

每周科研进度汇报

方向：多模态大模型与视觉对齐

张三 (Your Name)

人工智能与多模态实验室

2025 年 11 月 19 日

本次汇报目录

① 核心摘要

② 文献阅读

③ 方法与改进

④ 实验结果

⑤ 问题与计划

核心摘要 (Executive Summary)

- 本周重点 (Key Achievements):

- 复现了 LLaVA 的多模态对齐模块。
- 调试通了 Flickr30k 数据集的预处理代码。

- 关键数据 (Highlight):

- 在小样本测试中，图文匹配准确率提升了 **2.1%**。
- 训练显存占用优化，从 24GB 降至 16GB。

- 当前状态: **正常推进 (On Track)**

文献阅读: [论文标题]

发表于: CVPR 2024 | 机构: OpenAI/Google

研究动机 (Motivation):

- 现有的 VLM 在处理细粒度空间关系时表现不佳。
- 提出了一种新的 **Region-Aware Attention** 机制。

核心方法 (Method):

- 引入 Bounding Box 作为 Prompt。
- 损失函数设计:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{iou}$$

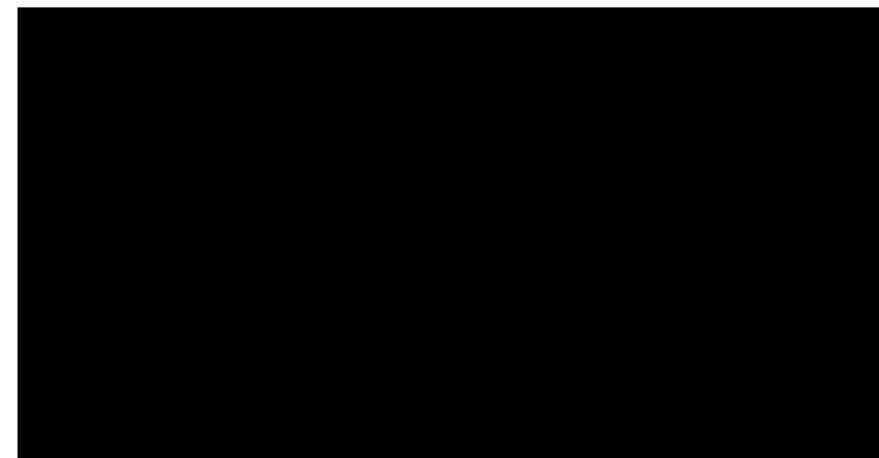


图: 论文提出的模型架构图

方法改进：跨模态注意力机制

1. 数学定义：为了增强图像区域 x 和文本 Token t 的交互，计算如下相似度：

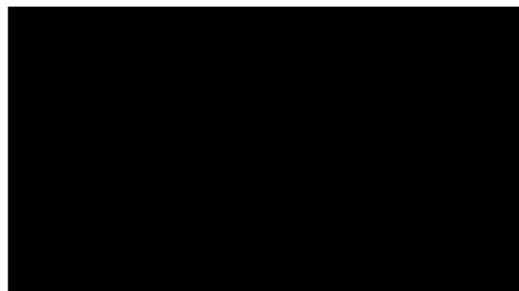
$$\text{Attention}(Q, K, V) = \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

2. PyTorch 实现片段：

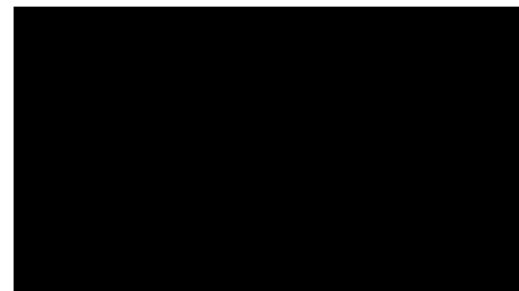
```
1 # 定义 Cross-Attention 层
2 class CrossAttention(nn.Module):
3     def forward(self, img_feat, text_feat):
4         # img_feat: [Batch, Img_Seq, Dim]
5         # text_feat: [Batch, Txt_Seq, Dim]
6
7         q = self.query(text_feat)
8         k = self.key(img_feat)
9
10        attn_output, _ = self.mha(q, k, k)
11        return attn_output
12
```

实验结果：定性分析 (Qualitative)

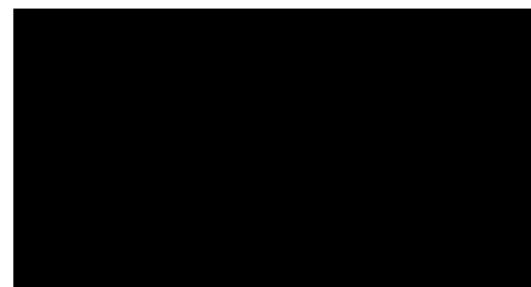
测试 Prompt: “一只戴着墨镜的赛博朋克风格猫咪”



(a) Baseline



(b) Ours (改进版)



(c) Ground Truth

当前问题与下周计划

遇到的困难 (Issues)

- ① OOM 报错：当 Batch Size 大于 32 时显存溢出。
- ② 数据脏乱：发现 LAION 数据集中有部分图片链接失效，导致 DataLoader 卡死。

下周计划 (Next Steps)

- ① 尝试使用 Gradient Checkpointing 技术节省显存。
- ② 增加异常捕获代码，跳过失效图片。
- ③ 跑完 COCO 数据集的对比实验。

Q & A

感谢聆听，请老师指导！