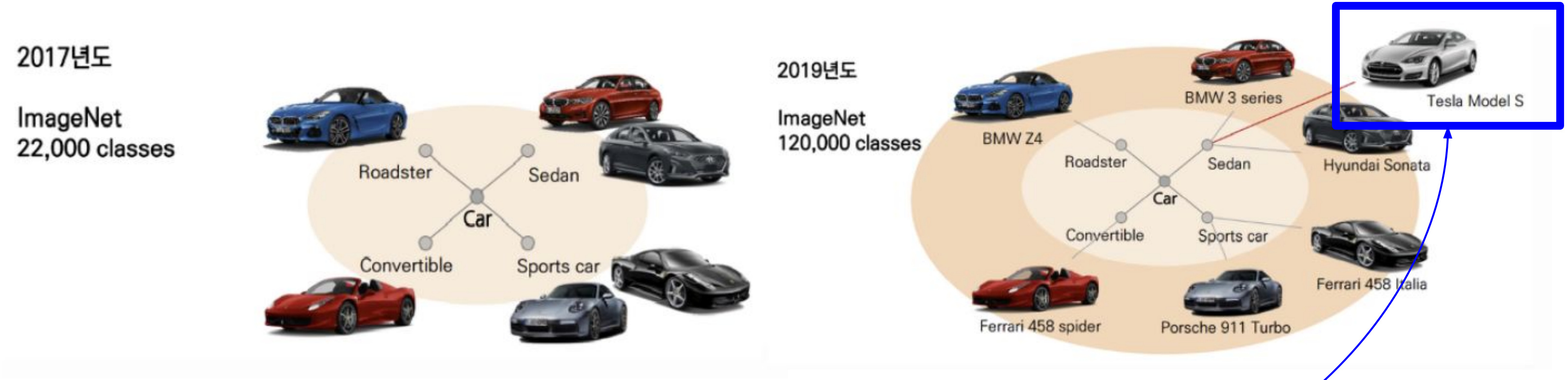
Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters

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
  - Alleviate long-term forgetting in incremental learning with vision-language models
- Contribution / Key idea
  - Parameter-efficient continual learning framework
    - Mixture-of-Experts (MoE) adapters
    - Distribution Discriminative Auto-Selector (DDAS)

- **Continual Learning**

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

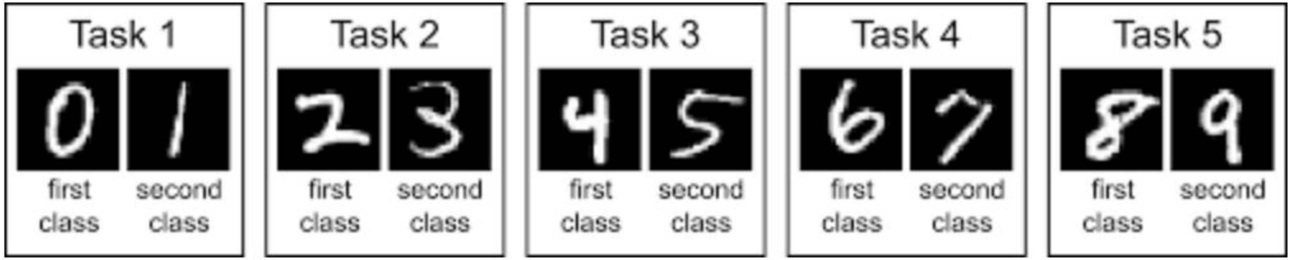


새로운 데이터가 나올때마다 처음부터 다시 학습하는 것이 아니라, 이미 학습된 모델에 새로운 데이터만 추가적으로 더 학습

- **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.

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 선행 연구들**

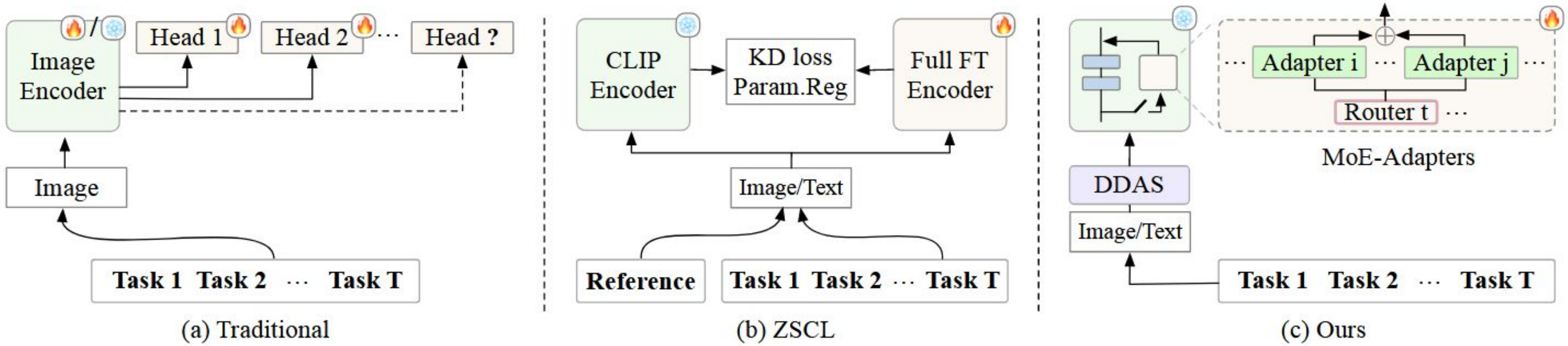


Figure 1. Comparison of various popular architectures to address CL. (a) Traditional dynamic expansion-based CL cannot distinguish unseen data. (b) Zero-shot CL [78] suffers from significant computational burdens. (c) The proposed MoE-Adapters and DDAS collaborate to form a parameter-efficient, zero-shot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

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

(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 및 확장성', (b) 의 장점인 'zero-shot 능력' 모두 살리겠다

전유진

- **Overall framework**

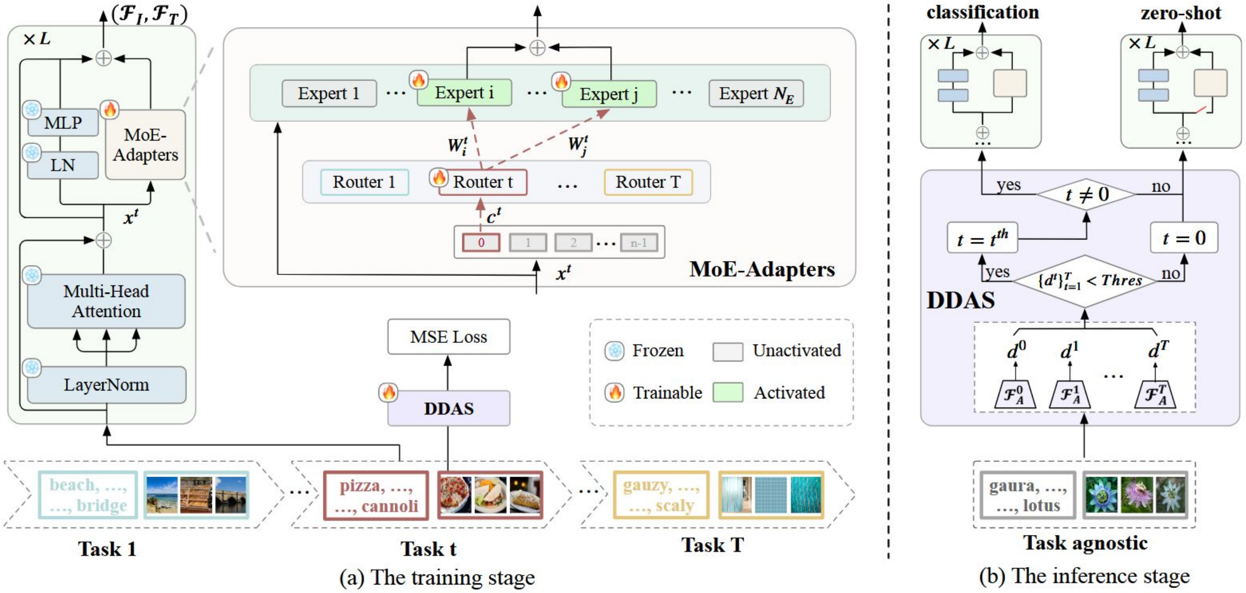


Figure 2. Overall framework of the proposed method. (a) At the training stage, CLIP's image and text encoders $(\mathcal{F}_I, \mathcal{F}_T)$ take input samples from **Task *t***. In each of transformer blocks, there is a MoE-Adapters, whose input is the tokens x^t from MHSA. The router takes the task-specific [CLS] token 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.

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.

CVPR 2024

•

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

-

전유진