2024 年（第 12 届）“泰迪杯”数据挖掘挑战赛——

B 题：基于多模态特征融合的图像文本检索

一、问题背景

随着近年来智能终端设备和多媒体社交网络平台的飞速发展，多媒体数据呈现海量增长的趋势，使当今主流的社交网络平台充斥着海量的文本、图像等多模态媒体数据，也使得人们对不同模态数据之间互相检索的需求不断增加。有效的信息检索和分析可以大大提高平台多模态数据的利用率及用户的使用体验，而不同模态间存在显著的语义鸿沟，大大制约了海量多模态数据的分析及有效信息挖掘。因此，在海量的数据中实现跨模态信息的精准检索就成为当今学术界面临的重要挑战。图像和文本作为信息传递过程中常见的两大模态，它们之间的交互检索不仅能有效打破视觉和语言之间的语义鸿沟和分布壁垒，还能促进许多应用的发展，如跨模态检索、图像标注、视觉问答等。

图像文本检索指的是输入某一模态的数据（例如图像），通过训练的模型自动检索出与之最相关的另一模态数据（例如文本），它包括两个方向的检索，即基于文本的图像检索和基于图像的文本检索，如图 1 所示。基于文本的图像检索的目的是从数据库中找到与输入句子相匹配的图像作为输出结果；基于图像的文本检索根据输入图片，模型从数据库中自动检索出能够准确描述图片内容的文字。然而，来自图像和来自文本的特征存在固有的数据分布的差异，也被称为模态间的“异构鸿沟”，使得度量图像和文本之间的语义相关性困难重重。

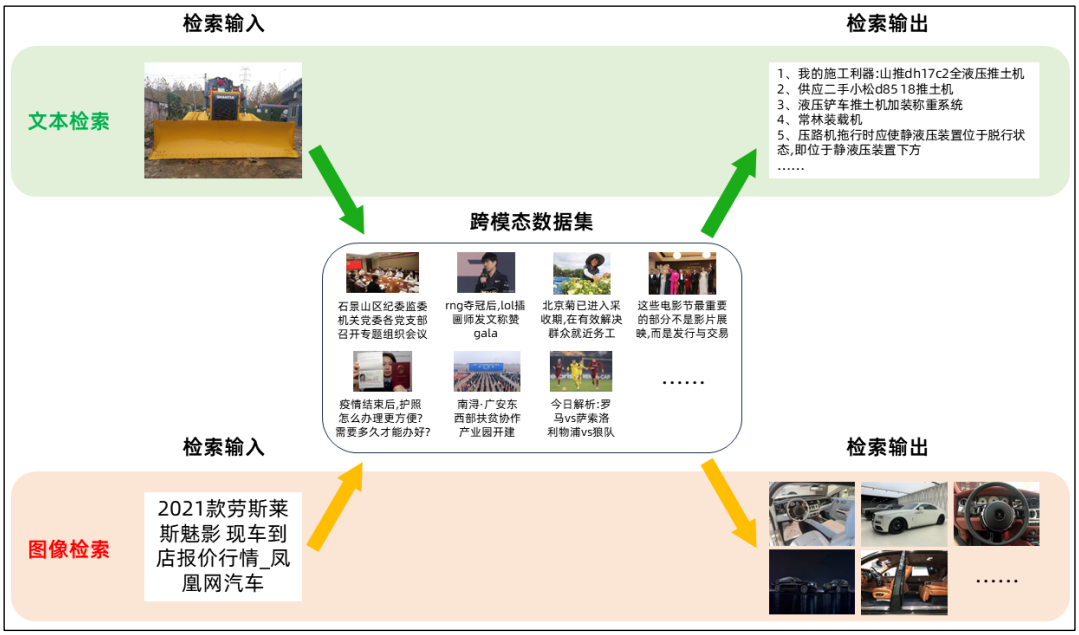

图 1 图像文本检索

二、解决问题

本赛题是利用附件 1 的数据集，选择合适方法进行图像和文本的特征提取，基于提取的特征数据，建立适用于**图像检索**的多模态特征融合模型和算法，以及建立适用于**文本检索**的多模态特征融合模型和算法。基于建立的“多模态特征融合的图像文本检索”模型，完成以下两个任务，并提交相关材料。

（1）**基于图像检索的模型和算法**，利用附件 2 中“word_test.csv”文件的文本信息，对附件 2 的 ImageData 文件夹的图像进行图像检索，并罗列检索相似度较高的前五张图像，将结果存放在“result1.csv”文件中（**模板文件详见附件 4 的 result1.csv**）。其中，ImageData 文件夹中的图像 ID 详见附件 2 的“image_data.csv”文件。

（2）**基于文本检索的模型和算法**，利用附件 3 中“image_test.csv”文件提及的图像 ID，对附件 3 的“word_data.csv”文件进行文本检索，并罗列检索相似度较高的前五条文本，将结果存放在“result2.csv”文件中（**模板文件见附件 4 的 result2.csv**）。其中，“image_test.csv”文件提及的图像 id，对应的图像数据可在附件 3 的 ImageData 文件夹中获取。

三、附件说明

附件 1、附件 2、附件 3 和附件 4 均含 csv 文件，采用 UTF-8 编码格式。

附件 1：图像文本检索的数据集，“ImageData”压缩包存储五万张图像，“ImageWordData.csv”文件存储图像数据对应的文本信息，如表 1 所示。其中，“image_id”为图像 ID，也是图像的文件名，可依据图像 ID 获取“caption”中图像对应的文本信息。

表 1 图像文本检索的数据集——CSV 文件示例内容

image_id	caption
Image14001001-0000.jpg	《绿色北京》摄影大赛胡子<人名>作品
Image14001001-0002.jpg	招聘计划学校现有教职工 1500 余人.
……	……

附件 2：本赛题任务（1）的数据信息，包含“word_test.csv”、“image_data.csv”两份 CSV 文件和 ImageData 文件夹。其中，“word_test.csv”属于测试集图像检索文本信息，记录了文本 ID 和文本内容，文件格式如表 2 所示；“image_data.csv”记录了 ImageData 文件夹中的图像 ID，文件格式如表 3 所示；ImageData 文件夹为任务（1）的图像数据库，存放了能与“image_data.csv”匹配的图像数据，如图 2 所示。

表 2 word_test.csv 示例内容

text_id	caption
Word-1000004254	后来美国历史学家及情报部高官说:金无怠的的间谍活动是导致韩战延迟
Word-1000030077	茶主题商业综合体的未来当下,如果专业市场只是安于做一个收商铺租赁
……	……

表 3 image_data.csv 示例内容

image_id
Image14001007-4040.jpg
Image14001007-4041.jpg
……

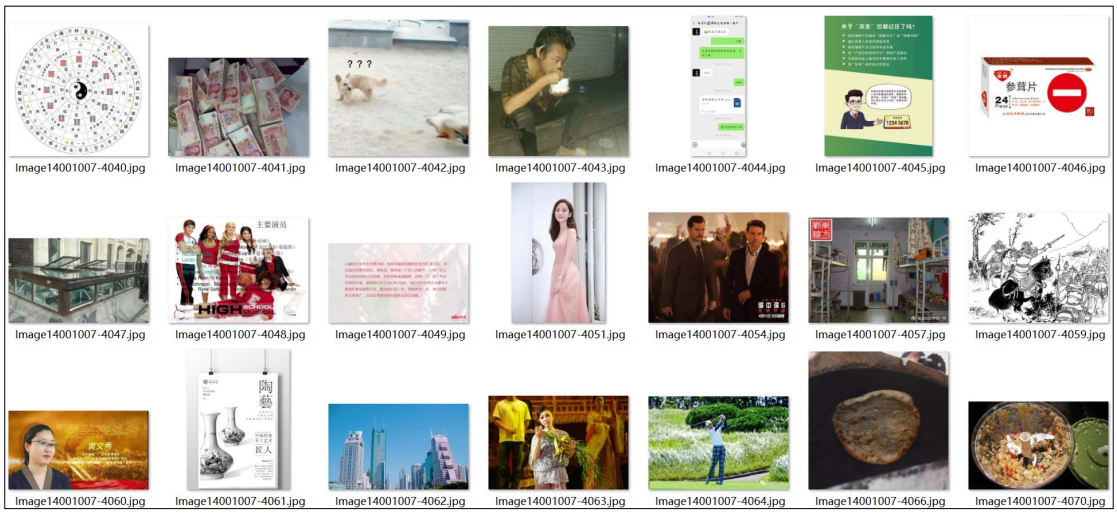

图 2 附件 2 的 ImageData 文件夹内容

附件 3：本赛题任务（2）的数据信息，包含“word_data.csv”、“image_test.csv”两份 CSV 文件和 ImageData 文件夹。其中，“word_data.csv”属于测试集文本检索文本信息，记录了文本 ID 和文本内容，文件格式如表 4 所示；“image_test.csv”记录了 ImageData 文件夹中的图像 ID，文件格式如表 5 所示；ImageData 文件夹为任务（2）的图像数据库，存放了能与“image_test.csv”匹配的图像数据，如图 3 所示。

表 4 word_data.csv 示例内容

text_id	caption
Word-1000050001	洛阳楼盘 老城区楼盘 道北楼盘 保利<人名>
Word-1000050002	大众大众(进口)途锐 2015 款 基本型
……	……

表 5 image_test.csv 示例内容

image_id
Image14001013-8213.jpg
Image14001013-8214.jpg
……

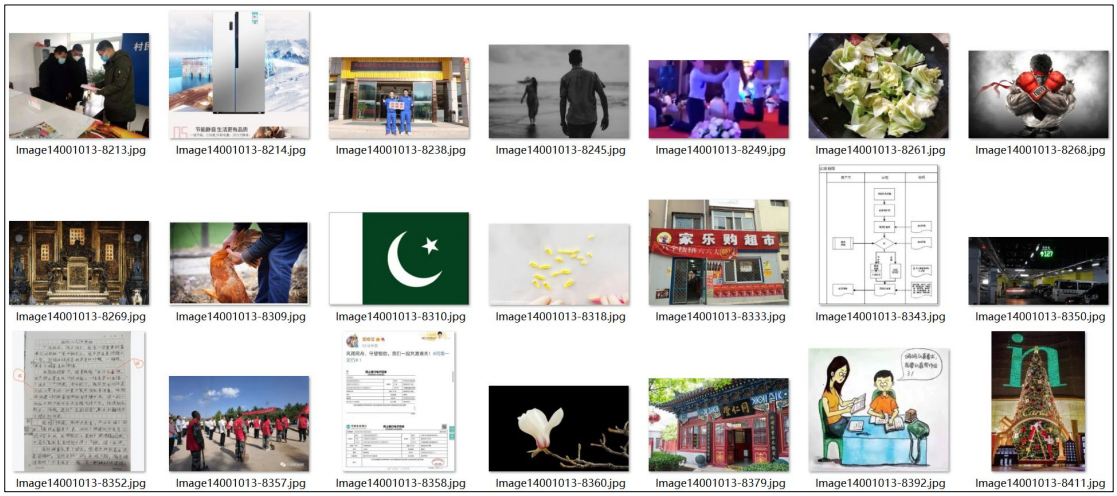

图 3 附件 3 的 ImageData 文件夹内容

附件 4：任务（1）和任务（2）结果文件的模板文件，具体字段名称和样例见表 6 和表 7。“result1.csv”中，text_id 是附件 2“word_test.csv”文件的文本 ID，similarity_ranking 是相似度排名，result_image_id 是相似度排名对应在“image_data.csv”文件的图像 ID；“result2.csv”中，image_id 是附件 2“image_test.csv”文件的图像 ID，similarity_ranking 是相似度排名，result_text_id 是相似度排名对应在“word_data.csv”文件的文本 ID。

表 6 result1.csv 示例内容

text_id	similarity_ranking	result_image_id
Word-1000000001	1	Image00010804-0898.jpg
	2	Image00015036-0854.jpg
	3	Image00018364-0375.jpg
	4	Image00042681-0598.jpg
	5	Image00038751-0658.jpg
Word-1000000002	1	Image00010804-0697.jpg
	2	Image00015036-0158.jpg
	3	Image00018364-0319.jpg
	4	Image00042681-0135.jpg
	5	Image00038751-0356.jpg
……	……	……

表 7 result2.csv 示例内容

image_id	similarity_ranking	result_text_id
Image00012212-0001.jpg	1	Word-1000001175
	2	Word-1000001658
	3	Word-1000001574
	4	Word-1000001359
	5	Word-1000001514
Image00012212-0002.jpg	1	Word-1000001124
	2	Word-1000001242
	3	Word-1000001425
	4	Word-1000001113
	5	Word-1000001854
……	……	……

四、评价标准

图像文本检索包括两个具体的任务，即文本检索（Image-to-Text，I2T），即针对查询图像找到相关句子；以及图像检索（Text-to-Image，T2I），即给定查询语句检索符合文本描述的图像。为了与现有方法公平地进行比较，在文本检索问题和图像检索问题中都采用了广泛使用的评价指标：召回率 Recall at K（R@K）。R@K 定义为查询结果中真实结果（ground-truth）排序在前 K 的比率，通常 K 可取值为 1、5 和 10，计算公式如式（1）所示。

$$R@K = \frac{Matched_{top-K}}{Groundtruth_{total}}$$

其中，$Groundtruth_{total}$ 表示真实匹配结果出现的总次数，$Matched_{top-K}$ 表示在排序前 K 个输出结果中出现匹配样本的次数。R@K 反映了在图像检索和文本检索中模型输出前 K 个结果中正确结果出现的比例。**本赛题的评价标准设定 K=5，即评价标准为 R@5。**