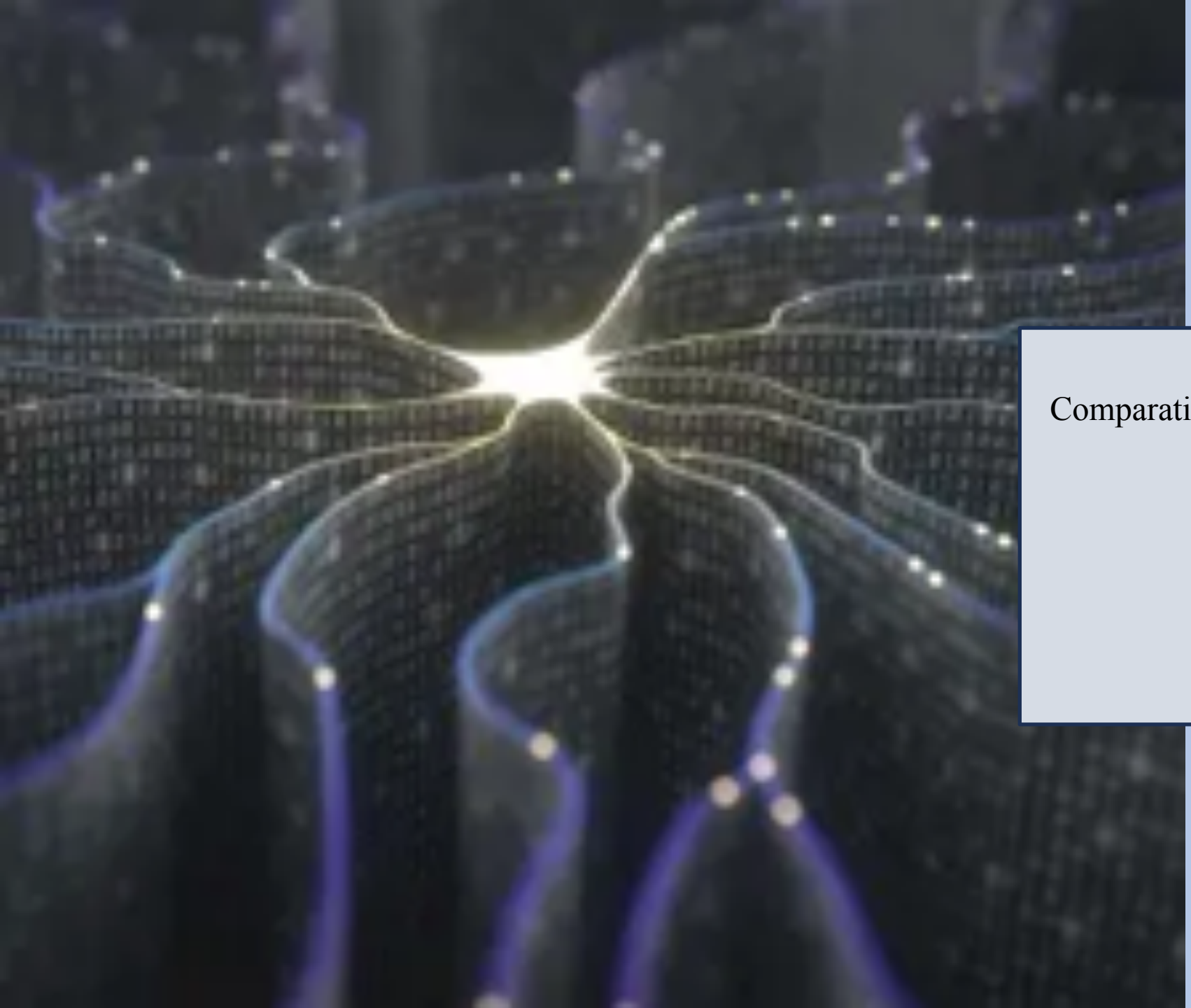


Comparative Study of Prompt Learning Methods in Vision–Language Models

Supervisor : Prof. Rajiv Soundarajan, IISC

Team : Rajendra Chandrasekhar
Rajani Jagarwal
Swathy Bhavani

Project Summary

Pre-trained vision–language models such as CLIP have demonstrated remarkable performance in zero-shot image classification through prompt engineering. However, handcrafted prompts are limited and static, often failing to adapt across datasets and domains. Recent advances in prompt learning (e.g., CoOp and CoCoOp) replace handcrafted prompts with learnable or conditional prompts, but their comparative performance, generalization, and domain transfer remain areas for deeper exploration.

This project aims to study, implement, and compare two prompt learning variants Context Optimization(CoOp) and Conditional Context Optimization(CoCoOp) using various datasets.

Dataset details

Caltech101 - https://www.kaggle.com/datasets/imbikramsaha/caltech-101 (Large scale dataset)
Flowers102 - https://www.kaggle.com/datasets/demonplus/flower-dataset-102 (Large scale dataset)
Sun397- https://huggingface.co/datasets/tanganke/sun397 (Large scale dataset)
Indian Food - https://www.kaggle.com/datasets/iamsouravbanerjee/indian-food-images-dataset **(**Small dataset**)**

Regional dataset : Indian Food Dataset
Total Classes : 36
Total Samples : 560

Class sample :

1. adhirasam
2. aloo_shimla_mirch
3. aloo_tikki
4. anarsa
5. ariselu
6. bandar_laddu
7. Basundi
8. Biryani
9. Chapati
10. Chak_hao_kheer … other 26 samples

Pre-processing steps

Dataset Setup

- Download the dataset (google drive in our case) and verify the folder structure of the dataset (Flowers102 - restructured the folder to class-wise)

Image Preprocessing

- **Resize**: Resize images to 224×224 pixels (CLIP's standard input size)
- **Center Crop**: Apply center crop if images aren't square
- **Normalization**: Normalize with ImageNet statistics:
  - Mean: [0.48145466, 0.4578275, 0.40821073], Std: [0.26862954, 0.26130258, 0.27577711]

Text Preprocessing

- **CLIP (Zero-shot):**
  - Use class names directly with prompt templates
  - Example: "a photo of {class name}" or “a photo of {class name}, a type of Indian food.”
  - Tokenize using CLIP's tokenizer (max length: 77 tokens)

- **CoOp (Context Optimization):**
  - Initialize learnable context vectors with random values
  - Use class names as suffix: "[V1] [V2] ... [Vn] {class name}"
  - Context length typically: 4, 8, or 16 tokens

- **CoCoOp (Conditional Context Optimization):**
  - Similar to CoOp but context is conditioned on image features with different meta net.

Model Architecture & Configuration

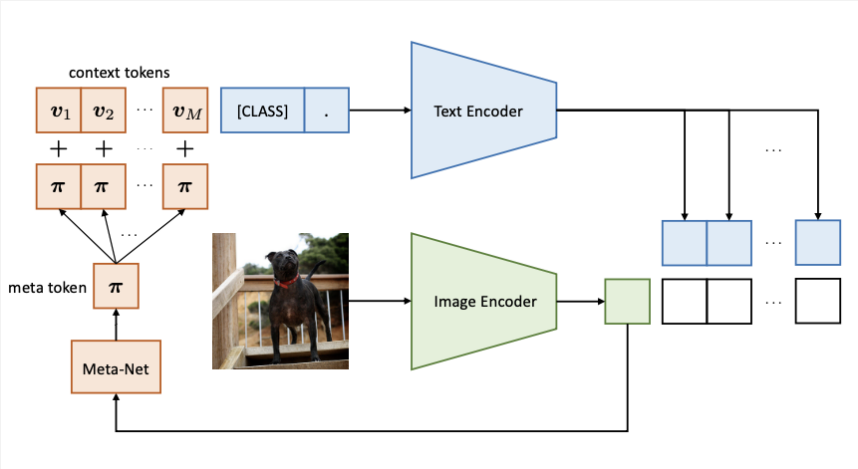
CLIP, CoOp, and CoCoOp for Few-Shot Learning

CLIP Architecture

Image → ViT-B-32 → Embeddings

Text → Encoder → Embeddings

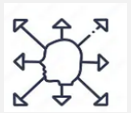
Base: ViT-B-32 (OpenAI)
224×224 input images



CoCoOp Architecture

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

Model Experiments



T1 : Prompt String Sensitivity
Effect of different prompt strings on Zero-shot CLIP across **Caltech-101** and **Indian Food** datasets

T2 : Model Performance Comparison
Zero-shot CLIP vs. CoOp vs. CoCoOp evaluated on Caltech-101, Indian Food, SUN397, Flowers-102

T3 : Prompt Context Vector Size Analysis
Impact of context vector length on model performance for **CoOp, CoCoOp** on **Caltech-101** and **Indian Food**

T4 : Meta-Net Layer Variations in CoCoOp
Performance changes with different meta-layer configurations on **Caltech-101** and **Indian Food**

T5: Base vs. New Class Performance
Comparative evaluation of CoOp and CoCoOp on **Caltech-101** for base-class and new-class generalization

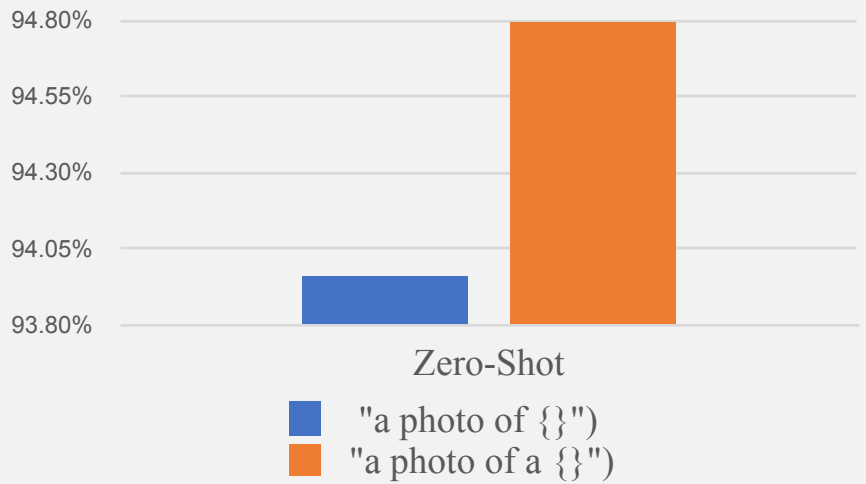
T1 : Prompt String Sensitivity

Summary:

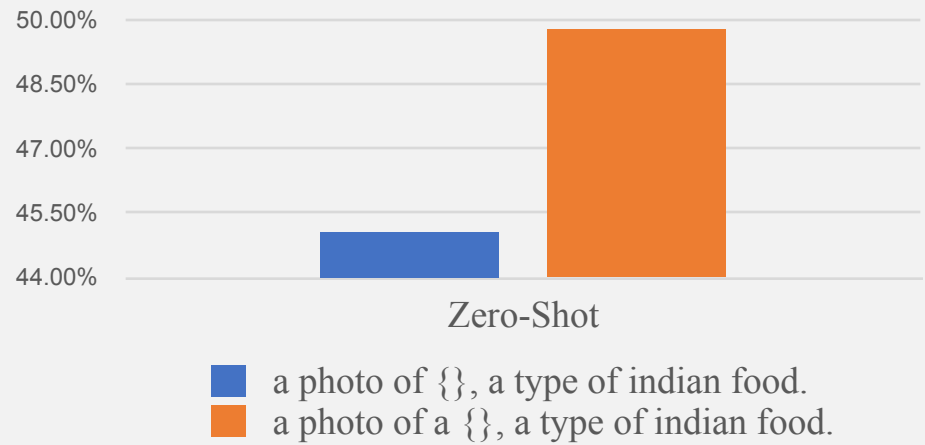
Evaluate Zero-Shot CLIP across 2 datasets : Caltech-101 and Indian Food with 30 classes to assess the effect of different prompt strings on model performance

Performance Comparison:

Caltech-101



Indian Food Images



Key Insights

- Addition of ‘a’ in prompt for Caltech-101 and Indian Food images is impacting the performance of Zero-Shot

T2 : Model Performance Comparison

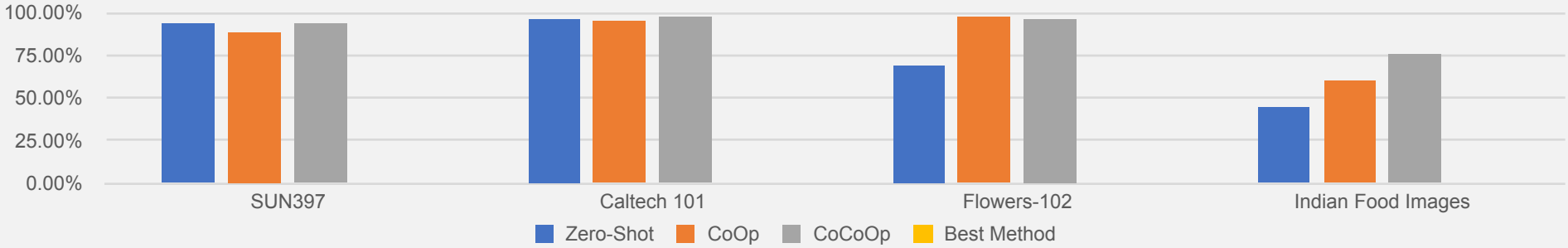
Summary:

Evaluate three methods — Zero-Shot CLIP, CoOp, and CoCoOp across 4 datasets : SUN397, Caltech-101, Flowers-102 and Indian Food Images to assess model changes impacting performance across datasets with 30 classes

Model Configuration Summary:

Method	Context Length	Epochs	Meta-Network
Zero-Shot CLIP	NA	NA	NA
CoOp	16	8–10	NA
CoCoOp	16	5–10	2NN

Performance Comparison:



Key Insights

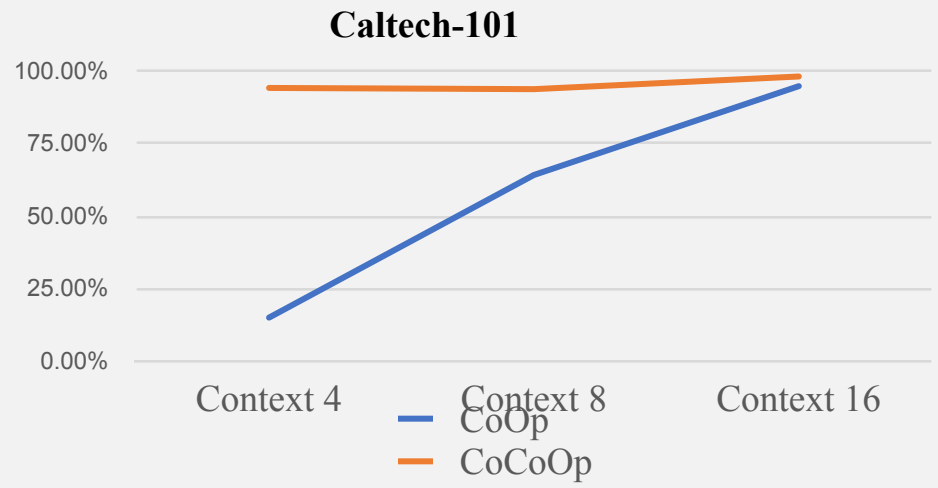
- CoCoOp is the most robust overall performer. Achieved highest accuracy in 3 of 4 datasets (SUN397, Caltech 101, Indian Food).
- CoOp excels on structured visual datasets. It outperformed CoCoOp on Flowers-102, where class patterns are visually distinct and benefit from static context optimization.
- Datasets with complex intra-class variation (e.g., Indian Food Images) showed the largest performance boost with conditional adaptation.
- CoCoOp's adaptive meta-network handled diversity better than static prompts. Dataset difficulty impacts model adaptability

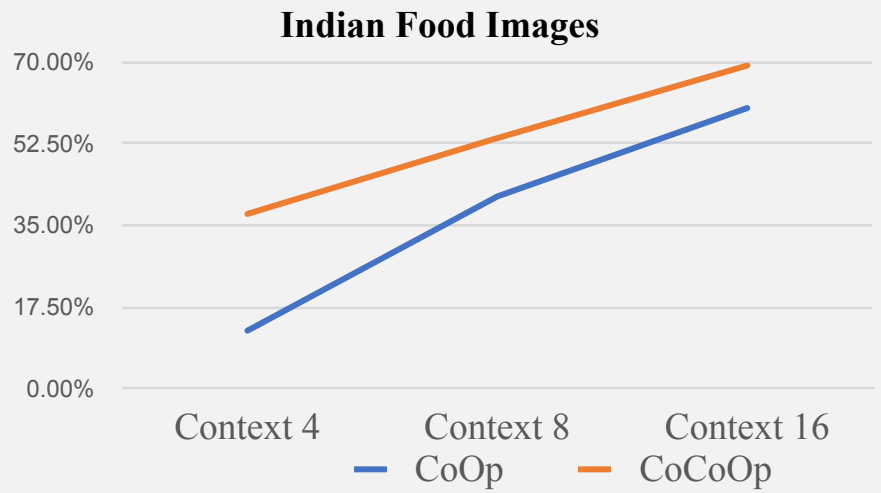
T3 : Prompt Context Vector Size Analysis

Summary:

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.

Performance Comparison:





Key Insights

- With increase in context length, improvement in accuracy of both models CoOp and CoCoOp is observed in both the datasets

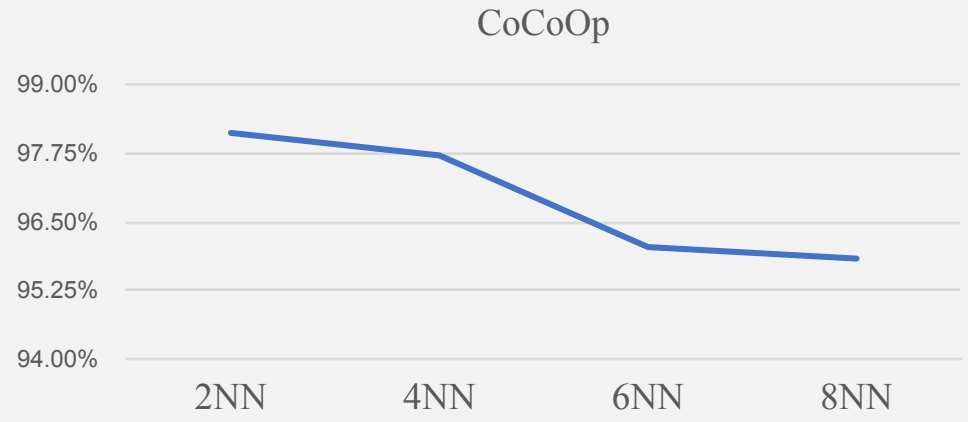
T4 : Meta-Net Layer Variations in CoCoOp

Summary:

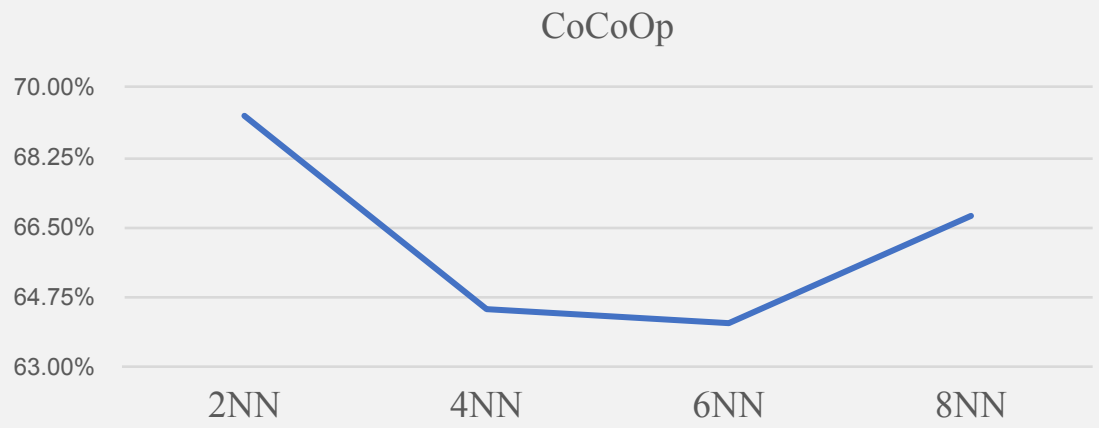
Evaluate performance changes on **Caltech-101** and **Indian Food** datasets using multiple experimental configurations of the CoCoOp meta-network.

Performance Comparison:

Caltech-101



Indian Food Images



Key Insights

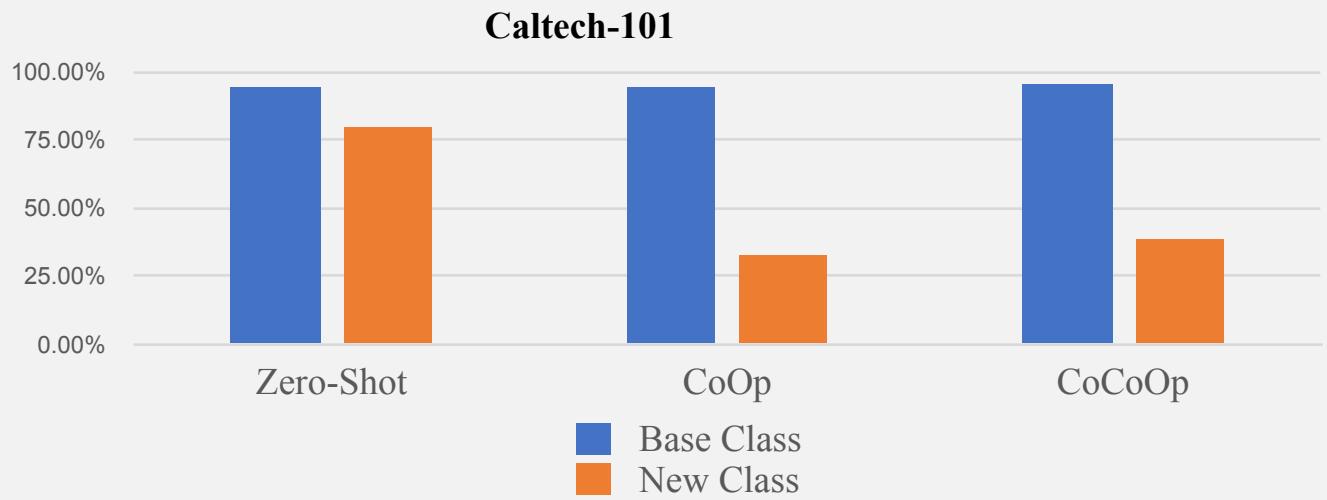
- Best performance is observed in 2NN layer for both datasets
- Performance improvement is inconsistent with increase in Meta-network layers

T5 : Base Vs New Class Performance

Summary:

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.

Performance Comparison:



Key Insights

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

Conclusion

Key Observations:

- CoCoOp performed well for all the datasets consistently however with exception of Flower-102 dataset
- In comparison to standard datasets like Caltech-101, Indian Food dataset showed degraded performance for Zero-Shot transfer, but with enhancement of the model to CoOp and CoCoOp it showed significant improvement in performance
- CoCoOp's superior performance comes at the cost of increased computational requirements
- Consistent improvement in performance was not observed with increase in meta layers across datasets
- Performance improvement is inconsistent with increase in Meta-network layers
- Zero-shot performs best on unseen classes, with CoCoOp outperforming CoOp but still falling short of zero-shot for new-class generalization.

Limitations:

- Full class execution of large scale datasets were restricted due to CPU and RAM limitations on Google Colab
- Team was working from 3 different time zones causing quite a few challenges

References:

[1] Zhou, K., Yang, J., Loy, C.C. and Liu, Z., 2022. Learning to prompt for vision-languagemodels. International Journal of Computer Vision, 130(9), pp.2337-2348.[22:21, 31/10/2025]

[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 computervision and pattern recognition* (pp. 16816-16825)

[3] [https://github.com/KaiyangZhou/CoOp]

Thank You