

Master Project: Evaluating Robustness of Vision Language Models in Open-set Scenarios

Team: TBD

Supervisor: Ahmed A. Ali, Manuel Günther

1 Introduction

Vision-Language Models (VLMs) like CLIP (Contrastive Language-Image Pre-training) have significantly advanced the field of computer vision by bridging the gap between visual and textual representations. These models are trained on large-scale datasets of image-text pairs, enabling them to associate images with textual descriptions effectively. This capability allows for **zero-shot learning**, where the model can recognize objects it has never seen during training, based solely on their textual descriptions. However, deploying VLMs in practical applications presents challenges, particularly in **open-set recognition** scenarios.

Prompt Engineering Challenges

One critical challenge is prompt engineering — the process of crafting textual inputs (prompts) that guide the model to produce desired outputs. Manual prompt engineering requires domain expertise and can be time-consuming (Zhou et al. , 2022), even the slightest variations in prompt wording can lead to substantial performance fluctuations.



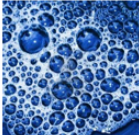

Dataset	Prompt	Accuracy
 Caltch101	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	[V] ₁ [V] ₂ ... [V] _M [CLASS].	91.83
(a)		
 Flowers102	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	[V] ₁ [V] ₂ ... [V] _M [CLASS].	94.51
(b)		
 Describable Textures (DTD)	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	[V] ₁ [V] ₂ ... [V] _M [CLASS].	63.58
(c)		
 EuroSAT	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	[V] ₁ [V] ₂ ... [V] _M [CLASS].	83.53
(d)		

Figure 1: Prompt engineering versus Context Optimization (CoOp). The former requires manual tuning of words, which is inefficient and may not generalize well. CoOp automates the process by learning optimal prompts from data, requiring only a few labeled images for learning. (Adapted from Zhou et al.)

Figure 1 represents traditional prompt engineering, where a human must manually find the best prompt (e.g., "a photo of a [CLASS]"). This process can be inefficient and may not yield the best performance across different tasks or domains. Therefore, **Context Optimization (CoOp)** is developed, a method that automates prompt learning by optimizing context tokens in the prompt. CoOp learns a set of continuous vectors (context tokens) that, when combined with the class name, form an optimal prompt for the model. This approach leverages a small number of labeled examples to fine-tune the prompt, improving both efficiency and performance.

¹ GitHub repository: <https://github.com/kaiyangzhou/coop>

Open-Set Recognition Challenges

Another significant challenge arises in **open-set recognition** — the ability of a model to handle instances from classes that were not seen during training. While VLMs are trained on vast and diverse datasets, they rely on a finite **query set** during inference. This means that even though the model has learned a rich representation of images and text, it can still misclassify unknown objects as one of the known classes with high confidence, as it lacks a mechanism to recognize and reject unknown classes.

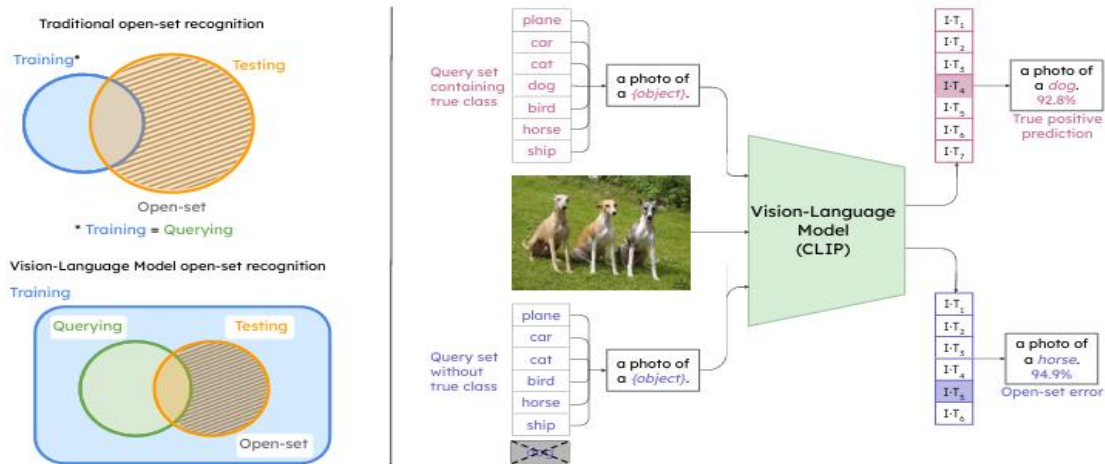


Figure 2: Traditional open-set recognition involves models trained on a finite set of classes and tested on unknown classes, leading to open-set errors. Vision-Language Models, despite being trained on large datasets, use a finite query set during inference. When presented with an object not in the query set, the VLM may incorrectly classify it as one of the known classes with high confidence. (Adapted conceptually from Miller et al.)

In Figure 2, the left side depicts a traditional open-set recognition scenario where models trained on specific classes encounter unknown classes during testing, often resulting in misclassifications (open-set errors). The right side shows that VLMs, during inference, compare image embeddings with text embeddings from a predefined query set of class labels. If an image contains an object not represented in this query set (an unknown class), the VLM has no option but to assign it to the most similar known class, potentially with high confidence. This limitation can lead to significant performance degradation in open-set conditions, as the model cannot recognize when it encounters something truly novel.

To address the prompt engineering challenge, methods like CoOp automate the process by learning optimal prompts directly from data. This reduces the reliance on manual tuning and can improve the model's adaptability to different tasks. For the open-set recognition challenge, recent studies have shown that VLMs impose closed-set assumptions through their finite query sets. They suggest that simply expanding the query set to include more classes does not solve the problem — in fact, it can worsen performance due to increased misclassifications and computational overhead (Miller et al., 2023). To mitigate these issues, approaches such as incorporating **predictive uncertainty measures** and **dedicated negative embeddings** have been proposed. These methods aim to help the model recognize when an input does not belong to any of the known classes and appropriately handle such cases

² GitHub repository: [dimitymiller/openset_vlms \(github.com\)](https://github.com/dimitymiller/openset_vlms)

2 Project

2.1 Project Objectives

The primary objective of this project is to assess the reliability and robustness of Vision-Language Models in open-set recognition scenarios and to improve their performance through automated prompt learning techniques. Specifically, the project aims to implement zero-shot open-set evaluation of VLMs and enhance this evaluation through prompt engineering and learnable prompts, transforming text tokens into learnable vectors. Additionally, it seeks to investigate the limitations of VLMs in open-set recognition, and explore methods to mitigate these challenges (Miller et al., 2023). If resources allow, the project will also explore fine-tuning techniques and the inclusion of dedicated negative embeddings. Students are invited to utilize our currently implemented toolbox for open-set recognition called OpenOSR (access can be provided).

2.2 Project Requirements and Responsibilities

To successfully complete this project, certain technical skills and responsibilities are required. A strong understanding of machine learning concepts, particularly in computer vision and natural language processing, is essential. Proficiency in Python programming and familiarity with the deep learning framework PyTorch (as used in the Deep Learning course) is necessary for implementing and experimenting with models. Experience with dataset curation and preprocessing will aid in managing and preparing the datasets used in this project. An interest in Vision-Language Models and prompt learning techniques is important, along with a willingness to learn if not already familiar. Familiarity with open-set recognition challenges and methods, as discussed in recent literature (Geng et. al, 2021), will be beneficial.

The candidates will be responsible for developing and implementing algorithms for open-set recognition using VLMs, ensuring that the code is of high quality and maintainable. This includes conducting literature reviews on prompt learning and open-set recognition to inform experimental design. The candidates will design experiments to evaluate model performance across closed-set, open-set, and zero-shot scenarios, and analyze the results to identify patterns, strengths, and weaknesses of different approaches. Finally, the candidates will document the findings in a comprehensive report and prepare presentations to communicate the results effectively.

2.3 Datasets

The project will utilize several datasets to evaluate the models:

- **ImageNet-1K²**: Serving as the set of known classes in open-set scenarios, ImageNet-1K provides a large-scale benchmark for evaluating classification models.
- **SUN397³**: This dataset will be used as the set of unknown classes in open-set scenarios, allowing assessment of the model's ability to handle unseen data.
- **iNaturalist Subset⁴**: This dataset contains 9,549 images of plants, insects, and animals. It will also be used as unknown samples to evaluate the models in a zero-shot setting.

²Dataset: [laurent/SUN397 · Datasets at Hugging Face](#)

³Dataset: [imagenet-1k | TensorFlow Datasets](#)

⁴Dataset: <https://huggingface.co/datasets/ba188/inaturalist>

2.4 Models

The project will focus on two Vision-Language Models:

- **CLIP (OpenAI)⁵**: Utilizing the ViT-B/32 architecture, CLIP is pre-trained on a large dataset of image-text pairs. The pre-trained weights from OpenAI's CLIP model will be used.
- **ViT-L/16 (COYO-Labeled-300M)⁶**: This model employs the ViT-L/16 architecture and is pre-trained on the COYO-Labeled-300M dataset, with fine-tuning on ImageNet-1K.

3 Modelling and Experimental Setup

3.1 Model Preparation

For the CLIP model, pre-trained weights provided by OpenAI will be loaded to initialize the model. The ViT-L/16 model will be initialized with weights pre-trained on the COYO-Labeled-300M dataset and fine-tuned on ImageNet-1K. This preparation ensures that both models have strong baseline performance for subsequent evaluations.

3.2 Evaluation Scenarios

The experiments will be conducted under two primary scenarios to comprehensively evaluate the models.

Closed-Set Evaluation

In the closed-set scenario, the models will be tested on the ImageNet-1K dataset. Performance metrics such as Top-1 and Top-5 accuracy will be measured to assess how well the models classify images when all test classes are known, i.e. present during network training, and included in the query set.

Open-Set Evaluation

For open-set evaluation, the models will be tested on images containing both known and unknown classes. Known classes will be from ImageNet-1K, while unknown classes will be from SUN397 and iNaturalist subsets. The evaluation will follow the protocol proposed by Miller et al., focusing on the models' ability to reject unknown classes not included in the query set. Recognition scores will be obtained through fine-tuning and zero-shot methods, with and without prompt learning, to assess the models' ability to generalize to unseen classes. Challenges in this scenario include the models' reliance on a finite query set during inference. When presented with images containing classes not in the query set, the models may misclassify them as known classes with high confidence. Simply expanding the query set to include more classes can degrade performance due to increased misclassifications and computational overhead. Performance metrics will include accuracy on known classes, as well as recent metrics including OpenAUC (Wang et al., 2022) and Operational Open-Set Accuracy (OOSA) (Cruz et al., 2024), all of which require to predict scores for known classes only, and exploit thresholds to determine if an input is from an unknown class.

⁵ Model: <https://huggingface.co/openai/clip-vit-base-patch32>

⁶ Model: <https://huggingface.co/kakaobrain/vit-l16-coyo-labeled-300m-i1k384>

3.3 Prompt Learning and Context Optimization

Building upon Zhou et al.'s work on "Learning to Prompt for Vision-Language Models," the project will implement Context Optimization (CoOp) to automate prompt engineering. CoOp models a prompt's context words with learnable vectors optimized by minimizing classification loss.

Two approaches will be explored:

- **Unified Context:** A shared context across all classes, represented as learnable vectors.
- **Class-Specific Context:** Independent context vectors for each class, allowing the model to tailor prompts to specific classes.

To address these challenges identified by Miller et al., the project will incorporate advanced methods from the literature to generate effective negative samples. Instead of using random words as negatives or Gaussian noise embeddings, which may not provide meaningful distinctions between known and unknown classes, we will utilize techniques like Manifold Mix-up.

Baseline Approaches to Open-Set Recognition:

- **Predictive Uncertainty:** Utilizing measures like softmax scores, cosine similarity, and entropy to identify unknown classes. This involves assessing the confidence of the model's predictions and using thresholds to determine if an input is from an unknown class.
- **Manifold Mixup:** Leveraging manifold mixup techniques to generate synthetic samples that interpolate between known class embeddings. This method helps the model learn smoother decision boundaries and improves its ability to recognize and reject unknown classes.

Various prompting strategies will be tested to determine their effectiveness in open-set recognition:

- **Standard CLIP Prompting:** "A photo of a [CLASS]."
- **Negative Prompting:** "The image does not show a [CLASS]." This approach helps the model learn to distinguish between classes by explicitly stating what the image does not contain.
- **Uncertainty-Aware Prompting:** This strategy enables the model to identify cases with lower confidence in known class assignments rather than assuming an 'other' category. "A photo of [CLASS] and not of something else. / An image that might be similar to but is not a [CLASS] "
- **Open-Ended Prompting:** "This image shows [DESCRIPTION]."

By testing these different prompting strategies, the goal is to improve the model's ability to recognize known classes while effectively identifying and rejecting unknown classes in an open-set scenario.

4 Schedule

The project is structured to span approximately 15 weeks of full-time work, assuming an average of 30 hours per week, aligning with the typical workload for a Master's project of this scope. The following schedule outlines the key phases, activities, and milestones tailored to the exploration of Vision-Language Models (VLMs) in open-set recognition scenarios.

Weeks 1–3: The initial phase focuses on setting up the work environment and ensuring all necessary tools and libraries are installed, including Python, PyTorch, and any required VLM frameworks. During this period, the students will familiarize themselves with the datasets — ImageNet-1K, SUN397, and the iNaturalist subset. This involves curating and preprocessing the data to suit both closed-set and open-set evaluation scenarios. The students will process target labels to align with the project's objectives and implement the initial version of the evaluation metrics, such as Top-1 and Top-5 accuracy for closed-set evaluation.

⇒ *Milestone 1:* The datasets are fully prepared, and a basic learning and evaluation framework is established, enabling initial experiments to be conducted.

Weeks 4–6: Building upon the foundational work, the students will conduct a guided literature review focusing on prompt learning techniques and open-set recognition challenges, particularly those highlighted by Zhou et al. and Miller et al. The students will explore available pre-trained VLMs, such as CLIP and ViT-L/16, and integrate them into the learning framework. Adaptations may be necessary to tailor these models to the specific tasks of closed-set and open-set evaluations within the project. Initial closed-set evaluations will be performed to establish baseline performance metrics.

⇒ *Milestone 2:* The first pre-trained VLM is successfully integrated into the framework, and baseline closed-set evaluation results are obtained, providing a reference point for further experimentation.

Weeks 7–11: The focus shifts to implementing and experimenting with open-set recognition methods. The students will fine-tune the integrated models by incorporating predictive uncertainty measures and dedicated negative embeddings to enhance open-set performance, as suggested by Miller et al. Context Optimization (CoOp) will be implemented to automate prompt engineering, exploring both unified and class-specific contexts. Various prompting strategies will be tested to assess their effectiveness in open-set scenarios. Additionally, the impact of scaling the query set size will be investigated to understand the trade-offs involved. Experiments will include zero-shot evaluations on the iNaturalist subset and combined datasets.

⇒ *Milestone 3:* Open-set recognition methods are incorporated into the framework, and extensive experiments are conducted, yielding insights into the effectiveness of different techniques and prompting strategies.

Weeks 12–15: In the final phase, the students will run comprehensive experiments using various versions of pre-trained models, different prompting techniques, and evaluation metrics. Cross-dataset analyses will be performed to assess the generalization capabilities of the models. An in-depth analysis will be conducted to identify patterns, strengths, and weaknesses in the models' performance, including error analysis of misclassifications and overconfident errors. The students will compile the findings, generate plots and visualizations, and finalize the report.

⇒*Milestone 4*: Experiments are completed, results are thoroughly analyzed, and the final report is drafted, encapsulating the project's methodology, findings, and conclusions.

Ongoing Tasks: Throughout the project, the students are encouraged to write the report, documenting methodologies, experimental setups, and preliminary findings as they progress. Keeping detailed records of what was done and when will facilitate a comprehensive final report. Regular meetings with the project supervisor will help guide the project's direction and ensure adherence to milestones.

Final Presentation: At the conclusion of the project, a presentation will be prepared to showcase the results and insights gained. This will be presented to the research group, providing an opportunity for feedback and discussion.

Additional Notes: The LATEX thesis template from the supervisor's webpage⁶ should be used for the final report. Starting the writing process early is highly recommended to allow ample time for revisions and to ensure clarity and coherence in the documentation of the project's outcomes.

⁶ <https://www.ifi.uzh.ch/en/aiml/theses.html>

5 Grading Criteria

- **Implementation (30%):** Quality of code, adherence to best practices, and maintainability.
- **Final Report (40%):** Comprehensive documentation of methodology, results, and analysis.
- **Presentation (30%):** Clarity and effectiveness in communicating project outcomes.

6 Contact

- **Project Supervisor:**
Prof. Dr. Manuel Günther
Email: guenther@ifi.uzh.ch
- **Primary Investigator:**
Ahmed A. Ali
Email: alali@ifi.uzh.ch

7 References

- Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). **Learning to Prompt for Vision-Language Models.** *International Journal of Computer Vision*, 130, 2337–2348. [arXiv:2109.01134](https://arxiv.org/abs/2109.01134)
- Miller, D., Sünderhauf, N., Kenna, A., & Mason, K. (2023). **Open-Set Recognition in the Age of Vision-Language Models.** *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1-17. [arXiv:2303.08166](https://arxiv.org/abs/2303.08166)
- Geng, C., Huang, S.-J., & Chen, S. (2021). **Recent advances in open set recognition: A survey.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3614–3631. doi:10.1109/TPAMI.2020.2981604
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). **ImageNet: A Large-Scale Hierarchical Image Database.** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). **SUN Database: Large-scale Scene Recognition from Abbey to Zoo.** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Radford, A., Kim, J. W., Hallacy, C., et al. (2021). **Learning Transferable Visual Models from Natural Language Supervision.** In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Cruz, S., Rabinowitz, R., Günther, M., & Boulton, T. E. (2024). **Operational open-set recognition and PostMax refinement.** In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*.
- Wang, Z., He, Y., Xu, Q., Cao, X., Yang, Z., & Huang, Q. (2022). **OpenAUC: Towards AUC-oriented open-set recognition.** In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*.