Continual Learning on Pretrained Foundation Models

Presentor: Zangwei Zheng

Outline

MTCL: A New Continual Learning Setup

Why **MTCL** for Foundation Models?

Why **Foundation Models** for CL?

Challenges of CL on Foundation Models

One-Step MTCL: **LLM Fine-tuning Revisited**

Multi-Step MTCL: CLIP and **ZSCL**

[ICCV'23] Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

MTCL: A New Continual Learning Setup

A review of continual learning setups

Setup	Input distribution	Data label space	Task label in testing	Task label in training
DIL	different	same	optional (limited domains)	optional
TIL	different	disjoint	required	required
CIL	different	disjoint	unavailable	required

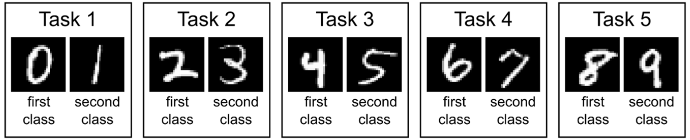


Figure 1: Schematic of split MNIST task protocol.

Table 2: Split MNIST according to each scenario.

Task-IL	With task given, is it the 1st or 2nd class? (e.g., 0 or 1)
Domain-IL	With task unknown, is it a 1st or 2nd class? (e.g., in $[0, 2, 4, 6, 8]$ or in $[1, 3, 5, 7, 9]$)
Class-IL	With task unknown, which digit is it? (i.e., choice from 0 to 9)

[1] Three scenarios for continual learning

Foundation models change the thing

Foundation model: a model that is pretrained on a **large-scale dataset** and can be **easily adapted** to downstream tasks. Often trained with **self-supervised learning**.

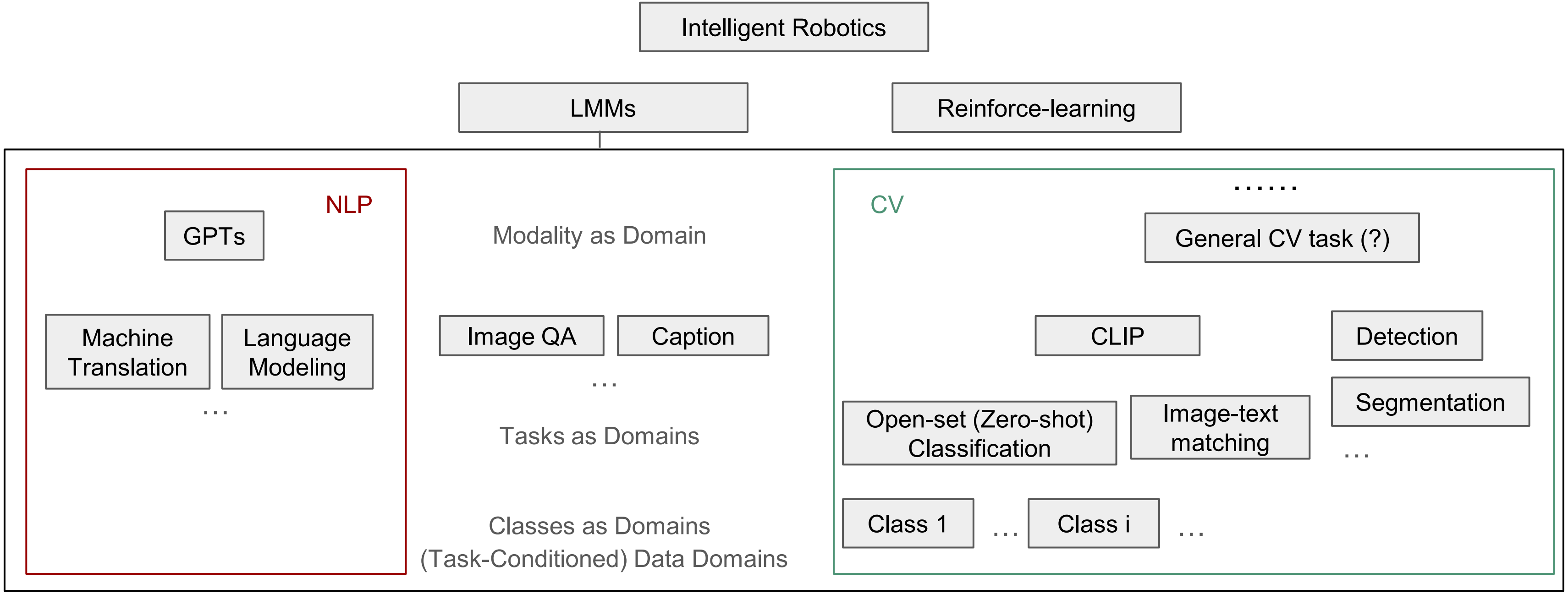
Tasks		**Tasks**
Language modeling, machine translation, sentiment classification, question answering ……	→	Next token prediction
	Unify	
Text retrieval ImageNet class classification Texture class classification (New class) classification ……	→	Image-text matching score prediction (Contrastive learning)

Foundation models change the thing

Muti-Domain Incremental Learning: learning new (sub)tasks which can be seen as a new domain for foundation models.

SubTasks		Tasks
Language modeling, machine translation, sentiment classification, question answering ……	prompt →	Next token prediction
Text/Image retrieval ImageNet class classification Texture class classification (New class) classification ……	Text template →	Image-text matching score prediction (Contrastive learning)

Task hierarchy



Muti-Domain Incremental Learning (MTIL)

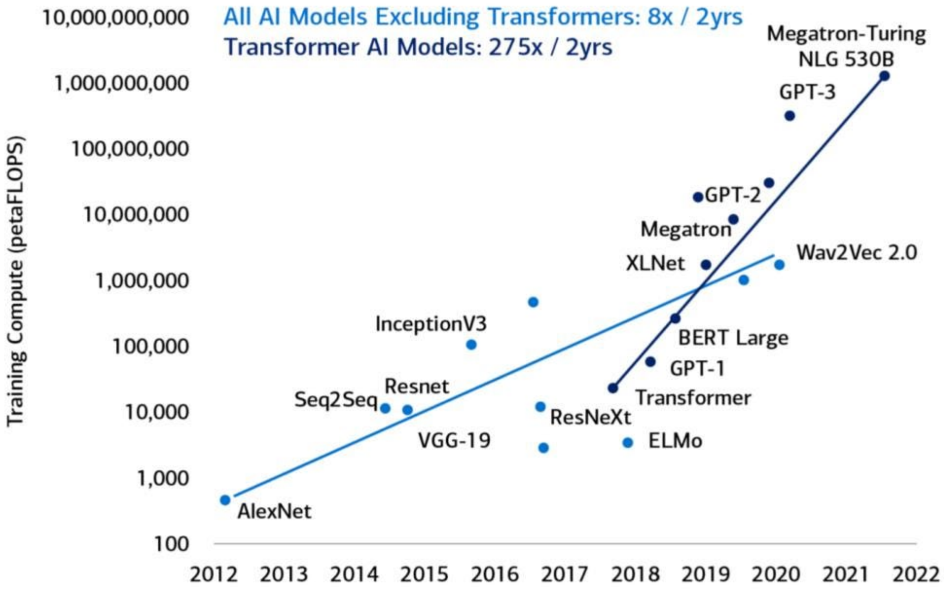
Setup	Input distribution	Data label space	Task label in testing	Task label in training
DIL	different	same	optional (limited domains)	optional
TIL	different	disjoint	required	required
CIL	different	disjoint	unavailable	required
MTIL	different	same	subtask encoded in input (unlimited domains)	subtask encoded in input

1. **Unlimited Domains**
2. **The first pre-training task contains overwhelming data**

Why MTCL for Foundation Model?

Foundation models are expensive

Exhibit 31: Transformer AI models require 275x more computing power every two years
Computational requirements for training transformers

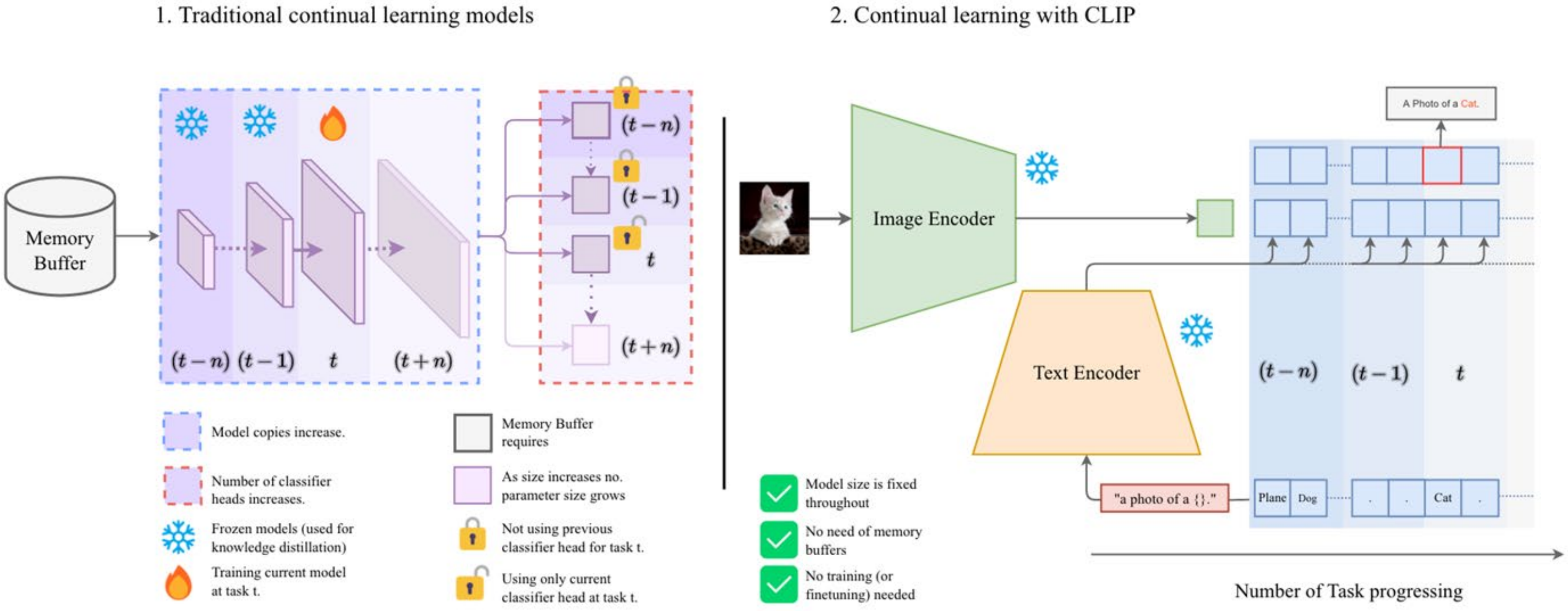


MTCL Applications for Foundation Models

1. **Adapt to the times and domains:** The foundation model needs to learn new knowledge as the world changes. In addition, applying foundation models into specific domains (e.g., medical, law) also requires the model to learn professional knowledge.
2. **Patching:** The foundation model can have factual errors, drawbacks, or bias. This is often caused by the dataset bias. For example, CLIP model has much worse performance on MNIST digits than a simple CNN model. We hope to add, delete, or modify the knowledge in the foundation model to fix the problem.
3. **Alignment:** Aligning foundation models (e.g., instruction tunning, multimodality fine-tunning) can also be viewed as a continual learning problem. The distribution of instruction data is different from the distribution of the pretraining dataset. We hope the model can learn the human values without forgetting the pre-training knowledge.

Why Foundation Model for MTCL?

Training from scratch cannot scale (params & data)



[1] CLIP Model is an Efficient Continual Learner

Training from scratch cannot scale

Traditional View of Continual Learning

Math

Physics

Art

……

In very early age of human beings, we are learning many things together (multi-task learning) and self-supervised.

Walking

Eating

Speaking

Recognizing

After learning general knowledge (pretraining), we learn specific ones

Math

Art

……

[1] CLIP Model is an Efficient Continual Learner

Training from scratch cannot scale

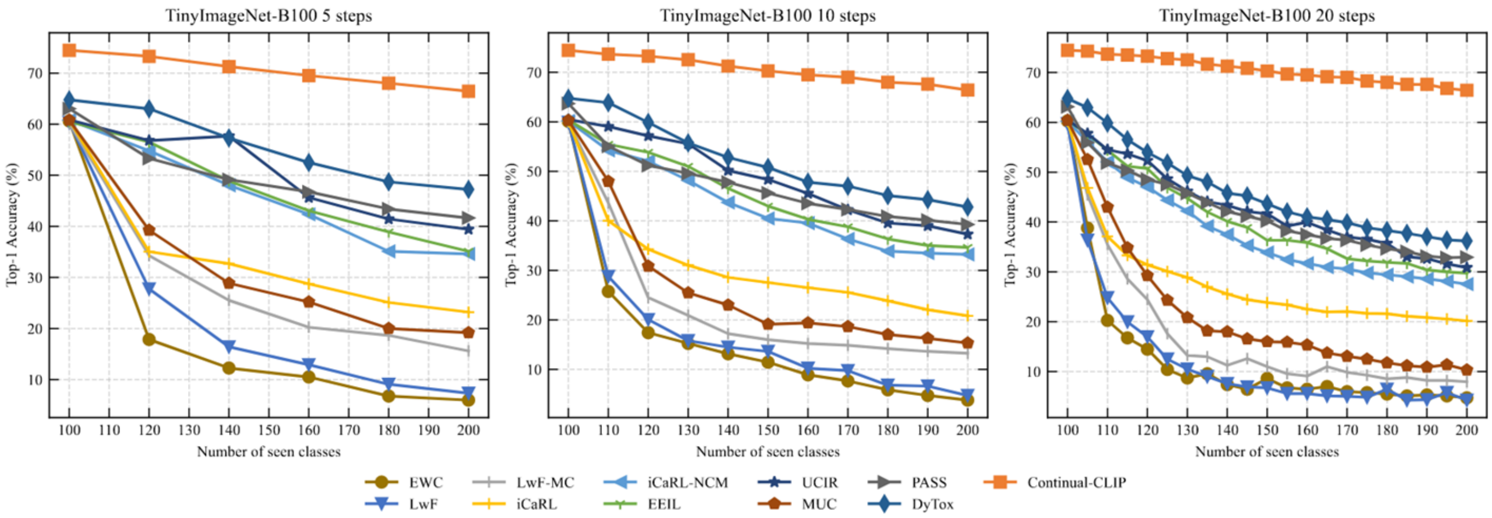


Figure 3: Accuracy trends on the TinyImageNet dataset in three different settings of step sizes. Note that the other competing methods require training at each stage, use memory buffers, may not apply to all CL settings and/or dynamically expand the architecture to learn new tasks.

[1] CLIP Model is an Efficient Continual Learner

Challenge of CL on Foundation Models

Catastrophic Forgetting

A great reduction in performance on old tasks when learning new tasks

Stability ⟷ **Plasticity**

Generality ⟷ **Speciality**

1. **Forgetting of General Knowledge:** Infinite domains
2. **Forgetting of Newly Learned Knowledge:** Multi-step setups
3. **Forgetting of Generalization Ability for Newly Learned Knowledge:** A special case of case 1. Here we care about if models can generalize well to the same task with different domains (e.g., ImageNet variants).

Evaluation Forgetting for one-step MTCL on LLM

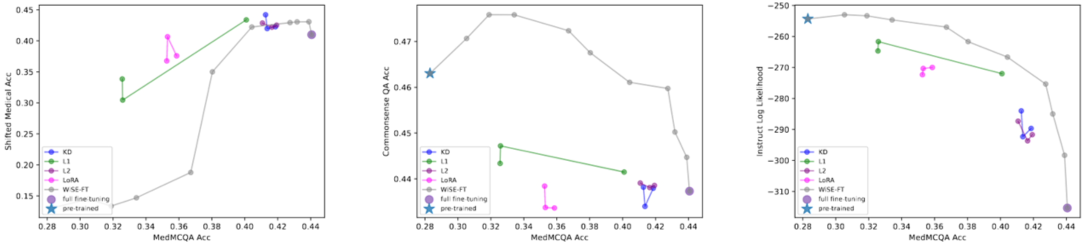
A great reduction in performance on old tasks when learning new tasks

Set	Elements
DK	STEM, Social, Human, Other
Rs	BoolQ, PIQA, Winogrande, Hellaswag, MathQA Mutual
RC	RACE-high, RACE-middle
Bias	Sexual Orientation, Physical Appearance, Religion Nationality, Race/Color, Gender, Socioeconomic Disability, Age

An better choice may be BIG-BENCH

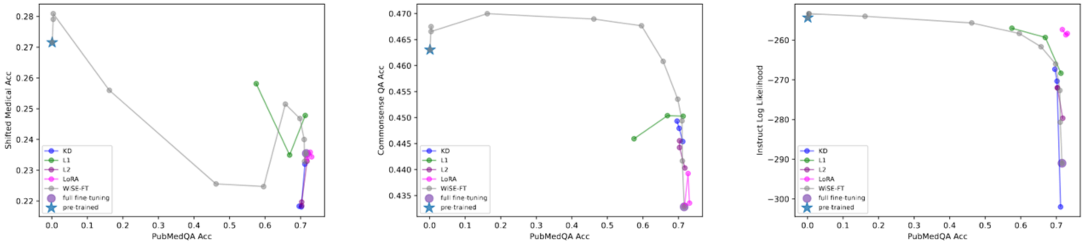
Task Type	Dataset Name	Example
Medical	PubMedQA [34]	*Context*: Middle third clavicular fracture ... ? *Question*: Does comminution play no role in treated middle third clavicular fracture? *Output*: yes
	MedMCQA [52]	*Question*: Severe painful sensorimotor and autonomic neuropathy along with alopecia may suggest poisoning with: (A) Thallium (B) Arsenic (C) Lead (D) Copper. *Output*: A
	MedQA-USMLE [33]	*Question*: A 23-year-old pregnant woman at 22 weeks... Which of the following is the best treatment for this patient? (A) Ampicillin, (B) Ceftriaxone, (C) Doxycycline, (D) Nitrofurantoin. *Output*:B
Common Sense	ARC Easy [16]	*Question*: What carries oxygen throughout the body? (A) white blood cells, (B) brain, (C) red blood cells, (D) nerves *Output*: C
	ARC Challenge [16]	*Question*: Which technology was developed most recently? (A) cellular telephone, (B) television, (C) refrigerator, (D) airplane. *Output*: A
	Race [39]	*Passage*: The rain had continued for a week, ... *Question*: What did Nancy try to do before she fell over? (A) Measure the depth, (B) Look for a tree trunk, (C) Protect her cows, (D) Run away *Answer*: C
	PIQA [5]	*Goal*: When boiling butter, when it's ready, you can (Sol1) Pour it onto a plate, (Sol2) Pour it into a jar, *Answer*:Sol1
Instruction	Alpaca [71]	*Instruction*: Give three tips for staying healthy. *Output:* 1. Eat a balanced diet. 2. Exercise regularly. 3. ....
	GPT4 instruct [54]	*Input*: Compare and contrast the effects of individual ...? *Output*: Individual performance refers to ...
	LMFlow [20]	*Human*: I think the biggest thing is that it's in her smile. *Assistant*: That sounds very comforting... *Human*: Ok, can you remind me to change scenes ? *Assistant*: Sure, it's important to change scenes every ...

[1] Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models
[2] An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning



(a) OOD Medical (b) Common Sense (c) Instruct

Figure 5: Fine-tune on MedMCQA. We evaluate the forgetting in terms of (a) distribution generality forgetting on the other two medical QA datasets including PubMedQA and MedQA-USMLE, (b) task generality forgetting on common sense tasks including ARC Easy and Challange, Race, and PIQA (c) instruction following tasks including Alpaca, GPT4 instruct and LMFlow.



(a) OOD Medical (b) Common Sense (c) Instruct

Figure 6: Fine-tune on PubMedQA. We evaluate the forgetting in terms of (a) distribution generality forgetting on the other two medical QA datasets including MedMCQA and MedQA-USMLE, (b) task generality forgetting on common sense tasks including ARC Easy and Challange, Race and PIQA (c) instruction following tasks including Alpaca, GPT4 instruct and LMFlow.

1. Different method may suit different setting.
2. The forgetting of LLM is more severe on the tasks that is significantly different from the fine-tuning task.

Evaluation Forgetting for one-step MTCL on LLM

Different models all suffer from catastrophic forgetting and larger models suffer more.

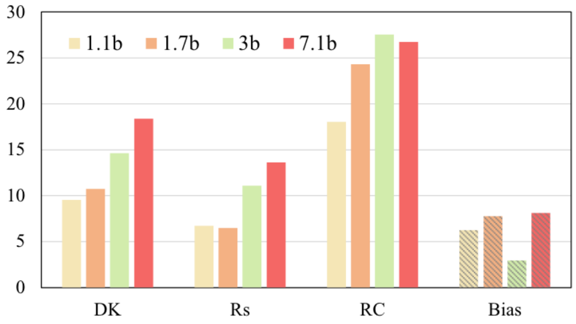
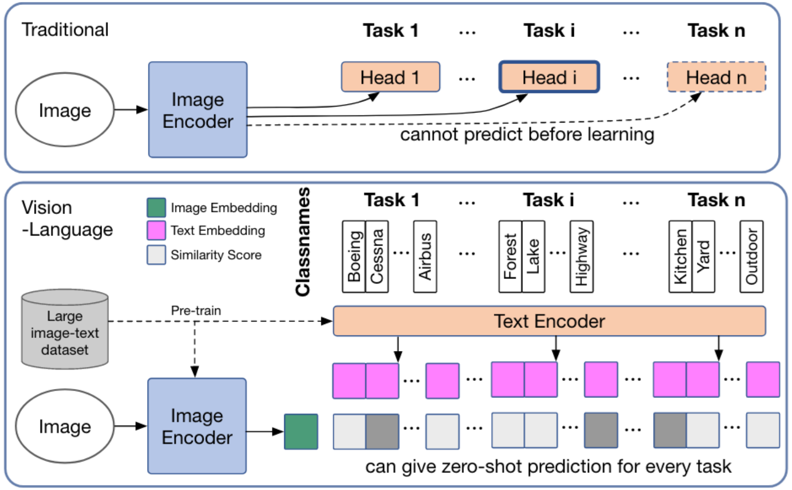


Figure 2: FG values of domain knowledge, reasoning, and reading comprehension in BLOOMZ with respect to different scales.

[1] Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models

Evaluation Forgetting for multi-step MTCL on LMM

1. **Forgetting of General Knowledge:** Infinite domains
2. **Forgetting of Newly Learned Knowledge:** Multi-step setups



(a) Comparison between traditional CL and CL with a pre-trained vision-language model

[1] Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

Evaluation Forgetting for multi-step MTCL on LMM

1. **Forgetting of General Knowledge:** Infinite domains
2. **Forgetting of Newly Learned Knowledge:** Multi-step setups

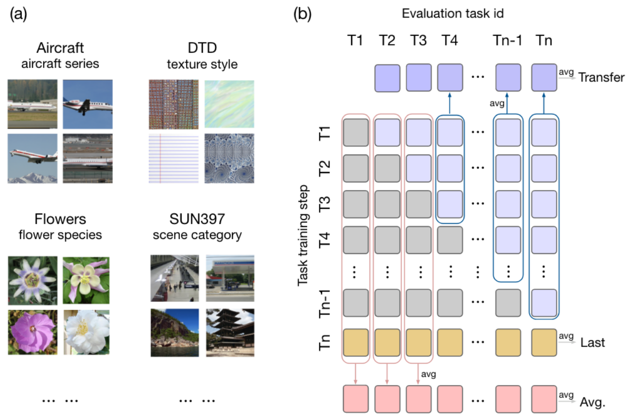


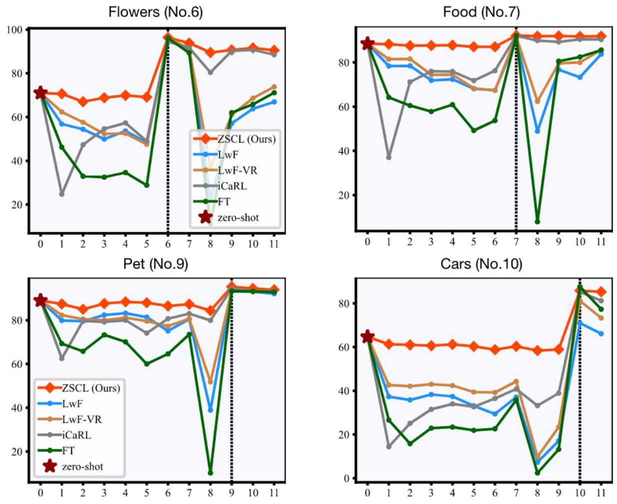
Figure 4. Fig.(**a**): examples of tasks from different domains in MTIL benchmark. Fig.(**b**): illustration of calculating metrics Transfer, Avg. and Last during continual learning.

Dataset	# classes	# train	# test	Recognition Task
Aircraft [44]	100	3334	3333	aircraft series
Caltech101 [17]	101	6941	1736	real-life object
CIFAR100 [31]	100	50000	10000	real-life object
DTD [6]	47	1880	1880	texture recognition
EuroSAT [20]	10	21600	5300	satellite location
Flowers [47]	102	1020	6149	flower species
Food [3]	101	75750	25250	food type
MNIST [10]	10	60000	10000	digital number
OxfordPet [50]	37	3680	3669	animal species
StanfordCars [30]	196	8144	8041	car series
SUN397 [71]	397	87003	21751	scene category
Total	1201	319352	97109	

[1] Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

Evaluation Forgetting for multi-step MTCL on LMM

1. **Forgetting of General Knowledge:** Infinite domains
2. **Forgetting of Newly Learned Knowledge:** Multi-step setups



(b) Performance of different methods on preventing forgetting phenonmenon

[1] Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

One-Step MTCL: LLM Finetuning Revisited

LLM fine-tuning v.s. CL method categories

Instruction tunning is a typical one-step continual learning setting. From this perspective, we can introduce more CL methods for instruction tunning.

CL Method Categories	SFT Method
Feature-based	KD, PTX
Weight-based	L1 ($\|\theta - \theta_0\|$) and L2 penalty, WiSE-FT
Architecture-based	LoRA, Prompt Tunning
Optimization-based	PPO

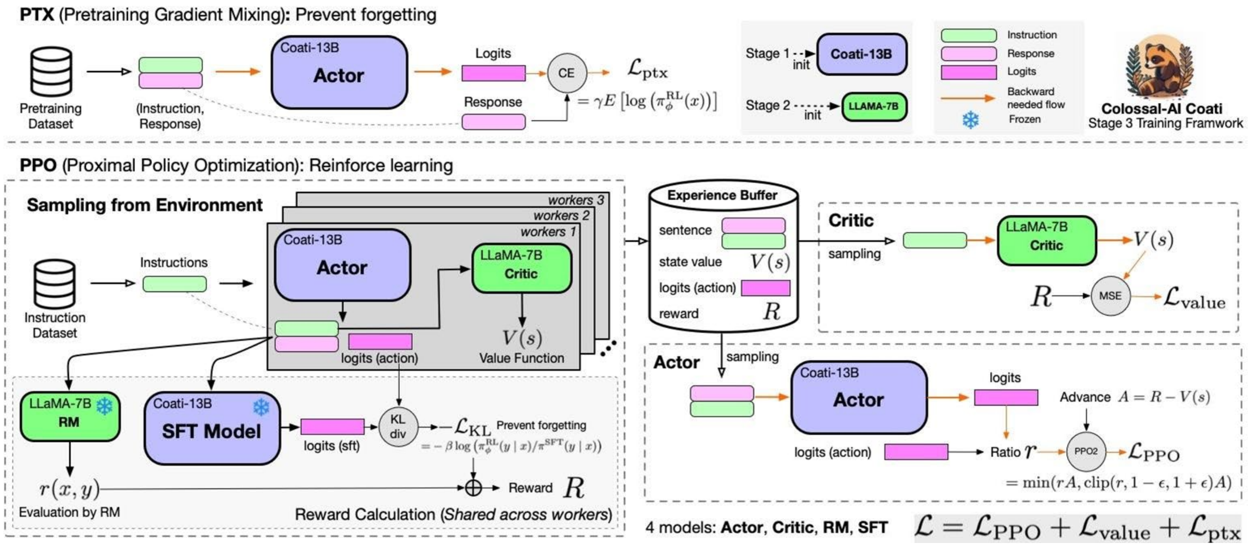
Quick Reviews

CL Method Categories	SFT Method
Feature-based	KD, PTX
Weight-based	L1 ($\lvert\theta - \theta_0\rvert$) and L2 penalty, WiSE-FT
Architecture-based	LoRA, Prompt Tunning
Optimization-based	PPO

Feature-based: stabilize in the output feature space

Knowledge Distillation (K $\|f_\theta(x) - f_{\theta_0}(x)\|_2^2$

Pretraining Gradient Mixing: KL divergence (in RLHF)

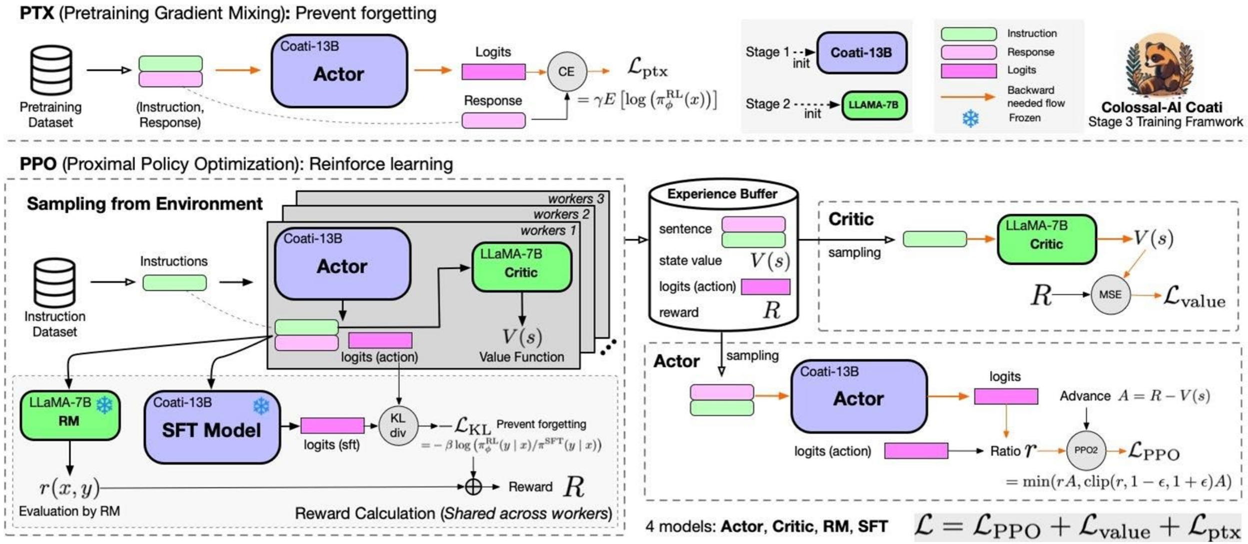


Quick Reviews

CL Method Categories	SFT Method
Feature-based	KD, PTX
Weight-based	L1 ($\|\theta - \theta_0\|$) and L2 penalty, WiSE-FT
Architecture-based	LoRA, Prompt Tunning
Optimization-based	PPO

Gradient-based:

PPO methods

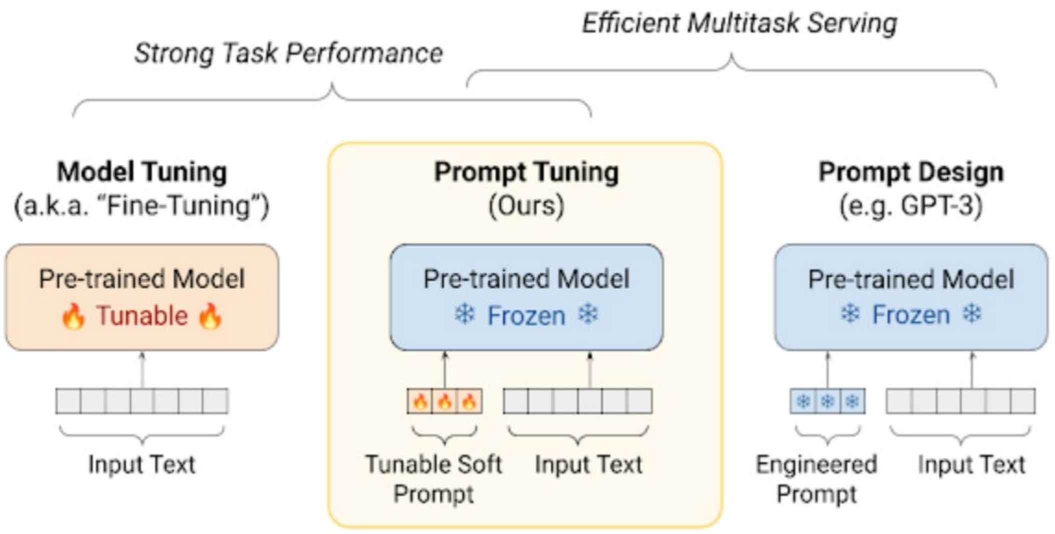


Quick Reviews

CL Method Categories	SFT Method
Feature-based	KD, PTX
Weight-based	L1 ($\|\theta - \theta_0\|$) and L2 penalty, WiSE-FT
Architecture-based	LoRA, Prompt Tunning
Optimization-based	PPO

Architecture-based: many parameter-efficient (PEFT) methods fall into this

LoRA, Prompt Tunning



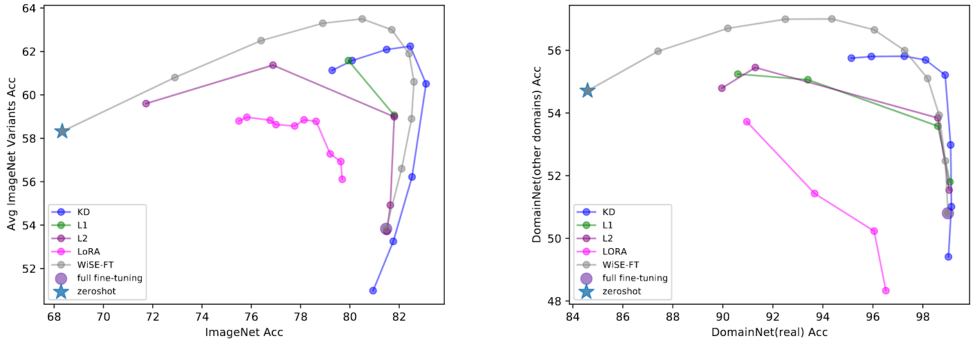
Quick Reviews

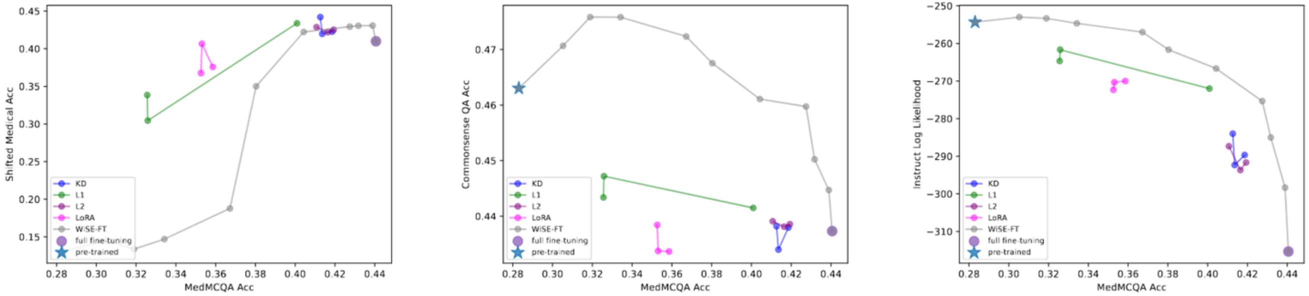
CL Method Categories	SFT Method
Feature-based	KD, PTX
Weight-based	L1 ($\|\theta - \theta_0\|$) and L2 penalty, WiSE-FT
Architecture-based	LoRA, Prompt Tunning
Optimization-based	PPO

Weight-based:

L1, L2 penalty

WiSE-$\theta = \alpha\theta_0 + (1-\alpha)\theta_1$





(a) OOD Medical (b) Common Sense (c) Instruct

Multi-Step MTCL: CLIP and ZSCL

ZSCL: Zero-shot Continual Learning Method

1. The first work to investigate the multi-step MTCL on LMMs.
2. Call for more attention on the prevention of zero-shot transfer degradation.
3. Combine **feature-based** and **weight-based** CL methods.

Feature-based Method

Knowledge Distillation Methods

- LwF: data from current task
- iCarl: data from previous task
- Ours: data from publicly available datasets

Table 1. **Ablation experiments**. Default settings are marked in gray, which uses image and text distillation loss with the initial CLIP model on 100k ImageNet images and texts generated from ImageNet classes with a simple template.

(a) **Continual learning loss**.

loss	Transfer	Avg.	Last
CE only	44.6	55.9	77.3
Feat. Dist.	47.6	58.7	77.1
Image-only	56.5	68.9	82.1
Text-only	56.7	69.0	82.6
Both	**56.8**	**69.2**	**83.0**

(b) **Data sources for replay**.

source	Transfer	Avg.	Last
current	56.7	66.5	80.2
ImageNet	56.8	**69.2**	**83.0**
CC	**57.2**	68.5	80.9
CIFAR100	55.2	65.9	80.7
Flowers	54.7	66.0	80.8

(c) **Text sources for replay**.

source	Transfer	Avg.	Last
current	51.8	64.9	82.0
prev. all	54.0	70.2	83.7
1k classes (IN)	56.8	69.2	83.0
13k Sent. (CC)	**58.9**	**70.5**	**84.0**
1k Rand. Sent.	58.7	70.2	83.8

(d) **Teacher model**.

source	Transfer	Avg.	Last
Initial	**56.8**	**69.2**	**83.0**
$n-1$	53.9	66.6	80.7
WiSE(0.5)	56.4	68.9	82.9
WiSE(0.8)	56.2	67.8	81.3

(e) **# images for replay**.

#image	Transfer	Avg.	Last
1M	**58.7**	**70.1**	**83.2**
100k	56.8	69.2	83.0
10k	57.8	68.7	81.2
1k	56.3	67.6	80.8

(f) **# image classes for replay**.

#class	Transfer	Avg.	Last
1000	**56.8**	**67.6**	**83.0**
100	56.7	67.3	82.3
10	53.8	66.4	81.0
1	53.1	65.5	80.5

Feature-based Method

Table 1. **Ablation experiments**. Default settings are marked in gray, which uses image and text distillation loss with the initial CLIP model on 100k ImageNet images and texts generated from ImageNet classes with a simple template.

(a) **Continual learning loss**.

loss	Transfer	Avg.	Last
CE only	44.6	55.9	77.3
Feat. Dist.	47.6	58.7	77.1
Image-only	56.5	68.9	82.1
Text-only	56.7	69.0	82.6
Both	**56.8**	**69.2**	**83.0**

(b) **Data sources for replay**.

source	Transfer	Avg.	Last
current	56.7	66.5	80.2
ImageNet	56.8	**69.2**	**83.0**
CC	**57.2**	68.5	80.9
CIFAR100	55.2	65.9	80.7
Flowers	54.7	66.0	80.8

(c) **Text sources for replay**.

source	Transfer	Avg.	Last
current	51.8	64.9	82.0
prev. all	54.0	70.2	83.7
1k classes (IN)	56.8	69.2	83.0
13k Sent. (CC)	**58.9**	**70.5**	**84.0**
1k Rand. Sent.	58.7	70.2	83.8

(d) **Teacher model**.

source	Transfer	Avg.	Last
Initial	**56.8**	**69.2**	**83.0**
$n-1$	53.9	66.6	80.7
WiSE(0.5)	56.4	68.9	82.9
WiSE(0.8)	56.2	67.8	81.3

(e) **# images for replay**.

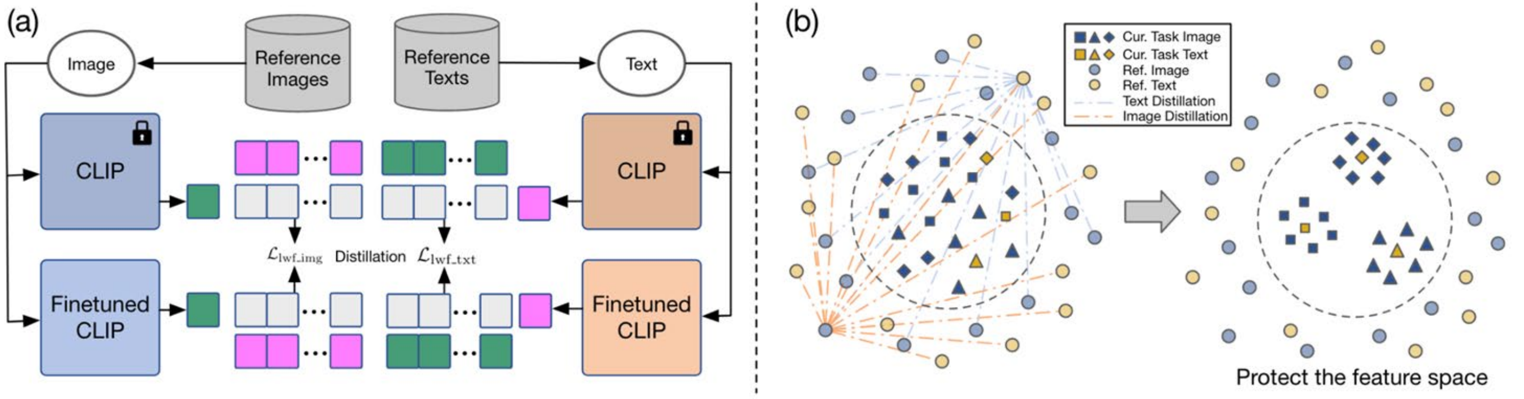
#image	Transfer	Avg.	Last
1M	**58.7**	**70.1**	**83.2**
100k	56.8	69.2	83.0
10k	57.8	68.7	81.2
1k	56.3	67.6	80.8

(f) **# image classes for replay**.

#class	Transfer	Avg.	Last
1000	**56.8**	**67.6**	**83.0**
100	56.7	67.3	82.3
10	53.8	66.4	81.0
1	53.1	65.5	80.5

1. Applying KD on both image and text is better than applying on one of them only.
2. Use the original model instead of newly trained model as the teacher model.
3. Images and texts with more diverse semantics are better for KD, even without the need of pairing them. This greatly reduces the cost of data collection.

Feature-based Method



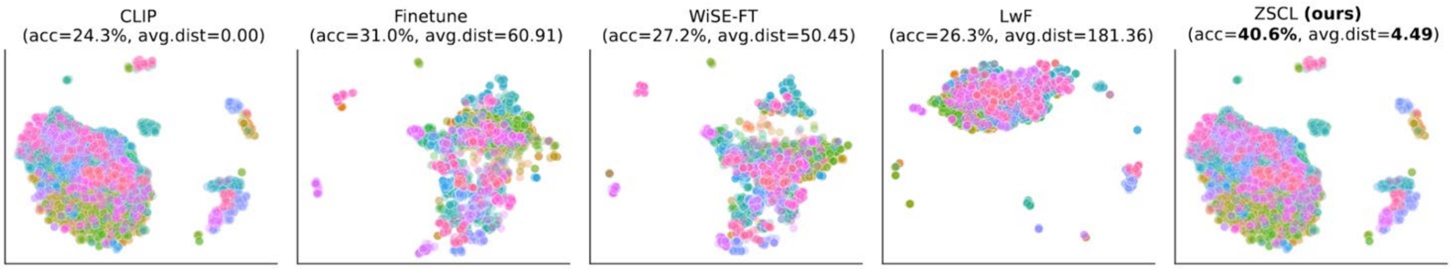
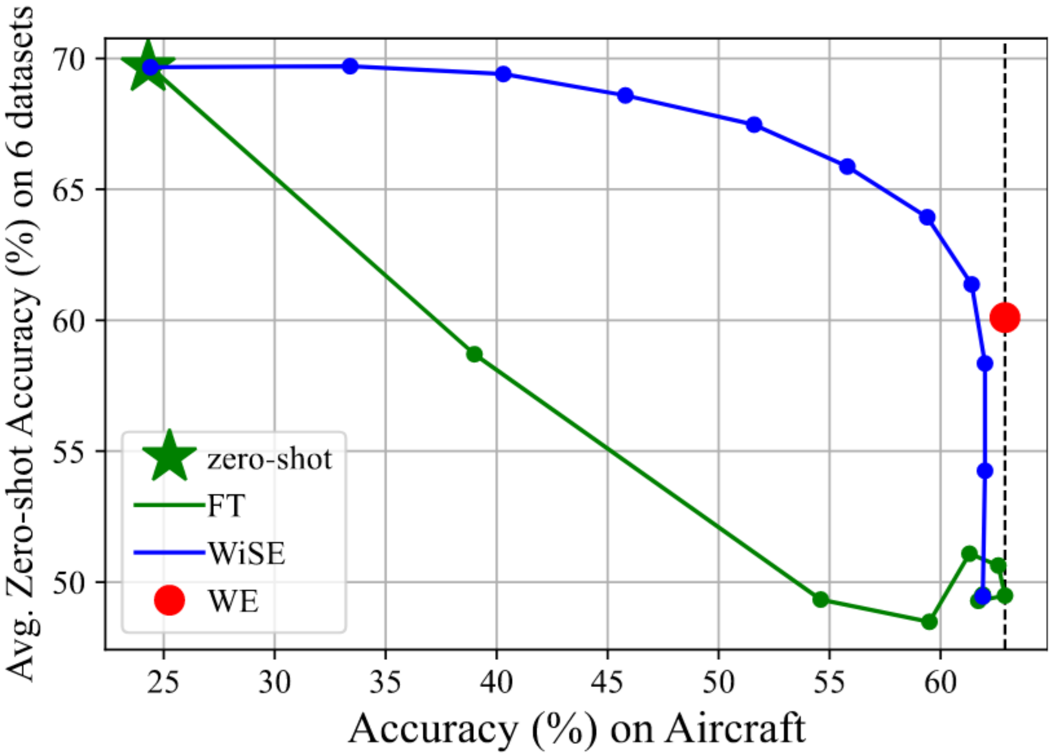


Figure 5. t-SNE on five models' outputs together of Aircraft datasets after MTCL training: only our model maintains a similar feature distribution to the original CLIP ones with minor shift, while the rest significantly distort the feature space.

Weight-based Method



ZSCL

Distillation on feature space + Weight Ensemble + L1-norm

Table 2. Ablation study of different components for ZSCL.

Method	Transfer	Δ	Avg.	Δ	Last	Δ
CLIP `ViT-B/16@224px`						
Zero-shot	69.4	0.0	65.3	0.0	65.3	0.0
Continual Learning	44.6	-24.8	55.9	-9.4	77.3	+12.0
+ Distillation	58.9	-10.5	70.5	+5.2	83.8	+18.5
+ WiSE-FT (best α)	61.7	-7.7	71.6	+6.3	83.3	+18.0
+ WE (ZSCL*)	62.2	-7.2	72.6	+7.3	**84.5**	**+19.2**
+ WC	67.6	-1.8	74.5	+9.2	83.2	+17.9
+ WiSE-FT	67.7	-1.7	74.2	+8.9	81.9	+16.6
+ WE (ZSCL)	**68.1**	**-1.3**	**75.4**	**+10.1**	83.6	+18.3

ZSCL

Distillation on feature space + Weight Ensemble + L1-norm

Table 3. Comparison of different methods on MTIL in Order I.

Method	Transfer	Δ	Avg.	Δ	Last	Δ
CLIP `ViT-B/16@224px`						
Zero-shot	69.4	0.0	65.3	0.0	65.3	0.0
Continual Learning	44.6	-24.8	55.9	-9.4	77.3	+12.0
LwF [39]	56.9	-12.5	64.7	-0.6	74.6	+9.0
iCaRL [57]	50.4	-19.0	65.7	+0.4	80.1	+14.8
LwF-VR [13]	57.2	-12.2	65.1	-0.2	76.6	+11.3
WiSE-FT [69]	52.3	-17.1	60.7	-4.6	77.7	+12.4
ZSCL* (Ours)	62.2	-7.2	72.6	+7.3	**84.5**	**+19.2**
ZSCL (Ours)	**68.1**	**-1.3**	**75.4**	**+10.1**	83.6	+18.3

Table 4. Comparison of different methods on MTIL in Order II.

Method	Transfer	Δ	Avg.	Δ	Last	Δ
CLIP `ViT-B/16@224px`						
Zero-shot	65.4	0.0	65.3	0.0	65.3	0.0
Continual Learning	46.6	-18.8	56.2	-9.1	67.4	+2.1
LwF [39]	53.2	-12.2	62.2	-5.2	71.9	+6.6
iCaRL [57]	50.9	-14.5	56.9	-8.4	71.6	+6.3
LwF-VR [13]	53.1	-12.3	60.6	-7.4	68.3	+0.9
WiSE-FT [69]	51.0	-14.4	61.5	-5.9	72.2	+6.9
ZSCL*	59.8	-5.6	71.8	+6.5	83.3	+18.0
ZSCL	**64.2**	**-1.2**	**74.5**	**+9.2**	**83.4**	**+18.1**

ZSCL

Distillation on feature space + Weight Ensemble + L1-norm

Table 6. Comparison of state-of-the-art CL methods on CIFAR100 benchmark in class-incremental setting.

Methods	10 steps Avg	10 steps Last	20 steps Avg	20 steps Last	50 steps Avg	50 steps Last
UCIR [23]	58.66	43.39	58.17	40.63	56.86	37.09
BiC [70]	68.80	53.54	66.48	47.02	62.09	41.04
RPSNet [56]	68.60	57.05	-	-	-	-
PODNet [15]	58.03	41.05	53.97	35.02	51.19	32.99
DER [72]	74.64	64.35	73.98	62.55	72.05	59.76
DyTox+ [16]	74.10	62.34	71.62	57.43	68.90	51.09
CLIP [54]	74.47	65.92	75.20	65.74	75.67	65.94
FT	65.46	53.23	59.69	43.13	39.23	18.89
LwF [39]	65.86	48.04	60.64	40.56	47.69	32.90
iCaRL [57]	79.35	70.97	73.32	64.55	71.28	59.07
LwF-VR [13]	78.81	70.75	74.54	63.54	71.02	59.45
ZSCL (Ours)	**82.15**	**73.65**	**80.39**	**69.58**	**79.92**	**67.36**
Impr	+7.68	+7.73	+5.19	+3.84	+3.95	+1.42

Table 7. Comparison of different methods on TinyImageNet splits in class-incremental settings with 100 base classes.

Methods	5 steps Avg	5 steps Last	10 steps Avg	10 steps Last	20 steps Avg	20 steps Last
EWC [29]	19.01	6.00	15.82	3.79	12.35	4.73
EEIL [5]	47.17	35.12	45.03	34.64	40.41	29.72
UCIR [23]	50.30	39.42	48.58	37.29	42.84	30.85
MUC [41]	32.23	19.20	26.67	15.33	21.89	10.32
PASS [77]	49.54	41.64	47.19	39.27	42.01	32.93
DyTox [16]	55.58	47.23	52.26	42.79	46.18	36.21
CLIP [54]	69.62	65.30	69.55	65.59	69.49	65.30
FT	61.54	46.66	57.05	41.54	54.62	44.55
LwF [39]	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL [57]	77.02	70.39	73.48	65.97	69.65	64.68
LwF-VR [13]	77.56	70.89	74.12	67.05	69.94	63.89
ZSCL (Ours)	**80.27**	**73.57**	**78.61**	**71.62**	**77.18**	**68.30**
Impr	+10.65	+8.27	+9.06	+6.03	+7.69	+3.00

Takeaways

1. MTCL is a new continual learning setup for foundation models. It is different from traditional continual learning setups as it has unlimited number of domains and blurred task boundaries.
2. Foundation models suffer from catastrophic forgetting during continual learning. LLM instruction tunning is a typical example.
3. LLMs and LMMs can benefit from both feature-based and weight-based CL methods.

Takeaways

1. MTCL is a new continual learning setup for foundation models. It is different from traditional continual learning setups as it has unlimited number of domains and blurred task boundaries.
2. Foundation models suffer from catastrophic forgetting during continual learning. LLM instruction tunning is a typical example.
3. LLMs and LMMs can benefit from both feature-based and weight-based CL methods.

Project homepage

Thank you for your listening!