《人工神经网络》大作业开题报告

张毅*
2022010387
工22
yizhang22@mails.tsinghua.edu.cn

苏博宇
2023011277
物理32
sby23@mails.tsinghua.edu.cn

杨弘毅
2023011638
英31
yang-hy23@mails.tsinghua.edu.cn

1 选题

我们的选题是：

- Jittor论文复现，题目为：
  - Grounding dino: Marrying dino with grounded pre-training for open-set object detection [1]

1.1 任务背景与定义

在本次大作业中，我们计划在jittor框架下对Grounding DINO模型进行复现。这是一种开放集（open-vocabulary）目标检测模型，旨在通过结合DETR 架构与基于图像-文本对的grounded 预训练，实现无需微调即可检测任意文本描述对象的能力。其核心思想是将DINO（DETR with Improved DeNoising anchOr boxes）检测器与对比学习驱动的图文对齐机制相结合[1]。

为了提升模型在开放词汇场景下的泛化能力，作者构建了大规模grounded 图文数据集（如GoldG），并利用COCO, Objects365 等强标注检测数据进行联合预训练。数学层面而言，可以将任务界定为：给定一张输入图像$I \in \mathbb{R}^{H\times W\times 3}$ 和一个文本提示集合$\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$，其中每个$t_i$ 表示一个自然语言短语，Grounding DINO 的目标是输出一组边界框-文本对$\{(b_j, s_j)\}_{j=1}^{M}$，其中$b_j \in [0,1]^4$ 表示归一化的边界框坐标，$s_j \in [0,1]$ 表示该框与对应文本的匹配置信度，并隐式地建立图像区域与文本语义之间的对齐关系。

本任务的形式化定义如下：设图像编码器为$f_I(\cdot)$，文本编码器为$f_T(\cdot)$，跨模态融合模块为$f_{fuse}(\cdot,\cdot)$，解码器为$f_{dec}(\cdot)$，则Grounding DINO 的整体映射可表示为：

$$\{(b_j, s_j)\}_{j=1}^{M} = f_{dec}\big(f_{fuse}(f_I(I), f_T(\mathcal{T}))\big).$$

在相关工作方面，我们的任务将主要基于Grounding DINO论文[1]，同时也将参考Detection Transformers、DETR模型和VLM模型（用于对比）领域的相关工作，具体信息如下：

- **Detection Transformers (DETR)**：Carion 等人[2] 首次将Transformer 引入目标检测任务，提出端到端的检测框架，摒弃了传统anchor 和NMS 后处理，为后续DETR 系列模型奠定基础。
- **DINO**：Zhang 等人[3] 在DETR 基础上引入对比去噪训练（Contrastive DeNoising）和混合查询选择机制，显著提升了收敛速度与检测精度，成为当前最先进的闭集检测器之一。

*请将组长放到第一个位置，推荐一组三位同学。

- **开放词汇检测模型**：如GLIP [4] 将目标检测统一为phrase grounding 任务，通过在大规模图文对上预训练实现开放词汇推理；而Grounding DINO 进一步融合DINO的检测能力与更强的图文对齐策略，在多个zero-shot 基准上取得SOTA 性能。

1.2 数据集

我们复现Grounding DINO所使用的核心数据集如下：

- **Objects365 (O365)** [5]：大规模检测数据集，包含365个类别，用于预训练以扩展模型的类别空间。数据集需转换为ODVG格式（使用官方提供的`coco2odvg.py`脚本）。
- **GoldG**：区域-短语对齐的接地气数据，包含Flickr30k [6]与GQA [7]数据集。用于建立区域-文本对齐能力，需通过`goldg2odvg.py`转换为ODVG格式。
- **LVIS** [8]：长尾分布实例分割数据集，包含1200+类别，用于微调及最终在MiniVal上的评测。
- **COCO** [9]：通用检测基准，部分实验将用于训练过程中的性能验证。

数据处理流程将严格遵循原论文与MMDetection代码库规范：

- 统一使用ODVG数据格式组织标注；
- 图像预处理：短边缩放至800像素，长边不超过1333像素，应用随机水平翻转与ImageNet归一化；
- 文本提示按论文所述构建子句级表征并施加注意力掩码。

1.3 基线结果

我们计划复现原论文中核心的框架可泛化性结果。原论文汇报，实验使用以Swin-T为主干网络的Grounding Dino T模型，在O365和GoldG数据集上进行预训练。训练过程中采用端到端方式联合优化检测与图文对齐目标。得到的模型在zero-shot场景下，在LVIS数据集上评估的平均准确度（AP）为25.6；在LVIS数据集上进行微调后，平均准确度达到52.1，其中分类别（稀有、常见、频繁类别）平均准确度分别为35.4、51.3和55.7。

我们注意到该项目有官方代码公开[2]与第三方实现[3]。我们计划以此为基础开展论文复现工作。

2 研究计划

我们计划复现原文核心的成果。重点包括：

- 在Jittor 框架下完整实现Grounding DINO-T（以Swin-T为主干）的模型架构，涵盖图像/文本编码器、跨模态特征融合模块及DINO 风格的检测头；
- 复现论文中zero-shot 与微调两种设置下在LVIS 数据集上的评估流程，并报告AP指标；
- 严格对齐原论文的预训练策略（如使用O365 + GoldG 联合训练），若计算资源受限，则优先复现微调阶段并在COCO/LVIS 上验证收敛性。

此外，我们将从以下额外工作中选择1–2 项深入展开，以体现工作量与创新性：

- 与主流VLM模型进行定性与定量对比：在相同输入图像与文本提示下，将Grounding DINO 与CLIP-based detector（如GLIP 或OWL-ViT）的检测结果进行可视化对比，并从模型设计目标（检测专用vs. 通用表征）解释性能差异；
- 展开更多的消融实验：以探究Grounding Dino 框架起效的核心机制（如：可变注意力机制、特征增强等）；

[2]链接：https://github.com/IDEA-Research/GroundingDINO
[3]链接: https://github.com/longzw1997/Open-GroundingDino

- 使用嵌入相似度和语义重合度等量化指标来量化与VLM模型的性能区别，并控制变量（如统一使用Swin-T 图像主干、相同预训练数据子集）以公平比较Grounding DINO 与VLM-based 方法的grounding 能力

2.1 挑战

1. **模型结构与算子复现。** Grounding DINO 包含多尺度可变形注意力、特征增强、语言引导query 等复杂模块，而Jittor 中缺乏完全对等的算子，需要手写高效CUDA/Kernel 并确保梯度正确性与数值对齐；同时需完成PyTorch 向Jittor 的权重格式转换和精度验证。

2. **多模态与VLM 对齐的复杂度。** 在Jittor 中集成BERT / CLIP 等文本或多模态编码器，需要自行封装tokenizer、位置编码与checkpoint 加载逻辑，保证与原论文严格一致。与主流VLM 对比时，还需在不同框架间对齐图像骨干、预训练数据与prompt 设计，带来额外的接口与数据格式适配成本。

3. **数据与评估管线的工程成本。** LVIS 的长尾标注复杂，需要在Jittor 中重写数据加载、采样策略与评估脚本（含频段AP、类别映射等）。任一实现细节偏差都会导致复现指标波动。新增的嵌入相似度、语义重合度等指标亦需在Jittor/NumPy 环境下单独实现和验证。

4. **大模型训练与调参不确定性。** Jittor 的动态图/静态图与内存复用机制与PyTorch 不同，可能在混合精度、多卡并行下产生新的数值稳定性或显存峰值问题，需要额外时间排查、调参和修正，从而增加整体实现难度。

5. **多数据集融合与格式对齐**：需将O365、GoldG、LVIS等多种异构数据集（检测、定位、实例分割）统一至ODVG格式，涉及复杂的标注映射与跨数据集ID协调。

6. **多模态数据加载与子句级文本处理**：需实现能够同时加载检测标注与区域-短语对的数据管道，并正确实现论文关键的"子句级注意力掩码"机制，阻断无关类别间的注意力传播。

7. **大规模数据存储与预处理**：完整数据集需数百GB存储空间，下载、转换及缓存流程在有限带宽与磁盘IO下可能成为瓶颈。

2.2 可行性

可行性方面，本项目整体难度偏高，但在课程算力支持下仍具可行性。

方法与数据上，Grounding DINO 具有较成熟的开源实现、结构模块化清晰；COCO、Objects365 与LVIS 数据集均公开可获得，处理和评估脚本亦可参考官方实现改写，因此不存在数据或指标获取层面的硬性障碍。

计算资源方面，完整复现与扩展实验预计需要数百GPU·h，仅依赖个人设备训练周期过长，难以保证调参与失败重跑空间。因此我们需要课程提供的多卡4090 支持。

组员方面，我们均非计算机相关专业，当前项目经验有限，本学期时间亦较紧张，不过我们会尽力完成任务。我们将压缩目标优先级：优先完成在Jittor 上的主线复现与关键消融，VLM对比等扩展任务视算力与时间情况选择性完成，以在有限精力下确保项目整体可行。

3 算力估计

本项目在Jittor 中重构并微调Grounding DINO-T，考虑到实现难度与调参开销，我们计划申请**4张RTX4090GPU**。

具体的算力估计为：主干LVIS 微调4 轮，每轮控制在12h 内（共约50 GPU·h）；9-12 组消融实验，每组8h（合计约96 GPU·h）；与VLM 的对齐与对比实验（含少量轻量微调及大规模推理）约60–70 GPU·h；其余60–80 GPU·h 用于Jittor 侧模型重构、权重转换验证、超参搜索及失败实验重跑。

我们并不具备其他计算资源。

参考文献

References

[1] Liu, S., Z. Zeng, T. Ren, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[2] Carion, N., F. Massa, G. Synnaeve, et al. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.

[3] Zhang, H., F. Liu, S. Li, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[4] Li, X., X. Yin, C. Li, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965. IEEE, 2022.

[5] Shao, S., Z. Li, T. Zhang, et al. Objects365: A large-scale, high-quality dataset for object detection. *ICCV*, 2019.

[6] Plummer, B. A., L. Wang, C. M. Cervantes, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*. 2015.

[7] Hudson, D. A., C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 2019.

[8] Gupta, A., P. Dollar, R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*. 2019.

[9] Lin, T.-Y., M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context. *ECCV*, 2014.