MiniGPT4-Finetuning

项目展示

夏子渊 金裕涵 林滨 程韬 潘越

NLP Group 33

Jun. 22nd, 2025

Outline

Outline

1. 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。

1. Project Introduction

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

Outline

2.1 理论知识

2.1.1 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。

2.1 理论知识

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。

2.1 理论知识

2.1.2 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，其核心思想是通过视觉编码器（如 CLIP、ViT 等）将图像提取为 image token，然后通过投影层将视觉特征映射到与文本 token 相同的向量空间，最终将图像和文本 token 拼接输入到统一的 Transformer 架构中实现联合建模，从而能够理解图片内容并根据文本指令生成相应描述，完成图像到文本的生成任务。

2.2 技术细节

2.2.1 评估指标

BLEU: 机器翻译评估指标，基于 n-gram 重叠程度，分数 0-1

- 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

CIDEr: 图像描述专用评估指标，分数 0-10

- 使用 TF-IDF 加权强调重要词汇
- 考虑 n-gram 共识程度，对罕见准确描述给予更高权重

ROUGE-L: 基于最长公共子序列的评估指标，分数 0-1

- 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

2.2 技术细节

2.2.2 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。

2.2.3 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。

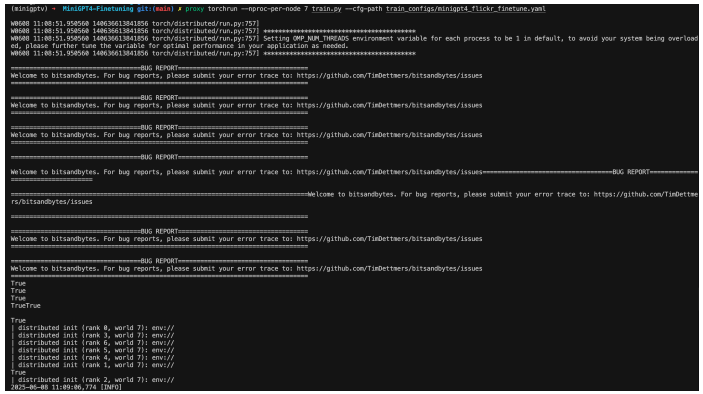
在自定义的 `eval_scripts/eval_flickr30k.py` 中，我们利用 `COCO API` 实现了上述评估指标的计算。

Outline

3.1 训练过程展示

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
```



3.1 训练过程展示

部分训练过程如图所示：

```
(minigptv) ➜  MiniGPT4-Finetuning git:(main) ✗ proxy torchrun --nproc-per-node 7 train.py --cfg-path train_configs/minigpt4_flickr_finetune.yaml

W0608 11:08:51.950560 140636613841856 torch/distributed/run.py:757]
W0608 11:08:51.950560 140636613841856 torch/distributed/run.py:757] *****************************************
W0608 11:08:51.950560 140636613841856 torch/distributed/run.py:757] Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overload
ed, please further tune the variable for optimal performance in your application as needed.
W0608 11:08:51.950560 140636613841856 torch/distributed/run.py:757] *****************************************

===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
================================================================================

===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
================================================================================

===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
================================================================================

===================================BUG REPORT===================================

Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues===================================BUG REPORT=============
======================

================================================================================Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettme
rs/bitsandbytes/issues

================================================================================

===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
================================================================================

===================================BUG REPORT===================================
Welcome to bitsandbytes. For bug reports, please submit your error trace to: https://github.com/TimDettmers/bitsandbytes/issues
================================================================================
True
True
True
TrueTrue

True
| distributed init (rank 0, world 7): env://
| distributed init (rank 3, world 7): env://
| distributed init (rank 6, world 7): env://
| distributed init (rank 5, world 7): env://
| distributed init (rank 4, world 7): env://
| distributed init (rank 1, world 7): env://
True
| distributed init (rank 2, world 7): env://
2025-06-08 11:09:06,774 [INFO]
```

```
=====  Running Parameters    =====
2025-06-08 11:09:06,775 [INFO] {
    "amp": true,
    "device": "cuda",
    "dist_backend": "nccl",
    "dist_url": "env://",
    "distributed": true,
    "evaluate": false,
    "gpu": 0,
    "init_lr": 3e-05,
    "iters_per_epoch": 500,
    "job_name": "minigpt4_flickr_finetune",
    "lr_sched": "linear_warmup_cosine_lr",
    "max_epoch": 5,
    "min_lr": 1e-05,
    "num_workers": 4,
    "output_dir": "output/minigpt4_flickr_finetune",
    "rank": 0,
    "resume_ckpt_path": null,
    "seed": 42,
    "task": "image_text_pretrain",
    "train_splits": [
        "train"
    ],
    "wandb_log": true,
    "warmup_lr": 1e-06,
    "warmup_steps": 200,
    "weight_decay": 0.05,
    "world_size": 7
}
2025-06-08 11:09:06,775 [INFO]
======  Dataset Attributes  ======
2025-06-08 11:09:06,775 [INFO]
======== flickr30k_grounded_detail =======
2025-06-08 11:09:06,776 [INFO] {
    "batch_size": 2,
    "build_info": {
        "ann_path": "/mnt/data4/home/ziyuan/MiniGPT4-Finetuning/flickr30k/flickr30k_grounded_detail.json",
        "image_path": "/mnt/data4/home/ziyuan/MiniGPT4-Finetuning/flickr30k/flickr30k-images"
    },
    "data_type": "images",
    "text_processor": {
        "train": {
            "name": "blip_caption"
        }
    },
    "vis_processor": {
        "train": {
            "image_size": 224,
            "name": "blip2_image_train"
        }
    }
}
2025-06-08 11:09:06,776 [INFO]
```

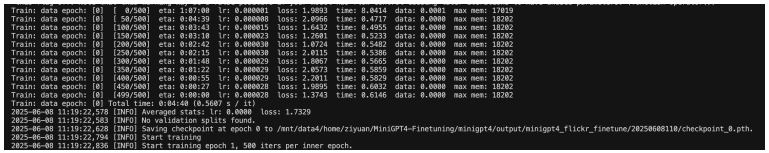
```
======  Model Attributes  ======
2025-06-08 11:09:06,777 [INFO] {
    "arch": "minigpt4",
    "ckpt": "/mnt/data4/home/ziyuan/MiniGPT4-Finetuning/prerained_minigpt4_7b.pth",
    "drop_path_rate": 0,
    "end_sym": "###",
    "freeze_qformer": true,
    "freeze_vit": true,
    "image_size": 224,
    "llama_model": "/mnt/data4/home/ziyuan/MiniGPT4-Finetuning/vicuna-7b",
    "max_txt_len": 300,
    "model_type": "pretrain_vicuna0",
    "num_query_token": 32,
    "prompt": "",
    "prompt_path": "prompts/alignment.txt",
    "prompt_template": "###Human: {} ###Assistant: ",
    "use_grad_checkpoint": false,
    "vit_precision": "fp16"
}
2025-06-08 11:09:06,777 [INFO] Building datasets...
2025-06-08 11:09:06,899 [INFO] Loading LLAMA
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:25<00:00, 12.62s/it]
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:25<00:00, 12.60s/it]
2025-06-08 11:11:48,189 [INFO] Loading LLAMA Done
2025-06-08 11:11:48,189 [INFO] Loading VIT
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:25<00:00, 12.55s/it]
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:26<00:00, 13.30s/it]
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:27<00:00, 13.67s/it]
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:27<00:00, 13.99s/it]
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2/2 [
00:28<00:00, 14.05s/it]
2025-06-08 11:12:43,597 [INFO] freeze vision encoder
2025-06-08 11:12:43,597 [INFO] Loading VIT Done
Loading Q-Former
2025-06-08 11:12:53,080 [INFO] freeze Qformer
2025-06-08 11:12:54,424 [INFO] load checkpoint from https://storage.googleapis.com/sfr-vision-language-research/LAVIS/models/BLIP2/blip2_pretrained_flant5xxl.pth
Loading Q-Former Done
Load 4 training prompts
Prompt Example
###Human: <Img><ImageHere></Img> Describe this image in detail. ###Assistant:
Load MiniGPT-4 Checkpoint: /mnt/data4/home/ziyuan/MiniGPT4-Finetuning/prerained_minigpt4_7b.pth
```

```
2025-06-08 11:14:38,584 [INFO] Start training
2025-06-08 11:14:42,182 [INFO] dataset_ratios not specified, datasets will be concatenated (map-style datasets) or chained (webdataset.DataPipeline).
2025-06-08 11:14:42,184 [INFO] Loaded 31014 records for train split from the dataset.
batch sizes [[2]]
module.module.llama_proj.weight
module.module.llama_proj.bias
2025-06-08 11:14:42,215 [INFO] number of trainable parameters: 3149824
2025-06-08 11:14:42,218 [INFO] Start training epoch 0, 500 iters per inner epoch.
```

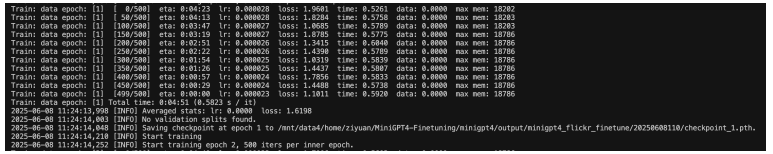
3.1 训练过程展示

部分训练结果如图所示：

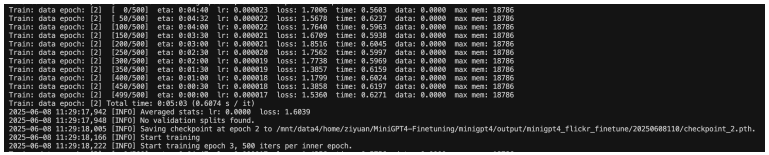
epoch0:



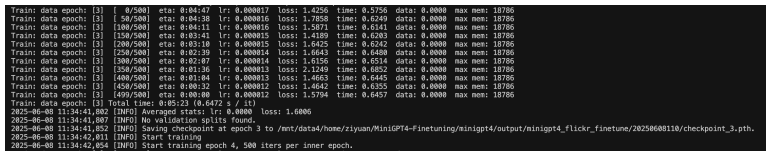
epoch1:



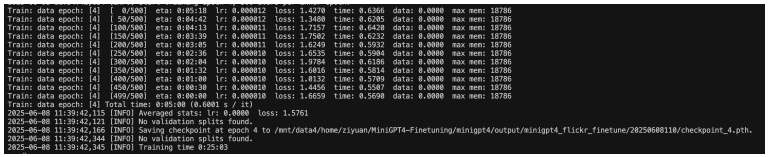
epoch2:



epoch3:



epoch4:



3.2 评估结果与分析

Table 1: 模型评估结果统计

	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**

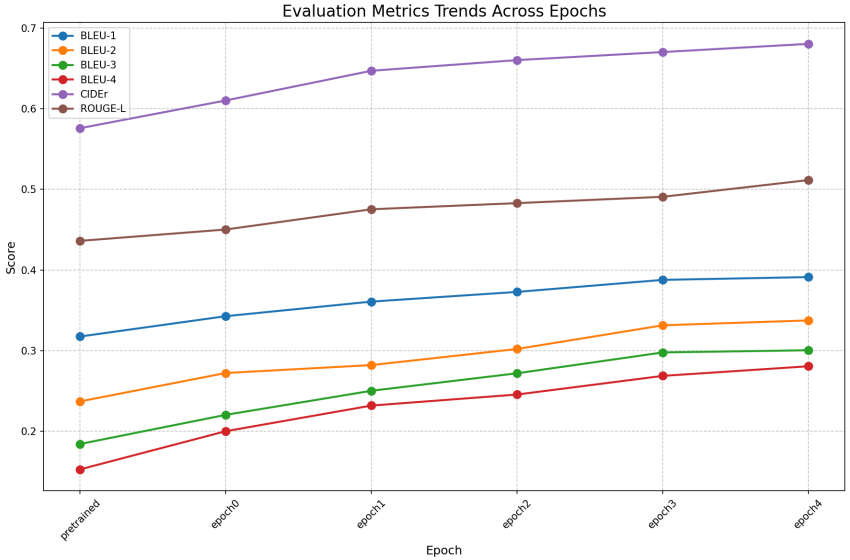


Figure 1: 模型评估结果折线图

3.2 评估结果与分析

3.2.1 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。

3.2 评估结果与分析

3.2.2 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。

3.2 评估结果与分析

3.2.3 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。

3.2 评估结果与分析

Thank you for your attention!