Computer Science and Application 计算机科学与应用, 2024, 14(1), 123-133
Published Online January 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.141014

Hans汉斯

基于深度学习的通用本地图像检索系统设计

张浩东，田春岐

同济大学电子与信息工程学院，上海

收稿日期：2023年12月25日；录用日期：2024年1月24日；发布日期：2024年1月31日

摘　要

随着大量数字图像数据的产生，高效准确的图像检索技术变得尤为重要。本文提出了一种结合深度学习和磁盘向量检索技术的通用本地图像检索系统，采用了深度神经网络模型作为特征提取的主要工具，通过深层网络结构捕获图像的高层语义信息，实现对图像内容的精细描述，旨在提升检索的准确性和效率，图像数据库的容量。由具体的实例数据验证说明了系统可用性，证明了其在实际应用中的广泛适用性，文中研究可对图像检索系统的进一步发展起到积极的参考作用。

关键词

图像检索，深度学习，磁盘向量检索，检索方法

Design of a General Local Image Retrieval System Based on Deep Learning

Haodong Zhang, Chunqi Tian

College of Electronic and Information Engineering, Tongji University, Shanghai

Received: Dec. 25th, 2023; accepted: Jan. 24th, 2024; published: Jan. 31st, 2024

Abstract

With the generation of a massive amount of digital image data, efficient and accurate image retrieval technology has become particularly important. This paper proposes a universal local image retrieval system that combines deep learning and disk vector retrieval technology, utilizing deep neural network models as the main tool for feature extraction. By capturing the high-level semantic information of images through deep network structures, the system achieves a fine-grained description of image content, aiming to enhance the accuracy and efficiency of retrieval, as well as the capacity of the image database. The usability of the system is demonstrated through specific instance data, proving its wide applicability in practical applications. The re-

文章引用: 张浩东, 田春岐. 基于深度学习的通用本地图像检索系统设计[J]. 计算机科学与应用, 2024, 14(1): 123-133.
DOI: 10.12677/csa.2024.141014

search presented in this paper can play a positive role in the further development of image retrieval systems.

Keywords

Image Retrieval, Deep Learning, Disk Vector Retrieval, Retrieval Measure

1. 引言

随着互联网多媒体的持续发展，网络传播的各类数字媒体数量呈现指数爆炸式增长，具有广泛的传播范围和快速的传播速度特点。在这其中，图像作为最基本的数字媒体形式之一，已在社会生活和工作领域得到广泛应用。然而，传统的关键字检索已无法满足当前丰富的信息形式需求，因此，各大门户网站纷纷针对自身业务推出相应的图像检索系统。得益于深度学习技术的进步，基于内容的图像检索方法已不再局限于人为划分，而是依赖深度学习模型对图像进行高层次特征提取，从而更准确地实现图像相似度检索，提高检索准确率。本文将介绍几种基于内容的图像检索方法，针对本地需求，应用最新的基于磁盘的向量检索技术，可以本地实现大规模图像向量检索数据库，并设计了一个图像检索系统模型，同时详细阐述了实现该模型的实现方法。

2. 基于内容的图像检索方法

基于内容的图像检索(Content-Based Image Retrieval, CBIR)，属于图像分析的一个研究领域，主要是对图像进行内容语义的分析和特征的提取，并基于这些特征在图像数据库中搜索并查找进行相似匹配的信息检索技术。根据不同的视觉表示方法，可以将基于内容的图像检索方法分为两类：基于 SIFT 特征和基于深度学习的[1]。

2.1. 基于 SIFT 的图像检索方法

受环境干扰较大的影响，对于相同物体的图像检索，我们通常会选择具有较强抗干扰性的不变性局部特征。尺度不变特征转换(Scale-Invariant Feature Transform, SIFT)是一种计算机视觉算法，用于检测和描述图像中的局部特征。该算法在不同的尺度空间寻找关键点，并计算关键点的方向。这些关键点具有显著性，不受光照、仿射变换和噪声等因素的影响，并能提取其位置、尺度和旋转不变性。

基于 SIFT 等局部特征，可以通过不同的编码方式构建图像的全局描述，代表性的方法有词袋模型(Bag of Words, BoW) [2]、局部特征聚合描述符(Vector of Locally Aggregated Descriptors, VLAD) [3]和 Fisher 向量(Fisher Vector, FV) [4]。这类基于 SIFT 的图像检索方法结合了 SIFT 不变性特性，并采用了从局部到全局的特征表达方式。在实际应用中，还可以使用 SIFT GPU 加速 SIFT 特征提取，从而获得较好的检索效果。然而，这类方法通常具有较高的特征维度。

SIFT 能够生成大量特征，为物体识别提供了丰富的图片特征信息。这些特征密集地覆盖了整个图像的尺度和位置。例如，对于一个 500 × 500 像素的图片，大约可以产生约 2000 个稳定的特征。最终，这些特征通过近似最近邻搜索(Approximate Nearest Neighbor Search, ANN)方法进行特征匹配与查找。

2.2. 基于深度学习的图像检索方法

在 2012 年，Krizhevsky 等人使用 AlexNet 在 ILSRVC 2012 上取得了当时世界上最高的识别准确率。从那时起，图像领域的研究重心逐渐转向基于深度学习，特别是卷积神经网络(CNN)的方法。

对于相同类别的图像检索，主要面临的问题是同一类别的图像内部变化巨大，而不同类别的图像之间差异较小。例如，在“湖泊”这一类别的图像中，同类别的图像在表现形式上存在很大的差异。而对于有些图像，尽管它们属于不同的类别，但如果采用低层次的特征进行描述，如颜色、纹理和形状等特征，它们之间的差异非常小，因此直接使用这些特征很难将它们区分开。因此，相同类别图像检索在特征描述上面临着较大的类内变化和较小的类间差异等挑战。

深度学习中的卷积神经网络(Convolutional Neural Network, CNN)在图像检索中能够自动学习图像的高层次特征表示，通过多层神经网络结构逐层提取图像的局部特征和全局信息，将图像映射到一个低维度的特征空间，从而实现对相似图像的有效检索和区分。这种特征表达方式相较于传统手工设计的特征具有更强的表达能力和鲁棒性，大幅提高了图像检索的准确性和效率。但是同时无论是相同物体图像检索还是相同类别图像检索，在使用 CNN 模型提取自动特征时，最终得到的特征维度通常为 4096 维，这仍然是一个相当高的维度。直接使用 PCA (Principal Component Analysis) [5]等降维方法，虽然能达到特征维度降低的目的，但在保持必要的检索精度的前提下，能够降低的维度仍然有限。因此，对于这类图像检索，有必要为其构建高效合理的快速检索机制，使其适应大规模或海量图像的检索需求。

2017 年提出的 Transformer 模型[6]最初是为了解决自然语言处理任务中的长距离依赖问题而设计的。随着 Transformer 在自然语言处理领域取得了巨大成功，研究人员开始尝试将其应用于计算机视觉任务。2020 年，Google Brain 团队提出了 Vision Transformer，证明了 Transformer 在图像分类任务上的有效性。

Vision Transformer (ViT) [7]将自然语言处理领域中的 Transformer 模型应用于图像处理任务。ViT 在图像检索中的应用表现出了很大的潜力，相比传统的 CNN 模型，它具有一些独特的优势。首先，ViT 能够捕捉长距离的依赖关系。在传统的 CNN 模型中，卷积操作主要关注局部信息的提取。尽管通过堆叠多个卷积层可以扩大感受野，但这种方式无法直接捕捉图像中的全局上下文信息。而 ViT 中的自注意力机制(Self-Attention)可以直接计算图像中任意两个位置之间的关系，从而有效捕捉长距离依赖关系。其次，ViT 具有更强的表达能力。由于 ViT 采用了 Transformer 结构，它可以对图像中的每个位置进行并行处理，从而提高计算效率。此外，ViT 通过多头自注意力(Multi-head Self-Attention)机制学习多种不同的上下文关系，这使得模型具有更丰富的表达能力。再者，ViT 具有更好的可扩展性。传统的 CNN 模型通常需要针对不同任务进行特定的设计和调整，而 ViT 的结构更加通用，可以很容易地扩展到不同尺寸的图像和不同领域的任务。这使得 ViT 在图像检索任务中具有更好的适应性。

在图像检索任务中，ViT 可以通过对输入图像进行分块，将每个图像块视为一个类似于文本中的单词。然后将这些图像块输入到 Transformer 模型中，提取高层次的特征表示。最后，通过对比不同图像的特征表示，实现相似图像的检索。

3. 高效向量检索方法

3.1. 最近邻搜索方法

最近邻搜索(Nearest Neighbor Search, NNS)是指在一个确定的距离度量和一个搜索空间内寻找与给定查询项距离最小的元素。这是许多机器学习和数据挖掘任务中的关键组件，如推荐系统、图像识别和模式匹配。最近邻搜索的经典方法如 KD 树[8]和球树[9]对于低维数据效果良好，但在高维数据中，由于所谓的“维度的诅咒”，它们的效果会显著降低。为了克服这些限制，研究者们已经提出了多种近似最

近邻搜索(Approximate Nearest Neighbor Search，简称 ANN)的技术，如局部敏感哈希(Locality-Sensitive Hashing, LSH) [10]层次导航小世界图(Hierarchical Navigable Small World Graphs, HNSW) [11]等，它们旨在在保持高查询精度的同时，提高搜索速度。随着深度学习和大数据的兴起，近似最近邻搜索技术已经成为了处理大规模高维数据的主流方法。

3.2. 基于磁盘的向量检索

DiskANN [12]是一个为大规模数据集设计的近似最近邻搜索库。与传统的内存中的 ANN 搜索方法不同，DiskANN 主要关注于如何高效地在磁盘上执行 ANN 搜索[12]，从而使得对于超大规模数据集的搜索成为可能。这对于那些因为 RAM 限制而无法完全加载到内存中的数据集尤为重要。

传统的最近邻搜索方法，如 LSH 和 HNSW，虽然在搜索速度和准确性上表现出色，但它们主要是为内存中的搜索优化的。当数据集增长到无法容纳在可用的 RAM 内时，这些方法的效率会显著下降。相比之下，DiskANN 针对磁盘的访问模式进行了优化，通过减少随机磁盘访问，从而大大提高了搜索速度[2]。DiskANN 的另一个显著优势是它的可扩展性。它允许用户按需扩展存储和搜索能力，无需进行昂贵的硬件升级。此外，DiskANN 还包括了多种索引和搜索参数的优化工具，使得用户可以根据其特定的应用需求进行定制。

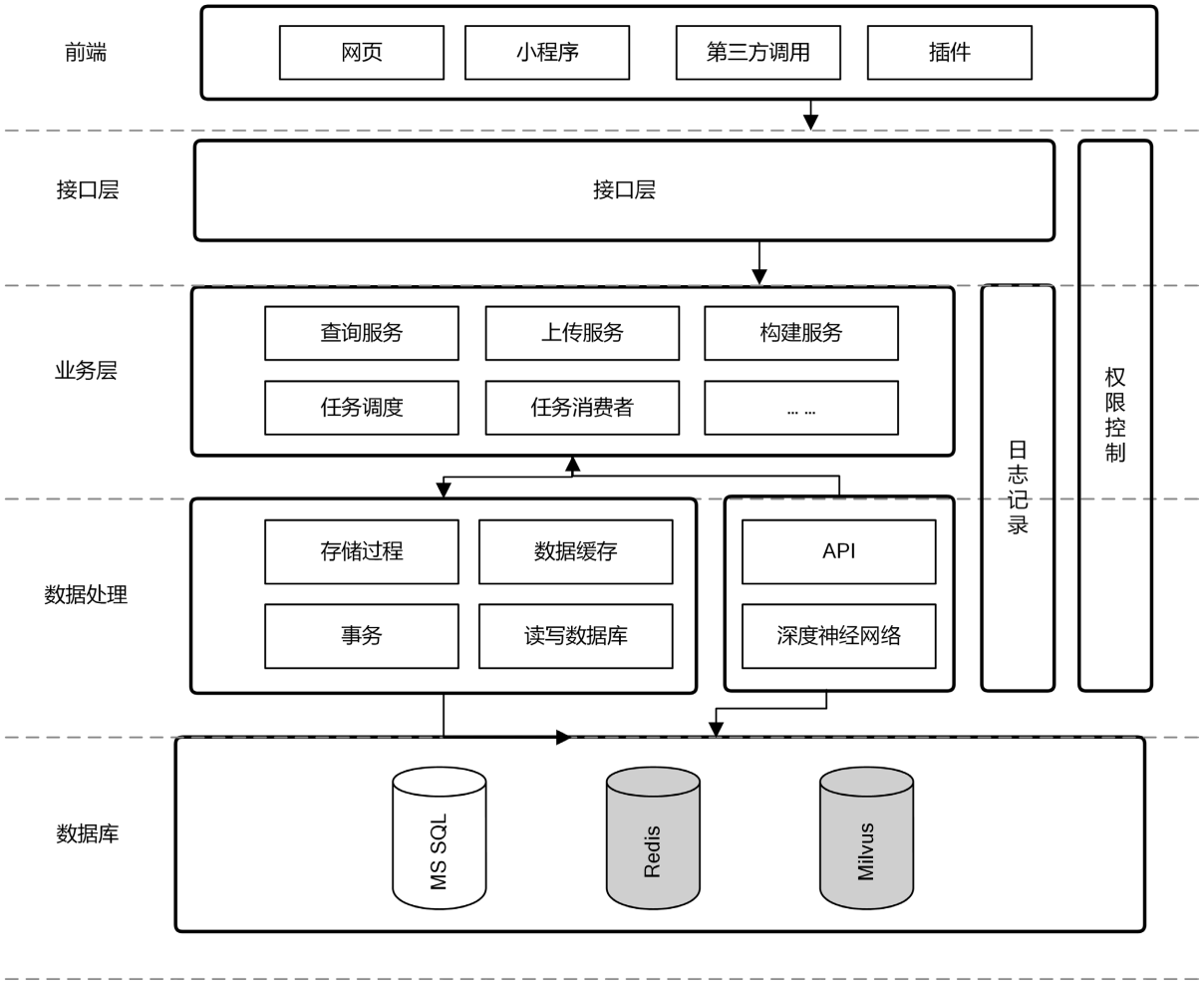

Figure 1. Diagram of image retrieval system architecture
图 1. 图像检索系统架构图

Milvus是一个开源的向量相似性搜索引擎，它支持多种最近邻搜索方法，包括但不限于HNSW、IVF、PQ (Product Quantization)等。DiskANN是一种为大规模数据集设计的近似最近邻搜索库，尤其针对在磁盘上执行搜索进行了优化。为了为用户提供更加高效和灵活的搜索能力，Milvus集成了DiskANN作为其支持的索引类型之一。选择DiskANN作为底层的索引和搜索方法，特别是当处理的数据集太大而不能完全加载到内存中时。通过这种集成，Milvus利用DiskANN提供的磁盘优化能力，能够更高效地处理和搜索大规模的数据集。

4. 图像检索系统模型

在本文中，我们提出了一个基于深度学习的图像检索系统模型，旨在实现高效、准确的图像检索功能。该系统模型整体结构分为三层：接口层、服务层和数据库层。架构图如图 1 所示通过将先进的图像特征提取方法与数据库结合，系统可以确保检索结果的准确性和实时性，同时保持良好的可扩展性和灵活性。

接口层作为系统的入口，负责处理用户请求和返回结果。用户可以通过该层与系统进行交互，实现图像查询和上传等功能。接口层将用户请求传递给服务层，以便进行后续处理。服务层是系统的核心部分，包含查询服务、上传服务和构建服务三个子模块。查询服务负责处理图像查询请求，将输入图片与数据库中的图像进行比对，返回与之最相似的图片的元信息。上传服务允许用户上传图片，以便将其纳入数据库并在日后的查询中使用。构建服务主要用于将给定的图像数据集构建成自定义的查询数据库，以满足不同用户的特定需求。

数据库层由三个数据库组成：Milvus、MySQL 和 Redis。Milvus 负责构建向量查询数据库，存储图像特征向量以供检索使用。MySQL 用于存储与 Milvus 中向量一一对应的图片元信息，以便在检索过程中返回相关信息。Redis 则负责缓存功能，提高系统在处理高频请求时的性能。

在服务层中，我们还引入了一个深度神经网络模型模块，负责提取图像特征。该模块与查询服务和构建服务相连，支持特征提取任务。为了保证系统的灵活性和可扩展性，我们使 Triton Inference Server 来部署神经网络模型，其 Triton Inference Server 模型推理服务化框架因其卓越的性能和灵活性受到了广泛的关注和采用。Triton 框架支持多种机器学习框架，包括但不限于 TensorFlow，PyTorch，ONNX，和 TensorRT，提供了一个统一且高效的推理服务平台。其内部实现了高性能的数据传输和计算优化机制，能够在不同硬件平台上自动实现最优的推理性能。除此之外，Triton 还支持模型版本管理，动态批处理，以及多模型并发执行等高级特性，极大地方便了开发者的使用并提升了服务的可用性。在通信接口方面，Triton 提供了丰富的接口选项，支持 gRPC 和 HTTP/REST 两种主要的通信协议。gRPC 接口提供了高性能的远程过程调用服务，通过 Protocol Buffers 定义服务接口，保证了通信的高效性和数据的一致性；HTTP/REST 接口则提供了一种更为简单和灵活的通信方式，开发者可以通过标准的 HTTP 请求与 Triton 服务进行交互。这两种通信方式都支持多种数据格式，包括 Numpy，TFRecord 等，满足不同应用场景的需求。

我们提出的这个基于深度学习的图像检索系统模型，通过整合先进的特征提取方法和数据库技术，实现了高效、准确的图像检索功能。该系统具有良好的可扩展性和灵活性，可广泛应用于各种图像检索场景。未来，我们将继续探索更多先进的深度学习模型和优化技术，以进一步提高系统性能和适应性。

5. 图像检索系统实现方法

5.1. 图像检索系统架构设计

在当今的计算机视觉领域，图像检索已经成为一项至关重要的研究和应用方向。面对大数据时代的

挑战，构建一个高效、可扩展且用户友好的图像检索系统变得尤为关键。为了应对这一需求，我们提出了一种基于前后端分离的现代化图像检索系统设计。

首先，考虑前端的设计。在当今多元化的互联网环境中，用户使用的设备和平台日益多样化。为了满足从桌面到移动设备的所有用户，我们的前端设计提供了 Web 界面以及专门为移动用户设计的小程序或移动应用。这些界面都有相似的功能，如上传图像、浏览检索结果等。

后端的架构核心的设计思想是模块化和解耦。Go 语言作为后端的主要开发语言，后端主要通过 API 的形式，为前端提供数据和服务，这种方式允许前端与后端之间的通信更为简洁高效。API 层还负责与其他后端组件的交互，如神经网络部署模块和数据库。

在神经网络部署方面，我们选择使用 Triton Inference Server。Triton 的主要优势在于其能够轻松部署和服务化各种深度学习模型，不仅限于图像处理。当用户上传图像进行检索时，API 层会调用 Triton 来进行图像特征提取，这些特征随后用于搜索和匹配。

为了高效地存储和检索图像特征，我们引入了 Milvus 作为向量搜索引擎。与传统的关系型数据库相比，Milvus 在大规模向量数据的存储和检索方面展现出了卓越的性能。此外，我们还使用 MySQL 来存储与图像相关的元数据，例如上传时间、图像来源等。而为了进一步优化性能，我们引入了 Redis 作为缓存系统，用于缓存高频请求的数据，从而实现更快的响应速度。

这样一个图片的离线处理的过程主要是对图像抽取特征/信息构建索引，如图 2 所示。在流程开始时，首先对数据库中存储的所有图像进行预处理操作，包括但不限于图像压缩、灰度化和去噪等，目的是提

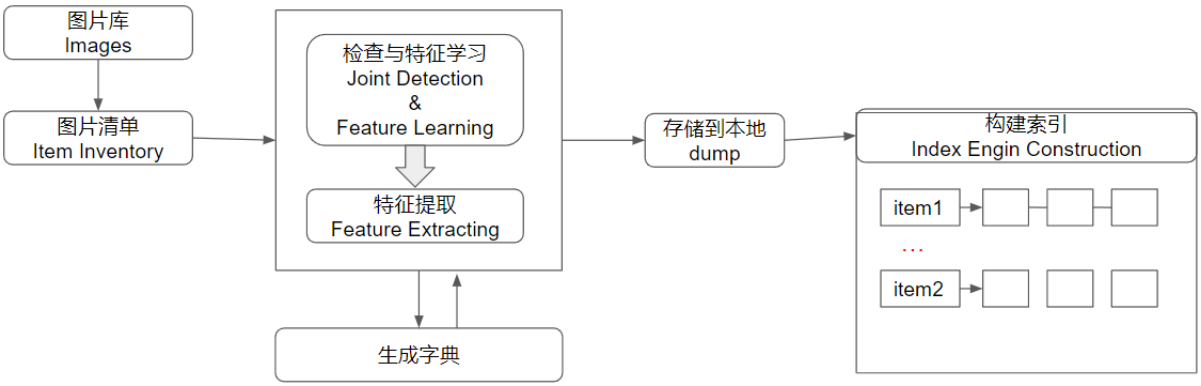

Figure 2. Image retrieval offline process
图 2. 图像检索离线流程

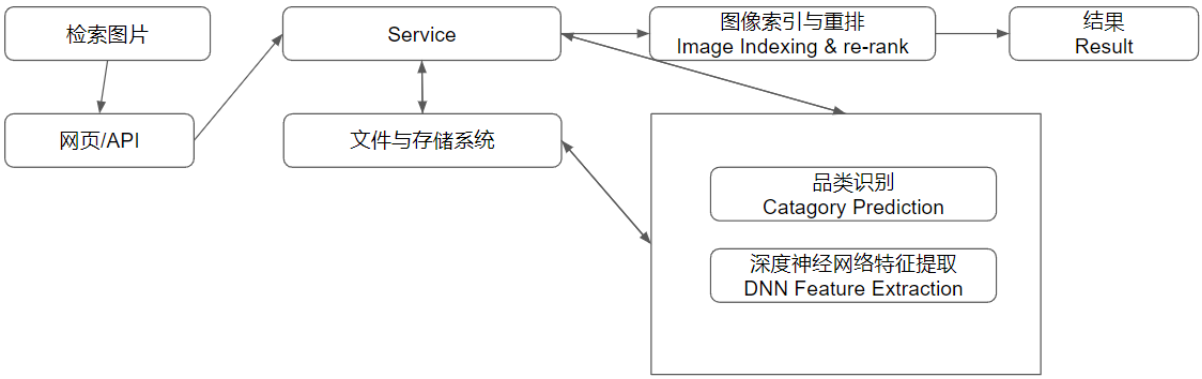

Figure 3. Image retrieval online process
图 3. 图像检索在线流程

高后续处理的效率并减小噪声干扰。接着，系统将对每张经过预处理的图像提取其视觉特征，这些特征通常包括颜色、纹理、形状等信息，并将其转换为数学表示以便于比较。随后，这些特征数据会被索引并存储到特征数据库中，比如 Milvus 等。

图像检索的在线流程从“应用终端”开始，用户通过应用或 API 提交查询。查询请求首先传递给服务模块，然后该模块将其转发到图像索引与重排序模块。其中，图像特征通过深度神经网络提取进行分析，并进一步用于类别预测。基于这些特征和预测结果，图像索引与重排序模块进一步处理并返回相应的结果给用户。

综上所述，我们提出的图像检索系统设计致力于实现高效、可扩展和用户友好的目标。通过前后端分离的架构，系统能够灵活应对不同的用户需求和设备。而后端的模块化设计确保了各个组件之间的低耦合性，同时提供了高度的可扩展性。

5.2. 特征提取深度神经网络算法设计

图像的内容特征提取是关键环节，该系统采用两种目前比较广泛的深度学习模型来进行特征提取，分别是基于卷积神经网络 CNN 的模型和基于 Transformer 的 Vison Transformer 模型。

5.2.1. 基于 CNN 的算法设计

随着深度学习技术的发展，卷积神经网络(Convolutional Neural Networks, CNN)已经在图像识别和特征提取领域表现出色。特别是深度残差网络(Deep Residual Networks)，为训练更深的网络模型提供了可能性。其中，ResNet-50 [13]作为案例进行表述。

ResNet-50 的核心思想是在每个模块中引入“残差连接”。这一设计目的是为了解决深度网络中的梯度消失和梯度爆炸问题。假设我们有一个浅层模型的输出 $h(x)$，理想化的映射应该为 $F(x)$，那么整体映射 $H(x)$ 就为 $h(x)+F(x)$。这里的 $F(x)$ 就是残差函数。

在数学表示上，我们可以描述为：

$$H(x)=h(x)+F(x) \tag{1}$$

$$F(x)=H(x)-h(x) \tag{2}$$

其中 $F(x)$ 表示残差映射，通常是由几个卷积层构成的。在实际实施时，$h(x)$ 可以是输入 x 本身，或是一些经过卷积、批量标准化(Batch Normalization)和 ReLU 激活函数处理后的版本。

ResNet-50 的架构包括五个阶段，每个阶段都有一个或多个残差块。每个残差块内部含有三个卷积层，分别具有 1 × 1、3 × 3 和 1 × 1 的卷积核大小。这种“瓶颈”设计的目的是减少计算复杂性。

为了特征提取，我们采用预训练的 ResNet-50 模型并移除其全连接层，从而获得全局平均池化层的输出。这样，对于每个输入图像，我们都可以得到一个维度为 2048 的特征向量。这些向量可以进一步用于下游任务的如图像检索、分类和聚类等。

Triplet Loss 三元组损失函数

在图像检索的应用场景下，当提供一张待查询图像时，中心任务是通过图像特征来精确地与图像数据库中的图片进行匹配。结合三元组的策略，目标是缩短查询图像与其在数据库中的相似图像之间的距离，同时增大查询图像与不同图像之间的距离。为实现此目标，我们采用了 Triplet Loss 函数，数学表达为：

$$\text{loss}\left(q,q^{+},q^{-}\right)=\left[L2\left(f(q),f\left(q^{+}\right)\right)-L2\left(f(q),f\left(q^{-}\right)\right)+\delta\right]_{+} \tag{3}$$

其中，$L2$ 代表两向量间的 $L2$ 范数，δ 是一个正的边界参数，f 是一个待学习的 CNN 特征提取函数，能

够通过端对端的训练方法进行学习。

为了进一步优化和减少由于训练图像内的噪声所带来的影响，我们对 Triplet Loss 进行了改进。新的损失函数为：

$$\begin{aligned}&\text{loss}=\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|N_q|}\sum_{q^-\in N_q}\left[L2\left(f(q),f\left(q^+\right)\right)-L2\left(f(q),f\left(q^-\right)\right)+\delta\right]_+\\&Q=\left\{q\mid\exists q^-,L2\left(f(q),f\left(q^+\right)\right)-L2\left(f(q),f\left(q^-\right)\right)+\delta>0\right\}\\&N_q=\left\{q^-\mid L2\left(f(q),f\left(q^+\right)\right)-L2\left(f(q),f\left(q^-\right)\right)+\delta>0\right\}\end{aligned}\tag{4}$$

该策略的核心是对同一查询图像的所有三元组样本计算平均损失，进而最大化地削弱由噪声三元组引起的影响。通过这一改进的三元组损失函数，我们能够更有效地学习 CNN 特征，并确保用户的查询图像与数据库中的图像在同一特征空间内被准确地匹配，实现了跨源图像的可靠匹配。

5.2.2. 基于 Transformer 的算法设计

基于 Vision Transformer(ViT)的模型的主要构建块。输入图像首先被分解为 M 个固定大小的块(例如 16×16)。每个 patch 被线性投影到 M 个向量形状的 token 中，并以排列不变的方式用作 Transformer 的输入。通过向输入 token 添加可学习的一维位置编码向量来合并位置先验。一个额外的可学习 CLS token 被添加到输入序列中，使其对应的输出 token 用作全局图像表示。Transformer 由 L 层组成，每一层由两个主要块组成：一个多头自注意(MSA)层，它将自注意操作应用于输入 token 的不同投影，以及一个前馈网络(FFN)。MSA 和 FFN 层之前都是层归一化，然后是 skip 连接。

从 ViT 主干中提取现成的特征，在 ImageNet 上进行预训练：首先，考虑一种简化方法 IRTO，该方法通过直接从在 ImageNet 上预训练的 Transformer 中抽取图像特征。这种策略与早期使用卷积网络进行图像检索的方法类似，它们都基于激活特征，构建了一个全局性的紧凑图像描述符。在 ViT 结构中，预分类层给出 $M+1$ 个向量，其中对应于 M 个图像块和一个特殊的类(CLS)嵌入。采用 CLS 池化方法，延续 BERT 和 ViT 模型的设计思路，将这个类嵌入作为图像的全局描述符。除此之外，还有其他几种全局池化策略，这些策略在卷积度量学习模型中经常被使用，如平均池化、最大池化和广义平均(GeM)池化，并将其应用到 M 个输出令牌上。最终描述符向量在池化后被归一化到单位球中。如果目标维度低于模型原始的维度，可以选择通过主成分分析(PCA)进一步减少向量的维度后再进行归一化处理。

使用度量学习微调 Transformer，特别是对比损失：还可以考虑一种用于图像检索的度量学习方法[14]，它是类别级别和特定对象检索的主要方法。将它与 Transformer 而不是卷积神经网络结合起来。采用带有跨 batch 记忆的对比损失，并为度量学习目标默认固定边距 $\beta=0.5$。对比损失最大化具有相同标签 y(或任何其他预定义相似性规则)的样本的编码低维表示 z_i 之间的相似性。同时，它最小化了具有不匹配标签的样本表示之间的相似性，这些标签被称为负样本。对于对比损失，只有相似度高于恒定边距 β 的负对对损失有贡献。这可以防止训练信号被简单的负值淹没。形式上，一个 batch 大小为 N 的对比损失定义为：

$$\mathcal{L}_{\text{contr.}}=\frac{1}{N}\sum_{i}^{N}\left[\sum_{j:y_i=y_j}\left[1-z_i^{\mathrm{T}}z_j\right]+\sum_{j:y_i\neq y_j}\left[z_i^{\mathrm{T}}z_j-\beta\right]_+\right]\tag{5}$$

表示 z_i 被假定为 $L2$ 归一化，因此内积等价于余弦相似度。

额外正则化输出特征空间以鼓励一致性：最近，一些工作研究了一组成对损失之间的联系以及学习表示 $Z=z_i$ 和相应的真实标签 $Y=\{y_i\}$ 之间的互信息最大化。本文对对比损失的特殊情况感兴趣。互信息

定义为：

$$\mathcal{I}(Z,Y)=\mathcal{H}(Z)-\mathcal{H}(Z\mid Y) \tag{6}$$

对比损失的正项导致条件微分熵 $\mathcal{H}(Z\mid Y)$ 的最小化，其中直观地，属于同一类别的样本表示被训练为更相似：

$$\mathcal{H}(Z\mid Y)\propto\frac{1}{N}\sum_{i}^{N}\sum_{j:y_i=y_j}\left[1-z_i^{\mathrm{T}}z_j\right] \tag{7}$$

另一方面，这种损失的负项负责防止所有样本表示都折叠到一个点的琐碎解决方案。因此，它最大化了学习表示的熵：

$$\mathcal{H}(Z\mid Y)\propto-\frac{1}{N}\sum_{i}^{N}\sum_{j:y_i=y_j}\left[z_i^{\mathrm{T}}z_j-\beta\right]_{+} \tag{8}$$

边际 β 在训练动态中起着重要作用。β 的低值允许探索更多的负样本。然而在这种情况下，简单的负数会主导训练并导致性能停滞不前。相反，较高的 β 值只会接受难负样本，可能会导致噪声梯度和不稳定的训练。添加了一个熵最大化项，该项与边缘接受的负样本无关，基于 Kozachenko & Leonenko 微分熵估计器：

$$\mathcal{L}_{\mathrm{KoLeo}}=-\frac{1}{N}\sum_{i}^{N}\log\left(\rho_i\right) \tag{9}$$

其中 $\rho_i=\min_{i\neq j}\left\|z_i-z_j\right\|$。换句话说，这种正则化最大化了每个点与其最近邻点之间的距离，从而缓解了崩溃问题。只需将正则化项添加到由正则化强度系数 λ 加权的对比损失中：$\mathcal{L}=\mathcal{L}_{\mathrm{contr}}+\lambda\mathcal{L}_{\mathrm{KoLeo}}$。

5.3. 图像特征提取模型部署

Triton 是C++语言开发的服务框架，通过 Backend 的方式集成各类推理框架，Backend 以动态 lib 库方式，集成到 Triton 框架中。Triton 抽象基础能力接口，暴露 Backend 的 C API 接口，各个 Backend 实现具体处理逻辑，进而达到框架与 Backend 库代码编译解耦，通过动态 lib 库方式，动态装载各类 Backend。模型管理本质是模型安装、卸载、执行。Triton Backend 设计抽象了 TRITONBACKEND_ModelInitialize/ModeFinalize 接口，规范了模型安装/卸载实现。Triton 框架代码对外 TRITONSERVER_ServerLoadModel C-API，支持模型 load/unload/update 管理能力。方便 Triton 的核心能力集成到不同业务。

Triton 架构允许多个模型和/或同一模型的多个实例在同一系统上并行执行，通过调度支持资源隔离，避免相互相互抢占，有利于提高资源利用率，并保证高优模型性能。Triton 支持多种调度和批处理算法，可以为每个模型独立选择，极大提高调度灵活性。对于无状态模型，如 CV 领域的图片分类、对象识别，支持动态批次处理，通过组合推理请求，从而动态创建批处理，创建一批请求通常会导致吞吐量，提高服务资源利用率。对有状态模型，需要维护推理请求之间的状态，希望多个推理请求一起形成一系列推理，这些推理必须路由到同一个模型实例，以便正确更新模型维护的状态；如 NLP 领域，支持通过关联 ID，对可端的请求进行串联，并送给同一个模型实例。

5.4. 系统评价指标

图像检索系统主要的评价标准通常包括准确性、召回率、平均精度(MAP)和查准查全曲线(Precision-Recall Curve)。准确性是指在检索到的前 k 个结果中，与查询图像相关的图像数量。召回率度量了与查询图像相关的图像中有多少被检索到。MAP 是一个常用的全局测量标准，它计算了不同召回率下精度的平均值，为我们提供了系统在整体上的性能概览。查准查全曲线则描绘了随着召回率的变化，

精确度是如何变化的，它有助于深入了解检索系统在不同召回率阈值下的性能。此外，为了进一步深入了解系统的性能，一些研究者还可能考虑使用其他评价指标，如归一化折扣累积增益(NDCG)或平均查准率。总的来说，选择合适的评价标准是确保 CBIR 系统有效性的关键，它为研究者和开发者提供了一个量化的方法来优化和改进检索效果。

$$\text{准确性 } P(Precision) = \frac{\text{检索到的相关图像数量}}{\text{检索到的总图像数量}} \tag{10}$$

$$\text{召回率 } R(Recall) = \frac{\text{检索到的相关图像数量}}{\text{数据库中所有相关图像数量}} \tag{11}$$

为检验系统的检索性能，随机抽取图片对向量数据库进行在单机配置：CPU：5600x RAM：32GB GPU：3080ti SSD：1 T 的条件下，图片最终的特征为 2048 维度的向量，在 100 K 向量数据库中以对候选集 L = 10，20，30，40，50，100 时进行检索，结果如表 1，recall@10 能达到 90%以上，检索的效率也在可接受的范围，可以看出该系统已经具有良好的检索性能。

Table 1. Retrieval accuracy and recall rate under different candidate sets
表 1. 不同候选集下检索准确率与召回率

候选集大小	每秒查询率	平均延时	平均 IO	准确率	召回率
10	3576.24	2271.92	23.80	92.03	91.79%
20	3119.61	4121.04	33.26	96.39	96.42%
30	2546.84	6147.14	42.64	98.82	98.78%
40	2148.29	8278.83	52.16	99.03	99.40%
50	1756.05	9913.28	61.76	99.43	99.63%
100	978.49	19107.81	110.61	99.92	99.91%

随着候选集的大小选择的更大，由于要进行更多的磁盘 IO，导致每秒查询率下降，但是可以看出在较小的候选集就可以达到很高的精度，这方面的参数还可以根据具体情况进行调整。

参考文献

[1] Zheng, L., Yang, Y. and Tian, Q. (2017) SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 1224-1244. https://doi.org/10.1109/TPAMI.2017.2709749

[2] Csurka, G., *et al*. (2004) Visual Categorization with Bags of Keypoints. *Workshop on Statistical Learning in Computer Vision*, *ECCV*, **1**. 1-22.

[3] Hervé, J., *et al*. (2010) Aggregating Local Descriptors into a Compact Image Representation. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 13-18 June 2010. https://doi.org/10.1109/CVPR.2010.5540039

[4] Perronnin, F., Sánchez, J. and Mensink, T. (2010) Improving the Fisher Kernel for Large-Scale Image Classification. In: Daniilidis, K., Maragos, P. and Paragios, N., eds., *Computer Vision—ECCV 2010*, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15561-1_11

[5] Urdinez, F. and Cruz, A. (2021) R for Political Data Science: A Practical Guide. Chapman and Hall/CRC, Boca Raton, FL, 375-393.

[6] Vaswani, A., *et al*. (2017) Attention Is All You Need. https://arxiv.org/abs/1706.03762

[7] Dosovitskiy, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929

[8] Bentley, J.L. (1975) Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM*, **18**, 509-517. https://doi.org/10.1145/361002.361007

[9] Omohundro, S.M. (1989) Five Balltree Construction Algorithms. International Computer Science Institute, Berkeley, 1-22.

[10] Gionis, A., Indyk, P. and Motwani, R. (1999) Similarity Search in High Dimensions via Hashing. *Very Large Data Bases Conference* (*VLDB*), **99**, 518-529.

[11] Malkov, Y.A. and Yashunin, D.A. (2018) Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 824-836. https://doi.org/10.1109/TPAMI.2018.2889473

[12] Subramanya, S.J., *et al*. (2019) DiskANN: Fast Accurate Billion-Point Nearest Neighbor Search on a Single Node. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, Canada, 31 August 2020.

[13] He, K.M., *et al*. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 27-30 June 2016. https://doi.org/10.1109/CVPR.2016.90

[14] El-Nouby, A., *et al*. (2021) Training Vision Transformers for Image Retrieval. https://arxiv.org/abs/2102.05644