Comparative Study of Prompt Learning Methods in Vision–Language Models

by

Team Member 1: Rajendra Chandrasekhar

Team Member 2: Rajani Jagarwal

Team Member 3: Swathy Bhavani

Report

Submitted in partial fulfillment of the requirements for the

DLFA Program

Centre for Continuing Education

Indian Institute of Science

Bangalore – 560 012 India

Abstract

Pre-trained vision–language models such as CLIP have demonstrated strong zero-shot image classification performance through prompt engineering. However, handcrafted prompts are limited and static, often failing to adapt across datasets and domains. To address this, Zhou et al. (2022a) introduced Context Optimization (CoOp), which replaces handcrafted prompts with learnable context vectors, enabling automatic adaptation of prompt context while keeping CLIP's pre-trained weights frozen [1] (https://arxiv.org/abs/2109.01134). Building on this, Zhou et al. (2022b) proposed **Conditional Context Optimization (CoCoOp), which adds a lightweight meta-network to generate instance-specific prompt vectors, improving generalization to unseen classes and domain shifts [2] (https://arxiv.org/abs/2203.05557). This project studies, implements, and compares CLIP, CoOp, and CoCoOp on multiple datasets — Caltech-101, Flowers-102, SUN397, and Indian Food — to evaluate prompt sensitivity, context length, meta-network variations, and base versus novel class performance. The findings reveal that CoCoOp consistently outperforms static prompt methods in complex datasets while maintaining robustness across domains, albeit with higher computational cost.

Acknowledgments

We would like to express sincere gratitude to professor Dr. Rajiv Soundararajan, our mentor Karthik and our faculty advisors and evaluators for their guidance and feedback throughout the course of this project. We are particularly thankful to the creators of the CLIP, CoOp, and CoCoOp frameworks, whose open source contributions made this research possible.

Table of Contents

Summary
12

Chapter I: Introduction

Problem Statement

Handcrafted prompts used in VLM model such as CLIP are dataset-dependent and non-adaptive, which limits generalization across unseen domains. There is a need to study prompt learning methods that can dynamically adapt prompts to improve model transferability and generalization.

Purpose of the Study

This study aims to:

- Compare the performance and generalization capabilities of CoOp and CoCoOp against baseline CLIP across multiple datasets
- Quantify the impact of prompt engineering variations, context length, and meta-network architectures
- Evaluate model robustness on both base classes and novel, unseen categories

Research Questions

Prompt Engineering Impact: How do different prompt formulations affect zero-shot classification performance across diverse datasets?
Learnable vs. Conditional Prompts: What performance gains do learnable prompts (CoOp) and conditional prompts (CoCoOp) provide over static approaches?
Architecture Parameters: How do context length and meta-network depth influence model accuracy and generalization?
Domain Transfer: How effectively do these models perform on base classes versus completely unseen categories?

Definition of Terms

CLIP: Contrastive Language–Image Pre-training model that aligns visual and textual representations.
CoOp: Context Optimization, a prompt learning technique that learns soft prompt embeddings.
CoCoOp: Conditional Context Optimization, extends CoOp by conditioning prompts on image features.

Meta-Network: A neural module that generates context token biases conditioned on image features.
Few-Shot Learning: Learning from a very limited number of labeled samples per class.
Domain Generalization: The ability of a model to perform well on unseen domains without retraining.
Vision-Language Model (VLM): A VLM, or Vision Language Model, is an advanced type of artificial intelligence model that combines computer vision and natural language processing capabilities. It can understand, interpret, and generate text about images or videos.

Assumptions and Limitations of the Study

- Computational limitations (Google Colab CPU/RAM) restricted experimental scope
- Large-scale datasets (e.g., ImageNet) could not be fully processed due to hardware constraints
- Time and resource limitations prevented exhaustive hyperparameter exploration

Overview

This report proceeds as follows: Chapter 2 reviews relevant literature on vision-language modeling and prompt learning. Chapter 3 details our experimental methodology, including data preparation and model configurations. Chapter 4 presents comprehensive results across all experiments. Chapter 5 synthesizes findings, discusses implications, and proposes future research directions

Chapter II: Related Work

Introduction

Zhou et al. (2022) introduced CoOp to address the static nature of handcrafted prompts by learning context vectors that improve model adaptability. Later, CoCoOp extended this by conditioning prompt vectors on image features using a meta-network, allowing better performance on unseen domains. These approaches represent a paradigm shift from manual prompt engineering to data-driven prompt learning.

Summary

Current research emphasizes that prompts play a pivotal role in VLM performance, particularly in zero-shot scenarios. While handcrafted prompts provide a baseline, their extreme sensitivity to phrasing variations limits their practical utility. Minor wording changes can dramatically impact classification accuracy, making prompt design a challenging and inconsistent process.

Learnable prompt techniques address these challenges by creating more robust and adaptive representations. However, several concerns persist:

- **Interpretability:** Learned prompts exist as continuous embeddings, making them difficult to interpret
- **Overfitting Risk:** Models may overfit to training distributions when context is learned
- **Computational Cost:** Conditional approaches require additional forward passes and meta-network training

.

Chapter III: Method/Experiment

Introduction

This chapter provides a description of the methodology used in our study of Conditional Prompt Learning for Vision-Language Models, emphasizing the CoCoOp model and its improvements over CLIP and CoOp. The primary goal is to enhance generalization to unseen classes and domains with minimal supervision. We explain our experimental setup, including data preprocessing, model selection, architectural modifications, training procedures, evaluation criteria, and identified limitations. The chapter concludes by summarizing the enhancements achieved and discussing the practical implications of the work.

Research Question(s)

- How does prompt engineering affect zero-shot performance across standard and new datasets?
- How do learnable prompts (CoOp) and conditional prompts (CoCoOp) improve generalization?
- What is the effect of context length and meta-network depth on performance?
- How do models perform on base versus unseen classes?

Data Preprocessing, Feature Engineering and Visualization

Dataset Selection and Structure

Four datasets were selected to represent varying levels of complexity and domain specificity:

- **Caltech-101:** General object recognition (30 classes used)
- **Flowers-102:** Fine-grained flower classification
- **SUN397:** Scene understanding and recognition
- **Indian Food:** Region-specific culinary dataset

Datasets were downloaded from cloud storage (Google Drive) and restructured to ensure consistent class-based directory organization.

Image Preprocessing Pipeline

All images underwent standardized preprocessing:

- **Resizing:** Uniform 224×224 pixel dimensions (CLIP standard input size)
- **Center Cropping:** Applied to non-square images to maintain aspect ratio
- **Normalization:** ImageNet statistics applied

- Mean: [0.48145466, 0.4578275, 0.40821073]
- Standard Deviation: [0.26862954, 0.26130258, 0.27577711]

Text Preprocessing Strategies

CLIP (Zero-Shot Baseline):

- Direct class name insertion into prompt templates
- Examples: "a photo of {class_name}" or "a photo of {class_name}, a type of Indian food"
- CLIP tokenizer with 77-token maximum length

CoOp (Context Optimization):

- Learnable context vectors initialized randomly
- Format: "[V1] [V2] ... [Vn] {class_name}"
- Context lengths tested: 4, 8, and 16 tokens

CoCoOp (Conditional Context Optimization):

- Similar structure to CoOp with instance-conditional context
- Meta-network generates context biases from image features

Configuration Parameter	**CoOp**	**CoCoOp**
Context Tokens	4, 8 and 16 learnable	4 , 8 and 16 learnable
Position	End	End
Conditioning	Static	Instance-conditional
Meta-Network	None	Multilayer perceptron (MLP) 2 ,4, 6 and 8
Epochs	10	10

Batch Size	32	32
Learning Rate	0.002	0.002
Weight Decay	0.0001	0.0001
Backbone Variant	VIT-B/32	VIT-B/32

Choice of Model

The study implemented three models: Pre-trained CLIP with frozen weights(zero-shot), CoOp (learnable prompts), and CoCoOp (conditional prompts). Experiments were run across Caltech-101, Flowers-102, SUN397, and Indian Food datasets.

Training the model, Performance of the Model and Metrics

Architecture:

- CLIP visual and text encoders remained frozen
- Only prompt learner (CoOp) or prompt learner + meta-network (CoCoOp) were trained
- Few-shot regime: 30 classes, 16 images per class

Evaluation Metrics:

- Base Class Accuracy: Performance on training classes
- New Class Accuracy: Performance on completely unseen classes

Experimental Variables:

- Prompt context length (4, 8, 16 tokens)
- Meta-network depth (2, 4, 6, 8 layers)
- Dataset complexity and domain specificity

Summary

CoCoOp demonstrated strong and consistent performance across the majority of tested datasets, with the exception of the Flower-102 dataset where its effectiveness was

limited. In comparison to established benchmarks such as Caltech-101, the Indian Food dataset exhibited reduced performance in zero-shot transfer tasks. Nevertheless, integrating adaptive prompt learning methods like CoOp and CoCoOp led to substantial improvements on this more challenging dataset. Notably, while CoCoOp delivered enhanced accuracy and generalization, it also required greater computational resources. Analysis of the meta-network architecture revealed that increasing the number of layers did not reliably improve results, indicating variable effects across different datasets. For unseen class recognition, zero-shot learning maintains a notable advantage, with CoCoOp surpassing CoOp; however, its generalization to new classes still does not fully match the capacity observed with standard zero-shot approaches.

Chapter IV: Results

Introduction

This chapter reports the experimental outcomes of our study on Conditional Prompt Learning for Vision-Language Models through CoOp and CoCoOp implementations. Our research addressed the challenge of achieving robust generalization to previously unseen classes and domains with minimal training examples. The experimental design aimed to quantify the performance gains of dynamic versus static prompt strategies and assess how architectural parameters—such as CLIP backbone selection and meta-network depth—affect classification results. We performed comprehensive evaluations on four diverse datasets: Caltech-101, SUN-397, Flowers-102, and Indian Food, which collectively span fine-grained recognition tasks and domain adaptation challenges.

Summary

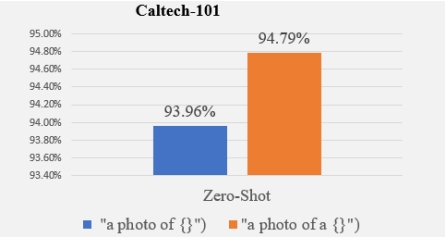
Experiment 1: Prompt Sensitivity Analysis

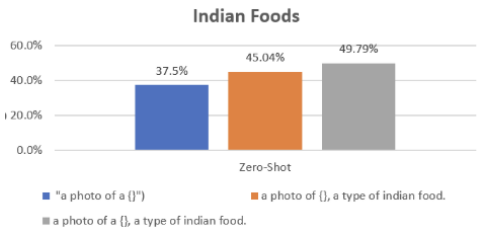
Objective: Assess the impact of prompt phrasing on zero-shot CLIP performance

Setup:

- Datasets: Caltech-101 and Indian Food (30 classes each)
- Prompt variations: With and without article "a" and further description.

Model Performance:





Key Observations:

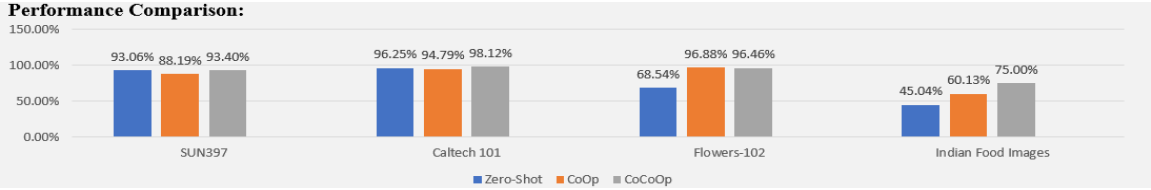
- The inclusion of the article "a" and adding detail for Indian Food dataset consistently increased zero-shot performance, demonstrating CLIP's sensitivity to minor linguistic variations.

Experiment 2: Cross-Dataset Model Comparison

Objective: Evaluate Zero-Shot CLIP across 2 datasets : Caltech-101 and Indian Food with 30 classes to assess the effect of different prompt strings on model performance

Setup: Four datasets with standardized evaluation protocol

Model Performance:



Key Observations:

- **CoCoOp Dominance:** Achieved highest accuracy on 3 of 4 datasets (SUN397, Caltech-101, Indian Food)
- **CoOp Specialization:** Outperformed CoCoOp on Flowers-102, where visually distinct patterns benefit from static context
- **Complexity Advantage:** Datasets with high intra-class variation (Indian Food) showed largest improvements with conditional adaptation
- **Adaptive Superiority:** CoCoOp's meta-network handled diversity more effectively than static prompt

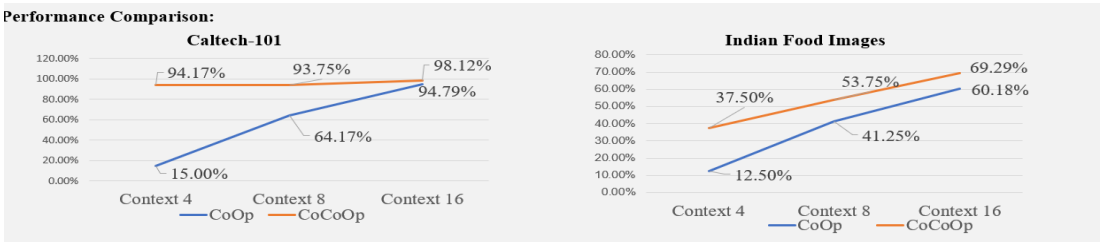
Experiment 3: Context Length Analysis

Objective: Evaluate CoOp, and CoCoOp — on Caltech-101(30 classes) and Indian Food(all classes) dataset, across multiple test configurations, to measure performance consistency and sensitivity to context vector length.

Setup:

- Models: CoOp and CoCoOp
- Datasets: Caltech-101 (30 classes) and Indian Food (all classes)
- Context lengths: 4, 8, 16 tokens

Model Performance:



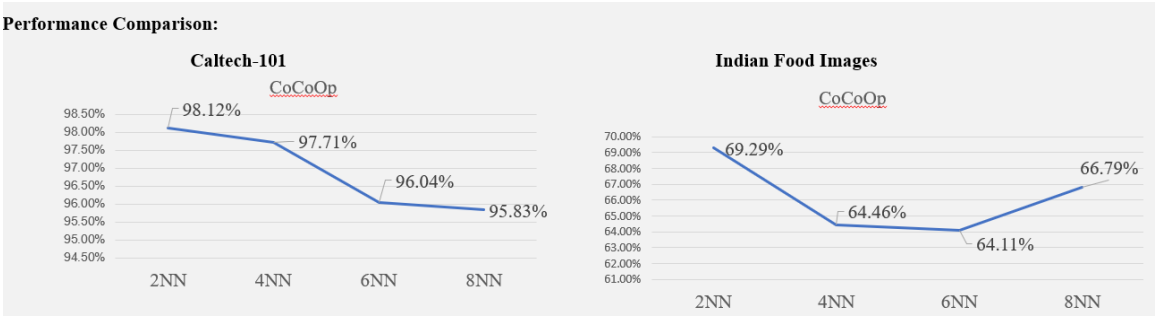
Key Observations:

- Both CoOp and CoCoOp demonstrated consistent accuracy improvements with increased context length across both datasets. The trend held regardless of dataset complexity, suggesting that richer learned representations enhance model expressiveness.

Experiment 4: Meta-Network Depth Analysis

Objective: Evaluate performance changes on **Caltech-101** and **Indian Food** datasets using multiple experimental configurations of the CoCoOp meta-network.

Model Performance:



Key Observations::

- Best performance is observed in 2NN layer for both datasets
- Performance improvement is inconsistent with increase in Meta-network layers

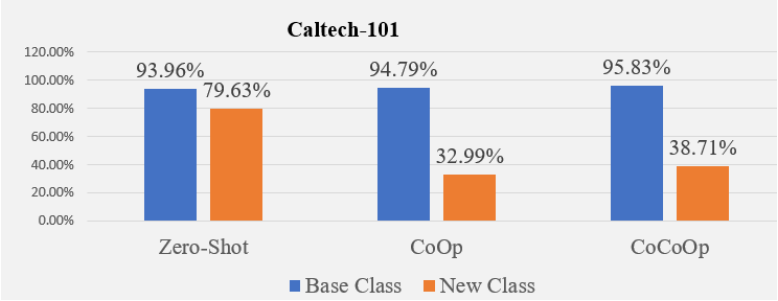
Experiment 5: Base vs. Novel Class Performance:

Objective: Evaluate three methods — Zero-Shot CLIP, CoOp, and CoCoOp for Caltech-101 dataset to assess model for base (30 classes) and new (71 classes) with 8NN layer for CoCoOp.

Setup:

- Dataset: Caltech-101
- Base classes: 30 (training)
- Novel classes: 71 (testing)
- CoCoOp configuration: 8 layers

Model Performance:



Key Observations:

- Zero-shot is performing best for new classes compared to CoOp and CoCoOp.
- CoCoOp is performing better than CoOp for new class.

Chapter V: Summary, Conclusions, and Recommendations

Introduction

In this work, we reproduced and extended the findings from papers [1] and [2] to deepen our understanding of prompt learning in Vision–Language Models (VLMs). We first examined the limitations of static prompts and then explored how dynamic prompt adaptation can improve few-shot image classification. Building upon the CLIP framework, we implemented CoOp and CoCoOp—methods that learn or condition prompts on image features to enhance generalization, particularly in low-data settings. To further evaluate model robustness, we incorporated a region-specific dataset, allowing us to compare the performance of CLIP and its extensions when applied to data distributions different from their pre-training domain.
Each experiment was designed to analyze how architectural components, such as prompt length (CoOp) and meta-network depth (CoCoOp), influence performance

Summary of the Results

Our study focused on the performance of CoOp and CoCoOp, which extend CLIP by learning adaptive text prompts. After validating our implementation by successfully reproducing baseline results, all experiments were conducted using the ViT-B/32 backbone for consistency.

Key findings include:

- **CoOp:** Increasing the prompt context length (4, 8, 16 tokens) consistently improved performance across datasets, demonstrating the advantage of richer learned prompts.
- **CoCoOp:** Using two meta-network layers generally provided strong performance across datasets, though Flowers-102 was an exception. Increasing the number of layers beyond the default did not show consistent improvements.
- **Indian Food Dataset:** Significant gains were observed when moving from CLIP to CoOp and CoCoOp, suggesting that adaptive prompts are particularly beneficial for region-specific datasets with limited standardization.
- **Unseen Classes:** Performance on unseen classes in Caltech-101 and Flowers-102 was weaker overall, though images containing highly distinctive concepts (e.g., “morning,” “rose,” “desert”) were still predicted correctly. This highlights persistent challenges in generalization.

Overall, the experiments emphasize the interaction between model architecture, dataset characteristics, and prompt-learning strategy in optimizing VLM performance.

Conclusions

The investigation into CoOp and CoCoOp demonstrates that conditional and learned prompts can substantially enhance few-shot classification performance. Reproducing baseline results confirmed the correctness and robustness of our implementation.

However, the experiments reveal important trade-offs:

- **Backbone Depth:** While the ViT-B/32 backbone improves performance on seen classes, it can reduce generalization to unseen classes, indicating a balance must be maintained between capacity and adaptability.
- **Meta-Network Complexity:** Increasing meta-network depth does not guarantee better performance; in our experiments, two layers performed best, suggesting diminishing returns beyond a certain level of model complexity.

These results collectively show that adaptive prompt learning is a powerful strategy for real-world, low-supervision scenarios, but careful architectural selection is crucial to maximize generalization.

Recommendations

Long-Context and Multimodal Prompting:

Modern vision-language models like Gemini 2.5 Pro support vastly larger context windows (millions of tokens, hundreds of images, extended video/audio) enabling more comprehensive analysis. Future research should explore:

- Extending CoOp/CoCoOp frameworks to long-context scenarios
- Integrating multiple modalities beyond vision and text
- Leveraging larger context for improved few-shot learning

Agentic and Interactive Prompting:

Recent VLM advances enable dynamic environmental interaction (GUIs, websites, applications) through sequential visual and textual input. Unlike CoOp/CoCoOp's single-pass classification:

- Agentic models perform multi-step reasoning and interaction
- Interactive prompting allows iterative refinement
- Sequential decision-making could enhance prompt learning

References

[1] Zhou, K., Yang, J., Loy, C.C. and Liu, Z., 2022. Learning to prompt for vision-language
models. *International Journal of Computer Vision*, *130*(9), pp.2337-2348.
[2] Zhou, K., Yang, J., Loy, C.C. and Liu, Z., 2022. Conditional prompt learning for vision-
language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816-16825).
[3] Reference implementation:
https://github.com/KaiyangZhou/CoOp