Revisiting Supervision for Continual Representation Learning

Daniel Marczak[1,2] Sebastian Cygert[1,3] Tomasz Trzciński [1,2,4] Bartłomiej Twardowski[1,5]

[1]IDEAS NCBR [2]Warsaw University of Technology [3]Gdańsk University of Technology [4]Tooploox [5]Computer Vision Center Barcelona

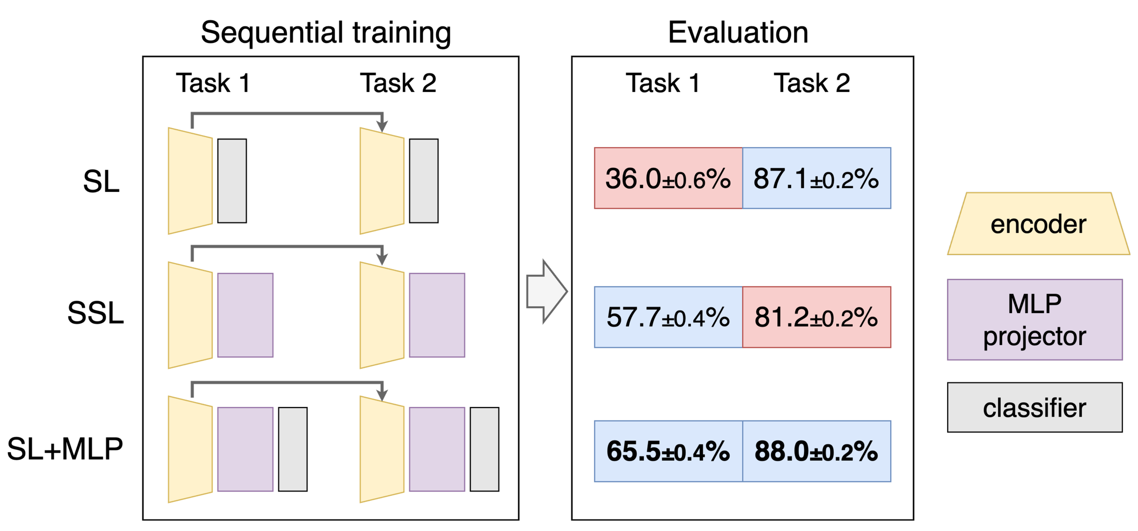
#TLDR

- We investigate Continual Representation Learning, the problem of training a feature extractor on a sequence of disjoint datasets.
- A number of studies show that unsupervised approaches outperform supervised approaches in this task [3, 2].
- We find it counter-intuitive and reckon that additional information, such as human annotations, should not deteriorate the quality of representations.
- Recent works identify that a multi-layer perceptron (MLP) projector is a crucial component responsible for superior transferability of SSL models [6, 1] and it can also improve the transferability of supervised models [5, 4].
- Encouraged by the advancements in improving the transferability of supervised models, we revisit supervision for continual representation learning. We are the first to show that **supervised models can continually learn representations of higher quality than self-supervised models** when trained with a simple MLP head.

Background

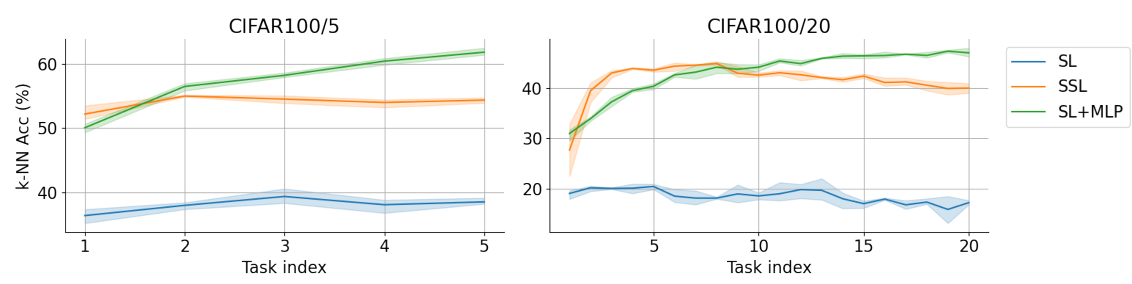
Imagine that you trained an image encoder, e.g. DINO [1], on a certain dataset. After some time you gathered much more data and you would like to **improve your image encoder using the new data**. You would like to improve your model whenever you gather a significant amount of new data, potentially an infinite number of times. The question is **how to do it efficiently**, ideally without accessing the old data which may be no longer accessible, e.g. due to the privacy reasons. This problem of training a backbone model on a sequence of disjoint datasets (tasks) is known as **Continual Representation Learning**.

Method



- We train the models on a sequence of two disjoint tasks.
- After the sequential training we evaluate the models separately on each task.
- Supervised learning (SL) results in representations that perform well on the second task but poorly on the first task.
- Representations trained with self-supervised learning (SSL) have higher first-task performance but they underperform on the second task.
- We show that adding a simple MLP projector to supervised learning (SL+MLP) yield representations that are **superior on the first task and on par with SL on the second task**.

Finetuning results



- SL+MLP achieves **strong performance after the initial task** compared to SL which indicates that it produces representations that are transferable to the unseen tasks.
- SL+MLP is the only method that is able to **accumulate knowledge** learned on a sequence of tasks.

Results with continual learning strategies

Method	CL strategy	C10/5	C100/5	C100/20	IN100/5
	Supervised Continual Learning				
SL	Finetune	56.9±1.4	38.5±0.4	17.2±0.3	35.3±1.3
	LwF	62.2±1.1	57.4±0.2	45.2±1.2	60.5±0.3
	PFR	68.5±1.5	57.7±0.4	44.4±1.3	58.7±0.2
SL+MLP	Finetune	65.9±0.7	61.9±0.5	47.1±0.7	62.4±0.4
	LwF	72.6±3.4	58.7±0.2	51.9±0.1	60.4±0.2
	PFR	76.3±1.0	**63.6±0.2**	**54.5±0.2**	65.2±0.1
t-ReX	Finetune	69.3±1.1	59.2±0.6	50.8±0.1	59.2±0.6
	LwF	74.5±0.7	58.3±0.4	50.4±0.1	58.6±1.0
	PFR	75.9±1.2	60.9±0.5	53.4±0.3	63.9±0.6
SupCon	Finetune	60.4±0.6	49.4±0.3	30.0±0.7	57.6±0.6
	CaSSLe	75.1±0.4	61.1±0.2	49.2±1.2	**70.4±0.6**
	PFR	**78.1±1.0**	57.0±0.2	51.2±0.8	68.0±0.7
	Unsupervised Continual Learning				
BarlowTwins	Finetune	76.2±1.2	54.1±0.3	40.0±0.8	57.0±0.4
	CaSSLe	**80.9±0.2**	**58.6±0.6**	49.3±0.1	**64.9±0.1**
	PFR	78.8±0.6	57.2±0.2	46.0±0.7	61.1±0.2
SimCLR	Finetune	72.4±1.3	48.9±0.4	33.4±0.5	54.7±0.4
	CaSSLe	80.6±0.5	55.9±0.5	48.2±0.4	59.3±0.5
	PFR	79.2±0.7	53.8±0.3	**49.4±0.1**	57.7±0.2

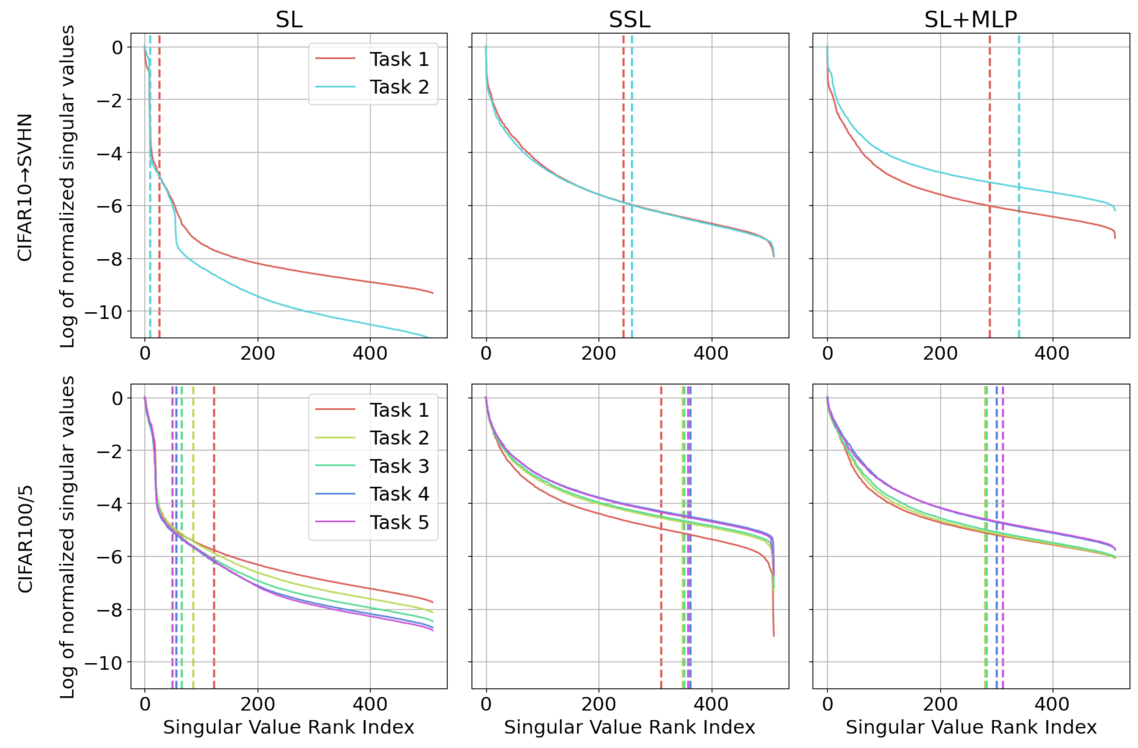
- All the supervised methods equipped with the projector significantly outperform simple SL.
- The positive effects of the MLP projector and CL strategy compound.
- The best models are those (1) trained in a supervised way (2) with the use of the MLP projector and (3) coupled with CL strategy based on temporal learnable projection, namely CaSSLe or PFR.

Forgetting

Training sequence	SL $Acc_{C10}\uparrow$	SL $F_{C10}\downarrow$	SSL $Acc_{C10}\uparrow$	SSL $F_{C10}\downarrow$	SL+MLP $Acc_{C10}\uparrow$	SL+MLP $F_{C10}\downarrow$
C10	92.6	-	88.8	-	**93.3**	-
C100	74.9	-	80.8	-	**84.5**	-
C10→C100	76.1	16.6	79.1	9.7	**88.8**	**4.5**
SVHN	21.8	-	**58.6**	-	56.3	-
C10→SVHN	22.6	70.1	54.9	33.8	**62.7**	**30.6**

- We observe high representation forgetting for SL, significantly lower for SSL, and the lowest for SL equipped with MLP projector.
- We can see that only SL+MLP is able to retain a significant part of pretraining features.

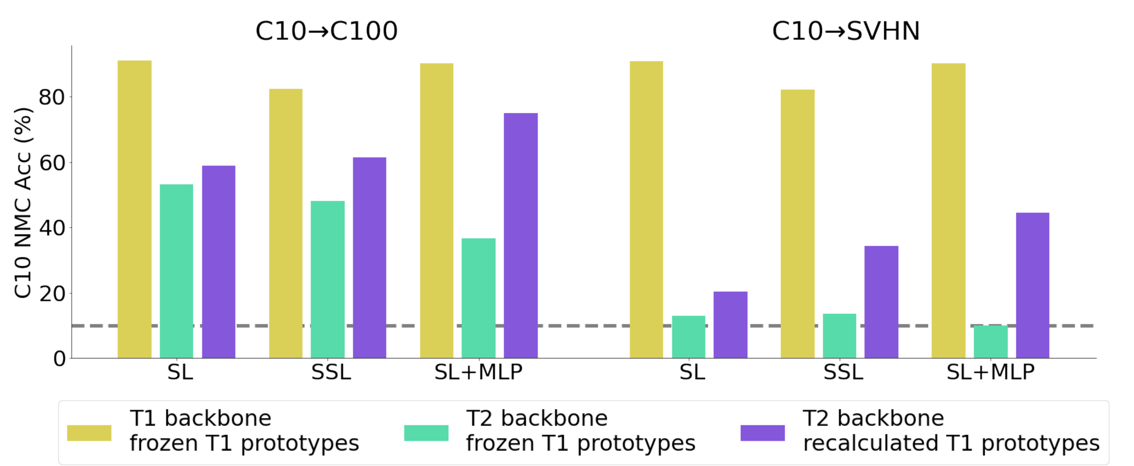
Spectral Analysis



Representations learned with SL+MLP (right) exhibit desirable properties from the continual learning point of view:

- they consist of a more diverse set of features (contrary to SL, left)
- they improve feature diversity when learning new tasks consistently across all the presented settings

Stability of representations



Take-Away Points

- Supervised learning can significantly outperform self-supervised learning in continual representation learning.
- The key is training a supervised model with a simple MLP projector discarded after the training, following the common practice from SSL.
- We shed some light on the reasons for improved performance when using MLP with SL: better transferability, lower forgetting, and higher diversity of learnt features.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *International Conference on Computer Vision (ICCV)*, 2021.

[2] Mohammad Reza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[4] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. No reason for no supervision: Improved generalization in supervised models. In *International Conference on Learning Representations (ICLR)*, 2023.

[5] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning (ICML)*, 2021.