软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3152−3164 [doi: 10.13328/j.cnki.jos.006620]

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

基于跨模态自蒸馏的零样本草图检索*

田加林, 徐 行, 沈复民, 申恒涛

(电子科技大学 计算机科学与工程学院, 四川 成都 611731)

通信作者: 申恒涛, E-mail: shenhengtao@hotmail.com

摘 要: 零样本草图检索将未见类的草图作为查询样本, 用于检索未见类的图像. 因此, 这个任务同时面临两个挑战: 草图和图像之间的模态差异以及可见类和未见类的不一致性. 过去的方法通过将草图和图像投射到一个公共空间来消除模态差异, 还通过利用语义嵌入 (如词向量和词相似度) 来弥合可见类和未见类的语义不一致. 提出了跨模态自蒸馏方法, 从知识蒸馏的角度研究可泛化的特征, 无需语义嵌入参与训练. 具体而言, 首先通过传统的知识蒸馏将预训练的图像识别网络的知识迁移到学生网络. 然后, 通过草图和图像的跨模态相关性, 跨模态自蒸馏将上述知识间接地迁移到草图模态的识别上, 提升草图特征的判别性和泛化性. 为了进一步提升知识在草图模态内的集成和传播, 进一步地提出草图自蒸馏. 通过为数据学习辨别性的且泛化的特征, 学生网络消除了模态差异和语义不一致性. 在 3 个基准数据集, 即 Sketchy、TU-Berlin 和 QuickDraw, 进行了广泛的实验, 证明了所提跨模态自蒸馏方法与当前方法相比较的优越性.

关键词: 零样本草图检索; 零样本学习; 跨模态检索; 知识蒸馏

中图法分类号: TP391

中文引用格式: 田加林, 徐行, 沈复民, 申恒涛. 基于跨模态自蒸馏的零样本草图检索. 软件学报, 2022, 33(9): 3152–3164. http://www.jos.org.cn/1000-9825/6620.htm

英文引用格式: Tian JL, Xu X, Shen FM, Shen HT. Cross-modal Self-distillation for Zero-shot Sketch-based Image Retrieval. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3152–3164 (in Chinese). http://www.jos.org.cn/1000-9825/6620.htm

Cross-modal Self-distillation for Zero-shot Sketch-based Image Retrieval

TIAN Jia-Lin, XU Xing, SHEN Fu-Min, SHEN Heng-Tao

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Zero-shot sketch-based image retrieval uses sketches of unseen classes as query samples for retrieving images of those classes. This task is thus faced with two challenges: the modal gap between a sketch and the image and inconsistencies between seen and unseen classes. Previous approaches tried to eliminate the modal gap by projecting the sketch and the image into a common space and bridge the semantic inconsistencies between seen and unseen classes with semantic embeddings (e.g., word vectors and word similarity). This study proposes a cross-modal self-distillation approach to investigate generalizable features from the perspective of knowledge distillation without the involvement of semantic embeddings in training. Specifically, the knowledge of the pre-trained image recognition network is transferred to the student network through traditional knowledge distillation. Then, according to the cross-modal correlation between a sketch and the image, cross-modal self-distillation indirectly transfers the above knowledge to the recognition of the sketch modality to enhance the discriminative and generalizable features of sketch features. To further promote the integration and propagation of the knowledge within the sketch modality, this study proposes sketch self-distillation. By learning discriminative and generalizable features from the data, the student network eliminates the modal gap and semantic inconsistencies. Extensive experiments conducted on three benchmark datasets, namely Sketchy, TU-Berlin, and QuickDraw, demonstrate the superiority of the proposed cross-modal self-distillation approach to the state-of-the-art ones.

Key words: zero-shot sketch-based image retrieval; zero-shot learning; cross-modal retrieval; knowledge distillation

---

* 基金项目: 国家自然科学基金 (61976049, 62072080, 61632007)
本文由“融合媒体环境下的媒体内容分析与信息服务技术”专题特约编辑汪萌教授、张勇东教授、俞俊教授以及张伟高级工程师推荐.
收稿时间: 2021-06-27; 修改时间: 2021-08-15; 采用时间: 2022-01-14; jos 在线出版时间: 2022-02-22

融媒体旨在整合存在共同点又存在互补性的媒体, 需要充分各种媒介载体, 实现“资源通融、内容兼融、宣传互融、利益共融”的新型媒体. 在这种需求之下, 针对各种媒体数据的智能处理是必然要面临的挑战. 近年来, 移动互联网的蓬勃发展带来了多媒体数据爆发式的增长. 这些数据不仅来源广泛, 而且内容和媒体形式也复杂多变. 在这种环境下, 如何更加精准地进行内容分析、建立不同媒体数据间的联系并服务于数据检索与分析等应用场景, 是实现融媒体的重要一环. 当今时代, 随着触摸屏设备的流行, 电子数据化的手绘草图变得越来越容易获取. 由于草图几乎可以由任何人费很小的代价画出, 且不涉及隐私和版权的问题, 对于融媒体实现具有很高的利用价值.

草图检索是利用草图的一个重要方向. 尽管草图表现出高度的抽象性, 它仍然包含足够的结构和外形信息来描述对象, 催生出使用草图从庞大的图像集中检索出所需内容的需求. 因此, 基于草图的图像检索 (sketch-based image retrieval, SBIR) 任务得到了越来越多的关注和研究. 现有的 SBIR 方法在可见类 (即训练时所用数据的类别集合) 数据上的检索效果表现良好, 但却难以应用到实际的应用场景. 第 1 原因在于“类别”这个概念广泛存在于现实场景中, 不可能收集到所有类别的数据. 第 2 原因在于这些方法在设计时只考虑训练数据的特点, 却未考虑方法的泛化性. 因此, 它们在零样本草图检索 (zero-shot sketch-based image retrieval, ZS-SBIR)[1]任务中被证实性能表现不佳.

对于 ZS-SBIR 任务, 模型训练于可见类数据, 但却测试于未见类数据. 这样的行为差异要求我们在模型设计和训练时, 既要考虑草图和图像数据形态上的模态差异, 也要考虑可见类和未见类的语义不一致性问题[2]. 最近, 一些工作大部分只专注于解决模态差异[3]问题, 对语义不一致性问题不够重视. 它们中的大部分工作[4–8]都采取深度生成模型作为主要框架, 学习从模态的原始表征到公共嵌入空间[9]的投影, 但忽略了之前由预训练模型获得的知识. 虽然 Liu 等人[10]率先尝试利用知识蒸馏过程来保留丰富的视觉特征, 但他们的方法依旧是基于单模态知识蒸馏的想法. 由于目前没有大规模预训练的草图识别模型, 单模态知识蒸馏方法只能针对图像模态, 忽略了对草图的泛化性的重要性.

此外, 这些 ZS-SBIR 方法 (除了 Kiran 等人的工作[5]) 期望通过简单地利用语义嵌入, 以消除可见类和未见类的语义不一致性问题. 它们或者是从词向量模型中提取类名的词向量[1,6,8], 或者通过分层模型衡量类名的词相似性[10], 或者以上二者的结合[4]. 然而, 这有两方面的问题. 一方面是, 语义嵌入编码的信息大部分是文本信息, 但 ZS-SBIR 是视觉任务, 它的引导作用不是最优的. 另一方面在于, 从类名提取语义嵌入需要预先定义准确的类名, 并且需要额外的语言模型和时间消耗, 导致训练资源获取方面的负担. 然而, 在一些实际的应用场景中, 数据只能被数字标记 (例如, 出于隐私原因), 或者类名是稀有词或复合词, 因而无法从语言模型中提取语义嵌入.

为了解决上述问题, 本文提出了一种新的方法, 即跨模态自蒸馏方法 (cross-modal self-distillation, CMSD), 用于零样本草图检索. 本文提出的 CMSD 方法可以通过跨模态知识迁移而无需语义嵌入来实现超越现有方法的性能. 如图 1 所示, CMSD 方法解决了现有的单模态知识蒸馏的限制, 将知识流通过跨模态迁移从图像模态引入其他模态. 图 2 展示了 CMSD 方法的具体流程框架, 在该框架中, 学生网络同时处理图像和草图, 通过特征的相似性和加权概率实现跨模态知识迁移.

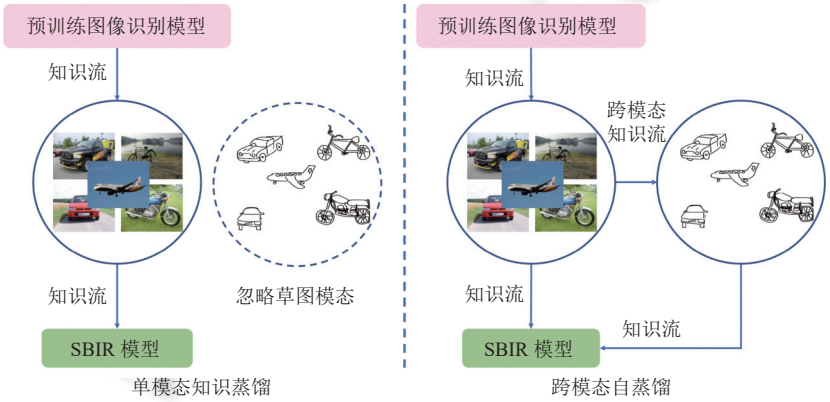


图 1 单模态知识蒸馏和跨模态知识蒸馏的区别

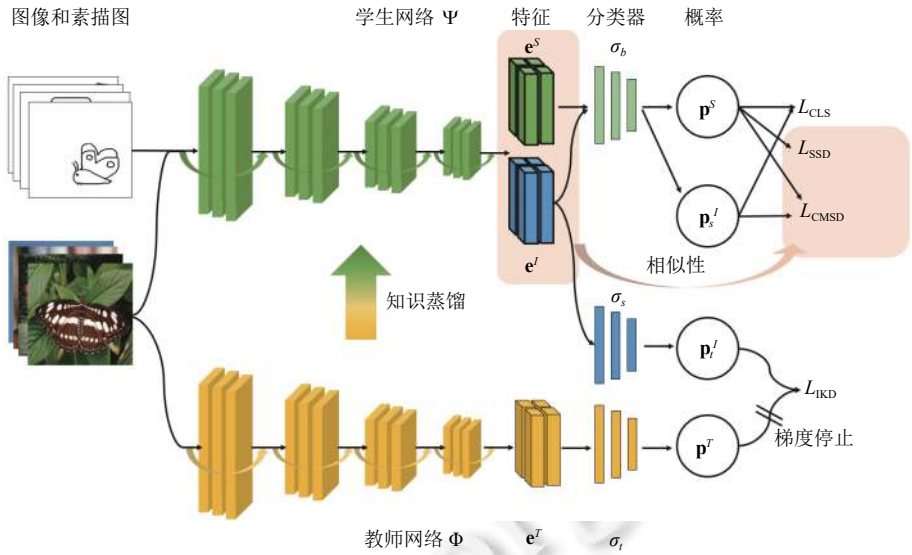


图 2　本文提出的基于跨模态自蒸馏的草图检索的架构图

为了从图像和草图数据中学习具有判别性、又不失泛化性的特征, 我们从分类和蒸馏两大方面进行模型训练. 首先, 图像和草图必须被正确的分类, 这将保证特征的判别性. 其次, 知识蒸馏将避免训练中的模型迅速地过拟合到特定于模态的信息, 造成预训练知识的遗忘. 我们采用和设计了 3 种知识蒸馏损失. 第一, 我们通过传统的知识蒸馏损失, 使学生网络模仿教师模型的图像分类能力. 第二, 我们提出跨模态自蒸馏损失. 跨模态自蒸馏利用草图和图像的特征相似度作为桥接模态差异的手段, 以集成和传播的方式, 将教师模型的图像分类知识间接地迁移到草图模态. 具体而言, 我们假设具有相近视觉特征的图像和草图也应该具有一致的分类预测概率; 因而对任意草图样本而言, 其他图像样本的分类预测概率可以加权、集成和传播, 形成软目标 (soft target), 作为该草图样本的分类监督信号. 通过优化跨模态自蒸馏损失, 既可以缩小特征的模态差异, 同时还提升模型在零样本场景下的泛化性. 最后, 我们进一步假设特征相近的草图也有上述的性质, 提出草图自蒸馏损失, 使得来自图像模态的知识得到更加有效的利用, 进一步提升特征的判别性和泛化性. 综上, 我们提出的 CMSD 方法既避免了语义嵌入所带来的资源获取负担, 同时还通过跨模态自蒸馏提升模型在零样本场景下的判别性和泛化性, 并通过大量的实验验证了 CMSD 的优越性.

综上所述, 本文的贡献有如下 3 个方面.

(1) 我们提出了一种新颖的 ZS-SBIR 方法, 简称为 CMSD. 该方法关注于视觉样本本身的信息, 不受语义嵌入的资源限制和性能限制.

(2) 我们设计了一种新颖的跨模态自蒸馏损失, 通过跨模态特征相似度, 间接地将教师网络的知识迁移到草图模态, 最终提升模型泛化性和判别性.

(3) 我们进一步提出草图自蒸馏损失, 以特征相似度为权重, 以集成和传播的方式对知识进行加权聚合, 使得知识蒸馏对于 ZS-SBIR 任务更加有效.

我们在用于 ZS-SBIR 任务的 3 个大规模基准数据集 Sketchy[11]、TU-Berlin[12]和 QuickDraw[8]上, 对本文提出的方法进行了广泛的实验对比和消融分析. 与十几种最先进的方法相比, 我们提出的 CMSD 方法始终取得了卓越的性能, 证明了本方法中草图自蒸馏和跨模态自蒸馏策略的有效性.

1　相关工作概述

本文研究工作为基于跨模态自蒸馏的零样本草图检索, 属于如下领域的交叉: 基于草图的图像检索 (sketch-based image retrieval, SBIR)、零样本学习 (zero-shot learning, ZSL) 和知识蒸馏 (knowledge distillation, KD).

1.1 基于草图的图像检索

SBIR 任务不做零样本假设, 只关注模态差异问题. 这个领域的方法可以大致地分为手工特征和深度模型方法. 早期的方法是基于手工特征的[13,14], 它们计算图像的边缘图作为图像的替代, 再使用词袋模型抽取边缘图和草图的特征, 以期获得特征的匹配. 当深度学习在识别任务中大获成功之后, 涌现了许多基于深度模型的方法. 其中, 以孪生网络为架构的方法[15–20]起到了重要作用. 它们以端到端的方式来解决这个问题, 并使用通用的排序损失来训练模型, 如对比损失[15]、三元组排序损失[16]和 HOLEF 损失[17].

1.2 零样本学习

ZSL 的开创性工作[21]提出可见类和未见类的概念, 期望模型根据语义特征和可见类的视觉特征进行推理, 并可直接应用于未见类数据的识别问题. 随后的工作大部分是基于投影的框架, 偏重于建立语义特征和视觉特征的联系. 它们要么直接学习从视觉空间到语义空间的映射[22], 或者反向投影[23,24]以避免前一种方法带来的枢纽点问题, 又或者学习另一个公共空间[25]. 最近, 很多工作[26–28]利用生成模型来合成未见类的特征, 解决单纯的投影所不能处理的领域偏差 (domain bias) 问题, 因而将 ZSL 问题转化为传统的监督分类问题. 值得说明的是, 所有的 ZSL 方法都依赖于额外信息, 包括词向量、层次信息和属性向量. 大多数 ZS-SBIR 方法[1–3]受此类方法的启发, 选择生成模型作为主体框架, 同样将语义嵌入引入到模型的训练流程中.

1.3 知识蒸馏

知识蒸馏指的是一种训练策略, 其中学生模型学习从预训练的教师网络中提取的各种知识. 这种技术最初应用于模型压缩领域[29], 此后广泛应用于对抗性防御[30]和特权学习[31]等方面. 从训练范式角度看, 教师网络的知识通常为类似概率的“软”标签, 这比通常的“硬”标签包含更多的类间关系. 理论上, 软标签也起到标签平滑和数据增强的作用. 最近, 已经有一些知识蒸馏工作在探索实例之间的关系[32–34]. 虽然我们提出的草图自蒸馏和跨模态自蒸馏的设计思想是基于特征相似度完成知识的集成和传播, 但也有许多不同之处: 我们专注于研究跨模态数据检索任务, 而不是单模态数据的分类任务; 我们考虑零样本设置下的判别性和泛化性的平衡问题, 提出跨模态蒸馏的知识迁移策略.

1.4 零样本草图检索

ZS-SBIR 任务结合 SBIR 和 ZSL 任务的特点, 研究如何同时处理草图和图像的模态差异问题以及可见类和未见类的语义不一致性问题. 如前所述, 大多数 ZS-SBIR 工作选择生成模型作为主体框架, 将语义嵌入引入到框架和损失函数的设计中, 并最终学习一个公共空间作为检索空间. 以生成散列模型[1,35], 自动编码器[5], 变量自动编码器[5,6]和生成对抗网络[6]为主要架构的方法是这类方法的典型. Liu 等人[10]从领域适应的角度看待 ZS-SBIR 任务, 并提出使用知识蒸馏避免灾难性遗忘. 他们中的大多数利用自然语言处理领域的语言模型, 提取词向量[1,6,8], 衡量词相似度, 或结合上述两者[4]. 尽管 Liu 等人[10]首先在 ZS-SBIR 任务中提出知识蒸馏的训练范式, 但他们的方法仍旧引入了语义嵌入, 且只关注了单模态的蒸馏. 尽管 Kiran 等人[5]未引入语义嵌入, 但却没有提出有效的方法去解决语义不一致性, 只实现了相对较差的性能. 我们的方法提出跨模态自蒸馏和草图自蒸馏, 取得了超越现有方法的性能表现.

2 基于跨模态自蒸馏的零样本草图检索

2.1 零样本草图检索定义

我们首先给出零样本草图检索的定义. 零样本草图检索的目的在于, 利用属于可见类的训练数据 (草图和图像) 训练一个模型, 并将其应用于检索属于未见类的草图相关的图像. 因此, 我们可以假定训练集为 $D_{\mathrm{tr}}=\{I^{\mathrm{seen}}, S^{\mathrm{seen}}\}$, 其中, $I^{\mathrm{seen}}=\{(\mathbf{I}_i,\mathbf{y}_i)|\mathbf{y}_i\in C^{\mathrm{seen}}\}_{i=1}^{n_1}$ 和 $S^{\mathrm{seen}}=\{(s_i,\mathbf{y}_i)|\mathbf{y}_i\in C^{\mathrm{seen}}\}_{i=1}^{n_2}$ 分别表示草图和图像数据所构成的集合, C^{seen} 为训练阶段所有数据所属类别构成的集合. 同样地, 由未见类 C^{un} 数据所构成的测试集可定义为 $D_{\mathrm{te}}=\{I^{\mathrm{un}},S^{\mathrm{un}}\}$. 在零样本领域, C^{seen} 和 C^{un} 集合之间的交集为空, 即 $C^{\mathrm{seen}}\cap C^{\mathrm{un}}=\emptyset$. 由于训练和测试阶段面临的不同类别数据, 本文

提出的方法构造各种软标签, 通过知识蒸馏训练模型.

2.2 CMSD 总体架构

本文所提出的 CMSD 模型的总体架构如图 2 所示. 它包含两个作为特征提取骨架的深度卷积网络 (教师网络 Φ 和学生网络 Ψ) 以及 3 个分类器 (σ_b, σ_s 和 σ_t). Φ 和 Ψ 在架构上几乎相同, 除了输出层的维度不同. 我们将教师网络和学生网络的输出特征的维度分别记为 d_1 和 d_2, 因此 3 个分类器的函数表示分别为如下公式: $\sigma_b(\cdot;\theta_b):\mathbb{R}^{d_2}\mapsto\{0,1\}^{|C^{\text{seen}}|}$, $\sigma_s(\cdot;\theta_s):\mathbb{R}^{d_2}\mapsto\{0,1\}^{|C^I|}$, 以及 $\sigma_t(\cdot;\theta_t):\mathbb{R}^{d_1}\mapsto\{0,1\}^{|C^I|}$. 其中, C^I 表示 ImageNet 数据集的类别集合, θ_b、θ_s 和 θ_t 分别表示相应分类器的参数. 图像通过学生网络和教师网络获得的特征分别用 $\mathbf{e}^I$ 和 $\mathbf{e}^T$ 表示. 草图只输入学生网络, 提取的特征用 $\mathbf{e}^S$ 表示. 如图 2 所示, 各个特征经过各个分类器得到相应的分类概率, 分别为 $\mathbf{p}^S$, $\mathbf{p}_s^I$, $\mathbf{p}_t^I$ 和 $\mathbf{p}^T$. 这些符号的上标的意义与特征符号的上标意义相同, 下标用于标识图像的分类概率. 概率向量 $\mathbf{p}_s^I$ 的维数为 C^{seen} 的类别数目, 而 $\mathbf{p}_t^I$ 的维数是前述 C^I 的类别数目 (固定为 1 000). 前者用于分类和构造草图的跨模态自蒸馏的软目标, 后者用于计算图像知识蒸馏损失.

具体而言, Φ 和 Ψ 由相同的 SE-ResNet-50[36]初始化 Φ 和 Ψ 的权重参数, 此外 σ_t 也初始化为 SE-ResNet-50 的分类器. 在训练阶段, Φ 和 σ_t 的参数不参与更新 (如图 2 所示“梯度停止”), 只用于监督学生网络 Ψ 及其分类器 σ_s 的训练过程. 由于学生网络要同时处理草图和图像, 我们在 SE-ResNet-50 的 Squeeze-and-Excitation 模块中添加了一个二进制编码, 用于指示输入是图像还是草图. 因此, 也可称特征提取网络为 CSE-ResNet-50 (C 为 conditional 的简写). 这样的框架微调有助于消除模态差异, 因为草图和图像的特征提取网络可视为参数共享的孪生网络.

2.3 单模态的图像知识蒸馏

教师网络是在非常大规模的图像数据集 (ImageNet) 上预训练完成的, 具有强大的辨别能力. 对于一张图像, 教师网络输出的概率向量提供了更加细粒度的语义信息, 而这通常是“硬”标签所包含的. 基于这一观察, Hinton 等人[26]提出, 让学生网络通过匹配教师网络给出的软标签来模仿和学习教师网络的分类能力.

给定一个图片样本 I_i, 将其输入 Φ 和 Ψ, 得到特征嵌入 $\mathbf{e}_i^I$ 和 $\mathbf{e}_i^I$; 再输入分类器 σ_s 和 σ_t, 通过 Softmax 归一化得到概率向量 $\mathbf{p}_{t,i}^I$ 和 $\mathbf{p}_i^T$. 知识蒸馏将学生和教师的预测结果之间的 KL 散度最小化, 使得学生网络模仿教师网络对图像模态的分类能力, 可公式化为如下:

$$L_{\text{IKD}}=D_{\text{KL}}(\mathbf{p}_i^T|\mathbf{p}_{t,i}^I) \tag{1}$$

公式 (1) 大体类同传统的知识蒸馏方法, 借助于在大规模图像数据集上预训练的深度卷积网络, 再结合分类损失, 驱使学生网络不仅能在训练集上学习到具有判别性的特征, 还能保留从大规模数据集中学习得到的知识, 从而使特征具有泛化性, 有助于消弭零样本设置下的语义不一致. 然而, 上述方法缺陷也很明显, 即需要预训练的深度卷积网络. 由于缺乏在大规模草图数据集上预训练的深度卷积网络, 因此无法应用于草图模态, 难以消弭多模态任务所面临的模态差异.

2.4 跨模态自蒸馏

为了解决上述模态差异问题, 本文提出跨模态自蒸馏解决此问题, 无需草图模态的预训练网络, 只需通过模态内和模态间的自蒸馏, 完成知识的二次迁移: 将图像模态的预训练模型的图像分类能力迁移为学生模型的图像分类能力, 再迁移为学生模型的草图分类能力. 通过知识的二次迁移, 学生模型在图像和草图两个模态上都得到监督引导, 同时消弭了模态差异和语义差异, 使得学生模型在零样本跨模态任务中表现更好.

图像和草图模态间的知识集成和传播, 即为前述跨模态自蒸馏. 给定任意一批样本, 假设图像和草图数目分别为 N_1 和 N_2, 对应的特征嵌入矩阵表示为 $\mathbf{E}^I=\left[\mathbf{e}_1^I,\mathbf{e}_2^I,\ldots,\mathbf{e}_{N_1}^I\right]^{\mathrm{T}}$ 和 $\mathbf{E}^I=\left[\mathbf{e}_1^S,\mathbf{e}_2^S,\ldots,\mathbf{e}_{N_1}^S\right]^{\mathrm{T}}$, 对应的概率矩阵表示为 $\mathbf{P}^I=\left[\mathbf{p}_{s,1}^I,\mathbf{p}_{s,2}^I,\ldots,\mathbf{p}_{s,N_1}^I\right]^{\mathrm{T}}$ 和 $\mathbf{P}^S=\left[\mathbf{p}_1^S,\mathbf{p}_2^S,\ldots,\mathbf{p}_{N_2}^S\right]^{\mathrm{T}}$, 其中 $\mathbf{p}_{s,i}^I$ 和 $\mathbf{p}_i^S$ 是通过带温度超参数 τ 的 Softmax 归一化得到的向量. 不失一般性地, $\mathbf{p}_i^S$ 的计算公式为:

$$\mathbf{p}_i^S = \frac{\exp\left(\sigma_b\left(\mathbf{e}_i^S\right)/\tau\right)}{\sum_{i=1}^{|C^{\text{seen}}|}\exp\left(\sigma_b\left(\mathbf{e}_i^S\right)/\tau\right)} \tag{2}$$

然后, 以余弦相似度为距离度量, 计算草图和图像特征的成对相似度, 构成相似矩阵 $\mathbf{R}\in\mathbb{R}^{N_1\times N_2}$:

$$\mathbf{R}_{i,j} = \frac{\left(\mathbf{e}_i^S\right)^T\cdot\mathbf{e}_j^I}{\left\|\mathbf{e}_i^S\right\|\cdot\left\|\mathbf{e}_j^I\right\|}) \tag{3}$$

接着, 将相似矩阵 $\mathbf{R}$ 的每一行进行归一化, 使得 $\mathbf{R}$ 的行和都为 1, 记为 $\widehat{\mathbf{R}}$, 即 $\sum_j\widehat{\mathbf{R}}_{i,j}=1$:

$$\widehat{\mathbf{R}}_{i,j} = \frac{\exp\left(\mathbf{R}_{i,j}\right)}{\sum_j\exp(\mathbf{R}_{i,j})} \tag{4}$$

对于任意草图样本, 可根据相似矩阵 $\mathbf{R}$ 加权集成图像样本的分类预测概率, 以形成草图的自蒸馏软目标, $\widehat{\mathbf{p}}_i^S=\sum_j\widehat{\mathbf{R}}_{i,j}\mathbf{p}_{s,j}^I$. 因此, 当草图和图像的特征近似时, $\mathbf{R}_{i,j}$ 接近于 1, $\widehat{\mathbf{p}}_i^S$ 也更接近于 $\mathbf{p}_{s,j}^I$. 然而, 单纯加权图像的分类预测概率, 也难以避免噪声和训练的不稳定性. 因此, 软目标可定为 $\sum_{j\neq i}\widehat{\mathbf{R}}_{i,j}\mathbf{p}_{s,j}^I$ 和 $\mathbf{p}_i^S$ 的滑动平均, 则最终的软目标可公式化如下:

$$\widehat{\mathbf{p}}_i^S = \omega\cdot\sum_j\widehat{\mathbf{R}}_{i,j}\mathbf{p}_{s,j}^I + (1-\omega)\cdot\mathbf{p}_i^S \tag{5}$$

其中, ω 为滑动平均的加权系数. 因此, 跨模态自蒸馏损失为:

$$L_{\text{CMSD}} = \tau^2\cdot D_{\text{KL}}\left(\widehat{\mathbf{p}}_i^S\mid\mathbf{p}_i^S\right) \tag{6}$$

2.5 单模态的草图自蒸馏

跨模态自蒸馏通过特征相似矩阵将草图和图像两个模态联系起来, 进而将草图模态和预训练模型联系起来, 完成知识从图像模态到草图模态的迁移. 我们进一步地提出草图自蒸馏损失, 目的在于将学到的知识在每一批草图样本中传播, 提升知识迁移的有效性.

类似地, 草图特征的成对相似度可定义为如下:

$$\mathbf{R}_{i,j}^S = \frac{\left(\mathbf{e}_i^S\right)^T\cdot\mathbf{e}_j^S}{\left\|\mathbf{e}_i^S\right\|\cdot\left\|\mathbf{e}_j^S\right\|} \tag{7}$$

与前述跨模态自蒸馏一样做归一化:

$$\widetilde{\mathbf{R}}_{i,j}^S = \frac{\mathbf{1}_{i\neq j}\cdot\exp\left(\mathbf{R}_{i,j}^S\right)}{\sum_j\mathbf{1}_{i\neq j}\cdot\exp\left(\mathbf{R}_{i,j}^S\right)} \tag{8}$$

其中, 当 $i\neq j$ 为真时 $\mathbf{1}_{i\neq j}$ 为 1, 否则为 0. 值得注意的是, 我们用 $\mathbf{1}_{i\neq j}$ 排除样本的自我比较情况, 以免影响该样本和其他样本的比较强度. 经滑动平均后得到的软目标为:

$$\widetilde{\mathbf{p}}_i^S = \omega\cdot\sum_j\widetilde{\mathbf{R}}_{i,j}^S\mathbf{p}_j^S + (1-\omega)\cdot\mathbf{p}_i^S \tag{9}$$

其中, ω 为滑动平均的加权系数. 由于公式 (8) 中的 $\widetilde{\mathbf{R}}^S$ 为对称方阵 (当 N_1 和 N_2 不相等时, 公式 (4) 中的 $\widehat{\mathbf{R}}$ 不为方阵), 可重复迭代公式 (9) 直至收敛, 可由矩阵形式解析解计算:

$$\widetilde{\mathbf{P}}_{(\infty)}^S = (1-\omega)\left(I-\omega\widetilde{\mathbf{R}}^S\right)^{-1}\mathbf{P}^S \tag{10}$$

因此, 草图自蒸馏损失可定义为如下:

$$L_{\text{SSD}} = \tau^2\cdot D_{\text{KL}}\left(\widetilde{\mathbf{p}}_{(\infty),i}^S\mid\mathbf{p}_i^S\right) \tag{11}$$

2.6 总体目标函数

综合公式 (1)、公式 (6) 和公式 (11), 以及草图和图像的交叉熵分类损失, 得到总体目标函数为:

$$L = \lambda_1 L_{\text{IKD}} + \lambda_2 L_{\text{CMSD}} + \lambda_3 L_{\text{SSD}} + \lambda_4 L_{\text{CLS}} \tag{12}$$

其中, λ_1、λ_2、λ_3 和 λ_4 为加权参数, L_{CLS} 为交叉熵分类损失. 对于图像模态, 其优化目标函数为:

$$L_{\text{IMG}} = \lambda_1 L_{\text{IKD}} + \lambda_4 L_{\text{CLS}} \tag{13}$$

对于草图模态, 其优化目标函数为:

$$L_{\text{SKT}} = \lambda_2 L_{\text{CMSD}} + \lambda_3 L_{\text{SSD}} + \lambda_4 L_{\text{CLS}} \tag{14}$$

2.7 优化过程

训练时, 同时采样草图和图像, 每批样本由 N_1 张图像和 N_2 张草图组成. N_1 和 N_2 的比例视数据集的情况而定: 对于图像和草图数量平衡的 Sketchy 和 QuickDraw, 我们将 N_1 和 N_2 的比例设为 2:1; 对于图像和草图数量极端不平衡的 TU-Berlin, 我们将比例设为 8:1. 这样的设置使得学生模型逐步学习草图的特征, 而不至于快速过拟合到特定于模态、特定于类别的状态, 从而提高模态的泛化性. 在实际优化过程中, 我们采用随机梯度下降对公式 (12) (公式 (13) 和公式 (14) 之和) 优化学生网络 Ψ 以及分类器 σ_b 和 σ_s 的参数:

$$\theta_\Psi, \theta_{\sigma_b}, \theta_{\sigma_s} = \arg\min_{\theta_\psi, \theta_{\sigma_b}, \theta_{\theta_s}} \lambda_1 L_{\text{IKD}} + \lambda_2 L_{\text{CMSD}} + \lambda_3 L_{\text{SSD}} + \lambda_4 L_{\text{CLS}} \tag{15}$$

模型收敛时, 停止优化过程. 学生网络的 θ_Ψ 最终用于提取未见类图像和草图的特征, 实现最后的检索.

3 实验结果与分析

3.1 数据集

3 个 ZS-SBIR 基准数据集用于度量我们提出的方法, 包括 Sketchy[11]、TU-Berlin[12]和 QuickDraw[8]. 原始的 Sketchy 数据集由 125 类别构成, 包含 12 500 张自然图像和 75 471 张草图. Liu 等人[8]对图像集合进行拓展, 最终得到包含 73 002 张图像的集合. 这个数据集有两种训练和测试数据的划分方法: 一种随机选择 25 个类别作为未见类 (Shen 等人[1]); 另一种固定选择 21 个类别 (Kiran 等人[5]), 这些类别确保与 ImageNet 类别集合没有交集. 我们在两种设置下都进行了实验.

TU-Berlin 数据集[12]由 250 个类别构成, 包含 20 000 张草图和 204 489 张自然图像. 正如草图数量只有图像的 1/10, 它存在草图和图像数量的极端不平衡, 使得模型在这个数据集上优化更加困难. 我们依照 Shen 等人[1]提出的划分方式, 随机选择 30 个类作为未见类.

QuickDraw 数据集[8]是 3 个数据集中最大的数据集, 由 110 个类别构成, 但包含了总共 33 万张草图和 20.4 万张图片. 另外, 这个数据集所包含的草图内容最抽象, 来源于业余用户的手绘, 而非专业人员的绘画. 同样地, 这个数据集划分出 30 个类别作为未见类, 并严格保证它们与 ImageNet 类别没有交集. 因此, 这个数据集也是 3 个数据集中最具有挑战性的数据集.

3.2 实现细节

我们的实验代码使用 PyTorch 实现, 在两块 RTX2080Ti GPU 上进行模型的训练. 我们选择 Adam 优化器作为模型的优化方法. 各损失函数加权系数 λ_1、λ_2、λ_3 和 λ_4 分别设为 1, 0.1, 0.1 和 1. 一批图像的数量 N_1 被设置为 64, N_2 按照前述比例进行设置. 学习率初始为 1E–4, 在训练过程中以指数衰减的方式降低到 1E–6. 蒸馏的温度超参数保持 $\tau = 0.1$. 滑动平均系数 ω 设为 0.5. 此外, 在训练期间周期性地冻结批量归一化层对算法的性能有提升作用. 为了公平比较, 我们采用前人提出的评价指标[6,7], 包括均值评价精度 (mAP@k) 和准确率 (Prec@k), 其中 k 表示前 k 个查询结果. 为未见类草图和图像提取特征后, 我们采用余弦相似性作为距离度量来进行检索.

3.3 与现有方法的比较

我们将提出的 CMSD 与 8 个现有的 ZS-SBIR 工作进行了比较: ZSIH[1], CAAE 和 CVAE[5], SEM-PCYC[4], Dey 等人[8], SAKE[10], LCALE[6], OCEAN[7]. 我们按照有无使用语义嵌入来分类这些方法. 其中, 除 CAAE 和

CVAE 外, 其余的算法都将语义嵌入引入至框架或损失函数的设计中. 为了更清楚地分析我们提出的方法的优越性, 我们根据 SAKE 的代码重新训练出无语义嵌入参与的模型, 并将其命名为 SAKE w/o s. 此外, 我们还分别将 SBIR 和 ZSL 领域的两篇论文纳入比较 (GN-Triplet[16]和 DSH[11], 以及 SAE[37]和 ZSH[38]), 以分析 ZS-SBIR 方法在零样本和跨模态设置下的优越性.

表 1 本文 CMSD 方法和 12 种比较方法在 Sketchy 和 TU-Berlin 上的总体比较

方法	语义嵌入	Sketchy		Sketchy (split 2)		TU-Berlin	
		mAP@all	Prec@100	mAP@200	Prec@200	mAP@all	Prec@100
GN-Triplet (TOG2016)[16]	×	0.211	0.310	0.083	0.169	0.189	0.241
DSH (CVPR2017)[11]	×	0.164	0.210	0.059	0.153	0.122	0.198
CAAE (ECCV2018)[5]	×	0.196	0.284	0.156	0.260	—	—
CVAE (ECCV2018)[5]	×	—	—	0.225	0.333	—	—
SAKE (w/o s)[10]	×	0.540	0.681	0.481	0.582	0.462	0.584
CMSD (Ours)	×	**0.620**	**0.733**	**0.504**	**0.601**	**0.489**	**0.620**
ZSH (ACM MM2016)[38]	√	0.165	0.217	—	—	0.139	0.174
SAE (CVPR2017)[37]	√	0.210	0.302	0.136	0.238	0.161	0.210
ZSIH (CVPR2018)[1]	√	0.254	0.340	—	—	0.220	0.291
SEM-PCYC (CVPR2019)[4]	√	0.349	0.463	—	—	0.297	0.426
Dey等人 (CVPR2019)[8]	√	—	—	0.369	0.370	0.110	0.121
SAKE (ICCV2019)[10]	√	0.547	0.692	0.497	0.598	0.475	0.599
LCALE (AAAI2020)[6]	√	0.476	0.583	—	—	—	—
OCEAN (ICME2020)[7]	√	0.462	0.590	—	—	—	—
CMSD (Ours)	×	**0.620**	**0.733**	**0.504**	**0.601**	**0.489**	**0.620**

注: —表示原始论文中没有报告相应指标的数字, 粗体和下划线分别表示最好和次好的结果

表 2 本文 CMSD 方法和 3 种比较方法在 QuickDraw 上的总体比较

方法	语义嵌入	QuickDraw			
		mAP@all	mAP@200	Prec@100	Prec@200
CVAE (ECCV2018)[5]	×	0.003	0.006	—	0.003
Dey等人 (CVPR2019)[8]	√	0.075	0.090	—	0.068
SAKE (w/o s)[10]	×	0.121	0.089	0.178	0.168
CMSD (Ours)	×	**0.142**	**0.122**	**0.227**	**0.215**

注: —表示原始论文中没有报告相应指标的数字, 粗体和下划线分别表示最好和次好的结果

在 Sketchy 和 TU-Berlin 数据集上的实验结果见表 1, 在 QuickDraw 数据集上的实验结果见表 2. 我们首先比较无语义嵌入参与训练的方法, 这包含 SBIR 方法和一些 ZS-SBIR 方法. 从表 1 可以观察到, 除了 CAAE 在 Sketchy 上表现不加, SBIR 方法的整体表现远远不如 ZS-SBIR 方法的整体表现. 原因在于 SBIR 方法只考虑了训练数据的分布, 而没考虑模型在未见类数据上的迁移性. 比较 ZS-SBIR 方法, 综合表 1 和表 2 可知, CAAE 和 CVAE 表现不佳, 我们提出的 CMSD 在所有实验设置下都取得了最好的结果. 以 mAP 指标为例, 我们在 Sketchy 上比 SAKE (w/o s) 高出 0.08, 在 Skethcy (split 2) 上高出 0.023, 在 TU-Berlin 上高出 0.027, 在 QuickDraw 高出 0.021. 这样一致且显著的提升证明 CMSD 能通过跨模态自蒸馏有效地完成知识迁移, 使得 CMSD 无需语义嵌入就能取得最好的结果. 而 CAAE 和 CVAE 都是生成式方法, 在没有将语义嵌入引入模型训练时, 很难实现将模型泛化至未见类的草图检索. SAKE 也是从知识蒸馏角度看待零样本草图检索任务, 然而它在取消语义嵌入后依旧出现了性能下降. 这进一步说明了我们方法的有效性.

接着, 我们将 CMSD 和使用语义嵌入的那一类方法进行比较, 它们属于 ZSL 和 ZS-SBIR 领域. 类似地, ZSL

方法的整体性能表现不如 ZS-SBIR, 因为它们并没有考虑多模态数据所面临的模态差异问题, 使得学习得到的特征依据保留较大的模态差异. 在 ZS-SBIR 方法中, CMSD 依旧表现出了一致且显著的提升, 超越了所有的现有方法. 因此, 这证明了 CMSD 无需语义嵌入也能同时处理模态差异和语义不一致问题.

综上实验观察, 我们可以得出结论, CMSD 在零样本草图检索任务上表现优异, 既减轻了对训练资源的需求, 也改善了现有方法在模型泛化性方面的不足, 提出的跨模态自蒸馏在解决模态差异和语义不一致性方面的价值.

3.4 CMSD 消融分析

我们通过消减相应的损失项来分析它们 (包括跨模态自蒸馏 CMSD、草图自蒸馏 SSD、图像知识蒸馏 IKD 和分类 CLS) 的效果. 我们选择 Sketchy 和 TU-Berlin 数据集做实验, 并在表 3 中显示了消融的模型. 这些模型的编号从模型 1 到模型 7. 模型 1 集成了所有损失项, 这是我们提出的方法. 模型 2、3、4 和 5 可以归为一类: 上述 4 个组件中的一个没有参与模型的训练和优化. 没有跨模态自蒸馏和草图自蒸馏的消融模型被称为模型 6. 我们进一步消减了所有的知识蒸馏组件, 并将其命名为模型 7.

表 3　在 Sketchy 和 TU-Berlin 数据集上每种组成部分的消融分析 (mAP@all)

模型	名称	Sketchy	TU-Berlin
1	CMSD+SSD+IKD+CLS	0.620	0.489
2	CMSD+SSD+CLS	0.521	0.453
3	CMSD+IKD+CLS	0.595	0.476
4	SSD+IKD+CLS	0.593	0.471
5	CMSD+SSD+IKD	0.184	0.199
6	IKD+CLS	0.571	0.460
7	CLS	0.483	0.422

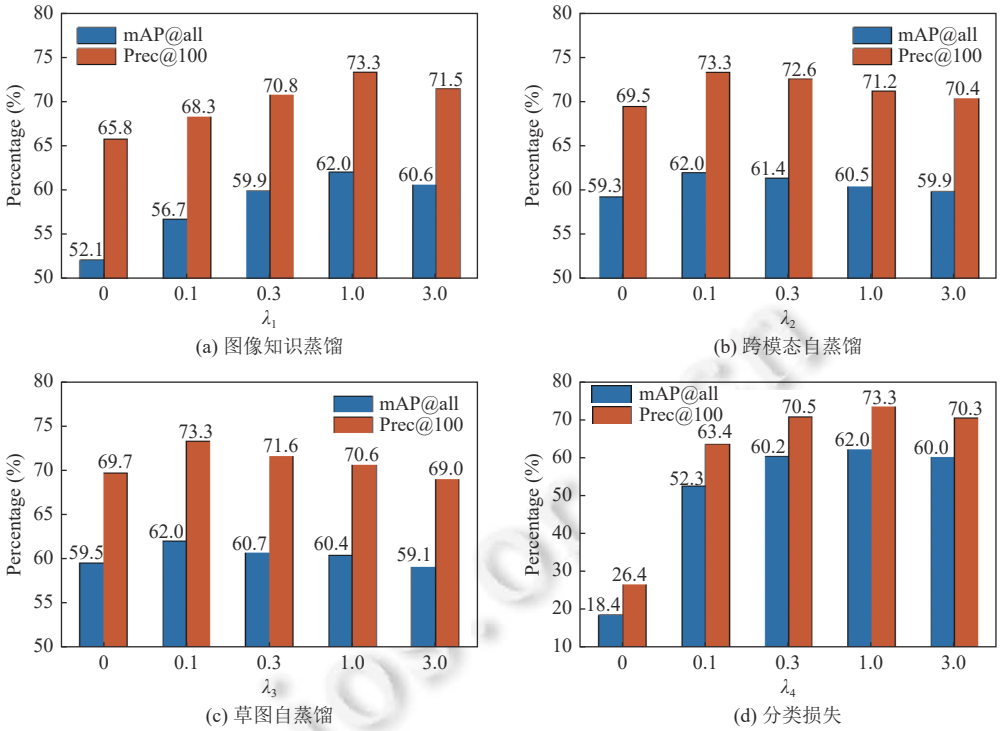
消融结果见表 3. 通过比较这些模型的实验结果, 我们可以得出以下结论: 1) 图像知识蒸馏是 3 个蒸馏损失项中最重要的, 它的消融将导致性能的明显下降, 这在模型 2、3、4 的比较中得到了体现. 2) 尽管跨模态自蒸馏和草图自蒸馏的重要性不如图像知识蒸馏, 但它们与图像知识蒸馏集合后可以显著提高模型的性能, 这可以在模型 3 和 6、模型 4 与 6 的比较中观察到. 3) 与模型 7 相比, 跨模态自蒸馏和草图自蒸馏的组合 (模型 4) 也取得了改善, 这表明跨模态自蒸馏和草图自蒸馏能有效地将知识从教师模型迁移到学生模型. 4) 模型 5 的性能严重退化, 因为在消减分类损失后, 模态差距很难缩小. 5) 模型 7 的性能明显下降, 因为所有的知识蒸馏损失在训练过程中不包括在内, 导致检索未见类的样本所需的泛化能力严重下降. 6) 在集成所有损失时, 我们的方法 (模型 1) 同时利用知识蒸馏和分类得到超过所有变体的结果.

3.5 CMSD 参数分析

在这个实验中, 我们通过改变 Sketchy 上的 λ_1、λ_2、λ_3 和 λ_4 的值进行了参数分析. 我们将这些系数的范围设定为 [0.0, 3.0], 并在图 3 中显示了结果. 我们可以观察到, 性能曲线在 λ_2、λ_3 为 0.1 时达到峰值, 在 λ_1 和 λ_4 为 1 时达到峰值. 当这些系数的值过大或过小时, 性能不可避免地会下降. 这可能是因为当这些系数小的时候, 训练过程对某一项组件关注过少, 影响了知识蒸馏的有效性; 而当它们太大时, 训练过程则对某一组件关注过多, 增加了优化的不稳定性. 因此, 适当加权有利于总体性能的提高.

3.6 定性实验

我们展示了 CMSD 在 Sketchy 上的可视化结果, 并在图 4 中与 SAKE 做了定性比较, 其中, 最左边的草图代表查询样本和它们的类别标签, 右边是正确的和错误的候选者, 分别用绿色的外框线和红色的叉号标记. 通过比较, 发现我们的模型可以成功地检索到与查询草图的相同类别的图像, 但 SAKE 的检索结果却有些错误. 通过观察检索结果, 我们还发现, 我们模型的检索能力是基于草图和图像中的视觉信息的, 即查询草图和候选图像在形状和结构模式上非常相似. 这样的结果在某种程度上是合理的: 草图是高度抽象的, 主要描绘物体的外观结构, 略去了细粒度的细节和复杂的背景, 这导致物体结构相似的图像更容易被检索.



(a) 图像知识蒸馏 (b) 跨模态自蒸馏

(c) 草图自蒸馏 (d) 分类损失

图 3 在 Sketchy 数据集上, 随参数 λ_1、λ_2、λ_3 和 λ_4 变化的实验结果

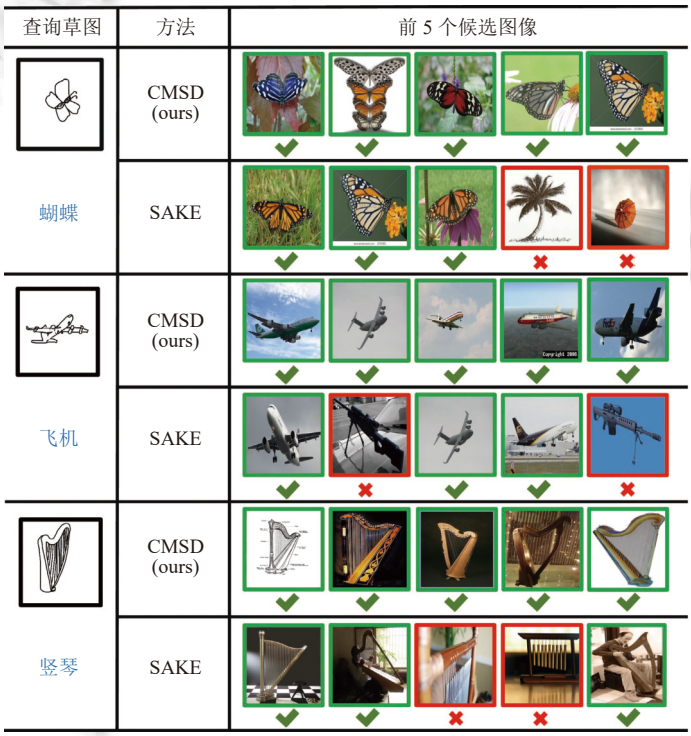


图 4 本文 CMSD 和比较模型 SAKE 在 Sketchy 数据集上的定性例子

4 结 论

本文提出了一种新颖的跨模态自蒸馏模型来解决零样本草图检索问题, 并在没有语义嵌入的情况下实现最好的性能. 一方面, 我们提出了跨模态自蒸馏损失, 使知识从预训练的图像识别模型流向草图模态. 另一方面, 我们进一步提出草图自蒸馏损失, 以集成和传播的方式使得知识在草图模态内得到更有效的利用. 在 3 个 ZS-SBIR 基准数据集上的广泛比较结果证明了我们模型的有效性. 消融实验和参数分析实验证明我们方法有效地完成知识迁移, 解决零样本草图检索所面临的模态差异问题和语义不一致性问题. 我们将在未来的工作中探索更有效的解决方案, 以最大限度地减少模态差异和语义不一致.

References:

[1] Shen YM, Liu L, Shen FM, Shao L. Zero-shot sketch-image hashing. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3598–3607. [doi: 10.1109/CVPR.2018.00379]

[2] Xu X, Lu HM, Song JK, Yang Y, Shen HT, Li XL. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. IEEE Trans. on cybeRnetics, 2020, 50(6): 2400–2413. [doi: 10.1109/TCYB.2019.2928180]

[3] Xu X, Wang T, Yang Y, Zuo L, Shen FM, Shen HT. Cross-modal attention with semantic consistence for image-text matching. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(12): 5412–5425. [doi: 10.1109/TNNLS.2020.2967597]

[4] Dutta A, Akata Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5084–5093. [doi: 10.1109/CVPR.2019.00523]

[5] Yelamarthi SK, Reddy SK, Mishra A, Mittal A. A zero-shot framework for sketch based image retrieval. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 316–333. [doi: 10.1007/978-3-030-01225-0_19]

[6] Lin K, Xu X, Gao LL, Wang Z, Shen HT. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2020. 11515–11522. [doi: 10.1609/aaai.v34i07.6817]

[7] Zhu JW, Xu X, Shen FM, Lee RKW, Wang Z, Shen HT. Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval. In: Proc. of the 2020 IEEE Int'l Conf. on Multimedia and Expo. London: IEEE, 2020. 1–6. [doi: 10.1109/ICME46284.2020.9102940]

[8] Dey S, Riba P, Dutta A, Lladós J, Song YZ. Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2174–2183. [doi: 10.1109/CVPR.2019.00228]

[9] Xu X, Lin KY, Yang Y, Hanjalic A, Shen HT. Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3030–3047. [doi: 10.1109/TPAMI.2020.3045530]

[10] Liu Q, Xie LX, Wang HY, Yuille A. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3661–3670. [doi: 10.1109/ICCV.2019.00376]

[11] Liu L, Shen FM, Shen YM, Liu XL, Shao L. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2298–2307. [doi: 10.1109/CVPR.2017.247]

[12] Zhang H, Liu S, Zhang CQ, Ren WQ, Wang R, Cao XC. SketchNet: Sketch classification with web images. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1105–1113. [doi: 10.1109/CVPR.2016.125]

[13] Eitz M, Hildebrand K, Boubekeur T, Alexa M. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. Computers & Graphics, 2010, 34(5): 482–498. [doi: 10.1016/j.cag.2010.07.002]

[14] Fan YC, Tan XH, Zhou MQ, Zheng X. A scale invariant local descriptor for sketch based 3d model retrieval. Chinese Journal of Computers, 2017, 40(11): 2448–2465 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.02448]

[15] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 539–546. [doi: 10.1109/CVPR.2005.202]

[16] Sangkloy P, Burnell N, Ham C, Hays J. The sketchy database: Learning to retrieve badly drawn bunnies. ACM Trans. on Graphics, 2016, 35(4): 119. [doi: 10.1145/2897824.2925954]

[17] Song JF, Yu Q, Song YZ, Xiang T, Hospedales TM. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5552–5561. [doi: 10.1109/ICCV.2017.592]

[18] Chen J, Bai C, Ma Q, Hao PY, Chen SY. Adversarial training triplet network for fine-grained sketch based image retrieval. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1933–1942 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5934.htm [doi: 10.13328/j.cnki.jos.005934]

[19] Xu P, Yin QY, Huang YY, Song YZ, Ma ZY, Wang L, Xiang T, Kleijn WB, Guo J. Cross-modal subspace learning for fine-grained sketch-based image retrieval. Neurocomputing, 2018, 278: 75–86. [doi: 10.1016/j.neucom.2017.05.099]

[20] Wang YF, Huang F, Zhang YJ, Feng R, Zhang T, Fan WG. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. Pattern Recognition, 2020, 100: 107148. [doi: 10.1016/j.patcog.2019.107148]

[21] Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453–465. [doi: 10.1109/TPAMI.2013.140]

[22] Romera-Paredes B, Torr PHS. An embarrassingly simple approach to zero-shot learning. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: ICML, 2015. 2152–2161.

[23] Zhang L, Xiang T, Gong SG. Learning a deep embedding model for zero-shot learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3010–3019. [doi: 10.1109/CVPR.2017.321]

[24] Wang ZQ, Yang W. Zero-shot learning based on semantic alignment and reconstruction. Computer Engineering and Design, 2021, 42(1): 70–75 (in Chinese with English abstract). [doi: 10.16208/j.issn1000-7024.2021.01.011]

[25] Long Y, Liu L, Shao L, Shen FM, Ding GG, Han JG. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6165–6174. [doi: 10.1109/CVPR.2017.653]

[26] Xian YQ, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5542–5551. [doi: 10.1109/CVPR.2018.00581]

[27] Chen Z, Wang S, Li JJ, Huang Z. Rethinking generative zero-shot learning: An ensemble learning perspective for recognising visual patches. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 3413–3421. [doi: 10.1145/3394171.3413813]

[28] Liu S, Shi CJ, Liu JY, Zhou WB, Chen QY. Zero-shot classification based on cycle-consistency. In: Proc. of the 14th National Conf. on Signal and Intelligent Information Processing and Application. Beijing, 2021. 500–507 (in Chinese with English abstract).

[29] Wang J, Bao WD, Sun LC, Zhu XM, Cao BK, Yu PS. Private model compression via knowledge distillation. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: IEEE, 2019. 1190–1197. [doi: 10.1609/aaai.v33i01.33011190]

[30] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 2016 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2016. 582–597. [doi: 10.1109/SP.2016.41]

[31] Gao ZF, Chung J, Abdelrazek M, Leung S, Hau WK, Xian ZC, Zhang HY, Li S. Privileged modality distillation for vessel border detection in intracoronary imaging. IEEE Trans. on Medical Imaging, 2020, 39(5): 1524–1534. [doi: 10.1109/TMI.2019.2952939]

[32] Peng BY, Jin X, Li DS, Zhou SF, Wu YC, Liu JH, Zhang ZN, Liu Y. Correlation congruence for knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 5006–5015. [doi: 10.1109/ICCV.2019.00511]

[33] Tung F, Mori G. Similarity-preserving knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1365–1374. [doi: 10.1109/ICCV.2019.00145]

[34] Ye HJ, Lu S, Zhan DC. Distilling cross-task knowledge via relationship matching. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12393–12402. [doi: 10.1109/CVPR42600.2020.01241]

[35] Shen HT, Liu LC, Yang Y, Xu X, Huang Z, Shen FM, Hong RC. Exploiting subspace relation in semantic labels for cross-modal hashing. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(10): 3351–3365. [doi: 10.1109/tkde.2020.2970050]

[36] Lu P, Huang G, Lin HY, Yang WM, Guo GD, Fu YE. Domain-aware SE network for sketch-based image retrieval with multiplicative euclidean margin softmax. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 3418–3426. [doi: 10.1145/3474085.3475499]]

[37] Kodirov E, Xiang T, Gong SG. Semantic autoencoder for zero-shot learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4447–4456. [doi: 10.1109/CVPR.2017.473]

[38] Yang Y, Luo YD, Chen WL, Shen FM, Shao J, Shen HT. Zero-shot hashing via transferring supervised knowledge. In: Proc. of the 24th ACM Int'l Conf. on Multimedia. Amsterdam: ACM, 2016. 1286–1295. [doi: 10.1145/2964284.2964319]

附中文参考文献:

[14] 樊亚春, 谭小慧, 周明全, 郑霞. 基于局部多尺度的三维模型草图检索方法. 计算机学报, 2017, 40(11): 2448–2465. [doi: 10.11897/SP.J.1016.2017.02448]

[18] 陈健, 白琮, 马青, 郝鹏翼, 陈胜勇. 面向细粒度草图检索的对抗训练三元组网络. 软件学报, 2020, 31(7): 1933–1942. http://www.jos.org.cn/1000-9825/5934.htm [doi: 10.13328/j.cnki.jos.005934]

[24] 王紫沁, 杨维. 基于语义对齐和重构的零样本学习算法. 计算机工程与设计, 2021, 42(1): 70–75. [doi: 10.16208/j.issn1000-7024.2021.

01.011]
[28] 刘帅, 史彩娟, 刘靖祎, 周文博, 程琦云. 基于循环一致性的零样本分类. 见: 第十四届全国信号和智能信息处理与应用学术会议论文集. 2021. 500–507.

田加林(1998－), 男, 硕士生, 主要研究领域为多媒体信息检索, 机器学习.

沈复民(1985－), 男, 博士, 教授, 博士生导师, 主要研究领域为计算机视觉, 人工智能.

徐行(1988－), 男, 博士, 副教授, 主要研究领域为多媒体信息检索, 模式识别, 计算机视觉.

申恒涛(1977－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体, 计算机视觉, 人工智能.