# Conditional Prompt Learning for Vision-Language Models

Vinit Raj - 19CS10065
Aryan Agarwal - 19CS30005

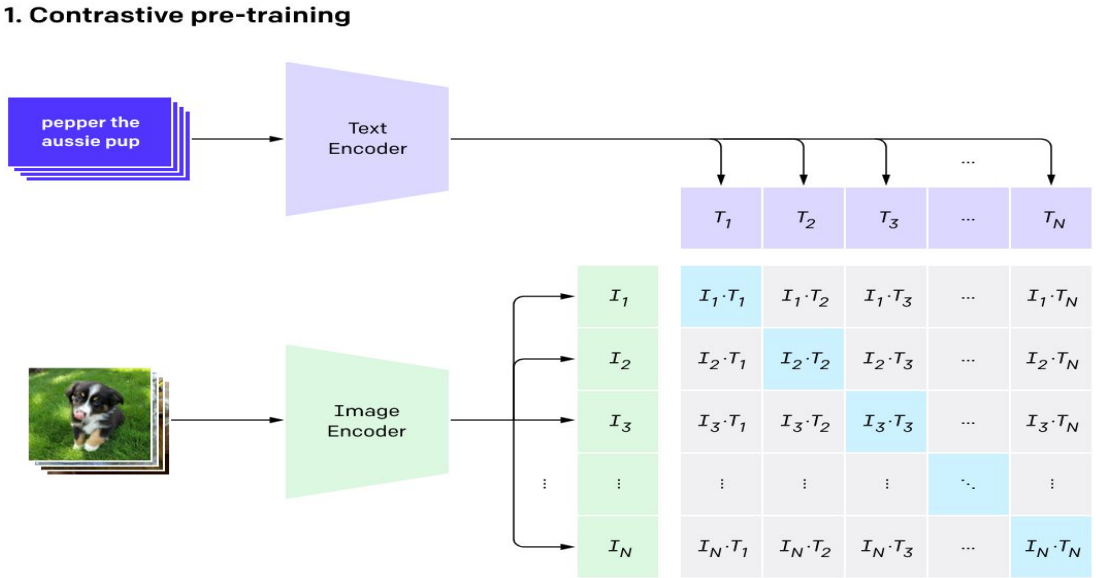
# Introduction

- Vision Language Models
- Task : Adapting pretrained models to other downstream datasets
- Prompt Learning and Context Optimization

# Vision Language Models

Textual as well as visual cues are used in learning process. Not just one word labels but often more descriptive.

## CLIP



# Task

- Classic vision models need lot of data but perform very good in their specified tasks
- But they need considerable effort to adapt to a new task or even same task with out of set inputs.
- Vision language models address this by not learning a benchmark but trying to learn useful embeddings that can be tuned efficiently for other downstream tasks.

# Prompt Learning and Context Optimization

- Handcrafted prompts are time consuming to generate and often have large implication of model performance
- So we automate prompt engineering using context words.

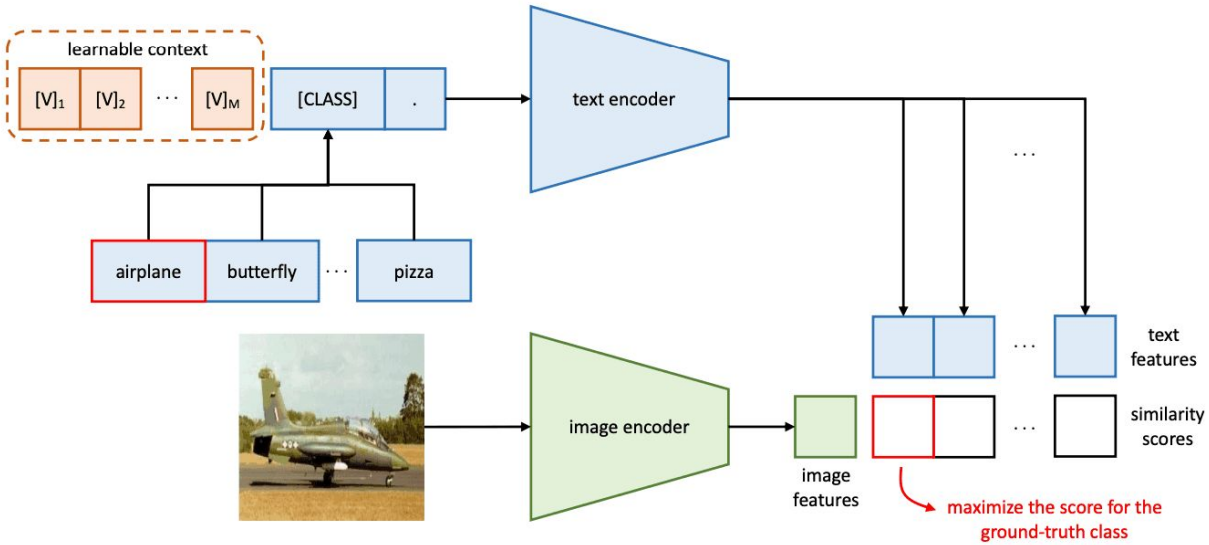
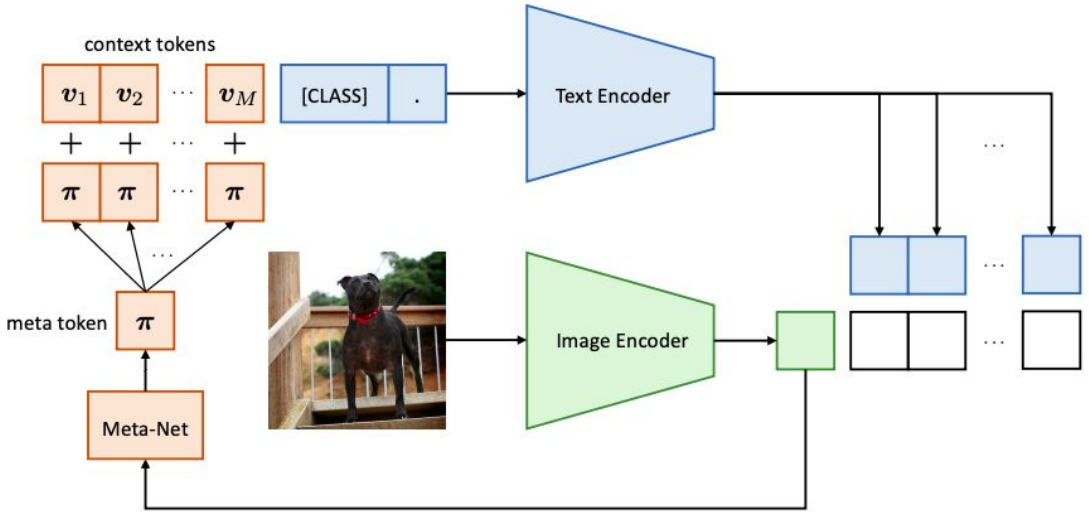



Figure 2: Overview of context optimization (CoOp).

# Methodology

Extra Meta net that produces input conditional token which helps in reducing the overfitting problem that CoOp experiences on the base classes.



# Methodology

- The Meta network produces input conditional values π from the feature map of the image and then adds it to context vectors.
- Then we find the similarity of each image with all the context vectors and then use softmax loss for each image and each text prompt.

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i(\boldsymbol{x}))/\tau)}.$$

# Evaluation

- Generalization from base to new classes within a dataset

(a) **Average over 11 datasets.**

| | Base | New | H |
|---|---|---|---|
| CLIP | 69.34 | **74.22** | 71.70 |
| CoOp | **82.69** | 63.22 | 71.66 |
| CoCoOp | 80.47 | 71.69 | **75.83** |

(b) ImageNet.

| | Base | New | H |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | **76.47** | 67.88 | 71.92 |
| CoCoOp | 75.98 | **70.43** | **73.10** |

- Cross dataset transfer

| | Source | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | FGVCAircraft | SUN397 | DTD | EuroSAT | UCF101 | *Average* |
| CoOp [62] | **71.51** | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | **46.39** | 66.55 | 63.88 |
| CoCoOp | 71.02 | **94.43** | **90.14** | **65.32** | **71.88** | **86.06** | **22.94** | **67.36** | **45.73** | 45.37 | **68.21** | **65.74** |
| Δ | **-0.49** | **+0.73** | **+1.00** | **+0.81** | **+3.17** | **+0.76** | **+4.47** | **+3.21** | **+3.81** | **-1.02** | **+1.66** | **+1.86** |

# Evaluation

- Domain Generalization

| | | Source | Target | | | |
|---|---|---|---|---|---|---|
| | Learnable? | ImageNet | ImageNetV2 | ImageNet-Sketch | ImageNet-A | ImageNet-R |
| CLIP [40] | | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 |
| CoOp [62] | ✓ | **71.51** | **64.20** | 47.99 | 49.71 | 75.21 |
| CoCoOp | ✓ | 71.02 | 64.07 | **48.75** | **50.63** | **76.18** |

# Conclusion

- Vision Language Models have vastly greater adaptive capabilities compared to classic CNN models that treat textual labels only as discrete values.
- These models extract semantic information from the text information provided and combining it with images provides deeper understanding of the subject matter.
- One such model CLIP uses contrastive learning to create a joint embedding space for the text and image, CoCoOp further improve the adaptation capability of the model by making the context vectors dependent on the image