21｜注意力机制、兴趣演化：推荐系统如何抓住用户的心？

2020-11-27 王喆

深度学习推荐系统实战 进入课程 >

讲述：王喆
时长 15:47 大小 14.46M

你好，我是王喆。

近几年来，注意力机制、兴趣演化序列模型和强化学习，都在推荐系统领域得到了广泛的应用。它们是深度学习推荐模型的发展趋势，也是我们必须要储备的前沿知识。

作为一名算法工程师，足够的知识储备是非常重要的，因为在掌握了当下主流的深度学习模型架构（Embedding MLP 架构、Wide&Deep 和 DeepFM 等等）之后，要想再进一步提高推荐系统的效果，就需要清楚地知道业界有哪些新的思路可以借鉴，学术界有哪些新的思想可以尝试，这些都是我们取得持续成功的关键。

所以，我会用两节课的时间，带你一起学习这几种新的模型改进思路。今天我们先重点关注注意力机制和兴趣演化序列模型，下节课我们再学习强化学习。

什么是“注意力机制”？

“注意力机制”来源于人类天生的“选择性注意”的习惯。最典型的例子是用户在浏览网页时，会有选择性地注意页面的特定区域，而忽视其他区域。

比如，图 1 是 Google 对大量用户进行眼球追踪实验后，得出的页面注意力热度图。我们可以看到，用户对页面不同区域的注意力区别非常大，他们的大部分注意力就集中在左上角的几条搜索结果上。

那说了这么多，“注意力机制”对我们构建推荐模型到底有什么价值呢？

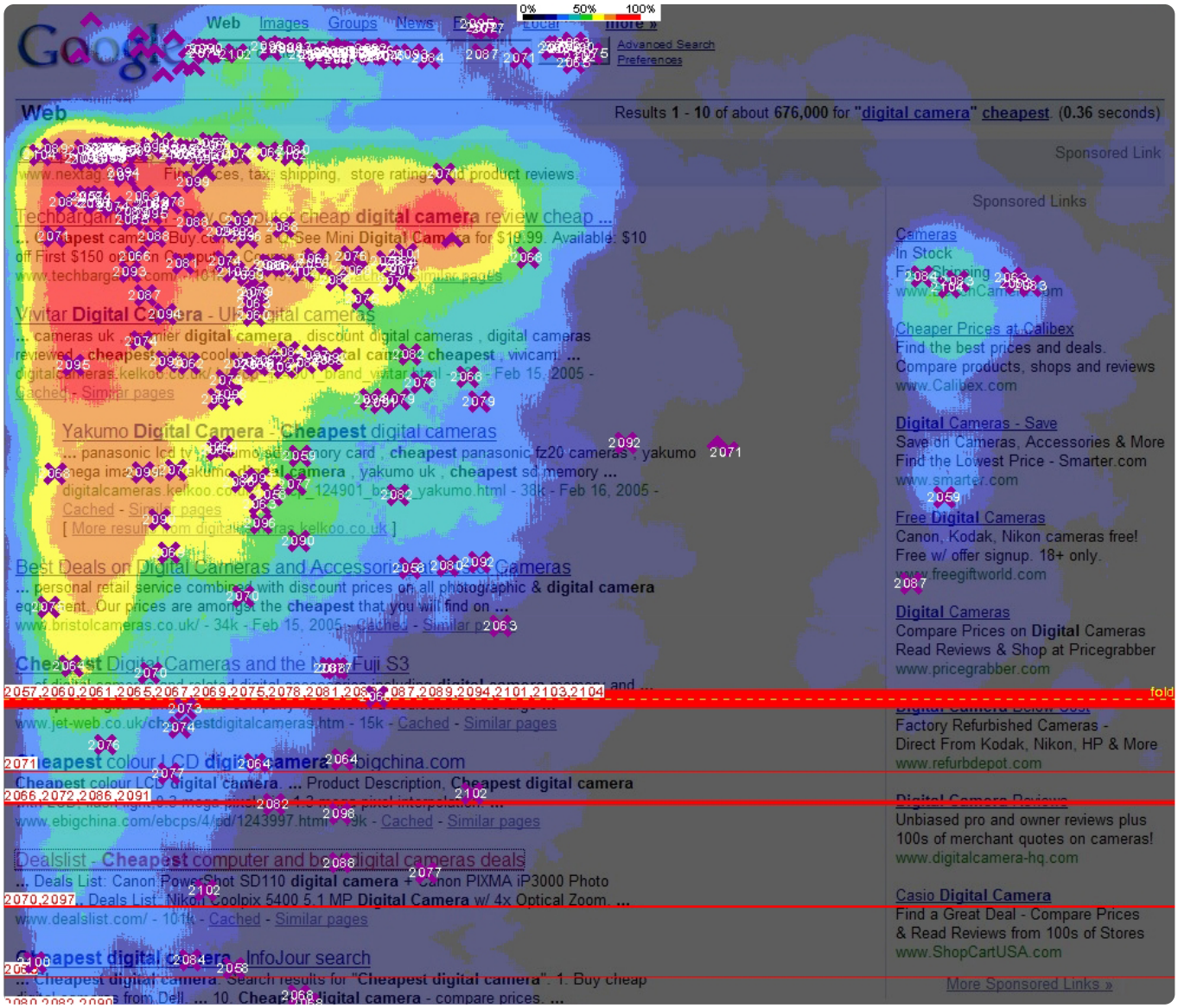

图1 Google搜索结果的注意力热度图

价值是非常大的。比如说，我们要做一个新闻推荐的模型，让这个模型根据用户已经看过的新闻做推荐。那我们在分析用户已浏览新闻的时候，是把标题、首段、全文的重要性设置成完全一样比较好，还是应该根据用户的注意力不同给予不同的权重呢？当然，肯定是

后者比较合理，因为用户很可能都没有注意到正文最后的几段，如果你分析内容的时候把最后几段跟标题、首段一视同仁，那肯定就把最重要的信息给淹没了。

事实上，近年来，注意力机制已经成功应用在各种场景下的推荐系统中了。其中最知名的，要数阿里巴巴的深度推荐模型，DIN（Deep Interest Network，深度兴趣网络）。接下来，我们就一起来学习一下 DIN 的原理和模型结构。

深度兴趣网络 DIN 的原理和结构

DIN 模型的应用场景是阿里最典型的电商广告推荐。对于付了广告费的商品，阿里会根据模型预测的点击率高低，把合适的广告商品推荐给合适的用户，所以 DIN 模型本质上是一个点击率预估模型。

注意力机制是怎么应用在 DIN 模型里的呢？回答这个问题之前，我们得先看一看 DIN 在应用注意力机制之前的基础模型是什么样的，才能搞清楚注意力机制能应用在哪，能起到什么作用。

下面的图 2 就是 DIN 的基础模型 Base Model。我们可以看到，Base Model 是一个典型的 Embedding MLP 的结构。它的输入特征有用户属性特征（User Proflie Features）、用户行为特征（User Behaviors）、候选广告特征（Candidate Ad）和场景特征（Context Features）。

用户属性特征和场景特征我们之前也已经讲过很多次了，这里我们重点关注用户的行为特征和候选广告特征，也就是图 2 中彩色的部分。

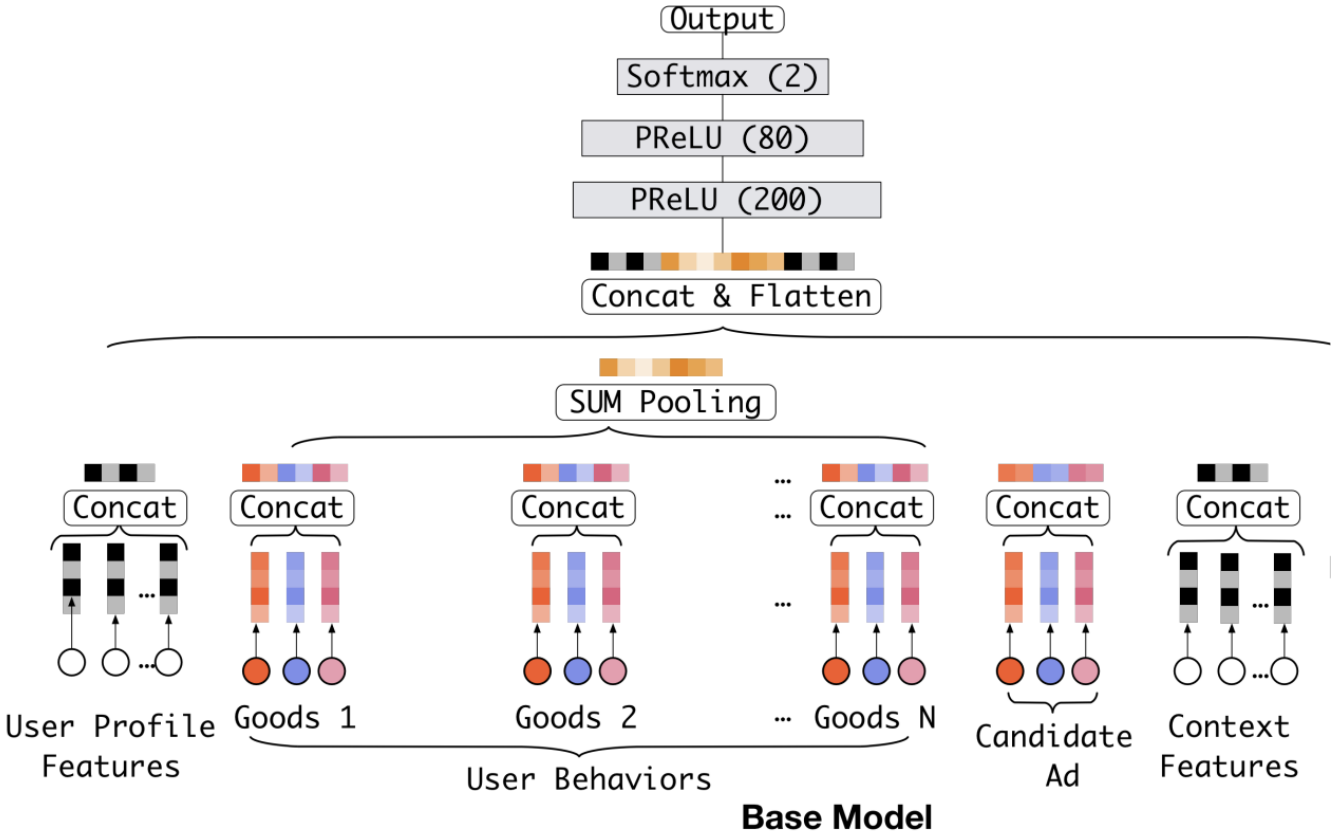

图2 阿里Base模型的架构图
(出自论文 Deep Interest Network for Click-Through Rate Prediction)

我们可以清楚地看到，用户行为特征是由一系列用户购买过的商品组成的，也就是图上的 Goods 1 到 Goods N，而每个商品又包含了三个子特征，也就是图中的三个彩色点，其中红色代表商品 ID，蓝色是商铺 ID，粉色是商品类别 ID。同时，候选广告特征也包含了这三个 ID 型的子特征，因为这里的候选广告也是一个阿里平台上的商品。

我们之前讲过，在深度学习中，只要遇到 ID 型特征，我们就构建它的 Embedding，然后把 Embedding 跟其他特征连接起来，输入后续的 MLP。阿里的 Base Model 也是这么做的，它把三个 ID 转换成了对应的 Embedding，然后把这些 Embedding 连接起来组成了当前商品的 Embedding。

完成了这一步，下一步就比较关键了，因为用户的行为序列其实是一组商品的序列，这个序列可长可短，但是神经网络的输入向量的维度必须是固定的，那我们应该怎么把这一组商品的 Embedding 处理成一个长度固定的 Embedding 呢？图 2 中的 SUM Pooling 层的结构就给出了答案，就是直接把这些商品的 Embedding 叠加起来，然后再把叠加后的 Embedding 跟其他所有特征的连接结果输入 MLP。

但这个时候问题又来了，SUM Pooling 的 Embedding 叠加操作其实是把所有历史行为一视同仁，没有任何重点地加起来，这其实并不符合我们购物的习惯。

举个例子来说，候选广告对应的商品是“键盘”，与此同时，用户的历史行为序列中有这样几个商品 ID，分别是“鼠标”“T 恤”和“洗面奶”。从我们的购物常识出发，“鼠标”这个历史商品 ID 对预测“键盘”广告点击率的重要程度应该远大于后两者。从注意力机制的角度出发，我们在购买键盘的时候，会把注意力更多地投向购买“鼠标”这类相关商品的历史上，因为这些购买经验更有利于我们做出更好的决策。

好了，现在我们终于看到了应用注意力机制的地方，那就是用户的历史行为序列。阿里正是在 Base Model 的基础上，把注意力机制应用在了用户的历史行为序列的处理上，从而形成了 DIN 模型。那么，DIN 模型中应用注意力机制的方法到底是什么呢？

我们可以从下面的 DIN 模型架构图中看到，与 Base Model 相比，DIN 为每个用户的历史购买商品加上了一个激活单元（Activation Unit），这个激活单元生成了一个权重，这个权重就是用户对这个历史商品的注意力得分，权重的大小对应用户注意力的高低。

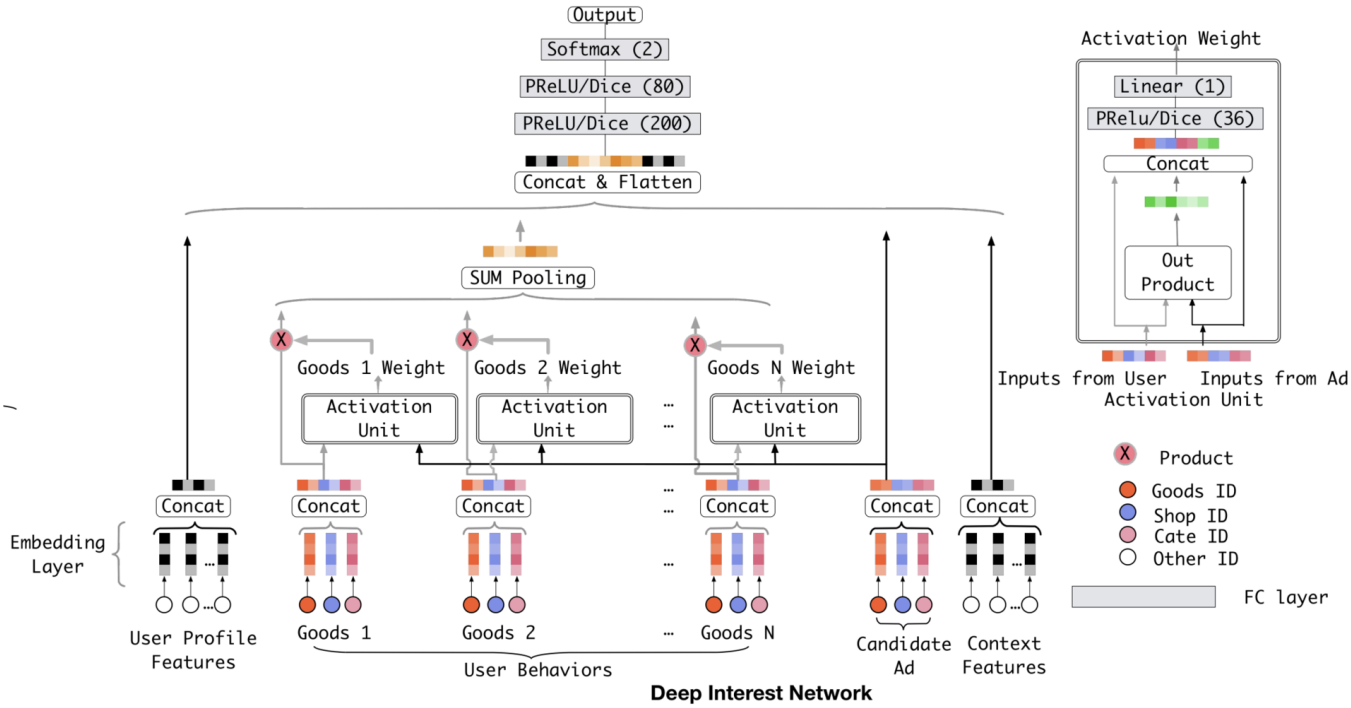

图3 阿里DIN模型的架构图
(出自论文 Deep Interest Network for Click-Through Rate Prediction)

那现在问题就只剩下一个了，这个所谓的激活单元，到底是怎么计算出最后的注意力权重的呢？为了搞清楚这个问题，我们需要深入到激活单元的内部结构里面去，一起来看看图 3 右上角激活单元的详细结构。

它的输入是当前这个历史行为商品的 Embedding，以及候选广告商品的 Embedding。我们把这两个输入 Embedding，与它们的外积结果连接起来形成一个向量，再输入给激活单元的 MLP 层，最终会生成一个注意力权重，这就是激活单元的结构。简单来说，**激活单元就相当于一个小的深度学习模型，它利用两个商品的 Embedding，生成了代表它们关联程度的注意力权重。**

到这里，我们终于抽丝剥茧地讲完了整个 DIN 模型的结构细节。如果你第一遍没理解清楚，没关系，对照着 DIN 模型的结构图，反复再看几遍我刚才讲的细节，相信你就能彻底消化吸收它。

注意力机制对推荐系统的启发

注意力机制的引入对于推荐系统的意义是非常重大的，它模拟了人类最自然，最发自内心的注意力行为特点，使得推荐系统更加接近用户真实的思考过程，从而达到提升推荐效果的目的。

从“注意力机制”开始，越来越多的对深度学习模型结构的改进是基于对用户行为的深刻观察而得出的。由此，我也想再次强调一下，**一名优秀的算法工程师应该具备的能力，就是基于对业务的精确理解，对用户行为的深刻观察，得出改进模型的动机，进而设计出最合适你的场景和用户的推荐模型。**

沿着这条思路，阿里的同学们在提出 DIN 模型之后，并没有停止其推荐模型演化的进程，而是又在 2019 年提出了 DIN 模型的演化版本，也就是深度兴趣进化网络 DIEN（Deep Interest Evolution Network），那这个 DIEN 到底在 DIN 基础上做了哪些改进呢？

兴趣进化序列模型

无论是电商购买行为，还是视频网站的观看行为，或是新闻应用的阅读行为，特定用户的历史行为都是一个随时间排序的序列。既然是和时间相关的序列，就一定存在前后行为的依赖关系，这样的序列信息对于推荐过程是非常有价值的。为什么这么说呢？

我们还拿阿里的电商场景举个例子。对于一个综合电商来说，用户兴趣的迁移其实是非常快的。比如，上一周一位用户在挑选一双篮球鞋，这位用户上周的行为序列都会集中在篮球鞋这个品类的各个商品上，但在他完成这一购物目标后，这一周他的购物兴趣就可能变成买一个机械键盘，那这周他所有的购买行为都会围绕机械键盘这个品类展开。

因此，如果我们能让模型预测出用户购买商品的趋势，那肯定会对提升推荐效果有益的。而 DIEN 模型，就正好弥补了 DIN 模型没有对行为序列进行建模的缺陷，它围绕兴趣进化这个点进一步对 DIN 模型做了改进。

图 4 就是 DIEN 模型的架构图，这个模型整体上仍然是一个 Embedding MLP 的模型结构。与 DIN 不同的是，DIEN 用“兴趣进化网络”也就是图中的彩色部分替换掉了原来带有激活单元的用户历史行为部分。这部分虽然复杂，但它的输出只是一个 h’(T) 的 Embedding 向量，它代表了用户当前的兴趣向量。有了这个兴趣向量之后，再把它与其他特征连接在一起，DIEN 就能通过 MLP 作出最后的预测了。

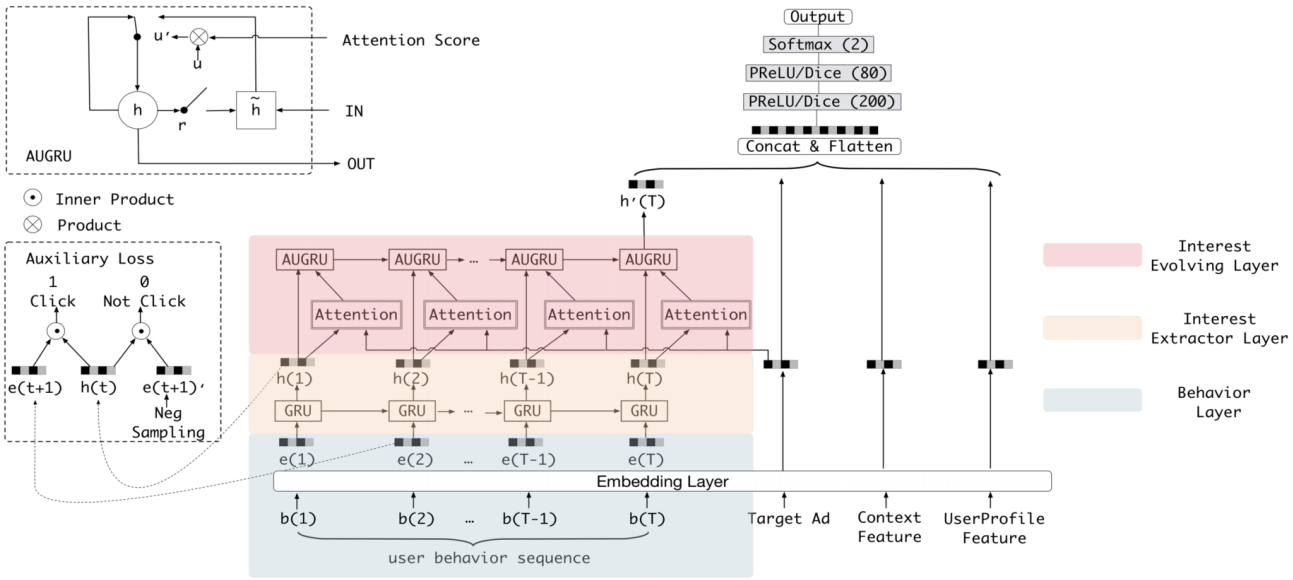

图4 DIEN模型的架构图 （出自论文 Deep Interest Evolution Network for Click-Through Rate Prediction）

好了，现在问题的焦点就在，DIEN 模型是如何生成这个兴趣向量的。关键就在于 DIEN 模型中彩色部分的三层兴趣进化网络，下面，我就按照从下到上的顺序，给你讲讲它们的名称和作用。

最下面一层是行为序列层（Behavior Layer，浅绿色部分）。它的主要作用和一个普通的 Embedding 层是一样的，负责把原始的 ID 类行为序列转换成 Embedding 行为序列。

再上一层是兴趣抽取层（Interest Extractor Layer，浅黄色部分）。它的主要作用是利用 GRU 组成的序列模型，来模拟用户兴趣迁移过程，抽取出每个商品节点对应的用户兴趣。

最上面一层是兴趣进化层（Interest Evolving Layer，浅红色部分）。它的主要作用是利用 AUGRU(GRU with Attention Update Gate) 组成的序列模型，在兴趣抽取层基础上加入

注意力机制，模拟与当前目标广告（Target Ad）相关的兴趣进化过程，兴趣进化层的最后一个状态的输出就是用户当前的兴趣向量 h’(T)。

不知道你发现了吗，兴趣抽取层和兴趣进化层都用到了序列模型的结构，那什么是序列模型呢？直观地说，图 5 就是一个典型的序列模型的结构，它和我们之前看到的多层神经网络的结构不同，序列模型是“一串神经元”，其中每个神经元对应了一个输入和输出。

那在 DIEN 模型中，神经元的输入就是商品 ID 或者前一层序列模型的 Embedding 向量，而输出就是商品的 Embedding 或者兴趣 Embedding，除此之外，每个神经元还会与后续神经元进行连接，用于预测下一个状态，放到 DIEN 里，就是为了预测用户的下一个兴趣。这就是序列模型的结构和作用。

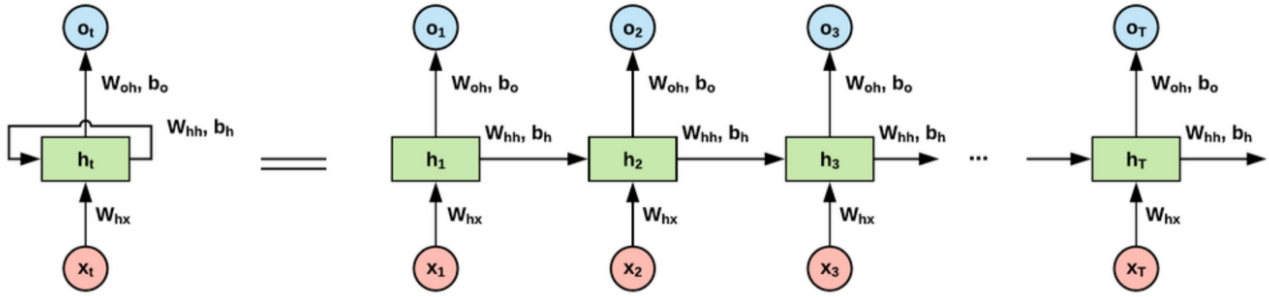

图5 RNN模型的经典结构

至于上面提到过的 GRU 序列模型，它其实是序列模型的一种，根据序列模型神经元结构的不同，最经典的有RNN、LSTM、GRU这 3 种。这里我们就不展开讲了，对理论感兴趣的同学，可以点击我给出的超链接，参考这几篇论文做更深入的研究。

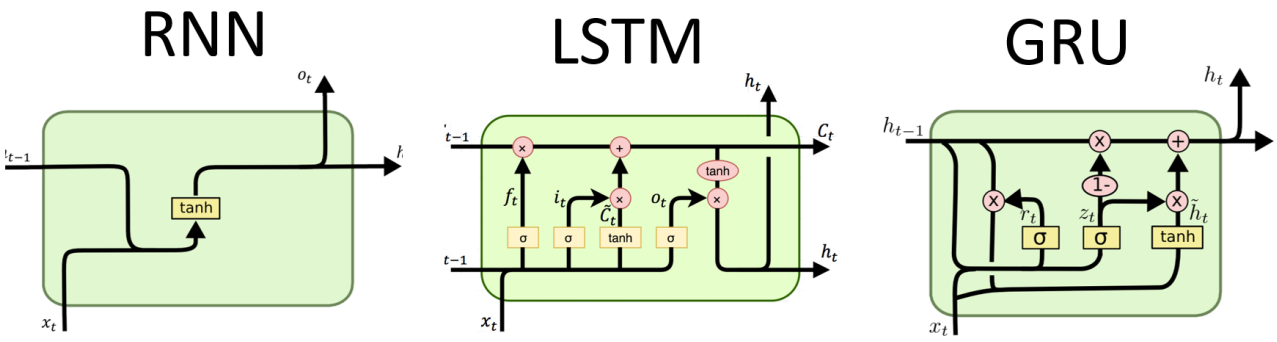

图6 序列模型中的不同单元结构

事实上，序列模型已经不仅在电商场景下，成功应用于推测用户的下次购买行为，在 YouTube、Netflix 等视频流媒体公司的视频推荐模型中，序列模型也用来推测用户的下次

观看行为（Next Watch）。除此之外，音乐类应用也非常合适使用序列模型来预测用户的音乐兴趣变化。所以，掌握 DIEN 模型的架构对于拓宽我们的技术视野非常有帮助。

小结

注意力机制和兴趣演化序列模型的加入，让推荐系统能够更好地抓住用户的心。

对于注意力机制来说，它主要模拟了人类注意力的天性。具体到阿里的 DIN 模型上，它利用激活单元计算出用户对于不同历史商品的注意力权重，针对当前广告商品，作出更有针对性的预测。

而序列模型更注重对序列类行为的模拟和预测。典型的例子是 DIEN 模型对用户购买兴趣进化过程的模拟。DIEN 模型可以应用的场景非常广泛，包括电商平台的下次购买，视频网站的下次观看，音乐 APP 的下一首歌曲等等。

总的来说，注意力机制的引入是对经典深度学习模型的一次大的改进，因为它改变了深度学习模型对待用户历史行为“一视同仁”的弊端。而序列模型则把用户行为串联起来，让用户的兴趣随时间进行演化，这也是之前的深度学习模型完全没有考虑到的结构。

最后，我把今天的重要概念总结在了表格中，方便你及时查看和复习。

知识点	关键概述
注意力机制	人类天生的“选择性注意”的习惯
DIN模型的结构	经典的Embedding MLP的框架， 加上应用了注意力机制的用户历史行为部分
DIN应用注意力机制的过程	根据历史行为商品和广告商品的Embedding， 使用了激活单元来预测注意力权重
DIEN模型的结构	经典的Embedding MLP的框架， 加上三层的序列模型结构
DIEN的三层序列模型结构	基于Embedding层的行为序列层 基于GRU序列模型的兴趣抽取层 基于AUGRU序列模型的兴趣进化层

课后思考

DIN 使用了一个结构比较复杂的激活单元来计算注意力权重，你觉得有没有更简单、更实用的方式来生成注意力权重呢？其实计算注意力权重就是为了计算历史行为物品和广告物品的相关性，在这个过程中，你觉得能不能利用到特征交叉的知识呢？为什么？

欢迎把你的思考和疑问写在留言区，如果你的朋友们也在关注注意力机制和兴趣演化序列模型的发展，那不妨也把这节课转发给他们，我们下节课见！

提建议

上一篇 20 | DeepFM：如何让你的模型更好地处理特征交叉？

下一篇 22 | 强化学习：让推荐系统像智能机器人一样自主学习

精选留言 (14)

写留言

范闲
2020-12-02

1.历史行为和广告物品直接dot
2.利用双塔模型，取最终的输出做特征

作者回复: 没问题

1

浣熊当家
2020-11-29

图3中DIN的激活单元里我们用到了“外积”， 之前的课程里感觉我们多数是用“内积”。请问老师，如何选择使用“内积”和“外积”， 有什么规则吗?

展开

作者回复: 本质上都是做特征交叉，计算相似性的方式，一般来说，外积因为输出是一个向量，所以表达能力更强一些。

1

freedom
2020-11-30

老师我想问一个问题，在训练模型过程中，在用到BN的时候，经常会出现BN层训练不好的情况，具体表现为训练时候(is _ training=TRUE)准召比较高。但是设置is_training为false的时候 准召降得比较厉害。moving_mean和moving_var也都有更新。请问这是因为什么原因呢，该如何解决呢?

展开

作者回复: 抱歉不能提供很多经验上的指导，我很少在推荐模型里面使用BN层，从原理上怀疑是某些类别型特征非常系数导致的，权当猜测。
如果其他同学有相关经验，欢迎讨论。

2

东格拉底
2020-11-30

DIN的激活单元相当于一个小型的深度学习模型，那这个激活单元是单独训练的嘛？如果是这样的话 那模型的y又是什么呢？还是说激活单元是作为整个DIN的一部分进行端到端的训练的？

作者回复: 端到端训练

浣熊当家
2020-11-29

想请教老师一个题外问题， 您作为面试官的话，对于MLE的候选人会更注重考察哪方面的能力（比如算法coding， 系统架构设计，机器学习的领域知识）？，然后对于机器学习的各种模型会期待候选人有多深的了解（比如说了解DIEN的各个层级的结构就够了，还是要知道GRU是具体如何实现的）。
随着老师课程，我对深度学习燃起了更大的热情，觉得想真正提高的话，最好的方法还...
展开

作者回复: 我在知乎有一篇专栏文章专门讲面试，可以参考 https://zhuanlan.zhihu.com/p/76827460

我在课程最后也会有一些总结。

Geek_e0d66a
2020-11-28

请问老师，DIN中的激活单元，的网络结构中，为什么用户Embedding和候选集的embedding的外积，再拼接在一起？？
展开

作者回复: 因为DIN的同学经过多次实验之后这个网络结构最好。曾经最早的DIN的激活单元是通过element wise minus做特征交叉，但是几经修改，所以都是实践的结果。

Leo Zhao
2020-11-28

思考题：广告和历史行为 相关性 其实就是 广告物品与历史物品的相关性，可以用一个dot层 或者通过物品embedding 算出similarity 当作feature 直接输入。

展开

作者回复: 非常好

张弛 Conor
2020-11-28

请问老师，DIN模型的Activation Unit中，两个首先会进行out product运算，两个向量的外积运算不应该是一个常数或者nxn的矩阵嘛，但是图示中得到的是和向量相同维度的向量，我查阅了老师的书籍，我看书中的激活单元中相应位置使用的是元素相减，请问老师这是为什么呢?

展开

作者回复: 元素相减是阿里第一版DIN的选择，正式论文中采用了这一讲里介绍的Activation Unit结构。

外积的运算结果是一个向量，向量方向是这个两个向量组成的平面的法向量方向。

1

Sebastian
2020-11-28

老师想问下，DIN模型在工业界的排序阶段使用的多吗？因为我在想在业界每个用户都有比较长的用户行为序列的场景可能还是少数，很多公司的场景可能是，用户进入app端后点击了2-3次后可能就没有后续行为了，那么这种场景下，DIN应该就不适用了吧?

展开

作者回复: DIN比DIEN的使用场景要求低很多，我知道很多团队在用，或者说很多团队在用DIN的思路来构建自己的模型。

就我自己的实践经验，attention机制是非常有价值的，推荐在自己的场景下尝试。

张弛 Conor
2020-11-27

课后疑问:
请问老师，对于DIN和DIEN这种包含N个历史商品的模型，如果用户历史商品数小于N，

那么这些位置应该如何去填充呢？如果用户商品数大于N，是否是选择最近的N个商品呢？

展开

作者回复: 大于N肯定要选择最近的N个， 小于N考虑使用masking layer
tf.keras.layers.Masking

张弛 Conor
2020-11-27

思考题：可以借鉴FM及DeepFM中特征交叉的计算方式，对两个向量直接计算内积，或者先通过Embedding层转换成维度相等的Embedding再求内积，又或者可以像双塔结构一样，设计一个历史行为物品塔和广告物品塔，在塔的最后通过求内积或者拼接后用全连接层输出权重。
课后疑问：...

展开

作者回复: 思考题的回答很赞同。

关于疑问，我不是很清楚你的问题，什么叫做“历史兴趣以外的兴趣”？

那时刻
2020-11-27

DIN 使用了一个结构比较复杂的激活单元来计算注意力权重，能否有其它方法来计算呢？我开始想到的上节课提到的特征交叉，我回顾了下特征交叉，它是加强模型对于特征组合和特征交叉的学习能力以及对于未知特征组合样本的预测能力。而注意力权重是计算历史物品和广告物品相关行。我觉得它们是不同的。

展开

作者回复: 其实只要是接收两个embedding，生成一个输出分数的结构理论上都可以作为激活单元。所以最简单的结构其实就是dot product或者cosin similarity。因为它们刚好揭示了两个item之间的相似度。

kenan
2020-11-27

王老师，候选集是这个意思，以点击预估为例，在排序阶段，我们通过把用户特征，item特征和上下文特征放入到模型里面，得到item的点击预估值。然后根据候选集里面每个item的点击预估值从大到小排序，然后取出topN。现在候选集只有item，是如何建模。有没有工程上的demo，或者资料推荐。

展开

作者回复: 不是很明白这个问题，可补充解释。

那时刻

2020-11-27

请问老师DIN模型里的concat&flatten作用是什么?

展开

作者回复: 就是把所有向量拼接在一起，成为一个向量。