

面向传统服饰的细粒度跨模态检索算法

赵海英¹, 向翔¹, 李婕¹, 张佳伟²

(1.北京邮电大学人工智能学院, 北京 100876;

2. 新疆师范高等专科学校新疆教育云技术与资源实验室, 乌鲁木齐 830043)

摘要: **目的** 由于跨模态数据集有限和模态异构表征问题, 利用跨模态检索算法解决实际问题一直是当前多模态研究中的一大研究方向。**方法** 提出了一种面向传统服饰的细粒度跨模态检索算法, 解决传统服饰跨模态检索的单模态表征和跨模态表征一致的问题。在单模态特征表征方面, 沿用 DCMH 使用深度学习的方法对初始数据进行特征提取; 在跨模态表征一致方面, 新增自监督语义网络, 以自监督的方式对应标签信息提取细粒度信息, 并将其用于图文哈希学习的监督, 从而得到更好的图文哈希表征。通过在传统服饰数据集上与其他方法进行对比实验, 验证算法的有效性。**结论** 有关此方面的应用探索, 有利于解决互联网时代中国传统服饰文本、图像处理等的保护性难题, 为未来纹样检索中的工作做铺垫, 实现中国传统服饰的创新性传承和发展。

关键词: 跨模态检索; 哈希; 细粒度; 自监督

中图分类号: TB472 **文献标识码:** A **文章编号:** 1001-3563(2021)22-0026-07

DOI: 10.19554/j.cnki.1001-3563.2021.22.005

Fine-grained Cross Modal Retrieval Algorithm for Traditional Clothing

ZHAO Hai-ying¹, XIANG Xiang¹, LI Jie¹, ZHANG Jia-wei²

(1.School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2.Xinjiang Laboratory of Education Cloud Technology and Resources, Xinjiang Teacher's College, Urumqi 830043, China)

ABSTRACT: Due to the limitation of cross modal data sets and the heterogeneous representation of modes, it is always a major research direction to solve the practical application problems by using cross modal retrieval algorithm. In this paper, a fine-grained cross modal retrieval algorithm for traditional clothing is proposed to solve the problem of consistency between single-mode representation and cross mode representation of traditional clothing cross modal retrieval. In the aspect of single-mode feature representation, DCMH is used to extract features from the initial data using deep learning method; In the aspect of cross modal representation consistency, a new self-supervised semantic network is added to extract fine-grained information from label information in a self-supervised way, and it is used to supervise the hash learning of images and texts, so as to obtain better hash representation of images and texts. The effectiveness of the proposed algorithm is verified by comparing with other methods on traditional clothing data sets. The application and exploration of this aspect will help solve the protective problems of the text and image processing of traditional Chinese clothing in the Internet era, and realize the innovative inheritance and development of traditional Chinese clothing.

KEY WORDS: cross modal retrieval; hashing; fine-grained; self-supervised

本文面向的数据集是传统服饰数据集, 是针对传统服饰文化保护性传承的有益探索。传统服饰文化是

中国传统文化的重要组成部分, 是中国古代劳动人民的智慧结晶, 更是中国古代社会手工艺、审美表现、

收稿日期: 2021-06-09

基金项目: 新疆维吾尔自治区重点实验室开放课题 (2017D04009)

作者简介: 赵海英 (1972—), 女, 山东人, 博士, 北京邮电大学人工智能学院副教授, 主要研究方向为文化计算与媒体数据挖掘研究。

通信作者: 向翔 (1997—), 男, 江苏人, 北京邮电大学人工智能学院硕士生, 主攻数字图像处理。

礼仪风范、价值风貌等文化内容的物质载体。然而，传统服饰因保存年代久远，会出现色泽消退、纹样不清、面料腐朽等问题，增加了传统服饰本体的研究难度。近些年来，基于传统服饰等相关非物质文化遗产的研究成为学界关注的重点。习近平总书记也多次强调要坚定文化自信，推动中华优秀传统文化的创新性发展。2017 年初，中央出台《关于实施中华优秀传统文化传承发展工程的意见》，首次以中央文件形式专题阐述中华优秀传统文化传承发展工作，并制定“到 2025 年，中国优秀传统文化传承发展体系基本完成”的方向性指导规划。传统服饰文化作为中华优秀传统文化的重要组成部分，对其进行探索性保护与传承是当前亟待关注的重点问题之一。

中国传统服饰品类丰富，在长期的社会实践中，古代劳动人民赋予了传统服饰丰富的文化内涵。目前流传于世的传统服饰除了部分收藏家收藏的传统服饰实物以外，大多为考古墓葬中发掘出来的服饰实物。因为传统服饰大多具有百年以上的历史，所以实物资料的获取十分难得，相关资料以图片和文字的信息进行记录仍是传统保护方式的重要途径。除了用图片来展示服饰本身外，服饰的历史背景、艺术特征、材料工艺、当代状况以及所蕴含的文化内涵往往需要文字来加以阐述。互联网时代计算机的介入对于探索更为便捷的图片及文字保护方式显得尤为重要。通过检索服饰图案获得其相关描述信息以及通过检索文本查看其在服饰上的具体表现形式，对于了解中国传统服饰文化具有非常重要的意义。对于跨模态检索研究，如何度量两个模态之间的关联性是解决跨模态检索问题的关键。与传统的检索方式相比，跨模态检索需要对不同模态数据建模，使其在公共空间的映射能进行相似性度量。国内外研究人员提出了多种发掘跨模态数据潜在关联的方法，希望能进一步提高跨模态检索的性能。与手动特征相比，深度学习特征能提取到多层语义信息，能更有效地捕捉模态间关联，而合理有效地利用数据原有的信息，将能提高检索精度。

1 国内外研究现状

跨模态检索相对而言是较为“年轻”的研究领域，国外的跨模态检索研究早于国内。基于空间学习的方法是当前跨媒体检索的主流，其主要的理念是：共享相同语义的数据往往具有潜在的相关性，从而能构建公共空间对此相关性进行表征。较为有代表性的跨模态检索方法有以下 5 种类别：基于传统统计、基于度量学习、基于图正则化、基于排序、基于深度哈希。最开始的跨模态检索可以认为是基于典型相关分析（Canonical Correlation Analysis, CCA）^[1]的跨模态检索，该方法是典型的基于传统统计的方法。度量学习方法是从给定的相似信息或不相似信息中学习输入特征的转换以获得更好的度量结果，此种方法被广

泛应用于单模态检索，X Zhai 等人提出的 JGRHML^[2]是较为典型的度量学习方法。除了从相似信息中学习外，JGRHML 还加入了一个联合图正则化项用于度量学习。在该联合图正则化中，不同的模态类型是互补的，联合优化它们可以使得求解过程更加容易。基于排序的方法通常将目标任务转化为排序任务。通过对跨模态检索的排序结果进行优化，使得相似度最高的实例排序得分最高。基于哈希的方法通常更适用于排序任务，其目的不仅是将数据压缩成不同的哈希码，还利用有限的二进制位来表达数据的内在属性，特别是语义相似关系和相邻结构。即如果元素在原始空间中具有语义相关性或拟合某些相邻关系，则它们在二值哈希空间中的距离也会很近。

针对跨模态数据的特征异构问题，通过对当前主流的跨模态检索方法的学习，本文提出了一种细粒度跨模态检索框架。该网络基于 DCMH^[3]框架进行改进，并基于该框架实现了自监督的细粒度跨模态检索网络。框架中加入自监督语义网络，相比普通的哈希学习有更好的细粒度语义信息提取，通过该信息对特征分布进行监督，更好地解决了跨模态数据的异构问题。通过构建传统服饰图文数据集，并在该数据集上进行跨模态检索实验，验证框架的有效性，解决实际应用问题。

2 细粒度跨模态检索算法

细粒度跨模态检索框架主要由图文模态、自监督语义和判别器学习 3 个部分组成。图文模态特征提取网络 ImgNet 和 TxtNet，自监督语义特征提取网络 LabNet，自监督细粒度跨模态检索框架见图 1。LabNet、ImgNet、TxtNet 的输入分别是标签、图像、文本，输出均为二进制码。其中，LabNet 从标签数据中提取标签语义特征，捕捉不同模式之间的数据语义相关性，然后使用语义特征来指导 ImgNet 和 TxtNet 的哈希学习过程。在公共语义空间中对图像和文本模态之间的语义相关性进行建模，同时学习语义特征的哈希函数。对于图文和标签的哈希学习有：

$$H^{l,v,t} = f^{l,v,t}(l, v, t; \theta^{l,v,t}) \quad (1)$$

其中， $f^{l,v,t}$ 是哈希函数， $\theta^{l,v,t}$ 是要学习优化的网络参数。通过对所学的值 $H^{l,v,t}$ 应用符号函数来得到二进制码 $B^{l,v,t}$ 。即：

$$B^{l,v,t} = \text{sign}(H^{l,v,t}) \in \{-1, 1\}^K \quad (2)$$

整个过程可直观表现为：

$$l, v, t \xrightarrow{f^{l,v,t}} H^{l,v,t} \xrightarrow{\text{sign}} B^{l,v,t} \quad (3)$$

2.1 特征提取网络

1982 年和 1984 年，学者 Shostack 发表了两篇关于“服务”可被设计的主题论文，首次提出了将具形的产品和无形的服务结合的概念，并首次提出了“服

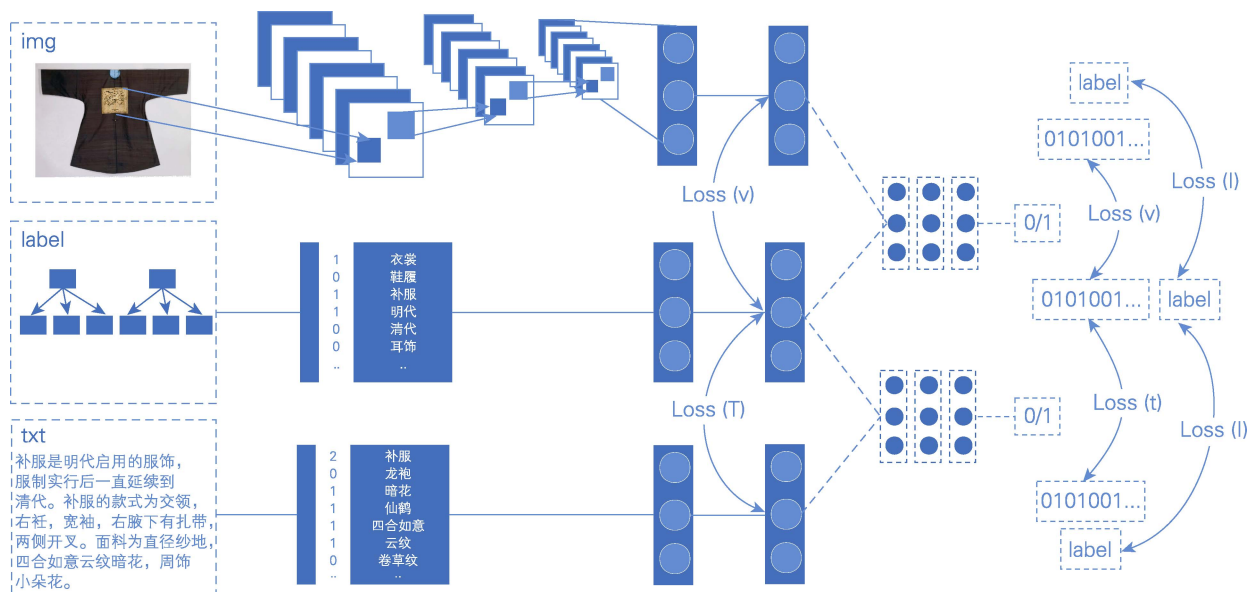


图1 自监督细粒度跨模态检索框架

Fig.1 Self-supervising fine-grained cross-modal retrieval framework

务蓝图”的设计方法。1986年学者 Heskett^[4]从组织的商业角度出发,定义了“服务”的对象为客户、雇员、利益相关者及借贷方四者。2000年 Johnson 提出了“新型服务发展”(New Service Development, NSD),意为整体发展新型服务的过程。同年 Edvardsson 在此基础上,扩大了“新型服务发展”的范围,其中包括战略、文化和服务政策的部署和实施,拓宽了服务设计的范围和领域,并提出了“服务”是由顺序、同时发生、重复或复发性3个因素组成的价值创造活动链,是一个整体的流程。1995年, Jean Gadrey 和合作者提出了“创造服务的意义是什么?”的问题,并作出回答:生产服务的定义即组织解决问题,并不是首先涉及提供一个有形的商品,而是将一系列参与者、技术、组织等各方面的能力综合运用起来,规划一种有效解决问题(对待方式、运营模式)的方案,完善了服务本身被设计的概念^[5]。

2.2 自监督语义特征提取网络和判别器学习

自监督学习^[6-7]主要思想是利用样本本身提供的信息来指导模型的学习。大多数机器学习任务都依赖于一个大规模而带有良好标签的数据集(如 ImageNet, Microsoft COCO),然后根据具体情况设计损失函数,建立神经网络。因此,有时机器学习算法的性能并不受其本身的影响,一个更好的数据集往往可以极大地提高算法的性能。然而,在许多其他的机器学习任务尤其是多模态任务中,标注良好的数据只是整个数据集的部分环节。因此,利用自监督的方式仅从现有数据提高性能是一种可行且有效的方法。

监督跨模态哈希检索方法中常用于监督的信息是与图文对应的多类标信息,本文处理的数据为传统服饰数据,由于数据集标签本身隐含层次约束,使用

自监督的方式能够直接从标签中提取更为丰富的细粒度特征,并使用该信息来监督图文两种模态的哈希学习,使得两种模态最后表征分布趋于一致。本文利用标签建立一个自监督的细粒度语义网络,学习的标签特性用于捕获语义相关性,以便在学习语义特征的哈希函数的同时,将图像和文本模态间的语义相关性也一同映射到建立的共同语义空间中。对于一个图像或一段文字,标签具有隐含的结构约束,且标签可以在语义上描述对象。模型通过计算并最小化语义损失函数,在标签语义特征的指导下,使得图像和文本特征的分布趋于一致。所用标签信息通过语义特征提取,能更好地在细粒度的层次上弥合模态之间的语义相关性。

为了进一步使得图文分布一致,以取得更好的检索结果,本文建立了图像和文本形态的两个判别器来发现它们的分布差异。对于图像/文本判别器,输入为图像/文本模态特征和语义特征,输出为0或1。该判别器可视作二分类器,其类别定义为将标签生成的语义特征的类别定义为1,将 ImageNet /TxtNet 生成的图像/文本语义模态特征的类别定义为0。将图像特征 F_v 和语义特征 F_l 输入到为图像设计的判别器,并将文本特征 F_t 和语义特征 F_l 输入到为文本设计的另一个判别器。当训练本文的模型时,这两个判别器分别处于两个对抗部分。

2.3 目标损失函数

为了详细说明,算法中涉及的问题有以下定义:用 $O = \{o_i\}_{i=1}^n$ 表示有 n 个实例的跨模态数据集,对于每个实例 $o_i = (v_i, t_i, l_i)$, 其中第 i 个实例的原始图像 $v_i \in R^{1 \times d_v}$, 第 i 个实例的原始文本特征 $t_i \in R^{1 \times d_t}$, $l_i =$

$[l_{i1}, l_{i2}, \dots, l_{ic}]$ 是分配给 o_i 的多标签标注, 其中 c 是类别标号。如果 o_i 标签中有第 j 类标签, 则 $l_{ij} = 1$, 如果 o_i 没有第 j 类标签, 则 $l_{ij} = 0$ 。对所有实例定义图像特征矩阵为 $\mathbf{V}$, 文本特征矩阵为 $\mathbf{T}$, 标签矩阵为 $\mathbf{L}$ 。使用两两多标签相似性矩阵 $S = \{s_{ij} \mid i = 1, 2, \dots, n; j = 1, 2, \dots, n\}$ 来描述两个实例之间的标签语义相似性, 即如果 o_i 与 o_j 在语义上相似, $s_{ij} = 1$, 否则 $s_{ij} = 0$ 。在多标签设置中, 两个实例 o_i 与 o_j 在语义上是否相似由多个标签决定, 即如果 o_i 和 o_j 至少共有一个标签, 则定义 $s_{ij} = 1$, 否则定义 $s_{ij} = 0$ 。

LabNet 的训练目标是使得最终生成的哈希码尽可能地保留原标签语义上的信息和图文两个模态间的相关性, 并能通过哈希码还原出最开始的标签。这将由损失函数约束实现, 如公式(4)。为了更容易理解, 使用了 $F^{l,v,t} \in R^{s \times n}$ 表示图像、文本和标签的共同语义空间中的语义特征, s 为语义空间的维数, n 为实例数。在实验中, $F^{l,v,t}$ 对应于深层神经网络 (ImgNet, TxtNet 和 LabNet)。LabNet 最终损失函数可以表示为:

$$\begin{aligned} \min_{B^l, \theta^l, \hat{L}} L^l = & -\alpha \sum_{i,j=1}^n \left(S_{ij} \Delta_{ij}^l - \log(1 + e^{\Delta_{ij}^l}) \right) - \\ & \beta \sum_{i,j=1}^n \left(S_{ij} \nabla_{ij}^l - \log(1 + e^{\nabla_{ij}^l}) \right) + \\ & \gamma \| H^l - B^l \|_F^2 + \mu \| \hat{L} - L \|_F^2 \\ \text{s.t. } B^l \in & \{-1, 1\}^K \end{aligned} \quad (4)$$

其中, $\Delta_{ij}^l = \frac{1}{2}(\mathbf{F}_i^l)^T (\mathbf{F}_j^l)$, $\nabla_{ij}^l = \frac{1}{2}(\mathbf{H}_i^l)^T (\mathbf{H}_j^l)$ 。 H^l 为预测哈希码, $\hat{L}$ 为预测标签。 α , β , γ , μ 是超参数。

公式(4)前两个量为两个负对数似然函数。第一个分量用于计算原本语义特征之间的相似损失, 第二个分量用于计算预测哈希码之间的相似损失。 Δ_{ij}^l 衡量的是样本 i 与样本 j 标签生成特征的余弦相似度, 而 ∇_{ij}^l 衡量的是样本 i 与样本 j 标签生成哈希码的余弦相似度。第三个分量是学习哈希码二值化的近似损失, 第四个分量是原始标签和预测标签的分类损失。

实现跨模态检索的一个重要理论基础是一个多模态实例的不同模态在语义上是相关的。为了保持这种语义相关性, 本文将用 LabNet 的语义信息对两种模式的特征学习过程进行监督, 包括对语义特征的监督和对学习到的二进制代码的监督。

对于图像模态的问题, 通过使用语义网络监督图像特征学习, 使 ImgNet 与语义网络保持相同的语义相关性。这是在 ImgNet 中将语义网络用于自监督。同样, 对于文本模态, 也使用语义网络来监督 TxtNet 的特征学习过程。因此, 图像模态 v 和文本模态 t 的自监督特征学习的目标函数可以写成:

$$\begin{aligned} \min_{B^{v,t}, \theta^{v,t}, \hat{L}^{v,t}} L^{v,t} = & -\alpha \sum_{i,j=1}^n \left(S_{ij} \Delta_{ij}^{v,t} - \log(1 + e^{\Delta_{ij}^{v,t}}) \right) - \\ & \beta \sum_{i,j=1}^n \left(S_{ij} \nabla_{ij}^{v,t} - \log(1 + e^{\nabla_{ij}^{v,t}}) \right) + \\ & \gamma \| H^{v,t} - B^{v,t} \|_F^2 \\ \text{s.t. } B^{v,t} \in & \{-1, 1\}^K \end{aligned} \quad (5)$$

其中, $\Delta_{ij}^{v,t} = \frac{1}{2}(\mathbf{F}_i^{v,t})^T (\mathbf{F}_j^{v,t})$, $\nabla_{ij}^{v,t} = \frac{1}{2}(\mathbf{H}_i^{v,t})^T (\mathbf{H}_j^{v,t})$ 。

$H^{v,t}$ 为预测哈希码。 α , β , γ 是超参数。

公式(5)前两个分量为两个负对数似然函数。第一个分量用于计算原本语义特征之间的相似损失, 第二个分量用于计算预测哈希码之间的相似损失, 第三个分量是学习哈希码二值化的近似损失。

公式(5)和公式(4)在形式上几乎一样, 但能从右上角的标签中看出这正是利用 LabNet 的 $\mathbf{F}_i^l$ 、 $\mathbf{H}_i^l$ 作为监督, 从而使得图文分布差距变小。

判别部分目标函数可以写成:

$$\begin{aligned} \min_{\theta^{v,t}} L_{adv}^{v,t} = & \sum_{i=1}^n \| D^{v,l}(x_i^{v,l}) - y_i^{v,l} \|_2^2 + \\ & \sum_{i=1}^n \| D^{t,l}(x_i^{t,l}) - y_i^{t,l} \|_2^2 \end{aligned} \quad (6)$$

其中, $D^{v,l}$, $D^{t,l}$ 为两个判别函数, $x_i^{v,l}$, $x_i^{t,l}$ 为判别器输入, $y_i^{v,l}$, $y_i^{t,l}$ 为输入类别值。

综上所述, 本文的框架会生成 3 种哈希码, 即 $B^{l,v,t} = \text{sign}(H^{l,v,t}) \in \{-1, 1\}^K$ 。在训练时, 将使用 $B = \text{sign}(H^l + H^v + H^t) \in \{-1, 1\}^K$ 来训练模型。最终的目标函数为 $L - L_{adv}$, 其中:

$$L = L^v + L^t + L^l \quad (7)$$

$$L_{adv} = L_{adv}^v + L_{adv}^t \quad (8)$$

3 实验结果与分析

3.1 数据集

MIRFLICKR-25K 数据集^[8]包含从 Flickr 收集的 25 000 个实例数据。每个图像都用其相关的文本标记进行标记。在实验中, 沿用 DCMH 对其做的处理, 处理完的数据总共有 20015 个, 其标签有 24 个, 文本词袋向量长度为 1386。

NUS-WIDE 数据集^[9]是一个包含 269 648 张 Web 图像的公共 Web 图像数据集, 具有 81 个手工标注的标签, 可以用于评估检索结果。在对没有任何标签或标记信息的数据进行清洗后, 选择一个 190 421 个图像文本对的子集作为本文使用的数据集, 这些图像文本对的类别都是一些较为常见的概念, 共 21 个类别。

对于两个公共数据集, 实验时在清理后的数据中随机抽取 10000 个图文对做训练和测试, 训练集和测

试集数量分别为 8000 和 2000。

自建传统服饰数据集是一个包含 4550 张传统服饰图像的多标签数据集, 总共有 40 种标签, 其标签主要为基于类别和本体描述的标签, 其标注参考 PatternNet 中标签体系字段^[13]。文本词袋向量长度为 1000。训练集和测试集数量分别为 3000 和 1550。

3.2 对比方法介绍

为了验证算法的有效性, 将分别在图像检索文本任务和文本检索图像任务上进行实验, 并选择 4 种常用的方法进行比较, 分别是 DCMH^[3]、SePH^[10]、STMH^[11]、SCM^[12]。下面是这 4 种方法的简单介绍。

(1) DCMH: 深度跨模态哈希设计一种端到端方法, 利用深度神经网络自动提取特征并进行哈希学习, 首次将特征提取与哈希学习整合在一个框架中。(2) SePH: 语义保留哈希将标签数据作为监督信息, 将训练数据的语义相似度转化为概率分布, 再用逻辑回归将特征映射到哈希码。(3) STMH: 通过对文本进行哈希, 对图像矩阵进行分解, 得到图像和文本的语义特征, 然后将其转换为公共子空间, 生成哈希码。(4) SCM: 语义相关最大化将语义标签信息无缝集成到大规模数据建模的哈希学习过程中, 最大化语义相关性。其中 DCMH 属于深度跨模态哈希方法, 其余都是浅层跨模态哈希。4 个对比方法的代码都由作者给出, 训练时参数均参照原文设置。

3.3 算法评价指标

评价跨模态检索方法需要进行两个方向的实验: 使用文本查询图像以及使用图像查询文本。为了评估本文的跨模态检索算法, 将使用以下两个常用的评价指标: 平均准确率均值 (MAP)、查准—查全曲线 (PR 曲线)。平均准确率均值是信息检索最常用的评价指标, 也是跨模态检索中使用最广泛的评价方法, 其根据查询实例和返回的查询结果之间的距离排序评价检索性能。本文计算的 MAP 实际上为 MAP@K, 其中 K 为检索序列的 top-K。平均准确率均值 (MAP) 由平均准确率 (AP) 求均值得到。平均准确率均值 (MAP) 越大, 检索效果越好。查准—查全曲线 (PR 曲线) 是以准确率 (Precision) 和查全率 (Recall) 为纵坐标和横坐标绘制的曲线, 通过 PR 曲线能直观看到在不同查全率下准确率的变化。对检索而言, 准确率表示检索结果中的相关样本数量的占比, 查全率表示检索到的相关样本数量与所有相关样本数量的比例。PR 曲线下与坐标轴围成的面积越大, 方法检索性能越好。

3.4 实验结果及实验分析

跨模态检索方法 MAP (MIRFLICKR-25K) 见表 1, 跨模态检索方法 MAP (NUS-WIDE) 见表 2, 跨模态检索方法 MAP (TCD) 见表 3, 分别展示了本文的方

表 1 跨模态检索方法 MAP (MIRFLICKR-25K)

Tab.1 Cross-modal retrieval method MAP (MIRFLICKR-25K)

Method@ Flickr-25	I->T			T->I		
	16bits	32bits	64bits	16bits	32bits	64bits
SCM	0.657	0.667	0.671	0.667	0.676	0.679
STMH	0.601	0.608	0.609	0.595	0.603	0.604
SePH	0.671	0.676	0.678	0.672	0.675	0.679
DCMH	0.732	0.737	0.746	0.773	0.778	0.786
Ours	0.768	0.773	0.781	0.792	0.790	0.794

表 2 跨模态检索方法 MAP (NUS-WIDE)

Tab.2 Cross-Modal Retrieval Method MAP (NUS-WIDE)

Method@ NUS-WIDE	I->T			T->I		
	16bits	32bits	64bits	16bits	32bits	64bits
SCM	0.523	0.540	0.552	0.467	0.476	0.478
STMH	0.519	0.520	0.532	0.495	0.513	0.534
SePH	0.478	0.488	0.490	0.472	0.475	0.479
DCMH	0.480	0.488	0.491	0.643	0.658	0.662
Ours	0.668	0.653	0.681	0.672	0.670	0.677

表 3 跨模态检索方法 MAP (TCD)

Tab.3 Cross-modal retrieval method MAP (TCD)

Method@ TCD	I->T			T->I		
	16bits	32bits	64bits	16bits	32bits	64bits
SCM	0.475	0.489	0.503	0.456	0.467	0.486
STMH	0.432	0.443	0.461	0.435	0.453	0.483
SePH	0.468	0.477	0.491	0.442	0.473	0.488
DCMH	0.506	0.510	0.518	0.501	0.505	0.509
Ours	0.553	0.558	0.560	0.542	0.550	0.551

法和其他对比方法在两种跨模态常用的公共数据集 (MIRFLICKR-25K^[13], NUS-WIDE^[14]) 和自建传统服饰数据集 (TCD)^[15]上的 MAP 结果。其中, 表中的 I->T 表示检索数据是图像类型, 检索数据库是文本类型; 表中的 T->I 表示检索数据是文本类型, 检索数据库是图像类型。

与基于神经网络的哈希方法如 STMH、SCM 相比, 本方法采用端到端网络结构, 能够自动提取各模态数据的特殊特征。与 DCMH 等深度哈希方法相比, 本方法设计了一个以自监督方式提取细粒度语义特征的网络, 并针对原 DCMH 中深度哈希网络存在的问题和不足, 采用了几种改进的方法。在 MIRFLICKR-25K 数据集上, 与 SCM、STMH、SePH 和同样用了深度哈希方法的 DCMH 相比, 本文的方法在 I->T 和 T->I 的 MAP 上都有更好的表现。对于 NUS-WIDE 数据集, 其相比 MIRFLICKR-25K 数据集而言实例更为多样复杂, DCMH 在 I->T 上并未取得更为优异的表现, 但本文的方法显示出了更好的性能。原因是: 在进行自监督标签语义学习的过程中, 加入判别器网络更加

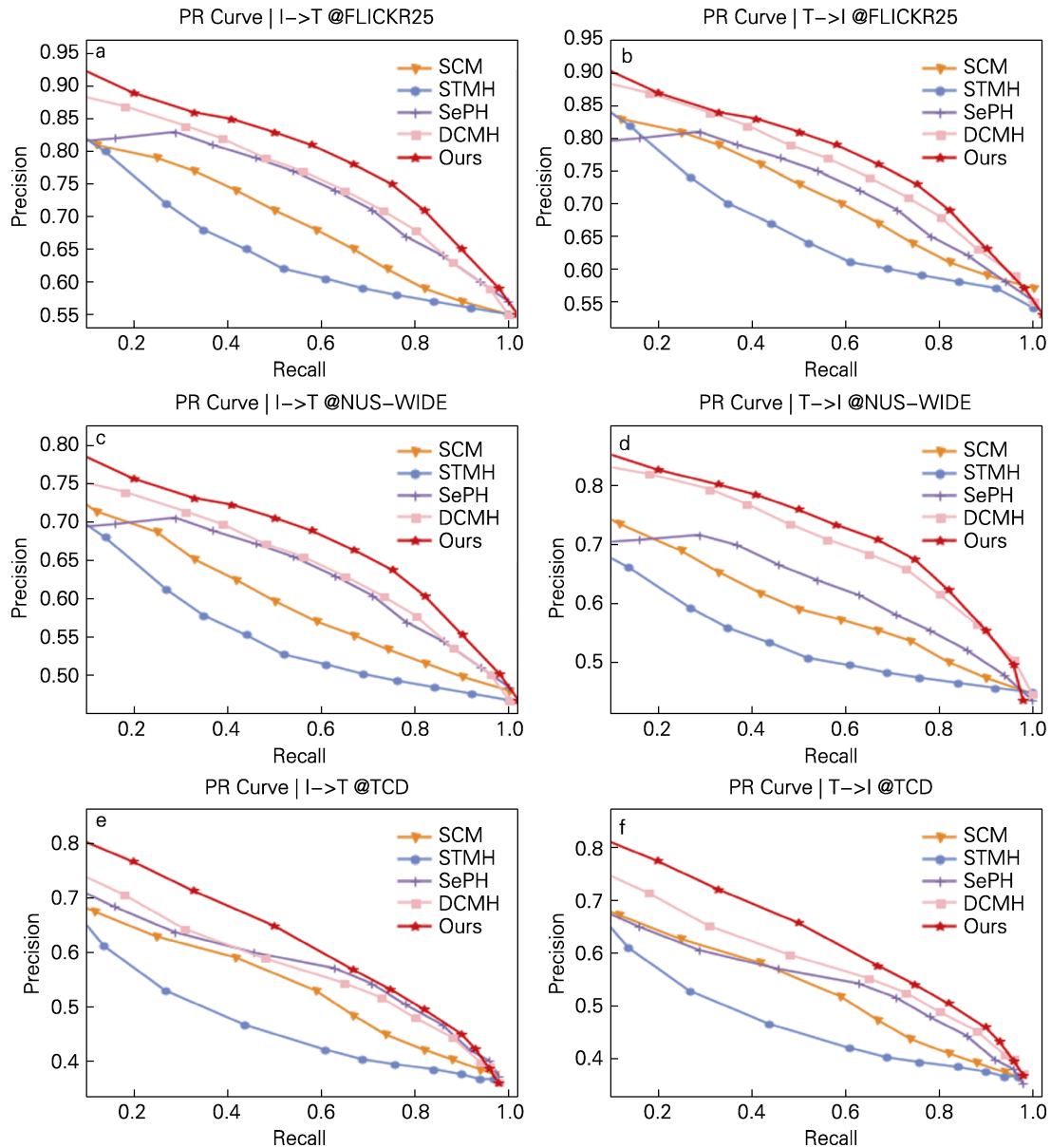


图 2 哈希码长度为 16 时的所有方法的 PR 曲线
Fig.2 The PR curve of all methods when hash code length is 16

有效地促进了两模态之间语义关联的学习,这意味着本文的方法能够学习到更多的具有区分性的表征。因此,本文的特征提取及哈希学习过程能够更准确地学习两模态之间的相关性。

本文的跨模态检索是基于哈希码的跨模态检索,基于哈希检索可以计算给定任何汉明半径的返回点的 PR 曲线。当汉明半径为 0~16,步长为 1 时,可以得到 PR 曲线。在 3 个数据集上哈希码长度为 16 时的所有方法的 PR 曲线见图 2。通过直观看 PR 曲线,可以看到本文的方法优于文中所选取的其他对比方法,这和 MAP 评价指标显示的结果一致。在浅度跨模态哈希方法中,SePH 方法的结果优于其他浅度方法。DCMH 是端到端的跨模态哈希学习,且使用的是深度图像特征提取方法,而本章方法是在 DCMH 上新增自监督语义网,因而 DCMH 和本章方法能取

得更好的检索性能。

4 结语

本文详细介绍了一种新的深跨模态哈希检索方法,该方法可称为自监督的细粒度跨模态哈希,将其用于更有效地解决实际跨模态检索的问题;并通过在两个公共数据集和一个自建数据集上的实验结果体现其性能。文中以传统服饰^[16]为检验对象,提出的自监督的细粒度跨模态哈希在 DCMH 的基础上新增了一个结合多标签信息的自监督语义网络,用于捕捉细粒度的语义信息,并通过判别器学习来最大化图文模式之间的语义相关性和特征分布一致性。文中通过实验表明,该方法在 3 个数据集上均取得了较好的检索效果,可应用于以计算机为媒介的传统服饰文本与图

像的信息快速处理与应用^[17],对将来的纹样检索工作有很大的帮助,是基于互联网时代对中国传统服饰文化保护性传承的有益探索^[18-19],后续也可以尝试更多不同模态的数据来构建更加完备的应用模型。

参考文献:

- [1] Rasiwasia N, Pereira J C, Coviello E, et al. A New Approach to Cross-Modal Multimedia Retrieval Proceedings of the 18th International Conference on Multimedia 2010[C]. Firenze: ACM, 2010.
- [2] X Zhai, Y Peng, J Xiao. Heterogeneous Metric Learning with Joint Graph Regularization for Cross-media Retrieval In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13)[C]. Bellevue: AAAI Press, 2013.
- [3] Q Jiang and W Li. Deep Cross-Modal Hashing[C]. Honolulu: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [4] K Chatfield, K Simonyan, A Vedaldi, et al. Return of the Devil in the Details: Delving Deep into Convolutional Nets[C]. British: BMVC, 2014.
- [5] A Krizhevsky, I Sutskever, G E Hinton. ImageNet Classification with Deep Convolutional Neural Networks[C]. Lake Tahoe: NIPS, 2012.
- [6] C Li, C Deng, N Li, et al. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval[C]. Salt Lake City: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
- [7] Chung S W, Chung J S, Kang H G. Perfect Match: Self-Supervised Embeddings for Cross-modal Retrieval[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 99: 1.
- [8] M J Huiskes and M S Lew. The Mir Flickr Retrieval Evaluation[C]. Niagara Falls: ACM CIVR, 2008.
- [9] T S Chua, J Tang, R Hong, et al. Nus-wide: a Real-world Web Image Database from National University of Singapore[C]. Geneva: ACM CIVR, 2009.
- [10] Z Lin, G Ding, M Hu, et al. Semantics-preserving Hashing for Cross-view Retrieval[C]. Boston: CVPR, 2015.
- [11] D Wang, X Gao, X Wang, et al. Semantic Topic Multimodal Hashing for Cross-media Retrieval[C]. Buenos Aires: IJCAI, 2015.
- [12] Zhang D, Li W. Large-scale Supervised Multimodal Hashing with Semantic Correlation Maximization[C]. Québec: AAAI, 2014.
- [13] M J Huiskes and M S Lew. The MIR Flickr Retrieval Evaluation[C]. Xi'an: Proceedings of the International Conference on Mechatronics and Intelligent Robotics, 2008.
- [14] Tat-Seng Chua, Jinhui Tang, Richang Hong, et al. Nus-wide: a Real-world Web Image Database from National University of Singapore[C]. Geneva: ACM CIVR, 2009.
- [15] 赵海英, 高子惠, 邓恋, 等. 基于图文混排的传统服饰图像以文标图算法[J/OL]. 图学学报, 2021, 6(10): 1-10.
ZHAO H Y, GAO Z Y, DENG L, et al. Method of Automatic Image Annotation for Image-text Mixed Domain Books[J/OL]. Journal of Graphics, 2021, 6(10): 1-10.
- [16] 刘姣姣, 梁惠娥. 管窥三十年来的服色文化研究[J]. 丝绸, 2019, 56(3): 98-104.
LIU Jiao-jiao, LIANG Hui-e. A Look at Three Decades of Research on Color Culture[J]. Silk, 2019, 56(3): 98-104.
- [17] 李晓兵. 我国传统服饰艺术的现代价值[J]. 丝绸, 2007(1): 48-49.
LI Xiao-bing. The Modern Value of Traditional Costume Art in China[J]. Silk, 2007(1): 48-49.
- [18] 许星. 中国传统服饰文化对现代时尚的影响[J]. 丝绸, 2001(12): 36-38.
XU Xing. The Influence of Traditional Chinese Dress Culture on Modern Fashion[J]. Silk, 2001(12): 36-38.
- [19] 吴春胜. 传统与时尚融合的中国风格服装设计手法[J]. 丝绸, 2011, 48(1): 37-39.
WU Chun-sheng. Traditional and Fashion Fusion of Chinese Style Clothing Design Techniques[J]. Silk, 2011, 48(1): 37-39.