CLIP学习

1 模型概述

1.1 基本定义

CLIP（Contrastive Language-Image Pre-Training，对比语言 - 图像预训练）是一种多模态模型，核心是构建视觉与语言的关联表示，能在无需训练数据的情况下制作高度特定且性能卓越的分类器，为多模态建模提供了创新思路。

1.2 适用人群

任何对计算机视觉、自然语言处理（NLP）或多模态建模感兴趣的人员，均可通过学习 CLIP 模型深入了解跨模态数据处理的方法与应用。

---

2 CLIP与传统图像分类器的对比

2.1 传统图像分类器

- **训练方式**：采用监督学习，以区分猫和狗的模型为例，需向模型输入大量猫和狗的图像，模型根据预测误差逐步调整参数，直至能准确区分二者。例如输入已知猫图像时，若初始输出为 “猫：50%，狗：50%”，则调整参数使输出趋近 “猫：100%，狗：0%”；输入已知狗图像时，若初始输出为 “猫：52%，狗：48%”，则调整参数使输出趋近 “猫：0%，狗：100%”。
- **局限性**：模型高度专业化，仅在训练数据覆盖的范围内表现良好，当面对包含相同类别但不同表示形式的类似数据集时，性能会大幅下降。

2.2 CLIP 模型优势

- **泛化能力强**：在不同数据集上表现稳健。如下表所示，与传统监督模型（ImageNet ResNet101）相比，CLIP 在多个数据集上展现出更优性能，尤其在 ImageNet-R 数据集上，性能提升高达 51.2%。

数据集	ImageNet ResNet101	Zero-Shot CLIP	性能差异（A Score）
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	-	-

- **创新分类思路**：摒弃传统 “预测图像所属类别” 的模式，转为 “预测图像是否与任意标题匹配”。例如对于 “穿着漂亮夹克的狗的图片” 这一文本，CLIP 能判断给定图像与该文本的匹配度，如某张狗的图像与该文本匹配度为 86%，而其他无关图像匹配度仅为 2%。

3 核心原理

3.1 核心思想

利用从互联网抓取的带字幕图像进行训练，让模型学习预测文本与图像是否匹配，通过对比学习建立图像和文本之间的关联。

3.2 对比学习机制

- **Embedding 学习**：CLIP 通过学习图像编码器和文本编码器，将图像和文本分别转换为 Embedding 向量（即数字列表）。在多模态 Embedding 空间中，匹配的图像和文本 Embedding 相似度高、距离近，不匹配的则相似度低、距离远。
- **训练目标**：训练过程中，最大化不匹配图文对的 Embedding 距离，最小化匹配图文对的 Embedding 距离，使模型能准确区分图文是否匹配。

3.3 相似度计算

CLIP 采用余弦相似度衡量两个 Embedding 向量的相似程度：

- 若两个向量夹角小，余弦相似度接近 1，表明两者高度相似；
- 若向量夹角为 90 度，余弦相似度为 0，说明两者无相似性；
- 若向量方向相反，余弦相似度为 - 1，代表两者完全不相似。
- 余弦相似度计算公式为：(s(\theta)=\frac{A\cdot B}{\vert A\vert\vert B\vert})，其中A和B为两个向量，(A\cdot B)是它们的点积，(\vert A\vert)和(\vert B\vert)分别是两个向量的模。

---

4 组成部分

CLIP 是一种高层框架，可灵活搭配不同子组件，核心由文本编码器和图像编码器构成。

4.1 文本编码器

- **功能**：将输入文本转换为能表示文本含义的 Embedding 向量。
- **基础架构**：基于标准 Transformer 编码器，Transformer 通过自注意力机制，对输入的单词序列进行重新表示和比较，生成具有上下文信息的抽象表示。
- **CLIP 优化**：与通用 Transformer 输出矩阵不同，CLIP 的文本编码器仅输出一个向量，直接提取输入文本序列中最后一个标记的向量作为整个文本序列的表示。
- **示例**：输入文本 “A picture of a dog wearing a fabulous jacket”，经过文本编码器处理后，输出对应的文本 Embedding 向量。

4.2 图像编码器

- **功能**：将输入图像转换为反映图像含义的 Embedding 向量。
- **可选架构**：CLIP 论文中提及多种图像编码器，本文以 ResNet-50 为例。ResNet-50 属于卷积神经网络（CNN），是一种成熟的图像处理架构。
- 工作原理：
    1. **卷积操作**：使用卷积核（小值矩阵）扫描图像，根据卷积核与输入图像的像素值计算新像素值，提取图像局部特征；

2. **下采样**：通过最大池化（Maxpool）等下采样操作，压缩图像尺寸，保留关键特征，减少计算量；
3. **特征转换**：经过多轮卷积和下采样，提取图像的抽象特征，再通过密集网络（Dense Layer）将这些特征转换为最终的图像 Embedding 向量。

- **示例**：输入一张图像，经图像编码器处理后，输出对应的图像 Embedding 向量，该向量可视为图像在高维向量空间中的抽象位置表示。

---

5 应用场景

1. **图像分类器构建**：提供多个文本描述（如 “一张猫的照片”“一张狗的照片”），让 CLIP 判断输入图像与哪个文本描述最相关，从而实现图像分类。例如在 CIFAR-100 数据集上，CLIP 对 “snake” 类别的识别准确率达到 38.02%，排名 1/100；在 ImageNetV2 Matched Frequency 数据集上，对 “beer bottle” 类别的识别准确率为 88.27%，排名 1/1000。
2. **图像搜索系统**：输入目标文本（如 “一张狗的照片”），CLIP 从众多图像中找出与该文本最匹配的图像，实现精准图像搜索。
3. Embedding 向量提取：
   - 利用图像编码器提取与文本相关的图像抽象信息（即图像 Embedding），该 Embedding 可用于其他机器学习模型的输入；
   - 借助文本编码器抽取文本 Embedding，为其他模型提供文本特征支持。

---

6 性能表现（部分数据集）

数据集	正确标签	正确排名	正确概率
CIFAR-100	snake	1/100	30.02%
ImageNetV2 Matched Frequency	beer bottle	1/1000	65.27%
FGVC Aircraft	Boeing 717	2/100	-
RESISC45	roundabout	1/45	96.39%
Stanford Cars	2012 Honda Accord Coupe	1/196	63.30%
SUN	kennel indoor	1/123	98.63%
Caltech-101	kangaroo	1/102	99.81%
UCF101	Volleyball Spiking	1/101	99.30%
ImageNet-R (Rendition)	Siberian Husky		

BLIP学习

1 概述

BLIP（Bootstrapping Language-Image Pretraining）是 Salesforce 于 2022 年提出的多模态框架，实现了视觉与语言理解和生成的统一。它引入跨模态的编码器和解码器，促进跨模态信息流动，在多项视觉和语言任务中取得 SOTA（State-of-the-Art）性能。在 AIGC 领域，BLIP 常被用于为图像生成 prompt，例如 ControlNet 中的 Automatic Prompt 便由 BLIP 生成，优质的 prompt 对交叉注意力的微调至关重要。

其名称中的 “Bootstrapping”（自举）源于训练数据的特性 —— 训练数据来自网络图文对，包含大量噪声。因此，BLIP 增加了在线数据打标签和清理任务，将处理后的优质数据用于迭代优化原模型，实现模型性能的逐步提升。

---

2 模型结构

BLIP 引入了**多模态混合结构 MED（Multimodal mixture of Encoder-Decoder）**，该结构能高效进行多任务预学习和迁移学习，具体组成如下表所示：

组件	功能描述
Image Encoder（图像编码器）	负责从输入图像中提取视觉特征，是单模态编码器的重要组成部分
Text Encoder（文本编码器）	对输入文本进行编码处理，提取文本的潜在特征，属于单模态编码器
Image-grounded Text Encoder（以图像为基础的文本编码器）	结合图像信息对文本进行编码，用于建模图文多模态信息的相关性
Image-grounded Text Decoder（以图像为基础的文本解码器）	在图像信息的引导下，以自回归的方式生成目标 caption（图像描述）

BLIP 通过三个核心损失函数联合进行预训练，以实现视觉与语言特征的有效对齐和模型性能优化，各损失函数详情如下：

1. 图像 - 文本对比损失 ITC（Image-Text Contrastive Loss）
    - 作用对象：图像编码器和文本编码器。
    - 原理：通过正负图文对的对比学习，将匹配的图文对（正样本）在潜在特征空间中拉近，将不匹配的图文对（负样本）推远，从而对齐图像和文本的潜在特征空间。
2. 图像 - 文本匹配损失 ITM（Image-Text Matching Loss）
    - 作用对象：以图像为基础的文本编码器。
    - 原理：将图文匹配性作为二分类任务（匹配为正类，不匹配为负类），训练模型判断给定的图像和文本是否匹配，进而建模图文多模态信息的相关性。
3. 语言建模损失 LM（Language Modeling Loss）
    - 作用对象：以图像为基础的文本解码器。
    - 原理：采用交叉熵损失进行优化，训练模型在图像信息的引导下，以自回归的方式生成符合图像内容的目标 caption，提升模型的文本生成能力。

---

3 训练方法

网络上获取的大量图文对通常存在不准确、错误等噪声信息，为高效利用这类数据，BLIP 提出了**caption 生成和过滤模块 CapFilt（Captioning and Filtering）**，通过 “学习 - 生成 - 过滤 - 迭代” 的流程优化训练数据，进而提升模型性能。

3.1 CapFilt 模块组成

CapFilt 包含两个关键子模块，二者协同工作实现对噪声数据的处理：

- Captioner（caption 生成器）
  - 初始化来源：从预训练的模型初始化。
  - 微调方式：在人工标注数据集上，以语言建模损失（LM）为目标进行微调。
  - 功能：对给定的网络图像生成合成 caption（synthetic texts），为后续数据扩充和过滤提供文本素材。
- Filter（过滤器）
  - 初始化来源：同样从预训练的模型初始化。
  - 微调方式：在人工标注数据集上，以图像 - 文本对比损失（ITC）和图像 - 文本匹配损失（ITM）为目标进行微调。
  - 功能：判断文本与图像是否匹配，过滤掉原始网络文本（web texts）和合成文本（synthetic texts）中的噪声 caption，保留高质量的图文对。

3.2 训练流程（Bootstrap 过程）

BLIP 的训练流程遵循自举思想，通过多轮迭代不断优化模型，具体步骤如下：

1. **初始预训练**：使用原始网络图文对（{(I_w, T_w)}，其中 I_w 为网络图像，T_w 为网络文本）和人工标注图文对（{(I_h, T_h)}，其中 I_h 为人工标注图像，T_h 为人工标注文本）对 MED 结构进行初始预训练。
2. 生成与过滤：
   - 利用微调后的 Captioner 对网络图像 I_w 生成合成文本 T_s，得到合成图文对 {(I_w, T_s)}。
   - 利用微调后的 Filter 对原始网络文本 T_w 和合成文本 T_s 进行过滤，得到过滤后的网络文本 T_w' 和过滤后的合成文本 T_s'。
3. **构建新数据集**：将过滤后的网络图文对 {(I_w, T_w')}、过滤后的合成图文对 {(I_w, T_s')} 以及人工标注图文对 {(I_h, T_h)} 组合，形成新的、高质量的训练数据集 D。
4. **迭代优化**：使用新数据集 D 重新预训练模型，重复步骤 2-4，通过多轮迭代实现模型性能的持续提升。

3.3 下游任务性能提升

通过 CapFilt 模块的处理和 Bootstrap 迭代训练，BLIP 在多种下游视觉 - 语言任务上均取得了稳定的性能提升，包括但不限于：

- 图像 - 文本检索（Image-Text Retrieval）
- 图像标题生成（Image Captioning）
- 视觉问答（Visual Question Answering, VQA）
- 视觉推理（Visual Reasoning）
- 视觉对话（Visual Dialogue）

LLaVA学习

1 模型结构

LLaVA 模型结构简洁，核心为 “CLIP 视觉编码器 + LLM（如 Vicuna）”，具体流程如下：

1. **图像特征提取**：利用 CLIP 视觉编码器（如 CLIP-ViT-L/14）将输入图像(X_v)转换为维度为([N=1, grid_H\times grid_W, hidden_dim])的 feature map，得到视觉特征(Z_v = g(X_v))。需注意，LLaVA 使用最后一层 Transformer 之前或之后的 grid features 作为图像表示，而非 CLIP 最后的输出层。
2. **特征维度对齐**：通过一个可训练的投影层（Projection W）将视觉特征(Z_v)转换为与语言模型中单词嵌入空间维度相同的语言嵌入标记(H_v)，公式为(H_v = W\cdot Z_v)，转换后得到维度为([N=1, grid_H\times grid_W = image_seqlen, emb_dim])的特征。
3. **模型输入构建**：将图像 token embedding（(H_v)）和文本 token embedding 合并，作为语言模型的输入，最终由语言模型生成描述文本。

与 InstructBLIP 或 Qwen-VL 相比，LLaVA 架构设计更简单。InstructBLIP 或 Qwen-VL 需在数亿甚至数十亿的图像文本配对数据上训练专门设计的视觉重新采样器，而 LLaVA 仅需在 600K 个图像 - 文本对上训练一个简单的完全连接映射层。LLaVA 在多模态指令跟随数据集上表现出色，与 GPT-4 相比分数达 85.1%，在 Science QA 上准确率刷新纪录，达 92.53%。

---

2 LLaVA 两阶段训练

阶段一：特征对齐预训练（Pre-training for Feature Alignment）

1. **数据处理**：为平衡概念覆盖度和训练效率，将 CC3M 数据集筛选为 595K 个图像 - 文本对，通过简单扩展方法将其转换为指令跟随数据，每个样本视为单轮对话。对于图像(X_v)，随机采样一个语言指令作为问题(X_q)，要求助手简要描述图像，原始标题作为真实答案(X_a)。
2. **训练策略**：冻结视觉编码器（Vision Encoder）和 LLM 模型的权重参数，仅训练投影层（Projection W）的权重，最大化真实答案(X_a)的似然性，使图像特征(H_v)与预训练 LLM 的单词嵌入对齐，此阶段可理解为为冻结的 LLM 训练兼容的视觉分词器。

阶段二：端到端训练（Fine-tuning End-to-End）

1. **训练策略**：依然冻结视觉编码器的权重，训练过程中同时更新投影层（Projection W）和 LLM 语言模型的权重。
2. 任务场景
    - **多模态聊天机器人（Multimodal Chatbot）**：在 158K 个语言 - 图像指令跟随数据上进行微调，数据包含三种响应类型，其中对话为多轮，另外两种为单轮，训练时均匀采样。
    - **Science QA 任务**：在 Science QA 基准数据集上进行研究，该数据集是首个大规模多模态科学问题数据集，每个问题提供自然语言或图像形式的上下文。助手需用自然语言提供推理过程，并在多个选项中选择答案。训练时将数据组织为单轮对话，问题和上下文作为指令输入(X_{instruct})，推理过程和答案作为输出(X_a)。