Deep Interest Network for Click-Through Rate Prediction

传统的基于深度学习的点击率预测算法是Embedding+MLP模块，通过低维嵌入拼接以及前馈网络的组合实现预测，尽管广告的特性是不会变化的，但是对于用户，其兴趣爱好可能会发生改变，与词的多义性相近，因此需要综合考虑用户的丰富的历史行为，自适应地学习针对不同广告下的用户的低维嵌入（不同广告下的嵌入不同）。此外，mini-batch aware正则化以及适应数据的激活函数的新方法被提出。

文章贡献

· 指出固定的用户嵌入向量的缺点，引入不同广告下对用户历史行为的不同响应来自适应调整用户的嵌入向量（local activation unit）。

· 引入减少深度网络的正则化计算量以及有利于防止过拟合的mini-batch aware正则化**和**通过考虑输入的分布来一般化PReLU数据自适应的激活函数。

· 实验证明算法很厉害！

特征表示

可使用独热编码和多热编码来表示各种特征：用户基本特征、用户行为特征、广告基本特征以及上下文（环境）特征。此外，模型将特征交互的过程在深度网络中进行实现，而非独立使用FM或FFM提取交互特征。

基本模型（嵌入+多层感知机）

·嵌入层

嵌入层遵循字典查找的规则方法，使用矩阵W作为嵌入字典，不同特征对应于不同的字典形式。针对不同的编码方式，独热编码只需找到一个嵌入向量，而多热编码则会返回一个嵌入向量的列表。（注意上面的示意图是用三个不同id：cate\_id/shop\_id/goods\_id的拼接而成的长向量作为最终用户产生行为的广告表示）

· 池化层和拼接层

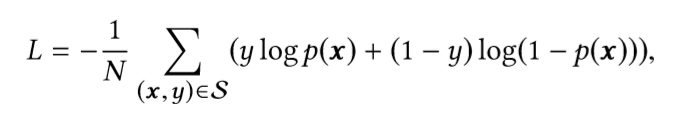
为了解决不同用户下多热编码返回的列表长度不一致的问题，对该列表进行池化操作从而产生相同长度的输出向量，常用求和和平均池化。

· 多层感知机

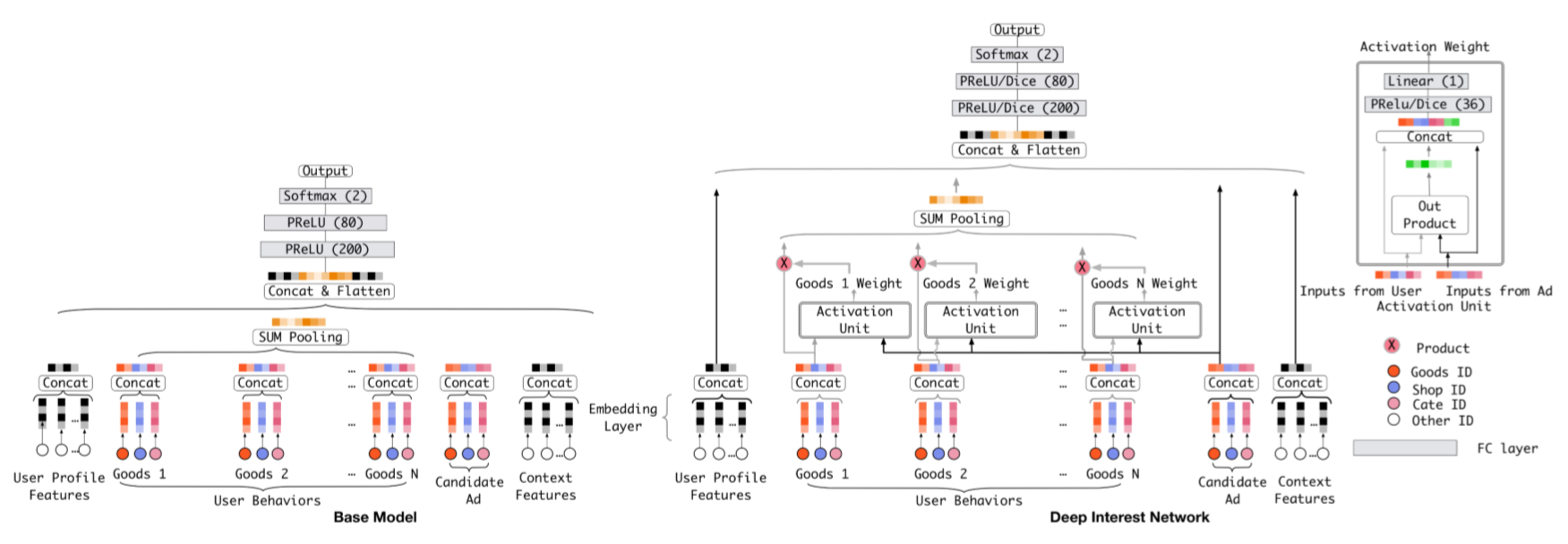
将最终得到的不同的嵌入向量进行拼接产生长向量，作为多层感知机的输入。

· 损失

负log似然函数



其中x为商品，对应图中的候选广告，同时y为0-1标签。

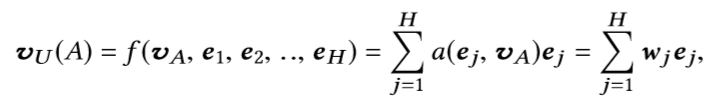


两种模型对比（base vs DIN）

DIN

不难发现，base模型下，不同的候选广告下，对应的用户嵌入向量是一致的，对模型表达能力产生了限制，为此，可以提高嵌入向量的维度来扩大用户的兴趣空间，但这会大大提升模型的复杂度和增加模型过拟合的风险，不利于构建线上模型。

为了更好地理解用户当前兴趣，DIN在base模型的基础上引入了局部激活单元，相当于给定候选广告，以加权求和的方式自适应地计算用户嵌入向量（即作用在有过历史行为的广告上）。



其中a(·)是前馈网络，输入两个嵌入向量并输出一个类概率值，但不需要求和为1，为了得到一个绝对的“兴趣强度值”。

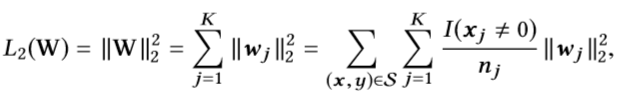
文章尝试了使用LSTM来对用户产生过历史行为的广告进行建模分析，发现并无提升，可能是因为用户的兴趣具有并发性，而非时序效应。

大数据下的训练技巧

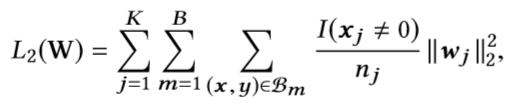
· mini-batch aware 正则化

由于特征的稀疏性，在添加正则化前，每个mini-batch只需更新那些非零的特征参数，但是在添加正则项后，则需要在每轮更新全部的参数，这在运算复杂度上是不可接受的。

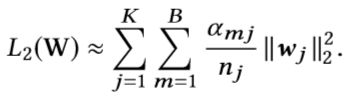
因此修改正则化项为下面形式，其中n代表某一维度下不为0的样本个数。



Mini-batch下的形式为

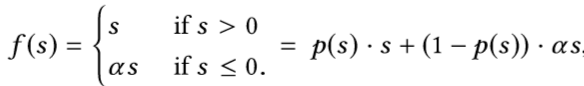


取其近似形式

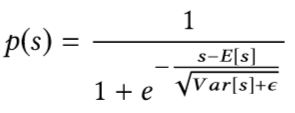


其中，用于记录某个mini-batch是否包含该特征。当mini-batch规模相对整个数据集规模较小时，该做法时合理有效的。在进行梯度下降的时候，可以利用前面这条简化后的式子来降低更新的复杂度。

· 数据自适应激活函数

PReLU：，其中p为指示函数，且alpha是一个需要学习的参数。

由于上述p函数在位置0有一处跳跃，非连续型，因此可能对不同分布的输入数据不具有普适性，因此提出Dice函数p



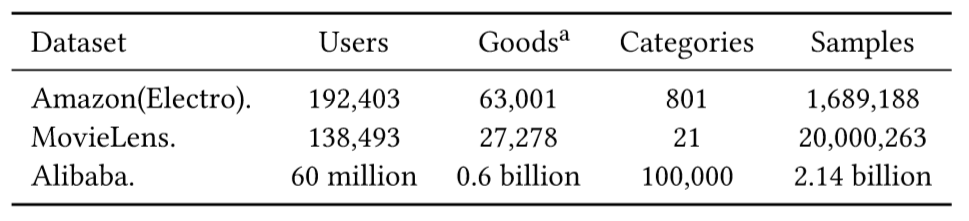
E，V在训练过程中分别是输入数据的均值和方差，在测试过程中，由E和V的滑动平均进行计算，eplison是很小的一个数。当E=V=0时，退化为指示函数。

实验部分

· Amazon数据集：以数据集的前n-1个历史行为作为训练集，最后一个作为测试集，使用学习率为1且衰减率为0.1的SGD算法且mini-batch的规模是32。

· MovieLens数据集：星级4-5标注为pos，其他星级标注为neg，以用户作为划分训练集测试集的标准，100000 vs 38493，使用与A相同的优化器。

· Alibaba数据集：两周的数据（20亿vs 1亿4千万），12维度的嵌入向量，192->200->80->2的MLP，使用Adam优化器（0.001，衰减率为0.9）。



对比算法：LR、baseModel（E&M）、Wide&Deep、PNN、DeepFM。

评价方式：AUC、RelaImpr

