Enhanced Continual Learning of Vision Language Models With Model Fusion

Haoyuan Gao [1]*, Zicong Zhang [1]*, Yuqi Wei [1], Linglan Zhao [4], Guilin Li [4], Yexin Li [3], Linghe Kong [1], Weiran Huang [1 2 3]†

[1] Shanghai Jiao Tong University [2] Shanghai Innovation Institute
[3] State Key Laboratory of General Artificial Intelligence, BIGAI [4] Tencent

Introduction

Limitation of Conventional Methods :

Conventional continual learning methods are insufficient for VLM fine-tuning, as they struggle to maintain the crucial zero-shot capabilities**.** Relatively few methods have been proposed for continual learning of VLMs.

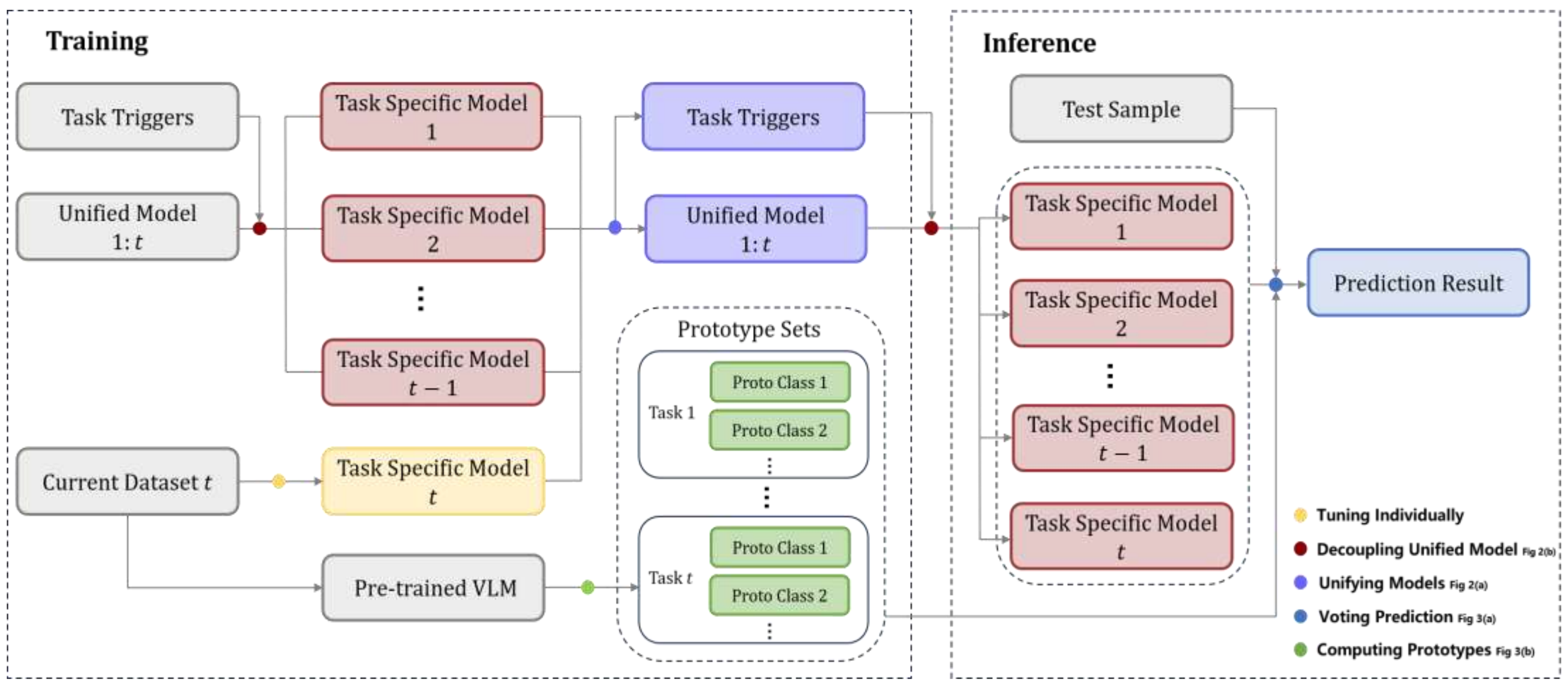
Optimization Objectives :

- Mitigating catastrophic forgetting
- Optimizing performance on current task
- Preserving zero-shot capabilities

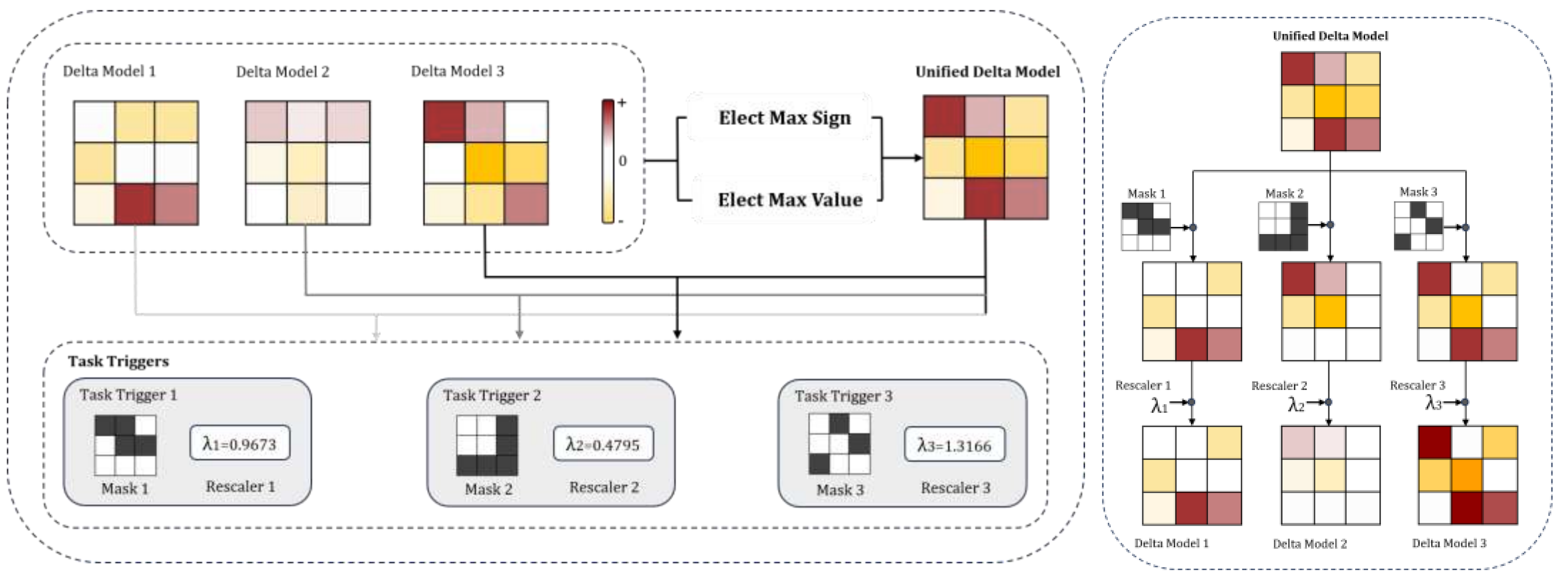
Methods

Framework of ConDU

ConDU maintains a unified model, a set of task triggers, and a series of prototype sets throughout the continual learning process.



Detailed Implementation of Unifying and Decoupling



$$\delta_j^{1:t} = \begin{cases} \max_i(\delta_j^i) & \text{if } \sum_{i=1}^t \delta_j^i > 0 \\ \min_i(\delta_j^i) & \text{if } \sum_{i=1}^t \delta_j^i < 0 \end{cases} \qquad M_j^i = \begin{cases} 1 & \text{if } \delta_j^i \cdot \delta_j^{1:t} > 0 \\ 0 & \text{if } \delta_j^i \cdot \delta_j^{1:t} < 0. \end{cases}$$

$$\lambda^i = \frac{\text{sum}(\text{abs}(\delta^i))}{\text{sum}(\text{abs}(M^i \odot \delta^{1:t}))}.$$

Training Stage : Continual Decoupling and Unifying

1. Tuning Individually :

Finetune pre-trained VLM on Current Dataset t to get θ^t

Subtracting θ^t from pre-trained model θ^0 to obtain delta model

$\delta^t = \theta^t - \theta^0$

2. Decoupling Unified Model :

Apply Task Triggers on Unified Model to reconstruct models

$\tilde{\delta}^i = \lambda^i M^i \odot \delta^{1:t}$ $\quad \tilde{\theta}^i = \tilde{\delta}^i + \theta^0$

3. Unifying Models :

Combine reconstructed models $\tilde{\delta}^i$and δ^t to get unified delta model

$\delta^{1:t} = \text{unify}(\tilde{\delta}^1, \tilde{\delta}^2 \cdots \delta^t)$

Inference Stage : Semantic-Based Voting Mechanisms

1. Computing Prototypes :

For each category in each task save its prototype during training

$P_k^i = f(y, \theta^0) + \frac{1}{|D_k^i|} \sum_{i=1}^{|D_k^i|} f(x_m, \theta^0)$

2. Aggregating Predictions :

a test image with task-id

reconstructed model to make prediction

a test image without task-id or from unseen tasks

Use pre-trained VLM to extract its image feature

Calculate cosine similarity between test feature with prototypes

Each task select highest similarity score then choose K-highest tasks

Weighted fuse the predictions of corresponding selected K models

Analysis

Theoretical Analysis

Theorem F.3 (Convergence of Iteration)**.** *Given n initial delta models δ^i, where $i \in [1, \ldots, n]$, after infinitely many iterations, if the relative order of λ^i values remains unchanged and $\forall i \neq j$, $\{k \mid M_k^i = 1 \text{ and } M_k^j = 1\} \neq \varnothing$, then these n delta models will converge to a uniquely determined set of n delta models.*

Theorem F.5. *If an initial delta model δ^1 is given, and during the n-th operation, a new delta model δ^n is added, and the current set of delta models $\{\delta^i(n) \mid i \in \{1, \ldots, n+1\}\}$ undergoes one iteration, then under the same conditions as Theorem F.3, and assuming all δ^i are independent and identically distributed, we have:*

1. *The probability of any $M_k^i(j)$ changing becomes negligible as n increases.*
2. *For each position in $\epsilon_{uni}(j)$, the probability of selecting a different corresponding delta model is small, and even if changes occur, their impact is minimal.*

Corollary F.6. *Under the same conditions as Theorem F.5, the following holds:*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left\| \delta^i(n) - \delta^i(n-1) \right\|_1 = 0.$$

t-SNE Visualization of Feature Space

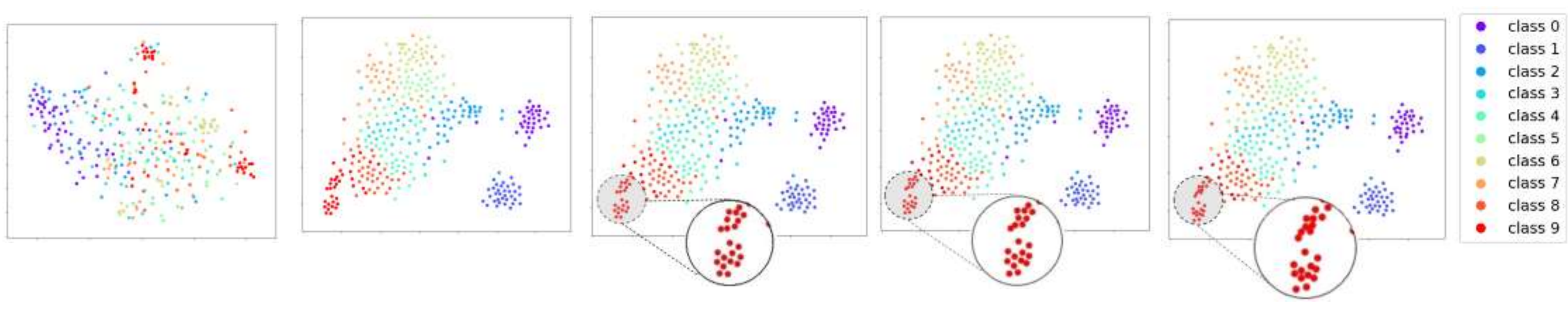


Fig 1:Pre-trained Model Fig 2:Session 1 Fig 3:Session 1 Fig 4:Session 6 Fig 5:Session 11

Results

Benchmark: Multi-domain Task Incremental Learning

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Individual FT	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
Transfer	ZSCL	-	86.0	67.4	45.4	50.4	69.1	87.6	61.8	86.8	60.1	66.8	68.1
	Dual-RAIL	-	88.4	68.2	44.6	54.9	71.0	88.5	59.6	89.0	64.7	65.2	69.4
	DPeCLIP	-	88.2	67.2	44.7	54.0	70.6	88.2	59.5	89.0	64.7	64.8	69.1
	MulKI	-	87.8	69.0	46.7	51.8	71.3	88.3	64.7	89.7	63.4	68.1	70.1
	ConDU(FT)	-	88.1	68.9	46.4	57.1	71.4	88.7	65.5	89.3	65.0	67.8	**70.8**
	ConDU(LoRA)	-	88.1	68.9	45.7	57.0	71.3	88.8	61.2	89.3	65.1	67.8	70.3
Average	ZSCL	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	68.0	75.4
	Dual-RAIL	52.5	96.0	80.6	70.4	81.3	86.3	89.1	73.9	90.2	68.5	66.5	77.8
	DPeCLIP	49.9	94.9	82.4	69.4	82.2	84.3	90.0	74.0	90.4	68.3	66.3	77.5
	MulKI	52.5	93.6	79.4	67.0	79.8	83.9	89.6	77.1	91.2	67.1	69.1	77.3
	ConDU(FT)	59.6	93.4	83.7	68.1	83.4	83.7	90.1	76.7	90.6	68.6	68.6	**78.8**
	ConDU(LoRA)	51.9	94.9	84.4	69.8	81.1	84.4	90.0	77.3	89.5	69.0	69.3	78.3
Last	ZSCL	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	93.9	85.3	80.2	83.6
	Dual-RAIL	52.5	96.8	83.3	80.1	96.4	99.0	89.9	98.8	93.5	85.5	79.2	86.8
	DPeCLIP	49.9	95.6	85.8	78.6	98.4	95.8	92.1	99.4	94.0	84.5	81.7	86.9
	MulKI	49.7	93.0	82.8	73.7	96.2	92.3	90.4	99.0	94.8	85.2	78.9	85.1
	ConDU(FT)	58.6	93.7	86.6	76.1	98.2	93.4	91.9	99.6	94.8	84.9	80.5	**87.1**
	ConDU(LoRA)	48.9	95.2	87.8	78.5	96.3	95.2	91.7	97.6	93.0	85.3	78.8	86.2

Contribution

- Introduce model fusion to continual learning for VLMs
- Propose a novel Decoupling-Unifying framework, compatible with PEFT and full-finetune paradigms.
- Propose a semantic-based voting mechanism for prediction in zero-shot scenarios.
- Extensive experiments on multiple benchmarks