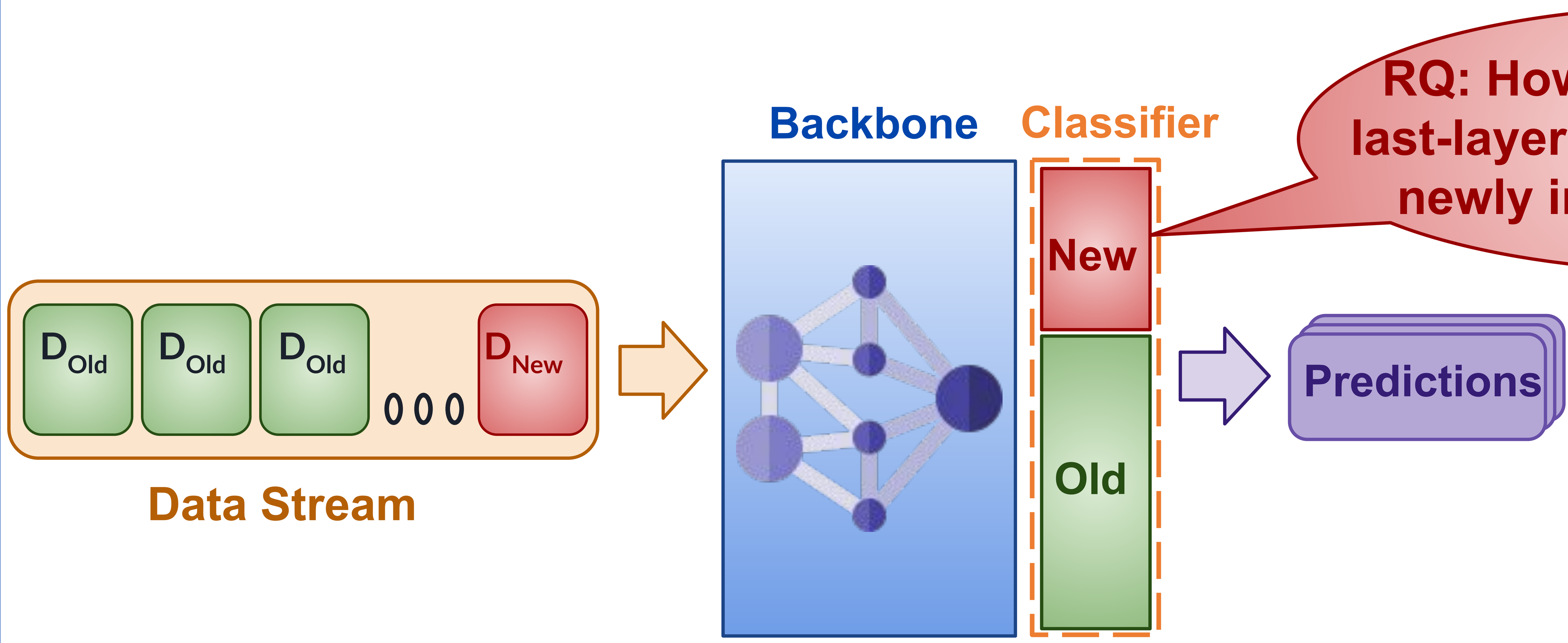
A Good Start Matters: Enhancing Continual Learning with Data-Driven Weight Initialization

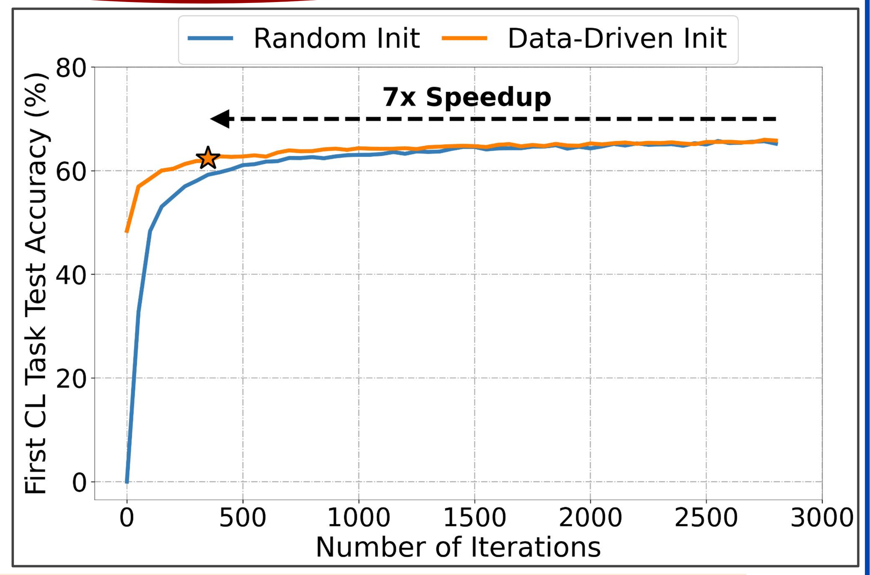
Md Yousuf Harun[1], Christopher Kanan[2]

[1]Rochester Institute of Technology, [2]University of Rochester

Motivation

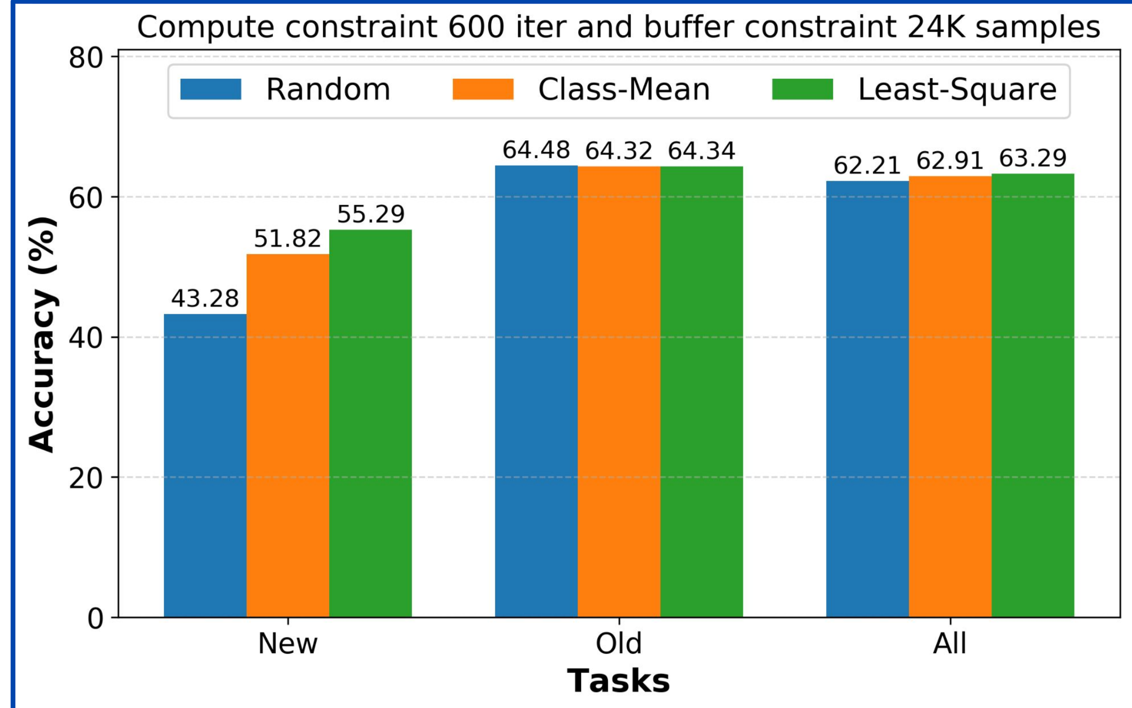
- **Problem:** In continual learning (CL), classifier weights for newly encountered classes are typically initialized *randomly*, leading to high initial training loss (spikes) and instability.
- Consequently, achieving optimal convergence requires prolonged training, increasing computational costs.

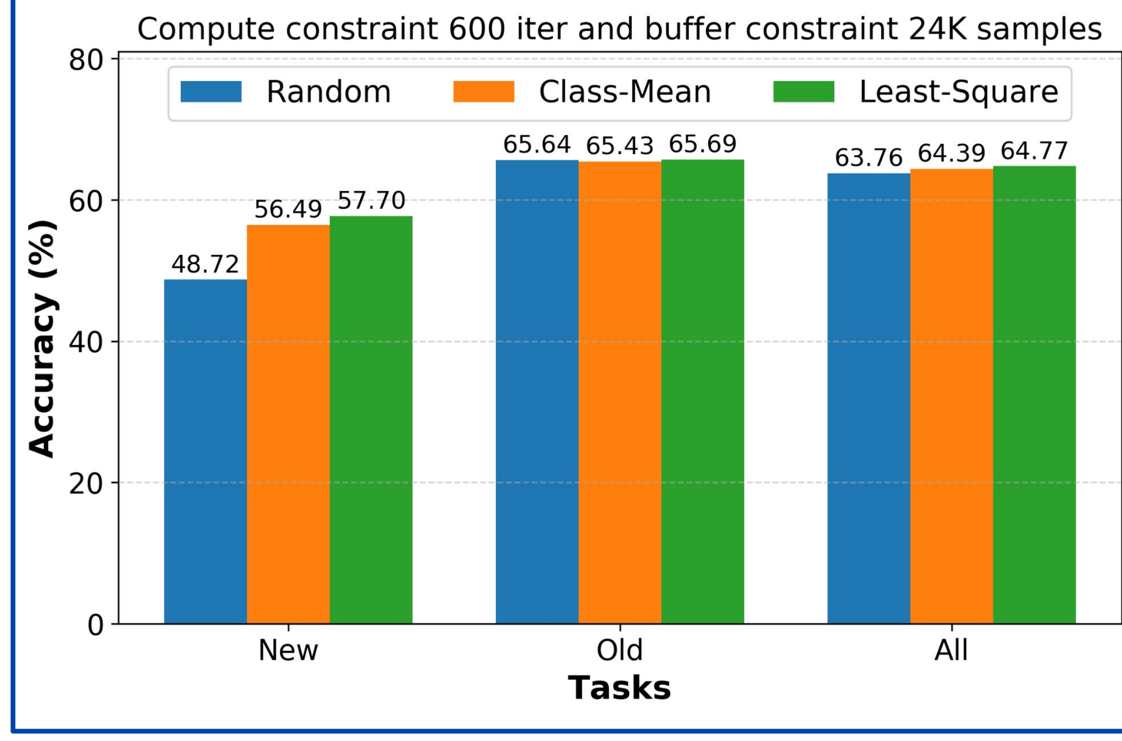




TL;DR: Inspired by **Neural Collapse**, we propose a **Least-Square-based weight initialization** method that optimally aligns classifier weights for newly introduces categories with their feature distribution.

Results





Goal: investigating how weight initialization impacts

- CL without changing pre-trained representations where we train the last-layer classifier while keeping the backbone frozen (top fig)
- continual representation learning where we selectively update the DNN backbone using LoRA (bottom fig)

Observation: LS initialization enhances CL performance and enables efficient adaptation to new tasks in both settings

Data-Driven Weight Initialization

- In DNNs trained with Mean-Squared-Error (MSE) loss, neural collapse gives rise to a Least-Square (LS) classifier in the last layer, whose weights can be analytically derived from learned features
- We leverage this LS formulation to initialize classifier weights in a data-driven manner, aligning them with the feature distribution

MSE loss:

$$L(\mathbf{W}) = \frac{1}{2N}\|\mathbf{WZ}-\mathbf{Y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2.$$

Differentiation:

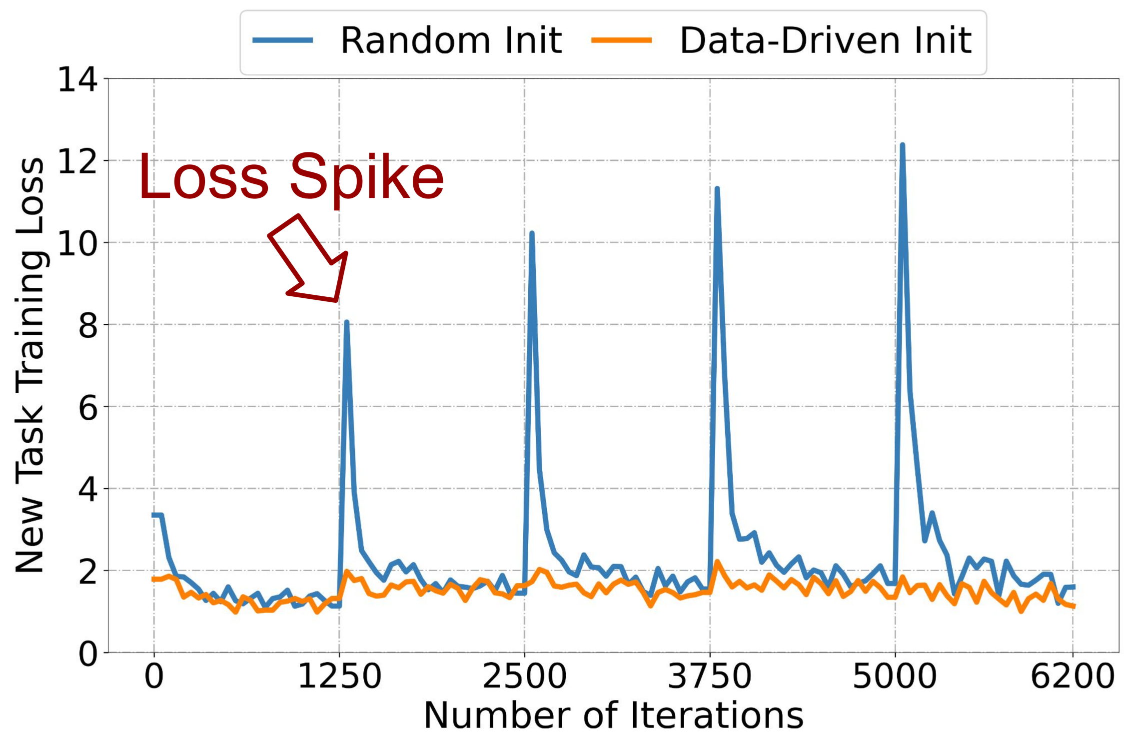
$$\frac{\partial L}{\partial \mathbf{W}} = \frac{1}{N}(\mathbf{WZZ}^\top - \mathbf{YZ}^\top) + \lambda\mathbf{W} = 0.$$

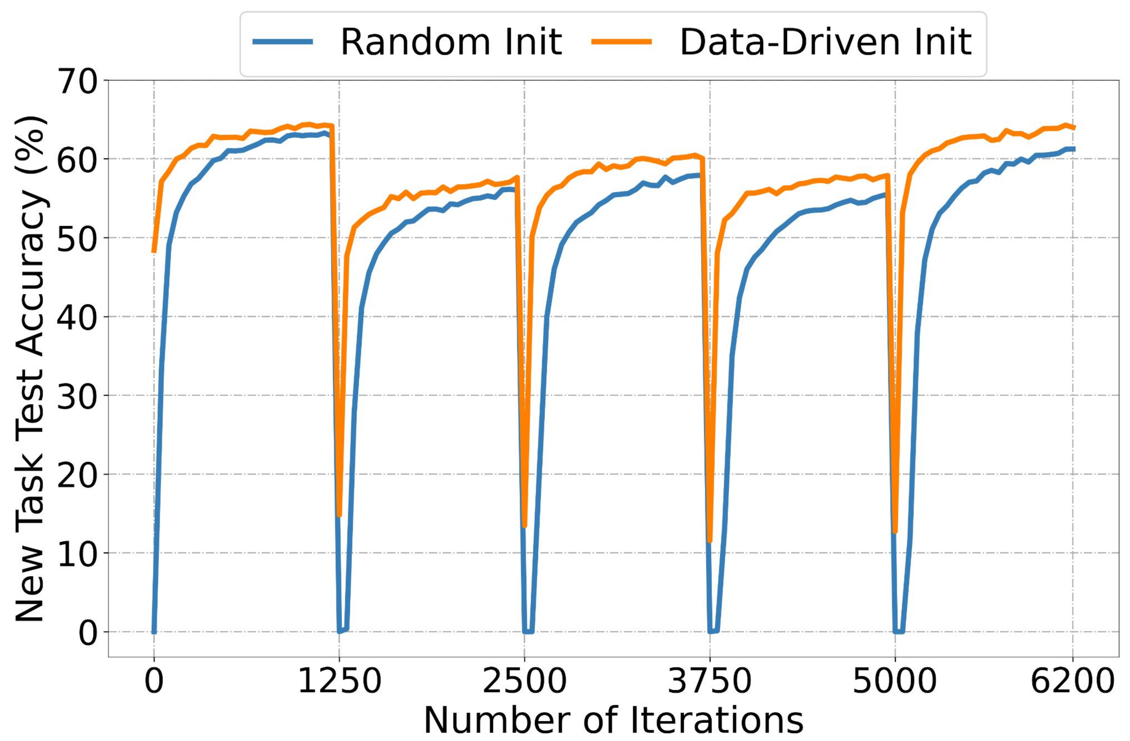
LS Weights:

$$\mathbf{W}_{LS} = \frac{1}{C}\mathbf{M}^\top(\mathbf{\Sigma}_T + \boldsymbol{\mu}_G\boldsymbol{\mu}_G^\top + \lambda\mathbf{I})^{-1}.$$

❑ LS weights *solely* depend on feature statistics and therefore can be analytically derived

Random Initialization Causes Loss Spike

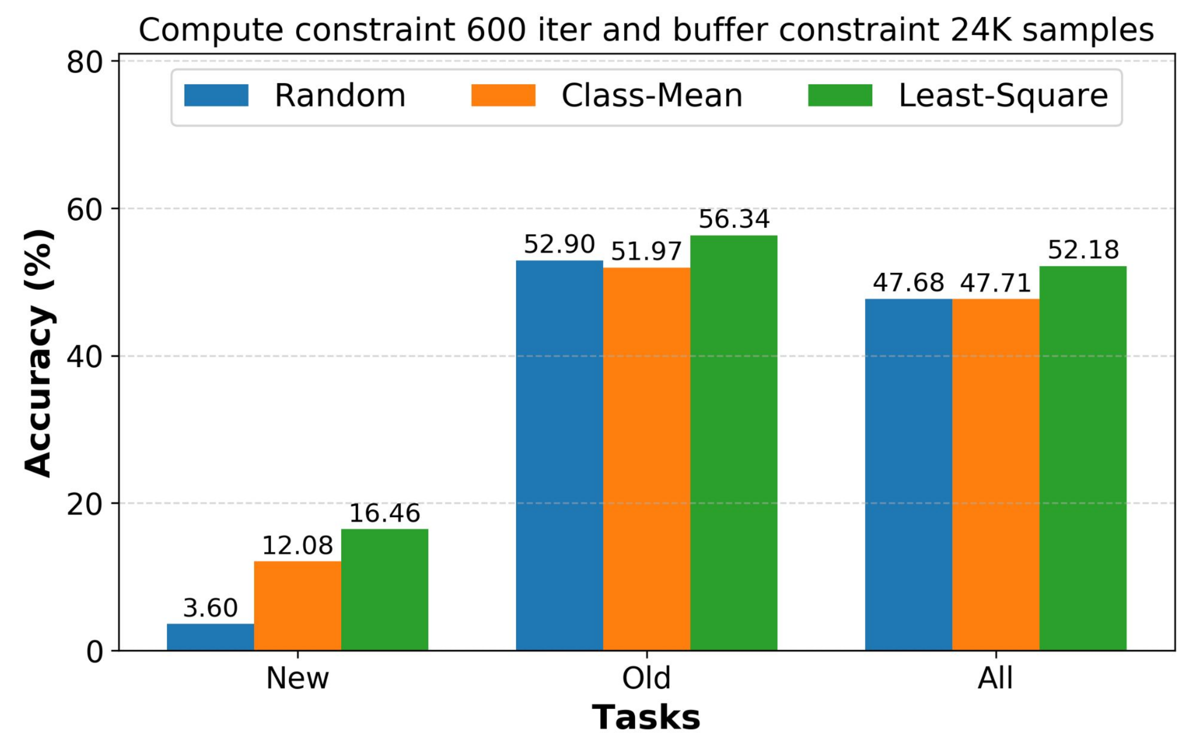




CL Setting: A ConvNeXt model pre-trained on ImageNet-1K incrementally learns 5 tasks (73 classes per task) from Places-365 dataset using rehearsal

Observations: Random initialization causes loss spikes whereas data-driven initialization prevents loss spikes and improves accuracy

Generality



ResNet18 Results

❖ **Goal:** investigating whether LS initialization remains effective when integrated with different CL methods and DNN architectures.

❖ **Observation:** LS enhances CL performance of experience replay, EWC, and DER++. It demonstrates efficacy for ConvNeXt and ResNet architectures.

Acknowledgments

We thank NSF for financially supporting this research