基于多模态特征融合嵌入的相似广告检索方法

冯 奕[1),2)] 周晓松[1),2)] 李传艺[1),2)] 王 挺[3)] 葛季栋[1),2)]
胡雨成[3)] 张小鹏[3)] 骆 斌[1),2)]

[1)](南京大学计算机软件新技术国家重点实验室 南京 210046)
[2)](南京大学软件学院 南京 210093)
[3)](深圳市腾讯计算机系统有限公司 广东 深圳 518000)

摘 要 随着互联网人工智能技术的飞速发展,学习用户特征并精准投放广告能够显著提升广告的点击率(Click-Through-Rate,CTR)与转化率(Conversion Rate,CVR).人群智能定向是解决广告投放问题中极其重要的一环,其业界主流方法是使用转化用户和非转化用户训练基于用户特征的判断其是否会成为转化用户的分类模型.这个分类器的优劣依赖广告的实际转化人群规模,规模越大,越能准确判断.但在实际应用中通常面临某些广告转化人群不足的问题,本文利用在学术与工业场景占据重要研究地位的基于内容的检索技术来扩充相似广告集合,从而扩充对应转化人群.现有的单模态检索方案只关注于单个模态的特征(文本/图像),忽视了不同模态间的内在共有联系,使得挖掘出的广告特征不全且包含大量噪声,最终导致相似广告的检索结果质量不高,从而导致相似转化人群的扩充质量低下.而近年来兴起的跨模态检索方案主要关注以文搜图或以图搜文,并且没有考虑到通用目标检测器并不适用于特定领域图像数据这一事实.为解决这些问题,本文提出一种以广告分类为基本训练目标的多模态商品广告特征融合建模方法,以提升相似广告检索的效果.具体来说,本文使用 Transformer 模型提取文本语义特征,使用目标检测 YOLO 模型挖掘图像中细粒度的视觉特征,并结合文本注意力机制识别图像中与商品相关的目标,以降低无关目标给广告特征带来的噪声影响.同时,本文提出了一种多模态融合注意力机制,以高效融合广告文本和图像特征.该模型命名为 ToTYEmb(Text oriented Transformer-Yolo fusion Embedding).另外,本文还提出了一种算法框架,将相似广告扩充、转化人群扩充加入到现有的人群智能定向工作流中.实验结果表明,较多个基线模型,本文方案有效提升了相似商品广告的检索质量,避免了很多由单模态信息带来的错误.同时离线人群定向更新实验表明本文提出的利用相似广告扩充转化人群确实能在很大程度上优化现有的人群智能定向算法.

关键词 多模态特征融合;相似广告检索;Transformer;注意力机制
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.01500

A Multi-Modal Feature Fusion Embedding Method for Similar Ad Retrieving

FENG Yi[1),2)] ZHOU Xiao-Song[1),2)] LI Chuan-Yi[1),2)] WANG Ting[3)] GE Ji-Dong[1),2)]
HU Yu-Cheng[3)] ZHANG Xiao-Peng[3)] LUO Bin[1),2)]

[1)](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210046)
[2)](*Software Institute, Nanjing University, Nanjing* 210093)
[3)](*Tencent, Shenzhen, Guangdong* 518000)

Abstract With the rapid development of Internet artificial intelligence technology, learning user characteristics and accurately placing advertisements can significantly increase the click-through rate (CTR) and conversion rate (CVR) of advertisements. Targeting the potential customers of a

收稿日期:2021-01-25;在线发布日期:2021-09-16.本课题得到国家自然科学基金青年科学基金(61802167)和腾讯公司资助.**冯 奕**,博士研究生,主要研究方向为自然语言处理、机器学习.E-mail:fengyinju@163.com.**周晓松**(共同第一作者),硕士研究生,主要研究方向为自然语言处理、机器学习.E-mail:realx_zhou@163.com.**李传艺**(通信作者),博士,助理研究员,中国计算机学会(CCF)会员,主要研究方向为自然语言处理、机器学习.E-mail:lcy@nju.edu.cn.**王 挺**,博士,主要研究方向为自然语言处理和机器学习.**葛季栋**,博士,副教授,主要研究方向为软件工程和自然语言处理.**胡雨成**,硕士,主要研究方向为推荐系统和自然语言处理.**张小鹏**,硕士,主要研究方向为推荐系统.**骆 斌**,博士,教授,主要研究领域为软件工程.

product from the crowd automatically is an extremely important part of solving the problem of advertising. The mainstream method in the industry is to use converted users and non-converted users to train a classification model based on user characteristics for judging whether a given use would be buy the product of the ads or not. The performance of this classifier largely depends on the actual converted population scale of advertising. The larger the scale, the more accurate the judgment can be. However, in practical applications, it usually faces the problem of insufficient specific converted users. This paper innovatively uses the widely studied content-based retrieval technology to expand the collection of similar ads of a given ad and treats the converted users of these similar advertisements as the converted users of the ad, thereby expanding the corresponding conversion population. However, on one hand, existing monomodal retrieval methods focus on the features of a single modal (text or image), and ignore the inherent relationship between different modals, resulting in insufficient features and a large amount of noise. On the other hand, the cross-modal retrieval schemes that have emerged in recent years mainly focus on searching images by text or searching text by images, and do not take into account the fact that general object detectors are not suitable for image data in specific fields. In order to solve these problems, this paper proposes a multi-modal feature fusion model with the training goal of classifiying ads to correct types to extract comprehensive features to improve the performance of similar ads retrieval. To be specific, we adopt Transformer to extract semantic information from text, and employ YOLO to mine fine-grained visual features from images. Besides, for mining text-image relation, a text-based attention mechanism is used to extract product-related objects in images to reduce the noise impact of irrelevant items. We also adopt a multimodal fusion layer to fuse text and image features. The proposed entire model names ToTYEmb (Text oriented Transformer-Yolo fusion Embedding). With classification target, ToTYEmb would generate useful embeddings for each ad and it will be used for searching similar ones. In addition, we also proposes an algorithmic framework that embeds similar advertising expansion and conversion crowd expansion into the existing crowd intelligent targeting workflow. In the experiments, we compared the classification accuracies and recall precisions of similar ads of the proposed ToTYEmb with the baseline classification models and multi modal pre-training models, and analyzed the impact of different modules and the number of selected image areas on the recall of similar ads. Experimental results prove that the proposed method effectively improves the retrieval quality of similar ads and avoids many errors caused by incorrect single-modal information. At the same time, the offline target crowd update experiment shows that the proposed method can indeed optimize the existing intelligent crowd targeting algorithm to a large extent.

Keywords　multi-modal feature fusion; similar ads retrieval; Transformer; attention mechanism

1　引　言

在大数据时代，互联网广告技术迎来了全新的变革，基于深度学习的广告推荐成为许多企业和个人的核心着力点. 而如何将特定广告投放在对应感兴趣的人群中，在减少开支的同时提升投放广告的点击率(Click-Through-Rate，CTR)、转化率(Conversion Rate，CVR)成为广告投放技术中所需要解决的核心问题之一. 其中工业界广泛应用的一种解决方案为人群智能定向，利用模型学习投放广告的转化人群的特征，从而智能更新人群，优化后续投放效果. 这种方案严重依赖于已投放广告背后的转化人群，而在实际应用中，经常出现转化人群数量极少的情况，直接影响了整个算法框架的性能. 例如，某些小众品牌广告主将一批广告投放在一个具有某些特征的人群包中，一个时间周期后，由于鲜有用户点击，其转化人群极少，也就无法有效学习用户特征，

从而更新投放人群.为了解决该问题,本文利用在学术与工业场景占据重要研究地位的基于内容的检索技术为广告扩充转化人群,从而优化现有的智能定向人群算法.由于本文提出方法的最终目的是通过扩充相似的广告从而扩充转化人群,我们称之为相似广告检索问题.

相似广告检索本身有着诸多应用场景,例如根据用户历史点击记录进行个性化推荐.通过检索用户历史点击广告的相似广告,可以为用户推荐其可能感兴趣的商品,从而提高被推荐商品广告的点击率和转化率.例如点击毛尖绿茶广告的用户可能会对碧螺春绿茶感兴趣,继而为其推荐其他绿茶类广告,促使用户产生购买行为.同样的,本文利用相似广告检索技术来扩充人群也是基于这样的事实:多模态语义上相似的广告其对应的转化人群特征也是相似的,这点将会在第 5 节中得到讨论.

事实上广告推荐模型还会结合用户行为、社交网络等信息进行建模,其侧重于推荐技术本身,但这不是本文重点.本文着重关注于深度融合图片模态与文本模态特征,构建多模态图文嵌入(Embedding)模型(类似于词嵌入、句嵌入),将该特征作为一条广告的内容特征,采用基于距离度量的方法召回最相似的广告集合,从而将扩充到的广告集合对应的转化人群也作为当前广告集合的转化人群,达到扩充转化人群的目的.

基于内容的检索技术被广泛应用在学术界和工业界的各项任务中,其中也包括相似广告检索.目前一种常用解决方案是利用广告文本嵌入技术,例如 Word2Vec[1]、GloVe[2]、Doc2Vec[3],乃至动态表征能力更强的 ELMo[4]、BERT[5]等,将广告文本映射到特征空间,使得相似广告在特征空间中聚集在一起,然后根据特征向量的距离来检索相似广告.

然而广告文本存在以下两个问题:(1)通常为短文本,且很多广告句式用词相同;(2)存在通用文本的情况(如"快来看看").仅依赖文本的语义信息在某些情况下无法区分广告.图 1 展示了相似广告检索问题难点示例及和公开图文跨模态检索数据集(Flickr30k[6])的差异对比,其中每一张图片都预先经过 YOLO 目标检测模型[7]提取高置信度目标并标注在图片中.如图 1 所示,(a1)和(a2)分别是棉被和羽绒服广告,但文本句式相似,用词相似,基于文本 Embedding 的方法极有可能造成假阳性结果.因为广告文案一般由文本和配图构成,所以除了基于文本的 Embedding 方法,也可以采用基于图像特征(这里也称之为 Embedding)策略.例如通过 VGG[8]、InceptionV4[9]等模型提取图像信息,利用池化层的输出作为图像的特征.同样,仅仅利用图像信息,也会造成许多错误.一些广告图像风格内容极为相似,且图像中含有大量商品无关噪声像素点,而商品相关像素点只占了图像的一小部分.在图 1(b1)和(b2)中,几乎无法通过图像区分这两个广告,对应商品的视野只占了图像一小块区域,而灰色背景噪声像素在图像 Embedding 过程起主导作用.因此,图像 Embedding 极有可能突出噪声特征,从而降低检索准确性.但是综合文本和图像进行决策,即融合文本的语义信息和图像的视觉信息,将有效弥补单一模态信息量不足的问题.人在理解广告过程中,是结合文本和图像,通过捕捉文本和图像的内在联系,综合

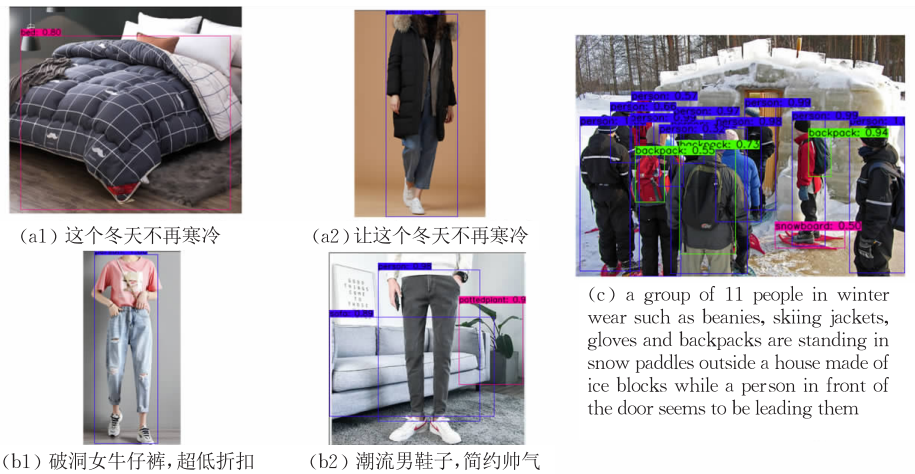

图 1 相似广告检索问题难点示例及和 Flickr30k 数据集的差异对比(其中(a1)、(a2)、(b1)、(b2)属于本文构建的电商广告数据集,(c)属于 Flickr30k 数据集.可以看出单从文字难以区分(a1)和(a2);单从图片难以区分(b1)和(b2).而作为公开跨模态检索数据集,(c)图片中被检测出众多高置信度对象,和文本关键字对应较多)

考量并得出结论.例如图 1(b1)中,图像噪声过多时结合文本中的词信息,根据关键词片段“女牛仔裤”判断图像是和牛仔裤相关,从而视野锁定在牛仔裤上.同样,在图 1(a1)中,文本特征不足时结合图片信息,可以得知对应文本在描述棉被商品.

另外,现有的跨模态检索模型主要关注以文搜图或以图搜文,以被广泛研究的 Flickr30k[6] 为例,如图 1(c)所示,其图片中被检测出众多相关目标,而在文本中有多处对应关系,例如“11 people”、“backpacks”等,这是因为该数据集和目标检测器训练所用的数据集的风格、目标类别相似;相比之下,在本文所构建的电商广告数据集中,单单凭借目标检测器自带的边框、类别置信度很难提取到足够且有用的目标特征.在图像 Embedding 过程中,一些细粒度的特征尤为重要.广告图像通常会有些关键的区域(Region),这些区域对应着广告的商品实体.在 Embedding 过程中,如果挖掘这些细粒度区域,不仅可以解决无关噪声像素问题,也可以突出商品特征像素点.基于目标检测的模型可以自动从图像中抽取关键区域,例如 YOLO[7]、Faster-RCNN[10] 等.但是已有目标检测模型抽取出的区域并不一定是和广告相关的.如图 1(b2)所提取到的沙发、盆栽都是无关目标.广告文本描述中,通常会包含商品词以及商品特征词,例如图 1(b2)中的“鞋子”、“帅气”,这些词汇有助于寻找图像中有效目标,并进一步优化相似商品广告检索.

为解决上述问题,本文创新地提出一种多模态图文特征融合模型,其能提取图文嵌入(Embedding)作为广告的融合内容特征,该模型命名为 ToTYEmb(Text oriented Transformer-Yolo fusion Embedding).广告图像中存在大量的无关噪声像素,为了忽视这些噪声,同时细粒度地理解图像,使用目标检测技术挖掘图像中的关键区域,并提出一种基于文本的注意力机制过滤无效目标.具体而言,我们使用预训练 Transformer[11] 的编码器和 YOLO[7] 分别作为文本和图像的特征提取器,并集成于同一个框架下.Transformer 对文本进行序列建模,挖掘上下文关系并获得单词的语义特征.YOLO 基于目标识别提取区域特征.在 YOLO 预测出目标区域后,使用商品名称词汇以及商品特征词汇作为线索指导注意力机制,从而筛选出广告相关预测目标并保留.接着通过融合注意力层,挖掘文本特征和区域特征的内在联系并进行融合,作为商品广告的特征表示.这种结合另一模态信息来提升下游任务性能的思路也被用于视觉信息和社交标签信息融合,跨模态社交图像聚类[12].另外,本文结合具体场景,开创性地利用文本线索信息指引图像模态,以一种主-从跨模态交互的形式提升了模型性能.

本文的贡献如下:

(1)本文提出一种多模态图文特征融合模型 ToTYEmb,其能提取图文嵌入(Embedding)作为广告的融合内容特征,同时融合文本语义信息和图像的视觉信息,弥补了单模态 Embedding 信息的不足;

(2)本文创新性地提出细粒度地挖掘图像区域特征,以商品广告文本中的特征词为线索筛选出图像中与商品相关的区域,从而降低与商品无关的噪声像素对特征提取的影响,并过滤掉无关目标.同时,为融合文本模态和图像模态,本文提出了一种融合注意力机制,通过挖掘文本与图像的内在联系,有效地综合两者特征;

(3)本文提出了一种算法框架,将相似广告扩充、转化人群扩充加入到现有的人群智能定向工作流中;

(4)本文在网络爬取构建的电商广告数据集上对 ToTYEmb 模型进行了评估,并与多种优秀基线模型(包括嵌入式、匹配式、预训练式)进行对比.实验结果表明,本文方法优于基线模型.

2　相关工作

本文研究跨模态语义融合问题,聚焦于提取图-文融合嵌入以改进目前常用的单模态嵌入方法,因此将从单模态的文本嵌入、图像嵌入角度对比分析现有工作.此外,目前在跨模态检索领域,也有一些关注图-文特征融合的工作,虽然任务不同,但模型架构上仍有可借鉴对比之处,故还从跨模态检索角度对比分析现有工作.

2.1　文本嵌入

早期的文本嵌入主要基于词袋模型,如 One-hot 编码、TF-IDF 等.词袋模型视词为独立特征,忽视了词与词的关联性,且无法体现语序特征.近些年来,研究者们基于深度网络建立语言模型以获得文本向量.Mikolov 等人[1] 提出 Word2Vec 模型,每个词会被映射到固定维度的特征空间,相似语义的词距离越近.然而 Word2Vec 的词向量是静态的,无法根据上下文的语境变化而改变.为了解决这一问题,Peters 等人[4] 提出一种动态文本嵌入技术,并提出了 ELMo 模型.该模型通过双向 LSTM 对语料库进

行建模，中间层的输出作为词向量. 为了更好得捕获长距离语言结果，OpenAI 团队提出了 GPT[13] 模型，他们利用 Transformer 学习一个通用的语义表示，并在下游任务上进行微调. BERT[5] 在 GPT 基础上进一步改进，通过遮挡部分预测词，从而达到双向语言模型的效果. 自 BERT 提出以来，越来越多的预训练语言模型被提出，并在各大复杂自然语义任务中取得卓越成果，例如基于 Transformer-XL 和Permutation语言模型的 XLNet[14]；基于动态掩蔽，移除下句预测，采用更大 batch size，更多语料的 RoBERTa[15]；基于 Knowledge Masking 的 ERNIE[16] 等.

大量实验证明基于 Transformer 架构的编码器能够很好地学习到文本的内在信息，本文同样采用 Transformer 抽取文本信息，并通过特定微调任务使其适用于广告的检索. 和已有文本嵌入解决方案不同的是，许多广告有着相同或相似的广告语，单从文本模态来区分它们是一件十分困难的事，因此本文融合了来自图像的信息.

2.2 图像嵌入

图像嵌入的策略与文本嵌入类似，通常使用深度网络抽取图像特征，中间层的输出作为最终的特征向量. 具体而言，利用卷积神经网(CNN)从像素级别挖掘图像的视觉信息，并使用不同尺寸的卷积核以及池化层，最终图像被压缩到固定维度的稠密向量. 目前，有多种不同的卷积网络模型可供使用. Simonyan 等人[8] 提出 VGG16 深度卷积网络，该网络采用 3×3 尺寸的卷积核以及 2×2 尺寸的最大池化，整个结构简单有效，具有很强的扩展性和泛化能力，可迁移到其他图片数据集上. 深度卷积网络在传递信息过程中，因为网络层数较多，可能会产生信息丢失，同时还存在梯度消失或者梯度爆炸问题. 为了解决这一问题，He 等人[17] 提出残差连接以缓解深度网络难以训练问题，并提出了 Resnet 卷积网络，该网络中有多条直连通路，直接将前层的输出连接到后层的输出. 不同于 VGG 和 Resnet 直接堆砌卷积核池化层，InceptionV4[9] 在宽度方向采用了多尺度的方式，同时使用不同大小的卷积核，扩展了网络的深度和宽度. 本文也使用深度卷积网络提取图像特征，但本文更加关注细粒度的区域特征，而非粗粒度的整体特征，因为广告图像包含大量的背景噪声，关键信息集中于小粒度的广告商品目标中. 本文采用 YOLO 作为目标检测器，从图像中提取目标，以减少无关像素点噪声影响. 同时本文融合文本模态的特征，更好地表征广告数据.

2.3 跨模态检索

跨模态检索关注于不同模态间的匹配问题，其通常输入文本(图像)，寻找匹配的图像(文本). 而本文关注于融合多模态，寻找多模态到多模态的对应关系. 虽然具体任务不同，但核心思路都在于融合图-文模态特征，故本节进行对比分析. 在跨模态检索领域，许多工作将文本和图像映射到同一个特征空间，直接根据向量的距离寻找匹配的文本或者图像. 这种策略过于粗粒度，忽视了一些局部特征，即文本的词和图像的区域. 为了更好地突出这些局部特征，一些工作切割文本和图片，并寻找它们之间的对应关系. Karpathy 和 Li[18] 使用双向 LSTM 挖掘文本的特征，利用 RCNN 抽取图像的区域特征，并计算每个单词和每个区域的相关度，最终得到匹配得分. Wang 等人[19] 在抽取区域特征过程中，加入区域的位置信息，并采用注意力机制使模型提取关键目标. 为了强化图像模态中的语义特征，Shi 等人[20] 利用知识图谱为图像扩充语义从而改进了图像的特征表示. Huang 等人[21] 同样利用图像的语义概念优化跨模态匹配，不同的是他们挖掘出图像概念后，为图像语义进行排序，突出语序特征.

近年来，预训练技术也被运用到多模态领域，为跨模态检索提供了新思路，并取得了出色的效果. Lu 等人[22] 提出 ViLBERT 模型，将 BERT 由单一的文本模态扩展为多模态双流模型，文本流和视觉流通过注意力 Transformer 层进行交互. Alberti 等人[23] 同样在同一个统一架构中利用了把单词指向图像中的一部分的参考信息的方法构建了一种简单但强有力的神经网络，在视觉常识推理数据集上表现优异. Tan 等人[24] 提出 LXMERT 模型，其由一个对象关系编码器、一个语言编码器和一个跨模态编码器组成，使用了 5 个不同的有代表性的预训练任务来提升模型性能. Su 等人[25] 提出基于单一编码器架构的 VL-BERT 模型，值得注意的是，为了保留图像全局信息，其将所有区域特征平均池化作为整体特征拼接到了模型中.

已有基于局部特征和预训练的跨模态相关工作在处理图像特征过程中，均通过切割图像或使用目标检测器提取区域特征，这些区域特征可能包含大量噪声信息. 本文利用 YOLO 作为图像目标特征提取器，以商品广告文本中的特征词为线索筛选图像中与商品相关区域，代替单纯使用边框置信度对目标进行排序，从而有效降低噪声目标的影响.

3　多模态特征融合的商品广告嵌入

本节将详细介绍基于多模态特征融合的商品广告嵌入模型(ToTYEmb). 模型主要由三个部分组成，分别是基于 Transformer 的文本特征提取器，基于 YOLO 的图像特征提取器，以及多模态融合组件，模型整体架构如图 2 所示. 具体而言，对于文本模态，本文使用词级预训练 Transformer 生成隐层特征向量. 对于图像模态，本文采用 YOLO 提取图中目标 region. 为了筛选出广告相关的 region，本文提出一种以文本为线索的注意力机制以给予关键 region 更大权重. 最后，本文利用一种 cross attention 策略，以融合文本语义和图像视觉语义，模型训练目标是细化商品分类预测，迭代训练完成后利用次顶层隐态向量作为一则广告的 fusion embedding 特征.

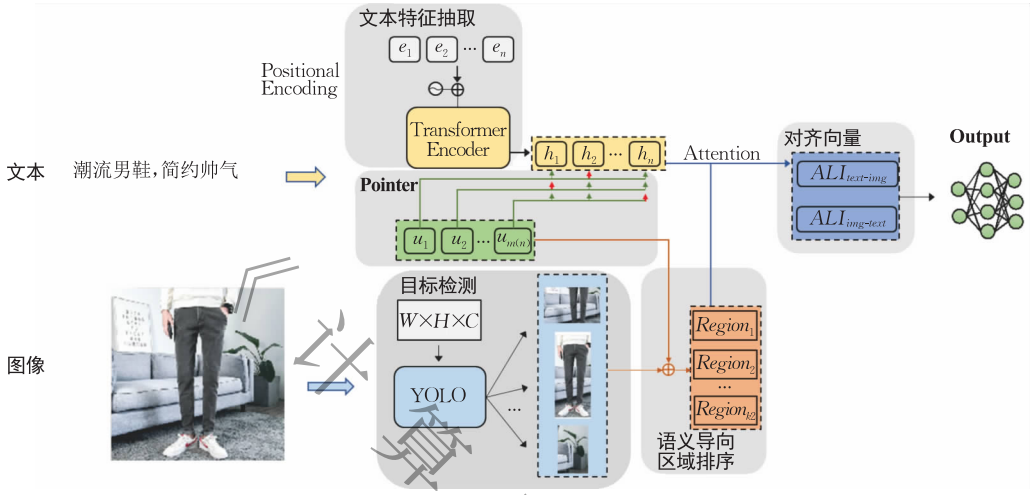

图 2　ToTYEmb 模型整体架构图

3.1　Transformer 编码器

广告文本多为短文本，为了能够有效挖掘其中的语义信息，从序列角度对文本进行建模，利用文本的语序信息，本文使用 Transformer[11] 编码器提取特征. Transformer 已被运用于多个预训练模型(如 BERT 等)的文本编码器，其可有效挖掘文本的语义、语序信息，且得益于 self-attention 机制，能够计算文本的关键信息. 给定一个广告文本序列 $X=\{w_1,w_2,\cdots,w_i\}$，其中 w_i 为输入序列中词语，$i\in[1,n]$，Transformer 编码过程如下：

$$\boldsymbol{e}_i=\boldsymbol{W}_e w_i \tag{1}$$

$$\boldsymbol{h}_i=\text{Transformer}(\boldsymbol{e}_i+\boldsymbol{p}_i) \tag{2}$$

其中，输入 w_i 由参数矩阵 $\boldsymbol{W}_e$ 转换成初始向量 $\boldsymbol{e}_i\in\mathbb{R}^{d_{\text{model}}}$，向量维度为 d_{model}. Transformer 不同于 RNN，非时序输入，需要单独的位置编码体现词的位置信息，本文采用和 Transformer 原文一致的位置编码. 接着将初始化词的 Embedding 和位置 Embedding 相加，通过 Multi-head attention 挖掘文本信息.

为了充分利用更多的无监督语料和一些原始标签信息(数据库中的广告行业一级类目)，本文采用类似 BERT[5] 的方式进行预训练，同时考虑到对于图文双模态双塔模型而言，太多的参数量会造成严重的收敛缓慢/过拟合现象，本文采用单层 Transformer Encoder 对文本部分进行建模，训练目标如：(1) 平均池化隐层输入向量后预测广告原始标签类目；(2) 文本输入模型时带有词性信息，以 15% 的概率遮蔽一些词性信息，利用模型预测缺失的词性. 如下式：

$$H_{\text{class}}=\text{Softmax}\left(\text{mean\_pooling}\left(\sum_{i\in\{1,\cdots,n\}}\boldsymbol{h}_i\right)\right) \tag{3}$$

$$Pos_i=\text{Softmax}(\text{Linear}(\boldsymbol{h}_i)) \tag{4}$$

3.2　YOLO 区域特征提取器

对于图像模态部分，本文更加关注于细粒度的 region 特征. 由于一幅图片中包含了大量像素点，如果采用类似 VGG、Resnet 等图像 Embedding 模型，将可能导致无关像素在特征向量中起主导作用，而真正关键的商品目标被忽略. 为了解决这一问题，本文提出利用目标检测模型 YOLO[7] 抽取图像的关键目标的区域，并基于区域表达特征信息. 相比于 Faster-RCNN 等两步计算模型，YOLO 能以更快的速度达到足够有竞争力的性能. 具体实现上，本文采用 YOLO-v3 版本进行实现，相比于前两代，其解决了小目标物体检测识别率低的问题. 对于一张图像，YOLO 可以识别出多个目标，本文取出这些目标对应的卷积特征，用全连接层映射到固定维度：

$$\boldsymbol{r}_j=F_v(\boldsymbol{v}_j) \tag{5}$$

其中 F_v 为全连接层，$\boldsymbol{v}_j$ 是 YOLO[7] 提取出来的卷积特征，$\boldsymbol{r}_j \in \mathbb{R}^{d_{\text{model}}}$ 是固定维度后的特征. 在提出来的特征中，仍然有许多噪声目标，因为识别出的目标，并不一定是商品相关的，极有可能是无效目标. 幸运的是广告文本中常常包含商品名词，以及商品相关特征词. 本文以文本信息为线索，对 YOLO 识别出的目标做筛选，保留商品相关目标，剔除无效目标.

值得注意的是，Transformer 编码器在预训练阶段并没有相应措施来寻找文本中对商品信息起决定性作用片段，因此单用 self-attention 来综合上下文信息是不够的，文本在文本编码信息指导图像目标信息筛选之前还利用了类似 pointer 的机制，以此选择原文中哪些词语是作为线索去指引图片模态重新排序目标候选集的，如下式：

$$\boldsymbol{v}_t = \tanh(\boldsymbol{W}_w \boldsymbol{h}_t + \boldsymbol{b}_w) \tag{6}$$

$$\alpha_t = \frac{\exp(\boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{v}_w)}{\sum_t \exp(\boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{v}_w)} \tag{7}$$

$$\boldsymbol{u}_i^{1:k} = concat(topk_{\alpha_t}(\boldsymbol{h}_t)) \tag{8}$$

然后用序列长度短得多的 $\boldsymbol{u}_i^{1:k}$ 来代替 Transformer 提取到的文本特征 $\boldsymbol{h}_i^{1:t}$，计算全局上下文特征：

$$\beta_i = \frac{\exp(\boldsymbol{W}_u \boldsymbol{u}_i)}{\sum \exp(\boldsymbol{W}_u \boldsymbol{u}_i)} \tag{9}$$

$$\boldsymbol{c} = \sum \beta_i \boldsymbol{u}_i \tag{10}$$

其中 β_i 是基于参数矩阵 $\boldsymbol{W}_u$ 得到的 $\boldsymbol{u}_i$ 权重. 并将它们的加权和作为全局上下文向量 $\boldsymbol{c}$. 下一步，基于该上下文向量从识别出的目标中筛选出商品相关目标，具体过程如下：

$$\gamma_j = \tanh(\boldsymbol{W}_c \boldsymbol{c} + \boldsymbol{W}_r \boldsymbol{r}_j) \tag{11}$$

其中 γ_j 代表了每个目标 $\boldsymbol{r}_j$ 和文本的相关度，本文依据 γ_j 值对目标进行排序，最终取 Top K 个目标 $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_K\}$ 作为图像的特征表示. 这样做有效避免了单纯凭借边框置信度取最高的区域特征在广告数据集上丢失信息的情况.

3.3 图文融合注意力机制

在得到文本特征 $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$ 和图像特征 $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_K\}$ 后，本文最终的目标是获得文本和图像的综合特征向量. 本文提出一种跨模态的融合注意力机制，能够融合不同模态的特征并挖掘不同模态间的相关性. 首先计算每个词和每个区域之间的相关度：

$$s_{ij} = \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{r}_j \tag{12}$$

其中 $\boldsymbol{h}_i^{\mathrm{T}}$ 为 $\boldsymbol{h}_i$ 的转置. 广告中图文是有内在关系的，文本描述着图像的视觉信息，而图像也是文本的语义体现. 为了凸显这种内在关系，本文用图片信息表征每个词，用文本信息表征每个图片区域：

$$\boldsymbol{h}_i' = \sum_{j=1}^{k} \frac{\exp(s_{ij})}{\sum_{m=1}^{k} s_{im}} \boldsymbol{r}_j \tag{13}$$

$$\boldsymbol{r}_j' = \sum_{i=1}^{n} \frac{\exp(s_{ij})}{\sum_{m=1}^{n} s_{mj}} \boldsymbol{h}_i \tag{14}$$

其中 $\boldsymbol{h}_i'$ 为每个词的图像信息表示，$\boldsymbol{r}_j'$ 为每个 region 的文本信息表示. 值得注意的是，这里的矩阵乘并取 Softmax 操作的部分本质上就是类似于 Transformer[11] 中的矩阵乘 Attention，不一样之处在于跨模态乘法将两个不同量纲的数值乘到了一起，而保证其有效性的前提是对各自模态的每一个分量都做了相同的乘法操作，并取 Softmax 作为 Attention 分数分布，对应式(13)、(14)中的分数部分.

使用统一的投影层对不同模态进行融合：

$$\bar{\boldsymbol{h}}_i = F_{\text{fusion}}^{\text{text}}([\boldsymbol{h}_i; \boldsymbol{h}_i']) \tag{15}$$

$$\bar{\boldsymbol{r}}_j = F_{\text{fusion}}^{\text{text}}([\boldsymbol{r}_i; \boldsymbol{r}_i']) \tag{16}$$

其中[;]表示两个向量的拼接. 使用最大池化层分别获得 $\{\bar{\boldsymbol{h}}_1, \bar{\boldsymbol{h}}_2, \cdots, \bar{\boldsymbol{h}}_n\}$ 和 $\{\bar{\boldsymbol{r}}_1, \bar{\boldsymbol{r}}_2, \cdots, \bar{\boldsymbol{r}}_k\}$ 的整体隐藏特征：

$$\bar{\boldsymbol{h}} = \text{maxpool}(\bar{\boldsymbol{h}}_1, \bar{\boldsymbol{h}}_2, \cdots, \bar{\boldsymbol{h}}_n) \tag{17}$$

$$\bar{\boldsymbol{r}} = \text{maxpool}(\bar{\boldsymbol{r}}_1, \bar{\boldsymbol{r}}_2, \cdots, \bar{\boldsymbol{r}}_k) \tag{18}$$

将所有特征向量组合成二维特征矩阵，最大池化大小为 $1 \times d_{\text{model}}$，步长为 1. 下一步，基于 $\bar{\boldsymbol{h}}$ 和 $\bar{\boldsymbol{r}}$ 预测广告对应商品类别：

$$\boldsymbol{o} = F_e([\bar{\boldsymbol{h}}, \bar{\boldsymbol{r}}]) \tag{19}$$

$$\hat{\boldsymbol{y}} = F_f(\boldsymbol{o}) \tag{20}$$

$\hat{\boldsymbol{y}}$ 经过 Softmax 后得到模型预测的广告商品概率分布，而在 inference 阶段，隐层向量 $\boldsymbol{o}$ 作为广告的特征向量表示，用于下游检索任务.

3.4 模型训练过程

本文基于一个基本假设：用户对同类商品有相同的兴趣(点击或购买)，即相似的商品具有相似的转化人群. 因此所提出的图文 fusion embedding 模型需要拟合商品细化类目. 细化类目由预先标注产生，详见第 5 节. 预测商品类别时，可使模型学习到商品类别的信息，以自动地从文本、图像中挖掘商品相关特征，同时学习如何融合特征，最终学会构建多模态特征的方法. 本文以多分类交叉熵为

损失函数：

$$loss=-\sum \boldsymbol{y}\log\hat{\boldsymbol{y}} \tag{21}$$

其中 $\boldsymbol{y}$ 为真实概率分布，$\hat{\boldsymbol{y}}$ 为模型预测概率分布. 在下游实际召回过程中，本文直接根据广告图文融合特征向量召回特征空间中最近的相似广告，而非根据分类结果召回. 因为测试集中有相当一部分商品类别从未在训练集中出现，在测试或者应用过程中，如果依据分类结果召回，范化能力较弱(没有学习过的类别永远不可能被分类正确). 本文模型侧重于自动地提取广告中多模态特征并获得图文融合嵌入向量，标签分类只是作为训练目标. 在真实应用场景中，对于任意一则商品广告，即使是新商品，只需根据特征向量在海量广告候选池中根据向量度量距离检索出相似广告即可.

4　相似广告检索下游框架

本节将详细介绍所提出的一种算法框架，将相似广告扩充、转化人群扩充加入到现有的人群智能定向更新工作流(pipeline)中，以提升投放广告的点击率、转化率等.

首先介绍现有的人群智能定向更新工作流，如图 3 中米黄色方框中的工作流程所示. 首先是广告投放. 由广告主经过广告创意模块设计合适的广告，然后根据一定经验规则筛选投放人群(例如选择 18～25 岁的使用华为手机的白领用户)，将对应广告投放到这批人群中，接外部模块进行后续流程，在图 3 中省略显示.

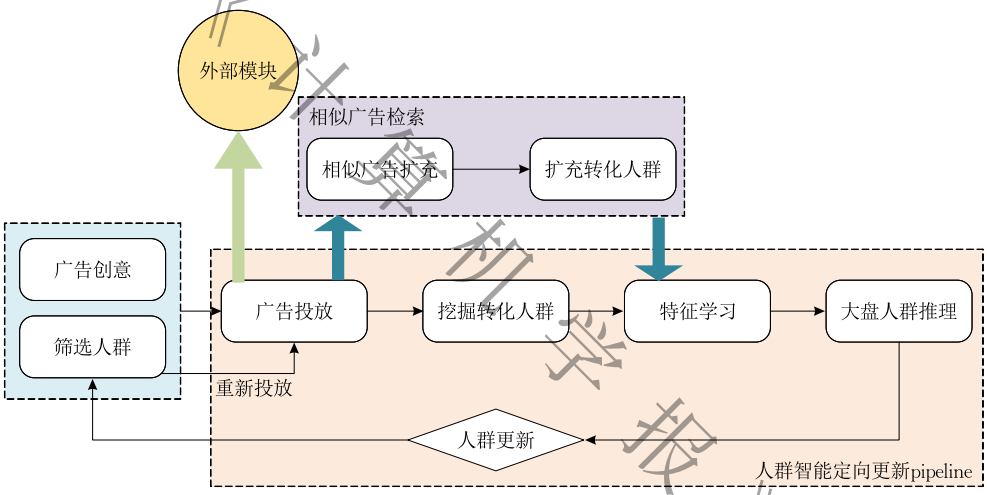

图 3　人群智能定向更新与可插拔相似广告检索的扩充框架

挖掘转化人群，即根据投放的广告在一定时间周期后，分析有哪些用户对这些广告进行了点击/收藏/购买等行为，并认为这批转化人群是真正容易对当前投放的广告产生兴趣的人群，将这些用户单独抽取出来.

特征学习，使用模型拟合这些用户的特征. 人群特征学习本文采用 XGBoost 模型，其为业内最常用点击预测模型之一，主要有以下优势：正则化项防止过拟合；二阶导数使损失更精确；并行计算优化提高模型效率；考虑了训练数据为稀疏值的情况，可以为缺失值或者指定的值指定分支的默认方向. 学习人群特征即根据人群中用户的离散特征训练预测其在该投放广告上点击/曝光未点击情况，正样本为抽取出的转化人群，负样本为在该广告的曝光下未发生点击行为的人群. 该 XGBoost 模型以机器学习评价指标中经典的 ROC 曲线下面积(AUC)来衡量模型的拟合情况.

大盘人群推理，人群特征学习器训练成功后，使用该模型对历史大盘内已注册的所有用户进行推断打分，根据模型预测分数从高到低排序，取出 Top$M$ 数量的用户(M 的值与最开始筛选出的人群大小一致).

人群更新，动态地将投放人群更新为这 M 个用户，然后将广告重新投放到这个人群中，进入下一个时间周期.

在这个过程中不难注意到，人群更新的效果严重依赖于特征学习器(这里指的 XGBoost 模型)，而在实际投放中经常会出现某些小众广告，在选定的人群中发生点击非常少，甚至只有不到 1000 个用户样本(在这样的情况下，人群更新会自动失效)，或只有

1000～5000 个样本(在这样的情景下,XGBoost 模型会存在非常严重的过拟合现象),模型的训练阶段 *AUC* 会飘高,而更新过后的人群在该批广告的下一周期投放时产生的转化比没有更新之前甚至更低.

针对这样的现状,本文所提出的相似广告检索模块可以根据历史投放的广告信息,从广告层面扩充相似广告,而进一步扩充转化人群,让人群特征学习器的性能更加鲁棒,同时让智能定向更新人群功能在更多场景下得以运用.如图 3 中紫色方框所示,假设当前广告集合的点击人群少于 500 人,则根据集合中每一条广告的图文融合嵌入,采用距离度量函数从当前周期所有投放的广告中召回最相似的广告集合,并将这些广告的点击人群也添加到原始人群中,用户数量从少于 500 来到了 10 万以上.相似广告检索模块在该框架中是以一种可插拔的形式存在的,即在带来收益的同时,完全不会影响到现有模块的运行.

5 实验及分析

本节将详细介绍实验数据集的构建、模型参数设置、基线模型、离线实验,并展示和分析实验结果.

5.1 数据集

为了使得实验数据的真实有效,从某电商平台上爬取电商广告数据,广告由图片和文本构成.每一条广告由人工标注其商品类别标签,筛选保留图片和文本都存在的广告后,一共 32 865 条广告,本文将该数据集随机拆分十等份,其中 6 份作为训练集,1 份作为验证集,3 份作为测试集.图 4 中左图表示广告文本经过分词后序列长度分布,右图表示广告图片经过 YOLO-v3 模型提取高置信度(阈值为 0.3)目标对象后对象数量分布.从图中可以得知广告文本序列长度普遍在 10～30 个词语之间,绝大部分的广告图片中包含的目标对象在 5 个以下.值得注意的是,标注的过程中尽可能以最细的粒度商品类别标注,例如,对于茶叶商品,细分成绿茶、红茶、普洱茶等等,即对应到一份在某电商平台三级类目体系上加以人工修正制定的类别表.完成标注后,一共有 624 个不同类别.这些经过标注的广告数据用以训练、测试本文所提出的图文跨模态特征融合 embedding 学习模型.相似广告检索下游人群智能定向更新实验采用在上述训练数据集中训练完毕的模型,在真实线上广告数据、用户点击流水记录表以及历史大盘用户信息.

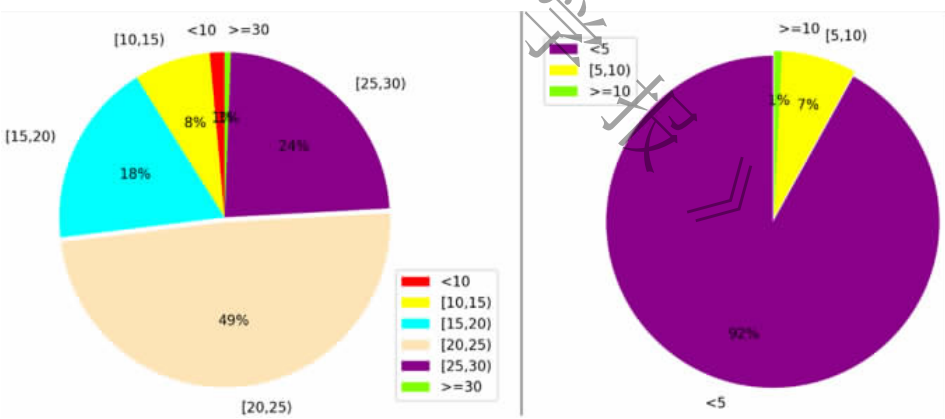

图 4 数据集中文本序列长度和图片对象数量分布

如第 3 节所述,Transformer Encoder 将先在大量未标注数据上进行预训练,因此从网络上爬取了约 384 万的广告文本用于预训练 Transformer 模块.对于 YOLO,使用 COCO(http://cocodataset.org/)数据集预训练目标检测.

5.2 实验设置

在广告文本中,通常形容词,动词,名词等词性表现出更多的特征.预训练 Transformer 时,将上述关键词性的预测作为训练目标.所有的特征向量大小被设置为 128,即 $d_{\text{model}}=128$.在使用全文信息寻找最相关的图像区域后,取 Top 5 区域来表示图像特征,即 $k=5$.训练过程中,Batch size 大小为 16,共训练 10 个 epoch.本文使用 Ada 优化器迭代参数.评价实验结果时,主要采用 *precision@k* 指标.具体而言,对于测试集,首先将所有广告的图文融合嵌入写入数据库中(真实应用也是同样流程)作为一个字段属性,然后对测试集中的每一条广告,用该广告的图文融合嵌入与测试集中其他所有广告计算欧式距

离，并召回 Top k 候选集，测试 k 个结果的精确率. 即这 k 条广告中有 m 条与待测样本属于同类商品（商品细化标签相同），则其 $precision@k=m/k$. 对于总体的 $precision@k$ 而言，则是将每一条样本的 $precision$ 加和取平均得到.

5.3　基线模型介绍

Word2Vec[1]. Word2Vec 通过建立神经网络语言模型学习词的特征向量. 本文计算词向量平均值作为文本的表示.

ELMo[4]. ELMo 模型通过双向 LSTM 对语料库进行建模，中间层的输出作为词向量. 本文计算词向量平均值作为文本的表示.

GPT[13]. GPT 是基于 Transformer 的单向语言模型，相较于基于 LSTM 结构的模型能捕获长距离依赖. 本文计算词向量的平均值作为文本的表示.

BERT[5]. BERT 是基于 Transformer 的一种双向语言模型，经微调后可应用于下游任务. 本文使用 400 万广告数据预训练 BERT 并在标注电商广告数据集微调，计算隐层 token 向量平均值作为文本的表示.

RoBERTa[15]. RoBERTa 是针对 BERT 进行的一项改进，其基于动态掩蔽，移除下句预测，采用更大 batch size，在更多语料上进行预训练，以获取更好的文本建模能力，对于该基线模型，同样采取隐层 token 向量平均值作为文本表示.

VGG16[8]. VGG16 是一种深度卷积网络，在图像分类问题中有较好的效果. 本文基于 ImageNet (http://www.image-net.org/) 预训练 VGG16 并在本文实验数据集上进行微调，使用池化层的输出作为图像的特征向量.

InceptionV4[9]. InceptionV4 也是一种深度卷积网络. 同样我们使用 ImageNet 预训练的 InceptionV4，并在本文实验数据集上进行微调，取池化层的输出作为图像的特征向量.

YOLO[7]. 本文还使用 YOLO 作为基线模型，从细粒度计算图像的相似度. 具体而言，在识别出图像目标后，筛选出前 10 个置信度最大的目标作为图像特征表示.

ViLBERT[22]. ViLBERT 作为多模态预训练模型的代表之一，其在视觉问答、视觉常识推理、指示表达和基于字幕的图像检索这四个视觉-语言任务上表现出优秀的性能，并且其证明了在 Flickr30k[6] 数据集上 Zero-Shot 推理能力，故本文将其引入做对比实验. 值得注意的是，这类预训练模型是为了几个公开挑战而设计的，其更关注跨模态检索，而不是融合模态检索，无法直接用来做对比. 因此我们对其进行了改进，保持其主干网络不变，输入文本和图片，取出其做关系预测前的隐层向量，作为类似于本文的图文融合嵌入，即可以进行实验对比，将图片和文本特征融合还分为加法融合和乘法融合（原文做法）.

ESIM[26]. ESIM 是一种相对传统的文本相似度计算模型，利用精心设计的链式 LSTM 顺序推理模型可以胜过以前很多复杂的模型，在多项文本匹配任务上达到非常好的性能.

ABCNN[27]. 不同于 ESIM 模型，该模型试图使用卷积神经网络（CNN）而不是 RNN 来对文本进行特征提取，同时结合了 CNN 与 attention 机制用来做文本匹配，同样在传统文本匹配任务上表现优秀.

DIIN[28]. DIIN 模型针对自然语言推断（NLI）任务提出一种 IIN 模型，其主要包括基础 Embedding 层、Encoding 层、交互层、特征提取层以及输出层，利用字向量＋动态更新词向量优化模型性能，在 NLI 任务上取得非常优秀的成绩.

值得注意的是，ESIM、ABCNN、DIIN 模型是在文本匹配任务中表现优异的经典模型，因此可以凭借计算测试集数据中两两匹配分数作为其距离度量指标并用来做相似广告召回实验对比，但这三者无法进行判别模型分类实验的对比.

5.4　实验及结果分析

5.4.1　判别模型分类实验

模型训练过程以商品细化类别预测为目标，因此本节对模型多模态分类实验进行对比. 模型输入为广告的文本描述以及图片描述，输出为模型预测的广告细化类别（共 624 个不同类别），当对比模型为文本建模方法时只针对文本信息进行分析，当对比模型为图片建模方法时只对图片信息进行分析. 不同的分类模型在测试集上的预测效果对比如表 1 所示. ELMo/GPT/BERT/RoBERTa 代表单独使用文本特征进行广告分类，VGG16/InceptionV4/YOLO 代表单独使用图片特征进行广告分类，Trans＋VGG16/Trans＋YOLO 代表文本部分采用 Transformer Encoder 编码特征，图片部分采用 VGG16/YOLO 进行区域特征提取，然后使用池化特征与文本特征拼接作为下游分类依据. 值得注意的是，这里采用的 VGG16 模型得到高维卷积特征后，采用了一种类似 YOLO 的物理切分方式，将卷积特征池化为 7×7 的小方格特征作为区域性特征与 YOLO 对齐.

表 1 本文方法和基线方法类别分类效果对比

Methods	*Accuracy*
ELMo[4]	0.8201
GPT[13]	0.8001
BERT[5]	0.8390
RoBERTa[15]	0.8451
VGG16[8]	0.7802
InceptionV4[9]	0.7989
YOLO[7]	0.5981
Trans+VGG16	0.8532
Trans+YOLO	0.8401
ToTYEmb(Ours)	**0.8641**

从广告分类实验结果来看，InceptionV4 的分类准确率比 VGG16 的分类准确率提升了约 2 个百分点，其展示了更优的图像特征提取能力. 同时，基于图片的分类性能普遍低于基于文本的分类方法，BERT 模型比 InceptionV4 模型的分类准确率高了接近 4 个百分点，这是因为图像的背景部分包含了大量的噪声，基于图像的检索往往会找到背景相似的广告，而不是真正商品相似的广告. 而广告文本中常常包含商品词、商品的特征词，可以直接体现商品特征. 除此之外，在多数情况下，不同商品的文本差异很大，所以基于文本的检索能够更好地捕捉广告特征，从而效果较图像检索好一些. RoBERTa 模型的效果好于 BERT 模型，这表明对于文本类别信息建模，基于动态 mask，移除无关任务，更多训练预热的 RoBERTa 的确更优. 单独使用 YOLO 模型的分类性能甚至远远低于平均水平，相比于粗粒度的 VGG16，InceptionV4 效果更差，这是因为 YOLO 抽取出的图像目标不一定是商品相关的，在分类模型中其会对图像目标按照置信度排序并取 Top$K$，如果不加以引导，极可能最后保留的 K 个图像目标全都与商品目标无关，从而造成检索错误. 本文所提出的模型在分类任务上性能最优，这表明模型中依据文本线索对 YOLO 抽取出的目标重新排序的确能缓解了噪声目标问题.

值得注意的是，表 1 中的最后三种图文融合建模的方法隐含了许多重要信息. Trans+VGG16 方式采用物理方式将图片划分为 7×7 的小方格特征；Trans+YOLO 方式采用 YOLO 模型的置信度分值对提取出的候选框进行排序并取 Top$K$；而本文提出的模型是在 Trans+YOLO 方式的基础上，将文本信息和 YOLO 模型置信度信息进行 Attention 融合，这一步即整体模型框图中的“语义导向区域排序”. 实验表明，Trans+VGG16 模型的性能优于不使用“语义导向区域排序”的 Trans+YOLO 模型，这表明在没有文本线索指引的前提下，从物理角度对高维卷积特征进行规则化区域划分能够在一定程度上缓解目标不相关的问题，因为在图文 cross-attention 模块，同样会对物理划分产生的小区间进行加权学习，这样的区间信息可能精度上会有一些不足，但不会出现完全丢掉正确目标信息的情况，并且这样做模型的参数量更小，在对推理速度有较高要求的情况下 Trans+VGG16 也是一种非常好的选择. 然而由于物理划分并不存在置信度分数，故而无法引入文本线索. 显然，在 YOLO 模型提取出的候选框基础上加入“语义导向区域排序”能够更加细粒度地提取文本相关的图像目标特征，因此本文所提出的模型能够达到最优的性能.

5.4.2 相似广告召回实验

本节进行相似广告召回(检索)实验对比. 对于图像/文本/图文融合嵌入的检索方法，首先利用训练好的模型对每一则广告根据其文本和图像，抽取图文融合嵌入向量，然后利用向量欧式距离，从测试集中除自己以外的所有广告中排序召回距离最小的 K 则广告，判断其是否与自己同属于一个细分商品类目，如果同一类目则召回成功，否则召回失败.

而对于基于匹配打分的检索方法，由于目前的多模态匹配式检索主要是以图搜文或以文搜图，所以这里对比了传统基于文本内容匹配的 ESIM、ABCNN 以及 DIIN 模型，基于文本匹配的模型检索阶段与基于嵌入的检索略有不同，输入为单个样本与测试集内其他所有样本组成的文本对(TextA，Text B)，经由模型打分计算匹配分数，定义语义距离为 1−匹配分数，代表 A 文本和 B 文本之间的语义距离. 然后取出语义距离最小的 K 则广告，如果同一类目则召回成功，否则召回失败. 基于匹配式的方法还存在一个数据集构建的问题，我们对训练、验证、测试集的构建采取同一种策略：在原先数据集的基础上，对每一条样本，随机采样三条同一类别的作为正样本，负样本则从商品细化类目与该样品商品细化类目不同，但文本描述上有交集的样本中随机采样三条，尽可能使模型能够学习到相近类目的商品广告特征. 具体而言，19719 条训练数据能够构造 114901 条文本对训练数据，验证集、测试集同理. 另外，对于基于匹配打分的方法，将测试集所有一一组成文本对测试在时间成本上难以满足要求(每一条文本对应 1 万条测试样本，一共 1 亿条样本)，故而随机采样了 1000 条测试样本进行试验对比. 表 2 详细描述了不同方法之间的召回对比效果. Matching *Acc* 表示在测试集上所有文本对的匹配/不匹配分类准确率，$P@K$ 表示相似广告检索 Top$K$ 精确率.

表 2　本文方法和基线方法召回实验效果对比

Methods	Matching *Acc*	*P*@1	*P*@5	*P*@10
Word2Vec[1]	—	0.4964	0.4152	0.3797
ELMo[4]	—	0.6961	0.5877	0.5539
GPT[13]	—	0.7337	0.6445	0.5923
GPT-fine tune[13]	—	0.7938	0.7507	0.7218
BERT[5]	—	0.7211	0.6358	0.5855
BERT_EN[5]	—	0.6869	0.5922	0.5379
RoBERTa[15]	—	0.7526	0.6720	0.6256
VGG16[8]	—	0.5802	0.5456	0.5258
InceptionV4[9]	—	0.5912	0.5613	0.5399
YOLO[7]	—	0.4704	0.4607	0.4559
ESIM[26]	0.9356	0.2920	0.4096	0.4305
ABCNN[27]	0.9218	0.2510	0.3216	0.3512
DIIN[28]	0.9418	0.4270	0.4121	0.4065
ViLBERT_Sum_Emb_EN	—	0.7154	0.6139	0.5604
ViLBERT_Mul_Emb_EN	—	0.7354	0.6388	0.5831
Trans+VGG16	—	0.8215	0.8022	0.7849
Trans+YOLO	—	0.8143	0.7875	0.7691
ToTYEmb (Ours)	**—**	**0.8343**	**0.8092**	**0.7925**

根据表 2 实验结果，整体来看，依据本文图文融合 embedding 思想所构建的最后三种模型在 *precision@k* 的指标上均远远高出单独从文本/图片角度依据 embedding/matching 方式召回相似广告，Top 10 精确率至少高出 15 个以上的百分点，这表明融合文本与图像信息对于相似广告检索而言是至关重要的. 同时，本文提出的模型在 *precision@k* 上的结果都明显优于 Trans+VGG16/Trans+YOLO 方式，其 Top 10 精确率高达 79.25%，明显优于第二好的模型. 基于文本特征 embedding 的 BERT/RoBERTa/GPT 模型效果比基于图像特征的 VGG16/InceptionV4/YOLO 效果高出 12 个百分点以上，这说明在广告中存在非常严重的图片同质化问题，仅从单一模态来召回相似广告的话，显然文本特征是更好的选择. 表 2 中的 GPT、BERT、RoBERTa 模型都是基于原始预训练模型进行特征提取，这么做是为了和近年来最好的通用文本特征提取器进行对比. 同时公平起见，表 2 中也对比了在本文数据集上进行微调后的 GPT，即 GPT-fine tune，从表中可以看到其性能仍明显逊于 ToTYEmb.

另外，表 2 中所列举的模型将基于 embedding 距离计算的方式和基于 matching 打分方式的方法进行了融合比较，并对比了在传统 text matching/NLI 任务上表现出色的 ESIM/ABCNN/DIIN 模型，实验结果表明，在测试集抽样文本对匹配分类准确率达到 92%以上的情况下，其相似广告召回精确率远远低于基于 embedding 的方法，分析主要原因如下：(1) 基于 matching 方式的模型训练严重依赖负采样方法，本文所使用的方法是在类目文本描述上有交集的样本中随机采样 3 则，而实际上这样的负采样本身也是一则非常困难的任务；(2) 匹配打分在模型推理阶段，面临很严重的过拟合/区分度小的问题，这体现在测试集文本对匹配/不匹配分类准确率达到 92%的情况下，其相似广告的召回性能低下，对于很多样本其打分集中在一个很小的分数区间内，区分度低，全局召回时很难保证精确率.

本文还对比了多模态预训练模型 ViLBERT，这里需要注意两点，首先该模型是为了视觉常识推理、以文搜图等任务设计的，无法直接用于本文相似广告检索问题，故而这里对其进行了一些改动，具体而言，保持其主干网络不变，输入文本和图片，取出其做关系预测前的隐层向量，作为类似于本文的图文融合嵌入. 其次，目前还没有一个中文多模态预训练工作，因此这里我们将所有广告的文本部分利用百度翻译接口翻译成了英文文本. 而为了验证翻译后的信息损坏情况，我们也对比了单纯基于英文文本的 BERT 嵌入对比，即表 2 中的 BERT_EN. 对于 ViLBERT 最后文本和图像隐层的交互，我们对比了基于加法的交互(ViLBERT_Sum_Emb_EN)和给予乘法的交互(ViLBERT_Mul_Emb_EN，原文做法). 从实验结果来看，BERT_EN 的 Top10 精确率为 0.5379，而 ViLBERT_Sum_Emb_EN 的 Top10 精确率为 0.5604，ViLBERT_Mul_Emb_EN 的 Top10 精确率为 0.5831，这说明，首先，中文转英文的过程中造成了一些信息缺失与错误，导致它们的性能比同架构直接基于中文的模型性能要差一些. 另外，再一次证明了融合文本与图像信息对于相似广告检索而言是至关重要的. 而遗憾的是由于 ViLBERT 模型直接从 Faster-RCNN 提取区域特征作为图像

侧的原始特征，没有考虑到广告领域图片的特性，其相比于 BERT_EN 的提升不是十分显著.

综上所述，本文所提出的模型能够在综合考虑图文特征的情况下学习到其融合特征表示，在相似广告检索任务上取得最优的性能.

5.4.3 人群更新离线实验

前述章节已验证本文所提出的模型在相似广告召回方面的性能，本节从下游人群更新离线实验角度验证其相似广告扩充结果对现有人群更新框架的正面影响. 如第 4 节所述，相似广告扩充能以一种可插拔的形式加入到现有的人群智能定向更新工作流中，其在原始广告转化人群不足时收益会最大化，因此本节专门针对转化人群相对不足的样本进行分析(广告投放后有效转化人群在 50000 以下). 其中人群包用户数据、绑定广告信息、用户点击流水均来自真实线上广告数据. 在总共 68 次人群更新转化 pipeline 中进行实验嵌入相似广告扩充模块，结果样例如表 3 所示. 人群包大小指的是最初依据一些特征选择的原始人群中用户数量；扩充前转化是指最初投放的广告在一定周期内发生点击/收藏/购买等行为的用户数量；扩充后转化是指把根据相似广告检索出来的新的广告在同样周期内发生的点击/收藏/购买用户加入到扩充前的转化人群中，新的用户数量；扩充前 *AUC* 是指根据扩充前的用户的离散特征(如身高、体重等)训练点击/未点击 XGB 模型，其 *AUC* 值；扩充后 *AUC* 是指在新的人群上重新训练该模型的 *AUC* 值；对照组 *CTR* 表示在原始人群更新方法进行人群更新后(或转化人群<1000，无法更新)，该人群在这批广告的投放下，下一个周期内的点击率，该点击率的计算即 *CTR*=点击数/曝光用户数；实验组 *CTR* 表示在相似广告扩充方式下人群更新后的点击率.

表 3 人群更新转化 pipeline 实验+相似广告扩充(Top-3)情况

人群包大小	扩充前转化	扩充后转化	扩充前 *AUC*	扩充后 *AUC*	对照组 *CTR*	实验组 *CTR*
49228410	1932	109908	0.95700	0.89440	0.010590	0.01172
20000000	2014	69652	0.99080	0.96516	0.007582	0.00761
67209738	4232	500000	0.86106	0.95229	0.070310	0.07450
34259	24688	308423	0.94398	0.97809	0.062700	0.07250
163583	3305	55328	0.96480	0.98734	0.052660	0.06198
49228410	932	138003	—	0.90960	0.010590	0.01181

首先，仅仅是对投放广告中每一条广告寻找其 Top-3 相似广告，就能使转化人群数量大大提升. 在原先转化人群很少(小于 2000)时会存在严重的过拟合现象，体现在 XGB 模型的 *AUC* 很高，而经过人群扩充后相应 *AUC* 降低. 而在原始转化人群数量不算太少时，扩充人群会得到模型 *AUC* 上的提升. 另外，如最后一个样例描述，原先转化人群数量不足 1000 时无法使用用户特征训练 XGB 模型，故无法启动人群智能定向更新，而采用了相似广告扩充-扩充对应转化人群的方法后，人群数量从 932 来到了 138003，启动人群智能定向更新后与原始人群投放效果对比，也能体现出更优的 *CTR* 情况.

整体而言，68 次人群智能定向更新实验中，实验组 *CTR* 高于对照组的占比为 48/65，约为 74%. 该结果表明将相似广告扩充模块嵌入到现有的人群更新转化框架中在大多数情况下都能取得正向的收益. 当然，从 *AUC* 上的现象可以看出本文提出的方法仍然存在一定缺陷，比如转化人群可能扩充过多，导致人群特征的发散，进而导致实验组 *CTR* 反而下降，这是在 bad case 中普遍存在的一个现象.

5.4.4 消融实验

为了透彻地研究本文所提出的多模态特征融合模型中各个组成部分对整体模型性能的影响，本节进行消融实验研究. 如表 4 所示，w/o Pointer 表示去除 Pointer 组件，以纯文本隐态特征指导图像区域排序；w/o 文本线索表示去除语义导向区域排序模块，即图像模态仅仅以 YOLO 预测置信度作为目标排序标准，直接取 Top k 的区域特征作为图像特征，即 5.4.2 节中所展示的 Trans+YOLO 方式；w/o 图像模态表示摒弃图像信息，单独从文本角度进行分类训练，以训练后的文本嵌入特征进行相似广告召回；w/o 文本模态表示摒弃文本信息，单独从图像角度进行分类训练即相似广告召回.

表 4 消融实验

方法	*P*@1	*P*@5	*P*@10
ToTYEmb (Ours)	**0.8343**	**0.8092**	**0.7925**
w/o Pointer	0.8307	0.8057	0.7887
w/o 文本线索	0.8143	0.7875	0.7691
w/o 图像模态	0.8023	0.7674	0.7426
w/o 文本模态	0.4704	0.4607	0.4559

表 4 显示，当删除 Pointer 组件时，总体 *P*@10 性能从 0.7925 下降到 0.7887，证明了在以文本为线索指引图像候选区域排序时，提炼兴趣点，缩减序列长度对于总体性能起到正面的影响. 当删除文本

线索模块时，总体 $P@10$ 性能下降到 0.7691，并在 $P@1$ 与 $P@5$ 性能上下降都超过 2 个百分点，这说明仅仅依靠 YOLO 模型自带的区域边框置信度排序取 Top$K$ 会损失许多有用的图像特征. 当删除图像模态，而仅仅依靠文本模态进行特征学习时，$P@10$ 性能下降到 0.7426，而删除文本模态，仅仅依靠图像模态进行特征学习时，$P@10$ 性能骤降至 0.4559，这说明了两点现象：(1) 从单个模态召回相似广告会损失不少有用信息，很多商品广告需要综合文本和图像来召回最相似的广告；(2) 相对于图片而言，大多数的商品广告文本已经具有较好的区分度，会出错的一些情况通常存在于同质化的广告文案，如“11.11 全场八折”，而对于图像来说，更多时候一整张图片中只有一个很小的部分是在描述目标商品，如本文图 1 中所示样例，这就造成了单独依赖图像模态精确度大大下降.

总的来说，本文模型取得了最好效果，得益于融合了多模态，挖掘出模态间的内在联系，使得特征向量更加健壮. 而且本文模型采用了 YOLO 挖掘细粒度广告图像特征，并以语义信息为线索抽取商品相关目标，使得抽取出的目标是语义相关的，所以模型效果更进一步.

5.4.5　目标数量 k 值分析

在目标检测过程中，本文根据文本语义对原始高置信度目标进行重排，取 Top k 目标作为图像特征，本节尝试不同 k 值分析不同目标数量对最终结果的影响，具体实验结果如表 5 所示.

表 5　ToTYEmb 模型图像目标 Top$k$ 取值实验

目标 k 值	$P@1$	$P@5$	$P@10$
Ours，$k=5$	**0.8343**	**0.8092**	**0.7925**
$k=1$	0.8258	0.8011	0.7881
$k=10$	0.8311	0.8061	0.7894
$k=15$	0.8316	0.8044	0.7896
$k=20$	0.8288	0.8032	0.7875

从表 5 中可以看出，当 k 取 5 时，ToTYEmb 模型能达到最优的性能，k 值过小或较大都会造成性能的损失. k 值较小时，检测出的目标过少，导致图像视觉特征部分丢失，从而造成性能下降；k 值较大时，检测出的目标更有可能包含背景噪声，也会造成性能下降. 总的来说，ToTYEmb 模型的性能会受到 k 值的影响，但由于数据集图像中包含目标对象数量普遍较少，且本文采用图文融合注意力机制对图文模态特征进行融合并使用最大池化缩减最终特征维度，不同的 k 值不会造成模型性能太大的波动.

5.4.6　实例分析

为了更好地展示本文模型在相似商品检索中的实际效果，从测试集中选取一条广告，并分别用 ToTYEmb（本文模型）、RoBERTa 和 InceptionV4 召回 Top 1 相似广告，并标注出在文本指引图像区域排序时最感兴趣的几个词语，如图 5 所示. 模型的

图 5　实例分析：输入是一则唇釉广告，本文模型检索出相似唇釉广告，RoBERTa 模型检索出男羊毛衫广告，而 InceptionV4 检索出文具/耗材-笔类广告

输入是一则唇釉商品广告，由描述文本以及图片组成，基于 RoBERTa 的检索只是检索出文本相似的广告，其是一则男羊毛衫的商品广告，检索失败. 两则广告中都表达了和冬天相关的句式，而输入广告的描述文本中对冬天的描述显然并不是关键信息，这些信息错误地被 RoBERTa 捕获. 同样，InceptionV4 模型只使用了图像信息，召回了图像物体轮廓相似的文具/耗材-笔类广告. 基于粗粒度像素级别的图像特征嵌入模型更关注于整体图像的特征以及一些轮廓特征，导致背景、轮廓噪声影响过大. 而 ToTYEmb（本文模型）成功检索出语义上有一定关联，图像目标特征上也较为相似的唇釉商品广告，这是因为其融合文本和图像模态召回相似广告，弥补单一模态的信息不足. 且文本信息指导下的目标识别可以识别出商品相关区域，从而减少了噪声影响.

6 总结与展望

本文提出一种多模态图文特征融合模型 ToTYEmb，其能提取图文嵌入（Embedding）作为广告的融合内容特征，同时融合文本语义信息和图像的视觉信息，解决相似商品广告检索问题，进而将其作为可插拔组件加入到现有人群智能定向更新框架中，提升广告投放推荐的效果. 本文使用 Transformer 提取文本的语义信息，使用 YOLO 抽取图像中的细粒度的视觉信息，接着利用文本为线索筛选出商品相关区域目标特征，最后基于融合注意力机制融合两个模态的信息. 相比于其他方法，本文方法具有明显优势：(1) 利用 YOLO 以及基于文本线索的注意力机制，可提取出商品对应区域目标特征，从而减少背景噪声和无关目标；(2) 考虑到广告商品中文本通常具有相对更好的区分度，以文本为线索，指引 YOLO 模块区域排序，避免丢失重要信息；(3) 多模态融合注意力机制有效融合了文本模态和图像模态，使得特征向量更加健壮. 本文方案有效提升了相似商品广告的检索质量，并且以可插拔的形式嵌入离线定向人群更新框架，能够在很大程度上提升商品广告转化效果.

本文方案仍然存在一些不足的地方亟待改进，在之后的工作中，我们将会进行以下尝试：(1) 除了区域化目标特征之外，尝试采用不同方法提取图像中的文本、节点信息，以丰富多模态特征；(2) 探究模型如何从现有的海量图文数据上学习模态之间的信息转化关系，以处理现实场景中可能单一模态缺失的情况. 此外，目前基于中文的多模态预训练工作比较缺失，未来我们将考虑在更大的数据集上采用预训练的方法训练模型，使其在具体领域的下游任务中具有更好的性能.

致 谢 感谢为本文原稿提供宝贵修改意见的匿名审稿人！

参 考 文 献

[1] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems. Nevada, USA, 2013: 3111-3119

[2] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1532-1543

[3] Le Q, Mikolov T. Distributed representations of sentences and documents//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 1188-1196

[4] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations//Proceedings of the North American Chapter of the Association for Computational Linguistics. Louisiana, USA, 2018: 2227-2237

[5] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 4171-4186

[6] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[9] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning//Proceedings of the Conference on Artificial Intelligence. San Francisco, USA, 2017: 4278-4284

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks//

Proceedings of the Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[12] Zhao Qi-Lu, Li Zong-Min. Cross-modal social image clustering. Chinese Journal of Computers, 2018, 41(1): 98-111 (in Chinese)
(赵其鲁，李宗民. 跨模态社交图像聚类. 计算机学报，2018，41(1)：98-111)

[13] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. URL https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf, 2018

[14] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 5753-5763

[15] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. URL https://arxiv.org/pdf/1907.11692.pdf, 2019

[16] Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding//Proceedings of the Conference on Artificial Intelligence. New York, USA, 2020: 8968-8975

[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[18] Karpathy A, Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3128-3137

[19] Wang Y, Yang H, Qian X, et al. Position focused attention network for image-text matching//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 3792-3798

[20] Shi B, Ji L, Lu P, et al. Knowledge aware semantic concept expansion for image-text matching//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 5182-5189

[21] Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6163-6171

[22] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 13-23

[23] Alberti C, Ling J, Collins M, et al. Fusion of detected objects in text for visual question answering//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong, China, 2019: 2131-2140

[24] Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from transformers//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong, China, 2019: 5099-5110

[25] Su W, Zhu X, Cao Y, et al. VL-BERT: Pre-training of generic visual-linguistic representations//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[26] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for natural language inference//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1657-1668

[27] Yin W, Schütze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272

[28] Gong Y, Luo H, Zhang J. Natural language inference over interaction space//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

FENG Yi, Ph. D. candidate. His research interests include natural language processing and machine learning.

ZHOU Xiao-Song, M. S. candidate. His research interests include natural language processing and machine learning.

LI Chuan-Yi, Ph. D., research assistant. His research interests include natural language processing and machine learning.

WANG Ting, Ph. D. His research interests include natural language processing and machine learning.

GE Ji-Dong, Ph. D., associate professor. His research interests include software engineering and natural language processing.

HU Yu-Cheng, M. S. His research interests include recommender system and natural language processing.

ZHANG Xiao-Peng, M. S. His main research interest is recommender system.

LUO Bin, Ph. D., professor. His main research interest is software engineering.

Background

Similar advertisement retrieval is of great significance in academic and industrial fields. This paper studies how to better retrieve similar advertisements from the comprehensive characteristics of the two modalities of text and images, and innovatively use the converted users corresponding to similar ads to optimize the existing automatic update framework of crowd packages in the industry. For similar ads retrieval, many scholars have used more complex text embedding models (e.g. BERT/RoBERTa/ERNIE) and image embedding models (e.g. VGG/InceptionV4/ResNet) to improve task effects. At the same time, many excellent cross-modal matching models have also been used to solve this problem. However, the embedding model for a single mode loses a lot of valuable information, causing the model to fail in some cases. The cross-modal matching model also has two serious problems: (1) The model is not highly distinguished in the inference stage, which causes the accuracy of Top$K$ to decrease; (2) In the advertising field, text usually has a guiding role. These models do not take this into consideration.

This paper proposes the fusion Embedding model ToTYEmb, which integrates text and image information, to make up for the lack of monomodal Embedding information. At the same time, it innovatively uses text information as a guide to improve the performance of image components and improve the overall performance of the model. Experiments show that ToTYEmb surpasses the multiple baseline models compared in similar ads retrieval tasks. At the same time, we innovatively propose a method that can integrate similar ads retrieval technology into the existing crowd package update framework, and experiments have proved that it improves the performance of existing recommendation tasks.

This work is supported by the National Natural Science Foundation of China (No. 61802167) and Tencent.