深圳大学考试答题纸

(以论文、报告等形式考核专用)

二〇 二四 ～二〇 二五 学年度第 二 学期

课程编号 1501990013 课序号 02 课程名称 数据挖掘 主讲教师 王祎乐 评分

学号 20221550 33 姓名 洪子敬 专业年级 2022 级软件工程（腾班）

教师评语：

1. 可以按照 PPT 展示内容稍作扩展整理完成；
2. 同一小组成员，题目可以一致，内容要有所区别（特别是总结思考项目启发部分），不要照抄他人；
3. **每位同学**都要提交留档，期末前完成提交；
4. 严格按照格式补充或删改高亮部分，每节的标题或顺序可适当改动。无具体字数页面要求，PPT 展示和该报告会一起计入成绩。

题目： **基于 CLIP 模型的图像分类和文搜图**

一、项目介绍

本项目采用 CLIP（Contrastive Language-Image Pre-training）模型，对该模型进行本地化部署，采用 ImageNet2012 的验证集，经过数据预处理得到目标输入格式后，分别对 ViT-B/32 和 RN50_X64 两个预训练模型在租借的服务器上进行测试，得到了相应的 TOP1 和 TOP5 测试精度及相应的平均耗时；此外，还对该模型进行了应用拓展，建立了图片的特征向量库，实现了图像的分类功能和文搜图功能，并做了可视化演示。

二、模型介绍

1. 模型背景

CLIP 是由 OpenAI 于 2021 年提出的跨模态预训练模型，主要用于连接自然语言文本和图像数据，通过对比学习实现图文语义对齐。其核心思想是：利用互联网海量图文对数据，训练模型使语义相关的文本和图像在特征空间中距离相近，反之则远离。这种能力使其在零样本迁移学习（即无需微调即可处理新任务）中表现出色，颠覆了传统计算机视觉模型的训练范式。

2．模型架构

CLIP 模型架构主要由文本编码器、图像编码器和对比学习损失函数三部分组成。

（1）文本编码器：通常采用 Transformer 架构（如 BERT 的变体），将输入文本转换为固定长度的语义特征向量（这里是 512 维）；

输入处理：文本先经过分词生成 token 序列，再添加位置编码后输入文本编码器；

（2）图像编码器：支持多种视觉主干网络，包括：

a. ResNet 系列：根据层数不同可分 RN50/101，根据训练的数据集不同的可分 openai、yfcc15m 和 cc12m，根据特征图的宽度不同可分初始、x4、x16 和 x64；

b. Vision Transformer（ViT）系列：根据规模不同可分为 Base、Small 和 Large，根据图像分块大小可分 16*16、32*32，根据训练数据集不同可分 openai、laionxxx 等；

（3） 对比学习损失函数

目标：使同一图文对的文本特征和图像特征在特征空间中尽可能接近，不同对的特征尽可能远离。

损失函数如下：

$$L_{i,j} = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{N}\exp(s_{i,k}/\tau)+\sum_{k=1}^{N}\exp(s_{k,j}/\tau)-\exp(s_{i,j}/\tau)}$$

- $L_{i,j}$：表示样本对(i,j)的损失值，具体是图像 i 和文本 j 构成匹配对的对比损失。
- $S_{i,j}$：样本 i 和样本 j 之间的相似度得分，具体是一般是图像 i 和文本 j 经过模型编码后，二者特征的相似度得分（比如点积 similarity），反映图像和文本语义匹配程度，得分越高越可能匹配。
- τ：温度系数（temperature parameter），是一个超参数，用于调整相似度得分的分布。它控制了 softmax 分布的 “尖锐程度”，τ值越小，分布越尖锐，模型对正确匹配的区分度要求越高。
- N：样本的数量，具体是参与对比的图像-文本对的数量，相当于 batch_size。

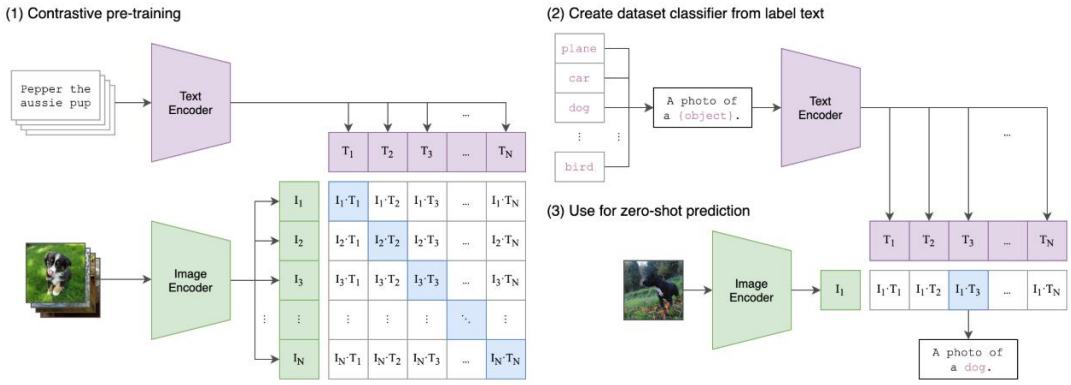


图 1. CLIP 整体架构

3. 模型的应用场景

（1）零样本图像分类

原理：将图像分类任务转化为文本匹配问题。例如，对 “识别狗” 的任务，构造文本

模板（如 “一张 {类名} 的照片”），生成所有类别对应的文本特征，再计算图像特征与这些文本特征的相似度，相似度最高的类别即为预测结果。

优势：无需标注图像数据，只需文本标签即可完成分类，尤其适合长尾类别或新兴概念

（2）图文检索与生成

a. 图像检索文本：输入图像，返回语义相关的文本描述。

b. 文本检索图像：输入文本查询，返回匹配的图像（如电商平台的文本搜图功能）。

c. 图文生成：结合扩散模型（如 DALL-E），根据文本描述生成图像（CLIP 用于评估生成图像与文本的语义一致性）。

（3）多模态下游任务迁移

通过微调 CLIP 的图文编码器，可适配至各类任务，如：

a. 目标检测：使用文本提示定位图像中的物体（如“找出图中的汽车”）。

b. 图像字幕生成：生成图像的自然语言描述。

c. 视频理解：结合时间序列建模（如 CLIP+Transformer），处理视频 - 文本对齐任务。

三、数据来源

本项目采了用 ImageNet2012 的验证集进行模型分类性能的测试，整个验证集大约 5 万张图像，图像存放格式为 JPEG，图像并没有直接按照类别分文件夹进行存放，而是存放在 ILSVRC2012_validation_ground_truth.txt 标注文件中，记录了每张图片对应的类别标签。从数据中截取的 demo 如图 2 所示：

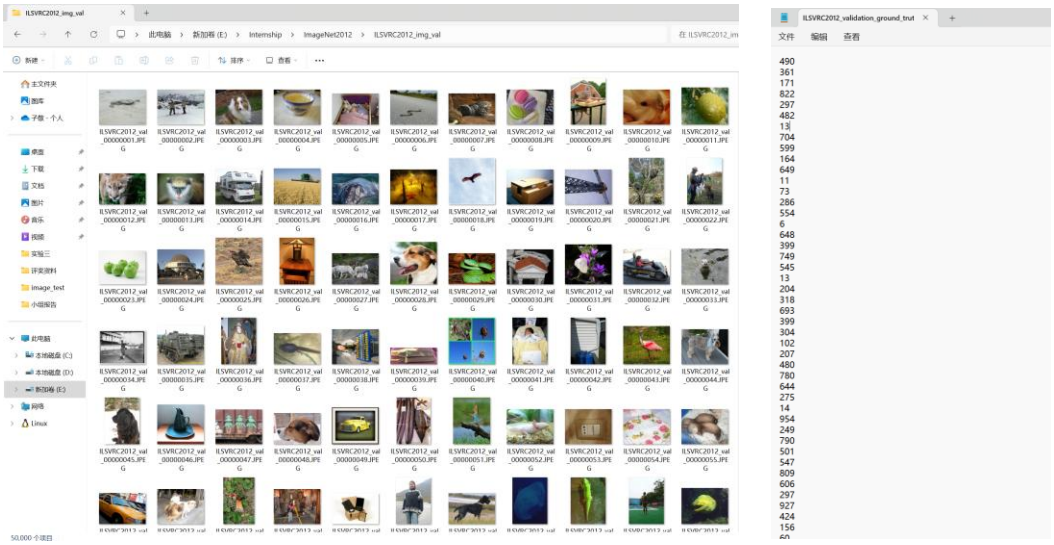
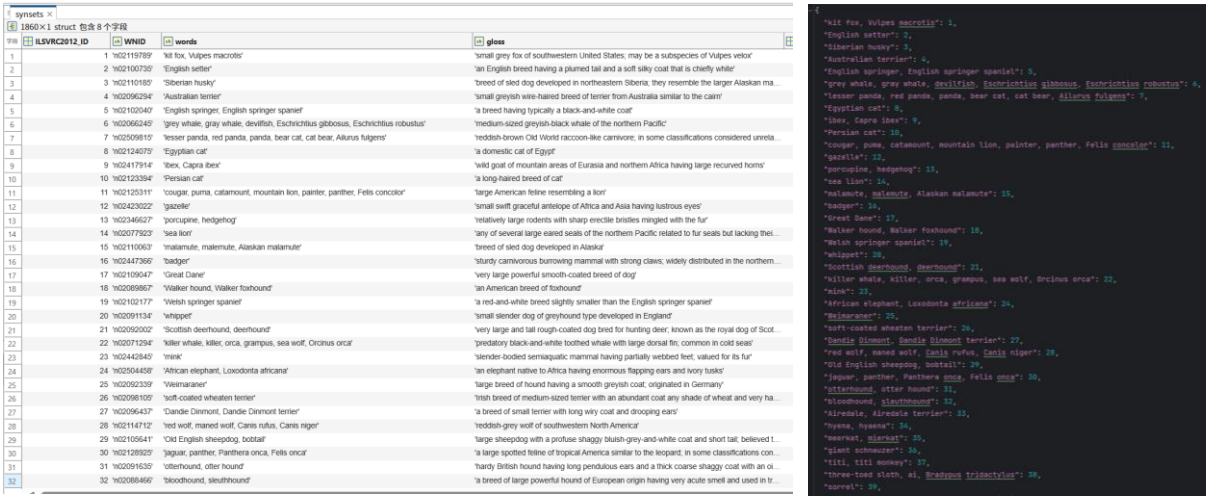


图 2. ImageNet2012 验证集 demo（左） groundtruth 标签 demo（右）

四、数据预处理

由于 CLIP 模型需要建立图像和对应文本之间的关联，因此需要找到 ImageNet2012 验证集图像对应的文本标签，而 meta.mat 存放着验证集中每个类的类别信息，包括唯一标识符、类别（words）和详细描述等等，为了方便后续计算 ACC 的方便，这里使用 pandas 库编写脚本分离 meta.mat 中的文本描述和对应的标签信息形成字典格式，并保留成 json 文件便于后续使用。从 meta 文件和处理后的 json 文件截取的 demo 如图 3 所示：



	ILSVRC2012_ID	WNID	words	gloss
1	1	'n02119789'	'kit fox, Vulpes macrotis'	'small grey fox of southwestern United States; may be a subspecies of Vulpes velox'
2	2	'n02100735'	'English setter'	'an English breed having a plumed tail and a soft silky coat that is chiefly white'
3	3	'n02110185'	'Siberian husky'	'breed of sled dog developed in northeastern Siberia; they resemble the larger Alaskan ma...
4	4	'n02096294'	'Australian terrier'	'small greyish wire-haired breed of terrier from Australia similar to the cairn'
5	5	'n02102040'	'English springer, English springer spaniel'	'a breed having typically a black-and-white coat'
6	6	'n02066245'	'grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus'	'medium-sized greyish-black whale of the northern Pacific'
7	7	'n02509815'	'lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens'	'reddish-brown Old World raccoon-like carnivore; in some classifications considered unrela...
8	8	'n02124075'	'Egyptian cat'	'a domestic cat of Egypt'
9	9	'n02417914'	'ibex, Capra ibex'	'wild goat of mountain areas of Eurasia and northern Africa having large recurved horns'
10	10	'n02123394'	'Persian cat'	'a long-haired breed of cat'
11	11	'n02125311'	'cougar, puma, catamount, mountain lion, painter, panther, Felis concolor'	'large American feline resembling a lion'
12	12	'n02423022'	'gazelle'	'small swift graceful antelope of Africa and Asia having lustrous eyes'
13	13	'n02346627'	'porcupine, hedgehog'	'relatively large rodents with sharp erectile bristles mingled with the fur'
14	14	'n02077923'	'sea lion'	'any of several large eared seals of the northern Pacific related to fur seals but lacking thei...
15	15	'n02110063'	'malamute, malemute, Alaskan malamute'	'breed of sled dog developed in Alaska'
16	16	'n02447366'	'badger'	'sturdy carnivorous burrowing mammal with strong claws; widely distributed in the northern...
17	17	'n02109047'	'Great Dane'	'very large powerful smooth-coated breed of dog'
18	18	'n02089867'	'Walker hound, Walker foxhound'	'an American breed of foxhound'
19	19	'n02102177'	'Welsh springer spaniel'	'a red-and-white breed slightly smaller than the English springer spaniel'
20	20	'n02091134'	'whippet'	'small slender dog of greyhound type developed in England'
21	21	'n02092002'	'Scottish deerhound, deerhound'	'very large and tall rough-coated dog bred for hunting deer, known as the royal dog of Scot...
22	22	'n02071294'	'killer whale, killer, orca, grampus, sea wolf, Orcinus orca'	'predatory black-and-white toothed whale with large dorsal fin; common in cold seas'
23	23	'n02442845'	'mink'	'slender-bodied semiaquatic mammal having partially webbed feet; valued for its fur'
24	24	'n02504458'	'African elephant, Loxodonta africana'	'an elephant native to Africa having enormous flapping ears and ivory tusks'
25	25	'n02092339'	'Weimaraner'	'large breed of hound having a smooth greyish coat; originated in Germany'
26	26	'n02098105'	'soft-coated wheaten terrier'	'Irish breed of medium-sized terrier with an abundant coat any shade of wheat and very ha...
27	27	'n02096437'	'Dandie Dinmont, Dandie Dinmont terrier'	'a breed of small terrier with long wiry coat and drooping ears'
28	28	'n02114712'	'red wolf, maned wolf, Canis rufus, Canis niger'	'reddish-grey wolf of southwestern North America'
29	29	'n02105641'	'Old English sheepdog, bobtail'	'large sheepdog with a profuse shaggy bluish-grey-and-white coat and short tail; believed t...
30	30	'n02128925'	'jaguar, panther, Panthera onca, Felis onca'	'a large spotted feline of tropical America similar to the leopard; in some classifications con...
31	31	'n02091635'	'otterhound, otter hound'	'hardy British hound having long pendulous ears and a thick coarse shaggy coat with an oi...
32	32	'n02088466'	'bloodhound, sleuthhound'	'a breed of large powerful hound of European origin having very acute smell and used in tr...


```
{
  "kit fox, Vulpes macrotis": 1,
  "English setter": 2,
  "Siberian husky": 3,
  "Australian terrier": 4,
  "English springer, English springer spaniel": 5,
  "grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus": 6,
  "lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens": 7,
  "Egyptian cat": 8,
  "ibex, Capra ibex": 9,
  "Persian cat": 10,
  "cougar, puma, catamount, mountain lion, painter, panther, Felis concolor": 11,
  "gazelle": 12,
  "porcupine, hedgehog": 13,
  "sea lion": 14,
  "malamute, malemute, Alaskan malamute": 15,
  "badger": 16,
  "Great Dane": 17,
  "Walker hound, Walker foxhound": 18,
  "Welsh springer spaniel": 19,
  "whippet": 20,
  "Scottish deerhound, deerhound": 21,
  "killer whale, killer, orca, grampus, sea wolf, Orcinus orca": 22,
  "mink": 23,
  "African elephant, Loxodonta africana": 24,
  "Weimaraner": 25,
  "soft-coated wheaten terrier": 26,
  "Dandie Dinmont, Dandie Dinmont terrier": 27,
  "red wolf, maned wolf, Canis rufus, Canis niger": 28,
  "Old English sheepdog, bobtail": 29,
  "jaguar, panther, Panthera onca, Felis onca": 30,
  "otterhound, otter hound": 31,
  "bloodhound, sleuthhound": 32,
  "Airedale, Airedale terrier": 33,
  "hyena, hyaena": 34,
  "meerkat, mierkat": 35,
  "giant schnauzer": 36,
  "titi, titi monkey": 37,
  "three-toed sloth, ai, Bradypus tridactylus": 38,
  "squirrel": 39,
```


图 3. meta 文件 demo（左） json 文件 demo（右）

五、模型测试

在租借的 RTX3090 24G 服务器上对 ViT-B/32 和 RN50_X64 两个模型分别在 ImageNet2012 验证集上进行测试，使用 TOP1 和 TOP5 两个指标进行衡量，通过 python 自带的 LOG 库保存在日志中。运行结果分别如图 4、图 5 所示：

```
2025-05-15 15:24:59,118 - Image: ILSVRC2012_val_00050000.JPEG, Most likely class: llama
2025-05-15 15:24:59,118 - Time used: 0.19508934020996094
2025-05-15 15:24:59,121 - TOP 1 Correct num: 27857
2025-05-15 15:24:59,121 - TOP 5 Correct num: 41360
2025-05-15 15:24:59,121 - TOP 1 Accuracy: 0.55714
2025-05-15 15:24:59,121 - TOP 5 Accuracy: 0.8272
2025-05-15 15:24:59,121 - Averaged Time: 0.20045136761665344
```

图 4. ViT-B/32 模型测试结果

```
2025-05-16 02:31:53,026 - Image: ILSVRC2012_val_00050000.JPEG, Most likely class: llama
2025-05-16 02:31:53,026 - Time used: 0.5629673004150391
2025-05-16 02:31:53,029 - TOP 1 Correct num: 32999
2025-05-16 02:31:53,029 - TOP 5 Correct num: 44248
2025-05-16 02:31:53,029 - TOP 1 Accuracy: 0.65998
2025-05-16 02:31:53,029 - TOP 5 Accuracy: 0.88496
2025-05-16 02:31:53,029 - Averaged Time: 0.5717206792593003
```

图 5. RN50_x64 模型测试结果

从结果不难看出 RN50_x64 的 TOP1 精度达到了 66%左右，TOP5 精度达到了 88%左右，ViT-B/32 的 TOP1 精度达到了约 55%，TOP5 精度达到了约 82%，前者的精度明

显是优于后者的，且高了 6%-11%；但前者的平均耗时是 0.57s，后者是 0.2s，前者精度虽然提高了，但是速度慢了 2.5 倍左右。

针对上述结果，其实是可解释的，因为 RN50x64 的参数量远大于 ViT-B-32，计算复杂度更高，精度更高，推理速度也就更慢，适合高精度研究工作；ViT-B-32 参数量较小，精度低一些，但效率更高，适合实时应用。

六、模型应用

1. 整体思想：

针对 CLIP 文本与图像关联的检测原理，本项目结合实际应用设计了此模型的两个拓展应用功能：

（1） 图像分类功能：给定待分类图片和类别区间，输出给定类别中关联概率最大类别；

（2） 文搜图功能：给定文本和图片库，输出与给定文本关联概率最大的图片；

不过需要注意的是每次检索时，若图像为新的，则需要保存图像的特征向量到向量知识库 features_library.pth 中，并根据索引建立特征向量和图片的映射，便于文搜图显示相应的图像；也就是说对于一个新的图片，需要将其保存到一个图片库中，并将模型输出的特征向量保存到向量库中，且两者需要建立联系，本项目是通过索引标号实现统一的；通过这个图片和向量库即可实现图片分类和文搜图的可视化效果。

2. 实际效果：

这里采用从视觉中国网站上下载的 10 张动物图片做一个 demo，详细如图 6 所示：

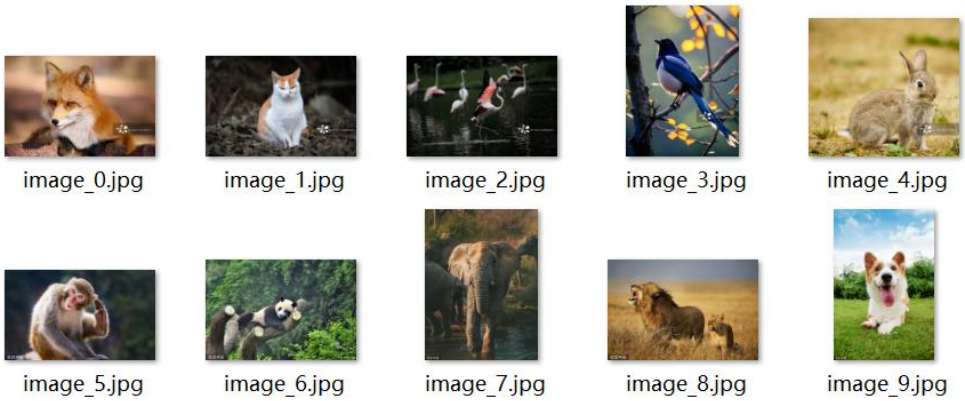


图 6. 测试图片

使用 python 的前端可视化库 gradio 编写可视化代码，整体分为上传图像进行分类和输入文字进行在当前图片库进行图像检索的功能，并提供了当前图像库的可视化效果；不过值得注意的是当前文搜图的功能只支持英文的输入，因为 CLIP 当前开源的模型大

多都是英文训练的，中文目前还不太支持，因此本项目的应用仅支持英文输入。

实现细节：对于输入的图片，经过 transform 处理后，使用 clip 模型提取图像特征，在计算完图像文本相似度进行后，从图片库中搜索输出相似度最高的图片到前端进行显示；最后需要判断当前的特征向量是否存在向量库中，若不存在，将该输入图片和特征向量分别保存到图片库和向量库中。

编写后的可视化 demo 分别如图 7（已建向量库）、图 8（未建向量库）所示，由于 word 不支持视频，详细的演示效果请见提交的 PPT：

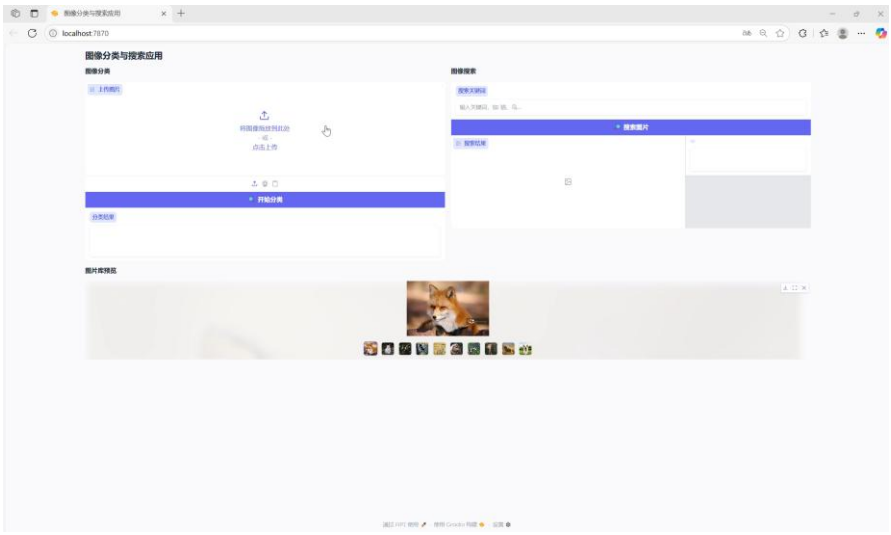


图 7. 已建向量库效果

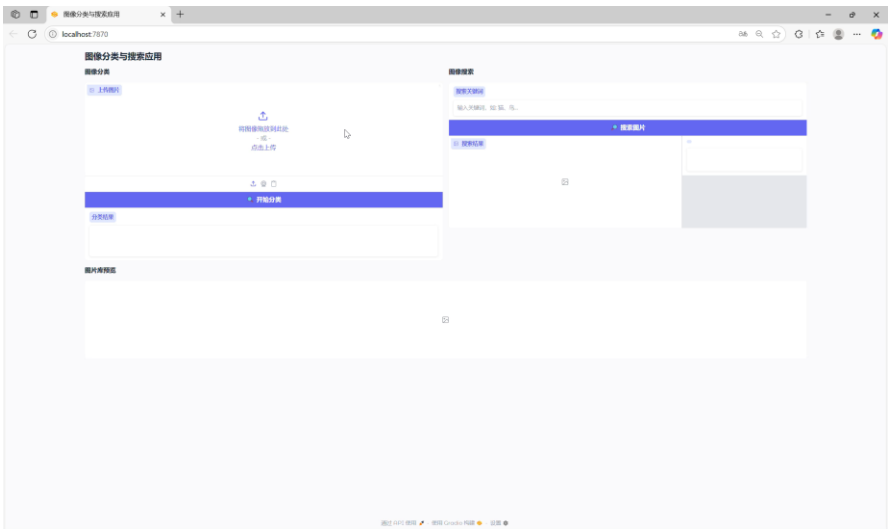


图 8. 新建向量库效果

七、总结思考

CLIP 是多模态领域的代表性模型，其核心价值在于**通过对比学习实现图文语义的深度对齐**，并以零样本能力突破传统监督学习的局限。

通过本次项目总结 CLIP 模型的优势和局限性如下：

1. 优势：

1) 摆脱数据标注依赖：传统 CV 模型需大量人工标注图像标签，而 CLIP 仅需图文对的弱监督信号，大幅降低数据成本；
2) 强大的泛化能力：在未见类别或领域（如艺术画作、医学影像）中表现优异，接近人类的语义理解能力；
3) 统一的跨模态框架：单一模型支持图文检索、分类、生成等多种任务，简化多模态系统开发。

2. 局限性

1) 训练成本极高：训练 CLIP 需数千块 GPU/TPU，消耗大量算力和能源，普通机构难以复现，这也是本项目采用预训练模型的原因；
2) 依赖英文数据：对非英文文本的理解精度较低，需额外优化（如引入多语言预训练）；
3) 语义偏差与伦理风险：训练数据可能包含偏见（如性别、种族刻板印象），导致模型输出存在潜在歧视。
4) 细粒度理解不足：对图像细节（如物体精确位置、纹理）的捕捉弱于纯视觉模型，适合高层语义任务，而非底层视觉任务。

八、项目启发

在本次项目的实施过程中，虽然对于模型的内部没有过多的深入，但是在模型的测试和 API 编写过程中启发性的东西还是不少，具体有：

1. 一开始本来是在我的 RTX4060、8G 的游戏本上测试的，但发现显存不够才去租借显存更大的服务器，后面才发现是使用了计算图的梯度导致显存爆炸，应该设置 model 为 eval 模式，同时设置不使用梯度 torch_no_grad 才不会占用从显存，但确实显存更大的机器跑起来确实比个人的游戏本更快；通过本项目，学会了如何租界服务器以及回顾复习了 Pytorch 使用过程的易漏点；
2. 在编写 API 的过程中，由于处理的不到位直接对图片进行 Resize 导致效果很差，后面经高人指点才发现是没做中心化，因为不同图片宽高比不同，不做中心化定位不到中心，直接 Resize 会丢弃调许多信息；
3. 通过此项目，发现了 CLIP 相比先前一些模型的优势在于零样本分类，通过引入先验知识（比如类别描述、文本标签等），使得模型能够泛化到未训练过的新类别，这点让我大开眼界，成功解决了数据集紧张的问题，这个与人脑的“举一反三”的思想还是比较接近了，个人觉得这点是现在 AI 比较智能的原因之一。