是"塔"！是"塔"!就是它，我们的双塔!

原创 十方 炼丹笔记 2020-12-23

收录于话题
#搜索推荐前沿算法

65个

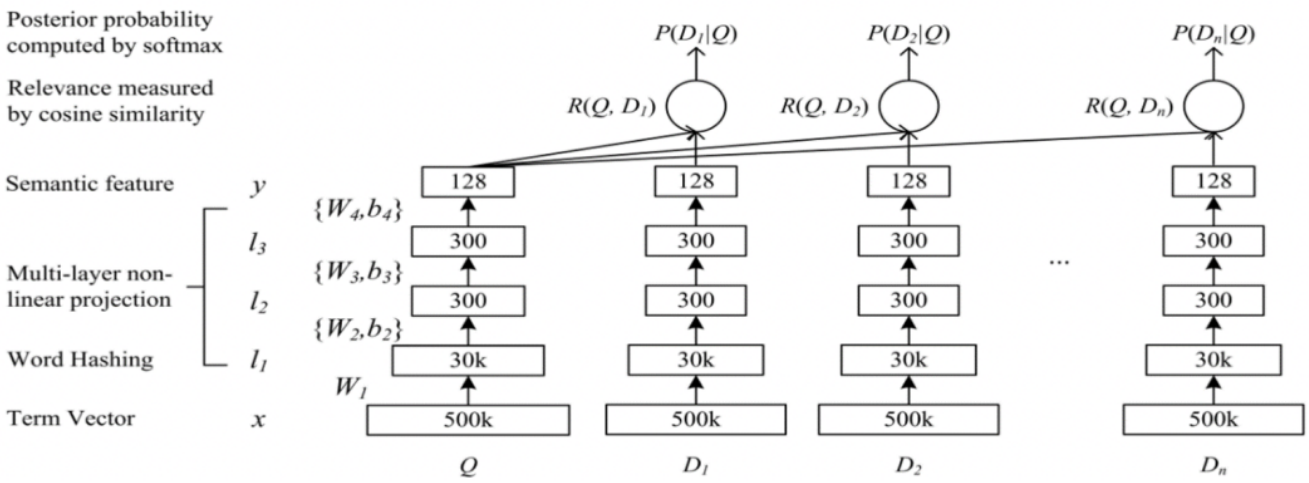
↑↑↑关注后"星标"**炼丹笔记**

炼丹笔记日常

作者：十方，一品炼丹师

对于基于向量召回，那就不得不提到双塔。为什么双塔在工业界这么常用？双塔上线有多方便，真的是谁用谁知道，user塔做在线serving，item塔离线计算embeding建索引，推到线上即可。**下面我就给大家介绍一些经典的双塔模型，快速带大家过一遍，如果想了解细节，强烈建议看论文。**

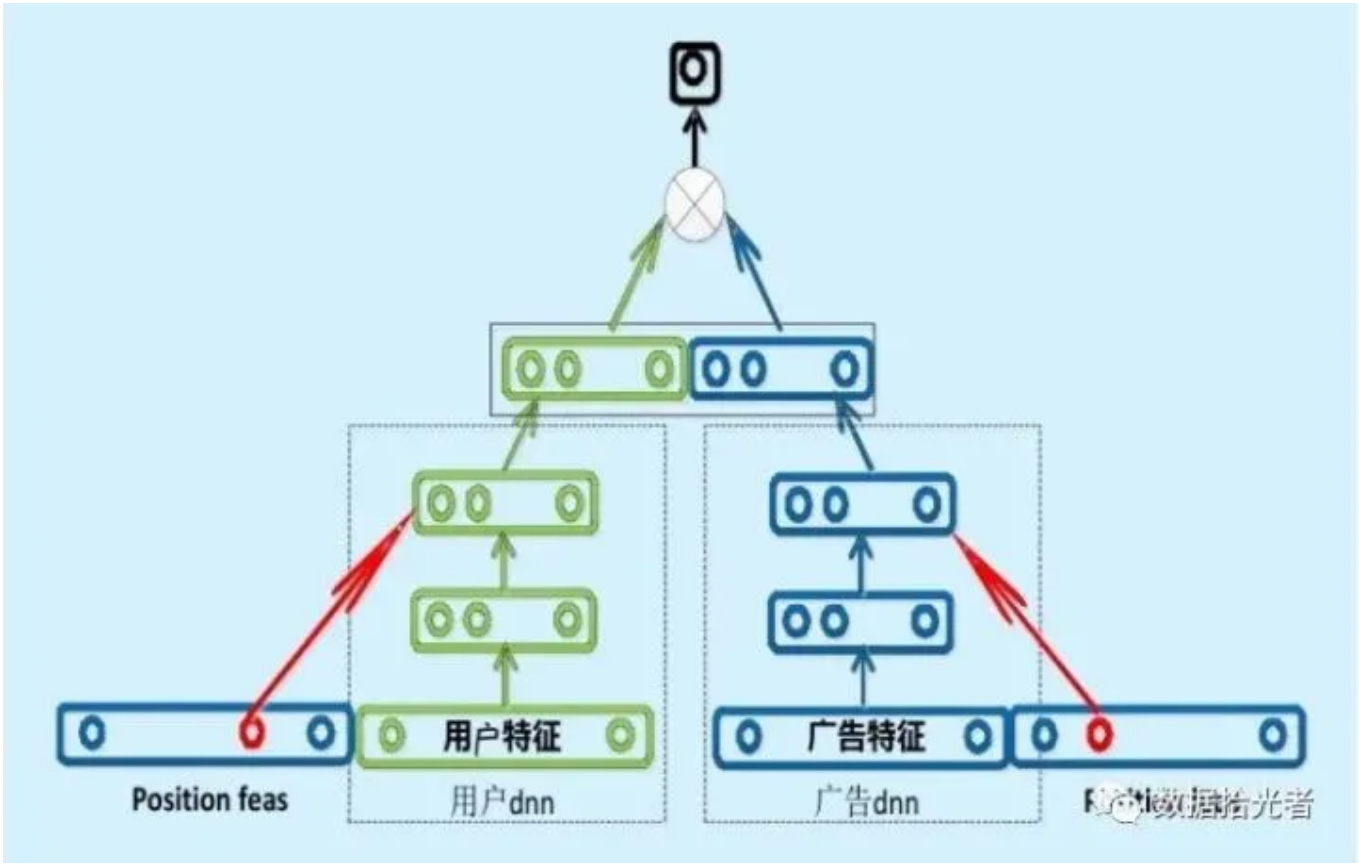
DSSM



先说双塔模型的鼻祖，这是微软在CIKM2013发表的一篇工作，它主要是用来解决NLP领域语义相似度任务的。word hashing真的是DSSM的骚操作了，不同于现有的RNN，Bert等模型，该方法直接把文本映射成了远低于vocab size的向量中，然后输入DNN，输出得到一个128维的低维语义向量。Query和document的语义相似度就可以用这两个向量的cosine相似度来表示，进一步我们可以通过softmax对不同的document做排序。这就是最初的DSSM。如果把把document换成item或是广告，就演变成了一个推荐模型。

论文地址：https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/cikm2013_DSSM_fullversion.pdf

看下百度双塔，左边user，右边广告，是不是很符合百度价值观，简单可依赖：

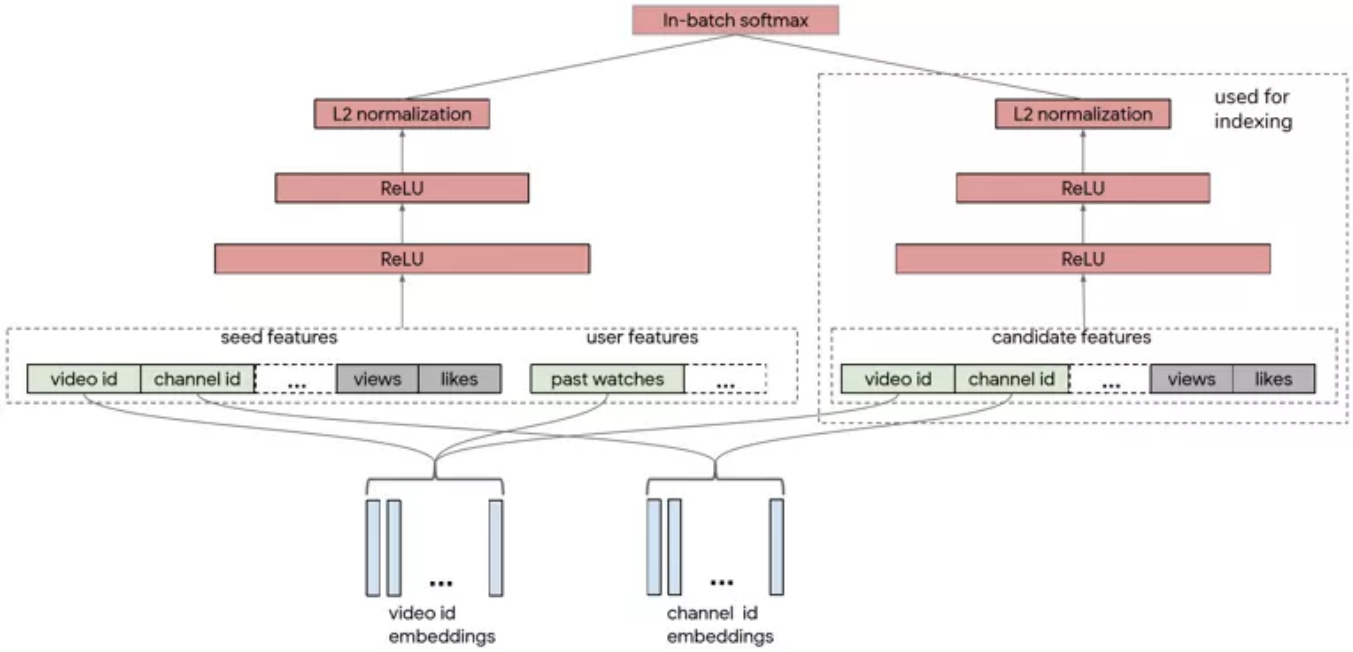


双塔模型的优势，总结如下:

- 可以离线计算item的embedding
- 线上计算user(&context)的embedding
- 线上计算相似度
- 实时性好

说来说去，主要就是实时性好，cos的表达是有限的，很难提取交叉特征，所以双塔还是比较适用于召回场景。

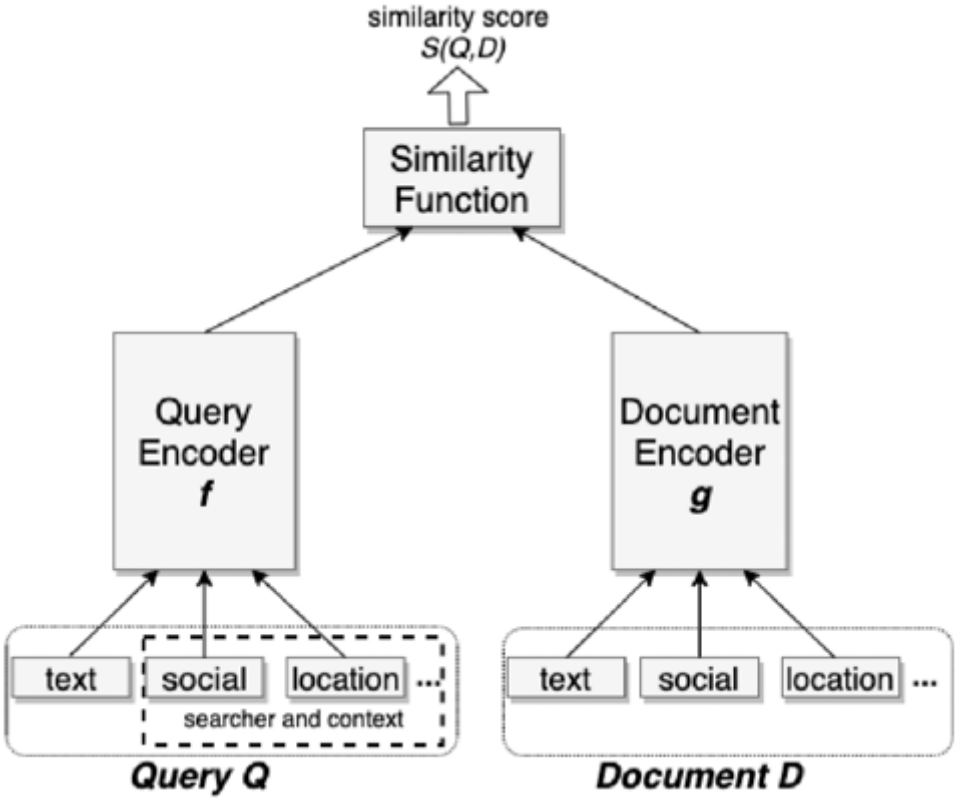
YouTube双塔



YouTube最新正在使用的视频召回双塔模型。这个模型在整体上就是最普通的双塔。左边是user塔，输入包括两部分，第一部分seed是user当前正在观看的视频，第二部分user的feature是根据user的观看历史计算的，比如说可以使用user最近观看的k条视频的id emb均值，这两部分融合起来一起输入user侧的DNN。右边是item塔，将候选视频的feature作为输入，计算item的 embedding。之后再计算相似度，做排序就可以了。YouTube这个模型最大的不同是，它的训练是基于流数据的，每一天都会产生新的训练数据。因此，负样本的选择只能在batch内进行，batch内的所有样本作为彼此的负样本去做batch softmax。这种采样的方式带来了非常大的bias。一条热门视频，它的采样概率更高，因此会更多地被当做负样本，这不符合实际。因此这篇工作的核心就是**减小batch内负采样带来的bias**。

论文地址:https://dl.acm.org/doi/pdf/10.1145/3298689.3346996

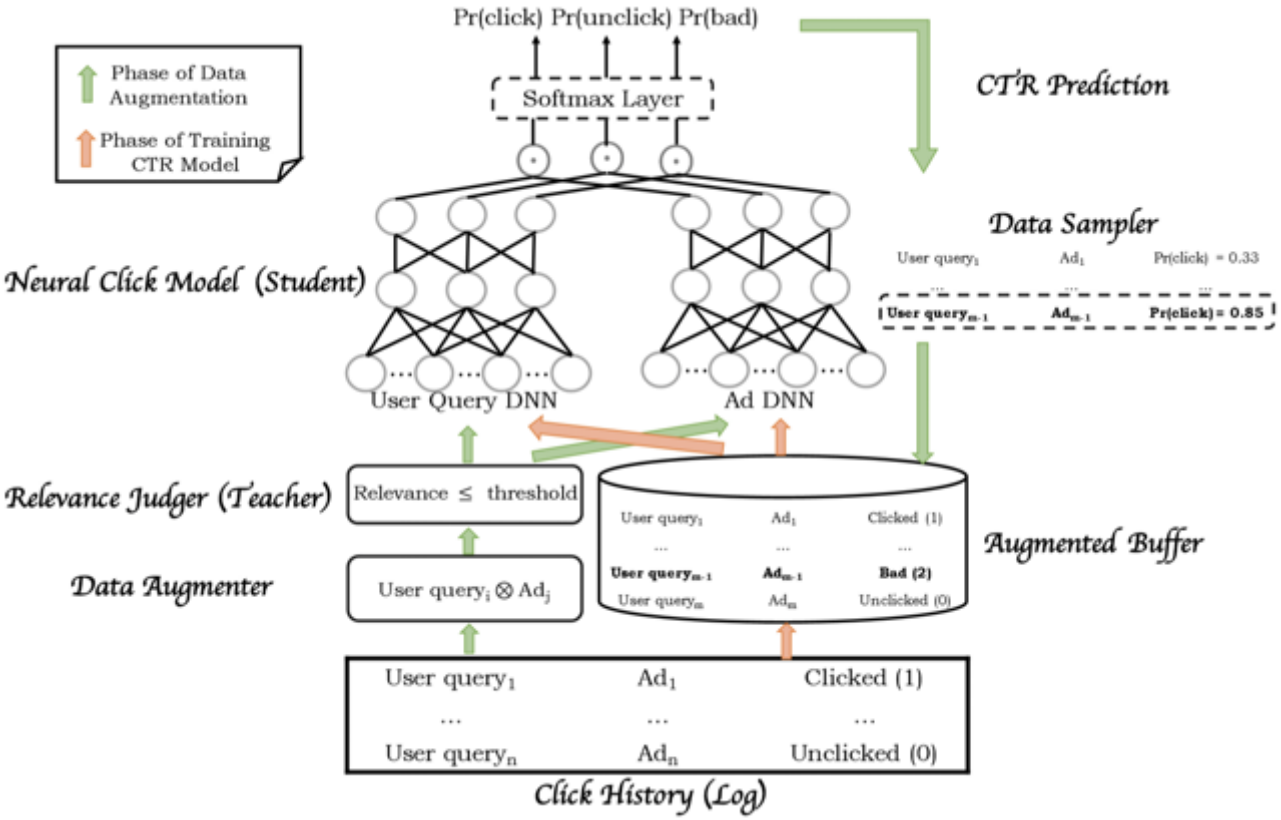
EBR (facebook kdd 2020)



这篇论文真的强推，模型结构没啥好说的，简单的双塔，两边塔的输入都是文本特征、社交特征和位置特征，其中社交特征和位置特征是他们在实验中发现对效果提升比较好的两种特征。这篇工作的**两个核心亮点**是hard negative mining和embedding ensemble。Hard negative mining是指，他们发现**如果将随机负样本这种比较easy的样本与上次召回中排名101-500名的比较hard的样本以100:1的比例去训练模型(为什么是101-500?)**，得到的效果会比较好。Embedding ensemble是指，可以将不同负样本训练得到的模型做融合来进行召回。融合的方式可以是相似度结果的直接加权或者是模型的串行融合，比如先用easy负样本训练模型进行初步的筛选，再用hard负样本训练模型进行最终的召回。另外他们还提到虽然使用unified的特征，就是输入中包含社交特征和位置特征，来进行召回效果会比较好，但是召回结果在一定程度上也会损失文本的匹配，因此也可以先通过只输入文本特征的模型来做筛选再用输入unified特征的模型来召回，这样可以保证文本的匹配。

论文地址:https://dl.acm.org/doi/pdf/10.1145/3394486.3403305

莫比乌斯 (百度 KDD2019)



百度可不止有简单可依赖的模型，也有复杂可依赖的。整个框架分为两个阶段，数据增强阶段是绿色箭头的部分，采样并利用样本中的用户请求与广告构造出更多样本，教师网络计算相似度后将低相似度的样本输入学生网络去预测CTR，通过采样的方式得到高CTR低相似度的样本存入buffer，这类样本我们称之为bad case。第二个阶段是橙色箭头表示的CTR模型训练阶段，将原先采样得到的原始样本也存入buffer，利用buffer中的三种样本去训练CTR模型。虽然百度提出了这样一种框架，但是召回和排序的直接统一在实现的过程中还是比较困难的，因为面临的候选广告集数量太大，在性能方面还是难以保证。但是Mobius的这种将商业指标提前引入召回阶段的思想是非常具有探索价值的，比如文章中提到将cosine相似度直接乘上一个商业指标作为系数，就是一个很简单的方式。

论文地址:http://research.baidu.com/Public/uploads/5d12eca098d40.pdf

大家对双塔有什么看法呢？加群讨论吧！