Python 期末大作业报告

MovieLens 数据集的跨模态对齐

沈千帆 2200013220

2024 年 5 月 31 日

备注：助教在测试的时候可以跳过 ipynb 文件里的模型训练部分。预训练模型权重单独放在一个网盘文件夹里，如果要使用下载后拖到文件夹里即可。

1 任务介绍

这次的任务主要是在已给的 2938 部电影对应的文本简介和海报的条件下，通过搭建神经网络探索图片内容和文本内容所存在的关联性。更加具体地说，我们要做的是对于海报和文本搭建一个距离矩阵，共享相似内容的海报和文本距离应该更近，反之应该更远。

2 算法思路

对于整个问题的建模，借鉴了 **CLIP** 论文 [1] 的思路。简单来说，就是用不同的 encoder 分别编码图像和文本两个模态，在我们这个任务里也就分别对应着海报图片和海报简介。接着把所得到的两个特征映射到同一个特征空间中进行**对比学习**。对于 N 个图像-文本对，我们可以得到一个 $N \times N$ 的矩阵，只有对角线上的特征对是匹配的（正样本），而其余的都是负样本。这是一种无监督的训练方式，我们对于得到的`image_feature`和`text_feature`做内积，得到海报特征向量和简介特征向量之间的 cosine 相似度矩阵，大小为 $N \times N$，图像和对应的文本嵌入越相似，那么他们的内积越大。最后用交叉熵训练，把相同电影的海报和简介特征映射成相似的状态 [2]。

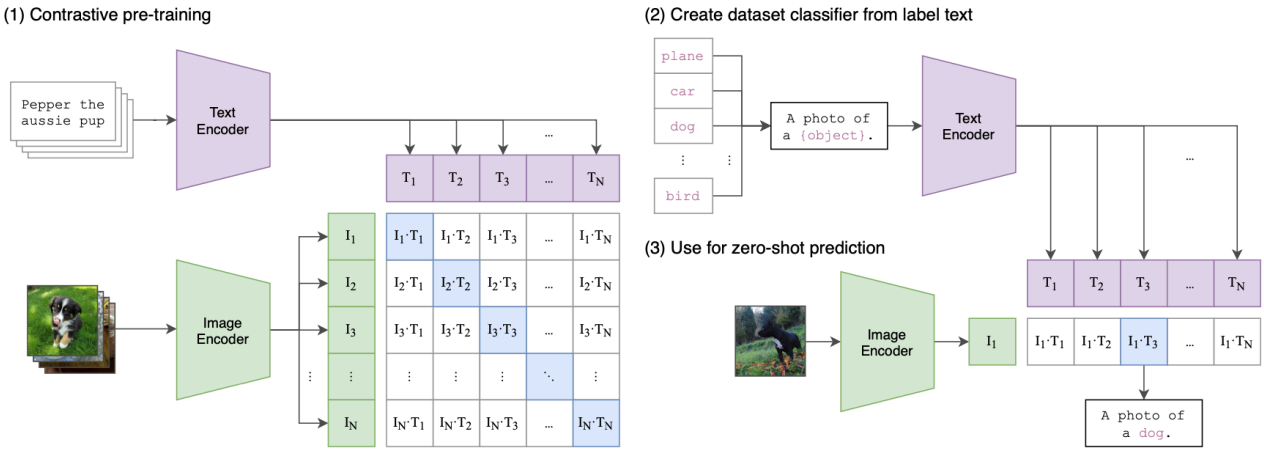


图 1: CLIP 流程图

3 训练和具体分析

对于海报的简介内容，使用了 BERT 预训练模型进行编码；对于海报图像，分别在开始时使用了 ViT 预训练模型和特征处理加 ResNet50 这两种编码方式。关于对比学习，按照原论文的伪代码定义了class NTXentLoss(nn.Module)求损失函数，具体代码如下：

```
def __init__(self, temperature=0.07):
    super(NTXentLoss, self).__init__()
    # temperature控制相似度量的尺度
    self.temperature = temperature
    self.logit_scale = nn.Parameter(torch.ones([]) * np.log(1 / temperature))

def forward(self, image_features, text_features):
    batch_size = image_features.size(0)
    # 归一化统一尺度
    image_features = nn.functional.normalize(image_features, dim=1)
    text_features = nn.functional.normalize(text_features, dim=1)
    # 生成标签用于计算交叉熵损失
    labels = torch.arange(batch_size).long().to(image_features.device)
    # 计算图像-文本和文本-图像的损失
    logit_scale = self.logit_scale.exp()
    logits_per_image = logit_scale * torch.matmul(image_features, text_features.T)
    logits_per_text = logit_scale * torch.matmul(text_features, image_features.T)
    loss_i2t = nn.CrossEntropyLoss()(logits_per_image, labels)
    loss_t2i = nn.CrossEntropyLoss()(logits_per_text, labels)
    return (loss_i2t + loss_t2i) / 2
```

最后训练 70 个 Epoch 得到的 loss 和 acc 曲线分别如下图所示。

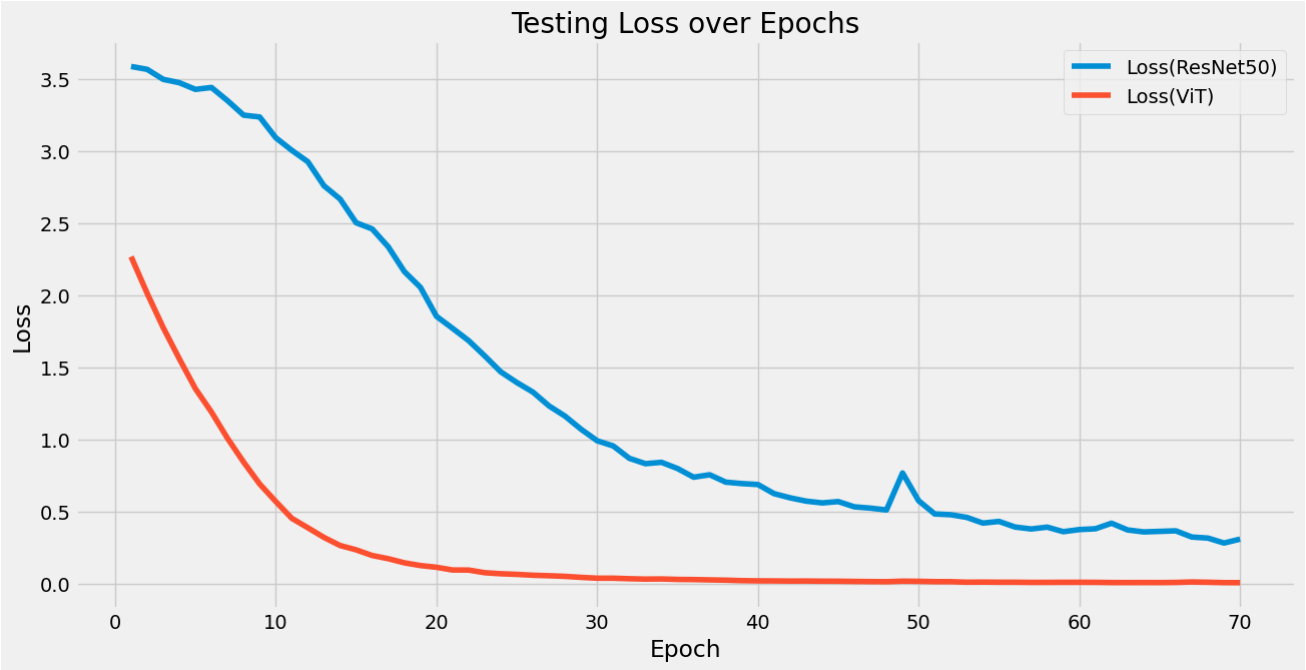


图 2: Loss

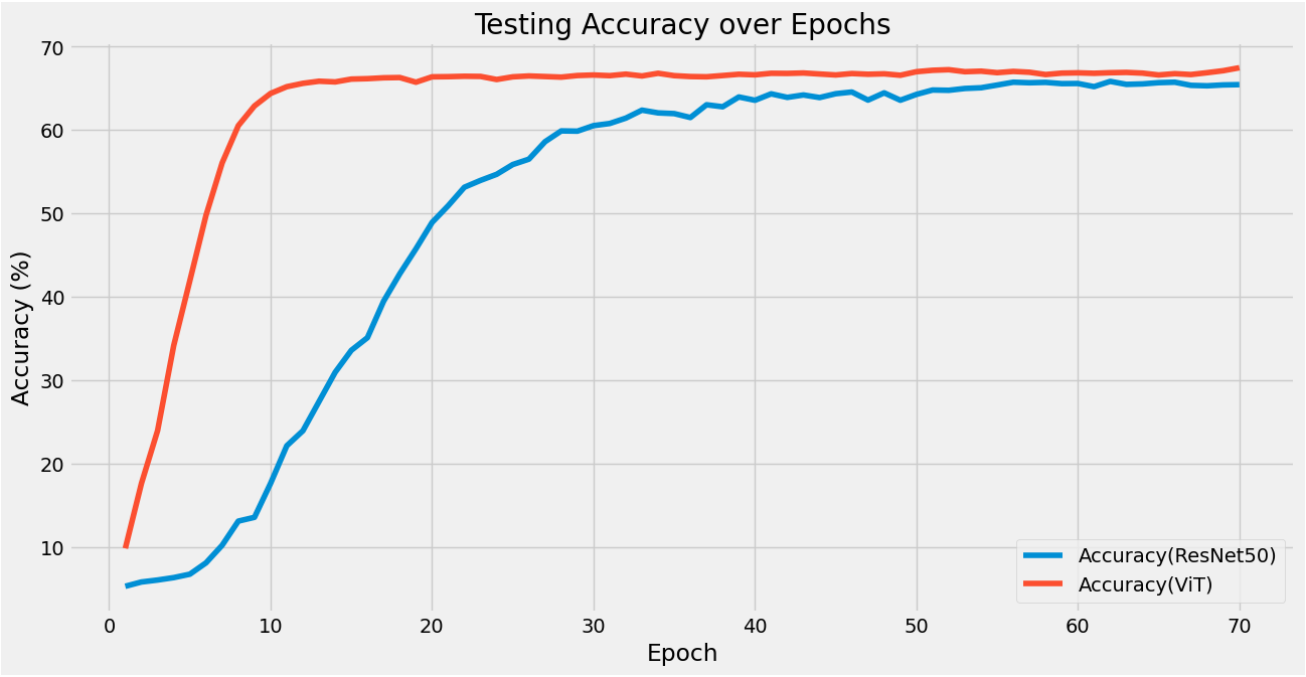


图 3: Accuracy

使用 ResNet50 和 ViT 分别编码海报的平均 accuracy 分别能达到 **65.83%** 和 **67.46%**。最优模型的权重分别存储在`bestmodel_resnet.pth`和`bestmodel_vit.pth`中。在训练的过程中使用`ReduceLROnPlateau`进行学习率调优。

由 loss 和 acc 图我们发现使用 ViT 模型的时候，收敛速度很快，但是到训练后期 loss 和 acc 都很难有提高，而使用 ResNet50 收敛的速度较为均匀和稳定。我认为这是由于 ResNet 通过非常深的卷积层能很好地提取整张图片的局部特征，并且残差神经网络可以避免梯度爆炸等问题，使得模型能保持稳定的提升；而 ViT 是运用了 self-attention 的架构，它更注重于捕捉全局特征的语义信息，因此在刚开始会很好地找到相似性关系。

因此，我尝试一开始用 ViT 的预训练权重去编码海报的特征，接着通过维度的转化放到 ResNet50 中进行训练，得到的效果如下图所示。从图中我们可以看出 hybrid 方式的整个训练过程确实介于了两者中间，并且在最后的几个 epoch 中 accuracy 还是有一些能继续提升的的空间，最终平均 accuracy 是达到 **67.17%**。最优模型的权重存储在`bestmodel_vitresnet.pth`中。

不过可以发现，如果只是使用预训练模型进行编码，不做更多的特征预处理，loss 和 acc 在最后基本上很难有更明显的提升。当然，我认为这也是由数据本身的特点造成的，比如海报的分辨率不是很高导致特征质量不够高；海报的数量还不够多等等。但是由于时间关系，就没有来得及进行更进一步的探索。除了准确度上的量化分析，接下来将展示一下在具体任务下呈现出的效果并进行分析。

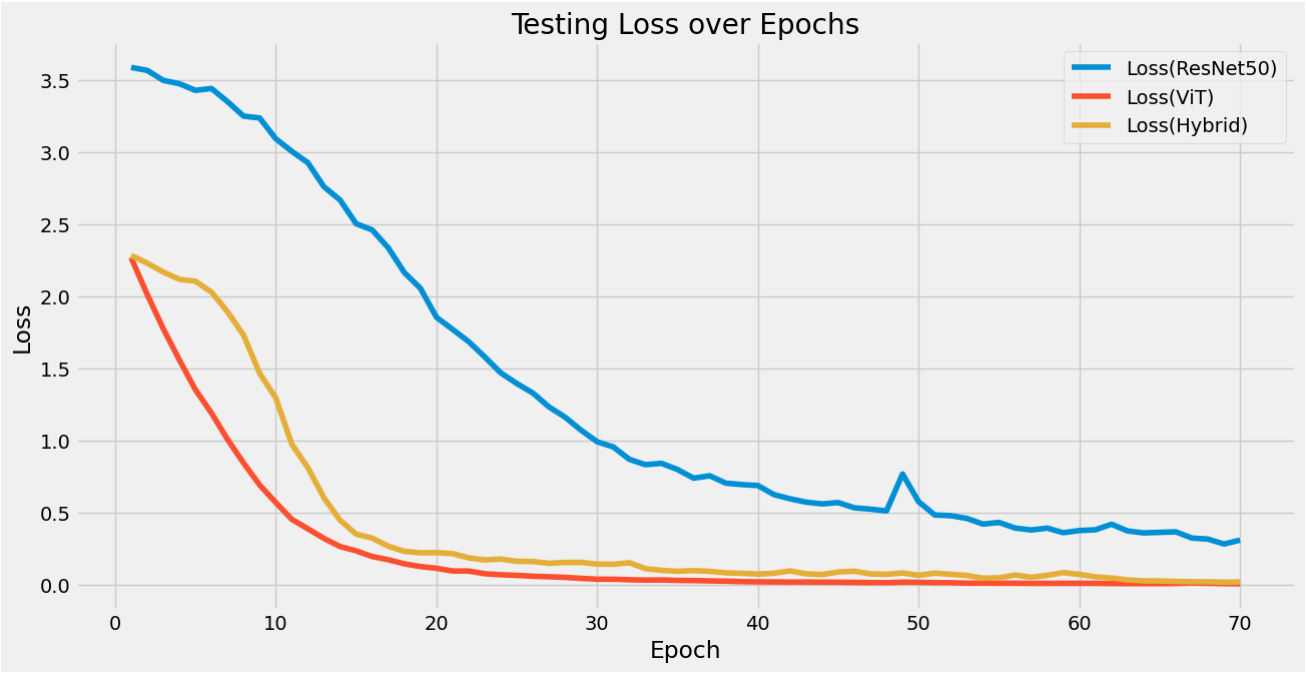


图 4: Loss

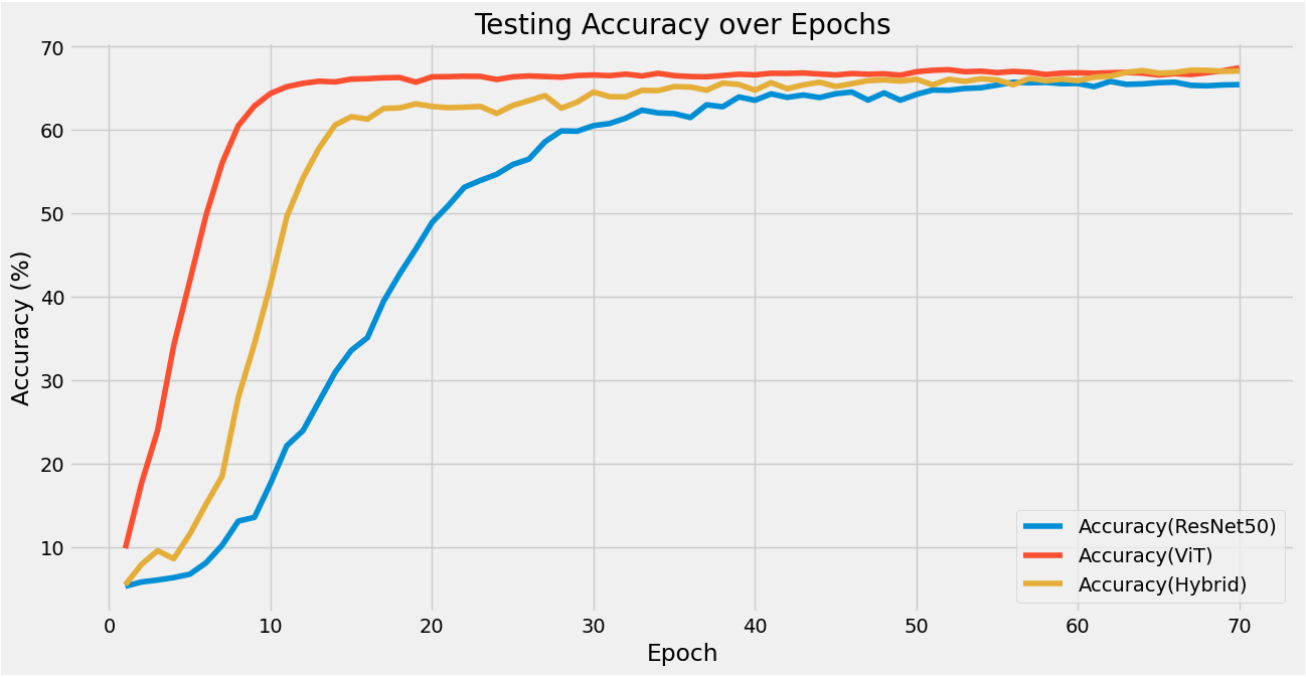


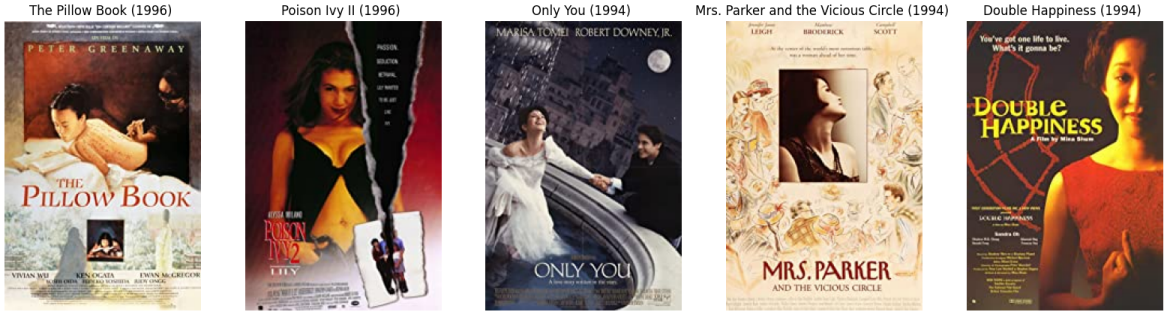
图 5: Accuracy

首先随机选择 6 部电影，用热力图来可视化他们的的距离矩阵，来展现文本（电影简介）和图像（电影海报）之间的相似度。接着，会随机选一部电影，去找它的简介对应的最相似的 5 张海报和它的海报对应的最相似的 5 个电影的简介。呈现的结果分别如下图所示：

图 6: Cosine Distance between Text and Images of 6 Movies

这幅图中呈现的是余弦距离，即 $\text{cos\_distance}(A,B) = 1 - \text{cos\_similarity}(A,B) = 1 - \dfrac{A \cdot B}{\|A\|\|B\|}$。所以越趋近于 0（对应颜色越深越蓝）说明该文本和该图像表达内容更相似，越趋近于 2（对应颜色越黄越亮），说明两者之间相似度越低，当然颜色越相近也说明相似程度是差不多的。我们可以看到对角线上 6 个中 4 个的颜色都很深，说明同一部电影的海报和简介相似度非常高（非常符合 test acc，有点巧）。然后我们可以单独取最后一列的第一，二行，这两个文本相对于这个名为 “Singin’ in the Rain” 的电影海报的相似度是相近的。那我们来看这两个对应的简介文本，分别是 “Shakespeare’s comedy of gender confusion, in which a girl disguises herself as a man to be near the count she adores, only to be pursued by the woman he loves.” 和 “On their way to Africa are a group of rogues who hope to get rich there, and a seemingly innocent British couple. They meet and things happen...” 因此，我们可以推测，由于原海报上比较显著的特征是两男一女和撑着伞，文本中恰好都有提到不同性别（作为提高相似度的因素），但是又都没有出现伞这个特征（作为降低相似度的因素）。我们可以再单独选择一部电影文本，看看与它特征最相似的 5 张海报：

```
movie name: The Pillow Book (1996)
movie intro:
A woman with a body writing fetish seeks to find a combined lover and calligrapher.
```



特别要强调的是，这里选的这部电影并不在训练集中，所以参考价值更大。首先，最相似的一张海报是这部电影自己的海报，这很赞！那么我们再看其他的四张，可以发现共同点都是有一个女性，这和简介中“woman”相符。然后或许这些海报包含一些元素或者风格可能和“lover”相关。我们再取一个电影的海报去找它最相似的一些文本简介：

```
movie name: Germinal (1993)
movie intro: In mid-nineteenth-century northern France, a coal mining town's workers are
exploited by the mine's owner. One day, they decide to go on strike, and the
authorities repress them.
```

- A group of professional bank robbers start to feel the heat from police when they unknowingly leave a clue at their latest heist.
- Three years ago, entomologist Dr. Susan Tyler genetically created an insect to kill cockroaches carrying a virulent disease. Now, the insects are out to destroy their only predator, mankind.
- The early life and career of Vito Corleone in 1920s New York City is portrayed, while his son, Michael, expands and tightens his grip on the family crime syndicate.
- A soldier convicted for murdering his commanding officer is dumped and left to die on a prison island inhabited by two camps of convicts.
- When tradition prevents her from marrying the man she loves, a young woman discovers she has a unique talent for cooking.

对于这样一个海报，该模型没有办法完全准确地找到对应的简介。这个海报的图片内容就是一群面露凶相的男性，手里还握着武器。那排名第一的这个文本简介中有“A group of professional bank robbers”，这很符合这张图片的内容和特点。之后的几则简介中也出现了诸如“kill”，“crime”，“prison”之类的词汇让模型认为该文本和这个海报是相似的。

那现在可以对助教在作业中提出的问题进行一些分析：

1. **（在这一数据集上）电影海报与电影简介内容是否具有某些相关性？**

   这个问题的答案显然是肯定的。我们在训练集上让同一张的电影和海报提取的特征尽可能相似，当扩展到整个测试集的时候，效果还是不错的，说明这存在一定的泛化能力。但是我认为限制模型表现的原因也和海报和简介内容有关。文本简介对于电影特征的描述相较海报本身能提供的是多余的。简介中很多词汇或者表达的意思在海报中其实没有很好地体现。这也间接地导致当我们有一个海报的时候，

很难从较少的信息中获取一个包含更多信息的文本，因为很多简介文本可能都共享这张海报所包含的这类特征。所以光从海报和简介文本中提取出的相关性其实是不足的。

2. **对于给定的海报/简介，最接近的前 K 个简介/海报是否共享相似的语义内容、主题、风格、或其他特征？**

   从前面举的例子来看，他们确实共享类似的语义内容（不过是局部的）、主题风格。这里可以再举一个例子：

```
movie name: Double Team (1997)
movie intro: An international spy teams up with an arms dealer to escape from a penal
colony and rescue his family from a terrorist.
```

   这个例子中，这几个得到的海报就共享类似的风格。不过就正如再上一个问题中所回答的，这个思路去做跨模态对齐本质上还是做了一个更大空间内的分类问题。它更多地是从特征中提取一些信息和特征。比如说海报里人物的占比非常大，所以判断标准里这方面的特征就体现的特别明显。但实际上在语义整体内容的理解上并没有做的特别好，数据或者特征处理上的不足会很大程度上影响模型的表现。

3. **跨模态对齐的特征可能在哪些下游任务上发挥怎么样的作用？**

   - **图像-文本检索**。这基本上和本次作业的任务一样可以用图像检索文本或者用文本检索图像。
   - **跨模态生成**。在上述基础上，可以用得到的特征信息进行文本-图像或者图像-文本的生成。
   - **零样本学习（Zero-Shot Learning）**。跨模态对齐使得模型能够在没有见过特定类别图像的情况下，通过自然语言描述来进行分类、从而经过训练之后可以根据文本描述来预测新的类别。
   - **图像标注**。模型可以更好地理解图像内容，并生成与之相关的文本描述。特征对齐之后有助于捕捉图像和文本之间的语义关联。

References

[1] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[2] Zyw2002. *CLIP 论文讲解和代码实操*. https://blog.csdn.net/zyw2002/article/details/129836180. 2024.