FashionERN: Enhance-and-Refine Network for Composed Fashion Image Retrieval

完成人：卢艺晗 2213583　　**日期：**2024年10月18日

《FashionERN: Enhance-and-Refne Network for Composed Fashion Image Retrieval》**论文笔记**

<table>
<tr><td rowspan="4">基本信息</td><td>发表刊物</td><td>第38届AAAI人工智能大会（AAAI-24）论文集</td><td>发表年份</td><td>2024</td><td>第一完成单位（国内）</td><td>北京大学王选计算技术研究所</td></tr>
<tr><td>作者</td><td colspan="5">Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng (Corresponding author), Jiahuan Zhou, Lele Cheng</td></tr>
<tr><td colspan="2">关键词（中文）</td><td colspan="4">组合时尚图像检索，参考图像主导现象，时尚增强和精炼，三分支修饰符增强模型，双引导视觉精炼模型</td></tr>
<tr><td colspan="2">关键词（英文）</td><td colspan="4">composed fashion image retrieval, reference image dominance phenomenon，FashionERN，TME，DVR</td></tr>
<tr><td>论文内容</td><td colspan="2">解决的问题（如有实际应用场景请说明）</td><td colspan="4">（1）背景知识——组合时尚图像检索
它是一种特定的图像检索任务，其目标是通过组合参考图像和修改文本，来找到与用户需求匹配的目标图像。
一个实际的例子：一个用户上传了一件他喜欢的连衣裙的图片，并希望找到颜色为红色的类似款式。他输入：“换成红色”。系统会根据参考图像的特征（如裙子的款式、长度）及修改文本的要求（红色），为用户推荐符合这些条件的裙子。
（2）关键问题
这篇论文致力于解决组合时尚图像检索任务中参考图像主导现象的问题：
很多时候，时尚图像检索任务中的输入是不平衡的，参考图像包含了丰富的视觉信息，而修改文本通常是简洁的、片段化的描述。现有的检索方法常用对称编码器（如CLIP）。它会导致视觉模态和文本模态之间的嵌入不平衡：参考图像对最终检索结果的作用加强，而修改文本的作用被弱化，导致结果偏向于与参考图像相似的图像，而未能准确反映用户输入的文本修改需求。
(3)应用场景
组合时尚图像检索任务有着广泛的实际应用，例如电商平台的商品搜索、个性化推荐、时尚搭配等。</td></tr>
</table>

解决问题的方法（采用什么模型框架等）

（1）方法框架

本文提出了FashionERN方法，致力于解决组合时尚图像检索任务中参考图像主导现象的问题。该方法由两个关键的模块：**三分支修饰符增强模型（Triple-branch Modifier Enhancement, TME）**和**双引导视觉精炼模型（Dual-guided Vision Refinement, DVR）**。

- **TME：**该模块的主要目的是增强文本编码器，通过参考图像的全局信息注入以及目标图像的语义对齐，来丰富修改文本的语义表示。它包括三个分支：常规编码分支、参考注入分支和目标对齐分支，分别提取原始文本特征、结合参考图像的语义信息以及实现文本与目标图像的跨模态对齐。
- **DVR：**该模块用于精炼视觉信息，提炼参考图像中与修改文本相关的关键信息。通过文本引导的精炼和自我引导的精炼这两个步骤，过滤掉无关信息，并聚焦于与修改文本匹配的视觉区域。

（2）具体原理

1. TME

TME的目的是通过多分支机制增强文本的语义表示，模块由三个分支构成：

a. **常规分支编码器：**

它是传统的Transformer编码器分支，用于处理文本输入。通过多头自注意力机制（MHSA）和全连接层（MLP），逐层提取文本的嵌入特征，形成基础的文本表征。该分支主要用于提取修饰文本的基本语义特征。

公式：$Zmod' = MHSA(LN(zmod)) + Zmod$

b. **参考注入分支：**

该分支将参考图像的全局信息注入到文本嵌入中，使文本能够借助参考图像中的视觉信息来增强其语义特征：通过多头交叉注意力（CA）机制，将参考图像的嵌入信息作为Key和Value，修饰文本嵌入作为Query，得到一个增强的文本表示。

公式：$Zinject = MLP(CA(zmod, zref, zref)) + Zmod$

c. **目标对齐分支：**

为了进一步缩小文本和目标图像模态之间的差距，目标对齐分支使用KL散度来引导修饰文本与目标图像的对齐。通过对比目标图像的语义分布，调整文本嵌入的语义，使其更加接近目标图像。

公式：$Zalign = MLP(MHSA(LN(zmod))) + Zmod$

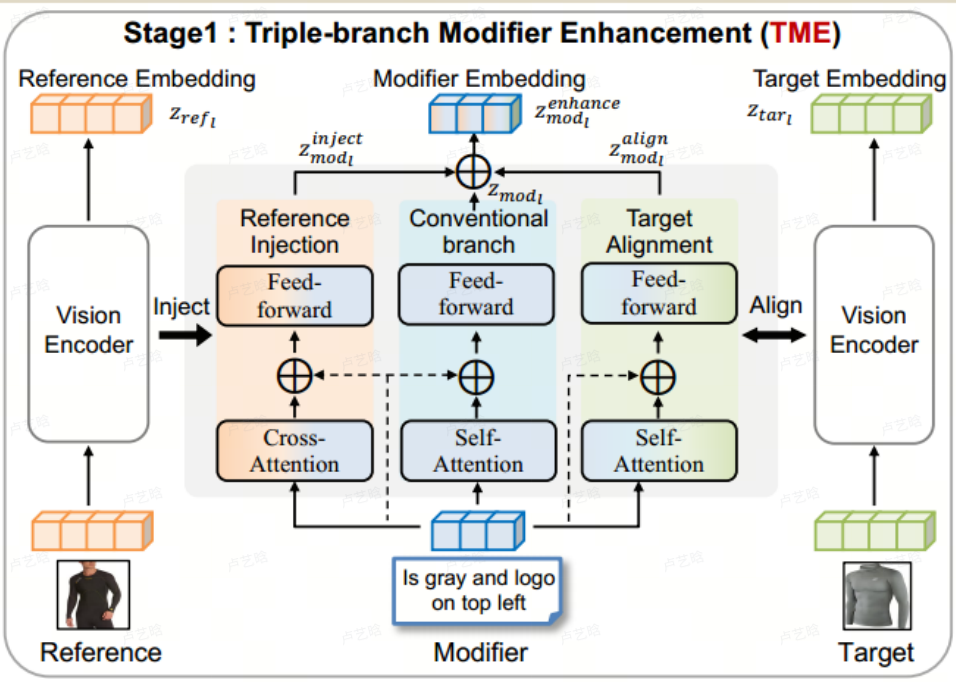

图1 TME框架

2. DVR

DVR的作用是从参考图像中精炼出与修改文本最相关的视觉信息，只保留关键的视觉语义信息而过滤掉不相关的其它信息，确保检索结果更符合用户需求。DVR主要包括两个步骤：

a. **文本引导的精炼：**

DVR首先通过文本模态引导视觉模态的精炼过程。通过将修饰文本的嵌入作为Query，参考图像的嵌入作为Key和Value，使用多头交叉注意力机制筛选出与文本相关的视觉区域。

公式：rMR=MLP(CA(rmod,rref,rref))+rmod

b. **自我引导的精炼：**

在文本引导的基础上，DVR进一步通过自我引导的方式对视觉信息进行精炼。系统通过视觉模态自身的上下文关系，过滤掉噪声信息，保留最关键的视觉语义。

实现细节：首先通过平均计算获取视觉模态的粗粒度特征，然后使用投影函数计算注意力权重，最终通过这些权重对视觉特征进行加权求和，得到自我精炼的视觉信息。

公式：rSR=（ΣWc）⊠（rMR,c）

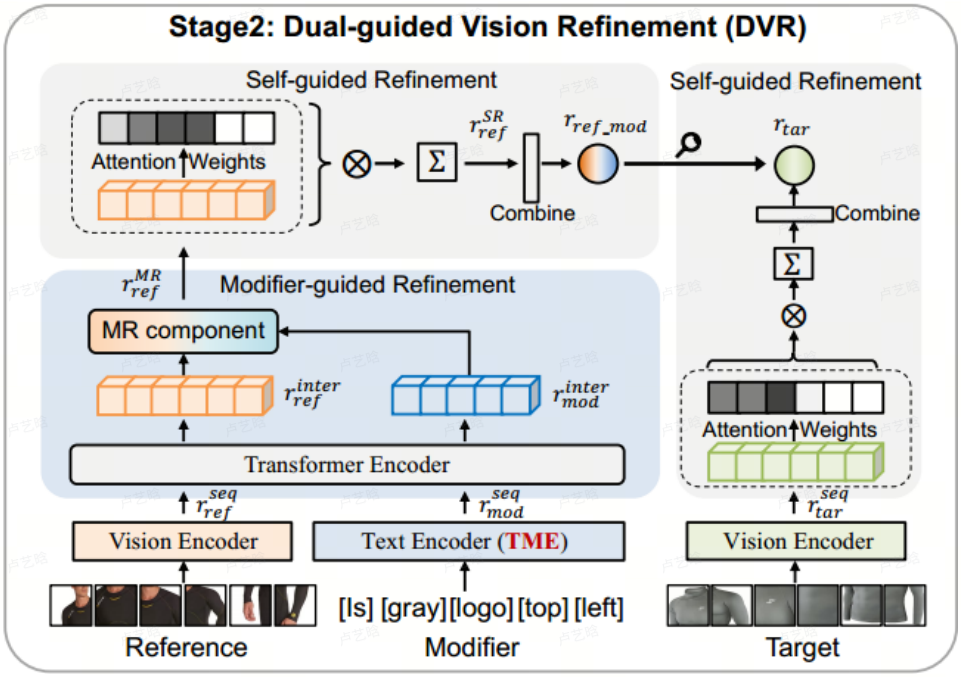

图2 DVR框架

3. 网络训练

a. **第一阶段**：首先训练TME模型，以获得增强的文本编码器。使用batch-based loss来处理每个小批次的图像-文本对，并通过KL散度对齐修饰文本与目标图像的模态。

b. **第二阶段**：在冻结视觉编码器和文本编码器的情况下，训练DVR模型，通过双重引导的精炼过程优化视觉信息。

仍旧存在的问题（注明论文中说明的问题或自己认为存在的问题）

（1）论文中说明的问题：

1. **多轮交互场景的适应性**：

虽然FashionERN显著改善了单轮的组合图像检索性能，但论文指出该模型在多轮交互检索场景下的应用仍然有限。即用户多次提供修改反馈时，系统能否保持高效的检索性能仍是一个未解决的问题。这一点在实际的电商购物场景中也很重要，因为用户可能会通过多次反馈逐步完善他们的需求。

2. **时间复杂度问题**：

论文中提到，虽然模型能够显著提高检索精度，但在大型数据集或实时场景下，时间复杂度仍是一个问题。随着图像和文本的复杂度增加，模型的推理时间可能会变长，影响实际的用户体验。

（2）自己认为存在的问题：

1. **泛化能力的问题：**

FashionERN的以时尚图像检索为例提出此方法，在论文中提到的几个时尚图像标准数据集上性能表现优异，但是该原理在其它非时尚的组合图像检索任务中是否也有效果，值得进一步探究。

实验采用的数据集

实验采用了FashionIQ、Fashion200K、CIRR和Shoes四个常用的时尚图像数据集。

实验内容	实验是否涉及实际应用场景	实验涉及了实际应用场景。该实验主要针对时尚电商平台的图像检索展开，特别是CIRR数据集，它包含21,552张真实世界的图片，能够在一定程度上反映该方法的实际效果。
	实验采用的对比方法	实验做了两方面的对比： 1. 将本方法与该领域当前一些比较先进的方法做对比，验证FashionERN的效果； 2. 消融研究：在ResNet和ViT两种骨干网络上增量地添加TME和DVR，对比添加前后的效果，来验证每个组件的价值。
	实验任务	（1）实验设计 1. 将 FashionERN与一些基线方法，比较它们在四个数据集上的效果，以验证FashionERN的性能。这些基线方法有DCNet、ARTEMIS、MGUR、FashionVLP 、Comquery、CLIP4Cir、FAME-MiL等等。 2. 消融研究：在ResNet和ViT两种骨干网络上增量地添加TME和DVR，比较添加前后的效果，验证本文实现的TME和DVR两个模型的效用。 （2）实现细节： 使用ResNet50x4和ViT-B/16作为骨干网络，Adam优化器（小批量大小为1024，初始学习率为4e-5，采用余弦退火策略），总共训练50个epochs，并使用8个Tesla V100 GPU进行模型训练。 （3）评估方式 1. 使用标准的top-K召回率作为评估指标，计算Recall@10、Recall@50和它们的平均值，来衡量效果； 2. 在FashionERN上使用两种评估协议： ◦ **VAL评估协议：**检索系统在进行检索时，所有图像（包括参考图像和目标图像）都被用作候选图像集。在这种协议下，由于候选集的图像较少，系统更容易在前K个结果中找到正确的目标图像，因此Recall@K的值会比较高。它更容易得到高分，适合衡量模型的基础能力。 ◦ **原始评估协议：**这是在FashionIQ数据集的初始研究中使用的评估方式。在这种协议下，候选集包含更多的图像（通常包括所有可能的图像），意味着系统需要在更大的图像集里查找目标图像。这种方式检索难度更大，更接近真实的应用场景，适合衡量模型的实际性能和在复杂情况下的表现。
	实验衡量指标	本实验采用图像检索领域标准的评估指标——**top-K召回率（Recall@K）**，采用**Recall@10、Recall@50和它们的平均值**作为衡量指标。 **概念理解：**Recall@K表示系统在前K个返回的结果中正确检索到目标的比例。即 Recall@K = (在前K个结果中检索到的正确相关项数) / (相关项的总数)。其中Recall@10 和 Recall@50 是在时尚图像检索任务中常用的评估指标。

		Recall@10：用于评估系统能否在前10个检索结果中返回大多数相关的时尚图片，体现系统的高精度表现。 **Recall@50**：则考察系统在前50个返回结果中是否能包含更多的相关项，体现系统的整体检索能力。
	实验说明所提出方法的优点	实验结果表明FashionERN的效果很多当前的前沿方法都要好，且FashionERN的两个组件TME和DVR都起到了一定的作用。 这表明，FashionERN通过增强文本编码器和精炼视觉信息，有效地解决了参考图像主导现象，提升了检索的精度和效果。
思考内容	**论文的主要优点是什么**	1. 发现关键问题，提出创新方法： 论文发现了组合图像检索中参考图像主导现象这个关键问题，从文本语义增强和参考图像精炼两个角度设计优化模型，弱化参考图像中无关信息的影响，强化修改文本的影响力，从而有效提升了组合图像检索的精准度。 2. 实验验证较为充分： 论文在四个标准数据集上与众多baseline进行效果对比，充分验证了FashionERN在提升检索效果上的优秀；此外，通过消融研究验证每个组件的作用，十分严谨。
	论文仍然可以改进的地方是什么	1. 时间开销： 该方法的实现细节较为复杂，在大规模数据集下可能耗时较多，在实际应用中可能影响用户体验。因此，可以考虑优化部分实现，降低时间复杂度，提升效率。 2. 多轮交互场景下的适应能力： 该方法仅针对单轮组合图像检索的性能提升，而在实际中，用户可能通过多次输入修改文本逐步精确目标。可以考虑放在多轮组合图像检索的场景下，提升FashionERN的应用范围。