

南昌 SMP 2016 表示学习论坛



表示学习：问题与挑战

邱锡鹏

复旦大学

2016年10月30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：需要借助人脑智能
- ▶ 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

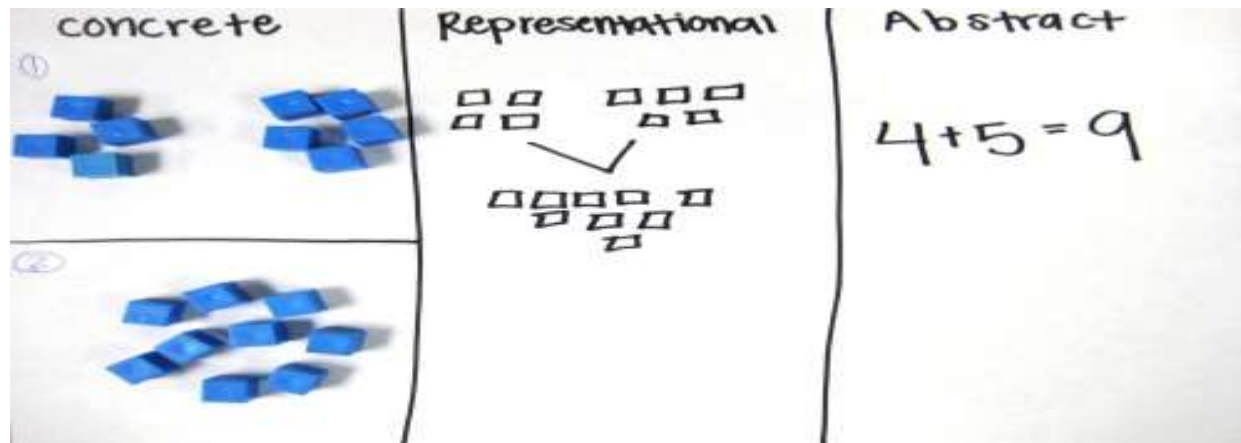
▶ 特征提取VS表示学习

- ▶ 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

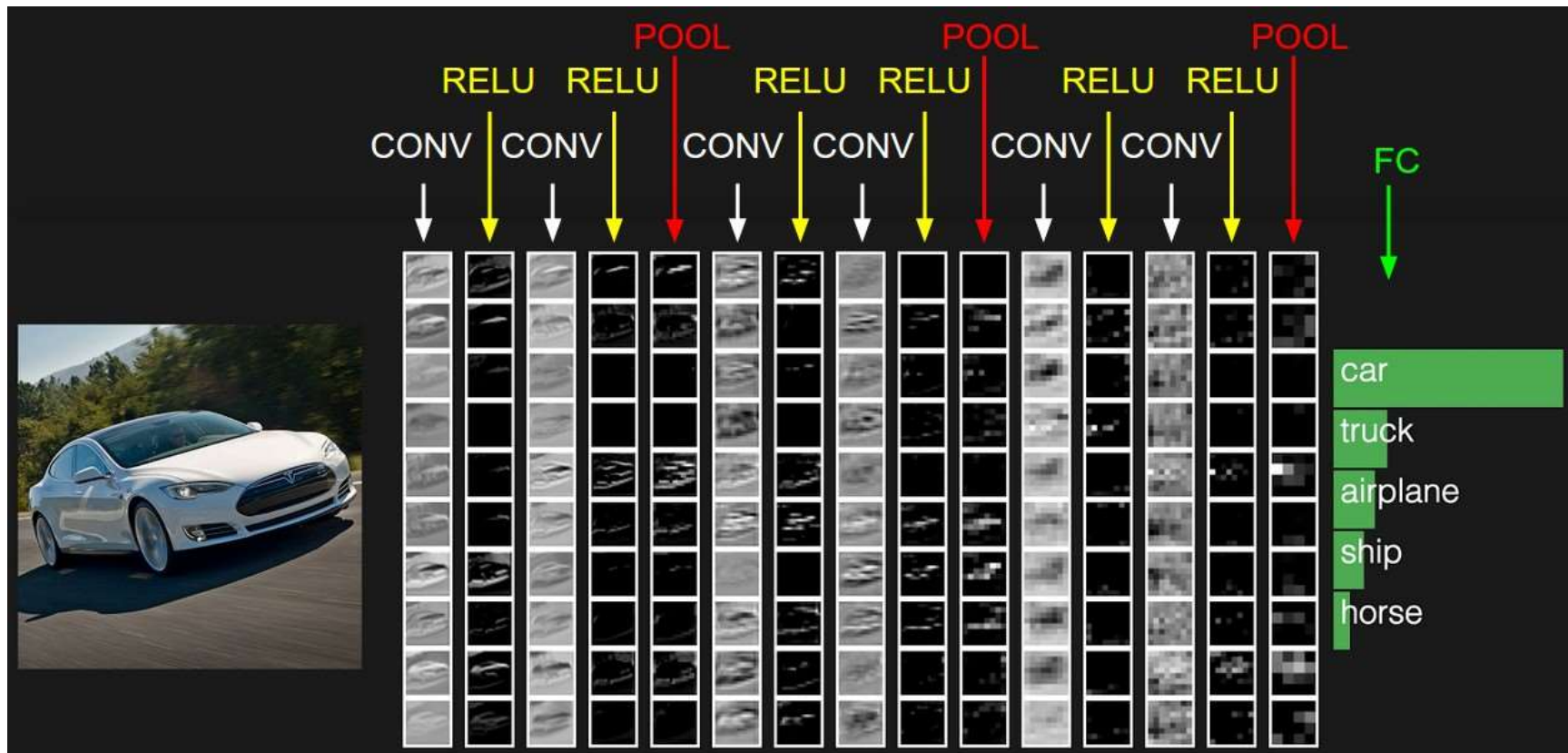
表示学习与深度学习

- ▶ 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/>

表示学习与深度学习



▶ 深度: 输入到输出之间的路径 (信用分配)

▶ 多层人工神经网络

- ▶ 层数大于1
- ▶ Schmidhuber: 10 is sure deep

▶ 表示学习

- ▶ 自动提取特征

1 billion connections

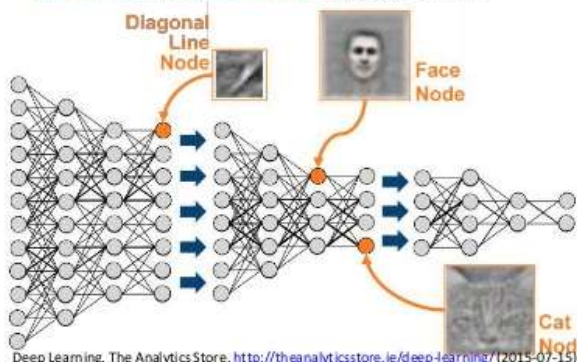
10 million

200x200 pixel

From Quoc Le, (ICML 2012)

Deep Learning

"the model has 1 billion connections, the dataset has 10 million 200 × 200 pixel images downloaded from the Internet...We train this network using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) for three days"



Building High-Level Features using Large Scale Unsupervised Learning. Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean and Andrew Y. Ng. In Proceedings of the Twenty-Ninth International Conference on Machine Learning, 2012.





深度学习

- ▶ 源于神经网络，走向人工智能
 - ▶ 控制（循环神经网络）
 - ▶ 记忆（记忆网络）
 - ▶ 注意力机制、内部记忆、外部记忆、寄存器
 - ▶ 计算模块（功能化、结构化）

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）。

表示形式

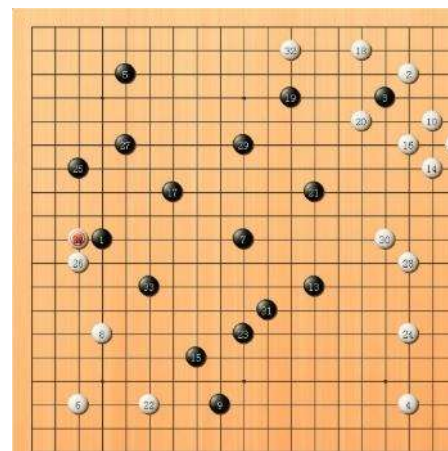
▶ 离散表示

- ▶ 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]

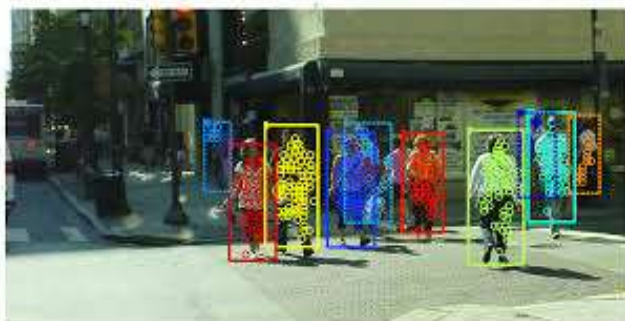
▶ 分布式(distributed)表示

- ▶ 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

几乎覆盖所有人工智能领域



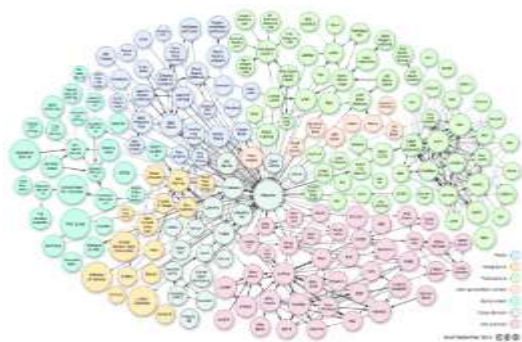
计算机视觉



自然语言处理



推荐系统



知识图谱



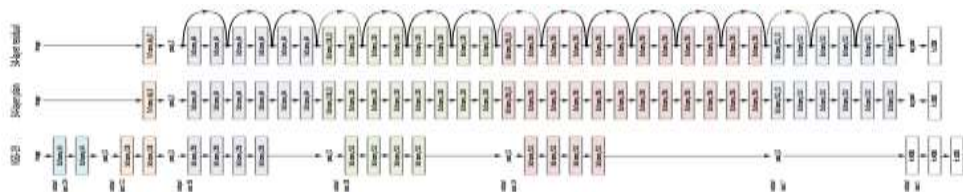
社交网络



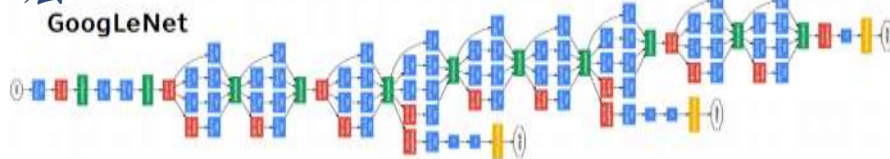
语言表示学习

为什么语言表示学习更难？

152 层



22层



Results:

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

计算机视觉中的深层网络模型

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 200X200	黑白: 128-256 彩色: 3 (128-256)	欧氏空间	纹理, 形状 彩色
文本	一维离散词符号序列 几千-几万个词	共250K (英文词类) 一般用几千个词	语法关系 句法关系 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

为避免歧义

- ▶ 分散式(distributed)表示
 - ▶ 压缩、低维、稠密向量
 - ▶ 和局部 (local) 表示对应

<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

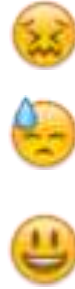
▶ 篇章

- ▶ 层次模型

语言表示学习

▶ 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!



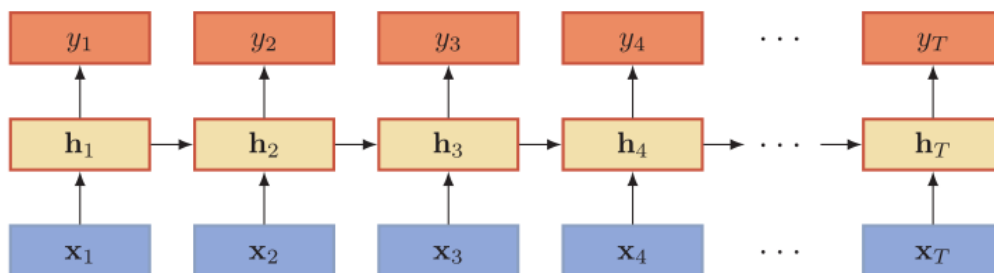
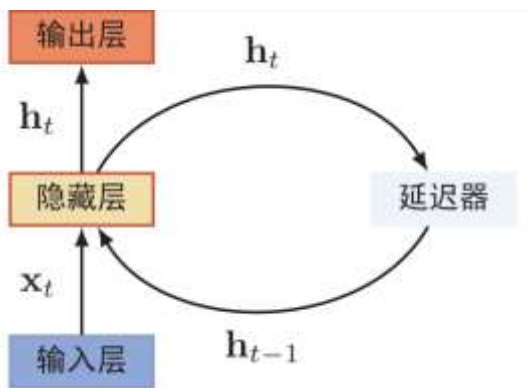
▶ 一切都是**概率**!

- ▶ $P(x_1, x_2, \dots, x_n)$
- ▶ $= \prod_i P(x_i | x_{i-1}, \dots, x_1)$
- ▶ $\approx \prod_i P(x_i | x_{i-1}, \dots, x_{i-n+1}) = g(h_t)$

h_t 为**内部状态** (或**记忆**)

循环神经网络 (RNN)

缺点：长距离依赖问题



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)。

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



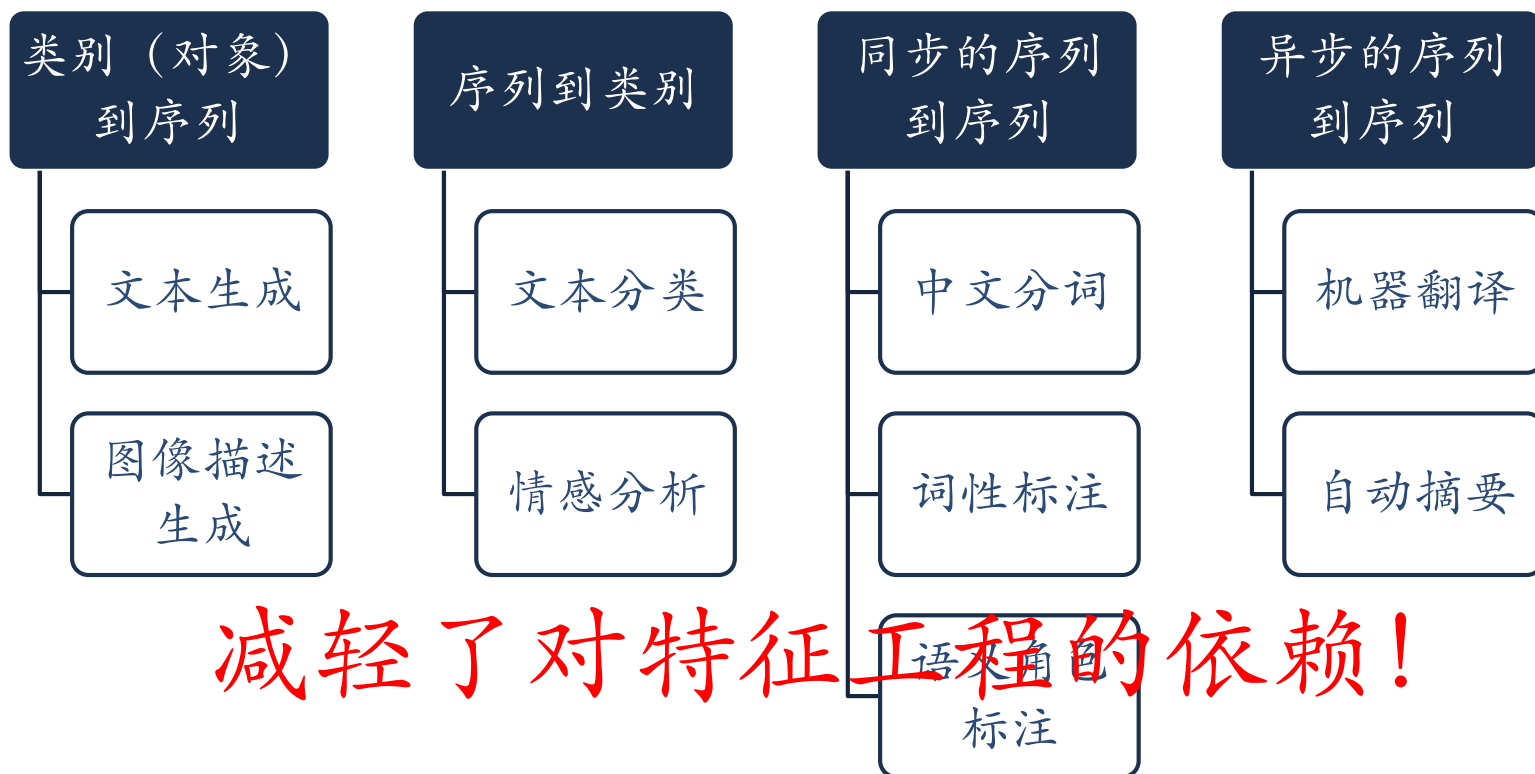
语言表示学习

		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 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 潜在狄利克雷分配	
	分散式表示	NNLM Skip-Gram模型 CBOW模型	连续词袋模型 序列模型 递归组合模型 卷积模型



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：





我们的工作 and 思考



我们的工作 and 思考

▶ 语言表示

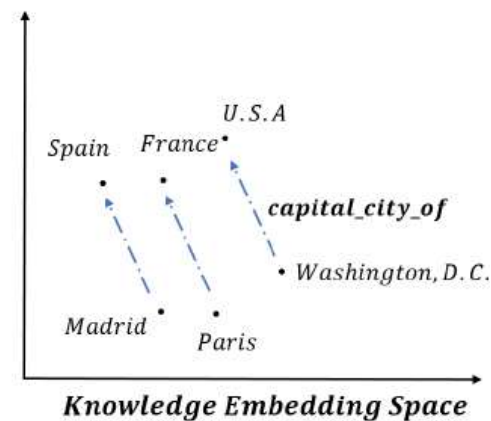
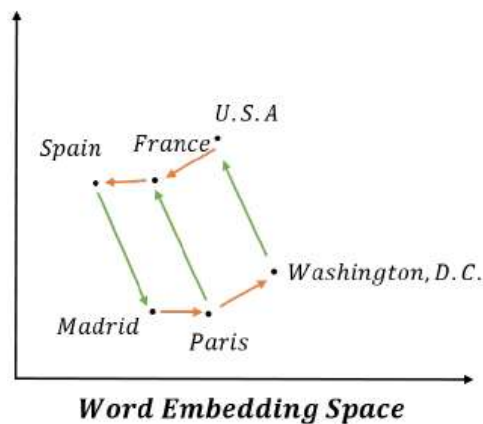
▶ 存储结构

▶ 层次化、功能化

▶ 符号表示与分布式表示之间的关系

▶ 长距离依赖问题：多尺度（周期）记忆

▶ 表示空间中“空白点”的意义





我们的工作 and 思考

▶ 学习机制

- ▶ 监督学习? 强化学习? 端到端?
- ▶ 从经验中直接学习
- ▶ 如何借鉴人脑学习机制
 - ▶ 人工记忆与人脑记忆的联系
 - ▶ Hebbian Learning
 - ▶ Zero-shot Learning



我们的工作 and 思考

- ▶ 表示共享：多任务学习
 - ▶ 内部记忆共享
 - ▶ 外部记忆共享
 - ▶ 记忆单元的可解释性

- ▶ 知识迁移
 - ▶ 不同任务之间的知识迁移



谢 谢