基于多模态特征融合的图像文本检索

李松泽　吴钰茹　王俊杰　何劲仪　曾雨琪　但松健
重庆第二师范学院，重庆 400000

摘　要：随着智能终端和多媒体社交网络的快速发展，多模态数据（如文本和图像）呈现爆炸式增长，导致对不同模态数据互相检索的需求日益增加。然而，模态之间的语义鸿沟限制了对海量多模态数据的有效分析和信息挖掘。因此，实现精准的跨模态信息检索成为学术界的重要挑战，尤其是在文本生成图像和图像生成文本的应用场景中。本文研究了基于 5000 条文本信息和 50000 张图片的文本生成图像检索，以及基于 50000 条文本信息和 5000 张图片的图像生成文本检索。通过计算归一化特征之间的余弦相似度，找出相似度排名前五的图像或文本。研究采用深度学习方法，特别是 CN-CLIP 模型，促进文本与图像的多模态特征融合，实现双向生成，提升用户的检索体验。CN-CLIP 模型在多模态表征学习中具有重要意义，采用对比学习方式在大规模图像 – 文本对数据集上进行预训练，成功建立视觉与语言之间的联系。该模型在视觉语言检索中表现优异，并在零镜头图像分类中展现出色性能。其简单有效的方法论推动了多模态表征学习和计算机视觉的研究进展，为跨模态信息检索、图像标注和视觉问答等应用提供了强大支持。随着技术的不断进步，CN-CLIP 模型将继续在多模态学习、图像标注、视觉问答等领域发挥重要作用，引领新的研究方向。

关键词：文本生成图像；图像生成文本；多模态特征融合；CN-CLIP 模型

前言

随着互联网和数字技术的发展，海量图像和文本数据被生成和传播，如何有效检索用户需求相关信息已成为研究热点。图像文本检索作为一种跨模态的信息检索任务，旨在根据文本描述从图像数据库中找到相关图像，广泛应用于搜索引擎、社交媒体和广告推荐等领域。

然而，由于图像和文本之间的语义鸿沟，传统的单模态特征检索方法效果有限。为此，研究人员开始探索多模态特征融合的方法，通过有效组合不同模态的特征，利用它们的互补信息来提高检索的准确性和鲁棒性。尽管此方法在提升检索性能方面取得进展，但依然面临特征融合、处理大规模数据集和提高检索速度等挑战。因此，基于多模态特征融合的图像文本检索具有重要的理论和实际价值。　随着智能终端和多媒体社交网络的快速发展，多媒体数据呈现海量增长，如何有效检索与用户需求相关的图像成为研究热点。CN-CLIP 模型在多模态特征融合方面展现出卓越性能，通过对比学习在大规模数据集上预训练，有效融合了图像和文本特征。该模型能够准确理解图像内容，并与文本信息精准匹配，实现高效的跨模态检索，同时具备强大的泛化能力，适应不同领域的多模态特征融合需求 [2]。在图像特征提取方面，CN-CLIP 使用预训练的卷积神经网络（CNN），如 ResNet 和 EfficientNet，提取关键特征（如边缘、纹理、形状和颜色）。图像经过多层卷积和池化后，输出高维特征向量，编码图像内容信息。文本特征提取则采用自然语言处理领域的 Transformer 结构，如 BERT 或 RoBERTa 变种，通过处理大量文本数据学习词语和句子间的语义关系。文本转换为词嵌入向量后输入 Transformer，利用自注意力机制生成代表文本语义的特征向量 [3]。CN-CLIP 通过对比学习计算图像和文本特征向量之间的相似度，优化模型参数，使相似内容的图像和文本在特征空间上相互靠近。特征融合方法包括加权求和、乘法融合和多层感知机等，研究提出了 Image-Text Matching Loss (ITM) 来激活基于图像的文本编码器，捕捉视觉和语言之间的细粒度对齐。相似度计算通过余弦相似度、欧氏距离等方法实现，使用 softmax 函数将相似度分数转化为概率分布，排序后找出与文本最相似的 5 个图像的索引，从 CSV 文件中读取相关信息。

1 关于跨模态特征的有效融合的解决方案

1.1 问题分析及思路

需要建立适用于图像检索的多模态特征融合模型和算法。

（1）图像和文本数据预处理

设置图像数据的文件夹路径、列出文件夹的所有文件，并将其中文本转化为模型。我们需要处理“word_test.csv”文件中的文本信息。这通常涉及文本预处理，如分词、去除停用词等，使用自然语言处理模型（如 word2vec、BERT 等）将文本转化为向量表示。

（2）提取图像和文本特征

初始化张量存储图像与文本的相似度，对于 ImageData 文件夹中的图像，我们需要使用图像特征提取算法（如 CNN、ResNet 等）来提取图像特征，并将这些特征转化为与文本向量可比较的向量表示。

（3）多模态特征融合模型和算法

分别得到图像和文本的特征后，建立一个多模态特征融合模型来整合这些特征。常见的模型包括：Chinese Clip 模型、向量拼接（Concatenation）、双向编码器（Bi-Encoder）、Transformer 模型等。计算特征向量之间的余弦相似度、欧氏距离或其他度量指标，根据计算出来的相似度，对图像进行排序，选择相似度前五的图像。

1.2 基于图像检索的模型和算法

1.2.1 模型建立

为了保证图像检索的准确性、效率性，我们经过大量的文献查阅和方法对比，发现 CN-CLIP 模型训练的效果较好。CN-CLIP 与传统的生成式预训练不同，它是一种基于对比学习的模型，其在一个大规模数据集上进行预训练的，该数据集包括从网络上收集的约 4 亿个图像 - 文本对数据。CN-CLIP 作为视觉基础模型，在一系列数据集的零镜头图像分类中表现出了最先进的性能。CN-CLIP 建立了视觉与语言之间的联系，改变了多模态表征学习和计算机视觉领域的研究。

我们主要采用 CN-CLIP 模型，其能够跨模态的匹配和检索任务。即给定一组图像描述（从文件夹中读取），代码计算每个图像描述与给定文本的相似度，并找出文本最相似的图像描述。初始阶段以预训练的方式设定了两种编码器，分别是 CLIP 的视觉编码器和中文版的 ROBERTA 文本编辑器。CN-CLIP 的主要思想是冻结 Image Encoder 使用 VIT 让 Text Encoder 能够从 Open AI 的 CLIP 的基础视觉模型中读出高质量的表示，接着将这些表示迁移到需要的数据中。CN-CLIP 使用的数据集包括 LAION-5B 中文子集、Wukong 的中文数据、COCO、Visual Genome 的翻译图文数据等等。[1]

（1）导入模型

采用“ViT-H-14”的 CN-CLIP 模型，具有 2 亿的参数规模，设置其为评估模式。Vision Transformer 是一种基于 Transformer 架构的视觉模型，它首先将图像分割成一系列的小块（patches），然后像处理文本序列一样处理这些小块。

（2）设置文件夹路径和文本

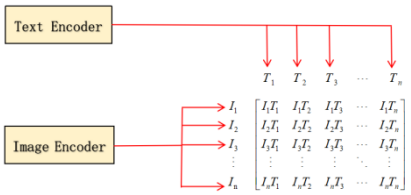
列出该文件夹中的所有文件的文件夹路径，使用 CN-CLIP.tokenize 将中文文本转化为模型可以理解的嵌入表示，并放到指定的 GPU/CPU 设备上。

（3）提取图像和文本特征

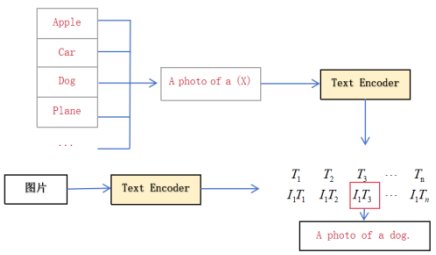
初始张量 features 来存储图像与文本的相似度分数，提取文本特征，并对其进行归一化。遍历文件夹的每个文件：检查文件是否为图像文件、使用预处理函数处理图像、提取图像特征进行归一化、计算图像与文本的相似度分数，将其添加到张量中。

（4）处理相似度分数并输出结果

使用 softmax 函数，将相似度分数转化为概率分布，对概率进行排序，并找出文本最相似的 5 个图像的索引。从 CSV 文件中读取图像信息，输出与文本最相似的 5 个图像的 ID 或相关信息。



（a）对比预训练



（b）从标签文本创建数据集分类器并用于预测

图 1 CN-CLIP 模型实现步骤

1.2.2 模型实现

要求实现，对 word_test.csv 中的每行文本，从 imageData 文件夹中检索出最相似的 5 张图片，并按相似度排序，用序

号表示。首先需要用 ImageWordData.csv 和 ImageData 作为训练集，训练多模态模型，接着用来测试数据。

表 1 基于图像检索的实现结果

text_id	similarity_ranking	result_image_id
《绿色北京》摄影大赛胡子＜人名＞作品	1	Image14001007-1973.jpg
	2	Image14001005-6093.jpg
	3	Image14001005-2940.jpg
	4	Image14001002-8748.jpg
	5	Image14001004-4091.jpg
招聘计划学校现有教职工 1500 余人 .	1	Image14001002-4678.jpg
	2	Image14001004-1574.jpg
	3	Image14001001-0002.jpg
	4	Image14001002-0364.jpg
	5	Image14001002-2729.jpg
...	...	...

提取文本信息“《绿色北京》摄影大赛胡子＜人名＞作品”的特征，通过 CN-CLIP 模型得到相似度排名前五的图片为如下五张图片。

图 2 图像相似度从大到小排列

2 针对文本检索的多模态特征融合模型和算法解决方案

2.1 问题分析及思路

（1）图像特征提取

通常使用预训练的卷积神经网络（CNN），如 ResNet，来提取图像的特征。这些特征被编码为一个向量表示。

（2）文本特征提取

使用 Transformer 架构（如 BERT）来编码文本。文本首先被转换为 token 序列，然后通过 Transformer 模型处理，最终得到文本的向量表示。

（3）对比学习

将图像和文本的向量表示送入一个对比损失函数，如对比损失（Contrastive Loss）或 InfoNCE 损失。这个损失函数会鼓励模型将匹配的图像和文本对拉近，而将不匹配的对推远。通过反向传播优化这个损失函数，更新图像编码器和文本编码器的参数。

（4）多模态特征融合模型和算法

分别得到图像和文本的特征后，建立一个多模态特征融合模型来整合这些特征。常见的模型包括：向量拼接（Concatenation）、双向编码器（Bi-Encoder）、Transformer 模型、多层感知机（MLP）、注意力机制（Attention）。

（5）特定的损失函数

在多模态的模型中，需要考虑对应的损失函数（如 Triplet Loss、Contrastive Loss 等）来训练模型，使得模型能够更好地学习多模态特征融合的表示能力。

2.2 基于文本检索的模型和算法

2.2.1 模型建立

问题二需要建立适用于文本检索的多模态特征融合模型和算法。

基于文本检索主要采用 CN-CLIP 模型，其能够跨模态的匹配和检索任务。即给定一组文本描述（从 CSV 文件中读取），代码计算每个文本描述与给定图像的相似度，并找出图像最相似的文本描述。CILP 模型是一个联合嵌入模型，具有多模态能力、通用性、零射击能力等。

（1）读取文本数据

使用 pandas 库读取 CSV 文件，对 ImageData 文件夹中的 caption 列进行文本特征提取，并利用 CN-CLIP 模型的 tokenize 函数将文本信息转换为模型，移动到计算 GPU/CPU 上。

（2）处理并提取图像特征

从附件 3 的 ImageData 文件夹中加载与图像 ID 对应的图像数据，使用预处理函数 CN-CLIP 库处理图像，将处理后的图像张量移动到 GPU/CPU 上。

（3）多模态特征融合模型和算法

分别得到图像和文本的特征后，建立一个多模态特征融合模型 CN-CLIP，采用模型的 encode_image 和 encode_text 分别提取图像和文本特征，对特征进行归一化，确保特征向量的长度为 1。

（4）计算并处理相似度结果

为使得模型能够更好地学习多模态特征融合的表示能力，采用模型的 get_similarity 方法计算图像和文本之间的相似度（通过计算归一化特征之间的余弦相似度实现）。

$$\text{sim} = \cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

其中，$\cos\theta$ 表示 $\vec{A}$ 和 $\vec{B}$ 之间的夹角的余弦值，$A \cdot B$ 表示 $\vec{A}$ 和 $\vec{B}$ 的内积，$\|A\|$ 和 $\|B\|$ 分别表示 $\vec{A}$ 和 $\vec{B}$ 的模长。

对 logits_per_image 应用 softmax 函数，得到每个文本描述作为图像描述的概率，根据概率，找出概率最高的 5 个文本描述索引。

2.2.2 模型实现

在模型实现过程中，我们采用 python 语言作为工具，使用 Pycharm 软件，基于深度学习实现算法。

要求实现对 image_test.csv 中的图像 ID，从 word_data.csv 文件夹中检索出最相似的 5 条文本，并按相似度排序。首先需要用 ImageWordData.csv 和 ImageData 作为训练集，训练多模态模型，用来测试数据。

表 2 基于文本检索的实现结果

image_id	similarity_ranking	result_text_id
Image14105004–6502.jpg	1	随意进行一个交易，弹出付款页面，支付密码输入错误后会弹出提示框
	2	反复付款不成功时立即停止操作 网购选用第三方支付平台进行付款时
	3	现在支付宝至少要绑定银行卡才能收
	4	国家下铁命令：微信 / 支付宝迎”大限”！付款改流程，资金更安全
	5	支钱百？共发，忙转八转账账骗生，六付党转付求
Image14105004–6485.jpg	1	日销 600 碗！火爆杭州 17 年的片儿川「扛把子」，年末放大招！
	2	拿好这份螺蛳地图，总有一家店让你神魂颠倒！
	3	溢碗葫芦头
	4	资阳这家新开业河鲜店，竟然敢这样做！
	5	十二道风味 3 折风暴登陆金融街万达，先下手为”抢”
…	…	…

初始化 ViT–H–14 视觉模型，以图 3（Image14105004–6502）为图像样本，检索所有文本对其的相似度排行（如表 2）。

图 3 图像 ID（Image14105004–6502.jpg）

根据运行结果数据，我们可知，与图片相似度最高的 5 条文字，分别是：随意进行一个交易，弹出付款页面，支付密码输入错误后会弹出提示框、反复付款不成功时立即停止操作，网购选用第三方支付平台进行付款时、现在支付宝至少要绑定银行卡才能收、国家下铁命令：微信 / 支付宝迎”大限”！付款改流程，资金更安全、支钱百？共发，忙转八转账账骗生，六付党转付求。

图 4 图像 ID（Image14105004–6485.jpg）

以图 4（Image14105004–6485.jpg）为样本，运行得出与图片相似度最高的 5 条文字，分别是：“日销 600 碗！火爆杭州 17 年的片儿川「扛把子」，年末放大招！”、“拿好这份螺蛳地图，总有一家店让你神魂颠倒！”、“溢碗葫芦头”、“资阳这家新开业河鲜店，竟然敢这样做！”、“十二道风味 3 折风暴登陆金融街万达，先下手为”抢”“。

3 结论

基于多模态特征融合的图像文本检索研究表明，通过有效融合图像和文本特征，可以显著提高检索的准确性和效率，利用互补信息更好地满足用户需求。CN–CLIP 模型在

这一领域展现出卓越性能，通过对比学习精准提取特征，实现高效匹配，且具有良好的类内相似性和泛化能力。该模型在不同数据集上表现出色，运算效率高、资源消耗低，未来具有广阔的应用前景。

基于多模态特征融合的图像文本检索论文专注于提升检索准确性和效率，主要通过 CN-CLIP 模型实现。这一模型通过设置图像数据路径和转化中文本为适合模型处理的格式，实现跨模态匹配。它通过对比学习有效融合图像和文本特征，快速计算相似度并选出最相似的图像。CN-CLIP 展现出强大的泛化能力，适用于不同场景，并且运算效率高，资源消耗低。

然而，该模型也存在一些缺陷，如需要大量高质量的多模态数据，数据不足会影响性能。此外，模型复杂性高可能导致过拟合，需要采取正则化措施。随着技术发展，新的特征融合方法不断涌现，CN-CLIP 需及时更新以保持竞争力。

参考文献：

[1] 李源凡 , 张丽红 . 基于 CLIP 模型和文本重建的人脸图像生成方法研究 [J]. 测试技术学报 ,2024,38(02):154-160.

[2] 肖佳涛 . 基于深度学习的中文文本生成国画图案方法研究 [D]. 景德镇陶瓷大学 ,2023.

[3] 张佳 . 基于深度学习的文本生成图像方法研究 [D]. 山西大学 ,2024.

基金项目：

2024年重庆第二师范学院大学生科研项目“基于多模态特征融合的图像文本检索技术的研究”（项目编号：KY20240048）