Analyzing and Improving OWL-ViT for Open-Vocabulary Object Detection

Pratyush Jena
International Institute of Information Technology
Hyderabad, India
Pratyush.jena@research.iiit.ac.in

Akshat Shah
International Institute of Information Technology
Hyderabad, India
akshat.shah@research.iiit.ac.in

I. Introduction

Open-vocabulary object detection (OVOD) has emerged as a critical area in computer vision, enabling models to detect objects beyond a fixed set of predefined categories. OWL-ViT is a state-of-the-art model that leverages vision-language pretraining to achieve open-vocabulary detection; however, its limitations in detecting certain objects and handling complex queries remain underexplored.

This Project aims to conduct a systematic analysis of OWL-ViT to identify cases where the model struggles, determine the underlying reasons for these difficulties, and propose improvements using ViT-Register, a refined transformer-based architecture. Specifically, we seek to:

- Analyze OWL-ViT's failure cases in open-vocabulary detection.
- Investigate difficult image-query pairs, focusing on challenging object attributes such as occlusion, scale, and ambiguity.
- Develop an improved approach leveraging ViT-Register to address these challenges and enhance detection performance.

II. Proposed Methodology

Our methodology follows a structured approach, beginning with failure case identification, followed by model refinement and enhancement, and culminating in benchmarking and evaluation.

A. *Overview of Approach*

We follow an *Input* $\rightarrow$ *Process* $\rightarrow$ *Output* framework:

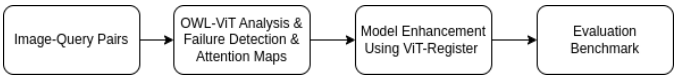

Fig. 1.

B. *Step-Wise Execution*

The project is structured into four primary phases:

*1) **Phase 1: Performance Analysis of OWL-ViT**:*

- Run OWL-ViT on benchmark datasets (LVIS, COCO, and Objects365).
- Collect detection metrics:
  - False positives (FP) and false negatives (FN).
  - Per-category mean Average Precision (mAP).
  - Zero-shot and one-shot detection performance.
- Generate attention maps to examine regions of focus within challenging queries.

*2) **Phase 2: Identification of Hard-to-Detect Queries and Images**:*

- Investigate cases where the model fails to detect objects or misclassifies categories.
- Study factors contributing to poor performance:
  - Occlusion, small object size, and background clutter.
  - Semantic ambiguity in textual descriptions.
  - Misalignment between image-text embeddings.

*3) **Phase 3: Enhancing Performance with ViT-Register**:*

- Fine-tune ViT-Register on detection datasets with a focus on failure cases.
- Implement query representation refinement for more robust text and image embeddings.
- Optimize the object localization pipeline, integrating multi-scale feature processing for improved spatial accuracy.

*4) **Phase 4: Benchmarking & Comparative Evaluation**:*

- Evaluate ViT-Register against OWL-ViT using the same test sets.
- Compare improvements in detection accuracy, robustness, and generalization.
- Conduct a qualitative assessment of failure case reductions through visualization techniques.

III. Ablation Studies

To quantify the impact of different components in both OWL-ViT and ViT-Register, we will conduct a series of ablation studies:

A. *Feature Ablations*

- **Text vs. Image Query Performance**: Compare detection accuracy when using only text-based queries vs. image-based queries for one-shot detection.

- **Impact of Query Length & Prompt Variations**: Test different textual descriptions for the same object (e.g., "a mug" vs. "a white ceramic coffee cup").
- **Attention Map Decomposition**: Measure how much each Transformer layer contributes to object localization.

B. Training Strategy Ablations

- **Fine-Tuning OWL-ViT Without Image-Level Pretraining**: Analyze how much pretraining affects object detection.
- **Regularization & Augmentation Effects**:
  - Evaluate the impact of image augmentations (cropping, scaling, mosaics) on generalization.
  - Compare zero-shot detection performance with vs. without pretraining on large-scale web images.

C. Architecture & Model Component Ablations

- **Replacing ViT with Hybrid CNN-ViT Architectures**: Compare pure Vision Transformers vs. CNN-ViT hybrids for detection.
- **Impact of Larger ViT Models (ViT-B/32 vs. ViT-H/14)**: Evaluate whether model size scaling leads to better generalization.

These ablation experiments will identify the most crucial components for robust OVOD performance and inform future architectural improvements.

IV. Evaluation Metrics

To assess the effectiveness of our approach, we define a comprehensive set of evaluation metrics, covering both quantitative performance indicators and qualitative analysis.

A. Machine Learning Metrics

- **Mean Average Precision (mAP)**: Measures overall detection accuracy.
- **Zero-Shot mAP**: Performance on unseen object categories.
- **One-Shot Image-Conditioned Detection Performance**: Measures the ability to detect objects given a reference image.
- **False Negative Rate (FNR)**: Quantifies how often objects are missed.

B. Qualitative Evaluation

- **Attention Map Visualization**: Examines whether the model correctly focuses on relevant regions.
- **Error Case Analysis**: Identifies recurring patterns in failure cases.

C. Computational Efficiency

- **Inference Latency**: Comparison between OWL-ViT and ViT-Register.
- **Memory and FLOP Usage**: Assesses model scalability.

A successful outcome would be an improvement in detection accuracy on difficult queries, a reduction in false negatives, and minimal additional compute overhead compared to OWL-ViT.

V. Work Allocation Among Team Members

Task	Pratyush Jena	Akshat Shah
Running OWL-ViT on benchmark datasets	✓	
Collecting failure case statistics	✓	✓
Attention map visualization & feature analysis	✓	
Implementing improvements in query representation	✓	
Enhancing object localization with ViT-Register		✓
Fine-tuning ViT-Register		✓
Running final evaluations and benchmarks	✓	✓
Writing documentation and research report	✓	✓

TABLE I
WORK ALLOCATION ARE EQUALLY DIVIDED.

VI. Compute Resource Availability

The primary compute resource for this project is the **ADA computing cluster**, which consists of **four NVIDIA RTX 2080 Ti GPUs (12GB VRAM each, totaling 48GB VRAM)**. This setup is **sufficient to fine-tune all target models**, including **ViT-B/32, ViT-B/16, ViT-L/14, and ViT-H/14**, using appropriate optimizations such as **gradient checkpointing and mixed precision training**.

A. Feasibility of Fine-Tuning on ADA Cluster

Given the total available **48GB VRAM**, we can effectively fine-tune various models with the following configurations:

Model	Fine-Tuning Feasibility	Batch Size Considerations	Optimization Requirements
ViT-B/32	Fully Feasible	Medium (32-64)	No major constraints
ViT-B/16	Fully Feasible	Small-Medium (16-32)	Mixed Precision, Gradient Checkpointing
ViT-L/14	Fully Feasible	Small ($\leq$16)	Gradient Checkpointing, DDP
ViT-H/14	Fully Feasible	Very Small ($\leq$8)	Mixed Precision, Reduced Image Resolution

TABLE II
FEASIBILITY OF FINE-TUNING VARIOUS VIT MODELS ON THE ADA CLUSTER.

B. Optimization Strategies for Efficient Compute Usage

To maximize training efficiency within the **ADA cluster's resources**, we will implement:

- **Gradient Checkpointing**: Reducing memory usage by storing only key activations.
- **Mixed-Precision Training (FP16/BF16)**: Minimizing VRAM consumption while maintaining computational accuracy.
- **Multi-GPU Training**: Using **Distributed Data Parallel (DDP)** to split workloads across GPUs.
- **Adaptive Batch Sizing**: Scaling batch sizes dynamically based on VRAM usage.

References

[1] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple Open-Vocabulary Object Detection with Vision Transformers," *arXiv preprint arXiv:2205.06230*, 2022. [Online]. Available: https://arxiv.org/pdf/2205.06230

[2] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision Transformers Need Registers," *arXiv preprint arXiv:2309.16588*, 2023. [Online]. Available: https://arxiv.org/pdf/2309.16588