一种基于可扩展表征学习的行人重识别方法

技术领域

本发明涉及多媒体信息检索领域，具体涉及一种基于可扩展表征学习的行人重识别方法。

背景技术

基于文本的行人重识别作为计算机视觉与自然语言处理交叉领域的关键技术，在智能安防、行人轨迹追踪等众多实际场景发挥着关键作用。其可通过文本描述检索不同场景下跨摄像头的行人图像，有力辅助安防人员快速锁定目标、构建人物行动轨迹，提升智能监控视频分析、检索的效率与精准度。

现有大多数研究都集中在探索图像-文本对在公共表征空间中的语义对齐，可以观察到这些方法在现有的公共基准数据集上使用的模型训练过程是基于一个强有力的假设：图像和文本的模态信息是完整的，并且数据集具有均匀分布，即模型可以提取图像和文本的有效特征表示，同时它们不存在偏差，这显然是不现实的。除此之外，目前多数基于语义表示的方法缺乏可扩展性，难以区分相似图像和有偏差的文本（如颜色偏差），从而导致行人重识别性能不佳。这主要源于两个方面：（1）首先，大量的跨模态行人重识别方法依赖于 CLIP 模型编码的特征，而这种与文本对齐的表征依赖于文本。也就是说，如果两张相似的行人图像都可以用相同的文本句子来描述，例如“这个男人穿着黄色衬衫，灰色短裤，网球鞋和背包”，那么它们的表征一定是相似的。（2）其次，当前使用的文本描述是存在偏差的，这可能会影响模型学习到的语义，例如根据某数据集词频统计，“黑色”占行人衣服颜色的比例为 38%，而“红色”只占 2%，类似分布不均的偏差会导致模型偏爱高频单词，从而影响模型的整体性能。因此，如何在建模过程中处理偏差，确保模型准确地捕获潜在的语义，并在多元场景中表现良好，仍有改进的空间，但很少被考虑。

因此，本发明针对上述问题，提出了一种基于可扩展表征学习的行人重识别方法。它由视觉表征学习模块、文本表征学习模块和行人检索模块组成。具体来说，视觉表征学习模块利用来自两个空间的 ViT 结构的视觉编码特征来增强图像判别性表征，并通过因果注意力机制进行表征融合。对于文本表征学习模块，通过示范样本消除数据集偏差是一种可靠的策略，同时采用因果注意力机制，消除数据集中存在的偏差，从而对跨模态输入进行因果干预，从而提高行人重识别的性能。这里使用具有 do 算子的结构因果模型，将数据集偏差视为混杂因

素，并随后消除它们的混杂效应。

发明内容

本发明提出了一种基于可扩展表征学习的行人重识别方法来解决上述问题，提升行人重识别的准确率。主要包括以下步骤：

步骤 S1：获取训练样本，每一个训练样本包括训练图像及与训练图像对应的训练文本描述；

步骤 S2：利用图像-文本编码器分别提取所述训练图像及训练文本描述的特征，得到训练图像特征和训练文本特征；

步骤 S3：利用非文本对齐的视觉编码器提取训练图像特征，优先考虑图像本身，在语义空间中有效地分离相似图像。然后通过向量余弦相似度获取与主干模型相对应的维度的图像以及文本特征，该特征即为示范样本特征；

步骤 S4：利用 do 算子的几何特征进行因果表征学习，它采用结构因果模型，将数据集偏差视为混杂因素，并随后消除它们的混杂效应；

步骤 S5：训练模型至收敛。然后将所述行人图像特征和文本特征输入训练后的行人重识别模型，得到行人重识别结果。

具体地，所述步骤 S2 包括：

通过 CLIP 模型的图像特征提取器和文本特征提取器分别从训练图像提取训练图像特征，对所述的训练文本描述提取训练文本特征；

所述步骤 S3 包括：

利用非文本对齐的视觉编码器 EVA 提取训练图像特征。利用 CLIP 的文本特征提取器提取示范样本特征；

具体地，所述步骤 S2 和步骤 S3 中，所述图像和文本特征提取器分别为 CLIP 模型的图像特征提取器和 CLIP 的文本特征提取器，以及 EVA 图像特征提取器。由于来自 CLIP 的与文本对齐的视觉图像编码器依赖于文本且难以区分相似图像，因此利用另一种同样具有 ViT 架构的非文本对齐的视觉编码器 EVA 作为语义空间的补充，并通过因果注意力机制来融合这两种表征，增强视觉特征的鲁棒性。

具体地，所述步骤 S4 的因果表征学习公式可以表示为：

$$\mathcal{P}(y \mid do(X)=x)=\sum_{z}\mathcal{P}(y \mid x,z)\mathcal{P}(z)$$

其中 X 是自变量（或原因），Y 是相关的因变量（或结果）；z 通常表示混杂因素；$P(y \mid do(X) = x)$ 是因果关系，表示在 $X = x$ 的干预下 Y 的分布，X 被设置为 x。

附图说明

为了能更进一步了解本发明的特征及技术内容，请参阅以下有关本发明的详细说明与附图，然而附图仅提供参考与说明用，并非用来对本发明加以限制。

图 1 为本发明实现的整个流程示意图；

图 2 为基于结构因果模型的干预因果示意图。

具体实施方式

为了更进一步阐释本发明所采取的技术手段及其效果，下面以智能安防场景应用及其附图进行详细描述。这些实施方式仅用于解释本发明的技术原理，并非限制本发明的应用范围。

请参阅图 1，本发明提供提出一种基于可扩展表征学习的行人重识别方法，具体步骤如下：

步骤 S1：获取训练样本，每一个训练样本包括训练图像及与训练图像对应的训练文本描述；

步骤 S2：利用图像-文本编码器分别提取所述训练图像及训练文本描述的特征，得到训练图像特征和训练文本特征；

具体地，所述步骤 S2 包括：

通过 CLIP 模型的图像特征提取器和文本特征提取器分别从训练图像提取训练图像特征，对所述的训练文本描述提取训练文本特征；

进一步地，CLIP 的图像编码器采用 ViT 架构。图像首先被划分为等大小的块，然后每个块被映射到一个具有固定维度的向量，以提取其嵌入。在与位置嵌入相结合后，图像嵌入输入到 ViT - L/14 网络架构中，以学习视觉上下文特征表示，该架构包含 12 个 Transformer 层。在一个批大小中采样 X 对，并将每个图像和文本的索引表示为 $i \in \{1,\ldots,N\}$，“N”是一个批大小中图像和文本的数量。

步骤 S3：利用非文本对齐的视觉编码器提取训练图像特征，优先考虑图像本身，在语义空间中有效地分离相似图像。然后通过向量余弦相似度获取与主干模型相对应的维度的图像以及文本特征，该特征即为示范样本特征；

具体地，所述步骤 S3 包括：

利用非文本对齐的视觉编码器 EVA 提取训练图像特征。利用 CLIP 的文本特征提取器提取示范样本特征。

具体地，所述步骤 S2 和步骤 S3 中，所述图像和文本特征提取器分别为 CLIP 模型的图像特征提取器和 CLIP 的文本特征提取器，以及 EVA 图像特征提取器。

进一步地，所述步骤 S3 包括：

由于来自 CLIP 的与文本对齐的视觉编码器依赖于文本且难以区分相似图像，因此，利用另一种同样具有 ViT 架构的非文本对齐的视觉编码器 EVA 作为语义空间的补充，从而获得用 EVA 编码的用于视觉表征的扩展嵌入。因此在这个模块中，分别获得图像特征 f_i^m 和扩展的视觉特征 $\tilde{f}_i^{\tilde{m}}$，它们可公式化为：

$$\begin{cases} f_i^m = V_{CLIP}\left(m_i\right) \\ \tilde{f}_i^{\tilde{m}} = V_{EVA}\left(m_i\right) \end{cases}$$

其中 $V_{CLIP}(\cdot)$ 和 $V_{EVA}(\cdot)$ 都是视觉投影函数。此外，需要考虑不同空间的 ViT 特征的适当融合策略，本发明采用上述所提及的因果注意力机制来融合这两种表征，增强视觉特征的鲁棒性。

更进一步地，所述步骤 S3 包括：

对于文本表示，采用 CLIP 模型的文本编码器作为骨干进行特征提取，分别获得文本特征 f_i^t 和示范样本特征 $\tilde{f}_i^{\tilde{t}}$，公式表示为：

$$\begin{cases} f_i^t = L(t_i) \\ \tilde{f}_i^{\tilde{t}} = L(\tilde{t}_i) \end{cases}$$

其中 $L(\cdot)$ 是文本投影函数。

步骤 S4：利用 do 算子的几何特征来进行因果表征学习，它采用结构因果模型将数据集偏差视为混杂因素，并随后消除它们的混杂效应；

具体地，所述步骤 S4 包括：

请参阅图 2，具体展示了因果图：Q（文本查询），I（图像），Y（预测）和 C（混杂因素，即数据集偏差），其中节点和边分别表示内生变量和因果关系。因果表征学习公式可以表示为：

$$\mathcal{P}(y \mid do(X) = x) = \sum_z \mathcal{P}(y \mid x, z)\mathcal{P}(z)$$

其中 X 是自变量（或原因），Y 是相关的因变量（或结果）。z 通常表示混杂因素。$P(y \mid do(X)=x)$ 是因果关系，表示在 $X=x$ 的干预下 Y 的分布，X 被设置为 x。

$Q,I \to Y$：在大多数方法中，模型使用文本描述 Q 和图像 I 来产生预测 Y。通常这个预测过程被认为是有偏差的，因为它忽略了数据集偏差 C，这是一个需要考虑的因素。

$C \to Q,I,Y$：数据集偏差表示为混杂因素 C，它可能导致文本输入和目标之间的虚假相关性。偏差在不同的数据集中通常会有所不同，例如空间偏差和时间偏差等，其中大多数偏差难以量化且通常不可观察。

基于这些，这里引入具有因果注意力机制的去混杂可扩展表征学习框架，用于对输入进行因果干预，并使用因果 do-算子技术推导无偏估计公式。

由于跨模态行人重识别任务主要利用文本检索图库中行人图像，所以来自 CLIP 模型的特征仍然需要处于主导地位，因此应用因果注意力进行特征交互采用一个单向过程，如下所示：

$$\mathcal{M}[i,j]=\begin{cases}0, & 0<i<\mathbb{M}, \mathbb{M}\le j<(\mathbb{M}+\mathbb{S}) \\ 1, & \text{others}\end{cases}$$

其中 M 是由 CLIP 模型图像编码器提取的视觉表征，S 是由非文本对齐的视觉编码器 EVA 编码的视觉表征。此外，i 和 j 是行和列索引，$M[i,j]$ 表示第 i 个图像在 Transformer 层中可以关注到第 j 个图像的表示。最后，所得到的双重视觉表示 $f_i^{'m}$ 的表达式为：

$$f_i^{'m}=[f_i^m, \tilde{f}_i^{\tilde{m}}]$$

此外，行人重识别中的数据集偏差（例如颜色信息）作为不可观察的混杂因素，通过利用类比推理来促进去混杂学习并减轻数据集偏差，从而在不明确建模混杂因素的情况下实现无偏估计。特别地，这个过程是有偏差的，因为它没有考虑由混杂因素 C 引起的虚假相关性。查询样本的表示和目标可以通过相似样本的输出来近似。这表明，为了计算对 Y 的因果关系，模型输入应该包含一个查询样本和相似的示范样本。因此，示范样本应该关注自身的视觉-文本信息，以引导查询样本进行类比学习，这可以通过因果注意力机制来表示。

类似地，文本表征学习进行因果注意力机制如下：

$$\mathcal{W}[i,j]=\begin{cases}0, & 0<i<\mathbb{D}, \mathbb{D}\le j<(\mathbb{D}+\mathbb{T}) \\ 1, & \text{others}\end{cases}$$

其中$\mathbb{T}$和$\mathbb{D}$是由 CLIP 文本编码器编码的文本表示。此外，i 和 j 是行和列索引，$W[i,j]$表示第 i 个文本在 Transformer 层中可以关注到第 j 个文本的表示。

步骤 S5：训练模型至收敛。然后将所述行人图像特征和文本特征输入训练后的行人重识别模型，得到行人重识别结果；

具体地，所述步骤 S5 包括：

使用相似性学习来确定特征向量是否属于同一个人。主要采用欧几里得距离来计算查询信息和图库图像之间的距离得分。其中，正样本对的行人图像将比负样本对产生更高的得分。

$$sim(\mathbb{M}_i,\mathbb{T}_i)=\mathbb{M}_i\cdot\mathbb{T}_i=e_M(m_i)\cdot e_T(t_i)$$

其中$e_M(\cdot)$和$e_T(\cdot)$是线性层，它们将嵌入投影到一个多模态嵌入空间中，损失总结如下。对于文本，文本到图像的对比损失L_{t2m}表示为：

$$\mathcal{L}_{t2m}=-log\frac{exp(sim(\mathbb{M}_i,\mathbb{T}_i))}{\sum_{a=1}^{N}exp(sim(\mathbb{M}_i,\mathbb{T}_a))}$$

使用获得的特征来计算图像到文本的损失L_{m2t}，

$$\mathcal{L}_{m2t}=-log\frac{exp(sim(T_i,M_i))}{\sum_{a=1}^{N}exp(sim(T_i,M_a))}$$

上述文本-图像的损失公式是来自匹配对的两个嵌入的相似度。因此，可以得到对比损失L_{Con}如下：

$$\mathcal{L}_{Con}=\mathcal{L}_{t2m}+\mathcal{L}_{m2t}$$

同样地，也考虑基于文本描述查询模态的检索任务。

$$\mathcal{L}_{MLM}=-\frac{1}{|\hat{C}||C|}\sum_{j\in\hat{C}}\sum_{h\in C}q_j^h log\frac{exp(c_j^h)}{\sum_{k=1}^{|C|}exp(c_k^h)}$$

其中$\hat{C}$表示被掩码的文本标记集，$|C|$表示词汇表C的大小。预测的标记概率分布用c^h表示，q^h是一个独热词汇分布，其中真实标记的概率为 1。相应地，行人重识别中 ID 损失L_{ID}，

$$\mathcal{L}_{ID} = -\sum_{i}^{G} y_i log(p_i)$$

其中 y_i 是真实标签的概率分布， p_i 是模型的预测概率分布。 $\hat{y}$ 是模型的预测输出，G 表示训练最小批次中的图像数量， i 是类别的索引，即 $1-G$ 中的 i 。最终，总体损失表示为：

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{Con} + \mathcal{L}_{MLM}$$

具体地，行人重识别模型通过待查询行人的文本描述特征检索行人图像特征，并输出与行人图像特征对应的预设 Rank-K 排序列表，作为行人重识别的结果。排名列表中的行人图像依据行人图像特征与待查询文本描述之间的相似度进行排列。