20th April 2023

Neural Collapse Terminus:
A Unified Solution for Class Incremental Learning and Its Variants

Yibo Yang[1#], Haobo Yuan[2#], Xiangtai Li[3], Jianlong Wu[4], Lefei Zhang[2], Zhouchen Lin[3], Philip H.S. Torr[5], Dacheng Tao[6], Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology, Jeddah, Saudi Arabia.
[2]School of Computer Science, Wuhan University, Wuhan, China.
[3]National Key Lab. of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, China.
[4]Harbin Institute of Technology (Shenzhen), China.
[5]University of Oxford, Oxford, United Kingdom.
[6]University of Sydney, Sydney, Australia.
[#]Equal Contribution

Possible TPAMI extension of their 2023 ICLR paper :
Neural Collapse Inspired Feature-Classifier Alignment for Few-Shot Class Incremental Learning

Course Instructor

Prof. Vineeth N Balasubramanian

Presenter

Rahul Vigneswaran K*

CS23MTECH02002

*Presented as part of the coursework for Advanced Topics in Machine Learning (CS6360)

- Contributions
- Problem setup
- Misalignment Dilemma
- Methodology
  - Neural Collapse
  - Flying to collapse
- Results
- Theoretical Setup
  - Problem
  - Definition
  - Theorem

Contributions

- Points out that the catastrophic forgetting in multiple class incremental learning tasks derive from the same origin, **the misalignment dilemma**.

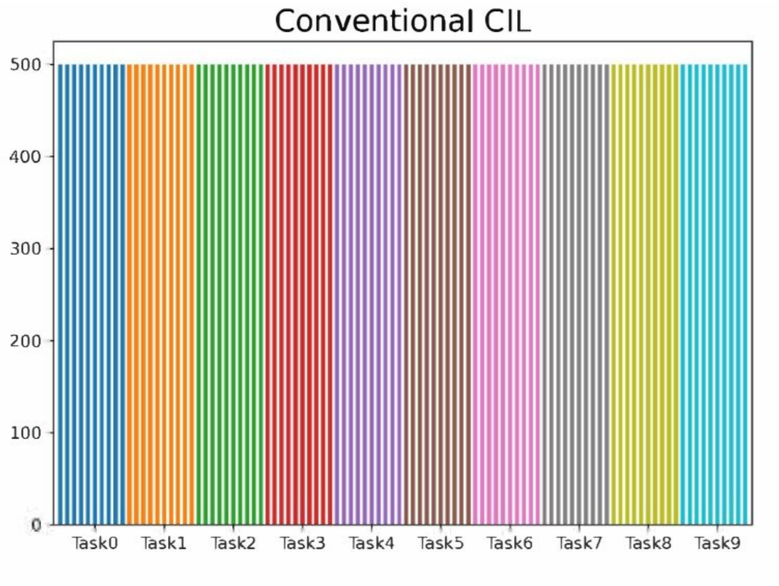
- First to propose a unified solution for CIL, LT-CIL, FS-CIL.

- For CIL and LTCIL, proposes a prototype evolving scheme named **flying to collapse** to avoid sharp shift.

- Proposes a novel misalignment loss.

- Performs theoretical analysis to show that the proposed method can induce the neural collapse optimality regardless of incremental training or data imbalance.

Problem setup



Train:

$\{\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(T)}\}$, where $\mathcal{D}^{(t)} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}^{(t)}|}$

$\mathcal{C}^{(t)} = \text{set}(\{y_i | (x_i, y_i) \in \mathcal{D}^{(t)}\})$

$\mathcal{C}^{(t)} \cap \mathcal{C}^{(t')} = \emptyset, \forall t' \neq t$

$\{\mathcal{D}^{(0)}\}$ is base session

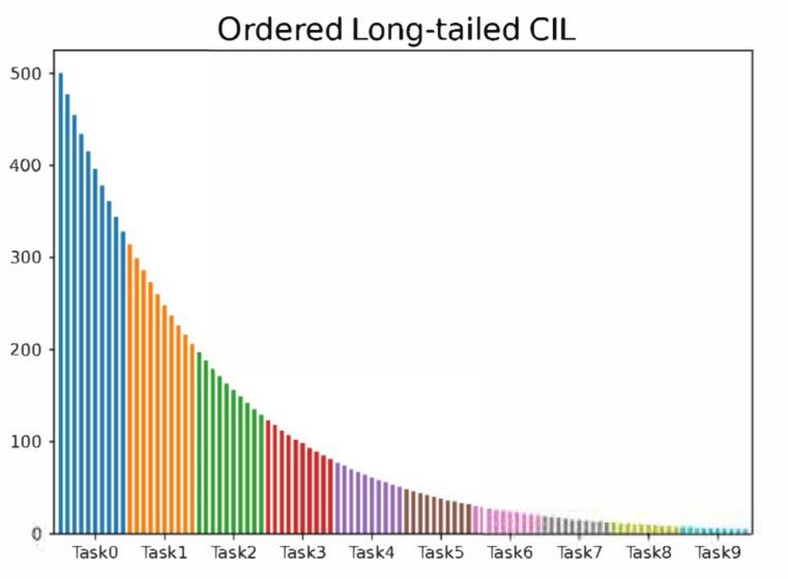
$\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(T)}\}$ are called incremental sessions

Test:

$\hat{\mathcal{C}}^{(t)} = \cup_{i=0}^{t} \mathcal{C}^{(i)}$

X. Liu, Y.-S. Hu, X.-S. Cao, A. D. Bagdanov, K. Li, and M.-M. Cheng, “Long-tailed class incremental learning,” in ECCV, 2022.

Problem setup



Train:

$\{\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(T)}\}$, where $\mathcal{D}^{(t)} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}^{(t)}|}$

$\mathcal{C}^{(t)} = \text{set}(\{y_i | (x_i, y_i) \in \mathcal{D}^{(t)}\})$

$\mathcal{C}^{(t)} \cap \mathcal{C}^{(t')} = \emptyset, \forall t' \neq t$

The number of training samples follows an exponential decay across all classes.

$\{\mathcal{D}^{(0)}\}$ is base session

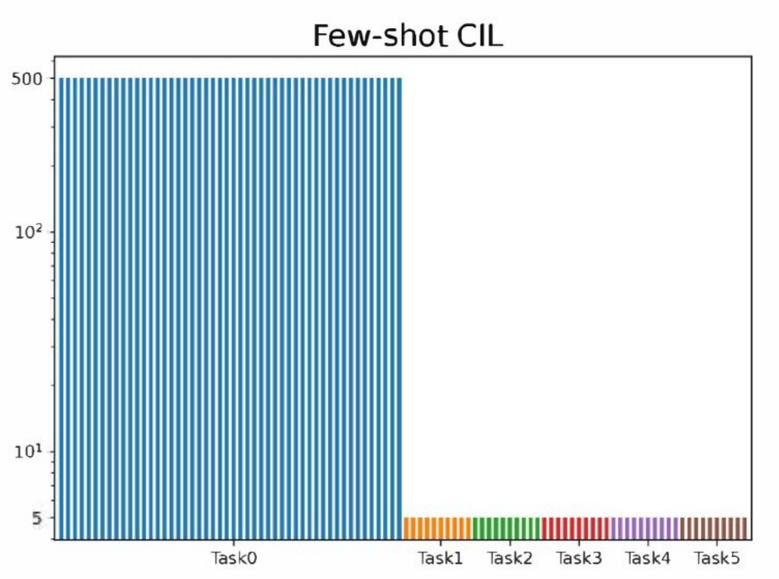
$\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(T)}\}$ are called incremental sessions

Test:

$\hat{\mathcal{C}}^{(t)} = \cup_{i=0}^{t} \mathcal{C}^{(i)}$

X. Liu, Y.-S. Hu, X.-S. Cao, A. D. Bagdanov, K. Li, and M.-M. Cheng, “Long-tailed class incremental learning,” in ECCV, 2022.

Problem setup



Train:

$\{\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(T)}\}$, where $\mathcal{D}^{(t)} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}^{(t)}|}$

$\mathcal{C}^{(t)} = \text{set}(\{y_i | (x_i, y_i) \in \mathcal{D}^{(t)}\})$

$\mathcal{C}^{(t)} \cap \mathcal{C}^{(t')} = \emptyset, \forall t' \neq t$

$\{\mathcal{D}^{(0)}\}$ is base session

$\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(T)}\}$ are called incremental sessions

$\mathcal{D}^{(t)}, t > 0$, only has a few labeled images

$|\mathcal{D}^{(t)}| = pq$, p-way q-shot

Not allowed to store exemplars

Test:

$\hat{\mathcal{C}}^{(t)} = \cup_{i=0}^{t} \mathcal{C}^{(i)}$

X. Liu, Y.-S. Hu, X.-S. Cao, A. D. Bagdanov, K. Li, and M.-M. Cheng, “Long-tailed class incremental learning,” in ECCV, 2022.

Misalignment Dilemma

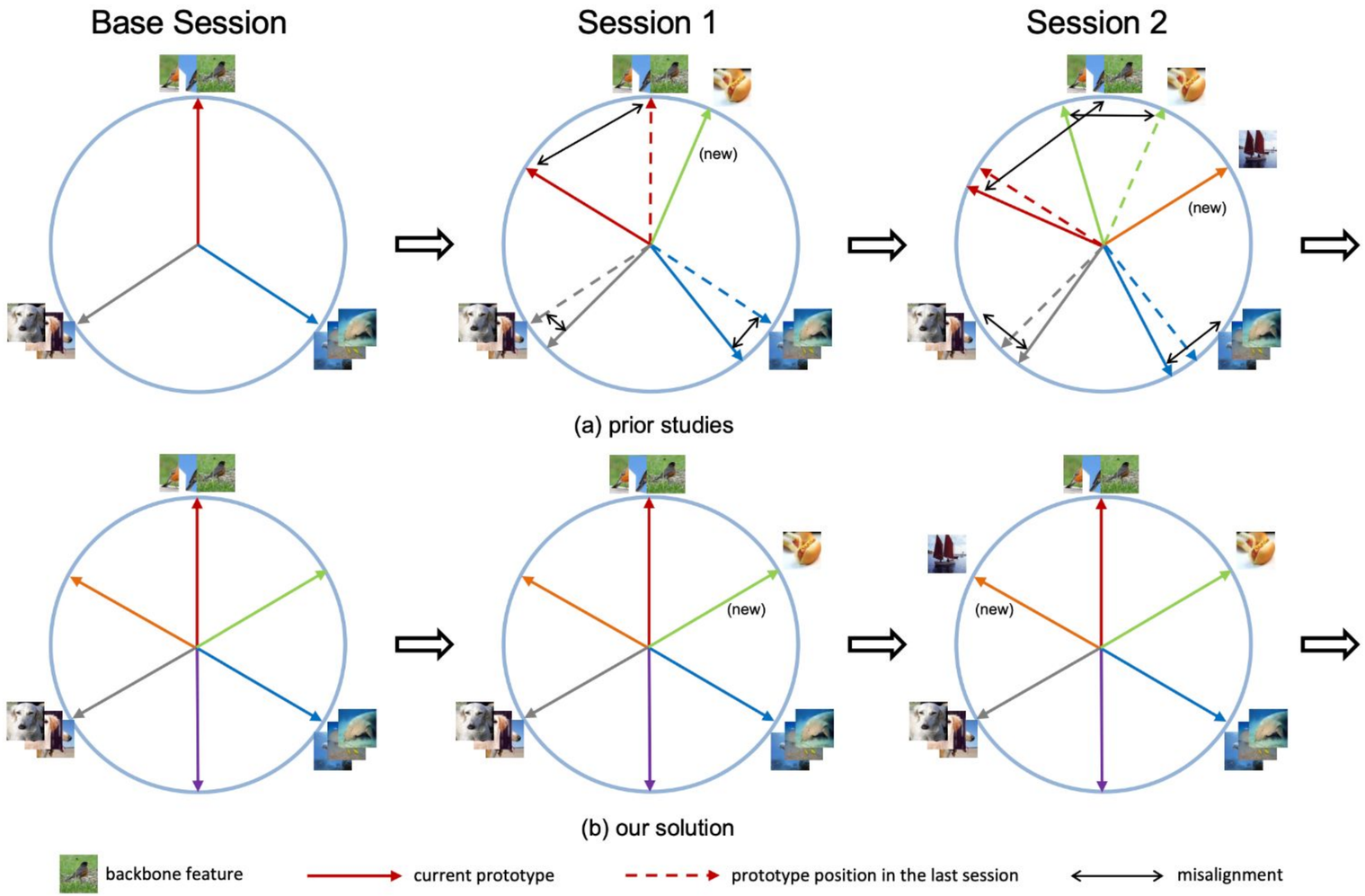


Fig. 1: A sketch comparison between prior studies and our solution. (a) Prior studies evolve the old-class prototypes via delicate loss or regularizer to keep them separated from the new-class ones, but will cause misalignment with the backbone features that are usually kept stable (and even fixed in FSCIL). (b) Our solution pre-divides the feature space as a consistent target and trains a model towards the same optimality to avoid dividing the feature space incrementally.

Neural Collapse

1. The last-layer features of the same class will collapse into their within-class mean.

   $\Sigma_W^{(k)} \to \mathbf{0}$, where $\Sigma_W^{(k)} = \mathbf{Avg}_i\{(\mu_{k,i} - \mu_k)(\mu_{k,i} - \mu_k)^T\}$,

   $\mu_{k,i}$ is the feature of sample i in class k,
   μ_k is the within-class feature mean of class k

2. The feature means of all classes centered by the global mean will converge to the vertices of a simplex ETF.

3. The feature means centered by the global mean will be aligned with their corresponding class prototypes (classifier vector), which means that the class prototypes will converge to the same simplex ETF,

   $\hat{\mu}_k = \mathbf{w}_k / \|\mathbf{w}_k\|, 1 \leq k \leq K$

   $\mathbf{w}_k$ is the class prototype of class k

V. Papyan, X. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” PNAS, 2020.

Neural Collapse

4. When **(1)** – **(3)** hold, the model prediction based on logits can be simplified to select the nearest class center.

V. Papyan, X. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” PNAS, 2020.

Flying to Collapse

$$\mathbf{w}_c^{(NCM)} = \mathbf{Avg}_i\{\hat{\mu}_i | y_i = c, c \in \mathcal{C}^{(t)}\}$$

c is a novel class,
$\hat{\mu}_i = \mu_i / \|\mu_i\|$,
$\mu_i = f(x_i, \theta_f)$ is the initial backbone feature of input x_i, and y_i is its label.

$$\mathbf{w}_c = \eta \hat{\mathbf{w}}_c^{(NCT)} + (1 - \eta)\hat{\mathbf{w}}_c^{(NCM)},\ \eta = \frac{e}{E}$$

$\hat{\mathbf{w}}_c^{(NCM)}$ is the NCM prototype ℓ_2-normalization,
$\hat{\mathbf{w}}_c^{(NCT)}$ is the prototype in our $\hat{\mathbf{W}}_{\mathrm{ETF}}$ for class c,
e is the epoch index,
E is the number of total epochs used to train this session,
η gradually evolves from 0 to 1 as training goes on.

Flying to Collapse

$$\mathcal{L}_{\text{align}}\left(\hat{\mu}_i, \hat{\mathbf{w}}_{y_i}\right) = \frac{1}{2}\left(\hat{\mathbf{w}}_{y_i}^T \hat{\mu}_i - 1\right)^2$$

$\hat{\mu}_i$ is the ℓ_2-normalized backbone feature of the i-th sample,
y_i is its class label,
$\hat{\mathbf{w}}_{y_i}$ is the ℓ_2-normalized class prototype for y_i

$$\mathcal{L}_{\text{distill}}\left(\hat{\mu}_i^{(t-1)}, \hat{\mu}_i^{(t)}\right) = \frac{1}{2}\left(\left(\hat{\mu}_i^{(t-1)}\right)^T \hat{\mu}_i^{(t)} - 1\right)^2$$

$\hat{\mu}_i^{(t-1)}$ is the feature generated by the old backbone network in the session $(t-1)$.

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda \cdot \mathcal{L}_{\text{distill}}$$

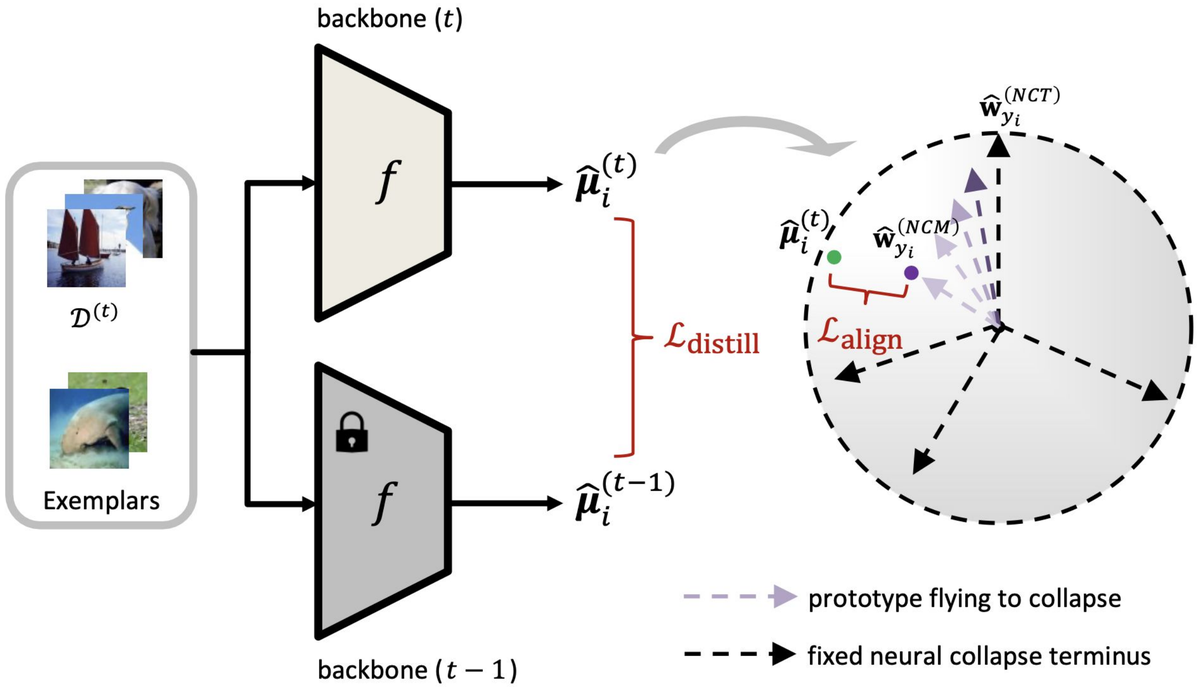


Fig. 2: An illustration of our architecture for CIL and LTCIL. The input includes the training data $\mathcal{D}^{(t)}$ of the current session t, and the exemplars allowable for CIL and LTCIL. The arrows moving from light to dark represent the prototypes using our flying to collapse strategy. It starts from the nearest class mean $\hat{\mathbf{w}}_{y_i}^{(NCM)}$ and terminates at our neural collapse terminus $\hat{\mathbf{w}}_{y_i}^{(NCT)}$ by Eq. (5). We adopt this simple pipeline for both CIL and LTCIL without bells and whistles.

Results

TABLE 2: Comparison of average incremental accuracy on ImageNet-1k. The base session has an equal number of classes to each incremental session (e.g., "10 steps" means 100 classes for base or each incremental session). The number of exemplars for each class is 20. The results of previous studies are reproduced by [23].

Method	ImageNet-1k	
	10 steps	20 steps
LwF [58]	40.86	27.72
iCaRL [6]	49.56	42.61
LUCIR [30]	56.40	52.75
PODNet [31]	57.01	54.06
AANet [16]	51.76	46.86
CwD [23]	58.18	56.01
NC-CIL (Ours)	**58.21**	**57.58**

Results

TABLE 3: Comparison of average incremental accuracy on CIFAR100-LT and ImageNet100-LT. The base session is with 50 classes. The number of exemplars for each class is 20. The results of LUCIR [30] and PODNet [31] are reproduced by [7]. LT-CIL [7] is a two-stage method with LUCIR and PODNet baseline. Our NC-LTCIL has the same architecture as NC-CIL **without** any change or hyper-parameter tuning. ↑ indicates improvement.

Method	Mode	CIFAR100-LT ($\rho = 0.01$)				ImageNet100-LT ($\rho = 0.01$)			
		5 steps	↑	10 steps	↑	5 steps	↑	10 steps	↑
LUCIR [30]	Ordered	42.69	**+14.00**	42.15	**+16.54**	52.91	**+10.89**	52.80	**+10.88**
PODNet [31]	Ordered	44.07	**+12.62**	43.96	**+14.73**	58.78	**+5.02**	58.94	**+4.74**
LT-CIL [7]	Ordered	45.88	**+10.81**	45.73	**+12.96**	58.82	**+4.98**	59.09	**+4.59**
NC-LTCIL (Ours)	Ordered	**56.69**	-	**58.69**	-	**63.80**	-	**63.68**	-
LUCIR [30]	Shuffled	35.09	**+10.54**	34.59	**+13.57**	45.80	**+6.67**	46.52	**+9.65**
PODNet [31]	Shuffled	34.64	**+10.99**	34.84	**+13.32**	49.69	**+2.78**	51.05	**+5.12**
LT-CIL [7]	Shuffled	39.40	**+6.23**	39.00	**+9.16**	52.08	**+0.39**	52.60	**+3.57**
NC-LTCIL (Ours)	Shuffled	**45.63**	-	**48.16**	-	**52.47**	-	**56.17**	-

Results

TABLE 4: Performance of FSCIL on miniImageNet and comparison with other studies. The top rows list class-incremental learning and few-shot learning results implemented by [8], [37] in the FSCIL setting. "Average Acc." is the average incremental accuracy. "Final Improv." calculates the improvement of our method after the last session over prior studies. The last row lists the improvements of our method over ALICE [76], which is a strong baseline in FSCIL.

Methods	**Accuracy in each session (%) ↑**									**Average**	**Final**
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Acc.**	**Improv.**
iCaRL [6]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	33.29	**+41.1**
NCM [30]	61.31	47.80	39.30	31.90	25.70	21.40	18.70	17.20	14.17	30.83	**+44.14**
D-Cosine [33]	70.37	65.45	61.41	58.00	54.81	51.89	49.10	47.27	45.63	55.99	**+12.68**
TOPIC [8]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64	**+33.89**
IDLVQ [40]	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84	51.16	**+16.47**
Self-promoted [81]	61.45	63.80	59.53	55.53	52.50	49.60	46.69	43.79	41.92	52.76	**+16.39**
CEC [37]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75	**+10.68**
LIMIT [80]	72.32	68.47	64.30	60.78	57.95	55.07	52.70	50.72	49.19	59.06	**+9.12**
Regularizer [39]	80.37	74.68	69.39	65.51	62.38	59.03	56.36	53.95	51.73	63.71	**+6.58**
MetaFSCIL [79]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	58.85	**+9.12**
C-FSCIL [38]	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41	61.61	**+6.90**
Data-free Replay [114]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	58.02	**+10.10**
ALICE [76]	80.60	70.60	67.40	64.50	62.50	60.00	57.80	56.80	55.70	63.99	**+2.61**
NC-FSCIL (ours)	**84.02**	**76.80**	**72.00**	**67.83**	**66.35**	**64.04**	**61.46**	**59.54**	**58.31**	**67.82**	
Improvement over ALICE	+3.42	+6.20	+4.60	+3.33	+3.85	+4.04	+3.66	+2.74	+2.61	**+3.83**	

Problem Setup

We consider the following problem,

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right), \ 0 \leq t \leq T, \qquad (16)$$
$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \leq 1, \quad \forall 1 \leq k \leq K^{(t)}, \ 1 \leq i \leq n_k,$$

where $\mathbf{m}_{k,i}^{(t)} \in \mathbb{R}^d$ denotes a feature variable that belongs to the i-th sample of class k in session t, n_k is number of samples in class k, $K^{(t)}$ is number of classes in session t, $N^{(t)}$ is the number of samples in session t, *i.e.*, $N^{(t)} = \sum_{k=1}^{K^{(t)}} n_k$, and $\mathbf{M}^{(t)} \in \mathbb{R}^{d \times N^{(t)}}$ denotes a collection of $\mathbf{m}_{k,i}^{(t)}$. $\hat{\mathbf{W}}_{\mathrm{ETF}} \in \mathbb{R}^{d \times K}$ refers to the neural collapse terminus for the whole label space, where $K = \sum_{t=0}^{T} K^{(t)}$.

$$\mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{w}}_k\right) = \frac{1}{2}\left(\hat{\mathbf{w}}_k^T \mathbf{m}_{k,i}^{(t)} - 1\right)^2$$

$$\mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right) = -\log \frac{\exp(\hat{\mathbf{w}}_k^T \mathbf{m}_{k,i}^{(t)})}{\sum_{j=1}^{K} \exp(\hat{\mathbf{w}}_j^T \mathbf{m}_{k,i}^{(t)})}$$

Definition

Definition 1 (Simplex Equiangular Tight Frame). *A simplex equiangular tight frame (ETF) refers to a collection of vectors* $\{\mathbf{e}_i\}_{i=1}^{K}$ *in* $\mathbb{R}^d$, $d \geq K-1$, *that satisfies:*

$$\mathbf{e}_{k_1}^T \mathbf{e}_{k_2} = \frac{K}{K-1}\delta_{k_1,k_2} - \frac{1}{K-1}, \quad \forall k_1, k_2 \in [1, K], \tag{1}$$

where $\delta_{k_1,k_2} = 1$ *when* $k_1 = k_2$, *and 0 otherwise. All vectors have the same* ℓ_2 *norm and any pair of two different vectors has the same inner product of* $-\frac{1}{K-1}$, *which is the minimum possible cosine similarity for* K *equiangular vectors in* $\mathbb{R}^d$.

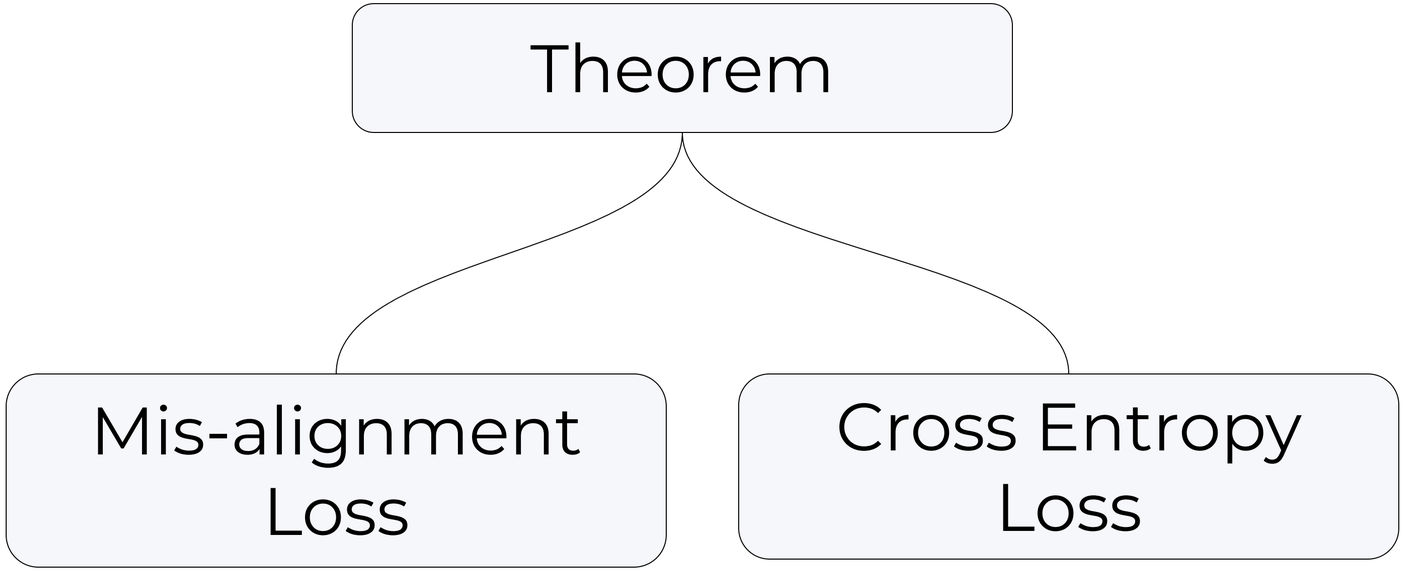
Theorem

Theorem. *Let $\hat{\mathbf{M}}^{(t)}$ denotes the global minimizer of Eq. (16) by optimizing the model incrementally from $t = 0$, and we have $\hat{\mathbf{M}} = [\hat{\mathbf{M}}^{(0)}, \cdots, \hat{\mathbf{M}}^{(T)}] \in \mathbb{R}^{d \times \sum_{t=0}^{T} N^{(t)}}$. No matter if $\mathcal{L}$ in Eq. (16) is CE or misalignment loss, for any column vector $\hat{\mathbf{m}}_{k,i}$ in $\hat{\mathbf{M}}$ whose class label is k, we have:*

$$\|\hat{\mathbf{m}}_{k,i}\| = 1, \quad \hat{\mathbf{m}}_{k,i}^{T} \hat{\mathbf{w}}_{k'} = \frac{K}{K-1} \delta_{k,k'} - \frac{1}{K-1}, \tag{17}$$

for all $k, k' \in [1, K]$, $1 \leq i \leq n_k$, where $K = \sum_{t=0}^{T} K^{(t)}$ denotes the total number of classes of the whole label space, $\delta_{k,k'} = 1$ when $k = k'$ and 0 otherwise, and $\hat{\mathbf{w}}_{k'}$ is the class prototype in $\hat{\mathbf{W}}_{\mathrm{ETF}}$ for class k'.

Proof Flow



References

- X. Liu, Y.-S. Hu, X.-S. Cao, A. D. Bagdanov, K. Li, and M.-M. Cheng, “Long-tailed class incremental learning,” in ECCV, 2022.

- V. Papyan, X. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” PNAS, 2020.

- Yang, Yibo, et al. "Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants." *arXiv preprint arXiv:2308.01746* (2023).