ISCAS 中国科学院软件研究所学术年会 2025 第十届
暨重点实验室科技活动周

学术论文

面向多模态大模型领域适应失配问题的因果前门校正方法

Rethinking Misalignment in Vision-Language Model Adaptation from a Causal Perspective

张雅楠*，李江梦*，刘立祥，强文文

2024 Conference and Workshop on Neural Information Processing Systems

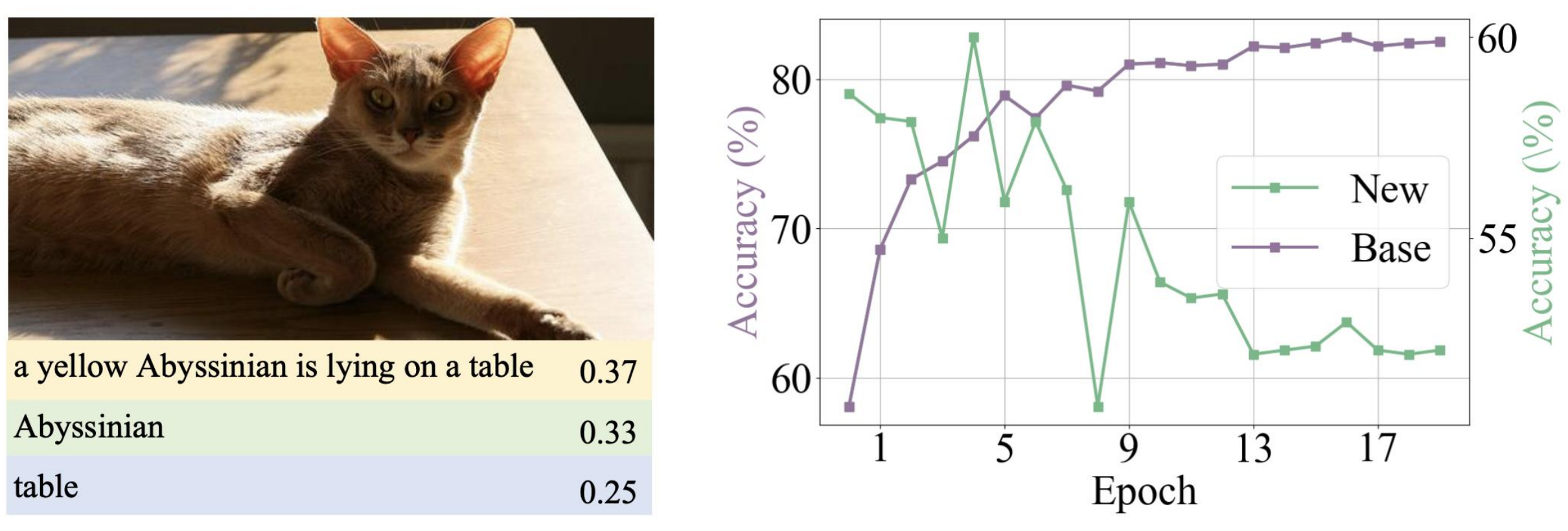
NeurIPS, **CCF-A**

李江梦，13121650625，jiangmeng@iscas.ac.cn

启发式探索

在CLIP与下游任务之间存在**两层次的失配问题**：

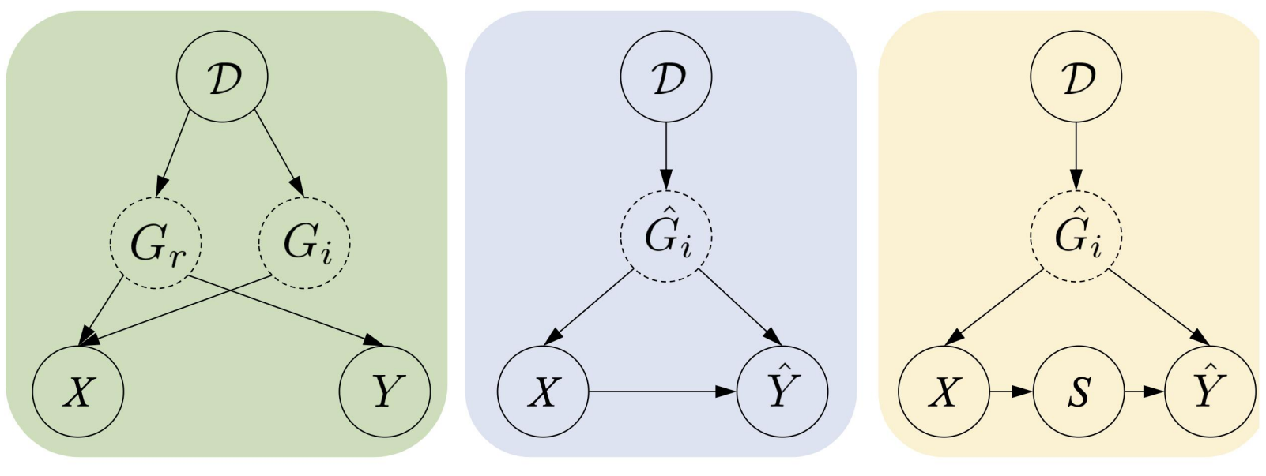
- **任务失配**：任务失配源于CLIP预训练目标与下游任务目标之间的差异。为缓解任务失配问题，引入了软提示（soft prompt）微调方法。
- **数据失配**：在软提示微调过程中，训练数据与测试数据之间存在不一致性。提示微调可能导致CLIP模型对基础类别（base classes）的过拟合现象。



(a) (b)

图示：(a) 一个关于任务失配的启发性示例，展示了在CLIP嵌入空间中，图像与各种文本描述之间的余弦相似度；(b) 一个关于数据失配的启发性实验，显示了在DTD数据集上，不同训练轮次中基础类别与新类别的准确率变化趋势。

问题分析



(a) (b) (c)

图示：结构因果模型（SCMs）。(a) 表示变量X和Y的真实数据生成过程；(b) 用于预测X标签的结构因果模型；(c) 展示前门调整方法的结构因果模型。

变量解释：

- $\mathcal{D}$：预训练阶段数据
- G_r/G_i：任务相关/任务无关生成因子集合
- X：输入数据
- $Y/\hat{Y}$：输入数据X的真实/预测类别
- $\hat{G}_i$：错误保留的任务无关生成因子集合

- X 和 Y 之间存在一个后门路径 $X \leftarrow \hat{G}_i \rightarrow Y$。
- **混淆因子**：$\hat{G}_i$
- **前门校正**：

$$P(\hat{Y}=y|do(x))=\sum_{s}P(s|x)\sum_{x'}P(y|x',s)P(x')$$

- **实现**：
  - ✓ **多模态双重语义解耦**：获得语义信息 s；
  - ✓ **解耦语义可信分类**：基于前门校正，估计样本 x 属于类别 y 的概率。

实验结果

我们所提出的方法在多个基准数据集上都展现出了强泛化性和强判别性。

Dataset	CoOp [4]			CoCoOp [5]			MaPLe [6]			CDC			
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM	Δ
Avg	82.69	63.22	71.66	80.47	71.69	75.83	82.28	75.14	78.55	**83.34**	**77.38**	**80.25**	+1.70
ImageNet	76.47	67.88	71.92	75.98	70.43	73.10	76.66	70.54	73.47	**77.50**	**71.73**	**74.51**	+1.04
Caltech	98.00	89.91	93.73	97.96	93.81	95.84	97.74	94.36	96.02	**98.20**	**94.37**	**96.25**	+0.23
Pets	93.67	95.29	94.47	95.20	97.69	96.43	95.43	97.76	96.58	**96.07**	**98.00**	**97.02**	+0.44
Cars	**78.12**	60.40	68.13	70.49	73.59	72.01	72.94	**74.00**	73.47	73.80	73.97	**73.88**	+0.41
Flowers	**97.60**	59.67	74.06	94.87	71.75	81.71	95.92	72.46	82.56	96.93	**75.07**	**84.61**	+2.05
Food	88.33	82.26	85.19	90.70	91.29	90.99	90.71	92.05	91.38	**90.87**	**92.33**	**91.59**	+0.21
Aircraft	**40.44**	22.30	28.75	33.41	23.71	27.74	37.44	35.61	36.50	37.47	**37.50**	**37.48**	+0.98
SUN	80.60	65.89	72.51	79.74	76.86	78.27	80.82	78.70	79.75	**82.37**	**80.03**	**81.18**	+1.43
DTD	79.44	41.18	54.24	77.01	56.00	64.85	80.36	59.18	68.16	**82.70**	**64.10**	**72.22**	+4.06
SAT	92.19	54.74	68.90	87.49	60.04	71.21	94.07	73.23	82.35	**95.10**	**82.33**	**88.26**	+5.91
UCF	84.69	56.05	67.46	82.33	73.45	77.64	83.00	78.66	80.77	**85.70**	**81.73**	**83.67**	+2.90

	Source	**Target**										
	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	SAT	UCF	Avg
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	**24.74**	67.01	46.49	48.06	**68.69**	66.30
CDC	**71.76**	**94.47**	**90.77**	**66.27**	**72.67**	**86.27**	24.50	**68.07**	**46.60**	**49.13**	68.60	**66.73**

	Source	**Target**				
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	**50.90**	76.98	60.28
CDC	**71.76**	**64.87**	**50.33**	50.40	**78.10**	**60.93**

因果引导语义解耦分类

多模态双重语义解耦

鉴于多模态大模型CLIP具有多样的提示模板，我们用 $t^m \in \{t^1, t^2, \ldots, t^M\}$ 来代表相应得到的不同语义集合。

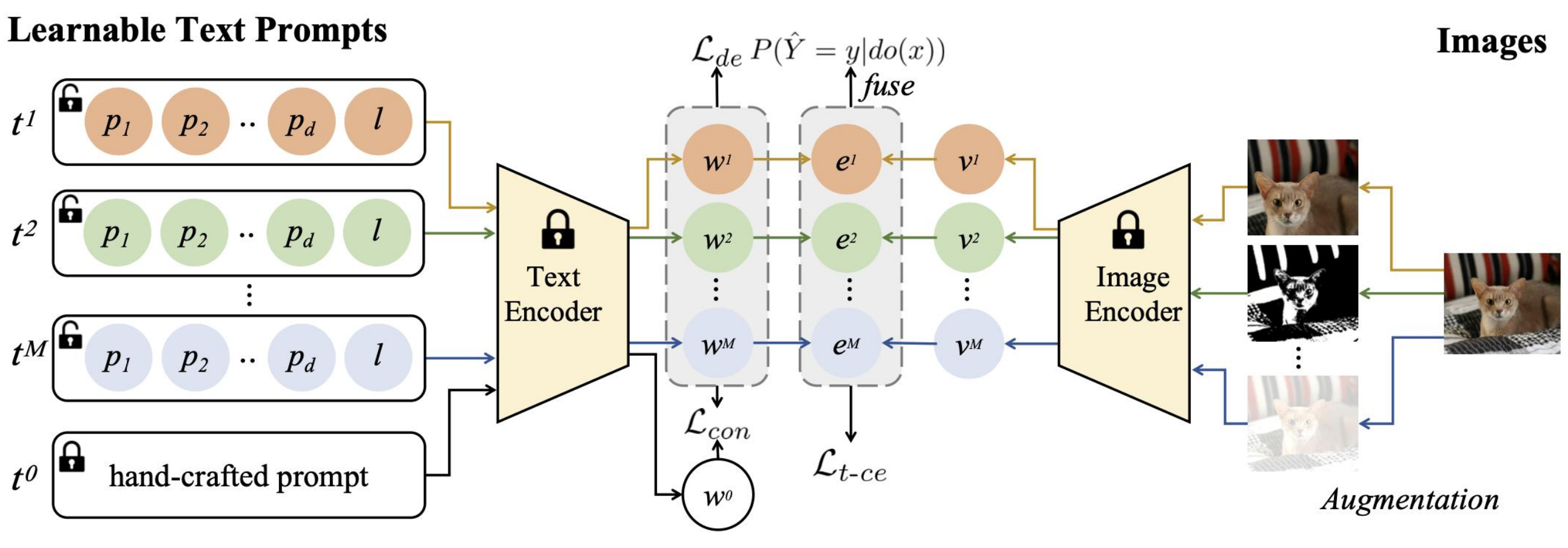
- **视觉**：利用多样化的数据增强方法生成图像输入，并为每个模板采用对应的增强方法。
- **语言**：
  - ✓ 旨在最大化不同模板对应嵌入的多样性。

$$\mathcal{L}_{de}=\frac{1}{M-1}\frac{1}{C}\sum_{m'=1,m'\neq m}^{M}\sum_{c=1}^{C}\sum_{\bar{c}=1}^{C}P(\bar{c}|w_c^m,w^{m'})\log P(\bar{c}|w_c^m,w^{m'})$$

  - ✓ 帮助每个模板捕获与手工设计模板相近的信息。

$$\mathcal{L}_{con}=-\frac{1}{C}\sum_{m=1}^{M}\sum_{c=1}^{C}\log P(c|w_c^m,w^0)$$

方法框架图



图示：**CDC框架**。t^m 表示单个模板，$p_1, p_2, \ldots, p_d$ 表示模板中的各个标记（token）。不同颜色表示不同的模板。“fuse”指的是从多个模板结果生成最终分类结果的过程。文本编码器和图像编码器被冻结，只有提示模板中的标记是可学习的。

总体损失函数：$\mathcal{L}_{CDC}=\mathcal{L}_{t\text{-}ce}+\beta\mathcal{L}_{de}+\gamma\mathcal{L}_{con}$

解耦语义可信分类

- 设存在一个 *Dirichlet* 分布 $a^m=[a_1^m, a_2^m, \ldots, a_C^m]$，其中 $a_c^m=e_c^m+1$，$e_c^m=\mathrm{h}(sim(w_c^m, v))$。
- 基于所有提示模板所得到的解耦语义信息，能够基于下列公式进行解耦语义信息融合：

$$B_c^m=\begin{cases} b_c^1, & \text{if } m=1 \\ \frac{1}{1-C}(B_c^{m-1}b_c^m+B_c^{m-1}u^m+b_c^mU^{m-1}), & \text{if } 1<m\leq M \end{cases},$$

$$U^m=\begin{cases} u^1, & \text{if } m=1 \\ \frac{1}{1-C}U^{m-1}u^m, & \text{if } 1<m\leq M \end{cases}.$$

- 推理时，基于前门校正的估计结果可以形式化为：

$$P(\hat{Y}=c|do(x))=\frac{B_c^M}{\Sigma_{c'=1}^{C}B_{c'}^M}$$

- 训练时，基于可信交叉熵损失的解耦语义信息融合损失可以形式化为：

$$\mathcal{L}_{t\text{-}ce}=\sum_{m=1}^{M}(\psi(A)-\psi(\alpha_y^m))$$