MaPLe: Multi-modal Prompt Learning
CVPR-23

[1]Muhammad Uzair Khattak [1]Hanoona Rasheed [1]Muhammad Maaz [1,2]Salman Khan [1,3]Fahad Shahbaz Khan
[1]Mohamed Bin Zayed University of AI, [2]Australian National University, [3]Linköping University

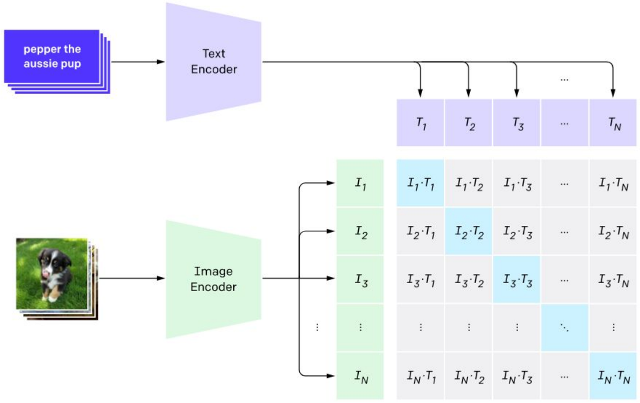
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background

- Foundational Vision-Language (VL) models
  - Pretrained on large image-text pairs
  - Typically trained using contrastive objectives

- Motivation to use
  - Open-vocabulary
  - Effectively transfer to downstream vision tasks
  - Generalizable



CLIP used for zero-shot classification (Radford et al., 2021)

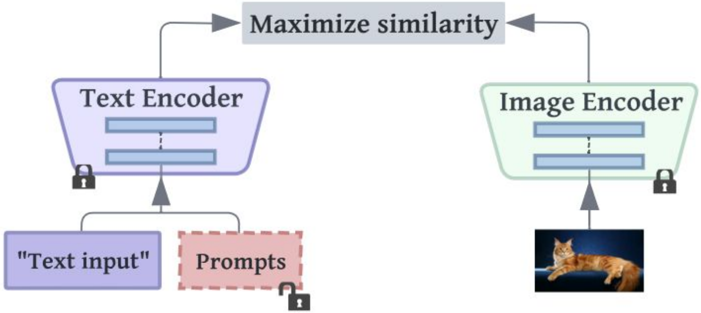
Problem statement

Adapting large scale Vision-Language models like CLIP for image-recognition tasks

- With limited downstream data (e.g few-/zero-shot)
- Without compromising on inherent generalization of CLIP

Existing solutions

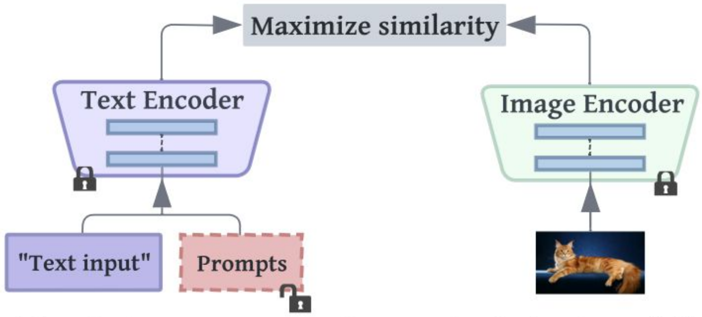
- Naive fine tuning (not trivial)
  - Data scarcity (few-shot/zero-shot) & training instability
  - Risk of losing generalization
- Inspired from NLP
  - Performs prompt tuning at language side



(a) Existing prompt tuning methods (Uni-modal)

Limitations in existing methods

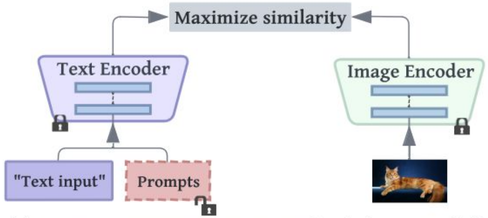
- Focus on uni-modal solutions
  - Ignores adapting the visual branch for better transfer
- Performs partial prompting
  - Prompting should instruct the model completely not partially



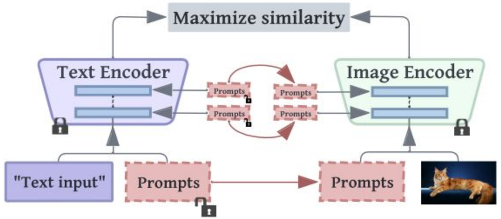
(a) Existing prompt tuning methods (Uni-modal)

Our contributions

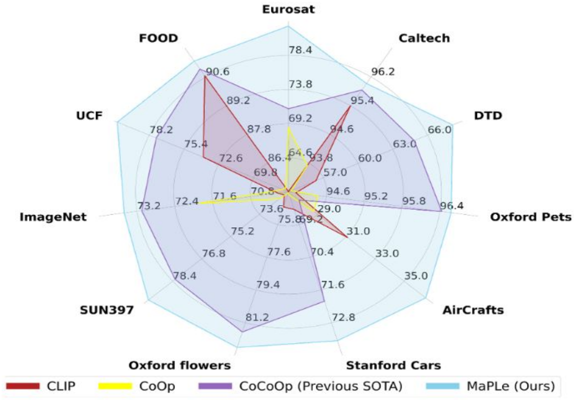
- We propose multi-modal prompt learning paradigm
  - Dynamically adjust both branches of CLIP for better transfer
- Hierarchical Vision-Language Deep Prompting
  - Learning of stage-wise feature relationships to allow rich context learning.
- Vision and Language Prompt Coupling
  - Ensure mutual synergy and discourages learning uni-modal solutions



(a) Existing prompt tuning methods (Uni-modal)



(b) Multi-modal Prompt Learning (MaPLe)

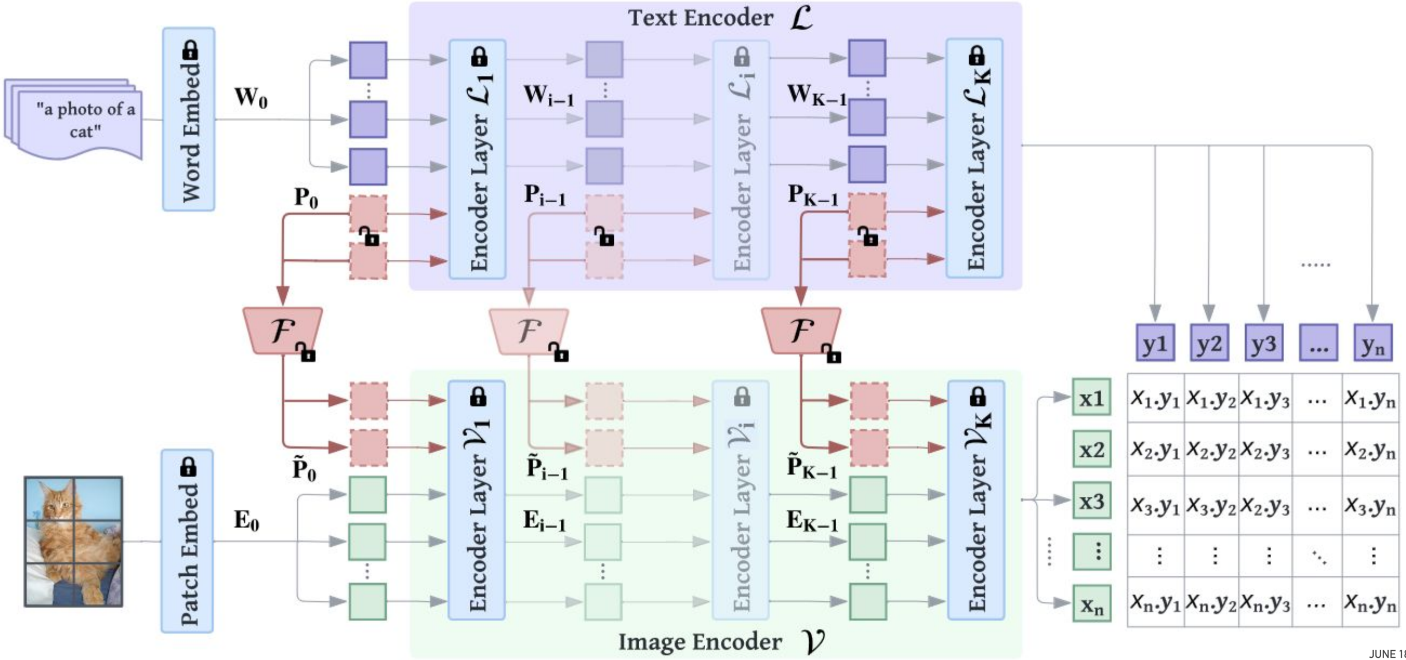


(c) Performance comparison on base-to-novel generalization

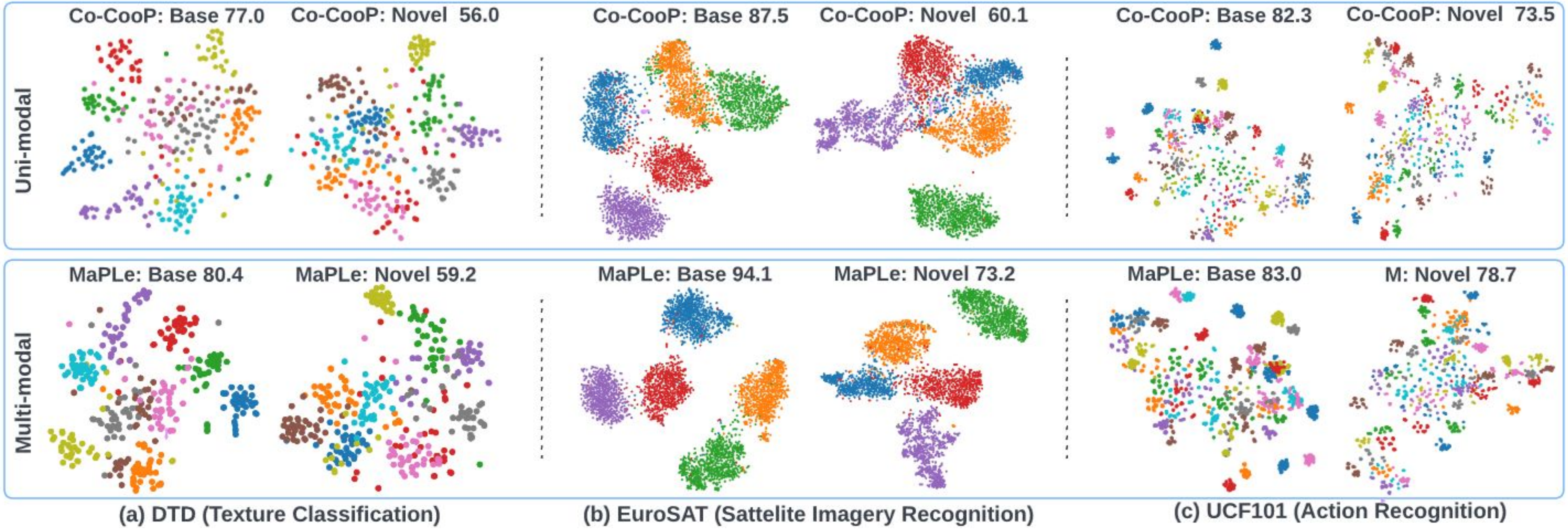
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

MaPLe architecture

- Hierarchical prompts in vision and language branches of CLIP
- Vision prompts conditioned on Language prompts via coupling function $\mathcal{F}(\cdot)$
- $\mathcal{F}(\cdot)$ Implemented as linear projection layer



Visualizations



Experiments

We conduct experiments on three different generalization benchmarks:

- Base-to-novel generalization (11 datasets)
    - Splits dataset in to base and novel classes
    - Train on base classes, evaluate on base and novel classes
- Cross-dataset evaluation (10 datasets)
    - Train on ImageNet source dataset
    - Directly evaluate on cross-datasets
- Domain generalization (4 datasets)
    - Train on ImageNet source dataset
    - Evaluate on out of distribution datasets

Experiments: Base-to-novel generalization

- Effect of individual components
- MaPLe provides optimal performance

Method	Base Acc.	Novel Acc.	HM	GFLOPS
1: MaPLe shallow ($J = 1$)	80.10	73.52	76.67	167.1
2: Deep vision prompting	80.24	73.43	76.68	18.0
3: Deep language prompting	81.72	73.81	77.56	166.8
4: Independent V-L prompting	82.15	74.07	77.90	167.0
5: MaPLe (Ours)	**82.28**	**75.14**	**78.55**	167.0

Experiments: Base-to-novel generalization

- Comparison with existing methods
- Co-CoOp is the previous state-of-the-art method

(a) **Average over 11 datasets**

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	**82.69**	63.22	71.66
Co-CoOp	80.47	71.69	75.83
MaPLe	82.28	**75.14**	**78.55**
	+1.81	+3.45	+2.72

Experiments: Cross-dataset transfer

- ImageNet trained model directly evaluated on cross-datasets

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	**90.49**	**65.57**	**72.23**	**86.20**	**24.74**	67.01	**46.49**	**48.06**	**68.69**	**66.30**

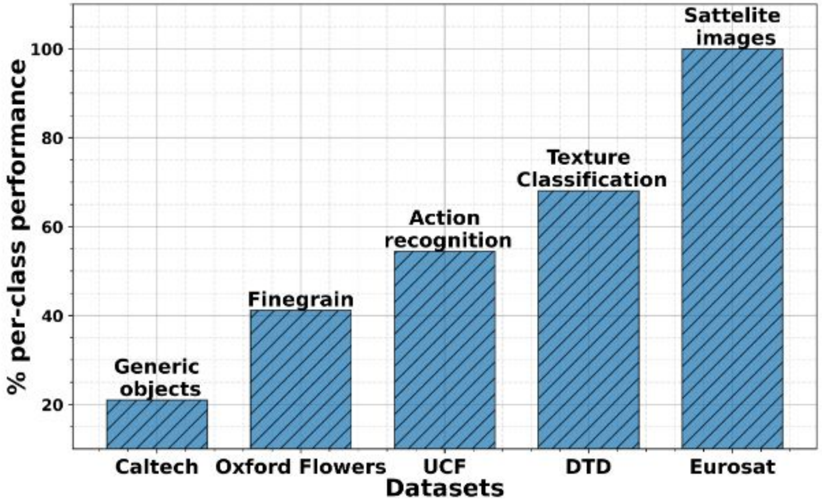
Experiments: Domain generalization

- ImageNet trained model directly evaluated on out-of-distribution datasets

	Source	Target			
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	**71.51**	**64.20**	47.99	49.71	75.21
Co-CoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	**49.15**	**50.90**	**76.98**

Further analysis on MaPLe

- Compared to Co-CoOp, MaPLe improves per-class performance as domain-shift of data increases (left to right)
- MaPLe is favourable for datasets with high domain-shifts



Conclusion

- Existing prompt learning methods
  - Adapts CLIPs partially
- We propose multi-modal prompting paradigm
  - Adapts both branches of CLIP for multi-modal representation learning
  - Integrates VL Deep Prompting for hierarchical context learning
  - VL Prompt Coupling for ensuring mutual synergies b/w VL modalities
- Improves generalization of CLIP towards 3 downstream tasks