深度之眼
deepshare.net

NLP文本分类挑战赛
词向量基础

导师：阿水

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

1 词向量介绍

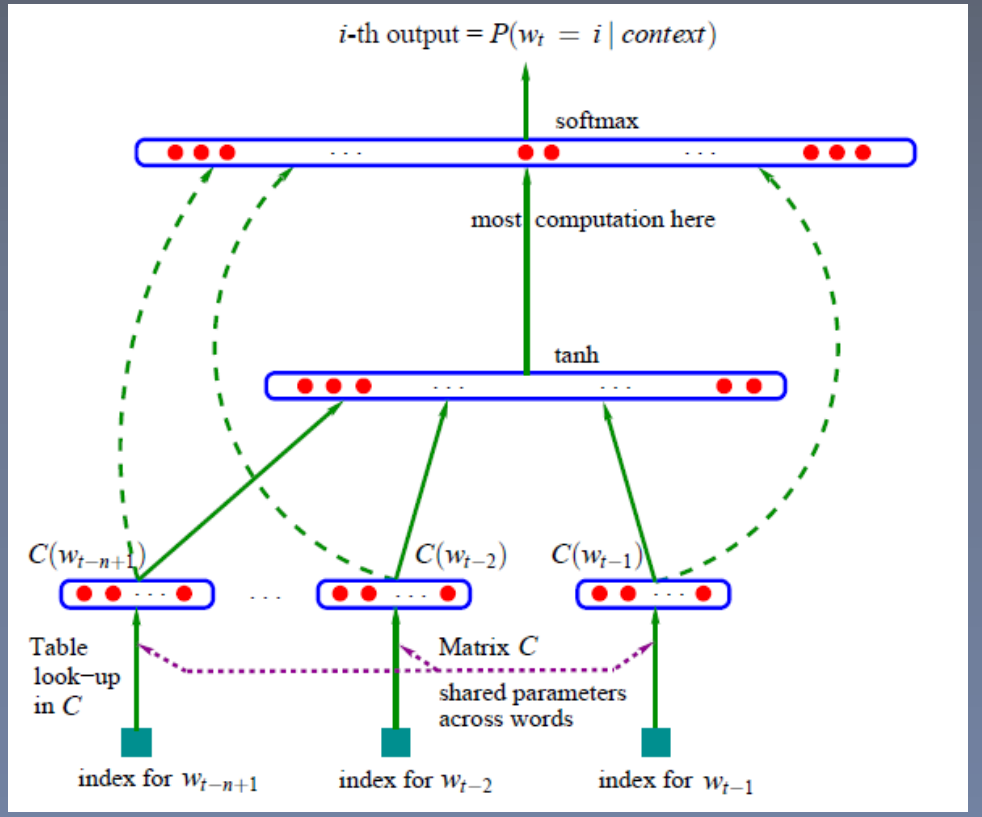
Introduction to Embedding

1 词向量介绍

Introduction to Embedding

词向量（Word Embedding）

- ✓ 低维稠密的表征方法；
- ✓ 代替了传统分布特征；
- ✓ 语言模型的副产物；



1 词向量介绍

Introduction to Embedding

为什么要有词向量？

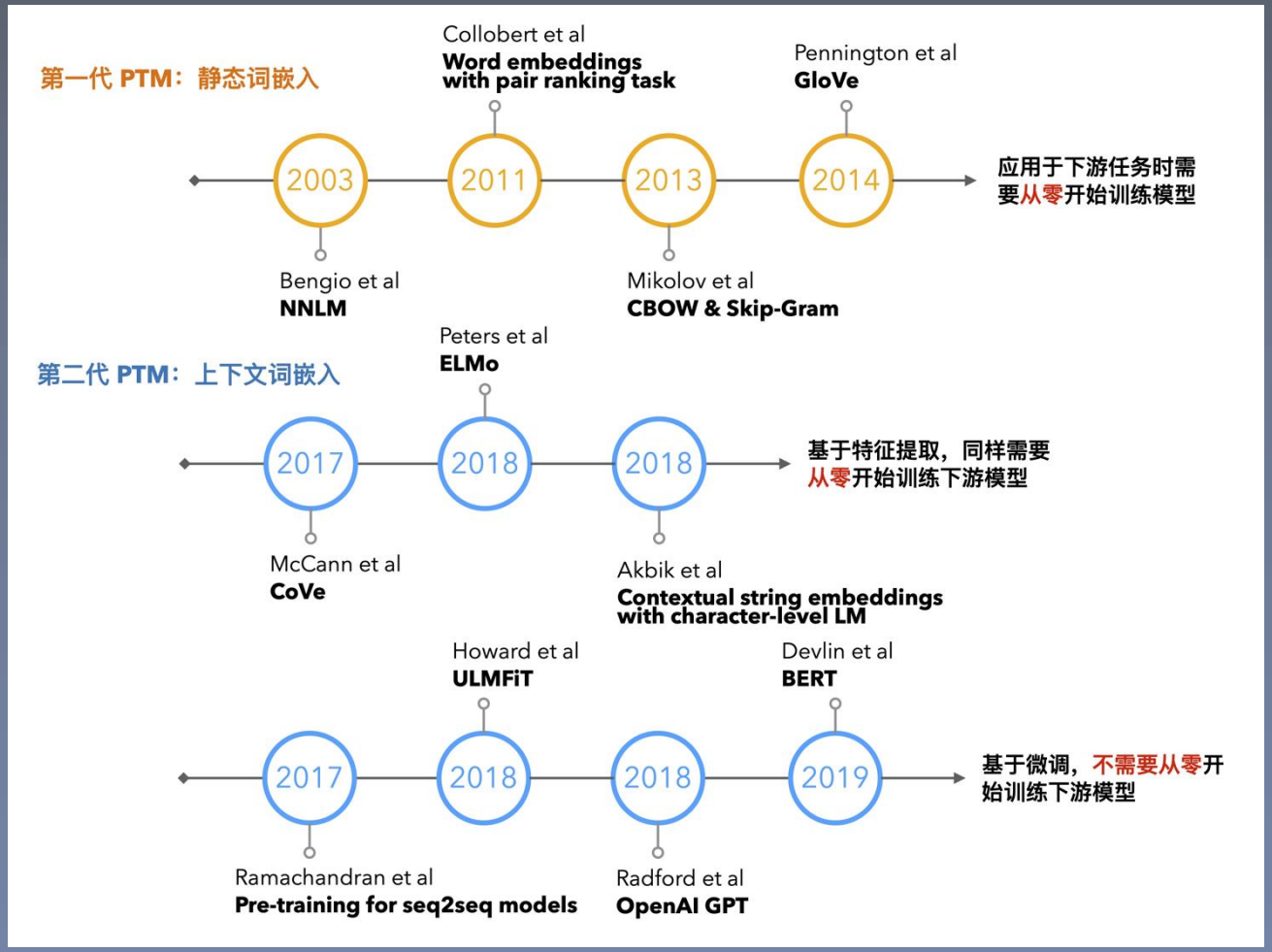
- ✓ 预训练可以有效帮助下游任务；
- ✓ 预训练提供模型初始化，有更好的泛化性能和更快的收敛速度；
- ✓ 预训练是一种避免过拟合的正则化方法；

1 词向量介绍

Introduction to Embedding

词向量的发展：

- ✓ 静态词嵌入（单一映射）；
- ✓ 动态词嵌入（上下文映射）；



1 词向量介绍

Introduction to Embedding

预训练任务：有监督学习、无监督学习、自监督学习；

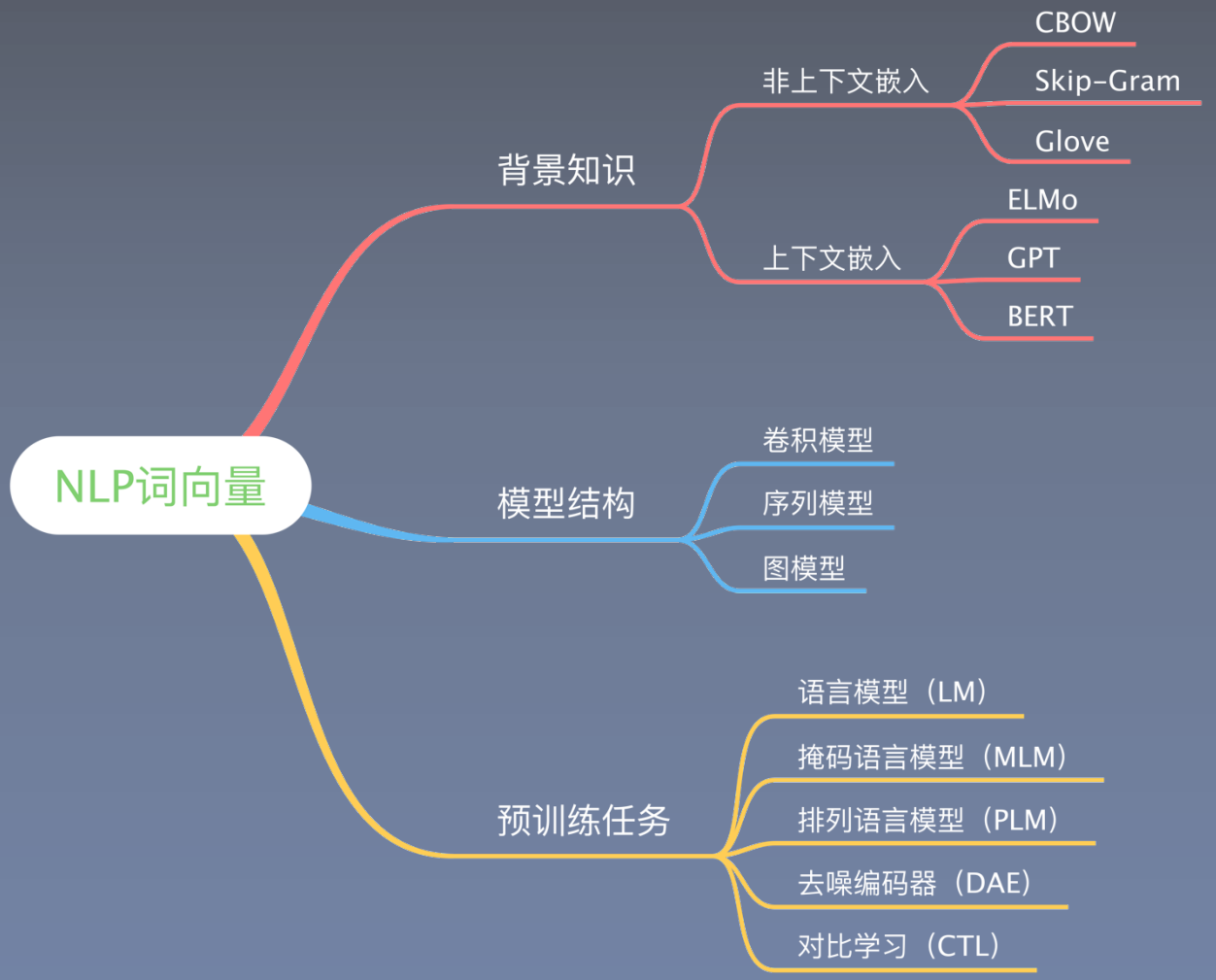
- ✓ 语言模型：预测词上下文或者自身；
- ✓ 掩码语言模型：预测被遮掉的词语；
- ✓ 对比学习：Replaced Token Detection、Next Sentence Prediction、Sentence Order Prediction

1 词向量介绍

Introduction to Embedding

词向量发展：

- ✓ 面向下游任务的预训练；
- ✓ 更好的知识迁移；
- ✓ 模型压缩；



1 词向量介绍

Introduction to Embedding

BERT原始论文：https://arxiv.org/abs/1810.04805

语言模型发展：https://arxiv.org/pdf/2003.08271.pdf

BERT相关论文：http://github.com/tomohideshibata/BERT-related-papers

深度之眼
deepshare.net

2 文本分类细节

Tricks in Text Classification

2 文本分类细节

Tricks in Text Classification

细节1：尽可能的使用到现有预料知识；

- ✓ Bert不做预训练，直接进行训练？
- ✓ 训练集和测试集一起加入pretrain；

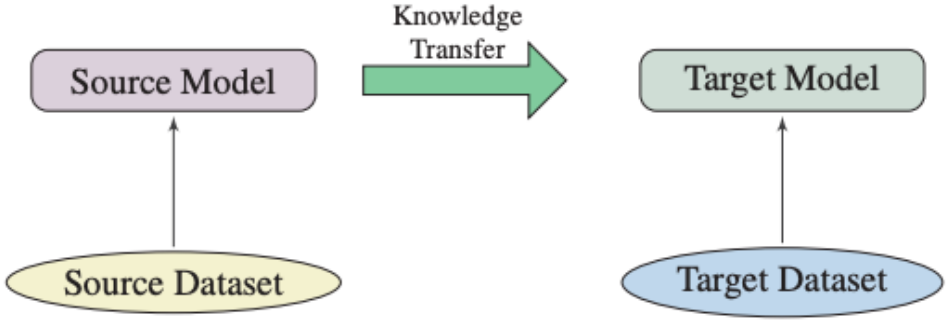


Figure 4: Transfer Learning

2 文本分类细节

Tricks in Text Classification

细节2：长语句如何解决；

✓ 训练集和测试集如何进行语句拆分？

pretrain阶段 vs finetune阶段

Method	IMDb	Sogou
head-only	5.63	2.58
tail-only	5.44	3.17
head+tail	**5.42**	**2.43**
hier. mean	5.89	2.83
hier. max	5.71	2.47
hier. self-attention	5.49	2.65

Table 2: Test error rates (%) on IMDb and Chinese Sogou News datasets.

2 文本分类细节

Tricks in Text Classification

细节3：Pooling层选择；

✓ Mean Pooling、Max Pooling、[Mean-Max] Pooling

细节4：Snapshot模型权重；

✓ pretrain和finetune不同step的权重，构建得到多个模型；

2 文本分类细节

Tricks in Text Classification

细节5：伪标签加入测试集A进行训练；

✓ 准确率大于90%的场景，都可以尝试伪标签；

✓ 尝试分析下TFIDF、Bert在不同类别的准确率；

细节6：Word2Vec / Glove + BILSTM，可以达到0.96分数；

✓ Bert-mini可能是更好的选择；

✓ 多个词向量如何共同构建模型？🤔

深度之眼
deepshare.net

3 内容检索

Content retrieval

3 内容检索

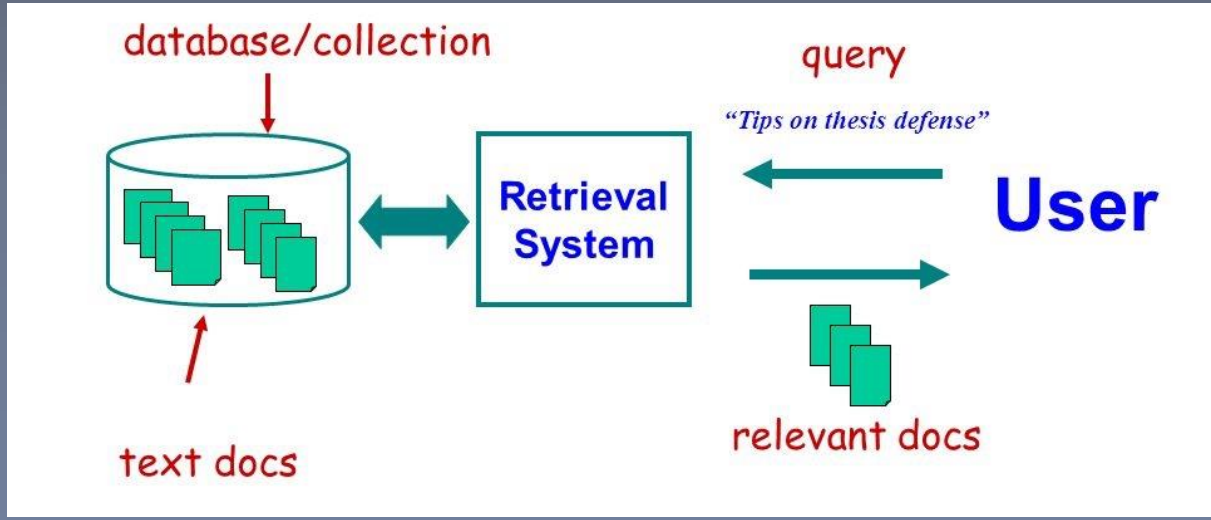
Content retrieval

文本检索：

✓ 将文本进行向量化，进行距离计算；

图像检索：

✓ 将图像进行向量化，进行距离计算；



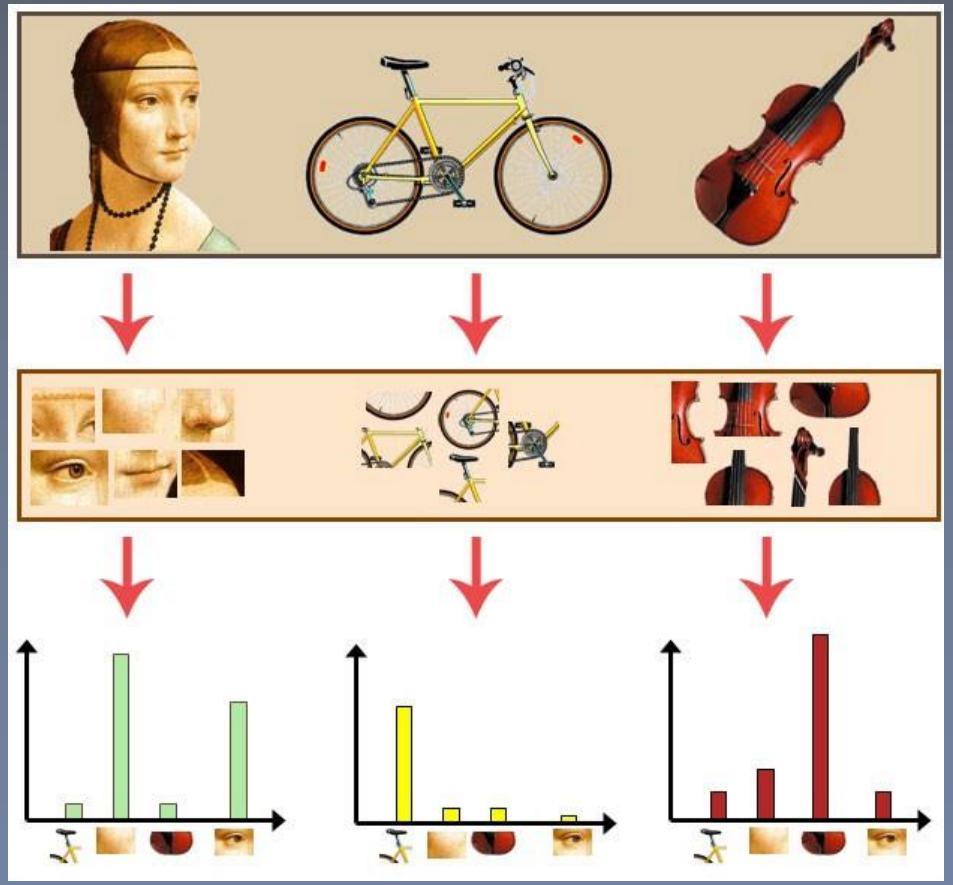
3 内容检索

Content retrieval

文本检索：TFIDF、BM25、SIF

✓ 局部词袋特征 vs 全局特征

图像检索：全局特征 vs 局部特征

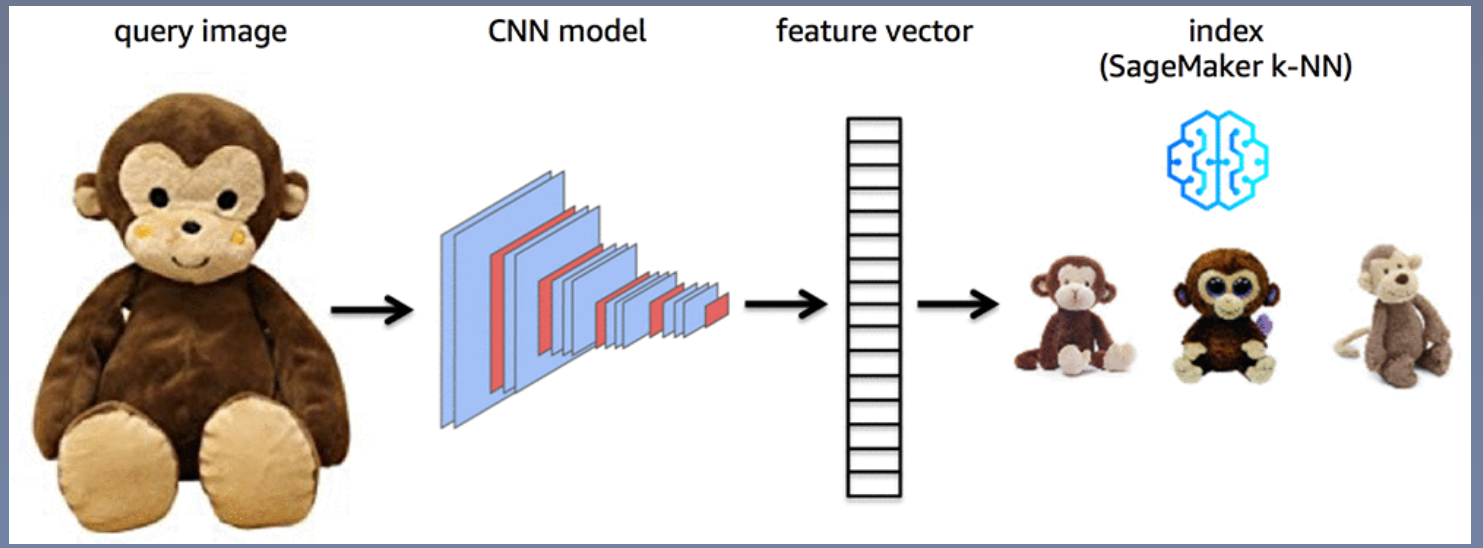


3 内容检索

Content retrieval

CNN网络也是一种Embedding方式：

- ✓ 好的预训练模型；
- ✓ 合理的特征提取方法；



深度之眼
deepshare.net

4 竞赛中的词向量

Embedding in Competition

4 竞赛中的词向量

Embedding in Competition

案例比赛：DC基于人工智能药物分筛选

比赛数据：分子蛋白质序列；

操作思路：Word2Vec + LGB；

4 竞赛中的词向量

Embedding in Competition

案例比赛：阿里云安全赛；

比赛数据：病毒API序列；

操作思路：Word2Vec + LGB；

4 竞赛中的词向量

Embedding in Competition

案例比赛：DCIC智慧海洋建设；

比赛数据：渔船经纬度；

操作思路：经纬度统计 + Word2Vec + LGB；

将渔船统计信息，通过分位点压缩到序列；

https://mp.weixin.qq.com/s/TvO2EPGJu04JvhX67uQsHw

4 竞赛中的词向量

Embedding in Competition

案例比赛：易观用户性别年龄预测；

比赛数据：用户基本信息+APP使用序列；

操作思路：APP序列的Word2Vec + TextCNN；

https://github.com/chizhu/yiguan_sex_age_predict_1st_solution

4 竞赛中的词向量

Embedding in Competition

案例比赛：融360金融算法赛；

比赛数据：用户APP信息+人脉关系；

操作思路：APP统计 + 人脉图Embedding

https://github.com/xSupervisedLearning/Rong360_feature_mining_1st_solution

4 竞赛中的词向量

Embedding in Competition

案例比赛：KDD2020多模态检索；

比赛数据：商品文本描述+图像特征；

操作思路：构建多模态Bert匹配模型；

https://github.com/steven95421/KDD_WinnieTheBest

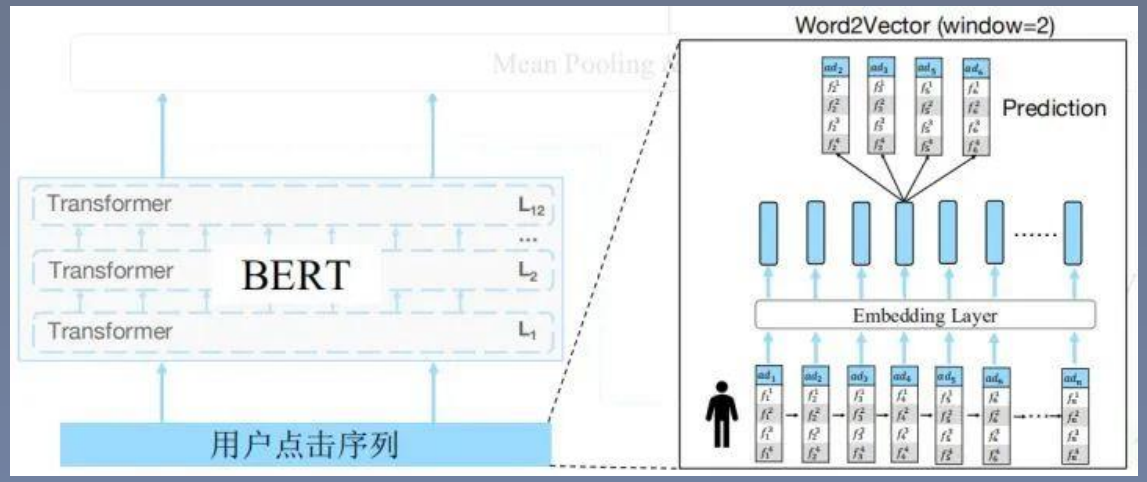
4 竞赛中的词向量

Embedding in Competition

案例比赛：腾讯广告算法大赛；

比赛数据：用户基本信息+用户广告点击序列；

操作思路：广告序列的Bert建模；



https://mp.weixin.qq.com/s/-lizDyP2y357plcG1M64TA

深度之眼
deepshare.net

5 Q&A

Ask me anything

Q&A

Ask me anything

1、万物皆可Embedding，图嵌入是未来的方向；

2、在具体问题中，预训练方法（如mask）需要与具体的问题对应；

3、掌握词向量训练方法，用途非常多；

请让我们一起立一个flag！

我承诺：

4周努力上TOP100！

结语

再小的细节，也值得被认真对待

深度之眼
deepshare.net

联系我们：

电话：18001992849

邮箱：service@deepshare.net

Q Q： 2677693114

公众号

客服微信