Enhanced Continual Learning of Vision-Language Models with Model Fusion

Haoyuan Gao *, Zicong Zhang *, Yuqi Wei , Linglan Zhao ,
Guilin Li , Yexin Li , Linghe Kong , Weiran Huang †

Speaker : Zicong Zhang

7th April 2025

SJTU MIFA LAB

Background

- Optimization Objectives for VLMs:

  mitigating catastrophic forgetting,

  optimizing performance on the current task

  preserving zero-shot capabilities
- Conventional continual learning approaches are insufficient for VLM fine-tuning, as they struggle to maintain the crucial zero-shot capabilities
- Continual learning methods designed for VLMs either requires reference dataset or the careful tuning of multiple hyperparameters

  .

Motivation

- Due to **limited memory resources for storing data and models**, continual learning faces the critical challenge of balancing stability and plasticity.

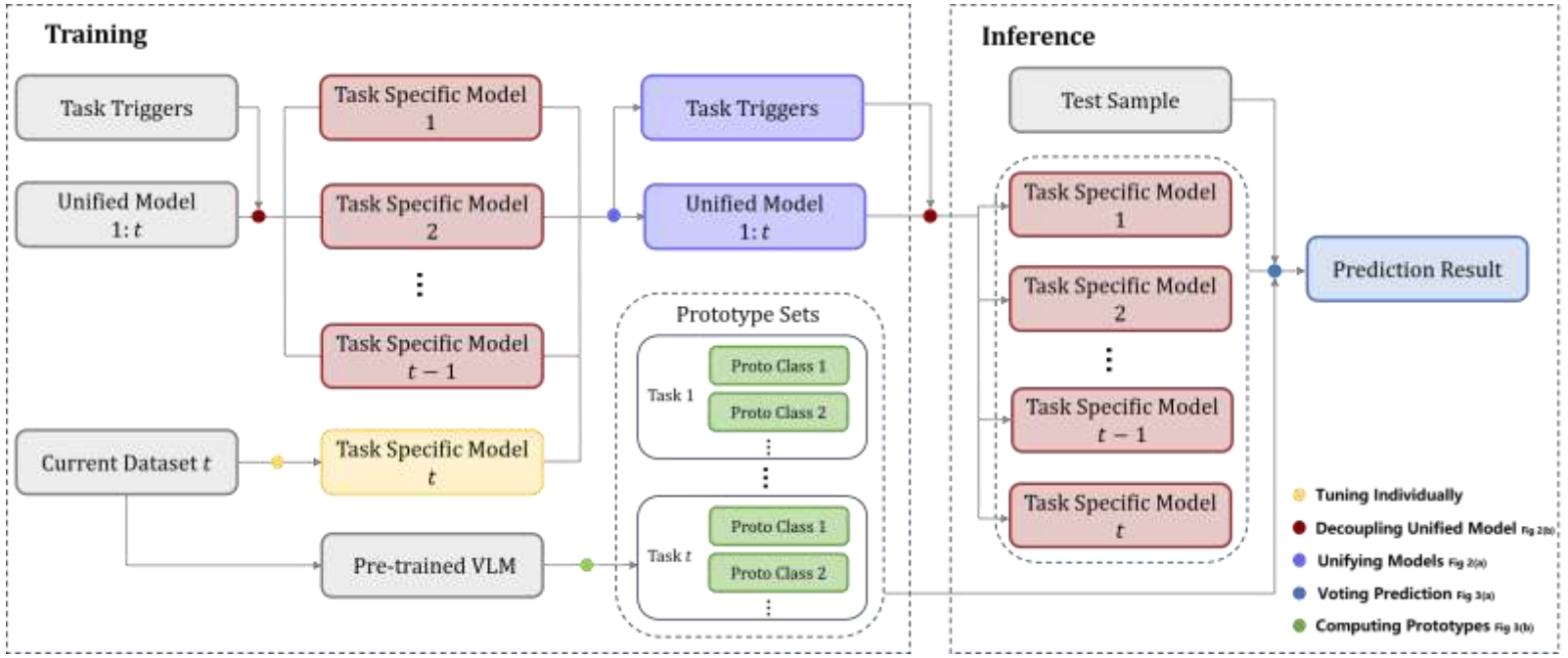
  What if we could “preserve all fine-tuned models” with minimal memory overhead?

- However, simply retaining past fine-tuned models falls short in addressing knowledge transfer and generalization.

  Could we leverage the capabilities of pretrained VLM to address these challenges?

Methods

- We introduce model fusion to VLMs and propose a novel **Decoupling-Unifying framework** compatible with PEFT and full-finetune paradigms.



Methods

Delta Models Continually Fusion at Training Stage :

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

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

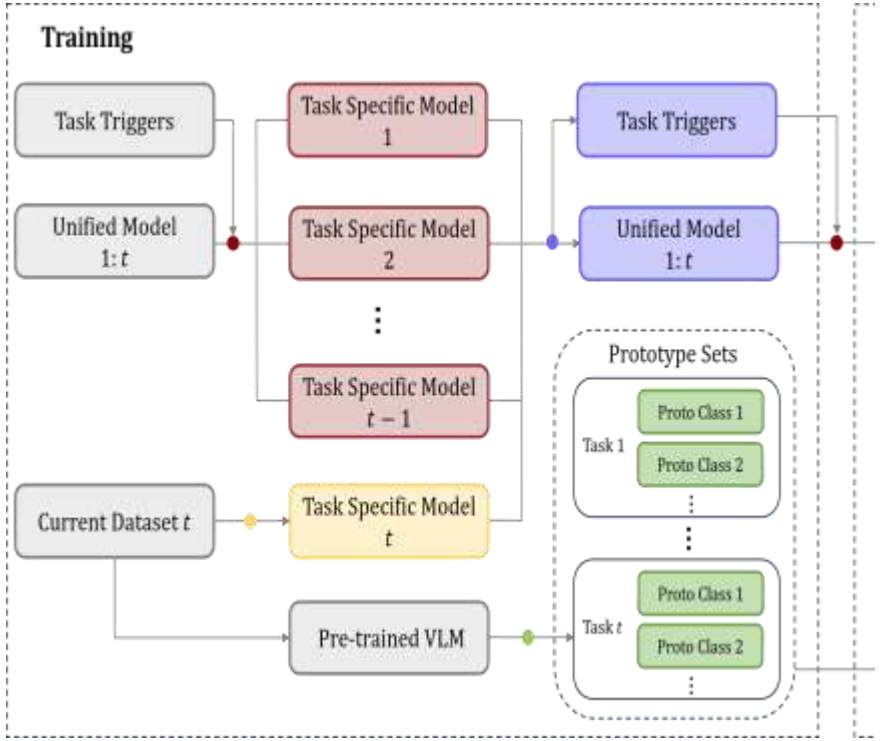
2. Decoupling Unified Model :

apply *Task Triggers* on *Unified Model* to reconstruct models

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

3. Unifying Models :

combine reconstructed models $\tilde{\delta}^i$and δ^t to get unified delta model $\delta^{1:t} = \text{unify}(\tilde{\delta}^1, \tilde{\delta}^2 \cdots \delta^t)$



Methods

Semantic-Based Aggregating Mechanism at Inference Stage:

1. Computing Prototypes :

for each category in each task, we save its prototype during training

2. Aggregating Predictions :

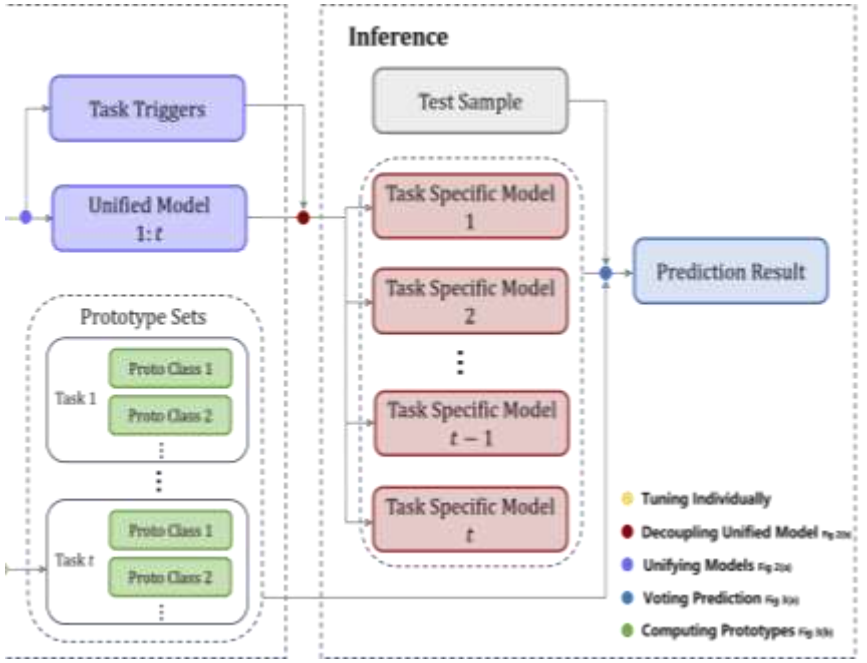
- for a test image with task-id we directly use the corresponding reconstructed model to make prediction
- for a test image without task-id or from unseen tasks

  use pre-trained VLM to extract its image feature

  calculate cosine similarity between test feature with prototypes

  for each task select highest similarity score then choose K-highest tasks

  weighted fuse the predictions of corresponding selected K models



Experiments : MTIL

We evaluate our method on Multi-domain Task Incremental Learning (MTIL) benchmark

	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

Experiments : few-shot MTIL

We evaluate our method on few-shot Multi-domain Task Incremental Learning (few-shot MTIL) benchmark

	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	59.6	89.0	64.7	65.2	65.3
	Individual FT	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
Transfer	Continual FT	-	72.8	53.0	36.4	35.4	43.3	68.4	47.4	72.6	30.0	52.7	51.2
	WiSE-FT	-	77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL	-	84.0	68.1	44.8	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	MoE	-	87.9	68.2	44.1	48.1	64.7	88.8	69.0	89.1	64.5	65.1	68.9
	Dual-RAIL	-	88.4	68.2	44.6	54.9	71.0	88.5	59.6	89.0	64.7	65.2	69.4
	ConDU(FT)	-	88.0	69.5	45.6	54.4	71.1	88.7	62.2	88.9	64.4	66.6	70.0
	ConDU(LoRA)	-	88.1	68.5	45.6	56.4	71.2	89.0	64.0	88.8	64.9	66.4	**70.3**
Average	Continual FT	28.1	86.4	59.1	52.8	55.8	62.0	70.2	64.7	75.5	35.0	54.0	58.5
	WiSE-FT	32.0	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
	ZSCL	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	MoE	30.0	89.6	73.9	58.7	69.3	79.3	88.1	76.5	89.1	65.3	65.8	71.4
	Dual-RAIL	36.0	94.2	70.9	58.8	70.6	84.3	88.5	70.3	89.7	66.5	65.8	72.3
	ConDU(FT)	33.1	90.5	74.1	58.3	76.2	81.0	87.9	73.4	88.0	64.8	67.1	72.3
	ConDU(LoRA)	32.4	92.1	75.4	58.8	75.1	82.9	87.3	74.0	89.3	65.1	67.0	**72.7**
Last	Continual FT	27.8	86.9	60.1	58.4	56.6	75.7	73.8	93.1	82.5	57.0	66.8	67.1
	WiSE-FT	30.8	88.9	59.6	60.3	80.9	81.7	77.1	94.9	83.2	62.8	70.0	71.9
	ZSCL	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	MoE	30.1	89.3	74.9	64.0	82.3	89.4	87.1	89.0	89.1	69.5	72.5	76.1
	Dual-RAIL	36.0	94.8	71.5	64.1	79.5	95.3	88.5	89.4	91.5	74.6	71.3	**77.9**
	ConDU(FT)	33.3	90.7	75.0	63.1	88.8	88.6	87.0	91.8	85.6	66.5	71.9	76.6
	ConDU(LoRA)	31.8	92.4	76.7	63.4	86.8	91.8	85.6	93.9	90.3	68.1	70.9	77.4

Experiments : task-agnostic MTIL

We evaluate our method on task-agnostic Multi-domain Task Incremental Learning benchmark

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
	Zero-shot	24.4	63.7	41.0	39.3	53.0	70.0	88.4	39.6	88.9	64.5	63.3	57.8
	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
Average	Continual-FT	25.5	81.5	59.1	53.2	64.7	51.8	63.2	64.3	69.7	31.8	49.7	55.9
Average	ZSCL	46.3	68.3	74.3	56.3	79.1	81.4	89.5	74.0	89.0	64.4	67.5	71.8
Average	MoE	37.2	65.3	79.5	67.6	19.7	83.1	80.5	74.0	88.5	67.5	65.3	66.2
Average	Primal-RAIL	42.4	88.5	57.1	55.7	64.7	80.7	83.0	62.9	84.8	68.7	63.7	68.4
Average	Dual-RAIL	45.0	88.8	57.8	56.8	66.2	81.0	85.2	63.4	87.8	68.9	64.7	69.6
Average	CoLeCLIP	48.2	77.8	71.7	65.7	76.8	83.8	89.6	72.2	90.3	68.0	66.4	73.7
Average	DPeCLIP	49.9	85.3	81.5	65.3	81.6	84.3	89.9	74.0	90.4	68.3	66.2	76.1
Average	ConDU(FT)	59.7	90.4	83.6	67.0	81.8	83.6	90.2	75.0	90.8	68.7	68.4	**78.1**
Average	ConDU(LoRA)	51.8	94.4	84.2	68.8	80.0	84.1	90.0	77.1	88.9	68.8	69.3	78.0
Last	Continual-FT	31.0	89.3	65.8	67.3	88.9	71.1	85.6	99.6	92.9	77.3	81.1	77.3
Last	ZSCL	42.5	64.4	67.2	54.8	89.7	90.4	91.7	95.8	93.4	85.2	78.3	77.6
Last	MoE	34.1	47.6	80.9	75.5	0.0	93.0	70.8	99.4	86.4	79.8	68.9	66.9
Last	Primal-RAIL	41.9	94.0	73.7	67.8	84.4	97.0	83.4	92.6	86.9	75.7	71.4	79.0
Last	Dual-RAIL	45.2	94.4	74.7	70.7	87.3	97.9	86.5	92.8	91.9	81.7	76.7	81.8
Last	CoLeCLIP	48.1	73.1	65.2	69.6	84.0	96.2	90.9	94.6	93.5	82.6	79.3	79.7
Last	DPeCLIP	49.9	84.2	83.2	71.1	97.0	95.8	92.0	99.4	93.9	84.5	80.2	84.6
Last	ConDU(FT)	58.6	90.8	86.3	74.0	96.3	93.4	91.9	99.6	94.7	84.9	80.1	**86.4**
Last	ConDU(LoRA)	48.4	94.4	87.3	77.1	94.1	94.3	90.8	96.2	90.8	84.3	78.1	85.1

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:*

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

Detailed Theoretical Analysis and Proof can be Found from Our Paper *https://arxiv.org/pdf/2503.10705*

Analysis : Visualization

We perform t-SNE visualization of features extracted from training data of 10 categories from Task1 (AirCraft)

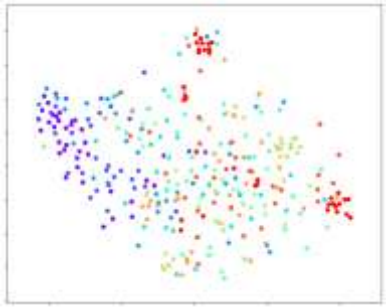


Fig 1:Pre-trained Model

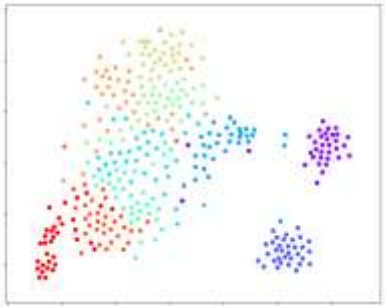


Fig 2:Session 1

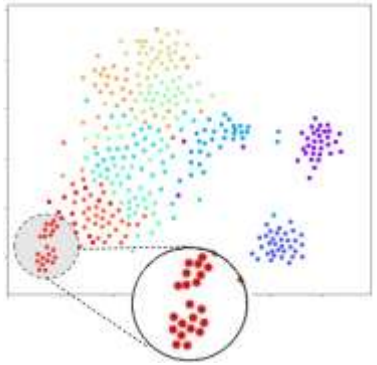


Fig 3:Session 1

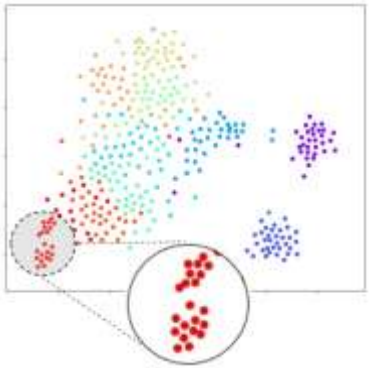


Fig 4:Session 6

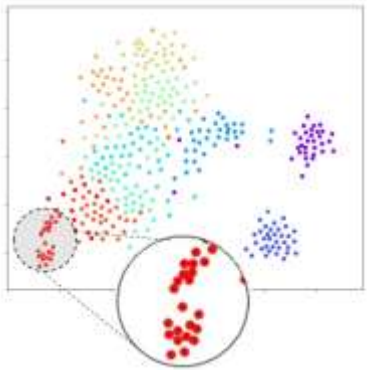


Fig 5:Session 11

- From Fig1 and Fig2, the fine-tuned task-specific model 1 shows significantly better data discrimination onTask1 compared to the pre-trained VLM
- Fig3 to Fig5 indicates that the task-specific model reconstructed by ConDU closely matches the representation ability of the model obtained through initial fine-tuning.

Thanks!