Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters

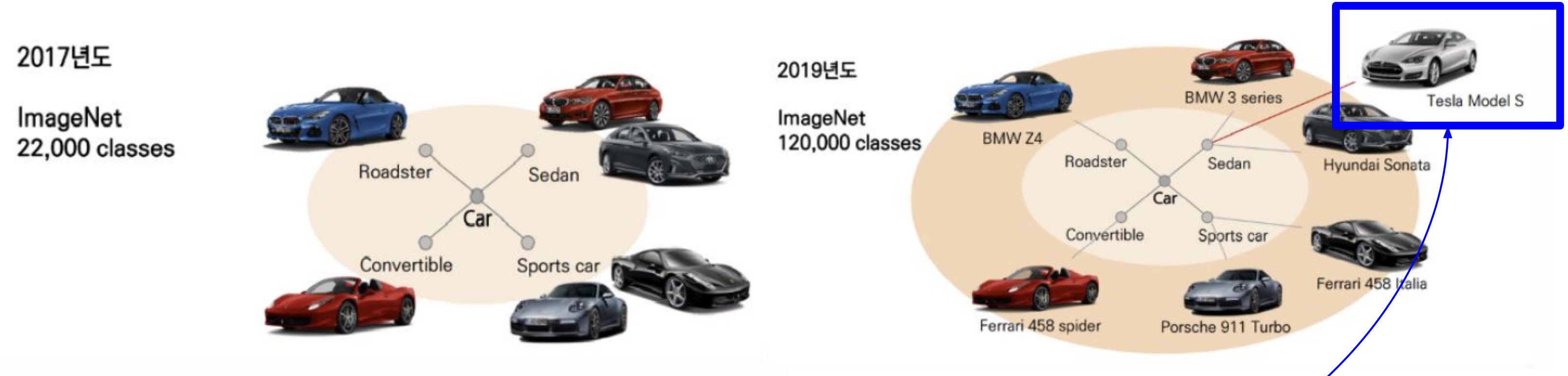
Jiazuo Yu[1], Yunzhi Zhuge[1], Lu Zhang[1,*], Ping Hu[2], Dong Wang[1], Huchuan Lu[1] and You He[3]
[1] Dalian University of Technology, China
[2] University of Electronic Science and Technology of China
[3] Tsinghua University, China
yujiazuo@mail.dlut.edu.cn, zhangluu@dlut.edu.cn

- Problem / objective
  - Empower the continual learning capabilities of vision-language models
    - Historical knowledge memorization
    - Zero-shot generalization
- Contribution / Key idea
  - Parameter-efficient continual learning framework
    - Mixture-of-Experts (MoE) adapters
    - Distribution Discriminative Auto-Selector (DDAS)

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

- **Continual Learning**

딥러닝 모델이 새로운 데이터에 대해 지속적으로 학습을 이어가며 지식을 확장하는 방식.



새로운 데이터가 나올때마다 처음부터 다시 학습하는 것이 아니라, 이미 학습된 모델에 새로운 데이터만 추가적으로 더 학습

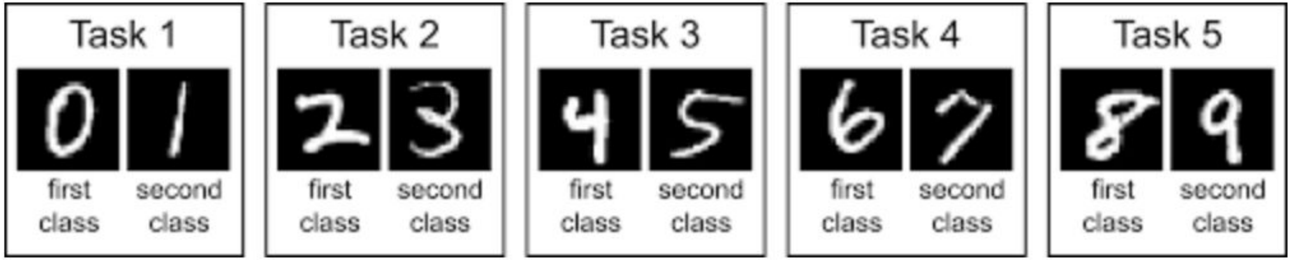
Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

CVPR 2024

- **Continual Learning 의 근본적인 문제 : "Catastrophic Forgetting"**

- 모델이 새로운 task 들을 점진적으로 학습함에 따라, 이전에 학습했었던 지식들을 점차 잊어버림.
- 이 문제를 해결하는 것이 Continual Learning 의 핵심.

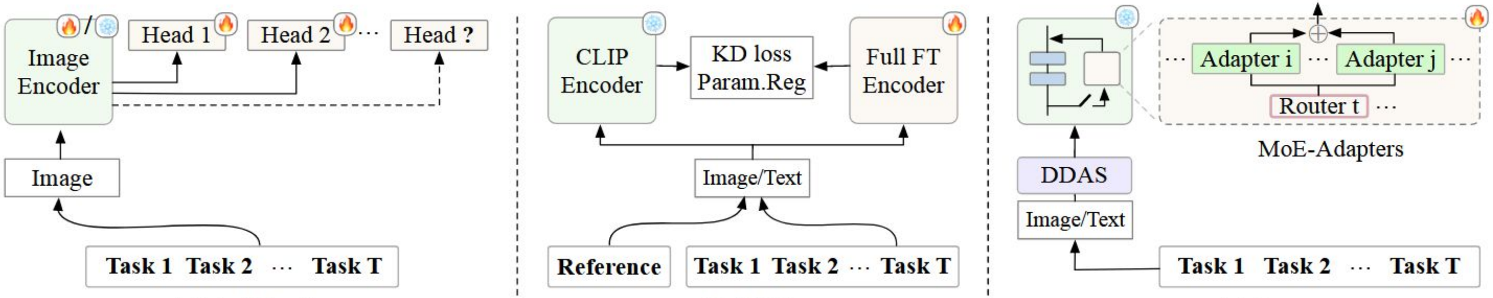
❑ [예시] MNIST 데이터셋을 5개의 task 로 나누어 continual learning 하였을때,



1. 처음에 task1 에서는 0과 1 구분하도록 모델을 학습.
2. 그다음 이 학습된 모델에 2와 3 구분하도록 또다시 학습.
3. ...
4. 8과 9 구분하는 마지막 task5 까지 모델 학습 마치고 나면,
5. 가장 맨 처음에 학습되었던 0과 1 구분하는 task1 에 대한 정확도가 상당히 낮아짐.

전유진

- **Continual Learning 선행 연구들**



(a) Traditional (b) ZSCL (c) Ours

Figure 1. Comparison of various popular architectures to address CL. (a) Traditional dynamic expansion-based CL cannot distinguish unseen data. (b) Zero-shot CL [78] suffers from significant computational burdens. (c) The proposed MoE-Adapters and DDAS collaborate to form a parameter-efficient, zero-shot CL.

(a) 대표적인 CL 방법: base model 에 추가로, task 마다 task-specific 모델 추가(dynamic expansion). -> 문제: zero-shot 못함.

(b) ZSCL: pretrained VLM 로부터 knowledge distillation 하여 zero-shot 능력 보완. -> 문제: 계산량 많고, 장기 기억 어려움.

(c) Ours: (b) 처럼 사전학습된 모델을 사용하여 (a) 의 dynamic expansion 기법을 적용하여, (a) 의 장점인 '기억력', (b) 의 장점인 'zero-shot 능력' 모두 살리겠다

전유진

- **Overall framework**

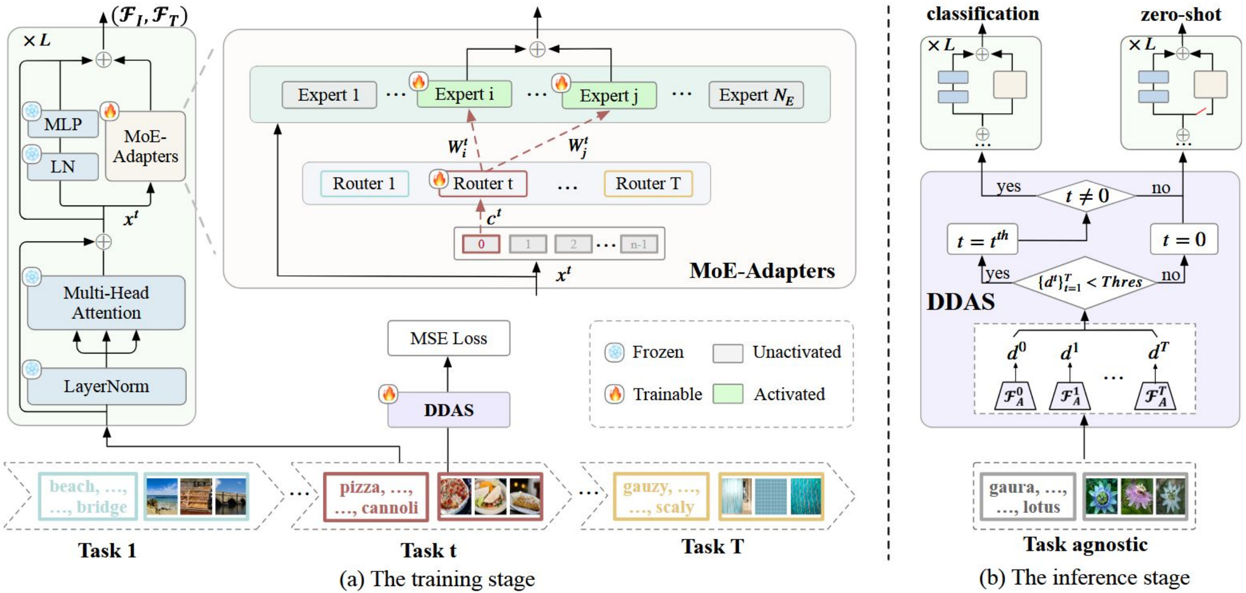
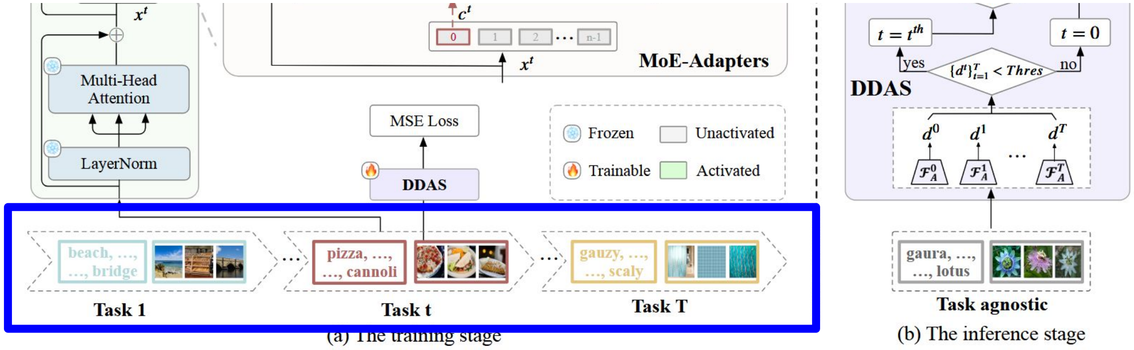


Figure 2. Overall framework of the proposed method. (a) At the training stage, CLIP's image and text encoders $(\mathcal{F}_I, \mathcal{F}_T)$ take input samples from **Task** $\boldsymbol{t}$. In each of transformer blocks, there is a MoE-Adapters, whose input is the tokens $\boldsymbol{x}^t$ from MHSA. The router takes the task-specific `[CLS]` token $\boldsymbol{c}^t$ as input and produces experts' weights W_i^t and W_j^t to combine the expert's output. DDAS is trained using only images via the MSE loss defined by Eq. 3. (b) At the inference stage, the proposed DDAS determines the data distribution by comparing the distribution $\{d^t\}_{t=1}^T$ in each autoencoder of the **task-agnostic** images. It can automatically assign the testing data into MoE-Adapters or original CLIP to predict with either seen or unseen data.

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

CVPR 2024

- **Continual Learning**

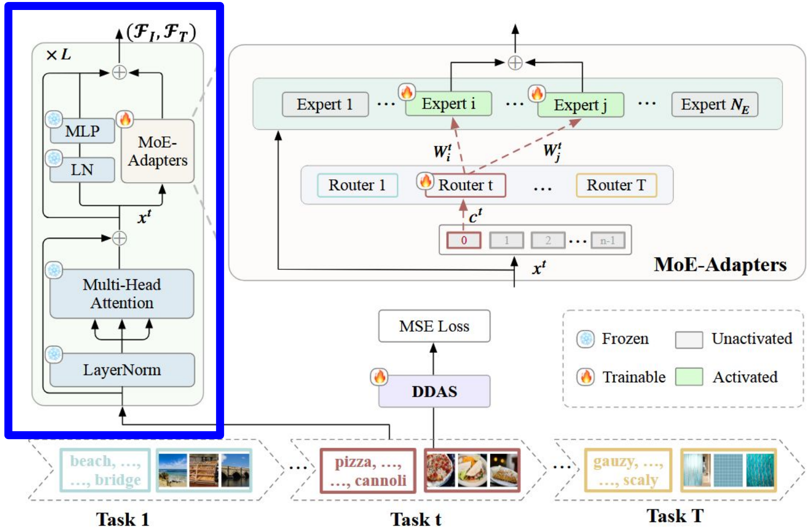


- 목표: 모든 task 에서 성능 잘 나오기
- T개의 task들: $\{\mathcal{T}^t\}_{t=1}^{T}$
  - t번째 task: $\mathcal{T}^t = \{\mathcal{D}^t, \mathcal{C}^t\}$ - 데이터: $\mathcal{D}^t = \{I_i^t, y_i^t\}_{i=1}^{N^t}$
  - 클래스들: $\mathcal{C}^t = \{c_j^t\}_{j=1}^{M^t}$
- 종류: TIL(Task Incremental Learning), CIL(Class Incremental Learning)
  - TIL: task-specific set $\mathcal{C}^t$ 내에서 예측.
  - CIL: 지금까지 본 클래스 집합 $\cup_{i=1}^{t}\mathcal{C}^i$ 내에서 예측

전유진

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

- **Incremental Mixture-of-Experts Adapters**



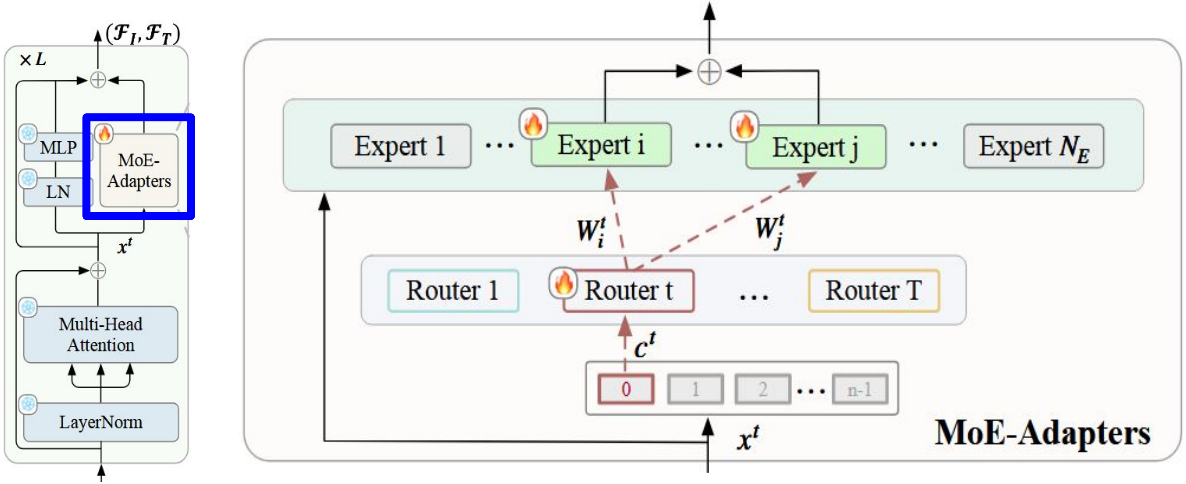
(a) The training stage

- "Catastrophic forgetting" 방지 위해, CLIP 내에 MoE 를 통한 확장 구조 사용.
- CLIP 의 Image, Text encoder 내 모든 transformer block들에서 MoE-Adapters 실행.
- MoE-Adapters: 1) Experts $\{\mathcal{E}_i\}_{i=1}^{N_E}$ 와 2) task-dependent Routers $\mathcal{R}^t, t \in [1, T]$ 로 구성.

[Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.]

CVPR 2024

- **Incremental Mixture-of-Experts Adapters**



1. 현재 task의 Router가 각 expert의 activation 정도 결정함.

: 각 transformer block 중간에서, [CLS] 토큰을 받아서, 모든 experts의 activation 확률 계산하여, top-K 개만 골라 gating weights 출력.

$$\boldsymbol{x}^t \in \mathbb{R}^{n\times d} \qquad \boldsymbol{c}^t \in \mathbb{R}^{1\times d} \qquad W^t = \{W_i^t\}_{i=1}^{N_E} \qquad W^t = Softmax(Topk(\mathcal{R}^t(\boldsymbol{c}^t))), \tag{2}$$

2. Top-K E

$$\boldsymbol{y}^t = \sum_{i=1}^{N_E} W_i^t \mathcal{E}_i(\boldsymbol{x}^t), \tag{1}$$

$$\boldsymbol{y}^t \in \mathbb{R}^{n\times d}$$

전유진

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

- **Incremental Mixture-of-Experts Adapters**

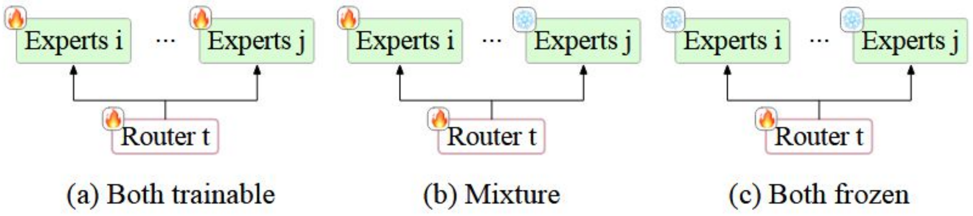


Figure 3. The three distinct combinations among activated experts (a) both trained, (b) trainable and frozen, (c) both experts are frozen, and only the router is trainable.

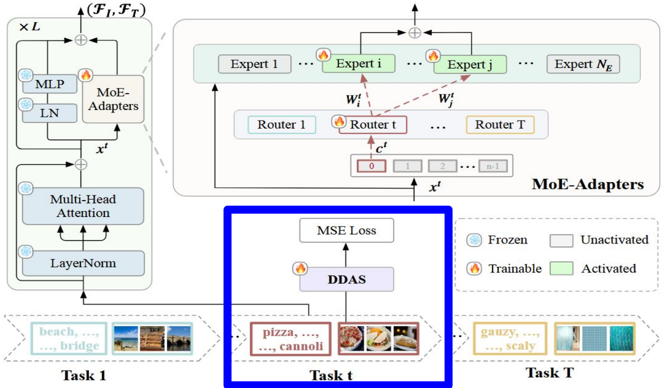
- 학습 방법: Incremental activate-freeze strategy

1. 이전 task에서 router의 아웃풋 분포를 통해, 학습된 top-K experts 파악.
2. 이전 task 에서 학습된 experts는 frozen하고, 학습되지 않았던 experts에 대하여 학습 진행.
3. 이렇게 함으로서, historical task에 대한 지식 이용 및 기억 + new task에 대한 새로운 지식 획득.

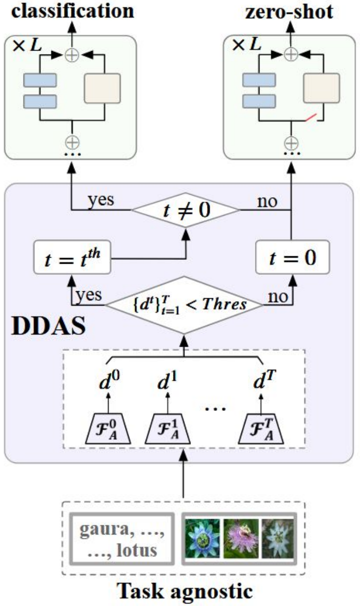
Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." nference on Computer Vision and Pattern Recognition. 2024.

CVPR 2024

- **Distribution Discriminative Auto-Selector**



(a) The training stage



(b) The inference stage

- 인풋 task에 맞는 적절한 router 결정해주는 역할.

- 학습 방법

  - 구조: 일련의 task-specific autoencoders + 추가적인 autoencoder (즉, 총 T+1 개의 autoencoders)

  - 각 autoencoder $\{\mathcal{F}_A^t\}_{t=1}^T$ 는 각 task $\{\mathcal{T}^t\}_{t=1}^T$ 의 분포를 MSE loss로 학습. $d^t = ||\boldsymbol{f}_i^t - \boldsymbol{f}_o^t||^2,$ (3)

  - 추가적인 autoencoder $\mathcal{F}_A^0$ 는 out-of-distribution data 파악 목적.

- 사용 방법

  - task-specific autoencoders 중 점수가 제일 낮은 task의 router 사용.

  - 만약, 모든 task-specific autoencoders 의 점수가 임계값보다 높다면, out-of-distribution data로 파악하고 zero-shot transfer함.

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

CVPR 2024

*MTIL: Multi-domain Task Incremental Learning
**CIL: Class Incremental Learning

- **Experiments**

  - **Datasets**

    1) *MTIL: 서로 다른 도메인의 데이터셋 11개 사용. … [1]

       i) Alphabet order (Order-I) (default)

       : Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, StanfordCars, SUN397.

       ii) Random order (Order-II)

       : StanfordCars, Food, MNIST, OxfordPet, Flowers, SUN397, Aircraft, Caltech101, DTD, EuroSAT, CIFAR100.

    2) **CIL: CIFAR100, TinyImageNet 데이터셋 사용. … [2]

       i) CIFAR100: 100개의 클래스가 {10, 20, 50} subsets 내 step으로 나뉨.

       ii) TinyImageNet: 100개의 클래스가 {5, 10, 20} subsets 내 step으로 나뉨.

[1] Zheng, Zangwei, et al. "Preventing zero-shot transfer degradation in continual learning of vision-language models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
[2] Douillard, Arthur, et al. "Dytox: Transformers for continual learning with dynamic token expansion." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- **Experiments**

  - **Metrics**

1) *MTIL: "Transfer", "Last", "Average" ··· [1]

i) Transfer: zero-shot transfer 능력 측정.

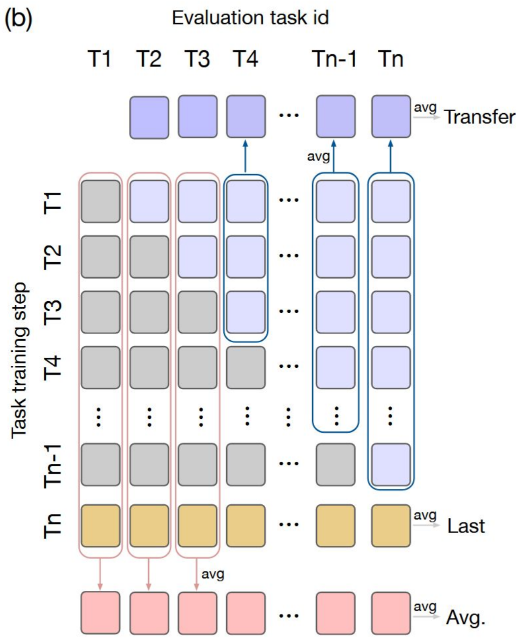
ii) Last: historical knowledge 기억력 측정.

iii) Average: 모든 timestamp에서의 정확도 평균값.

2) **CIL: "Last", "Average" ··· [2]

i) Last: 마지막 step이 끝난 후 정확도.

ii) Average: 각 step이 끝날때마다의 정확도 평균값.



[1] Zheng, Zangwei, et al. "Preventing zero-shot transfer degradation in continual learning of vision-language models." *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.
[2] Douillard, Arthur, et al. "Dytox: Transformers for continual learning with dynamic token expansion." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

- **Experiments**

  - **Implementation details**

    1. Model: CLIP model with ViT-B/16
    2. MoE-Adapters
       a. Experts: 22개
       b. Router: single MLP
       c. top-K 에서 K=2
    3. DDAS
       a. Reference dataset: TinyImageNet
       b. Threshold: 0.065 (full-shot), 0.06 (few-shot)
       c. Autoencoder: pretrained AlexNet with MLP and a non-linear layer
    4. Iterations
       a. TIL: 1k (full-shot), 0.5k (few-shot) for each task
       b. DDAS: 1k (reference datasets), 0.3k (incremental tasks)

[Yu, Jiazuo, et al. "Boosting continual l Pattern Recognition. 2024.]

- **Experiments**

  - MTIL 성능
  - Order-I 순서로 학습 및 테스트

	Method	Aircraft [49]	Caltech101 [21]	CIFAR100 [38]	DTD [9]	EuroSAT [25]	Flowers [54]	Food [4]	MNIST [13]	OxfordPet [58]	Cars [37]	SUN397 [69]	*Average*
CLIP	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
	Full Fine-tune	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
	Fine-tune Adapter	56.8	92.6	89.4	79.0	98.4	97.0	92.9	99.2	94.1	89.1	82.7	88.3
Transfer	Continual-FT		67.1	46.0	32.1	35.6	35.0	57.7	44.1	60.8	20.5	46.6	44.6
	LwF [42]		74.5	56.9	39.1	**51.1**	52.6	72.8	60.6	75.1	30.3	55.9	58.9
	iCaRL [61]		56.6	44.6	32.7	39.3	46.6	68.0	46.0	77.4	31.9	60.5	50.4
	LwF-VR [15]		77.1	61.0	40.5	45.3	54.4	74.6	47.9	76.7	36.3	58.6	57.2
	WiSE-FT [67]		73.5	55.6	35.6	41.5	47.0	68.3	53.9	69.3	26.8	51.9	52.3
	ZSCL [78]		86.0	67.4	**45.4**	50.4	69.1	87.6	**61.8**	86.8	60.1	**66.8**	68.1
	Ours†		**87.9**	**68.2**	42.2	41.4	68.7	**88.7**	59.4	**89.1**	**64.5**	64.0	67.4(-0.7)
	Ours		**87.9**	**68.2**	44.4	49.9	**70.7**	**88.7**	59.7	**89.1**	**64.5**	65.5	**68.9**(+0.8)
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
	LwF [42]	36.3	86.9	72.0	59.0	73.7	60.0	73.6	74.8	80.0	37.3	58.1	64.7
	iCaRL [61]	35.5	89.2	72.2	60.6	68.8	70.0	78.2	62.3	81.8	41.2	62.5	65.7
	LwF-VR [15]	29.6	87.7	74.4	59.5	72.4	63.6	77.0	66.7	81.2	43.7	60.7	65.1
	WiSE-FT [67]	26.7	86.5	64.3	57.1	65.7	58.7	71.1	70.5	75.8	36.9	54.6	60.7
	ZSCL [78]	45.1	**92.0**	80.1	64.3	**79.5**	81.6	**89.6**	**75.2**	88.9	64.7	**68.0**	75.4
	Ours†	**54.3**	91.1	**85.1**	**69.7**	77.5	**84.5**	89.1	73.8	89.2	**69.0**	65.8	**77.2**(+1.8)
	Ours	50.2	91.9	83.1	69.4	78.9	84.0	89.1	73.7	**89.3**	67.7	66.9	76.7(+1.3)
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
	LwF [42]	26.3	87.5	71.9	66.6	79.9	66.9	83.8	**99.6**	92.1	66.1	80.4	74.6
	iCaRL [61]	35.8	**93.0**	77.0	70.2	83.3	88.5	90.4	86.7	93.2	81.2	81.9	80.1
	LwF-VR [15]	20.5	89.8	72.3	67.6	85.5	73.8	85.7	**99.6**	93.1	73.3	80.9	76.6
	WiSE-FT [67]	27.2	90.8	68.0	68.9	86.9	74.0	87.6	**99.6**	92.6	77.8	81.3	77.7
	ZSCL [78]	40.6	92.2	81.3	70.5	94.8	90.5	**91.9**	98.7	**93.9**	85.3	80.2	83.6
	Ours†	**54.3**	90.8	**88.8**	**80.3**	**98.1**	**97.5**	89.6	99.1	89.5	**89.2**	**83.8**	**87.4**(+3.8)
	Ours	49.8	92.2	86.1	78.1	95.7	94.3	89.5	98.1	89.9	81.6	80.0	85.0(+1.4)

Table 1. Comparison with state-of-the-art methods on MTIL benchmark in terms of "Transfer", "Average", and "Last" scores (%). "Ours†" and "Ours" indicate our method trained on 3k and 1k iterations, respectively. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

CVPR 2024

- ## Experiments

- Few-shot MTIL 성능

- Order-I 순서로 학습 및 테스트

	Method	Aircraft [49]	Caltech101 [21]	CIFAR100 [38]	DTD [9]	EuroSAT [25]	Flowers [54]	Food [4]	MNIST [13]	OxfordPet [58]	Cars [37]	SUN397 [69]	*Average*
CLIP	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
	5-shot Full Fine-tune	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
	5-shot Fine-tune Adapter	29.7	90.0	75.3	63.9	81.1	94.2	87.8	90.4	89.0	68.2	72.5	76.6
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
	LwF [42]		72.1	49.2	35.9	44.5	41.1	66.6	50.5	69.0	19.0	51.7	50.0
	LwF-VR [15]		82.2	62.5	40.1	40.1	56.3	80.0	60.9	77.6	40.5	60.8	60.1
	WiSE-FT [67]		77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL [78]		84.0	68.1	**44.8**	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	Ours		**87.9**	**68.2**	44.1	**48.1**	**64.7**	**88.8**	**69.0**	**89.1**	**64.5**	**65.1**	**68.9**(+3.6)
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
	LwF [42]	23.5	77.4	43.5	41.7	43.5	52.2	54.6	63.4	68.0	21.3	52.6	49.2
	LwF-VR [15]	24.9	89.1	64.2	53.4	54.3	70.8	79.2	66.5	79.2	44.1	61.6	62.5
	WiSE-FT [67]	**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 [78]	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	Ours	30.0	**89.6**	**73.9**	**58.7**	**69.3**	**79.3**	**88.1**	**76.5**	**89.1**	**65.3**	**65.8**	**71.4**(+7.0)
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
	LwF [42]	22.1	58.2	17.9	32.1	28.1	66.7	46.0	84.3	64.1	31.5	60.1	46.5
	LwF-VR [15]	22.9	**89.8**	59.3	57.1	57.6	79.2	78.3	77.7	83.6	60.1	69.8	66.9
	WiSE-FT [67]	**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 [78]	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	Ours	30.1	89.3	**74.9**	**64.0**	**82.3**	**89.4**	**87.1**	89.0	**89.1**	**69.5**	**72.5**	**76.1**(+4.2)

Table 2. Comparison with state-of-the-art methods on few-shot MTIL benchmark in terms of "Transfer", "Average", and "Last" scores (%). Ours converges in 500 iterations on few-shot. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.

Yu, Jiazuo, et al. "Boosting continual learning of vision-language models via mixture-of-experts adapters." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

CVPR 2024

• Experiments

- CIL 성능
- Router 수: 1개, Expert 수: 2개. (∵ Single domain)

Method	10 step		20 step		50 step	
	Avg.	Last	Avg.	Last	Avg.	Last
UCIR [26]	58.66	43.39	58.17	40.63	56.86	37.09
Bic[68]	68.80	53.54	66.48	47.02	62.09	41.04
PODNet[18]	58.03	41.05	53.97	35.02	51.19	32.99
DER [70]	74.64	64.35	73.98	62.55	72.05	59.76
DyTox+[19]	74.10	62.34	71.62	57.43	68.90	51.09
DNE [29]	74.86	70.04	-	-	-	-
CLIP Zero-shot	74.47	65.92	75.20	65.74	75.67	65.94
Fine-tune	65.46	53.23	59.69	43.13	39.23	18.89
LwF [42]	65.86	48.04	60.64	40.56	47.69	32.90
iCaRL [61]	79.35	70.97	73.32	64.55	71.28	59.07
LwF-VR [15]	78.81	70.75	74.54	63.54	71.02	59.45
ZSCL [78]	82.15	73.65	80.39	69.58	79.92	67.36
Ours	**85.21**	**77.52**	**83.72**	**76.20**	**83.60**	**75.24**

Table 3. Comparison of different methods on CIFAR100 in class-incremental setting. We label the best and second-best methods with **bold** and underline styles.

Method	5 step		10 step		20 step	
	Avg.	Last	Avg.	Last	Avg.	Last
EWC [36]	19.01	6.00	15.82	3.79	12.35	4.73
EEIL [6]	47.17	35.12	45.03	34.64	40.41	29.72
UCIR [26]	50.30	39.42	48.58	37.29	42.84	30.85
MUC [46]	32.23	19.20	26.67	15.33	21.89	10.32
PASS [80]	49.54	41.64	47.19	39.27	42.01	32.93
DyTox [19]	55.58	47.23	52.26	42.79	46.18	36.21
CLIP Zero-shot	69.62	65.30	69.55	65.59	69.49	65.30
Fine-tune	61.54	46.66	57.05	41.54	54.62	44.55
LwF [42]	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL [61]	77.02	70.39	73.48	65.97	69.65	64.68
LwF-VR [15]	77.56	70.89	74.12	67.05	69.94	63.89
ZSCL [78]	80.27	73.57	78.61	71.62	77.18	68.30
Ours	**81.12**	**76.81**	**80.23**	**76.35**	**79.96**	**75.77**

Table 4. Comparison of different methods on TinyImageNet dataset in class-incremental settings with 100 base classes. We label the best and second methods with **bold** and underline styles.

- **Experiments**

Method	Train Params ↓	GPU ↓	Times ↓
LWF [42]	149.6M	32172MiB	1.54s/it
LWF-VR [15]	149.6M	32236MiB	1.51s/it
ZSCL [78]	149.6M	26290MiB	3.94s/it
MoE-Adapters	51.1M	19898MiB	1.37s/it
DDAS	8.7M	2461MiB	0.21s/it
Ours	59.8M	22358MiB	1.58s/it
Δ	**-60.03%**	**-14.95%**	**-59.90%**

Table 5. Comparison of computational cost during training between our method and others in terms of training parameters, GPU burdens and training times of each iteration. And the Δ is the improvement relative to the SOTA ZSCL [78].