智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一致性协议匹配的跨模态图像文本检索方法

宫大汉， 陈辉， 陈仕江， 包勇军， 丁贵广

引用本文:
宫大汉, 陈辉, 陈仕江, 等. 一致性协议匹配的跨模态图像文本检索方法[J]. 智能系统学报, 2021, 16(6): 1143–1150.
GONG Dahan, CHEN Hui, CHEN Shijiang, et al. Matching with agreement for cross-modal image-text retrieval[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(6): 1143–1150.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202108013

DOI: 10.11992/tis.202108013
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210922.1142.004.html

一致性协议匹配的跨模态图像文本检索方法

宫大汉[1,2], 陈辉[2,3], 陈仕江[4], 包勇军[5], 丁贵广[1,2]

（1. 清华大学 软件学院, 北京 100084; 2. 清华大学 北京信息科学与技术国家研究中心, 北京 100084; 3. 清华大学 自动化系, 北京 100084; 4. 涿溪脑与智能研究所, 浙江 杭州 311121; 5. 京东集团, 北京 100176）

摘　要: 跨模态图像文本检索的任务对于理解视觉和语言之间的对应关系很重要, 大多数现有方法利用不同的注意力模块挖掘区域到词和词到区域的对齐来探索细粒度的跨模态关联。然而, 现有的方法没有考虑到基于双重注意力会导致对齐不一致的问题。为此, 本文提出了一种一致性协议匹配方法, 旨在利用一致性对齐来增强跨模态检索的性能。本文采用注意力实现跨模态关联对齐, 并基于跨模态对齐结果设计了基于竞争性投票的跨模态协议, 该协议衡量了跨模态对齐的一致性, 可以有效提升跨模态图像文本检索的性能。在 Flickr30K 和 MS COCO 两个基准数据集上, 本文通过大量的实验证明了所提出的方法的有效性。

关键词: 人工智能; 计算机视觉; 视觉和语言; 跨模态检索; 一致性协议匹配; 注意力; 卷积神经网络; 循环神经网络; 门控循环单元
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2021)06−1143−08

中文引用格式: 宫大汉, 陈辉, 陈仕江, 等. 一致性协议匹配的跨模态图像文本检索方法 [J]. 智能系统学报, 2021, 16(6): 1143–1150.
英文引用格式: GONG Dahan, CHEN Hui, CHEN Shijiang, et al. Matching with agreement for cross-modal image-text retrieval[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1143–1150.

Matching with agreement for cross-modal image-text retrieval

GONG Dahan[1,2], CHEN Hui[2,3], CHEN Shijiang[4], BAO Yongjun[5], DING Guiguang[1,2]

(1. School of Software, Tsinghua University, Beijing 100084, China; 2. Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; 3. Department of Automation, Tsinghua University, Beijing 100084, China; 4. Zhuoxi Institute of Brain and Intelligence, Hangzhou 311121, China; 5. Jd.Com, Inc, Beijing 100176, China)

Abstract: The task of cross-modal image-text retrieval is important to understand the correspondence between vision and language. Most existing methods leverage different attention modules to explore region-to-word and word-to-region alignments and study fine-grained cross-modal correlations. However, the inconsistent alignment problem based on attention has rarely been considered. This study proposes a matching with agreement (MAG) method, which aims to take advantage of the alignment consistency, enhancing the cross-modal retrieval performance. The attention mechanism is adopted to achieve the cross-modal association alignment, which is then used to perform a cross-modal matching agreement with a novel competitive voting strategy. This agreement evaluates the cross-modal matching consistency and effectively improves the performance. The extensive experiments on two benchmark datasets, namely, Flickr30K and MS COCO, show that our MAG method can achieve state-of-the-art performance, demonstrating its effectiveness well.
Keywords: artificial intelligence; computer vision; vision and language; cross-modal retrieval; matching with agreement; attention; convolutional neural network; recurrent neural network; gated recurrent unit

随着社交媒体的空前发展, 互联网上积累了大量的用户数据, 比如图像、文本、语音等。利用这些跨模态数据挖掘用户需求, 提升产品服务, 成为了工业界的迫切需求之一。跨模态图像文本检索是实现跨模态数据挖掘的关键技术之一。它旨

收稿日期: 2021−08−13.　**网络出版日期:** 2021−09−23.
基金项目: 国家自然科学基金项目 (61925107, U1936202); 中国博士后科学基金创新人才支持计划项目 (BX2021161).
通信作者: 丁贵广. E-mail: dinggg@tsinghua.edu.cn.

在探索图像和文本之间的对应关系，实现图像和文本的跨模态信息理解，以及文本检索图像或图像检索文本的智能服务。图像和文本的跨模态检索在实际社交媒体领域有广泛的技术应用价值，比如在人机交互、商业化广告文案推荐以及内容推荐等领域，因此吸引了众多研究人员的注意力。

在实际场景中，跨模态图像文本检索面临两大挑战：1) 不同模态数据的异质性阻碍了模型学习到优异的跨模态表示；2) 视觉和语言之间关联丰富而复杂，准确挖掘两者的对应关系十分困难。为了解决以上挑战，前人工作提出了特征嵌入表示技术来联合学习图像和文本的特征。Wang 等[1] 使用双视图网络分别将图像和文本映射到共享嵌入空间中，然后采用一个保结构的双向目标函数来优化网络。Faghri 等[2] 提出使用难负例来增强嵌入空间的学习，实现了性能的显著性提升。然而特征嵌入表示方法将图像和文本信息映射到统一的向量空间中，忽视了图像和文本信息的复杂性，以及两者之间信息关联的多样性和复杂性。近年来，研究人员提出了许多方法来挖掘图像和文本之间的细粒度的跨模态关联信息。Karpathy 等[3] 将图像中的每个区域与文本中的每个单词对齐，提出了一种基于片段的匹配方法。Nam 等[4] 使用注意力机制[5] 和记忆机制来动态探索图像和文本之间的微妙交互。Lee 等[6] 提出了一种堆叠交叉注意力模型，称为 SCAN，取得了先进的图像文本检索性能。

1　本文工作

SCAN 的成功很好地展示了细粒度跨模态关联关系挖掘的优势。然而，这种方式依靠注意力来实现片段 (即区域和单词) 之间的匹配，只关注两者的一阶关系，并不能反映两种不同匹配方式之间的一致性。具体来说，SCAN 分别用注意力构建了区域−单词和单词−区域的两种对齐方式，其中区域−单词是计算所有单词跟给定区域的相似性得分，并经过规范化操作得到相似性分布，同样地，单词−区域是计算所有区域跟给定单词的规范化后的相似性得分。由于规范化操作，单词和区域的相似性度量在两种对齐方式中会得到不一样的得分，使得出现不同的情况。比如在图 1 中，在区域−单词匹配方式中，和区域 r_2 最相关的词是 dress，而和 beautiful 的相关性较弱，而在单词−区域匹配方式中，区域 r_2 却是和 beautiful 最相关的区域。这种矛盾说明了两种方式不一致的问题。

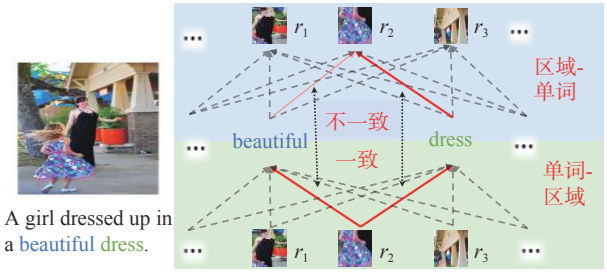


图 1　对齐不一致问题

Fig. 1　Inconsistent alignment problem

本文旨在充分挖掘不同对齐方式的一致性信息，来增强跨模态图像和文本的匹配准确性。本文提出了一种一致性协议匹配的方法 (matching with agreement, MAG)，如图 2 所示，在使用注意力机制得到对齐上下文特征后，构建了区域−单词关联信息表示和单词−区域关联信息表示，并在此基础上，提出关联信息一致性协议的匹配策略，提升图像和文本的跨模态检索性能。

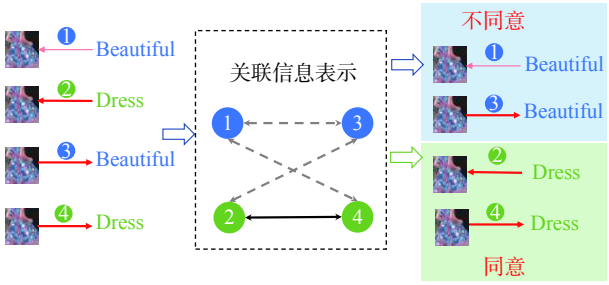


图 2　一致性协议匹配

Fig. 2　Matching with agreement

所提出的一致性协议匹配 (MAG) 方法包含 4 个层，即表示层、对齐层、协议层和匹配层。其中，在表示层，本文使用卷积神经网络 (convolutional neural network, CNN) 和循环神经网络 (recurrent neural network, RNN) 分别提取了图像区域和单词的特征；在对齐层，使用注意力机制得到区域−单词对齐和单词−区域对齐；接着在协议层中，两种不同的对齐可以通过融合注意力机制里的上下文特征得到关联信息表示，并采用竞争性投票的方案得到对齐一致性得分；最后在匹配层通过聚合不同的匹配线索，获得图像文本对之间的相似性。

本文在两个大型的跨模态图像文本检索的基准数据集 (Flickr30K 和 MS COCO) 上进行了大量的实验，实验结果表明，相比于一系列先进的跨模态图像文本检索基线模型，本文提出的一致性协议匹配方法在两个数据集上都具有显著的性能优势，进一步的模型分析和实例分析很好地验证了所提出方法的有效性和优越性。

2　相关工作

跨模态图像文本检索的相关工作通常使用深度神经网络来避免使用手工制作的特征。它们可以大致分为两类：基于嵌入的方法和基于细粒度对齐的方法。

基于嵌入的方法通常学习一个共享的嵌入空间，并通过计算嵌入空间中图像和文本特征之间的距离来估计图像和文本之间的相似性。Frome 等[7]尝试通过 CNN[8] 和 Skip-Gram 模型[9]来学习跨模态表示。类似地，Kiros 等[10]采用 CNN 来提取视觉特征，并采用门循环单元 (gated recurrent unit, GRU)[11]来提取文本特征。Faghri 等[2]提出了一种难负例挖掘的三元组损失函数，取得了显著的性能提升，并成为跨模态图像文本检索领域广泛使用的目标函数。

基于细粒度对齐的方法旨在探索图像和文本之间潜在的细粒度对应关系。Karpathy 等[3]将图像和文本的片段对齐到公共空间中，并通过聚合局部对齐来计算图像和文本的全局相似度。Niu 等[12]提出了一种分层模型，其中图像和文本通过分层策略实现实例到特征的全局和局部联合映射。Lee 等[6]提出了一个堆叠交叉注意力模型，旨在发现图像区域和文本词之间的完整潜在对齐，并在多个基准数据集上实现先进的性能。

3　一致性协议匹配

本节讨论所提出的一致性协议匹配方法，如图 3 所示。

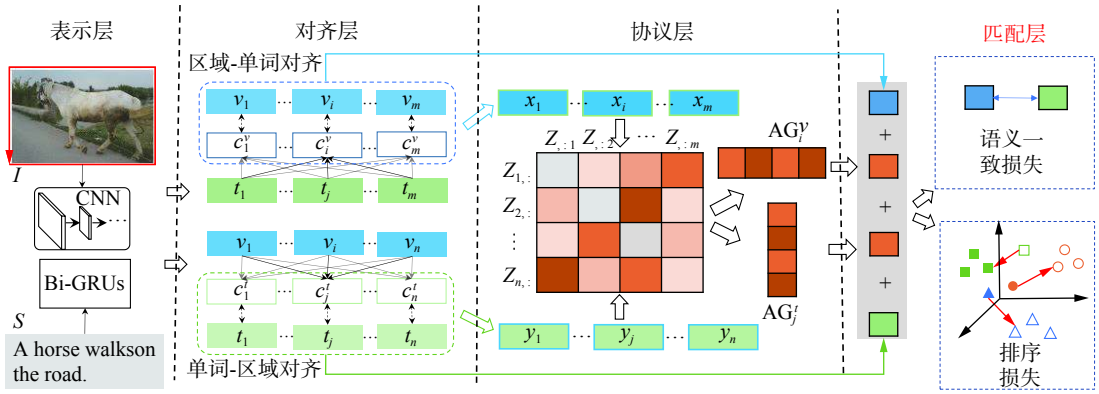


图 3　一致性协议匹配方法框架

Fig. 3　Framework of the proposed MAG method

3.1　表示层

在表示层，本文的目标是在潜在共享特征空间中对图像和文本的信息进行特征表达，以估计它们之间的相似性。给定一个包含 N 个图像文本对的数据集 $\mathcal{D}=\left\{(I_k,S_k)_{k=1}^{N}\right\}$，其中，$I$ 代表图像，S 代表文本。本文使用两个不同的编码器分别提取图像的视觉信息和文本的文本信息。

具体来说，给定图像 I，本文使用一个预训练的对象检测模型 Faster R-CNN[13] 来充当图像编码器。Faster R-CNN 会推断出图像中的显著性对象信息，并以包围框的方式定位到对象的区域，记为 r_i，并将经过区域池化操作得到对象的特征记为 f_i。接着，本文使用一个线性变换层将区域特征映射到一个 d 隐层特征空间中：

$$v_i = W_v f_i + b_v \tag{1}$$

式中：v_i 是区域 r_i 在隐层特征空间中的特征表示；W_v 和 b_v 是线性变换的可学习参数。为了方便描述，假定图像 I，Faster R-CNN 检测到 m 个对象区域，则最终可以得到 m 个特征来表示图像 I，本文用 $V=\{v_i|i=1,2,\cdots,m;v_i\in R^d\}$ 来表示图像特征。

给定有 n 个词的文本 $S=\{w_1,w_2,\cdots,w_n\}$，本文使用一个双向门控循环单元 (bidirectional GRU, Bi-GRU) 充当文本编码器。首先，本文首先将每个离散的单词 w_j 使用独热码进行表示，接着使用一个可学习的嵌入矩阵将独热码转化为一个词向量 $\boldsymbol{e}_j$。然后使用一个 Bi-GRU 分别从左到右 (前向) 和从右到左 (后向) 两个方向对词向量进行处理：

$$\underset{\boldsymbol{h}_j}{\rightarrow}=\underset{\mathrm{GRU}}{\rightarrow}\left(\boldsymbol{e}_j,\underset{\boldsymbol{h}_{j-1}}{\rightarrow}\right);\underset{\boldsymbol{h}_j}{\leftarrow}=\underset{\mathrm{GRU}}{\leftarrow}\left(\boldsymbol{e}_j,\underset{\boldsymbol{h}_{j+1}}{\leftarrow}\right) \tag{2}$$

式中 $\overrightarrow{\boldsymbol{h}}$ 和 $\overleftarrow{\boldsymbol{h}}$ 分别表示前向 GRU 和后向 GRU 的隐藏状态向量。最后，单词 w_j 的上下文表示可以通过公式获得：$t_j=(\overrightarrow{\boldsymbol{h}}_j+\overleftarrow{\boldsymbol{h}}_j)/2$。为了使单词特征 t_j 和图像区域特征 v_i 可以在特征空间中进行计算，这里设置 Bi-GRU 的隐藏状态向量维度和 v_i 一样，有 $t_j\in\mathbf{R}^d$。为了方便下文描述，本文用 $T=\{t_j|j=1,2,\cdots,n;t_j\in\mathbf{R}^d\}$ 来表示文本 S 的特征。

3.2　对齐层

对齐层旨在探索视觉信息和文本信息之间的

细粒度关联关系。和前人工作[14]一样，我们采用双向注意力机制将图像中的区域和文本中的单词巧妙地对齐。具体来说，给定图像特征 V 和文本特征 T，首先计算图像中每个区域特征 $\boldsymbol{v}_i$ 和每个单词特征 t_j 之间的相关性：具体来说，给定图像特征 V 和文本特征 T，首先计算图像中每个区域特征 $\boldsymbol{v}_i$ 和每个单词特征 t_j 之间的相关性：

$$\boldsymbol{A}_{ij}=\frac{\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{t}_j}{\|\boldsymbol{v}_i\|\cdot\|\boldsymbol{t}_j\|},\quad \forall i\in[1,m],\forall j\in[1,n] \tag{3}$$

这里，本文用余弦距离来度量两个向量间的相似性，向量上标表示向量转置。$\boldsymbol{A}\in\mathbf{R}^{m\times n}$ 为相似性矩阵。本文使用注意力机制计算每个区域的文本上下文特征和每个单词的图像上下文特征。对于区域 r_i，在相似性矩阵 $\boldsymbol{A}$ 中的第 i 行表示该区域和文本 S 的每个单词的相似性，为此，对应的文本上下文特征可以对文本特征 T 和相似性 $\boldsymbol{A}_i$ 进行加权和得到：

$$c_i^t=\sum_{j=1}^{n}\alpha_{ij}\boldsymbol{t}_j,\quad \alpha_{ij}=\frac{\exp(\lambda a_{ij})}{\sum_k\exp(\lambda a_{ik})} \tag{4}$$

式中：λ 是一个温度因子；c_i^t 是区域 r_i 对应的文本上下文特征；a_{ij} 是相似性矩阵 $\boldsymbol{A}$ 按列规范后的元素，即

$$a_{ij}=\sigma(\boldsymbol{A}_{ij})\Big/\sqrt{\sum_k\sigma(\boldsymbol{A}_{kj})^2} \tag{5}$$

同理，可以为每个单词计算它对应的图像上下文特征：

$$c_j^v=\sum_{i=1}^{m}\beta_{ij}\boldsymbol{v}_i,\quad \beta_{ij}=\frac{\exp(\lambda b_{ij})}{\sum_k\exp(\lambda b_{kj})} \tag{6}$$

式中：c_j^v 是单词 w_j 对应的图像上下文特征；b_{ij} 是相似性矩阵 $\boldsymbol{A}$ 按行规范后的元素，即

$$b_{ij}=\frac{\sigma(\boldsymbol{A}_{ij})}{\sqrt{\sum_k\sigma(\boldsymbol{A}_{ik})^2}} \tag{7}$$

和 Chen 等[14]工作一样，给定一个图像文本对 (I,S)，可以通过聚合每个区域特征和其对应的文本上下文特征的相似性以及聚合每个文本特征和其对应的图像上下文特征的相似性得到图像和文本的相似性：

$$F_{\mathrm{aln}}(I,S)=\frac{1}{m}\sum_i^m\frac{\boldsymbol{v}_i^{\mathrm{T}}c_i^t}{\|\boldsymbol{v}_i\|\cdot\|c_i^t\|}+\frac{1}{n}\sum_j^n\frac{\boldsymbol{t}_j^{\mathrm{T}}c_j^v}{\|\boldsymbol{t}_j\|\cdot\|c_j^v\|} \tag{8}$$

因为 $(\boldsymbol{v}_i,c_i^t)$ 和 (t_j,c_j^v) 是成对存在的，分别表示区域−单词对齐和单词−区域对齐，因此这里定义 $F_{\mathrm{aln}}(I,S)$ 为图像文本对 (I,S) 的对齐分数。

3.3　协议层

从式 (4) 和式 (6) 可以看出，对齐层利用相似性矩阵 $\boldsymbol{A}$ 的不同维度来计算注意力权重，使得同一个区域和单词计算得到的区域−单词对齐和单词−区域对齐可能被赋予不同的重要性，导致对齐不一致 (如图 1 所示)。本文旨在利用这种不一致的特点来强化对图像和文本的相似性的建模。为此，本文提出了一种基于协议的匹配策略，以利用这种对齐不一致的特点。本文首先将对齐层的对齐操作进行特征实例化，并使用竞争性投票的策略将不同对齐在特征空间中进行一致性度量，度量结果作为协议层的输出，表征图像和文本之间的一致性分数。

具体来说，首先定义对齐操作的特征表示为每个区域或者单词和其对应上下文特征的加和：

$$\boldsymbol{x}_i=\boldsymbol{v}_i+\boldsymbol{c}_i^t,\quad \boldsymbol{y}_j=\boldsymbol{t}_j+\boldsymbol{c}_j^v \tag{9}$$

式中：$\boldsymbol{x}_i$ 表示区域−单词对齐 $(\boldsymbol{v}_i,\boldsymbol{c}_i^t)$ 的特征表示；$\boldsymbol{y}_j$ 表示单词−区域对齐 $\left(\boldsymbol{t}_i,\boldsymbol{c}_j^v\right)$ 的特征表示。遍历 i 和 j，可以得到一组区域−单词对齐特征实例 $X=\{\boldsymbol{x}_i|i=1,2,\cdots,m,\boldsymbol{x}_i\in\mathbf{R}^d\}$ 和单词−区域对齐特征 $Y=\{\boldsymbol{y}_j|j=1,2,\cdots,n,\boldsymbol{y}_j\in\mathbf{R}^d\}$。

其次，使用余弦距离来衡量两种对齐特征的相似性：

$$Z_{ij}=\frac{\boldsymbol{x}_i^T\boldsymbol{y}_j}{\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{y}_j\|},\quad \forall i\in[1,m],\forall j\in[1,n] \tag{10}$$

式中：Z_{ij} 衡量以区域 r_i 为核心的区域−单词对齐特征和以单词 w_j 为核心的单词−区域对齐特征之间的相似性。如果区域 r_i 和单词 w_j 在对方的对齐方式中同等重要，即对齐一致，那么 Z_{ij} 会很大，反之，则是对齐不一致的问题，则 Z_{ij} 会较小。因此，Z_{ij} 刻画了两种对齐方式是否一致，我们称式 (10) 为协议操作。

为了鼓励一致性的局部对齐 (即区域−单词对齐和单词−区域对齐) 能够在后续计算图像和文本的相似性过程中被赋予更高的重要性，本文对 Z_{ij} 的每一行和每一列分别选取最大值，得到两种协议得分：

$$\mathrm{AG}_i^v=\max_j Z_{ij},\quad \mathrm{AG}_j^t=\max_i Z_{ij} \tag{11}$$

按行取最值可以让每个单词−区域对齐互相竞争，胜者跟区域−单词对齐 $(\boldsymbol{v}_i,\boldsymbol{c}_i^t)$ 最一致。同理，按列取最值可以竞争出跟单词−区域对齐 $(\boldsymbol{t}_i,\boldsymbol{c}_j^v)$ 最一致的区域−单词对齐。

最后，本文将所有的协议得分进行平均，得到图像和文本的协议分数：

$$F_{\mathrm{agr}}(I,S)=\frac{1}{m}\sum_i^m\mathrm{AG}_i^v+\frac{1}{n}\sum_j^n\mathrm{AG}_j^t \tag{12}$$

对比对齐分数 (见式 (8))，协议分数 $F_{\mathrm{agr}}(I,S)$ 可以看成区域和单词的二阶对齐分数，因此作为

对齐分数的补充, 可以更好地衡量图像和文本之间的相似性。

3.4　匹配层

匹配层的目的是累积所有匹配线索以估计图像和文本之间的相似性。本文将对齐层的对齐分数和协议层的协议分数结合起来计算给定图像–文本对 (I,S) 的相似度:

$$F(I,S)=F_{\text{aln}}(I,S)+F_{\text{agr}}(I,S) \tag{13}$$

训练时, 本文采用 Faghri 等[2] 提出的基于难负例的三元组排序损失函数来训练模型:

$$\begin{aligned}\mathcal{L}_{\text{rank}}(I,S)=&[\Delta-F(I,S)+F(I,S')]_+ +\\ &[-F(I,S)+F(I',S)]_+\end{aligned} \tag{14}$$

式中: (I,S) 表示一对正例样本; I' 和 S' 分别是文本 S 和图像 I 的负例样本; $[x]_+=\max(0,x)$; Δ 表示排序间隔, 即希望查询样本和正例样本之间的相似性比查询样本和最难负例样本之间的相似性大一个 Δ。

此外, 考虑到在协议层中本文希望能够挖掘更多的一致性对齐来增强对图像和文本相似性的度量, 为此本文采用 Chen 等[14] 的方法引入语义一致性损失函数:

$$\mathcal{L}_{\text{aln}}(I,S)=\left(\frac{1}{m}\sum_{i}^{m}\frac{v_i^{\mathrm{T}}c_i^t}{\|v_i\|\cdot\|c_i^t\|}-\frac{1}{n}\sum_{j}^{n}\frac{t_j^{\mathrm{T}}c_j^v}{\|t_j\|\cdot\|c_j^v\|}\right)^2 \tag{15}$$

在训练过程中, 从数据集中采样一批次图像文本对进行训练, 即 $\{(I_k,S_k)\}^{N_b}\sim\mathcal{D}$, 最终的损失函数是排序损失和一致性损失的加权和:

$$\mathcal{L}=\sum_{k}^{N_b}\mathcal{L}_{\text{rank}}(I_k,S_k)+\lambda_{\text{aln}}\sum_{k,l}^{N_b}\mathcal{L}_{\text{aln}}(I_k,S_l) \tag{16}$$

式中 λ_{aln} 是一个可调节平衡超参数。

4　有效性验证

4.1　实验配置

1) 数据集。本文采用了两个跨模态图像文本标准基线数据集来验证所提出的一致性协议匹配方法。①Flickr30K[15]。这个数据集由 31 000 张图片组成, 每张图片都至少标注了 5 个英文文本。本文采用 29 000 张图片作为训练集, 1 000 张图片作为验证集, 剩下的 1 000 张图片作为测试集, 这也是标准的数据划分。②MSCOCO[16]。COCO 数据集大概有 123 000 张图片, 每张图片标注了至少 5 个英文句子。和前人工作[1]一样, 本文将 123 287 张图片划分为 113 287、5 000 和 5 000, 分别构成了训练集、验证集和测试集。为了能够公平地评价模型的结果以及跟别人的工作进行对比, 本文同时展示在 5 000 张测试图片上的整体性能 (用 MS COCO(5K) 表示) 以及 5 次实验 (每次 1 000 张图片 (用 MS COCO(1K) 表示) 的平均值。

2) 评价指标。本文进行了图像检索文本和文本检索图像两类不同的检索任务, 采用前 K 召回率 (R@K) 来评测两种跨模态检索性能, 并且和前人的工作进行对比。具体来说, 本文展示 R@1、R@5 和 R@10 的结果, 并且, 跟 Chen 等[14] 的工作一样, 本文将所有的指标加起来来综合评价模型的性能, 该指标用 R@sum 表示。

3) 实现细节。本文使用 Pytorch1.0[17] 来实现所提出的方法。在构建模型时, 本文将图像区域特征的维度设置为 1 024。Bi-GRU 的隐藏向量的维度也是 1 024, 使得图像区域特征和单词特征的维度一致。训练过程中, 三元组损失函数中的排序间隔设置为 0.2, 即式 (14) 中的 Δ 默认为 0.2, 式 (4) 中的温度因子 λ 默认设置为 9。在更新网络参数时, 本文采用 Adam[18] 优化器来优化, 并且每批次数据容量为 128 张图像文本对。

4.2　模型对比分析

为了验证所提出的一致性协议匹配方法的先进性, 本文引入了目前相关的先进的跨模态图像文本检索方法, 并在 Flickr30K 和 MS COCO 两个数据集上都进行了模型对比。本文对比的基线模型有 DVSA[3]、VSE++[2]、DPC[19]、SCO[20]、SCAN[6]、PFAN[21]、PVSE[22] 和 SC[14]。其中, SCAN、PFAN、PVSE 和 SC 跟本文一样, 都是致力于挖掘图像和文本之间的细粒度跨模态关联来提升跨模态检索性能。表 1、2、3 分别给出了本文的方法和基线模型在 Flickr30K 和 MS COCO 上的对比结果, 其中, 表格中第 1 列中带*标记的方法表示该结果是采用模型集成的结果, "—"表示该结果未在原始论文中给出。

表 1　Flickr30K 上对比结果

Table 1　Comparison with state-of-the-art methods on Flickr30K

方法	图像检索文本			文本检索图像			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA	22.2	48.2	61.4	15.2	37.7	50.5	235.2
VSE++	52.9	—	87.2	39.6	—	795	—
DPC	55.6	81.9	89.5	39.1	69.2	80.9	416.2
SCO	55.5	82.0	89.3	41.1	70.5	80.1	418.0
SCAN*	67.4	90.3	95.8	48.6	77.7	85.2	465.0
PFAN*	70.0	91.0	95.0	50.4	78.7	86.1	472.0
SC*	69.7	91.7	96.4	54.0	79.7	87.2	478.7
MAG(本文)	**72.1**	**92.8**	**96.7**	**52.8**	**80.2**	**87.1**	**481.8**
MAG*(本文)	**74.4**	**93.0**	**96.8**	**54.3**	**81.0**	**87.9**	**487.4**

表 2　MS COCO(1K) 上对比结果

Table 2　Comparison with state-of-the-art methods on MS COCO(1K)

方法	图像检索文本			文本检索图像			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DVSA	38.4	69.9	80.5	27.4	60.2	74.8	351.2
VSE++	64.6	—	95.7	52.0	—	92.0	—
DPC	65.6	89.8	95.5	47.1	79.9	90.0	467.9
SCO	69.9	92.9	97.5	56.7	87.5	94.8	499.3
SCAN*	72.7	94.8	98.4	58.8	88.4	94.8	507.9
PVSE	69.2	91.6	98.3	55.2	86.5	93.7	492.8
SC*	73.8	95.3	98.3	59.9	88.9	94.9	511.1
MAG(本文)	**75.2**	**95.4**	**98.3**	**59.1**	**87.9**	**94.3**	**510.2**
MAG*(本文)	**76.1**	**95.7**	**98.5**	**60.6**	**88.9**	**95.8**	**514.8**

表 3　MS COCO（5K）上对比结果

Table 3　Comparison with state-of-the-art methods on MS COCO(5K)

方法	图像检索文本			文本检索图像			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	41.3	—	81.2	30.3	—	72.4	—
DPC	41.2	70.5	81.1	25.3	53.4	66.4	337.9
SCO	42.8	72.3	83.0	33.1	62.9	75.5	369.6
SCAN*	50.4	82.2	90.0	38.6	69.3	80.4	410.0
PVSE	45.2	74.3	84.5	32.4	63.0	75.0	374.4
MAG(本文)	**52.0**	**81.3**	**90.0**	**37.2**	**65.4**	**77.9**	**404.8**
MAG*(本文)	**54.1**	**82.7**	**90.8**	**38.6**	**67.8**	**79.0**	**413.0**

从表 1、2、3 中可以看到，本文提出的一致性协议匹配方法在两个数据集上都取得了比基线模型更优异的跨模态图像文本检索性能。具体来说，1) 在 Flickr30K 上，本文的 MAG*取得了比最好的基线模型 SC*更好的性能，特别是在图像检索文本任务的 R@1 上提高了 4.7%，在文本检索图像的 R@1 上取得了 0.3% 的性能提升，整体上提升了 8.7%(R@sum)；2) 在 MS COCO(1K) 中，本文的 MAG*获得了更先进的性能，与 SC* 相比，本文的 MAG*在图像检索文本任务的 R@1 上可以获得 76.1% 的召回率，提高了 2.3%，在文本检索图像任务的 R@1 上，MAG*取得了 60.6% 的性能，提升了 0.7%；3) 在 MS COCO(5K) 中，本文的 MAG*在大多数评价指标上也获得了优于最佳基线 SCAN*的卓越性能。相比于 SCAN*，MAG*在图像检索文本任务上最多可以获得 3.7%(R@1) 的性能提升，整体提升 3%。这些结果证明了所提出的一致性匹配方法的有效性。

4.3　模块分析

本节对所提出的一致性协议匹配方法中的重要因素进行分析。首先分析匹配层中语义一致性损失的作用。语义一致性损失可以驱动两个独立注意力模块的对齐保持一致，有助于后续基于协议的匹配过程。因此，本文通过调节式 (16) 中的 λ_{aln} 超参数来观察模型性能的变化，结果展示在表 4 中。从表 4 可以看出，当没有对模型施加语义一致性约束时 ($\lambda_{\text{aln}}=0$)，模型的性能较低，当 $\lambda_{\text{aln}}>0$,模型都有一定的性能的提升，当 $\lambda_{\text{aln}}=1$ 时，模型取得了最好的性能。

表 4　语义一致性损失函数的影响

Table 4　Effect of the semantic consistency objective

λ_{aln}	图像检索文本		文本检索图像	
	R@1	R@10	R@1	R@10
0.0	70.6	96.6	50.6	85.6
0.5	71.2	96.7	52.0	86.8
1.0	**72.1**	**96.7**	**52.8**	**87.1**
1.5	70.9	96.2	52.3	86.7

本文接着对图像和文本的相似性度量进行分析。为了观察所提出的协议层的影响，本文通过逐步擦除 $F_{\text{agr}}(I,S)$ 的组成来分析各个项对模型的影响。分析结果见表 5，第 1 行是本文提出的 MAG 的默认使用方式，即 $F_{\text{aln}}+F_{\text{agr}}$，第 2 行是去掉了 F_{agr} 中的右边一项，只保留 AG_i^v 那一项 (见式 (12))，第 3 行是去掉了 F_{agr} 中的左边一项，只保留 AG_j^t 那一项，最后一行是把 F_{agr} 全部去掉得到的模型，即去掉整个协议层。可以看出，跟去掉协议层的模型 (最后一行) 相比，不管是仅保留 AG_i^v、仅保留 AG_j^t 还是两者都保留，只要有协议层存在，模型都能取得显著的性能提升，特别是，AG_i^v 和 AG_j^t 都保留的话，模型取得了最好的性能。这些结果显示了所提出的一致性协议匹配的有效性。

4.4　实例分析

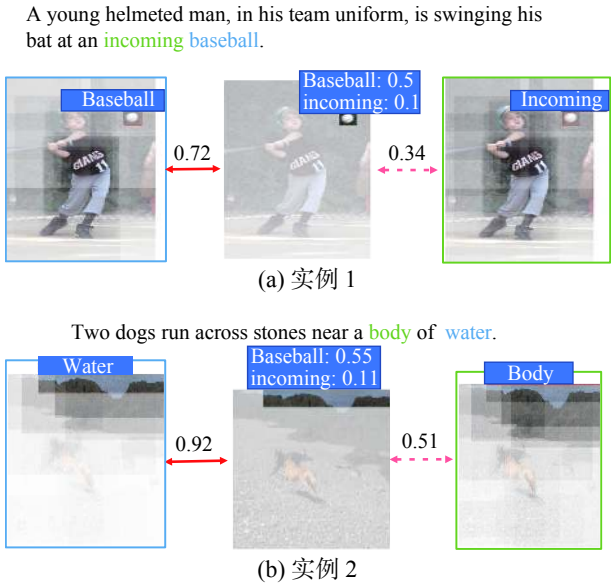
本文对模型进行进一步的实例分析。在图 4 中，本文展示了两个实例，在每个可视化示例中，分别在左侧和右侧的图像展示了给定文本中的两个单词 (分别用蓝色和绿色标记) 及其在图像区域上的注意力结果，这种注意力结果可以被看作是单词–区域对齐。对于中间的图像，本文展示了一个显著性区域跟文本中单词的注意力结果，这可以看作是区域–单词对齐。这里用红色的双

向箭头表示两种对齐之间的一致性得分。

表 5　协议层的影响

Table 5　Effect of the agreement layer

$F(I,S)$	图像检索文本		文本检索图像	
	R@1	R@10	R@1	R@10
$F_{\text{aln}}+F_{\text{agr}}$	**72.1**	**96.7**	**52.8**	**87.1**
$F_{\text{aln}}+\text{AG}_i^v$	70.7	96.0	50.8	86.1
$F_{\text{aln}}+\text{AG}_j^t$	70.8	96.2	52.3	86.4
F_{aln}	69.7	95.8	51.1	85.9



(a) 实例 1

(b) 实例 2

图 4　一致性协议匹配实例分析

Fig. 4　Examples of the proposed MAG method

可以看到，在两个单词–区域对齐 (左和右) 中，对应的词都与红色框中的区域高度相关。而在区域–单词对齐 (中间) 中，左侧的单词与区域的相关性更高，导致左侧的对齐一致性得分高于右侧。例如，在图 4(a) 中，左边和右边的图像中的注意力结果表明“Baseball”和“incoming”都与红框中的区域有很强的相关性。然而，中间的图像显示红色区域与单词“Baseball”的相关性高于单词“incoming”，使得左侧单词–区域对齐和中间区域–单词对齐之间的一致性得分更高。未来，本文将继续在跨模态行人再识别[23-24]，跨模态哈希检索[25] 等其他跨模态任务挖掘这种关联一致性问题，并将本文的方法进行应用扩展，促进跨模态学习的发展。

5　结束语

本文针对跨模态图像文本任务提出了一种一致性协议匹配方法。与之前的工作一样，首先使用注意力机制充分探索了图像中区域和文本中单词之间的单词–区域和区域–单词的对齐方式，接着提出跨模态协议来估计对齐的一致性。本文将协议的推导过程实例化为模型的协议层，并采用了一种新颖的竞争性投票方案，为细粒度跨模态关联关系提供强有力的协议准则，促进模型对图像文本之间的相似性的准确建模。本文在两个基准数据集 (Flickr30K 和 MS COCO) 上进行了广泛的实验。实验结果表明，本文提出的方法取得了先进的跨模态图像文本检索性能，很好地验证了方法的有效性。

参考文献：

[1] WANG Liwei, LI Yin, LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 5005−5013.

[2] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives [EB/OL]. (2018-07-29)[2021-07-30] https://arxiv.org/pdf/1707.05612.

[3] KARPATHY A, LI Feifei. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3128−3137.

[4] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2156−2164.

[5] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. Sydney, Australia, 2015: 2048−2057.

[6] LEE K H, CHEN Xi, HUA Gang, et al. Stacked cross attention for image-text matching[M]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. Proceedings of the 15th European Conference on Computer Vision-ECCV 2018. Munich, Germany: Springer, 2018: 201−216.

[7] FROME A, CORRADO G S, SHLENS J, et al. DeViSE: A deep visual-semantic embedding model[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Nevada, USA, 2013: 2121–2129.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10). https://arxiv.org/pdf/1409.1556.

[9] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-09-07)[2021-07-30] https://arxiv.org/pdf/1301.3781.

[10] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unify-

ing visual-semantic embeddings with multimodal neural language models[EB/OL]. (2014-11-10). https://arxiv.org/pdf/1411.2539.

[11] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2021-07-30] https://arxiv.org/pdf/1412.3555.

[12] NIU Zhenxing, ZHOU Mo, WANG Le, et al. Hierarchical multimodal LSTM for dense visual-semantic embedding[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1899−1907.

[13] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 91−99.

[14] 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. Nice, French, 2019: 1749−1757.

[15] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the association for computational linguistics, 2014, 2(1): 67–78.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//13th European Conference on Computer Vision-ECCV 2014. Zurich, Switzerland, 2014: 740−755.

[17] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in PyTorch[C]//31st Conference on Neural Information Processing Systems. Long Beach, USA, 2017.

[18] KINGMA D P, BA J L. Adam: A method for stochastic optimization[EB/OL]. (2015-04-23)[2021-08-01] https://arxiv.org/pdf/1412.6980.

[19] 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.

[20] HUANG Yan, WANG Wei, WANG Liang. Instance-aware image and sentence matching with selective multimodal LSTM[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 2310−2318.

[21] WANG Yaxiong, YANG Hao, QIAN Xueming, et al. Position focused attention network for image-text matching[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China, 2019: 3792−3798.

[22] SONG Yale, SOLEYMANI M. Polysemous visual-semantic embedding for cross-modal retrieval[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019.

[23] 陈丹, 李永忠, 于沛泽, 等. 跨模态行人重识别研究与展望 [J]. 计算机系统应用, 2020, 29(10): 20–28.
CHEN Dan, LI Yongzhong, YU Peizhe, et al. Research and prospect of cross modality person re-identification[J]. Computer systems & applications, 2020, 29(10): 20–28.

[24] 刘天瑜, 刘正熙. 跨模态行人重识别研究综述 [J]. 现代计算机, 2021, 27(7): 135–139.
LIU Tianyu, LIU Zhengxi. Overview of cross modality person Re-identification research[J]. Modern computer, 2021, 27(7): 135–139.

[25] 姚伟娜. 基于深度哈希算法的图像—文本跨模态检索研究 [D]. 北京: 北京交通大学, 2018.
YAO Weina. Image-text cross-modal retrieval based on deep hashing method[D]. Beijing: Beijing Jiaotong University, 2018.

作者简介：

宫大汉，博士研究生，主要研究方向为图像语义理解、卷积神经网络压缩加速。

陈辉，助理研究员，博士，主要研究方向为图像语义理解、多媒体信息处理。

丁贵广，副教授，博士，主要研究方向为多媒体信息处理、计算机视觉感知。主持基金委重点项目、重点研发项目等国家级项目数十项。曾获国家科技进步二等奖、吴文俊人工智能科技进步一等奖、中国电子学会技术发明一等奖等。发表学术论文近百篇，引用量近 7 000 次。