软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(12):4838–4850 [doi: 10.13328/j.cnki.jos.006368]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向视频冷启动问题的点击率预估*

章磊敏[1], 董建锋[1], 包翠竹[1], 纪守领[2], 王 勋[1]

[1](浙江工商大学 计算机与信息工程学院, 浙江 杭州 310014)

[2](浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

通信作者: 董建锋, E-mail: dongjf24@gmail.com

摘 要: 视频的点击率预估是视频推荐系统中的重要任务之一, 推荐系统可以根据点击率的预估调整视频推荐顺序以提升视频推荐的效果. 近年来, 随着视频数量的爆炸式增长, 视频推荐的冷启动问题也变得愈发严重. 针对这个问题, 提出了一个新的视频点击率预估模型, 通过使用视频的内容特征以及上下文特征来加强视频点击率预估的效果; 同时, 通过对冷启动场景的模拟训练和基于近邻的替代方法提升模型应对新视频点击率预估的能力. 提出的模型可以同时对旧视频和新视频进行点击率预估. 在两个真实的电视剧(Track_1_series)和电影(Track_2_movies)点击率预估数据集上的实验表明: 提出的模型可以显著改善对旧视频的点击率预估性能, 并在两个数据集上均超过了现有的模型; 对于新视频, 相比于不考虑冷启动问题的模型只能获得0.57左右的AUC性能, 该模型在两个数据集上分别获得0.645和0.615的性能, 表现出针对冷启动问题更好的鲁棒性.

关键词: 视频推荐; 点击率预估; 冷启动问题; 内容特征; 上下文特征

中图法分类号: TP391

中文引用格式: 章磊敏, 董建锋, 包翠竹, 纪守领, 王勋. 面向视频冷启动问题的点击率预估. 软件学报, 2022, 33(12): 4838–4850. http://www.jos.org.cn/1000-9825/6368.htm

英文引用格式: Zhang LM, Dong JF, Bao CZ, Ji SL, Wang X. Click-through Rate Prediction for Video Cold-start Problem. Ruan Jian Xue Bao/Journal of Software, 2022, 33(12): 4838–4850 (in Chinese). http://www.jos.org.cn/1000-9825/6368.htm

Click-through Rate Prediction for Video Cold-start Problem

ZHANG Lei-Min[1], DONG Jian-Feng[1], BAO Cui-Zhu[1], JI Shou-Ling[2], WANG Xun[1]

[1](College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310014, China)

[2](College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: Video click-through rate (CTR) prediction is one of the important tasks in the context of video recommendation. According to click-through prediction, recommendation systems can adjust the order of the recommended video sequence to improve the performance of video recommendation. In recent years, with the explosive growth of videos, the problem of video cold start has become more and more serious. Aim for this problem, a novel video click-through prediction model is proposed which utilizes both the video content features and context features to improve CTR prediction; a simulation training of the cold start scenario and neighbor-based new video replacement method are also proposed to enhance the model's CTR prediction ability for new videos. The proposed model is able to predict CTR for both old and new videos. The experiments on two real-world video CTR datasets (Track_1_series and Track_2_movies) demonstrate the effectiveness of the proposed method. Specifically, the proposed model using both video content and contextual information improves the performance of CTR prediction for old videos, which also outperforms the existing models on both datasets. Additionally, for new videos, a baseline model without considering the cold start problem achieves an AUC score of about 0.57. By contrast, the proposed model gives much better AUC scores of 0.645 and 0.615 on Track_1_series and Track_2_movies, respectively, showing the better robustness to the cold start problem.

Key words: video recommendation; click-through rate prediction; cold-start problem; content feature; context feature

* 基金项目: 国家自然科学基金(61902347); 浙江省自然科学基金(LQ19F020002, LGF21F020010)
收稿时间: 2021-01-09; 修改时间: 2021-04-11; 采用时间: 2021-05-06; jos 在线出版时间: 2021-11-24

随着互联网技术和智能终端设备的快速发展, 观看在线视频已经成为大众分享日常生活、获取信息、娱乐的重要媒介. 除了优酷、爱奇艺、腾讯视频等传统视频分享平台外, 抖音、快手等短视频分享平台近几年也迅速崛起, 吸引了数量庞大的用户上传和观看视频, 平台视频数量也快速增长. 以快手为例, 截止到 2019 年 1 月, 快手短视频库存已达 80 亿个; 2018 年, 快手每天上传的短视频超过 1 500 万个 (https://www.donews.com/news/detail/2/3035932.html). 面对这些海量的视频, 如何准确而有效地从中找到用户感兴趣的视频, 是巨大的挑战. 为了缓解海量数据带来的信息过载, 视频推荐系统被广泛应用于各大视频分享平台. 推荐系统通过分析用户的观看数据, 从海量视频中挑选用户可能感兴趣的候选视频, 并将其推送给用户.

视频的点击率预估是视频推荐中的主流方法之一, 其根据用户的视频观看记录来预估用户点击某个视频的概率(点击率). 基于视频点击率预估模型, 视频推荐系统可根据预估的视频点击率对视频的推荐顺序进行调整, 将预估点击率高的视频进行优先推荐以提高推荐的效果. 但主流的视频点击率预估模型通常依赖于用户行为数据, 无法处理视频的冷启动问题[1,2], 即: 当有新视频被上传时, 由于该视频缺少用户的交互信息, 因此无法将该新视频推荐给相关用户. 对于主流的视频分享平台来说, 每天都有海量的新视频被上传, 这也导致视频的冷启动问题愈发严重. 近年来, 深度学习技术不仅在计算机视觉和自然语言处理等领域使用, 也被广泛应用于推荐领域[3−7]. 虽然基于深度学习的方法表现出良好的视频点击率预估性能, 但其仍未能很好地解决视频的冷启动问题.

为了解决视频冷启动问题, Sachdeva 等人利用视频的文本标注信息, 例如视频的文本简介、标题等进行建模[8]. 但并不是所有视频都包含文本信息, 文本信息不易获取, 人工标注又费时费力; 另一方面, 视频的文本信息可能并不准确可靠, 一些视频上传者为了吸引用户, 可能会编写虚假的标题和视频介绍(例如标题党现象). 视频的文本信息并不能准确地反映视频的语义内容且不易获取, 利用文本信息解决冷启动的方式不能取得令人满意的推荐效果. 另一些工作[9,10]从视频自身包含的音频和视觉信息出发, 对视频内容进行建模分析. 相比于视频的文本信息, 视频的音频和视觉信息不易被篡改, 能够直接有效地反映视频本身的信息, 表现出更好的性能. 基于音频和视觉内容的分析的方法能在一定程度缓解视频的冷启动问题, 但由于视频的音频和视觉信息无法很好地反映视频被用户喜好的程度, 其性能也受到了一定程度的制约. 因此, 本文认为, 仅仅利用视频的内容信息是不够的. 考虑视频分享平台存在大量用户的视频浏览记录, 同时, 这些记录在一定程度上记录了视频被用户喜好的程度. 受无监督的自然语言处理方法的启发, 本文利用无监督的序列建模方法,从用户的视频浏览记录中为每个视频学习一个新的视频特征. 为了便于描述, 本文将通过该方式学到的特征称为视频的上下文特征. 本文设计了两种不同的方式学习视频的上下文特征, 并在所提出的模型中同时对视频的内容特征和上下文特征进行建模, 有效提升了视频点击率预估的性能.

虽然上下文特征可作为视频内容特征的补充, 但对于一个刚上传的新视频, 由于其没有在任何用户的视频浏览记录中出现过, 因此无法获得新视频的上下文特征, 同样也无法解决视频的冷启动问题. 假如能通过某种方法获得新视频的上下文特征, 就可以在一定程度上缓解视频的冷启动问题. 一种直接方式是用全零的特征向量作为上下文特征, 但这会导致模型的训练和预测的不一致, 影响模型对新视频的推荐效果. 针对该问题, 本文提出了一种冷启动场景的模拟训练方法, 该方法在模型的训练过程中, 以一定的概率随机地将旧视频视为新视频, 用全零的特征向量作为新视频的上下文特征, 从而使得训练得到的模型能够更均衡地兼顾旧视频和新视频. 此外, 考虑到两个内容上接近的视频其上下文特征也可能相似, 本文还提出基于近邻的替代方法来获得新视频的上下文特征. 该方法无需改变模型的结构和训练策略, 只需在预测阶段用与新视频内容相近的若干视频的上下文特征来替代新视频的上下文特征, 同样能较好地对新视频的点击率进行预估. 总的来说, 本文的主要贡献如下:

(1) 除了视频的内容特征外, 本文额外使用了从用户的视频浏览记录中学习到的上下文特征. 设计了两种不同的方式学习视频的上下文特征, 并同时使用视频的内容特征和上下文特征, 有效提升了视频点击率预估的性能;

(2) 针对视频的冷启动问题, 本文提出两种方法: 冷启动场景的模拟训练方法和基于近邻的替代方法.这

两种方法无需改变模型的结构, 通过新颖的训练策略和预测方法, 显著提升模型针对视频冷启动问题的鲁棒性. 理论上, 这两种方法可应用到任何点击预测的模型中, 并提升模型对新视频的点击率预估能力;

(3) 在两个真实的电视剧(Track_1_series)和电影(Track_2_movies)视频点击率预估数据集上的实验表明, 本文提出的同时利用视频内容信息和上下文信息的模型可以提升对旧视频的点击率预估. 在 Track_ 1_series 上, 将 AUC(area under curve)性能从 0.712 1 提升到 0.739 5; 在 Track_2_movies 上, 将 AUC 从 0.680 3 提高到 0.689 7; 并在两个数据集上超过了现有模型. 此外, 对于新视频的推荐场景, 相比于不考虑冷启动问题的模型只能获得 0.57 左右的 AUC 性能, 本文所提出方法在两个数据集上分别获得 0.645 和 0.615 的 AUC 性能, 表现出针对视频冷启动问题更好的鲁棒性.

1 相关工作

推荐系统被广泛应用于各个领域, 用以解决信息过载问题. 在视频推荐中, 视频的点击率预估是主流方法之一, 其根据用户的视频观看记录来预估用户点击某个视频的概率. 主流方法之一是基于会话(session)来预测点击率并进行推荐[11−13]. 考虑到用户隐私等原因, 基于会话的方法一般只使用用户的历史行为进行建模, 比如浏览记录, 但无法获取用户的基本信息, 比如性别、年龄等. Hidasi 等人[13]将点击率预测视为二分类问题, 其将用户的浏览数据看作序列, 使用门控循环单元(GRU)对浏览数据进行编码, 并基于多层感知机预测浏览数据和候选视频是否相关, 从而判断用户是否会点击候选视频. He 等人[14]较早将注意力机制引入推荐模型, 从而显著提升推荐的性能. 基于类似的思想, Kang 等人[11]利用自注意力机制[15]来捕获序列中的浏览视频之间的关系并对其进行编码, 同时还添加了位置编码, 以弥补自注意力机制忽略时序信息的缺点. 在文献[16]中, 赵等人利用词向量[17]模型分析用户浏览历史序列, 将视频映射成特定维度的特征向量, 并基于特征向量计算视频间的相关性, 最终根据与用户观看视频相关性高低来判断用户是否会点击特定视频. 为了更充分地挖掘浏览记录中数据之间的依赖关系, 图神经网络近年来也被广泛应用于推荐中[5,18,19]. 其中, Wu 等人[5]提出用图神经网络来处理用户的浏览记录. 更进一步地, Wei 等人[18]在图神经网络上挖掘用户的隐式反馈, 从而提升推荐的性能. 虽然基于会话的方法对于旧视频推荐有较好的效果, 但其主要依赖于用户的交互数据, 无法处理没有交互数据的新上传视频.

基于内容的方法从视频本身的内容评估不同视频之间的相似性, 该类方法在预测时不依赖于用户的交互数据, 因此在一定程度上可以缓解视频的冷启动问题. 比如, 王娜等人[20]利用视频的标签来计算视频之间的相似度, 为用户推荐与其观看过的视频中相似度较高的视频. 但该方法依赖于标签质量, 其性能会因标签质量降低而受到影响. Yang 等人[21]利用视频的色调、运动强度特征估计视频之间的相关性, 并将与用户观看过的视频相关的视频推荐给用户. 基于类似的思想, 在用户交互信息缺失时, Van 等人[22]利用神经网络抽取音频信号的内容特征, 来更好地对新歌曲进行推荐. 正如在引言中提到的, 每天都会有大量的新视频被上传到视频分享平台, 因此视频推荐的冷启动问题尤为严重. 针对这一问题, Hulu 在国际多媒体顶级会议 ACM MM 2018 上举办了基于视频内容的推荐挑战赛. 挑战赛提供了用预训练过的神经网络模型提取的视频视觉和音频特征, 让参赛者利用这些特征来挖掘视频之间的相关性. 针对这一任务, Dong 等人[23]通过特征重学习的方法学习到更符合视频推荐的内容特征, 而不使用直接提供的内容特征. Chen 等人[24]通过挖掘视频之间的二阶相关性, 更精准地衡量视频的相关性. 2019 年的 ACMMM 会议上, Hulu 再次举办了基于内容的视频点击率预估比赛[25]. 针对这个比赛, Xu 等人[26]提出了 TSE 模型, 其在视频特征向量中自适应地引入一个时间衰退因子, 在计算候选视频与视频浏览历史中的视频的相似度时, 更多考虑用户最近观看的视频, 相对地减弱较远的视频的影响. Wang 等人[27]则把视频点击率预估看作视频相关性问题, 提出了 CMN 模型, 通过计算用户浏览历史中的视频和目标视频的相关度来取代视频点击率预估的结果. 在文献[27]中, Wang 等人在深度兴趣网络 DIN[4]的基础上提出了 REDIN 模型. 该模型额外加入一个辅助任务, 通过将相关视频之间的距离在公共空间上拉近, 从而使得视频特征更适应点击率预估的任务. Chen 等人[3]提出了 MMDIN 模型, 该模型通过两层的注意力层来捕获用户浏览历史中的视频与目标视频之间的相关性, 通过模型的预训练来提升点击预估的能力.

不同于上述文献从模型结构的角度来缓解冷启动问题, 本文从模型的训练和预测角度提出了两种新的方法, 通过新颖的训练策略和预测方法, 显著提升了模型针对视频冷启动问题的鲁棒性.

2 本文方法

给定用户的视频浏览历史, 即观看过的 n 个视频序列 $V=\{v_1,v_2,\ldots,v_n\}$和候选视频 v_c 作为输入, 视频的点击率预估任务要求根据用户浏览历史预测用户会点击给定候选视频的概率 $p(V,v_c)$. 概率越高, 表明该用户对候选视频感兴趣的可能更高; 反之, 则越低. 正如引言中提到的, 视频的文本标注信息(元数据描述)容易被篡改, 因此本文没有使用视频的元数据描述, 而是利用更可靠的视觉和音频的内容特征. 基于点击率预估的视频推荐系统则会根据该预测的概率, 将概率高的若干视频推荐给目标用户, 从而实现个性化推荐. 针对视频的冷启动问题, 本文设计了基于视频内容和上下文特征的视频点击率预估模型, 并提出冷启动场景的模拟训练和基于近邻的替代方法缓解模型处理新视频的能力. 本节接下来依次介绍如何获得视频的特征表达、点击率预估模型的结构以及冷启动场景的模拟训练和基于近邻的替代方法.

2.1 视频的特征表示方法

2.1.1 视频的内容特征

对于一个视频来说, 其本身视觉内容和音频内容蕴含了视频丰富的信息, 这些信息对于视频的推荐是非常有帮助的. 本文直接使用了数据集提供的视觉特征和音频特征.

视觉特征通过在 ImageNet 数据集预先训练好的 Inception 模型进行抽取. 具体来说, 给定的一个视频 v, 每隔 1 秒从视频中提取一个视频帧, 并将提取的视频帧送到预训练好的 Inception-v3 模型, 将在分类层之前的最后一个隐藏层的 ReLU 激活的输出作为视频帧的内容特征. 因此, 视频的每一帧都表示成 2048 维的特征向量. 进一步地, 对视频帧的特征从时间维度进行平均池化, 得到视频级的视觉特征向量, 并通过主成分分析(PCA)将其降维到 64 维.

音频特征则通过在 AudioSet 数据集预训练过的 VGGish 模型[28]进行抽取. 具体地, 给定的一个视频 v, 首先用 FFmpeg 多媒体处理工具提取视频中的音频, 并将其切割成 0.96 秒不重叠的音频片段, 然后分别将其输入到预训练好的 VGGish 抽取特征, 将在分类层前的最后一个隐藏层的输出作为音频片段的音频特征(特征维度为 512). 同样地, 音频片段的特征从时间维度进行平均池化, 得到视频级的音频特征向量, 并通过 PCA 将其降维到 64 维.

2.1.2 视频的上下文特征

除了视频的内容特征外, 本文还使用视频的上下文特征来表示视频. 本文中的上下文是指视频在视频浏览记录中与其他视频的上下文关系. 一个用户的视频浏览记录中同时出现的视频很有可能具有一定的相关性, 比如一个用户可能会连续看完《指环王》三部曲. 受此启发, 本文基于大量的视频浏览记录学习得到上下文特征, 该特征可以在一定程度上反映视频之间的相关性. 同时, 用户可能会对一部与之前看过的视频相关的视频感兴趣, 因此本文认为, 上下文特征能增强推荐效果. 本文设计了基于两种自然语言处理的无监督模型来提取视频的上下文特征, 分别是词向量(Word2vec)方法和基于变化的双向编码器表示(BERT)方法.

Word2vec 词向量方法[17]将每一个词用一个稠密向量进行表示, 使得语义上相似的词在词向量空间中距离相近, 反之则远离. 根据输入输出的不同, 词向量方法可以分成两种: 跳字模型和连续词袋模型. 跳字模型(skip-gram)是用一个词语作为输入, 来预测它周围的上下文; 连续词袋模型(continuous bag of words, CBOW)则用一个词语的上下文作为输入, 预测这个词语本身. 在 Word2vec 词向量训练中, 句子的每个单词用一个编号进行表示, 并在大量文本语料库中进行训练. 训练完成后, 每个单词可以表示成一个词特征向量. 类似地, 本文把用户的视频浏览历史当成一个句子, 每个视频编号当作单词进行训练, 并用所有的用户浏览记录训练词向量模型. 通过大量浏览记录学习得到每个视频的上下文特征, 该特征可以在一定程度上反映视频之间的相关性. 在初步实验中发现, 连续词袋模型相比于跳字模型表现出更好的性能. 因此, 本文最终使用连续词袋模型获取视频的上下文特征. 在下文中, 通过该方式得到的特征被称为 Word2vec 特征.

基于变化的双向编码器表示(BERT)是近年来谷歌提出的自然语言处理模型[29], 在多个不同的自然语言处理相关的任务上都达到了先进的水平. 不同于循环神经网络(recurrent neural network, RNN), BERT 可以避免词之间的相互关系随距离递减的问题, 从而表现出更好的序列学习能力. 受文献[30,31]启发, 本文利用 BERT 模型学习视频的上下文特征, 具体的预训练模型如图 1 所示. BERT 提供了多种训练方式, 本文用掩码的语言模型(masked LM)进行训练, 其从输入用户浏览记录中随机去掉一个视频, 让模型通过上下文信息预测去掉的视频, 最终得到的视频特征能够反映其与其他视频的上下文关系. 具体地, 模型的输入为用户的视频浏览记录, 以概率 p 随机地将输入序列中的一个视频去掉并用特殊符号掩码([*mask*])表示, 然后用一个视频嵌入层(embedding layer)和位置嵌入层对视频进行编码. 在图 1 中, f_i 和 p_i 分别表示视频 v_i 的视频嵌入特征和第 i 个位置的位置嵌入特征. 对于视频 v_i, 首先用独热编码(one-hot encoding)表示其视频 id(独热编码的维度等于所有视频的数量, 每一维代表一个特定的视频), 用视频嵌入层将其表示成 d 维的视频嵌入特征向量; 视频的位置信息同样先用独热编码对进行编码, 并用位置嵌入层将其表示成 d 维的位置嵌入特征向量. 值得注意是:本文将输入 transformer 的视频嵌入特征和位置嵌入特征的维度设置成较小的 64 维, 从而使得模型的复杂度不至于过于庞大. 进一步地, 两个嵌入向量进行相加并做 $L2$ 归一化处理后输入到两层的 Transformer 模块, 最终通过一个全连接层和 *Softmax* 层预测被去掉视频的编号. 模型通过交叉熵损失函数进行预测, 模型预训练完成后, 视频嵌入特征和位置嵌入特征的和作为视频的上下文特征. 为了便于描述, 本文将该方式得到的特征称为 BERT 特征.

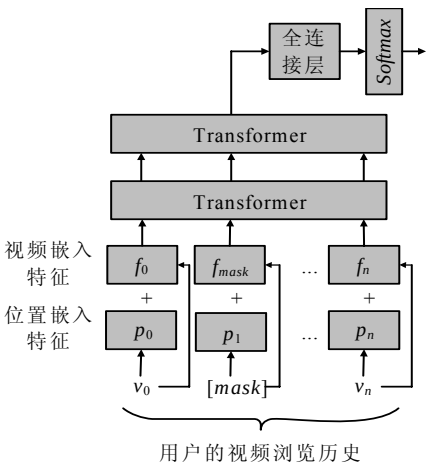


图 1 用于视频表示的 BERT 预训练模型结构

2.2 模 型

2.2.1 模型结构

本文模型是基于 Wang 等人在文献[27]里提出的相关性增强的深度兴趣网络(REDIN)进行改进. 为了缓解冷启动问题, 该模型只是用视频的内容特征对视频进行表达, 并通过额外的相关损失函数约束低层特征之间的相关性. 考虑视频的上下文信息蕴含视频被用户喜好的程度, 本文额外使用视频的上下文特征对视频进行表达, 并提出了两种方法以提升模型应对新视频点击率预估的能力.

如图 2 所示, 给定用户的视频浏览历史 $V=\{v_1,v_2,\ldots,v_n\}$和候选视频 v_c, 用第 2.1 节的视频特征表示方法得到各个视频的内容特征和上下文特征, 并将两种特征进行拼接输入到网络中. 对于输入的拼接特征, 首先利用一个全连接层对其进行非线性变化, 使得变化后的特征更适合于视频的点击预测任务. 与 REDIN 模型一样, 本文使用注意力模块来聚合用户的视频浏览记录. 对于注意力层的实现, 首先计算候选视频与历史记录中视频的相关度,也就是注意力权重, 然后将历史记录中所有视频的加权特征作为用户的兴趣特征, 该兴趣特征反映了用户对于候选视频的兴趣, 同时作为注意力层的输出. 更正式地, 注意力层输出特征向量表示成:

$$A(V)=\sum_{v_i\in V}a(v_i',v_c')v_i' \tag{1}$$

其中, v_i' 和 v_c' 分别表示 v_i 和 v_c 通过全连接层变换后的视频特征向量 $a(v_i',v_c')$. 表示通过多层感知机(MLP)得到的注意力权重. 具体来说, 首先将输入的两个特征 v_i' 和 v_c' 进行拼接, 然后通过带两层隐藏层的 MLP 预测注意力权重. 之后, 将通过注意力层的特征向量 $A(V)$和变换后的候选视频特征进行拼接, 并将其输入到含两层隐藏层的多层感知机进行二分类. 二分类模型中的隐藏层后跟一个 *ReLU* 非线性激活函数和 dropout, 最后的输出层通过一个 *Sigmoid* 激活函数将输出分数归一化到 0~1 之间, 最终的输出记为 $p'(V,v_c)$. 该模型基于大量历史数据进行端到端的训练, 通过学习使模型具备对点击概率 $p'(V,v_c)$进行预测的能力, 具体训练方式见下节.

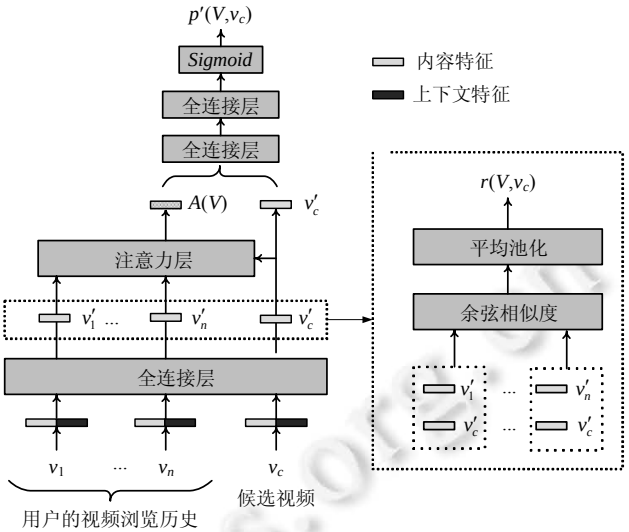


图 2 相关性增强的深度兴趣网络

2.2.2 模型训练

对于模型的训练, 除了使用在点击率预估模型中常用的二分类交叉熵损失函数(binary cross entropy loss), 本文还使用了在文献[23]中引入的视频相关性损失作为额外的约束. 该约束作用于模型中间层的输出, 即变换后的视频特征 v'_*, 通过约束使得变换后的特征仍保留视频的拓扑属性, 使得相似的视频在变换后的特征空间具有较小的距离, 反之则具有较大的距离. 此外, 额外添加的损失函数让梯度更容易传到前面的层, 同时助于后续层的训练. 具体地, 对于一条用户的视频观看记录(V,v_c,y), 其损失函数定义为:

$$L(V,v_c,y)=-[y\log p'(V,v_c)+(1-y)\log(1-p'(V,v_c))]+\alpha[y\max(0,m_1-r(V,v_c))+(1-y)\max(0,r(V,v_c)-m_2)] \quad (2)$$

公式(2)的损失函数由两部分组成: 前半部分表示交叉熵损失函数, 后半部分表示相关性损失函数. 其中: y 表示用户是否点击了候选视频, 其值为 1 表示点击, 0 表示不点击; α表示平衡两个不同损失的权重, 其取值大于等于 0, α小于 1 代表前半部分损失更重要, 而α大于 1 代表后半部分损失更重要; $r(V,v_c)$表示整个视频记录 V 和候选视频 v_c 的整体相关性, 实际使用候选视频与浏览记录 V 中的每一个视频相关度的平均值作为最终的整体相关性, 计算公式如下:

$$r(V,v_c)=\frac{1}{n}\sum_{v_i\in V} cs(v'_i,v'_c) \quad (3)$$

其中: $cs(\cdot,\cdot)$表示特征之间的余弦相似度, 相似度越高, 表明两个视频越相关; n 表示用户浏览记录中的视频个数. m_1 和 m_2 为常数阈值并满足 $m_1>m_2$, 对于用户点击了候选样本的数据(即正样本 y=1), 本文认为整个视频记录 V 和候选视频 v_c 应该具有较高的相关性, 该目标通过最小化 $y\max(0,m_1-r(V,v_c))$保证, 使得整体相关大于阈值 m_1; 否则, 产生损失对模型进行惩罚. 同样地, 对用户未点击候选样本的数据(即负样本 y=0), 希望整个视频记录 V和候选视频 v_c的整体相关性较小, 该目标通过最小化$(1-y)\max(0,r(V,v_c)-m_2)$保证, 使得整体相关大于阈值 m_2. 最后, 通过最小化公式(2), 在所有训练样本上的损失和来训练模型.

2.2.3 模型预测

考虑到如果候选视频与用户观看过的视频序列中的视频相关性比较高, 用户很有可能对该候选视频感兴趣, 因此在模型训练完成之后, 除了考虑多层感知机的输出, 还额外利用候选视频和视频浏览历史序列的相关性来预测视频被点击的概率. 具体来说, 对于观看过 $V=\{v_1,v_2,\ldots,v_n\}$视频序列的用户, 其点击候选视频 v_c 的概率 $p(V,v_c)$计算如下:

$$p(V,v_c)=\beta p'(V,v_c)+(1-\beta)r(V,v_c) \quad (4)$$

其中, β为权衡常数值.

2.3 针对视频冷启动问题的处理方法

上述介绍模型以视频的内容特征和上下文特征作为输入, 因此需要同时获取这两种特征才能对视频的点击率进行预估. 对于一个新上传的视频, 其视觉和音频的内容特征较容易获取. 但由于该视频是新上传的, 未在任何用户的视频浏览记录中出现, 因而无法获得新视频真实的上下文特征, 这导致模型无法处理新上传的视频. 对于这一问题, 一种简单的解决方法是用全零向量替代视频的上下文特征, 但这种方式会导致训练和预测不一致, 从而导致模型性能迅速下降. 针对该问题, 本文提出两种处理方法, 分别是模拟冷启动场景的训练方法和基于近邻的替代方法, 以提升模型对新视频点击率预估的鲁棒性.

2.3.1 模拟冷启动场景的训练方式

视频点击率预估模型通常基于大量用户的浏览记录进行训练, 同时浏览记录中的每个视频都作为旧视频(旧视频的内容特征和上下文特征可同时获得)进行训练. 由于这种方式未在训练过程中考虑新上传视频的情况, 因此基于这种训练方法得到的模型不能很好地处理新视频的点击率预估. 为了缓解这一问题, 本文提出了模拟冷启动场景的训练方法, 该方法在训练过程中以一定概率将出现在用户浏览记录中的最后一个视频当作新视频. 通过在训练过程中模拟新视频出现的情况, 从而提高模型对于新视频的点击率预估能力.

具体来说, 给定一个用户浏览记录$\{v_1,v_2,\ldots,v_n,v_c\}$, 在训练过程中, 以概率 q 将视频 v_c 当作新上传的视频, 以概率 $1-q$ 仍将 v_c 视频当作旧视频看待. 由于新视频的上下文特征无法获取, 因此用全零向量替代视频的上下文特征; 而对于旧视频, 则使用第 2.1.2 小节中描述的上下文特征. 通过这种方式训练的模型, 能够更好地兼顾新视频和旧视频的点击率预估能力. 在预测阶段, 如果给定的视频为新视频, 本文用全零的特征向量作为其上下文特征; 如果是旧视频, 则使用真实的上下文特征. 值得注意的是: 在本文提出的模拟冷启动场景的训练方法中, 概率 q 为超参数. 当 q 为 0 时, 模拟冷启动场景的训练就退化成普通的训练方法, 即训练过程做所有的视频都为旧视频; 而 q 为 1 时, 将所有用户浏览记录中的候选视频都当成新视频. 为了让模型同时兼顾新视频和旧视频, 本文将 q 设为 0.5. 在第 3.4.1 小节的实验给出了具体超参数 q 对模型性能的影响.

2.3.2 基于近邻的替代方法

考虑到两个内容上接近的视频, 其上下文特征也可能相似, 本文提出的第 2 种方法是利用新视频的近邻视频作为辅助来帮助模型对新视频进行点击率预估. 具体来说, 在预测阶段, 给定一个新视频, 根据视频的内容特征计算其和训练集中其他所有视频的余弦相似度, 并将相似度最高的 k 个视频的上下文特征进行平均池化, 池化后的特征作为新视频的上下文特征, 而保持内容特征不变. 不同于模拟冷启动场景的训练方法通过重新训练模型来提升模型对于新视频的点击率预估能力, 基于近邻的替代方法无需重新训练模型, 仅仅通过改变模型的预测方式提升对新视频的点击率预估能力.

3 实验结果与分析

在本节中, 我们将对本文提出方法的有效性进行验证. 第 3.1 节介绍实验的基本设置, 包括采用的实验数据集、性能指标以及实现细节. 在第 3.2 节, 本文进行了不考虑视频冷启动的点击率预估实验, 首先对本文模型进行消融实验, 并与其他已有模型进行性能比较. 第 3.3 节展示了考虑视频冷启动的点击率预估实验, 即被推荐的候选视频为新视频; 在实验中, 我们首先测试了模型的超参数对性能的影响, 并与基线方法比较来证明所提出的模拟冷启动场景的训练方法和基于近邻的替代方法对模型冷启动问题的有效性.

3.1 实验设置

- 数据集:

本文使用了两个真实的视频点击率预估数据集, Track_1_series 和 Track_2_movies[25], 前者是电视剧视频, 后者是电影视频. 两个数据集均来自 HULU 平台真实用户的浏览记录, 因此在这两个数据集的性能好坏也能一定程度上反映模型在真实应用中的表现. 数据集的每条序列都是以$\{v_1,v_2,\ldots,v_n,v_c,y\}$的形式给出, 其中: $v_1,v_2,\ldots,v_n$ 表示用户的视频浏览历史; v_c 是候选视频; y 表示用户在浏览了视频浏览历史中的视频后是否点击了

击候选视频, 其值为 1 表示点击, 0 表示不点击. 由于数据集的测试集没有公开标注信息, 本文的实验性能均在验证集上进行测试. Track_1_series 数据集共有 2 642 个不同的视频, 每条序列的用户浏览视频数量为 10 个, 其中, 训练集有 5 221 221 条用户历史序列, 验证集上则有 931 820 用户历史序列. Track_2_movies 数据集共有 6 283 个不同的视频, 每条序列的用户浏览视频数量为 5 个, 其中, 训练集和验证集的用户历史序列分别为 1 123 786 条和 552 577 条.

- 性能指标:

与之前的文献[26,27]一样, 本文采用 AUC 作为性能评价指标. AUC 的数值越大, 表明模型性能越好. 此外, 仿照之前的文献[6], 本文还使用了对数损失函数(LogLoss)作为额外的性能指标. Logloss 越小, 表明模型性能越好.

- 实现细节:

在用 Word2vec 词向量训练视频上下文特征时, 本文使用了 CBOW 模型, 并采用负采样的方法训练, 窗口大小为数据集序列长度, 视频上下文特征维度设为 64 维, 其他采用默认的参数. 在用 BERT 模型训练上下文特征时, 批大小(batch size)设置为 128, 视频嵌入特征和位置嵌入特征的维度都是 64 维, Transformer 层中多头注意力[15]中的头(head)数量设置为 2, 特征维度为 64; 在 Track_1_series 数据集上, 初始学习率为 0.001, 随机置 *mask* 的概率 p 为 0.1; 在 Track_2_movies 数据上, 初始学习率为 0.000 1, 随机置 *mask* 的概率 p 为 0.2. 对于相关增强深度兴趣网络, 单特征输入时维度为 64, 双特征输入时维度为 128, 两个全连接层的维度分别为 512 和 256; 注意力层中, MLP 的两层隐藏层维度分别为 2 048 和 512, 其中, 隐藏层后的非线性激活函数是 *sigmod* 函数. 公式(2)中的 m_1 经验性地设为 0.8, m_2 设为 0.2, 公式(4)中的β设为 0.7; 在训练时, 学习率设置为 0.0001,批大小为 64. 在训练 BERT 模型和相关增强兴趣模型时, 使用 PyTorch 训练框架, Adam 梯度下降算法, 当模型性能在 2 个 epoch 没有提升时, 学习率变为原来的二分之一; 当连续 5 个 epoch 性能没有提升时, 提前结束训练.

3.2 不考虑视频冷启动的点击率预估实验

3.2.1 消融实验

在本实验中, 本文从视频的特征选择、注意力以及损失函数这 3 个角度对模型进行了消融实验. 表 1 展示了在两个数据集上, 本文模型使用不同视频特征的性能比较. 不管是 ACU 还是 Logloss 性能, 当使用单个视频特征时, 使用视频的上下文特征(Word2vec 或 BERT 特征)的模型性能明显优于使用视频内容特征的模型. 比较两个使用不同上下文特征的模型, 两者在两个数据集上的表现并不一致: BERT 特征在 Track_1_series 数据集上的表现比 Word2vec 特征好; 而在 Track_2_movies 数据集上, Word2vec 特征优于 BERT 特征. 我们推测, 这个不一致现象是由于两个数据集中用户浏览的历史序列长度不一样: Track_1_series 数据集的序列长度是 10, 而 Track_2_series 是 5. 由于 BERT 相比于 Word2vec 具有更强的数据拟合能力, 在序列较短的 Track_2_series 数据集上容易出现过拟合, 从而导致性能变差. 表 1 的下半部分展示了同时使用视频内容特征和上下文特征的性能. 实验结果显示, 使用两种特征的模型超过其对应使用单个特征的性能. 这个结果说明, 视频内容特征与上下文特征对于视频的点击率预测具有良好的互补性. 特别是 BERT 特征和音频特征的组合, 在 Track_1_series 上比单纯使用 BERT 特征性能从 0.712 1 的 AUC 性能提高到 0.739 5. 因此, 本文后续实验中采用使用视频内容特征与上下文特征的方案.

为了验证模型中注意力层的有效性, 本文进行了有无注意力层的性能测试. 其中, 无注意力层的模型用平均池化对历史浏览视频的特征进行聚合, 即认为每一个视频的重要程度是一样.

表 2 展示了在 Track_1_ series 数据集上的性能, 其中, ×表明模型不使用注意力层, √表明模型使用注意力层. 可以发现: 不管是 ACU 还是 Logloss 性能, 使用注意力层的模型都明显优于不使用注意力层的模型, 表明了注意力层在模型中的重要性.

表 1 使用不同视频特征的性能比较

(AUC 越高, 表明性能越好; Logloss 越低, 则性能越好)

视频特征	Track_1_series		Track_2_movies	
	AUC	Logloss	AUC	Logloss
语音特征	0.610 5	0.493 6	0.603 9	0.699 2
视觉特征	0.632 8	0.480 7	0.624 5	0.655 6
Word2vec 特征	0.673 9	0.457 0	**0.680 3**	**0.633 0**
BERT 特征	**0.712 1**	**0.449 3**	0.670 3	0.648 5
Word2vec 和语音特征	0.691 4	0.413 6	**0.689 7**	0.638 9
BERT 和语音特征	**0.739 5**	**0.396 5**	0.683 8	0.644 6
Word2vec 和视觉特征	0.687 0	0.409 6	0.688 6	**0.637 0**
BERT 和视觉特征	0.733 4	0.398 4	0.689 5	0.641 5

表 2 注意力层的有效性

视频特征	AUC		Logloss	
	×	√	×	√
Word2vec 和语音特征	0.643 8	**0.691 4**	0.426 3	**0.413 6**
BERT 和语音特征	0.725 0	**0.739 5**	0.415 0	**0.396 5**
Word2vec 和视觉特征	0.644 1	**0.687 0**	0.422 2	**0.409 6**
BERT 和视觉特征	0.725 4	**0.733 4**	0.411 8	**0.398 4**

此外, 本文还对损失函数中相关性损失函数的有效性进行验证. 具体来说, 本文在 Track_1_series 数据集上比较了没有相关性损失函数(α=0)和有相关性损失函数(α=1)的性能差异, 结果见表 3. 对于使用不同视频特征的所有模型, 有相关性损失函数约束的模型明显表现出更好的性能. 该实验表明了相关性损失函数对于视频点击率预测的有效性.

表 3 相关性损失函数的有效性

视频特征	AUC		Logloss	
	α=0	α=1	α=0	α=1
Word2vec 和语音特征	0.638 1	**0.691 4**	0.443 5	**0.413 6**
BERT 和语音特征	0.720 2	**0.739 5**	0.423 2	**0.396 5**
Word2vec 和视觉特征	0.628 2	**0.687 0**	0.449 4	**0.409 6**
BERT 和视觉特征	0.719 8	**0.733 4**	0.427 6	**0.398 4**

3.2.2 与其他模型的对比

为了验证本文提出的模型在不考虑视频冷启动情况下的有效性, 本文和在 Track_1_series 和 Track_2_movies 数据集上性能最好的几个模型进行了比较. 表 4 展示了本文模型和其他先进模型在两个数据集上的性能. TSE, mDIN, REDIN 和 MMDIN 将视频的点击预测任务看作二分类问题, 根据浏览历史的视频和候选视频的相关性来判断用户是否点击候选视频. 这类模型通过端到端的方式进行训练, 模型简单且表现出较好的性能. 其中: TSE 认为用户最近观看的视频相比于很早前观看的视频对于视频的点击率预测更具参考价值, 因此引入了时间衰减系数来增加最近观看视频的权重; REDIN 在 DIN 的基础上加入了内容特征相关性模块约束特征学习, 从而提升了模型的性能; MMDIN 通过两层注意力层来更好地捕获候选视频和视频序列视频中的相关性. 不同于上述模型基于二分类模型, CMN 直接将视频点击预测问题转化成候选视频和历史视频序列中的视频的相关性计算问题, 根据视频间的相关性来推测视频的点击率. 但该模型比较依赖于训练数据的数量, 在训练数据更多的 Track_1_series 数据集上表现出比以上 4 个模型更好的性能; 但在 Track_2_movies 数据集上, 表现比 DIN 和 REDIN 差. 本文用在各个数据集上性能最好的单模型, 即在 Track_1_series 数据集上为使用 BERT 和视觉特征的模型, 在 Track_2_movies 数据集上为使用 Word2vec 和语音特征的模型, 与以上模型进行比较. 表 4 总结了在两个数据集上本文提出的方法与现有方法的性能比较(现有方法没有报告 Logloss 性能, 本文额外报告了该指标的性能以便于后续工作比较). 如表 4 所示, 本文提出的单模型在两个数据集上都超过了已有模型. 这是因为本文提出的模型同时考虑视频的内容特征和上下文特征, 而对比模型主要考虑了视频的内容特征. 这个结果显示了额外地对视频的上下文特征进行建模, 对视频的点击率预估是有帮助的.

表 4 与其他先进模型的性能比较

模型	Track_1_series		Track_2_movies	
	AUC	Logloss	AUC	Logloss
TSE[26]	0.602 0	–	–	–
DIN[4]	0.616 0	–	0.620 0	–
REDIN[27]	0.653 3	–	0.642 8	–
MMDIN[3]	0.672 3	–	0.586 4	–
CMN[27]	0.685 6	–	0.607 1	–
本文模型(单模型)	0.739 5	0.396 5	0.689 7	0.638 9
本文模型(模型融合)	**0.742 5**	**0.372 7**	**0.699 7**	**0.636 6**

此外, 本文还将多个不同的模型进行融合, 即将不同模型的点击预测概率取平均作为最终的概率. 在 Track_1_series 上, 将 BERT 和语音特征的模型以及 BERT 与视觉特征的模型进行了融合, 得到了 0.742 5 的 AUC 性能和 0.372 7 的 Logloss 性能. 在 Track_2_movies 上, 融合基于 word2vec 与语音特征、word2vec 与视觉特征、BERT 与语音特征和 BERT 与视觉特征的 4 个模型达到了 0.699 7 的 AUC 性能和 0.636 6 的 Logloss 性能.在两个数据集上, 融合后的模型都表现出比单模型更好的性能, 这说明模型融合是有帮助的.

3.3 考虑视频冷启动的点击率预估实验

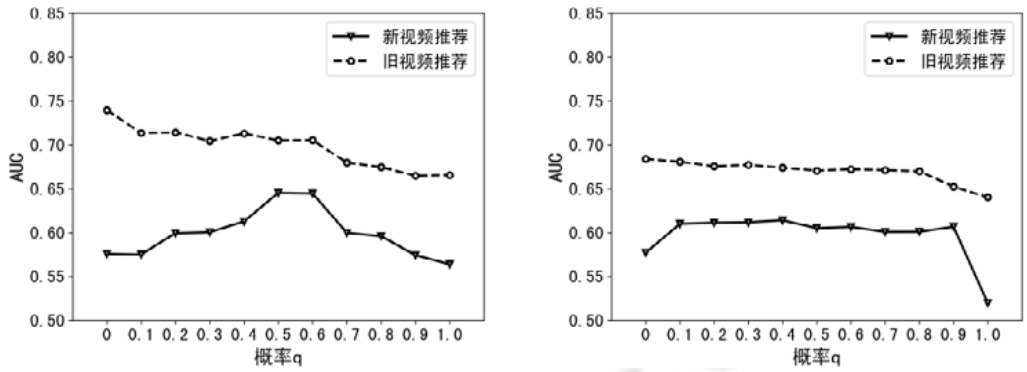
为了验证本文所提出的冷启动场景的模拟训练方法和基于近邻的替代方法对于视频冷启动问题的鲁棒性, 本实验使用 BERT 特征和音频特征的模型作为基准模型, 测试不同方法对于新视频的点击率预估能力. 由于 Track_1_series 和 Track_2_movies 数据集不能直接用于冷启动测试, 本文将两个数据集中所有测试序列中的候选视频视为新视频. 针对新视频, 模型无法获取这些视频的上下文特征.

3.3.1 模型的超参数对性能的影响

图 3 展示了在 Track_1_series 和 Track_2_movies 数据集上, 冷启动场景的模拟训练方法中超参数对性能的影响. 值得注意的是: 当概率 q 为 0 时, 表示在训练阶段不使用模拟冷启动场景的训练方法; 大于 0 时, 表示使用该训练方式. 对于新视频的推荐, 不使用模拟训练的模型在两个数据集上的性能分别只有 0.579 和 0.577; 而使用该训练方式时(q>0), 大部分的模型的性能都得到了提升, 显示了模拟训练对于新视频推荐的有效性. 但是当 q 取为 1 时, 模拟训练方法反而起了负面影响, 此时模型性能比不使用该方法还差. 我们推测: 当 q 取为 1 时, 模型将所有的候选视频都当作新视频进行训练, 在训练过程中完全忽略了候选视频的上下文特征, 从而导致性能变差. 该结果也在一定程度上也说明上下文特征对新视频的推荐是有帮助的. 当 q 取为 0 到 1 之间的值时, 模型在训练过程中同时考虑新视频和旧视频; 当 q<0.5 时, 更多地考虑旧视频; 当 q>0.5 时, 更多地考虑新视频. 总体上, 当 q 取中间值 0.5 时, 即模型更均衡地兼顾新视频和旧视频时, 模型在两个数据集上都表现出整体更好的性能.

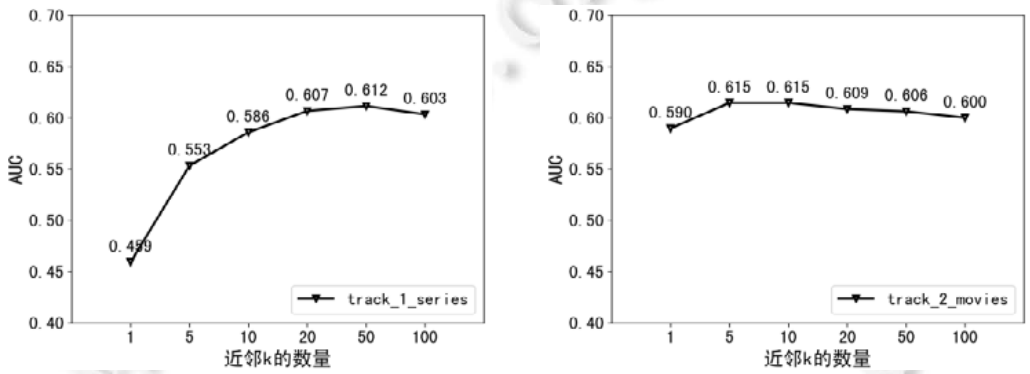
此外, 图 3 还展示了通过模拟训练后的模型对于旧视频的点击率预估能力, 即模型在预测阶段可以直接使用视频的上下文特征. 在两个数据集上, 模型对于新视频的点击率预估性能低于旧视频的预估, 这说明推荐新视频相比于旧视频更具挑战. 除此之外, 我们还发现: 当 q 越小时, 模型对于旧视频的推荐性能越高. 这是因为当q越小时, 模拟训练方法会更加的关注旧视频, 从而有助于对旧视频的推荐. 虽然通过模拟冷启动场景的训练模型对旧视频的点击率预估性能会略微下降(q>0 的性能比 q=0 的差), 但对于新视频的预估比不使用该训练方法的模型表现出明显的优势.

图 4 展示了超参数 k 对于基于近邻的替代方法的性能影响. 在 Track_1_series 上, 该方法对于 k 比较敏感; 而在 Track_2_movies, 该方法表现出更好的稳定性. 我们推测, 这可能和数据集的视频数量有关: Track_1_series 相比于 Track_2_movies 数据集的视频数量更少, 因此不容易在较少的视频找到合适的近邻视频, 从而导致对超参数更为敏感. 在下面实验中, 本文将性能最好的参数作为默认参数, 即在 Track_1_series 上 k 取 50, 在 Track_2_movies 上 k 取 5.



(a) Track_1_series 数据集 (b) Track_2_movies 数据集

图 3 不同的概率 q 对于模型性能的影响



(a) Track_1_series 数据集 (b) Track_2_movies 数据集

图 4 不同 k 的对于模型性能的影响

3.3.2 性能比较

为了验证第 2.3.1 节和第 2.3.2 节中所提出的两个方法的有效性, 本文将其与基线方法进行比较. 基线方法表示不使用模拟训练的模型, 在预测时, 用全零的向量作为新视频的上下文特征. 表 5 汇总了不同方法对于新视频点击率预估的性能. 实验结果显示: 在两个数据集上, 本文提出的两种方法都明显优于基线方法, 这表明这两种方法对于新视频点击率预估的有效性. 比较模拟冷启动场景的训练方法和基于近邻的替代方法, 前者在 Track_1_series 上表现得更好, 而后者则在 Track_2_movies 上表现出更好的性能. 我们推测, 这不一致的性能与两个数据集视频数量有关: 由于 Track_1_series 数据的视频较少, 将导致基于近邻的替代方法不能找到合适的近邻视频从而影响性能. 因此, 基于近邻的替代方法在 Track_1_series 数据集上比模拟冷启动场景的训练方法差.

表 5 不同方法对于新视频点击率预估的性能

模型	Track_1_series		Track_2_movies	
	AUC	Logloss	AUC	Logloss
基线方法	0.579	0.576	0.577	0.763
模拟冷启动场景的训练方法	**0.645**	**0.418**	0.605	0.690
基于近邻的替代方法	0.612	0.438	**0.615**	**0.658**

4 结束语

本文通过同时使用视频的内容特征和上下文特征来加强视频点击率预估模型的性能, 并设计了两种不同获取视频上下文特征的方法. 在两个真实电视剧和电影推荐上的实验表明: 两种特征有很好的互补性, 有助

于提升模型的性能. 针对视频的冷启动问题, 本文提出了冷启动场景的模拟训练方法和基于近邻的替代方法. 实验表明: 这两种方法都能明显提升模型对于新视频的点击率预估能力, 对于视频的冷启动问题表现出更好的鲁棒性. 此外, 当训练视频较少时, 推荐使用冷启动场景的模拟训练方法; 反之, 则使用基于近邻的替代方法. 现在的模型还有很大的提升空间, 在后续的研究中, 我们将对视频的内容特征和上下文特征之间做更深层次的融合而不是简单的拼接; 同时利用用户的个性化信息, 进一步提升模型的性能.

References:

[1] Schein AI, Popescul A, Ungar LH, *et al*. Methods and metrics for cold-start recommendations. In: Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2002. 253−260.

[2] Li SJ, Lei WQ, Wu QY, *et al*. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. ACM Trans. on Information Systems, 2021. 1−29.

[3] Chen ZY, Xu K, Zhang W. Content-Based video relevance prediction with multi-view multi-level deep interest network. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: ACM, 2019. 2607−2611.

[4] Zhou G, Zhu XQ, Song CR, *et al*. Deep interest network for click-through rate prediction. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. New York: ACM, 2018. 1059−1068.

[5] Wu S, Tang YY, Zhu YQ, *et al*. Session-Based recommendation with graph neural networks. In: Proc. of the AAAI Conf. on Artificial Intelligence. Vol.33. Menlo Park: AAAI, 2019. 346−353.

[6] Chen JH, Zhang Q, Wang SL, *et al*. Click-Through rate prediction based on deep belief nets and its optimization. Ruan Jian Xue Bao/Journal of Software, 2019, 30(12): 3665−3682 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/30/3665.htm [doi: 10.13328/j.cnki.jos.005640]

[7] Shao YW, Zhang M, Ma WZ, *et al*. Integrating latent item-item complementarity with personalized recommendation systems. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1090−1100 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5925.htm [doi: 10.13328/j.cnki.jos.005925]

[8] Sachdeva N, Mcauley JJ. How useful are reviews for recommendation? A critical review and potential improvements. In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2020. 1845−1848.

[9] Geng X, Zhang H, Bian J, *et al*. Learning image and user features for recommendation in social networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE, 2015. 4274−4282.

[10] Chen X, Zhang YF, Ai QY, *et al*. Personalized key frame recommendation. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2017. 315−324.

[11] Kang WC, McAuley JJ. Self-Attentive sequential recommendation. In: Proc. of 2018 IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2018. 197−206.

[12] Chen Q, Zhao H, Li W, *et al*. Behavior sequence transformer for e-commerce recommendation in alibaba. In: Proc. of the 1st Int'l Workshop on Deep Learning Practice for High-Dimensional Sparse Data. 2019. 1−4.

[13] Hidasi B, Karatzoglou A. Recurrent neural networks with top-*k* gains for session-based recommendations. In: Proc. of the 27th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM, 2018. 843−852.

[14] He XN, He ZK, Song J, *et al*. NAIS: Neural attentive item similarity model for recommendation. IEEE Trans. on Knowledge and Data Engineering, 2018, 30(12): 2354−2366.

[15] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Proc. of the Neural Information Processing Systems Conf. New York: NIPS, 2017. 5998−6008.

[16] Zhao N, Pi WC, Xu CQ. Video recommendation algorithm for multidimensional feature analysis and filtering. Computer Science, 2020, 47(4): 103−107 (in Chinese with English abstract).

[17] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. In: Proc. of the ICLR 2013. 2013. Workshop Track Proceedings.

[18] Wei YW, Wang X, Nie LQ, *et al*. Graph-Refined convolutional network for multimedia recommendation with implicit feedback. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: ACM, 2020. 3541−3549.

[19] Ge Y, Chen SC. Graph convolutional network for recommender systems. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1101−1112 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/31/1101.htm

[20] Wang N, He X, Liu Z, *et al*. Personalized video recommendation strategy based on user's playback behavior sequence. Chinese Journal of Computers, 2020, 43(1): 123−135 (in Chinese with English abstract).

[21] Yang B, Mei T, Hua XS, *et al*. Online video recommendation based on multimodal fusion and relevance feedback. In: Proc. of the 6th ACM Int'l Conf. on Image and Video Retrieval. New York: ACM, 2007. 73−80.

[22] Van AVD, Dieleman S, Schraumen B. Deep content-based music recommendation. In: Proc. of the Neural Information Processing Systems Conf. New York: NIPS, 2013. 2643−2651.

[23] Dong JD, Wang X, Zhang LM, *et al*. Feature re-learning with data augmentation for video relevance prediction. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(5): 1946−1959.
[24] Chen XS, Zhao R, Ma SJ, *et al*. Content-Based video relevance prediction with second-order relevance and attention modeling. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. New York: ACM, 2018. 2018−2022.
[25] Wang P, Jiang YS, Xu CX, *et al*. Overview of content-based click-through rate prediction challenge for video recommendation. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: ACM, 2019. 2593−2596.
[26] Xu Q, Xu HC, Chen WL, *et al*. Time-Aware session embedding for click-through-rate prediction. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: ACM, 2019. 2617−2621.
[27] Wang X, Du YL, Zhang LM, *et al*. Exploring content-based video relevance for video click-through rate prediction. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: ACM, 2019. 2602−2606.
[28] Hershey S, Chaudhuri S, Ellis DPW, *et al*. CNN architectures for large-scale audio classification. In: Proc. of the 2017 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2017. 131−135.
[29] Devilin J, Chang MW, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL-HLT, 2019. 4171−4186.
[30] Zhang XR, Yuan X, Li YW, *et al*. Cold-Start representation learning: A recommendation approach with bert4Movie and movie2Vec. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: ACM, 2019. 2612−2616.
[31] Qiu ZP, Wu X, Gao JY, *et al*. U-BERT: Pre-training user representations for improved recommendation. In: Proc. of the AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2021. 1−8.

附中文参考文献:

[6] 陈杰浩, 张钦, 王树良, 等. 基于深度置信网络的广告点击率预估的优化. 软件学报, 2019, 30(12): 3665−3682. http://www.jos.org.cn/1000-9825/30/3665.htm [doi: 10.13328/j.cnki.jos.005640]
[7] 邵英玮, 张敏, 马为之, 等. 融合商品潜在互补性发现的个性化推荐方法. 软件学报, 2020, 31(4): 1090−1100. http://www.jos.org.cn/1000-9825/5925.htm [doi: 10.13328/j.cnki.jos.005925]
[16] 赵楠, 皮文超, 许长桥. 一种面向多维特征分析过滤的视频推荐算法. 计算机科学, 2020, 47(4): 103−107.
[19] 葛尧, 陈松灿. 面向推荐系统的图卷积网络. 软件学报, 2020, 31(4): 1101−1112. http://www.jos.org.cn/1000-9825/5928.htm [doi: 10.13328/j.cnki.jos.005928]
[20] 王娜, 何晓明, 刘志强, 等. 一种基于用户播放行为序列的个性化视频推荐策略. 计算机学报, 2020, 43(1): 123−135.

章磊敏(1994－), 男, 硕士, 主要研究领域为人工智能, 推荐系统.

纪守领(1986－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为人工智能与安全.

董建锋(1991－), 男, 博士, 研究员, CCF 专业会员, 主要研究领域为多媒体理解, 计算机视觉.

王勋(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为可视媒体计算, 模式识别.

包翠竹(1990－), 女, 博士, 讲师, CCF 专业会员, 主要研究领域为图像处理.