

Fair Class-Incremental Learning using Sample Weighting

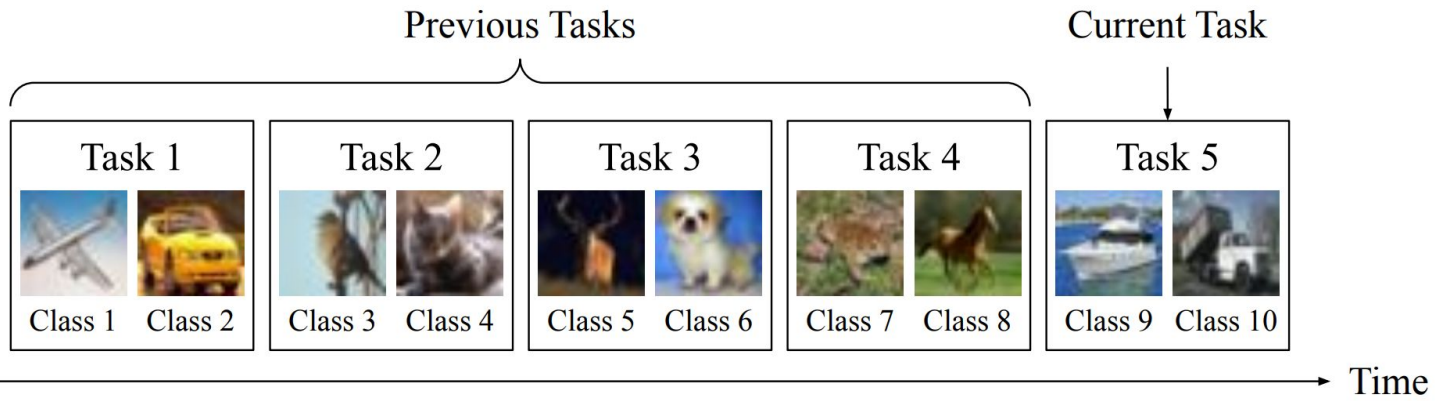
Jaeyoung Park*, Minsu Kim*, and Steven Euijong Whang

Data Intelligence Lab, KAIST

KDD 2026

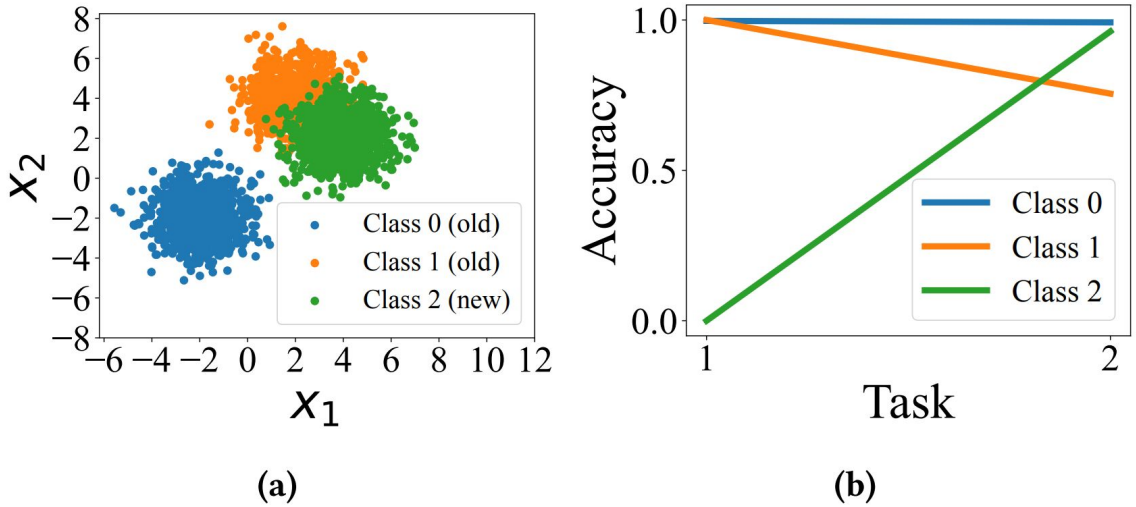
Class-Incremental Learning

- Setting: Model incrementally learns new classes of data over tasks
- Problem: Model forgets knowledge of previously-learned classes while learning new classes (i.e., catastrophic forgetting)
- Solution: Store a small portion of buffer data for previous tasks and use them jointly with the current task data for training (i.e., data replay)



Unfair Catastrophic Forgetting

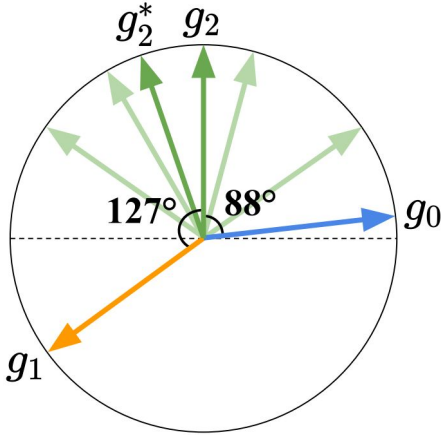
- Using all the samples of the current task for training results in unfair catastrophic forgetting for certain sensitive groups including classes
- Occurrence of unfair catastrophic forgetting on a synthetic dataset



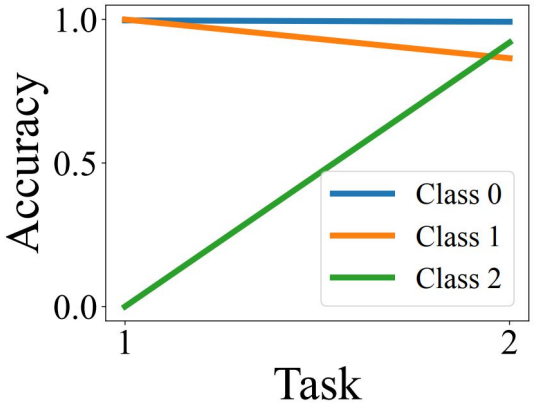
(a) (b)

Theoretical Analysis

- Catastrophic forgetting occurs when the gradient vectors of different tasks point in opposite directions
- If a current task sample's gradient opposes only the underperforming group's average gradient, training on the sample amplifies unfairness between groups



(c)



(d)

Fairness-aware Sample Weighting (FSW)

- Optimize **sample weights** to learn new classes without unfair catastrophic forgetting

$$\min_{\mathbf{w}_l} \quad L_{\text{fair}} + \lambda L_{\text{acc}},$$

λ: fairness-accuracy trade-off parameter

$$\text{where} \quad L_{\text{fair}} = \frac{1}{|\mathbb{Y}|\,|\mathbb{Z}|} \sum_{y\in\mathbb{Y}} \sum_{z\in\mathbb{Z}} \left| \tilde{\ell}(f_\theta, G_{y,z}) - \tilde{\ell}(f_\theta, G_y) \right|,$$

L_{fair}: depends on targeted group fairness metric

$$L_{\text{acc}} = \frac{1}{|\mathbb{Y}_c|\,|\mathbb{Z}|} \sum_{y\in\mathbb{Y}_c} \sum_{z\in\mathbb{Z}} \tilde{\ell}(f_\theta, G_{y,z}),$$

$$\tilde{\ell}(f_\theta, G_{y,z}) = \ell(f_\theta^{l-1}, G_{y,z}) - \eta\, \nabla_\theta \ell(f_\theta^{l-1}, G_{y,z})^\top \left(\frac{1}{|T_l|} \sum_{d_i \in T_l} \mathbf{w}_l^i\, \nabla_\theta \ell(f_\theta^{l-1}, d_i) \right).$$

*Linear in **w** → solvable by Linear Programming*

Algorithm 1 Fair Class-Incremental Learning

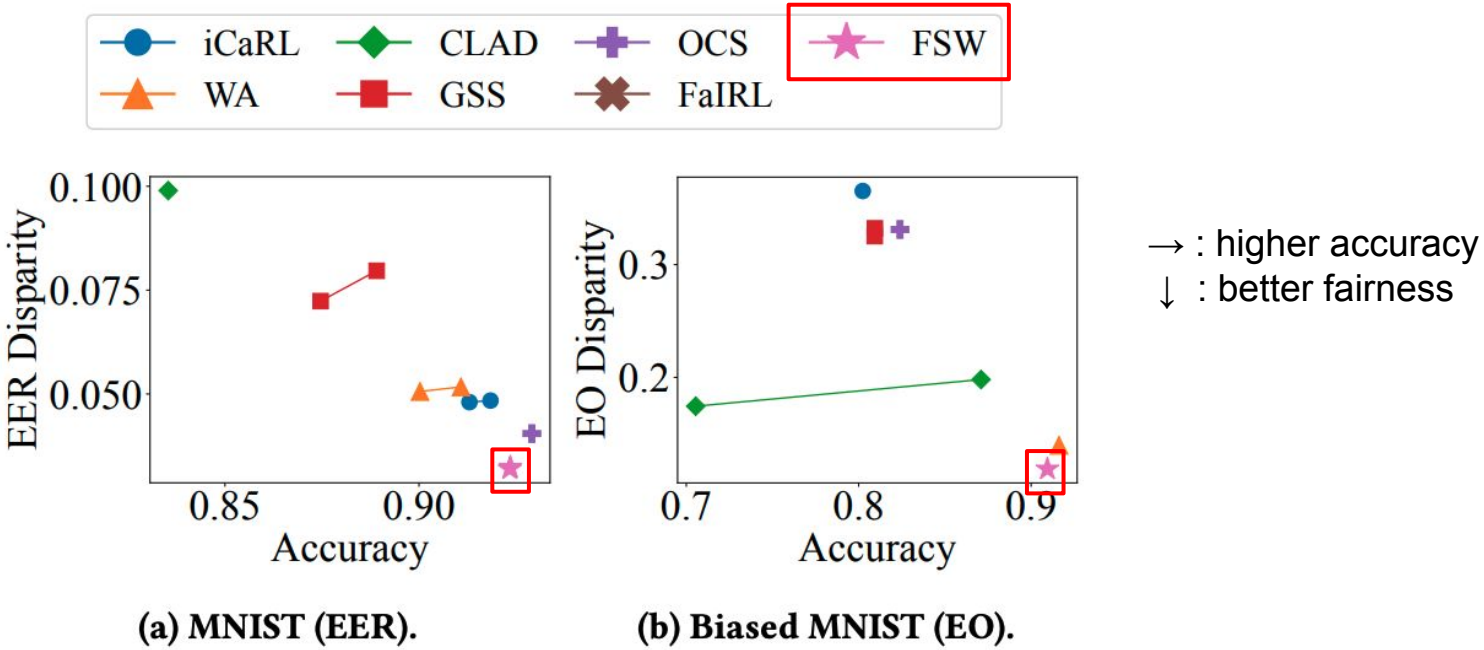
Input: Current task data T_l, previous buffer data $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{l-1}\}$, previous model f_θ^{l-1}, loss function ℓ, learning rate η, hyperparameters $\{\alpha, \lambda, \tau\}$, and fairness measure F.

1: **for** each epoch **do**
2: $\quad \mathbf{w}_l^* \leftarrow \textit{FSW}(T_l, \mathcal{M}, f_\theta^{l-1}, \ell, \alpha, \lambda, F)$
3: $\quad g_{\text{curr}} \leftarrow \frac{1}{|T_l|} \sum_{d_i \in T_l} \mathbf{w}_l^{*i}\, \nabla_\theta\, \ell(f_\theta^{l-1}, d_i)$
4: $\quad g_{\text{prev}} \leftarrow \nabla_\theta\, \ell(f_\theta^{l-1}, \mathcal{M})$
5: $\quad \theta \leftarrow \theta - \eta\, (g_{\text{curr}} + \tau\, g_{\text{prev}})$
6: $\mathcal{M}_l \leftarrow \textit{Buffer Sample Selection}(T_l)$
7: $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{M}_l$

Optimization → Sample Weights → Gradient Update

Accuracy and Fairness Results

- FSW achieves **a better fairness-accuracy trade-off** than baselines



(a) MNIST (EER). **(b) Biased MNIST (EO).**

Conclusion

- Naïvely training on new tasks leads to **unfair catastrophic forgetting**
- This unfairness arises when average gradients conflict across groups
- FSW mitigates unfair forgetting via **fairness-aware sample weighting**
- Achieves a **better fairness-accuracy trade-off** across diverse datasets

KAIST