文献综述

摘要：运用多模态来对视频的特征进行提取表示已得到国内外学者的广泛研究，相比于单模态，多模态能更加全面和丰富的对视频进行描述。相比于显式反馈，隐式反馈能以更低的成本在更大的范围收集。然而传统协同过滤视频推荐算法大多忽略用户与多媒体内容交互的隐式反馈，多模态视频推荐算法能充分分析视频内容，结合用户的隐式反馈对用户进行更加精确的视频推荐。本文首先对多模态视频推荐进行介绍，然后介绍不同部分的国内外多模态视频推荐研究现状，最后进行总结。

1. 介绍

随着计算机与互联网技术的发展，各类信息呈现出指数级别的增长趋势。各种多媒体设备的广泛应用、计算机处理能力的不断提高与无线网络的大量普及使得视频信息的快速获取、大量存储和广泛传播成为了可能，视频也因其丰富的表现形式越来越受到人们的关注。然而，海量视频资源为用户检索自己感兴趣的视频带来了挑战，由此视频推荐应运而生。在“信息爆炸”的时代，推荐系统在主动提供用户可能感兴趣的视频，节约用户手动检索时间上的优势显得愈发明显。

传统的推荐算法主要分为基于内容的推荐算法、协同过滤推荐算法以及混合推荐算法。然而，视频数据具有数据维度较高、表现形式迥异、含有大量冗余信息等特点。因此，传统推荐算法无法对视频内容本身进行分析以此获得准确的视频信息，无法直接应用于视频推荐中。近年来，随着深度学习在各个领域的广泛应用，其有效性也体现在信息检索和推荐系统的研究中。现有大多数基于深度学习的推荐算法[1][2][3]都仅使用视频的文本信息来提高推荐效果。文本信息多为视频标题、视频标签和简介等，然而许多视频的标签都是不准确甚至缺失的，许多的视频题目及简介也会因为上传者为了吸引眼球而夸张化，从而导致视频内容与视频题目不符使得推荐效果欠缺。针对文本内容不可用的问题，研究学者也提出了利用非文本内容来进行推荐的模型。He[4]等人提出了一种视觉贝叶斯个性化排序算法（Visual Personalized Ranking，VBPR），分析产品图像，将其视觉特征纳入矩阵分解中进行推荐。类似的，Liang[5]等人将音频内容信息整合到协同过滤中并应用于音乐推荐系统。然而，类似于仅使用文本信息的模型，这些方法都仅探索某一特定模态特征，在他们依赖的内容有所损失或噪声太大时，这些模型的性能就会下降。

相比于仅使用视频的单模态信息，多模态内容充分利用视频的图像、音频、文本信息，对视频的表示更加丰富与全面，且当某一模态信息不可用或噪声较大时，另外两个模态就有补充的功能，弥补该模态造成的缺失问题。屈雯等人[6][7]提出首先在单个模态下对电影评分进行预测，且若用户历史评分数量低于阈值，

则进行用户喜好增强，更新各模态评分后对三个模态评分进行合并，得到最后的预测结果。然而该方法中对用户喜好的增强本质上还是通过已有的用户评分信息预测用户行为，因此并不适用于用户评分信息大量缺失的情况。Du[8]等人利用丰富的视频内容特征，提出了可同时用于冷启动和热启动的个性化视频推荐方法，并提出多特征融合方法 PRI 计算不同模态特征的权重，然而该方法只有在不同模态影响有很大不同时才有效。

2. 相关工作

2.1 基于深度学习的推荐

针对推荐中无法获取用户行为中评分信息的情况，研究学者从对依赖评分信息的视频推荐方法的研究转向对视频排序算法的设计。排序问题主要有点级（Point-Wise）、对级（Pair-Wise）、列表级（List-Wise）三大类型[10]，在早期工作中，Badrul Sarwar 等人提出奇异值矩阵分解（Singular Value Decomposition，SVD）来学习特征矩阵，但该矩阵分解模型存在过拟合的问题[11]。随后，Steffen Rendle 等人提出了贝叶斯个性化排序模型（Bayesian Personalized Ranking，BPR），针对每一个用户自己的喜好进行排序，在海量数据中选择极少量数据做推荐的应用场景下，该模型有很大优势[12]。文献[13]提出双相似度的个性化排序模型（DSBPR），将异质信息网络中的相似度信息融入到基于 BPR 模型的矩阵分解算法中，结合对级方法进行模型的训练，相比于矩阵分解（Matrix Factorization，MF）和 BPR 算法表现出了更好的性能。然而视频数据同时具有文本、图像和音频的特点，根据用户视频间关系进行推荐的算法并不能完全适用于视频推荐。

2.2 注意力机制

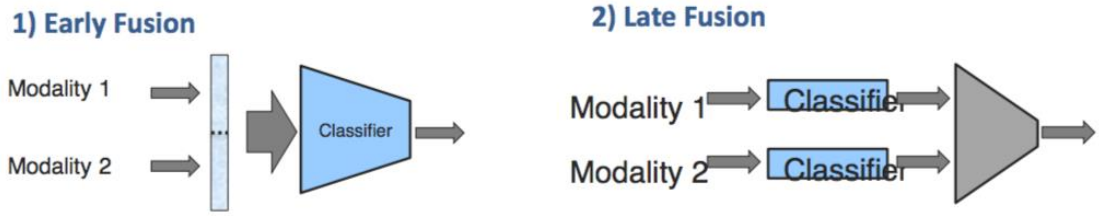
注意力机制[14]的核心思想是指在学习过程中对样本中的不同内容赋予不同权重。注意力机制计算当前输入序列与输出向量的匹配程度，匹配度高则相对应的得分越高，反之亦然。该机制能够使处理系统更专注于捕获输入数据中显著的与当前输出相关的有用信息[15]。注意力机制的最终目的是帮助类似编解码器的框架，探索多模态表示之间的关联关系[16]，克服编码器由于无法解释从而很难设计的缺陷，因此，注意力机制非常适合用于推理任务，诸如多模态融合[17]、多模态相似度学习[18]等多种不同模态数据之间的相互映射关系。

自注意力机制作为注意力机制的一个分支，在一些序列化推荐任务上反映出比卷积神经网络与循环神经网络更好的性能，在推荐任务上也有多种广泛的应用。Zhang[18]等人提出了一个通过充分利用自注意力和度量学习的序列感知模型，提升了序列推荐的性能。Zhou 等人[19]将自注意力机制用在用户异构行为建模上。

2.3 视频多模态特征表示

传统的图像特征提取方法[6]主要依赖人工提取如图像的颜色直方图、纹理特征等特征后再进行下一步的处理。近年来，随着机器学习的快速发展，卷积神经网络（Convolutional Neural Network , CNN）成为提取图像特征的首选方式。当前具有代表性的深度神经网络包括 AlexNet[20]、VGGNet[21]、GoogLeNet[22]和ResNet[23]等。描述音频特征的指标常有梅尔频率倒谱系数、过零率、短时能量等。在提取其特征上，SoundNet[24]结构利用视频中视觉和声音的同步性以及未标记的视频数据，通过迁移学习将预训练好的模式识别模型应用到音频识别的领域。GloVe（Global Vectors for Word Representation）模型常用于提取文本特征。GloVe[25]是一个基于全局词频统计的词表征工具，即用可以表征单词语义特征的实值向量来表示单词。Glove 模型克服了 LSA（Latent Semantic Analysis）复杂度高、计算代价大以及所有单词具有同一权重[26]等缺点，同时解决了 word2vec 无法充分利用所有语料的缺点，在语义准确度、语法准确度以及总体准确度等方面上得到了比较好的结果。

得到各模态特征后，需将各模态以特定的方式融合为统一的特征表示以进行下一步的推荐任务。常见的融合方式有早期融合（Early Fusion）以及后期融合（Late Fusion），如下图所示，早期融合旨在将多个内容空间映射到共享的同质空间，而后期融合则是在单个模态经过预测之后在对各模态进行映射。

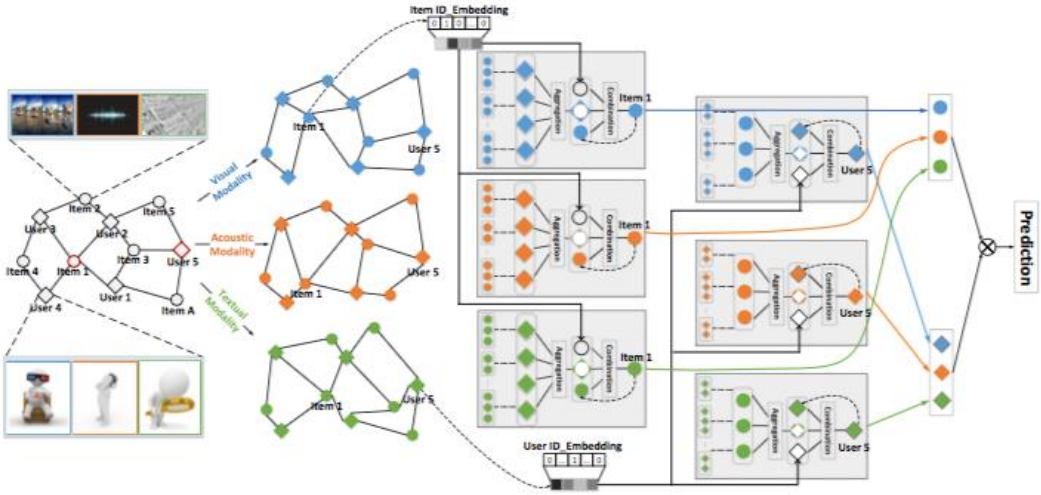


2.4 多模态推荐

由于协同过滤方法在推荐系统的成功，早期的多模态[27][28][29]推荐算法都主要基于协同过滤的方法。而基于协同过滤的方法主要利用用户的反馈来预测用户项目间的交互，所以这些方法对于具有足够反馈的问题效果很好，但并不适用于反馈较少的情况。

为了弥补基于协同过滤模型的弊端，研究人员开发了混合方法[30]，将多媒体项目背景和内容与协调过滤模型进行推荐。Chen[31]等人探索了在组件级别上用户对项目偏好，并引入了一种新颖的注意力机制来解决多媒体推荐中具有挑战性的项目和组件级反馈。在这种方法中，用户特征由协同过滤和交互项目的内容信息刻画，尽管此方法已经将用户偏好用两种范围级别表示，但它无法表示用户对不同模态的偏好程度。在表示单个模态特征方面，屈雯[6]分别提取视频的文本、音频、图像特征，在单个模态下对电影评分之后，对三个模态评分

进行合并得到预测结果。然而该方法主要依靠人工提取特征，灵活性不如适用神经网络端到端学习，且其给予三个模特评分权重一致，无法动态调整权重。提出利用用户-项目交互来指导每种模态的表示学习，并进一步个性化微视频推荐。Wei[32]等人设计了一个基于图神经网络消息传递的多模态图卷积网络框架，该框架可以生成用户和微视频的模态特定表示，以更好地捕获用户的偏好。具体来说，在每个模态中构造一个用户-项目双向图，并用其邻居的拓扑结构和特征丰富每个节点的表示。如下图所示：



4. 结论

相对于传统的视频推荐技术，多模态视频推荐技术是一个比较新的研究方向，随着深度学习技术的发展，也有越来越多的学者开始研究并已有可观的研究成果，然而多模态视频研究依然面临着一些挑战，有以下几点：

（1） 对视频各模态特征的提取计算量大、耗时久，如何在保证特征提取质量的前提下降低计算复杂度成为一个亟待解决的问题。

（2） 目前的多模态视频推荐算法大多应用于短视频甚至微视频中，将其应用于规模更大、时长更高的视频推荐任务上也是未来工作的一个重点。

参考文献

[1] 姚静静. 基于协同过滤的电影推荐算法研究与实现. 北京: 北京邮电大学, 2018.
(Yao Jingjing. Research and implementation of movie recommendation algorithm based on collaborative filtering. Beijing: Beijing University of Posts and Telecommunications, 2018.)

[2] 王建洋. 基于深度学习的电影推荐系统研究与实现. 成都: 西南交通大学, 2018.
(Wang Jianyang. Research and implementation of film recommendation system based on deep learning. Chen Du: Southwest Jiaotong University, 2018.)

[3] 李靖怡. 基于内容分发平台的短视频产品推荐算法研究与实现. 北京: 首都经济贸易大学, 2018.
(Li Jingyi. Research and implementation of short video product recommendation algorithm based on content distribution platform. Beijing: Capital University of Economics and Business, 2018.)

[4] He R , Mcauley J . VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback[J/OL]. [2019-9-

11]. https://arxiv.org/pdf/1510.01784.pdf.

[5] Liang D, Zhan M, Ellis Daniel PW , et al. Content-Aware Collaborative Music Recommendation Using Pre-trained Neural Networks[C/OL]. [2019-9-11].https://dawenl.github.io/publications/LiangZE15-ccm.pdf.

[6] 屈雯. 基于多模态内容分析的多视角视频推荐技术研究. 沈阳: 东北大学,2015.
(Qu Wen. Research on Multi-view Video Recommendation Approaches based on Multimodal Content Analysis, 2015.)

[7] Qu Wen, Song Kaisong, Zhang Yifei, et al. A novel approach based on multi-view content analysis and semi-supervised enrichment for movie recommendation. Journal of Computer Science and Technology, 2013, 28(5): 776–787.

[8] Du XZ, Yin HZ, Chen L, et al. Personalized video recommendation using rich contents from videos. IEEE Transaction on Knowledge and Data Engineering, 2018.

[9] Chen J, Song X, Nie L, et al. Micro tells macro: predicting the popularity of micro-videos via a transductive model // Proc. of the 24th ACM international conference on Multimedia, New York: ACM, 2016: 898–907.

[10] Li Hang. A short introduction to learning to rank. IEICE TRANSACTIONS on Information and Systems, 2011, 94(10): 1854–1862.

[11] Sarwar B, Konstan J, Riedl J. Incremental singular value decomposition algorithms for highly scalable recommender systems // Fifth international conference on computer and information science, 2002: 27-28.

[12] Rendle S , Freudenthaler C , Gantner Z , et al. BPR: Bayesian Personalized Ranking from Implicit Feedback // Proc. of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal: AUAI Press, 2009: 452-461.

[13] 史龙飞. 基于 BPR 模型的情景感知推荐算法的研究与实现. 北京:北京邮电大学, 2018.

[14] (Shi Longfei. The research and implementation of contextaware recommender based on Bayesian Personalized Ranking, 2018.)

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017: 5998–6008.

[16] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. Computer Science, 2014.

[17] Gao L, Guo Z, Zhang H, et al. Video Captioning with Attention-Based LSTM and Semantic Consistency. IEEE Trans on Multimedia, 2017: 19(9): 1-1.

[18] Hori C, Hori T, Lee TY, et al. Attention-Based Multimodal Fusion for Video Description. Proc of the IEEE international conference on computer vision, 2017, 4193-4202.

[19] Zhang S , Tay Y , Yao L , et al. Next Item Recommendation with Self-Attention[C/OL]. [2019-9-11]. https://arxiv.org/pdf/1808.06414v2.pdf .

[20] Zhou C , Bai J , Song J , et al. ATRank: An Attention-Based User Behavior Modeling Framework for Recom-mendation[C/OL]. [2019-9-11].https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16216/16770.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks // Proc. International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: ACM, 2012: 1097-1105.

[22] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C/OL]. [2018-12-15].https://arxiv.org/pdf/1409.1556.pdf.

[23] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions // Proc. the 28th IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1-9.

[24] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition // Proc. the 28th IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: Taylor & Francis, 2015: 70-778.

[25] Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video // Advances in neural information processing systems, 2016: 892–900.

[26] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation // Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP), Stroudsburg: ACL, 2014 1532–1543.

[27] Altszyler E, Sigman M, Slezak D F. Corpus specificity in LSA and Word2vec: the role of out-of-domain documents[J/OL]. [2019-9-11]. https://arxiv.org/pdf/1712.10054.pdf.

[28] He XN, He Z, Du X, et al. Adversarial personalized ranking for recommendation. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018: 355-364.

[29] Wang X, He X Nie L, et al. Item silk road: Recommending items from information domains to social users // Proc. of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, New York: ACM, 2017.

[30] Wang X, He X, Feng F, et al. Tem: Tree-enhanced embedding model for explainable recommendation // Proc. of the 2018 World Wide Web Conference, Switzerland: International World Wide Web Conferences Steering Committee，2018: 1543-1552.

[31] Van den Oord, Aaron, Sander Dieleman, Benjamin Schrauwen. Deep content-based music recommendation. Advances in neural information processing systems. 2013: 2643-2651

[32] Chen Jingyuan, Zhang Hanwang, He Xiangnan, et al. Attentive Collaborative Filtering: Multimedia recommendation with item- and Component-Level attention // Proc. of the 40th International ACM SIGIR conference on Research and Development in Information Retrievall, New York: ACM，2017: 335-344

[33] Wei Y, Wang X, Nie L, et al. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video[C]//Proceedings of the 27th ACM International Conference on Multimedia. ACM, 2019: 1437-1445.