文章编号：2095-2163(2020)06-0272-06 中图分类号：TP391.4 文献标志码：A

基于分治思想粗匹配和精微匹配相结合的跨模态检索算法

苏 林[1]，卜 巍[2]，邬向前[1]
（1 哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001；2 哈尔滨工业大学 媒体技术与艺术学院，哈尔滨 150001）

摘 要：跨模态检索是人工智能领域的一个重要研究方向，在社会生活中应用广泛，有着巨大的应用价值和经济价值。随着深度学习的兴起，跨模态检索也取得了长足发展。本文借鉴了分治思想和混合推荐的方法，在一个算法框架中构建两个检索模型，分别负责粗匹配和精微匹配。通过特征值取平均值的方式将两个检索模型整合在一起，通过同时使用两个检索模型的检索能力来提升算法的检索效果，增强算法的抗干扰性。

关键词：跨模态；检索；混合

Based on divide and conquer thought combining coarse matching and fine matching cross-modal retrieval algorithm

SU Lin[1], BU Wei[2], WU Xiangqian[1]
（1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2 School of M edia Technology and Art, Harbin Institute of Technology, Harbin 150001, China）

【Abstract】 Cross-modal retrieval is an important research direction in the field of artificial intelligence. It is widely used in social life and has huge application value and economic value. With the rise of deep learning, cross-modal retrieval has also made great progress. This paper draws on the idea of divide and conquer and the method of mixed recommendation. Two retrieval models are constructed in an algorithm framework, which are responsible for rough matching and fine matching respectively. The two retrieval models are integrated by averaging feature values, and the retrieval capabilities of the two retrieval models are used at the same time to improve the retrieval effect of the algorithm and enhance the anti-interference ability of the algorithm.

【Key words】 Cross-modal; Retrieval; Mixing

0 引 言

信息检索一直是信息时代推动社会发展的重要引擎，信息检索可以使人们高效地获取需要的信息，在日常生活和生产中有着重要的价值。随着人工智能的崛起，跨模态图文检索的研究也获得了快速发展，由于其广阔的应用场景和巨大的社会价值，受到越来越多的重视。以往研究跨模态图文检索主要是研究如何提升神经网络的学习能力，进而提高检索的效果。本文不同于以往的研究思路，从检索的整体结构入手，借鉴分治思想，将特征信息中的粗粒度信息和精微信息分开，分别独立学习和匹配。在一个算法框架中，实现两个检索模型，一个负责粗粒度信息学习和检索匹配，即粗匹配；另一个负责精微信息学习和检索匹配，即精微匹配。借鉴混合推荐算法，通过取平均值来整合多个推荐模型，将两个检索模型（粗匹配和精微匹配）整合在一起。粗匹配和精微匹配相结合的检索方法，显著提升了图像检索文本的效果，并且增强了算法在大规模检索中的抗干扰性。本文的主要贡献总结如下：

（1）将混合推荐方法带入检索，提出了粗匹配和精微匹配相结合的跨模态检索方法。

（2）提升了图像检索文本的效果。

（3）增强了算法在大规模检索中的抗干扰性。

1 相关工作

在多层神经网络方面，David Rumelhart 等提出了 BP 网络的误差反向后传 BP（Back propagation）学习算法；RUCK D W 等于 1990 年提出了多层感知机；Hinton 等于 2006 年提出了深度学习的概念。在

基金项目：国家自然科学基金（61672194）；国家重点研究与发展计划（2018YFC0832304）；中国黑龙江省杰出青年科学基金（JC2018021）；国家机器人与系统国家重点实验室项目（SKLRS-2019-KF-14）；中兴通讯产学研合作论坛合作项目。

作者简介：苏 林（1994-），男，硕士研究生，主要研究方向：跨模态检索、深度学习；卜 巍（1977-），女，博士，副教授，主要研究方向：数字媒体技术、数字图像处理、医学图像分析等；邬向前（1973-），男，博士，教授，博士生导师，主要研究方向：数字图像处理、模式识别、生物特征识别等。

通讯作者：卜 巍 Email：buwei@ hit.edu.cn

收稿日期：2020-03-09

损失函数方面，Florian Schroff 等提出了的 triple loss 损失函数并应用于人脸识别；Olivier Chapelle 等提出了一种旨在直接优化一种流行测量指标的算法；Jun Xu 等提出了基于排序指标直接优化的方法；Fartash Faghri 等提出了使用最难分辨的错误项的 triple ranking loss。在跨模态检索方面，Jiuxiang Gu 等提出了将文本生成和图像生成融合进特征嵌入的方法，以学习到具体接地的表示；Kunpeng Li 等提出了简单且可解释的推理模型，以通过区域关系推理和全局语义推理生成增强的视觉表示方法；Sijin Wang 等提出了基于 scene graph 的对象匹配和关系匹配的方法。

2　算法详述

本文提出的 MVSRN 算法是在 VSRN[1] 的基础上加入混合检索，算法流程如图 1 所示。算法是由两个检索匹配任务和一个图像描述生成任务共 3 个任务构成。两个匹配任务分别是粗粒度信息的检索匹配即粗匹配，和精微信息的检索匹配即精微匹配。

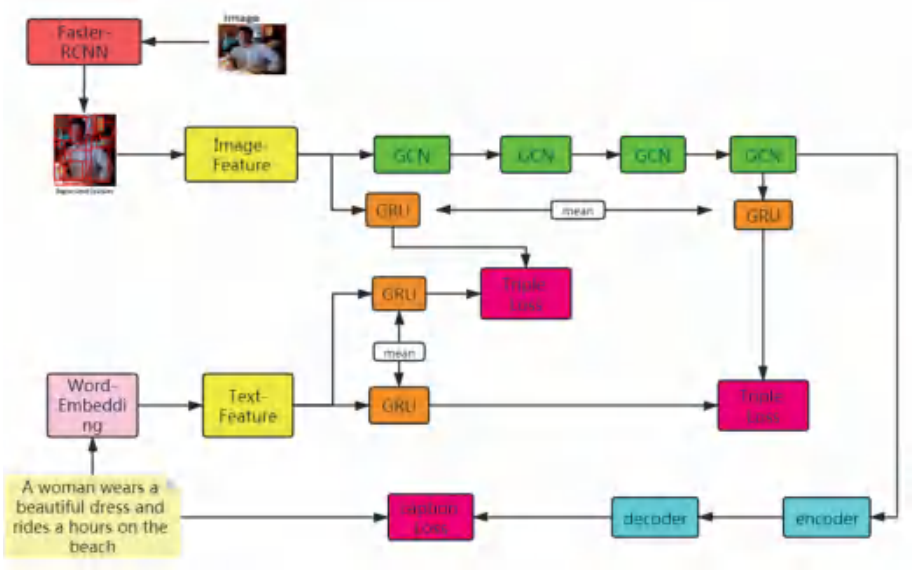

图 1　MVSRN 流程图

Fig. 1　The flow chart of MVSRN

首先图像通过 Faster R-CNN 得到一组在图像中选定的区域，对于每个选定的区域 i，在平均池化层之后提取特征，从而得出 2048 维的 f_i，使用全连接将 f_i 转换为 D 维空间中的特征向量（D 为 2048）如公式（1）所示。

$$v_i = W_f f_i + b_f. \tag{1}$$

得到的是一组图像特征 $V=\{v_1,\cdots,v_k\}$，$v_i \in R^D$ 表示，每个特征 v_i 编码此图像中的一个对象或此图像中的一个区域。用图卷积对图像特征集合 V 中的各个图像特征 v_i 进行图卷积运算。通过公式（2）来构建特征空间中图像区域之间的成对亲和力。

$$R(v_i,v_j) = \varphi(v_i)^{\mathrm{T}}\varphi(v_i). \tag{2}$$

$\varphi(v_i) = W_\varphi v_i$ 和 $\varphi(v_j) = W_\varphi v_j$ 是两个特征向量。权重参数 W_φ 和 W_φ 可以通过反向传播来学习。一个全连接关系图 $G_r = (V,E)$，其中 V 是检测到的区域的集合，而边缘集 E 由亲和力矩阵 R 描述。R 是通过使用公式（2）计算每对区域之间的亲和力而获得的。然后，在此全连接图上进行图卷积（GCN），将残余连接添加到原始 GCN 上，如公式（3）所示。

$$V^* = W_r(RVW_g) + V. \tag{3}$$

其中，W_g 代表尺寸为 D×D 的 GCN 层权重矩阵，W_r 代表残余结构的权重矩阵，R 代表形状为 k×k 的亲和矩阵。对亲和矩阵 R 进行逐行归一化，输出的 $V^*=\{v_1^*,\cdots,v_k^*\}$，$v_i^* \in R^D$ 是经过图卷积后的各个图像特征。

通过将图像特征的序列 $V^*=\{v_1^*,\cdots,v_k^*\}$，$v_i^* \in R^D$ 逐一放入 GRU 中来对所有的图像特征进行融合，得到整张图像的特征表示。在输入第 i 个图像特征时，更新门 z_i，分析当前输入区域特征 v_i^* 和最后一步的 m_{i-1}，以决定该单元对其存储单元进行多少更新。更新门的计算方法如公式（4）所示。

$$z_i = \sigma_z(W_z v_i^* + U_z m_{i-1} + b_z). \tag{4}$$

其中，W_z 代表权重参数，U_z 代表权重参数，σ_z 代表 sigmoid 激活函数，b_z 代表偏差。用 v_i^* 和 m_{i-1} 计算得到重置门 r_i，用来决定要忘记的内容。与更新门类似地计算出 r_i，如公式（5）所示。

$$r_i = \sigma_r(W_r v_i^* + U_r m_{i-1} + b_r). \tag{5}$$

生成候存储单元，候选存储单元 $\tilde{m}_i$ 中的信息用

于更新存储单元 m_{i-1} 中的信息。候选存储单元的计算方式如公式(6)所示。

$$\tilde{m}_i = \sigma_m(W_m v_i^* + U_z(r_i \circ m_{i-1}) + b_m). \quad (6)$$

其中 σ_m 表示 tanh 激活函数，$\circ$ 表示逐元素乘法，r_i 代表重置门。然后，用当前的 $\tilde{m}_i$ 对上一步的 m_{i-1} 进行更新，得到当前的 m_i 计算公式(7)所示。

$$m_i = (1 - z_i) \circ m_{i-1} + z_i \circ \tilde{m}_i. \quad (7)$$

其中，$\circ$是逐元素乘法。将序列 V^* 末尾的存储单元 m_k 作为整个图像的最终表示 I_1，其中 k 是 V^* 的长度。这是混合检索中第一个检索模型(精微信息学习和精微匹配)的图像处理过程。第二个检索模型(粗粒度信息学习和粗匹配)的图像处理是在 Faster-RCNN 模型输出特征集合 $V = \{v_1, \cdots, v_k\}$，$v_i = R^D$ 后，直接将其输入到 GRU 中进行图像特征融合，得到整张图像的特征表示 I_2，这里的 GRU 和前面的 GRU 各自独立。

文本处理是先将文本转换为 one-hot 编码，再经过 word-embedding 变成高维特征向量，再使用基于 GRU 的文本编码器对文本特征进行融合，得到与图像特征表示 I 相同的 D 维特征表示 $C \in R^D$ (D 为 2048)，C 便是最终用于匹配的文本特征表示。GRU 可以学习到句子中的语义上下文。在文本处理部分，使用了两个独立的 GRU，分别得到两个文本特征表示 C_1 和 C_2。第一个检索模型(精微信息学习和精微匹配)最终得到的文本特征表示是 C_1，第二个检索模型(粗粒度信息学习和粗匹配)最终得到的文本特征表示是 C_2。用 I_1 和 C_1，I_2 和 C_2 分别实现两个检索模型的匹配任务。匹配任务采用的是基于铰链的三元组排名损失，使用最有挑战性的错误项来计算损失，即最接近每个查询的错误项。损失定义如公式(8)所示。

$$L_M = [\alpha - S(I,C) + S(I,\hat{C})]_+ + [\alpha - S(I,C) + S(\hat{I},C)]_+. \quad (8)$$

其中，α 用作余量参数，$[x]_+ \equiv \max(x,0)$。此铰链损失包括两项，一项为 I，一项为 C。$S(\cdot)$ 是联合嵌入空间中的相似函数。在实验中通常使用内积作为相似函数的计算方式。$\hat{I} = \arg\max_{j \neq I} S(j,C)$ 和 $\hat{C} = \arg\max_{d \neq C} S(I,d)$ 是与 (I,C) 最接近的错误项。为了提高计算效率，在每个 batch 中都找到 $\hat{I}$ 和 $\hat{C}$，而不是在整个训练集中找到 $\hat{I}$ 和 $\hat{C}$。这里的 I 可以是 I_1 或 I_2，C 可以是 C_1 或 C_2。I_1 和 C_1 的损失函数表示为 L_{M1}，I_2 和 C_2 的损失函数表示为 L_{M2}。

对于文本生成部分，学习到的视觉表示还应该具有生成接近真实字幕句子的能力。具体来说，使用具有注意机制的序列对模型即图 1 中的 encoder-decoder 模型来实现此目的，最大化预测生成句子的对数似然性。损失函数定义如公式(9)所示。

$$L_G = -\sum_{t=1}^{l} \log p(y_t \mid y_{t-1}, V^*; \theta). \quad (9)$$

其中，l 是输出单词序列的长度 $Y = (y_1, \cdots, y_l)$，θ 是序列到序列模型的参数。最终损失函数定义如公式(10)所示，以实现联合优化。

$$L = L_{M1} + L_{M2} + L_G. \quad (10)$$

3 实 验

3.1 数据集与评价指标

本文采用了和 VSRN[1] 相同的数据集和评价指标。在 Microsoft COCO 数据集[2] 和 Flickr30K 数据集[3] 上评估本文的方法。MS-COCO 包括 123,287 张图像，每张图像带有 5 个文本描述。对于 MSCOCO，本文采用了和 VSRN 相同的划分方法，其中包含 113 287 张用于训练的图像，1 000 张用于验证的图像和 5 000 张用于测试的图像。每个图像带有 5 个字幕。最终结果是通过对 5 倍的 1K 测试图像的结果进行平均或在完整的 5K 测试图像上进行测试而获得的。Flickr30K 包含从 Flickr 网站收集的 31783 张图像。每个图像都有 5 个带有人工注释的文字说明。本文使用标准的训练，验证和测试分割[4]，分别包含 28 000 张训练图像，1 000 张验证图像和 1 000 张测试图像。对于评估，通过在 K(R @ K)处的召回率来衡量性能，其定义为在距查询最近的 K 点中检索到正确项的查询占所有查询的比例。

3.2 实验参数设置

将单词嵌入大小设置为 300，将联合嵌入空间的维度设置为 2048。同样采用与 VSRN 相同的设置来设置视觉自下而上的注意模型。基于 GRU 的全局语义推理的区域顺序由自下而上的注意力探测器生成的类别检测置信度得分的降序确定。本文使用 Adam 优化器训练了 30 个 epoch，开始的 15 个 epoch 以学习率 0.000 2 进行训练，将其余 15 个 epoch 的学习率降低到 0.000 02。在等式中设置边距 α 为 0.2，使用的最小批量为 128。对于测试集的评估，通过选择在验证集上表现最佳的模型来解决过度拟合问题。根据验证集中的召回总和选择最佳模型。实验环境会对实验结果有很大影响，尤其是 python、pytorch 和其它库的版本对实验结果影响很

大。

3.3　实验结果与分析

首先，用 VSRN 作者公布的代码，做了 VSRN 的实验。实验中采用了和作者一致的环境、数据集和参数配置。实验结果如表 1 所示。从表中可以看到本文实际做出的结果和作者论文中发布的结果整体差了 3 个百分点左右，作者给出的说明是在后期继续研究过程中对代码做了改动，虽然算法一样，但具体实现上有一些差异，导致了实验结果上有些许差距。

表 1　VSRN 实验结果

Tab. 1　The results of VSRN

data	Methods	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Flick30k	VSRN	71.3	90.6	96.0	54.7	81.8	88.2
	VSRN(ours)	68.1	88.9	93.7	52.1	78.7	86.1
COCO(1k)	VSRN	76.2	94.8	98.2	62.8	89.7	95.1
	VSRN(ours)	73.0	94.1	97.8	60.3	88.4	94.2
COCO(5k)	VSRN	53.0	81.1	89.4	40.5	70.6	81.1
	VSRN(ours)	48.9	78.0	87.4	37.2	68.0	79.2

本文的实验代码是在作者公布的代码上改进的，实验环境和参数设置等都和作者的完全一样，只修改了算法中对应的神经网络结构部分。将 MVSRN 的实验结果和用作者代码真实做出的实验结果进行了对比分析。

首先，在 Flick30k 数据集上进行了 MVSRN 的实验，实验结果如表 2 所示。从表 2 中的实验结果可以看出，在 Flick30k 数据集上，MVSRN 图像检索文本的效果要明显好于 VSRN(ours)，R@1 提高了 3 个百分点，R@5 提升了 0.5 个百分点，R@10 提升了 1.2 个百分点。这说明混合检索可以明显提高图像检索文本的效果。这可能是因为文本部分得到的两个文本特征差异不大，因为文本部分得到两个文本特征所使用的神经网络结构是一样的，都是 GRU，而且 GRU 的输入是同一个 word-embedding，这就使得其学到的两个文本特征差异不会很大，所以文本特征从整体态势来看具有一定的一致性。

表 2　MVSRN 在 Flick30k 上的实验结果

Tab. 2　The results of MVSRN on Flick30k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$[8]	42.5	71.9	81.5	30.2	60.4	72.3
$VSE{+}{+}_{BMVC'18}$[5]	52.9	79.1	87.2	39.6	69.6	79.5
$SCO_{CVPR'18}$[7]	55.5	82.0	89.3	41.1	70.5	80.1
$SCAN_{ECCV'18}$[6]	67.4	90.3	95.8	48.6	77.7	85.2
VSRN(ours)	68.1	88.9	93.7	52.1	78.7	86.1
MVSRN(ours)	71.1	90.4	94.9	47.8	76.8	84.5

两种态势的叠加，由于一致性较好，因此会有态势上的互相增强，使得取平均后得到的文本特征会更加利于检索。因此图像检索文本的效果会得到提升。但是文本检索图像的效果反而有所下降，这很可能是因为 Flick30k 数据集上的数据存在主题性的原因。Flick30k 数据集中的数据来自 Flicker 网站，人们在 Flicker 网站上分享个人日常生活信息，人们分享的生活信息往往具有一定的主题性。例如：饮食、衣服、以及一些活动等等，这类话题往往是人们分享最多的生活信息，也是人们谈论最多、最广泛的内容。由于数据集中的数据具有主题性，因此数据集中会出现大量数据属于同一个主题的现象，同一个主题的数据在内容上往往很接近。MVSRN 的图像特征学习部分比 VSRN 多出了一个对象特征融合的分支。在 Faster-RCNN 得到图像的对象特征后，直接用对象特征进行融合，得到整张图片的特征表示。这个图像的整体特征表示是在对象特征的基础上得到的，包含的都是图像中的内容信息。由于同

一主题下大量数据在内容上很接近,因此这一分支得到的图像特征表示就很相似,在特征空间上距离也很近,很难分辨。当这个图像特征表示整合进入最终用于检索的图像特征表示后,就导致了最终的图像特征表示也比较相近,难以分辨。因此文本检索图像的效果就出现了下降的现象。

之后,又在 coco 数据集上进行了训练和测试,实验结果如表 3 和表 4 所示。从表 3 中的实验结果可以看到,在 coco1k(1k 是指测试集的数据量)上相比于 VSRN(ours)检索效果有了整体提升。图像检索文本任务上 R@1 提升了 0.8 个百分点,R@5 提升了 0.3 个百分点,R@10 提升了 0.2 个百分点。其中 R@1 提升最多,这说明 MVSRN 确实在图像检索文本任务上实现了效果的提升。在 coco1k 上文本检索图像也同样获得了效果的提升,R@1 提升了 1.1个百分点,R@5 提升了 0.5 个百分点,R@10 提升了 0.4 个百分点。这是因为 coco 数据集中的数据不存在明显的主题性,不会出现很多数据属于同一主题非常接近的现象。在 coco5k 上,MVSRN 的实验结果同样整体好于 VSRN(ours),这说明 MVSRN 混合检索通过将两个检索模型整合在一起,确实可以提升检索的效果。从表 3 和表 4 可以看出,MVSRN 在 coco5k 上的提升幅度要高于在 coco1k 上的提升幅度。可以看到从 coco1k 到 coco5k 的测试结果都有一个明显的下降,这是因为测试集变大的原因。当测试集的数量变大时,在检索的时候面对的检索对象就会变多,干扰项就会变多,因此检索效果会出现下降的现象。但是当用两个特征向量取平均来进行检索时,用于检索的特征向量是由两个检索模型共同产生的,是两个检索模型共同在起作用,因此抗干扰能力就会更强。当检索的数据量增大时,对其产生的影响会更小,检索效果下降的就会更少。因此导致了 MVSRN 在 coco5k 上的提升幅度要高于在 coco1k 上的提升幅度。

表 3 MVSRN 在 coco 上的实验结果

Tab. 3 The results of MVSRN on coco1k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$[8]	53.2	83.1	91.5	40.7	75.8	87.4
$VSE++_{BMVC'18}$[5]	64.6	89.1	95.7	52.0	83.1	92.0
$SCO_{CVPR'18}$[7]	69.9	92.9	97.5	56.7	87.5	94.8
$SCAN_{ECCV'18}$[6]	72.7	94.8	98.4	58.8	88.4	94.8
VSRN(ours)	73.0	94.1	97.8	60.3	88.4	94.2
MVSRN(ours)	73.8	94.4	98.0	61.2	88.9	94.6

表 4 MVSRN 在 coco 上的实验结果

Tab. 4 The results of MVSRN on coco5k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$VSE++_{BMVC'18}$[5]	41.3	69.2	81.2	30.3	59.1	72.4
$SCO_{CVPR'18}$[7]	42.8	72.3	83.0	33.1	62.9	75.5
$SCAN_{ECCV'18}$[6]	50.4	82.2	90.0	38.6	69.3	80.4
VSRN(ours)	48.9	78.0	87.4	37.2	68.0	79.2
MVSRN(ours)	50.1	79.5	88.4	38.2	69.4	80.2

4 结束语

本文中从整体结构的角度来进行跨模态检索研究,借鉴分治思想将粗粒度信息学习和精微信息学习分开独立进行,并且各自进行检索匹配即粗匹配和精微匹配。将混合推荐的方法带入到检索中,在一个算法中同时构建了两个检索模型,分别负责粗粒度信息的学习和粗匹配以及精微信息的学习和精微匹配。通过特征值取平均的方式将两个检索模型整合在一起,将粗匹配和精微匹配相结合。借助两个检索模型的检索能力在图像检索文本任务上显著提升了检索效果,并且增强了算法在大规模检索中的抗干扰性。

(下转第 284 页)