NEURAL INFORMATION PROCESSING SYSTEMS

Overcoming Recency Bias of Normalization Statistics in Continual Learning: Balance and Adaptation

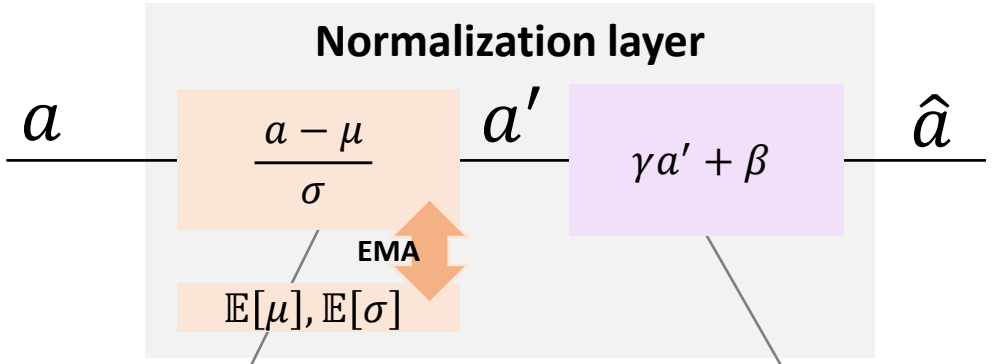
Yilin Lyu[1*], Liyuan Wang[2*], Xingxing Zhang[2†], Zicheng Sun[1], Hang Su[2], Jun Zhu[2], Liping Jing[1†]

[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
[2] Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab, Tsinghua-Bosch Joint Center for ML, Tsinghua University

[*]Equal contribution [†]Corresponding authors

Background & Motivation

- Key challenge in CL: Catastrophic forgetting (recency bias)



Statistics parameters
(Statistics-based update)

Unaddressed issue

Affine transformation parameters
(Gradient-based optimization)

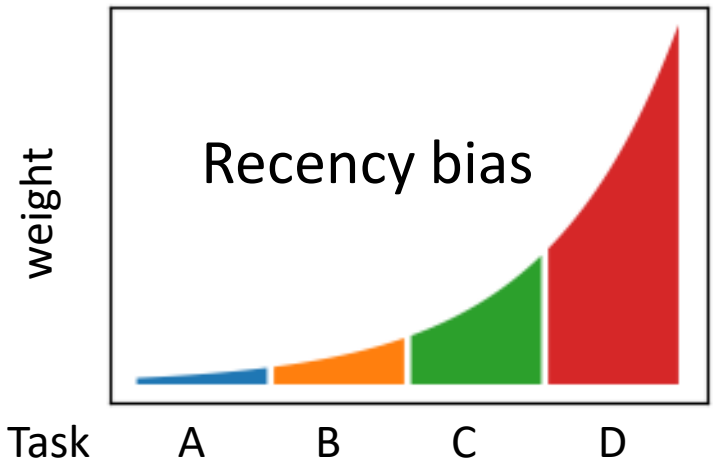
Existing efforts

Background & Motivation

- The update rule of BN's **statistics parameters**: exponential moving average (EMA)

$$\hat{\mathbb{E}}[S_m] := \hat{\mathbb{E}}[S|a_1, \cdots, a_m] = (1-\eta)\hat{\mathbb{E}}[S|a_1, \cdots, a_{m-1}] + \eta S_m$$

where $\eta \in (0, 1)$ and $S(\cdot) = [\mu(\cdot), \sigma^2(\cdot, \mu(\cdot))]^\top$



Imbalanced task-wise contributions to statistics

- To what degree?
- How does replay help?
- How to balance?

Analysis

- **Thm. 1**: Statistical weight for EMA

$$w_t = \left[\sum_{i=m_{t-1}+1}^{m_t} \eta_i r_i \prod_{j=i+1}^{m_T} (1-\eta_j) + \sum_{i=m_t+1}^{m_T} \frac{\eta_i (1-r_i)}{T-1} \prod_{j=i+1}^{m_T} (1-\eta_j) \right] / Z$$

This exposes a dilemma between *balance* and *adaptation.*

- **Cor. 2**: *Adaptation* of BN statistics

$$w_t = \frac{\bar{\eta}^{m_T - t} - \bar{\eta}^{m_T - t + 1}}{1 - \bar{\eta}^{m_T}} / Z$$

Without the use of replay, the statistical weights of past tasks decrease exponentially.

- **Cor. 3**: *Balance* of BN statistics

An improved strategy is needed to reconcile *balance* and *adaptation*.

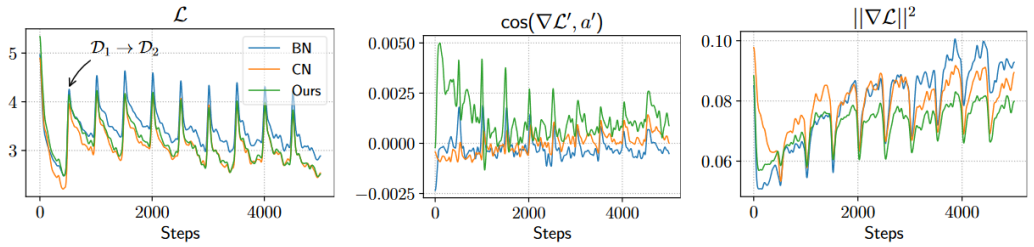
- **Cor. 4**: Statistical weight for CMA

An improved strategy needs to take into account the parameters of the current model

- **Thm. 5:** BN may fail to stabilize and improve generalization in continual learning

$$\|\nabla_{a_{[:,j]}} \mathcal{L}\|^2 \leq \frac{\gamma^2}{\sigma^2_{[:,j]} C} \left[C \|\nabla_{a_{[:,j]}} \mathcal{L}'\|^2 - (\mathbf{1}^\top \nabla_{a_{[:,j]}} \mathcal{L}')^2 - (\nabla^\top_{a_{[:,j]}} \mathcal{L}' \cdot a'_{[:,j]})^2 \right]$$

The upper bound of the gradient magnitude will be loosened by the increase of gradient similarity, which potentially affects the benefit of BN.



(a) Loss (b) Gradient Similarity (c) Gradient Magnitude

Figure 1: Dynamics of the loss of training batches, gradient similarity (i.e., cosine similarity between gradients and normalized representations) and gradient magnitude.

Solution (AdaB^{2}N)

- Training aspect
  - A Bayesian-based strategy:

$$a'_m = \frac{a_m - \mathbb{E}[\mu|a_m]}{\sqrt{\mathbb{E}[\sigma^2|a_m] + \varepsilon}}$$

$$\hat{\mathbb{E}}[S|a_m] = \sum_\tau \sum_t \mathcal{P}_\phi(\tau|a_m)\mathcal{P}_\theta(t|\tau, a_m)S_m^t = \sum_t \frac{\phi_t + N_t}{\bar{\phi} + N} S_m^t$$

where ϕ is learnable

  - Objective function

$$\min_{\theta,\psi} \mathcal{L}_{CL}(\theta) + \lambda \sum_\ell \mathcal{L}_{ada}^{(\ell)}(\theta, \psi)$$

$$\mathcal{L}_{ada}^{(\ell)}(\theta, \psi) = \|\hat{\mathbb{E}}[S|a_m] - \hat{\mathbb{E}}[S_m]\|^2$$

- Testing aspect
  - Striking an balance between EMA and CMA

$$\eta_i := \eta(i, \eta_{i-1}) = \frac{\eta_{i-1}}{\eta_{i-1} + (1 - \tilde{\eta})^\kappa}, \quad \eta_0 := \tilde{\eta}^\kappa$$

- Performance
  - Achieves statistically significant improvements in both **online** and **offline** scenarios ($p < 0.05$)

Table 1: Performance of **online task-incremental learning** with batch size $|B| = 10$. We report final average accuracy of all seen tasks (↑) with ± standard deviation.

Method	Split CIFAR-10		Split CIFAR-100		Split Mini-ImageNet	
	$\|\mathcal{M}\|$=500	$\|\mathcal{M}\|$=2000	$\|\mathcal{M}\|$=2000	$\|\mathcal{M}\|$=5000	$\|\mathcal{M}\|$=2000	$\|\mathcal{M}\|$=5000
ER-ACE w/ BN	86.34±2.35	89.17±1.21	61.21±1.63	63.83±1.94	63.62±3.14	64.60±1.73
ER-ACE w/ LN	78.01±6.78	81.59±4.39	42.29±0.29	44.32±0.35	45.91±1.63	45.82±1.91
ER-ACE w/ IN	84.57±1.42	86.03±1.99	49.46±2.25	50.15±1.71	41.07±1.66	40.93±4.38
ER-ACE w/ GN	81.46±3.74	82.85±2.14	44.84±1.41	42.50±2.06	43.45±3.29	45.43±0.62
ER-ACE w/ CN	88.32±1.43	90.01±0.95	61.00±1.58	63.42±0.65	64.23±1.22	65.14±1.48
ER-ACE w/ Ours	**88.74±1.77**	**90.84±2.01**	**63.88±1.58**	**67.01±2.90**	**66.78±1.91**	**68.03±0.41**
DER++ w/ BN	87.61±1.67	90.42±1.83	65.53±1.17	66.23±0.94	61.70±2.40	61.28±1.29
DER++ w/ LN	81.35±4.25	82.80±5.48	43.24±0.86	44.42±2.90	40.05±3.67	37.64±4.26
DER++ w/ IN	84.81±2.99	87.04±2.60	47.39±1.46	49.17±2.13	32.88±1.16	35.23±6.11
DER++ w/ GN	82.05±1.24	83.34±3.34	43.54±1.47	45.12±1.15	40.65±1.76	38.26±3.26
DER++ w/ CN	86.92±0.89	89.75±0.76	66.20±0.38	67.39±1.88	65.09±1.76	66.14±1.40
DER++ w/ Ours	**90.15±2.52**	**91.99±0.81**	**70.33±0.49**	**71.70±0.84**	**66.42±3.62**	**69.12±2.51**

Table 4: Performance of **offline task-incremental learning** and **class-incremental learning** on Split Mini-ImageNet with memory size $|\mathcal{M}| = 2000$.

Method	Task-IL			Class-IL		
	ER-ACE	DER++	LiDER	ER-ACE	DER++	LiDER
BN	36.36±2.33	35.91±1.59	36.96±1.92	10.85±0.54	12.69±1.23	**10.99±0.74**
CN	35.83±1.43	35.46±1.42	36.23±2.62	10.00±0.62	12.13±0.92	9.97±0.36
Ours	**37.64±0.85**	**36.98±1.52**	**37.36±0.79**	**11.05±0.10**	**13.07±1.30**	*10.96±0.67*

Thank you!

Paper link:
https://arxiv.org/abs/2310.08855

Code link:
https://github.com/lvyilin/AdaB2N