基于循环生成对抗网络的跨媒体信息检索算法

聂为之[1)] 王 岩[1)] 杨 嵩[1)] 刘安安[1)] 张勇东[2)]

[1)](天津大学电气自动化与信息工程学院 天津 300072)

[2)](中国科学技术大学信息科学技术学院 合肥 230026)

摘 要 随着近年来智能终端设备和多媒体社交网络平台的飞速发展,平台用户海量增加的同时,多媒体数据同样呈现海量增长的趋势.使当今主流的社交网络平台充斥着海量的文本、图像等多模态媒体数据,有效的信息检索和分析可以大大提高平台多模态数据的利用率及用户的使用体验,而不同模态间存在显著的语义鸿沟,大大制约了海量多模态数据的分析及有效信息挖局,因此,如何在海量的多模态数据中实现跨模态信息的精准检索就成为当今学术界和工业界面临的重要挑战.本文提出了一种基于循环生成对抗网络的跨媒体信息检索算法.方法基于对抗网络模型框架,通过约束条件的设计,实现了跨模态数据表征的一致性和信息的完整性.首先,该方法构建生成模型实现了文本和图像模态间的互相转换,并基于对抗学习理论,实现跨模态数据在独立空间下语义的一致性约束,保证跨模态数据信息表征的完整性;其次,为了进一步缩小跨模态数据的语义鸿沟,提出了循环交叉熵损失函数,增强跨模态数据在独立空间下表征的一致性,进一步确保信息表征的完整性;最后,通过多模异构数据共嵌特征空间构造引导跨模态数据在共嵌空间下的一致性表征,消除跨模态数据的语义鸿沟,实现跨媒体数据的精准检索.本文针对算法的优势在公开数据库 Flickr30k 和 MSCOCO 上的对比实验和相关实验分析,相关的实验结果也证明了本文所提算法的优越性和合理性.

关键词 跨模态检索;图文检索;生成对抗网络

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.01529

Cross-Media Information Retrieval Based on Cycle Generative Adversarial Networks

NIE Wei-Zhi[1)] WANG Yan[1)] YANG Song[1)] LIU An-An[1)] ZHANG Yong-Dong[2)]

[1)](*School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072)

[2)](*School of Information Science and Technology, University of Science and Technology of China, Hefei* 230026)

Abstract In recent years, with the development of smart devices and social networks, the number of users in these social networks is becoming more and more. The multi-media data is also showing a trend of massive growth. Thus, these current social platforms store large scales of multi-modal data, such as text, video, and images. Effective information retrieval and analysis can greatly improve the utilization of multi-modal data and user experience. However, there exists a significant semantic gap between the two different modalities. This condition greatly restricts the analysis of massive multi-modal data and the mining of effective information, which seriously reduces the user experience. Thus, How to handle accurate cross-modal information retrieval based on the massive multi-modal data has become an important challenge in both academia and industry. Many approaches have been proposed to handle the cross-modal information retrieval problem. However, there is an innate semantic gap between cross-modal data, which leads that many current approaches pay more attention to the consistency of the cross-modal data representation

收稿日期:2020-12-10;在线发布日期:2021-09-22.本课题得到国家自然科学基金项目(U21B2024,61525206,61872267)、天津市新一代人工智能重大专项(19ZXZNGX00110,18ZXZNGX00150)、天津市青年基金(19JCQNJC00500)资助.**聂为之**,博士,副教授,研究方向为计算机视觉、机器学习等.E-mail:weizhinie@tju.edu.cn.**王 岩**,博士研究生,研究方向为跨媒体内容分析和检索、机器学习.**杨 嵩**,博士研究生,研究方向为跨媒体内容分析和检索、机器学习.**刘安安**(通信作者),博士,教授,研究领域为计算机视觉、机器学习等.E-mail:anan0422@gmail.com.**张勇东**,博士,教授,研究领域为多媒体内容分析和理解、多媒体内容安全、视频编码和流媒体技术等.

and ignore the completeness of information representation. The lack of completeness leads to a decrease in the accuracy of cross-modal information retrieval. In this paper, we propose a cross-media retrieval method based on the cycle generative adversarial networks, which design the effective loss functions to guarantee the consistency and completeness of cross-modal data representation based on the generative adversarial network framework. First, we design the generative model to handle the transformation between the text and image data. Moreover, it implements the semantic consistency representation of cross-modal data in the independent feature space based on the adversarial networks. This design is inspired by the GAN, which can guarantee consistency in semantic space. We hope this consistency can guide feature learning in the co-embedding feature space. Second, we propose the cycle cross-entropy loss function to further narrow the cross-modal semantic gap, which can further benefit enhancing the feature-level consistency constraint and representation completeness in the independent feature space. This design can be used to assist the adversarial network to further pull in the consistency of the representation of cross-modal data in independent space, and guarantee the completeness of final cross-modal data features in the co-embedding feature space. Finally, we construct the co-embedding feature space of multi-modal heterogeneous data to guide the consistent representation learning of cross-modal data. This method can effectively bridge the semantic gap between multiple modalities to achieve more accurate cross-media retrieval. The consistency of cross-modal features in independent space will guide the feature learning in the co-embedding feature space to guarantee the completeness of the final cross-modal data representation. The popular Flickr30k and MSCOCO datasets are utilized for the evaluation and analysis. Some classic cross-modal information retrieval methods have been selected as comparison methods. We make the corresponding experiments and analyses according to the completeness and consistency of the cross-modal data representation respectively. We also make the qualitative experimental analysis to discuss the advantages and disadvantages of our approach. In general, the final comparison experiments demonstrate the superiority and effectiveness of the proposed method.

Keywords cross-modal retrieval; image-text retrieval; generative adversarial network

1 引 言

近年来,随着移动互联网和智能终端设备的发展和普及,多媒体数据,特别是文本和图像数据,呈现海量增长的趋势.传统的基于文本信息的检索方式,已经无法满足多样化的用户需求和网络信息的多媒体特性.如何通过跨媒体信息的检索,实现海量多媒体信息的精准匹配,成为诸多社交网络平台和相关科研机构面临的首要问题.多模态数据之间天然的语义鸿沟为实现有效跨模态信息检索带来了巨大的挑战[1].其中,图文检索是当前跨模态检索领域的研究热点.图文检索目的是根据一段文本描述,检索出带有相同语义信息的图像数据,或者根据一幅图像,检索出带有相同语义信息的文本描述[2-3].图文检索的核心科学问题是如何准确地发现图像和文本之间潜在的语义对应关系,并计算图像和文本的语义相似度[4-5],进而实现跨模态的精准检索[6-8].

当前主流图文检索方法是利用多模态数据语义信息的一致性来指导多模态信息在共嵌空间的特征学习,进而将两种模态的向量化表征映射至共同特征空间[9-10],通过在共嵌空间中对多模态数据相似度的有效度量来解决跨模态的检索问题.典型相关分析(Canonical Correlation Analysis,CCA)[11]是早期代表性方法之一,通过最大化两组不同模态数据之间的成对相关性来学习共嵌空间下的向量化表征.因为多模态数据相关性复杂多样,仅通过线性投影无法实现跨模态数据在共嵌空间下的一致性表征,所以研究人员提出了一系列基于核函数的方法[12-13],但这些方法本质上仍然没有脱离相关分析的范式.因此,在面对跨模态数据复杂语义信息的情

况下，基于核函数的跨模态信息映射能力仍然有限. 近年来，随着深度神经网络理论的发展，大量深度学习方法被用来解决跨模态信息的表征. Jiang 等人[14]利用深度卷积网络，并使用标签信息学习不同模态数据在共嵌空间下的表征. Wang 等人[15]提出了一种基于深度卷积神经网络和神经语言模型的多模态深度卷积神经网络，以利用语义相关性分别学习图像和文本在共嵌空间下的语义表征. Wang 等人[16]提出将跨模态的分类信息用于学习模态类内特征和语义空间特征映射. Guo 等人[17]提出了新的交叉型生成对抗跨模态网络来学习不同模态在共嵌空间下的向量化表征. 以上方法多专注于构建跨模态数据在共嵌空间下的特征学习网络，实现跨模态数据在共嵌空间下表征的一致性，而忽略了不同模态数据在共嵌空间下信息表征的完整性.

针对上述问题，本文提出了一种基于循环生成对抗网络的跨模态信息检索算法. 如图 1 所示，模型包含两个紧密相关的生成对抗网络，分别用来处理文本到图像、图像到文本的模态映射. 该模型引入了对抗网络判决机制，实现了跨模态数据在独立空间下语义的一致性约束. 同时，提出了循环交叉熵损失函数，实现跨模态数据在独立空间下表征的一致性，进而通过图像和文本在独立空间下表征的一致性来引导跨模态数据在共嵌空间下的特征学习，确保信息表征的完整性. 此外，通过构建相关损失函数，对多模态数据在共嵌空间下的表征进行一致性约束，进一步消除跨模态数据的语义鸿沟，实现跨模态数据的精准检索，方法的最终目的是最大限度地发现跨媒体数据之间的共性信息，并利用响应的约束条件是实现语义标准的一致性，进而解决跨媒体的检索问题. 本文在公开的数据集(Flickr30k 和 MSCOCO)上评估了模型性能，并与图文检索代表性的方法进行了比较分析，相关实验结果证明了本文提出模型的优越性.

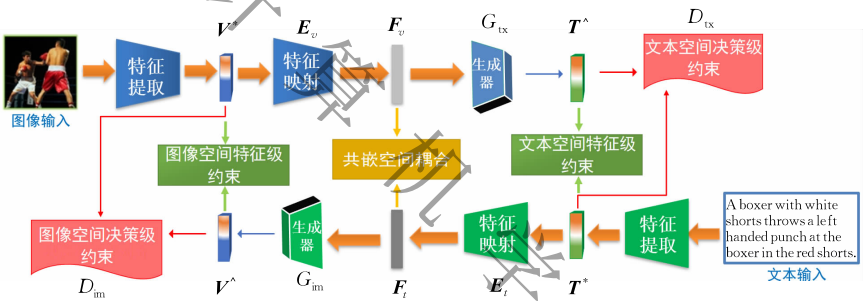

图 1　循环生成对抗网络模型

2　相关工作

本节针对当前代表性跨模态信息检索算法进行简介. 目前主流方法是基于图像和文本数据的语义信息，将图像和文本特征映射到共嵌空间后计算图像和文本特征之间的相似性，进而解决跨模态的信息检索问题. Faghri 等人[18]通过引入最大负样本对(The Hardest Negative Pairs)来计算信息损失，改善了排序误差，进而实现跨模态的信息检索. Zheng 等人[19]探索了利用文本的卷积神经网络和实例损失(Instance Loss)，学习文本和图像更具辨识度的语义表征. Zhang 等人[20]使用投影分类损失对表征向量从一种模态到另一种模态的投影特征进行分类，并利用改进的 Norm Softmax 损失来学习跨模态数据的一致性. Niu 等人[21]利用经典的递归卷积神经网络学习视觉和文本在共嵌空间下的语义表征. Huang 等人[22]提出了一种基于文本序列化信息来进行表征学习的语义模型，通过文本元素信息与视觉信息的匹配，实现图像和文本信息在共嵌空间下表征的一致性. 此外，一些论文使用无监督、哈希编码等方法[23-27]来解决图文检索问题，其中 Peng 等人[23,25]使用了 RNN-CNN 卷积网络和无监督网络对特征进行映射从而学习其相关性，Zhan 等人[24]和 Ye 等人[27]使用哈希编码来表征图文特征从而学习图文特征的相关性来解决跨模态提取问题.

近年来，生成对抗网络受到越来越多的关注，被广泛用于解决跨模态信息检索问题[28]. Sarafianos 等人[29]提出了文本和图像模态对抗匹配算法(Text-Image Modality Adversarial Matching，TIMAM)，采用对抗表示学习(Adversarial Representing Learning，ARL)来约束跨模态语义的一致性，以实现更有效的图文数据在共嵌空间下的表征. ARL 框架在共嵌空间中优化了一个两层全连通网络对抗判别器，从而提升跨模态检索性能. TIMAM 还在 LSTM 分支

前增加了来自 Transformer[30] 的双向编码器表示，以优化文本特征编码. Liu 等人[31] 提出了一种新的对抗图注意力机制卷积网络模型(Adversarial Graph Attention Network，AGANet)，从设计的图中学习高级结构的语义视觉特征，特定的联合嵌入层通过对抗性学习模块连接图像和文本特征. 这些方法一定程度上解决了跨模态检索的问题，并在部分公开数据库上取得了较好的检索结果，但这些方法都过度关注于图像和文本特征在共嵌空间上表征的一致性问题，忽略了信息表征的完整性及图像文本原始特征对共嵌特征学习过程的约束作用.

3 循环生成对抗网络模型

本文提出了一种基于循环生成对抗网络模型的跨模态信息检索算法(如图 1 所示). 该模型采用了经典的循环对抗网络架构，包括两个主要模块：(1) 文本生成对抗网络. 通过对文本的预处理得到文本数据的原始特征 $\boldsymbol{T}^*$，利用两个生成器来分别得到共嵌空间特征 $\boldsymbol{F}_t$ 和基于文本信息生成的图像特征 $\boldsymbol{V}^\wedge$；(2) 图像生成对抗网络. 利用卷积网络提取图像数据的原始特征 $\boldsymbol{V}^*$，利用两个生成器分别得到共嵌空间特征 $\boldsymbol{F}_v$ 和基于图像信息生成的文本特征 $\boldsymbol{T}^\wedge$. 基于此架构，创新性的提出了三种特征学习约束机制：(1) 共嵌空间一致性约束. 实现跨模态数据在共嵌空间下表征($\boldsymbol{F}_t$ 和 $\boldsymbol{F}_v$)的一致性；(2) 决策级一致性约束. 利用对抗网络的判决机制，实现对跨模态生成数据($\boldsymbol{V}^\wedge$ 和 $\boldsymbol{T}^\wedge$)与原始模态数据($\boldsymbol{V}^*$ 和 $\boldsymbol{T}^*$)的独立空间下的语义一致性约束；(3) 特征级一致性约束. 利用跨模态数据特征的相似性度量来进一步约束跨模态生成数据($\boldsymbol{V}^\wedge$ 和 $\boldsymbol{T}^\wedge$)与原始模态数据($\boldsymbol{V}^*$ 和 $\boldsymbol{T}^*$)在独立空间下表征的一致性. 下面对其进行详细介绍.

3.1 数据预处理

该网络模型输入分别为文本信息和图像信息，提取句子中 N 个单词的特征，并使用单词级特征作为文本特征 $\boldsymbol{T}=[\boldsymbol{t}_1,\cdots,\boldsymbol{t}_N]\in R^{d\times N}$，同时提取图像中 R 个区域的区域级特征作为视觉特征 $\boldsymbol{V}=[\boldsymbol{v}_1,\cdots,\boldsymbol{v}_R]\in R^{d\times R}$. 区域信息的提取有利于描述视觉的局部信息，保证视觉信息的完整性. 然后，采用注意力机制分别实现视觉区域特征和文本特征的融合. 具体来说，当给定初始图像的区域特征 $\boldsymbol{V}$ 和文本局部特征 $\boldsymbol{T}$，通过线性映射和分类归一化计算注意力权重，如式(1)所示，用求得的权重来融合区域特征，进而得到视觉和文本的融合特征表示.

$$\boldsymbol{a}_v=\mathrm{Softmax}\left(\frac{\boldsymbol{W}_v\boldsymbol{V}}{\sqrt{d}}\right)^{\mathrm{T}},\ \boldsymbol{V}^*=\boldsymbol{V}\boldsymbol{a}_v,$$

$$\boldsymbol{a}_t=\mathrm{Softmax}\left(\frac{\boldsymbol{W}_t\boldsymbol{T}}{\sqrt{d}}\right)^{\mathrm{T}},\ \boldsymbol{V}^*=\boldsymbol{V}\boldsymbol{a}_t \tag{1}$$

其中 $\boldsymbol{W}_v,\boldsymbol{W}_t\in R^{1\times d}$ 代表了可学习的线性映射参数，$\boldsymbol{a}_v\in R^R$ 代表了 R 个区域可视化特征的注意力权重，$\boldsymbol{a}_t\in R^N$ 代表了 N 个单词文本特征的注意力权重，$\boldsymbol{V}^*\in R^d$ 代表了 $\boldsymbol{V}$ 的融合特征，$\boldsymbol{T}^*\in R^d$ 代表了 $\boldsymbol{T}$ 的融合特征. 注意力机制的运用，可以有效去除图像和文本的冗余信息，提高信息表征的鲁棒性. 这里需要注意的是，对于文本和图像数据，本文采用预训练的模式，与整体模型分开进行训练，这样可以保证文本和图像数据表征($\boldsymbol{V}^*$ 和 $\boldsymbol{T}^*$)的有效性和完整性，有利于指导跨模态数据在共嵌空间下的特征学习.

3.2 模型构建

本模型首先使用两个映射函数 $\boldsymbol{E}_t$ 和 $\boldsymbol{E}_v$ 的学习使图像和文本信息可以映射到同一特征空间，从而实现图像和文本数据匹配. 其中 $\boldsymbol{E}_t:\boldsymbol{T}^*\rightarrow\boldsymbol{F}_t$，$\boldsymbol{E}_v:\boldsymbol{V}^*\rightarrow\boldsymbol{F}_v$. $\boldsymbol{F}_v$ 为共嵌空间下的图像特征，$\boldsymbol{F}_t$ 为共嵌空间下的文本像特征. 然后，引入了生成器 G_{im} 和 G_{tx}，以保证共嵌空间下图像和文本数据表征的有效性和完整性. G_{im} 实现文本共嵌特征到图像特征的转换，G_{tx} 实现图像共嵌特征到文本特征的转换. 通过交叉监督来确保转换信息的一致性，引导跨模态数据在共嵌空间下特征学习的完整性. 最后，为了保证 $\boldsymbol{E}_t$ 和 $\boldsymbol{E}_v$ 能生成相近的特征并使 G_{im} 和 G_{tx} 可以生成相对完整的模态信息，本模型在共嵌空间和原始特征空间分别提出了三种约束机制，以下将进行逐一介绍.

3.2.1 共嵌空间一致性约束

为了保证特征 $\boldsymbol{F}_v$ 和 $\boldsymbol{F}_t$ 在共嵌空间中的距离分别小于 $\boldsymbol{F}_v$ 和文本负样本 $\boldsymbol{F}_t'$ 以及 $\boldsymbol{F}_t$ 和图像负样本 $\boldsymbol{F}_v'$ 之间的距离，消除两个特征 $\boldsymbol{F}_v$ 和 $\boldsymbol{F}_t$ 间的差异，通过引入三元组损失函数(Triple Loss)对其一致性进行约束：

$$\ell_{\mathrm{triple}}(\boldsymbol{F}_v,\boldsymbol{F}_t)=\sum_{\boldsymbol{F}_t'}[\sigma-s(\boldsymbol{F}_v,\boldsymbol{F}_t)+s(\boldsymbol{F}_v,\boldsymbol{F}_t')]_+ +\sum_{\boldsymbol{F}_v'}[\sigma-s(\boldsymbol{F}_v,\boldsymbol{F}_t)+s(\boldsymbol{F}_v',\boldsymbol{F}_t)]_+ \tag{2}$$

其中 σ 为边界阈值；$s(\cdot)$ 为特征相似度度量函数，本文采用欧氏距离来计算两两特征的距离，计算公式如下：

$$s(\boldsymbol{F}_v,\boldsymbol{F}_t)=1/\|\boldsymbol{F}_v-\boldsymbol{F}_t\|_2 \tag{3}$$

但仅仅使用共嵌空间中跨模态特征相似度损失，易

导致特征映射过程中各模态特性信息的损失. 为了保证信息的完整性，本文引入 G_{im} 和 G_{tx} 两个生成模型，在独立空间下对跨模态数据的特征学习进行进一步的约束.

3.2.2　决策级一致性约束

为了进一步实现跨模态数据在独立空间下语义表征的一致性，设计了判别器 D_{im} 和 D_{tx}，其中 D_{tx} 用来约束文本特征 $\boldsymbol{T}^*$ 和生成文本特征 $\boldsymbol{T}^\wedge$ 语义的一致性，D_{im} 用来约束图像特征 $\boldsymbol{V}^*$ 和生成图像特征 $\boldsymbol{V}^\wedge$ 语义的一致性. 具体的损失函数如下所示：

$$\ell_{\text{gan}}^{v}(G_{\text{im}},D_{\text{im}},\boldsymbol{V}^*,\boldsymbol{F}_t)=$$
$$E[\log D_{\text{im}}(\boldsymbol{V}^*)]+E[\log(1-D_{\text{im}}(G_{\text{im}}(\boldsymbol{F}_t)))] \quad (4)$$

其中 G_{im} 所生成的图像特征 $\boldsymbol{V}^\wedge$ 应该近似于初始图像特征 $\boldsymbol{V}^*$，因为二者有相似的语义信息. 在这个损失函数中，训练生成器 G_{im} 的最小化损失函数，训练判别器 D_{im} 的最大化损失函数，即 $\min_{G_{\text{im}}}\max_{D_{\text{im}}}\ell_{\text{gan}}^{v}(G_{\text{im}},D_{\text{im}},\boldsymbol{V}^*,\boldsymbol{F}_t)$. 对于生成器 G_{tx} 以及判别器 D_{tx}，损失函数如下所示：

$$\ell_{\text{gan}}^{t}(G_{\text{tx}},D_{\text{tx}},\boldsymbol{T}^*,\boldsymbol{F}_v)=$$
$$E[\log D_{\text{tx}}(\boldsymbol{T}^*)]+E[\log(1-D_{\text{tx}}(G_{\text{tx}}(\boldsymbol{F}_v)))] \quad (5)$$

其中 G_{tx} 所生成的文本特征 $\boldsymbol{T}^\wedge$ 应该近似于初始文本表示 $\boldsymbol{T}^*$，原因与前者一致. 训练生成器 G_{tx} 的最小化损失函数，训练判别器 D_{tx} 的最大化损失函数，即 $\min_{G_{\text{tx}}}\max_{D_{\text{tx}}}\ell_{\text{gan}}^{t}(G_{\text{tx}},D_{\text{tx}},\boldsymbol{T}^*,\boldsymbol{F}_v)$.

3.2.3　特征级完整性约束

为了进一步保证跨模态数据表征在独立空间的一致性，设计了特征级完整性约束. 该约束实现跨模态数据在独立空间下原始特征表征的完整性，进而通过图像和文本在独立空间下表征的完整性来引导跨模态数据在共嵌空间下的特征学习，保证信息表征的鲁棒性和完整性. 当图像特征被 G_{tx} 生成模型转换到文本模态空间 $\boldsymbol{T}^\wedge$，应与原始输入的文本特征保持完整性，确保语义信息完整. 同理，当文本特征被 G_{im} 生成模型转换到视觉模态空间 $\boldsymbol{V}^\wedge$，应该与原始输入的图像特征保持完整性. 特征映射过程可以表示为：(1) $\boldsymbol{E}_v:\boldsymbol{V}^*\rightarrow\boldsymbol{F}_v$ 和 $G_{\text{tx}}:\boldsymbol{F}_v\rightarrow\boldsymbol{T}^\wedge$；(2) $\boldsymbol{E}_t:\boldsymbol{T}^*\rightarrow\boldsymbol{F}_t$ 和 $G_{\text{im}}:\boldsymbol{F}_t\rightarrow\boldsymbol{V}^\wedge$. 因为 $\boldsymbol{T}^*$ 和 $\boldsymbol{V}^*$ 有相同的语义信息，因此 $\boldsymbol{T}^\wedge$ 和 $\boldsymbol{V}^\wedge$ 在特征上与前者具有一致性. 损失函数如下：

$$\ell_x^t(\boldsymbol{V}^*,\boldsymbol{T}^*)=E[\|G_{\text{tx}}(\boldsymbol{E}_v(\boldsymbol{V}^*))-\boldsymbol{T}^*\|_2],$$
$$\ell_x^v(\boldsymbol{V}^*,\boldsymbol{T}^*)=E[\|G_{\text{im}}(\boldsymbol{E}_t(\boldsymbol{T}^*))-\boldsymbol{V}^*\|_2] \quad (6)$$

综合上述分析，本文所提出循环交叉对抗网络模型的损失函数为

$$\ell_{\text{all}}=\ell_{\text{triple}}+\alpha(\ell_{\text{gan}}^{v}+\ell_{\text{gan}}^{t})+\beta(\ell_x^v+\ell_x^t) \quad (7)$$

其中，参数 α 和 β 为可训练的超参数，通过统计实验，将参数值 α 和 β 分别设定为 0.6 和 0.8. 训练两个生成器和两个映射函数使用自适应矩估计优化器[32]，训练两个判别器使用随机梯度下降优化器[33]. 具体训练方法如算法 1 所示.

算法 1.　模型训练方法.

输入：输入文本和图像的原始特征 $\boldsymbol{T}^*$ 和 $\boldsymbol{V}^*$

输出：生成器：G_{im} 和 G_{tx}，判决器：D_{im} 和 D_{tx}，特征映射矩阵 $\boldsymbol{E}_v$，$\boldsymbol{E}_t$

1. 首先，初始化 $\boldsymbol{E}_v$，$\boldsymbol{E}_t$，G_{im} 和 G_{tx}，D_{im} 和 D_{tx} 的参数；
2. 固定 $\boldsymbol{E}_v$，$\boldsymbol{E}_t$ 和 D_{im}，D_{tx} 的模型参数，基于式(4)～(6)分别优化生成器 G_{im} 和 G_{tx} 的模型参数；
3. 固定 $\boldsymbol{E}_v$，$\boldsymbol{E}_t$ 和 G_{im}，G_{tx} 的模型参数，基于式(4)～(6)分别优化判决器 D_{im} 和 D_{tx} 的模型参数；
4. 固定 G_{im}，G_{tx}，D_{im} 和 D_{tx} 模型参数，基于式(2)优化 $\boldsymbol{E}_v$ 和 $\boldsymbol{E}_t$ 参数；
5. 重复 2～4 的步骤，优化式(7)；
6. 返回 $\boldsymbol{E}_v$，$\boldsymbol{E}_t$，G_{im} 和 G_{tx}，D_{im} 和 D_{tx}.

4　实　验

4.1　数据集

本文采用经典的图文检索数据集 Flickr30k[34] 和 MSCOCO[35] 对模型性能进行评测. Flick30k 数据集包含了 31 783 张图像，其中每张图像有 5 个人工生成的描述文本. 参照经典的数据集划分设置[18]，采用 1000 个图像进行交叉验证，另 1000 个图像进行测试.

MSCOCO 数据集包含了 123 287 张图像，每幅图像有 5 个人工生成的描述文本. 采用经典的 Karthy split[36] 划分方法，包含 113287 个训练图像，5000 个交叉验证图像和 5000 个测试图像.

4.2　实验设置

参照流行方法[37]，使用 Faster R-CNN[38] 来提取图像的 36 个区域特征. 通过平均池化层来计算每个图像区域的特征. 对于包含 N 个单词的文本，首先将每个单词嵌入为 300 维特征向量，然后用单层双向 GRU[39] 来处理整个句子，每个单词的特征表示为前向 GRU 和后向 GRU 中隐藏状态的平均值. 设置 G_{tx} 和 $\boldsymbol{E}_t$ 输入 1024 维特征向量，输出 512 维特征；同理，设置 G_{im} 和 $\boldsymbol{E}_v$ 输入 1024 维特征向量，输出 512 维特征向量. 本文使用随机梯度下降法对模型

进行优化，训练轮次一共 30 轮，训练过程中生成器和判别器的优化中设置前 15 个训练轮次学习率为 0.0002，后 15 个训练轮次学习率为 0.00002. 模型在第 25 轮训练后损失收敛，模型参数值 α 和 β 分别设定为 0.6 和 0.8.

4.3 评测指标

对于图文检索任务，采用 k 点召回率($R@k$)进行定量评测. k 点召回率表示在前 k 个检索结果中正确匹配的比例. 本文主要采用了第 1 名召回率($R@1$)、前 5 名召回率($R@5$)和前 10 名召回率($R@10$).

4.4 对比实验

表 1 和表 2 分别给出了所提出方法在 MSCOCO 和 Flickr30k 数据集上的性能对比实验. CMPM、JGCAR、RRF-NET 和 CSE 方法均属于利用共嵌空间特征一致性解决图文检索问题，和本文相比它们都忽略了信息表征的完整性和独立空间下特征的一致性. 在 MSCOCO 数据集中，该方法相对于共嵌特征空间学习最优方法 RRF-NET 具有更优性能，图搜文任务在 $R@1$ 上提升 10.3%，在 $R@5$ 上提升 5.5%，在 $R@10$ 上提升 2.7%. 在文搜图任务中，相比于共嵌特征空间学习最优的 CSE 方法，本文模型在 $R@1$ 上提升 6.9%，在 $R@5$ 上提升 2.5%，在 $R@10$ 上提升 0.1%. 在 Flickr30k 上，相比于共嵌特征空间最优算法 CMPM，本文模型在图搜文任务中平均提升 11.2%，在文搜图任务中平均提升 8.7%. 而相比于同样采用对抗学习思想的方法 TIMAM，该方法在图搜文任务上 $R@1$ 提升 10.7%，$R@5$ 提升 9.8%，$R@10$ 提升 6.1%；在文搜图任务上 $R@1$ 提升 4%，$R@5$ 提升 3.6%，$R@10$ 提升 0.8%. 对比实验结果表明，所提出方法能够有效提升图像和文本表征的一致性，因此所提出方法较当前两类代表性方法具有更优性能.

表 1 MSCOCO 数据集上的对比实验

类别	方法	MSCOCO 数据集					
		图搜文			文搜图		
		$R@1$/%	$R@5$/%	$R@10$/%	$R@1$/%	$R@5$/%	$R@10$/%
共嵌空间特征学习	JGCAR[40]	52.7	82.6	90.5	40.2	74.8	85.7
	CMPM[41]	56.1	86.3	92.9	44.6	78.8	89.0
	CSE[42]	56.3	84.4	92.2	45.7	81.2	90.6
	RRF-NET[43]	56.4	85.3	91.5	43.9	78.1	88.6
其他模型	PFAN[44]	51.2	84.3	89.2	41.4	70.9	79.0
循环生成对抗	本文方法	**66.7**	**90.8**	**94.2**	**52.6**	**83.7**	**90.7**

表 2 Flickr30k 数据集上的对比实验

类别	方法	Flickr30k 数据集					
		图搜文			文搜图		
		$R@1$/%	$R@5$/%	$R@10$/%	$R@1$/%	$R@5$/%	$R@10$/%
共嵌空间特征学习	CSE[42]	44.6	74.3	83.8	36.9	69.1	79.6
	JGCAR[40]	44.9	75.3	82.7	35.2	62.0	72.4
	RRF-NET[43]	47.6	77.4	87.1	35.4	68.3	79.9
	CMPM[41]	49.6	76.8	86.1	37.3	65.7	75.5
生成对抗网络	TIMAM[29]	53.1	78.8	87.6	42.6	71.6	81.9
其他模型	VRA[45]	55.1	81.8	88.5	41.7	71.5	80.4
	SMAN[46]	57.3	85.3	92.2	43.4	73.7	**83.4**
循环生成对抗	本文方法	**63.8**	**88.6**	**93.7**	**46.6**	**75.2**	82.7

表 3 给出了本文方法和 X-GACMN 模型在 Flickr30k 数据集和 NUS-WIDE-10k 数据集上的性能对比结果. 由于验证 X-GACMN 模型性能使用的数据库和本文的数据库不同，其使用的数据库不包含标签信息，所以本文从两个方面和 X-GACMN 模型进行性能比较. 对于 Flickr30k 数据库，本文所提的方法相对于误分类损失的 X-GACMN 模型具有更优性能，图搜文任务在 $R@1$ 上提升 10.4%，在 $R@5$ 上提升 1.3%，在 $R@10$ 上提升 1.1%；文搜图任务中，该方法在 $R@1$ 上提升 7.9%，在 $R@5$ 上提升 4.6%，在 $R@10$ 上提升 5.2%. 对于包含标签数据的 NUS-WIDE-10k 数据集，该方法引入了分类损失函数，从表 3 对比可知，本文加入分类损的模型相对于 X-GACMN 模型具有更优性能，图搜文任务在 mAP 上提升 2.2%，文搜图任务在 mAP 上提升 2.7%.

表 3　与 X-GACMN 方法对比实验

Flickr30k 数据集							
类别	方法	图搜文			文搜图		
		$R@1/\%$	$R@5/\%$	$R@10/\%$	$R@1/\%$	$R@5/\%$	$R@10/\%$
生成对抗网络	X-GACMN(无分类损失)[17]	53.4	87.3	92.6	38.7	70.6	77.5
循环生成对抗	本文方法	**63.8**	**88.6**	**93.7**	**46.6**	**75.2**	**82.7**

NUS-WIDE-10k 数据集			
类别	方法	图搜文	文搜图
		mAP/%	mAP/%
循环生成对抗	X-GACMN[17]	50.1	52.6
生成对抗网络	本文方法(加入分类损失)	**52.3**	**55.3**

实验结果表明，虽然分类损失可以更好地指导跨模态数据的表征的一致性学习，但是对于不带标签的数据本文所提出方法相对于 X-GACMN 模型可以更好地引导跨模态数据在共嵌空间表征的一致性，同时保证了共嵌空间特征的完整性，进而提升跨模态信息检索的准确率.

4.5　网络结构分析

基于多模态数据的特点和模态相关性，本文所提出方法的目标函数包含三个重要的损失函数项，通过对不同损失函数项对应网络结构的调整，分析了不同损失函数组合对模型性能的影响. 表 4 给出了本文模型在 Flickr30k 数据集的实验结果. 为了说明循环生成对抗网络设计的有效性，对三类约束机制进行了相应实验分析，同时对其三者的不同组合也进行了实验，从而对比分析各项的作用.

表 4　消融实验结果

Flickr30k 数据集							
方法		图搜文			文搜图		
		$R@1/\%$	$R@5/\%$	$R@10/\%$	$R@1/\%$	$R@5/\%$	$R@10/\%$
共嵌空间损失	ℓ_{triple}	54.1	81.8	89.0	39.4	69.2	79.1
	ℓ_{gan}	52.9	81.6	87.2	39.6	69.3	78.5
决策级损失	$\ell_{\text{triple}}+\ell_{\text{gan}}$	58.3	82.1	89.4	40.8	70.3	81.7
	ℓ_x	53.1	81.7	87.7	39.8	69.0	79.3
特征级损失	$\ell_{\text{triple}}+\ell_x$	57.8	81.5	88.4	40.3	69.9	81.2
	$\ell_{\text{gan}}+\ell_x$	57.9	80.4	88.6	39.9	69.7	80.6
	$\ell_{\text{triple}}+\ell_{\text{gan}}+\ell_x$	63.8	88.6	93.7	46.6	75.2	82.7

由表 4 可知，采用单独的损失函数得到的评价指标值要远远低于多损失函数联合优化的性能. 其中，单独采用 L_{sim} 的检索结果性能最差. 通过约束机制组合的实验结果观察，加入决策级损失后，图搜文任务中 $R@1$ 提升 4.2%，文搜图任务中 $R@1$ 提升 1.4%；加入特征级损失后，图搜文任务 $R@1$ 提升 3.7%，文搜图任务 $R@1$ 提升 0.9%；两者都加入时，图搜文任务 $R@1$ 提升 9.7%，文搜图任务 $R@1$ 提升了 7.2%. 该结果表明，决策级约束和特征及约束在跨模态数据表征一致性学习上起到积极作用，在共嵌空间得到了更优的数据表征.

4.6　定性对比分析

图 2 和图 3 展示了本文所提出方法在图文检索任务上的定性对比实验结果. 如图 2 所示，对于图搜文任务，通过和 TIMAM 对比，所提出方法在前三检索结果命中 2 个，且排序更高，直观地证明了该方法的优越性. 但仔细分析错误的检索结果发现，这些结果确实有一些有意义的描述，只是一些关键信息错误而已. 该案例表明，挖掘并融合高排序检索结果的关键信息有助于模型优化.

如图 3 所示，对于文搜图任务，特意采用短文本描述进行检索，增加检索难度. 显然，所提出方法在该挑战性条件下性能依然优于代表性的 TIMAM 方法. 但是，本文方法的效果仍具有改进空间，仔细观察反馈图像，其实这些图均与输入的待检索文本描述存在不同程度的相关性. 比如第一幅图，非常像一个人坐在地铁，只有仔细分辨，才知道这人坐在车厢外. 这个无法严格界定成功还是失败的案例说明，对于这个基于主观标注的评测，不能单纯地依赖量化性能的提升，还需定性分析检索结果，来探索更优的评价机制.

图 2 图搜文任务举例

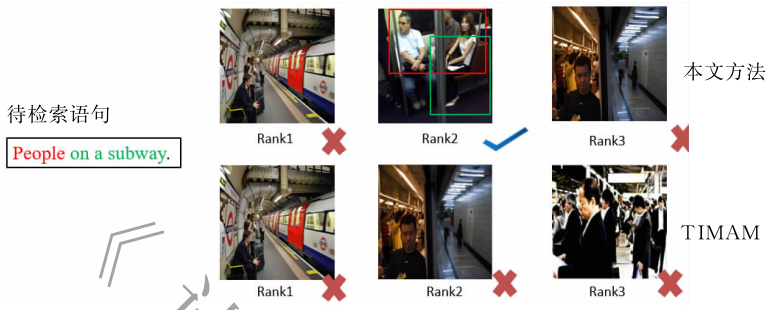

图 3 文搜图任务举例

5 总 结

本文提出了一种基于循环生成对抗网络的跨媒体信息检索方法. 该方法构建了创新的循环对抗生成网络，实现图像和文本数据模态的交互映射，并提出了共嵌空间一致性约束、原始模态特征空间下的决策级约束和特征级约束，增强了图文数据在共嵌空间表征的完整性，降低不同模态间的异构差异，从而提高图文检索准确性. 通过在两个流行数据集的评测，证明所提出方法在图文检索任务中的优越性和网络结构合理性.

参 考 文 献

[1] Karpathy A, Joulin A, Fei-Fei L F. Deep fragment embeddings for bidirectional image sentence mapping//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, 2014: 1889-1897

[2] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN)//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2014: 1-8

[3] Zheng F, Tang Y, Shao L. Hetero-manifold regularisation for cross-modal hashing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 40(5): 1059-1071

[4] Hu M, Yang Y, Shen F, et al. Hashing with angular reconstructive embeddings. IEEE Transactions on Image Processing, 2017, 27(2): 545-555

[5] Hendricks L A, Venugopalan S, Rohrbach M, et al. Deep compositional captioning: Describing novel object categories without paired training data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1-10

[6] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2015: 2048-2057

[7] Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2015: 2425-2433

[8] Kazemi V, Elqursh A. Show, ask, attend, and answer: A strong baseline for visual question answering//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2017: 1704-1711

[9] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1-12

[10] Wang L, Li Y, Huang J, et al. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 394-407

[11] Hotelling H. Relations between two sets of variates//Kotz S, Johnson N L, eds. Breakthroughs in Statistics. New York, USA: Springer, 1992: 162-190

[12] Akaho S. A kernel method for canonical correlation analysis//Proceedings of the International Meeting of the Psychometric

Society. Osaka, Japan, 2001: 1-9

[13] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-442

[14] Jiang Q Y, Li W J. Deep cross-modal hashing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2017: 3232-3240

[15] Wang W, Yang X, Ooi B C, et al. Effective deep learning-based multi-modal retrieval. The VLDB Journal, 2016, 25(1): 79-101

[16] Wang B, Yang Y, Xu X, et al. Adversarial cross-modal retrieval//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA, 2017: 154-162

[17] Guo Weikuo, Liang Jian, Kong Xiangwei, et al. X-GACMN: An X-shaped generative adversarial cross-modal network with hypersphere embedding//Proceedings of the Asian Conference on Computer Vision. Cham, Germany, 2018: 513-529

[18] Faghri F, Fleet D J, Kiros J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives//Proceedings of the British Machine Vision Conference. Newcastle, UK, 2017: 1-5

[19] Zheng Z, Zheng L, Garrett M, et al. Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(2): 1-23

[20] Wang Haoran, Zhang Ying, Ji Zhong, et al. Consensus-aware visual-semantic embedding for image-text matching// Proceedings of the European Conference on Computer Vision. Springer, Cham, 2020: 18-34

[21] Niu Z, Zhou M, Wang L, et al. Hierarchical multimodal LSTM for dense visual-semantic embedding//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2017: 1881-1889

[22] 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

[23] Peng Y, Qi J. Reinforced cross-media correlation learning by context-aware bidirectional translation. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1718-1731

[24] Zhan Y W, Luo X, Wang Y, et al. Supervised hierarchical deep hashing for cross-modal retrieval//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 3386-3394

[25] Peng Y, Ye Z, Qi J, et al. Unsupervised visual-textual correlation learning with fine-grained semantic alignment. IEEE Transactions on Cybernetics, 2018, 30(67): 1059-1064

[26] Peng Y, Qi J, Ye Z, et al. Hierarchical visual-textual knowledge distillation for life-long correlation learning. International Journal of Computer Vision, 2021, 129(4): 921-941

[27] Ye Z, Peng Y. Sequential cross-modal hashing learning via multi-scale correlation mining. ACM Transactions on Multimedia Computing, Communications and Applications, 2019, 15(4): 1-20

[28] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Palais des Congrès de Montréal, Canada, 2014: 2672-2680

[29] Sarafianos N, Xu X, Kakadiaris I A. Adversarial representation learning for text-to-image matching//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5814-5824

[30] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1-13

[31] Liu J, Zha Z J, Hong R, et al. Deep adversarial graph attention convolution network for text-based person search// Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 665-673

[32] Sharma M, Pachori R B, Rajendra A U. Adam: A method for stochastic optimization. Pattern Recognition Letters, 2017, 94(3): 172-179

[33] Zou D, Cao Y, Zhou D, et al. Gradient descent optimizes over-parameterized deep ReLU networks. Machine Learning, 2020, 109(3): 467-492

[34] 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(25): 67-78

[35] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Cham, Germany, 2014: 740-755

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

[37] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[38] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[39] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling// Proceedings of the Neural Information Processing Systems. QC, Canada, 2014: 1-12

[40] Wang S, Chen Y, Zhuo J, et al. Joint global and co-attentive representation learning for image-sentence retrieval//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea, 2018: 1398-1406

[41] Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 686-701

[42] You Q, Zhang Z, Luo J. End-to-end convolutional semantic embeddings//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5735-5744

[43] Liu Y, Guo Y, Bakker E M, et al. Learning a recurrent residual fusion network for multimodal matching//Proceedings of the IEEE International Conference on Computer Vision. Columbus, USA, 2017: 4107-4116

[44] Wang Yaxiong, Yang Hao, Bai Xiuxiu, et al. PFAN++: Bi-directional image-text retrieval with position focused attention network. IEEE Transactions on Multimedia, 2020, 34(99): 1-14

[45] Guo Yutian, Chen Jingjing, Zhang Hao, Jiang Yu-Gang. Visual relations augmented cross-modal retrieval//Proceedings of the 2020 International Conference on Multimedia Retrieval. New York, USA, 2020: 9-15

[46] Ji Z, Wang H, Han J, et al. SMAN: Stacked multimodal attention network for cross-modal image-text retrieval. IEEE Transactions on Cybernetics, 2020, 10(99): 1-12

NIE Wei-Zhi, Ph. D., associate professor. His research interests include computer vision and machine learning.

WANG Yan, Ph. D. candidate. His main research interests include cross-media content analysis and retrieval, machine learning.

YANG Song, Ph. D. candidate. Her main research interests include cross-media content analysis and retrieval, machine learning.

LIU An-An, Ph. D., professor. His current research interests include computer vision and machine learning.

ZHANG Yong-Dong, Ph. D., professor. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology.

Background

Text-image cross-modal retrieval is a challenging task in the field of language and vision. Most previous approaches independently embed images and sentences into a joint embedding space and compare their similarities. However, previous approaches rarely explore the interactions between images and sentences before calculating similarities in the joint space. In this paper, based on the consistency of cross-modal data modalities in the cyclic adversarial generation network, an innovative adversarial loss function and cyclic cross-loss function are proposed to effectively utilize the interaction between text and image, to reduce the differences in cross-modal data representation, and ensure the final cross-media information Retrieval accuracy and relevance. For the text-to-image task, the *Rank*@10 value of the most advanced algorithms in COCO and Flickr databases are 90.6 and 81.9 respectively, the *Rank*@10 value of this paper in COCO and Flickr databases are 90.7 and 82.7. For the image-to-text task, the *Rank*@10 value of the most advanced algorithms in COCO and Flickr databases are 92.9 and 87.6, the *Rank*@10 value of this paper in COCO and Flickr databases are 94.2 and 93.7.

The project of this paper is the National Natural Science Foundation of China, the grant of Tianjin New Generation Artificial Intelligence Major Program. With the rapid development of image acquisition equipment and Internet technology in recent years, especially the development of social networking platforms, multimedia data has shown a trend of massive growth. How to achieve accurate retrieval and recommendation of information in massive multimedia data is now several technology companies and urgent problems facing scientific research institutions. This paper proposes an innovative cross-modal circulation adversarial production network model. The algorithm can effectively solve the consistency of cross-media information representation between pictures and text data, effectively improve the flexibility of retrieval methods, and increase the diversity of expression forms of retrieval results and improve the accuracy of search results.