计算机科学
COMPUTER SCIENCE

基于渐进原型匹配的文本-动态图片跨模态检索算法

彭姣, 贺月, 商笑然, 胡塞尔, 张博, 常永娟, 欧中洪, 卢艳艳, 姜丹, 刘亚铎

引用本文

彭姣, 贺月, 商笑然, 胡塞尔, 张博, 常永娟, 欧中洪, 卢艳艳, 姜丹, 刘亚铎. 基于渐进原型匹配的文本-动态图片跨模态检索算法[J]. 计算机科学, 2025, 52(9): 276-281.

PENG Jiao, HE Yue, SHANG Xiaoran, HU Saier, ZHANG Bo, CHANG Yongjuan, OU Zhonghong, LU Yanyan, JIANG dan, LIU Yaduo. Text-Dynamic Image Cross-modal Retrieval Algorithm Based on Progressive Prototype Matching [J]. Computer Science, 2025, 52(9): 276-281.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200204

基于渐进原型匹配的文本-动态图片跨模态检索算法

彭 姣[1] 贺 月[1] 商笑然[2] 胡塞尔[2] 张 博[1] 常永娟[1] 欧中洪[3] 卢艳艳[1] 姜 丹[1] 刘亚铎[1]

1 国网河北省电力有限公司信息通信分公司 石家庄 050000
2 北京邮电大学计算机学院 北京 100876
3 北京邮电大学网络与交换技术全国重点实验室 北京 100876
(p2010015645@163.com)

摘 要 在社交和聊天场景中,用户不再局限于使用文字或 emoji 表情符号,而是采用语义更加丰富的静态或动态图片来进行交流。尽管现有的文本-动态图片检索算法取得了一定效果,但仍存在模态内和模态间缺乏细粒度交互,以及原型生成过程中缺乏全局引导的问题。为了解决上述问题,提出了一种全局敏感的渐进原型匹配模型(Global-aware Progressive Prototype Matching Model,GaPPMM)用于文本-动态图片跨模态检索,采用三阶段渐进原型匹配的方法来实现跨模态细粒度交互,并提出了全局敏感的时间原型生成方法,利用全局分支产生的预览特征作为注意力机制的查询,引导局部分支关注到最相关的局部特征,实现了动态图片的细粒度特征提取。实验结果表明,提出的模型在公开数据集上的召回率总和超越了现有的 SOTA 模型。

关键词:跨模态检索;动态图片检索;渐进原型匹配;注意力机制;全局敏感性分析
中图分类号 TP391

Text-Dynamic Image Cross-modal Retrieval Algorithm Based on Progressive Prototype Matching

PENG Jiao[1],HE Yue[1],SHANG Xiaoran[2],HU Saier[2],ZHANG Bo[1],CHANG Yongjuan[1],OU Zhonghong[3],LU Yanyan[1],JIANG dan[1] and LIU Yaduo[1]

1 Information & Telecommunications Branch,State Grid Hebei Electric Power Co.,Ltd.,Shijiazhuang 050000,China
2 School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876,China
3 State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract In social and chat scenes,users are no longer limited to using text or simple emoji,but are more inclined to use static or dynamic images with richer semantic meaning to communicate. Although existing text-dynamic image retrieval algorithms have been achieved,there are still problems such as lack of fine-grained intra-modal and inter-modal interactions,and lack of global guidance in the prototype generation process. In order to solve the above problems,this paper proposes a Global-aware Progressive Prototype Matching Model(GaPPMM) for text-dynamic image cross-modal retrieval. A three-stage progressive prototype matching method is used to achieve cross-modal fine-grained interaction. In addition,a globally sensitive temporal prototype generation method is proposed,which uses the preview features generated by the global branch as the query of the attention mechanism to guide the local branch to pay attention to the most relevant local features,so as to realize the fine-grained feature extraction of dynamic images. The experimental results demonstrate that the proposed model surpasses state-of-the-art in terms of recall rate on the publicly available dataset.

Keywords Cross-modal retrieval,Text-dynamic image retrieval,Progressive prototype matching,Attention mechanisms,Global sensitivity analysis

1 引言

近年来,随着微信、微博等社交平台和抖音、快手等短视频平台的兴起,数千万用户每天向互联网上传了大量的静态图片、动态图片和短视频等多模态数据。在社交和聊天场景中,用户(尤其是年轻用户)不再仅仅局限于使用文字或者简单的 emoji 表情符号,而是更倾向于采用语义更加丰富的静态 JPG 或者动态 GIF 图片来交流和表达自己的情绪和态度,

到稿日期:2024-12-30 返修日期:2025-03-31
基金项目:国网河北省电力有限公司(SGHEXT00SJJS2310134)
This work was supported by the State Grid Hebei Information and Telecommunication(SGHEXT00SJJS2310134).
通信作者:欧中洪(zhonghong.ou@bupt.edu.cn)

这使得表情包图片的数量在互联网中快速增长。随着用户习惯的变化，互联网上出现了各种表情包生成和定制工具，用户可以制作个性化的静态或动态表情图片，一些互联网厂商甚至聘请专业人员设计并发布表情包图片。由于可供用户选择的表情包数量呈几何倍数增加，因此在社交聊天场景中，用户根据需要准确找到自己想要的表情包图片变得越来越困难，一些用户常常遇到在大量候选集中找不到自己需要的表情包图片的问题。因此，如何分析用户的文本查询，从大量包含静态和动态图片的候选集中快速检索出关联性最强的表情包图片成为一个亟需解决的问题。

目前，文本-动态图像的检索方法虽然取得了一定进展，实现了较高的检索召回率，但缺乏模态内和跨模态细粒度交互，没有充分实现文本-动态图片的细粒度匹配。此外，现有的检索模型的原型生成过程完全自下而上进行，缺少全局理解的引导，不能有效关注到最相关的局部特征来去除局部噪声，未能充分实现细粒度匹配。针对上述问题，本文提出了一种基于渐进原型匹配的跨模态检索算法，主要贡献如下：

1）提出了全局敏感的渐进原型匹配模型（GaPPMM），利用目标-短语原型匹配、事件-句子原型匹配和全局原型匹配的三阶段匹配方法，实现了跨模态细粒度交互。

2）提出了全局敏感的时间原型生成方法，利用注意力机制和全局分支产生的全局预览特征作为查询，引导局部分支关注到最相关的局部特征，去除局部噪声，充分实现动态图片的细粒度特征提取。

3）设计并实施了大量实验，模型在公开数据集 TGIF[1] 和 Taiwan[2] 上的检索指标超过现有 SOTA 模型，验证了所提方法的有效性。

2　相关工作

早期的跨模态检索方法，针对图像和文本间复杂的语义交互作用，主要采用统计分析方法，如典型相关性分析方法和跨模态因子分析方法，但该类方法对实际应用场景中不同模态数据的复杂相关性难以建模。随着深度学习的兴起，基于语义表征的图文跨模态检索方法得到了广泛关注，这些方法利用深度学习模型替代了基于统计的人工定义特征，利用模型挖掘图像和文本中的语义信息，解决了文本图像两种模态由于异质特性带来的语义鸿沟问题。

2.1　文本-图像跨模态检索技术

文本到图像检索，即给定一个文本查询，从图像候选集中检索出与该查询语义相关的图像，反之亦然。直觉上，可以简单地使用一个图像帧来代表整个动态图像，从而把文本-动态图像检索任务转换为文本-图像跨模态检索任务。随着 Transformer 在自然语言处理领域取得巨大成功，不少研究者提出了基于 Transformer 的文本-图像跨模态检索算法。Chen 等[3] 提出了基于循环注意记忆的迭代匹配方法，通过迭代匹配方法逐步更新跨模态注意力核心。Zhang 等[4] 提出了语境感知注意力网络，根据全局上下文有选择地关注信息量最大的局部片段，综合了模态间和模态内注意过程。Zheng 等[5] 提出了 CABIR 模型，该模型基于遥感图像的区域级语义特征和跨注意力机制来提升跨模态文本-图像检索的性能。Song 等[6] 提出了一个多义视觉语义嵌入（Polysemous Visual-Semantic Embedding，PVSE）模型，该模型将每种模态映射到多个表示空间，允许表达文本或图像的多种含义，从而捕捉到文本和 GIF 之间的不同对应关系。Wang 等[7] 融合了 GIF 的 3 条信息，包括 GIF 标题、对象名称和区域特征，以增强 GIF 的视觉表现力，并最大限度地学习文本和 GIF 之间的语义一致性。这些工作虽然没有利用 GIF 标记信息，但是标记信息可以在此类检索场景中提供“免费”监督。最近几年，随着文本-图像预训练模型的出现，更多研究者将目光转移到大规模预训练模型上。Li 等[8] 提出了一个通用的用于图文表示的模型 Unicoder-VL，通过预训练方式让模型学习视觉和语言的联合表示。该模型借鉴了跨语言预训练模型的思想，如 XLM[9] 和 Unicoder[10]，视觉和语言内容都被输入一个多层 Transformer[11] 中用于跨模态预训练。Li 等[12] 提出了一种预训练模型 OSCAR，使用图像中检测到的对象作为锚点，简化了语义对齐的学习过程。

虽然近年来提出的方法很好地解决了文本-图像检索面临的语义异质鸿沟问题，但上述方法不能很好地捕捉动态图片在时序上的语义特征，直接用于文本-动态图片检索时的准确率仍有待提升。

2.2　文本-视频跨模态检索技术

与文本-图像跨模态检索任务类似，文本-视频跨模态检索任务指的是给定一个文本，系统从视频候选集中检索出与该查询语义相关的视频，反之亦然。动态图像可以看作没有音频的视频数据，两者均可表示为一系列的图像帧，所以研究文本-视频跨模态检索任务有助于解决文本-动态图片跨模态检索任务。解决文本-视频跨模态检索任务的关键在于，找到一种特征提取方法和一种跨模态相似度计算方法来衡量查询与候选集间的语义相关性。随着概率图模型的兴起，Chen 等[13] 提出了 HGR 模型，使用图卷积网络，设计了 3 个分支来捕获视频中的 3 个分层语义层，分别负责捕获全局事件、局部动作和实体。Song 等[14] 提出了一个将视频建模为时空图的框架，其中节点对应视觉对象，边对应对象间的关系。随着基于大规模预训练模型的视频文本检索技术的快速发展，Miech 等[15] 提出了一个包含 1.36 亿网络教学视频片段的大规模数据集 HowTo100M，并在该数据集上进行了预训练，展示了大规模预训练在跨模态表征方面的巨大潜力。Luo 等[16] 研究了基于预训练 CLIP 的 3 种相似度计算机制，进一步在有噪声的 HowTo100M 数据集上对 CLIP 进行预训练，以学习更好的文本-视频嵌入。最近，Peng 等[17] 从博弈论的角度提出可以把视频帧和文本词视为合作博弈的参与者，以跨模态相似性度量作为合作博弈的特征函数，使用 Banzhaf 交互来表示任意一组特征之间的合作趋势。Dong 等[18] 提出了一个带有动态知识蒸馏的双分支学习框架，该框架利用大型视觉语言模型的知识作为教师来指导学生模型。

虽然上述方法使得文本-视频跨模态检索领域的研究取得了长足进步，但这些模型不能同时兼顾跨模态细粒度交互和检索效率，没有充分利用大规模预训练模型的知识，没有充分实现文本-动态图片的细粒度匹配。

3 方法详述

本文提出了一种新的基于渐进原型匹配的文本-动态图像跨模态检索算法，整体框架如图1所示，其中全局原型生成(Global Prototype Generation，GPG)模块负责生成全局原型，局部分支模块则利用全局原型进行细粒度特征提取。GPG模块通过处理动态图像的全局特征，生成全局原型，然后将全局原型传递给时间原型匹配模块和全局原型匹配模块，用于后续的匹配计算。空间原型生成(Spatial Prototype Generation，SPG)模块用于生成空间原型，时间原型匹配(Temporal Prototype Matching，TPM)模块用于计算时间原型的匹配度，全局原型匹配(Global Prototype Matching，GPM)模块用于计算全局原型的匹配度。与图像帧级别原型生成时直接使用随机初始化的方法不同，本文模型使用了全局-局部双分支设计，并利用多头跨模态注意力来进行第三阶段的全局特征匹配。该模型可捕捉文本-动态图像间的细粒度交互关系，从而更有效地捕捉图像帧的局部特征信息。

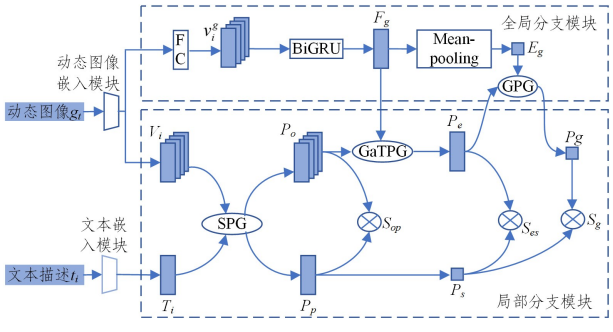

图1 基于渐进原型匹配的文本-动态图片跨模态检索结构

Fig. 1 Structure of text-dynamic image cross-modal retrieval based on progressive prototype matching

3.1 动态图像嵌入模块

动态图像嵌入模块是模型的基础部分，负责将动态图像转换为可用于后续处理的特征向量。对于动态图像的特征，本文使用预训练的Clip Vit-b/32模型作为教师模型的主干网络。对于一张给定的动态图片$\boldsymbol{v}_i$，从中均匀地提取L个图像帧作为关键帧，记为$\boldsymbol{f}_1^i,\boldsymbol{f}_2^i,\cdots,\boldsymbol{f}_L^i$。对于每一个图像帧$\boldsymbol{f}_l^i$，通过Clip图像编码器提取区域级图像特征，记为$\boldsymbol{f}_l^i=\{v_c,v_1,v_2,\cdots,v_k\}\in\mathbb{R}^{(K+1)*D}$，其中，$D$是特征向量的维度，$\boldsymbol{v}_c^l$是标记[CLS]对应的输出，表示图像帧的全局特征。最终，对于一张动态图像$\boldsymbol{v}_i$，得到一组动态图片特征向量：

$$\boldsymbol{V}_i=\{\boldsymbol{v}_c^l,\boldsymbol{v}_1^l,\boldsymbol{v}_2^l,\cdots,\boldsymbol{v}_k^l\}_{l=1}^{l=L}\in\mathbb{R}^{L*(k+1)*D} \tag{1}$$

3.2 文本嵌入模块

文本嵌入模块的作用是将输入的文本描述转换为能够与图像特征进行比较的特征向量。对于一段文本描述t_i，本文使用CLIP的文本编码器，得到文本特征向量：

$$\boldsymbol{T}_i=\{t_s,t_1,t_2,\cdots,t_m,t_e\}\in\mathbb{R}^{(M+2)*D} \tag{2}$$

其中，D是特征向量的维度，t_s和t_e分别是标记[SOT]和[EOT]对应的输出，t_e表示文本描述的全局特征。

3.3 局部分支模块

局部分支模块的主要作用是通过全局敏感的时间原型生成方法，利用全局分支提供的全局特征，引导局部分支关注最相关的局部特征，去除局部噪声，从而实现动态图像的细粒度特征提取。这种方法能够有效捕捉动态图像中的局部语义信息，提高跨模态检索的精度。

通过全局分支和局部分支的协同工作，模型能够实现跨模态的细粒度交互，从而更有效地捕捉动态图像的局部特征信息，提高检索精度。

3.4 全局分支模块

全局分支模块负责提供全局特征，为局部分支的细粒度特征对齐提供全局引导。本文使用了轻量化的模型作为全局分支，为局部分支的细粒度特征对齐提供全局引导。对于上文提到的特征$\boldsymbol{V}_i$，利用线性层映射得到$\boldsymbol{V}_i^g\in\mathbb{R}^{L*(k+1)*D}$，利用双向门控循环单元(Bidirectional Gated Recurrent Unit，BiGRU)来实现对图像帧序列的全局特征提取。对于每一个时间步t，输入$\boldsymbol{v}_k^l$，输出$\boldsymbol{h}_t$，每个时间步的输出视为当前帧的全局特征：

$$\boldsymbol{F}_g=\{\boldsymbol{h}_t\}_{t=1}^{L}\in\mathbb{R}^{L*D} \tag{3}$$

接着，使用平均池化层获得整张图片的全局特征：

$$\boldsymbol{E}_g=\frac{1}{L}\sum_{t=1}^{L}\boldsymbol{h}_t\in\mathbb{R}^{D} \tag{4}$$

3.4.1 空间原型生成

给定一组动态图片的特征$\boldsymbol{V}_i$，对于其中每一帧的特征$\boldsymbol{v}_k^l\in\mathbb{R}^{(k+1)*D}$，本文使用了两个线性层和ReLU层作为空间原型生成网络，来预测区域级的权重矩阵$\boldsymbol{W}_o^l\in\mathbb{R}^{(k+1)*N_o}$，其中$N_o$表示目标原型的数量，对应的空间目标原型可以表示为：

$$\boldsymbol{P}_o^l=(\boldsymbol{W}_o^l)^{\mathrm{T}}\boldsymbol{v}_k^l\in\mathbb{R}^{N_o*D} \tag{5}$$

$$\boldsymbol{P}_o=\{\boldsymbol{P}_o^l\}_{l=1}^{l=L}\in\mathbb{R}^{L*N_o*D} \tag{6}$$

类似地，给定一段文本描述特征$\boldsymbol{T}_i\in\mathbb{R}^{(M+2)*D}$，可以得到对应的短语原型为：

$$\boldsymbol{P}_p=(\boldsymbol{W}_p^l)^{\mathrm{T}}\boldsymbol{T}_i\in\mathbb{R}^{N_p*D} \tag{7}$$

其中，$\boldsymbol{W}_p^l\in\mathbb{R}^{(m+2)*N_p}$是可学习的权重矩阵。

3.4.2 空间原型匹配

给定一张动态图片$\boldsymbol{v}_i$和一段文本描述$\boldsymbol{t}_i$，根据上述特征提取和空间原型生成方法，可计算得到目标原型$\boldsymbol{P}_o$和短语原型$\boldsymbol{P}_p$。对于每一个目标原型$\boldsymbol{P}_o$，本文使用余弦距离计算出某一帧中该目标原型$\boldsymbol{P}_o^l$与短语原型最匹配的相似度，表示目标-短语匹配相似度。计算式如下：

$$S_{op}=\frac{1}{N_o}\sum_{j=1}^{N_o}\max_{l=1}^{L}\max_{i=1}^{N_p}[\cos\langle\boldsymbol{P}_p\cdot\boldsymbol{P}_o^l\rangle]_{i,j} \tag{8}$$

其中，N_o是目标原型数量，N_p是短语原型数量，L是动态图像的帧数，函数$\cos\langle*\rangle$是余弦相似度。

3.4.3 全局敏感的时间原型生成

对于文本特征$\boldsymbol{T}_i$，直接取标记[EOT]对应的输出$\boldsymbol{t}_e$来作为句子原型：

$$\boldsymbol{P}_s=\boldsymbol{t}_e\in\mathbb{R}^{D} \tag{9}$$

对于动态图片特征$\boldsymbol{V}_i$，根据上述方法可得到对应的目标原型。本文将前两维展开得到$\boldsymbol{P}_o'\in\mathbb{R}^{(L*N_o)*D}$，利用全局敏感注意力和掩码得到图像帧级原型$\boldsymbol{P}_f\in\mathbb{R}^{L*D}$，表示一个图像帧的整体特征。

$$\boldsymbol{P}_f=\mathrm{softmax}(\boldsymbol{M}_j+\boldsymbol{Q}_f\boldsymbol{K}_o^{\mathrm{T}})\boldsymbol{V}_o+\boldsymbol{Q}_f \tag{10}$$

$$\boldsymbol{Q}_f=\boldsymbol{F}_g\cdot\boldsymbol{W}_1 \tag{11}$$

$$\boldsymbol{K}_o=\boldsymbol{P}_o'\cdot\boldsymbol{W}_2 \tag{12}$$

$$\boldsymbol{V}_o=\boldsymbol{P}_o'\cdot\boldsymbol{W}_3 \tag{13}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{W}_3$是可学习的权重矩阵，$\boldsymbol{F}_g\in\mathbb{R}^{L*D}$是全局分支的图像帧级特征，$\boldsymbol{M}_f\in\mathbb{R}^{L*(L*N_o)}$是一个掩码矩阵，用于过滤掉不符合时序的图像帧关系，其定义为：

$$\boldsymbol{M}_f(i,j)=\begin{cases}0, & i*N_o\leqslant j<(i+1)*N_o\\ -\infty, & \text{else}\end{cases} \tag{14}$$

$\boldsymbol{P}_f^l$再加上标记[CLS]对应的输出$\boldsymbol{v}_c^l$，即$\boldsymbol{P}_f^l=(\boldsymbol{P}_f^l+\boldsymbol{v}_c^l)/2$。

接着，利用随机初始化查询的注意力模块，根据图像帧级原型$\boldsymbol{P}_f$提取出事件原型$\boldsymbol{P}_e\in\mathbb{R}^{N_e*D}$，包含若干图像帧组成的事件级特征。

$$\boldsymbol{P}_e=\text{softmax}(\boldsymbol{Q}_e\boldsymbol{K}_f^{\mathrm{T}})\boldsymbol{V}_f+\boldsymbol{Q}_e \tag{15}$$

$$\boldsymbol{K}_f=\boldsymbol{P}_f\cdot\boldsymbol{W}_4 \tag{16}$$

$$\boldsymbol{V}_f=\boldsymbol{P}_f\cdot\boldsymbol{W}_5 \tag{17}$$

其中，$\boldsymbol{Q}_e\in\mathbb{R}^{N_e*D}$是一个随机初始化的可学习的查询矩阵，$\boldsymbol{W}_4$，$\boldsymbol{W}_5$是权重矩阵，$N_e$是事件个数。

3.4.4　空间原型匹配

给定一张动态图片$\boldsymbol{v}_i$和一段文本描述$\boldsymbol{t}_i$，根据上述方法可以计算出句子原型$\boldsymbol{P}_s$和事件原型$\boldsymbol{P}_e$，可通过余弦相似度计算出与文本的句子原型匹配度最高的事件原型，用其相似度来表示空间原型匹配的相似度。计算式如下：

$$\boldsymbol{S}_{\text{es}}=\max_{i=1}^{N_e}(\cos\langle\boldsymbol{P}_e^i,\boldsymbol{P}_s\rangle) \tag{18}$$

3.4.5　全局原型生成及匹配

给定一张动态图片的特征$\boldsymbol{V}_i$，对于其中的每一帧$\boldsymbol{v}_f^l\in\mathbb{R}^{(K+1)*D}$，本文取标记[CLS]对应的输出$\boldsymbol{v}_c^l$组成$\boldsymbol{V}_i^g=\{\boldsymbol{v}_c^l\}_{l=1}^{l=L}\in\mathbb{R}^{L*D}$，使用多头跨模态注意模块来生成事件原型：

$$\boldsymbol{P}_g=\underset{h=1}{\overset{H}{Avg}}\left(\text{Softmax}\left(\frac{\boldsymbol{Q}_h(\boldsymbol{K}_h)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_h\right)\in\mathbb{R}^D \tag{19}$$

$$\boldsymbol{Q}_h=\boldsymbol{E}_g\cdot\boldsymbol{W}_q^h \tag{20}$$

$$\boldsymbol{K}_h=\boldsymbol{V}_i^g\cdot\boldsymbol{W}_k^h \tag{21}$$

$$\boldsymbol{V}_h=\boldsymbol{V}_i^g\cdot\boldsymbol{W}_v^h \tag{22}$$

其中，$\boldsymbol{W}_q^h\in\mathbb{R}^{D*d_k}$，$\boldsymbol{W}_k^h\in\mathbb{R}^{D*d_k}$，$\boldsymbol{W}_v^h\in\mathbb{R}^{D*d_k}$是可学习的权重矩阵；$Avg(*)$表示平均池化；$H$表示注意力头的数量，是一个训练的超参数。

最终，得到文本描述和动态图片的全局相似度：

$$\boldsymbol{S}_g=\cos\langle\boldsymbol{P}_s\cdot\boldsymbol{P}_g\rangle \tag{23}$$

3.5　模型训练

在一个批次中，采用 InofNCE 损失函数来训练本文模型。

$$\mathcal{L}_{\text{t2v}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\boldsymbol{S}^{ii}/\sigma)}{\sum_{j=1}^{B}\exp(\boldsymbol{S}^{ij}/\sigma)} \tag{24}$$

$$\mathcal{L}_{\text{v2t}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\boldsymbol{S}^{ii}/\sigma)}{\sum_{j=1}^{B}\exp(\boldsymbol{S}^{ji}/\sigma)} \tag{25}$$

其中，B是批大小，σ是一个温度控制参数，S表示上述的一种相似度计算方法。

对于目标-短语匹配：

$$\mathcal{L}_{\text{opm}}=\frac{L_{\text{t2v}}(\boldsymbol{S}_{\text{op}})+L_{\text{v2t}}(\boldsymbol{S}_{\text{op}})}{2} \tag{26}$$

对于事件-句子匹配：

$$\mathcal{L}_{\text{esm}}=\frac{L_{\text{t2v}}(\boldsymbol{S}_{\text{es}})+L_{\text{v2t}}(\boldsymbol{S}_{\text{es}})}{2} \tag{27}$$

对于全局特征匹配：

$$\mathcal{L}_{\text{g}}\frac{L_{\text{t2v}}(\boldsymbol{S}_g)+L_{\text{v2t}}(\boldsymbol{S}_g)}{2} \tag{28}$$

模型的整体损失可表示为：

$$\mathcal{L}=(\mathcal{L}_{\text{opm}}+\mathcal{L}_{\text{esm}}+\mathcal{L}_{\text{g}})/3 \tag{29}$$

在推理阶段，可直接计算三阶段的相似度：

$$S=\alpha\boldsymbol{S}_{op}+\beta\boldsymbol{S}_{es}+\gamma\boldsymbol{S}_g \tag{30}$$

其中，α，β和γ是相似度控制因子。

4　实验验证

为验证本文提出的全局敏感的渐进原型匹配模型(GaP-PMM)的有效性，在公开数据集 TGIF[1] 和 Taiwan[2] 上进行了实验验证。

4.1　常用评价指标

对于跨模态检索任务来说，召回率 Recall@K、平均排序数 MnR 和排序中位数 MdR 是 3 个重要的评价指标。以文本检索图像为例，Recall@K 反映了在规模为 N 的测试集上进行 N 次检索实验，检索模型得到的预测结果排序中，正确匹配的图片出现在前 K 个结果中的概率，MnR 和 MdR 分别表示正确匹配图片在预测结果排序中排名数的平均数和中位数。

$$R@K=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{F}(t_i,K) \tag{31}$$

$$F(t_i,K)=\begin{cases}1, & t_i\in\{\boldsymbol{p}_1^i,\boldsymbol{p}_2^i,\cdots,\boldsymbol{p}_k^i\}\\ 0, & t_i\notin\{\boldsymbol{p}_1^i,\boldsymbol{p}_2^i,\cdots,\boldsymbol{p}_k^i\}\end{cases} \tag{32}$$

其中，$F(*)$表示每次检索是否成功，t_i表示第 i 组样本对应的正确图片，$\boldsymbol{p}_i$表示本次测试模型的预测结果。本文中的 Rsum 表示 Recal@1，Recall@5 以及 Recall@10 的和。

4.2　数据集

本实验采用了两个数据集：TGIF[1] 和 Taiwan[15] 数据集。TGIF 数据集是一个包含 100000 张 GIF 动态图片和 120000 条描述这些图片视觉内容的句子的跨模态检索数据集，这些动态 GIF 图片是从社交平台 Tumblr 上随机选取的，时间跨度为 2015 年 5 月至 6 月。该数据集通过众包方式收集了高质量的图像描述，训练集包含 800000 张图片，每张图片对应一句文本描述；验证集包含 10708 张图片，每张图片同样对应一句文本描述；测试集则包含 11360 张图片，每张图片对应三句文本描述。TGIF 数据集因其广泛使用而在跨模态检索领域中占有重要地位。

Taiwan 数据集最初是用来检测“诱导文本情感”的数据集，在本文中用于验证不同的模型。这是一个由 19077 个文本-反应 GIF 对组成的小型数据集，涵盖了 43 种 GIF 标记。实验使用其中的 15254 个文本-反应 GIF 对进行训练，1923 对进行验证，1900 对进行测试。通过这两个数据集，全面评估模型在不同类型和规模的动态图片数据上的表现，验证模型的有效性和泛化能力。

4.3　实验设置

实验在 TGIF 数据集和 Taiwan 数据集上进行训练和测试，采用跨模态检索领域常用的评估指标前 k 召回率 R@k、检索排序平均数 MnR 和检索排序中位数 MdR 来进行实验

评估。为了与现有方法进行对比，本文选择了 CLIP Vit-b/32 版本作为模型的主干网络。在实验设置中，设定文本的最大长度为 32，图像帧数的最大值为 12，模型向量的维度为 512 维，批大小设置为 128。初始化学习率为 1×10^{-4}，并采用 Adam 优化器，设置衰减率为 0.9。在推理阶段，设置相似度控制因子 $\alpha=1,\beta=1.5,\gamma=1$，所有实验均在一张 GEFORCE GTX 3090 GPU 上完成。

本文还进行了消融实验来评估模型中不同组件的效果，结果如表 1 和表 2 所列。实验结果表明，随着模型中组件的增加，特别是全局原型匹配的引入，模型性能得到了显著提升。其中，包含所有组件的完整模型（GaPPMM）展现出了最优的检索性能。

表 1 在 TGIF 数据集上的消融实验结果

Table 1 Ablation experiment results on TGIF dataset

空间原型匹配	时间原型匹配	时间原型匹配＋全局分支	全局原型匹配	R@1↑	R@5↑	R@10↑	Rsum↑
√				21.6	41.7	54.0	117.3
	√			**25.6**	45.4	54.6	125.6
√	√			23.4	44.1	54.9	122.4
√		√		25.6	45.4	54.6	125.6
√		√	√	25.1	**45.9**	**55.3**	**126.3**

表 2 在 Taiwan 数据集上的消融实验结果

Table 2 Ablation experiment results on Taiwan dataset

空间原型匹配	时间原型匹配	时间原型匹配＋全局分支	全局原型匹配	R@1↑	R@5↑	R@10↑	Rsum↑
√				22.1	40.2	48.3	110.6
	√			23.0	41.0	49.1	113.1
√	√			21.8	39.5	47.2	108.5
√		√		23.0	41.0	49.1	113.1
√		√	√	**26.8**	**47.0**	**55.5**	**129.3**

4.4 实验结果

将本文模型与最近几年文本-动态图片跨模态检索领域的先进模型进行了对比。为了公平地进行实验对比，本文利用 TGIF 和 Taiwan 训练集对 CLIP 编码器进行了微调，微调后的模型记为 CLIP-FT。在两个数据集上的实验结果如表 3 和表 4 所列。

表 3 在 TGIF 数据集上的对比实验结果

Table 3 Comparison experiment results on TGIF dataset

模型	Text to GIF Retrieval					
	R@1↑	R@5↑	R@10↑	Rsum↑	MnR↓	MdR↓
JE	18.7	37.5	47.1	103.3	—	—
W2VV++	22.0	42.8	42.7	107.5	—	—
SEA	16.4	33.6	42.5	92.5	—	—
CLIP-FT	21.5	40.6	49.9	112.0	101.2	14
MMT	22.1	42.2	51.7	116.0	—	—
LAFF	24.1	44.7	54.3	123.1	98.6	10
LAFF-ml	24.5	45.0	54.5	124.0	97.3	9
Ours (GaPPMM)	**25.1**	**45.9**	**55.3**	**126.3**	**94.7**	**8**

表 4 在 Taiwan 数据集上的对比实验结果

Table 4 Comparison experiment results on Taiwan dataset

模型	Text to GIF Retrieval					
	R@1↑	R@5↑	R@10↑	Rsum↑	MnR↓	MdR↓
W2VV++	23.5	43.1	50.6	117.2	—	—
SEA	20.1	38.2	46.3	104.6	—	—
CLIP-FT	22.3	41.5	49.8	113.6	95.4	12.0
MMT	24.2	44.3	52.1	120.6	—	—
LAFF	25.0	45.5	53.2	123.7	90.5	9.0
LAFF-ml	25.4	46.0	53.8	125.2	89.7	8.5
Ours (GaPPMM)	**26.8**	**47.8**	**55.6**	**130.2**	**88.1**	**7.0**

如表 3 所列，本文模型（GaPPMM）在各项指标上均取得了最佳性能，具体表现为在 R@1，R@5，R@10 以及 Rsum 指标上分别达到了 25.1，45.9，55.3 和 126.3，在 MnR 和 MdR 指标上分别取得了 94.7 和 8 的较低值。如表 4 所列，本文模型在 R@1，R@5，R@10 以及 Rsum 指标上分别达到了 26.8，47.8，55.6 和 130.2，在 MnR 和 MdR 指标上分别取得了 88.1 和 7 的较低值。

通过对比实验结果可以看出，本文方法在 TGIF 数据集和 Taiwan 数据集上文本到图像的检索任务中召回率总和、排序平均数均超过了基线模型。实验证明，本文提出的基于渐进原型匹配的文本-动态图像检索算法实现了跨模态的细粒度匹配，取得了更好的检索精度。

此外，为了进一步验证本文提出的三阶段原型匹配和全局敏感的时间原型匹配结构的有效性，进行了消融实验，旨在验证模型中每个组件的作用。实验结果表明，相较于传统的时间原型匹配结构，本文采用的双分支设计的全局敏感时间原型匹配结构能显著提升检索效果。全局原型匹配的加入本应进一步提升模型的检索性能，然而，在引入全局原型匹配模块后模型性能出现了下降，这可能归因于模块间的协同不足或特征融合策略不当。

结束语 本文提出了一种全局敏感的渐进原型匹配模型（GaPPMM），用于文本-动态图片跨模态检索任务。该模型采纳了三阶段匹配策略，包括目标-短语原型匹配、事件-句子原型匹配以及全局原型匹配，以实现跨模态间的细粒度交互。引入全局敏感的时间原型生成方法，该方法借助注意力机制及全局分支导出的预览特征，指导局部分支聚焦于最为相关的局部特征，以此降低局部噪声并充分挖掘动态图片的细粒度特征。在公开数据集上的实验结果表明，GaPPMM 模型在召回率总和与平均排序数等关键指标上超越了现有的最先进

模型，实现了更优的检索精度。但是，模型复杂度的增加可能导致训练时间增加及过拟合现象，且模型在不同数据集上的泛化能力仍需进一步的实证研究来验证。同时，模型对计算资源的相对高需求，以及对推理速度和模型可解释性的优化，均是未来研究中需要关注和改进的方面，这些探讨将为后续工作提供明确的方向。

参 考 文 献

[1] LI U C, SONG Y, CAO L L, et al. TGIF: A New Dataset and Benchmark on Animated GIF Description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 4641-4650.

[2] SHMUELI B, RAY S, KU L W. Happy dance, slow clap: Using reaction GIFs to predict induced affect on Twitter[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL, 2021: 395-401.

[3] CHEN H, DING G, LIU X, et al. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 12655-12663.

[4] ZHANG Q, LEI Z, ZHANG Z, et al. Context-Aware Attention Network for Image-Text Retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 3536-3545.

[5] ZHENG F, LI W, WANG X, et al. A Cross-Attention Mechanism Based on Regional-Level Semantic Features of Images for Cross-Modal Text-Image Retrieval in Remote Sensing[J]. Applied Sciences, 2022, 12(23): 12221.

[6] SONG Y, SOLEYMANI M. Polysemous visual-semantic embedding for cross-modal retrieval[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1979-1988.

[7] WANG X, JURGENS D. An animated picture says at least a thousand words: selecting gif-based replies in multimodal dialog[C]// Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, PA: ACL, 2021: 3228-3257.

[8] LI G, DUAN N, FANG Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 11336-11344.

[9] CONNEAU A, LAMPLE G. Cross-lingual Language Model Pretraining[C]// NeurIPS: Advances in Neural Information Processing Systems. Curran Associates Inc., 2019.

[10] HUANG H, LIANG Y, DUAN N, et al. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks[J]. arXiv: 1909.00964, 2019.

[11] ZHANG K, MAO Z, WANG Q, et al. Negative-Aware Attention Framework for Image-Text Matching[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 15661-15670.

[12] LI X, YIN X, LI C, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks[C]// Proceedings of 16th European Conference on Computer Vision (ECCV 2020). Springer, 2020: 121-137.

[13] CHEN S, ZHAO Y, JIN Q, et al. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 10638-10647.

[14] SONG X, CHEN J, WU Z, et al. Spatial-Temporal Graphs for Cross-Modal Text2Video Retrieval[J]. IEEE Transactions on Multimedia, 2022, 24: 2914-2923.

[15] MIECH A, ZHUKOV D, ALAYRAC J B, et al. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 2630-2640.

[16] LUO J, LI Y, PAN Y, et al. CoCo-BERT: Improving Video-Language Pre-training with Contrastive Cross-modal Matching and Denoising[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 5600-5608.

[17] PENG J, HUANG J, XIONG P, et al. Video-Text As Game Players: Hierarchical Banzhaf Interaction for Cross-Modal Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2023: 2472-2482.

[18] DONG J F, ZHANG M, ZHANG Z, et al. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2023: 11302-11312.

PENG Jiao, born in 1991, postgraduate, engineer. Her main research interests include NLP image processing and big data analysis.

OU Zhonghong, born in 1982, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 69730S). His main research interests include small sample learning, cross-domain adaptation and small target detection.

（责任编辑：何杨）