doi:10.11887/j.cn.202105012 http://journal.nudt.edu.cn

跨模态检索中的相似性漂移问题*

郑奇斌[1,2],刁兴春[3,4],王彦臻[3,4],曹建军[5],刘 艺[3,4],秦 伟[3,4]

(1. 陆军工程大学 指挥控制工程学院,江苏 南京 210007;2. 军事科学院,北京 100089;
3. 军事科学院 国防科技创新研究院,北京 100071;4. 天津(滨海)人工智能创新中心,天津 300450;
5. 国防科技大学 第六十三研究所,江苏 南京 210007)

摘 要:为了降低"相似性漂移"问题的影响,提出一种基于"邻域传播"的匹配策略,将待查询项的模态内近邻映射到目标空间中,并将它们在目标空间中的最近邻作为查询项的跨模态近邻。基于邻域传播的匹配策略在不改变跨模态映射函数的条件下,可以有效地降低"相似性漂移"带来的误匹配现象。理论和实验分析证明,跨模态映射函数的"相似性漂移"问题广泛存在,而基于"邻域传播"的匹配策略可以有效降低其影响,提高匹配的准确率。

关键词:跨模态检索;相似性漂移;邻域传播;深度神经网络

中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2021)05-099-08

Similarity drifting problem in cross-modal retrieval

ZHENG Qibin[1,2], *DIAO Xingchun*[3,4], *WANG Yanzhen*[3,4], *CAO Jianjun*[5], *LIU Yi*[3,4], *QIN Wei*[3,4]

(1. Command and Control Engineering College, Army Engineering University, Nanjing 210007, China;
2. Academy of Military Sciences, Beijing 100089, China; 3. National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing 100071, China; 4. Tianjin Artificial Intelligence Innovation Center, Tianjin 300450, China;
5. The Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China)

Abstract: In order to reduce the impact of the "similarity drift" problem, a matching strategy based on "neighborhood propagation" was proposed, which maps the intra-modal neighbours of the query items onto the target space, and takes their nearest neighbours in the target space as cross-modal neighbours of the query term. Experiments on real data sets prove that the similarity drifting problem exists widely, and the proposed matching strategy can effectively reduce its impact and improve the accuracy of matching.

Keywords: cross-modal retrieval; similarity drifting; neighborhood propagation; deep neural networks

随着多媒体、互联网和大数据等技术的迅速发展,文本、图像等不同模态的数据迅速涌现[1]。不同模态的数据结合在一起,显示出较单模态数据更加丰富的自然和社会属性[2]。而近年来机器学习等技术的发展使得综合利用多模态数据成为可能,特别是得益于深度学习技术的发展,跨模态检索[3]、视觉问答[4]、跨模态推理[5]等多模态应用取得了巨大的进步。

跨模态检索旨在发现不同模态(除少数工作,如文献[6]涉及两种以上模态的数据,大部分研究都聚焦于文本和图像两种模态)数据对象间的相似关系,例如通过文本描述检索具有相似语义的图像,或通过图像检索具有相似语义的文本[7]。由于不同模态数据的表征是异构的,其相似度难以直接计算,通常需要将文本、图像等数据映射到目标表示空间或一个公共表示空间中[2-3]。现有研究通过典型相关分析[7-11]、主题模型[12-14]、稀疏表示[15-16]等方法实现跨模态映射,而近年来基于深度学习的方法[6,17-22]由于其优异的性能成了主流。

尽管以上方法各不相同,但其中采用的映射函数形式几乎是线性变换或深度神经网络,并通过相应的损失函数学习其具体参数。其中,最常见的损失函数是最大边界损失[17,23],此外还有对抗型损失[24]、最大似然估计损失[25]等。这些损失函数的目的是使跨模态映射函数能够同时保持对象的模态内和模态间近邻关系。然而实际中由于训练数据不足等原因,并不能保证学习到的映

* **收稿日期**:2020-02-25
基金项目:国家自然科学基金资助项目(91648204,61532007,61371196)
作者简介:郑奇斌(1990—),男,助理研究员,博士,E-mail:zqb1990@hotmail.com;
刘艺(通信作者),男,助理研究员,博士,E-mail:albertliu20th@163.com

射函数可以完全跨越模态间的障碍。Collell 和 Moens[26]对线性变换和深度神经网络的跨模态映射能力进行测试,发现其对模态内近邻关系的保持较好,而对模态间近邻关系的保持存在缺陷。

在此基础上,本文发现常见跨模态函数存在“相似性漂移”问题——映射函数对模态间近邻关系的保持能力与邻域的大小相关,在较小的邻域内近邻结构与真实近邻保持一致;而当邻域变大时,映射函数的近邻保持能力迅速降低。“相似性漂移”问题的存在会增大跨模态检索中误匹配的概率,降低其准确性。为了降低其影响,本文提出了一种基于“邻域传播”的匹配策略——通过样本的模态内近邻替代它自身,在映射空间中的较小邻域中进行跨模态相似样本的匹配。

本文首先介绍常见的跨模态映射函数,并引出其“相似性漂移”问题;然后,提出基于“邻域传播”的匹配策略,在不改变跨模态映射函数的条件下,降低“相似性漂移”问题对跨模态检索精度的影响;最后,通过在真实数据集上的实验分析,对“相似性漂移”问题的存在性以及匹配策略的有效性进行验证。

1 映射函数与“相似性漂移”问题

跨模态检索任务是找到待查询对象 $x_i \in X$ 在目标集合的跨模态近邻 $y_i \in Y$,为计算任意跨模态对象间的相似度,可以通过映射函数 $f: X \to Y$ 或 $g: Y \to X$ 将源对象映射到目标对象的表示空间[7]。构建跨模态映射函数的过程中,通常需要 f 和 g 能够同时保持对象的模态内近邻关系和模态间近邻关系。以 f 为例,为了在目标空间中保持模态间近邻关系,对任意 $x_i \in X$ 和 $y_i \in Y$,映射 f 需要满足:

$$x_i \approx y_i \leftrightarrow \| f(x_i) - y_i \|_Y \leqslant \delta_Y \tag{1}$$

式(1)表示如果不同模态的样本 x_i 和 y_i 相似,则通过 f 将 x_i 映射到 Y 中后,$f(x_i)$ 和 y_i 的距离应小于 δ。

同时,为了保持模态内对象的近邻关系,映射 f 还需要保持 X 的模态内近邻关系:

$$\| x_i - x_j \|_X \leqslant \delta_X \leftrightarrow \| f(x_i) - f(x_j) \|_Y \leqslant \delta \tag{2}$$

式(2)表示如果同模态对象在原始表示空间中距离较小,在映射后它们的距离仍然要保持足够小;反之,如果同模态对象在原始表示空间中差异较大,在映射后它们的距离仍然要保持足够大。由式(2)可以进一步导出:

$$\| f(x_i) - f(x_j) \|_Z \leqslant K_X \| x_i - x_j \|_X \tag{3}$$

式(3)说明为了保持模态内关系,f 必须为 Lipschitz 连续的,其中 $K_X > 0$ 为 Lipschitz 常数。现有研究中最常用的线性变换

$$f(x) = \boldsymbol{W}_0 x + b_0 \tag{4}$$

以及深度神经网络

$$f(x) = \boldsymbol{W}_1 \sigma(\boldsymbol{W}_0 x + b_0) + b_1 \tag{5}$$

都满足上述条件。其中,$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 为线性映射矩阵,b_0 和 b_1 为偏置,σ 为非线性映射函数。为了使映射 f 能保持对象间的近邻关系,实际中经常通过最小化最大边界损失(max-margin loss)[17]来学习 f:

$$L = \max\{0, \theta + \| f(x) - y \| - \| f(x') - y \|\} \tag{6}$$

其中:θ 为边界;x, y 为相似的样本对;x', y 为不相似样本对。

尽管深度神经网络具有强大的学习能力,但是因为训练数据不足等原因,想要在跨模态映射中“完美”地保持模态内近邻关系和模态间近邻关系并非容易。Collell 和 Moens[26]通过在不同多模态数据集上的实验证明,尽管现有方法都致力于在跨模态映射中保持样本的近邻关系,但是最终学习到的映射函数并不能很好地保持样本的模态间近邻关系:主流的线性变换和深度神经网络更倾向于保持模态内近邻关系,而对模态间近邻关系的保持较差。因此,相似的跨模态样本经过映射后不一定保持靠近,而不相似的跨模态样本却可能接近,从而导致检索的准确率下降。

Collell 和 Moens[26]提出平均近邻覆盖率(mean Nearest Neighbor Overlap,mNNO)来度量跨模态映射对近邻关系的保持能力,给定两个一一配对的对象集合 V 和 Z,mNNO 定义为:

$$\begin{aligned} mNNO^K(V,Z) &= \frac{1}{KN}\sum_{i=1}^{N} NNO^K(v_i, z_i) \\ &= \frac{1}{KN}\sum_{i=1}^{N} |NN^K(v_i) \cap NN^K(z_i)| \end{aligned} \tag{7}$$

其中,索引相同的 $v_i \in V$ 及 $z_i \in Z$ 为匹配的对象,N 为数据集 V 和 Z 的对象总数,$NN^K(v_i)$ 和 $NN^K(z_i)$ 分别为 v_i 和 z_i 的 K 近邻对象索引集合。*mNNO* 通过计算映射前后对象的平均 K 近邻结构覆盖率来度量映射 f 对近邻结构的保持能力,*mNNO*(X, $f(X)$)表示模态内近邻覆盖率,*mNNO*(Y, $f(X)$)表示模态间近邻覆盖率。*mNNO* 越高,则通过 f 进行映射后,匹配的准确率也会越高。

mNNO 对不同粒度(由 K 值体现)的近邻覆盖率取平均值,从整体上度量跨模态映射对近邻关系的保持能力,而本文发现在不同的 K 值下,映射函数的模态间近邻保持能力是变化的:当 K

较小时,样本在映射空间中的跨模态近邻和真实近邻的覆盖率较高;随着 K 变大,跨模态近邻覆盖率迅速下降。本文将这种映射函数对模态间近邻关系的保持能力随邻域变化的现象称为“相似性漂移”,如图 1 所示。

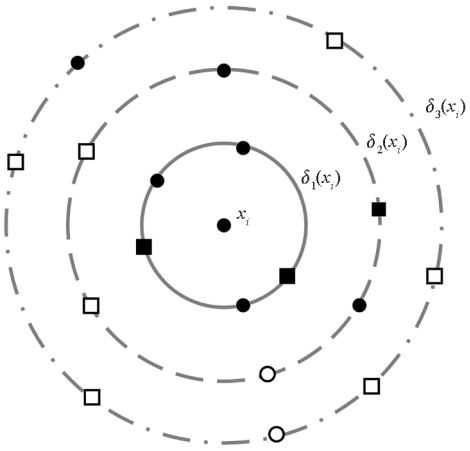


图 1　相似性漂移问题示意

Fig. 1　Illustration of similarity drifting

图 1 中展示的是 x_i 在映射空间 Y 中的近邻结构,其中圆点表示 x_i 同模态近邻在 Y 中的象,方形表示其跨模态近邻,实心表示真匹配,空心表示误匹配。由于映射函数的“相似性漂移”问题,在映射空间中的同模态近邻大部分为真匹配;而跨模态近邻中误匹配较多。此外,图 1 中随着邻域 δ 的增大,发生误匹配的概率逐渐变大。这是由于对象间的相似性经过跨模态映射 f 后难以完全保持,并且其失真程度随着相似性判定的粒度增长(也就是邻域的扩大)而迅速升高。

跨模态映射函数的“相似性漂移”问题显然会增大误匹配发生的概率,并降低跨模态检索的准确性。

2　基于邻域传播的匹配方法

由于跨模态映射函数的“相似性漂移”问题,映射空间中样本的模态间近邻关系难以保持。而相对模态间近邻关系,包括线性变换和深度神经网络在内的映射函数都可以较好地保持样本的模态内近邻关系。此外,映射函数对模态间近邻结构的保持能力是随着邻域的增大而降低的,当邻域较小时,映射函数可以较好地保持模态间的近邻关系。因此,可以借助样本 x_i 同模态近邻在映射空间的象,在其较小邻域中进行近邻匹配,进而降低“相似性漂移”造成的影响,得到更加准确的匹配结果。综合上述讨论,本节提出一种基于“邻域传播”的匹配方法,其基本思想如图 2 所示。

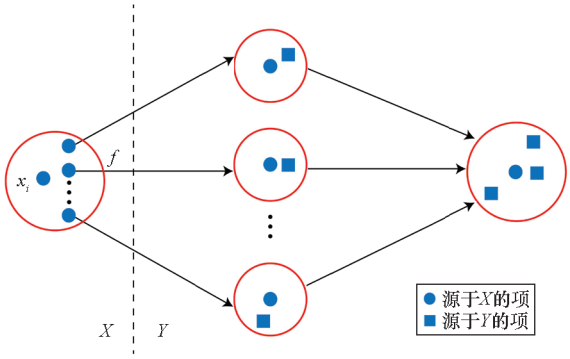


图 2　基于邻域传播的匹配示意

Fig. 2　Illustration of neighbor-propagation matching

不同于传统方法直接通过 f 将 x_i 映射到 Y 空间中后再进行相似度匹配,图 2 为了寻找样本 x_i 的跨模态相似样本,首先通过给定的阈值 τ 在表征空间 X 中筛选同模态相似样本;然后利用 f 将这些样本投影到 Y 空间中,并在 Y 空间中这些样本的邻域内进行近邻匹配,选择每个样本的最近邻作为各自的跨模态相似样本;最后将上述结果求并集,得到样本 x_i 的所有跨模态相似样本。上述过程可以形式化为:

$$\bigcup_{s(x_i,x_j)>\tau} \underset{y_l}{\operatorname{argmax}} s(f(x_j), y_l) \tag{8}$$

其中,$y_l \in Y$ 为目标项集中第 l 项,s 为相似度函数(本文中使用余弦相似度)。模态内相似度阈值 τ 决定了匹配的粒度,本文中由用户根据其对准确率、召回率的偏好,以及数据的分布来决定。详细步骤见算法 1。

算法 1　邻域传播匹配

Alg. 1　Neighbor-propagation matching

输入:查询集合 Q,目标集合 T,阈值 τ

输出:匹配结果 S

```
学习跨模态映射 f: Q→T
S = ∅
for all q_i in Q
    for all q_i ≠ q_j
        if s(q_i, q_j) > τ
            max = 0
                for all t in T
                if s(t, f(q_j)) > max
                    max = s(t, f(q_j))
                    t_j = t
                end if
            end for
            S←S∪{(q_i, t_j)}
        end if
    end for
end for
return S
```

首先,利用现有方法学习跨模态映射函数 f(例如,通过式(6)中的最大边界损失学习式(5)中的深度神经网络作为映射函数 f);然后对每个待查询对象 q_i,筛选所有相似度大于阈值 τ 的模态内近邻 $q_j(j\neq i)$,并利用学习到的函数 f 将其映射到目标空间中后,选择 T 中与 q_j 的相似度最高的 t_j 作为查询对象的匹配对象,并将(q_i, t_j)加入匹配结果集合中。

设 $|Q|=n$, $|T|=m$ 不考虑步骤 1 中跨模态映射学习的复杂度,上述算法的复杂度为 $O(n^2m)$。其中,查询集中共有 n 个待查询项;对每个待查询项,根据阈值 τ 过滤其近邻的复杂度为 $n-1$,相似度高于阈值的最多为 $n-1$ 个,而查询每个近邻在目标项集中的最近邻复杂度为 m,则整个算法的复杂度为 $O(n\times(n-1)\times m)=O(n^2m)$。

3 实验分析

为了验证“相似性漂移”问题以及基于邻域传播的匹配策略,本节在真实数据集上对二者进行了实验分析。

3.1 数据集和实验设置

数据集及特征提取:IAPR TC-12[27],Wikipedia[7],训练集和测试集按照 4∶1 的比例划分。其中,图像的特征通过预训练神经网络模型 VGG[28] 提取,而文本的特征通过双向门限循环单元网络(Bi-directional Gated Recurrent Unit, Bi-GRU)[29] 提取。

跨模态映射:分别通过式(4)中的线性变换[21](记为 Linear),及式(5)中的前馈神经网络完成。与文献[26]一样,$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 的初始参数产生自均匀分布$[-1,1]$,b_0 和 b_1 初始化为 0,非线性映射 σ 采用分别采用 ReLU[20]、TanH[30]、Sigmoid[23] 三种激活函数,网络的深度为五层。

实验中主要验证两个问题:

1)“相似性漂移”问题验证:利用线性变换和深度神经网络对不同数据集上的文本和图像数据进行跨模态映射,通过计算不同邻域的平均最近邻覆盖率[26],分析跨模态映射对模态间关系保持能力和相似性粒度之间的关系,验证“相似性漂移”问题的存在。

2)“文本-图像”匹配方法验证:对相似性匹配方法在文本和图像的双向匹配任务中的表现进行比较,验证邻域传播匹配的有效性。其中,直接通过对象自身相似度阈值进行匹配的方法记为 TH,本文提出的邻域传播匹配方法记为 NP。

通过线性变换和深度神经网络进行跨模态映射,然后通过余弦相似度执行文本到图像以及图像到文本的相似度计算和匹配,并通过准确率(Precision)、召回率(Recall)指标进行对比:

$$Precision=\frac{TP}{TP+FP} \tag{9}$$

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

其中,TP 指正匹配对象的数量,FP 指误匹配对象的数量,FN 指未匹配到的正确对象数量。此外,为了更加直观地体现方法间的性能差异,还计算了曲线的 AUC(area under curve)值,也就是曲线与坐标轴围成的面积。

3.2 最近邻覆盖率测试

本节的实验中,分别测试在给定不同最近邻参数 K 的条件下,线性变换(记为 Linear)和深度神经网络(同文献[26],激活函数使用 ReLU,记为 NN)在跨模态映射中对模态内关系(记为 $f(X)$, X)和模态间关系(记为 $f(X)$, Y)的保持能力,包括文本到图像(记为 I2T)以及图像到文本(记为 T2I)两个方向,以余弦距离(记为 Cos)和欧式距离(记为 Euc)为相似性度量。两个数据集上的平均最近邻覆盖率测试结果如图 3~6 所示。

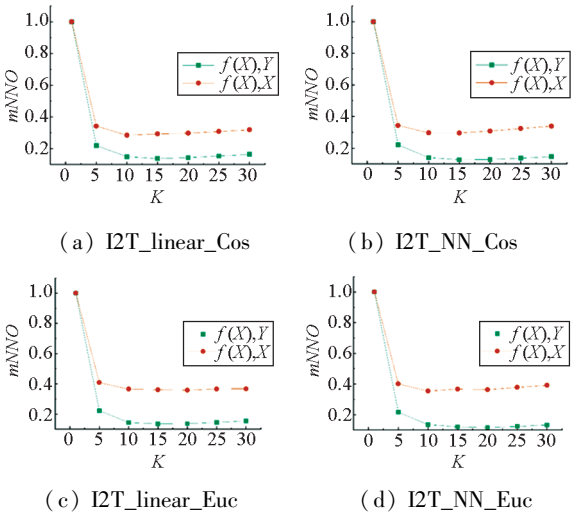
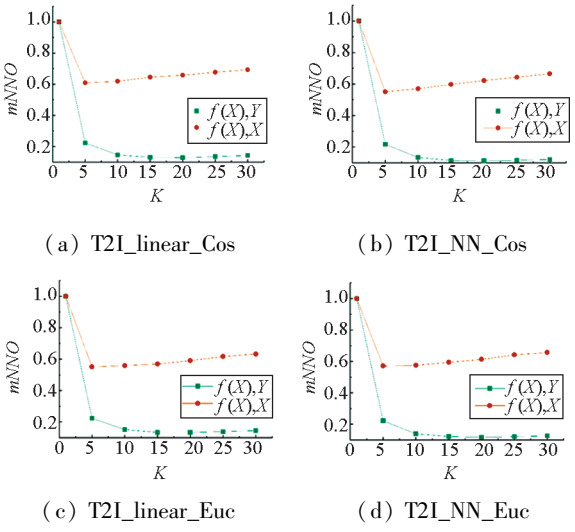


图 3 Wikipedia 图像-文本最近邻覆盖率

Fig. 3 Wikipedia image-text nearest neighbor overlap

图 3 所示为 Wikipedia 数据集中图像-文本的平均近邻覆盖率,经过两种跨模态映射后,图像模态内对象近邻结构覆盖率较高,而图像到文本的跨模态近邻覆盖率较低。此外,模态内近邻和模态间近邻的覆盖率随着 K 的增大,呈现降低的趋势。

图 4 所示为 Wikipedia 数据集上文本-图像



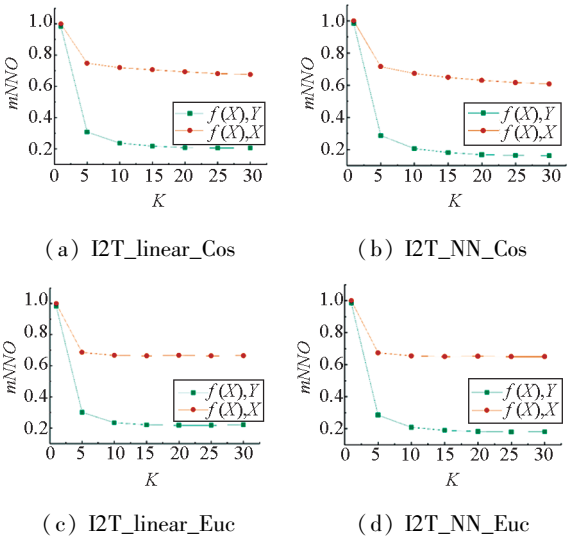
(a) T2I_linear_Cos　(b) T2I_NN_Cos

(c) T2I_linear_Euc　(d) T2I_NN_Euc

图 4　Wikipedia 文本 – 图像最近邻覆盖率

Fig. 4　Wikipedia text-image nearest neighbor overlap

的平均近邻覆盖率结果。跨模态映射对 Wikipedia 数据集的文本数据的模态内近邻结构保持能力更高(高于图像数据约 0.2)。此外,当 $K=1$ 时,该数据集的模态内和模态间的近邻覆盖率同样保持最高,而随着 K 的增大,模态间平均近邻覆盖率仍然随之降低,但模态内近邻覆盖率在降低之后有轻微回升。



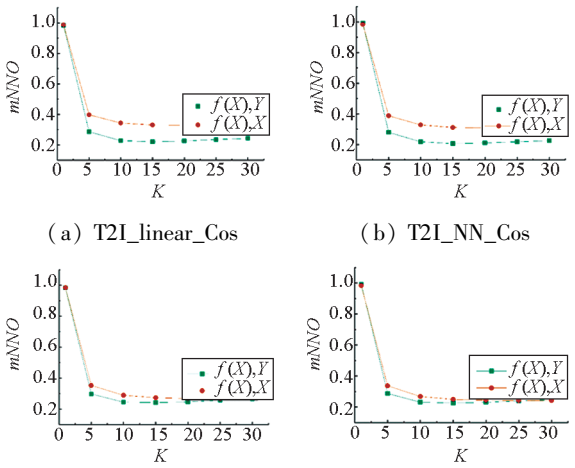
(a) I2T_linear_Cos　(b) I2T_NN_Cos

(c) I2T_linear_Euc　(d) I2T_NN_Euc

图 5　IAPR TC – 12 图像 – 文本最近邻覆盖率

Fig. 5　IAPR TC – 12 image-text nearest neighbor overlap

图 5 为 IAPR TC – 12 数据集中图像 – 文本的平均近邻覆盖率,可以发现,通过线性变换或者神经网络将图像数据映射到共同空间中后,无论使用余弦距离还是欧式距离,两种跨模态映射对模态内关系的保持能力要高于对模态间关系的保持能力。并且,无论对图像到图像的模态内保持,还是图像到文本的模态间保持,其平均覆盖率当 $K=1$ 时最大,而随着 K 的增长,很快下降到一个稳定的值。



(a) T2I_linear_Cos　(b) T2I_NN_Cos

(c) T2I_linear_Euc　(d) T2I_NN_Euc

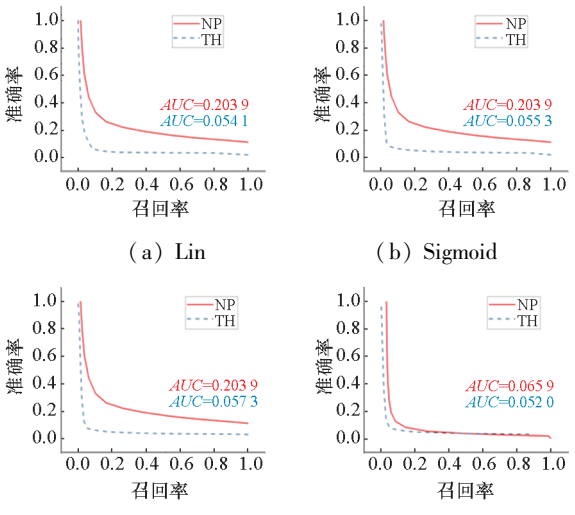
图 6　IAPR TC – 12 文本 – 图像最近邻覆盖率

Fig. 6　IAPR TC – 12 text- image nearest neighbor overlap

在图 6 的 IAPR TC – 12 数据集上文本 – 图像的平均近邻覆盖率测试中,线性变换和深度神经网络同样倾向于保持模态内近邻关系,但是二者的差距较小。此外,当 K 为 1 时,两种跨模态映射函数的近邻保持能力仍最高,并且当 $K>1$ 时迅速下降达到较低水平。

3.3　跨模态匹配验证

在 IAPR TC – 12 和 Wikipedia 两个数据集上执行双向(图像到文本,记为 I2T;图像到文本,记为 T2I)匹配,其准确率 – 召回率曲线如图 7 ~ 10 所示。



(a) Lin　(b) Sigmoid

(c) TanH　(d) ReLU

图 7　IAPR TC – 12 数据集图像 – 文本准确率 – 召回率

Fig. 7　I2T PR curve of IAPR TC – 12

图 7 为 IAPR TC – 12 数据集的图像 – 文本匹配结果,其中基于邻域传播的匹配方法在线性变换以及 Sigmoid 和 TanH 作为激活函数的深度神

经网络中均取得了更高的准确率，而在以 ReLU 作为激活函数的神经网络中准确率较低，但仍然高于通过阈值直接匹配的方法。

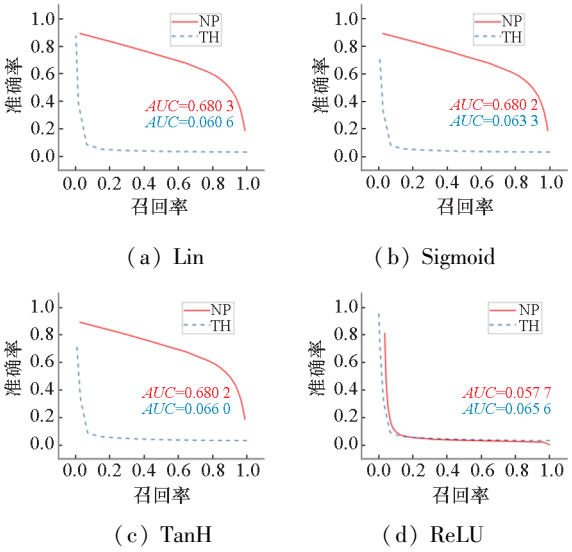


图 8　IAPR TC－12 数据集文本－图像准确率－召回率

Fig. 8　T2I PR curve of IAPR TC－12

图 8 为 IAPR TC－12 数据集的文本－图像匹配结果，在线性变换和以 TanH、Sigmoid 为激活函数的深度神经网络中，基于邻域传播的匹配方法取得了更高的准确率，其 AUC 值远高于直接通过阈值进行匹配的方法；而在采用 ReLU 的深度神经网络中，两种匹配方法近似，其 AUC 值都较低。

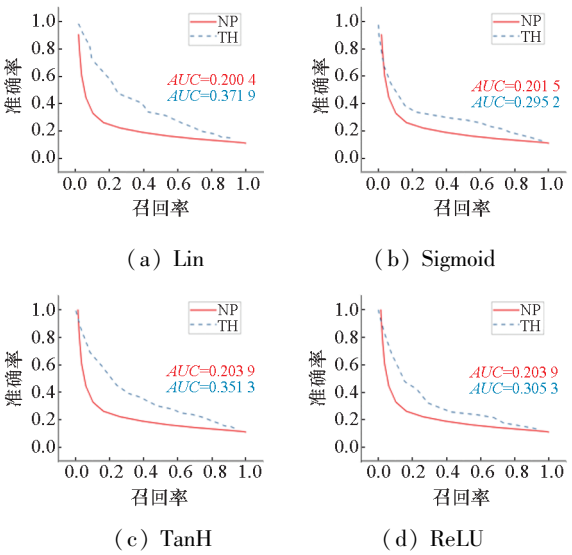


图 9　Wikipedia 数据集图像－文本准确率－召回率

Fig. 9　I2T PR curve of Wikipedia

图 9 为 Wikipedia 数据集图像－文本匹配结果，其中基于邻域传播的匹配方法表现不佳，在四种跨模态映射函数中其准确率始终低于基于阈值的匹配方法。

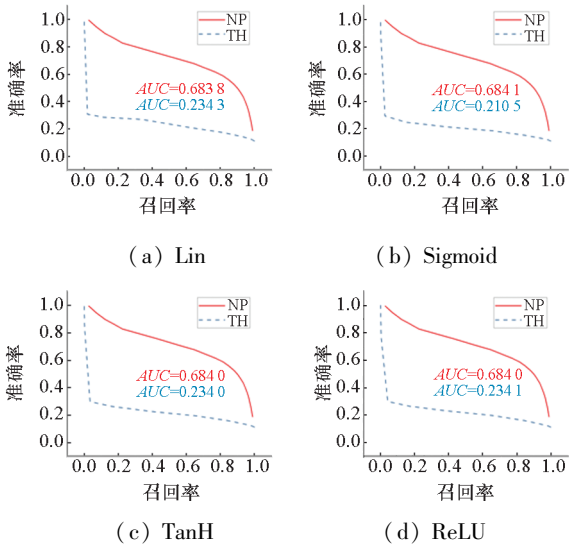


图 10　Wikipedia 数据集文本－图像准确率－召回率

Fig. 10　T2I PR curve of Wikipedia

图 10 中，基于邻域传播的匹配方法在 Wikipedia 数据集文本－图像匹配任务中，准确性远远超过了基准方法。其 AUC 值高出基于阈值匹配的方法约 0.45。

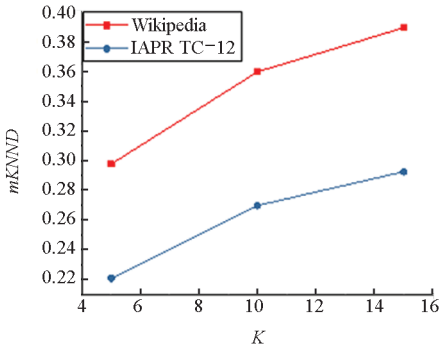
为验证基于邻域传播的匹配方法在部分情况下失效的原因，实验还通过计算样本与其近邻之间的距离，对数据集中文本和图像数据样本的近邻结构进行分析。其中，样本与其近邻的距离通过平均 K 近邻距离（mean K Nearest Neighbor Distance，mKNND）进行度量，其定义如下：

$$mKNND(x_i) = 1/K\sum_{j=1}^{K} d(NN^j(x_i), x_i) \quad (11)$$

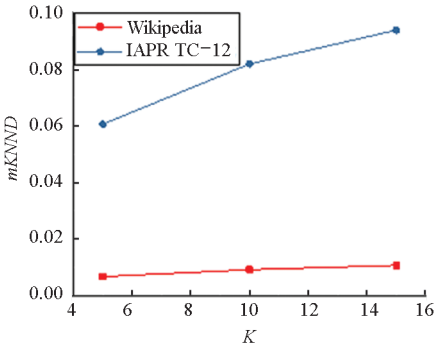
其中，$NN^j(x_i)$ 表示 x_i 的第 j 近邻，d 表示距离（实验中采用余弦距离）。

图 11（a）中，Wikipedia 数据集图像数据的平均 K 近邻距离明显高于 IAPRTC－12 数据集，并且随着 K 的增大而增大。而在图 11（b）中，两个数据集的文本数据平均 K 近邻距离明显低于图像数据，其中 Wikipedia 数据集的平均 K 近邻距离更低并且随着 K 的增大增长较慢。根据图 11 的结果，可以推断图 9 中基于邻域传播的匹配方法失效的原因之一是 Wikipedia 数据集的图像样本间差别较大，在邻域传播的过程中误差增大，导致匹配失效。此外，两个数据集上文本数据的近邻结构更加紧凑，这也是文本－图像匹配准确度高于图像－文本匹配的主要原因。

通过上述实验可以说明，尽管基于邻域传播的匹配方法在特殊情况下会失效，但是在大部分条件下都能有效地提升跨模态匹配的准确率，特别是当模态内和模态间近邻保持能力差别较大



(a) 图像
(a) Image



(b) 文本
(b) Text

图 11　图像和文本数据的平均 K 近邻距离结果

Fig. 11　mKNND results of images and text data

时。因此,本文提出的基于邻域传播的匹配方法对提升跨模态检索准确率具有重要意义。

4　结论

现有跨模态检索问题的研究中,通常通过深度神经网络或线性变换对不同模态的文本和图像数据进行跨模态映射,在此基础上进行相似度计算。而本文发现跨模态映射函数对近邻关系保持能力随着相似性判定的粒度增大而衰减,即存在“相似性漂移”问题。该问题导致误匹配的概率上升,进而降低检索的准确性。

为降低相似性漂移问题的影响,本文提出基于邻域传播的匹配方法,利用同模态近邻样本来发现待匹配对象的跨模态近邻。通过实验验证可以证明,该匹配方法对降低“相似性漂移”问题的影响,提高跨模态检索的准确率具有明显效果。尽管其有效性受到模态内近邻结构的影响,但是这不影响其具有重要参考意义。在未来的工作中,可以通过与普通的匹配方法结合来克服其局限性。例如设定一个阈值,当查询样本和其模态内近邻的距离小于阈值时采取邻域传播的匹配方法,当距离大于阈值时仍然通过该样本自身来进行匹配。

参考文献(References)

[1] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展, 2019, 56(1): 183 – 208.
PENG Yuxin, QI Jinwei, HUANG Xin. Current research status and prospects on multimedia content understanding[J]. Journal of Computer Research and Development, 2019, 56(1): 183 – 208. (in Chinese)

[2] PENG Y X, ZHU W W, ZHAO Y, et al. Cross-media analysis and reasoning: advances and directions[J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(1): 44 – 57.

[3] PENG Y X, HUANG X, ZHAO Y Z. An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges[EB/OL]. [2019 – 10 – 10]. https://arxiv.org/abs/1704.02223.

[4] ANTOL S, AGRAWAL A, LU J S, et al. VQA: visual question answering[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015: 2425 – 2433.

[5] HUANG X, PENG Y X, WEN Z. Visual-textual hybrid sequence matching for joint reasoning[C]// Proceedings of IEEE Transactions on Cybernetics, 2020: 1 – 14.

[6] PENG Y X, QI J W. Quintuple-media joint correlation learning with deep compression and regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(8): 2709 – 2722.

[7] RASIWASIA N, COSTA PEREIRA J, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia, 2010: 251 – 260.

[8] RANJAN V, RASIWASIA N, JAWAHAR C V. Multi-label cross-modal retrieval[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015: 4094 – 4102.

[9] YAN F, MIKOLAJCZYK K. Deep correlation for matching images and text[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3441 – 3450.

[10] WANG L Q, SUN W C, ZHAO Z C, et al. Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval[J]. Signal Processing, 2017, 131: 249 – 260.

[11] MROUEH Y, MARCHERET E, GOEL V. A symmetrically weighted CCA and hierarchical kernel sentence embedding for image & text retrieval[EB/OL]. [2019 – 10 – 10]. https://arxiv.org/abs/1511.06267.

[12] ROLLER S, WALDE S S I. A multimodal LDA model integrating textual, cognitive and visual modalities[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013: 1146 – 1157.

[13] WANG Y F, WU F, SONG J, et al. Multi-modal mutual topic reinforce modeling for cross-media retrieval[C]// Proceedings of the 22nd ACM International Conference on Multimedia, 2014: 307 – 316.

[14] WANG D, GAO X, WANG X, et al. Semantic topic multimodal hashing for cross-media retrieval[C]// Proceedings of the 24th International Joint Conference on

Artificial Intelligence, 2015.

[15] XU X, YANG Y, SHIMADA A, et al. Semi-supervised coupled dictionary learning for cross-modal retrieval in Internet images and texts [C]//Proceedings of the 23rd ACM International Conference on Multimedia, 2015: 847 – 850.

[16] DENG C, TANG X, YAN J C, et al. Discriminative dictionary learning with common label alignment for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2016, 18(2): 208 – 218.

[17] PENG Y X, QI J W, YUAN Y X. Modality-specific cross-modal similarity measurement with recurrent attention network[EB/OL]. [2019 – 10 – 10]. https://arxiv.org/abs/1708.04776.

[18] XU P, YIN Q Y, HUANG Y Y, et al. Cross-modal subspace learning for fine-grained sketch-based image retrieval[EB/OL]. [2019 – 10 – 10]. https://arxiv.org/abs/1705.09888.

[19] ZHAN Y, YU J, YU Z, et al. Comprehensive distance-preserving autoencoders for cross-modal retrieval [C]// Proceedings of the 26th ACM International Conference on Multimedia, 2018: 1137 – 1145.

[20] ZHU B, NGO C W, CHEN J J, et al. R^2GAN: cross-modal recipe retrieval with generative adversarial network [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 11469 – 11478.

[21] GU J X, CAI J F, JOTY S, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7181 – 7189.

[22] PENG Y X, CHIJ Z. Unsupervised cross-media retrieval using domain adaptation with scene graph [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(11): 4368 – 4379.

[23] SOCHER R, KARPATHY A, LE Q V, et al. Grounded compositional semantics for finding and describing images with sentences [J]. Transactions of the Association for Computational Linguistics, 2014, 2: 207 – 218.

[24] WANG B K, YANG Y, XU X, et al. Adversarial cross-modal retrieval [C]//Proceedings of the 25th ACM international conference on Multimedia, 2017: 154 – 162.

[25] LAZARIDOU A, DINU G, BARONI M. Hubness and pollution: delving into cross-space mapping for zero-shot learning[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, 2015, 1: 270 – 280.

[26] COLLELL G, MOENS M F. Do neural network cross-modal mappings really bridge modalities? [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 462 – 468.

[27] The IAPR TC – 12 benchmark: a new evaluation resource for visual information systems [EB/OL]. [2019 – 10 – 20]. http://Thomas.deselaers.de/publication s/papers/grubinger_lrec06.pdf.

[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of International Conference on Learning Representations, 2015.

[29] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. [2019 – 10 – 20]. https://arxiv.org/abs/1406.1078v3.

[30] WANG W, OOI B C, YANG X Y, et al. Effective multi-modal retrieval based on stacked auto-encoders [C]// Proceedings of the VLDB Endowment, 2014, 7(8): 649 – 660.