Optimizing Mode Connectivity for Class Incremental Learning

Haitao Wen, Haoyang Cheng, Heqian Qiu, Lanxiao Wang, Lili Pan, Hongliang Li

University of Electronic Science and Technology of China

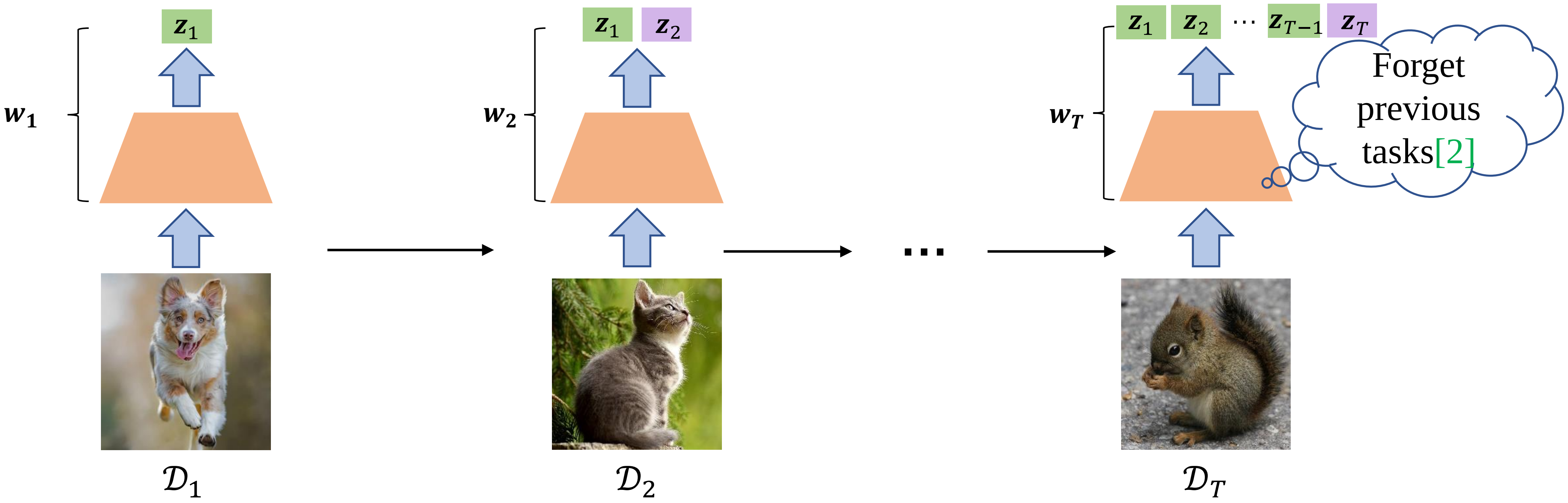
Email: haitaowen@std.uestc.edu.cn

https://github.com/HaitaoWen/EOPC

Introduction

Backbone   Classifier head   New classifier head

$\mathcal{D}_i$: dataset of the i-th task $\boldsymbol{w}_i$: model parameters the i-th task $\boldsymbol{z}_i$: new classifier parameters for i-th task

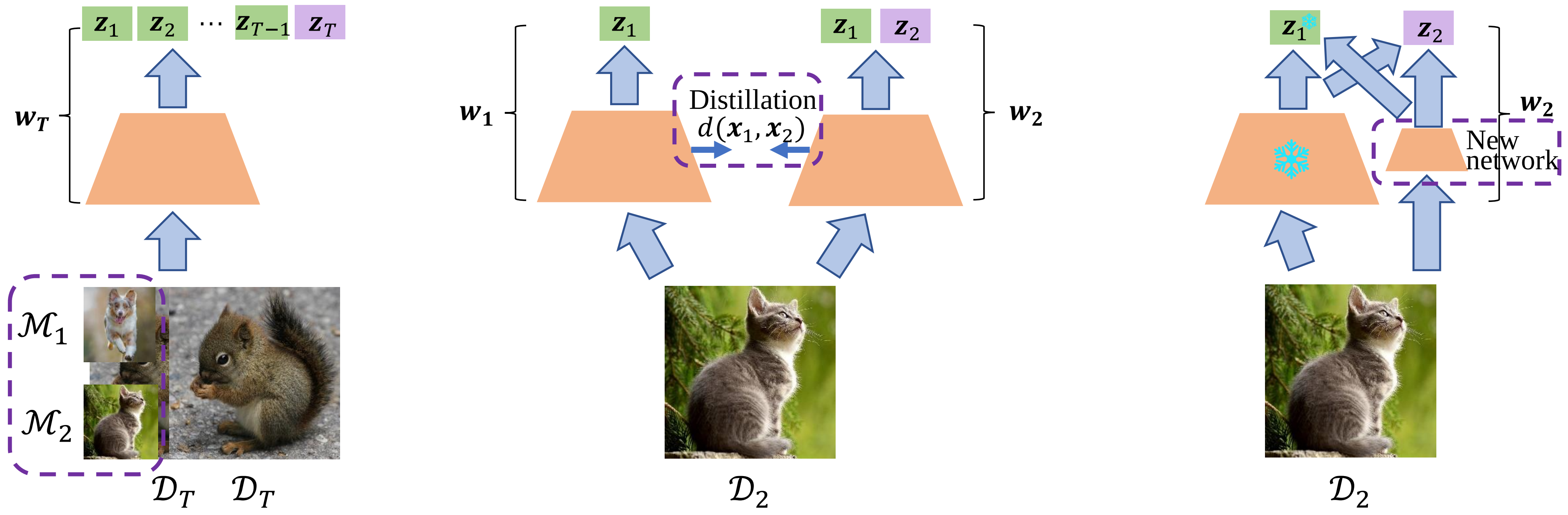


The process of class incremental learning (CIL)[1].

[1] Van de Ven, Gido M., and Andreas S. Tolias. "Three scenarios for continual learning." arXiv preprint arXiv:1904.07734 (2019).
[2] McCloskey, Michael, and Neal J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem." Psychology of learning and motivation. Vol. 24. Academic Press, 1989. 109-165.

Existing Work

$\mathcal{M}_i$: memory of the i-th task $\boldsymbol{x}_i$: feature of the i-th task $d(\boldsymbol{x}, \boldsymbol{y})$: similarity metric ❄: frozen



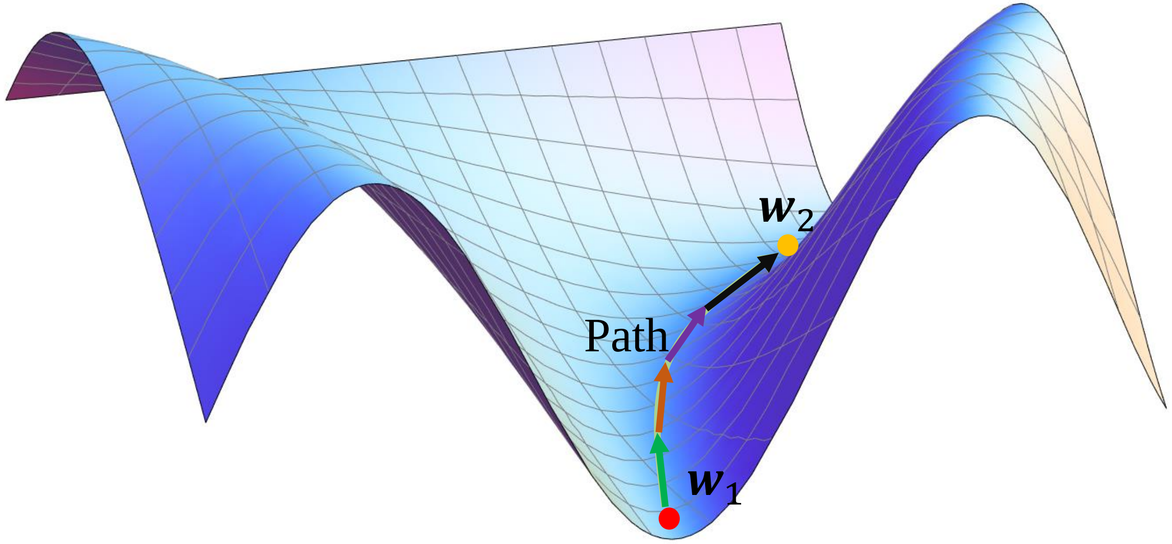
Memory Replay[3] Regularization[4] Dynamic Architecture[5]

[3] Liu, Yaoyao, et al. "Mnemonics training: Multi-class incremental learning without forgetting." Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2020.

[4] Douillard, Arthur, et al. "Podnet: Pooled outputs distillation for small-tasks incremental learning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer International Publishing, 2020.

[5] Liu, Yaoyao, Bernt Schiele, and Qianru Sun. "Adaptive aggregation networks for class-incremental learning." Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2021.

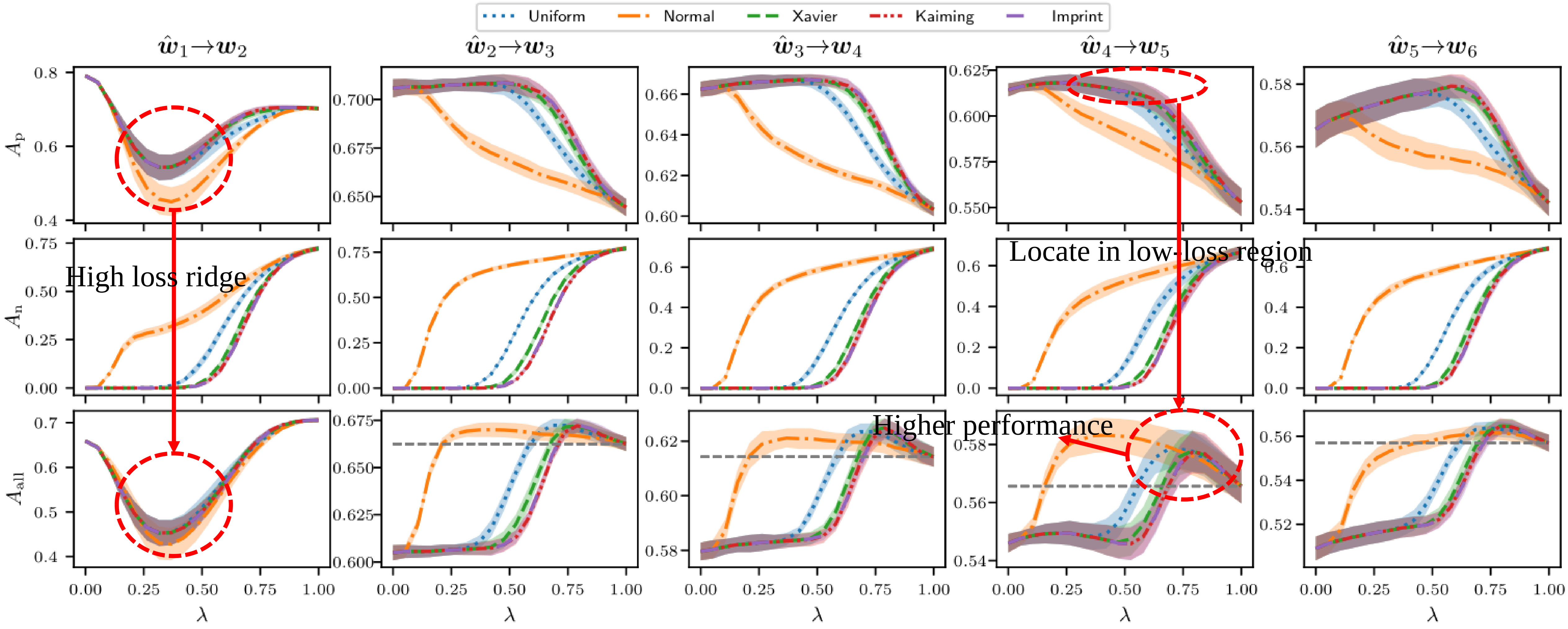
Mode Connectivity



Loss landscape[6]

[6] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." Advances in neural information processing systems 31 (2018).

Linear Connectivity in CIL



Testing accuracy curves along the linear connection between two adjacent continual minima of PODNet [4] for 5 steps of increments (i.e., 6 tasks in total) on CIFAR-100. A p , A n , and A all denote accuracy on previous tasks, on the new task, and on all learned tasks respectively. λ is the interpolation factor.

OPC: Optimizing Connectivity between Minima

Let $\boldsymbol{p_\theta}(\lambda)$: $[0,1] \rightarrow \mathbb{R}^n$ be the parameterized arbitrary path between minima $\widehat{\boldsymbol{w}}_{t-1}$ and $\boldsymbol{w}_t$, such that

$$\boldsymbol{p_\theta}(0) = \widehat{\boldsymbol{w}}_{t-1} \quad \text{and} \quad \boldsymbol{p_\theta}(1) = \boldsymbol{w}_t$$

We commonly use the expected loss $\hat{\ell}(\boldsymbol{\theta})$ along the path to characterize its quality, i.e.,

$$\hat{\ell}(\boldsymbol{\theta}) = \int_0^1 \mathcal{L}\big(\boldsymbol{p_\theta}(\lambda)\big) d\lambda = \mathbb{E}_{\lambda \sim U(0,1)} \left[\mathcal{L}\big(\boldsymbol{p_\theta}(\lambda)\big)\right],$$

where $\mathcal{L}$ is the task loss, such as cross-entropy loss, NCA loss [4], or embedding loss [7], $U(0,1)$is the uniform distribution on the interval [0,1].

we can randomly sample points λ between [0,1] and minimize loss $\mathcal{L}\big(\boldsymbol{p_\theta}(\lambda)\big)$ with respect to $\boldsymbol{\theta}$ to optimize the path, i.e.,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \gamma \boldsymbol{\nabla_\theta} \boldsymbol{\mathcal{L}}\big(\boldsymbol{p_\theta}(\lambda)\big), \lambda \sim U[0,1]$$

However, there are significant differences between continual minima because of catastrophic forgetting, we can not directly connect them each other. We redefine a low-loss path taking named switching point (SP) as a bridge, where the part between the previous minimum and SP is for previous tasks, and the part between SP and the new minimum is for the new task.

[7] Hou, Saihui, et al. "Learning a unified classifier incrementally via rebalancing." Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2019.

OPC: Optimizing Connectivity between Minima

The expected loss along the path can be reformulated for continual learning as follows,

$$\ell(\boldsymbol{\theta})=\int_0^{\lambda^*} \mathcal{L}_{1:t-1}\big(\boldsymbol{p}_{\boldsymbol{\theta}}(\lambda)\big)d\lambda + \int_{\lambda^*}^{1} \mathcal{L}_t\big(\boldsymbol{p}_{\boldsymbol{\theta}}(\lambda)\big)d\lambda,$$

where λ^* corresponds to SP in the interval $[0,1]$, $\mathcal{L}_{1:t-1}$ is loss on previous tasks and $\mathcal{L}_t$ is loss on the new task.

OPC: Optimizing Connectivity between Minima

Connectivity Modeling:

Path modeled by Fourier series:

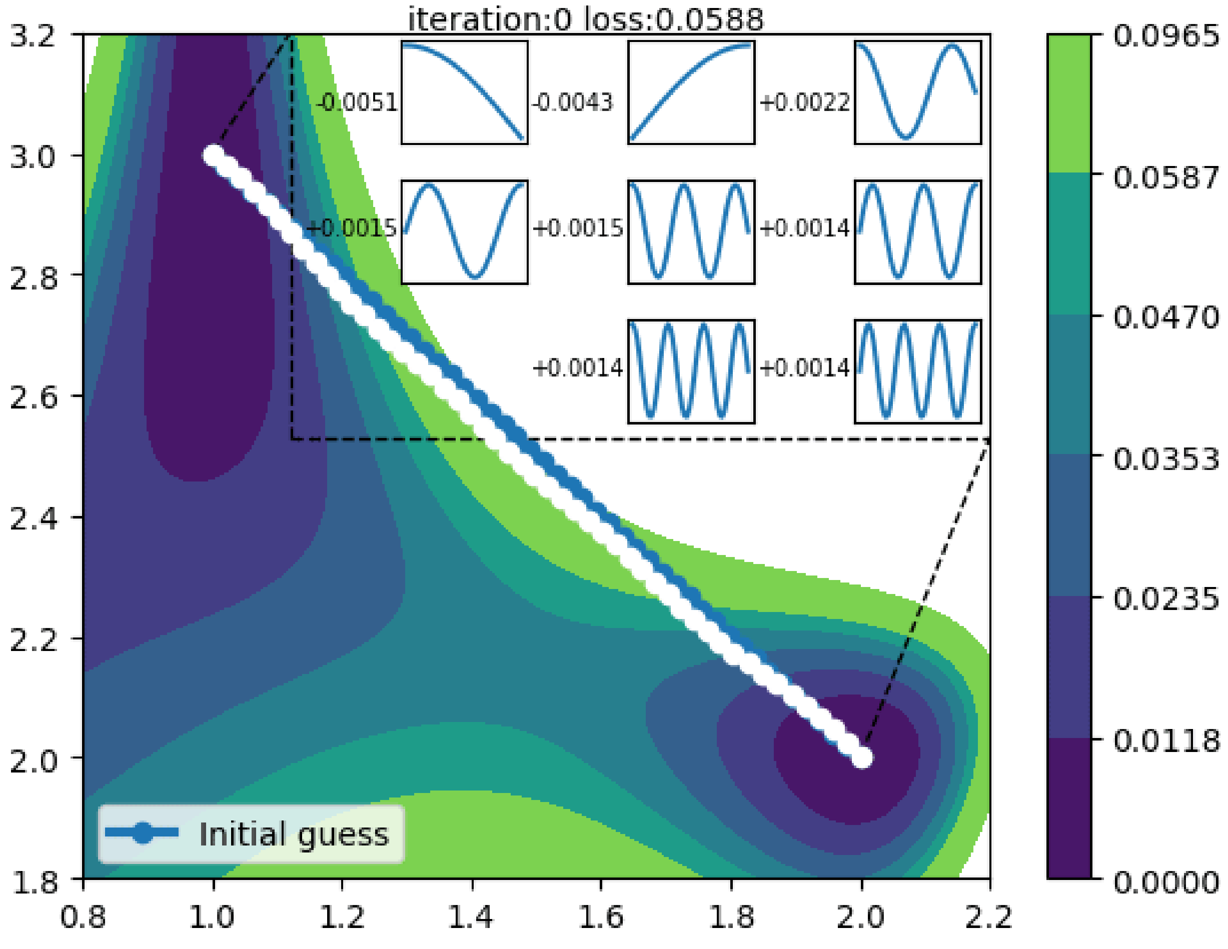
$$\boldsymbol{p_\theta}(\lambda) = $$
$$(\boldsymbol{AC} + (1-\lambda)\boldsymbol{1}_L)\cdot \widehat{\boldsymbol{w}}_{t-1} + (\boldsymbol{BS} + \lambda\boldsymbol{1}_L)\cdot \boldsymbol{w}_t$$

$$\boldsymbol{A} = \begin{bmatrix} \alpha_{1,1} & \cdots & \alpha_{1,N} \\ & \vdots & \\ \alpha_{L,1} & \cdots & \alpha_{L,N} \end{bmatrix} \boldsymbol{C} = \begin{bmatrix} \cos(\frac{\pi}{2}\lambda) \\ \vdots \\ \cos(\frac{(4N-3)\pi}{2}\lambda) \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,N} \\ & \vdots & \\ \beta_{L,1} & \cdots & \beta_{L,N} \end{bmatrix} \boldsymbol{S} = \begin{bmatrix} \sin(\frac{\pi}{2}\lambda) \\ \vdots \\ \sin(\frac{(4N-3)\pi}{2}\lambda) \end{bmatrix}$$

where $\boldsymbol{\theta} = [\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{B}^{\mathrm{T}}]^{\mathrm{T}}$, it must meet the following Condition to make path still pass through endpoints:

$$\boldsymbol{\theta 1}_N = \boldsymbol{0}_{2L}$$



A toy example of optimizing connectivity between minima in the 2-dimensional plane.

OPC: Optimizing Connectivity between Minima

Connectivity Regularization:

- The tangent of the path is

$$\boldsymbol{p}'_{\boldsymbol{\theta}}(\lambda) = (\boldsymbol{A}'\boldsymbol{S} - \boldsymbol{1}_L) \cdot \hat{\boldsymbol{w}}_{t-1} + (\boldsymbol{B}'\boldsymbol{C} + \boldsymbol{1}_L) \cdot \boldsymbol{w}_t$$

Where $A', B' \in \mathbb{R}^{L \times N}$, specifically,

$$\boldsymbol{A}' = \begin{bmatrix} -\frac{\pi}{2}\alpha_{1,1} & \cdots & -\frac{(4N-3)\pi}{2}\alpha_{1,N} \\ & \vdots & \\ -\frac{\pi}{2}\alpha_{L,1} & \cdots & -\frac{(4N-3)\pi}{2}\alpha_{L,N} \end{bmatrix}$$

$$\boldsymbol{B}' = \begin{bmatrix} \frac{\pi}{2}\beta_{1,1} & \cdots & \frac{(4N-3)\pi}{2}\beta_{1,N} \\ & \vdots & \\ \frac{\pi}{2}\beta_{L,1} & \cdots & \frac{(4N-3)\pi}{2}\beta_{L,N} \end{bmatrix}$$

- Then, randomly sample noise from the normal distribution and orthogonalize it with the tangent direction, i.e.,

$$\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}} - \frac{\boldsymbol{p}'_{\boldsymbol{\theta}}(\lambda)^{\mathrm{T}}\hat{\boldsymbol{\epsilon}}\boldsymbol{p}'_{\boldsymbol{\theta}}(\lambda)}{\|\boldsymbol{p}'_{\boldsymbol{\theta}}(\lambda)\|^2}, \quad \hat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

- Next, add normalized noise scaled by radius r to the path and obtain the point on the surface of the cylinder,

$$\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda) = \boldsymbol{p}_{\boldsymbol{\theta}}(\lambda) + r\frac{\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|}$$

- Replace $\boldsymbol{p}_{\boldsymbol{\theta}}(\lambda)$ with $\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda)$ in expected loss along the path, we can get the flattening connectivity regularization, i.e.,

$$\ell(\boldsymbol{\theta}) = \int_0^{\lambda^*} \mathcal{L}_{1:t-1}\big(\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda)\big)d\lambda + \int_{\lambda^*}^1 \mathcal{L}_t\big(\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda)\big)d\lambda$$

OPC: Optimizing Connectivity between Minima

Connectivity Optimization:

- The parameters of the path $\boldsymbol{\theta} = [\boldsymbol{A}^{\mathrm{T}}, \boldsymbol{B}^{\mathrm{T}}]^{\mathrm{T}}$ must meet the following condition to make path keep pass through endpoints:

$$\boldsymbol{\theta}\mathbf{1}_N = \mathbf{0}_{2L}$$

- We adopt gradient projection to update parameters along the direction orthogonal to the normal of equation, i.e.,:

$$\begin{aligned}\boldsymbol{\Delta}(\lambda) &= \nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda))(\boldsymbol{I}_{N\times N} - \mathbf{1}_N(\mathbf{1}_N^{\mathrm{T}}\mathbf{1}_N)^{-1}\mathbf{1}_N^{\mathrm{T}}) \\ &= \nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda)) - \frac{1}{N}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\lambda))\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}.\end{aligned}$$

- Then, the iterative rule for parameters of the path is as follows,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \gamma\boldsymbol{\Delta}(\lambda), \quad \lambda \sim U[0,1]$$

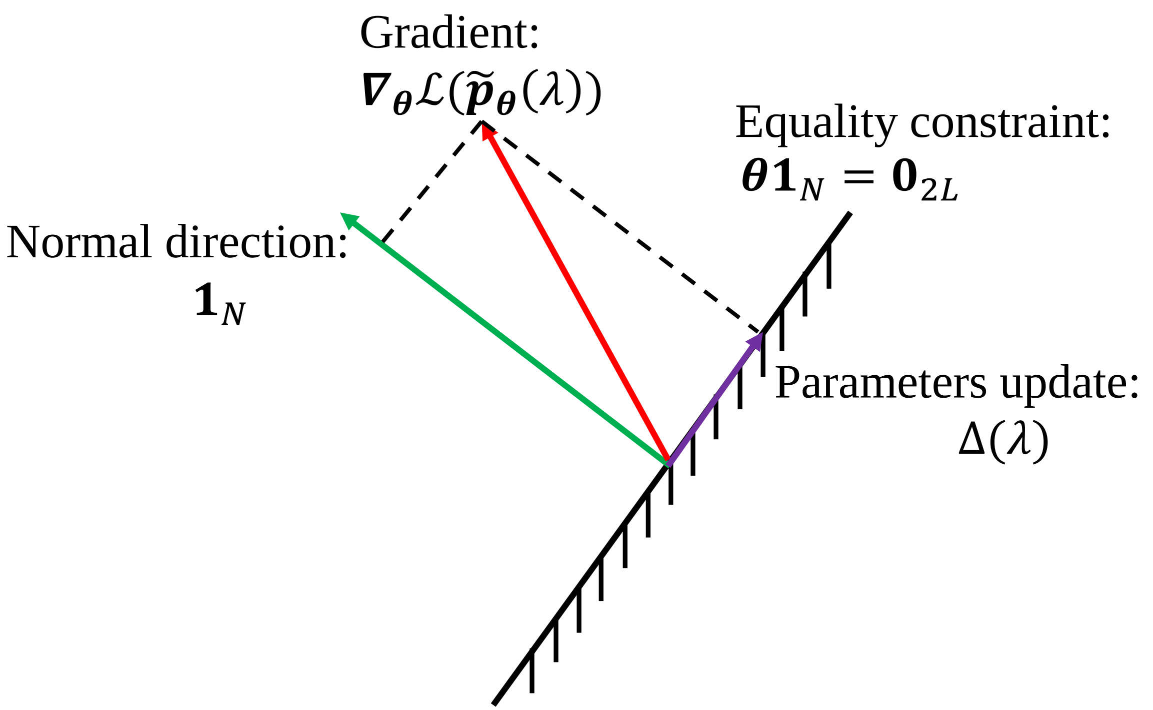


Diagram of gradient projection

OPC: Optimizing Connectivity between Minima

EOPC: Ensembling with OPC

- The optimized path provides infinite low-loss solutions on both sides of the switching point (SP), i.e., $\boldsymbol{p_\theta}(\lambda^*)$. To further improve performance on learned tasks, we propose EOPC to ensemble points within a local bent cylinder around SP.

- The cylinder is constructed according to the tangent of the path, let S be the set of points within this cylinder and can be formulated as follows,

$$S = \{\boldsymbol{w} | (\boldsymbol{w} - \boldsymbol{p_\theta}(\lambda))^{\mathrm{T}} \boldsymbol{p}'_{\boldsymbol{\theta}}(\lambda) = 0, \|\boldsymbol{w} - \boldsymbol{p_\theta}(\lambda)\|_2 \leq r; \\ \lambda \in [\lambda^* - \tau/2, \lambda^* + \tau/2]\},$$

- We adopt ensembling in parameter space by averaging points within S, i.e.,

$$\overline{\boldsymbol{w}} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{w}_i, \quad \boldsymbol{w}_i \sim S$$

where M is the number of total sampling points.

- We take $\overline{\boldsymbol{w}}$ as the minimum of the current task and the initial parameters of model in the next task.

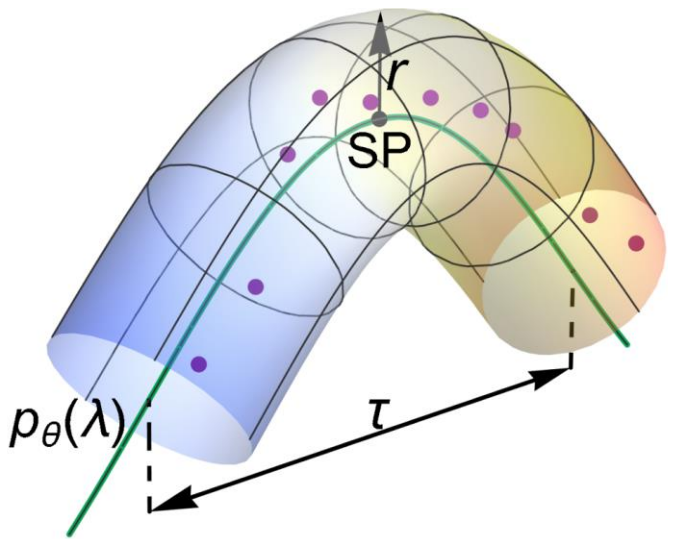
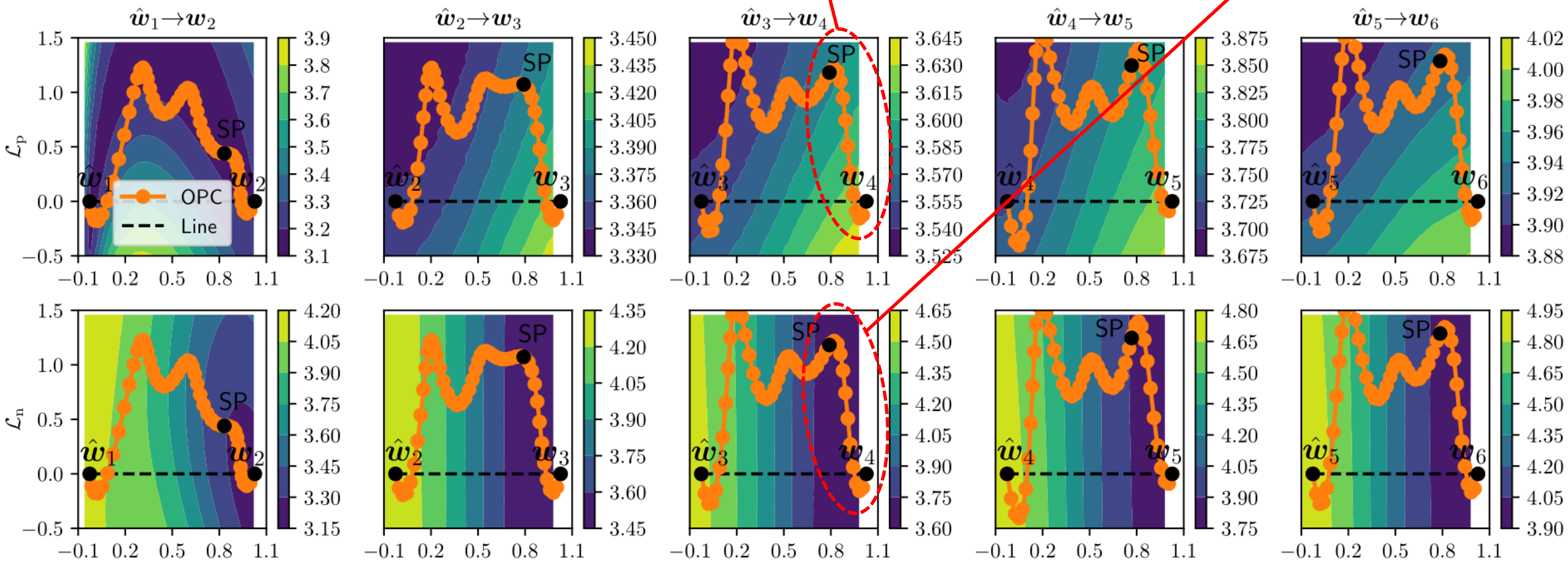


Diagram of local bent cylinder

Experiments

Compared with $\boldsymbol{w}_4$, SP locates in a lower-loss region in landscape of previous tasks.

Compared with $\boldsymbol{w}_4$, SP locates in the same low-loss region in landscape of the new task.



Visualization of paths found by OPC in loss landscape of previous tasks ($\mathcal{L}_p$) and the new task ($\mathcal{L}_n$).

Experiments

Method	CIFAR-100			ImageNet-100			ImageNet-1K		
$\mathcal{A}$ (%) $\uparrow$	5	10	25	5	10	25	5	10	25
iCaRL (Rebuffi et al., 2017)	57.83	52.63	49.02	64.75	58.80	52.46	51.60	47.42	41.03
BiC† (Wu et al., 2019)	59.36	54.20	50.00	70.07	64.96	57.73	62.65	58.72	53.47
LUCIR (Hou et al., 2019)	63.62	60.95	57.79	71.93	69.43	63.51	66.13	61.63	54.05
Mnemonics† (Liu et al., 2020)	63.34	62.28	60.96	72.58	71.37	69.74	64.63	63.01	61.00
GeoDL† (Simon et al., 2021)	65.14	65.03	**63.12**	73.87	73.55	71.72	65.23	64.46	62.20
AFC (Kang et al., 2022)	65.87	64.45	62.05	77.27	**75.47**	**72.41**	69.07	66.85	**63.40**
PODNet (Douillard et al., 2020)	65.47	63.13	59.85	76.32	73.54	63.05	68.33	65.35	58.62
w/ EOPC	$66.68_{+1.21}$	$64.94_{+1.81}$	$62.36_{+2.51}$	$77.12_{+0.8}$	$74.53_{+0.99}$	$68.18_{+5.13}$	**69.72**$_{+1.39}$	**67.57**$_{+2.22}$	$62.35_{+3.73}$
AANet (Liu et al., 2021)	66.53	64.63	61.05	77.98	74.70	68.65	68.87	65.65	60.07
w/ EOPC	**67.55**$_{+1.02}$	**65.54**$_{+0.91}$	$61.82_{+0.77}$	**78.95**$_{+0.97}$	$74.99_{+0.29}$	$70.10_{+1.45}$	$69.47_{+0.6}$	$67.35_{+1.7}$	$62.20_{+2.13}$

The adaptation results of EOPC and comparison results with existing incremental learning methods on CIFAR-100, ImageNet-100, and ImageNet-1K.

Thank you!