每周进度汇报

杨溟天

2025 年 11 月 20 日

本次汇报目录

1. 核心摘要
2. CLIP 微调
3. IRRA
4. UFineBench
5. 计划

核心摘要 (Executive Summary)

- **本周重点 (Key Achievements):**
  - 3 篇模型微调的论文，均出自一个作者
  - IRRA:2023 年的 SOTA
  - UFineBench: 一个新的数据集，同时也是第一个使用 LLM 来做描述增强的数据集

CSKT：CLIP-based Synergistic Knowledge Transfer for Text-based Person Retrieval

发表于: ICASSP 2024

研究动机 (Motivation):

- 第一篇对 CLIP 在行人检索任务上进行微调的论文。

核心方法 (Method):

- 1.Adapter 对 encoder 的 MLP 进行微调。
- 2.Prompt Tuning 对 transformer 各层的 Prompt 做调整

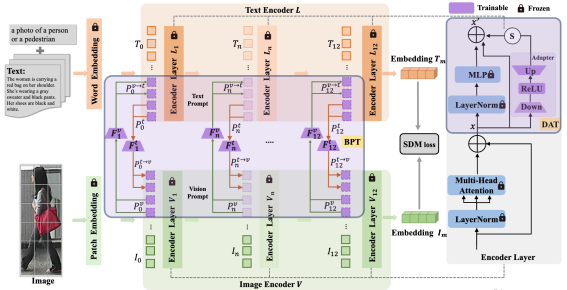


图: 论文提出的模型架构图

具体的微调算法

MLP 微调:

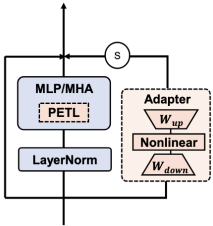


图: MLP-Adapter 的结构图

- 对 attention 后的 MLP 做微调

$$x_{out} = x + x_{MLP} + s \cdot \mathrm{ReLU}\left(\mathrm{LN}(x) \cdot W_d\right) \cdot W_u$$

Prompt Tuning:

- 对 transformer 各层的 Prompt 做调整

$$[P_0^t,\ P_0^{v\to t},\ T_0]$$

- P_0^t 是初始的文本 token，对于下一层来说，是前一层的输出。
- $P_0^{v\to t}$ 是初始的视觉 token 经过 projection 后的，
- T_0 是可学习的文本 token。
- 上述三个向量拼接得倒最终送入当前 layer 做注意力计算的输入

DM-Adapter: Domain-Aware Mixture-of-Adapters for Text-Based Person Retrieval

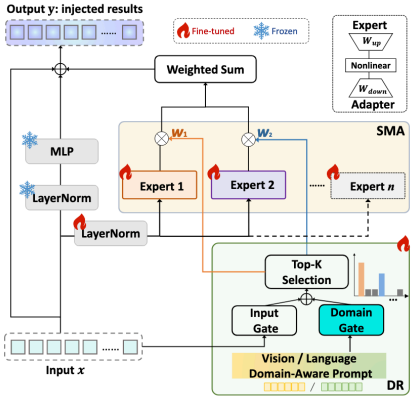
发表于: AAAI 2025

研究动机 (Motivation):

- 引入 LLM 中的 MOE 结构，在推理时选择最合适的两个 Adapter 进行激活

核心方法 (Method):

- 引入 MOE 架构，实现 Adapter 的动态选择。
- 引入 Load-Balance Loss，确保每个 Adapter 都得到充分训练。



具体的 MOE 算法

- **MOE 激活**：

$$y = h_o + \sum_{i=0}^{n-1} \mathrm{Softmax}\left(\mathrm{TopK}\left(xW + pW_d\right)_i\right) \cdot \mathrm{Adapter}_i(x)$$

- h_o 是原始的输出。
- x 是输入。
- W 是权重，用于计算每个专家的激活值。
- p 是一个可学习的参数，文中作者称之为“Domain-Aware Prompt”。
- W_d 是门控权重。
- TopK 用于取激活值最高的专家参与运算。

均衡 MOE 负载

- 在 MOE 中，我们存在负载不均衡的问题：某些专家大量被激活导致失去 MOE 的 sparse 特点，退化为 dense 结构。
- **解决办法：Load-Balance Loss**

$$l_{aux} = \alpha \sum_{i=1}^{n} f_i \cdot p_i$$

- f_i 是第 i 个专家处理的 token 比例。
- p_i 是第 i 个专家的平均参数。

UP-Person: Unified Parameter-Efficient Transfer Learning for Text-based Person Retrieval

发表于: CAST 2025

研究动机 (Motivation):

- 引入 Lora 到 CLIP 微调中
- 对 attention 的 K、V 进行 prefix 填充
- 改进 Adapter 以均衡分布

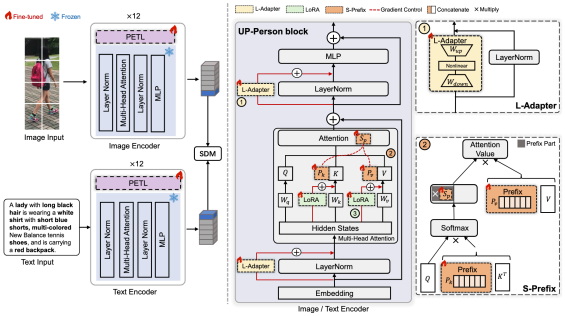


图: 论文提出的模型架构图

Lora 对 Attention 权重进行微调

- **Lora 原理**

$$W_k = W_k + \alpha \cdot W_{\text{loraA}} W_{\text{loraB}}$$

- 其中 W_k 是用于计算 K 的 projection 矩阵
- 由于 W_{loraA} 和 W_{loraB} 的秩远远小于 W_k，我们可以实现高效的微调

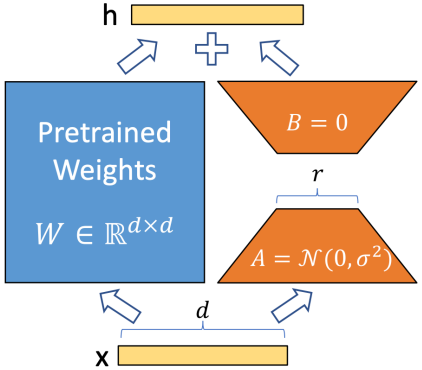


图: Lora 原理

改进 Adapter

研究动机 (Motivation):

- 之前：并行 Adapter:

$$output = x + MLP(x) + Adapter(x)$$

- 现在：合并到输入

$$output = x + MLP(x + Adapter(x))$$

- **为什么要这么做**
- 作者认为传统 Adapter 位于 Attention 层容易与 LoRA/Prefix 产生结构重叠，将其移至 LayerNorm 层实现了空间解耦，有效消除了组件间的优化干扰。

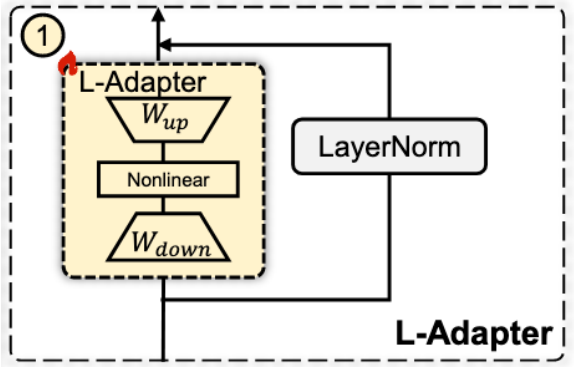


图: 论文提出的模型架构图

S-Prefix

研究动机 (Motivation):

- 在文本和视觉 Transformer 的每一层的 Key 和 Value 中引入可学习的前缀 Token。
- K $= [P_k : \mathrm{K}]$ P_K 是 prefix，K 是原始的 Key，二者拼接。
- V $= [P_v : \mathrm{V}]$ P_V 是 prefix，V 是原始的 Value，二者拼接。
- 我个人的理解是注入行人检索特定的先验知识

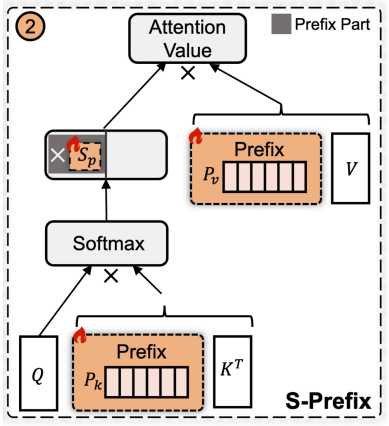


图: 论文提出的模型架构图

Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval

发表于: CVPR 2023

核心贡献:

- 引入了 cross attention，实现图像区域和文本 token 的细粒度交互
- 引入了 Masked Langua Model(MLM), 随机 MASK token 后做预测，提升模型的理解能力
- 引入了 Similarity Distribution Match(SDM)，改进了 InfoNCE loss

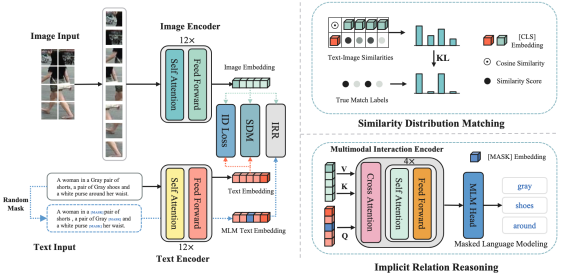


图: 论文提出的模型架构图

Cross Attention

- 把经过视觉 encoder 的图像区域特征 f_{vision} 和文本 encoder 的文本 Token 特征 f_{text} 进行 cross attention 计算
- 文本特征作为 Query，图像区域特征作为 Key 和 Value
- 计算得到的注意力特征 f_{mask} 送入到下面会提到的 Masked Language Model

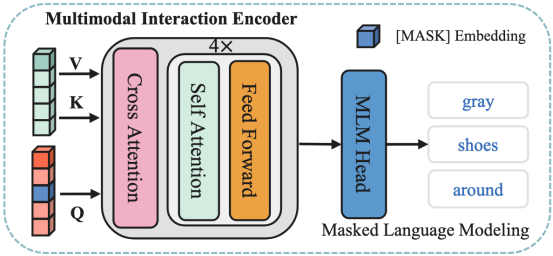


图: Cross Attention

Masked Language Model

- 我们会以 15% 的概率随机 MASK 掉文本 Token，然后利用 cross attention 计算得到的 f_{mask} 来预测被 MASK 掉的 Token
- 这样可以提升模型对文本的理解能力，理解图像的含义，作者称之为 Implicit Relation Reasoning (IRR)
- 作者提出 IRR loss：

$$\mathcal{L}_{irr} = -\frac{1}{|\mathcal{M}|\,|\mathcal{V}|} \sum_{i\in\mathcal{M}} \sum_{j\in|\mathcal{V}|} y_j^i \log \frac{\exp(m_j^i)}{\sum_{k=1}^{|\mathcal{V}|} \exp(m_k^i)}.$$

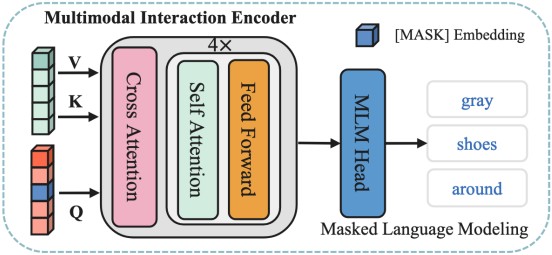


图: Cross Attention

Similarituy Distribution Matching

- 传统的 InfoNCE loss 只关注正样本和负样本。在 CLIP 预训练中图文是一一对应的，但行人检索里一个图像对应多个文本描述，InfoNCE loss 无法刻画这种一对多关系。
- 我们把相似度转换为概率分布，用 KL 散度拉近模型给出的分布与真实分布：

$$KL(\mathbf{p_i} \parallel \mathbf{q_i}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{i,j} \log \left(\frac{p_{i,j}}{q_{i,j} + \epsilon} \right)$$

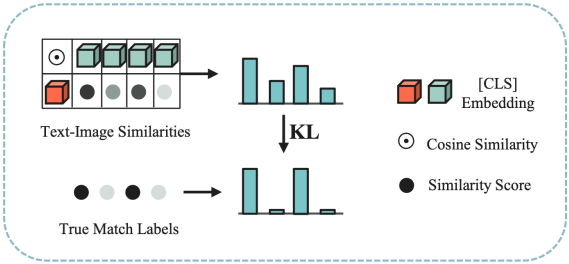


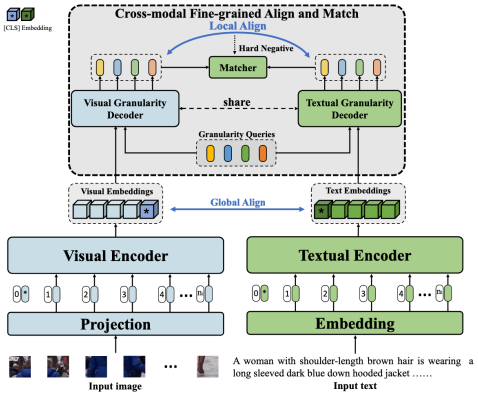
图: Cross Attention

UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity

发表于: CVPR 2024

核心贡献:

- 提出了一个全新的行人检索训练集，提供更加细粒度、描述更加丰富的文本描述
- 提出了一个综合性的评测基准，首次使用 LLM 来对数据集描述做改写增强
- 提出了一种新的评估指标, 优于 mAP 、rank@k
- 提出一种基于 decoder 的细粒度对齐方法



新的数据集:UFine6926

- 人工标注的全新数据集，平均每段描述用了 80 个词，远高于 CUHK-PEDS(27 个)
- 使用对比实验证明了在这个数据集上训练的模型 rank@k, mAP 等指标均有显著提升

Training Sets	CFAM					IRRA [12]					PLIP [55]				
	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD	R@1	R@5	R@10	mAP	mSD
CUHK-PEDES	53.80	71.05	78.25	50.40	38.26	50.06	67.98	75.46	47.57	36.50	40.45	57.51	65.20	38.94	30.82
ICFG-PEDES	36.79	54.64	62.93	34.21	25.47	30.57	47.61	55.87	28.38	21.24	34.32	50.52	57.94	32.59	24.88
RSTPReid	29.85	49.08	58.54	29.66	21.82	21.62	39.53	49.38	21.90	16.09	25.25	40.70	48.30	24.62	18.18
UFine6926	**62.84**	**77.82**	**83.23**	**59.31**	**46.04**	**56.34**	**72.17**	**78.47**	**54.24**	**42.92**	**64.59**	**80.16**	**85.63**	**60.43**	**47.76**

图: UFine6926 数据集效果

新的验证集:UFine3C

- 提供了不同分辨率、亮度、背景、场景的人物图片
- 借助 LLM(Qwen llama) 来对描述进行改写，提供不同风格，不同粒度的文本描述

图: UFine6926 数据集效果

新的评价指标：mSD（mean similarity distribution）

- 右图所示的情况，三组相似度不同，但是 rank@k,mAP 得出的结果不同
- 说明 mAP 与 rank@k 不能很好的捕捉到正负样本之间的相对值大小带来的差异
- 即 [100(正) 99(负)] 和 [100(正) 0（负）] 显然后者是更好的相似度

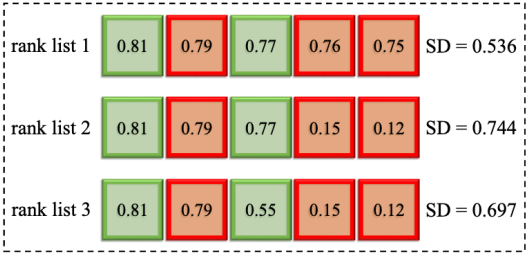


图: UFine6926 数据集效果

使用 decoder 实现的细粒度对齐

- 人们常常使用「CLS」token 放在图片 patch 的首部，取这个 token 的 hidden 来作为我们的图像信息
- 相似的：「EOS」token 放在文本末尾，作为文本信息的总结。
- 但是在计算相似度的过程中，其余的 token 的信息完全不会被用到

使用 decoder 实现的细粒度对齐

- 训练公共的 decoder 和一个公共的 Query 向量，来做完整的图文 embedding 对齐
- 即：把图文的完整 embedding，分别和这个公共 Query 送入 decoder，得倒新的图、文特征表达
- 用这个新的图、文特征来做对齐

ing

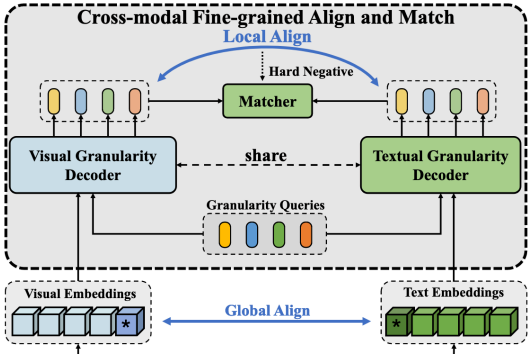


图: UFine6926 数据集效果

下周计划

下周计划 (Next Steps)

1. 继续阅读行人检索领域的论文
2. 阅读这周所看部分论文作者开源的代码，尝试理解其中的代码实现

Q & A

感谢聆听，请老师指导！