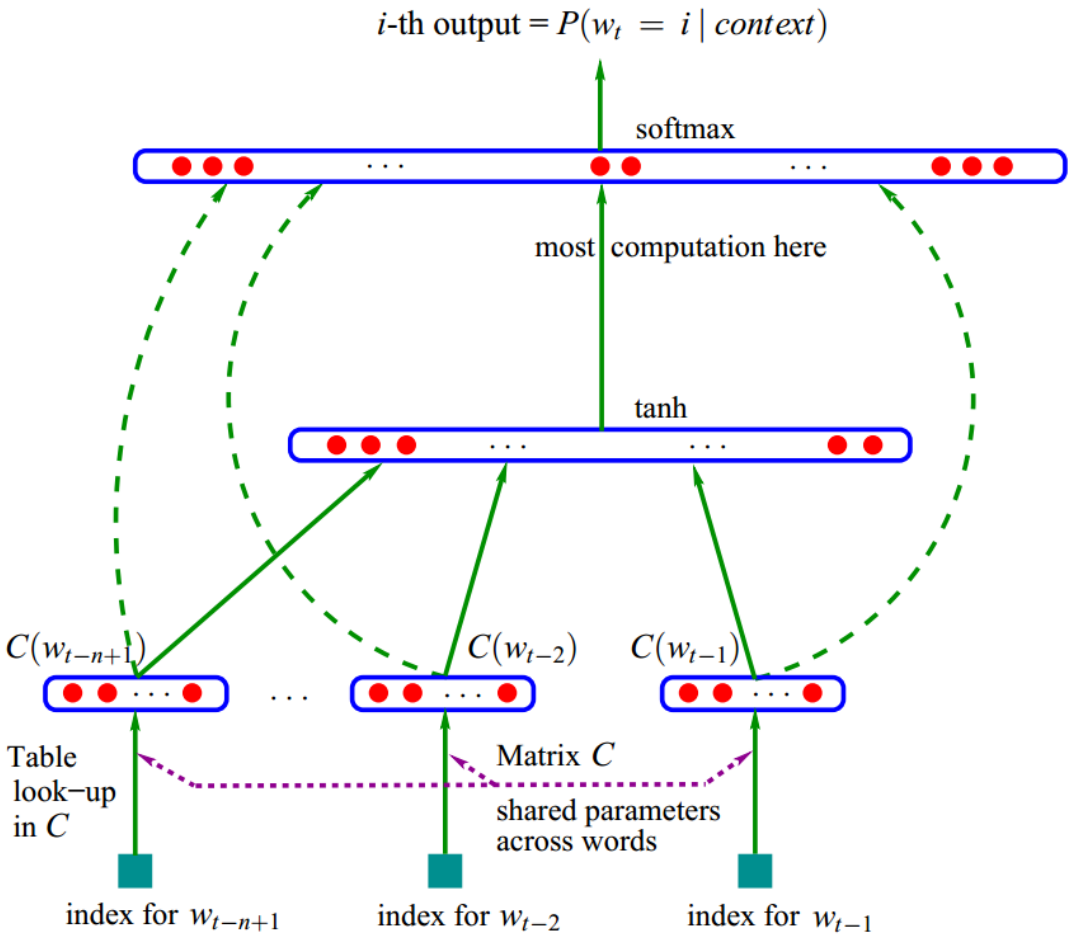


$z_j = \sum_i w_{ji} x_i$ $\quad$ $a_j = \tanh(z_j)$

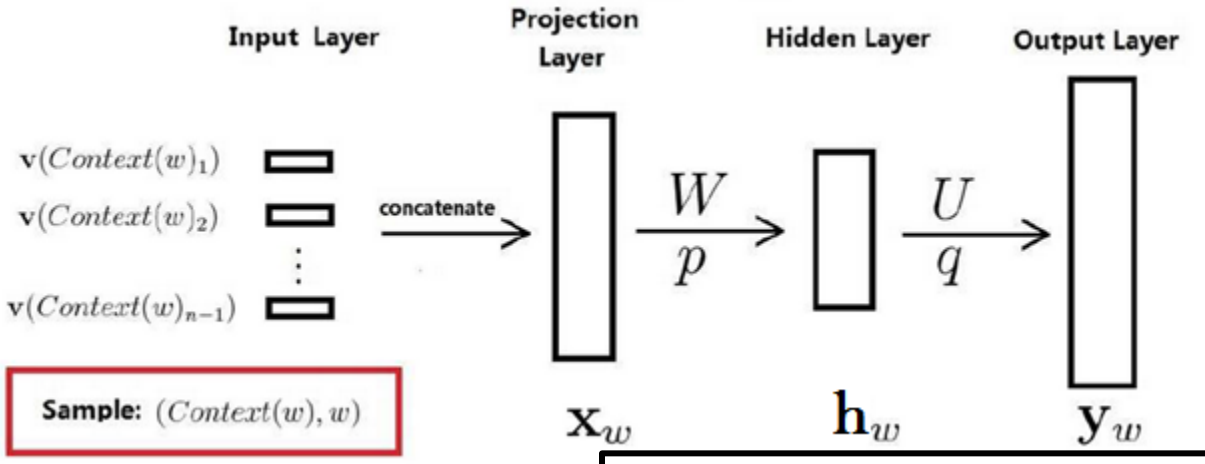
$w_{11} = 0.2$	$w_{21} = 0.3$	$w_{31} = 0.1$	$p_1 = 0.1$	$z_1 = 0.25$	$a_1 = 0.24$
$w_{12} = 0.3$	$w_{22} = 0.4$	$w_{32} = 0.2$	$p_2 = 0.2$	$z_2 = 0.2$	$a_2 = 0.19$
$w_{13} = 0.7$	$w_{23} = 0.4$	$w_{33} = 0.6$	$p_3 = 0.3$	$z_3 = 0.4$	$a_3 = 0.37$
$w_{14} = 0.1$	$w_{24} = 0.6$	$w_{34} = 0.8$	$p_4 = 0.4$	$z_4 = 0.38$	$a_4 = 0.36$

正向传播算法

2 基础篇——词向量的相关研究



2 基础篇——词向量的相关研究



$h_w = \tanh(Wx_w + p)$

$z^o = Uh_w + q$

$\bar{y}_w = softmax(z^o)$ $\quad \bar{y}_{w_i} = \frac{e^{z_i^o}}{\sum_j e^{z_j^o}}$

误差函数：$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$

待求参数：$\theta = (x_w, W, p, U, q)$

The backpropagation algorithm

1. Input x_w: the vector of context(w)
2. Feedforward:

$$h_w = \tanh(Wx_w + p)$$
$$z^o = Uh_w + q$$
$$\bar{y}_w = softmax(z^o)$$

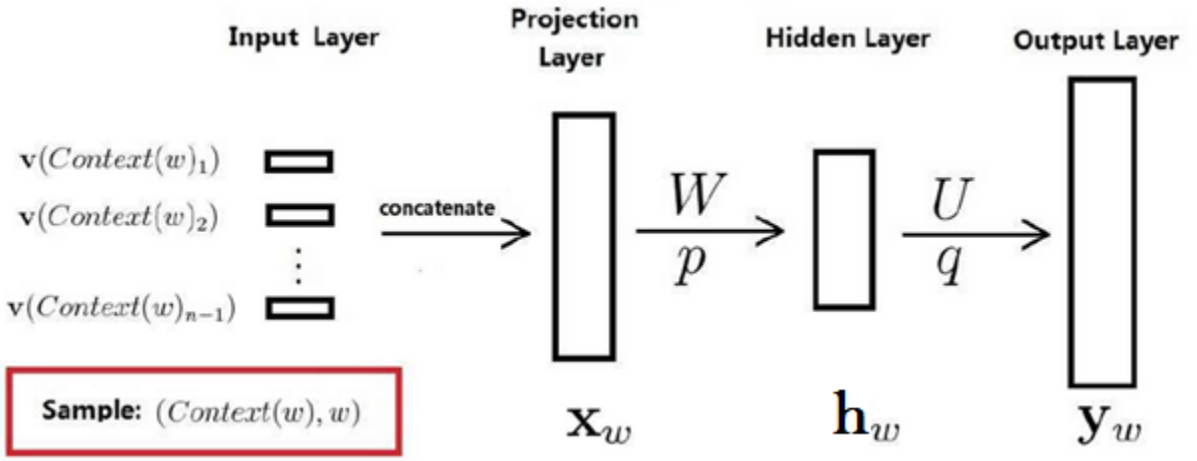
3. Output error:

$$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$$

3. Backpropagate:

$$\theta = \theta - \eta \frac{\partial C}{\partial \theta}$$

2 基础篇——词向量的相关研究



$h_w = \tanh(Wx_w + p)$

$z^o = Uh_w + q$

$\bar{y}_w = softmax(z^o)$ $\quad \bar{y}_{w_i} = \dfrac{e^{z_i^o}}{\sum_j e^{z_j^o}}$

误差函数：$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$

待求参数：$\theta = (x_w, W, p, U, q)$

问题：如何计算$\frac{\partial C}{\partial \theta}$？

(1)首先求解$\frac{\partial \bar{y}_{w_i}}{\partial z_j^o}$

If i =j:

$$\frac{\partial \bar{y}_{w_i}}{\partial z_j^o} = \frac{\partial}{\partial z_j^o}\frac{e^{z_i^o}}{\sum_j e^{z_j^o}} = \frac{(e^{z_i^o})'\sum_j e^{z_j^o} - e^{z_i^o} e^{z_i^o}}{(\sum_j e^{z_j^o})^2} = \bar{y}_{w_i}(1 - \bar{y}_{w_i})$$

If i ≠j:

$$\frac{\partial \bar{y}_{w_i}}{\partial z_j^o} = \frac{\partial}{\partial z_j^o}\frac{e^{z_i^o}}{\sum_j e^{z_j^o}} = \frac{0\sum_j e^{z_j^o} - e^{z_j^o} e^{z_i^o}}{(\sum_j e^{z_j^o})^2} = -\bar{y}_{w_i}\bar{y}_{w_j}$$

基础篇——词向量的相关研究

$$h_w = \tanh(Wx_w + p)$$

$$z^o = Uh_w + q$$

$$\bar{y}_w = softmax(z^o)$$

误差函数：$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$

待求参数：$\theta = (x_w, W, p, U, q)$

(2)求解$\frac{\partial C}{\partial q_j}$:

$$\frac{\partial C}{\partial q_j} = \frac{\partial C}{\partial \bar{y}_{w_j}} \frac{\partial \bar{y}_{w_j}}{\partial z_j} \frac{\partial z_j}{\partial q_j}$$

$$= \frac{\partial C}{\partial \bar{y}_{w_j}} \frac{\partial \bar{y}_{w_j}}{\partial z_j} \frac{\partial \sum_i u_{ji} h_i + q_j}{\partial q_j}$$

$$= \frac{\partial}{\partial \bar{y}_{w_j}} \left(-\sum_k y_{w_k} \log(\bar{y}_{w_k})\right) \frac{\partial \bar{y}_{w_j}}{\partial z_j}$$

$$= -\sum_k y_{w_k} \frac{1}{\bar{y}_{w_k}} \frac{\partial \bar{y}_{w_i}}{\partial z_j}$$

$$= -y_{w_j} \frac{1}{\bar{y}_{w_j}} \frac{\partial \bar{y}_{w_j}}{\partial z_j} - \sum_{k \neq j} y_{w_k} \frac{1}{\bar{y}_{w_k}} \frac{\partial \bar{y}_{w_i}}{\partial z_j}$$

$$= -y_{w_j} \frac{1}{\bar{y}_{w_j}} \bar{y}_{w_j} \left(1 - \bar{y}_{w_j}\right) - \sum_{k \neq j} y_{w_k} \frac{1}{\bar{y}_{w_k}} \left(-\bar{y}_{w_k} \bar{y}_{w_j}\right)$$

$$= -y_{w_j} + y_{w_j} \bar{y}_{w_j} + \sum_{k \neq j} y_{w_k} \bar{y}_{w_j}$$

$$= \bar{y}_{w_j} - y_{w_j}$$

(3)求解$\frac{\partial C}{\partial u_{ji}}$:

$$\frac{\partial C}{\partial u_{ji}} = \frac{\partial C}{\partial \bar{y}_{w_j}} \frac{\partial \bar{y}_{w_j}}{\partial z_j} \frac{\partial z_j}{\partial u_{ji}}$$

$$= \frac{\partial C}{\partial \bar{y}_{w_j}} \frac{\partial \bar{y}_{w_j}}{\partial z_j} \frac{\partial \sum_i u_{ji} h_i + q_j}{\partial u_{ji}}$$

$$= \left(\bar{y}_{w_j} - y_{w_j}\right) h_i$$

$$\delta^o \equiv \frac{\partial C}{\partial z^o}$$

$$\frac{\partial C}{\partial q_j} = \delta^o$$

$$\frac{\partial C}{\partial u_{ji}} = \delta^o h_i$$

基础篇——词向量的相关研究

$$h_w = \tanh(Wx_w + p)$$

$$z^o = Uh_w + q$$

$$\bar{y}_w = softmax(z^o)$$

误差函数：$C = -\sum_k y_{w_k} \log(\bar{y}_{w_k})$

待求参数：$\theta = (x_w, W, p, U, q)$

(4) 求解$\frac{\partial C}{\partial p_j}, \frac{\partial C}{\partial w_{ji}}$

$$z^h = Wx_w + p$$

$$\delta^h \equiv \frac{\partial C}{\partial z^h}$$

$$\delta^h = (u^T \delta^o) \odot tank'(z^h)$$

- $\frac{\partial C}{\partial q_j} = \delta^h$
- $\frac{\partial C}{\partial u_{ji}} = \delta^h x_i$

(5) 求解$\frac{\partial C}{\partial x_w}$

$$\delta^i \equiv \frac{\partial C}{\partial z^i} = w^T \delta^h$$

$$z^i = x_w$$

$$\frac{\partial C}{\partial x_w} = \delta^i$$

$$\delta^o \equiv \frac{\partial C}{\partial z^o}$$

$$\frac{\partial C}{\partial q_j} = \delta^o$$

$$\frac{\partial C}{\partial u_{ji}} = \delta^o h_i$$

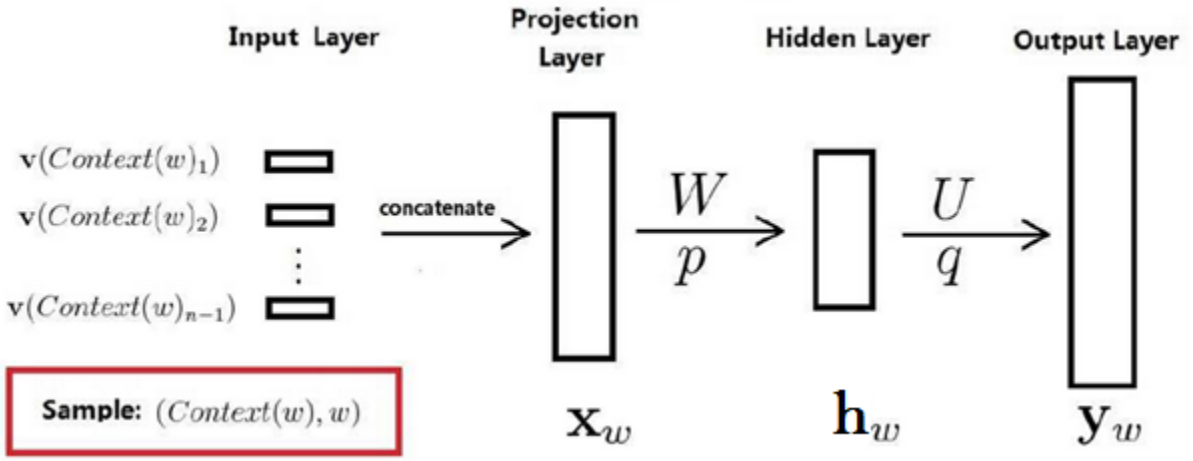
$$\delta_j^h \equiv \frac{\partial C}{\partial z^h} = \sum_k \frac{\partial C}{\partial z_k^o} \frac{\partial z_k^o}{\partial z_j^h} = \sum_k \delta_k^o \frac{\partial z_k^o}{\partial z_j^h}$$

$$z_k^o = \sum_i u_{ki} h_i + q_k = \sum_i u_{ki} tank(z_i^h) + q_k$$

$$\frac{\partial z_k^o}{\partial z_j^h} = u_{kj} tank'(z_i^h)$$

$$\delta_j^h = \sum_k \delta_k^o u_{kj} tank'(z_i^h)$$

2 基础篇——词向量的相关研究



- $v(w) \in \mathbb{R}^m$，m表示词向量的长度，一般不大于10^3这个量级；
- $X_w \in \mathbb{R}^{m(n-1)}$ ，n表示上下文的范围，一般不超过5；
- $W \in \mathbb{R}^{n_h \times (n-1)m}$，表示投影层到隐含层的参数。 n_h表示隐含层节点的个数，一般不大于10^4这个量级；
- $p \in \mathbb{R}^{n_h}$，表示投影层到隐含层的偏置；
- $h_w \in \mathbb{R}^{n_h}$，表示隐含层的输出；
- $U \in \mathbb{R}^{N \times n_h}$，表示隐含层到输出层的参数。 N表示词的个数；
- $q \in \mathbb{R}^N$，表示隐含层到输出层的偏置；

$10^3 \times (2 \times 10^5) + 10^3 \times 5 \times 10^4 + 10^4 + (2 \times 10^5) \times 10^4 + 2 \times 10^5$

2 基础篇——词向量的相关研究

问1：与one-hot-represent相比，词向量有什么优点？

- 解决了维数灾难的问题。
- 词语间的相似性可以通过词向量来体现。
- 将离散空间上的词转换到连续空间上的向量，便于计算。

问2：与n-gram语言模型相比，神经网络语言模型有什么优点？

- 模型参数更少。
- 自带平滑特性。

S1 = “A dog is running in the room.”
S2 = “A cat is running in the room.”

Count(S1) = 1000
Count(S2) = 1
P(S1) >> S(S2)

2 基础篇——词向量的相关研究

问1：与one-hot-represent相比，词向量有什么优点？

- 解决了维数灾难的问题。
- 词语间的相似性可以通过词向量来体现。
- 将离散空间上的词转换到连续空间上的向量，便于计算。

问2：与n-gram语言模型相比，神经网络语言模型有什么优点？

- 模型参数更少。
- 自带平滑特性。

A dog is running in the room
A cat is running in the room
The cat is running in a room
A dog is walking in a bedroom
The dog was walking in the room
...

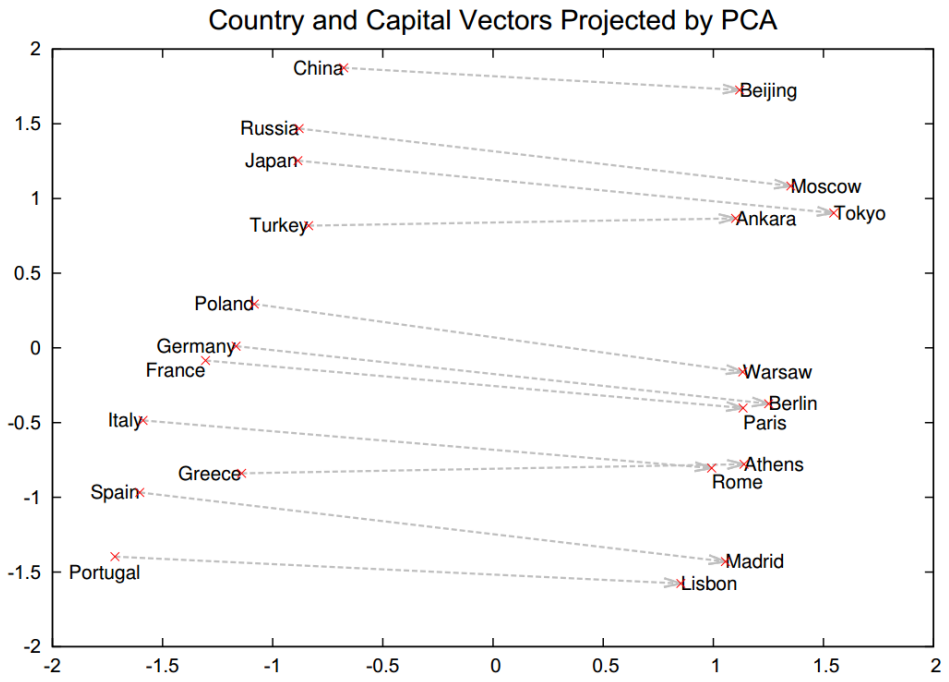
2 基础篇——词向量的相关研究

- 词向量的研究最早可以追溯到Hinton 在 1986 年的论文《Learning distributed representations of concepts》
- Hinton等人在2007 年 ICML 发表了《Three new graphical models for statistical language modelling》
- Jason Weston 等人在 2008 年的 ICML 上发表的《A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning》
- Mikolov 等人在2010 年的INTERSPEECH 上发表的《Recurrent neural network based language model》

Word2Vev 的讲解

3 进阶篇——Word2Vev的讲解

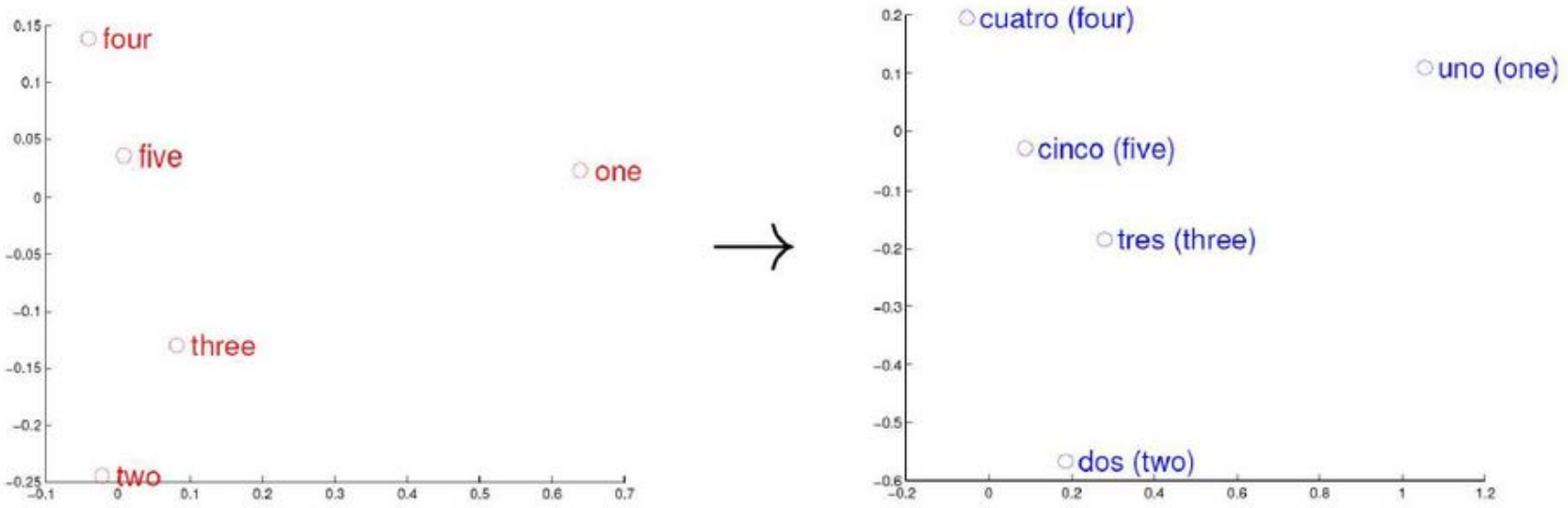
- Mikolov 等人在2013 年的NIPS上发表的《Distributed Representations of Words and Phrases and their Compositionality》



$$v(\text{China}) - v(\text{Beijing}) \approx v(\text{Russia}) - v(\text{Moscow})$$

3 进阶篇——Word2Vev的讲解

英语和西班牙语中1,2,3,4,5的分布



3 进阶篇——Word2Vev的讲解

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

3 进阶篇——Word2Vev的讲解

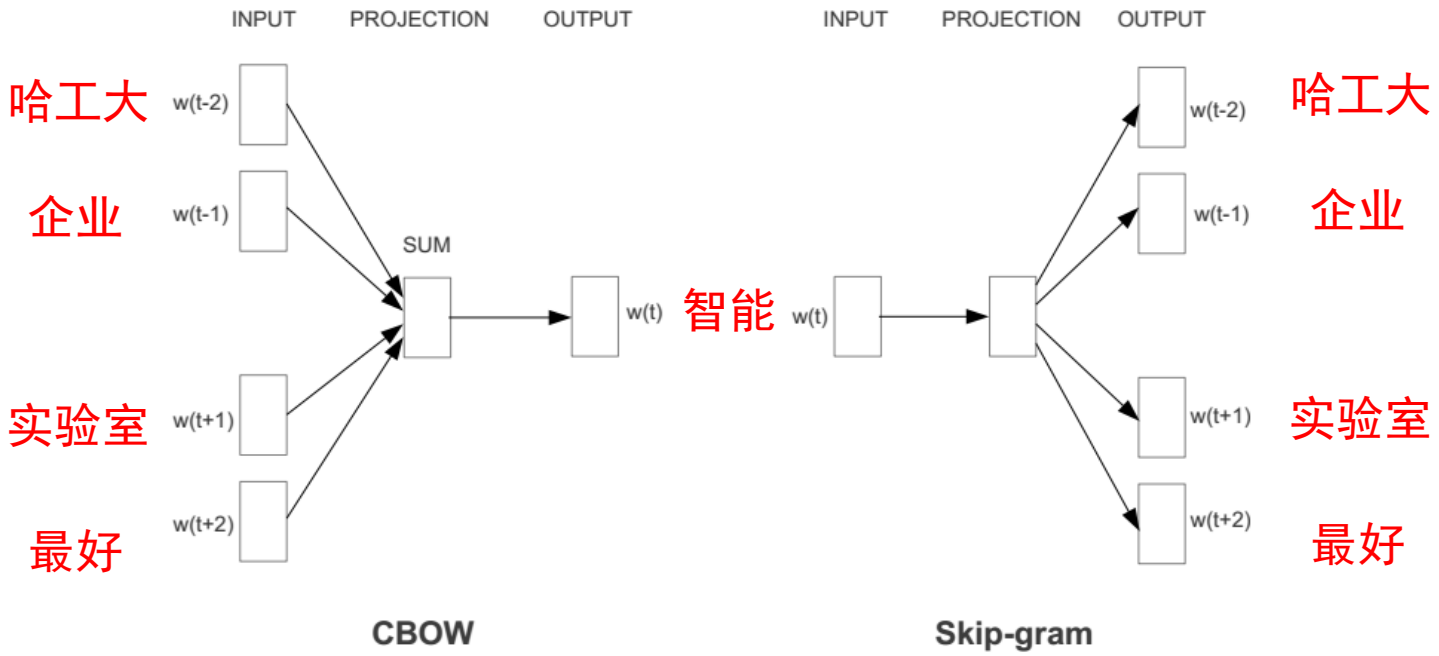


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

$$\mathcal{L} = \sum_{w \in C} \log p(w|context(w))$$

$$\mathcal{L} = \sum_{w \in C} \log p(context(w)|w)$$

3 进阶篇——Word2Vev的讲解

为了实现对函数的构造，作者设计了两个框架：

- Hierarchical Softmax
- Negative Sampling

3 进阶篇——Word2Vev的讲解

- 基于Negative Sampling的模型

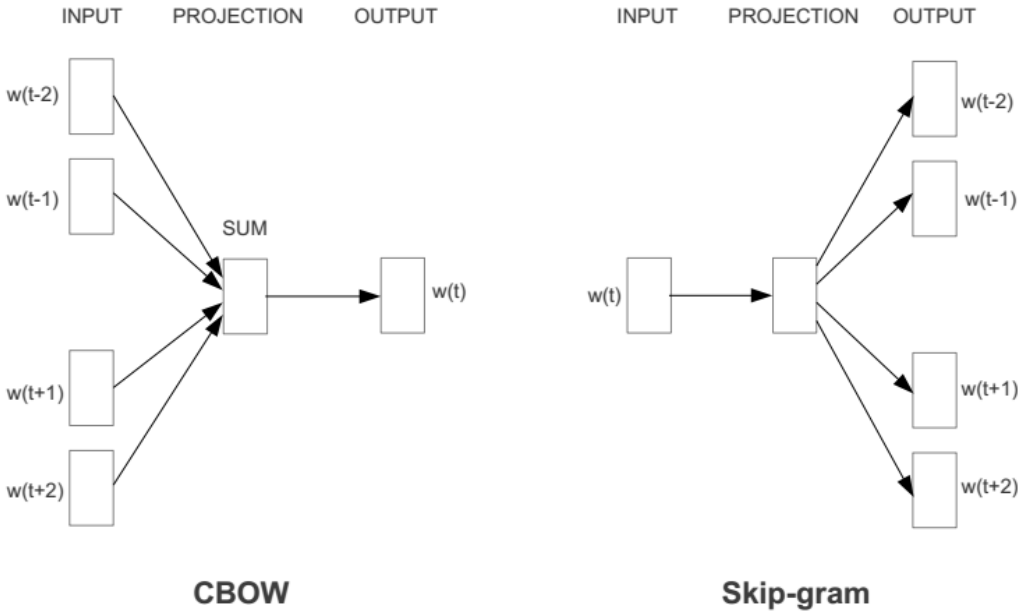


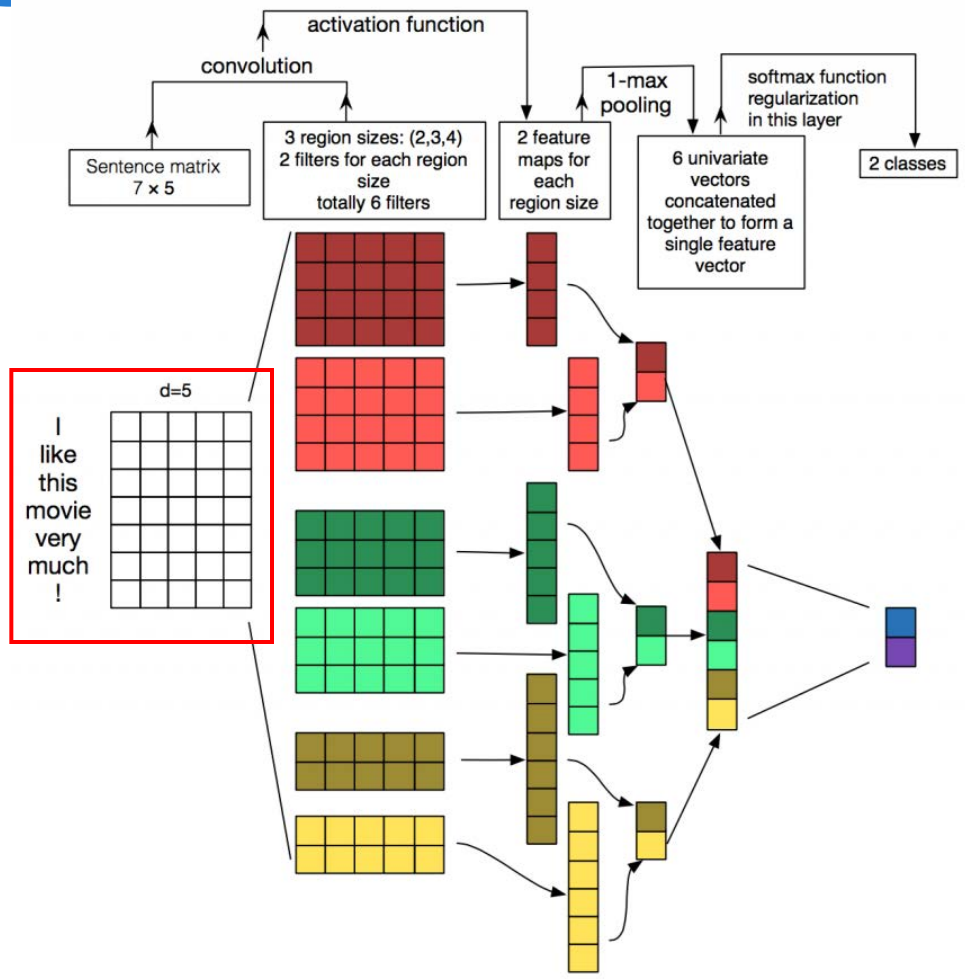
Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

$$\mathcal{L} = \sum_{w \in C} \log p(w|context(w))$$

$$\mathcal{L} = \sum_{w \in C} \log p(context(w)|w)$$

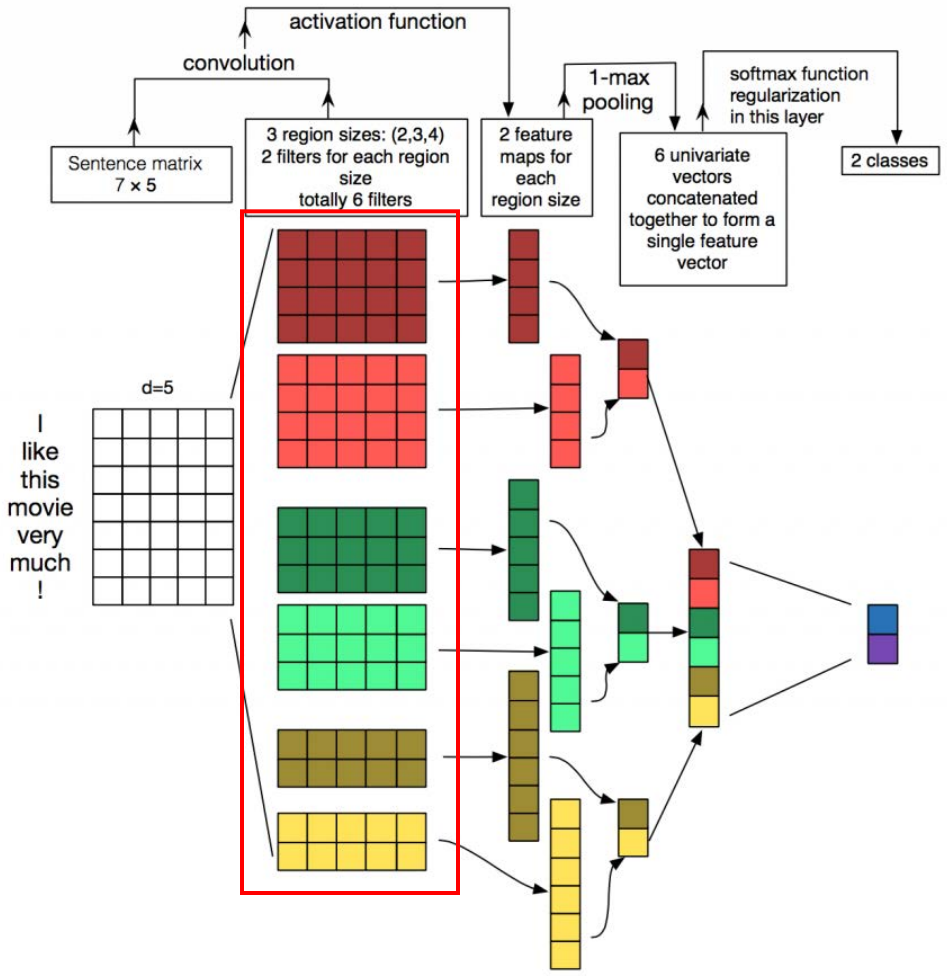
词向量的应用

4 应用篇——词向量的应用



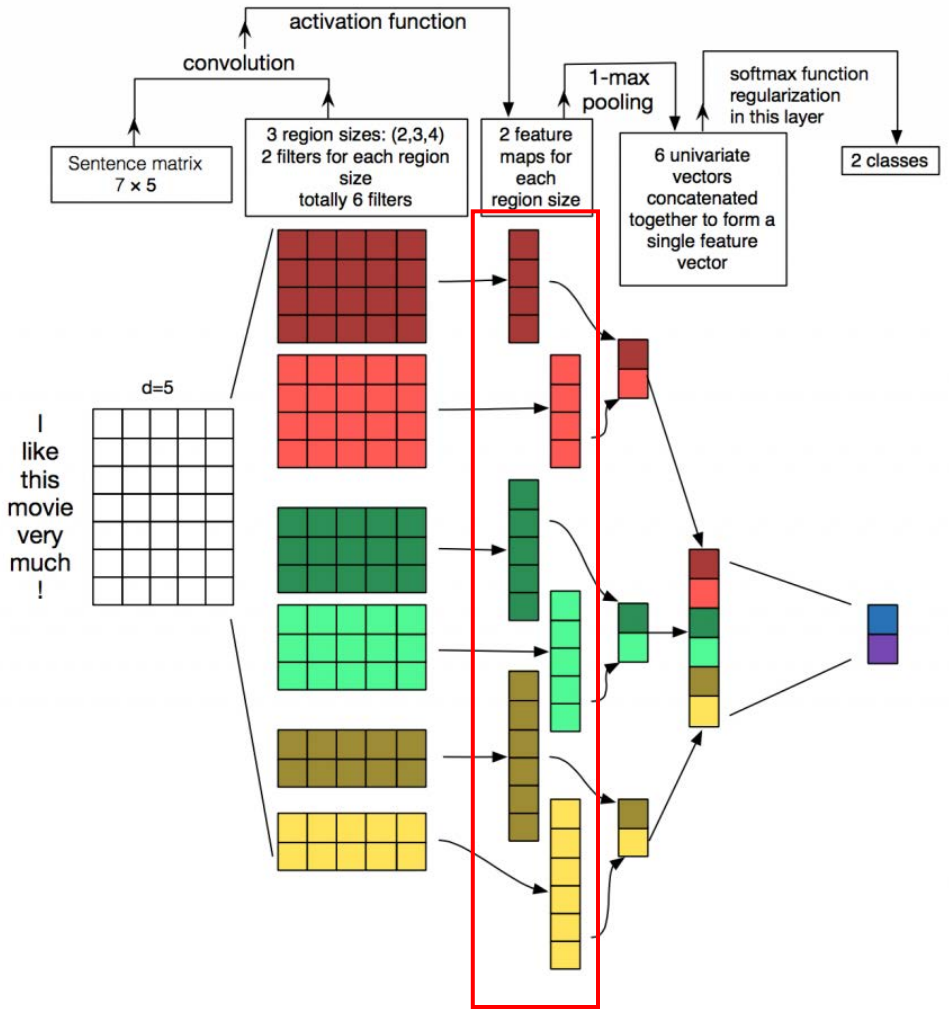
- 第一列：把每个单词的word embedding拼接起来组成的矩阵，我们可以把它当做图像处理中的一幅图像。

4 应用篇——词向量的应用



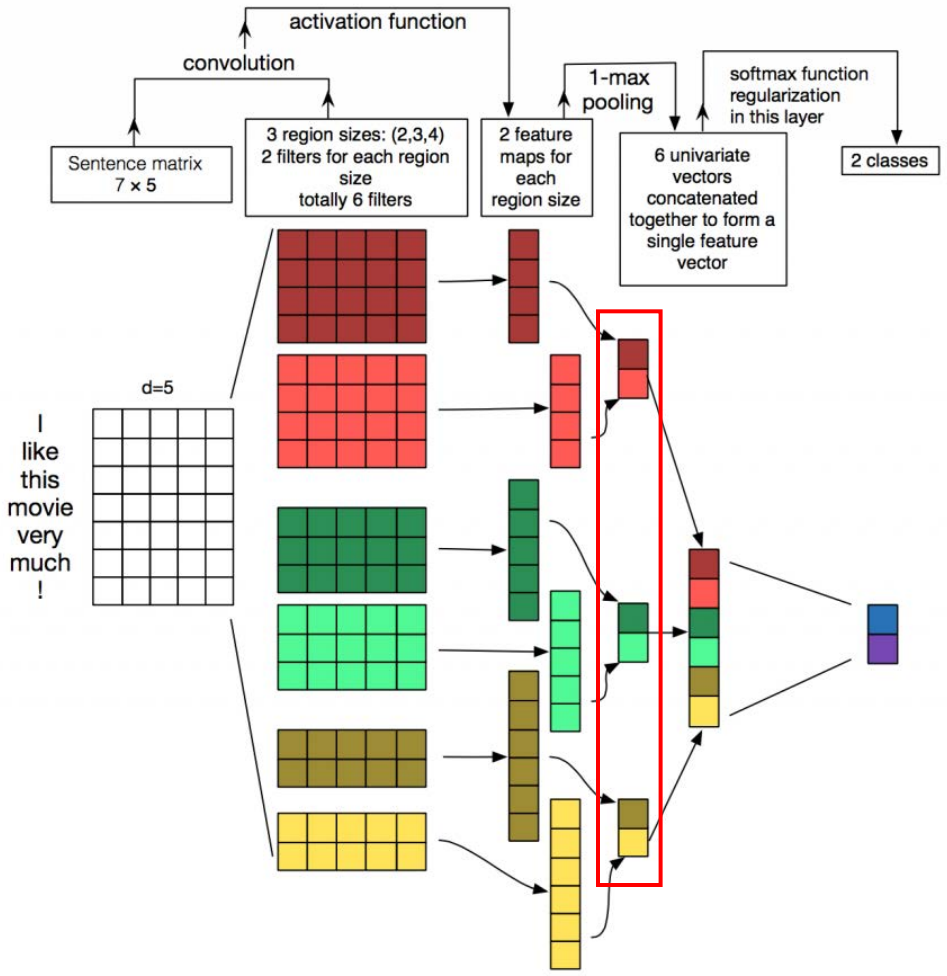
- 第一列：把每个单词的word embedding拼接起来组成的矩阵，我们可以把它当做图像处理中的一幅图像。
- 第二列：这里我们对滤波器设置了三种尺寸：2、3和4行，每种尺寸各有两种滤波器。每个滤波器对句子矩阵做卷积运算，得到不同程度的特征。

4 应用篇——词向量的应用



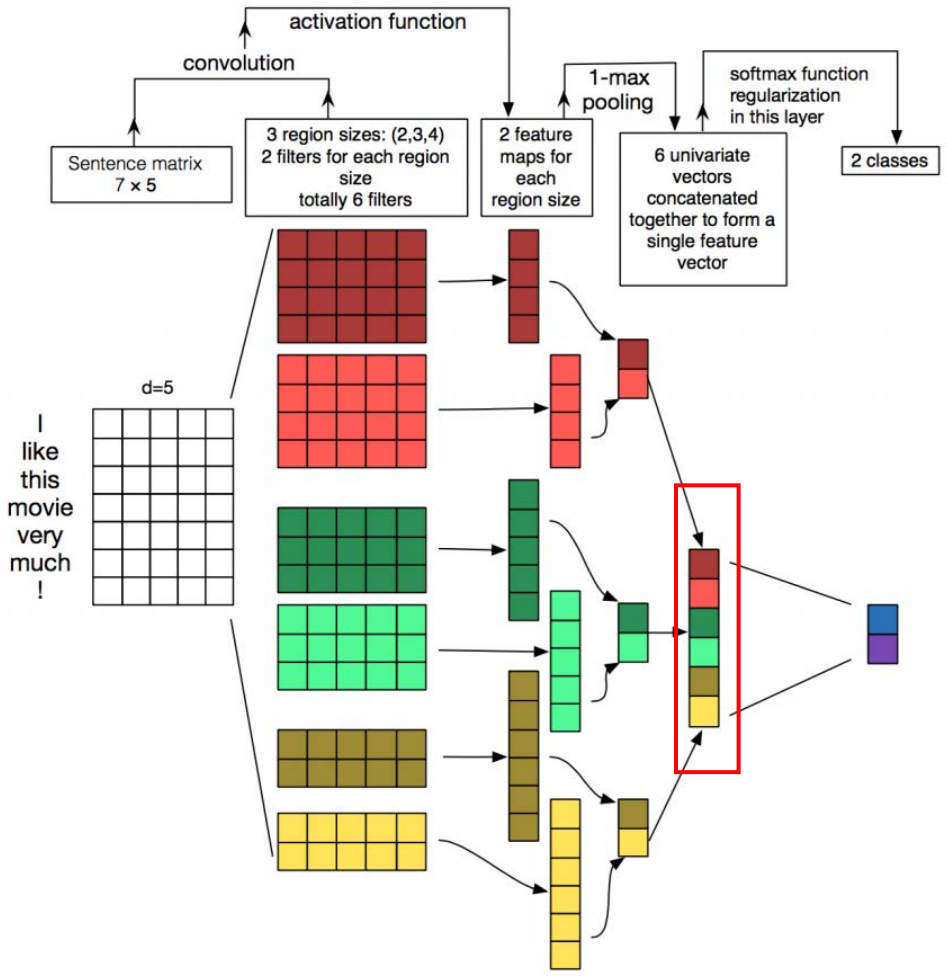
- 第一列：把每个单词的word embedding拼接起来组成的矩阵，我们可以把它当做图像处理中的一幅图像。
- 第二列：这里我们对滤波器设置了三种尺寸：2、3和4行，每种尺寸各有两种滤波器。每个滤波器对句子矩阵做卷积运算，得到不同程度的特征。
- 第三列：使用激励函数将2维的特征装换成1维的特征。

4 应用篇——词向量的应用



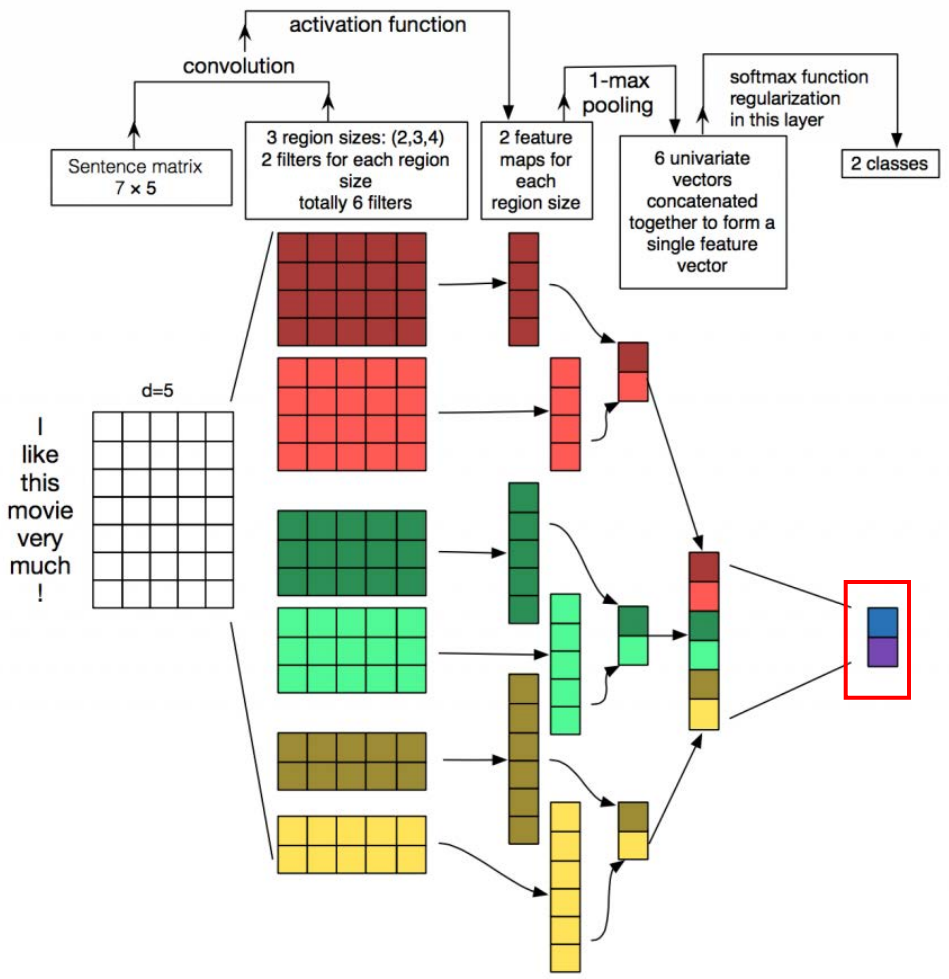
- 第一列：把每个单词的word embedding拼接起来组成的矩阵，我们可以把它当做图像处理中的一幅图像。
- 第二列：这里我们对滤波器设置了三种尺寸：2、3和4行，每种尺寸各有两种滤波器。每个滤波器对句子矩阵做卷积运算，得到不同程度的特征。
- 第三列：使用激励函数将2维的特征装换成1维的特征。
- 第四列：然后对每个特征做最大值池化，也就是只记录每个特征的最大值。

4 应用篇——词向量的应用



- 第一列：把每个单词的word embedding拼接起来组成的矩阵，我们可以把它当做图像处理中的一幅图像。
- 第二列：这里我们对滤波器设置了三种尺寸：2、3和4行，每种尺寸各有两种滤波器。每个滤波器对句子矩阵做卷积运算，得到不同程度的特征。
- 第三列：使用激励函数将2维的特征装换成1维的特征。
- 第四列：然后对每个特征做最大值池化，也就是只记录每个特征的最大值。
- 第五列：我们假设这里是二分类问题，最终的输出层有两个节点。

4 应用篇——词向量的应用


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

谢谢您的聆听