基于深度监督学习的零样本跨模态检索方法

曾素佳，庞善民，郝问裕
（西安交通大学电子与信息学部，710049，西安）

摘要：针对当前零样本跨模态检索的研究中未兼顾类别匹配和对应匹配的问题，提出一种基于深度监督学习的零样本跨模态检索方法。对 3 种类型的图文数据对进行了区分，分别是来自同一类别并且匹配的数据对，来自同一类别但不匹配的数据对，以及来自不同类别的数据对；在保持图文类别匹配关系的条件下，为了进一步实现两者的对应匹配，构造了两种基于掩码的匹配约束条件，一种是隐藏同一类别但不匹配的另一模态数据，约束不同类别的图文数据之间的匹配关系，另一种是隐藏其他类别的另一模态数据，约束同一类别内的图文数据之间的对应匹配关系；通过对齐视觉空间和语义空间中对应特征的分布结构，再次约束图文间的类别匹配和对应匹配关系；为了增强文本语义的表征能力，以注意力池化从词序列特征中获得语义显著的句子深度表征。实验结果表明，在 CUB 数据集上，所提方法对图像检索文本和文本检索图像的效果相较基线模型分别提升了 5.9%和 2.2%；在 FLO 数据集上的检索效果分别比现阶段表现最佳的方法高 4.2%和 1.7%。

关键词：零样本；跨模态检索；匹配；注意力
中图分类号：TP391 **文献标志码**：A
DOI：10.7652/xjtuxb202211016 **文章编号**：0253-987X(2022)11-0156-11

Zero-Shot Cross Modal Retrieval Method Based on Deep Supervised Learning

ZENG Sujia, PANG Shanmin, HAO Wenyu
(Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A novel zero-shot cross modal retrieval method based on deep supervised learning is proposed as category matching and corresponding matching are not considered in current research. Firstly, three types of image-text pairs are distinguished, including pairs from the same category that match correspondingly, pairs from the same category that do not match correspondingly, and pairs from different categories. Secondly, with the category of images and texts matched, to further realize the corresponding matching between them, two matching constraints are constructed based on different masking patterns. One is to mask samples that are of another modality and of the same category but do not match with each other, restraining the matching relations among images and texts of different categories. The other is to mask samples that are of another modality and of different categories, restraining the corresponding matching relations between images and texts of the same category. Finally, by aligning distribution structures of visual features and their corresponding semantic features in each space, the category matching and corresponding matching relations between images and texts are constrained again.

收稿日期：2022-04-26。 作者简介：曾素佳(1996—)，女，硕士生；庞善民(通信作者)，男，副教授，博士生导师。 基金项目：国家自然科学基金资助项目(61972312)；陕西省重点研发计划资助项目(2020GY-002)。
网络出版时间：2022-08-05 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220805.1143.002.html

In addition, to enhance the representation of text semantics, the attention mechanism is also utilized to obtain more significant sentence semantic features from word sequences. Experimental results show that on CUB dataset, the proposed method improves the image-based text retrieval and text-based image retrieval effects by 5.9% and 2.2% compared with baseline model respectively; on FLO dataset, these figures are 4.2% and 1.7% higher compared with the current best-performing methods, respectively.

Keywords: zero-shot; cross modal retrieval; matching; attention

以图像来检索与视觉特征最相关的文本，或者以文本来检索与语义最相关的图像，称为图像-文本的跨模态检索。图像-文本跨模态检索是信息检索领域新的重要发展方向，也为实现图像、文本、音频和视频等多模态检索提供了研究基础[1-2]。传统的图像-文本检索以典型相关分析为主要方法，随着深度学习的发展，出现了基于哈希[3-4]、余弦相似度[5-7]等相关性度量的深度神经网络建模方法，并逐渐成为图像-文本检索研究的主流[1-2]。随着零样本学习[8-9]的兴起，出现了图文跨模态检索与零样本学习相结合的应用场景，即零样本的图文跨模态检索[10-11]。

零样本的图文跨模态检索关键在于保持图像的视觉特征和文本的语义特征之间的类别匹配和对应匹配关系，检索的质量受限于深度特征提取的方法，以及类别相关的图文特征之间的关联学习。当前的研究[12-15]大多依赖于从深度预训练的词向量，例如 Word2Vec、BERT[16]等中获得一个较好的文本语义初始化表示，缺乏对文本本身语义显著性的理解。此外，训练阶段的模型也只约束了图文数据对间的类别关系，而没有区分过同一类别并且匹配的另一模态数据、同一类别但不匹配的另一模态数据，以及类别不同的另一模态数据。

本文选取文献[12]的研究作为基线模型，提出一种基于深度监督学习的零样本图文跨模态检索方法，在保持类别匹配的同时又兼顾图文的对应匹配关系。在以 KL 散度(Kullback-Leibler divergence)约束图文数据类别间的匹配关系时，也采用掩码的方式，从两种不同的角度构造了新的图文对应匹配约束条件；同时，采用两种分类网络进一步实现视觉空间和语义空间结构的对齐，并在深度特征的嵌入学习中引入自注意力机制，有效提升了零样本图文跨模态检索性能。

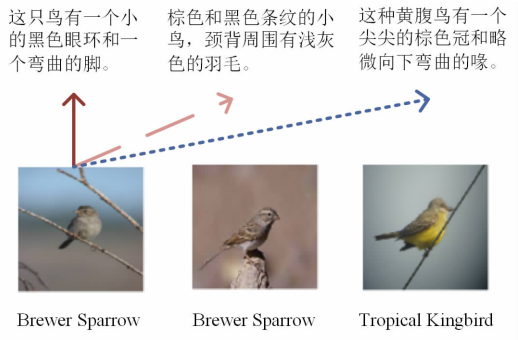
1　相关工作

文献[8]最早将零样本学习的原理引入图像-文本跨模态检索，在训练集图文数据的类别和测试集图文数据的类别不相交的条件下，建立自然语言描述和视觉特征之间的细粒度联系，并对研究动机进行了阐述：在零样本学习中，由于测试集图像的类别未在模型训练过程中出现过(称为未见类别)，测试图像通常要借助专家标注的属性向量作为语义辅助信息进行类别的区分。尽管具有较高的准确率，但是属性向量标注的成本也较高，标注者也需要具备相关专业知识，耗费大量人力、财力和时间。自然语言则提供了一种更灵活和紧凑的方式，通过描述显著的视觉特征就能区分出图像类别。因此，该文献收集了 Caltech-UCSD Birds-200-2011(CUB)和 Oxford Flowers-102(FLO)两个数据集图像的细粒度文本描述，提出通过一个零样本的图文跨模态检索任务，从细粒度的自然语言中训练得到语义辅助信息，来克服属性向量存在的局限性。此后，CUB 和 FLO 被普遍作为评估图文跨模态检索方法在零样本学习方向上的性能表现公共数据集，文献[12-15]均在其上进行方法验证。

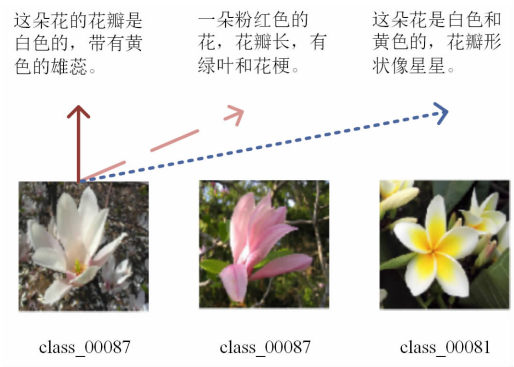
文献[12]提出跨模态的投影学习，采用 KL 散度来度量图文数据分布的跨模态相关性，避免了三元组形式的损失函数[17]需要人为选定合适边界值的不足，检索效果也超过了三元组损失。文献[13]以 BERT 预训练的词向量初始化词序列，以 ResNet-101 为图像的视觉特征提取网络，在文献[12]的研究基础上，采用生成对抗网络的训练方式消除视觉特征和语义特征间的异质性，进一步改进了零样本图文跨模态检索的性能。文献[14]基于 BERT 提出多级文本(包括句子、短语和单词)语义学习网络，同时利用耿贝尔注意力(Gumbel attention)强化这些多级语义和图像区域视觉特征之间的联系。文献[15]则利用图注意力关系网络建模单词之间的语义关系，再与图像整体和区域的视觉特征分别进行匹配学习。

在类别相关的零样本图文跨模态检索研究中，通常存在 3 种类型的图文数据对，分别为来自同一类别并且匹配的数据对、来自同一类别但不匹配的

数据对以及来自不同类别的数据对。然而，相关研究或者将前两者都视为正样本对，或者将后两者都视为负样本对，都会给图像和文本之间的细粒度匹配带来误差，如图 1 所示。



(a)CUB



(b)FLO

图 1 CUB 和 FLO 数据集中 3 种不同类型的图文数据对

Fig. 1 Three different types of image-text pairs on CUB and FLO dataset

由图 1(a)可知，尽管前两对图像和文本都来自“Brewer Sparrow”类，但是从第 1 张图像中几乎看不到第 2 个文本描述中的“黑色条纹”特征；同样，对于第 2 张图像，第 1 个文本描述中的“弯曲的脚”视觉特征也不明显，可知每张图像的视觉特征只与其对应的文本描述最相关，而不一定与同一类别的其他图像的文本描述相关。因此，仅以类别关系作为图像-文本跨模态样本对之间的匹配关系是片面的、不够准确的。

如果将每一个图文数据对视为一个单独的类别，例如文献[18]，虽然能够保持跨模态数据间的对应匹配关系，但是却丢失了有用的类别监督信息。如图 1(b)所示，虽然第 1 张图像和第 2 张图像中的花卉来自同一类别，具有相似的形态特点，只是颜色不同，但是如果没有类别信息作为监督，神经网络便无法学到形态才是区分“class_00087”类花卉的主要特征，而颜色是次要特征。在文本检索图像时，同一类别的图像便不能排到检索结果中靠前的位置，而这是检索排序中不期待出现的情况，因为理想的排序结果是类内的图像总是排在其他类别图像的前面。因此，基于实例的匹配也是不完美的，对于类别相关的零样本图文跨模态检索任务中，既要保证图文的对应匹配关系，又要保持图文间的类别匹配关系。

2 零样本的跨模态检索方法

本文训练阶段的方法框架如图 2 所示，从左到右主要包括 3 个模块。

(1)最左侧为基于注意力的特征学习模块，文本和图像分别通过一个深度神经网络提取其深度嵌入特征；采用注意力的方式使获得的语义特征和视觉特征信号表达更加显著；以 1×1 的卷积网络将两者变换到同一维度的公共空间中进行度量学习。

(2)中间为分类网络学习模块，设计了语义特征分类网络和视觉特征分类网络，分别用于对语义特征和视觉特征进行分类，输出两种特征的类别分布信息，再共享两者的分布信息。对于成对的图像和文本，在这一模块中既能对齐类别分布信息，又能保持对应匹配关系。

(3)最右侧为深度跨模态特征匹配学习模块，利用视觉特征和语义特征计算得到图像和文本间跨模态的匹配概率分布，用于约束跨模态的图文数据对之间的类别关系和对应匹配关系。其中，浅紫色的色块表示同一类别并且匹配的跨模态数据对之间的匹配概率，浅红色的色块表示同一类别但不匹配的跨模态数据对之间的匹配概率，浅灰色的色块表示类别不同的跨模态数据对之间的匹配概率。

2.1 基于注意力的特征学习模块

对于数据集中的文本描述，采用随机初始化的方式，获得其低维词向量序列表示，接着双向长短时记忆网络(bi-directional long short term memory network，Bi-LSTM)被用于学习这些初始化词向量序列的深度语义嵌入。不同于先前的研究需要人为选定平均池化或最大池化的方式来将单词序列特征转换为句子的语义特征，本文提出以自注意力池化[19]的方式将 Bi-LSTM 网络输出的词序列隐藏状态转换为句子的语义特征，表达式如下

$$\boldsymbol{a}_i = \mathrm{softmax}(\boldsymbol{W}_{\mathrm{out}}\tanh(\boldsymbol{W}_{\mathrm{hid}}\boldsymbol{h}_i + \boldsymbol{b}_{\mathrm{hid}}) + \boldsymbol{b}_{\mathrm{out}}) \tag{1}$$

$$\boldsymbol{t} = \sum_{i=1}^{l}(\boldsymbol{a}_i \circ \boldsymbol{h}_i) \tag{2}$$

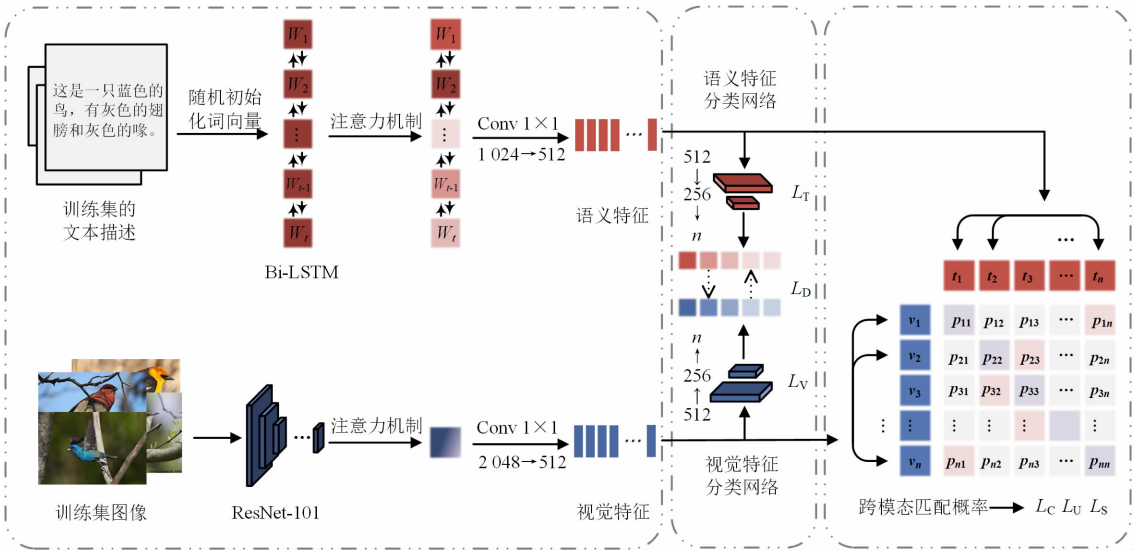


图 2　训练阶段的方法框架

Fig. 2　Framework of proposed method

式中：$\boldsymbol{h}_i$ 为第 i 个单词的隐藏状态特征；$\boldsymbol{a}_i$ 为 h_i 的注意力向量；$\boldsymbol{W}_{\text{hid}}$、$\boldsymbol{W}_{\text{out}}$ 分别为线性变换层的权重向量；$\boldsymbol{b}_{\text{hid}}$、$\boldsymbol{b}_{\text{out}}$ 为线性变换层的偏置量；$\boldsymbol{t}$ 为句子的语义特征；l 为每个句子的单词数；$\circ$ 为哈达玛积操作。

对于数据集中的图像，采用深度预训练神经网络提取图像的深度视觉特征 $\boldsymbol{V}$，并且对称地采用自注意力的方式使重要的视觉特征更加显著。

为了在同一公共空间对视觉特征和语义特征进行度量学习，训练过程中采用一个 1×1 的卷积网络分别将视觉特征和语义特征变换到相同维度的空间中，获得视觉特征 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_N]\in\mathbf{R}^{N\times d}$ 和语义特征 $\boldsymbol{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_M]\in\mathbf{R}^{M\times d}$。其中，$d$ 为公共空间的维度，N 和 M 分别为图像数和文本数。

2.2　分类网络学习模块

视觉特征来源于二维图像数据，而语义特征来源于文本序列，两者分别从不同的深度特征嵌入网络中获得，因此两者在公共空间中的流形结构具有差异，给跨模态的匹配带来了困难。为了对齐视觉空间和语义空间的结构，视觉特征分类网络和语义特征分类网络分别被用于学习两种特征在各自空间中的类别分布信息，同时，通过共享两者的类别分布信息，实现视觉空间和语义空间结构的对齐。

视觉特征分类网络和语义特征分类网络均由两层全连接神经网络构成，对于视觉特征及其对应匹配的语义特征，分类网络计算其类别分布的数学表达式如下

$$\hat{\boldsymbol{p}}(\boldsymbol{v}_i)=\sigma(\boldsymbol{W}_{\text{ov}}\text{ReLU}(\boldsymbol{W}_{\text{iv}}\boldsymbol{v}_i+\boldsymbol{b}_{\text{iv}})+\boldsymbol{b}_{\text{ov}}) \tag{3}$$

$$\hat{\boldsymbol{p}}(\boldsymbol{t}_i)=\sigma(\boldsymbol{W}_{\text{ot}}\text{ReLU}(\boldsymbol{W}_{\text{it}}\boldsymbol{t}_i+\boldsymbol{b}_{\text{it}})+\boldsymbol{b}_{\text{ot}}) \tag{4}$$

式中：$\hat{\boldsymbol{p}}(\boldsymbol{v}_i)\in\mathbf{R}^{1\times K}$，$\hat{\boldsymbol{p}}(\boldsymbol{t}_i)\in\mathbf{R}^{1\times K}$ 分别为视觉特征 $\boldsymbol{v}_i$ 和语义特征 $\boldsymbol{t}_i$ 分到 K 个类别的概率分布；σ 为 Sigmoid 函数；$\boldsymbol{W}_{\text{ov}}$、$\boldsymbol{W}_{\text{iv}}$、$\boldsymbol{W}_{\text{ot}}$ 和 $\boldsymbol{W}_{\text{it}}$ 为构成视觉和语义特征分类网络的线性层权重向量；$\boldsymbol{b}_{\text{iv}}$、$\boldsymbol{b}_{\text{ov}}$、$\boldsymbol{b}_{\text{it}}$ 和 $\boldsymbol{b}_{\text{ot}}$ 为线性层的偏置量。

对于分类网络输出的类别分布，在类别标签的监督下，采用均方差损失函数计算网络的预测损失，将视觉特征和语义特征分类网络的预测损失分别记为 L_{V} 和 L_{T}，具体计算公式如下

$$L_{\text{V}}=\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{p}}(\boldsymbol{v}_i)-y_i)^2 \tag{5}$$

$$L_{\text{T}}=\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{p}}(\boldsymbol{t}_i)-y_i)^2 \tag{6}$$

式中：n 为每次迭代的样本批量大小；$y_i\in\mathbf{R}^{1\times K}$ 为第 i 个样本的类别标签，为独热编码的形式。

此外，本文也通过计算对应匹配的视觉特征和语义特征之间的类别分布差异来实现两者的类别信息共享和空间结构对齐，因为当任意一个视觉特征及其对应的语义特征在各自空间中的类别分布状态相同时，视觉空间和语义空间应具有相同的结构，如下式所示

$$L_{\text{D}}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\boldsymbol{p}}(\boldsymbol{v}_i)-\hat{\boldsymbol{p}}(\boldsymbol{t}_i)\right| \tag{7}$$

式中：$|\cdot|$ 为绝对值误差损失函数。

2.3　深度跨模态特征匹配学习模块

零样本的图文跨模态检索是类别相关的检索任务，因此，在训练过程中，对于某一批次的采样，视觉特征可能遇到 3 种不同类型的语义信息，分别为与

其对应匹配的语义特征、与其属于同一类别但不匹配的语义特征以及与其属于不同类别的语义特征，对于文本的语义特征亦是如此。

为了度量图文数据之间的类别匹配关系，受启发于文献[12]的跨模态投影学习，本文将视觉特征 $\boldsymbol{v}_i$ 匹配到来自同一类别的语义特征都视为等概率事件，并将该类事件的概率总和记为 1，而将 $\boldsymbol{v}_i$ 匹配到来自不同类别的语义特征的事件概率均记为 0，对于文本的语义特征亦是如此，获得类别匹配约束下的概率分布，如图 3 所示。

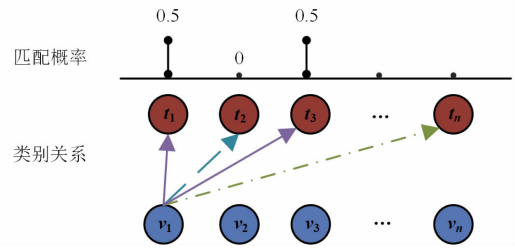


图 3 类别匹配约束下的概率分布

Fig. 3 Distributions in category matching constraint

同时，计算视觉特征和语义特征分别在各自空间中的匹配概率，表达式如下

$$\boldsymbol{q}(\boldsymbol{v}_i \mid \boldsymbol{t}_j)=\exp\left(\boldsymbol{v}_i^{\mathrm{T}}\frac{\boldsymbol{t}_j}{\|\boldsymbol{t}_j\|}\right)\Big/\sum_{k=1}^{n}\exp\left(\boldsymbol{v}_i^{\mathrm{T}}\frac{\boldsymbol{t}_k}{\|\boldsymbol{t}_k\|}\right) \tag{8}$$

$$\boldsymbol{q}(\boldsymbol{t}_i \mid \boldsymbol{v}_j)=\exp\left(\boldsymbol{t}_i^{\mathrm{T}}\frac{\boldsymbol{v}_j}{\|\boldsymbol{v}_j\|}\right)\Big/\sum_{k=1}^{n}\exp\left(\boldsymbol{t}_i^{\mathrm{T}}\frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}\right) \tag{9}$$

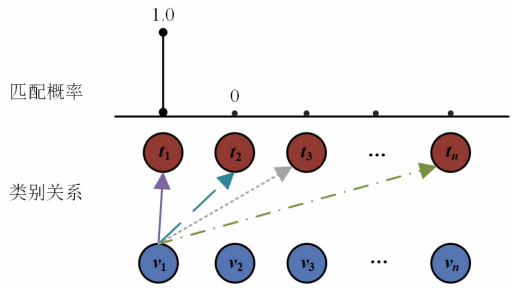
式中：$\boldsymbol{q}(\boldsymbol{v}_i|\boldsymbol{t}_j)$ 为视觉特征 $\boldsymbol{v}_i$ 在语义空间中的投影，作为与 $\boldsymbol{t}_j$ 之间的匹配概率；$\boldsymbol{q}(\boldsymbol{t}_i|\boldsymbol{v}_j)$ 为语义特征 $\boldsymbol{t}_i$ 在视觉空间中的投影，作为与 $\boldsymbol{v}_j$ 之间的匹配概率。

再以反向 KL 散度计算训练的模型对图 3 中数据分布的拟合误差 L_{C}，表达式如下

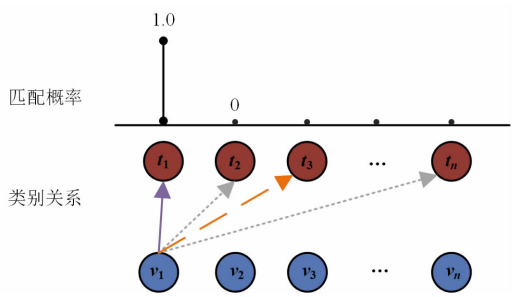
$$L_{\mathrm{C}}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{q}(\boldsymbol{v}_i \mid \boldsymbol{t}_j)\ln\frac{\boldsymbol{q}(\boldsymbol{v}_i \mid \boldsymbol{t}_j)}{p_{ij}+\varepsilon}+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{q}(\boldsymbol{t}_i \mid \boldsymbol{v}_j)\ln\frac{\boldsymbol{q}(\boldsymbol{t}_i \mid \boldsymbol{v}_j)}{p_{ij}+\varepsilon} \tag{10}$$

式中：p_{ij} 为类别匹配约束下的概率分布，若 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$ 属于同一类别，则 $p_{ij}=\frac{1}{m}$，m 为该批次样本中与 $\boldsymbol{v}_i$（或 $\boldsymbol{t}_j$）属于相同类别的语义（或视觉）特征数；$\varepsilon=1\times10^{-8}$，避免除零错误。

为了度量图文数据间的对应匹配关系，不同于文献[18]将每一张图像（或文本描述）视为一个单独的类别，本文利用两种掩码的方式，在保持数据集中原始类别关系的同时，也保持视觉特征和语义特征间的对应匹配关系，两种掩码隐藏方式（以灰色点线箭头表示）如图 4 所示。



(a)隐藏同一类别但不匹配的另一模态样本



(b)隐藏其他类别的另一模态样本

图 4 两种掩码隐藏方式

Fig. 4 Two masking patterns

在图 4(a)中，若 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$（或 $\boldsymbol{t}_i$ 和 $\boldsymbol{v}_j$）来自同一类别但不匹配，则将模型预测的概率值 $\boldsymbol{q}(\boldsymbol{v}_i|\boldsymbol{t}_j)$（或 $\boldsymbol{q}(\boldsymbol{t}_i|\boldsymbol{v}_j)$）替换为负无穷，以避免经过 softmax 函数变换后，再经过自然对数函数变换时真数为 0，分别记为 $\boldsymbol{q}^{\mathrm{s}}(\boldsymbol{v}_i|\boldsymbol{t}_j)$（或 $\boldsymbol{q}^{\mathrm{s}}(\boldsymbol{t}_i|\boldsymbol{v}_j)$）。同时，也将该概率值对应的标签 p_{ij} 替换为 0，相当于不使用 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_i$ 之间的概率匹配信息更新神经网络的参数，以避免同类跨模态样本对约束类别匹配关系产生不利影响。在隐藏了同一类别但不匹配的图文数据之间的匹配概率后，以交叉熵损失函数构造该隐藏条件下的匹配约束条件 L_{S}，如下式所示

$$L_{\mathrm{S}}=-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ij}\ln(\mathrm{softmax}(\boldsymbol{q}^{\mathrm{s}}(\boldsymbol{v}_i \mid \boldsymbol{t}_j)))-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ij}\ln(\mathrm{softmax}(\boldsymbol{q}^{\mathrm{s}}(\boldsymbol{t}_i \mid \boldsymbol{v}_j))) \tag{11}$$

式中：$\boldsymbol{q}^{\mathrm{s}}$ 为隐藏同一类别但不匹配的图文数据之后的跨模态匹配概率；x_{ij} 为 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$ 是否来自同一类别的标签，若 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$ 来自同一类别，则 $x_{ij}=1$，否则 $x_{ij}=0$。

在图 4(b)中，若 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$（或 $\boldsymbol{t}_i$ 和 $\boldsymbol{v}_j$）来自不同类别，则将模型预测的概率值 $\boldsymbol{q}(\boldsymbol{v}_i|\boldsymbol{t}_j)$（或 $\boldsymbol{q}(\boldsymbol{t}_i|\boldsymbol{v}_j)$）替换为负无穷，以避免经过 softmax 变换后，再经过自然对数函数变换时真数为 0，记为 $\boldsymbol{q}^{\mathrm{u}}(\boldsymbol{v}_i|\boldsymbol{t}_j)$（或

$\boldsymbol{q}^{\mathrm{u}}(\boldsymbol{t}_i \mid \boldsymbol{v}_j))$。同时，将该概率值对应的标签 p_{ij} 替换为 0，相当于其他类别的数据不参与神经网络的参数更新，只对来自同一类别的跨模态数据按照是否对应匹配进行二元分类。由于隐藏了其他类别的另一模态样本，在这种方式下能够学到同一类别内跨模态样本之间的区分性。同样地，以交叉熵损失函数构造该隐藏条件下的匹配约束条件 L_{U}，如下式所示

$$L_{\mathrm{U}} = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}\ln(\mathrm{softmax}(\boldsymbol{q}^{\mathrm{u}}(\boldsymbol{v}_i \mid \boldsymbol{t}_j))) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}\ln(\mathrm{softmax}(\boldsymbol{q}^{\mathrm{u}}(\boldsymbol{t}_i \mid \boldsymbol{v}_j))) \quad (12)$$

式中：$\boldsymbol{q}^{\mathrm{u}}$ 为隐藏其他类别的图文数据之后的跨模态匹配概率；y_{ij} 为 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$ 是否对应匹配的标签，若 $\boldsymbol{v}_i$ 和 $\boldsymbol{t}_j$ 对应匹配，则 $y_{ij}=1$，否则 $y_{ij}=0$。

在式(11)和式(12)中，本文将掩码处理后的匹配概率均进行了一次 softmax 变换，以获得未隐藏的跨模态样本之间新的归一化的匹配概率。

综上，通过基于注意力的特征学习模块、分类网络学习模块以及深度跨模态特征匹配学习模块的工作后，得出训练阶段两种优化的目标函数。

方法 1　优化 L_1，基于隐藏同一类别但不匹配的另一模态数据约束匹配关系，如下式所示

$$L_1 = L_{\mathrm{S}} + \alpha_1 L_{\mathrm{C}} + \alpha_2 L_{\mathrm{D}} + \alpha_3 L_{\mathrm{T}} + \alpha_4 L_{\mathrm{V}} \quad (13)$$

式中：α_1、α_2、α_3 和 α_4 为各项约束条件的权重系数。

方法 2　优化 L_2，基于隐藏类别不同的另一模态数据约束匹配关系，如下式所示

$$L_2 = L_{\mathrm{U}} + \beta_1 L_{\mathrm{C}} + \beta_2 L_{\mathrm{D}} + \beta_3 L_{\mathrm{T}} + \beta_4 L_{\mathrm{V}} \quad (14)$$

式中：β_1、β_2、β_3 和 β_4 为各项约束条件的权重系数。

在测试阶段，语义特征由测试集全部的文本描述依次通过训练阶段保存的文本处理模块得到，并且按类别取平均作为文本类语义；视觉特征由图像依次通过 ResNet-101[20] 网络，以及训练阶段保存的图像处理模块得到。接着，计算得到测试图像的视觉特征和文本类语义特征之间的余弦相似度，作为测试阶段检索排序的依据，如图 5 所示。

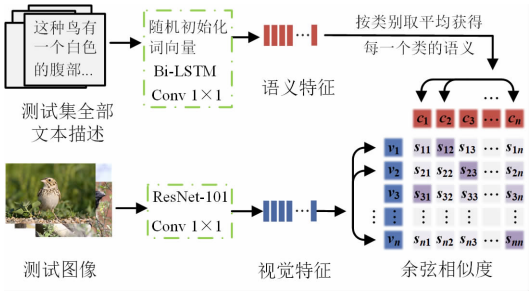


图 5　测试阶段流程图

Fig. 5　Flowchart of testing stage

3　实验结果与分析

3.1　数据集与参数设置

CUB、FLO、Flickr30k[21] 和 Microsoft COCO[22] 均为图文跨模态检索常用的公共数据集。其中，CUB 数据集含有 200 种鸟类的图像，共 11 788 张；FLO 数据集共有 102 种花卉的 8 089 张图像。此外，CUB 数据集的每一类图像也具有 312 个专家系统标注的属性向量，可作为图像分类时的语义原型。文献[8]利用 Amazon Mechanical Turk (AMT)众包平台收集了这两个数据集每一张图像的 10 个细粒度的句子描述，这些句子只描述视觉外观信息，不含比喻，不含命名，也不包括背景、地名等信息。与相关研究[8,12-15,23-25]一致，本文采用这两个数据集公开的零样本划分方式。对于 CUB 数据集，训练集共有 150 个类别的 8 821 张图像和88 210 个文本描述；测试集共有 50 个类别的 2 967 张图像和 29 670 个文本描述；对于 FLO 数据集，训练集含有 7 034 张图像和 70 340 个文本描述，共 82 种花卉；测试集含有 1 155 张图像和 11 550 个文本描述，共 20 种花卉，训练集数据的类别和测试集数据的类别是不相交的。

Flickr30k 数据集包含 30 000 多张图像以及每张图像对应的 5 个句子描述，而 Microsoft COCO 数据集含有的图像更多，共 123 287 张，两者都是复杂生活场景图像，对一张图像通常不只用一个类别标签来标注。然而，零样本学习的设定中单个图像只对应一个类别标签，所以这两个数据集不太适合零样本学习相关的任务，况且当前训练集和测试集这两个数据集的数据类别也有重叠，也不符合零样本学习数据集的划分方式。相比而言，CUB 和 FLO 数据集更适合本文进行方法验证。

使用 PyTorch 搭建模型框架，在 NVIDIA GEFORCE GTX 2080 GPU 服务器上进行训练，使用 Adam 优化器，初始学习率为 2×10^{-4}，在迭代 30 000次和 80 000 次后分别将学习率调整为 2×10^{-5} 和 2×10^{-6}；权值衰减系数设置为 4×10^{-4}；每次迭代的批量大小为 64；训练阶段迭代总次数为 200 000。实验中，式(13)中的超参数设置为 $\alpha_1=1.0$，$\alpha_2=1.0$，$\alpha_3=1.0$，$\alpha_4=1.0$；式(14)中的超参数设置为 $\beta_1=1.0$，$\beta_2=1.0$，$\beta_3=1.0$，$\beta_4=1.0$。

对于图像，采用预训练的 ResNet-101 网络提取全局平均池化后的 2 048 维深度视觉表征，再使用

一层 1×1 的卷积将视觉特征变换到 512 维;对于文本,将每个句子的单词数截取或填充(补零)到 30,采用随机初始化的方式,获得单词的 512 维随机向量,再通过 Bi-LSTM 获得 1 024 维的深度语义表征,同样地,使用一层 1×1 的卷积将语义特征变换到 512 维。使用注意力机制时,将注意力层用在 1×1卷积层之前。

3.2 评价指标

在测试阶段,与相关研究[8,12-15,23-25]在这两个数据集上的处理方式一样,将所有测试类别文本的语义特征按类别取平均,作为参与检索的文本类语义。对于图像检索文本,采用召回率($R@1$)作为评价指标,即对于某一张图像,计算检索到的排名第一的文本类语义与之属于相同类别的比例;对文本检索图像,采用准确率($P@50$)作为评价指标,即对于某一个文本类语义,计算检索到的前 50 个图像中与之属于同一类别的比例。

3.3 实验结果

本文方法在 CUB 和 FLO 数据集上进行验证,并且与近年来相关的零样本图文本检索方法[8,12-15,23-25]进行比较,对图像检索文本以及文本检索图像的评估结果如表 1 所示。

表 1 在 CUB 和 FLO 数据集上图像检索文本的召回率以及文本检索图像的准确率结果比较

Table 1 Comparison of recall for image-to-text and precision for text-to-image on CUB and FLO dataset

方法	图像检索文本的召回率/%		文本检索图像的准确率/%	
	CUB	FLO	CUB	FLO
GMM+HGLMM[23]	36.5	54.8	35.6	52.8
Attributes[24]	50.4		50.0	
Word CNN[8]	51.0	60.7	43.3	56.3
Word CNN-RNN[8]	56.8	65.6	48.7	59.6
Triplet[25]	52.5	64.3	52.4	64.9
Latent Co-attention[25]	61.5	68.4	57.6	70.1
CMPM[12]	62.1	66.1	64.6	67.7
CMPM+CMPC[12]	64.3	68.9	67.9	69.7
TIMAM[13]	67.7	70.6	70.3	73.7
HGAN[14]	68.9	72.4	71.7	75.1
GARN[15]	69.7	71.8	69.4	72.4
本文方法 1	69.8	76.6	69.7	76.8
本文方法 2	70.2	76.6	70.1	76.0

注:表内空白代表未测或无此项。

由表 1 可知,在 CUB 数据集上,本文方法在图像检索文本上的召回率取得了最佳表现,其中,方法 1 和方法 2 的召回率分别高于基线模型 CMPM+CMPC[12] 5.5%和 5.9%。对于文本检索图像,和基线模型 CMPM+CMPC[12] 相比,方法 1 和方法 2 的准确率分别提升了 1.8%和 2.2%,但是比现阶段表现最佳的 HGAN[14] 效果略低,分别相差 2.0%和 1.6%。

在 FLO 数据集上,本文提出的方法 1 和方法 2 在图像检索文本和文本检索图像上都取得了最佳表现。其中,方法 1 的召回率和准确率比 CMPM+CMPC[12] 提高了 7.7%和 7.1%,同时也高于 HGAN[14] 4.2%和 1.7%。实验证明,在含有类别信息的细粒度数据集上,本文提出的方法有效提升了图像-文本跨模态检索的效果。

此外,与基于属性向量的类语义相比,如表 1 中 Attributes[24],从自然语言描述中得到的类语义特征具有远高于其的识别效率。采用专家系统标注数据集的属性向量往往耗费大量人力、财力和时间,而自然语言的获取则容易得多,同样是在零样本学习的方式[8,12-15,23-25]下,本文对于从自然语言中学习零样本的图像分类也具有参考意义。

本文方法 1 和方法 2 的区别主要在于掩码隐藏的方式不同。由于保留了不同类别间的关系,方法 1 中的掩码匹配损失 L_S 也能够在本文任务中单独使用。方法 2 中的掩码匹配损失 L_U 不能在本文任务中单独使用,因为 L_U 只用于对比类内的图文匹配关系,不能对不同类别的跨模态样本进行比较。

由表 1 可知，方法 2 在 CUB 数据集上的检索效果略高于方法 1，而方法 2 在 FLO 数据集上的检索效果略低于方法 1，也反映出两个数据集本身的特性。同一种鸟类的图像可能因为图像的拍摄角度、背景等不同，而形成视觉特征上的干扰差异。同一种花卉图像的差别可能不仅来自图像的拍摄角度、背景等，还可能来自颜色、形态等因素。因此，FLO 数据集类内的方差更大，在本文任务中其类别间的比较更加重要，因此方法 1 对 FLO 数据集上的检索更有利；而 CUB 数据集图像的粒度更细，方法 2 可以对类内的匹配关系进行比较，所以方法 2 对 CUB 数据集上的检索更有利。

3.4　消融分析

为了探究注意力机制、分类网络，以及基于掩码匹配的约束条件对检索效果的影响，本文对方法 1 和方法 2 分别进行了消融实验，如表 2 和表 3 所示。

表 2　本文方法 1 的消融实验结果

Table 2　Results of ablation study for method 1

L_S	L_C	注意力机制	分类网络	图像检索文本的召回率/%		文本检索图像的准确率/%	
				CUB	FLO	CUB	FLO
√				66.6	73.8	64.0	74.7
√	√			69.1	76.6	69.1	75.2
√	√	√		69.5	76.6	69.8	76.8
√	√	√	√	69.8	75.2	69.7	77.1

注：√表示有该项约束条件。

表 3　本文方法 2 的消融实验结果

Table 3　Results of ablation study for method 2

L_U	L_C	注意力机制	分类网络	图像检索文本的召回率/%		文本检索图像的准确率/%	
				CUB	FLO	CUB	FLO
√	√			69.5	76.0	69.6	75.7
√	√	√		70.2	76.6	70.1	75.8
√	√	√	√	70.1	76.6	69.6	76.0

其中，未使用注意力机制时，对词序列的深度特征使用全局平均池化生成句子的语义特征；分类网络包括使用 L_D、L_T 和 L_V 3 种约束条件；使用 L_U 时，由于隐藏了其他类别的跨模态数据，只能约束同一类别内的跨模态数据，因此不能单独使用，只给出与 L_C 共同使用时的结果。

由表 2 可知，L_S 单独使用时，在 CUB 数据集上，图像检索文本的结果高于 CMPM[12] 4.5%，而文本检索图像的结果仅低 0.6%；在 FLO 数据集上，两个方向上的检索表现都超过了基线模型 CMPM+CMPC[12]。L_S 和文献[18]相似，都是将一个样本视为单独的一个类；不同的是，L_S 只是不使用来自同一类别但不匹配的另一模态样本的信息，而不是像文献[18]那样将其视为负样本。因此，L_S 保留了真实的类别信息，保证了某一批次样本中只有一个正样本，即使单独使用也能发挥作用，而基于实例的损失由于破坏了真实的类别关系，单独使用时对测试任务几乎不起作用。

L_S 和 L_C 共同使用时进一步提升了在两个数据集上的图像-文本检索效果。其中，在 CUB 数据集上，图像检索文本和文本检索图像比单独使用 L_S 分别提高了 2.5%和 5.1%，在 FLO 数据集上分别提高了 2.8%和 0.5%。使用注意力机制后，均能改进在两个数据集上的检索表现，证明了使用注意力的有效性。加入分类网络后，在 CUB 数据集上对图像检索文本的效果有 0.3%的提高；而 FLO 数据集上的结果显示，分类网络更有利于文本检索图像的任务。

由表 3 可知，未使用注意力机制时，L_U+L_C 在 CUB 数据集上两个方向的检索结果略胜过 L_S+L_C（表 2），分别高了 0.4%（召回率）和 0.5%（准确率）；加入注意力后，与之相比仍高了 0.7%（召回率）和 0.2%（准确率）。在 FLO 数据集上，使用注意力机制和分类网络均能进一步改进 L_U+L_C 的检索效果；当同时使用注意力机制和分类网络时，达到了方法 2 在 FLO 数据集上的最佳表现。

3.5 定性分析

利用训练过程中保存的文本和图像处理模块分别获得测试集中图像的视觉特征和文本的类语义特征，并根据两者间的余弦相似度大小对检索结果从左到右排序，错误结果以红框标记，进行定性分析。由于在测试阶段对句子特征按类别取平均处理作为类语义特征，因此，当图像检索文本时，也相当于零样本的图像分类。本文对文本检索图像的排序结果进行可视化，如图 6 所示。

(a)CUB

(b)FLO

图 6 不同数据集上文本检索图像的结果

Fig. 6 Results for text-to-image retrieval on different datasets

由图 6(a)可知，在 CUB 数据集上，由于拍摄角度不同，有些图像不能体现鸟类的主要识别特征，同时，背景中其他物体的颜色、形态等也会形成干扰，从而导致错误的排序。例如第 2 行第 3 张图像的视角中，也可以看到“麻雀形状”“白色的喉颈”“黄色的”等视觉特征，因此被误认为属于相关实例。

当图像能全面体现鸟类的主要特征，并且背景干扰较小时，返回的排序结果就比较可靠，例如：第 1 行的图像，“深棕色斑点”的特征都比较明显；第 3 行的图像，视觉外观上都体现了“黑色的头”“黑色的颈”“黑色覆盖其身体的其余部分”，背景环境中也没有和鸟类特征相似的干扰信息。

由图 6(b)可知，对于 FLO 数据集，第 1 行第 3 张图像被误认为有很高的相关度，可能与其同样具有“尖尖末端”的花瓣形态，以及图像背景中有大量的与“长长的绿色和橙色萼片”相似的视觉干扰信息有关。第 2 行的第 3 张图像中的花由于具有类似于“粉红色和黄色斑点”的外观特征，也被误认为属于相关实例，但是在形态上它与真实类别的花卉相差较大。与之相反，第 3 行的第 2 张图像，由于图中的花整体形态与该类别的花卉非常相似，而被排名为响应第二的检索结果，说明花卉形态可能是区分该类别的重要特征，而网络在训练过程中对颜色的影响关注较小。

根据图 6 的分析可以得出，为了提高图文跨模态检索的性能，可以进一步解决主要特征和次要特征判定不清、判断规则具有片面性等问题，以及抵抗背景信息的干扰等。

4 结 论

本文提出一种基于深度监督学习的零样本跨模态检索方法。首先，注意力机制被用于优化语义特征以及视觉特征的深度特征嵌入。其次，在语义空间和视觉空间中分别构造了分类网络，以使成对的图文既能对齐类别分布，又能保持对应匹配关系。此外，针对当前零样本图文跨模态检索的研究中未能同时兼顾图文间的类别匹配关系和对应匹配关系，基于掩码的方式，有选择地利用同一类别但不匹配的另一模态数据，以及类别不同的另一模态的数据，以此构造了两种新的匹配约束条件。在 CUB 和 FLO 数据集上的实验证明，所提方法有效提升了零样本图文跨模态检索的性能。在后续研究中，也可以引入生成对抗网络的训练方式来消除图像和文本之间的异质性，进一步提升检索表现。

参考文献：

[1] 尹奇跃，黄岩，张俊格，等. 基于深度学习的跨模态检索综述 [J]. 中国图象图形学报，2021，26(6)：1368-1388.
YIN Qiyue，HUANG Yan，ZHANG Junge，et al. Survey on deep learning based cross-modal retrieval [J]. Journal of Image and Graphics，2021，26(6)：1368-1388.

[2] 刘颖，郭莹莹，房杰，等. 深度学习跨模态图文检索研究综述 [J]. 计算机科学与探索，2022，16(3)：489-511.
LIU Ying，GUO Yingying，FANG Jie，et al. Survey of research on deep learning image-text cross-modal re-

trieval [J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(3): 489-511.

[3] HU Di, NIE Feiping, LI Xuelong. Deep binary reconstruction for cross-modal hashing [J]. IEEE Transactions on Multimedia, 2019, 21(4): 973-985.

[4] 李慧琼，王永欣，陈振铎，等. 基于排序的监督离散跨模态哈希 [J]. 计算机学报，2021，44(8)：1620-1635.
LI Huiqiong, WANG Yongxin, CHEN Zhenduo, et al. Ranking-based supervised discrete cross-modal hashing [J]. Chinese Journal of Computers, 2021, 44(8): 1620-1635.

[5] WANG Liwei, LI Yin, HUANG Jing, et al. Learning two-branch neural networks for image-text matching tasks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 394-407.

[6] 王红斌，张志亮，李华锋. 基于堆叠交叉注意力的图像文本跨模态匹配方法 [J]. 信号处理，2022，38(2)：285-299.
WANG Hongbin, ZHANG Zhiliang, LI Huafeng. Image-text cross-modal matching method based on stacked cross attention [J]. Journal of Signal Processing, 2022, 38(2): 285-299.

[7] CHEN Hui, DING Guiguang, LIN Zijia, et al. Cross-modal image-text retrieval with semantic consistency [C]// Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 1749-1757.

[8] REED S, AKATA Z, LEE H, et al. Learning deep representations of fine-grained visual descriptions [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 49-58.

[9] WANG Wei, ZHENG V W, YU Han, et al. A survey of zero-shot learning: settings, methods, and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 13.

[10] 张桂梅，龙邦耀，曾接贤，等. 基于去冗余特征和语义关系约束的零样本属性识别 [J]. 模式识别与人工智能，2021，34(9)：809-823.
ZHANG Guimei, LONG Bangyao, ZENG Jiexian, et al. Zero-shot attribute recognition based on de-redundancy features and semantic relationship constraint [J]. Pattern Recognition and Artificial Intelligence, 2021, 34(9): 809-823.

[11] LIN Kaiyi, XU Xing, GAO Lianli, et al. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11515-11522.

[12] ZHANG Ying, LU Huchuan. Deep cross-modal projection learning for image-text matching [C]// Proceedings of the 2018 European Conference on Computer Vision. Berlin, Germany: Springer International Publishing, 2018: 707-723.

[13] SARAFIANOS N, XU Xiang, KAKADIARIS I. Adversarial representation learning for text-to-image matching [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 5813-5823.

[14] ZHENG Kecheng, LIU Wu, LIU Jiawei, et al. Hierarchical Gumbel attention network for text-based person search [C]// Proceedings of the 28th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2020: 3441-3449.

[15] JING Ya, WANG Wei, WANG Liang, et al. Learning aligned image-text representations using graph attentive relational network [J]. IEEE Transactions on Image Processing, 2021, 30: 1840-1852.

[16] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL, 2019: 4171-4186.

[17] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 815-823.

[18] ZHENG Zhedong, ZHENG Liang, GARRETT M, et al. Dual-path convolutional image-text embeddings with instance loss [J]. ACM Transactions on Multimedia Computing Communications and Applications, 2020, 16(2): 51.

[19] MERKX D, FRANK S L, ERNESTUS M. Language learning using speech to image retrieval [EB/OL]. [2019-09-09]. https://arxiv.org/abs/1909.03795.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.

[21] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: new similarity met-

rics for semantic inference over event descriptions [C] // Transactions of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2014: 67-78.

[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]// Proceedings of the 2014 European Conference on Computer Vision. Berlin, Germany: Springer International Publishing, 2014: 740-755.

[23] KLEIN B, LEV G, SADEH G, et al. Associating neural word embeddings with deep image representations using Fisher vectors [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 4437-4446.

[24] AKATA Z, REED S, WALTER D, et al. Evaluation of output embeddings for fine-grained image classification [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 2927-2936.

[25] LI Shuang, XIAO Tong, LI Hongsheng, et al. Identity-aware textual-visual matching with latent co-attention [C] // 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 1908-1917.

（编辑 武红江 刘杨）