Computer Science and Application 计算机科学与应用, 2021, 11(10), 2529-2537
Published Online October 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1110256

Hans汉斯

TopN成对相似度迁移的三元组跨模态检索

谭钜源[1,2]，何国辉[1,2]，袁文聪[1,2]

[1]五邑大学智能制造学部，广东 江门
[2]江门市智能数据分析与应用工程技术研究中心，广东 江门

收稿日期：2021年8月18日；录用日期：2021年10月13日；发布日期：2021年10月20日

摘 要

随着科技的快速发展，网络上的信息呈现出多模态共存的特点，如何存储和检索多模态信息成为当前的研究热点。其中，跨模态检索就是使用一种模态数据去检索语义相关的其它模态数据。目前大部分研究都聚焦于如何在公共子空间中使相关的样本尽可能靠近，不相关的样本尽可能分离，没有过多考虑相关样本的排序情况。因此提出一种TopN成对相似度迁移的三元组跨模态检索方法，其利用三元组损失和局部保持投影构建多模态共享的公共子空间，同时将原始空间中样本之间的高相似度关系迁移到公共子空间，以构建合理的排序约束。最后在两个经典跨模态数据集上证明了方法的有效性。

关键词

跨模态检索，子空间学习，三元组损失，局部保持投影，成对相似度迁移

Triplet Cross-Modal Retrieval Based on TopN Pairwise Similarity Transfer

Juyuan Tan[1,2], Guohui He[1,2], Wencong Yuan[1,2]

[1]Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong
[2]Jiangmen Intelligent Data Analysis and Application Engineering Technology Research Center, Jiangmen Guangdong

Received: Aug. 18th, 2021; accepted: Oct. 13th, 2021; published: Oct. 20th, 2021

Abstract

With the rapid development of science and technology, information on the Internet shows the characteristics of multi-modal coexistence. How to store and retrieve multi-modal information has become a current research hotspot. Cross-modal retrieval is to use one type of modal data to re-

文章引用: 谭钜源, 何国辉, 袁文聪. TopN 成对相似度迁移的三元组跨模态检索[J]. 计算机科学与应用, 2021, 11(10): 2529-2537. DOI: 10.12677/csa.2021.1110256

trieve semantically related data of other modalities. Most of the current research focuses on how to bring related samples as close as possible and how to separate unrelated samples as much as possible in the common subspace, but ignores the ranking of related samples. Therefore, a triplet cross-modal retrieval method based on TopN pairwise similarity transfer is proposed. It uses triplet loss and Locality Preserving Projections to construct a multi-modal shared common subspace. Meanwhile, it transfers the high similarity relation from origin subspace to common subspace to construct reasonable ordering constraints. Finally, the effectiveness of the method is proved on two classical cross-modal datasets.

Keywords

Cross-Modal Retrieval, Subspace Learning, Triplet Loss, Locality Preserving Projections, Pairwise Similarity Transfer

1. 引言

随着科技的进步和社会的发展，网络数据呈现多模态共存、多模态融合的趋势。例如，在新闻网站上用于从不同角度描述同一新闻事件的新闻图像和详细文本，其中图像和文本是不同形式的数据，但是它们共同描绘着一个同一的新闻事件，也就是说图像和文本通过一个新闻事件关联在一起。这类多模态数据呈现底层特征异构、高层语义相关等特点[1]。如何存储和检索多模态数据成为研究热点。

跨模态检索是利用一种模态数据去检索语义相关的其它模态的数据，其核心问题是如何测量不同模态数据的相似度[2]。主流的思路是将异构的数据映射到相同的潜在子空间中，建立异构数据之间的联系，将异构数据转换为同构数据进行相似度测量。根据是否利用标签约束，主要分为有监督的子空间学习和无监督的子空间学习两种方法。在无监督的子空间学习中，最为经典的方法就是原理简单、使用广泛的典型关联分析(Canonical Correlation Analysis, CCA) [3]。CCA 将成对的异构数据分别进行线性变换并投影到公共子空间中，以最大化成对数据之间的相关性为目标优化各自的线性变换矩阵。而在有监督的子空间学习中，人们充分利用标签信息，从多模态数据中学习到更优的关联关系。例如，Deng 等人[4]利用语义标签生成语义相似矩阵，并借此构建跨模态三元组学习异构数据之间的关联关系。目前大多数跨模态检索方法只聚焦于如何在公共子空间中使相关的异构数据尽可能相近，不相关的数据尽可能相离，却忽略了相关检索结果的排序情况，以至于虽然返回了相关的检索结果，但排名靠前的却不是最终匹配的结果，非常影响搜索体验。

为此，本文提出 TopN (前 N 项)成对相似度迁移的三元组跨模态检索方法，用于提高排名靠前检索结果的匹配准确度。本方法首先采用一种单模态三元组配合局部保持投影的方法构建异构数据之间的关联关系，然后引入成对相似度迁移方法捕获原始数据之间的高相似度关系并迁移到公共子空间中，使得公共子空间中相邻特征之间能保持原始特征的高相似度关系，以实现检索结果的高匹配度。

2. 相关概念

2.1. 三元组损失

三元组损失是一种常用的分类损失，广泛应用于人脸识别领域[5]。三元组损失需要三个输入，分别

是锚、正样本和负样本，其中锚可以是任意样本，正样本是与锚类别相同的样本，负样本是与锚类别不同的样本。随后将三元组输入通过共享参数的训练网络，得到各自的嵌入特征。三元组损失的思想就是拉近正样本到锚的距离，同时尽可能推远负样本到锚的距离。具体来说，使负样本嵌入特征到锚嵌入特征的距离大于正样本嵌入特征到锚嵌入特征的距离。通过三元组损失可以有效地区分正样本和负样本，让类别相同的样本尽可能靠近，类别不同的样本尽可能远离。

2.2. 局部保持投影

局部保持投影(Locality Preserving Projections, LPP) [6]是一种经典的降低特征维度方法，其思想也被广泛应用于特征提取领域[7]。局部保持投影原理是先在原始空间中构建样本对之间的远近关系矩阵，并在投影的过程中也保持这种关系，使得在低维空间中的样本特征也保持原本的远近关系。具体来说，通过远近关系矩阵辨别低维空间样本之间的近邻关系，使近邻的样本之间的距离尽可能小，以保持原始的近邻结构。

3. TopN 成对相似度迁移的三元组跨模态检索

TopN 成对相似度迁移的三元组跨模态检索主要由两部分构成。第一部分是子空间学习，主要利用三元组损失和局部保持投影方法构建公共子空间；第二部分是排名学习，引入成对相似度迁移方法将原始特征中的高相似度关系迁移到公共子空间中，以保持检索结果的高匹配度。为了论述的方便，将围绕图像和文本两种模态数据进行论述，相关方法可以方便地扩展到其它模态数据。

3.1. 符号定义

假定 n 为数据集的总数量，$\chi=\left\{\left(X_i^{\upsilon},X_i^{t}\right)\right\}_{i=1}^{n}$ 表示 n 对图像文本对，其中 X_i^{υ} 表示第 i 项图像原始特征，X_i^{t} 表示第 i 项文本原始特征；并且 $X_i^{\upsilon}\in\mathbb{R}^{d_1}$ 和 $X_i^{t}\in\mathbb{R}^{d_2}$，$d_1$ 和 d_2 分别表示图像和文本特征空间的维度。$Y=\left\{y_i\right\}_{i=1}^{n}$ 表示相应的类别标签信息，其中 $y_i\in\{0,1\}^{d_3}$ 是类别标签的独热编码表示，一共由 d_3 位 $\{0,1\}$ 表示。假定有映射到公共子空间的图像特征与文本特征，分别为 $H^{\upsilon}=f_{\upsilon}\left(X^{\upsilon};\theta_{\upsilon}\right)$ 和 $H^{t}=f_{t}\left(X^{t};\theta_{t}\right)$，其中 θ_{υ} 和 θ_{t} 分别表示图像映射模型和文本映射模型的参数，并且 $H^{\upsilon}\in\mathbb{R}^{d_4}$ 和 $H^{t}\in\mathbb{R}^{d_4}$，d_4 表示公共子空间的维度。假定映射后的公共子空间中的距离适用于 Frobenius 范数，如(1)所示。

$$dist\left(H^{\upsilon},H^{t}\right)=\left\|H^{\upsilon}-H^{t}\right\|_F \tag{1}$$

3.2. 三元组损失和局部保持投影

不同于 TDH 模型[5]中在两个模态上构建三元组损失，本方法分别在图像模态和文本模态构建各自的三元组损失，然后通过局部保持投影方法进行联合。该做法可以更好地关联图像文本对中对位的图像和文本信息，以提升检索的准确性。

为了尽可能使相关的图像数据靠近，不相关的图像数据分离，在公共子空间中引入图像三元组 $\left(H_i^{\upsilon,anchor},H_i^{\upsilon,postive},H_i^{\upsilon,negative}\right)$ 构建图像三元组损失。其中 $H_i^{\upsilon,anchor}$ 是随机选取的图像特征，作为锚；$H_i^{\upsilon,positive}$ 表示与锚语义相关的图像特征，作为正样本，而 $H_i^{\upsilon,negative}$ 表示与锚语义不相关的图像数据特征，作为负样本。三元组损失希望与锚语义相关的正样本在公共子空间上与锚的距离较近；与锚语义不相关的负样本在公共子空间上与锚的距离较远。具体来说，锚到负样本的距离应该比锚到正样本的距离要更大，距离的差值为 α^{υ}，如公式(2)所示。

$$l^{\upsilon}=\sum_{i=1}^{n}\left[\left\|H_i^{\upsilon,anchor}-H_i^{\upsilon,positive}\right\|_F^2-\left\|H_i^{\upsilon,anchor}-H_i^{\upsilon,negative}\right\|_F^2+\alpha^{\upsilon}\right]_+ \tag{2}$$

同理，文本特征在公共子空间上也构建相应的文本三元组损失。随机选取的三元组 $\left(H_i^{t,anchor},H_i^{t,postive},H_i^{t,negative}\right)$ 分别作为锚、正样本和负样本；α^t 表示为锚到负样本的距离与锚到正样本的距离之间的距离差值。损失如公式(3)所示。

$$l^{t}=\sum_{i=1}^{n}\left[\left\|H_i^{t,anchor}-H_i^{t,positive}\right\|_F^2-\left\|H_i^{t,anchor}-H_i^{t,negative}\right\|_F^2+\alpha^{t}\right]_+ \tag{3}$$

其中 $[\cdot]_+$ 表示如果计算结果大于 0，则输出计算结果；否则输出 0。

为了关联不同模态的语义相关数据，引入局部保持投影方法。具体来说，语义相关的跨模态特征，在公共子空间中的距离应该靠近。其中 H_i^{υ} 和 H_j^{t} 分别表示公共子空间中选取的图像特征和文本特征。$R\in\{0,1\}^{n\times n}$ 表示语义相关矩阵，如果图像 H_i^{υ} 和文本 H_j^{t} 的标注信息是一样的，即图像和文本语义相关，令 $R_{ij}=1$；否则认为图像和文本语义不相关，令 $R_{ij}=0$。如公式(4) (5)所示。

$$l^{c}=\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|H_i^{\upsilon}-H_j^{t}\right\|_F^2R_{ij} \tag{4}$$

$$R_{ij}=\begin{cases}1 & y_i=y_j\\ 0 & y_i\neq y_j\end{cases} \tag{5}$$

3.3. TopN 成对相似度迁移

多数研究仅仅实现了如何检索语义相关的跨模态数据，但是语义相关的跨模态数据的检索排名对用户的检索体验同样重要。在实际检索中，与搜索项相似度越高的检索结果排位应该越靠前。我们发现在原始特征空间中，相似度越高的特征对，其实际的数据也越相似，但是仅限于相似度最高的前 N 项。受 Kang 等[8]文献的启发，本文引入成对相似度迁移方法，将原始特征之间的高相似度关系迁移到公共子空间中，构建相关特征的排位关系。与上述文献中迁移所有相似度关系不同，本文仅仅迁移前 N 高的相似度关系，以提升公共子空间排名学习鲁棒性。为了实现成对相似度迁移，利用每个模态的原始特征相似度关系构建相似度矩阵 $S^{m,origin}\in\mathbb{R}^{n\times n},m\in\{\upsilon,t\}$，和处理后的相似度矩阵 $S^{m}\in\mathbb{R}^{n\times n},m\in\{\upsilon,t\}$，计算公式如下。

$$S^{m,origin}=\cos\left(H^m,H^m\right)\quad m\in\{\upsilon,t\} \tag{6}$$

其中 cos 为余弦相似度计算公式：

$$\cos\left(H_i^m,H_j^m\right)=\frac{H_i^m\cdot\left(H_j^m\right)^T}{\left\|H_i^m\right\|_2\cdot\left\|H_j^m\right\|_2} \tag{7}$$

因每个样本的可辨别性，原始的样本之间的相似度计算结果不会太高，但是又需要将较高的相似度关系迁移至公共子空间才能提升检索匹配度，所以本文对原始相似度矩阵进行了增强处理。处理原理是利用反比例函数在区间 $[0,1]$ 之间的陡峭变化以增大原来数值的区间，然后进行归一化处理。处理方法的特点为：1) 因为处理方法是递增函数，原来的相似度排序依然相同；2) 处理后的特征之间的相似度整体上会增大，只有少部分较小的相似度会变小，这样有利于提取样本之间的高相似度关系。处理方法如(8)所示。其中 $S_{i*}^{m,origin}$ 代表原始相似度矩阵的第 i 行，S_{i*}^{m} 表示处理后的相似度矩阵的第 i 行。

$$S_{i*}^{m}=1-\frac{\frac{1}{S_{i*}^{m,origin}}-1-\min\left(S_{i*}^{m,origin}\right)}{\max\left(S_{i*}^{m,origin}\right)-\min\left(S_{i*}^{m,origin}\right)} \tag{8}$$

针对图像模态数据构建 TopN 的排序损失，其中 $top(i,j,N)\in\{0,1\}$ 表示原始图像特征 X_i^{υ} 和 X_j^{υ} 的相似度是否属于前 N 高相似度，如果属于前 N 高相似度，则 $top(i,j,N)=1$，否则 $top(i,j,N)=0$。$\left(S_{ij}^{\upsilon}-\cos\left(H_i^{\upsilon},H_j^{\upsilon}\right)\right)^2$ 表示原始图像空间中的相似度迁移至公共子空间。图像模态的 TopN 排序损失如公式(9)所示。

$$l^{\upsilon,rank}=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}top(i,j,N)\times\left(S_{ij}^{\upsilon}-\cos\left(H_i^{\upsilon},H_j^{\upsilon}\right)\right)^2 \tag{9}$$

同理，在文本模态数据中同样构建 TopN 的排序损失，如公式(10)所示。

$$l^{t,rank}=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}top(i,j,N)\times\left(S_{ij}^{t}-\cos\left(H_i^{t},H_j^{t}\right)\right)^2 \tag{10}$$

至此，本方法的总体损失函数如公式(11)所示，由局部保持投影损失、各模态的三元组损失和 TopN 排序损失组成。超参数 β 和超参数 λ 分别控制三元组损失和 TopN 排序损失对整体损失的贡献程度。

$$l=l^{c}+\beta\left(l^{\upsilon}+l^{t}\right)+\lambda\left(l^{\upsilon,rank}+l^{t,rank}\right) \tag{11}$$

3.4. 优化

本模型需要训练的参数有 $\left\{\theta^{\upsilon},\theta^{t}\right\}$，分别是图像映射网络和文本映射网络的参数。采取随机梯度下降方法(Stochastic Gradient Descent, SGD) [9]和分批次交替优化的策略优化网络的参数。具体来说，对于每一批输入的数据，只训练优化一个网络的参数，即第 $2k$ 批输入优化参数 θ^{υ}，第 $2k+1$ 批输入优化参数 θ^{t}，如此类推。优化过程如下：

算法 1 TopN 成对相似度迁移的三元组跨模态检索优化

输入：图像文本对 χ，标注信息 Y，批次大小 $size$，超参数 $\{\beta,\lambda\}$，网络 $\{f_{\upsilon},f_t\}$，训练次数 T，学习率 lr；

输出：参数 $\left\{\theta^{\upsilon},\theta^{t}\right\}$；

1) 初始化参数 $\left\{\theta^{\upsilon},\theta^{t}\right\}$；

2) 循环 T 次：

3) For $k=1$ to $\lceil n/size\rceil$:

4) 批输入经过网络得到公共子空间特征 $H^{\upsilon}=f_{\upsilon}\left(X^{\upsilon};\theta_{\upsilon}\right)$ 和 $H^{t}=f_t\left(X^{t};\theta_t\right)$；

5) 通过公式(11)计算损失 l;

6) If k% 2 == 0:

7) 反向传播优化更新参数 θ^{υ}；

8) Else：

9) 反向传播优化更新参数 θ^{t}；

10) End for;

11) 结束循环。

4. 实验

4.1. 实验准备

本文方法是在开源的深度学习框架 TensorFlow 上实现的，其中设置实验参数为 $\alpha^{v}=1.0$、$\alpha^{t}=1.0$ 和 $N=3$，超参数 $\beta=0.5$ 和 $\lambda=0.15$，学习率 $lr=0.001$。在构建三元组损失阶段选取三元组时，选取的三元组 $\left(H^{anchor},H^{positive},H^{negative}\right)$ 满足 $H^{anchor}\neq H^{positive}$ 条件。因为一般数据集中正样本的数量少于负样本的数量，当允许 $H^{anchor}=H^{positive}$ 时，会导致在公共子空间中不相关的数据分离情况变差。

4.2. 数据集

Wikipedia 数据集[10]是从维基百科中搜集的 2866 对图像文本对，一共分为 10 种类别。其中每一对图像文本对由一张图像加上一段关于这张图片的描述构成。将图像通过视觉几何组(Visual Geometry Group, VGG)预先训练的模型提取 4096 维的原始图像特征；并且将文本描述通过词袋模型(Bag of Word, BoW) [11]提取出 100 维的原始文本特征。最后将数据集划分 2006 对图像文本对作为训练集和 860 对图像文本对作为测试集。

PascalSentence 数据集[12]由 1000 对图像文本对构成，总共分为 20 类，每一类有 50 对图像文本对。其中每一对图像文本对由一张图像和 5 段相似的描述构成，为了拟合实际检索情况，只使用第一段文本描述作为实验文本。同样分别提取 4096 维 VGG 特征作为原始图像特征和提取 1386 维 BoW 特征作为原始文本特征。同时将数据集划分 700 对图像文本对作为训练集和 300 对图像文本对作为测试集。

在训练模型时只使用训练集样本；在测试时将测试集作为搜索项，测试集加上训练集作为被检索项。

4.3. 评价指标

为评价模型的跨模态检索效果，采用两种检索任务进行检索测试，其一是用图像检索文本，记作 I→T；其二是用文本检索图像，记作 T→I。本实验采用检索领域常用的平均精度均值(Mean Average Precision, MAP) [13]作为评价指标，因为 MAP 不仅仅考虑检索结果的准确率，还考虑检索结果的排名信息。MAP 是多次检索结果 AP 值的平均值，公式如(12)所示。

$$AP=\frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k}\times rel_k \tag{12}$$

其中 n 是指检索结果的前 n 项；k 是指理想的相关项排位情况；R_k 是指实际的相关项排位情况；rel_k 是指该检索结果与搜索项是否相关，如果搜索项和检索结果相关，则 $rel_k=1$，否则 $rel_k=0$；R 是指前 n 项检索结果与搜索相关的数量。默认情况下，MAP 代表取全部检索结果进行评测；MAP@n 代表取检索结果的前 n 项进行评测。

4.4. 各方法的 MAP 对比

为了验证本方法的有效性，引入多个跨模态方法，在公开数据集上进行对比实验。

典型关联分析 CCA 使用线性方法，将高维空间的异构数据投影到低维的公共子空间中，目的是使投影后的原始跨模态数据之间的相关性最大。

核典型关联分析(Kernel Canonical Correlation Analysis, KCCA) [14]针对现实数据大多都是非线性的情况，使用核函数将异构数据映射到高维核空间，再使用 CCA 方法对核空间中的跨模态数据进行关联。

深度典型关联分析(Deep Canonical Correlation Analysis, DCCA) [15]使用深度神经网络代替线性映射，将异构数据映射到公共子空间，目的是使原始跨模态数据之间的相关性最大。

语义相关最大化方法(Semantic Correlation Maximization, SCM) [16]计算标注信息的余弦相似度作为语义信息，目标是最大化语义信息和模态特征之间的相关性。

联合特征学习方法(Joint Representation Learning, JRL) [17]同时对五种模态数据进行公共子空间的学习，并引入 $l_{2,1}$ 范数对参数矩阵进行约束，提升参数矩阵的稀疏性，以提高学习到的参数矩阵的质量。

对抗跨模态检索(Adversarial Cross Modal Retrieval, ACMR) [18]利用特征提取网络作为生成器，模态分类器作为判别器巧妙地设计对抗网络，结合三元组损失，模态分类器损失和标签预测损失共同减少不同模态特征之间的差距。

基于三元组的深度哈希(Triplet-based Deep Hashing, TDH) [4]引入三元组损失使用三元组负对数似然函数同时构建多模态之间的关联关系，以减少分类损失和提高多模态数据的关联性。

图表征学习(Graph Representation Learning, GRL) [19]将多模态数据映射到各自的图子空间中，使用无向边连接语义相关的图节点，然后用图卷积网络提取特征，利用分类器和图拉普拉斯正则来优化模型。

表 1 呈现各种模型在两个数据集上的 MAP 值比较情况。从表中有以下发现：

1) 深度跨模态检索方法一般优于线性跨模态检索方法。现实数据大多都是复杂的、非线性的，而深度网络在处理非线性数据时效果要比线性方法要更好。

2) 本文方法在两个数据集上 MAP 结果优于其它模型，原因可能在于单模态三元组加上局部保持投影的联合作用能更好地关联相关特征和分离不相关特征。

Table 1. MAP Comparison of all methods
表 1. 各方法在两个数据集上的 MAP 对比

数据集	方法	I→T	T→I	Average
Wikipedia	CCA	0.224	0.282	0.253
Wikipedia	KCCA	0.329	0.329	0.329
Wikipedia	SCM-Orth	0.340	0.335	0.338
Wikipedia	ACMR	0.439	0.361	0.400
Wikipedia	DCCA	0.444	0.396	0.420
Wikipedia	JRL	0.520	0.568	0.544
Wikipedia	GRL	0.567	0.552	0.560
Wikipedia	本文方法	**0.600**	**0.609**	**0.605**
PascalSentence	CCA	0.272	0.237	0.254
PascalSentence	KCCA	0.315	0.316	0.315
PascalSentence	SCM-Orth	0.465	0.335	0.400
PascalSentence	ACMR	0.434	0.416	0.425
PascalSentence	JRL	0.439	0.464	0.452
PascalSentence	DCCA	0.556	0.653	0.605
PascalSentence	GRL	0.716	**0.709**	0.713
PascalSentence	本文方法	**0.751**	0.689	**0.720**

4.5. 前 *n* 项检索结果的 MAP

考虑到前 n 项检索结果对用户搜索体验非常重要，因此以下分别计算模型在 Wikipedia 数据集和 PascalSentence 数据集的前 n 项检索结果的 MAP 值，即 MAP@n，其中 n 分别取值为 1、5、10、20、50。随着 n 的减小，前 n 检索结果的 MAP 值增大，说明本模型不仅能有效地检索到相关跨模态数据，而且排

位越靠前检索准确率越高。在 Wikipedia 数据集实验中，MAP@1 比 MAP@50 提升 9.7 个百分点，而在 PascalSentence 数据集实验中，MAP@1 比 MAP@50 也提升 5.9 个百分点。结果如表 2 所示。

Table 2. MAP@n results from two datasets
表 2. 两个数据集 MAP@n 结果

数据集	前 *n* 项	I→T	T→I	Average
Wikipedia	1	0.698	0.819	0.759
Wikipedia	5	0.655	0.768	0.711
Wikipedia	10	0.640	0.746	0.693
Wikipedia	20	0.631	0.726	0.678
Wikipedia	50	0.625	0.700	0.662
PascalSentence	1	0.873	0.829	0.851
PascalSentence	5	0.850	0.800	0.825
PascalSentence	10	0.837	0.788	0.813
PascalSentence	20	0.818	0.767	0.792
PascalSentence	50	0.751	0.689	0.720

5. 结束语

本文提出 TopN 成对相似度迁移的三元组跨模态检索方法，用于提升跨模态检索结果的匹配度。首先引入三元组损失和局部保持投影构建多模态共享的公共子空间，然后利用成对相似度迁移策略将原始空间样本之间的高相似度关系迁移到公共子空间中，保持相似样本在公共子空间中近邻结构，以达到提高检索匹配度的结果。通过 Wikipedia 和 PascalSentence 两个公开数据集的大量实验表明本文方法的有效性和优越性。未来工作将考虑引入用户检索的实际点击数来共同优化跨模态检索结果的排名。

基金项目

国家自然科学基金项目(61771347)；广东省基础与应用基础研究基金(No. 2019A1515010716)；广东省普通高校基础研究与应用基础研究重点项目(No. 2018KZDXM073)。

参考文献

[1] 欧卫华, 刘彬, 周永辉, 等. 跨模态检索研究综述[J]. 贵州师范大学学报: 自然科学版, 2018, 36(2): 114-120.

[2] Wang, K., Yin, Q., Wang, W., *et al*. (2016) A Comprehensive Survey on Cross-Modal Retrieval. arXiv:1607.06215.

[3] Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J. (2004) Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, **16**, 2639-2664. https://doi.org/10.1162/0899766042321814

[4] Deng, C., Chen, Z., Liu, X., *et al*. (2018) Triplet-Based Deep Hashing Network for Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, **27**, 3893-3903. https://doi.org/10.1109/TIP.2018.2821921

[5] Schroff, F., Kalenichenko, D. and Philbin, J. (2015) Facenet: A Unified Embedding for Face Recognition and Clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 815-823. https://doi.org/10.1109/CVPR.2015.7298682

[6] He, X. and Niyogi, P. (2004) Locality Preserving Projections. *Proceedings of the* 16*th International Conference on Neural Information Processing Systems*, Whistler, Columbia, 9-11 December 2003, 153-160.

[7] Zhang, W., Kang, P., Fang, X., *et al*. (2019) Joint Sparse Representation and Locality Preserving Projection for Feature Extraction. *International Journal of Machine Learning and Cybernetics*, **10**, 1731-1745. https://doi.org/10.1007/s13042-018-0849-y

[8] 康培培, 林泽航, 杨振国, 等. 成对相似度迁移哈希用于无监督跨模态检索[J]. 计算机应用研究, 2021, 38(10):

3025-3029.

[9] Zhu, Z., Li, Y. and Liang Y. (2018) Learning and Generalization in Overparameterized Neural Networks, Going beyond Two Layers. arXiv preprint arXiv:181104918.

[10] Pereira, J.C., Coviello, E., Doyle, G., *et al*. (2013) On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 521-535. https://doi.org/10.1109/TPAMI.2013.142

[11] Mikolov, T., Chen, K., Corrado, G., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv e-prints, arXiv:1301.3781.

[12] Rashtchian, C., Young, P., Hodosh, M., *et al*. (2010) Collecting Image Annotations Using Amazon's Mechanical Turk. *Proceedings of the NAACL HLT* 2010 *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, June 2010, 139-147.

[13] Peng, Y., Zhai, X., Zhao, Y., *et al*. (2015) Semi-Supervised Cross-Media Feature Learning with Unified Patch Graph Regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, **26**, 583-596. https://doi.org/10.1109/TCSVT.2015.2400779

[14] Blaschko, M.B. and Lampert, C.H. (2008) Correlational Spectral Clustering. *Proceedings of the* 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587353

[15] Andrew, G., Arora, R., Bilmes, J., *et al*. (2013) Deep Canonical Correlation Analysis. *Proceedings of the International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. 28, Atlanta, 16-21 June 2013, 1247-1255.

[16] Zhang, D. and Li, W.-J. (2014) Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization. *Proceedings of the AAAI Conference on Artificial Intelligence*, Québec, 27-31 July 2014, 2177-2183.

[17] Zhai, X., Peng, Y. and Xiao, J. (2013) Learning Cross-Media Joint Representation with Sparse and Semisupervised Regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, **24**, 965-978. https://doi.org/10.1109/TCSVT.2013.2276704

[18] Wang, B., Yang, Y., Xu, X., *et al*. (2017) Adversarial Cross-Modal Retrieval. *Proceedings of the Proceedings of the* 25*th ACM International Conference on Multimedia*, Mountain View, 23-27 October 2017, 154-162. https://doi.org/10.1145/3123266.3123326

[19] Cheng, Q. and Gu, X. (2021) Bridging Multimedia Heterogeneity Gap via Graph Representation Learning for Cross-Modal Retrieval. *Neural Networks*, **134**, 143-162. https://doi.org/10.1016/j.neunet.2020.11.011