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

Zangwei Zheng[1] Mingyuan Ma[2] Kai Wang[1] Ziheng Qin[1] Xiangyu Yue[3] Yang You[1]
[1]National University of Singapore [2]UC Berkeley [3]The Chinese University of Hong Kong
[1]{zangwei, kai.wang, zihengq, youy}@comp.nus.edu.sg [2]mamingyuan2001@berkeley.edu [3]xyyue@ie.cuhk.edu.hk

- Problem / objective
  - Zero-shot transfer degradation in the continual learning of vision-language models
- Contribution / Key idea
  - Method (ZSCL: Zero-Shot Continual Learning)
    - Distillation in the feature space
    - Weight ensemble in the parameter space
  - Benchmark (MTIL: Multi-domain Task Incremental Learning)

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.

- **Continual Learning**

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

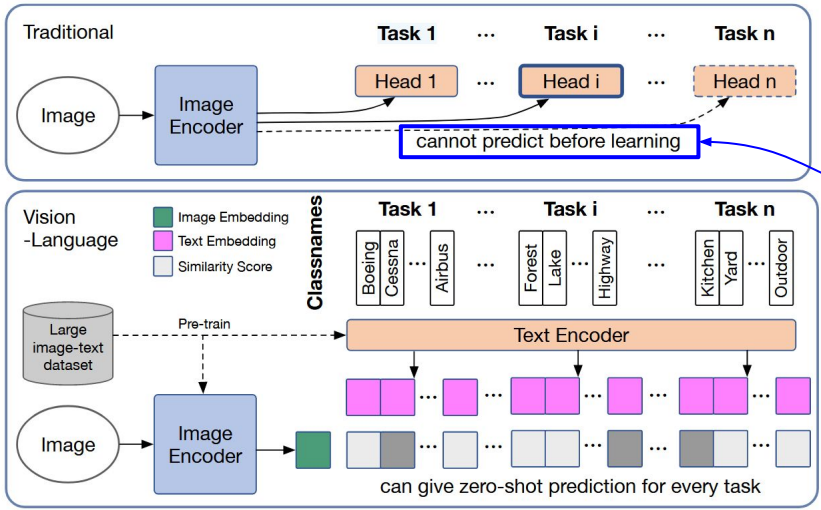
새로운 데이터가 나올때마다 처음부터 다시 학습하는 것이 아니라, 이미 학습된 모델에 새로운 데이터만 추가적으로 더 학습

이전에 학습된 지식을 잊어버리지 않고 보존하는 능력도 중요
("Catastrophic Forgetting")

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.
ICCV 2023

- **Continual Learning with VLM**
  - ❏ 장점: 전통적인 CL에서는 못했던 Zero-shot transfer ability 굿.
  - ❏ 단점: 고민거리 하나 더 생김.

    1) 원래의 고민: previously learned task knowledge 계속 잊어버려. ("Catastrophic Forgetting")
    2) 추가된 고민: pretrained VLM의 knowledge도 잊어버려. ("Degradation of zero-shot transfer ability")



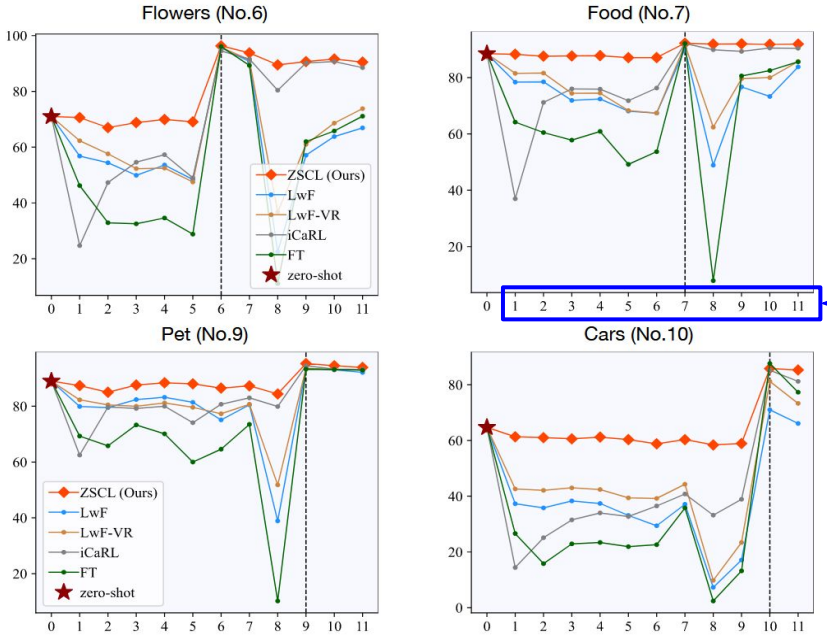
(a) Comparison between traditional CL and CL with a pre-trained vision-language model

기존 CL의 한계

Figure 1. a) Conventional CL learns distinct task-specific heads, while CL with vision-language models can predict both learned tasks and out-of-distribution tasks. b) Accuracy (%) changes during CL of four datasets on 11 datasets. Our method is superior to others in preventing the forgetting of both zero-shot transfer ability and new knowledge.

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.
ICCV 2023

- **Objective: Let's protect the Zero-Shot transfer ability during Continual Learning.**
  - ❏ 본 논문은, 1) zero-shot transfer ability 와 2) learned knowledge 둘다 제일 잘 보존해요.
  - ❏ 아래 그림에서 1) 은 점선 앞 그래프, 2) 는 점선 뒤 그래프 추이 보면 됨.



(b) Performance of different methods on preventing forgetting phenonmenon

Task1~11: Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, StanfordCars, SUN397

Figure 1. a) Conventional CL learns distinct task-specific heads, while CL with vision-language models can predict both learned tasks and out-of-distribution tasks. b) Accuracy (%) changes during CL of four datasets on 11 datasets. Our method is superior to others in preventing the forgetting of both zero-shot transfer ability and new knowledge.

- **Preliminaries**
  - ➢ Continual Learning
    - 목표: 모든 task에서 성능 잘 나오기
    - n개의 task들: $[\mathcal{T}^1, \mathcal{T}^2, \cdots, \mathcal{T}^n]$

      각 task: $\mathcal{T}^i = (\mathcal{D}^i, C^i), i = 1, \ldots, n$

      데이터셋: $\mathcal{D}^i = \{(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i\}_{j=1}^{N_i}$

      클래스들: $C^i = \{c_j^i\}_{j=1}^{m_i}$
    - 종류: TIl (Task-Incremental Learning), CIL (Class-Incremental Learning)
      - TIL: C^t 내에서 클래스 예측
      - CIL: $C = \bigcup_{i=1}^{n} C^i$ 내에서 클래스 예측
  - ➢ CLIP model
    - Image / Text embedding 간 유사도 점수 for task $\mathcal{T}^i$: $\boldsymbol{s}_{k,j}^i = \langle f_i(\boldsymbol{x}_k), \boldsymbol{t}_j^i \rangle$
    - Fine-tuning for downstream tasks:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\tau \cdot \boldsymbol{s}_i, \boldsymbol{y}_i), \qquad (1)$$

- **Distillation in Feature Space**
  - ❏ 문제: Downstream task 에 바로 fine-tuning 하면 VLM 의 feature space 의 분포가 왜곡되어, VLM 의 zero-shot 성능 많이 하락함.
  - ❏ 해결: 1) CE loss for fine-tuned feature space && 2) distillation loss for out-of-distribution feature space

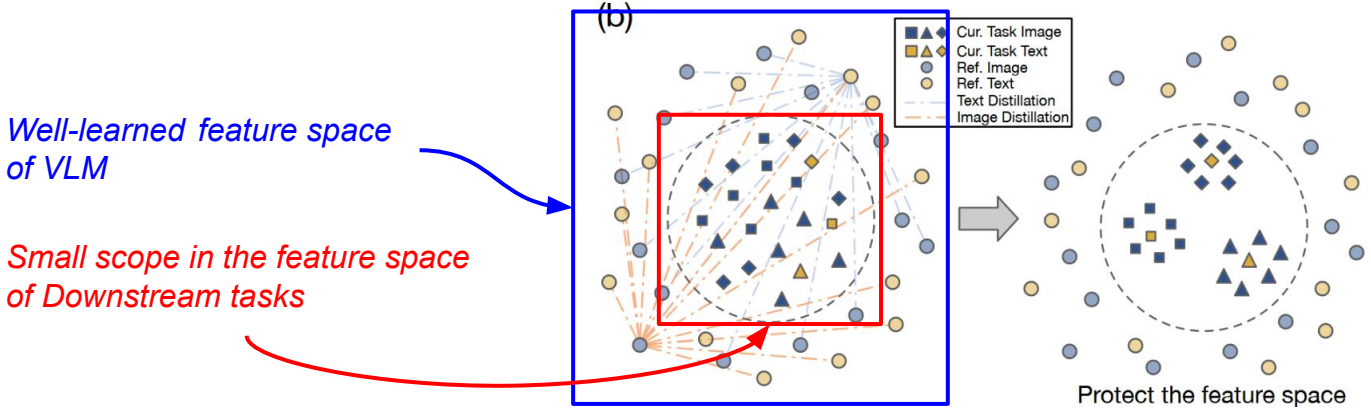
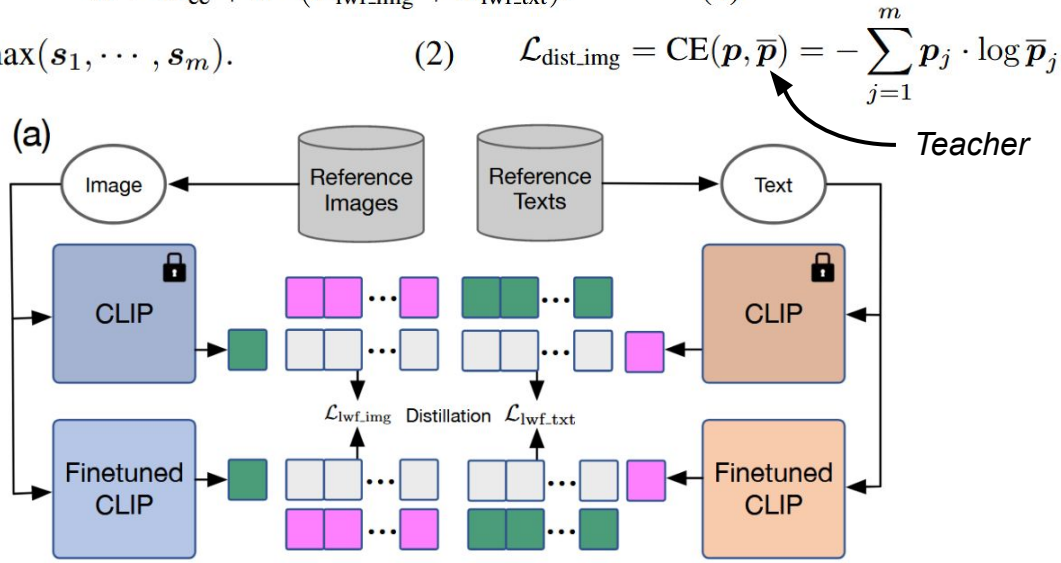


Figure 2. Illustration of ZSCL in feature space. Fig. 3(a) shows the pipeline of distillation. The original and the current model encode the reference images and texts, respectively. The probability distribution of images and texts with respect to each other is distilled. Fig. 3(b) displays how distillation loss preserves the feature space. Compared with the reference dataset, the features of fine-tuning tasks lie in a small subspace. The distillation loss preserves the structure of the feature space by maintaining relative distances.

- **Distillation in Feature Space**
  - 실험적으로 아래 세팅 설정.
    a. **Data source.** Reference dataset: publicly available image dataset with enough semantics.
    b. **Teacher model.** Pre-trained model.
    c. **Loss design.**

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \cdot (\mathcal{L}_{\text{lwf\_img}} + \mathcal{L}_{\text{lwf\_txt}}). \quad (4)$$

$$\boldsymbol{p} = \text{Softmax}(\boldsymbol{s}_1, \cdots, \boldsymbol{s}_m). \quad (2) \qquad \mathcal{L}_{\text{dist\_img}} = \text{CE}(\boldsymbol{p}, \overline{\boldsymbol{p}}) = -\sum_{j=1}^{m} \boldsymbol{p}_j \cdot \log \overline{\boldsymbol{p}}_j, \quad (3)$$



- **Weight Ensemble (WE) in Parameter Space**
  - ❑ 문제: 그냥 Fine-tuning 하면 zero-shot transfer ability 급격한 하락.
  - ❑ 목표: zero-shot transfer ability 와 downstream task performance 간 적절한 trade-off 지점 찾기.
  - ❑ 해결: 학습 도중 모델들의 가중치 평균내겠다.

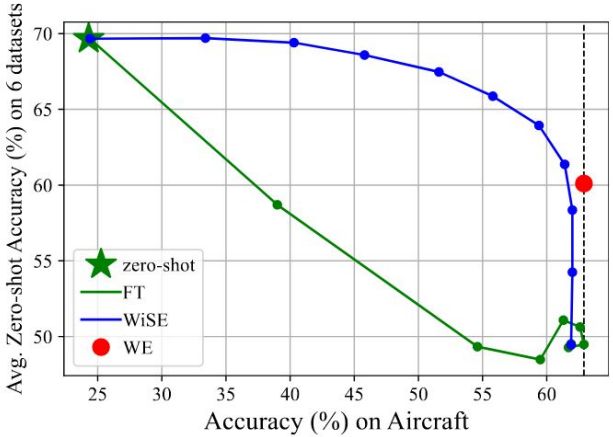


Figure 3. Models during training contain different tradeoffs between zero-shot and new task performance. Points for FT are sampled every 100 iterations, and the ones for WiSE-FT means different α choices. WE ensembles models during training and achieve better performance.

- ➢ FT: Fine-Tuning
- ➢ Wise-FT [1]: Zero-shot 모델과 fine-tuned 모델의 가중치 ensemble

$$f(x; (1-\alpha)\cdot\theta_0 + \alpha\cdot\theta_1), \qquad (6)$$

- ➢ WE (Ours): 학습 도중 모델들의 가중치 ensemble

$$\hat{\theta}_t = \begin{cases} \theta_0 & t = 0 \\ \frac{1}{t+1}\theta_t + \frac{t}{t+1}\cdot\hat{\theta}_{t-1} & \text{every I iterations} \end{cases}. \qquad (7)$$

[1] Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

- **Multi-domain Task Incremental Learning**

❑ 전통적인 Continual Learning Benchmark.

a. Single dataset 내에 클래스 분류하여 task 정의.

(예를들어, m-classes 갖는 데이터셋에서 k-step 세팅 사용한다면, 각 task 당 m/k개의 클래스 학습한다는 의미.)

❑ MTIL Benchmark 새롭게 제안.

a. 서로 다른 datasets 로부터 서로 다른 tasks 정의.

b. 구체적으로, 11 datasets, 1201 classes.

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.

Table 8. Dataset description of multi-domain task incremental 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	

- **Multi-domain Task Incremental Learning**

❑ MTIL Benchmark 새롭게 제안.

c. 평가 메트릭

i. Transfer

: zero-shot transfer 능력 측정.

ii. Last

: historical knowledge 기억력 측정.

iii. Average

: 모든 timestamp에서의 정확도 평균값.

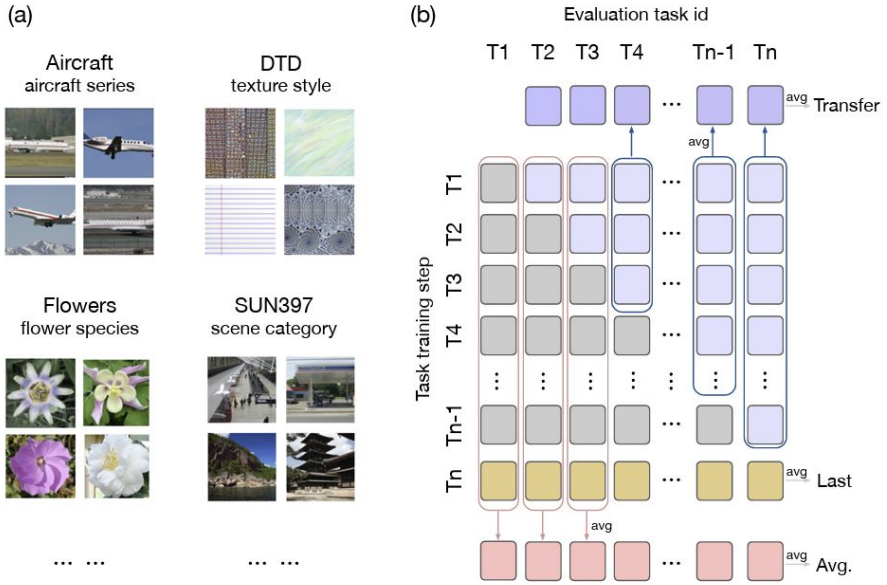


Figure 4. Fig.(**a**): examples of tasks from different domains in MTIL benchmark. Fig.(**b**): illustration of calculating metrics Transfer, Avg. and Last during continual learning.

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.
ICCV 2023

- **Experiments**

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**

- **Experiments**

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

	10 steps		20 steps		50 steps	
Methods	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**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.

	5 steps		10 steps		20 steps	
Methods	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**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**