Enhancing Pre-trained ViTs for Downstream Task Adaptation: A Locality-Aware Prompt Learning Method

ACM Multimedia 2024 Melbourne, Australia

Paper ID: 2070

Shaokun Wang, Yifan Yu, Yuhang He, Yihong Gong*
Xi'an Jiaotong University
China

西安交通大學 XI'AN JIAOTONG UNIVERSITY

Motivation & Goal & Solution

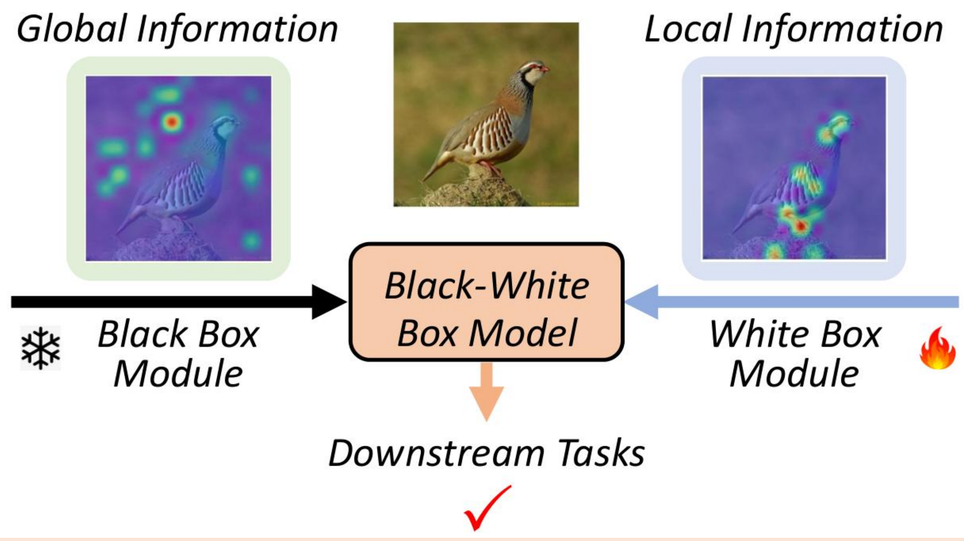
1.1 Limitation of pre-trained ViTs

Locality vanishing problem: *In downstream tasks, fully supervised pre-trained ViT tends to focus on global information while neglecting local information in critical regions.*

1.2 Goal

*We aim at improving the **limited local information incorporating capacity** of pre-trained ViTs, thereby **enhancing the adaptation of pre-trained ViTs to downstream tasks**.*

1.3 Solution



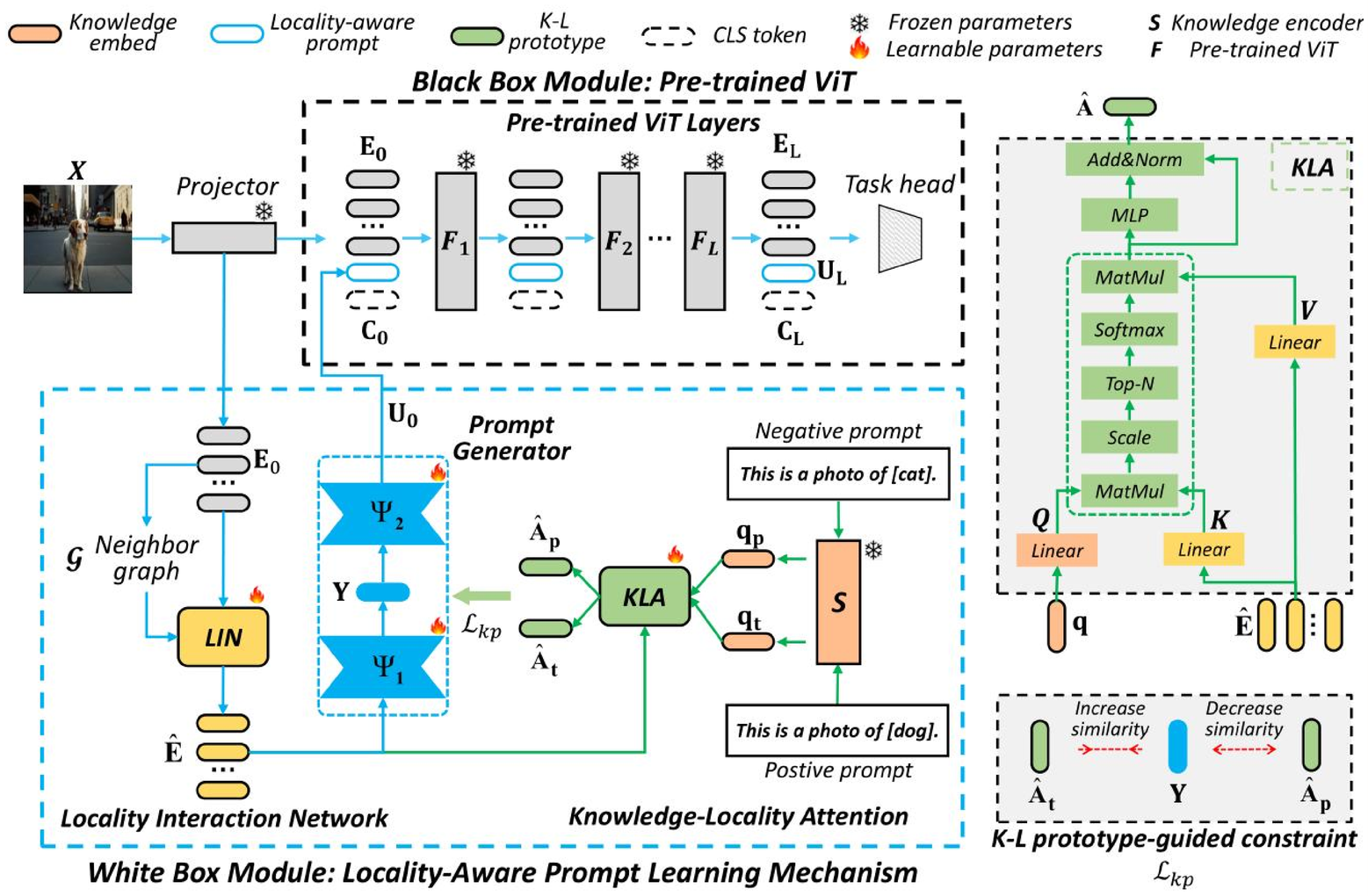
Working logic: *Our White Box module compensates for the Black Box module's (i.e., pre-trained ViT) local information incorporating capacity to better adapt to downstream tasks.*

Method

2.1 The framework of our LORE

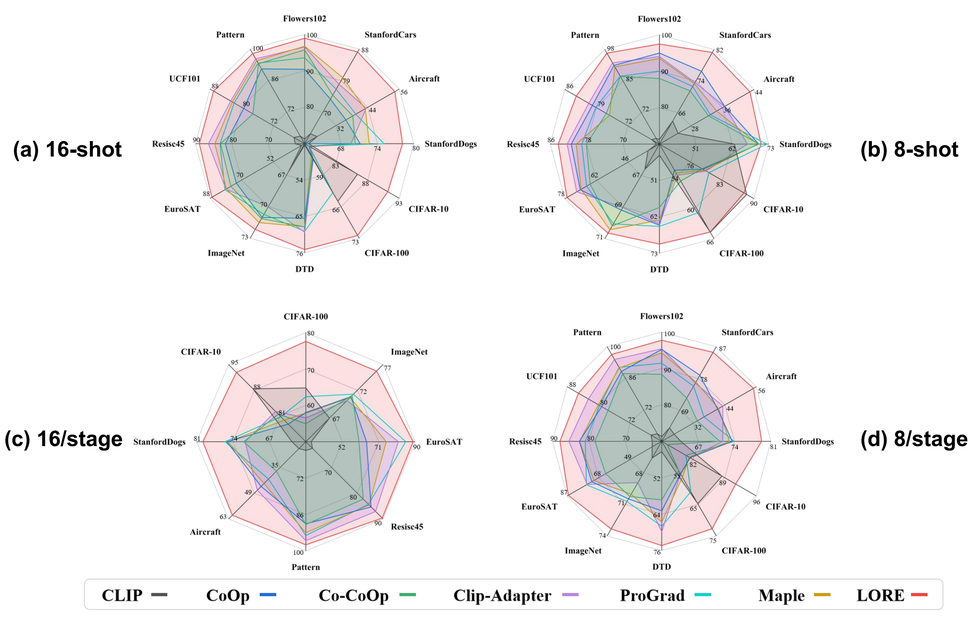
*Our LORE consists of a data-driven **Black Box module** and a knowledge-driven **White Box module**.*

- *The White Box module consists of a Locality Interaction Network (**LIN**), a Knowledge-Locality Attention (**KLA**), and a Prompt Generator (**PG**).*
- ***LIN** **enhances information interaction within image local regions.***
- ***KLA** **captures critical local regions under the guidance of semantic knowledge.***
- ***PG** **generates locality-aware prompts** with a K-L prototype-guided constraint.*
- *The workflow indicated by the **green** lines is **not necessary for the inference phase.***



Results

3.1 Classification task



3.2 Other downstream tasks

Task	Image Retrieval								Point Correspondences		Video Object Segmentation	
Dataset	ROxford5k				RParis6k				SPair-71k		Davis	
Metric	CLIP		ViT		CLIP		ViT		CLIP	ViT	CLIP	ViT
	M	H	M	H	M	H	M	H	PCK@0.1		J&F M	
B-B	0.397	0.107	0.302	0.094	0.708	0.482	0.603	0.358	18.25	16.61	54.38	58.12
LORE	**0.418**	**0.171**	**0.449**	**0.172**	**0.750**	**0.552**	**0.720**	**0.523**	**20.71**	**18.07**	**55.62**	**59.46**

3.3 Attention map visualization