



Stanford NLP中的 深度学习方法研究

Krse Lee
2016/4/3

神经网络&深度学习



深度学习

- + 机器学习的一个分支
- + 复杂结构或由多重非线性变换 ➡ 多个处理层
- + 常见的深度学习框架:
 - + 深度神经网络(DNN)
 - + 卷积神经网络(CNN)
 - + 深度信念网络(DBN)
 - + 递归神经网络(RNN)
- + 应用领域:
 - + 计算机视觉、语音识别、自然语言处理、音频识别、生物信息学等

利用深度学习将照片变为画



Input Photo



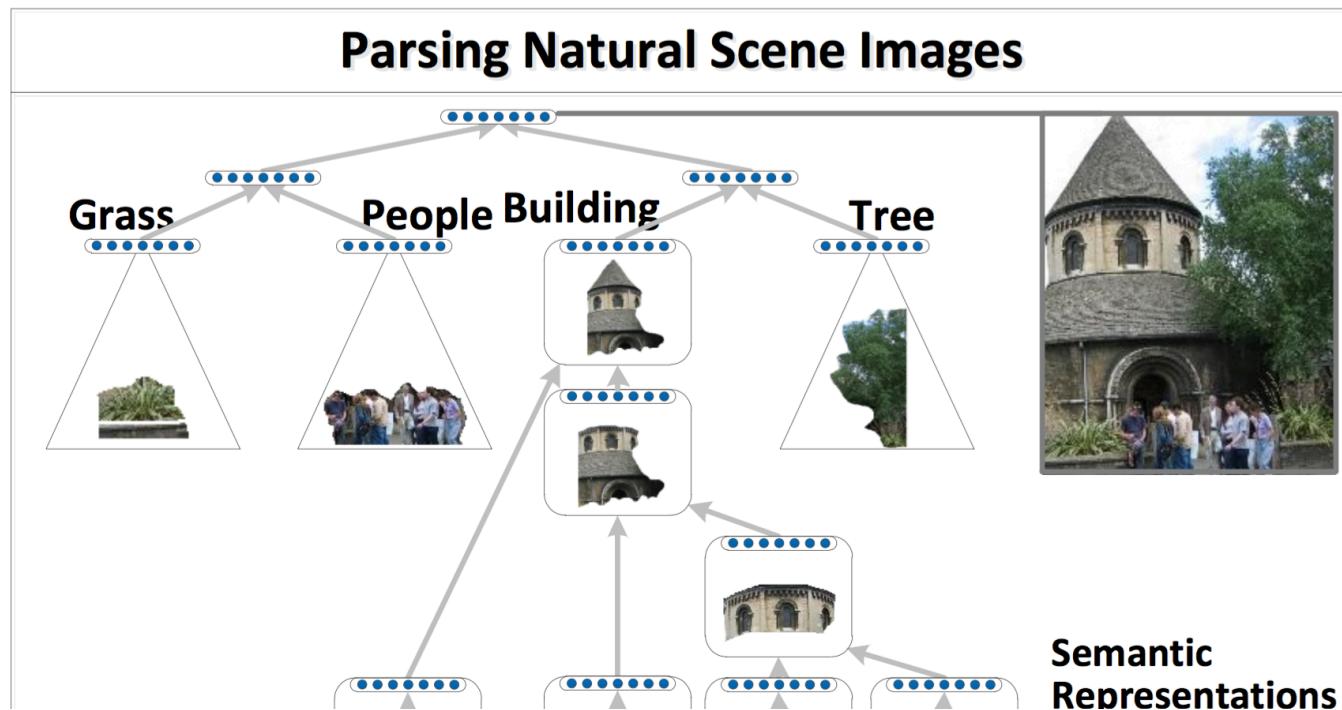
Pencil Drawing



Color Pencil Drawing

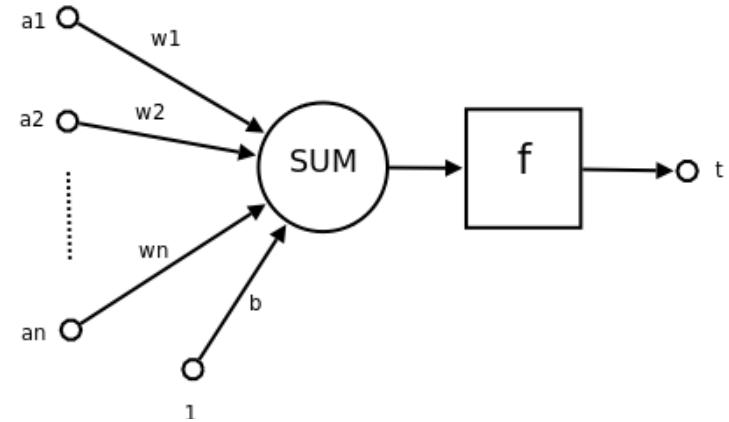
利用深度学习进行图片分解

Same Recursive Neural Network as for natural language parsing!
(Socher et al. ICML 2011)

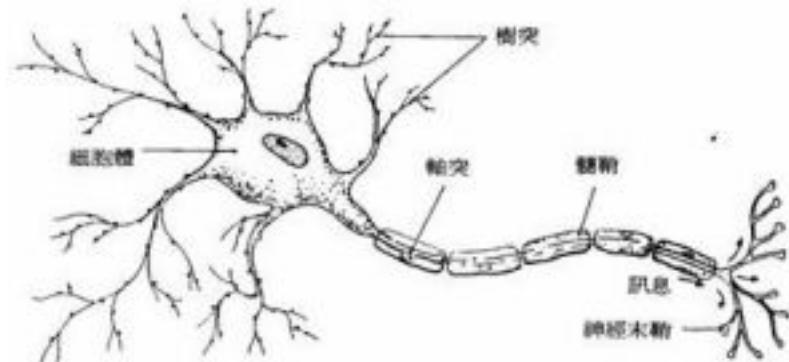


神经元

- + 组成神经网络的基本单位
 - + 输入值a
 - + 权重值w
 - + 偏置b
 - + 激活函数f
 - + 输出t

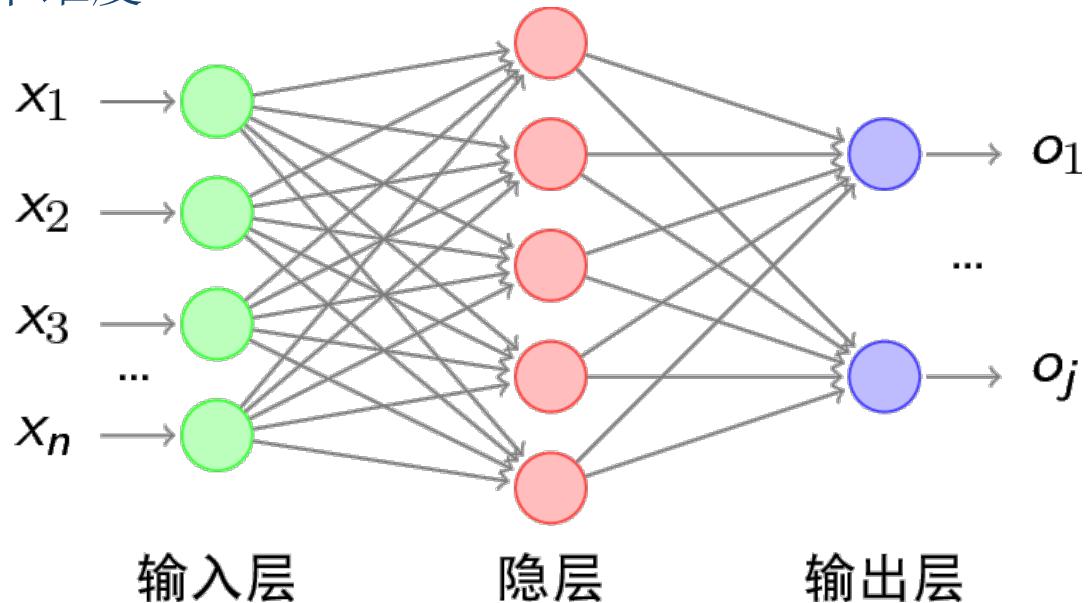


$$t = f(s_j) = f\left(\sum_{i=1}^n a_i w_{ij} - b_j\right)$$

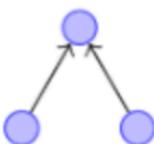
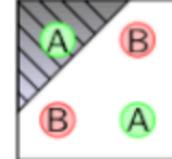
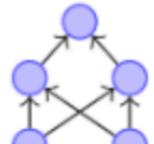
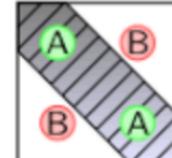
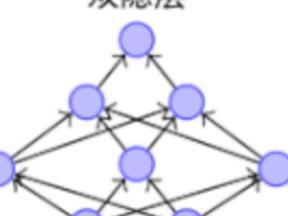
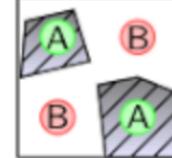


神经网络结构

- + 最简单的神经网络有3层
 - + 输入层（跟输入样本维度有关）
 - + 隐藏层
 - + 输出层



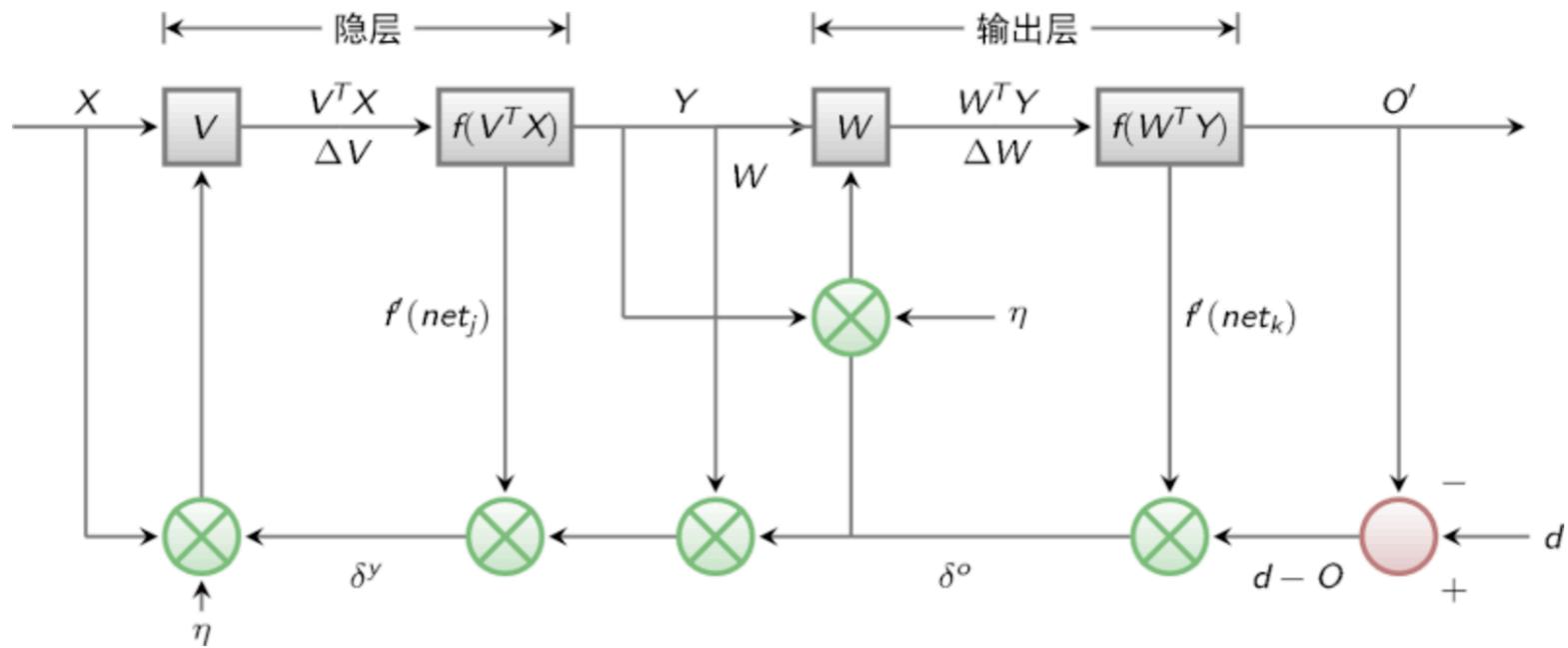
感知机的分类能力

结构	决策区域类型	区域形状	异或问题
无隐层 	由一超平面分成两个		
单隐层 	开凸区域或闭凸区域		
双隐层 	任意形状（其复杂度由单元数目确定）		

误差反向传播训练（BP）算法

- + BP算法的基本思想是,学习过程由信号的正向传播与误差的反向传播两个过程组成。
- + 正向传播时,输入样本从输入层传入,经各隐层逐层处理后,传向输出层。若输出层的实际输出与期望的输出(教师信号)不符,则转入误差的反向传播阶段。
- + 反向传播时, 将输出以某种形式通过隐层向输入层逐层反传,并将误差分摊给各层的所有单元,从而获得各层单元的误差信号,此误差信号即作为修正各单元权值的依据。
- + 采用梯度下降方法, 需要设定梯度下降速率

BP算法的信号流图



其他常用的神经网络训练方法

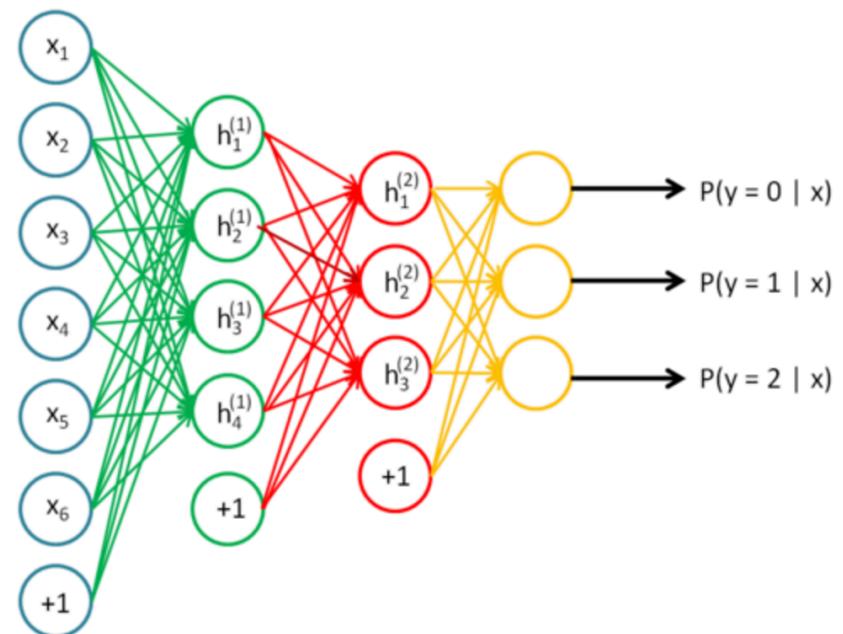
- + 曼哈顿更新规则
 - + 梯度下降值只用来决定更新权重矩阵的方向，需要设置学习速率
- + 快速训练（QPROP）
 - + 基于牛顿方法，需要设置学习速率
- + 弹性传播训练（RPROP）
 - + 与曼哈顿更新规则类似，但是梯度下降速率会改变，各个神经元权重更新速率也不一样，不需要设定参数
- + 量化共轭梯度（SCG）
 - + 基于共轭梯度方法的最优化函数，不需要人工设置参数
- + 莱文贝格—马夸特算法（LMA）
 - + 结合牛顿方法和梯度下降

深度学习网络



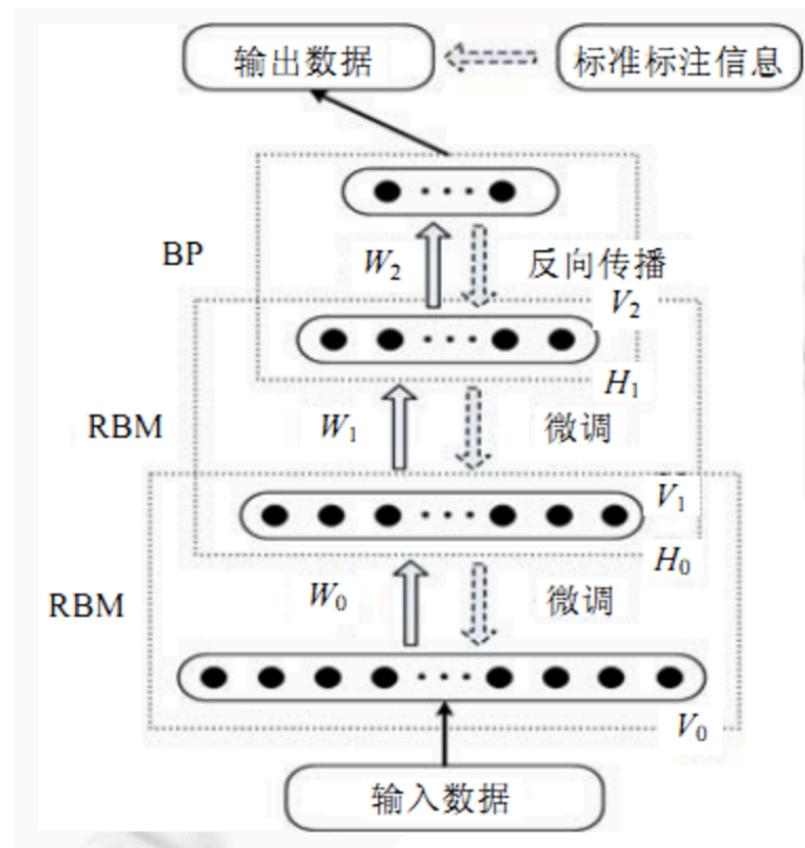
深度神经网络 (DNN)

- + 级联神经网络
- + 上一层的输出（学习到的特征）作为下一层的输入
- + 一般采用BP方法训练



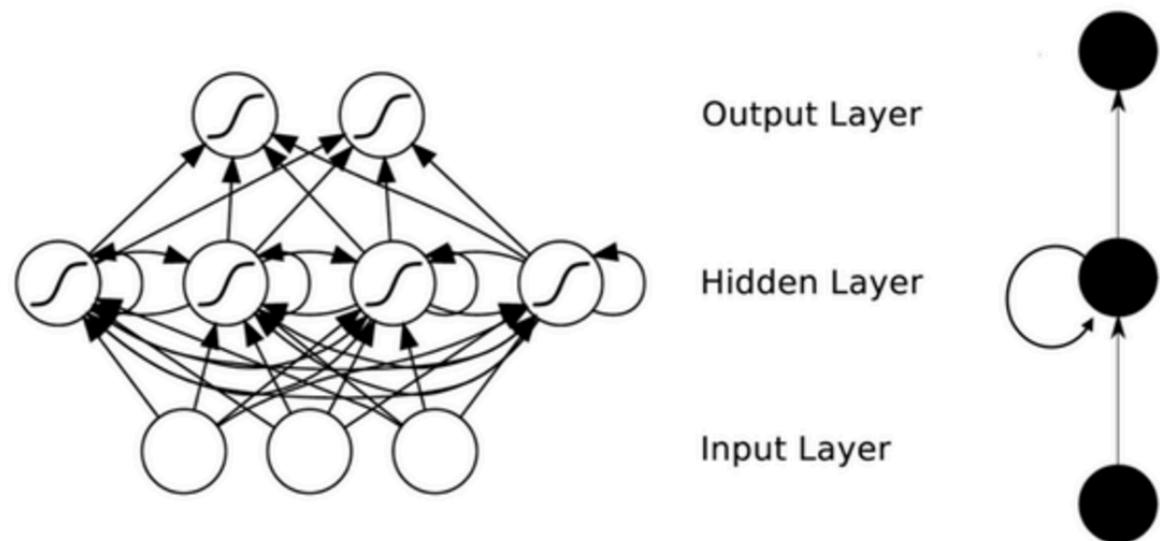
深度信念网络 (DBN)

- + 基于受限玻耳兹曼机(RBM)
- + 若干层RBM加上一层BP
- + 训练过程:
 - + 分别训练每层RBM (无监督)
 - + BP训练 (有监督)
- + 优点:
 - + 克服BP网络因随机初始化权值参数而容易陷入局部最优和训练时间长的缺点。



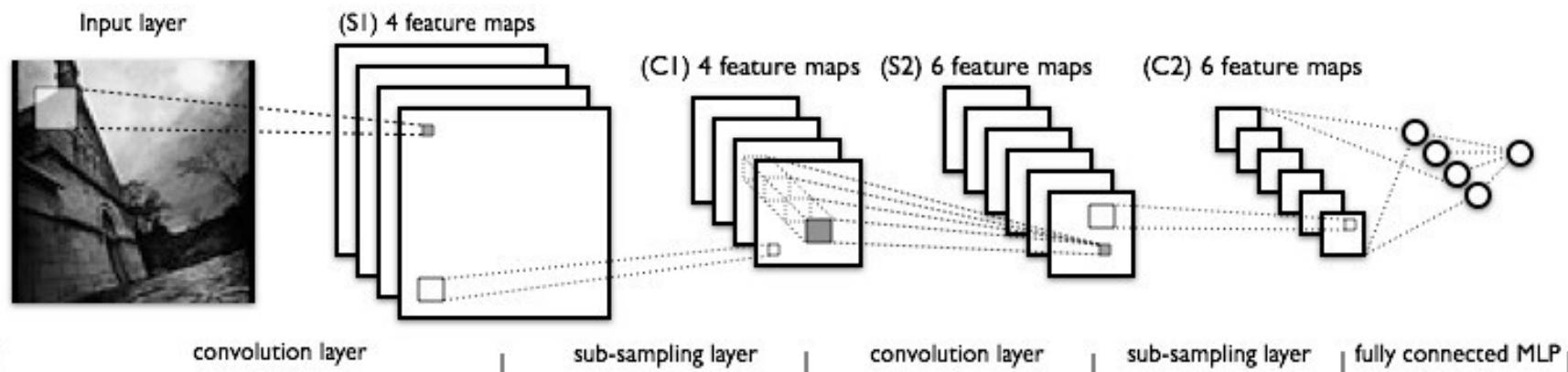
递归神经网络 (RNN)

- + 对时间序列上的变化进行建模
- + 每层神经元的输出可以在下一个时间戳作用到自身
- + 运用在NLP、语音识别、手写体识别



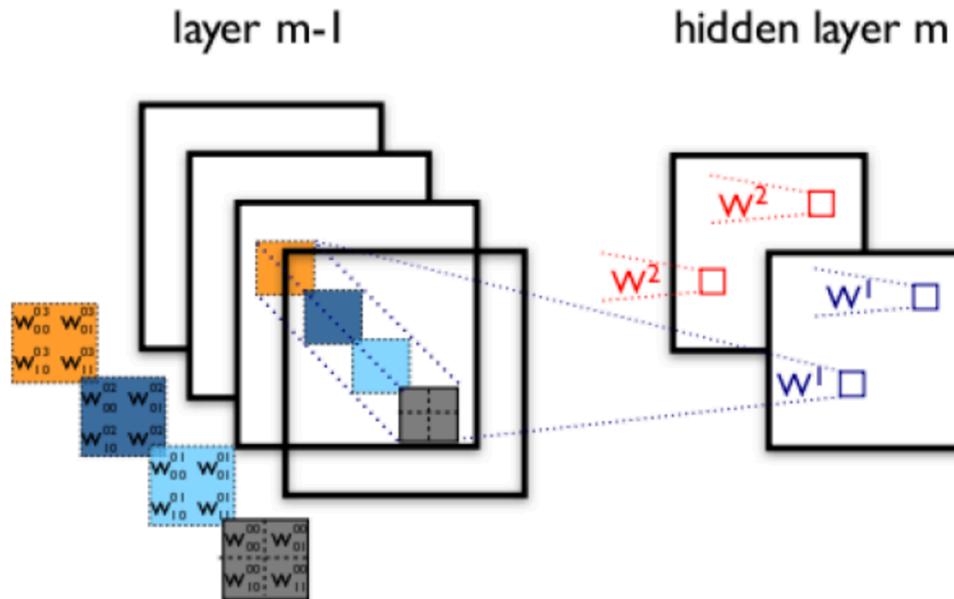
卷积神经网络 (CNN)

- + 并非所有上下层神经元都直接相连，以“卷积核”作为中介
- + 图像处理：减少神经元数目，加快训练速度
- + 同一个卷积核在所有图像内是共享的，卷积操作后仍保留原先的位置关系



卷积过程

- + 卷积：提取图像特征
- + 每种卷积核会产生一副新的图像
- + 四通道（ARGB），双卷积核

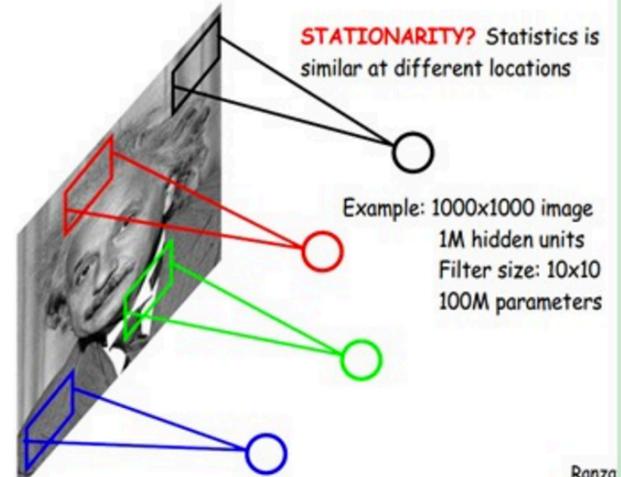


1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Image

Convolved Feature

LOCALLY CONNECTED NEURAL NET

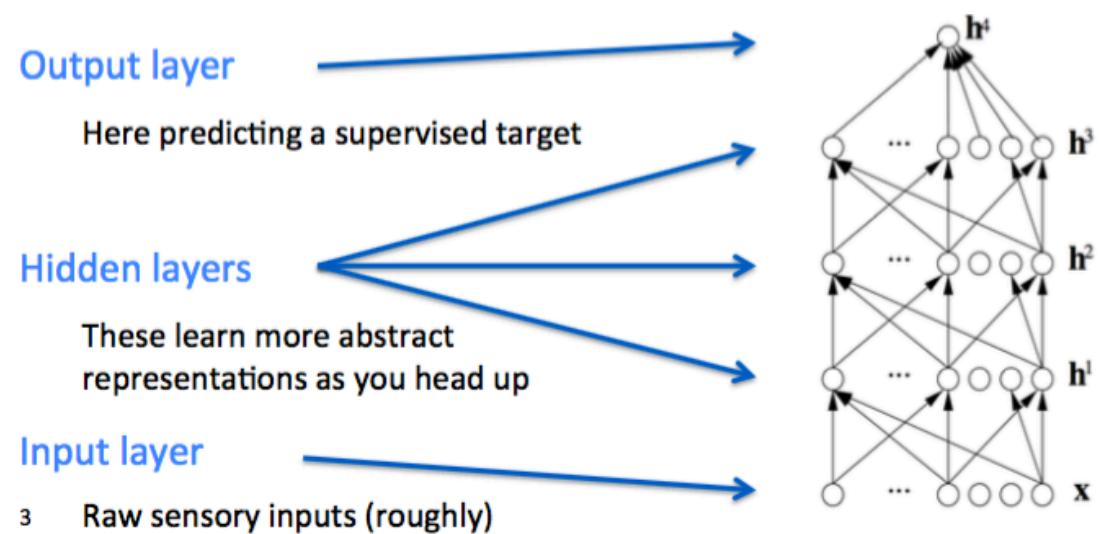


Stanford NLP中的 深度学习应用



深度学习网络架构

- + 使用基于受限玻耳兹曼机的深度信念网络
- + 多层次马尔可夫随机域
- + 包括多个隐含层，多种类型的神经网络



从最大化分类器到神经网络

神经网络 = 同时运行多个Logistic回归

+ NLP中的最大化分类器

$$P(c|d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c' \in C} \exp \sum_i \lambda_i f_i(c', d)}$$

+ 通过监督学习，得到输入向量形式

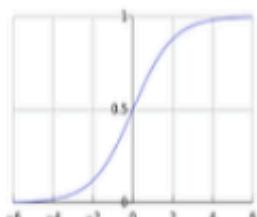
$$P(c|d, \lambda) = \frac{e^{\lambda^\top f(c, d)}}{\sum_{c'} e^{\lambda^\top f(c', d)}}$$

Make two class:

$$P(c_1|d, \lambda) = \frac{e^{\lambda^\top f(c_1, d)}}{e^{\lambda^\top f(c_1, d)} + e^{\lambda^\top f(c_2, d)}} = \frac{e^{\lambda^\top f(c_1, d)}}{e^{\lambda^\top f(c_1, d)} + e^{\lambda^\top f(c_2, d)}} \cdot \frac{e^{-\lambda^\top f(c_1, d)}}{e^{-\lambda^\top f(c_1, d)}}$$

$$= \frac{1}{1 + e^{\lambda^\top [f(c_2, d) - f(c_1, d)]}} = \frac{1}{1 + e^{-\lambda^\top x}} \quad \text{for } x = f(c_1, d) - f(c_2, d)$$

$$= f(\lambda^\top x)$$

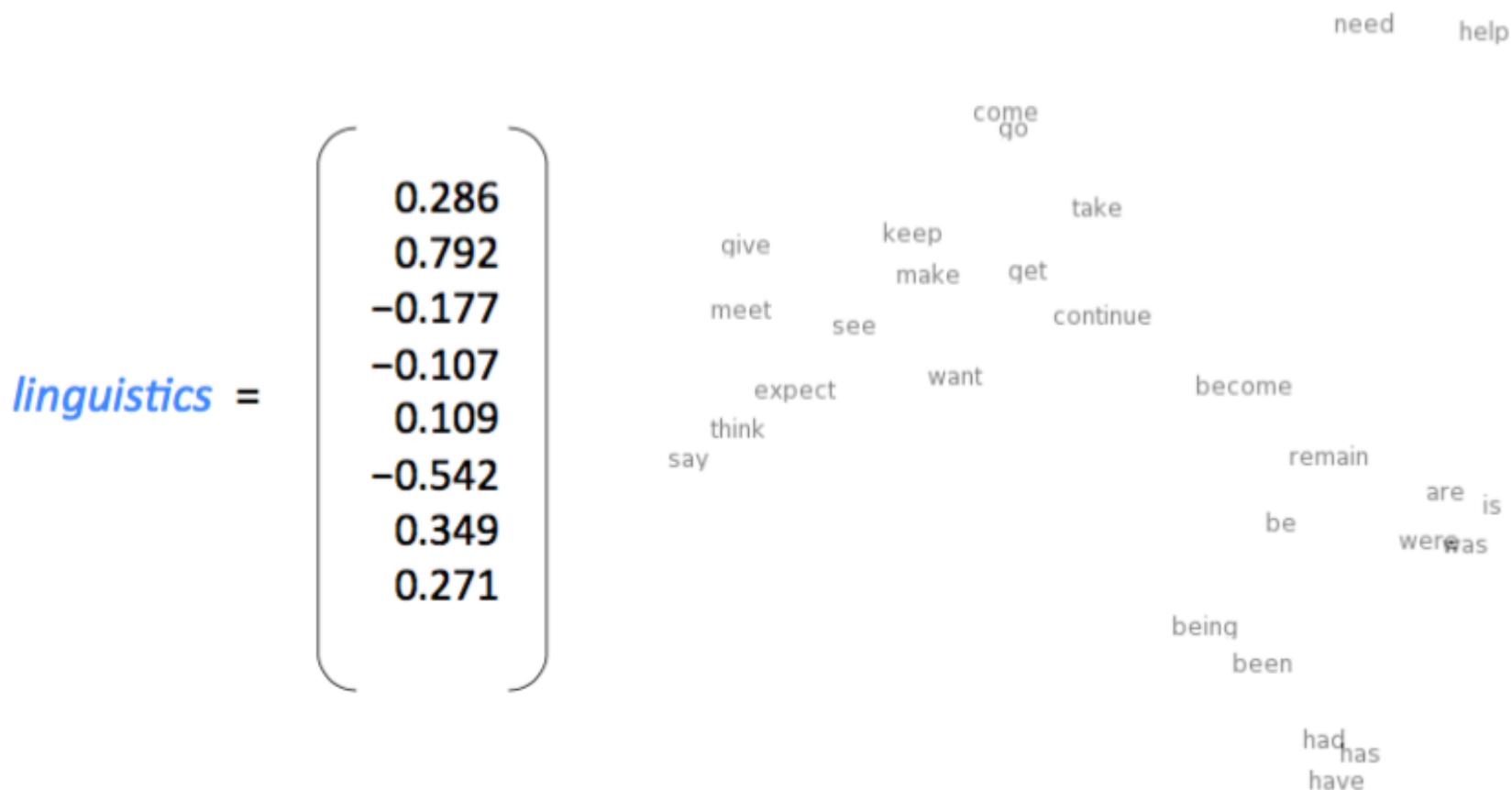


for $f(z) = 1/(1 + \exp(-z))$, the logistic function – a sigmoid non-linearity.

motel [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

单词向量 与 语法解析

单词向量映射(Word Vector Map)



训练单词向量

- + 随机初始化单词向量嵌套矩阵，单词向量 $x=Le$, e 是一个one-hot向量
- + $\text{score}(\text{ cat chills on a mat }) = L =$
- + 连接单词向量

$$\begin{bmatrix} \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \text{the} & \text{cat} & \text{mat} & \cdots & \vdots & \vdots \end{bmatrix}_n^{|V|}$$



Then concatenate them to $5n$ vector:

$$x = [\quad \bullet\bullet\bullet\bullet \quad \bullet\bullet\bullet\bullet \quad \bullet\bullet\bullet\bullet \quad \bullet\bullet\bullet\bullet \quad \bullet\bullet\bullet\bullet \quad]$$

训练单词向量

+ 使用一个三层神经网络

$$s = U^T f(Wx + b) \quad x \in \mathbb{R}^{20 \times 1}, W \in \mathbb{R}^{8 \times 20}, U \in \mathbb{R}^{8 \times 1}$$

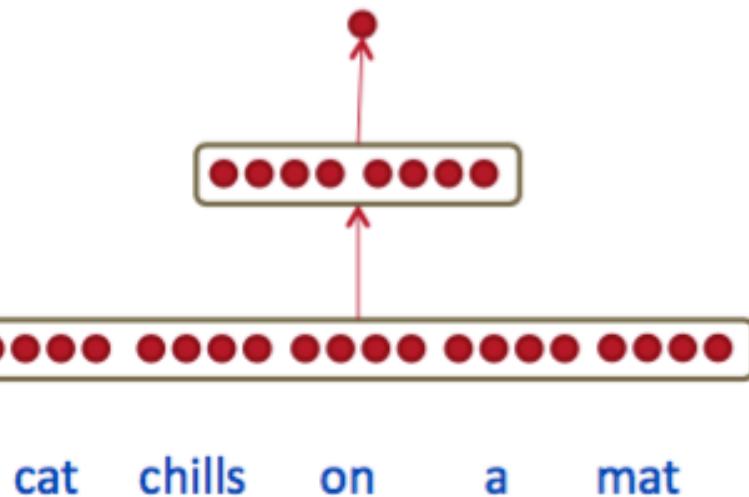
$$s = U^T a$$

$$a = f(z)$$

$$z = Wx + b$$

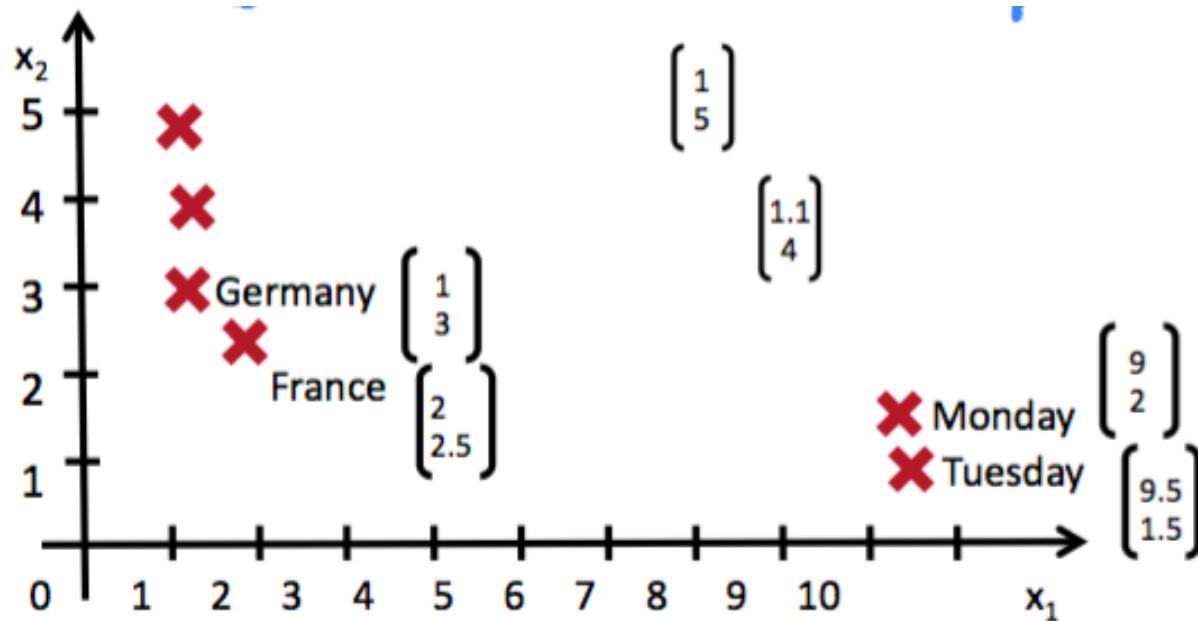
$$x = [x_{cat} \ x_{chills} \ x_{on} \ x_a \ x_{mat}]$$

$$L \in \mathbb{R}^{n \times |V|}$$



构建单词向量空间模型(Word Vector Space Model)

+ 问题：如何表示长句的语义向量？

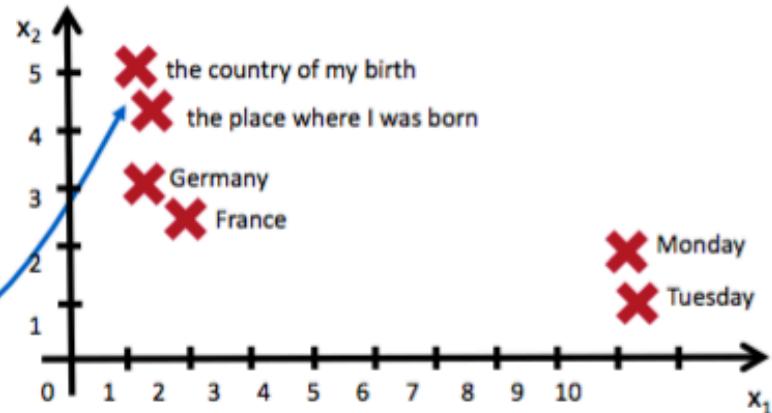
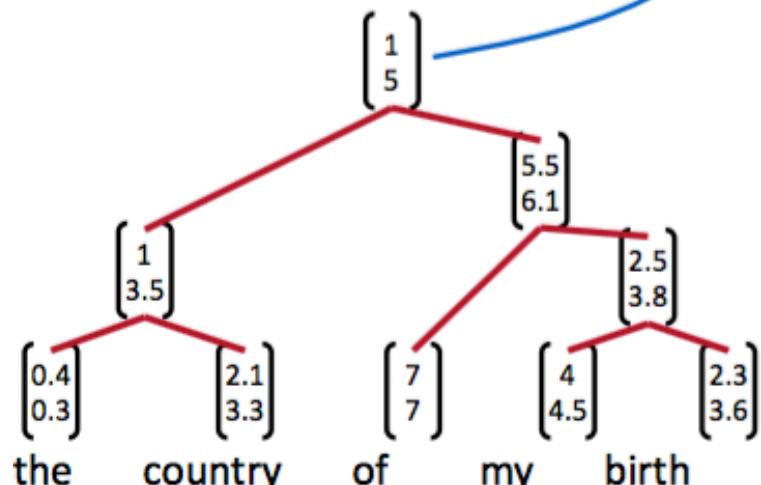


the country of my birth
the place where I was born

构建单词向量空间模型(Word Vector Space Model)

+ 方法：将句子和单词映射到同一个向量空间。

Use principle of compositionality
The meaning (vector) of a sentence is determined by
(1) the meanings of its words and
(2) the rules that combine them.



Models in this section can jointly learn parse trees and compositional vector representations

自然语言的语法树

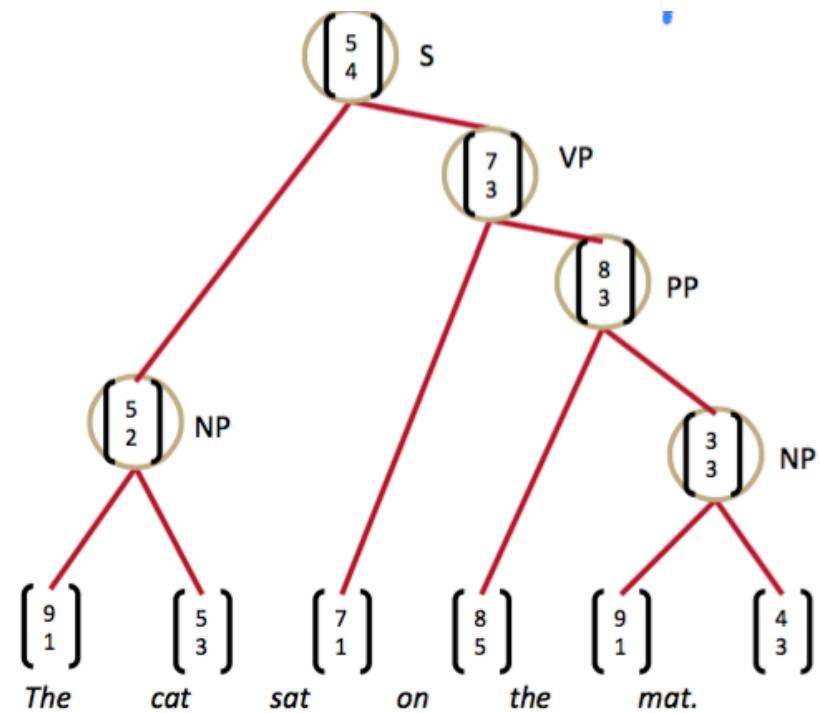
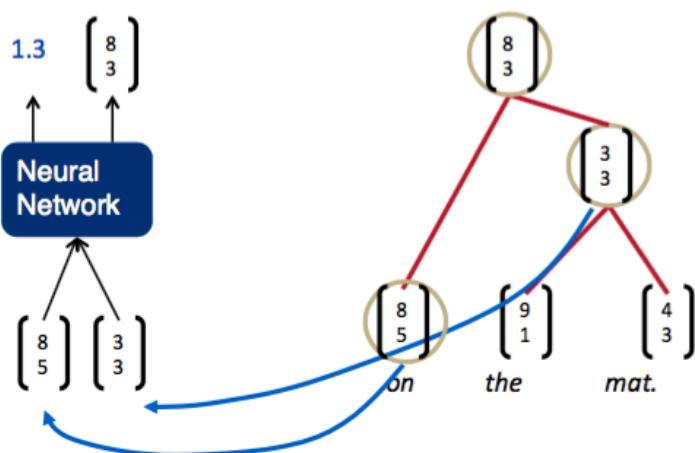
+ 如何建立语法树?

+ 递归神经网络

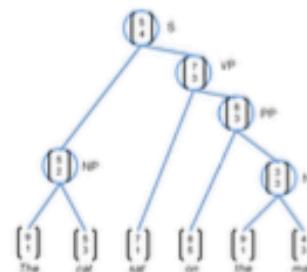
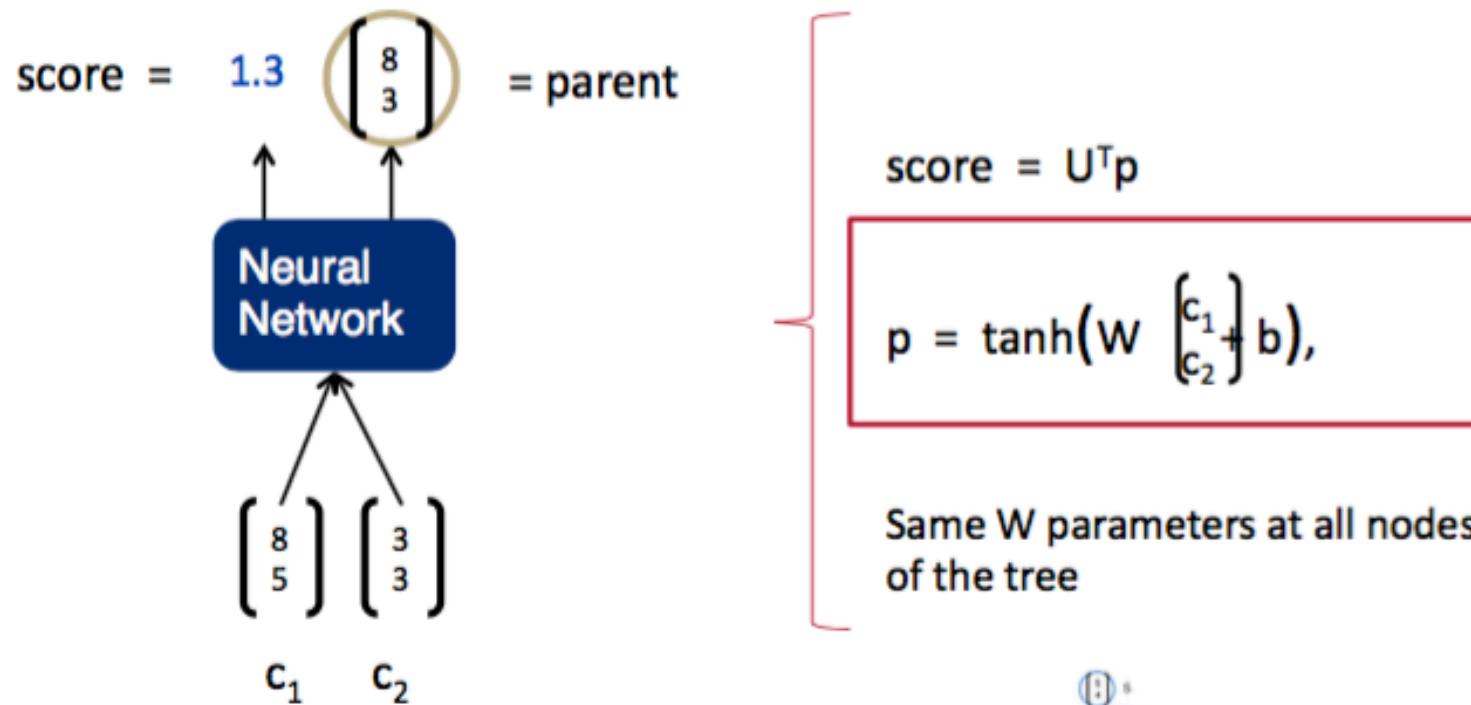
+ 输出:

+ 两个节点合并后的向量值

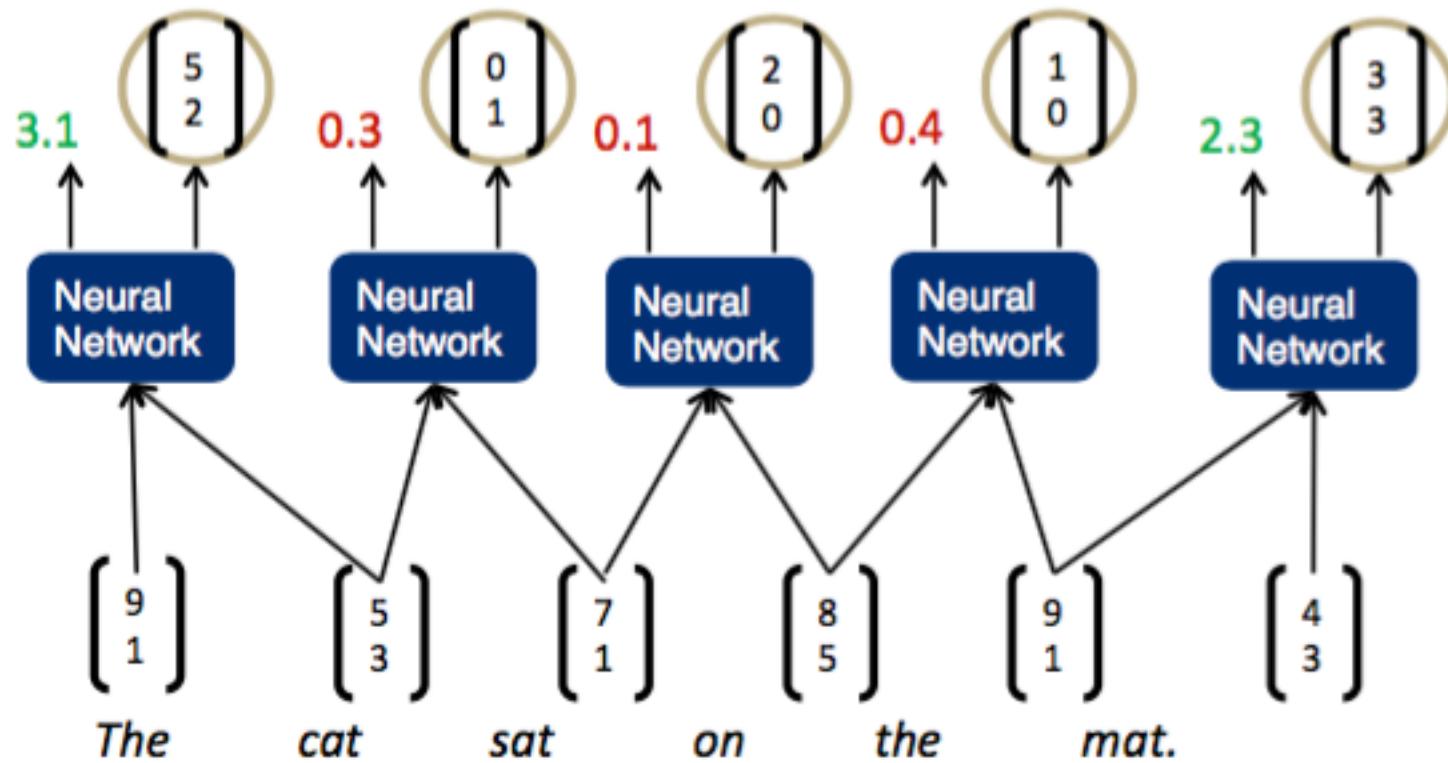
+ 合并得分



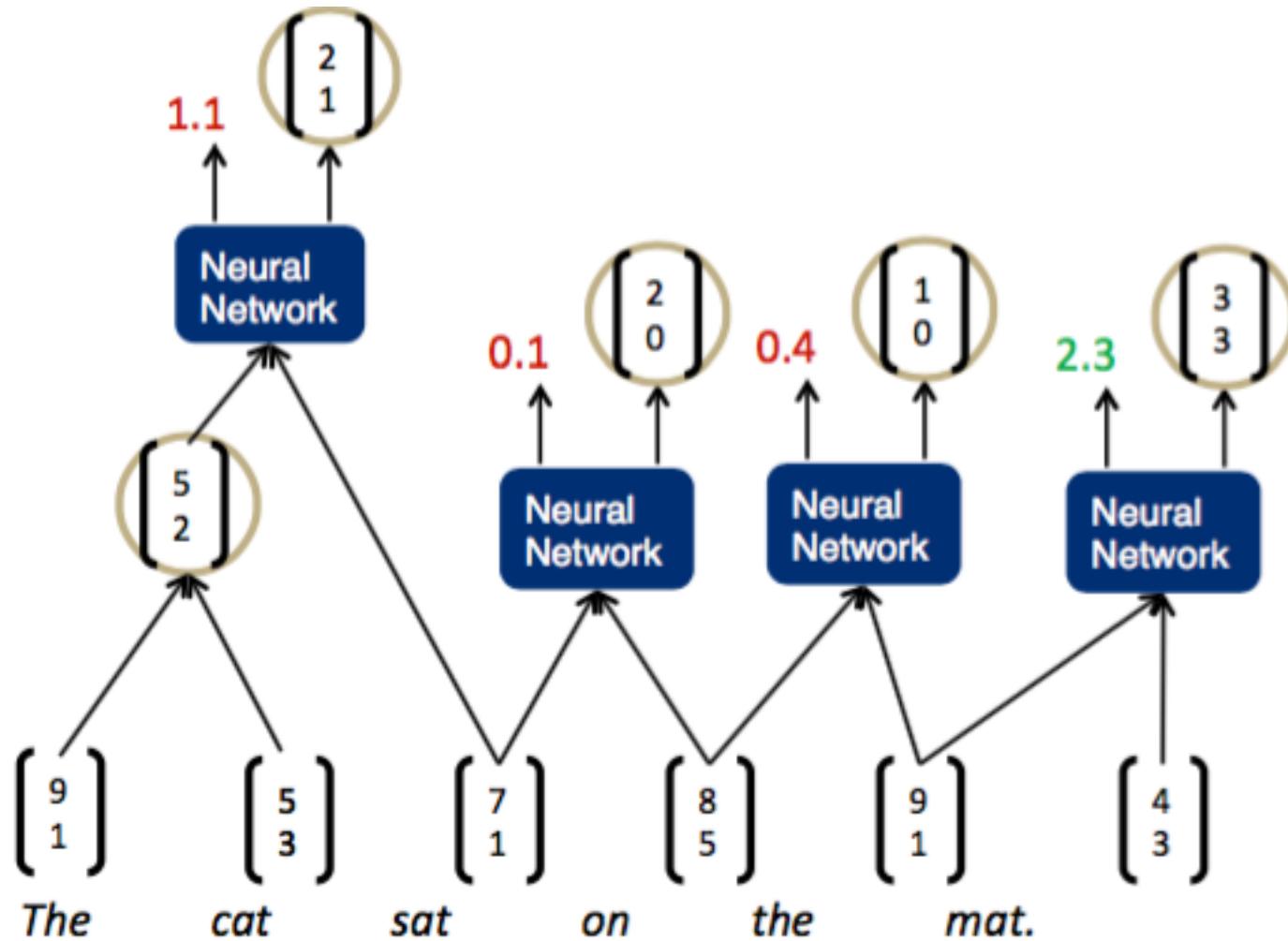
递归神经网络 (Recursive neural Network)



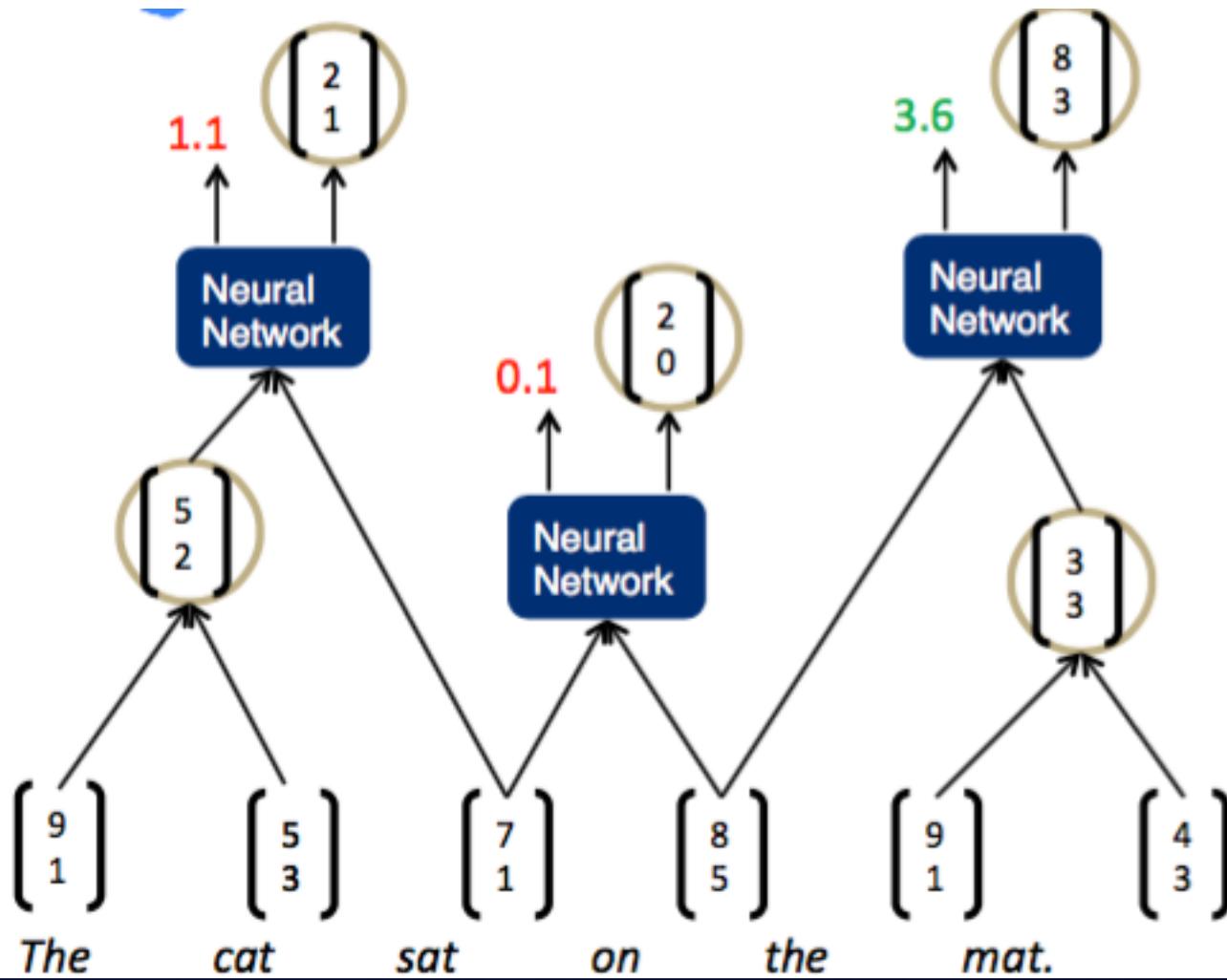
构建语法树过程



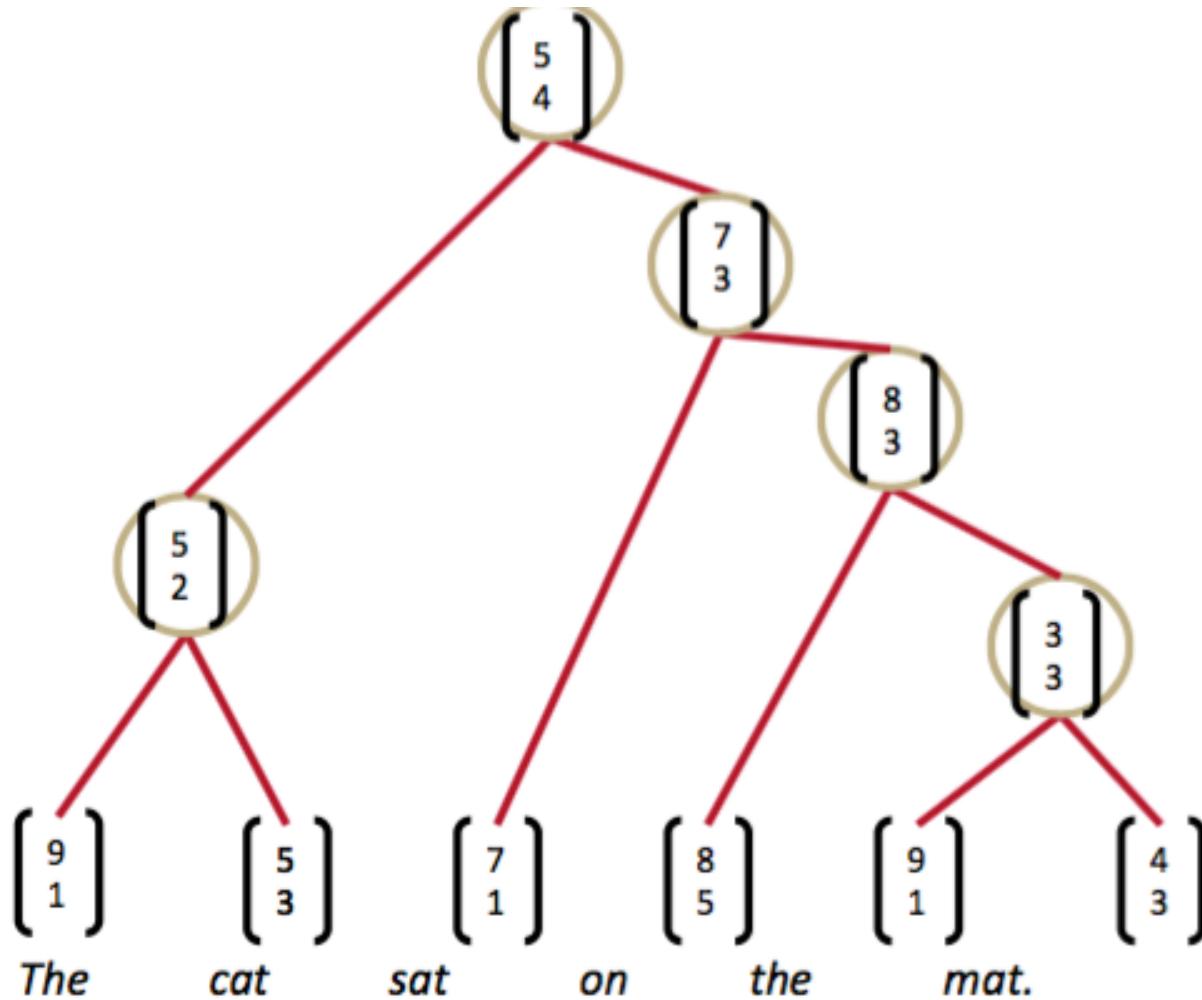
构建语法树过程



构建语法树过程



构建语法树过程

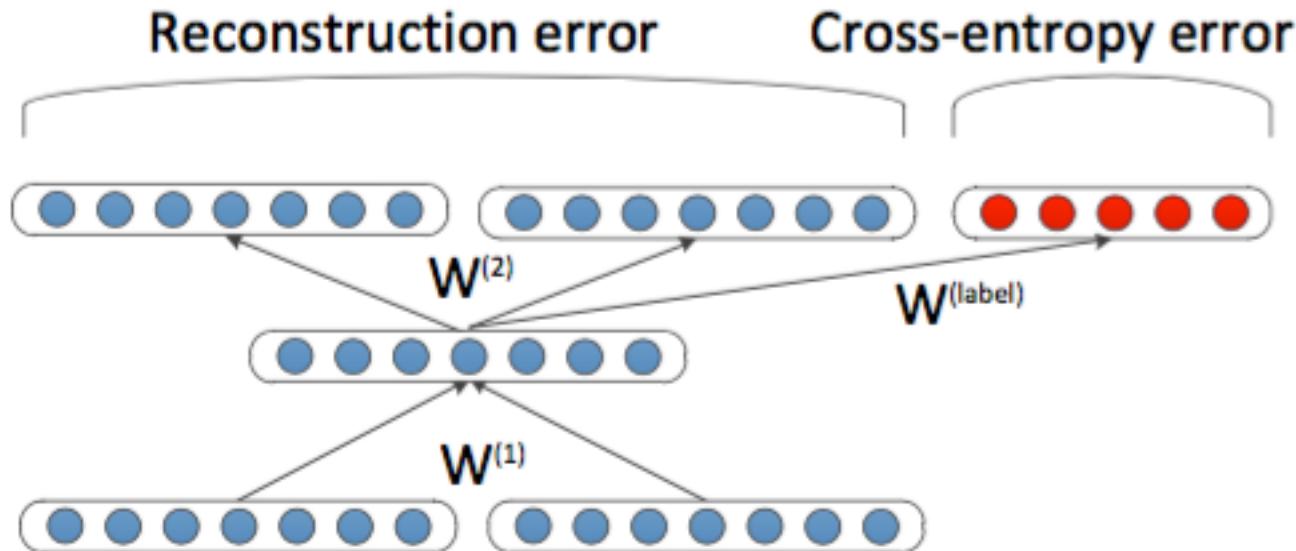


How to compare
the meaning
of two sentences?

递归自动编码、句子语义比较、情感分析

递归自动编码技术

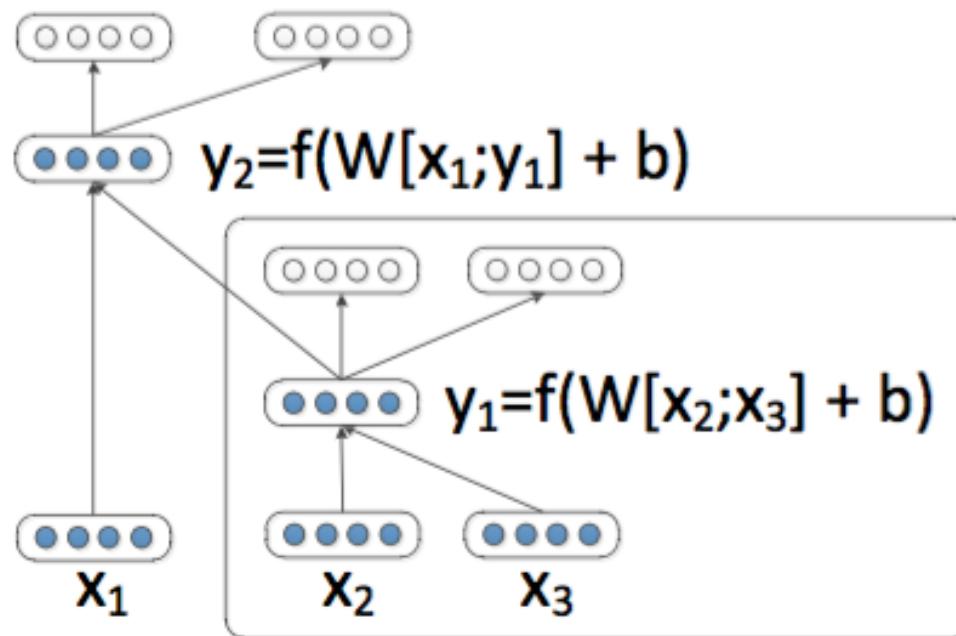
- + 目的：捕获情感，解决反义词问题
- + 误差选取：重构误差+交叉熵



递归自动编码（RAE）

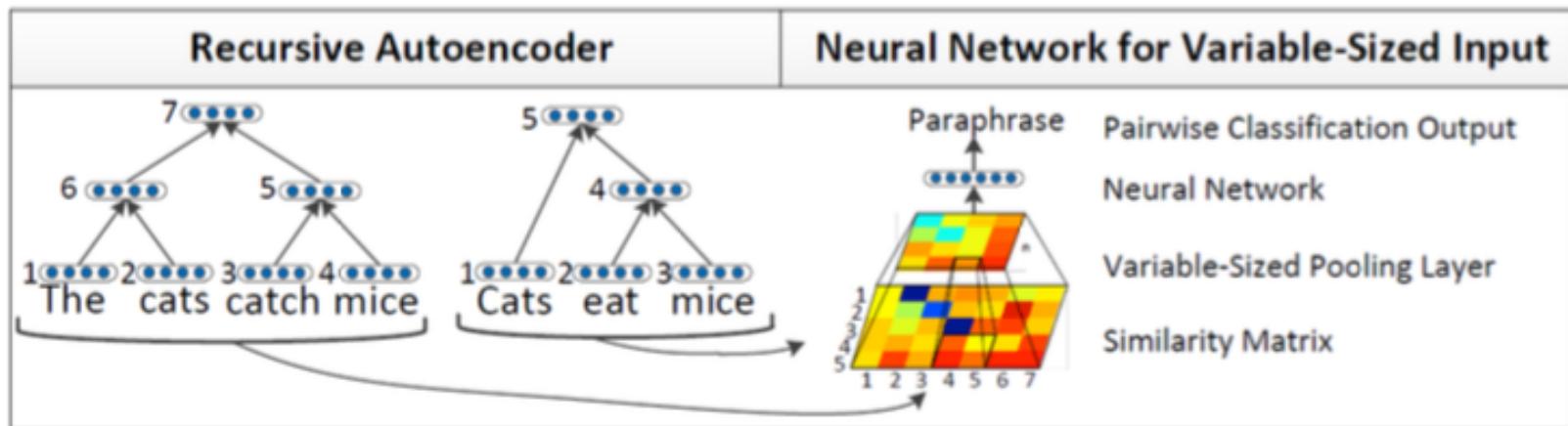
- + 与RNN相似，但是将节点合并得分(Score)换成了重构误差(reconstruction error)

$$E_{rec}([c_1; c_2]) = \frac{1}{2} \left\| [c_1; c_2] - [c'_1; c'_2] \right\|^2$$



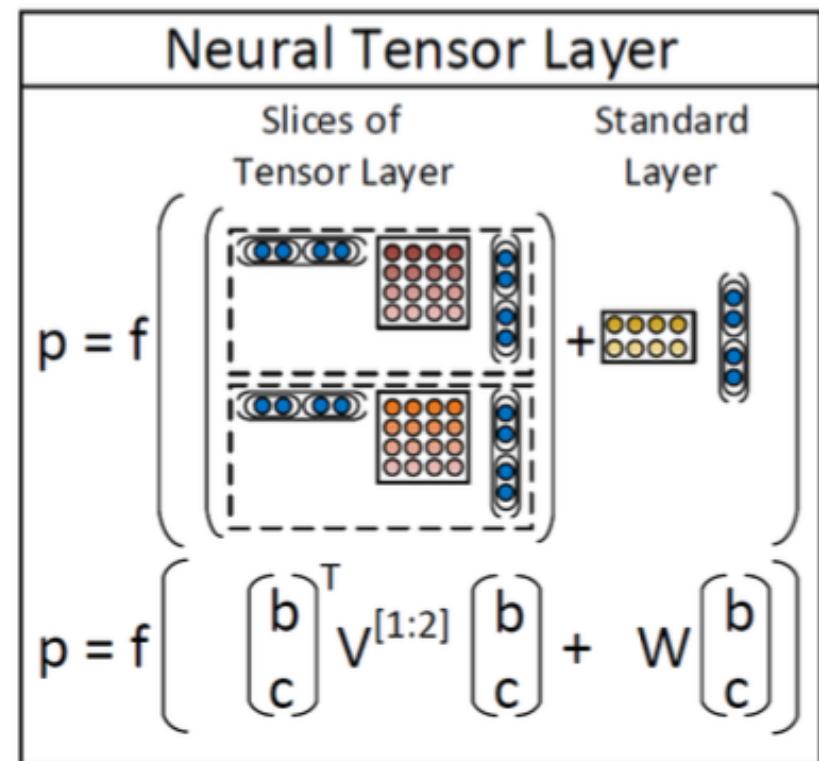
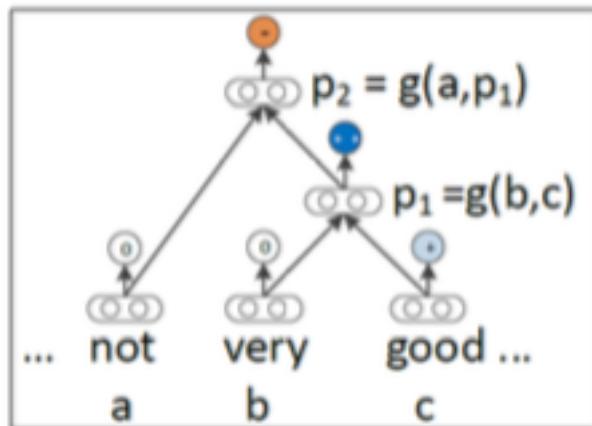
将不同结构的语法树映射到同一空间

- + RAE将语法树的每个节点都进行编码
- + 语法树结构不同，无法直接比较语义



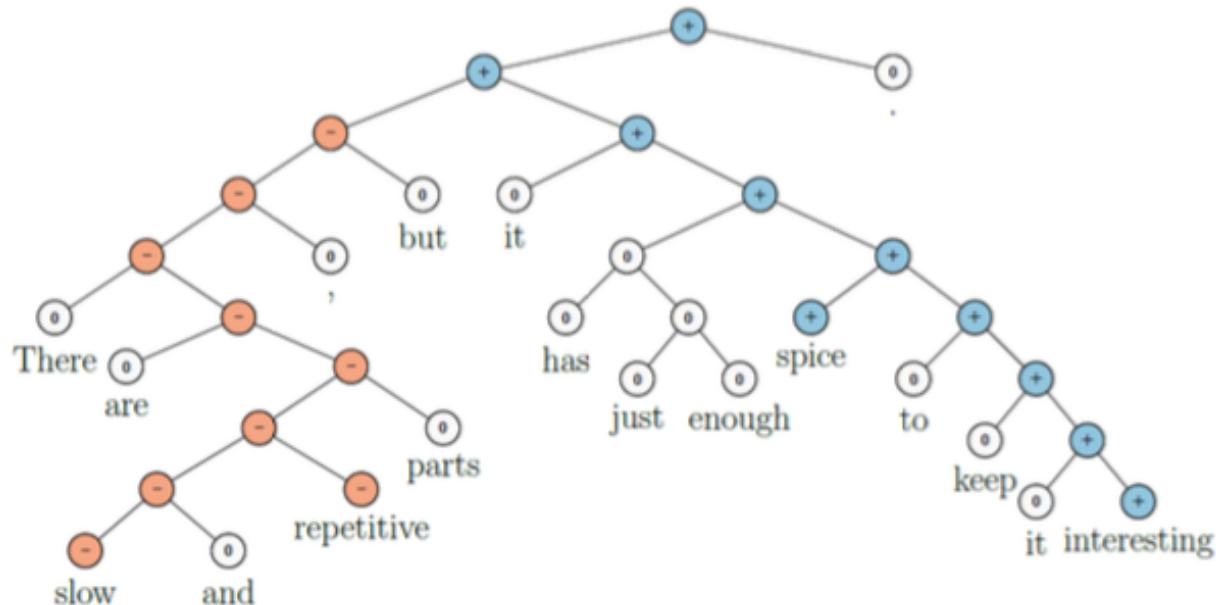
递归神经伸张网络

- + 递归神经网络 + 伸张层
- + 每层的偏置量b变成了伸张层
- + 输入是编码后的语法树



情感分析

- + 从叶子节点开始，递归向上使用同一个复合函数 $g = (x_1, x_2)$ 计算上一个节点的情感向量值，直到根节点
- + 监测到反义关键词时，情感取相反数



RNTN情感分析的准确性

- + 从统计数据上看，基于RNTN方法的情感分析的准确率是目前最高的

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	80.7	45.6	87.6	85.4

