Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning

Siteng Huang Biao Gong Yutong Feng Min Zhang Yiliang Lv Donglin Wang✉

Zhejiang University & Alibaba Group & MiLAB, Westlake University

CVPR SEATTLE, WA JUNE 17-21, 2024

Overview

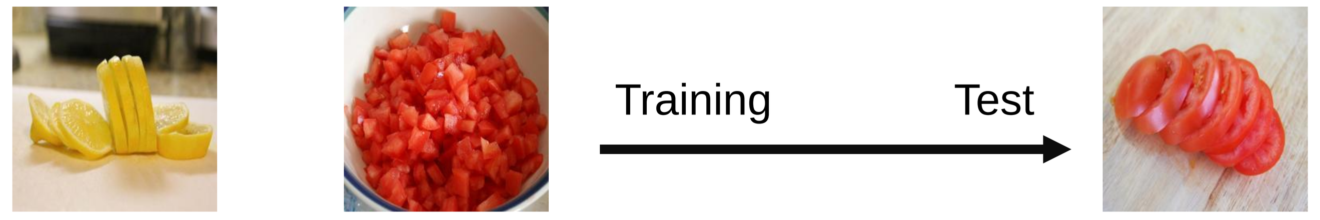
Research Question

Can we improve CLIP-based CZSL solutions with a focus on the universality?

Contributions

- propose a novel Multi-Path paradigm for CZSL with VLMs, which is flexible enough to derive new approaches.
- devise a model named Troika that effectively aligns the branch-specific prompt representations and decomposed visual features.
- achieves the SOTA performance on three CZSL benchmark datasets for both closed-world and open-world settings.

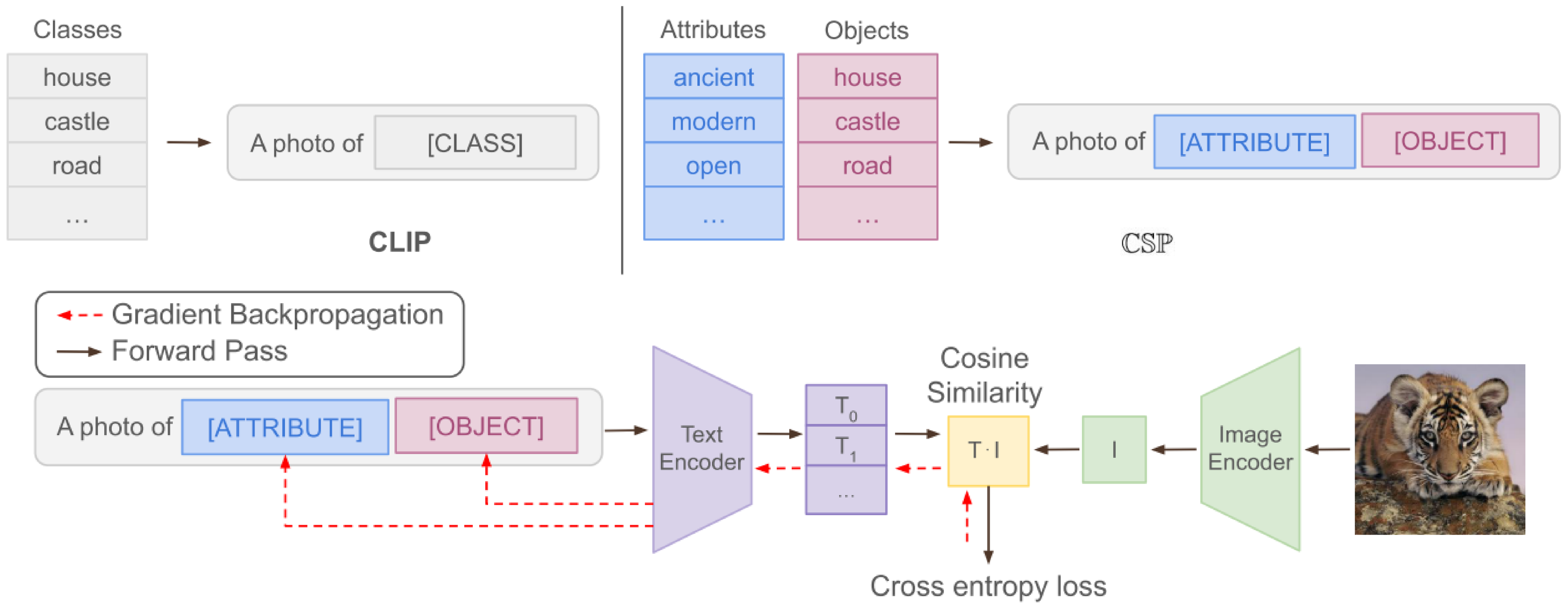
Background



Sliced Lemon Diced Tomato Sliced Tomato

Compositional Zero-Shot Learning (CZSL) studies to recognize unseen compositions at test time, while states and objects (*i.e.*, primitives) are presented in seen compositions during training.

CLIP-based solutions (*e.g.*, CSP in ICLR 2023):

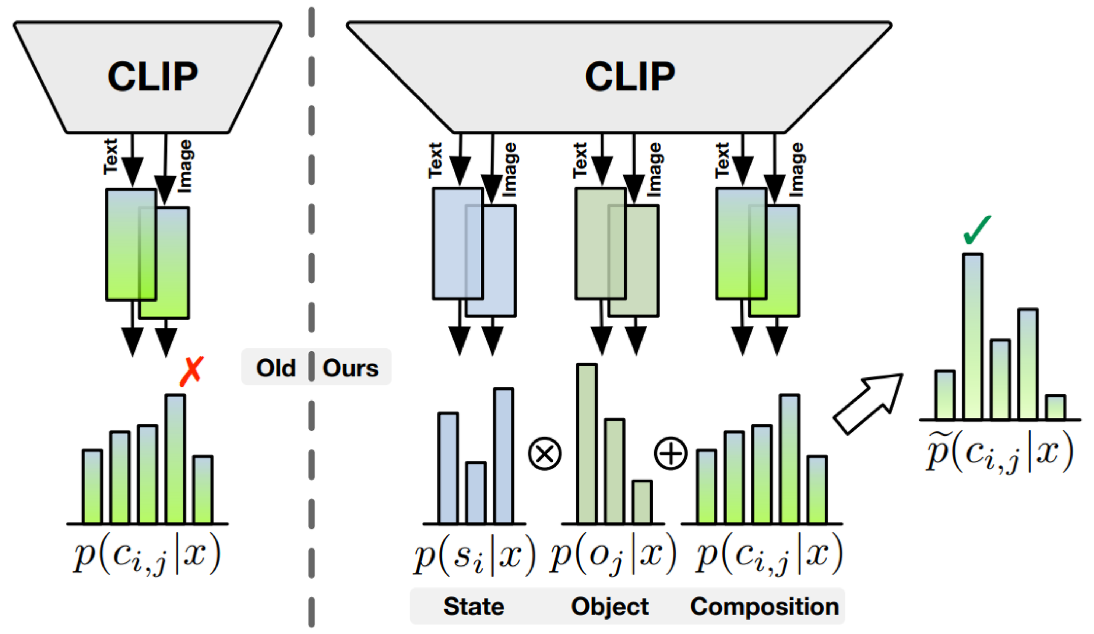


Lacking of independent and explicit primitive modeling:

- **The full leveraging of pre-trained knowledge fails**, since cross-modal knowledge is not only tied to the compositions, but also related to the single primitive.
- **The difficulty of generalizing to unseen compositions increases**, since the model easily over-rely on a limited number of seen compositions.

Multi-Path Paradigm

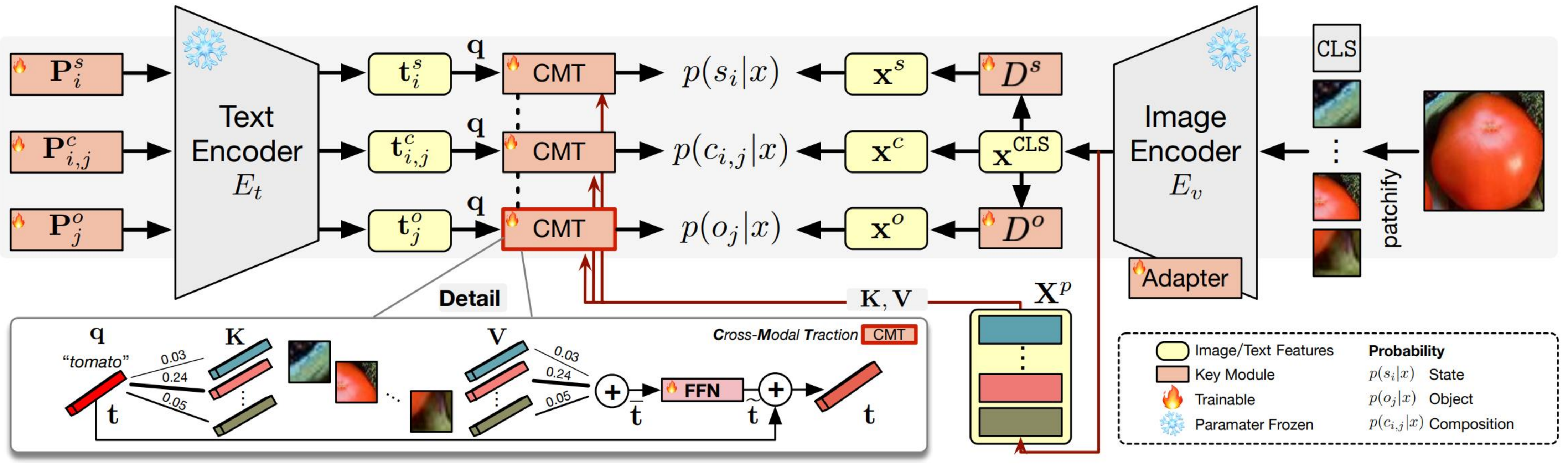
- **Training**: explicitly constructs vision-language alignments for the **state**, **object**, and **composition**.
- **Test**: integrates the predictions of all semantic components for the final decision.



Improvements by extending existing baselines with the paradigm:

	CLIP [33]				CoOp [43]			
	S	U	HM	AUC	S	U	HM	AUC
w/o MP	15.8	49.1	15.6	5.0	52.1	49.3	34.6	18.8
w/ MP	**24.3**	**49.6**	**21.9**	**8.2**	**62.5**	**58.1**	**41.9**	**28.3**

Troika: An Efficient Implementation



Learning Visual Representations:

- apply Adapter to adapt the image encoder in a parameter-efficient manner, avoiding updating its original parameters.
- introduce state and object disentanglers to decompose the state and object visual representations from the [CLS] features.

Learning Textual Representations:

- respectively conduct prompts for the state, object, and composition branches to maximize the exploitation of pre-trained knowledge.
- employ an independent prompt prefix for each branch to introduce different priors through special contexts.
- maintain the same primitive vocabulary as a cue of semantic compositionality.

red tomato

Method	Prefix	Vocabulary	Prompt
CLIP	*a photo of*	*red tomato*	$\mathbf{P}_{\text{red,tomato}}$
CoOp	$\mathbf{p}_1,\cdots,\mathbf{p}_m$	*red tomato*	$\mathbf{P}_{\text{red,tomato}}$
CSP	*a photo of*	$\mathbf{v}^s_{\text{red}}\ \mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}_{\text{red,tomato}}$
DFSP	$\mathbf{p}_1,\cdots,\mathbf{p}_m$	$\mathbf{v}^s_{\text{red}}\ \mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}_{\text{red,tomato}}$
Troika (Ours)	$\mathbf{p}^s_1,\cdots,\mathbf{p}^s_m$	$\mathbf{v}^s_{\text{red}}$	$\mathbf{P}^s_{\text{red}}$
	$\mathbf{p}^o_1,\cdots,\mathbf{p}^o_m$	$\mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}^o_{\text{tomato}}$
	$\mathbf{p}^c_1,\cdots,\mathbf{p}^c_m$	$\mathbf{v}^s_{\text{red}}\ \mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}^c_{\text{red,tomato}}$

Cross-Modal Traction:

- **Motivation**: compared to diverse visual presentations, learning only a fixed textual representation is intuitively insufficient to match all corresponding images from different domains.
- **Solution**: adaptively shift the prompt representation to accommodate the content diversity and diminish the cross-modal discrepancies.
- **Implementation**: apply cross-attention module to update the textual representation, taken the textual representation as query, the patch features as key and value. In practice, all three branches share the same module to reduce the parameter overhead.

Main Results with CLIP ViT-L/14

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Closed-world Results												
CLIP [33]	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
CoOp [43]	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
CSP [29]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
PromptCompVL [37]	48.5	47.2	35.3	18.3	64.4	64.0	46.1	32.2	-	-	-	-
DFSP(i2t) [23]	47.4	52.4	37.2	20.7	64.2	66.4	45.1	32.1	35.6	29.3	24.3	8.7
DFSP(BiF) [23]	47.1	52.8	37.7	20.8	63.3	69.2	47.1	33.5	36.5	32.0	26.2	9.9
DFSP(t2i) [23]	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
***Troika* (Ours)**	**49.0**$\pm$0.4	**53.0**$\pm$0.2	**39.3**$\pm$0.2	**22.1**$\pm$0.1	**66.8**$\pm$1.1	**73.8**$\pm$0.6	**54.6**$\pm$0.5	**41.7**$\pm$0.7	**41.0**$\pm$0.2	**35.7**$\pm$0.3	**29.4**$\pm$0.2	**12.4**$\pm$0.1
Open-world Results												
CLIP [33]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
CoOp [43]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
CSP [29]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
PromptCompVL [37]	48.5	16.0	17.7	6.1	64.6	44.0	37.1	21.6	-	-	-	-
DFSP(i2t) [23]	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	1.95
DFSP(BiF) [23]	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6	36.4	7.6	10.6	2.39
DFSP(t2i) [23]	47.5	18.5	19.3	6.8	**66.8**	60.0	44.0	30.3	38.3	7.2	10.4	2.40
***Troika* (Ours)**	**48.8**$\pm$0.4	**18.7**$\pm$0.1	**20.1**$\pm$0.1	**7.2**$\pm$0.1	66.4$\pm$1.0	**61.2**$\pm$1.0	**47.8**$\pm$1.3	**33.0**$\pm$1.0	**40.8**$\pm$0.2	**7.9**$\pm$0.2	**10.9**$\pm$0.3	**2.70**$\pm$0.1

For more experimental results, please refer to our paper.