对话式图像检索：基于 ChatIR 模型的复现验证研究

王淑怡

2025 年 1 月 5 日

摘要

本研究围绕对话式图像检索系统 ChatIR 的复现与分析展开，旨在探讨其在多轮对话中的应用潜力。ChatIR 模型通过与用户进行对话，逐步获取信息并明确搜索意图，从而优化图像检索过程。系统设计包括对话构建和图像搜索两个主要部分，其中对话构建阶段通过问题生成器和视觉对话模型引导用户反馈，而图像搜索阶段则利用文本编码器在视觉特征空间中检索目标图像。实验结果显示，该模型在 50,000 张图像中实现了 78% 的成功率，显著优于传统单次文本到图像检索的 64%。本文的研究不仅展示了 ChatIR 的有效性，还为未来图像检索系统的设计与优化提供了新的思路。

关键词：对话式图像检索；模型复现；ChatIR；图像文本匹配；视觉问答

1 引言

用户始终是信息检索的核心，对话式搜索有助于提升搜索效果与效率。随着可搜索视觉媒体数量剧增，快速可靠的检索系统需求凸显。目前计算机视觉领域的图像检索方法主要集中于图像到图像、文本到图像和合成图像检索，但单一查询可能无法完全传达用户的搜索意图，需多次尝试才能获取满意结果。对话搜索为信息检索提供了新的机会，可以提高搜索的有效性和效率。

Matan Levy 等人 [1] 受大语言模型（LLM）的启发，提出了一种新的基于对话的图像检索模型 ChatIR，通过与用户对话去获取更多信息，以此来明确用户的搜索意图，并利用大型语言模型来生成初始图像描述的后续问题。这些问题与用户形成对话，以便从大型语料库中检索所需的图像。

本研究旨在探讨对话式图像检索系统的能力和局限性，为进一步提高图像检索的效率和准确性提供新的思路。根据现有研究，ChatIR 模型在 50,000 张图像中实现了 78% 的成功率，相较于传统单次文本到图像检索的 64% 有着显著的提升 [1]。这一结果表明，对话式图像检索不仅能够提高检索效率，更能增强用户的搜索体验。

本报告将详细描述 ChatIR 模型的设计思路、复现过程以及评估结果，为未来的研究和应用提供参考。

2

2 相关工作

随着技术的进步，人机对话得到大力发展，基于大型知识库的聊天机器人逐渐从简单的任务扩展到更复杂的场景，例如多模态交互与动态的信息处理。以下将详细探讨视觉模态与对话任务的相关研究进展及其应用。

2.1 视觉对话：从单一交互到多轮对话的演变

在视觉领域，许多传统应用仅涉及单轮的用户与图像互动。例如，视觉问答 [2–4] 允许用户基于图像提出问题，而系统则根据图像生成答案；图像检索及其多种变体 [5, 6] 支持用户通过查询信息找到目标图像；而组合图像检索 [7, 8] 通过结合图像和文本信息帮助用户更准确地描述需求。然而，这些方法普遍存在输出单一化的缺点，难以支持多轮互动。

为了解决这一问题，视觉问题生成技术 [9, 10] 应运而生，其目标是根据图像生成有针对性的问题。然而，与其相比，视觉对话任务则更进一步 [11, 12]，要求系统能够与用户基于图像内容展开多轮对话。在这一任务中，用户通常扮演提问者的角色，而系统通过“观察”图像并结合对话历史生成回答。

近年来，生成模型的兴起进一步推动了这一领域的发展。例如，Mittal 等人 [13] 开发了逐步生成图像的技术，通过基于场景描述的图形序列逐步构建视觉内容。Wu 等人 [14] 设计了多功能对话系统 VisualChatGPT，通过整合视觉与语言能力，实现图像的生成与处理。此外，研究还尝试将 ChatGPT 与 BLIP2[15] 结合，以增强对图像内容的描述能力。这些进展表明，视觉与语言的深度融合为智能系统的发展提供了更多可能。

尽管以上方法在图像理解和生成方面取得了显著成果，它们大多未直接涉及图像检索的任务。与此同时，生成视觉对话技术 [16–18] 旨在让系统生成关于图像的自然对话的领域为此问题提供了独特的视角。该领域的研究通常训练两个对话代理，分别负责提问与回答，通过协同完成任务。这种方法的目标是测试机器生成多轮自然对话的能力，训练过程中采用强化学习策略，而评估则通过“合作图像猜测”等任务完成。基础视觉与语言模型近年来在多方面表现卓越，包括回答准确性 [12]、问题生成多样性以及其他下游任务 [19–21]，显著超越了这些早期方法。

2.2 视觉搜索的转型：从静态查询到动态对话

视觉搜索作为信息检索的核心任务之一，其发展离不开用户的参与。在传统图像检索中，研究主要通过结合用户的反馈来优化检索性能 [22–24]。这些反馈形式从最初的二元选择（如相关或不相关）[25] 逐渐演变为基于预定义属性的多分类反馈 [24, 26]。最近，自然语言处理技术的进步使得用户可以通过开放式的自然语言描述进行反馈，从而诞生了组合图像检索 [21, 27] 这一新形式。该任务允许用户通过图像和文本的组合输入，描述需求并找到目标图像。

在这一领域，部分研究开始尝试通过多轮交互提升检索效果。例如，系统可以结合用户提供的文本反馈，逐步优化结果。然而，这些方法通常要求用户主动描述每次检索的细节，而这些描述彼此独立，未能充分利用对话历史信息。这种模式在复杂检索任务中容易加重用户负担，影响交互效率。

为了解决这种局限性，ChatIR 系统采用了一种全新的检索策略。与传统方法不同，ChatIR 主动引导用户通过提问明确需求，同时将对话的上下文和历史信息整合到每一次检索中。这

种方法不仅减轻了用户的输入压力，还显著提升了复杂场景下的检索精度和对话连贯性。

2.3 未来发展的潜力与挑战

在未来，视觉对话与视觉搜索的进一步融合将会开启一个全新的研究领域。通过结合多模态模型、强化学习以及动态交互技术，可以进一步提高系统的智能化水平和应用能力。此外，如何利用用户行为特征与对话历史，构建更自然、更高效的智能系统，也是值得探索的重要方向。ChatIR 的研究不仅为视觉检索提供了创新的思路，也为多模态技术在其他复杂任务中的应用提供了借鉴。随着技术的不断进步，视觉与语言的结合将为更多实际场景创造价值，例如虚拟助手、电子商务推荐及教育领域的交互系统等。

综上，视觉模态与人机对话的结合正在迅速改变传统信息检索的方式。未来发展中，如何进一步提升模型的多轮交互能力和跨模态理解能力，将成为推动这一领域的重要课题。

3 本文方法

3.1 本文方法概述

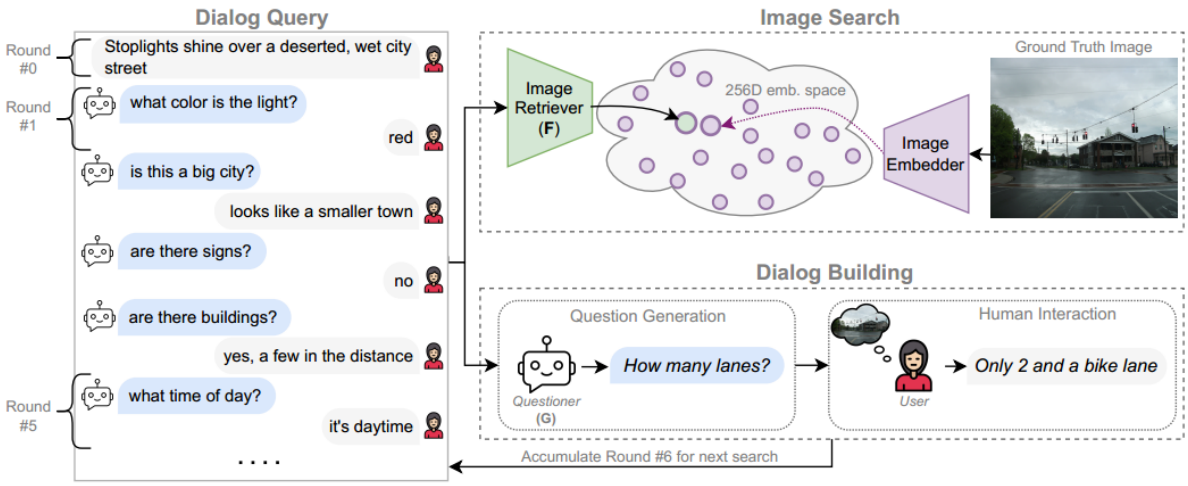

图 1. ChatIR 系统概述。图片来源：文献 [1]

这篇文章 [1] 提出了一个名为 ChatIR 的对话式图像检索系统，目的是在通过对话的方式使图像搜索过程更加高效，包括两个主要部分: 对话构建 (Dialog Building, DB) 和图像搜索 (Image Search, IS)。这篇文章还设计了相应的评估方法，对不同模块进行了深入分析和比较。如图 1 的系统概述所示，IS 阶段将正在进行的对话（由图像标题和几轮问答组成）作为输入，以找到目标图像。需要我们注意的是，长度为 0 的对话框仅仅是图像标题，相当于文本到图像检索任务。DB 阶段向当前对话框提供后续问题。

3.2 对话构建模块

DB 部分包括问题生成器 G 和答案提供者 (人类用户或者预训练的视觉对话模型 BLIP2)。G 负责根据对话 D_i 历史生成下一个问题 Q_{i+1}，作用是引导用户逐步缩小搜索范围。用多

种预训练的大语言模型 (LLM) 作为 G，如 FLAN-T5-XXL、FLAN-ALPACA-XL、FLAN - ALPACA-XXL、ChatGPT 等，并通过特定提示和少量示例指导模型生成问题。同时还采用了一种"Unanswered Questioner" 方法，即让 ChatGPT 仅根据图像描述生成 10 个问题，不参考答案信息此研究答案对问题生成及检索性能的影响。此外，还从 VisDial 数据集中提取人类问题作为"Human" 问题生成器，以模拟人类提问方式。

为了避免依赖人工回答问题，作者使用 BLIP2 作为答案提供者。在实验最后也进行了少量的人机对话实验，以评估 BLIP2 与人类回答之间的差异。为便于大规模实验，在多数情况下使用 BLIP2 模型作为答案提供者回答 G 生成的问题。

3.3 图像检索模块

IS 部分使用一个文本编码器模型 F，将对话序列编码 D_i 映射到视觉特征空间，然后通过余弦相似度对图像库进行检索。F 是基于预训练的 BLIP 图像/文本编码器，并通过对比学习针对基于对话的检索任务进行微调。

图像嵌入模块将语料库中的所有图像编码为单个特征表示 $\bar{f} \in \mathbb{R}^d$，其中 d 为图像嵌入空间维度。在训练 F 时，将对话元素与特殊分隔符 [SEP] 和 [CLS] 连接后输入模型。

为了评估系统性能，作者设计了合适的评估方法，使用 VisDial 数据集作为基准。他们测量了在不同对话轮次下成功检索目标图像的概率，并比较了不同问题生成器 G 的性能。

4 复现细节

4.1 与已有开源代码对比

1. **demo 复现**
   demo 复现的代码用于在 Google Colab 环境中进行设置，与已于开源的代码不同。代码首先检查是否在 Google Colab 环境中运行（'google.colab' in sys.modules)。如果是在 Colab 环境中运行，它会继续使用 pip3 安装特定版本的 Python 包。然后，它通过 git clone 命令克隆名为"BLIP" 的 GitHub 代码仓库。最后，代码使用%cd 命令将当前工作目录更改为"BLIP" 代码仓库的目录。

2. **复现 Image-Text Retrieval**
   在复现图像检索模型时，根据以下实验步骤操作：

   (a) **从 GitHub 上下载 BLIP 源代码：**
   使用以下命令从 GitHub 克隆源代码：

   ```
   git clone https://github.com/salesforce/BLIP.git
   ```

   (b) **下载 COCO 数据集：**
   下载 COCO 数据集，包括 `train2017.zip`、`val2017.zip` 和 `test2017.zip`，并将其解压到同一个文件夹（例如 `datasets/coco/`)。以下是下载链接：

   - train2017.zip
   - val2017.zip

- test2017.zip

(c) **下载预训练模型：**
下载预训练模型 `model_base_retrieval_coco.pth`：

- 下载链接：model_base_retrieval_coco.pth

(d) **修改 `train_retrieval.py` 文件：**
在 `train_retrieval.py` 文件中，修改以下配置部分：

```
--config ./configs/retrieval_coco.yaml \
--output_dir ./output/retrieval_coco
```

(e) **修改配置文件 `retrieval_coco.yaml`：**
修改 `image_root` 路径：设置为 COCO 数据集的本地路径。例如：

```
image_root: ./datasets/coco/
```

修改 `pretrained` 路径：设置为预训练模型的本地路径。例如：

```
pretrained: ./models/model_base_retrieval_coco.pth
```

(f) **修改 BERT 模型路径：**
在代码中找到 `init_tokenizer()` 函数，并修改如下代码：

```
tokenizer = BertTokenizer.from_pretrained('./models/
    bert-base-uncased')
```

需要先从 Hugging Face 下载 `bert-base-uncased` 模型，并保存在本地路径（如 `models/bert-base-uncased/`）。

(g) **确保所有路径为本地路径：**
Overleaf 无法直接访问外部网络，因此所有路径需要使用相对路径并上传到项目中。如果未修改路径，可能会报错 `ConnectionError`，表示尝试从外网下载数据集或模型失败。

(h) **根据 GPU 配置修改**

(i) **发现 yaml 文件里面出现一处错误的地方，对其进行修改**

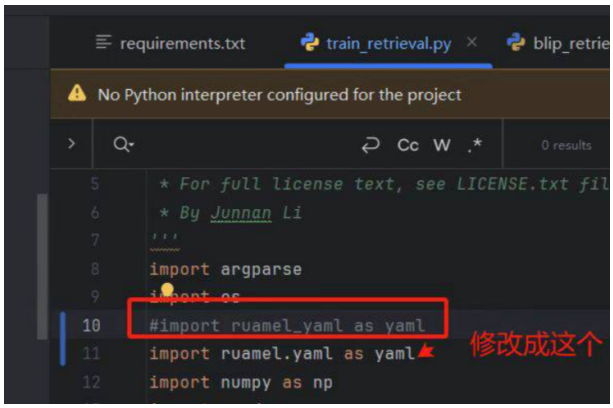

图 2. yaml 文件修改

(j) **安装 ruamel.yaml**

因为外网资源限制，必须挂 vpn 去跑。

(k) **出现 CPU 与 GPU 分配不均问题**

把全部资源转到 GPU 中，避免 CPU 与 GPU 分配不均多次报错。

```
106         image_embed = F.normalize(image_embed,dim=-1)
107
108         image_feats.append(image_feat)
109         # image_feats.append(image_feat.cpu())
110         image_embeds.append(image_embed)
111
112     # image_feats = torch.cat(image_feats,dim=0)
113     image_feats = torch.cat(image_feats, dim=0).to(device)
114     image_embeds = torch.cat(image_embeds,dim=0)
115
116     sims_matrix = image_embeds @ text_embeds.t()
117     score_matrix_i2t = torch.full((len(data_loader.dataset.image),len(texts)),-100.0).to(device)
118
119     num_tasks = utils.get_world_size()
120     rank = utils.get_rank()
121     step = sims_matrix.size(0)//num_tasks + 1
122     start = rank*step
123     end = min(sims_matrix.size(0),start+step)
124
125     for i,sims in enumerate(metric_logger.log_every(sims_matrix[start:end], 50, header)):
126         topk_sim, topk_idx = sims.topk(k=config['k_test'], dim=0)
127         topk_idx = topk_idx.to(device)
128 # 加了一条
129         encoder_output = image_feats[start+i].repeat(config['k_test'],1,1).to(device)
130         encoder_att = torch.ones(encoder_output.size()[:-1],dtype=torch.long).to(device)
```

图 3. 解决 GPU 分配不均问题

4.2 实验环境搭建

1. **对配置包版本进行更新：**

   项目的基本依赖可以通过以下 `requirements.txt` 文件进行安装：

```
timm==0.4.12
transformers==4.15.0
fairscale==0.4.4
pycocoevalcap
```

   使用以下命令安装上述依赖项：

```
pip install -r requirements.txt
```

2. **解决 Torch 版本不兼容的问题：**

   默认情况下，BLIP 项目使用 `torch==1.7.1+cu110` 和 `torchvision==0.8.2+cu110`。但在实验过程中，如果 Python 版本为 3.7，则 `torch 1.7.1` 存在兼容性问题。为解决此问题，将 `torch` 和相关包更新至兼容的版本 `1.10.1`。可以通过以下命令安装兼容版本：

```
pip install torch==1.10.1+cu102 torchvision==0.11.2+cu102
    torchaudio==0.10.1
```

```
    -f https://download.pytorch.org/whl/cu102/torch_stable.html
    -i https://pypi.tuna.tsinghua.edu.cn/simple
```

其中：

- `torch==1.10.1+cu102`：指定 PyTorch 版本，并使用 CUDA 10.2。
- `https://pypi.tuna.tsinghua.edu.cn/simple/`：使用清华镜像源加速安装过程。

3. **下载镜像源进行更新：**
在实验过程中发现，GitHub 提供的镜像源版本较老，直接使用默认的 `pip` 安装会报错。因此，将镜像源更新至清华源，确保后续安装无报错。以下是更新后的源配置：

```
    -f https://download.pytorch.org/whl/torch_stable.html
    -f https://dl.fbaipublicfiles.com/detectron2/wheels/cu110/
       torch1.7/index.html
```

此外，建议在安装前清空旧的缓存，避免因缓存导致的环境冲突：

```
pip cache purge
```

4.3 创新点

1. **优化模型配置和兼容性：**在实验环境中更新了配置包版本，尤其是调整了 PyTorch 的兼容性问题。这种对环境配置的优化使得模型可以在不同的硬件和软件环境下更稳定地运行。

2. **本地化资源路径：**通过手动修改配置文件和模型路径，避免了默认从外网下载的限制。这不仅提高了模型加载的速度，还增强了实验的可控性和稳定性。

3. **增强模型的训练效率：**对 GPU 的使用进行了优化，确保资源分配合理，避免了 CPU 和 GPU 使用不均的问题。这种优化能够有效提高训练效率，使模型训练过程更加高效。

4. **自定义模型微调：**通过对 BLIP 模型的微调，尤其是在对话任务上的应用，展示了其在复杂多模态任务中的优越性。微调过程的创新在于针对特定数据集和任务需求调整模型，以提高其泛化能力。

5. **对比分析实验：**通过对比 CLIP 和 BLIP 模型在不同对话长度下的表现，提供了定量化的分析和结果。这一分析帮助识别了各模型在特定任务中的优劣势，为进一步的模型改进提供了方向。

6. **多任务支持的展示：**在演示过程中展示了 BLIP 模型在多种任务上的应用，如图像标题生成、视觉问答、特征提取等。这种多任务的展示不仅验证了模型的多样性和灵活性，也为其在实际应用中提供了丰富的使用场景。

7. **改进的训练和评估策略：**在训练过程中，采用分布式训练策略并结合动态学习率调度，这对于提升模型性能和训练效果具有显著作用。同样，基于余弦相似度的评估方法提供了更加细致的性能分析。

5 实验结果分析

5.1 图像检索模型训练

在本实验中，我们对图像检索模型的训练过程及其结果进行全面分析。我们使用了 BLIP (Bootstrapping Language-Image Pre-training)模型进行图像-文本检索任务，数据集包括 COCO 和 Flickr30k。

在训练过程中，我们采用了分布式训练策略，使用了 4 张 NVIDIA A100 GPU。训练模型时，首先下载并准备数据集，确保图像根目录在配置文件中正确设置。模型的预训练权重通过指定网址加载，以确保能够在已有基础上进行微调。

数据加载：通过 `create_dataset` 和 `create_loader` 函数构建训练、验证和测试数据集的加载器。

模型构建：实例化 BLIP 模型，并将其移动到指定设备上（GPU）。

优化器设置：采用 AdamW 优化器，结合学习率调度策略，动态调整学习率以提高训练效果。

训练循环：在指定的最大训练轮数内，不断进行模型训练并评估性能，记录训练过程中的损失和学习率。

在训练完成后，我们对模型的性能进行了评估。评估过程中，我们计算了每个图像对应的文本嵌入向量，并使用余弦相似度矩阵来衡量图像和文本之间的相关性。表 1 中展示了模型在验证集和测试集上的评估结果。

表 1. 模型评估结果

评估指标	文本检索结果	图像检索结果
R@1	80.58%	63.24%
R@5	94.84%	85.49%
R@10	97.34%	91.77%
R_mean	90.92%	80.17%

从图 4 中的评估结果来看，BLIP 模型在文本检索方面表现优异，前 1 名的准确率达到 80.58%。而在图像检索方面，相较于文本检索的性能稍显不足，前 1 名的准确率为 63.24%。综合检索结果，即结合文本和图像的检索表现 R_mean 值为 85.54%，这可能与数据集的特性和图像特征的提取方式有关。

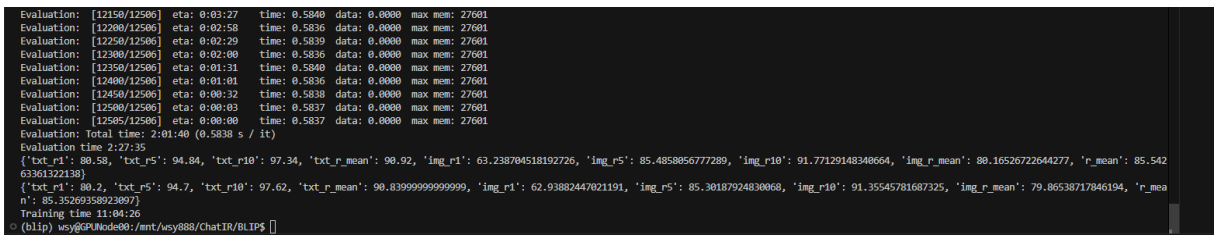

图 4. 模型评估结果图

5.2 ChatIR 任务模型测试结果

本节将对在 ChatIR 任务中使用不同基线模型的测试结果进行分析。通过运行提供的脚本，我们评估了 CLIP zero-shot 基线和经过对话微调的 BLIP 基线在不同对话长度下的性能。测试结果为模型的有效性和检索能力提供了量化依据。

在测试过程中，我们采用了以下的实验配置：

基线选择：在 eval.py 脚本中可以选择使用 CLIP zero-shot 基线或 BLIP 基线。为了进行对比分析，我们分别执行了这两种基线模型的评估。

评估指标：主要评估指标为 Hits@10，表示在给定的对话长度下，模型能够正确检索到目标图像的比例。

运行结果如表 2 所示，在使用 CLIP 模型进行评估时，随着对话长度的增加，模型的检索效果也有所提升。从结果可以看出，CLIP 模型在短对话长度下的检索效果较低，随着对话长度的增加，模型的表现得到了逐步提升，但整体准确率仍然有限。

相较于 CLIP，BLIP 模型在经过对话微调后，表现出更优越的检索能力，运行结果如表 2 所示。BLIP 模型的结果展示了其面对更复杂的对话内容时，能够更有效地检索相关图像。整体来看，BLIP 模型的检索准确率显著高于 CLIP 模型，并且随着对话长度的增加，检索能力持续提升。

表 2. 评估结果 Hits@10

对话长度	CLIP zero-shot Hits@10	微调 BLIP Hits@10
0	43.94%	63.42%
1	48.84%	67.54%
2	51.74%	70.54%
3	53.15%	72.97%
4	54.36%	74.85%
5	55.38%	76.94%
6	56.10%	78.34%
7	56.54%	79.60%
8	56.64%	80.57%
9	56.69%	81.01%
10	56.73%	81.93%

通过对比这两种基线模型的测试结果，可以得出以下结论：

- **CLIP 模型的局限性：**由于 CLIP 模型是为图像标题生成训练的，因此在处理对话数据时，表现较为欠缺，对话长度的限制进一步影响了其检索能力。
- **BLIP 模型的优势：** BLIP 模型经过对话数据的专门微调，显著提高了其在图像与对话匹配任务中的表现，能够更好地理解对话上下文，从而实现更高的召回率。

综上所述，模型的选择在 ChatIR 任务中对检索性能有直接影响，而经过微调的模型通常能获得更优的表现。未来的工作应继续关注模型的优化与改进，以适应更复杂的应用场景。

5.3 demo 演示 ChatIR 工作流程

在本章节，我们实现了对话图像检索系统（ChatIR）的演示，展示了模型在标题生成、开放式视觉问答、多模态特征提取以及图像-文本匹配任务上的应用。

1. **环境配置：**首先，可以通过安装所需的库和克隆 BLIP 仓库来设置环境。

2. **图像标题生成：**使用 BLIP 模型生成图像标题。模型成功识别出图像中的内容，并生成标题，例如“一个女人和她的狗在海滩上”，如图 5 所示。

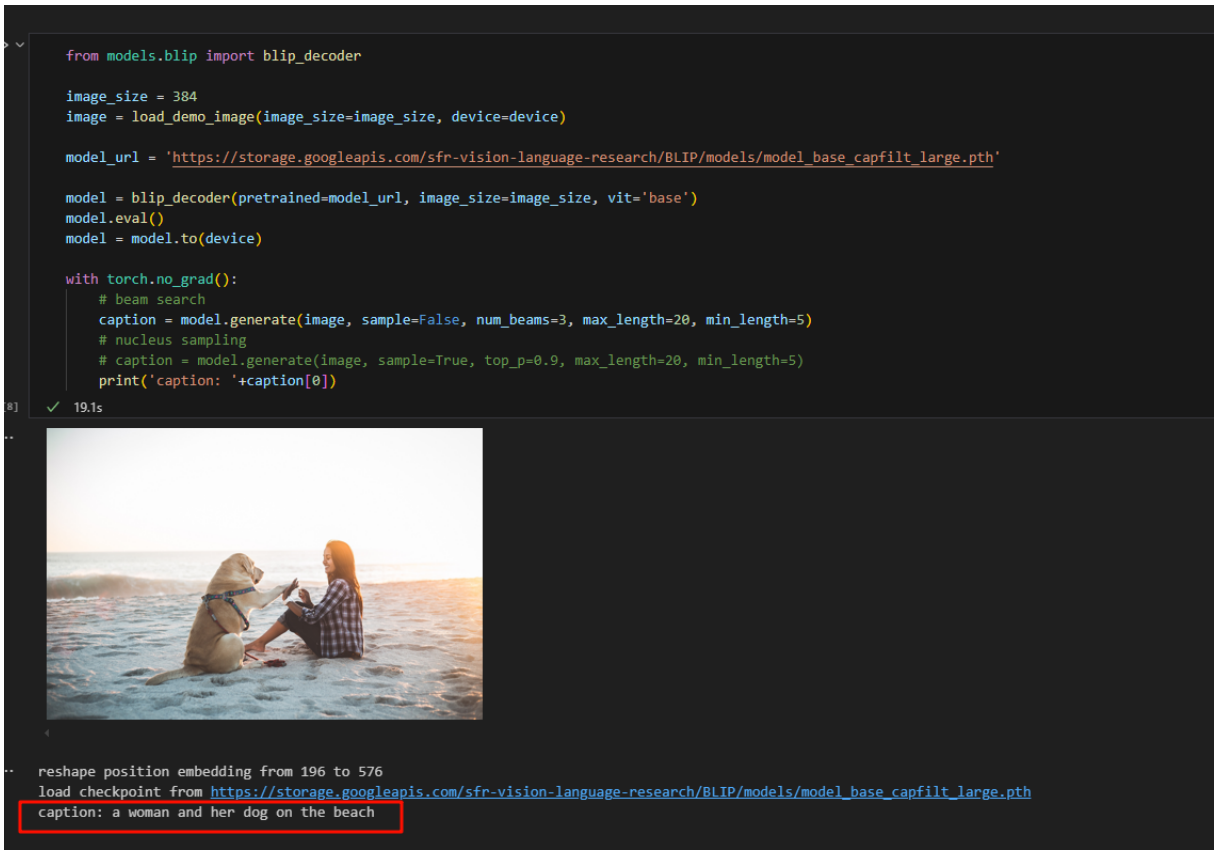

图 5. 图像标题生成结果

3. **视觉问答：**BLIP 模型支持视觉问答功能，用户可以提出问题，模型将基于图像提供答案。在本例中，用户询问“女人坐在哪里?”模型正确回答“在海滩上”，如图 6 所示。

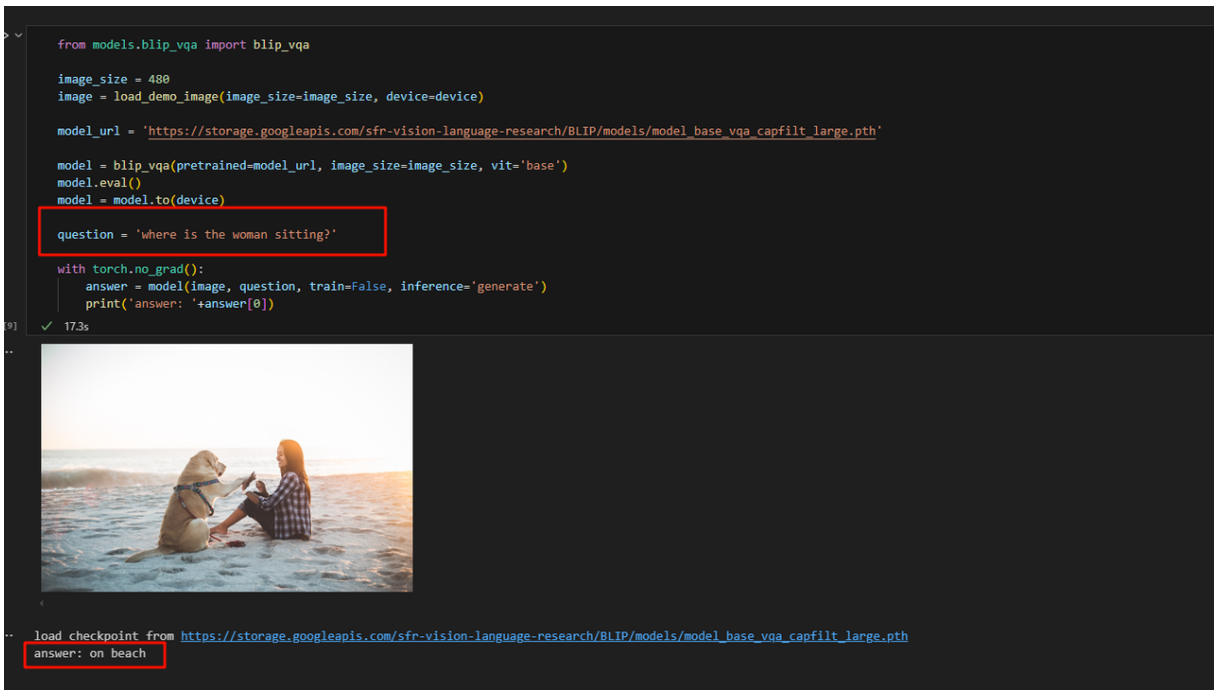

图 6. 视觉问答结果

4. **特征提取：** BLIP 模型能够提取图像和文本的特征，这些特征可以用于后续的图像检索任务，如图 7 所示。

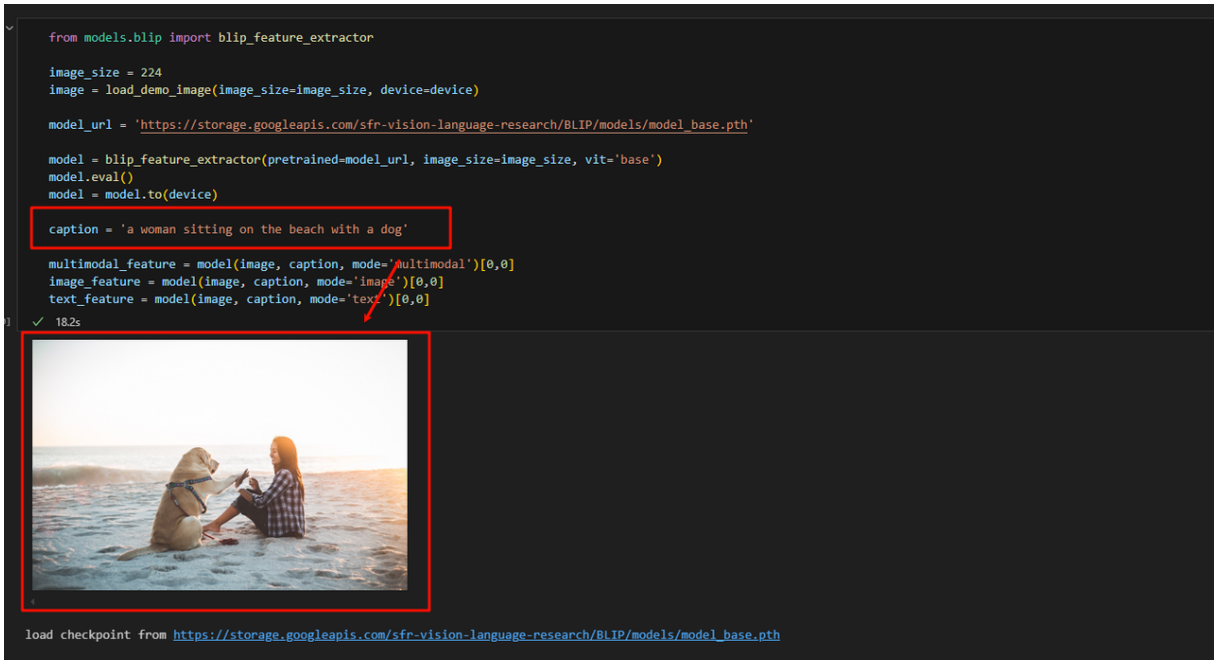

图 7. 特征提取结果

5. **图像与文本匹配：** 最后，使用 BLIP 模型进行图像与文本的匹配，计算图像特征与文本特征之间的余弦相似度，计算匹配概率。在这个例子中，模型输出匹配概率为 0.9960，表明图像与文本之间的匹配程度非常高。而余弦相似度为 0.5262，表示图像特征与文本特征之间的相似度，如图 8 所示。

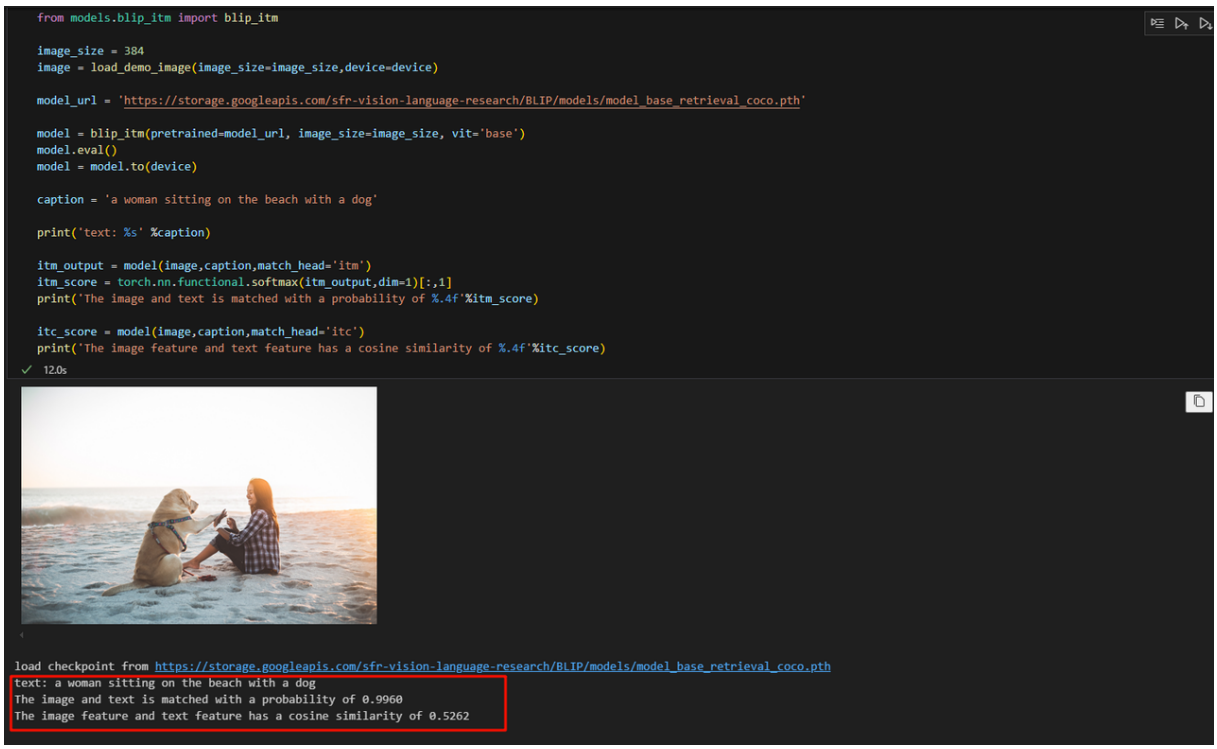

图 8. 图文匹配结果

6 总结与展望

本文基于《Chatting Makes Perfect: Chat-based Image Retrieval》进行复现研究，这篇文章提出了一种全新的图像检索模型 ChatIR，它通过与用户的对话式交互逐步优化检索结果，使得检索过程更加自然和高效。ChatIR 的核心特点是通过提问和回答的多轮对话，将用户的意图精准转化为检索条件，从而改进搜索结果的相关性。通过复现实验表明，该模型利用基础模型可以在性能上接近人类提问者，充分展示了基于对话的检索在智能系统中的巨大潜力。

通过实验复现和结果分析，本文发现有效问题的生成是模型成功的关键所在，而无法持续生成新的高质量问题是某些失败案例的主要原因之一。目前，ChatIR 仅依赖对话历史作为生成后续问题的依据，未充分利用检索结果或候选项来提取显著特征以进一步聚焦用户需求。在理想情况下，一个更优化的提问者应能够综合考虑检索结果与用户意图，提出更针对性的问题以缩小选择范围。

尽管还有问题的存在，但 ChatIR 为将聊天形式应用于图像检索提供了一个创新框架。它不仅展示了通过对话改进检索的可行性，还为人机交互技术的发展提供了新的研究方向。未来的工作可以进一步探索如何结合检索反馈与候选信息，改进模型的提问逻辑，提升交互性能。总体而言，ChatIR 为复杂检索场景中的多轮对话和动态优化奠定了重要基础，同时也为基于对话的智能检索技术提供了新的思路。

参考文献

[1] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 61437–61449. Curran Associates, Inc., 2023.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[6] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pages 25994–26009, 2022.

[7] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21434–21442, 2022.

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

[9] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6116–6124, 2018.

[10] Badri Patro, Vinod Kurmi, Sandeep Kumar, and Vinay Namboodiri. Deep bayesian network for visual question generation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1566–1576, 2020.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12363 of *Lecture Notes in Computer Science*, pages 336–352. Springer, 2020.

[13] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive image generation using scene graphs. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023.

[16] Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2951–2960, 2017.

[17] Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. Improving generative visual dialog by answering diverse questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[18] Zipeng Xu, Fandong Meng, Xiaojie Wang, Duo Zheng, Chenxu Lv, and Jie Zhou. Modeling explicit concerning states for reinforcement learning in visual dialogue. *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International*

Conference on Machine Learning (ICML), volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

[21] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023.

[22] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogério Schmidt Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 676–686, 2018.

[23] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[24] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2973–2980. IEEE Computer Society, 2012.

[25] Lorenzo Putzu, Luca Piras, and Giorgio Giacinto. Convolutional neural networks for relevance feedback in content-based image retrieval: A content-based image retrieval system that exploits convolutional neural networks both for feature extraction and for relevance feedback. *Multimedia Tools and Applications*, 79:26995–27021, 2020.

[26] Devi Parikh and Kristen Grauman. Relative attributes. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[27] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2105–2114, 2021.