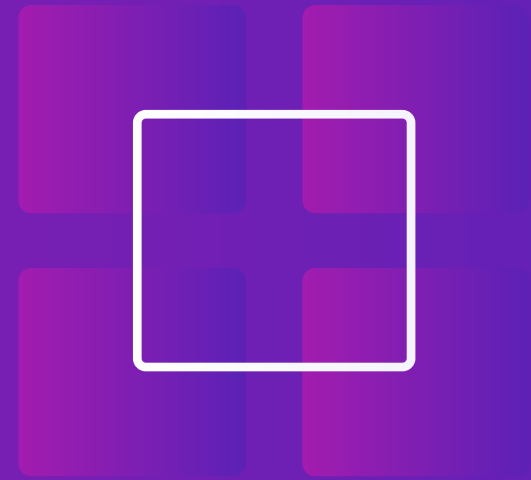
Deep Learning for Click-Through Rate Estimation

乔梁 2022.7.15

一、背景介绍
二、从浅层模型到深度模型
三、特征交互
四、用户行为建模
五、自动化架构搜索
六、总结与展望
七、学习收获



一、背景介绍

CTR

点击率（CTR）估计作为各种个性化在线服务的核心功能模块发挥作用，包括**在线广告**、**推荐系统**和**网络搜索**等。从2015年起，深度学习的成功有利于提升CTR模型的估计性能，现在深度CTR模型已广泛应用于许多行业平台。

本综述全面回顾了CTR估计任务的深度学习模型。

1. 回顾了从浅层到深层CTR模型的转换，并解释了为什么深度模型是一个必要的发展趋势。
2. 介绍了最近出现的深度CTR模型体系结构设计的自动化方法。
3. 总结了综述内容并讨论了该领域的未来前景。

背景

个性化服务在各种在线信息系统中变得非常重要，例如电子商务中的商品推荐、在线广告中的拍卖、web搜索中的页面排名等等。将点击视为用户偏好的代表性行为，基于对记录的行为数据进行学习的点击率（CTR）估计是这些个性化服务的核心功能模块。

对于模型训练，CTR估计任务被表述为一个二元分类问题，交叉熵损失函数为：

$$l(x, y, \Theta) = -y \log \sigma\left(f_\Theta(x)\right) - (1-y) \log\left(1 - \sigma\left(f_\Theta(x)\right)\right)$$

选取一些有代表性的模型来展示CTR估计模型的发展趋势，如图2所示。模型的发展可以概括为两个方面，即特征工程复杂性和模型容量。对于早期CTR模型，受计算能力限制，主要致力于通过采用简单模型来设计更好的特征。随后，引入了更复杂的模型（具有深层架构和更好的建模能力），以释放特征工程中人类工作的复杂性。最近的一个趋势是再次关注使用一些可学习方法的特征工程，因为仅通过设计更复杂的深度模型，它已经达到了性能瓶颈。将`复杂模型`和`可学习特征工程`相结合是新的发展方向。

背景

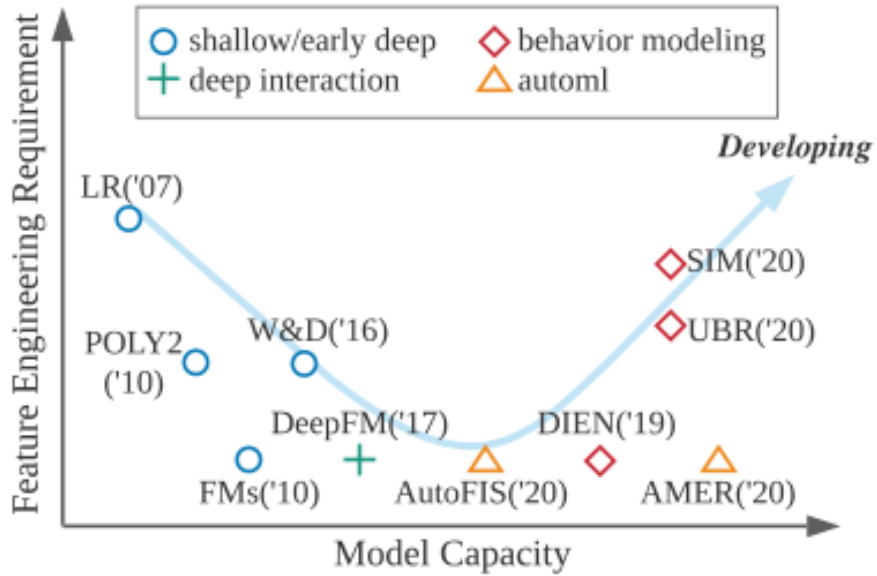


Figure 2: Development trend of CTR estimation models.

二、从浅层模型到深度模型

处理二元分类任务，逻辑回归（LR）是最基本的模型，具有效率高、易于快速部署的优点。LR的logit值计算如下：

$$f_{\Theta}^{\mathrm{LR}}(x)=\theta_0+\sum_{i=1}^{m}x_i\theta_i$$

其中 $\Theta=\{\theta_i\}_{i=0}^{m}$，$m$ 表示特征尺寸。然而，点击预测的许多判别模式是组合特征（或称为交叉特征），例如，用户是学生（职业：学生），位置是LA（城市：LA），将提升迪斯尼乐园广告（广告：迪斯尼乐园）的预测CTR，这对应于三阶组合特征。因此，一种简单的方法是手动选择和设计许多有用的组合特征，这需要大量的人力，如图2所示。

- POLY2 为每个二阶组合特征分配权重，这需要 $O(m^2)$ 参数空间。不幸的是，当数据稀疏时，POLY2的性能可能较差。关键问题是在稀疏数据中无法正确估计与特征交互相关的参数，其中许多特征从未或很少同时出现。
- 提高 LR 的另一种方法是利用特征选择和梯度提升决策树（GBDT）组合的能力来自动学习特征交互。然而，GBDT很难并行训练，只能利用一小部分可能的特征交互，这限制了其在大规模场景中的性能和应用。
- 因子分解机（FM）为每个特征 i 分配一个 k 维可学习嵌入向量 v_i，这使得模型能够以灵活的方式探索组合特征的有效性：

$$f_{\Theta}^{\mathrm{FM}}(x) = \theta_0 + \sum_{i=1}^{m} x_i \theta_i + \sum_{i=1}^{m} \sum_{j=i+1}^{m} x_i x_j v_i^{\top} v_j$$

其中 $\Theta = (\theta, v)$。可以直观地看到，如果组合特征 $(i，\ j)$与 $P(y=1)$ 正（负）相关，则嵌入内积 $v_i^T v_j$ 将自动学习为正（负）。FM 可进一步扩展。最显著的扩展是 FFM，它为特征分配多个独立嵌入，以明确建模特征与不同场的交互。

自2015年起，深度学习的成功开始通过转移经典架构或开发新架构来提高CTR估计性能。神经网络的普遍逼近特性和具有GPU计算堆栈的深度学习编程库使训练深度神经网络模型以捕捉高阶特征交互模式和实现更好的CTR估计性能成为可能。利用每个稀疏（或分类）特征的向量表示，可以通过连接这些向量来构建实例密集向量，这些向量可以简单地馈送到具有sigmoid输出的多层感知器（MLP）中。在工业上，这种结构被称为DNN或稀疏神经网络（SNN）。

Wide&Deep network：

$$f_{\Theta}^{\mathrm{W\&D}}(x)=\theta_0+\sum_{i=1}^{m}x_i\theta_i+\mathrm{MLP}_{\phi}\left([v_1,v_2,\ldots,v_m]\right)$$

其中 $\Theta=(\theta,v,\phi)$。请注意，手动设计的交叉特征也可以包含在特征向量中。因此，DNN部分可以被视为学习LR的残余信号以进一步接近标签。此外，DeepCross扩展了残差网络，以隐式方式执行自动特征交互学习。这种早期的深度CTR模型通过合并简单的MLP减轻了人们在特征工程中的努力。

虽然这些早期基于MLP的深度模型确实实现了性能改进，但要训练出一个好的模型可能需要付出相当大的努力。Qu等人[2018]指出了DNN在捕捉离散特征交互方面的不敏感梯度问题，并根据经验表明，不同大小的DNN不能很好地拟合POLY2函数。Rendle等人[2020]发现，与FM中的点积相比，MLP更难有效学习CTR估计任务中的高阶组合特征模式。

从浅层模型到深度模型

1. 特征交互学习主要关注实例内的高效模式挖掘；
2. 用户行为建模探索用户的多个实例之间的依赖关系；
3. 自动化架构搜索方法旨在以免提方式设计上述两种深度模型。

三、特征交互

特征交互

输入的多字段分类数据通常表示为高维独热向量。嵌入层通常用于将这些 onehot 特征压缩为低维实值向量。

特征交互建模表示多个特征的组合关系，是建立预测模型的关键。DNN具有强大的特征表示学习能力，在自动特征交互建模中具有潜力。

然而，单个DNN很难有效地学习高阶特征交互。因此，近年来提出了许多与DNN明确结合特征交互的工作。现有模型的特征交互学习可以表述为：

$$f_{\Theta}^{\mathrm{FIL}}(x) = f_{\psi}\left([v_1, v_2, \ldots, v_m]\right) + \mathrm{MLP}_{\phi}\left([v_1, \ldots, v_m]\right)$$

其中 $\Theta = (\phi, \psi)$是参数集，f_{ψ} 是显式特征交互学习函数，MLP_{ϕ} 是只有双塔模型才具有的可选函数。

特征交互算子

为显式特征交互学习开发了多个算子，主要可分为三类，即乘积算子、卷积算子和注意力算子。

乘积操作

PNN、NFM、CIN、KPNN、PIN

卷积操作

除了用于特征交互建模的产品操作外，还探讨了卷积神经网络（CNN）和图卷积网络（GCN）用于特征交互建模。卷积点击预测模型（CCPM）[Liu等人，2015]重复执行卷积、池和非线性激活，以生成任意阶特征交互，如图4（b）所示。然而，CCPM只能学习相邻特征之间的部分特征交互，因为它对字段顺序很敏感。特征生成卷积神经网络（FGCNN）[Liu等人，2019a]通过引入重组层来建模非相邻特征，改进了CCPM。然后，它将CNN生成的新特征与原始特征相结合，以进行最终预测。FGCNN验证了CNN生成的特征可以扩展原始特征空间，降低现有深部结构的优化难度。

特征交互图神经网络（FiGNN）[Li等人，2019]认为，使用简单非结构化特征场组合的现有深度模型对复杂特征交互建模的能力有限。受GCN成功的启发，它将多字段分类数据视为完全连通图，其中不同字段作为图节点，不同字段之间的交互作为图边，然后通过图传播对特征交互进行建模。

注意力操作

AFM、AutoInt

特征交互算子

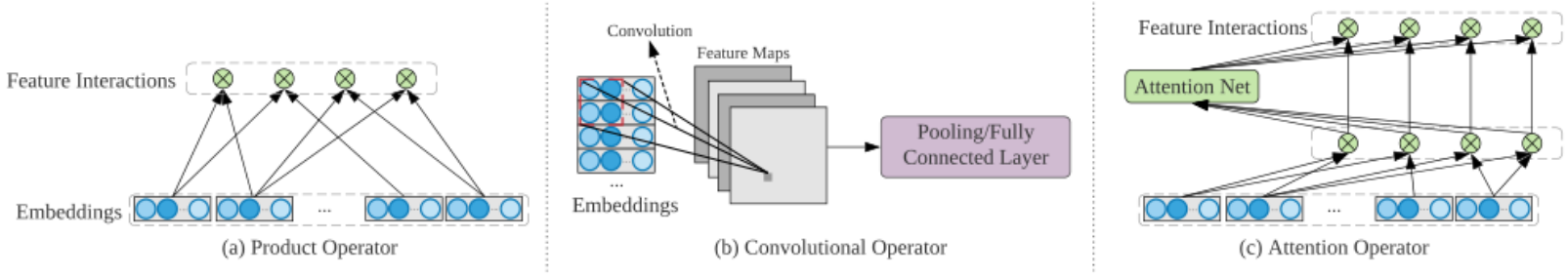


Figure 4: Illustration of three typical interaction operators.

DNN在深度CTR模型中的作用

特征交互算子可用于根据DNN和特征交互层在架构中的相对位置构建`单塔网络`或`双塔网络`。

- 单塔模型在架构中依次放置特征交互和深度网络，如图3（a）所示。这些模型可以有效地捕捉高阶特征相互作用，但低阶特征相互作用的信号可能在以下DNN中消失。
- 为了更好地捕捉低阶特征交互，提出了双塔网络。如图3（b）所示，它平行放置特征交互层和DNN。特征交互层负责显式捕获低阶交互，而高阶交互由DNN隐式捕获。两个模块的输出用于生成最终预测。

NFM 和 PIN 等单塔模型具有更强的建模能力，具有更复杂的网络结构。但是，它们通常存在很差的局部极小值，并且严重依赖于参数初始化。Wide&Deep、DeepFM、DCN、DCN V2、xDeepFM 和 Autoint 都是双塔模型。

DNN部分始终可以被视为学习特征交互层的残余信号以接近标签的补充，从而产生稳定的训练和改进的性能。

单塔模型和双塔模型

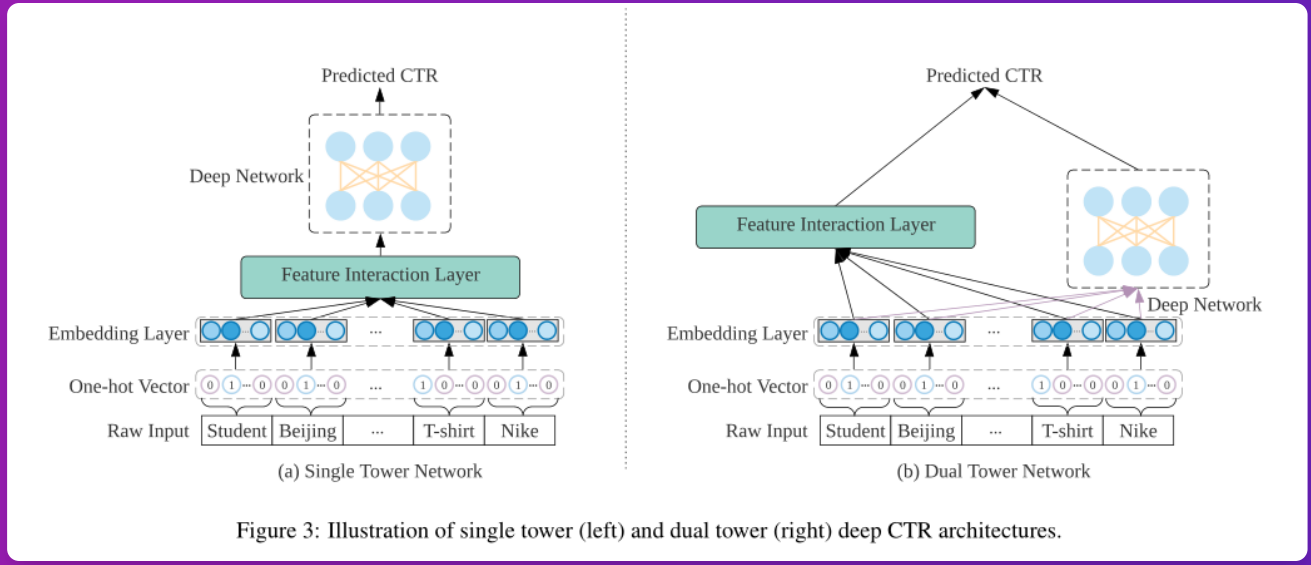


Figure 3: Illustration of single tower (left) and dual tower (right) deep CTR architectures.

四、用户行为建模

用户行为建模

用户行为包含用户兴趣的关键模式。近年来，对用户行为进行建模已成为CTR估计任务的一个重要主题。 用户行为通常被组织为多值分类特征，每个值都是一个行为标识符。行为特征通常是连续的，按时间顺序排序。项目ID和相应的项目特征（如果有）形成一个行为序列。用户行为建模的总体框架如图5所示。

$$f_{\Theta}^{\mathrm{UBM}}(x)=\mathrm{MLP}_{\phi}\left(\left[v_{1}, v_{2}, \ldots, v_{m-1}, \mathcal{U}_{\psi}\left(x_{m}\right)\right]\right)$$

其中 $x_m = \{b_1, b_2, \ldots, b_T\}$是多值行为特征，$T$ 是序列长度，$\Theta = (\phi, \psi)$ 是参数集，$\mathcal{U}_{\psi}$是用户行为建模函数。

关键是设计一个有效的 $\mathcal{U}_{\psi}$ 函数来学习行为特征的稠密表示。在获得行为表示后，将其与其他特征嵌入进行聚合（级联），并将聚合的特征反馈到MLP中以获得最终预测。用户行为建模算法可分为三类：

- 基于注意力的模型
- 基于记忆网络的模型
- 基于检索的模型

在顺序推荐任务中也有许多用户行为建模方法，但本文专注于CTR估计任务，因此这些方法不包括在本节中。

用户行为建模

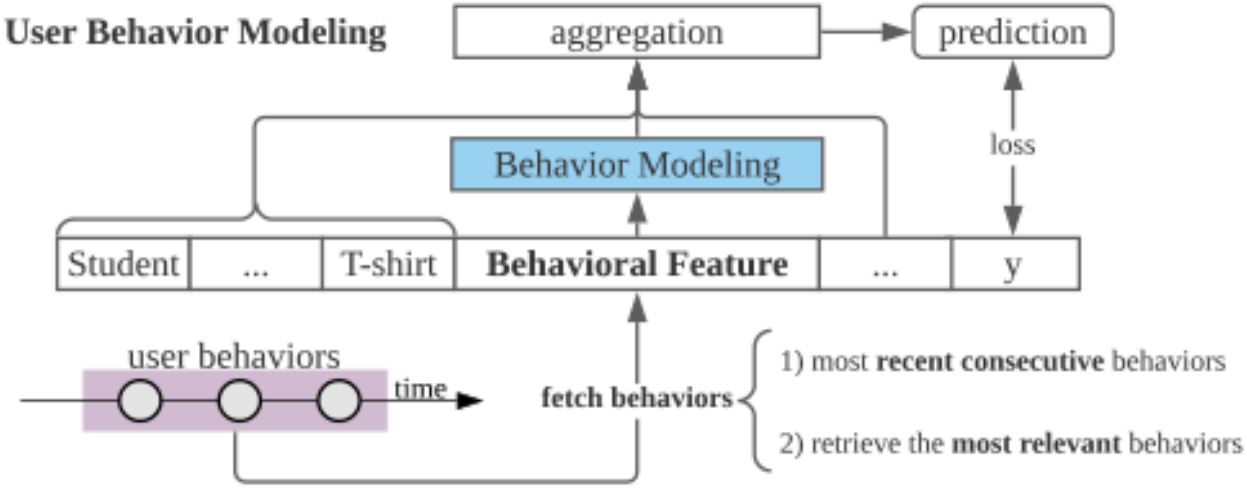


Figure 5: General framework of user behavior modeling.

用户行为建模

基于注意力的模型

DIN、DIEN、BST、DSIN

基于内存网络的模型

由于在大型电子商务平台上积累了大量的用户行为数据，因此处理非常长的行为序列并在更遥远的用户历史中挖掘模式更为重要。DIN或DIEN等模型受时间复杂度的限制，在建模很长的序列时不可行。

因此，提出了基于内存网络的模型，并进行了适当的系统设计。Ren等人[2019]提出了分层周期记忆网络（HPMN）和终身序列建模框架。

它使用终身个性化内存来存储用户兴趣表示，HPMN负责增量更新内存状态。HPMN是一种多层GRU架构，每层具有不同的更新频率。上层更新的频率低于下层，因此HPMN可以捕捉长期但多尺度的时间模式。出于类似动机，Pi等人[2019]设计了一个用户兴趣中心（UIC）模块，该模块将耗时的用户行为建模过程与实时预测过程解耦。在推理时，用户行为表示直接从UIC获得，UIC离线计算。UIC提出了一种多通道用户兴趣记忆网络（MIMN）来计算用户行为表示，因为用户兴趣通常是多通道的。与HPMN相比，UIC从工业角度为长序列建模提供了更为系统的解决方案，而HPMN是第一个将内存概念纳入处理长序列建模的CTR估计任务。

用户行为建模

基于检索的模型

尽管当前的用户行为建模解决方案取得了巨大的成功，但仍然存在一些缺点。以上所有模型都使用了最新的连续行为，并试图结合复杂的体系结构来捕获更长序列中的序列模式，这给系统开销带来了沉重负担。用户行为检索（UBR4CTR）[Qin等人，2020]提出了一种新的框架，使用搜索引擎技术从整个用户行为序列中检索最相关的行为。每次推理时只检索少量最相关的行为，而不是使用大量连续的行为。

这种检索方法不仅解决了处理超长序列的时间复杂度问题，而且还缓解了长连续序列中的噪声信号，因为只有最相关的行为被输入到模型中。

UBR4CTR使用搜索引擎技术检索topk相关的历史行为。选择推理目标的特征来生成查询。查询生成过程是参数化的。因此，它使检索过程可学习，并使用增强算法对其进行优化。

基于搜索的兴趣模型（SIM）[Qi等人，2020]是另一种基于检索的模型，提出了硬搜索和软搜索方法。对于硬搜索，它使用预定义的ID（如用户ID和类别ID）来构建索引。对于软搜索，SIM通过局部敏感哈希（LSH）使用嵌入来检索相关行为。

本节中提到的大多数用户行为模型都部署在真实的在线系统中。这些用户行为建模算法和系统在经济回报和学术价值方面为CTR估计技术的发展做出了重大贡献。

五、自动化架构搜索

自动化架构搜索

由于一些自动设计的网络在计算机视觉中达到了与人类开发的模型相当的性能，推荐系统的研究人员也提出了设计深度连接时序分类体系结构设计的自动化方法。

根据这些方法关注的不同部分，它们可以分为三类：

1. 为单个特征自动设计灵活且适应性强的嵌入维度；
2. 自动选择或生成有效的特征交互；
3. 网络架构的自动设计。

自动化架构搜索

嵌入维度搜索

嵌入表示是深度连接时序分类模型的关键因素，因为它是模型参数的主要组成部分，自然对模型性能有很大影响。因此，一些工作通过自适应自动搜索不同特征的嵌入维数来优化嵌入。

- NIS[Joglekar等人，2020年]和ESAPN[Liu等人，2020c]执行强化学习（RL），以自动搜索混合特征嵌入维度。NIS[Joglekar等人，2020年]首先将通用维度空间划分为由人类专家预定义的几个块，然后应用RL生成关于为不同特征选择此类维度块的决策序列。RL算法的奖励函数同时考虑了推荐模型的准确性和内存开销。
- ESAPN[Liu等人，2020c]还为不同的特征预定义了一组候选嵌入维度。每个字段使用策略网络来动态搜索不同特征的嵌入大小。策略网络将特征的频率和当前嵌入大小作为输入，并输出是否扩大该特征当前维度的决策。

自动化架构搜索

嵌入维度搜索

上述两种方法从一组小的离散候选维度中选择特征维度，而DNIS[Cheng等人，2020]和PEP[Liu等人，2021]使候选维度连续。更具体地说，DNIS[Cheng等人，2020]为每个特征块引入了一个二进制指标矩阵（特征根据其频率分组为特征块，以减少搜索空间），以指示相应维度的存在。然后提出了一个软选择层，将二元指标矩阵的搜索空间松弛为连续。然后使用预定义的阈值来过滤软选择层中的不重要维度。然而，该阈值在实践中很难调整，因此可能会导致模型性能不理想。受这一观察结果的启发，PEP[Liu等人，2021]修剪嵌入参数，其中阈值可以从数据中自适应学习。

上述工作在搜索空间中执行硬选择，即仅为每个特征选择一个嵌入维度。相反，AutoEmb[Zhao等人，2020c]和AutoDim[Zhao等人，2020b]中提出了软选择策略。软选择策略使用可学习的权重对候选维度的嵌入进行求和，其中此类权重通过可微搜索算法（例如DART）进行训练。AutoDim为不同的特征字段分配不同的嵌入大小，而AutoEmb为单个特征搜索不同的嵌入大小。

自动化架构搜索

功能交互搜索

如前所述，特征交互建模在深度CTR模型中至关重要，因此自动设计有效的特征交互具有很高的潜在价值。

AutoFIS[Liu等人，2020b]自动识别并选择因子分解模型的重要特征交互。AutoFIS枚举所有特征交互，并利用一组架构参数来指示单个特征交互的重要性。这些架构参数通过梯度下降（灵感来自DART[Liu等人，2019b]）和GRDA优化器[Chao和Cheng，2019]进行优化，以获得稀疏解，从而自动过滤掉那些不重要的特征交互（即架构参数为零值）。然而，AutoFIS将交互功能限制为内积。

如PIN[Qu等人，2018]所述，作为交互功能，微网络可能比内积更有效。受这一观察的启发，除了识别重要的特征交互之外，确定合适的交互功能也是有益的，这激发了SIF[Yao等人，2020]和AutoFeature[Khawar等人，2020]。SIF[Yao等人，2020]自动为不同数据集中的矩阵分解（仅考虑用户和项目标识符，不考虑任何其他特征）设计合适的交互函数。交互函数的搜索空间由微观空间和宏观空间组成，其中微观空间指元素级MLP，宏观空间包括五个线性代数运算。AutoFeature[Khawar等人，2020]提出利用具有不同架构的微网络来建模每对场之间的特征交互。每个微网络的架构都是从一个具有五个预定义操作的搜索空间中搜索的。搜索过程由具有朴素贝叶斯树的进化算法实现。

自动化架构搜索

功能交互搜索

然而，AutoFIS和AutoFeature都不能对高阶特征交互进行建模，因为它们需要事先枚举所有可能的特征交互。为了避免这种效率低下的枚举，AutoGroup[Liu等人，2020a]建议生成一些特征组，以便它们在给定顺序下的交互是有效的。在任何此类组中，每个特征的初始概率为0.5。此类概率值由GumbelSoftmax技巧[Jang等人，2017]从CTR估计任务的监督信号中参数化和学习。

上述工作为特征场选择或生成有效的特征交互。BP-FIS[Chen等人，2019b]是第一个通过贝叶斯变量选择识别不同用户的重要特征交互的工作，提供了比以前工作更精细的特征交互选择粒度。具体来说，建立了贝叶斯生成模型，其导出的下限可以通过有效的随机梯度变分贝叶斯方法进行优化，以学习参数。

自动化架构搜索

整体架构搜索

最后一类工作研究了深度连接时序分类模型的整体架构。

AutoCTR[Song等人，2020]通过将最先进的CTR估计架构（即MLP、点积和因子分解机）中的代表性结构抽象为虚拟块，设计了两级分层搜索空间，其中这些块以与DART中类似的方式连接为有向无环图（DAG）。外部空间由块之间的连接组成，而内部空间由不同块中的详细超参数组成。

使用进化算法作为搜索算法，并对效率进行了一些优化。

AMER[Zhao等人，2020a]搜索架构，从序列特征（即用户行为）中提取序列表示，并自动同时探索非序列特征之间的不同特征交互。一方面，行为建模的搜索空间包括规范化、激活和层选择（例如卷积、递归、池、注意层）。在验证集上随机抽取几个架构进行评估，并且仅选择顶部架构进行进一步评估。另一方面，通过逐步增加交互顺序来搜索有效的特征交互。候选特征交互由单个非序列特征初始化，并通过与所有可能的非序列特征交互并保持最佳交互和最高验证性能来更新。

六、总结与展望

总结与展望

本文简要回顾了CTR估计任务深度学习模型的发展。通过特征交互算子，深度模型更能够捕捉多领域分类数据中的高阶组合特征模式，并产生更好的预测性能。通过注意力机制、记忆网络或基于检索的方法，可以有效地学习用户行为历史的表示，从而进一步提高预测性能。由于用于CTR估计的深度架构可以是多种多样的，因此已经进行了一些试验，以自动搜索深度架构，使整个过程免提。所有上述进展都是在过去五年内取得的。

尽管CTR评估的深度学习发展迅速，取得了巨大成功，但在这一领域还有一些重大挑战需要解决。

- 深度学习理论。关于连接时序分类模型的设计有很多工作，但很少有工作关注这些模型的深度学习理论，包括样本复杂性分析、特征交互层的学习行为、梯度分析等。
- 表征学习。与其他离散数据（如文本和图形）一样，可以合理预期，多字段分类数据的表示学习（或称为预训练）将大大提高CTR预测性能。然而，到目前为止，关于这一观点的现有工作很少。
- 学习多模态数据。在现代信息系统中，存在项目或浏览环境的各种多媒体元素。因此，设计CTR估计模型来处理多模态数据上的特征交互具有很大的潜力。
- 战略性数据处理。用户历史行为获取策略（如用户行为检索和排序）的最新进展表明，探索数据处理与深度模型设计相结合的方法具有巨大潜力。使这种数据处理易于学习将具有很高的研究价值。

七、学习收获

学习收获

1. 回顾之前学过的浅层模型与深度模型；
2. 更加深刻地理解CTR模型；
3. 指明下一阶段的学习方向。