将门网络直播

复旦大學
FUDAN UNIVERSITY

# 面向自然语言处理的分布式表示学习

邱锡鹏
复旦大学

2016年12月21日

http://nlp.fudan.edu.cn/xpqiu

# 语义鸿沟：人工智能的挑战之一

- 底层特征 VS 高层语义
  - 人们对文本、图像的理解无法从字符串或者图像的底层特征直接获得


床前明月光，
疑是地上霜。
举头望明月，
低头思故乡。

# 表示学习

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

- 数据表示是机器学习的核心问题。
  - 特征工程：需要借助人类智能
- 表示学习
  - 如何自动从数据中学习好的表示
- 难点
  - 没有明确的目标

# 什么是好的数据表示？

- 数据分布有很多个不同的潜在因子决定
  - 分布式表示的假设
  - 这些因子在不同任务中共享

- 目标：解构变化背后的潜在因子
  - 尽可能解构更多的因子
  - 尽可能少地丢失信息

- 万变不离其宗
  - 发现多变性中的不变性

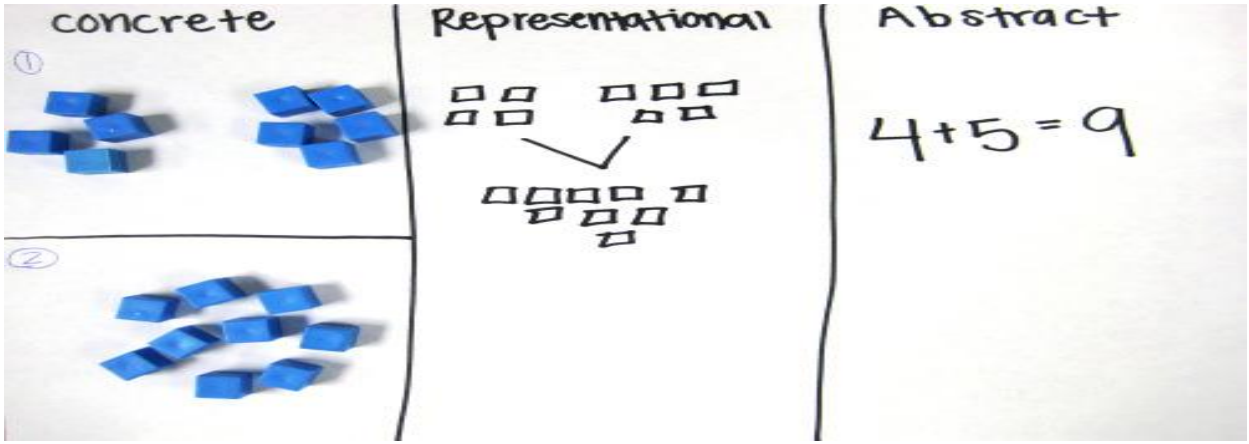
# 传统的特征提取

- 特征提取
  - 线性投影（子空间）
    - PCA、LDA
  - 非线性嵌入
    - LLE、Isomap、谱方法
  - 自编码器

- 特征提取VS表示学习
  - 特征提取：基于任务或先验对去除无用特征
  - 表示学习：解构潜在因子

# 表示学习与深度学习

- 一个好的表示学习策略必须具备一定的深度
  - 特征重用
    - 指数级的表示能力
  - 抽象表示与不变性
    - 抽象表示需要多步的构造



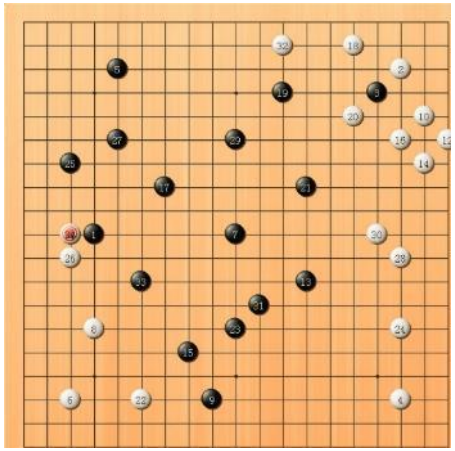

https://mathteachingstrategies.wordpress.com/2008/11/24/concrete-and-abstract-representations-using-mathematical-tools/

# 表示形式

- 离散表示
  - 局部表示、符号表示
  - One-Hot向量

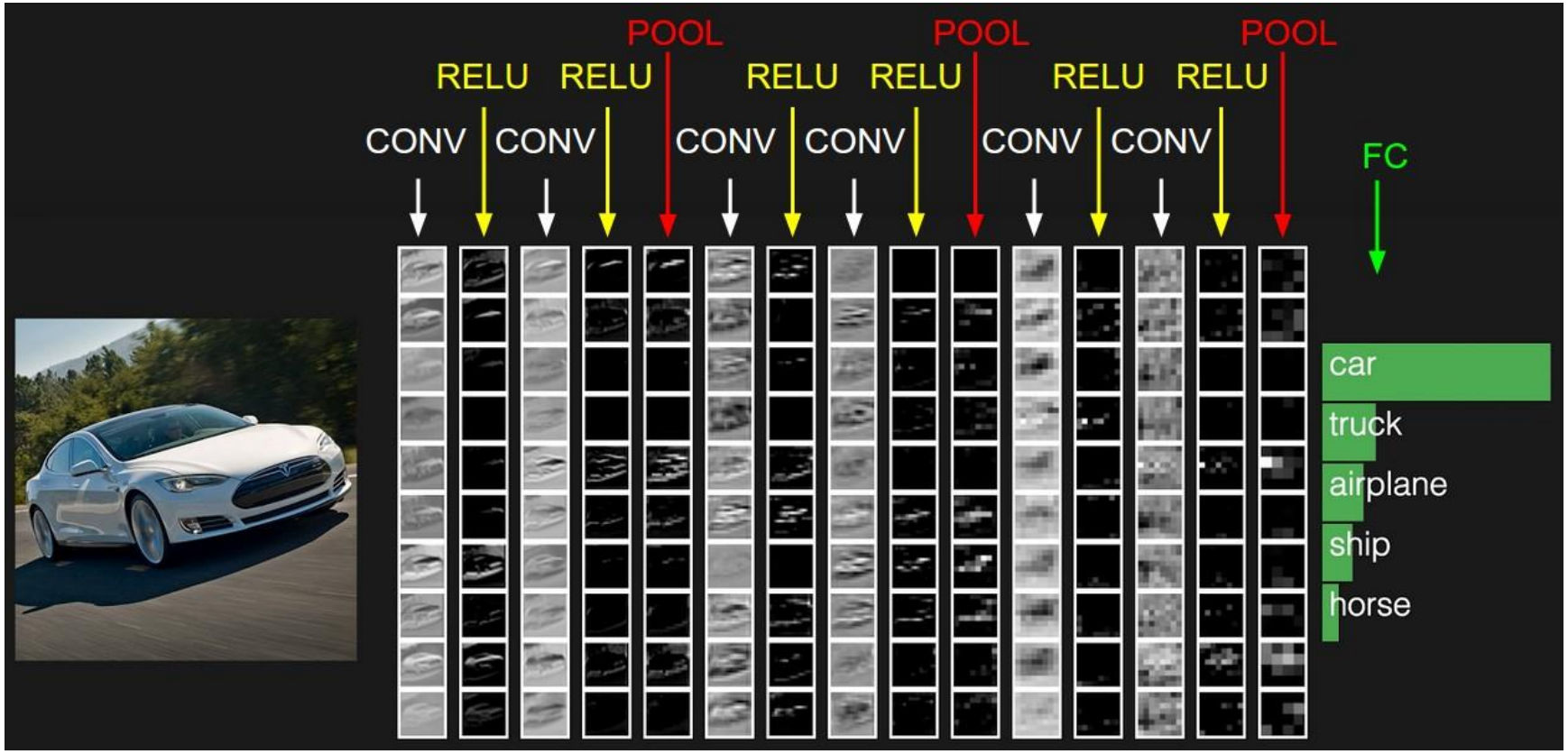
| | 离散表示 | 连续表示 |
|---|---|---|
| A | [1 0 0 0] | [0.25 0.5] |
| B | [0 1 0 0] | [0.2 0.9] |
| C | [0 0 1 0] | [0.8 0.2] |
| D | [0 0 0 1] | [0.9 0.1] |

- 分布式(distributed)表示
  - 压缩、低维、稠密向量
  - 用O(N)个参数表示 $O(2^k)$ 区间
    - k为非0参数，k<N

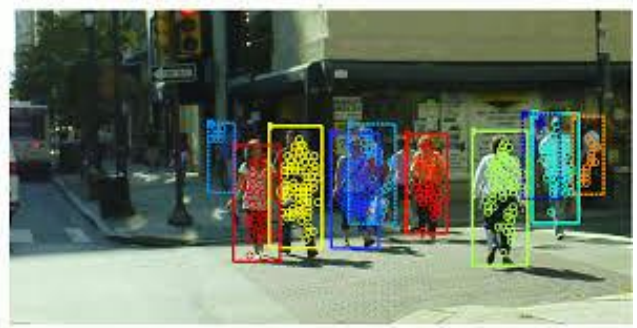


分布式表示

# 表示学习与深度学习


POOL
POOL
POOL
RELU
RELU
RELU
RELU
RELU
RELU
CONV
CONV
CONV
CONV
CONV
CONV
FC
car
truck
airplane
ship
horse

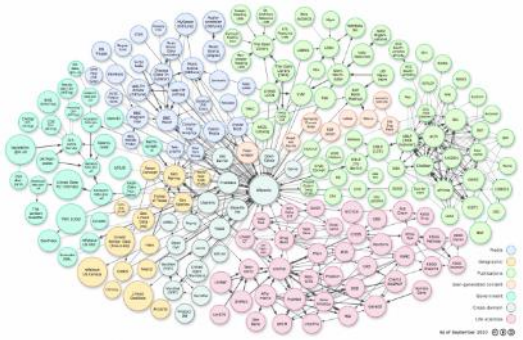
# 几乎覆盖所有人工智能领域



计算机视觉


自然语言处理

推荐系统



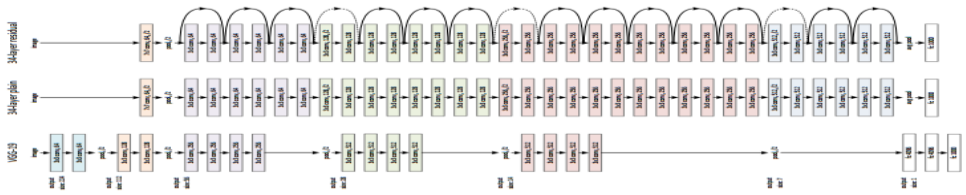
知识图谱

社交网络

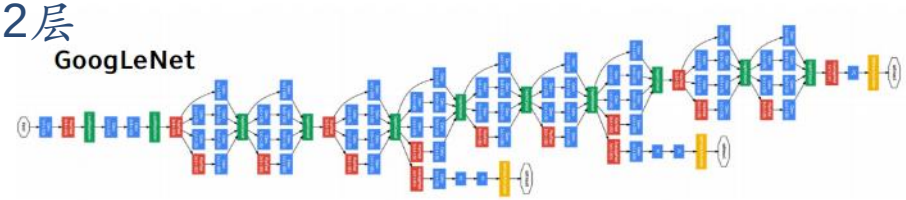
# 语言表示学习

# 为什么语言表示学习更难？

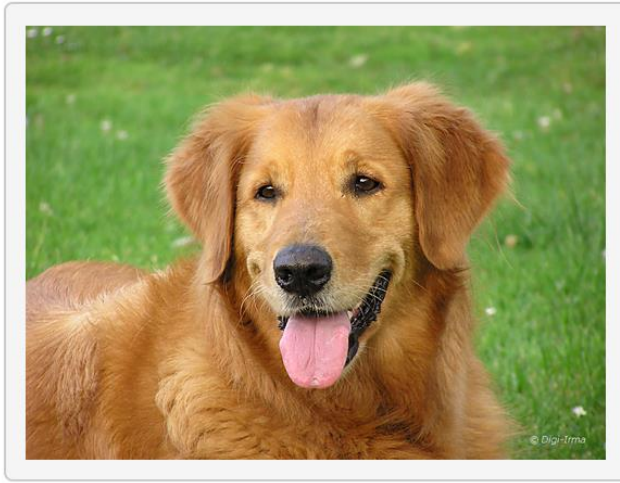
152 层



22层




计算机视觉中的深层网络模型



Results:

- golden retriever: 0.97293
- Tibetan mastiff: 0.01576
- Irish setter: 0.00364
- redbone: 0.00152
- standard poodle: 0.00127

对应NLP的最底层：词汇

# 文本与图像信息的差异

| | 输入量 | 信息量 | 关系 | 底层特征 |
|---|---|---|---|---|
| 图像 | 二维像素集<br>200X200 | 黑白：128-256<br>彩色：3（128-256） | 欧氏空间 | 纹理，形状<br>彩色 |
| 文本 | 一维离散词符号序列<br>几千-几万个词 | 共250K（英文词类）<br>一般用几千个词 | 语法关系<br>句法关系<br>语义关系 | 句子长度，句子在段落中的位置，段落在文章中的位置，.. |

# 语言表示学习

- 词
  - 分布式(distributional)表示
    - 基于分布式假设
    - 共现矩阵
  - 分散式(distributed)表示
    - 压缩、低维、稠密向量
    - 和局部（local）表示对应

为避免歧义

https://zhuanlan.zhihu.com/p/22386230

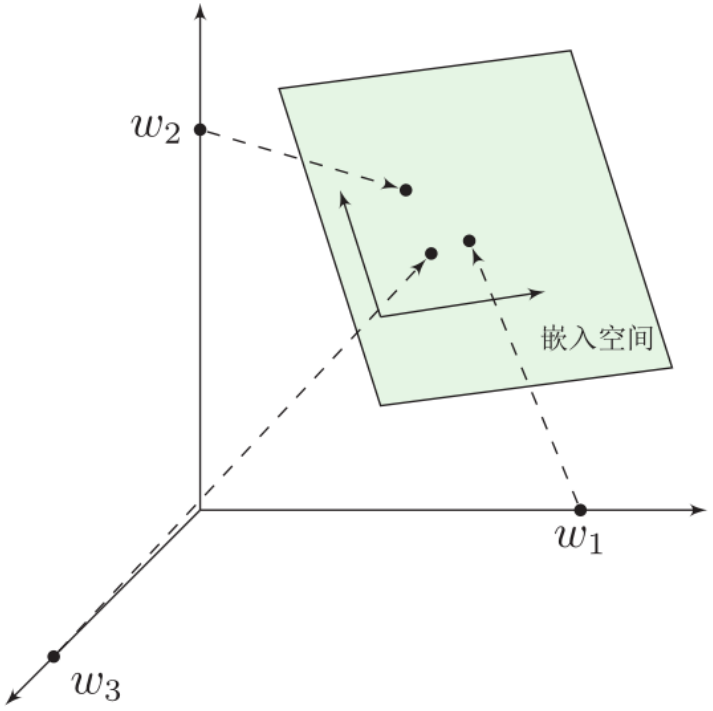
- 短语
  - 组合语义模型
- 句子
  - 连续词袋模型
  - 序列模型
  - 递归组合模型
  - 卷积模型
- 篇章
  - 层次模型

# 分布式表示
## --来自神经科学的证据


http://www.nature.com/nature/journal/v532/n7600/full/nature17637.html

# 词嵌入

# 词嵌入




如何获得一个好的词嵌入呢?

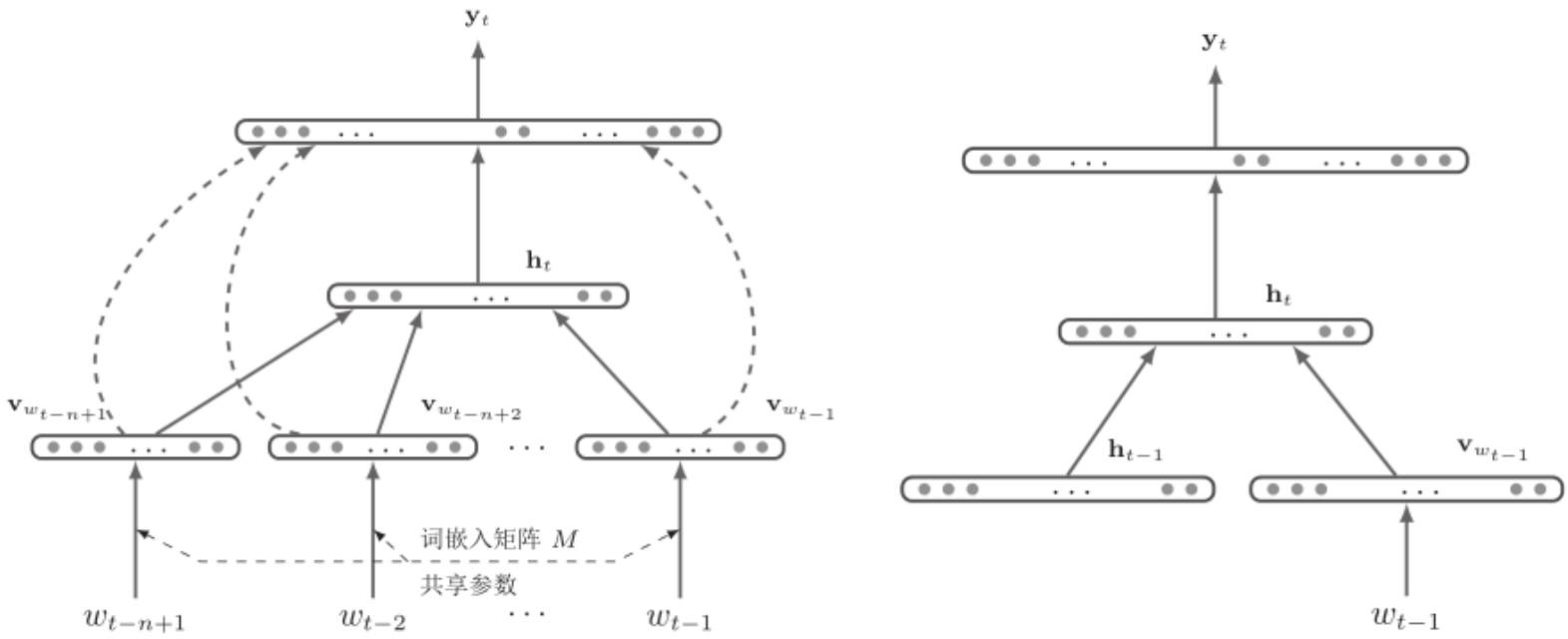
# 语言模型

- 自然语言理解 → 一个句子的可能性/合理性
  - ！ 在报那猫告做只
  - 那只猫在作报告！
  - 那个人在作报告！


- 一切都是概率!
  - $P(x_1, x_2, \cdots, x_n)$
  - $= \prod_i P(x_i | x_{i-1}, \cdots, x_1)$
  - $\approx \prod_i P(x_i | x_{i-1}, \cdots, x_{i-n+1}) = g(h_t)$

$h_t$为内部状态（或记忆）

# 神经网络语言模型




(a) 前馈神经网络语言模型 Bengio et al.[2003a]

(b) 循环神经网络语言模型 Mikolov et al. [2010]

# 神经网络语言模型

- 不足
  - 最后一层为扁平的softmax函数
  - 词典大小

- 改进
  - 层次化softmax
  - 重要性采样

详细介绍见： https://nndl.github.io/

# 进一步改进

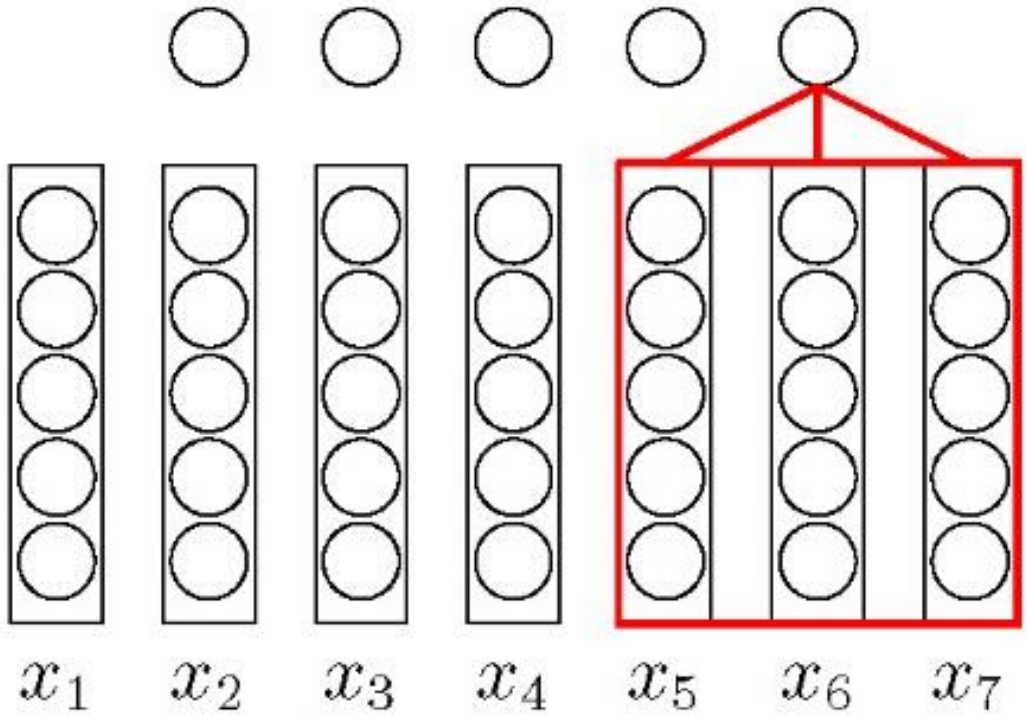
- Word2Vec
  - 抛弃困惑度指标，直接学习词嵌入
  - 删除隐藏层
  - 使用Hierarchical softmax 或negative sampling
  - 去除小于minCount的词
  - 选取邻近词的窗口大小不固定

# 句子表示

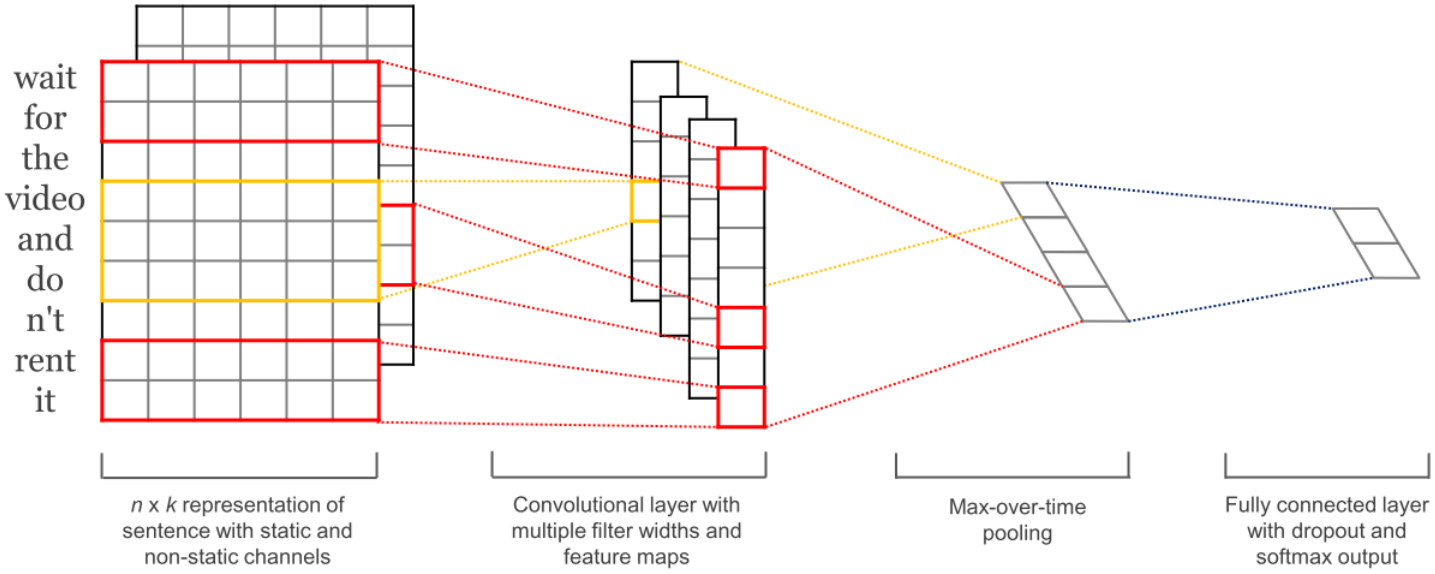
# 句子的分布式表示

- 连续词袋模型
- 序列模型
- 递归组合模型
- 卷积模型
- …

# 文本序列的卷积

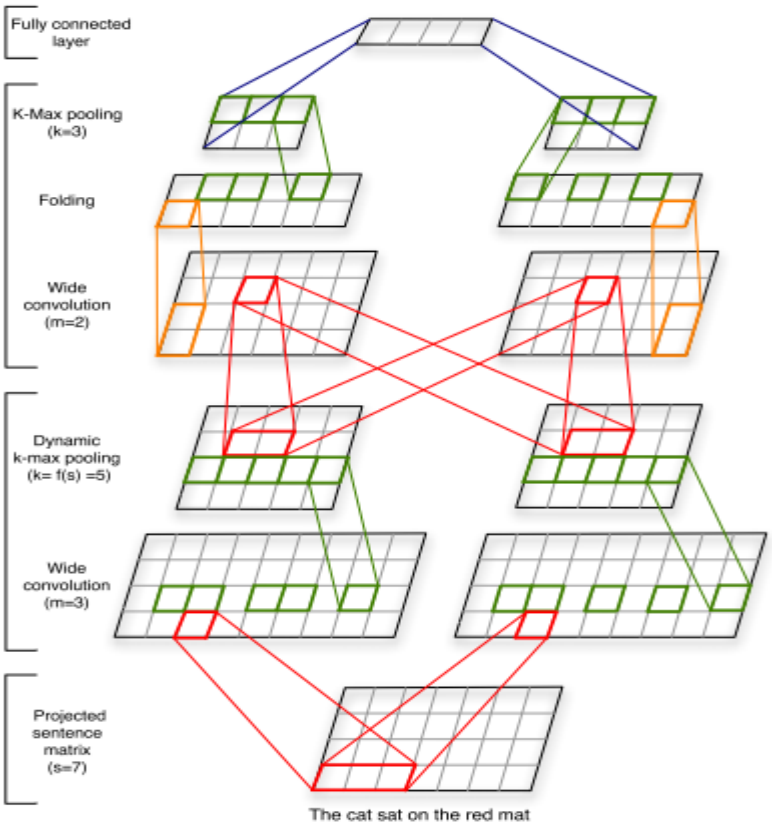


# 基于卷积模型的句子表示



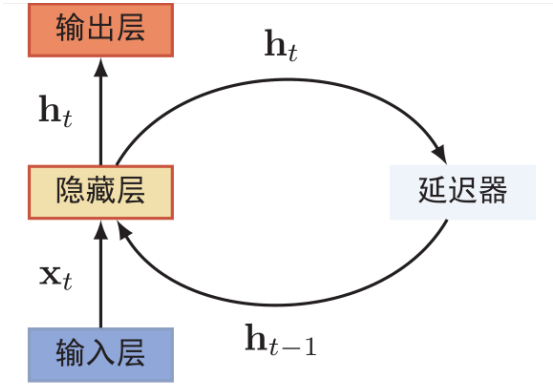

Y. Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

# 基于卷积模型的句子表示

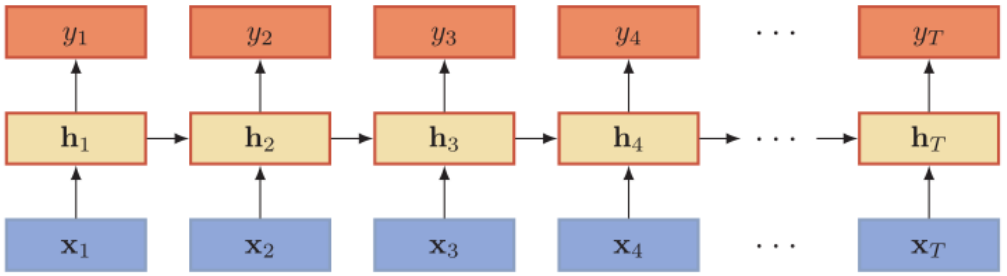



N. Kalchbrenner, E. Grefenstette, and P. Blunsom. "A Convolutional Neural Network for Modelling Sentences". In:
*Proceedings of ACL*. 2014

# 循环神经网络 (RNN)




缺点：长距离依赖问题

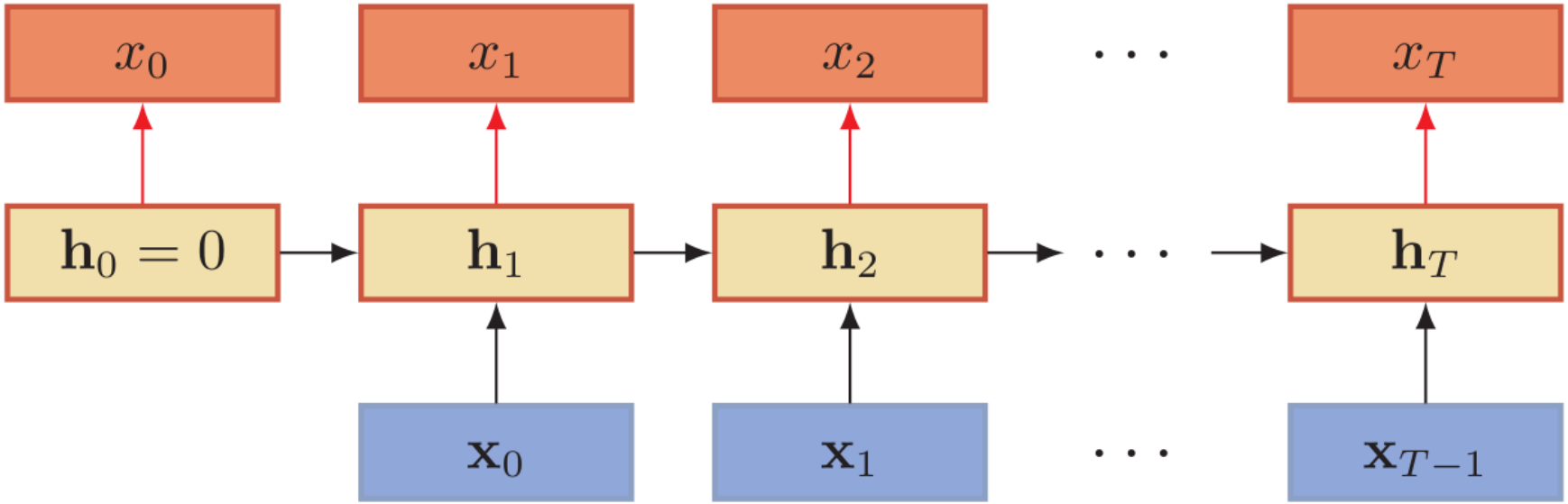



$$\mathbf{h}_t = \begin{cases} 0 & t = 0 \\ f(\mathbf{h}_{t-1}, \mathbf{x}_t) & \text{otherwise} \end{cases}$$

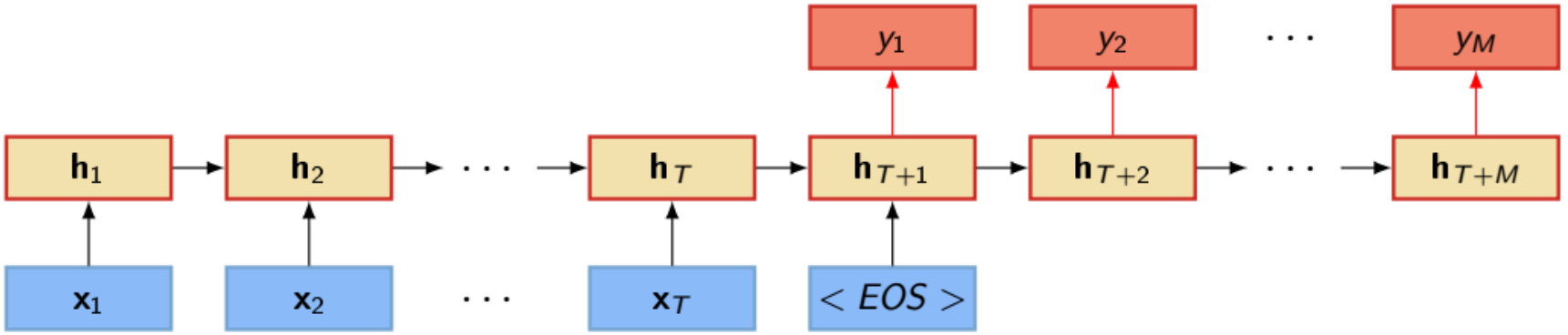
RNN是图灵完全等价的 (Siegelmann and Sontag, 1995)

FNN：模拟任何函数
RNN：模拟任何程序（计算过程）。

# 基于RNN的语言模型

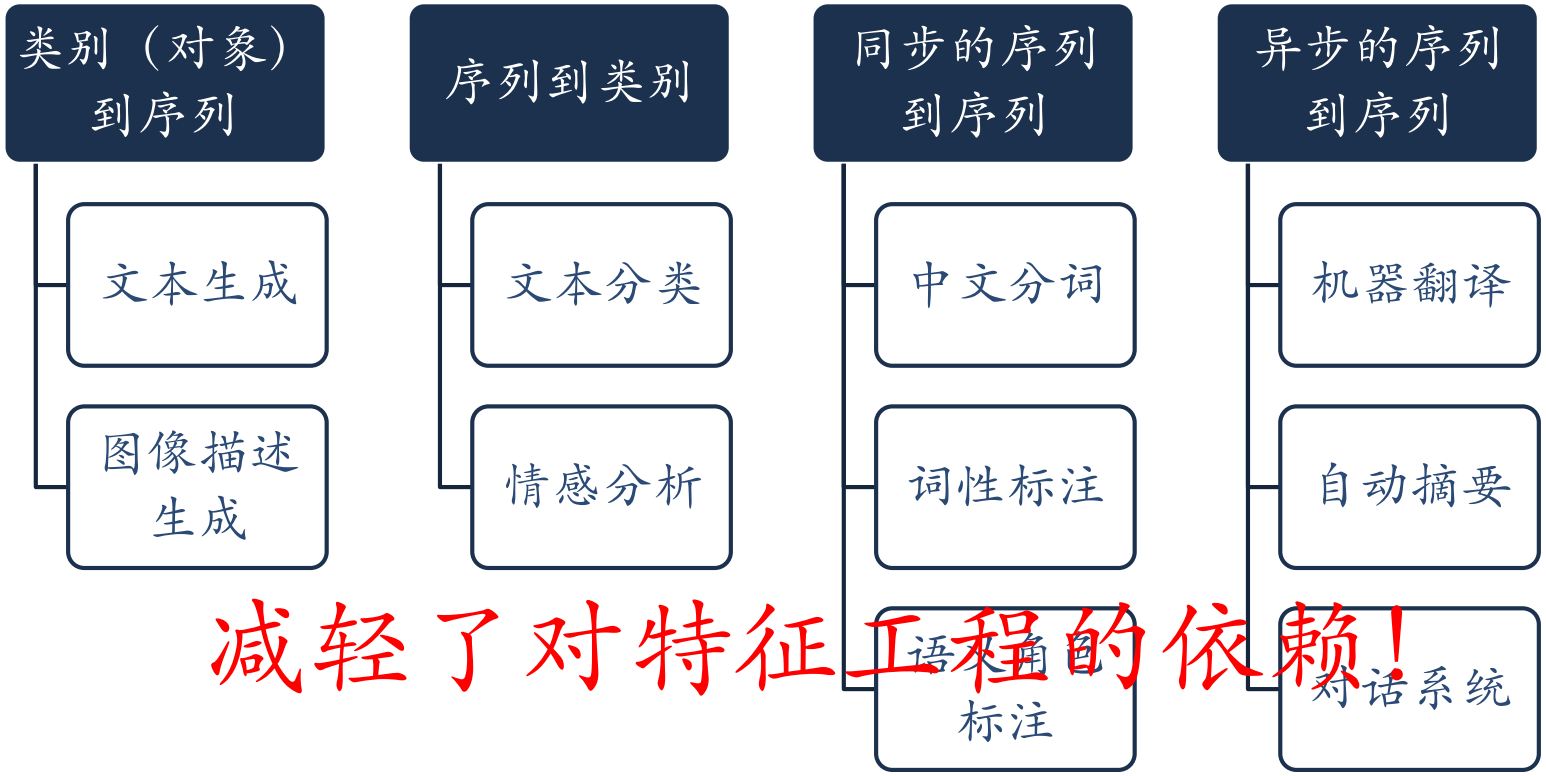

$x_0$
$x_1$
$x_2$
$\cdots$
$x_T$
$\mathbf{h}_0 = 0$
$\mathbf{h}_1$
$\mathbf{h}_2$
$\cdots$
$\mathbf{h}_T$
$\mathbf{x}_0$
$\mathbf{x}_1$
$\cdots$
$\mathbf{x}_{T-1}$

# 序列到序列模型


$y_1$
$y_2$
$y_M$
$\mathbf{h}_1$
$\mathbf{h}_2$
$\mathbf{h}_T$
$\mathbf{h}_{T+1}$
$\mathbf{h}_{T+2}$
$\mathbf{h}_{T+M}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_T$
$< EOS >$

# 自然语言处理任务

- 在得到字、句子表示之后，自然语言处理任务类型划分为：




减轻了对特征工程的依赖！

# 注意力机制与外部记忆

# 长期依赖

- 简单循环神经网络存在长期依赖问题
  - （LSTM网络）引入一个近似线性依赖的记忆单元来存储远距离的信息。
  - 记忆单元的存储能力和其大小相关。如果增加记忆单元的大小，网络的参数也随之增加。

- 改进方法
  - 注意力机制
  - 外部记忆

# 外部记忆

- 外部记忆定义为矩阵$M \in R^{K \times D}$
  - K 是记忆片段的数量，D 是每个记忆片段的大小
- 外部记忆类型
  - 只读
    - Memory Network
    - RNN中的$h_t$
  - 可读写
    - NTM
- 如何读写？

# 注意力

- 当一个人在吵闹的鸡尾酒会上和朋友聊天时，尽管周围噪音干扰很多，他还是可以听到朋友的谈话内容，而忽略其他人的声音。

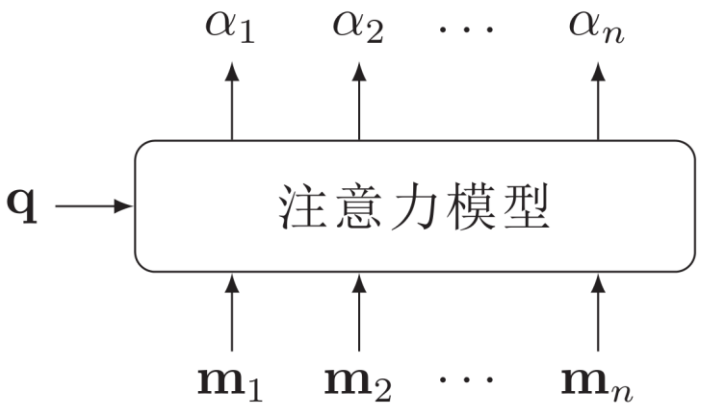
- 同时，如果未注意到的背景声中有重要的词（比如他的名字），他会马上注意到。

鸡尾酒会效应

# 注意力

给定一个上下文输入 $\mathbf{q}$，注意力 $\alpha_i(\mathbf{q})$ 为

$$\alpha_i(\mathbf{q}) = \text{softmax}(e(\mathbf{m}_i, \mathbf{q})) = \frac{\exp((\mathbf{m}_i, \mathbf{q}))}{\sum_{j=1}^{n} \exp(e(\mathbf{m}_j, \mathbf{q}))}$$



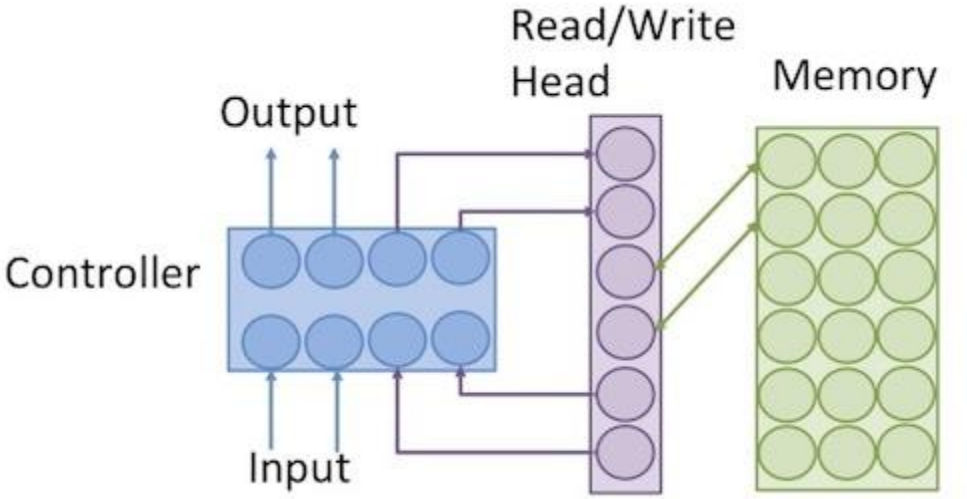

$$e(\mathbf{m}_i, \mathbf{q}) = \mathbf{v}^T \tanh(\mathbf{W}\mathbf{m}_i + \mathbf{U}\mathbf{q}),$$

$$e(\mathbf{m}_i, \mathbf{q}) = \mathbf{m}_i^\top \mathbf{q},$$

$$e(\mathbf{m}_i, \mathbf{q}) = \mathbf{m}_i^\top M\mathbf{q},$$

# 神经图灵机

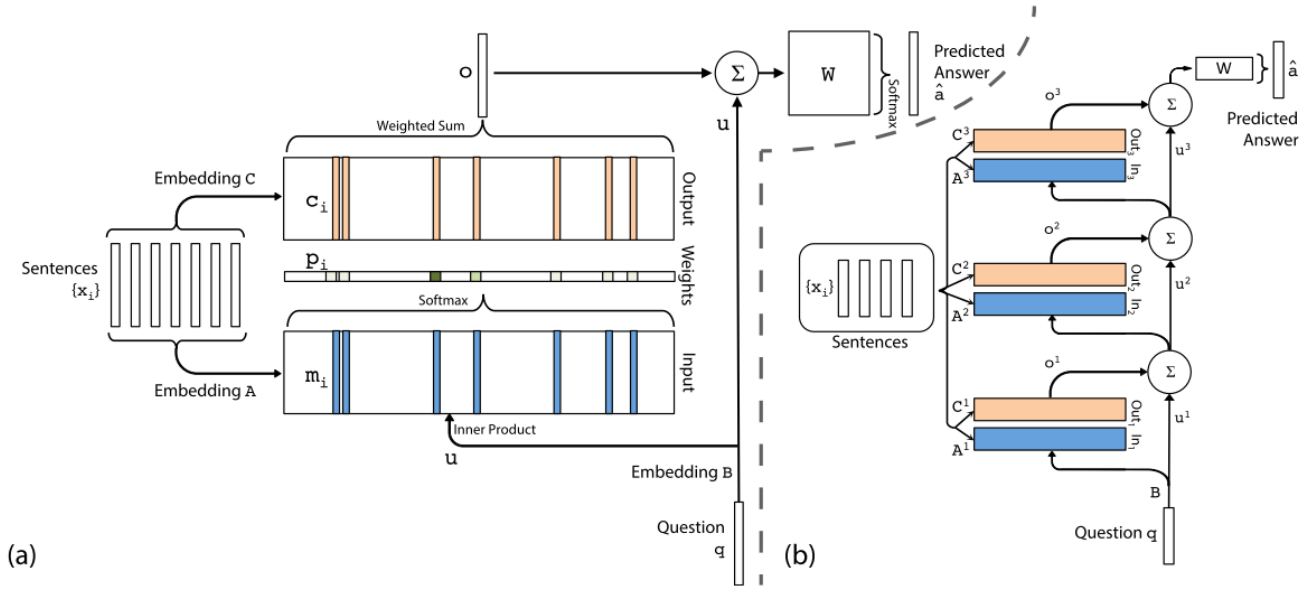
- 组件
  - 控制器
  - 外部记忆
  - 读写操作
- 整个架构可微分




图片来源：
http://cpmarkchang.logdown.com/posts/279710-neural-network-neural-turing-machine

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. Arxiv, 1–26. http://arxiv.org/abs/1410.5401

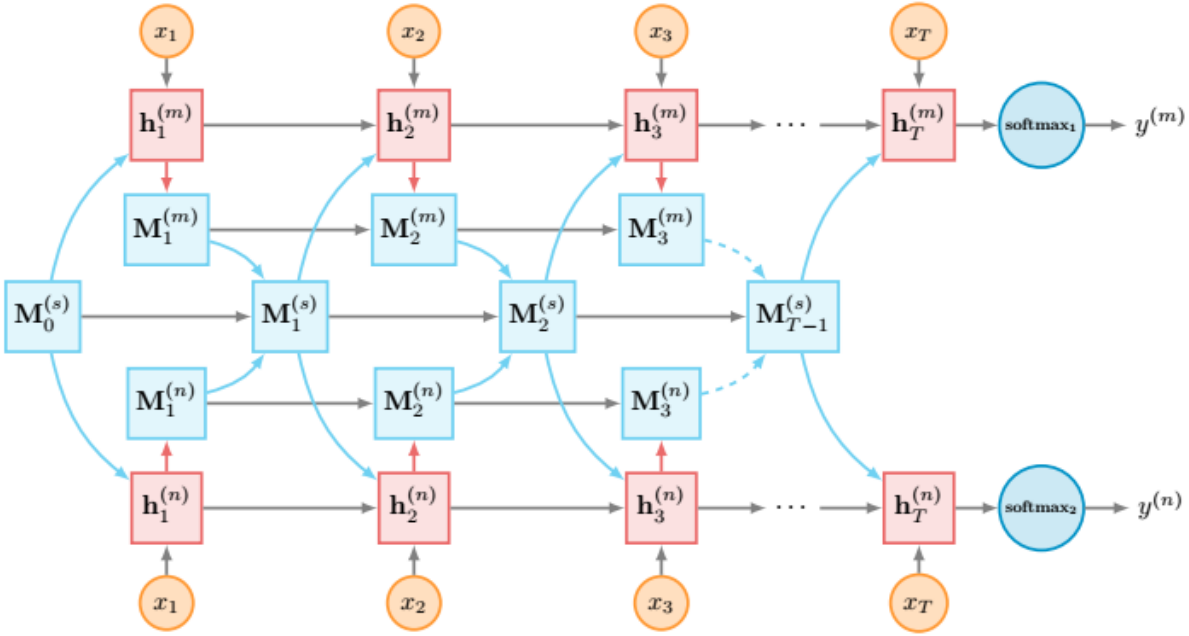
# 记忆网络




Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks, 1–11. http://arxiv.org/abs/1503.08895

# 记忆共享

- 基于神经图灵机
- 内部、外部记忆深度融合
- 外部记忆共享机制



# 文本匹配

# 文本匹配

- 定义
  - 给定两段文本A和B，判断A和B之间的关系
- 应用
  - 信息检索
  - 自动问答
  - 文本蕴涵
  - 机器翻译

# 文本匹配应用

- 自动问答

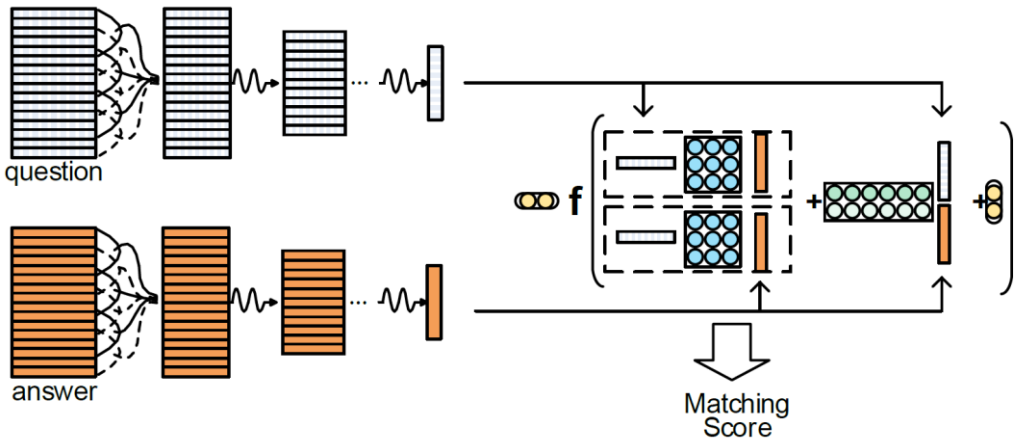
| |
|---|
| **Query**:<br>Q: Why is my laptop screen blinking? |
| **Expected**:<br>Q1: How to troubleshoot a flashing screen on an LCD monitor? |
| **Not Expected**:<br>Q2: How to make text blink on screen with Power-Point? |

# 文本匹配应用

- 文本蕴涵
  - 前提: 香港的主权和领土是在1997 年由英国归还给中国的。
  - 假设: 1997 年香港回归中国。
  - 关系： 蕴涵

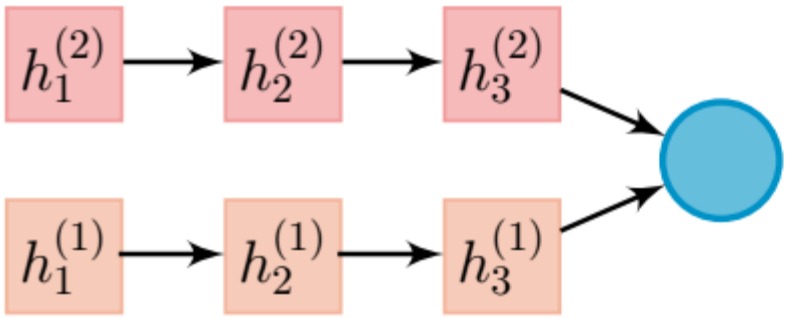
# 弱交互的匹配

- 给定两个文本的分布式表示a和b，如何计算相似度？
  - 拼接 $a \oplus b \longrightarrow ANN$
  - 双线性 $a^T M b \longrightarrow ANN$
- 改进方案

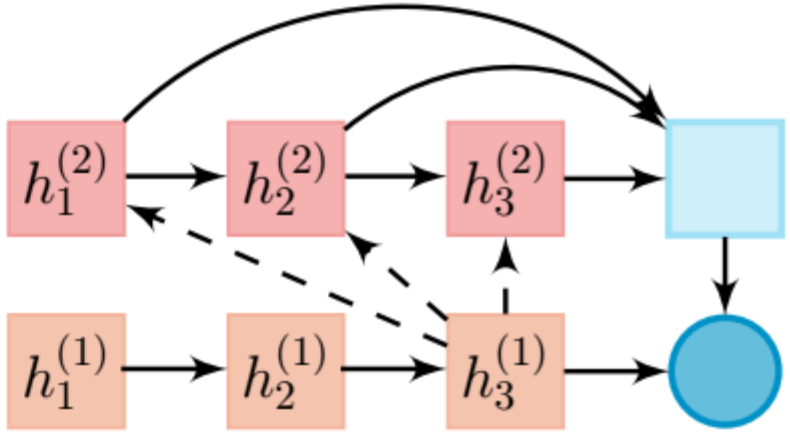



Xipeng Qiu, Xuanjing Huang, Convolutional Neural Tensor Network Architecture for Community-based Question Answering, In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2015.

# 半交互的匹配

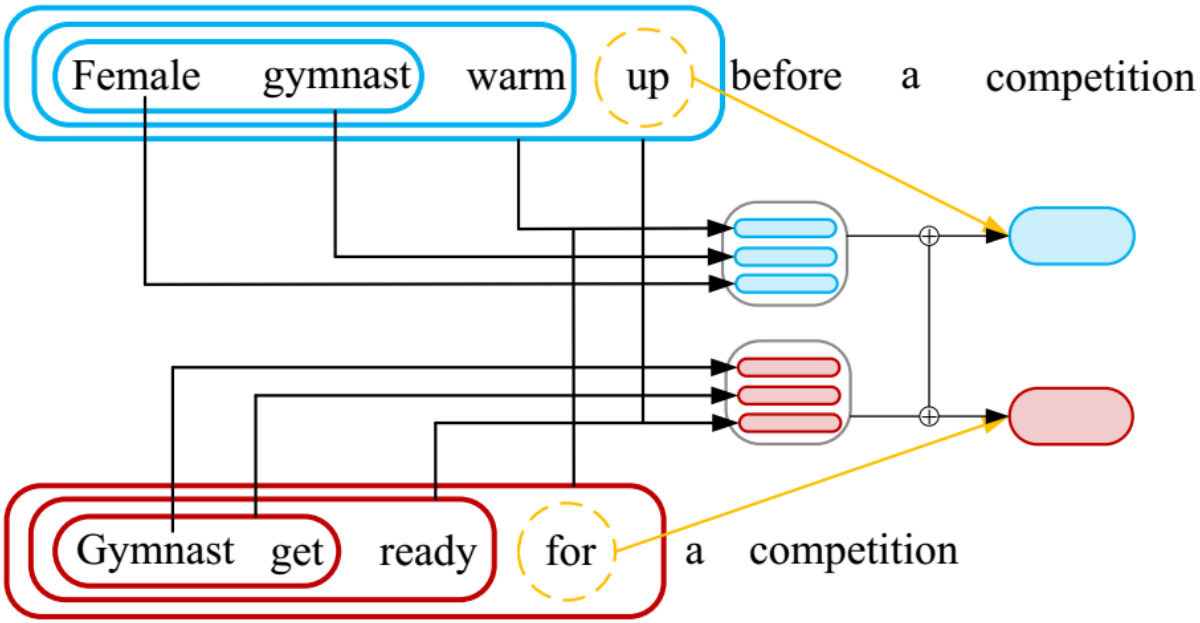



弱交互的匹配




半交互的匹配

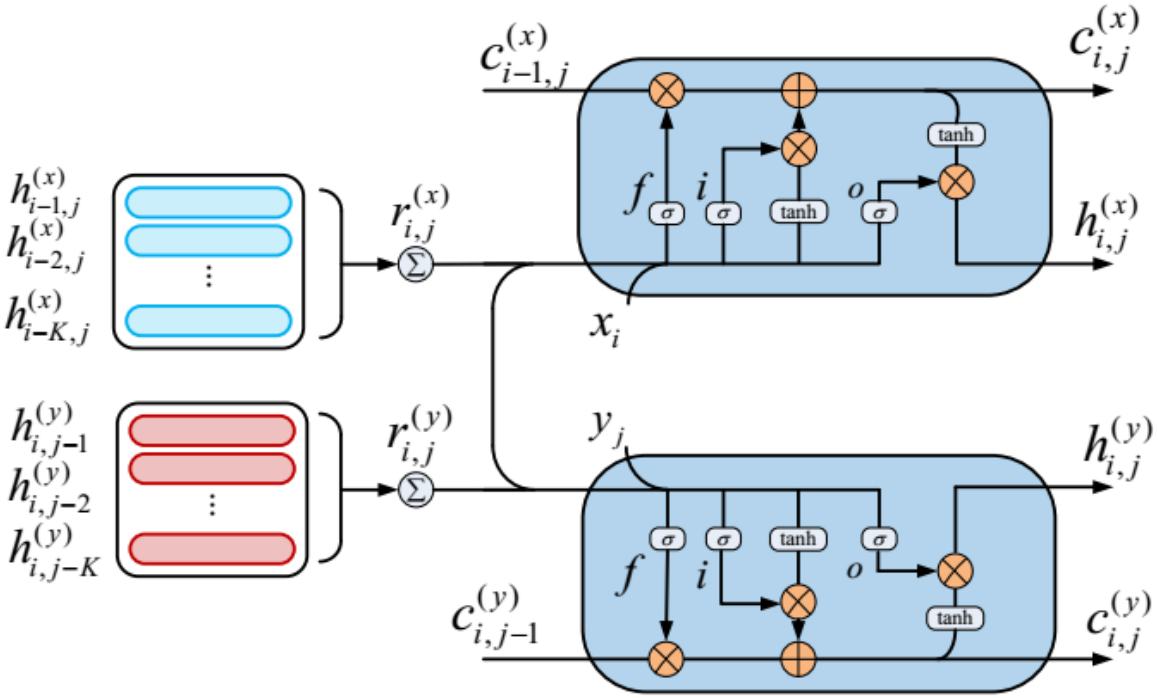
# 强交互的匹配




基于递归的文本匹配策略

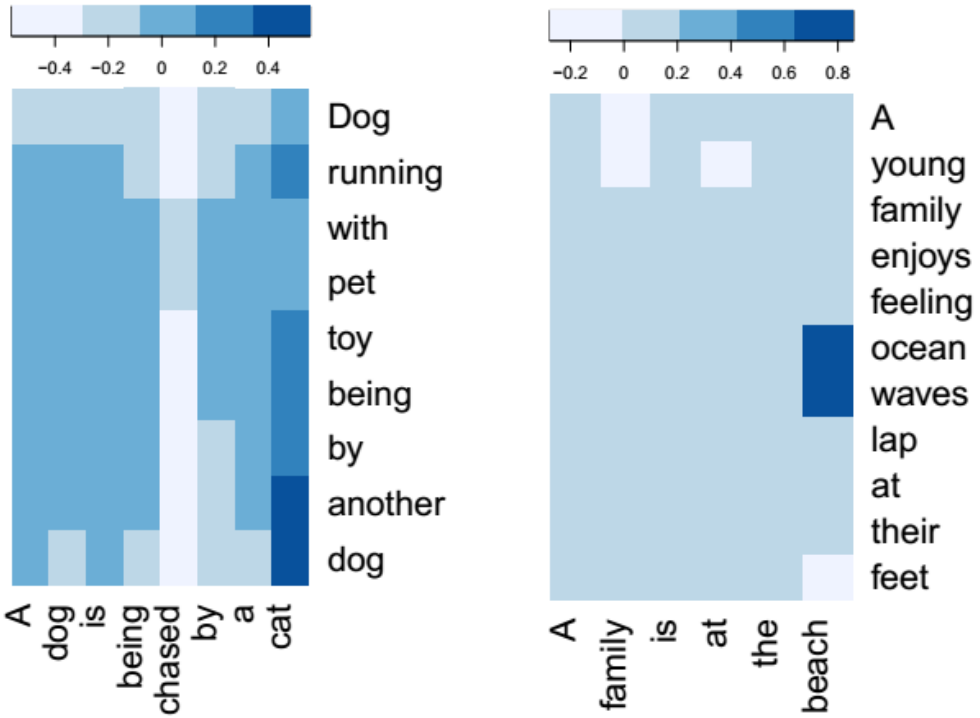
Pengfei Liu, Xipeng Qiu, Jifan Chen, Xuanjing Huang, Deep Fusion LSTMs for Text Semantic Matching, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

# Deep Fusion LSTMs




Pengfei Liu, Xipeng Qiu, Jifan Chen, Xuanjing Huang, Deep Fusion LSTMs for Text Semantic Matching, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

# 示例：神经元




(a) 5-th neuron (b) 11-th neuron

# 示例：文本对

| Index of Cell | Word or Phrase Pairs | Explanation |
|---|---|---|
| 5-th | (jeans, shirt), (dog, cat)<br>(retriever, cat), (stand, sitting) | different entities or events of the same type |
| 11-th | (pool, swimming), (street, outside)<br>(animal, dog), (grass,outside) | word pair related to lexical entailment |
| 20-th | (skateboard, skateboarding), (running, runs)<br>(advertisement, ad), (grassy, grass) | words with different morphology |
| 49-th | (blue, blue), (wearing black, wearing white),<br>(green uniform, red uniform) | words related to color |
| 55-th | (a man, two other men), (a man, two girls)<br>(Two women, No one) | subjects with singular or plural forms |

# Lexical Decomposition and Composition

- similarity matrix

$$a_{i,j} = \frac{s_i^T t_j}{\|s_i\| \cdot \|t_j\|} \qquad \forall s_i \in S, \forall t_j \in T.$$

- semantic matching functions

$$f_{match}(s_i, T) = \begin{cases} \frac{\sum_{j=0}^{n} a_{i,j} t_j}{\sum_{j=0}^{n} a_{i,j}} & global \\ \frac{\sum_{j=k-w}^{k+w} a_{i,j} t_j}{\sum_{j=k-w}^{k+w} a_{i,j}} & local\text{-}w \\ t_k & max \end{cases}$$

- decomposition

$$\alpha = \frac{s_i^T \hat{s}_i}{\|s_i\| \cdot \|\hat{s}_i\|}$$
$$s_i^+ = \alpha s_i$$
$$s_i^- = (1-\alpha) s_i$$

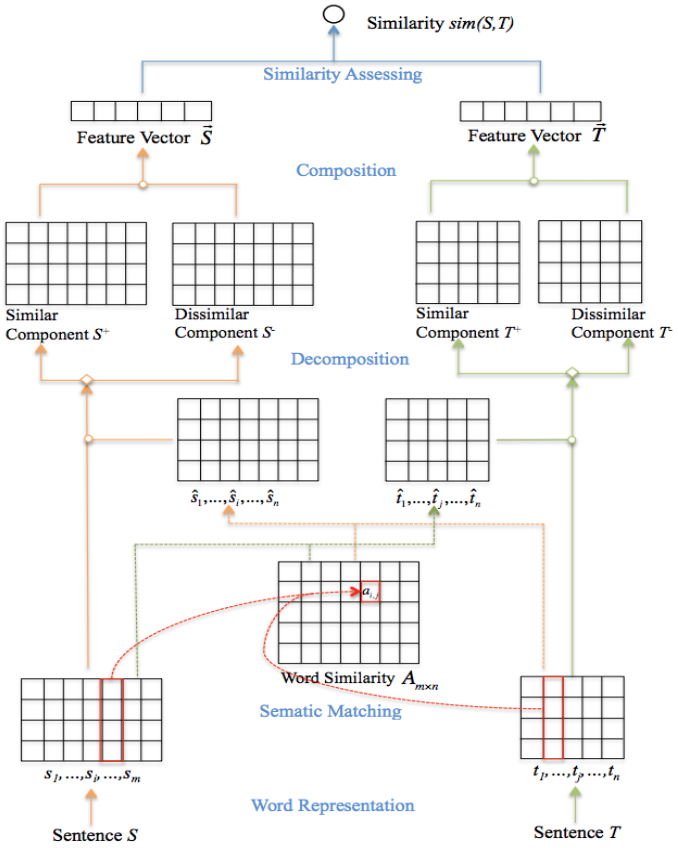



**Figure 1:** Model overview.

Wang, Zhiguo, Haitao Mi, and Abraham Ittycheriah. "Sentence Similarity Learning by Lexical Decomposition and Composition." arXiv preprint arXiv:1602.07019 (2016).

# Attention-over-Attention

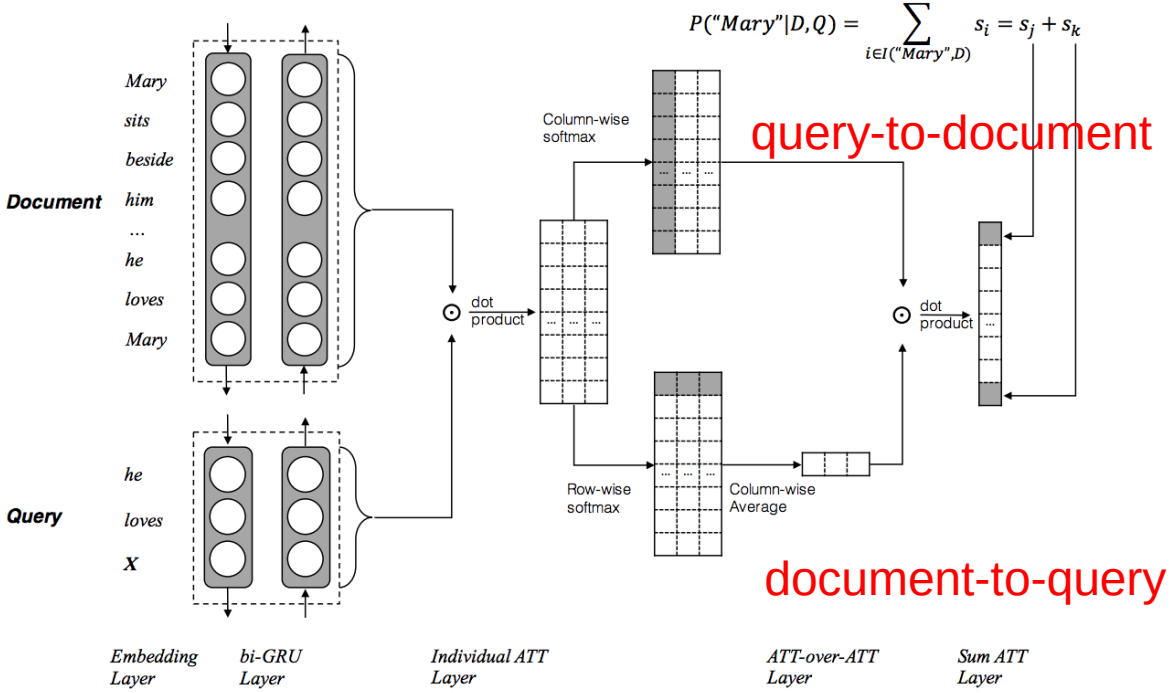



Figure 1: Neural network architecture of the proposed Attention-over-Attention Reader (AoA Reader).

Cui, Yiming, et al. "Attention-over-attention neural networks for reading comprehension." arXiv preprint arXiv:1607.04423 (2016).

# 总结

# 语言表示学习

| | | 表示学习模型 | |
|---|---|---|---|
| | | 词 | 句子、篇章 |
| 离散表示 | 符号表示 | One-Hot表示 | 词袋模型<br>N元模型 |
| | 基于聚类的表示 | Brown聚类 | K-means聚类 |
| 连续表示 | 分布式表示 | 潜在语义分析<br>潜在狄利克雷分配 | |
| | 分散式表示 | NNLM<br>Skip-Gram模型<br>CBOW模型 | 连续词袋模型<br>序列模型<br>递归组合模型<br>卷积模型 |

# 谢 谢

https://nndl.github.io/