Apprentissage continu de représentations visuelles

ENSIMAG
2023-2024

Karteek Alahari & Diane Larlus

Apprentissage continu

https://project.inria.fr/bigvisdata/

Summary: Continual Learning

- **Flavour of different approaches:**
  1. Regularization based: LwF, EBLL, EWC, SI, MAS, IMM, …
  2. Rehearsal / Replay: iCaRL, DGR, GEM, …
  3. Architecture based: PackNet, progressive nets , HAT, …

- Other learning frameworks, e.g., self-supervised

- Takeaways

A Comparative Analysis

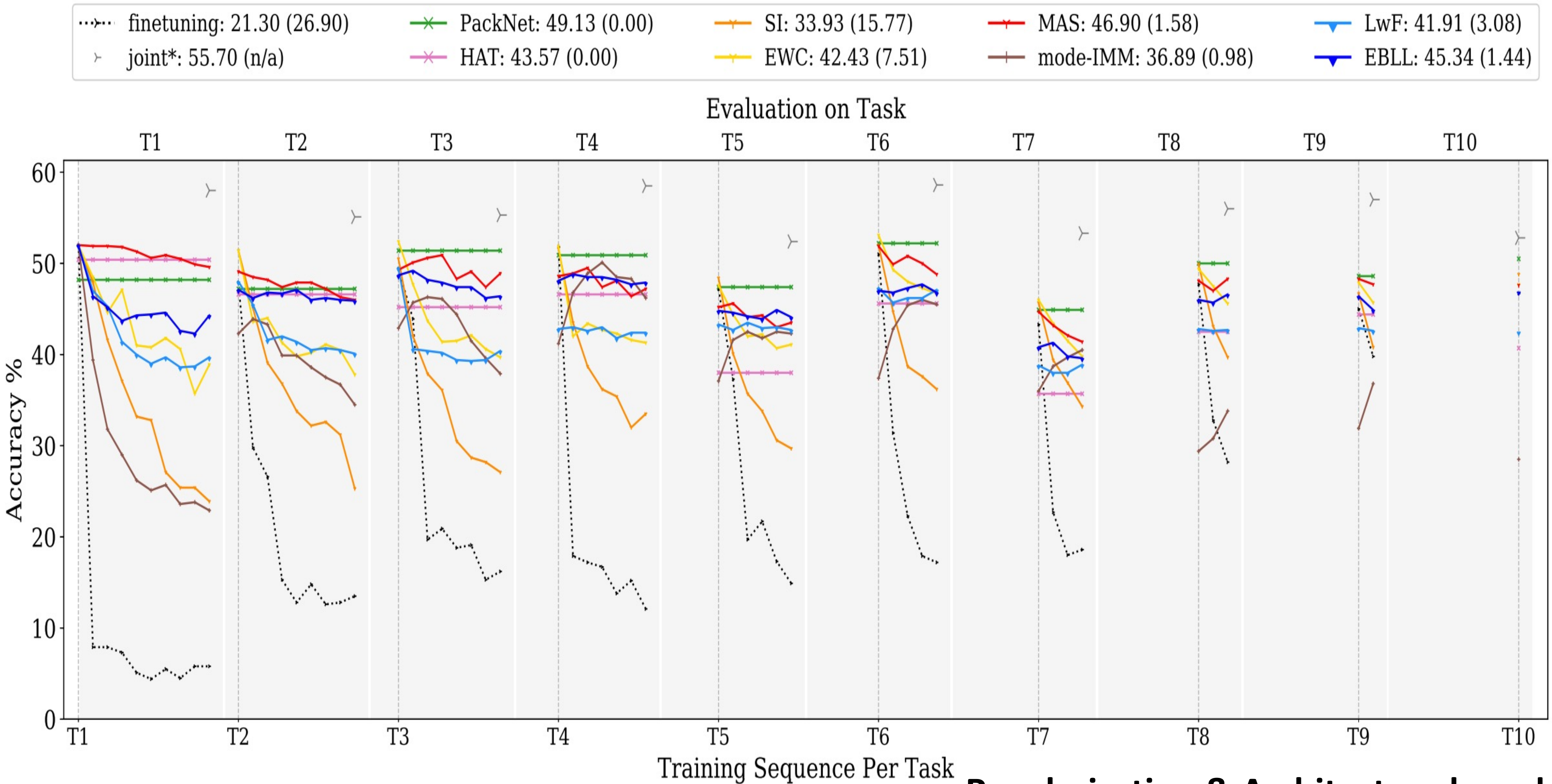
- TinyImagenet: small, balanced, class-incremental
- iNaturalist: large-scale, unbalanced, task-incremental

	Tiny Imagenet	iNaturalist
Tasks	10	10
Classes per task	20	5 to 314
Training data per task	8k	0.6k to 66k
Validation data per task	1k	0.1k to 9k
Task Constitution	random class selection	supercategory

- Fair way of setting hyperparameters (stability-plasticity tradeoff)

[Lange et al., 2020]

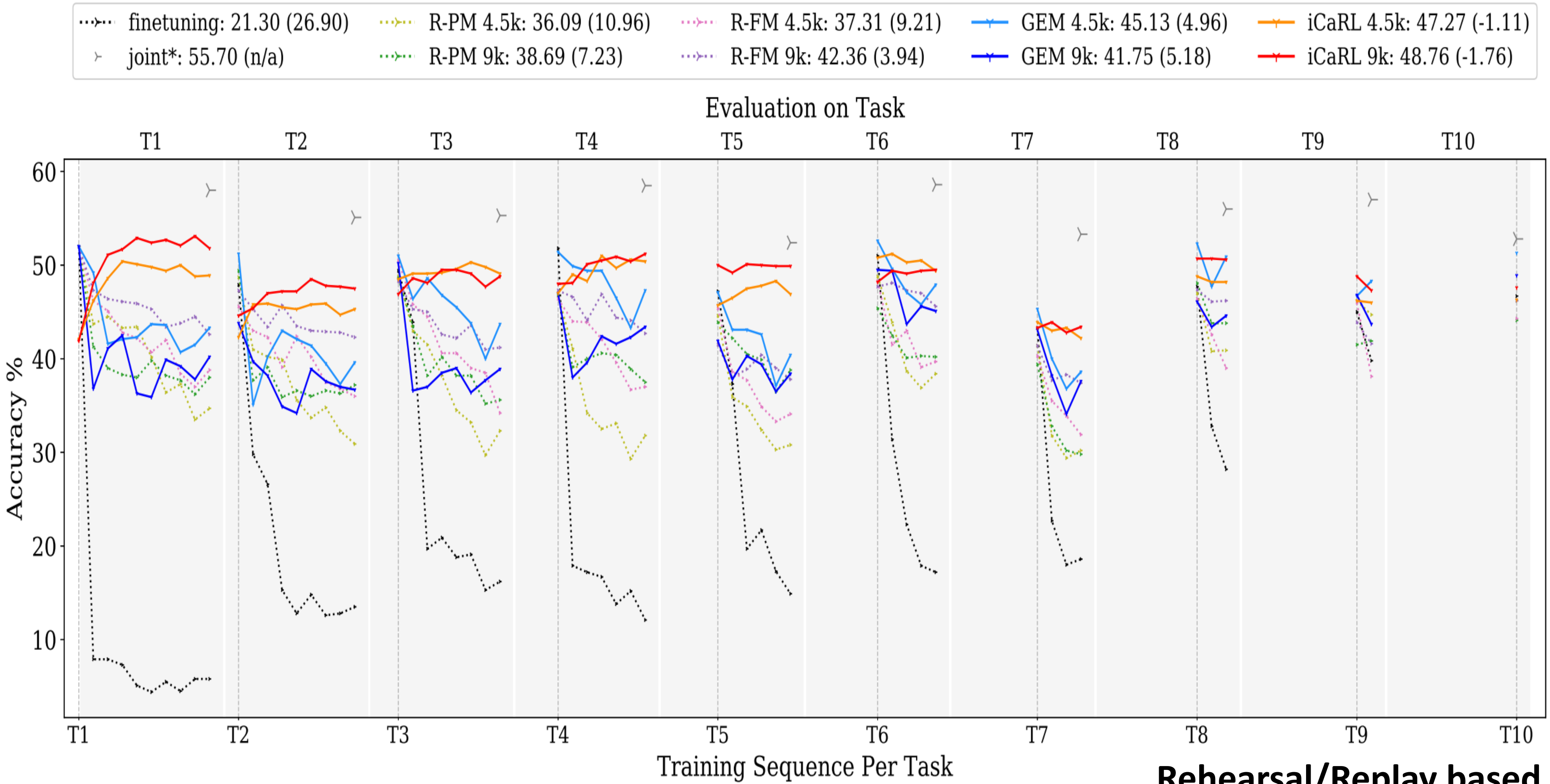
Comparative Evaluation (TinyImagenet)



Regularization & Architecture based

Image credit: [Lange et al., 2020]

Comparative Evaluation (TinyImagenet)



Rehearsal/Replay based

Image credit: [Lange et al., 2020]

What else will we see in the class?

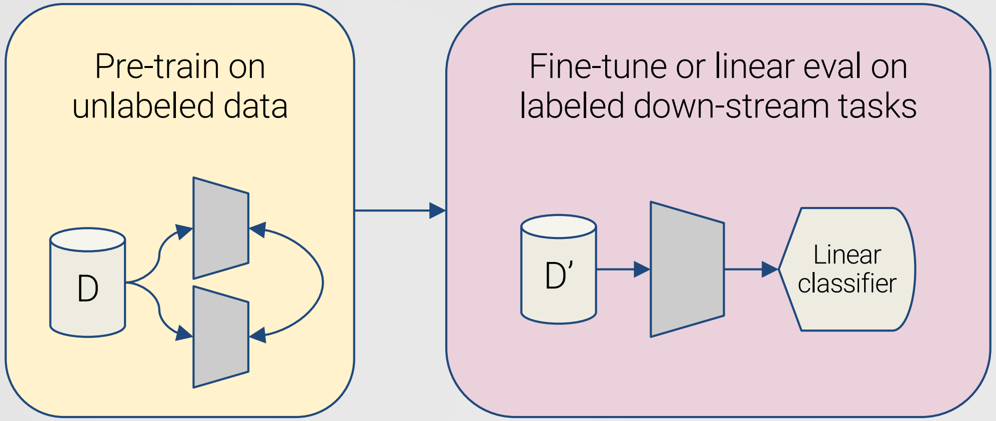
- Flavour of different approaches:
  1. Regularization based: LwF, EBLL, EWC, SI, MAS, IMM, …
  2. Rehearsal / Replay: iCaRL, DGR, GEM, …
  3. Architecture based: PackNet, progressive nets , HAT, …

- **Other learning frameworks, e.g., self-supervised**

- Takeaways

Self-Supervised Learning (SSL)

➔ Self-Supervised Learning exploits the intrinsic structure of the data to pretrain strong feature extractors



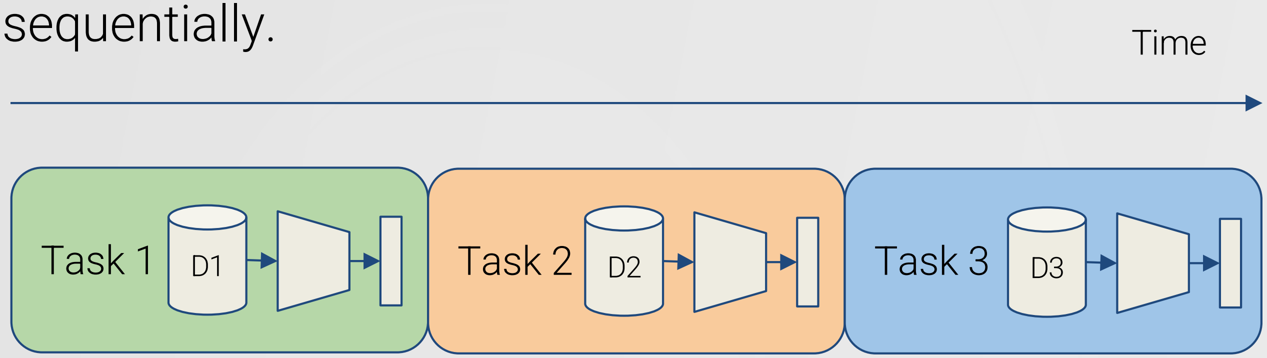
➔ The pre-training objective is to find parameters such as:

$$\underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathcal{L}_{SSL}\left(\boldsymbol{z}^{A},\boldsymbol{z}^{B}\right)\right]$$

Slide courtesy: E. Fini

Continual Learning (CL)

➔ Continual learning tackles the problem of learning tasks sequentially.



➔ More formally, the objective is to find parameters such as:

$$\underset{\theta'}{\operatorname{argmin}} \sum_{t=1}^{T} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_t} \left[\mathcal{L}_{CL}\left(\boldsymbol{p},\boldsymbol{y}\right)\right]$$

Slide courtesy: E. Fini

A new perspective

- Most of the literature studies CL with the following assumptions:
  - Availability of **Labels** (supervised learning)
  - Focus on learning a **classifier** that solves all tasks

- What if we looked at CL from a **different perspective**?

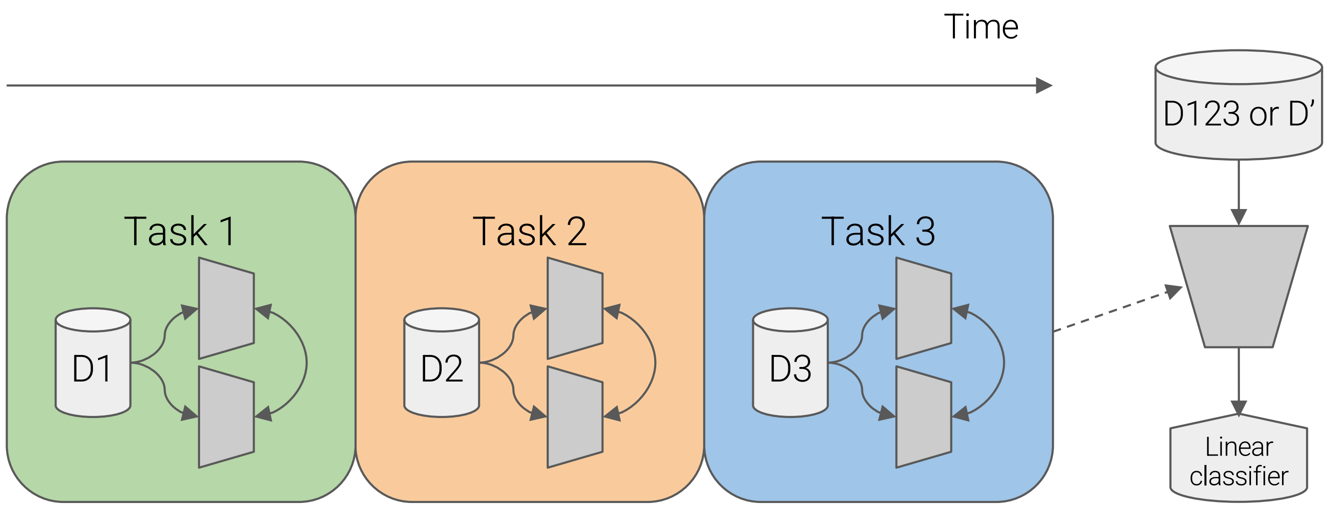
Assumption	Motivation
No labels	Availability of labels when learning online or sequentially is unlikely
Focus on learning **representations**	Training a linear classifier or fine-tuning a pre-trained feature extractor on a down-stream task is very simple and inexpensive with modern hardware

[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Continual Self-Supervised Learning (CSSL)

➔ Continual Self-Supervised Learning is the problem of learning strong feature extractors from streams of unlabeled data



➔ The continual pre-training objective is to find parameters such as:

$$\underset{\theta}{\operatorname{argmin}} \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_t} \left[\mathcal{L}_{SSL} \left(\boldsymbol{z}^A, \boldsymbol{z}^B \right) \right]$$

[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Continual Self-Supervised Learning (CSSL)

- → **Self-Supervised** fine-tuning sometimes **outperforms supervised** fine-tuning
- → **Plasticity** of representations is fundamental in CSSL
- → SSL benefits from **longer** training, the evolution of representations should not be overly **constrained**
- → SSL methods exhibit different losses and feature normalizations that **interfere** with CL regularization losses and vice-versa

Methods	Loss	Equation	
SimCLR [13] MoCo [28] NNCLR [19]	InfoNCE	$-\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_i^B)/\tau)}{\sum_{\boldsymbol{z}_j \in \eta(i)} \exp(\mathrm{sim}(\boldsymbol{z}_i^A, \boldsymbol{z}_j)/\tau)}$	(6)
BYOL [26] SimSiam [15] VICReg [3]	MSE	$-\|\|\boldsymbol{q}^A - \boldsymbol{z}^B\|\|_2^2$	(7)
SwAV [7] DCV2 [7] DINO [8]	Cross-entropy	$-\sum_d \boldsymbol{a}_d^B \log \frac{\exp(\mathrm{sim}(\boldsymbol{z}^A, \boldsymbol{c}_d)/\tau)}{\sum_k \exp(\mathrm{sim}(\boldsymbol{z}^A, \boldsymbol{c}_k)/\tau)}$	(8)
Barlow Twins [58] VICReg [3]	Cross-correlation	$\sum_u (1 - \mathcal{C}_{uv})^2 + \lambda \sum_u \sum_{v \neq u} \mathcal{C}_{uv}^2$	(9)

[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

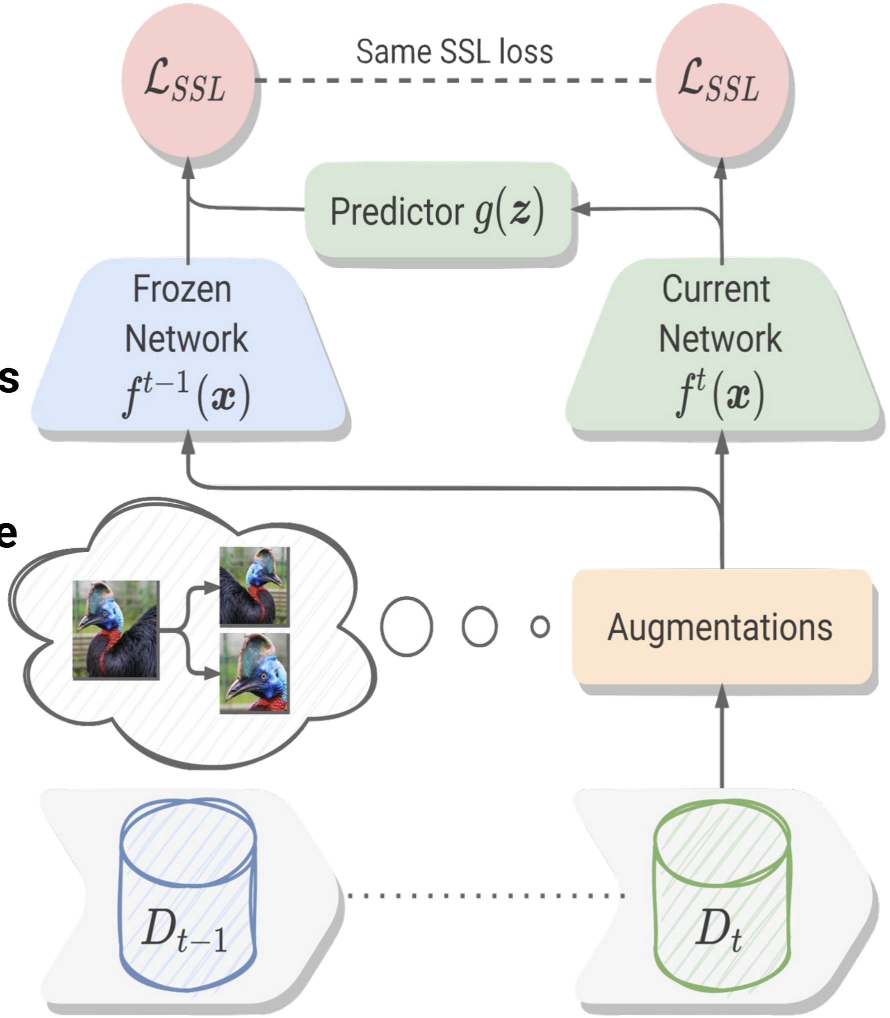
Proposed Method (CaSSLe)

Basically, two simple ideas:

- ➔ a **predictor** network that maps the current state of the representations to their past state
- ➔ a **family** of adaptable **distillation losses** inherited from the SSL literature

… and an important key insight: use the **same loss** for distillation and representation learning!

$$\mathcal{L} = \mathcal{L}_{SSL}(\boldsymbol{z}^A, \boldsymbol{z}^B) + \mathcal{L}_D(\boldsymbol{z}^A, \bar{\boldsymbol{z}}^A)$$
$$= \mathcal{L}_{SSL}(\boldsymbol{z}^A, \boldsymbol{z}^B) + \mathcal{L}_{SSL}(g(\boldsymbol{z}^A), \bar{\boldsymbol{z}}^A)$$



[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Experiments

➔ We train six SSL models:

- **Barlow Twins**
- **SwAV**
- **BYOL**
- **VICReg**
- **MoCoV2+**
- **SimCLR**

➔ We evaluate three CL settings:

- **Class-incremental**: each task contains a new set of classes
- **Data-incremental**: each task contains new samples of the same classes
- **Domain-incremental**: each task contains new domain

➔ On three widely used datasets:

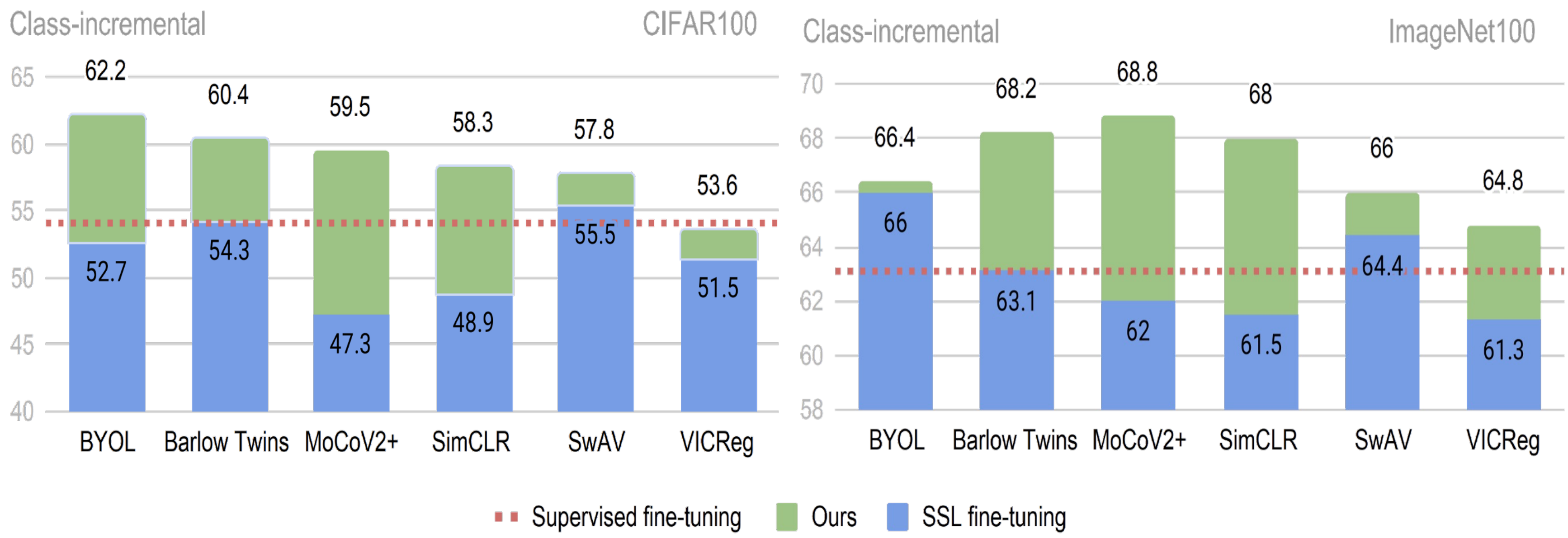
- **CIFAR100**
- **ImageNet100**
- **DomainNet**

[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Experiments

→ It turns out that by just using this simple technique you can obtain significant improvements:

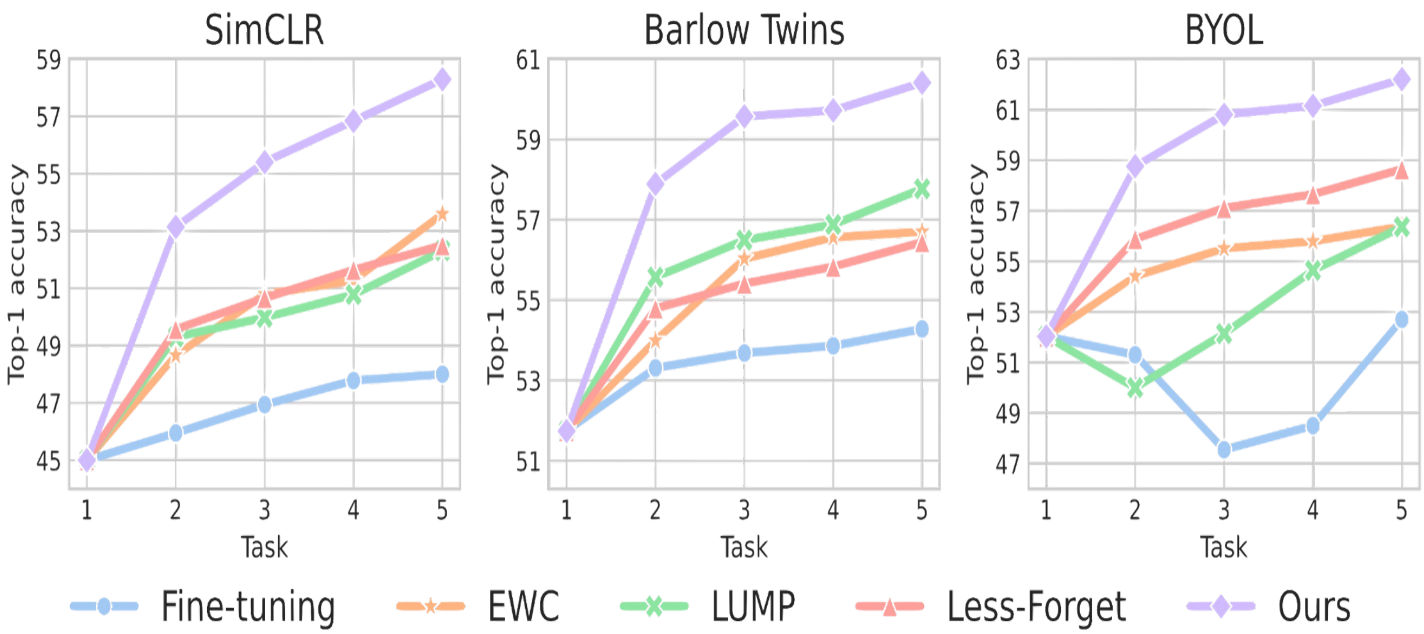


[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Experiments

➔ …and outperform other CL methods by large margins throughout the whole training trajectory:



[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

Experiments

➔ …and yield better forward transfer:

[Fini et al., CVPR 2022]

Slide courtesy: E. Fini

ICCV23
PARIS

A soft nearest-neighbor framework for continual semi-supervised learning

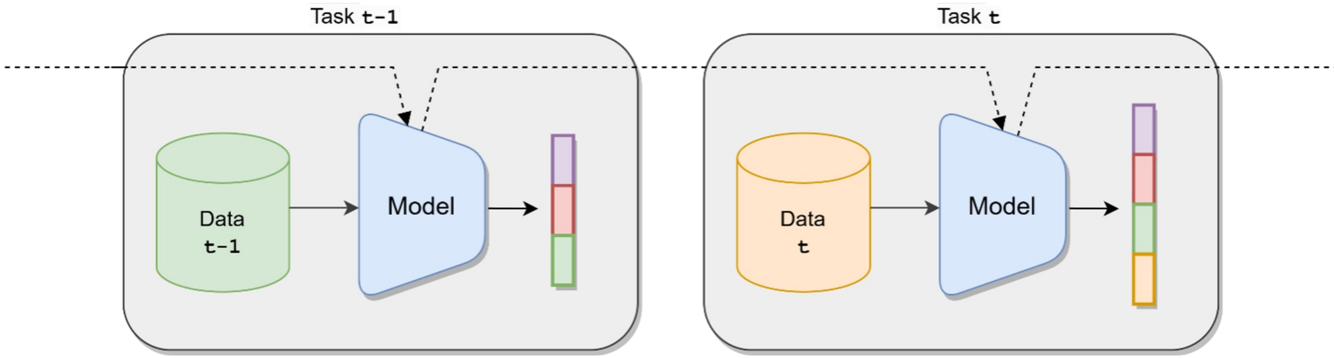
Zhiqi Kang*, Enrico Fini*, Moin Nabi, Elisa Ricci and Karteek Alahari

SCAN ME

Inria
UNIVERSITÀ DI TRENTO
SAP
FONDAZIONE BRUNO KESSLER

Continual semi-supervised learning

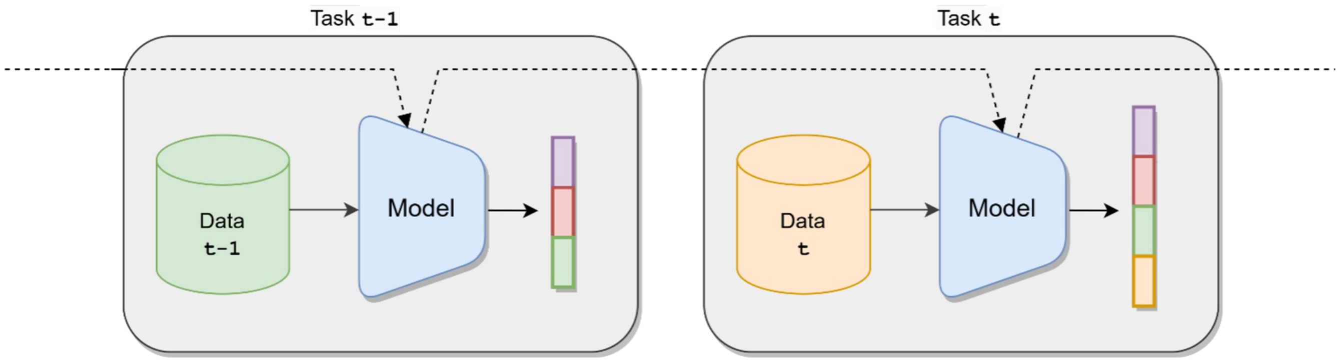
Continual Learning (CL) deals with a dynamic learning scenario.



Kang, Zhiqi, et al. *"A soft nearest-neighbor framework for continual semi-supervised learning."*

Continual semi-supervised learning

Continual Learning (CL) deals with a dynamic learning scenario.



Why is the **semi-supervised setting** realistic and interesting?

Annotating the entire dataset (**fully supervised**) can be:

- Expensive
- Impractical

Training on unannotated datasets (**self-supervised**) can be:

- Expensive
- Complex

Semi-supervised setting allows for a good trade-off

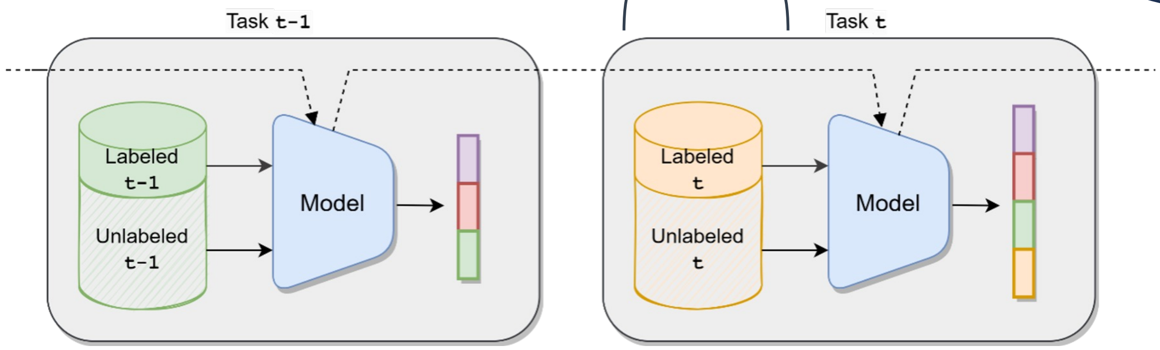
Kang, Zhiqi, et al. "A soft nearest-neighbor framework for continual semi-supervised learning."

Continual semi-supervised learning

New scenario [1, 2], new challenges:

Conventional semi-supervised learning methods?

Overfitting the small labeled set



Forgetting the previous tasks

Conventional continual learning methods?

[1] Wang, L., Yang, K., Li, C., Hong, L., Li, Z., and Zhu, J. (2021). Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In CVPR.
[2] Boschini, M., Buzzega, P., Bonicelli, L., Porrello, A., and Calderara, S. (2022). Continual semi-supervised learning through contrastive interpolation consistency. Pattern Recognition Letters.