强化学习驱动的组合图像检索多轮查询优化

2213583　卢艺晗

1　背景与意义

1.1　组合图像检索

随着互联网上多媒体数据的激增，图像检索已成为计算机视觉的基本任务，应用广泛。传统图像检索的目标是在大型数据集中查找最相关的图像。然而，由于单模态输入包含的信息通常比较有限，有时无法完全反映用户的搜索意图。因此，Vo et al. [8] 首先将传统的单模态输入扩展为组合图像检索（CIR）的多模态输入。

组合图像检索（Composed Image Retrieval, CIR）是一种利用多模态信息的图像检索方法。用户输入一张图片（reference）和一段文字描述（modification），文字描述中指定修改图片上的某个或某几个实体，要求返回和根据文字修改后图片最相似的图片。组合图像检索广泛应用于计算机视觉领域、时尚推荐、电商平台、图像创意设计等场景中，帮助用户高效地找到更精确的图像内容，提升用户体验。

1.2　支持多轮查询优化的意义

在真实的应用场景中，由于文字描述不全面等多种因素，检索系统有时不能够经过一次查询就得到用户心中的目标图像，特别是当用户对检索图像具有细粒度的特征要求时。用户可能需要多次输入文字描述，让系统不断调整检索方向，才能逐步趋近于用户最想要的目标图像。

本研究旨在设计一个支持多轮查询的组合图像检索方法，通过用户的多轮查询输入，不断优化系统检索目标，从而提升检索的准确性和用户的满意度。

2　现状分析

2.1　组合图像检索研究现状

组合图像检索领域的研究自提出以来，一直十分热门。现有的方法大多针对单轮查询（即输入一个图像、一段文本，一次性检索）。一类流行方法主要致力于研究如何将组合输入的特征正确地组合成一个联合表示，即多模态融合问题，ComposeAE [1] 将整个图像和文本特征融合在一个复杂的空间中，以得出组合的输入表示。除了全局合成模块之外，一些方法提出在局部级别用视觉描述符来合成修改文本。例如，陈 [3] 等人利用分层注意力机制融合局部特征图和文本特征。为了有效捕获参考图像和修改文本之间的语义关系，LBF [5] 通过 Faster-RCNN 将图像检测为一组局部实体，并通过跨模式注意力融合组合输入。其他流行的方法如 [6] 等侧重于用不同的策略改进公共匹配空间。此外，最近的一些方法如 [2]、[7] 研究零样本合成图像检索（ZS-CIR）任务，该任务可以在无需对三元组数据集进行训练的情况下执行合成图像检索。

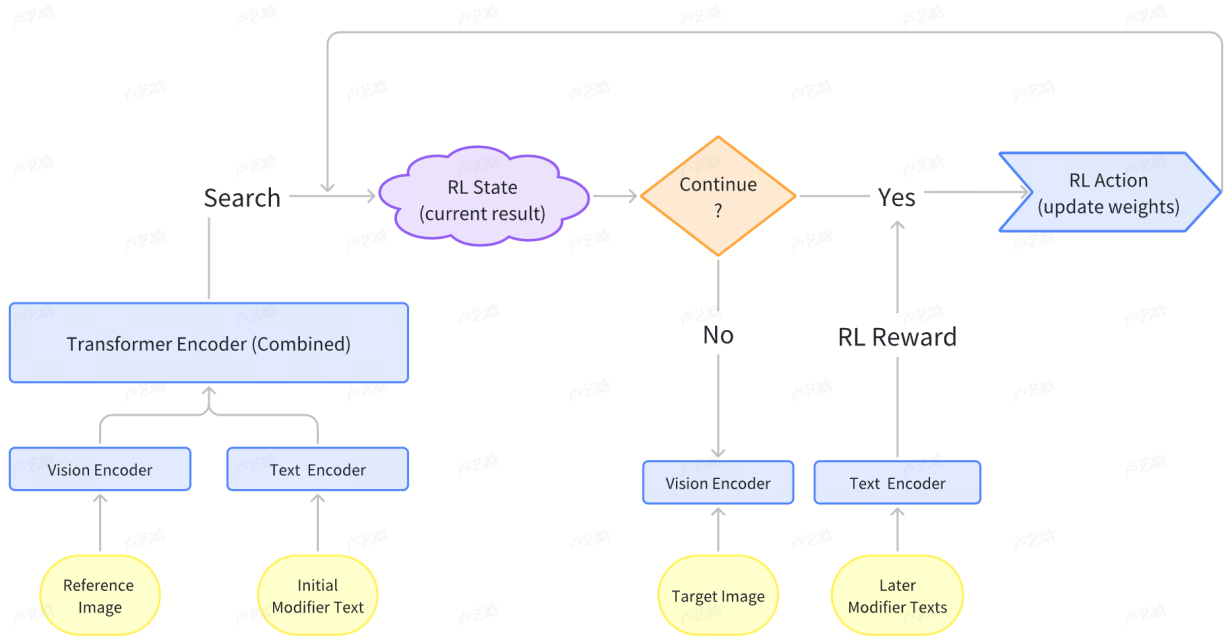

图 3.1: 方法流程图

2.2 问题矛点

Fashion-ERN 方法 [4] 致力于解决组合图像检索中参考图像主导问题，这篇论文作者指出多轮查询优化场景应用的重要性，只是 Fashion-ERN 并未能支持多轮查询优化。其难点在于如何将多次查询输入的文本嵌入到多模态融合的特征之中。

于是，在这篇文章中，我立足于对多轮查询优化的支持，设计了强化学习驱动的组合图像检索多轮查询优化方法，将用户的多轮查询输入作为反馈，提取其语义信息，用来调整特征权重，从而解决多模态融合与多次查询嵌入的接洽难点。

3 方案设计

本方法旨在设计一个强化学习驱动的多轮查询优化的组合图像检索系统，方法流程图如图3.1所示。我的方法可以简要分为三个模块：特征提取模块、强化学习驱动的检索模块、动态特征融合模块。

3.1 特征提取模块

特征提取模块负责将用户的输入——参考图像和描述性文本转换为模型可处理的多模态特征表示。通过视觉和文本编码器分别提取图像和文本的特征，生成查询嵌入。

- **视觉编码器** (Vision Encoder)：用于提取参考图像的深层视觉特征。例如，可以采用预训练的卷积神经网络（如 ResNet 或 Vision Transformer）将参考图像转化为特征向量。
- **文本编码器** (Text Encoder)：用于处理用户的描述性文本，生成捕捉用户意图的文本嵌入。如常用的 BERT 或其他语言模型。

上述视觉和文本特征通过 Transformer 编码器组合生成多模态嵌入，作为初始查询的特征表示。可以举个简单的例子：用户输入了一张穿着红色短裙的女人照片，输入的初始查询文本是“换成蓝色，长

裙”。那么系统对输入图像和文本分别采用视觉编码器、文本编码器编码后，在 transformer 编码器中融合，得到一些表示着裙子、颜色、长度等特征的嵌入。

由于这部分在组合图像检索领域研究广泛，我们不对该部分进行特别的创新，图中只是一个最简单的示意。在真正的系统开发时，可以采用一些先进的方法进行平衡、强化。

3.2 强化学习驱动的检索模块

强化学习驱动的检索模块是本系统的核心创新点，通过在检索系统中引入 Q-learning 算法，在多轮反馈中动态优化检索策略。该模块将用户的多轮查询输入的描述性文本作为反馈，根据用户反馈更新特征权重，使系统能够逐渐接近用户检索的目标特征。

- **状态表示：**将每轮查询生成的检索结果作为当前状态，根据用户输入情况指示下一步采取的动作。继续前面提到的例子，假设系统返回的结果图像中，基本都是蓝色裙子，但是长度参差不齐（用户想要的是全长的裙子），那么当前状态未完全满足作者的要求，且检索到的裙子花纹全都和参考图像一样，而这是不需要的。

- **动作选择：**若用户继续输入查询文本，则根据输入信息采取下一步动作。例如，用户继续输入“颜色正确，花纹不需要一样，裙子要全长”。那么系统将会采取下一步动作，如改变特征权重等，并再次进行检索。在这个例子中，系统会将花纹的权重降低，而长度的权重增大，然后再次检索。若用户什么都不输入，则默认用户接受了这个结果，系统返回这些图片。

- **奖励/惩罚计算：**根据用户反馈评价动作效果，若检索结果更接近用户需求，系统获得正向奖励；若未满足需求，则施加负向奖励。奖励值用于 Q-learning 更新，以逐渐找到最优的检索策略。

在这个模块，用户继续输入查询文本的行为是驱动系统不断更新、调整策略的动因。只要用户继续输入，系统就会继续迭代，直至用户关闭此次查询。

3.3 动态特征融合模块

动态特征融合模块用于在每轮检索中将当前检索结果特征与历史查询的特征向量进行融合，使系统能够整合用户所有历史查询信息，从而更全面地掌握用户对目标图像的偏好。通过特征融合，系统在每一轮查询中都能在既有基础上进一步调整，避免信息丢失或偏差。该模块可以通过加权平均或注意力机制实现特征融合。当前轮次的嵌入表示将与之前的特征表示结合，使系统能够有效利用多轮反馈中累积的信息，以更全面和稳定的方式接近目标。

4 技术路线

这部分我们主要对本方法的创新之处——强化学习驱动的检索模块和动态特征融合模块进行详细的技术讲解。

4.1 强化学习驱动的检索模块

强化学习驱动的检索模块是系统支持多轮查询优化的核心之一，负责在多轮查询过程中动态调整特征权重，以逐步更接近用户的目标。我们采用 Q-learning 算法，逐步优化检索策略，以更加适应用户的需求。

(1) Q-learning **要素定义**

- **状态** (State, S)：状态主要包含当前检索结果的嵌入表示，并加入用户在当前轮次的反馈特征。这些反馈特征（例如“颜色正确但长度不够”）通过文本编码器生成嵌入，并与检索结果的特征向量组合。通过将用户反馈嵌入到状态表示中，系统能够在下一轮检索中动态调整检索方向。

- **动作** (Action, A)：这里，动作主要指系统对不同特征的权重调整。在这个例子中，系统可以增加长度特征的权重，或降低花纹特征的权重。动作集合具体包括：增加某特征权重、减少某特征权重、保持特征权重不变等动作。每次执行一个动作后，系统会生成新的查询嵌入，从而改变检索结果。

- **奖励** (Reward, R)：奖励机制根据用户的反馈来计算系统的动作效果。为了捕捉用户逐步变化的需求，需要灵活设计奖励函数。具体而言，当用户反馈系统满足部分需求（如“颜色正确”）时，系统获得部分奖励；若系统满足所有需求，则获得更高的正向奖励；若系统未满足用户的关键信息（如长度需求），则会受到惩罚。在每轮检索中，系统会根据用户反馈自动调整奖励函数的参数。

- **经验回放** (Experience Replay)：在传统 Q-learning 中，经验回放用于加速收敛。在此系统中，每一轮用户反馈和相应的 Q 值更新被存储，并在随后的轮次中随机采样，从而让系统在多轮查询优化中逐渐提高稳定性。

(2) Q-learning 的调整和优化

在将 Q-learning 应用于组合图像检索时，需要根据系统的多轮查询特点对 Q 值更新公式进行调整。标准的 Q-learning 公式为：

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(R + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a) \right)$$

我们对其进行以下调整，以适应多轮查询优化的组合图像检索：

- **动态学习率调整**: 在标准 Q-learning 中，学习率 a 通常是固定的。然而在多轮检索任务中，用户的提示重点可能逐轮发生变化。为此，我们引入动态学习率：逐轮衰减学习率，初始轮次保持较高学习率，便于系统快速收敛；后续轮次逐渐降低学习率，增强系统的稳定性。

- **自适应折扣因子**: 折扣因子 γ 用于控制未来奖励的重要性。为应对用户的动态反馈，我们引入自适应折扣因子 γ，使其与用户满意度关联。例如，当系统较接近用户目标时，增大 γ，强调未来动作的重要性；而当用户反馈较差时，减小 γ，强化对当前反馈的重视。

4.2 动态特征融合模块

动态特征融合模块旨在确保系统在多轮查询过程中不仅能够灵活响应用户的每次反馈，还能将新的嵌入与旧的特征多模态融合，在搜索时全面地掌握用户需求，从而更加接近用户的目标。

在每轮查询中，系统通过用户的反馈信息动态更新特征权重。每轮查询生成的特征表示会结合当前反馈的嵌入与强化学习模块输出的权重调整，形成一个全局特征表示。该全局特征表示不仅反映了用户当前的需求，还保留了系统对用户偏好的整体理解，从而生成更符合用户期望的检索结果。

(1) 全局特征生成：在用户输入新的反馈（如“颜色正确，但长度不够、花纹不必相同”）后，系统通过强化学习模块调整特征权重，增加长度的权重，降低花纹的权重。更新后的权重应用于多模态嵌入生成新的全局特征表示，并直接用于当前轮次的检索。

(2) 短期记忆机制：为了保留最近反馈的影响力，我们引入短期记忆机制，以滑动窗口的形式存储最近一轮反馈的关键信息。具体而言，每轮查询生成的全局特征表示将基于当前和上轮反馈的特征进行加权融合。在第 $t+1$ 轮中，系统的全局特征表示 Z_{t+1} 可以通过公式计算（一个简单的公式示例）：

$$Z_{t+1} = \alpha \cdot z_{t+1} + (1-\alpha) \cdot z_t$$

其中，α 为自适应参数，根据用户对当前轮次的反馈满意度调整。当用户对当前结果的满意度较高时，α 取较大值，使当前反馈主导全局特征；若满意度一般，α 则偏小，从而让系统保留一定的历史信息。

通过强化学习驱动的检索模块和动态特征融合模块的结合，该系统能够在多轮交互中有效利用用户的最新反馈和历史记忆，逐步接近用户想要的目标图像。

5 总结

在这篇文章中，我讲述了基于强化学习让组合图像检索系统支持多轮查询优化的方法，其关键有两点: 一是 Q-learning 算法的引入和动作、奖励函数的设计，其中利用用户查询输入进行特征权重的调整是思想核心；二是通过动态特征融合让模型兼顾旧的特点与新的反馈，避免注意力过度偏颇，从而全面地、逐步地趋近用户心中的目标图像。

当然，此方法未经实践，也存在诸多不足之处，如 Q-learning 和全局特征的具体公式仅给出样板示意，尚未完善；当用户查询次数过多时，如何管理历史特征记忆的内存空间，也是一个问题；强化学习的奖励函数针对不同反馈的自适应调整，也是一个需要精细化琢磨的细节问题。

整体而言，通过这次探索和思考，我不仅对组合图像检索领域的研究有了较为广泛的了解，还发现了将机器学习与图像检索结合的诸多可能，拓宽了思路。

参考文献

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1139–1148, 2021.

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 15292–15301, 2023.

[3] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2998–3008, 2020.

[4] Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, Jiahuan Zhou, and Lele Cheng. Fashionern: Enhance-and-refine network for composed fashion image retrieval. In AAAI Conference on Artificial Intelligence, 2024.

[5] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3593–3602, 2020.

[6] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Xin Liu, and Heng Tao Shen. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 34(4):2959–2972, 2024.

[7] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19305–19314, 2023.

[8] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval: An empirical odyssey. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6432–6441, 2019.