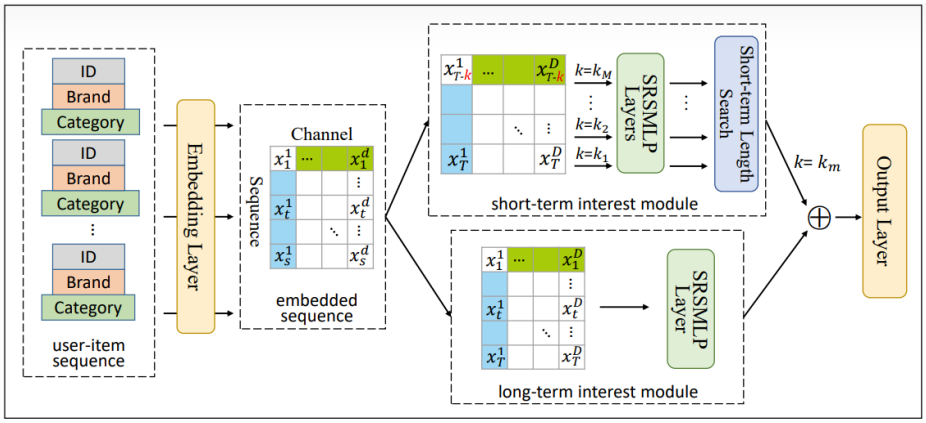
知乎



网易2023-AutoMLP技术，自动选择序列长度，建模推荐系统中的长期和短期序列

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

13 人赞同了该文章

Introduction

序列推荐在推荐系统中扮演着重要角色，用于购买预测、网页推荐和下一个兴趣点推荐等场景。它利用用户过去的行动模式建立序列依赖模型，从而比静态推荐模型更优。

现有序列推荐方法常用RNN或带注意力机制+的RNN建模item间序列依赖，但面临长序列时易退化，需借助LSTM/GRU等技术。Transformer虽表现优异，但对序列顺序不敏感，需辅以定位嵌入学习序列信息，且其计算复杂度随序列长度+呈平方增加。此外，用户长期行为可能异于短期行为，现有工作大多训练两个模型分别捕获两者，但依赖于固定长度的近期交互或固定时段内的选择，缺乏自动化适应性。

- 我们的新推荐模型采用AutoMLP结构，表现优秀，且结构简单、线性复杂度低，超越了最先进的方法。
- 设计自动短期会话长度学习器，搜寻给定上下文相关最优短期兴趣长度，提高模型泛化能力。
- 我们在多个数据集及行业私有数据集上实验证明了该方法的有效性。

Framework

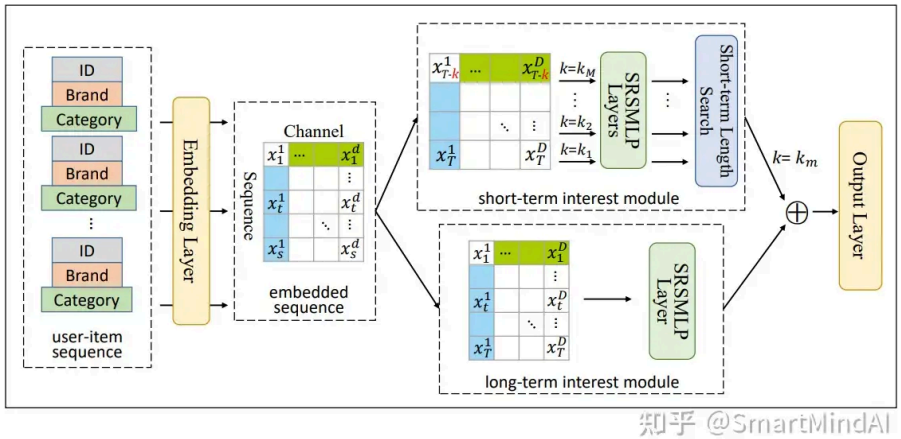


Figure 2: Framework overview of AutoMLP.

Problem Formulation

目标是为用户u的历史交互序列S_u找出推荐下一项的函数$f(S_u)$。

Framework Overview

知乎

序列和最近k次交互。近期交互次数k通过DARTS神经架构搜索算法确定，该算法优化可微的连续放松神经架构搜索空间。最终，通过全连接层融合，分别预测下一个交互项。

Detailed Architecture

Embedding Layer

使用查找表将物品ID与其他特征转换为嵌入向量。全连接层融合不同特性的嵌入向量为具有预定义维度的嵌入向量$\boldsymbol{x_t}$。对于用户交互序列为T时，得到一个大小为$T \times D$的嵌入序列。

Long-term Interest Module

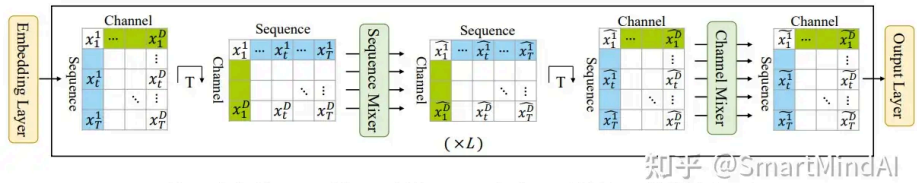


Figure 3: Architecture of Sequential Recommender System MLP Layer (SRSMLP Layer).

$$\hat{\boldsymbol{x}}^{\boldsymbol{d}} = \boldsymbol{x}^{\boldsymbol{d}} + \boldsymbol{W}^{\boldsymbol{2}} g^{l}(\boldsymbol{W}^{\boldsymbol{1}} LayerNorm(\boldsymbol{x}^{\boldsymbol{d}}))$$

其中，$d = 1, 2, \ldots, D$。$\boldsymbol{x}^{\boldsymbol{d}}$是从嵌入表中提取的第$d$个示例。$\hat{\boldsymbol{x}}^{\boldsymbol{d}}$是序列混合器块的输出，$g^{l}$是非线性激活函数在层$l$上的应用，$\boldsymbol{W}^{\boldsymbol{1}}$是一个可学习权重矩阵，表示序列混合器中的第一全连接层，$\boldsymbol{W}^{\boldsymbol{2}}$是一个可学习权重矩阵，表示序列混合器中的第二全连接层，R_s是序列混合器的可调隐藏大小。我们使用了与MLP-mixer相同的层标准化和残差连接。

$$\hat{\boldsymbol{x}}^{\boldsymbol{t}} = \boldsymbol{x}^{\boldsymbol{t}} + \boldsymbol{W}^{\boldsymbol{4}} g^{l}(\boldsymbol{W}^{\boldsymbol{3}} LayerNorm(\boldsymbol{x}^{\boldsymbol{t}}))$$

文中方程如下：$t = 1, 2, \cdots, T$，$\boldsymbol{x}^{\boldsymbol{t}}$是输入，$\hat{\boldsymbol{x}}^{\boldsymbol{t}}$是通道混合器的输出，它们之间存在相关性。此外，$\boldsymbol{W}^{\boldsymbol{3}} \in \mathbb{R}^{R_c \times D}$和$\boldsymbol{W}^{\boldsymbol{4}} \in \mathbb{R}^{D \times R_c}$分别是第一层和第二层全连接层的学习权重，其中$R_c$为通道混合器可调大小的隐藏层。

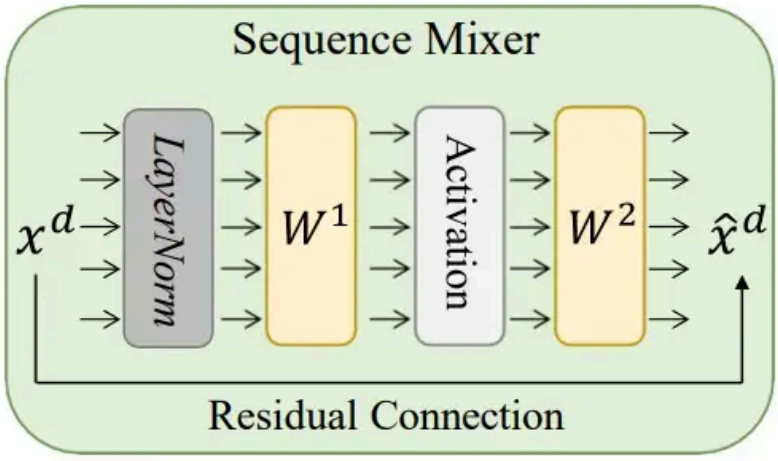


Figure 4: Sequence mixer. Channel mixer has a similar architecture, except for the input x^d and output $\hat{x}^d$.

Short-term Interest Module.

针对这个问题，我们提出了一种基于贪心算法的策略。该策略通过在不同长度上多次训练，并使用反向传播和梯度下降来调整权重，从而得到局部最优解。这样不仅可以提高模型的性能，而且大大减少了计算时间。

知乎

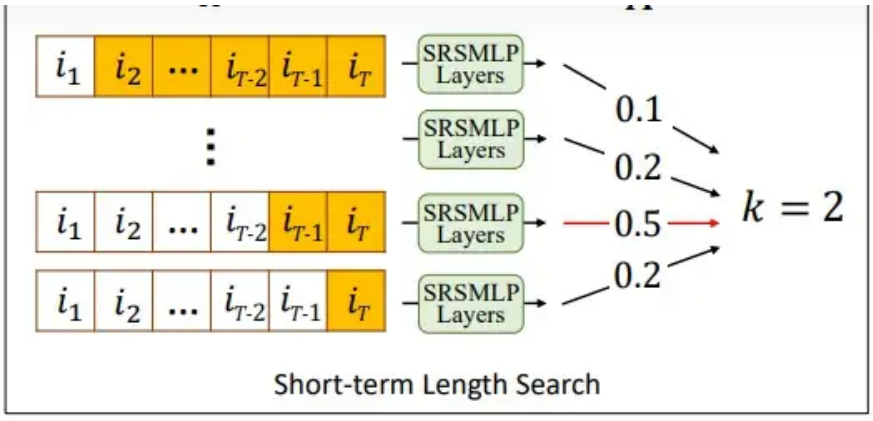


Figure 5: Short-term session length search process. K is the set of candidate lengths, represented by the highlight part.

Output Layer

找到最优短期序列长度后，从长期和短期利率模块获取输出。接下来，使用全连接层学习其联合表示如下： $\boldsymbol{h_T = W^o LayerNorm(x_T^s; x_T^l)) + b^o}$ $h_T = W^o[x_T^s, x_T^l] + b^o$是时间步数为$T$的最终隐藏表示，其中$W^o$是一个可学习的权重矩阵，它将组合后的$x_T^s$和$x_T^l$投影到更低维度的表示，$W^o \in \mathbb{R}^{D \times 2D}$，$b^o \in \mathbb{R}^D$是可学习的偏置向量。最后，$h_T$ 将用于预测用户的下一个交互。

Training and Inference

训练损失函数⁺为交叉熵⁺。

$$\mathcal{L} = -\sum_{S_u \in S} \sum_{t \in [1,\dots,T]} [log(\sigma(r_{i_t,t})) + \sum_{j \notin S_u} log(1 - \sigma(r_{i_j,t}))]$$

Training

$$\min_{A} \mathcal{L}_{val}(W^*(A), A)$$
$$s.t.\ W^*(A) = \arg\min_{W} \mathcal{L}_{train}(W, A)$$

优化内部参数一般涉及大量计算。为优化A，需先训练W至收敛；或使用一步近似法。

$$W^*(A) \approx W - \xi \nabla_W \mathcal{L}_{train}(W, A)$$

使用学习率为ξ的近似方法⁺，通过单个训练步骤来逼近$W^*(A)$，而不需要彻底训练。该方法的搜索阶段详细训练算法在算法.中给出。

Input: 嵌入表 x 和实际交互序列S

Output: 已经学得的参数 W^* 和 A

在搜索阶段，我们把所有候选的短期序列都融入到AutoMLP框架中，其中最短的子优化序列可能会对模型性能产生负面影响。因此，我们需要用最优的短期序列来重训AutoMLP。 具体来说，在优化搜索阶段之后，我们得到了一个很好地训练过的A^*。我们选择最高α_m的最优短时长度，并且丢弃其他所有的候选长度。然后，我们根据长期兴趣模块和最优短期兴趣模块的输出按照常见的深度推荐系统训练范式（即通过梯度下降优化损失$\min_W \mathcal{L}_{train}(W)$）来重训AutoMLP框架。

Inference

我们使用了在序列推荐中常见的推理过程来计算模型输出与所有候选内容之间的余弦相似度⁺。具体来说，我们首先获取输出层的隐藏表示hT作为整个序列的最终隐藏表示，然后计算hT与所有候

知乎

Complexity Analysis

AutoMLP的复杂度主要由两个MLP块-序列混合器和通道混合器决定。其中，序列混合器的复杂度为$O(T \times R_s + R_s \times T)$，而通道混合器的复杂度为$O(D)$。由于短期模块具有相同的结构，除了序列长度之外，我们还设定了短期序列混合器的复杂度为$O(k)$和短期通道混合器的复杂度为$O(D)$。因此，总复杂度为$O(T + D + k)$，这是一个由三个决定变量决定的线性函数。

Experiments

Datasets

MovieLens 是推荐系统常用的数据集，包括用户对电影的评分和时间戳；Amazon Beauty是一个包含商品评论和评级的数据集，我们使用类别；我们删除了在两个数据集中交互次数少于10次的用户；

Table 1: Statistics of the datasets.

Data	MovieLens	Amazon Beauty
# Interactions	1,000,209	2,023,070
# Items	3,952	249,274
# Users	6,040	1,210,271

Experimental Setting

接下来，我们将按照在序列推荐中常见的场景，即留一评估来进行下一个项预测。具体来说，我们将使用用户最后的交互作为测试集，其次近的交互作为验证集，其余所有交互作为训练集。此外，对于每个真实内容的100个负样本，我们将对其进行采样，并基于其流行度进行评估。

Evaluation Metrics

本文使用HR、NDCG和MRR这三种常见的评估指标来衡量推荐系统的性能。其中HR表示前N个推荐结果中包含目标项的比例；NDCG衡量的是目标项在前N推荐结果中的位置；MRR则是前N个推荐结果中目标项的排名倒数。

Baselines

与几种流行的序列推荐模型（包括MC基于、RNN基于和Transformer基于）相比，我们提出的方法更有效。这些模型根据设计动机和适应的技术分为三类：一类是通用的序列推荐系统。

- FPMC 代表先前在序列推荐系统中的工作，其中使用马尔可夫链来模型项之间的顺序依赖性。
- Hidasi等人提出了使用RNN进行序列推荐，并通过引入GRU和PPR改进了基础的RNN模型。
- Sun等人使用双向自注意力技术来改进基于Transformer的序列推荐，并且有研究显示BERT4Rec优于传统的自注意力方法（例如SASRec）。

LSTM-LSTM模型结合了长短期记忆网络的优点，能够在保持长期序列依赖性的同时捕捉到短时序列的依赖性。

功能级序列推荐系统是一种利用内容特征信息进行下一步预测的序列推荐模型。

- GRU4Rec+：改进版的GRU4Rec，通过使用项特征在场景丰富的环境下进行更准确的预测，并使用并行RNN来提高效率。

知乎

Implementation Details

我们的基准方法和AutoMLP基于RecBole框架实现，后者是一个开源的推荐系统库，可以提供公平的性能评估环境。基准方法采用原始论文推荐的超参数，对于未指定的参数，通过网格搜索进行调优，选择验证集表现最优的参数。我们采用早停策略，在接下来的10个epoch中，若验证性能无提升则停止训练，使用当前最佳验证性能模型。对于大模型如BERT4Rec，考虑到其高空间复杂度可能超过GPU内存限制，我们定义了 空间效率 设置，限制搜索范围在我们的32GB GPU内存范围内。优化器采用Adam，学习率为1e-3，动量参数分别为0.9和0.999。

Overall Performance

Table 2: Overall performance comparison. Best performances are bold, next best performances are underlined

Methods	MovieLens			Beauty			Param
Metrics	MRR@10	NDCG@10	HR@10	MRR@10	NDCG@10	HR@10	
FPMC	0.2453	0.3088	0.5156	0.0991	0.1251	0.2098	100 M
GRU4Rec	0.3893	0.4553	0.6666	0.1162	0.1435	0.2324	33.5 M
BERT4Rec	0.3535	0.4289	0.6695	0.0907	0.1198	0.2154	17 M
NextItNet	0.2085	0.2642	0.4455	0.1087	0.1393	0.2385	16.8 M
GRU4Rec+	0.3736	0.4412	0.6578	0.1325	0.1638	0.2657	34.1 M
FDSA	0.3725	0.4409	0.6594	0.1305	0.1595	0.2536	34.3 M
AutoMLP	**0.3912***	**0.4593***	**0.6767***	**0.1438***	**0.1754***	**0.2779***	16.8 M

"*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the strongest baseline.
"Param" refers to the number of trainable model parameters on Beauty dataset, and M = million.

- 特征级模型能获得优于不能处理特征的模型的性能，说明特征在序列推荐中的重要性。
- AutoMLP在所有数据集上都保持了竞争性的性能。具体来说，AutoMLP超越了现有长/短序列推荐系统如NextItNet，显示出它不仅是高效的选项，也是准确性方面的有力替代品。
- AutoMLP在序列推荐系统中具有更好的空间效率。

AutoMLP在公共数据集上达到了与流行基准相近的性能，并且具有较低的空间复杂性，证实了其有效性。

Efficiency Analysis

本文研究AutoMLP的时间/空间效率。AutoMLP通过自动化短期兴趣长度搜索来提升推荐系统的性能。对比了AutoMLP和AutoMLP-s，发现AutoMLP-s虽然搜索算法更复杂，但在较小的搜索空间下却能更快地找到最佳性能。随着搜索空间增大，AutoMLP的增长速度较慢，说明其具有更好的可扩展性。使用了图a来比较当候选搜索空间为5时，两种方法的搜索时间；图b则展示了随着搜索空间增大，AutoMLP的增长比例远小于AutoMLP-s，进一步验证了AutoMLP的时间/空间效率优势。

如图所示，AutoMLP相较于最先进的Transformer-based方法，在时间使用和内存消耗方面表现出了更高的训练效率。尽管BERT4Rec通过采用复杂的训练技术如Cloze目标函数和双向自注意力等实现了更平衡的训练效率，但由于其训练成本高，训练效率相对较低。而FDSA则通过结合自注意力和vanilla注意力来实现更好的训练效率，但在训练成本上仍高于AutoMLP。

Hyper-parameters Analysis

如图(a)所示，最优的层数为8；图(b)表明，小的嵌入大小会导致性能下降，而随着嵌入大小增大，AutoMLP能更好地学习用户-物品交互表示。

Conclusion

本文介绍了一种名为AutoMLP的长期短期序列推荐系统，它使用多层感知器（MLP）架构并在公开发布的基准数据集和实际应用的数据集上表现出了与最先进的方法相当的竞争性性能。此外，我们还设计了一个自动短期兴趣长度搜索算法，能够高效地学习最优的短期兴趣长度，并且结合了MLP架构的线性复杂度，因此我们的方法具有更高的效率和更大的可能性进行改进。

论文原文《AutoMLP: Automated MLP for Sequential Recommendations》

知乎

编辑于 2024-02-19 15:31 · IP 属地北京

时间序列分析及应用（书籍） 工业级推荐系统 长短期记忆

▲ 赞同 13 ▼ 添加评论 分享 喜欢 收藏 申请转载 ···

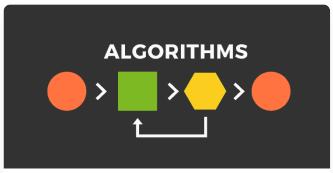
理性发言，友善互动

还没有评论，发表第一个评论吧

推荐阅读

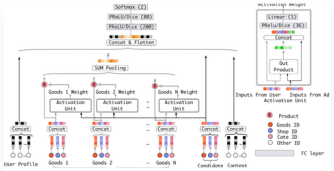
微软算法面试题：如何找最长的增长子序列?

蜗牛学苑



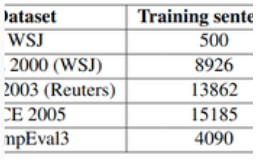
快速排序及其优化超详细解答+代码（真正理解）

进击的st... 发表于数据结构与...



《推荐系统》系列之五：序列推荐

朱勇椿 发表于数据挖掘小...



序列标注方法BIO、BIOS IOB、BILOU、BMEWO

mount... 发表于机器学