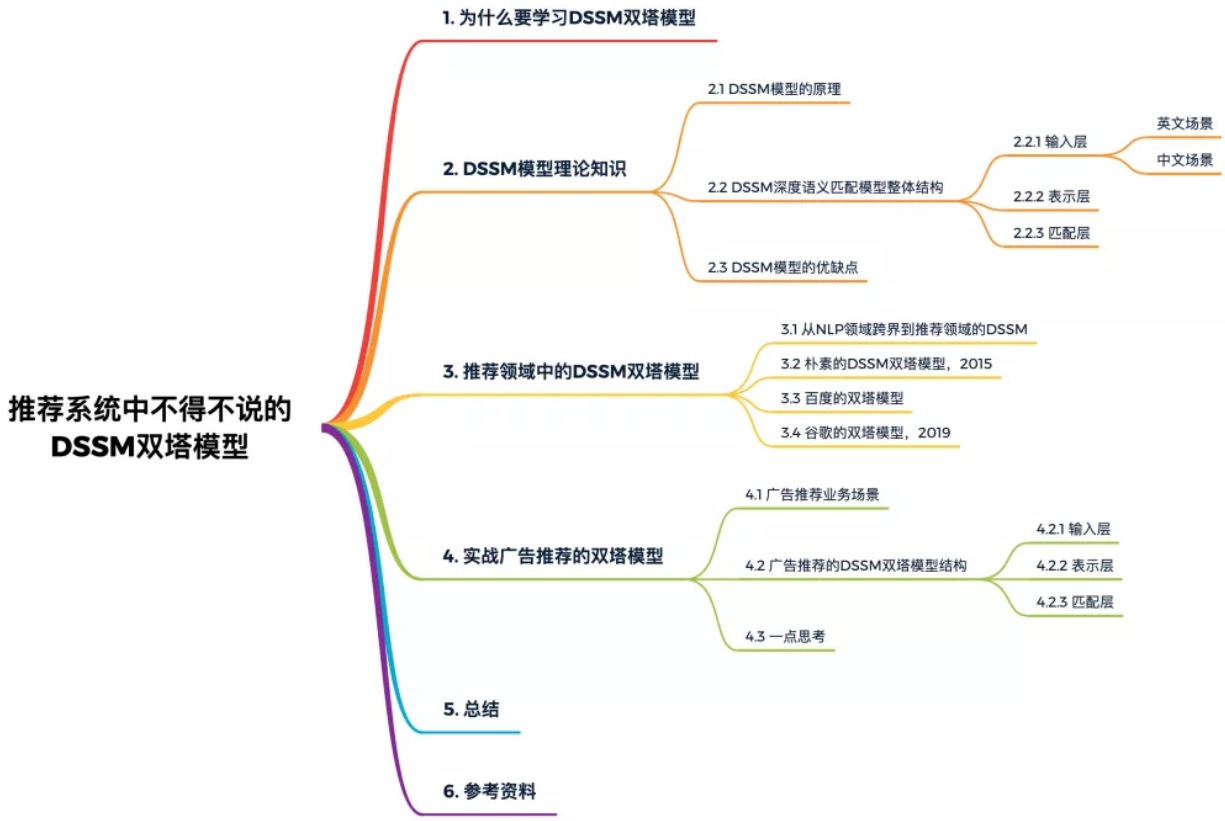
推荐系统中不得不说的DSSM双塔模型

数据拾光者 Microstrong 2020-11-13

近日读到一篇非常不错的文章，忍不住分享给大家，同时也方便自己日后学习查阅。Microstrong为这篇文章写了一段推荐语：首先，详细讲解了最早在NLP领域中用于语义相似度任务的DSSM语义匹配模型的理论知识，并分析了该模型的优缺点；然后，由于都是排序问题，进而引入该模型到推荐领域，并概述了从朴素的DSSM双塔模型到各大厂的双塔模型；最后，分享了作者使用DSSM双塔模型实战到广告推荐场景的案例。

本文在原文的基础上，添加了相关论文的引用，并为了提高阅读性，对文章排版稍有修改。

本文概览：



本文主要介绍项目中用于商业兴趣建模的DSSM双塔模型。作为推荐领域中大火的双塔模型，因为效果不错并且对工业界十分友好，所以被各大厂广泛应用于推荐系统中。

通过构建user和item两个独立的子网络，将训练好的两个“塔”中的user embedding 和item embedding各自缓存到内存数据库中。线上预测的时候只需要在内存中计算相似度运算即可。DSSM双塔模型是推荐领域中不得不会的重要模型。

1. 为什么要学习DSSM双塔模型

我们标签组主要的服务对象是广告主，服务目标是为广告主提供更好的广告转换效果。这里涉及到两种建模：

- 一种是自然兴趣建模，根据用户操作终端行为获得user-item关联，给不同的数据源打标获得item-tag关联，最后将上面两种关联进行join操作得到user-tag的关联实现给用户打上兴趣标签，这里相当于是从标签维度为广告主推荐人群；
- 另一种就是商业兴趣建模，在自然兴趣建模的基础上，从广告维度为广告主推荐人群，那么就需要目前大火的DSSM双塔模型了。

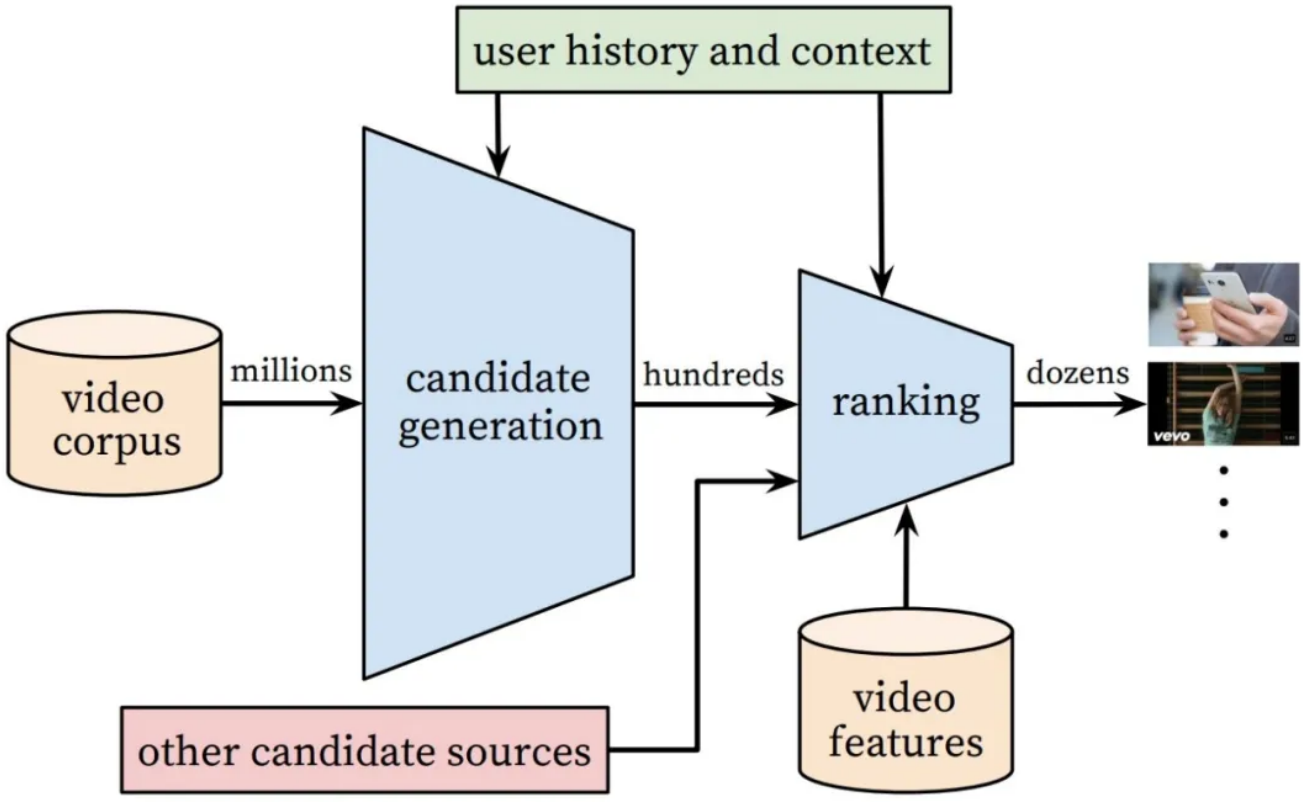


图1 YouTube的推荐系统架构图

拿YouTube视频推荐系统举例，一般推荐系统中有两个流程：

- 第一步是召回模型，主要是进行初筛操作，从海量视频资源池中初步选择一部分用户可能感兴趣的视频数据子集，从数量上看可能是从千万级别筛选出百级别；
- 第二步是精排模型，主要作用是对上面找到的百级别的视频子集进一步精筛，从数量上看可能是从百级别筛选出几十级别。然后根据得分高低排序，生成一个排序列表作为用户的候选播放列表从而完成视频推荐任务。

我们广告推荐领域中使用的DSSM双塔模型是从广告维度为广告主推荐一定数量的人群，从数量上看是从百亿级别人群中找出百万级人群用于投放广告，所以是召回模型。

【相关论文】

- YouTube推荐经典论文：Covington P , Adams J , Sargin E . Deep Neural Networks for YouTube Recommendations[C]// Acm Conference on Recommender Systems. ACM, 2016:191-198.

2. DSSM模型理论知识

2.1 DSSM模型的原理

DSSM(Deep Structured Semantic Models)也叫深度语义匹配模型，最早是微软发表的一篇应用于NLP领域中计算语义相似度任务的文章。

DSSM深度语义匹配模型原理很简单：获取搜索引擎中的用户搜索query和doc的海量曝光和点击日志数据，训练阶段分别用复杂的深度学习网络构建query侧特征的query embedding和doc侧特征的doc embedding，线上infer时通过计算两个语义向量的cos距离来表示语义相似度，最终获得语义相似模型。**这个模型既可以获得语句的低维语义向量表达sentence embedding，还可以预测两句话的语义相似度。**

【相关论文】

- 微软发表的DSSM论文：Huang P S , He X , Gao J , et al. Learning deep structured semantic models for web search using clickthrough data[C]// Proceedings of the

22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

2.2 DSSM深度语义匹配模型整体结构

DSSM模型总的来说可以分成三层结构，分别是输入层、表示层和匹配层。结构如下图所示：

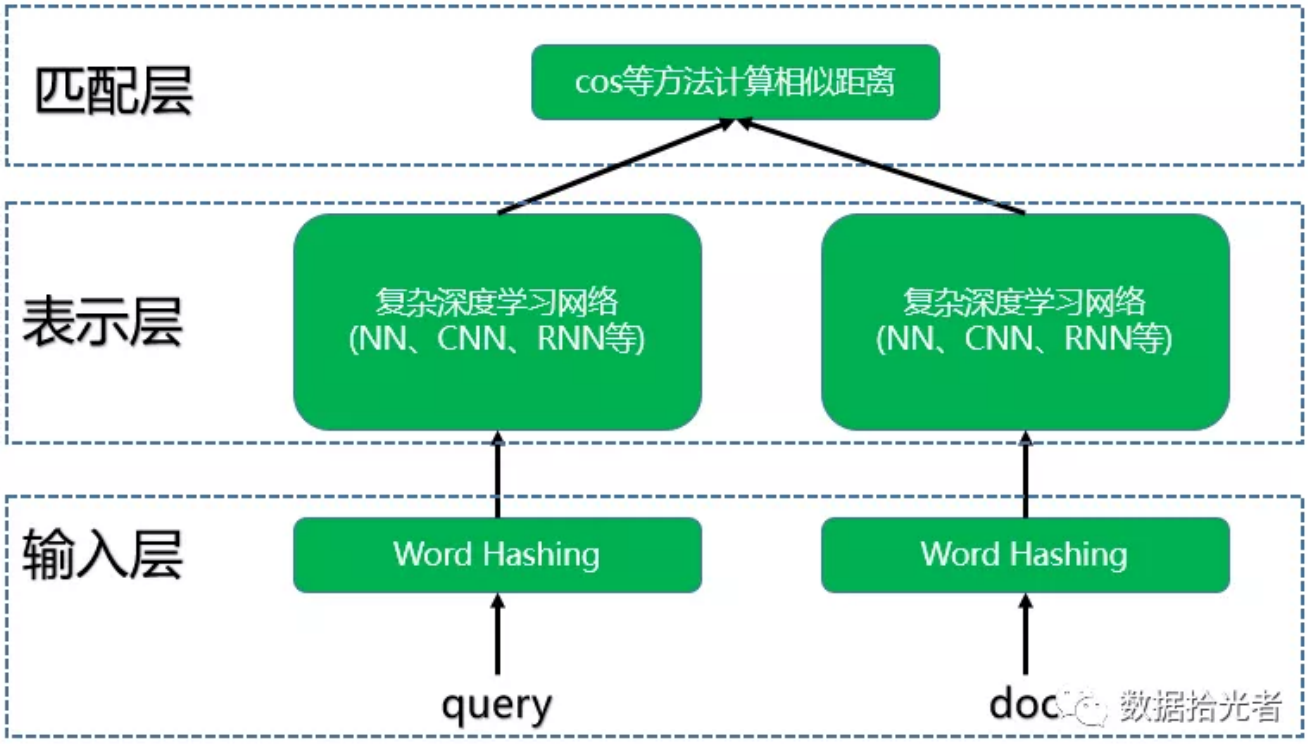


图2 DSSM模型结构图

2.2.1 输入层

输入层主要的作用就是把文本映射到低维向量空间转化成向量提供给深度学习网络。NLP领域里中英文有比较大的差异，在输入层处理方式不同。

(1) 英文场景

英文的输入层通过Word Hashing方式处理，该方法基于字母的n-gram，主要作用是减少输入向量的维度。举例说明，假如现在有个词boy，开始和结束字符分别用#表示，那么输入

就是(#boy#)。将词转化为字母n-gram的形式，如果设置n为3，那么就能得到(#bo,boy,oy#)三组数据，将这三组数据用n-gram的向量来表示。

使用Word Hashing方法存在的问题是可能造成冲突。因为两个不同的词可能有相同的n-gram向量表示。下图是在不同的英语词典中分别使用2-gram和3-gram进行Word Hashing时的向量空间以及词语碰撞统计：

	Letter-Bigram		Letter-Trigram	
Word Size	Token Size	Collision	Token Size	Collision
40k	1107	18	10306	2
500k	1607	1192	30621	22

图3 不同词典下n-gram向量空间和词语碰撞统计

可以看出在50W词的词典中如果使用2-gram，也就是两个字母的粒度来切分词，向量空间压缩到1600维，产生冲突的词有1192个(这里的冲突是指两个词的向量表示完全相同，因为单词储量实在有限，本来想找几个例子说明下，结果没找到)。如果使用3-gram向量空间压缩到3W维，产生冲突的词只有22个。综合下来论文中使用3-gram切分词。

(2) 中文场景

中文输入层和英文有很大差别，首先要面临的是分词问题。如果要分词推荐jieba或者北大pkuseg，不过现在很多模型已经不进行分词了，比如BERT中文的预训练模型就直接使用单字作为最小粒度了。

2.2.2 表示层

DSSM模型表示层使用的是BOW(bag of words)词袋模型，没有考虑词序的信息。不考虑词序其实存在明显的问题，因为一句话可能词相同，但是语义则相差十万八千里，比如“我爱女朋友”和“女朋友爱我”可能差距蛮大的(这个小伙伴们自己体会)。

下图是DSSM表示层的结构：

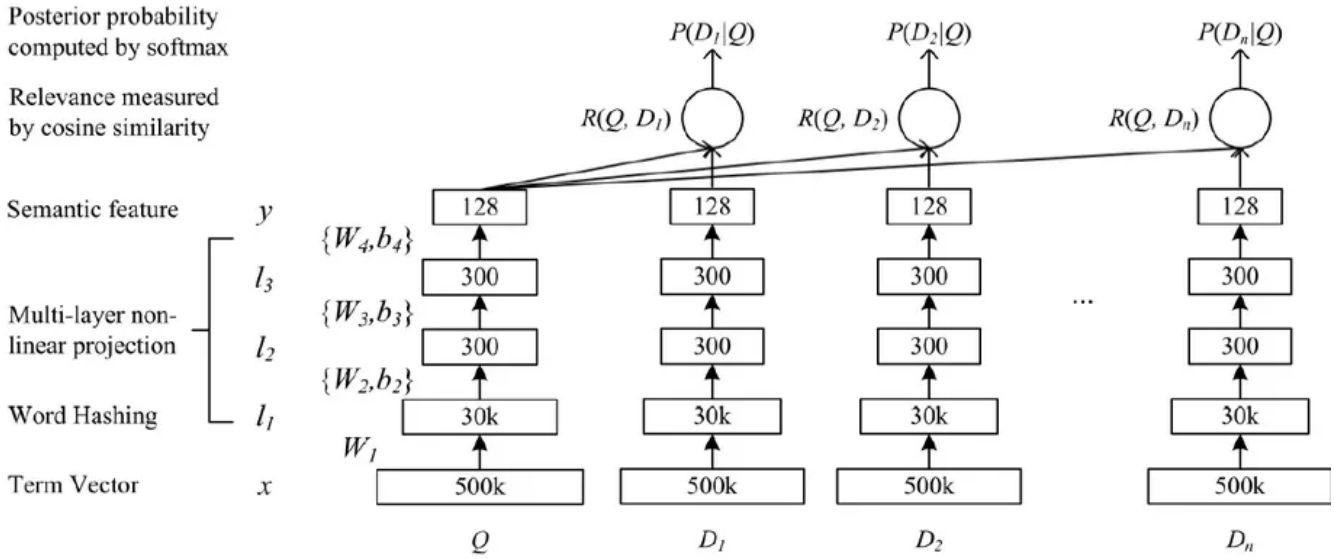


图4 DSSM模型表示层结构图

最下面的Term Vector到Word Hashing将词映射到3W维的向量空间中。然后分别经过两层300维度的隐藏层，最后统一输出128维度的向量。

2.2.3 匹配层

现在我们把query和doc统一转换成了两个128维的语义向量，接下来如何计算它们的语义相似度呢？通过cos函数计算这两个向量的余弦相似度就可以了，公式如下：

$$R(Q,D) = cosine(y_Q, y_D) = \frac{y_Q^T y_D}{||y_Q|| ||y_D||}$$

2.3 DSSM模型的优缺点

先说说**DSSM**模型的优点：

- 解决了LSA、LDA、Autoencoder等方法存在的字典爆炸问题，从而降低了计算复杂度。因为英文中词的数量要远远高于字母n-gram的数量；
- 中文方面使用字作为最细切分粒度，可以复用每个字表达的语义，减少分词的依赖，从而提高模型的泛化能力；
- 字母的n-gram可以更好的处理新词，具有较强的鲁棒性；
- 使用有监督的方法，优化语义embedding的映射问题；
- 省去了人工特征工程；

- 采用有监督训练，精度较高。传统的输入层使用embedding的方式(比如Word2vec的词向量)或者主题模型的方式(比如LDA的主题向量)做词映射，再把各个词的向量拼接或者累加起来。由于Word2vec和LDA都是无监督训练，会给模型引入误差。

再说说**DSSM**模型的缺点：

- Word Hashing可能造成词语冲突；
- 采用词袋模型，损失了上下文语序信息。这也是后面会有CNN-DSSM、LSTM-DSSM等DSSM模型变种的原因；
- 搜索引擎的排序由多种因素决定，用户点击时doc排名越靠前越容易被点击，仅用点击来判断正负样本，产生的噪声较大，模型难以收敛；
- 效果不可控。因为是端到端模型，好处是省去了人工特征工程，但是也带来了端到端模型效果不可控的问题。

3. 推荐领域中的DSSM双塔模型

3.1 从NLP领域跨界到推荐领域的DSSM

DSSM深度语义匹配模型最早是应用于NLP领域中计算语义相似度任务。因为语义匹配本身是一种排序问题，和推荐场景不谋而合，所以DSSM模型被自然的引入到推荐领域中。**DSSM模型分别使用相对独立的两个复杂网络构建用户相关特征的user embedding和item相关特征的item embedding，所以称为双塔模型。**

3.2 朴素的DSSM双塔模型，2015

双塔模型最大的特点是user和item是独立的两个子网络，对工业界十分友好。将两个塔各自缓存，线上预测的时候只需要在内存中进行相似度运算即可。下面是2015年朴素的DSSM双塔模型结构：

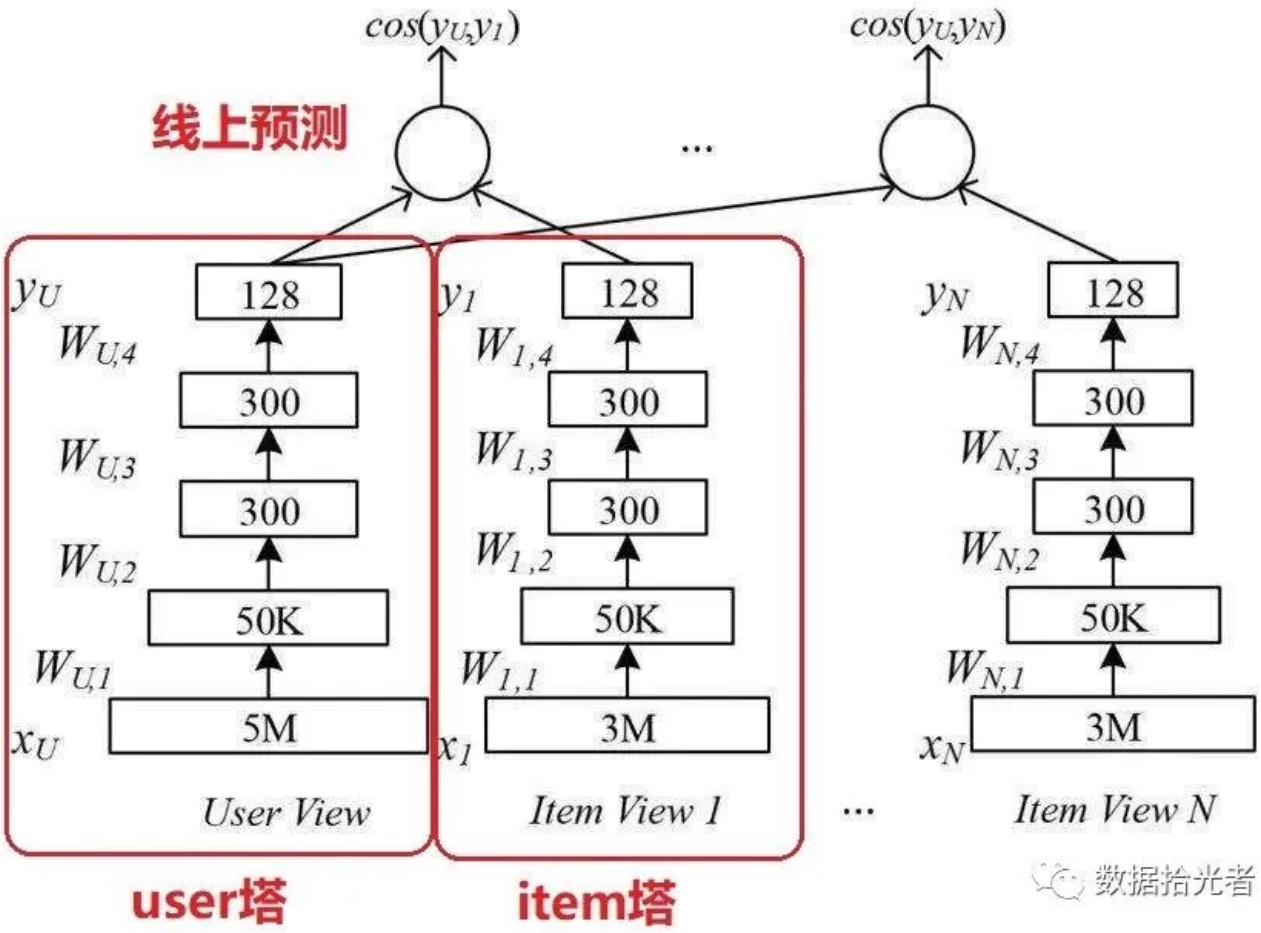


图5 朴素的DSSM双塔模型

3.3 百度的双塔模型

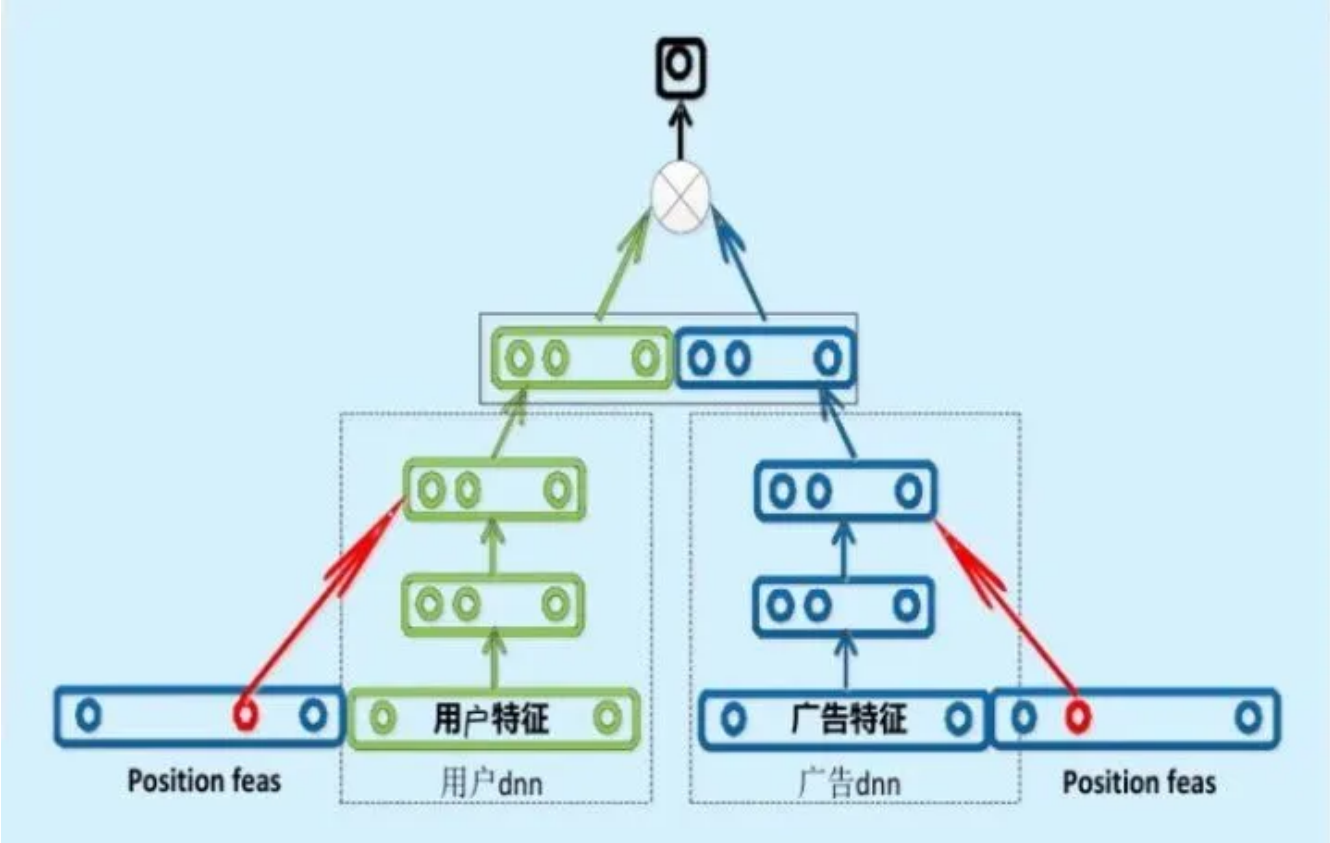


图6 百度的双塔模型

百度的双塔模型分别使用复杂的网络对用户相关的特征和广告相关的特征进行embedding，分别形成两个独立的塔，在最后的交叉层之前用户特征和广告特征之间没有任何交互。这种方案就是训练时引入更多的特征完成复杂网络离线训练，然后将得到的user embedding和item embedding存入redis这一类内存数据库中。线上预测时使用LR、浅层NN等轻量级模型或者更方便的相似距离计算方式。这也是业界很多大厂采用的推荐系统的构造方式。

3.4 谷歌的双塔模型，2019

2019年谷歌推出自己的双塔模型，文章的核心思想是：在大规模的推荐系统中，利用双塔模型对user-item对的交互关系进行建模，从而学习【用户，上下文】向量和【item】向量的关联。针对大规模流数据，提出in-batch softmax损失函数与流数据频率估计方法更好的适应item的多种数据分布。利用双塔模型构建Youtube视频推荐系统，对于用户侧的塔根据用户观看视频特征构建user embedding，对于视频侧的塔根据视频特征构建video emebdding。两个塔分别是相互独立的网络。

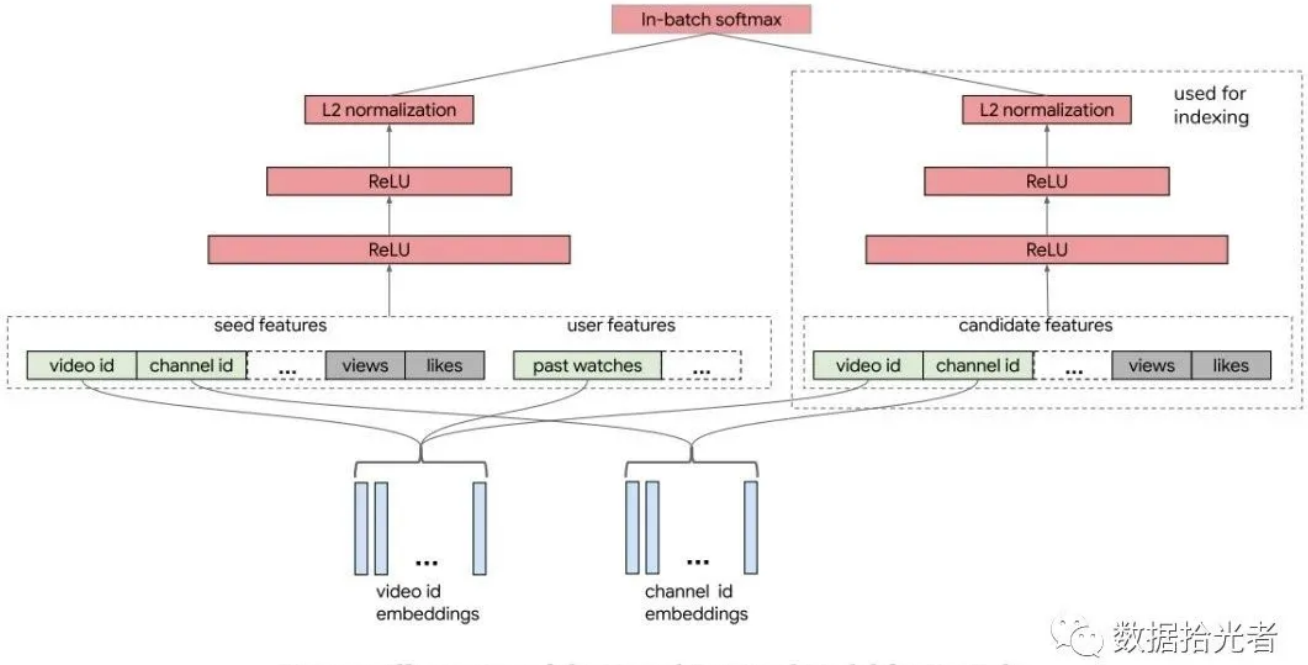


Figure 2: Illustration of the Neural Retrieval Model for YouTube.

图7 谷歌的双塔模型

【相关论文】

- 谷歌双塔模型：Yi X , Yang J , Hong L , et al. Sampling-bias-corrected neural modeling for large corpus item recommendations[C]// the 13th ACM Conference. ACM, 2019.

4. 实战广告推荐的双塔模型

4.1 广告推荐业务场景

讲了上面一大堆，就是为了这一节构建咱们广告推荐的DSSM双塔模型。对应到咱们的广告业务就是构建DSSM双塔模型，用户侧输入用户对广告的历史行为特征(包括点击、下载、付费等)从而得到固定长度的user embedding，同理广告侧输入广告特征得到相同长度的ad embedding，分别存入redis内存数据库中。

线上infer时给定一个广告ad，然后分别和全量用户求相似度，找到“距离最近”的user子集，对这部分人群投放广告从而完成广告推荐任务。

4.2 广告推荐的DSSM双塔模型结构

模型整体结构如下图所示，也分成三层：输入层、表示层和匹配层。

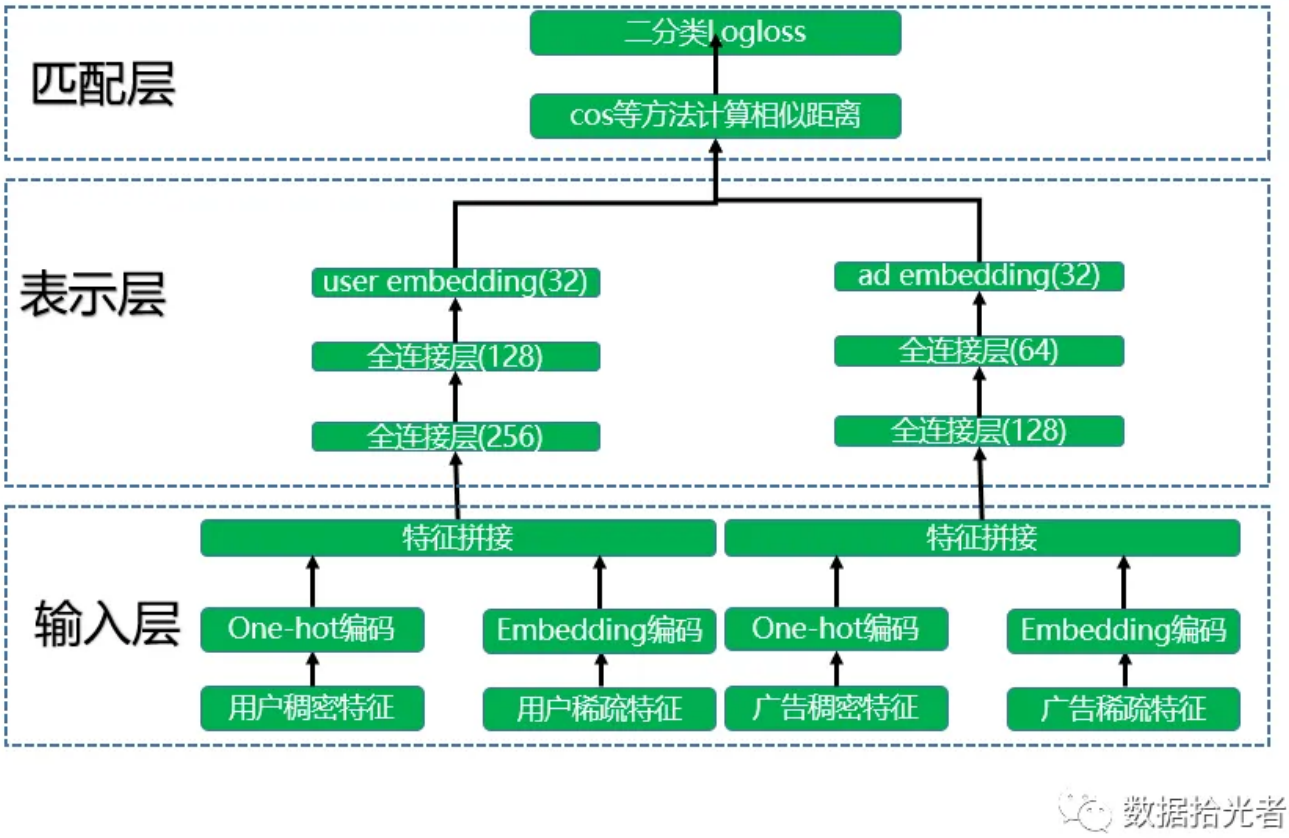


图8 广告推荐的DSSM双塔模型结构

4.2.1 输入层

模型训练分成两座不同的“塔”分别进行，其实也就是两个不同的神经网络。其中一座塔是用于生成user embedding。输入用户特征训练数据，用户特征包括用户稠密特征和用户稀疏特征，其中用户稠密特征进行one-hot编码操作，用户稀疏特征进行embedding降维到低维空间(64或者32维)，然后进行特征拼接操作。广告侧和用户侧类似。

关于里面的特征，不在于你要什么，而在于你有什么。整个工程超级复杂的就是这块的特征工作。这里不再赘述。

4.2.2 表示层

得到拼接好的特征之后会提供给各自的深度学习网络模型。用户特征和广告特征经过各自的两个全连接层后转化成了固定长度的向量，这里得到了维度相同的user embedding和ad embedding。各塔内部的网络层数和维度可以不同，但是输出的维度必须是一样的，这样才能在匹配层进行运算。项目中user embedding和ad embedding 维度都是32。

4.2.3 匹配层

模型训练好了之后会分别得到user embedding和ad embedding，将它们存储到redis这一类内存数据库中。如果要为某个特定的广告推荐人群，则将该广告的ad embedding分别和所有人群的user embedding计算cos相似度。选择距离最近的N个人群子集作为广告投放人群，这样就完成了广告推荐任务。模型训练过程中将cos函数得到的结果进入sigmoid函数和真实标签计算logloss，查看网络是否收敛。模型评估主要使用auc指标。

小结下，本节讲了下我们使用DSSM双塔模型完成广告推荐任务。模型整体结构分成输入层、表示层和匹配层。首先在输入层处理数据获取特征；然后在表示层通过深度学习网络得到user embedding和ad embedding；最后在匹配层进行广告推荐。

4.3 一点思考

DSSM双塔模型有很多变种，比如CNN-DSSM、LSTM-DSSM等等。项目中表示层使用了两层全连接网络来作为特征抽取器。现在深度学习领域公认最强的特征抽取器是Transformer，后续是否可以加入Transformer。

5. 总结

本篇主要介绍了项目中用于商业兴趣建模的DSSM双塔模型。作为推荐领域中大火的双塔模型，最大的特点是效果不错并且对工业界十分友好，所以被各大厂广泛应用于推荐系统中。

通过构建user和item两个独立的子网络，将训练好的两个塔中的user embedding 和item embedding各自缓存到内存数据库中。线上预测的时候只需要在内存中进行相似度运算即可。

首先，讲了下DSSM语义匹配模型的理论知识，最早是应用于NLP领域中用于语义相似度任务；然后，因为都是排序问题，所以引入到推荐领域。从朴素的DSSM双塔模型到各大厂的双塔模型；最后，讲了下我们使用DSSM双塔模型实战到广告推荐场景。

6. 参考资料

【1】LearningDeep Structured Semantic Models for Web Search using Clickthrough Data
【2】Sampling-bias-corrected neural modeling for largecorpus item recommendations

最后给出该文作者的微信公众号：数据拾光者。