



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

新媒体大数据分析

New Media Big Data Analysis

第二章 数据分析

黄振亚，朱孟潇，张凯

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html>

助教：齐畅，朱家骏

bigdata_2025@163.com

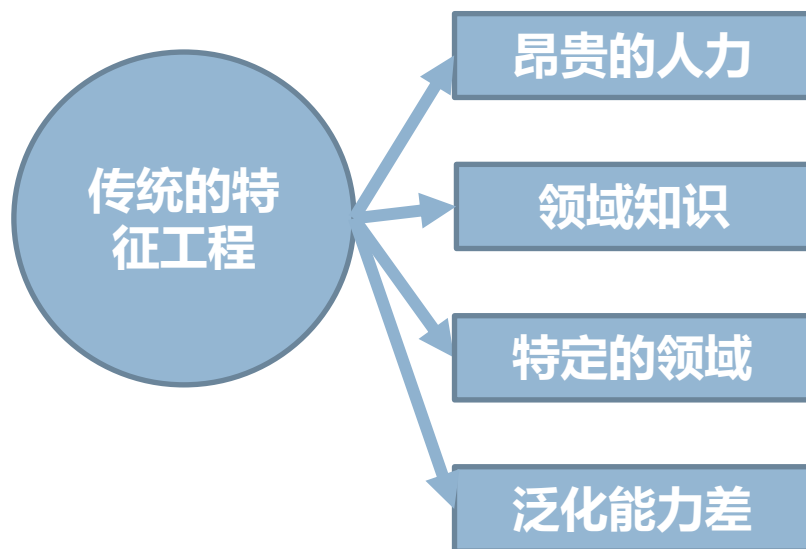
11/5/2025



传统特征工程的缺点

29

□ 传统特征工程的缺点



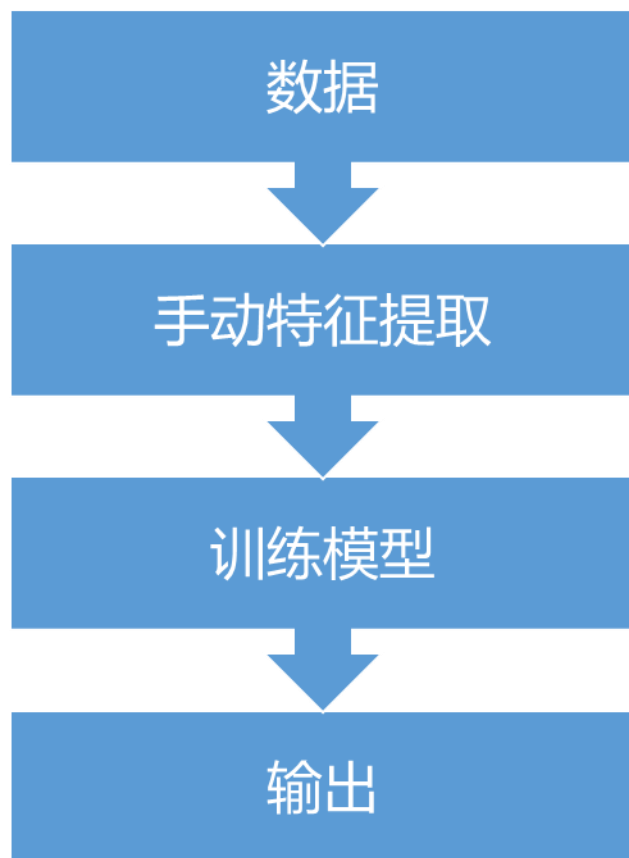


传统特征工程的缺点

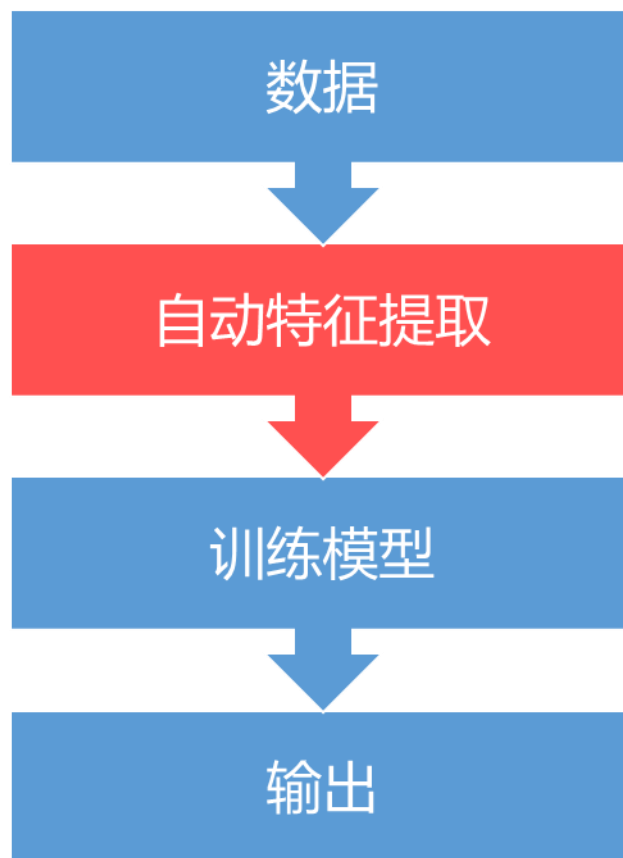
30

□ 传统特征工程的缺点

标准机器学习



深度学习





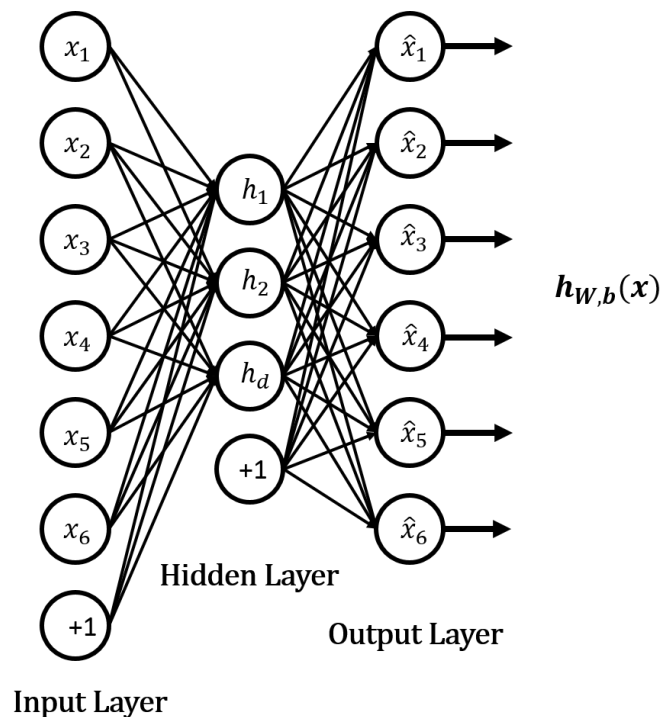
特征学习

31

□ 特征学习

如何从数据中能够自主的学习特征，在这里我们主要介绍在深度学习中常用的三种网络结构。

□ 自编码结构(Auto-Encoder)



将数据的特征 X 作为Input Layer输入
同样将原始数据特征 X' 作为Output Layer的输出来重构出原数据。

$$\text{Encoder: } H = f(A * X + b)$$

$$\text{Decoder: } X' = f(A' * H + b')$$

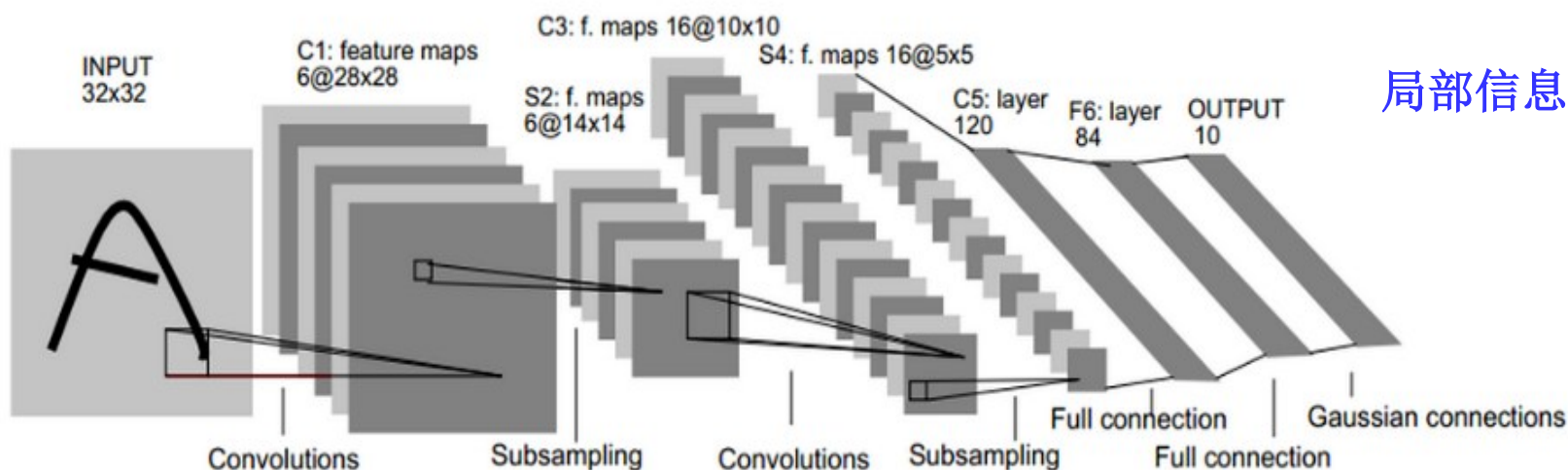
将中间的隐含层 H 的输出作为学习到的数据特征。



特征学习

32

□ 卷积神经网络(CNN): 常用于图像特征提取



卷积层: 通过局部平移, 利用不同的卷积核来提取图像中不同的特征

池化层: 计算某个区域的特征, 提高模型的泛化能力

全连接层: 通过多层的神经网络, 抽取更高阶的特征。

最终**全连接层的输出**即为该图像的特征向量表示。

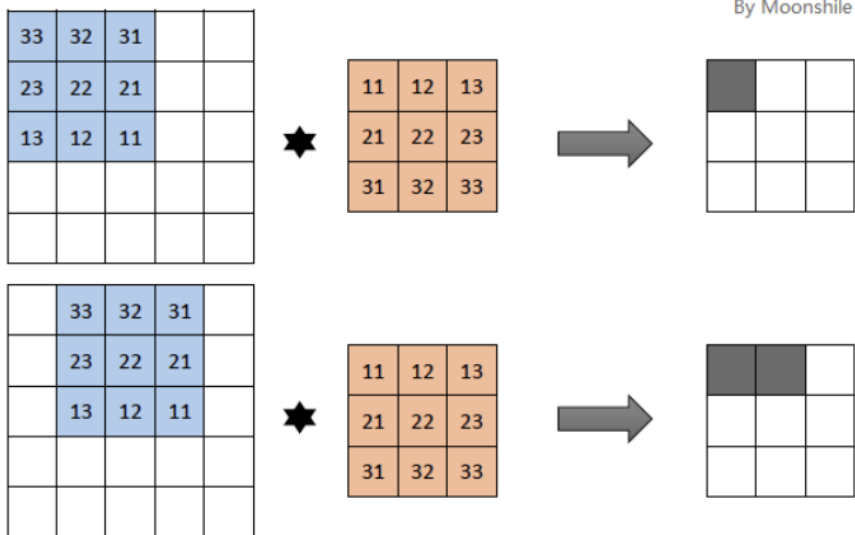


特征学习

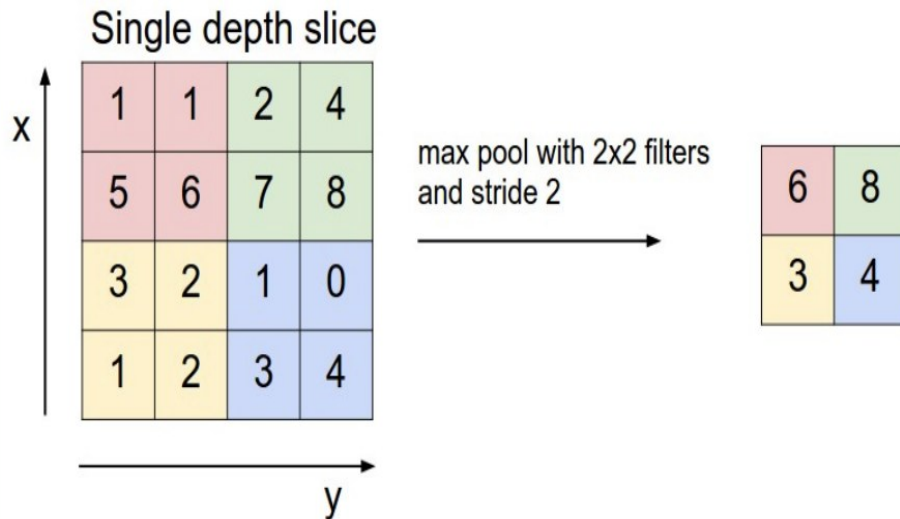
33

□ 卷积神经网络(CNN): 常用于图像特征提取

卷积操作



池化操作



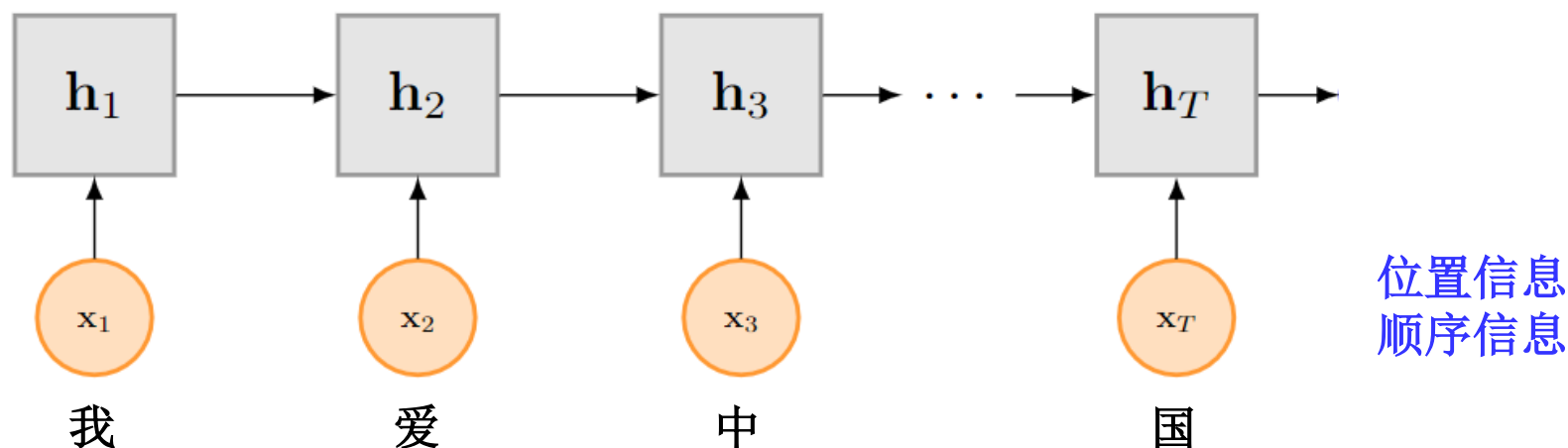
思考: 卷积神经网络能否用于文本处理? 如何做?



特征学习

34

- 循环神经网络(RNN): 常用于序列数据的特征提取



将序列中的每个数据依次作为RNN的输入，如上图中的文本数据‘我’、‘爱’、‘中’、‘国’，并将最后一层网络的输出 h_T 作为最终序列数据的特征向量



特征学习

35

□ 利用标准数据集进行特征学习（特征预训练）

□ 作用：模型效果验证 & 应用问题中的模型预训练

□ **图像数据**预训练：ImageNet

■ <http://www.image-net.org/>

■ 1400万张图片数据，2万类别，已标注

■ 常用模型：ResNet, AlexNet, VGG等

■ 常见应用：图像分类、目标检测、目标定位

□ **文本数据**预训练：Twitter, Wiki

■ <https://nlp.stanford.edu/projects/glove/>

■ 2 Billion tweets, 27 Billion 词数, 1.2M 词表

■ 常用模型：CBOW, Skip-gram, Glove等Word2vec模型

■ 常见应用：文本分类, 文本推理, 翻译等

训练好的特征即可直接作为其它模型的输入来使用



特征学习

Efficient estimation of word representations in vector space

[T Mikolov](#), [K Chen](#), [G Corrado](#), [J Dean](#) - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing ...

☆ 被引用次数: 24421 相关文章 所有 43 个版本

36

□ Word2Vec—自然语言处理的预训练

□ 哪句话更像自然语句

S1: 语言模型的本质是对一段自然语言的文本进行预测概率的大小

S2: 语言模型的本质是对自然一段语言的文本进行预测概率的大小

S3: 语言模型的本质是对自然语言一段的文本进行预测概率的大小

□ 计算词构成句子的概率—最大化

$$P(w_1, w_2, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1, \dots, w_{n-1})$$

$$L = \sum_{w \in C} \log P(w|\text{context}(w))$$



特征工程: 可解释性

Factorization machines

[S Rendle](#) - 2010 IEEE International conference

In this paper, we introduce Factorization Machines (FM). It combines the advantages of Support Vector Machines (SVMs), FM is a general predictor working

☆ 被引用次数: 1862 相关文章

特征的含义

- 人工设计的特征
- 特征的构造基于先验知识，天然具有可解释性

Feature vector $\mathbf{x}$																Target y						
$\mathbf{x}^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y^{(1)}$
$\mathbf{x}^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y^{(2)}$
$\mathbf{x}^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y^{(2)}$
$\mathbf{x}^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y^{(3)}$
$\mathbf{x}^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y^{(4)}$
$\mathbf{x}^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y^{(5)}$
$\mathbf{x}^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y^{(6)}$
A B C ...				TI NH SW ST ...				TI NH SW ST ...				Time		TI NH SW ST ...								
User				Movie				Other Movies rated						Last Movie rated								

$$S = \{(A, TI, 2010-1, 5), (A, NH, 2010-2, 3), (A, SW, 2010-4, 1), \\ (B, SW, 2009-5, 4), (B, ST, 2009-8, 5), \\ (C, TI, 2009-9, 1), (C, SW, 2009-12, 5)\}$$



特征工程: 可解释性

38

□ 特征的含义

- Word2Vec—自然语言处理的预训练
- 学习语言的语义特性: 解决“语义鸿沟”

“dog”

3

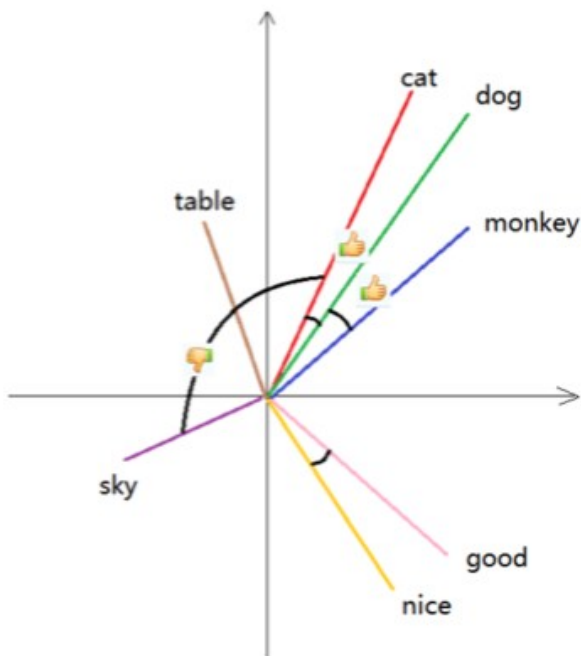
$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

“canine”

399,999

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix}$$

词的相似性



词的类比性

King – Queen $\sim$ Man – Woman

China – Beijing $\sim$ UK – London $\sim$ Capital

Efficient estimation of word representations in vector space

[T Mikolov](#), [K Chen](#), [G Corrado](#), [J Dean](#) - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing ...

☆ 被引用次数: 24421 相关文章 所有 43 个版本

11/5/2025



特征工程: 可解释性

39

□ 特征的含义—可解释性

□ CNN Feature map——特征图可视化

□ 学习图像的特点：图形图像的“颜色，边角，轮廓，图形”等

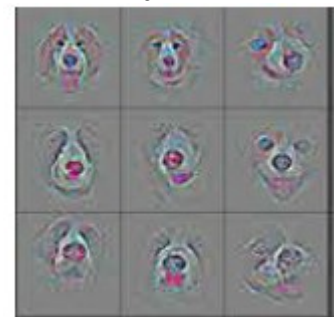
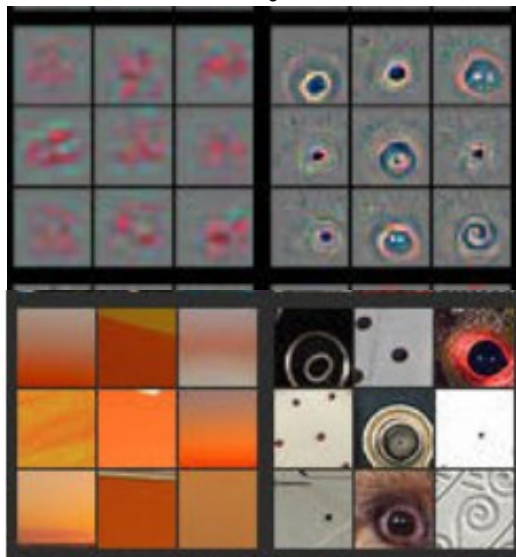
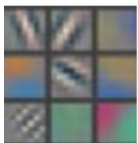
Layer 1

Layer 1

Layer 2

Layer 3

Layer 4



Visualizing and understanding convolutional networks

MD Zeiler, R Fergus - European conference on computer vision, 2014 -

Abstract Large Convolutional Network models have recently demonstrated high classification performance on the ImageNet benchmark. However, there is no clear understanding of why they perform so well, or how they might be improved. In this paper we explore both issues. We introduce a novel visualization technique that allows us to gain insight into the function of intermediate feature layers and the operation of the network. Used in a diagnostic role, these visualizations allow us to find model archi-

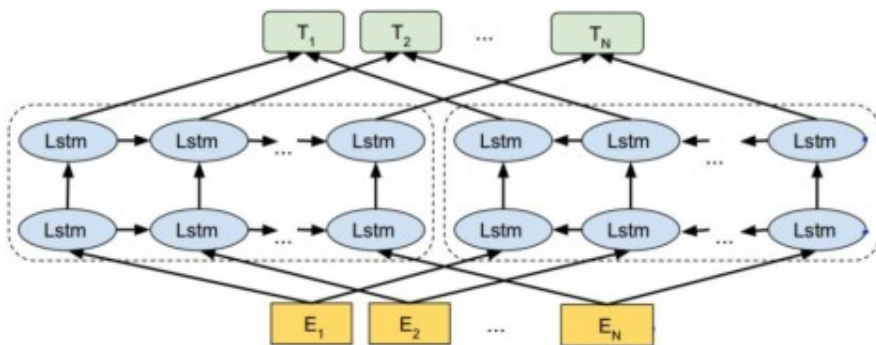
☆ 99 被引用次数: 13612 相关文章 所有 18 个版本



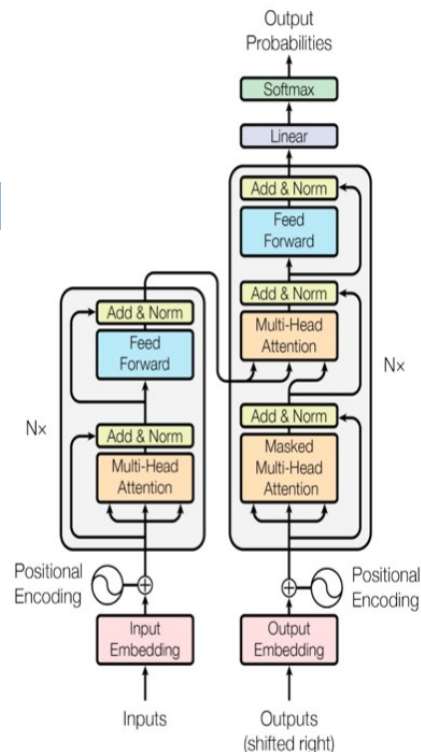
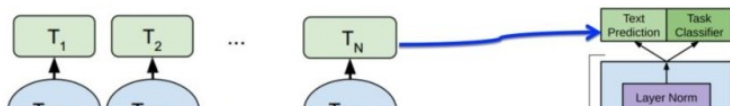
特征学习

40

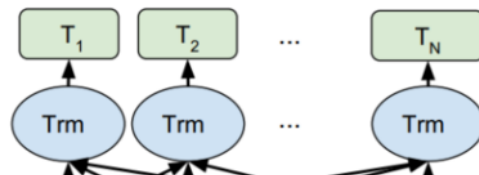
- 自然语言处理的预训练模型
 - Elmo, GPT, Transformer, BERT



OpenAI GPT



BERT (Ours)



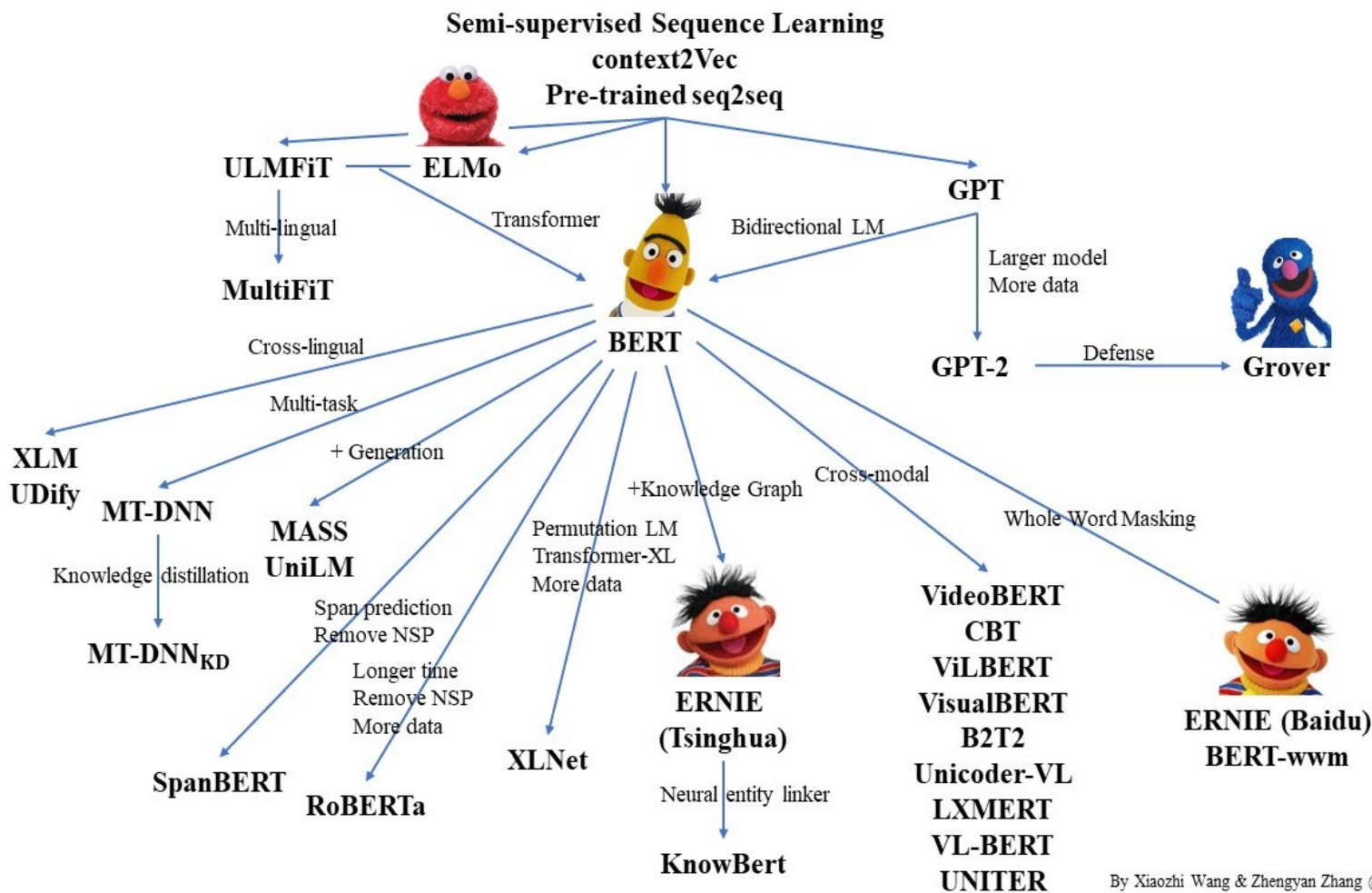
特征学习过程已经不局限于人工的思考、构造、统计等方法。它已经成为一个重要的研究方向，专门的特征学习模型已经在 CV、NLP, graph 等领域取得重要的突破。



特征学习

41

前沿：自然语言处理的预训练模型





特征学习

42

□ 前沿：大语言模型

GPT

无监督预训练，有监督微调

5G文本数据 | 1.17亿模型参数

在9/12任务上最优，包括问答、语义相似度、文本分类

2018

GPT-2

多任务、零样本学习 (zero-shot)

40G文本数据 | 15亿模型参数

在7/8任务上最优，包括阅读理解、翻译、问答

2019

GPT-3

小样本学习 (few-shot)

45T文本数据 | 1750亿模型参数

在阅读理解任务上超越当时所有 zero-shot模型

2020

大规模预训练模型

GPT-4o

多模态，可处理图像和文本输入

GPT-4的升级版模型，其中“O”是Omni的缩写，意为“全能”。其在响应速度、多模态能力、实时交互性方面较GPT-4能力有极大的提升

2024.5

GPT-4

多模态，可处理图像和文本输入

在大多数专业和学术考试中表现出人类水平，且能通过律师资格考试，排名考生中前10%，相较之下GPT-3.5排名低于后10%

2023.3

ChatGPT (3.5)

基于InstructGPT进行优化

能生成更翔实的回复：标注数据质量更高

更擅长连续对话：源于标注人员标注的多轮对话数据

2022.11

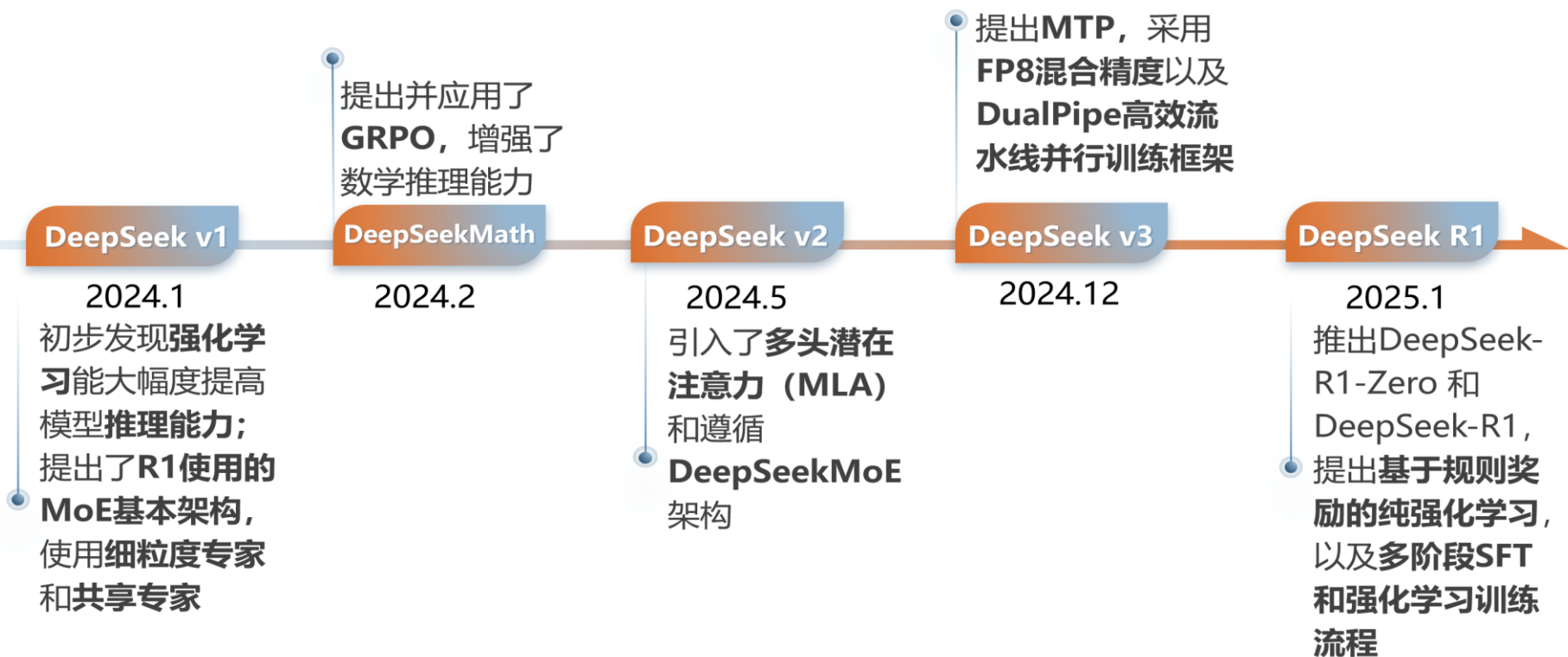
捕获人类意图进一步优化



特征学习

43

□ 前沿：大语言模型





参考文献

44

□ 书籍

- 数据挖掘导论
- 机器学习

□ 论文

- 《An Introduction to Variable and Feature Selection》
- 《特征选择常用算法综述》

□ 实战经验

- Sklearn官方文档
- Kaggle和天池比赛论坛



第二章数据分析基础小结

45

□ 数据采集

Data Collection

- 信息检索
- 网络爬虫

□ 数据预处理

Data Preprocessing

- 数据清洗
- 数据集成
- 数据变换

□ 特征工程

Feature Engineering

- 特征设计
- 特征理解