自然语言处理导论

MiniGPT4-Finetuning 项目报告

Author: 夏子渊 金裕涵 林滨 程韬 潘越

Date: 2025-06-20

2025 春夏 Semester

Table of Contents

1. 摘要

MiniGPT4-Finetuning 项目旨在基于视觉语言模型 MiniGPT-4[1]，对 **Flickr30k**[2] 数据集开展指令微调，以增强图像-文本理解与生成能力。本文档总结了项目背景、环境配置、模型下载、微调流程、评估结果与未来工作等内容，为后续复现与迭代提供参考。

2. Project Introduction

MiniGPT-4 通过结合视觉编码器与 Vicuna 语言模型，实现了图像到文本的高质量对齐。为了进一步提升其在 Image Captioning 任务上的理解与表现，我们开展了面向 Flickr30k 数据集的指令微调。

实验环境:

- 操作系统: Ubuntu 18.04 LTS
- 显卡: NVIDIA GeForce RTX 3090 × 7
- Python: 3.10.13
- CUDA: 11.7
- PyTorch: 2.0.1

3. Technical Details

3.1. 理论知识

3.1.1. 指令微调

指令微调（Instruction Tuning）是指在大规模预训练语言模型的基础上，利用带有明确指令（instruction）的数据对模型进行再训练，使其能够更好地理解和执行各种自然语言指令。其核心思想是通过多样化的指令-响应对，提升模型的泛化能力和任务适应能力。

具体来说，假设我们有一个预训练模型 f_θ，输入为图像 I 和文本指令 x，输出为文本 y。指令微调的目标是最小化如下损失函数:

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(I,x,y)\sim\mathcal{D}}[-\log P_\theta(y|I,x)]$$

其中 $\mathcal{D}$ 表示带有指令的训练数据集。

例如，给定一张图片 I 及指令 x:

- 指令 x: “请详细描述这张图片的内容。”
- 期望输出 y: “一只小狗在草地上奔跑，背景有蓝天和白云。”

通过指令微调，模型不仅学习到图片与文本的对应关系，还能理解“详细描述”这类指令的语义，从而在遇到不同指令时做出相应的生成。

3.1.2. 视觉语言模型

视觉语言模型（Vision-Language Model, VLM）是一类能够同时处理图像和文本信息的深度学习模型。其核心思想是将图像通过视觉编码器（如 CLIP、ViT 等）提取为一组高维特征向量，这些特征向量通常被称为 image token。具体来说，输入的原始图像首先经过视觉编码器，被分割成若干 patch，每个 patch 经过编码后生成一个 token，所有 token 共同表征整张图片的语义信息。

随后，这些 image token 会被送入与文本模型（如大语言模型 LLM）对齐的投影层，将视觉特征映射到与文本 token 相同的向量空间。这样，模型就可以将图像 token 与文本 token 拼接在一起，输入到统一的 Transformer 或多模态架构中，实现图像与文本的联合建模。通过这种方式，视觉语言模型不仅能够理解图片内容，还能根据输入的文本指令生成相应的文本描述，实现图像到文本的生成任务（如 Image Captioning、视觉问答等）。

以 MiniGPT-4 为例，其流程为：图像输入后，首先由视觉编码器生成 image token，经过线性投影层后与文本 token 拼接，最终输入到 Vicuna 语言模型中进行推理和生成。这种设计使得模型能够充分融合视觉和语言信息，提升多模态理解与生成能力。

3.1.3. 评估指标

3.1.3.1. BLEU

BLEU (Bilingual Evaluation Understudy) 是一种广泛使用的机器翻译评估指标。它通过比较生成文本与参考文本之间的 n-gram 重叠程度来计算分数。BLEU 分数范围从 0 到 1,1 表示完全匹配。主要特点:

- 计算 n-gram 精确率
- 使用简短惩罚因子
- 支持多个参考文本
- 对词序敏感

3.1.3.2. CIDEr

CIDEr (Consensus-based Image Description Evaluation) 是专门为图像描述任务设计的评估指标。它的主要特点包括:

- 使用 TF-IDF 加权来强调重要词汇
- 考虑 n-gram 的共识程度
- 对罕见但准确的描述给予更高权重
- 分数范围从 0 到 10,越高越好

3.1.3.3. ROUGE-L

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) 是一种基于最长公共子序列的评估指标。其特点:

- 不要求严格连续匹配

- 考虑词序信息
- 计算召回率和精确率
- 对句子结构敏感
- 分数范围从 0 到 1

3.2. 技术细节

3.2.1. 转换数据集

我们使用 `prepare_flickr30k.py` 脚本将 Flickr30k 数据集转换为适合指令微调的格式。该脚本读取原始的 `flickr_annotations_30k.csv` 文件，并生成一个 JSON 文件。对于每一个样本，我们将其 5 个参考描述（`caption` 列表）用空格连接起来，并和其对应的图像 ID 一起存储在 JSON 文件中。

3.2.2. 定义测评指标

我们使用上述 BLEU、CIDEr 和 ROUGE-L 等指标评估模型性能。

在自定义的 `eval_scripts/eval_flickr30k.py` 中，我们利用 `COCO API` 实现了上述评估指标的计算。

4. Experiment Results

4.1. 实验流程

4.1.1. Conda 环境

```
conda env create -f environment.yml
conda activate minigptv
```

4.1.2. 预训练模型获取

下载 Vicuna-7B 语言模型：

```
git clone https://huggingface.co/Vision-CAIR/vicuna-7b
cd vicuna-7b && git lfs pull
```

下载 MiniGPT-4 模型权重[3]，并配置 `train_configs/minigpt4_flickr_finetune.yaml`。

4.1.3. 数据准备

执行脚本生成注解文件：

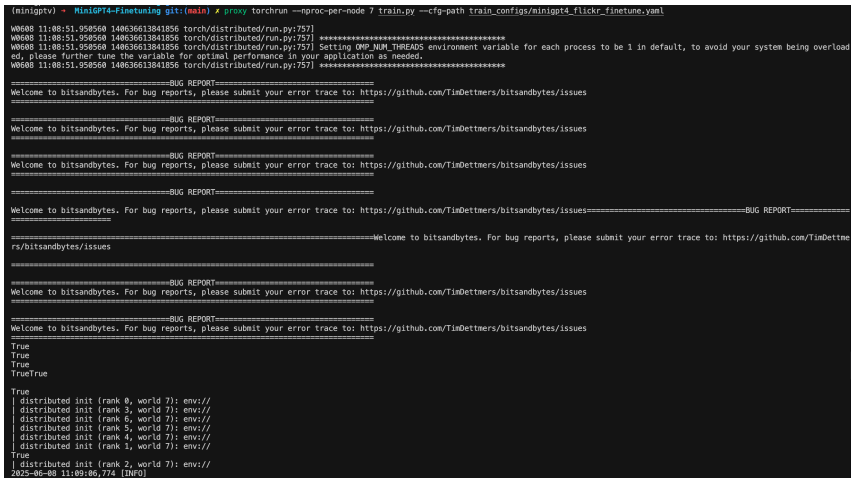
```
python prepare_flickr30k.py
```

4.1.4. 指令微调

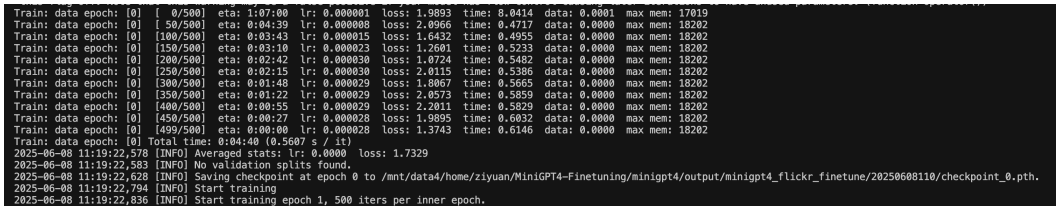
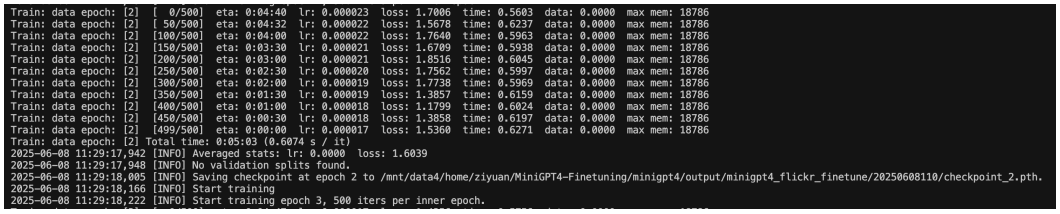
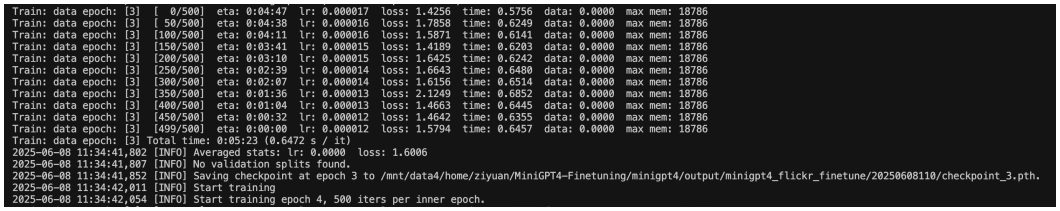
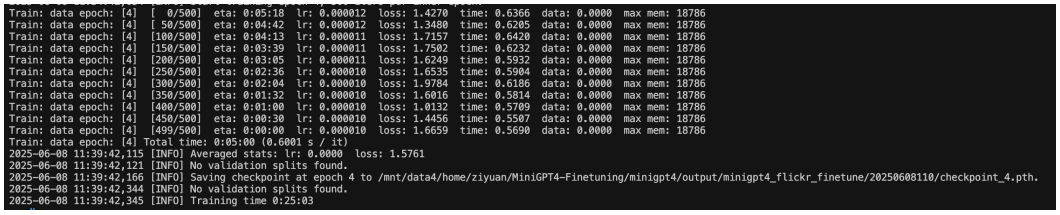
使用单卡 GPU 训练线性映射层：

```
torchrun --nproc-per-node 1 train.py --cfg-path train_configs/
minigpt4_flickr_finetune.yaml
```

训练过程部分如图所示：



训练结果部分如图所示：

epoch0:	
epoch1:	
epoch2:	
epoch3:	
epoch4:	

4.1.5. 评估

运行如下脚本在多模型间对比性能：

```
bash evaluate.sh
```

4.2. 评估结果与分析

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
pretrained	0.3175	0.2371	0.1843	0.1529	0.4361	0.5758
epoch0	0.3427	0.2723	0.2204	0.2002	0.4502	0.6101
epoch1	0.3608	0.2822	0.2503	0.2320	0.4753	0.6469
epoch2	0.3728	0.3021	0.2720	0.2456	0.4829	0.6602
epoch3	0.3878	0.3314	0.2978	0.2688	0.4907	0.6701
epoch4	**0.3912**	**0.3375**	**0.3005**	**0.2807**	**0.5115**	**0.6803**

表 1　模型评估结果统计

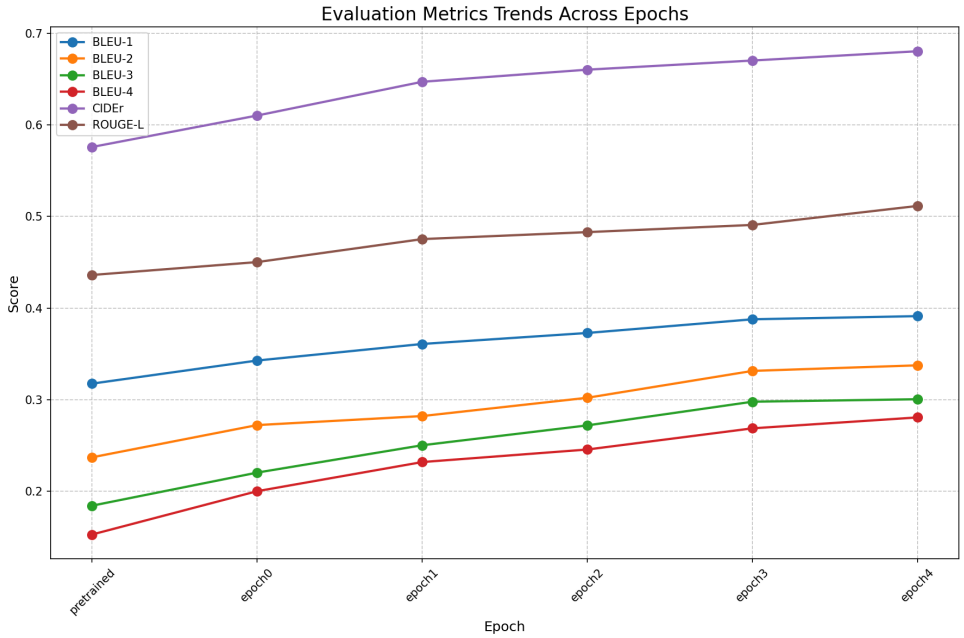


图 1　模型评估结果折线图

4.2.1. 模型性能持续提升

随着训练轮次（Epoch）的增加，所有评价指标均呈现稳定上升趋势，表明模型通过迭代学习有效捕捉了文本生成任务的核心规律。其中：

- 语义与连贯性优化显著：CIDEr 和 ROUGE-L 的增速远超其他指标（详见图表斜率），说明模型在生成内容的语义相关性、上下文连贯性上提升最为突出，逐渐接近人类语言表达模式。
- 局部一致性稳步改进：BLEU 系列指标增长平缓但持续（BLEU-1 至 BLEU-4 增幅约 50%-80%），反映模型在局部词汇匹配和短语结构的准确性上逐步完善。

4.2.2. 训练动态揭示关键拐点

- 早期快速收敛：pretrained 至 epoch1 阶段所有指标快速跃升，验证预训练权重提供了高质量初始化。

- 中后期差异化优化：epoch2 后 CIDEr 与 ROUGE-L 仍保持陡峭上升，而 BLEU 系列进入平缓增长期，表明模型后期更侧重于语义整体性而非局部词序精确度，符合文本生成任务的本质目标。
- 持续训练价值：截至 epoch4，各曲线仍未出现平台期，建议扩展训练轮次（如至 epoch6）以挖掘性能潜力。
- 重点优化方向：可针对性设计长依赖文本和抽象语义的增强训练模块（如注意力机制改进），进一步发挥 CIDEr 与 ROUGE-L 的优势。

4.2.3. 局限性与后续工作

- 评估维度补充：需增加人工评价或SPICE等细粒度指标，验证模型在视觉语义对齐上的表现（若为多模态任务）。
- 泛化能力检验：当前结果基于单一数据集，需在跨领域数据上验证鲁棒性。
- 探索更大的语言模型后端。
- 优化推理速度与显存占用。

References

[1] D. Zhu, J. Chen, X. Shen, X. Li, 和 M. Elhoseiny,《MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models》. [在线]. 载于: https://arxiv.org/abs/2304.10592

[2] P. Young, A. Lai, M. Hodosh, 和 J. Hockenmaier,《From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions》, *Transactions of the Association for Computational Linguistics*, 卷 2, 页 67–78, 2014.

[3] D. Zhu, J. Chen, X. Shen, X. Li, 和 M. Elhoseiny, 《GitHub - Vision-CAIR/MiniGPT-4: Open-sourced codes for MiniGPT-4 and MiniGPT-v2 — github.com》. [在线]. 载于: https://github.com/Vision-CAIR/MiniGPT-4