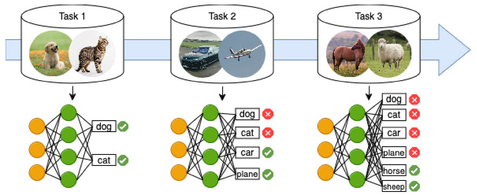
Memory-efficient Continual Learning with Neural Collapse Contrastive

Trung-Anh Dang[1], Vincent Nguyen[1], Ngoc-Son Vu[2], Christel Vrain[1]
[1]Université d'Orléans, INSA CVL, LIFO UR 4022, Orléans, France
[2]ETIS - CY Cergy Paris University, ENSEA, CNRS, France

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Context

- In real-life situations, ML models need to learn continuously from non-stationary data.
- Human learning is adaptive, ongoing, and expands on previous knowledge.
- Current DNNs are learned in batches, requiring huge amount of data to retrain with new classes.
- Challenge: **Catastrophic forgetting (CF)** problem.



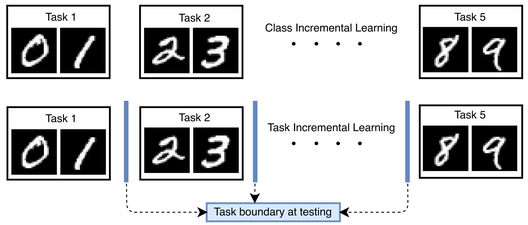
- Models need to consider memory systems trade-off between storing old knowledge and acquiring new one.

Related works

Current prevalent approaches:

- **Regularization-based** (e.g.: Knowledge Distillation (KD)): Force the current model's parameters to be sufficiently close to the past model.
- **Replay or rehearsal based**: A small portion of previous seen samples are stored and mixed with current data

Different scenarios: Task-IL, Domain-IL, Class-IL, Data-IL



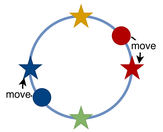
- In some settings, existing task boundaries can be provided or not to the model during training/test time.
- The construction of splits between tasks depends on different scenario.

Preliminaries

- **Supervised Contrastive Learning:** Learning objectives tending to maximize similarity between representations of same-class samples while minimizing similarity between different-class samples.

$$\mathcal{L}_{SupCon} = \sum_{i=1}^{2N} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log\left(\frac{e^{\langle \mathbf{z}_i \cdot \mathbf{z}_j \rangle / \tau}}{\sum_{\substack{k=1 \\ k \neq i}}^{2N} e^{\langle \mathbf{z}_i \cdot \mathbf{z}_k \rangle / \tau}}\right) \quad (1)$$

- **Neural Collapse:** Behaviour of last layer features that collapse to their class means, aligned with a simplex equiangular tight frame (ETF).



Representations tend to move towards prototype corresponding to their class.

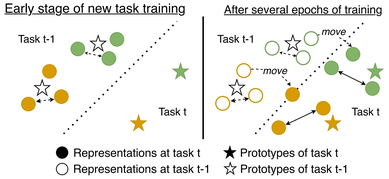
Hard learning & Soft learning

In learning new tasks:

- "Soft" learning relies solely on inter-sample relationships, which lead to representation drift and overlap with new tasks.
- "Hard" learning focuses only on sample-prototype relationships, which can reduce diversity, disrupt within-class data distribution, and lead to forgetting as older representations shift towards current task prototypes.

In preserving previous knowledge:

- "Hard" stability hinders the learning of new knowledge when starting training a new task.
- "Soft" stability diminishes its effectiveness as training progresses.



Proposed Method

- **Main idea:** Integrating "hard" and "soft" relationships in both $\mathcal{L}_{plasticitiy}$ ($\mathcal{L}_{FNC^2}$) and $\mathcal{L}_{stability}$ ($\mathcal{L}_{HSD}$).

Plasticity loss

$$\mathcal{L}_{FNC^2} = -\sum_{i=1}^{2N} \frac{1}{|P_i|+1} \left(\sum_{\mathbf{z}_j \in P_i} (1 - c_{ij})^\gamma \log(c_{ij}) + (1 - r_i)^\gamma \log(r_i) \right) \quad (2)$$

where

$$\begin{cases} c_{ij} = \frac{e^{\langle \mathbf{z}_i \cdot \mathbf{z}_j \rangle / \tau}}{A}; r_i = \frac{e^{\langle \mathbf{z}_i \cdot \mathbf{p}_{\mathbf{z}_i} \rangle / \tau}}{A} \\ A = \sum_{k \neq i} e^{\langle \mathbf{z}_i \cdot \mathbf{z}_k \rangle / \tau} + \underbrace{\sum_{\mathbf{p}_m \in \mathbf{P}_{1:t-1}} e^{\langle \mathbf{z}_i \cdot \mathbf{p}_m \rangle / \tau}}_{\text{cross-task}} \end{cases}$$

Stability loss

$$\mathcal{L}_{HSD} = (1-\alpha) \underbrace{\sum_{i=1}^{2N} -\mathbf{o}^{t-1}(\mathbf{x}_i) \log(\mathbf{o}^t(\mathbf{x}_i))}_{\mathcal{L}_{IRD}} + \alpha \underbrace{\sum_{i=1}^{2N} -\mathbf{q}^{t-1}(\mathbf{x}_i; \mathbf{P}_{1:t}) \log(\mathbf{q}^t(\mathbf{x}_i; \mathbf{P}_{1:t}))}_{\mathcal{L}_{S-PRD}} \quad (3)$$

where

$$\begin{cases} \mathbf{o}^t(\mathbf{x}_i) = [o^t_{i,1}, ..., o^t_{i,i-1}, o^t_i, o^t_{i,i+1}, ..., o^t_{i,2N}], o^t_{i,j} = \frac{e^{\langle \mathbf{z}^t_i \cdot \mathbf{z}^t_j \rangle / \kappa}}{\sum_{k \neq i} e^{\langle \mathbf{z}^t_i \cdot \mathbf{z}^t_k \rangle / \kappa}} \\ \mathbf{q}^t(\mathbf{x}_i; \mathbf{P}_{1:t}) = [q^t_{i,1}, q^t_{i,2}, ..., q^t_{i,S}], q^t_{i,j} = \frac{e^{\langle \mathbf{z}^t_i \cdot \mathbf{p}_j \rangle / \zeta}}{\sum_{s=1}^{S} e^{\langle \mathbf{z}^t_i \cdot \mathbf{p}_s \rangle / \zeta}} \end{cases}$$

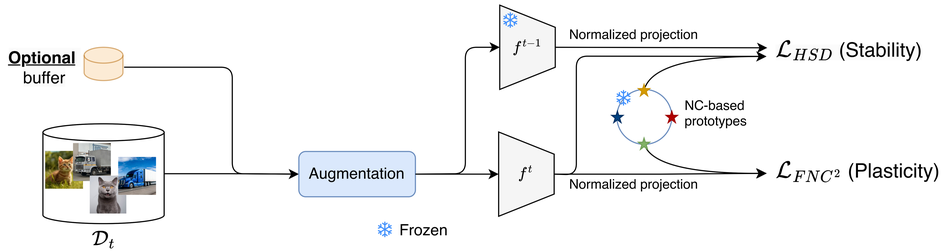


Figure: Overall architecture of our method.

Results

- By exploring both hard and soft relationships in NC-based CL, we achieve SoTA performance in memory-free settings while remaining competitive with limited buffer size scenarios.

Buffer	Dataset / Scenario	Seq-Cifar-10 Class-IL	Seq-Cifar-10 Task-IL	Seq-Cifar-100 Class-IL	Seq-Cifar-100 Task-IL	Seq-Tiny-ImageNet Class-IL	Seq-Tiny-ImageNet Task-IL
0	Co^{2}L (ICCV'21)[2]	58.89±2.61	86.65±1.05	26.89±0.78	51.91±0.63	13.43±0.57	40.21±0.68
	Ours	**69.26±0.32**	**94.41±0.43**	**32.57±0.55**	**57.87±0.62**	**14.54±0.52**	**43.81±0.47**
200	iCaRL (CVPR'17) [3]	49.02±3.20	88.99±2.13	28.00±0.91	51.43±1.47	7.53±0.79	28.19±1.47
	DER (NeurIPS'20)[1]	61.93±1.79	91.40±0.92	31.23±1.38	63.09±1.09	11.87±0.78	40.22±0.67
	Co^{2}L (ICCV'21)[2]	65.57±1.37	93.43±0.78	27.38±0.85	53.94±0.76	13.88±0.40	42.37±0.74
	GCR (CVPR'22)[4]	64.84±1.63	90.80±1.05	33.69±1.40	**64.24±0.83**	13.05±0.91	42.11±1.01
	CILA (ICML'24)[5]	67.06±1.59	94.29±0.24	-	-	14.55±0.39	44.15±0.70
	Ours	**72.63±0.78**	**95.31±0.32**	**34.04±0.42**	59.46±0.65	**15.52±0.53**	**44.59±0.72**

Table: Results of our method compared with other supervised baselines using the Average Accuracy (AA) metric.

Ablation studies

- Both $\mathcal{L}_{FNC^2}$ and $\mathcal{L}_{HSD}$ show effectiveness, especially when they are combined concurrently.

Plasticity	Stability	Buffer size 0	Buffer size 200
$\mathcal{L}_{FNC^2}$	✗	53.59±0.63	53.62±0.81
$\mathcal{L}^{asym}_{SupCon}$	✗	53.25±1.70	53.57±1.03
$\mathcal{L}^{asym}_{SupCon}$	$\mathcal{L}_{IRD}$	58.89±2.61	65.57±1.37
$\mathcal{L}_{FNC^2}$	$\mathcal{L}_{IRD}$	63.65±0.55	70.54±0.95
$\mathcal{L}_{FNC^2}$	$\mathcal{L}_{S-PRD}$	64.17±0.41	69.20±0.58
$\mathcal{L}_{FNC^2}$	$\mathcal{L}_{HSD}$	**69.26±0.32**	**72.63±0.78**

Table: Performance comparison in Class-IL setting on the Seq-Cifar-10 dataset.

Limitations & Future work

- Like other NC-inducing methods in CL, our approach is limited by the need to pre-define prototypes, which is impractical when the number of prototypes is unknown.
- Propose predefined maximum prototypes to address the need to know the exact number in advance.
- Explore alternative memory-free evaluation methods, as current approaches including Co^{2}L [2], CILA [5] rely on buffers, which are less effective with limited samples for old classes.
- Identify easily forgotten samples and focus on distilling only the core knowledge.

References

[1] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. "Dark Experience for General Continual Learning: a Strong, Simple Baseline". *NeurIPS*. 2020.
[2] H. Cha, J. Lee, and J. Shin. "Co2L: Contrastive Continual Learning". *ICCV*. 2021.
[3] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. "iCaRL: Incremental Classifier and Representation Learning". *CVPR*. 2017.
[4] R. Tiwari, K. Killamsetty, R. Iyer, and P. Shenoy. "GCR: Gradient Coreset Based Replay Buffer Selection For Continual Learning". *CVPR*. 2022.
[5] Y. Wen, Z. Tan, K. Zheng, C. Xie, and W. Huang. "Provable Contrastive Continual Learning". *ICML*. 2024.

Contact Information

Please see more details in our paper or by scanning the project QR code: