

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2023)05-1384-12

论文引用格式: Jiang D and Ye M. 2023. Transformer network for cross-modal text-to-image person re-identification. Journal of Image and Graphics, 28(05):1384-1395(姜定, 叶茫. 2023. 面向跨模态文本到图像行人重识别的Transformer网络. 中国图象图形学报, 28(05):1384-1395)[DOI: 10.11834/jig.220620]

面向跨模态文本到图像行人重识别的 Transformer 网络

姜定, 叶茫*

武汉大学计算机学院, 武汉 430072

摘要: 目的 文本到图像的行人重识别是一个图像文本跨模态检索的子任务, 现有方法大都采用在全局特征匹配的基础上加入多个局部特征进行跨模态匹配。这些局部特征匹配的方法都过分复杂且在检索时会大幅减慢速度, 因此需要一种更简洁有效的方法提升文本到图像的行人重识别模型的跨模态对齐能力。对此, 本文基于通用图像文本对大规模数据集预训练模型, 对比语言—图像预训练(contrastive language-image pretraining, CLIP), 提出了一种温度投影匹配结合 CLIP 的文本到图像行人重识别方法。**方法** 借助 CLIP 预训练模型的跨模态图像文本对齐的能力, 本文模型仅使用全局特征进行细粒度的图像文本语义特征对齐。此外, 本文提出了温度缩放跨模态投影匹配(temperature-scaled cross modal projection matching, TCMPM)损失函数来进行图像文本跨模态特征匹配。**结果** 在本领域的两个数据集上与最新的文本到图像行人重识别方法进行实验对比, 在 CUHK-PEDES(CUHK person description)和 ICFG-PEDES(identity-centric and fine-grained person description)数据集中, 相比于现有性能较好的局部匹配模型, 本文方法 Rank-1 值分别提高了 5.92% 和 1.21%。**结论** 本文提出的基于双流 Transformer 的文本到图像行人重识别方法可以直接迁移 CLIP 的跨模态匹配知识, 无须冻结模型参数训练或接入其他小模型辅助训练。结合提出的 TCMPM 损失函数, 本文方法仅使用全局特征匹配就在检索性能上大幅超过了现有局部特征方法。

关键词: 跨模态检索; 行人重识别; 行人搜索; Transformer; 图文检索

Transformer network for cross-modal text-to-image person re-identification

Jiang Ding, Ye Mang*

School of Computer Science, Wuhan University, Wuhan 430072, China

Abstract: Objective Text-to-image person re-identification is a sub-task of image-text retrieval, which aims to retrieve the target person images corresponding to the given text description. The main challenge of the text-to-image person re-identification task is the significant feature gap between vision and language. The fine-grained matching between the semantic information of the two modalities is restricted by modal gap as well. The mixture of multiple local features and global feature are often adopted for cross-modal matching recently. These local-level matching methods are complicated and suppress the retrieval speed. Insufficient training data is still challenged for text-to-image person re-identification tasks

收稿日期: 2022-06-24; 修回日期: 2022-11-06; 预印本日期: 2022-11-13

* 通信作者: 叶茫 yemang@whu.edu.cn

基金项目: 国家自然科学基金项目(62176188); 湖北省重点研发计划项目(2021BAA187)

Supported by: National Natural Science Foundation of China(62176188); Key R&D Program of Hubei Province, China (2021BAA187)

as well. To alleviate this insufficiency, conventional methods are typically initialized their backbone models with weights pre-trained on single-modal large-scale datasets. However, this initialization method cannot be used to learn the information of fine-grained image-text cross-modal matching and its semantic alignment. Therefore, an easy-to-use method is required to optimize the cross-modal alignment for the text-to-image person re-identification model. **Method** We develop a transformer network with a temperature-scaled projection matching method and contrastive language-image pre-training (CLIP) for text-to-image person re-identification. The CLIP is a general multimodal foundation model pre-trained on large-scale image-text datasets. The vision transformer is used as the visual backbone network to preserve fine-grained information, which can resolve the convolutional neural network (CNN)-based constraint of long-range relationships and down-sampling. To optimize the cross-modal image-text alignment capability of the pre-trained CLIP model, our model is focused on fine-grained image-text semantic feature alignment using global features only. In addition, a temperature-scaled cross-modal projection matching (TCMPM) loss function is developed for image-text cross-modal feature matching as well. The TCMPM loss can be used to minimize the Kullback-Leibler(KL) divergence between temperature-scaled projection distributions and normalized true matching distributions in a mini-batch. **Result** Extensive experiments are carried out on two datasets in comparison with the latest text-to-image person re-identification methods. We adopt the two popular public datasets, CUHK person description(CUHK-PEDES) and identity-centric and fine-grained person description(ICFG-PEDES), to validate the effectiveness of the proposed method. Rank- K ($K = 1, 5, 10$) are adopted as the retrieval evaluation metrics. On the CUHK-PEDES dataset, the Rank-1 value is improved by 5.92% compared to the best performing existing local-level matching method, and it is improved by 7.09% for existing global-level matching method. On the ICFG-PEDES dataset, the Rank-1 value is improved by 1.21% for local-level matching model. The ablation studies are also carried out on the CUHK-PEDES and ICFG-PEDES dataset. Compared to original CPM loss, the Rank-1 value of the TCMPM loss is improved by 9.54% on the CUHK-PEDES dataset, and the Rank-1 value is improved by 4.67% on the ICFG-PEDES dataset. Compared to the InfoNCE loss, a commonly-used loss in cross-modal comparative learning, the Rank-1 value can be improved by 3.38% on the CUHK-PEDES dataset in terms of the TCMPM loss, and the Rank-1 value is improved by 0.42% on the ICFG-PEDES dataset. **Conclusion** An end-to-end dual Transformer network is developed to learn representations of person images and descriptive texts in the text-to-image person re-identification. We demonstrate that the global-level matching method has its potential to outperform current state-of-the-art local-level matching methods. The transformer network can resolve the problem that CNN cannot model the long-range relationship and detailed information-loss for down-sampling. In addition, our proposed method can benefit from the powerful cross-modal alignment capability of CLIP, and together with our further designed TCMPM loss, our model can thus learn more discriminative image-text features.

Key words: cross modal retrieval; person re-identification; person search; Transformer; text-image retrieval

0 引言

文本到图像行人重识别(text-to-image person re-identification)是行人重识别和跨模态图像检索的子问题,旨在利用自然语言描述从大规模图像或视频库中检索最符合文本描述形容的行人,如图 1 所示。相比于基于属性的行人重识别,文本描述需要的专业知识和先验知识更少,且可以提供比属性更全面的描述。此外,文本到图像行人重识别技术可以很好地弥补传统的基于图像的行人重识别技术至少需要一幅行人图像的局限性。



图 1 文本到图像行人重识别示例

Fig. 1 Example of text-to-image person re-identification

自 Li 等人(2017b)首次提出使用文本描述检索对应的行人图像以来,人们提出了许多探索文本到

图像行人重识别任务的方法。虽然这些方法取得了一定成功,但是检索性能还不足以应对真实世界应用场景。造成文本到图像行人重识别任务检索准确率不高的主要原因是图像和文本两个模态之间的差异,由于图像和文本的表现形式不一致,两个模态的语义信息之间很难做到精准的细粒度匹配。而且现有方法的图像骨干网络都采用了卷积神经网络(convolutional neural network, CNN)结构,而 CNN 的下采样操作会导致细粒度的特征丢失,从而影响跨模态的细粒度匹配。现有文本到图像行人重识别方法可以分成全局特征匹配和局部特征匹配两类。全局特征匹配方法主要集中在全局视觉和文本表征的学习上,通过优化目标将图像和文本特征映射到统一的特征空间。现有的全局特征匹配方法使用的 CNN 和 LSTM/GRU (long short-term memory/gated recurrent unit)骨干网络无法有效提取有辨识度的全局特征。近年来,在基于图像的行人重识别方法中,提出了越来越多的局部特征匹配方法(郑鑫等, 2020)。一些局部特征匹配的方法在全局特征的基础上,引入了额外的局部特征对行人图像和文本描述进行细粒度的匹配,这种局部特征匹配配合全局特征匹配可以取得一定的检索性能提升,但其中一些方法引入了额外的外部模型,如人体姿势估计、语义分割或自然语言工具包,带来了额外的计算开销和不确定性,并且使网络无法进行端到端的学习。局部特征匹配方法需要提取存储图像或文本全局特征和多个局部特征,检索时需计算全局和局部相似度,消耗了额外存储空间和计算时间,不利于模型的实际部署。缺乏足够的训练数据是文本到图像行人重识别任务长期以来面临的一大挑战,一直缺乏有效的解决方案。为了减轻缺乏数据造成的影响,现有方法通常使用在单一模态大规模数据集如 ImageNet(Deng等, 2009)上进行预训练的主干网络来增强模型的特征提取能力,例如 ResNet(residual network)(He等, 2016)和 BERT(bidirectional encoder representations from transformers)(Devlin等, 2019),但是这种基于单模态数据预训练的主干网络只能学习到图像或文本两个模态内部的信息,无法学习到图像文本跨模态匹配和语义对齐的信息。

针对上述文本到图像行人重识别领域面对的挑战,本文提出了一种仅使用全局特征的基于 Transformer(Vaswani等, 2017)的双流文本到图像行人重

识别框架。由于基于 CNN 的算法无法挖掘长距离的关系,且会因为下采样操作丢失细粒度信息,因此本文算法采用 Transformer 作为视觉骨干网络。为了解决文本到图像行人重识别任务中缺少高质量标注数据的问题,本文使用了 CLIP(contrastive language-image pre-training)(Radford等, 2021)模型权重对主干网络参数进行初始化。CLIP 是一个在包含 4 亿个图像文本对的 WIT(WebImageText)数据集上训练得到的视觉语言预训练模型。相比使用在单一模态的大规模数据集上进行预训练的主干网络,使用 CLIP 能够利用预训练模型的图像文本跨模态匹配能力,从而获得更好的图像和文本语义表征。本文的主要贡献如下:1)提出一个端到端的双流 Transformer 网络来学习文本到图像行人重识别任务中行人图像和描述文本的表征,仅使用全局特征就可以超过目前使用全局特征+局部特征的最先进算法。2)设计了一个温度缩放跨模态投影匹配(temperature-scaled cross modal projection matching, TCMPM)损失函数。在 TCMPM 中,温度参数 τ 通过控制 softmax 函数内跨模态投影数值的分布,使模型更新的重点聚焦到难分负样本上,并对它们施加相应的惩罚,从而使模型学习到更有区分度的图像文本特征。本文将温度参数 τ 设置为一个可学习参数,直接通过训练阶段进行更新,避免将其视为一个超参数进行手动调参。3)本文方法在 CUHK-PEDES(CUHK person description)(Li等, 2017b)和 ICFG-PEDES(identity-centric and fine-grained person description)(Ding等, 2021)数据集进行了大量实验,文本到图像的检索结果大幅领先目前的 SOTA(state-of-the-art)模型。

1 相关工作

1.1 文本到图像行人重识别

Li 等人(2017b)第 1 个提出文本到图像行人重识别这个任务,并提出了 GNA-RNN(recurrent neural network with gated neural attention)框架,其包含了一个用于提取视觉特征的 VGG-16(Visual Geometry Group 16-layer network)网络和一个用于提取文本特征的 LSTM 网络。Li 等人(2017a)又提出了一个身份可知的两阶段网络架构,通过两阶段网络联合最小化图像和文本特征的身份间距离和跨模态距离。Zheng 等人(2020)提出了一个双路视觉 CNN+ 文本

CNN网络架构,使用了标准的Ranking损失函数和提出的Instance损失函数来联合优化图像和文本特征。Zhang和Lu(2018)提出了一个跨模态投影匹配损失和一个跨模态投影分类损失。Sarafianos等人(2019)提出通过对抗性学习在一个共享空间中学习模态不变的视觉和文本表征。Han等人(2021)设计了一个迁移学习方法从大规模粗粒度通用图像文本数据集预训练的CLIP模型中迁移知识,并设计了一个跨模态动量对比学习框架,更好地利用数据量有限的文本到图像行人重识别数据集。

还有一些方法采用了局部特征对齐的方法为全局特征对齐提供补充信息,学习图像区域与文本描述中的短语或单词之间的语义相关性。Aggarwal等人(2020)从行人文本描述中自动抽取属性词作为行人图像属性识别的类别,通过学习到的属性空间局部特征和潜在空间全局特征共同完成检索,提出了一个基于姿态的多粒度注意网络。Jing等人(2020)使用人类姿势估计工具分割出行人图像的人体局部特征,提出了一个由姿势指导的图像文本特征对齐网络。Wang等人(2020)同时使用了自然语言分词和语义分割工具提取文本和图像局部特征,设计了一个跨模态对齐网络来匹配图像文本全局特征和属性局部特征。Chen等人(2021)使用了在基于图像的行人重识别方法中流行的PCB(part-based convolutional baseline)(Sun等,2018)策略进行局部特征提取,提出了一个多阶段跨模态匹配策略对双流网络中低层特征和高层特征进行多模态匹配,最后检索时只使用全局图像文本特征进行相似度计算。Ding等人(2021)收集了一个新的文本到图像行人重识别数据集ICFG-PEDES,基于非局部注意力机制提出了一个多视角非局部网络,学习人体各个部位之间的关系,检索时使用局部特征、非局部特征和全局特征一起计算相似度。

国内最早出现的文本到图像行人重识别研究是李晟嘉(2019)提出的基于注意力机制的跨模态融合行人检索DSFA-Net(description-strengthened and fusion-attention network)。随后,霍昶伟(2020)提出了基于截断式注意力机制和堆叠损失函数的网络框架。王玉煜(2020)提出了一个多连接分类损失函数来优化模型。张鹏(2021)专注于设计有效的损失函数,提出了基于乘性角度余量的损失函数和基于样本对权重赋值的损失函数。殷雪朦(2021)为了实现

细粒度的特征匹配,采用局部特征对齐方法,提出了一个结合全局和局部特征匹配的文本到图像行人重识别网络框架。

这些局部匹配方法有的需要使用额外的外部模型如人类姿势估计、语义分割或自然语言分词,带来了大量的计算复杂度且无法实现端到端训练。这些模型都采取了粗粒度全局特征加多个细粒度局部特征匹配的做法,大部分在实际检索阶段需要重复计算多次图像文本局部特征之间的相似度,这些方法由于过高的复杂度无法满足现实场景的要求。Han等人(2021)虽然设计了一个迁移学习和动量对比学习方法,从CLIP预训练模型中转移跨模态对齐的知识,仅使用全局特征就超过现有的局部特征对齐方法,但是没有充分利用预训练模型的强大跨模态匹配知识。相比而言,本文提出的双路Transformer网络方法简单又有效,并且能够进行端到端的训练,在检索时仅需计算全局特征余弦相似度即可快速从文本检索行人图像。

1.2 视觉语言预训练

基于Transformer的大规模预训练语言模型BERT在自然语言处理(natural language processing, NLP)领域取得了巨大成功。而在计算机视觉领域,预训练的大模型如ResNet和ViT(vision Transformer)(Dosovitskiy等,2021)也已广泛用做图像分类和分割等任务的骨干网络。受到NLP和计算机视觉中大规模预训练模型取得成功的启发,近年来视觉语言预训练(vision language pre-training, VLP)在多模态任务中成为了学习多模态表征的主流。在图文检索(Kiros等,2014)和视觉回答(Antol等,2015)等任务上取得了巨大成功。VLP模型结构可以分为单流结构和双流结构两种。单流模型采取单个Transformer同时处理多模态输入,使用注意力机制来融合多模态特征。代表性工作有Oscar(Li等,2020)和UNITER(universal image-text representation)(Chen等,2020)等。双流模型使用单独的两个Transformer分别提取图像和文本特征,两者之间权重不共享。虽然单流模型效果很好,但是单流模型的跨模态注意力机制在训练和推理时都不可避免地需要大量的计算时间。而双流模型所提取的图像文本特征在推理阶段仅需要简单的点乘计算余弦相似度,所以双流模型更适合图文检索任务。

本文方法在文本到图像行人重识别中引入了双

流结构模型,在大规模图像文本对数据集上使用对比学习预训练的 CLIP 模型,并设计了一个新的对比学习损失函数,使模型学到更强大的跨模态图像文本特征对齐能力,极大地减少了图像文本间的跨模态差异。

2 本文方法

本文提出了一个基于双流 Transformer 的文本到图像行人重识别方法网络模型,整体框架如图 2 所示,实现了端到端的方式同时学习视觉和文本表

征,且保证文本检索图像时的检索速度。与其他行人重识别方法一样,本文框架也由特征提取和监督学习两个阶段组成。特征提取阶段包括一个用于处理图像输入的视觉 Transformer 和一个用于处理文本描述句子的文本 Transformer。整个网络直接接受数据集中原始图像和文本描述对 $\{I_i, T_i\}_{i=1}^N$ 作为输入,其中, N 表示小批量中图像文本对的个数,无需使用其他现有的图像分割模型或 BRRT 模型对图像或文本进行预处理。相比现有方法,端到端的微调学习策略在学习跨模态对齐关系时更有优势。

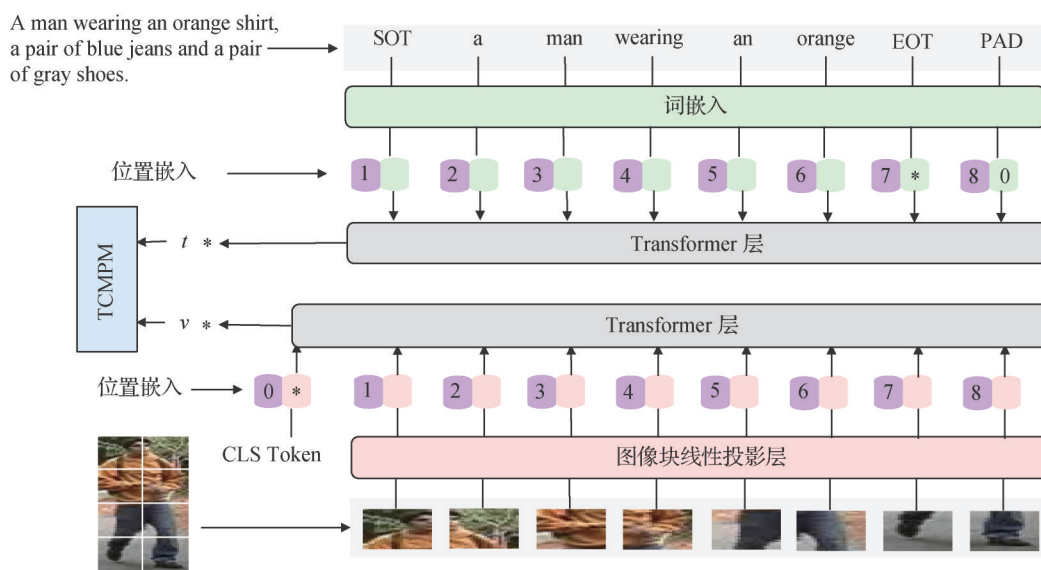


图 2 基于双流 Transformer 的文本到图像行人重识别方法框架图

Fig. 2 The framework of the proposed dual-stream Transformer baseline

2.1 视觉表征学习

本文采用了视觉 Transformer (ViT) 作为图像特征提取器。如图 2 所示,给定一幅行人图像 $I \in \mathbb{R}^{H \times W \times C}$, 其中 H, W, C 分别表示图像的高度、宽度和通道数,将图像均分成 N 个 16×16 像素的不重叠的图像块 $\{I_p^i | i = 1, 2, \dots, N\}$ 。之后将 N 个图像块通过一个线性投影层投影得到 N 个 D 维的图像块向量,并且在输入图像块向量序列前插入一个可学习的 [CLS] 嵌入向量。为了学习到各图像块向量之间的相对位置关系,输入图像块向量序列会加上一个位置嵌入 $P \in \mathbb{R}^{(N+1) \times D}$ 。最终输入到 ViT 网络的图像块向量序列可以表示为

$$f_v^0 = [I_{\text{cls}}; F(I_p^1); F(I_p^2); \dots; F(I_p^N)] + P \quad (1)$$

式中, f_v^0 表示图像输入序列的嵌入向量, F 表示将小

块图像映射到 D 维向量的线性投影层。ViT 网络的最后一层输出的图像特征嵌入向量可以表示为 $f_v = \{f_v^{\text{cls}}, f_v^1, f_v^2, \dots, f_v^N\}$ 。输出的 [CLS] 嵌入向量代表 f_v^{cls} 会作为行人图像的全局特征 f_v 。

由于文本到图像行人重识别任务所使用数据集的图像分辨率 and WIT 数据集上原始图像的分辨率不匹配,预训练的 CLIP 的 ViT 模型中的位置嵌入不能直接导入。本文采用了 TransReID (Transformer-based object re-identification) (He 等, 2021) 中使用的线性 2D 插值法来调整位置嵌入的尺寸以适应不同的图像分辨率。

2.2 文本表征学习

文本特征提取器是一个由 Radford 等人 (2019) 修改过的 12 层文本 Transformer。与 CLIP 中的文本特征提取器保持一致,输入的文本描述 T , 采用了词

库为 49 152 的小写字节对编码(byte pair encoding, BPE)(Sennrich 等, 2016)进行分词和编码处理。如图 2 所示,分词后的文本描述在开头和结尾分别插入了 [SOS] 和 [EOS] 嵌入向量来标识文本描述句子的开头和结尾。且为了保证计算效率,最大的文本描述序列长度设置为 77。为了学习句子中的单词相对位置关系,位置嵌入 $P \in \mathbf{R}^{77 \times D}$ 也会加到单词向量输入序列中。最终输入到文本 Transformer 的序列可以表示为

$$\mathbf{f}_t^0 = [\mathbf{E}(\mathbf{T}_{\text{sos}}); \mathbf{E}(\mathbf{T}_1); \mathbf{E}(\mathbf{T}_2); \cdots; \mathbf{E}(\mathbf{T}_{\text{eos}})] + \mathbf{P} \quad (2)$$

式中, $\mathbf{f}_t^0$ 表示文本输入序列的嵌入向量, $\mathbf{E}$ 表示将单词向量映射到 D 维向量的投影层。网络的最后一层输出的文本特征嵌入向量可以表示为 $\mathbf{f}_t = \{\mathbf{f}_t^{\text{sos}}, \mathbf{f}_t^1, \mathbf{f}_t^2, \cdots, \mathbf{f}_t^{\text{eos}}\}$ 。输出的 [SOS] 嵌入向量 $\mathbf{f}_t^{\text{eos}}$ 会作为文本描述的全局特征 $\mathbf{f}_t$ 与 $\mathbf{f}_v$ 进行跨模态匹配以最小化二者之间的距离。

2.3 损失函数设计

为了最小化文本模态到图像模态之间的距离,受到跨模态投影匹配(cross-modal projection matching, CMPM)损失(Zhang 和 Lu, 2018)和 InfoNCE (information noise contrastive estimation) (van den Oord 等, 2019)损失函数的启发,本文设计了一个温度缩放跨模态投影匹配(temperature-scaled cross modal projection matching, TCMPM)损失函数。给定一个由 N 个行人图像和描述文本表征向量对组成的小批量,对每个图像表征 $\mathbf{v}_i$,这个小批量可以表示为 $\{(\mathbf{v}_i, \mathbf{t}_j), \mathbf{y}_{i,j}\}_{i,j=1}^N$, 其中 $\mathbf{y}_{i,j} = 1$ 表示 $\mathbf{v}_i$ 和 $\mathbf{t}_j$ 属于同一个行人, $\mathbf{y}_{i,j} = 0$ 表示这个图像文本对不匹配。 $\mathbf{v}_i$ 和 $\mathbf{t}_j$ 匹配的概率可以定义为

$$p_{i,j} = \frac{\exp\left(\frac{\mathbf{v}_i^T \bar{\mathbf{t}}_j}{\tau}\right)}{\sum_{k=1}^N \exp\left(\frac{\mathbf{v}_i^T \bar{\mathbf{t}}_k}{\tau}\right)} \quad (3)$$

式中, $\bar{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|}$ 是归一化后的文本表征向量, τ 是一个可学习的温度参数,初始值设为 0.07。 $\mathbf{v}_i^T \bar{\mathbf{t}}_j$ 在几何上代表图像表征向量在文本表征向量上的投影标量值。 $p_{i,j}$ 可以视为 $(\mathbf{v}_i, \mathbf{t}_j)$ 的投影值在小批量中所有图像对文本 $\{(\mathbf{v}_i, \mathbf{t}_j)\}_{j=1}^N$ 投影值中的比率。 τ 用来控制 $p_{i,j}$ 的取值,起到对难分负样本进行惩罚的作用。在一个小批量中,匹配的正样本个数可能不止一个,

本文对于真实匹配概率采用了 softmax 函数进行归一化处理,真实匹配概率表示为

$$q_{i,j} = \frac{\exp\left(\frac{\mathbf{y}_{i,j}}{\tau}\right)}{\sum_{k=1}^N \exp\left(\frac{\mathbf{y}_{i,k}}{\tau}\right)} \quad (4)$$

式中, $q_{i,j}$ 表示真实标签的分布概率, τ 是和上面计算图像文本对投影值比率中共用的温度参数。相比原始的 CMPM 损失函数使用的线性函数归一化方法,使用 softmax 函数这种非线性的归一化方法可以起到标签平滑的作用。

通过计算图像到文本的投影概率和真实匹配概率的 KL(Kullback-Leibler)散度,可以得到一个小批量的图像到文本的匹配损失函数,具体为

$$L_{\text{I2T}} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N p_{i,j} \log \frac{p_{i,j}}{q_{i,j} + \varepsilon} \quad (5)$$

式中, ε 是一个非常小的数用来防止出现数值溢出问题。通过最小化 $KL(p_i \| q_i)$,可以使 p_i 分布曲线的形状接近 q_i 分布曲线的形状,从而使匹配的图像文本对的投影值最大,不匹配的图像文本对投影值最小。

图像到文本的匹配损失函数 L_{I2T} 在图像模态到文本模态的这个方向上拉近视觉表征与其匹配的文本表征之间的距离。在一般的图像文本表征学习过程中,图像文本匹配函数需要考虑两个方向。由于损失函数的对称性,通过计算文本表征到图像表征上的投影,拉近文本表征与其匹配的视觉表征之间距离的文本到图像匹配损失 L_{T2I} 可以通过将式(3)中的 $\mathbf{v}$ 和 $\mathbf{t}$ 交换后代入式(5)进行计算。最后,双向的 TCMPM 损失函数可以表示为

$$L_{\text{TCMPM}} = L_{\text{I2T}} + L_{\text{T2I}} \quad (6)$$

3 实验

3.1 数据集

CUHK-PEDES 数据集包含 13 003 个行人的 40 206 幅图像和 80 440 个文本描述,大部分图像都有两个文本描述。每个文本描述中平均单词长度为 23.5,最大单词长度为 96,最小单词长度为 12。数据集划分为训练集、验证集和测试集。训练集包括 11 003 个行人的 34 054 幅图像和 68 126 个文本描述。测试集包含 1 000 个行人的 3 074 幅图像和

6 156个文本描述,剩下的属于验证集。测试集的文本描述和行人图像分别作为查询集和图像库。

ICFG-PEDES数据集包含4 102个行人的54 522幅图像描述文本对,即每幅图像有一个对应的文本描述。ICFG-PEDES中所有的图像来源于MSMT17 (multi-scene multi-time) (Wei等, 2018)数据集。数据集划分为训练集和测试集。训练集包含3 102个行人的34 674个图像文本对,测试集包含剩下的1 000个行人的19 848个图像文本对。

3.2 评估指标

本文遵循文本到图像行人重识别的标准评估指标,通过Rank- K ($K = 1, 5, 10$)准确率来评估模型性能。具体来说,给定一个文本描述的查询,计算图库中所有图像与这个文本描述的相似度,然后根据相似度值进行排序。Rank- K 是指真实匹配图像在相似度排序表中出现在前 K 个的概率。

3.3 实现细节

在实验中,遵循文本到图像行人重识别工作的实验设置(Han等, 2021),所有的行人图像分辨率都重新调整到 384×128 像素,并且在训练阶段使用了随机水平翻转、随机裁剪和随机擦除的图像数据增广方法。最大文本长度设置为77。视觉表征向量 f_v 与文本表征向量 f_t 的维度为512。使用Adam优化器进行模型优化,基础学习率为 $1E-5$,权重衰减为 $4E-5$ 。模型训练60轮,在开始的5轮中采用warm-up策略使学习率从 $1E-6$ 线性增长到 $1E-5$ 。学习率在20~50轮中每10轮衰减为原值的0.1。 L_{TCMPM} 中的 τ 和 ε 的初始值分别设置为0.07和 $1E-8$ 。

所有实验均在单张NVIDIA GeForce RTX 3090 GPU上进行,使用了Pytorch深度学习平台进行半精度训练。视觉Transformer和文本Transformer的初始化使用了CLIP(ViT-B/16)预训练模型。

3.4 实验结果分析

3.4.1 定量分析

表1和表2给出了本文算法和现有的文本到图像行人重识别算法的在CUHK-PEDES数据集和ICFG-PEDES数据集上的实验结果。表中列出的所有算法结果都没有包括使用重排序在内的后处理方法。所有的仅使用全局特征匹配的方法在表格的类型列中标注为“G”,使用了局部特征匹配的方法则标注为“L”。此外,表中特征维度标注了算法所使

用的特征维度大小。

从表1可以看出,文本到图像行人重识别的关注重点在于使用局部特征匹配,并且取得了相对较好的性能,但是也可以看出局部特征匹配方法已经处于瓶颈期。在CUHK-PEDES数据集上,本文算法大幅超过了现有的文本到图像行人重识别方法,特别是对于最重要的指标Rank-1来说,本文算法相比现有最好的局部匹配方法和全局匹配方法分别提升了5.92%和7.09%。在ICFG-PEDES数据集上,本文算法也取得了同样的检索性能提升。相比目前最好的局部特征匹配方法提升了1.21%。

除了优越的检索性能,本文算法由于仅使用了全局特征匹配和较小的特征维度,相比现有的局部特征匹配方法,在训练效率、检索效率和离线特征存储方面均有巨大优势。

3.4.2 消融实验

为了验证本文提出的温度缩放跨模态投影匹配(TCMPM)损失函数的有效性,在CUHK-PEDES和ICFG-PEDES数据集上进行验证TCMPM有效性的消融实验,对比完整的TCMPM、原始的CMPM、原始的CMPM + CMPC (cross-modal projection classification)、在原始的CMPM中的softmax中加入温度参数、多模态自监督学习中常使用的infoNCE这5组损失函数,并且加入了不使用损失函数的CLIP预训练模型做零样本学习(无损失函数)的对照组。

表3给出了在CUHK-PEDES和ICFG-PEDES数据集上进行消融实验的结果。可以看出,相比原始的CMPM损失函数,使用其他3种加入温度参数的损失函数可以显著提升模型的跨模态特征匹配能力,从而大幅增强文本到图像的检索能力。为了更好地分析温度参数的作用,图3给出了可学习的温度参数随着训练轮数的变化曲线。可以看出,温度参数 τ 从初始值0.07开始的20轮中快速增长并在20轮之后趋于稳定。

CLIP零样本在CUHK-PEDES和ICFG-PEDES数据集上的文本到图像检索结果不理想,但是经过使用和CLIP训练时一样的infoNCE损失函数进行微调之后的效果已经超过了现有的所有文本到图像行人重识别方法,说明了文本到图像行人重识别这种细粒度的检索任务也可以有效地从CLIP这种通用图像文本跨模态大模型中迁移图像文本跨模态对齐知识。

为了验证本文算法在不同特征提取能力的视觉

表 1 本文算法与其他文本到图像行人重识别算法在 CUHK-PEDES 数据集上的实验结果对比
Table 1 Comparison with state-of-the-art text-to-image Re-ID methods on CUHK-PEDES dataset

算法	类型	文献来源	视觉网络	文本网络	特征 维度	Rank-1/%	Rank-5/%	Rank-10/%
GNA-RNN(Li 等,2017b)	G	CVPR17	VGG	LSTM	512	19.05	–	53.64
Dual Path(Zheng 等,2020)	G	TOMM20	ResNet	ResNet	2 048	44.40	66.26	75.07
CMPM/C(Zhang 和 Lu,2018)	G	ECCV18	ResNet	LSTM	512	49.37	71.69	79.27
MIA(Niu 等,2020)	L	TIP20	VGG	GRU	1 024	53.10	75.00	82.90
PMA(Jing 等,2020)	L	AAAI20	ResNet	LSTM	1 024	54.12	75.45	82.97
TIMAM(Sarafianos 等,2020)	G	ICCV19	ResNet	BERT	512	54.51	77.56	84.78
CMKA(Chen 等,2021)	G	TIP21	ResNet	LSTM	512	54.69	73.65	81.86
ViTAA(Wang 等,2020)	L	ECCV20	ResNet	LSTM	256	55.97	75.84	83.52
CMAAM(Aggarwal 等,2020)	L	WACV20	Mobilenet	LSTM	512	56.68	77.18	84.86
NAFS(Gao 等,2021)	L	arXiv21	ResNet	BERT	768	59.94	79.86	86.70
NAFS+LapsCore(Wu 等,2021)	L	ICCV21	ResNet	BERT	768	63.40	–	87.80
SSAN(Ding 等,2021)	L	arXiv21	ResNet	LSTM	2 048	61.37	80.15	86.73
Han 等人(2021)	G	BMCV21	CLIP ResNet	CLIP-TE	256	64.08	81.73	88.19
TIPCB(Chen 等,2022)	L	Neurocomputing22	ResNet	BERT	2 048	64.26	83.19	89.10
Li 等人(2022)	L	ICASSP22	ViT	BERT	768	64.13	82.62	88.40
LGUR(Shao 等,2022)	L	ACM MM'22	DeiT-Small	BERT	512	65.25	83.12	89.00
本文	G	本文	CLIP ViT	CLIP-TE	512	71.17	87.95	93.00

注:加粗字体表示各列最优结果。GNA-RNN: recurrent neural network with gated neural attention, CMPM/C: cross-modal projection matching/classification, MIA: multi-granularity image-text alignments model, PMA: pose-guided multi-granularity attention network, TIMAM: text-image modality adversarial matching, CMKA: cross-modal knowledge adaptation, ViTAA: visual-textual attributes alignment model, CMAAM: cross modal attribute aided matching, NAFS: non-local alignment over full-scale representations, LapsCore: language-guided person search based on color reasoning, SSAN: semantically self-aligned network, TIPCB: text-image part-based convolutional baseline, LGUR: learn granularity-unified representations。CLIP-TE代表CLIP文本编码器,“–”表示论文作者未报告数据。

表 2 本文算法与其他文本到图像行人重识别算法在 ICFG-PEDES 数据集上的实验结果对比
Table 2 Comparison with state-of-the-art text-to-image Re-ID methods on ICFG-PEDES dataset

算法	类型	文献来源	视觉网络	文本网络	特征 维度	Rank-1/%	Rank-5/%	Rank-10/%
Dual Path(Zheng 等,2020)	G	TOMM20	ResNet	ResNet	2 048	38.99	59.44	68.41
CMPM/C(Zhang 和 Lu,2018)	G	ECCV18	ResNet	LSTM	512	43.51	65.44	74.26
MIA(Niu 等,2020)	L	TIP20	VGG	GRU	1 024	46.49	67.14	75.18
ViTAA(Wang 等,2020)	L	ECCV20	ResNet	LSTM	256	50.98	68.79	75.78
SSAN(Ding 等,2021)	L	arXiv21	ResNet	LSTM	2 048	54.23	72.63	79.53
TIPCB(Chen 等,2022)	L	Neurocomputing22	ResNet	BERT	2 048	54.96	74.72	81.89
LGUR(Shao 等,2022)	L	ACM MM'22	DeiT-Small	BERT	512	59.02	75.32	81.56
本文	G	本文	CLIP ViT	CLIP-TE	512	60.23	78.09	84.23

注:加粗字体表示各列最优结果,CLIP-TE代表CLIP文本编码器。

表3 不同损失函数的消融实验结果对比
Table 3 Comparison results of different loss ablation studies

实验设置	CUHK-PEDES 数据集			ICFG-PEDES 数据集		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
CLIP 零样本(无损失函数)	9.57	21.56	29.09	3.85	10.70	16.17
CMPM	61.63	81.12	87.67	55.86	73.73	80.22
CMPM+CMPC	61.47	80.85	87.77	56.31	74.01	80.24
InfoNCE	67.79	85.51	91.16	59.81	77.02	83.13
CMPM+温度参数	70.87	87.48	92.67	59.32	77.59	93.94
TCMPM	71.17	87.95	93.00	60.23	78.09	84.23

注:加粗字体表示各列最优结果。

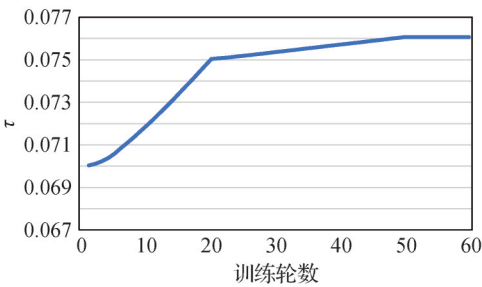


图3 温度参数 τ 随训练轮数变化图
Fig. 3 The temperature τ curve with training epoch

骨干网络中的效果,在 CUHK-PEDES 数据集上进行 CLIP ResNet50, CLIP ResNet101, CLIP ViT-B/32, CLIP ViT-B/16 这 4 个特征提取能力越来越强的视觉骨干网络的消融实验,结果如表 4 所示。可以看出,在使用同等特征提取能力的视觉骨干网络时,本文算法已经超过 Han 等人(2021)方法,并且在方法简洁性上更是远超目前所有文本到图像行人重识别方

表4 不同视觉骨干网络消融实验结果对比
Table 4 Comparison results of ablation studies

方法	视觉骨干网络	CUHK-PEDES 数据集		
		Rank-1	Rank-5	Rank-10
Han 等人(2021)	CLIP RN50	61.65	80.98	86.78
Han 等人(2021)	CLIP RN101	64.08	81.73	88.19
本文	CLIP RN50	60.88	81.69	88.19
本文	CLIP RN101	64.68	84.08	89.62
本文	CLIP ViT-B/32	68.50	86.70	91.85
本文	CLIP ViT-B/16	71.17	87.95	93.00

注:加粗字体表示各列最优结果。

法。随着视觉骨干网络的特征提取能力增强,本文算法在文本到图像检索的 Rank-1, Rank-5, Rank-10 值都有提升。

3.4.3 定性分析

图 4 展示了本文算法在 CUHK-PEDES 测试集上文本到图像检索的可视化结果。图 4 左边为输入的文本描述,右边为根据相似度从高到低排序检索出的行人图像。黑色边框为真实标签图像,绿色边框为检索出的正确匹配的行人图像,红色边框为检索出的不正确匹配的行人图像。从图 4 第 1 行的结果可以看出,本文算法成功检索出了满足“black and white striped top”、“khaki pants”和“black backpack”条件的行人图像,且在检索出所有的正确匹配行人后,检索出了满足这些条件的其他行人的图像。第 2 行的检索结果也满足了其文本查询的“tan coat with a hood”、“black backpack”、“jeans”和“white sneakers”条件,成功检索出了正确匹配的行人图像。证明了本文算法可以利用文本描述中的细粒度信息检索出符合描述的目标行人图像。第 3、4 行是同一个行人图像标注的不同的两个文本描述的检索结果。第 3 行是一个失败检索案例,从第 3 行的文本描述的属性可以看出,描述都是很粗粒度的形容短语,文本描述过于模糊。第 3 行的检索结果虽然都符合文本描述,但是文本描述不具备独特性,无法准确地匹配目标行人图像。而第 4 行的文本描述标注相比第 3 行有着更细粒度的属性信息,使用第 4 行的文本描述进行检索,能够检索出正确匹配的行人图像。从图 4 可以得到两个结论:1)本文算法可以区分给定文本查询中的细粒度属性信息并匹配到具体的图

The man is stepping up onto a curb. He is wearing a black and white striped top, khaki pants and black shoes. He is also wearing a black backpack.



The woman is wearing a tan coat with a hood, carrying a black backpack, wearing jeans and white sneakers.



He has short black hair. He is wearing a black shirt and a backpack as well.



The boy is wearing a long-sleeved black shirt with a collar and jeans. He is carrying a dark-colored backpack over both shoulders.



图 4 本文提出的文本到图像行人重识别算法检索结果可视化图

Fig. 4 Visualization of text-to-image person re-identification results by our method

像信息。2) 当文本描述信息比较模糊时, 本文方法仍然可以给出符合文本描述的结果。

4 结 论

本文提出一种新的文本到图像行人重识别方法, 从通用图像文本对数据集大规模预训练的 CLIP 模型中迁移通用跨模态特征对齐能力, 弥补了文本到图像行人重识别可使用数据不足的问题。同时本文融合了 CMPM 和 infoNCE 损失函数的优点, 提出了 TCMPM 损失函数来更好地对齐图像文本特征到共有特征空间。本文算法在两个公开的平均数据集上与现有的先进方法 LGUR (learning granularity-unified representations) 相比都有明显提升, 在 CUHK-PEDES 和 ICFG-PEDES 数据集上的 Rank-1 值分别提升了 5.92% 和 1.21%。此外, 最近的文本到图像行人重识别方法都采取了手动分割多个局部图像和文本特征进行多层次匹配的做法, 这种做法虽然会带来一定的性能提升, 但也会破坏模态内信息的完整性, 并且多层次匹配会降低检索速度。本文算法采用全局特征匹配, 仅采用强大的特征提取骨干网络和设计的 TCMPM 损失函数, 就超越了近年的局部匹配方法, 成为新的先进的文本到图像行人重

识别方法。在两个数据集上文本方法优秀的检索准确率充分证明了本文方法的有效性。值得注意的是, 本文工作采取 Transformer 作为骨干网络, 因此无法采用近期行人重识别领域流行的基于 CNN 的局部特征匹配策略, 未考虑在细粒度的跨模态特征上进行对齐。因此, 设计一种细粒度的基于 Transformer 局部特征匹配文本到图像行人重识别方法, 使模型可以更精准地对齐图像和文本模态之间的细粒度特征, 将是文本到图像行人重识别的主要研究方向之一。

参考文献 (References)

- Aggarwal S, Radhakrishnan V B and Chakraborty A. 2020. Text-based person search via attribute-aided matching//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 2606-2614 [DOI: 10.1109/wacv45572.2020.9093640]
- Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick C L and Parikh D. 2015. VQA: visual question answering//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/iccv.2015.279]
- Chen Y C, Huang R, Chang H, Tan C Q, Xue T and Ma B P. 2021. Cross-modal knowledge adaptation for language-based person search. IEEE Transactions on Image Processing, 30: 4057-4069

- [DOI: 10.1109/tip.2021.3068825]
- Chen Y C, Li L J, Yu L C, El Kholy A, Ahmed F, Gan Z, Cheng Y and Liu J J. 2020. UNITER: universal image-text representation learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 104-120 [DOI: 10.1007/978-3-030-58577-8_7]
- Chen Y H, Zhang G Q, Lu Y J, Wang Z X and Zheng Y H. 2022. TIPCB: a simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171-181 [DOI: 10.1016/j.neucom.2022.04.081]
- Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/cvpr.2009.5206848]
- Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/1810.04805.pdf>
- Ding Z F, Ding C X, Shao Z Y and Tao D C. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/2107.12666.pdf>
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/2010.11929.pdf>
- Gao C Y, Cai G Y, Jiang X Y, Zheng F, Zhang J, Gong Y F, Peng P, Guo X W and Sun X. 2021. Contextual non-local alignment over full-scale representation for text-based person search [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/2101.03036.pdf>
- Han X, He S, Zhang L and Xiang T. 2021. Text-based person search with limited data [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/2110.10807.pdf>
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/cvpr.2016.90]
- He S T, Luo H, Wang P C, Wang F, Li H and Jiang W. 2021. TransReID: transformer-based object re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14993-15002 [DOI: 10.1109/iccv48922.2021.01474]
- Huo C W. 2020. Research on Person Re-Identification Algorithm Based on Natural Language Description. Tianjin: Hebei University of Technology (霍昶伟. 2020. 基于自然语言描述的行人再识别算法研究. 天津: 河北工业大学)
- Jing Y, Si C Y, Wang J B, Wang W, Wang L and Tan T N. 2020. Pose-guided multi-granularity attention network for text-based person search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 11189-11196 [DOI: 10.1609/aaai.v34i07.6777]
- Kiros R, Salakhutdinov R and Zemel R S. 2014. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/1411.2539.pdf>
- Li C J. 2019. Deep Attention Based Cross-Modal Person Search Via Natural Language Descriptions. Tianjin: Tianjin University (李晟嘉. 2019. 基于深度注意力的自然语言描述跨模态行人检索. 天津: 天津大学)
- Li S, Xiao T, Li H S, Yang W and Wang X G. 2017a. Identity-aware textual-visual matching with latent co-attention//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1908-1917 [DOI: 10.1109/iccv.2017.209]
- Li S, Xiao T, Li H S, Zhou B L, Yue D Y and Wang X G. 2017b. Person search with natural language description//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5187-5196 [DOI: 10.1109/cvpr.2017.551]
- Li S P, Cao M and Zhang M. 2022. Learning semantic-aligned feature representation for text-based person search//Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 2724-2728 [DOI: 10.1109/icassp43922.2022.9746846]
- Li X J, Yin X, Li C Y, Zhang P C, Hu X W, Zhang L, Wang L J, Hu H D, Dong L, Wei F R, Choi Y and Gao J F. 2020. Oscar: object-semantics aligned pre-training for vision-language tasks//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 121-137 [DOI: 10.1007/978-3-030-58577-8_8]
- Niu K, Huang Y, Ouyang W L and Wang L. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29: 5542-5556 [DOI: 10.1109/tip.2020.2984883]
- Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): #9
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/2103.00020.pdf>
- Sarafianos N, Xu X and Kakadiaris I. 2019. Adversarial representation learning for text-to-image matching//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5813-5823 [DOI: 10.1109/iccv.2019.00591]
- Sennrich R, Haddow B and Birch A. 2016. Neural machine translation of rare words with subword units [EB/OL]. [2022-05-21]. <https://arxiv.org/pdf/1508.07909.pdf>
- Shao Z Y, Zhang X Y, Fang M, Lin Z F, Wang J and Ding C X. 2022. Learning granularity-unified representations for text-to-image per-

- son re-identification [EB/OL]. [2022-05-21].
<https://arxiv.org/pdf/2207.07802.pdf>
- Sun Y F, Zheng L, Yang Y, Tian Q and Wang S J. 2018. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 501-518 [DOI: 10.1007/978-3-030-01225-0_30]
- van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2022-05-21].
<https://arxiv.org/pdf/1807.03748.pdf>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wang Y Y. 2020. Language-based Person Re-Identification. Dalian: Dalian University of Technology (王玉煜. 2020. 基于语言信息的行人重识别算法研究. 大连: 大连理工大学)
- Wang Z, Fang Z Y, Wang J and Yang Y Z. 2020. VITAA: visual-textual attributes alignment in person search by natural language//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 402-420 [DOI: 10.1007/978-3-030-58610-2_24]
- Wei L H, Zhang S L, Gao W and Tian Q. 2018. Person transfer GAN to bridge domain gap for person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 79-88 [DOI: 10.1109/cvpr.2018.00016]
- Wu Y S, Yan Z Z, Han X G, Li G B, Zou C Q and Cui S G. 2021. LapsCore: language-guided person search via color reasoning//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1604-1613 [DOI: 10.1109/iccv48922.2021.00165]
- Yin X M. 2021. Research on Cross-Modal Pedestrian Re-Recognition Technology Based on Natural Language. Chengdu: University of Electronic Science and Technology of China (殷雪朦. 2021. 基于自然语言的跨模态行人重识别技术研究. 成都: 电子科技大学)
- Zhang P. 2021. Research on Pedestrian Image Search Based on Natural Language Description. Chengdu: University of Electronic Science and Technology of China (张鹏. 2021. 基于自然语言描述的行人检索研究. 成都: 电子科技大学)
- Zhang Y and Lu H C. 2018. Deep cross-modal projection learning for image-text matching//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 707-723 [DOI: 10.1007/978-3-030-01246-5_42]
- Zheng X, Lin L, Ye M, Wang L and He C L. 2020. Improving person re-identification by attention and multi-attributes. Journal of Image and Graphics, 25(5): 936-945 (郑鑫, 林兰, 叶茂, 王丽, 贺春林. 2020. 结合注意力机制和多属性分类的行人再识别. 中国图象图形学报, 25(5): 936-945) [DOI: 10.11834/jig.190185]
- Zheng Z D, Zheng L, Garrett M, Yang Y, Xu M L and Shen Y D. 2020. Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(2): #51 [DOI: 10.1145/3383184]

作者简介

姜定,男,硕士研究生,主要研究方向为多媒体检索。

E-mail: jiangding@whu.edu.cn

叶茫,通信作者,男,教授,主要研究方向为计算机视觉、多媒体检索、安防大数据分析。E-mail: yemang@whu.edu.cn