Progressively Overcoming Catastrophic Forgetting in Kolmogorov–Arnold Networks

Evgenii Ostanin
eostanin@torontomu.ca

Nebojsa Djosic
nebojsa.djosic@torontomu.ca

Fatima Hussain
fatima.hussain@torontomu.ca

Salah Sharieh
salah.sharieh@torontomu.ca

Alexander Ferworn
aferworn@torontomu.ca

Malek Sharieh
malek.sharieh@hts.on.ca

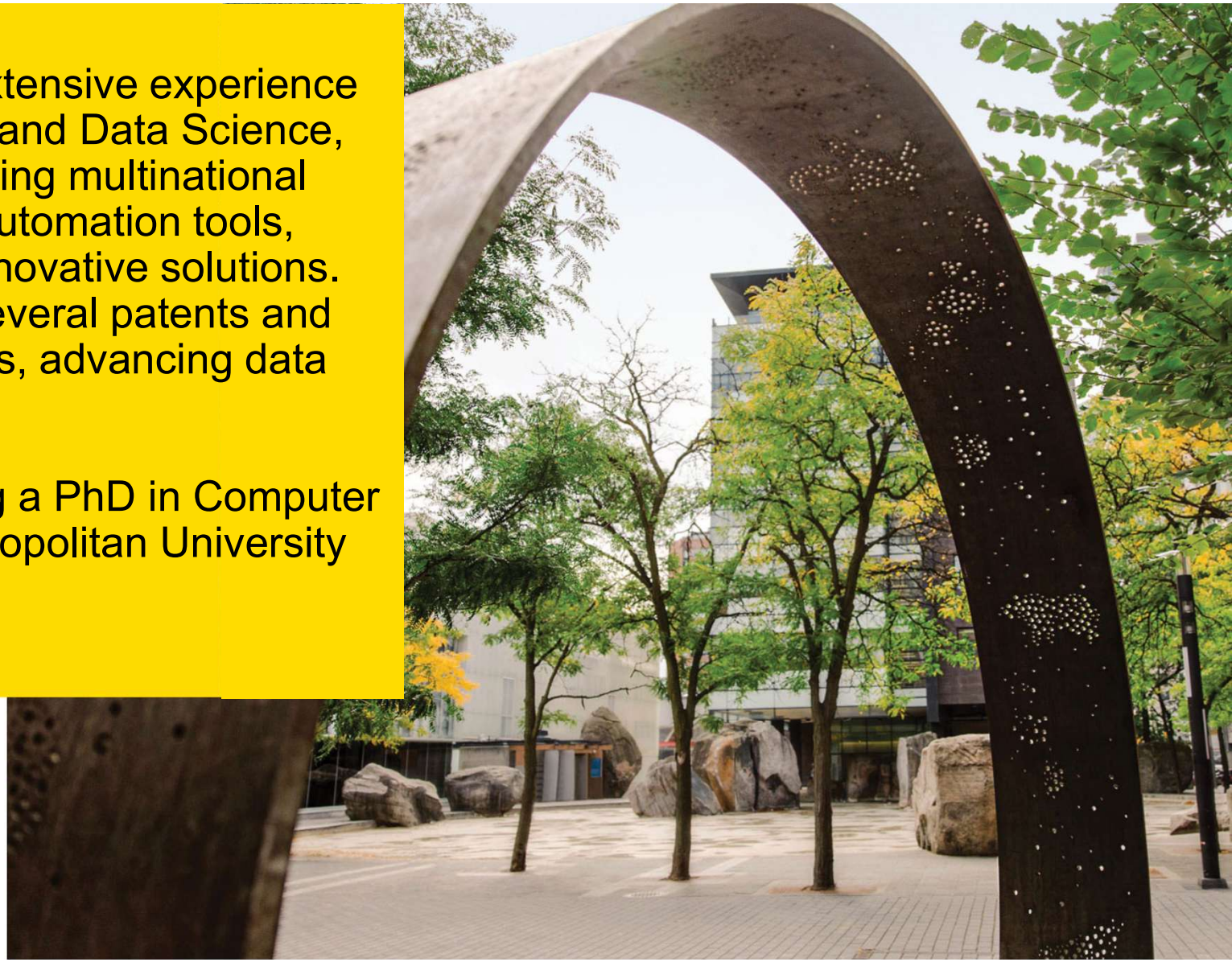
Toronto Metropolitan University, Toronto, Canada

Presenter
Evgenii Ostanin

IARIA Congress 2025, July 06, 2025 to July 10, 2025 - Venice, Italy

Evgenii Ostanin has extensive experience in Economics, Finance, and Data Science, having worked with leading multinational companies to develop automation tools, financial models, and innovative solutions. He has contributed to several patents and proof-of-concept projects, advancing data science and technology.

Currently, he is pursuing a PhD in Computer Science at Toronto Metropolitan University (TMU).



Toronto Metropolitan University

The authors are members of the Computer Science Department at the Toronto Metropolitan University (formerly Ryerson).

They are actively involved in research and applied projects centered on leveraging Artificial Intelligence and Machine Learning for automation in key domains including cybersecurity, governance, and public safety.

Toronto Metropolitan University

Presentation Outline

- Introduction and Problem Statement
- Kolmogorov-Arnold Networks (KANs) introduction
- Methodology and Architecture
- Experimental Results
- Conclusions and Future Work

Toronto Metropolitan University

Key Contributions

- A comparison of KANs and MLP in continual learning using the Split-MNIST benchmark.
- Testing experience replay (random sampling) and stratified (class-balanced) replay strategies for mitigating forgetting.
- Introduction and evaluation of two novel KAN-specific freezing techniques, targeting spline control points and entire spline tensors.

Toronto Metropolitan University

Introduction

- Traditional models like MLPs forget earlier tasks when trained sequentially (shared parameters overwritten during sequential training resulting in catastrophic forgetting)
- Kolmogorov–Arnold Networks (KANs) use adaptive spline activations that may help preserve prior knowledge
- KANs offer interpretable, modular transformations. Could they also improve retention in continual learning?

Problem Statement

- **Goal 1:** Evaluate KAN vs. MLP on continual learning with replay strategies
- **Goal 2:** Goal 2: Explore KAN-specific freezing techniques to improve memory retention
- **Key Questions:**
  - Do KANs retain prior knowledge better?
  - Does spline freezing help mitigate forgetting?

Toronto Metropolitan University

Kolmogorov-Arnold Network Architecture

What are KANs?

- Neural networks based on the *Kolmogorov-Arnold* representation theorem
- Use *spline-based* learnable activation functions instead of traditional fixed activations

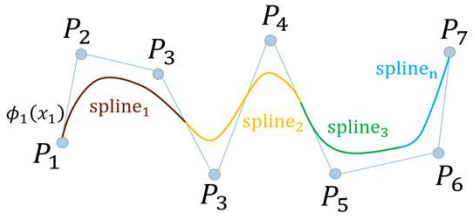
$$y = F(x_1,\ x_2,\ x_3,\ \cdots,x_n)$$

$$\phi_1(x_1) + \phi_2(x_2) + \phi_3(x_3) + \cdots + \phi_n(x_n)$$

univariate functions

Why KANs?

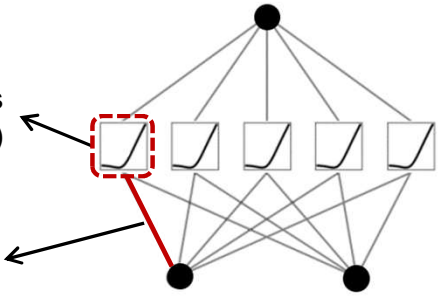
- Enhanced **flexibility** and **interpretability**
- Adaptability to complex, non-linear relationships in data

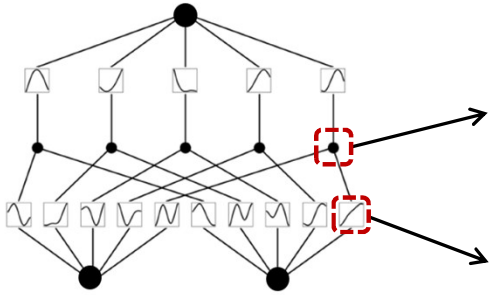


Traditional MLPs:

fixed activation functions on nodes (like ReLU)

learnable weights on edges





KANs:

sum operation on nodes

learnable activation functions on edges

Toronto Metropolitan University

Z. Liu et al., KAN: Kolmogorov-Arnold Networks, 2024.

Methodology

Models: MLP and KAN

Forgetting Mitigation Strategies:

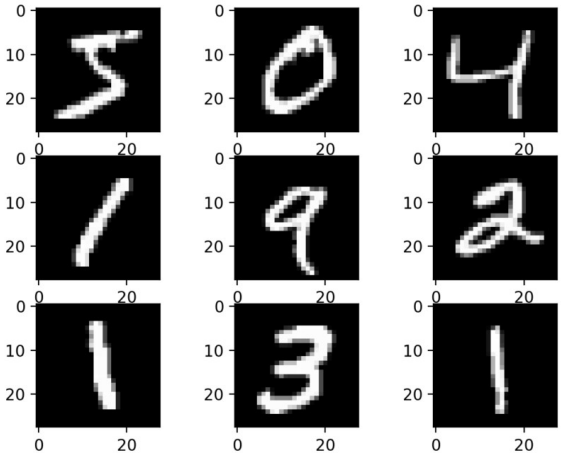
Replay and Stratified (balanced) replay

Spline Freezing Strategies:

Tensor-level and Point-level freezing

Dataset: MNIST (handwritten digits)

Evaluation Metrics: Accuracy, Forgetting



KAN vs. MLP Architectures

MLP

Input Layer: Takes the raw input data (28x28 pixel image, or 784 features.).

Hidden Layer 128-dimensional hidden layer applies fixed activation functions (e.g., ReLU), and every neuron is fully connected to neurons in adjacent layers.

Output Layer: Produces the final classification or prediction (10 - one neuron per class for MNIST digits).

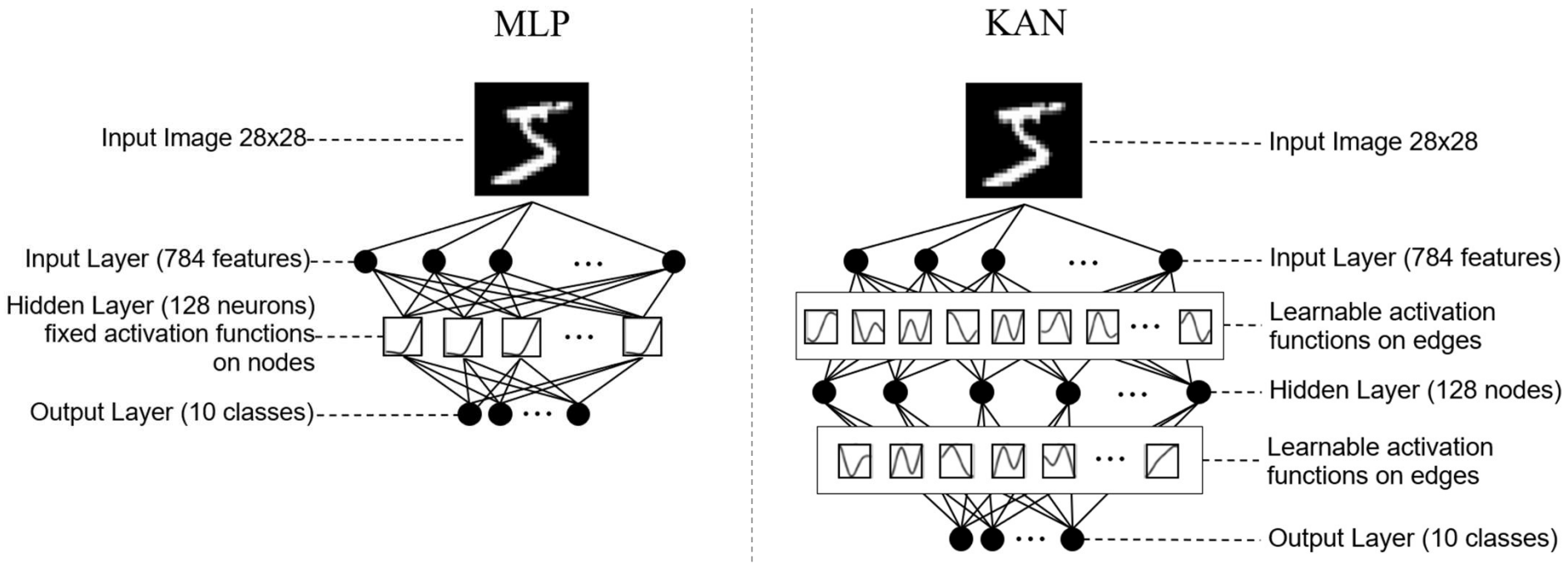
KAN

Input Layer: Similar to MLP, takes raw input data.

KANLinear Layers: Replaces standard hidden layers with layers using learnable spline-based activation functions, allowing flexibility in how activations adapt during training.

Output Layer: Similar to MLP, outputs predictions (e.g., digits 0–9 for classification).

KAN vs. MLP Architectures



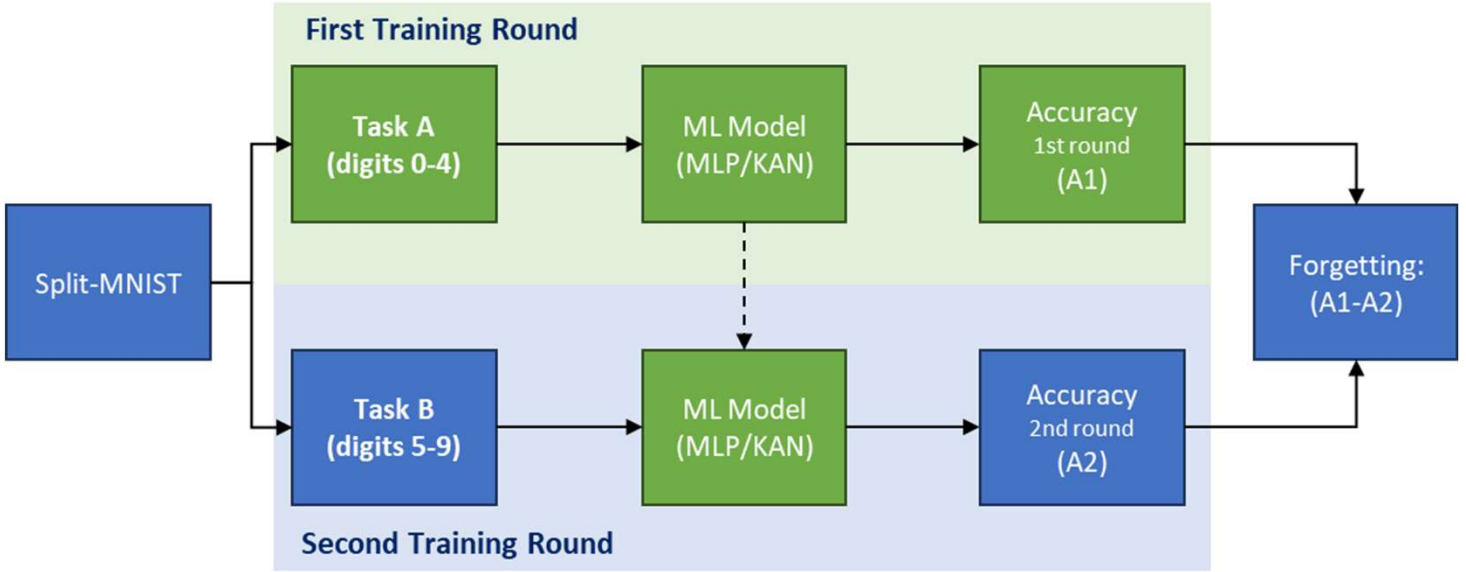
MLP: Fixed weights and ReLU activations.

KAN: Spline-based activations on edges.

Toronto Metropolitan University

Continual Learning Setup

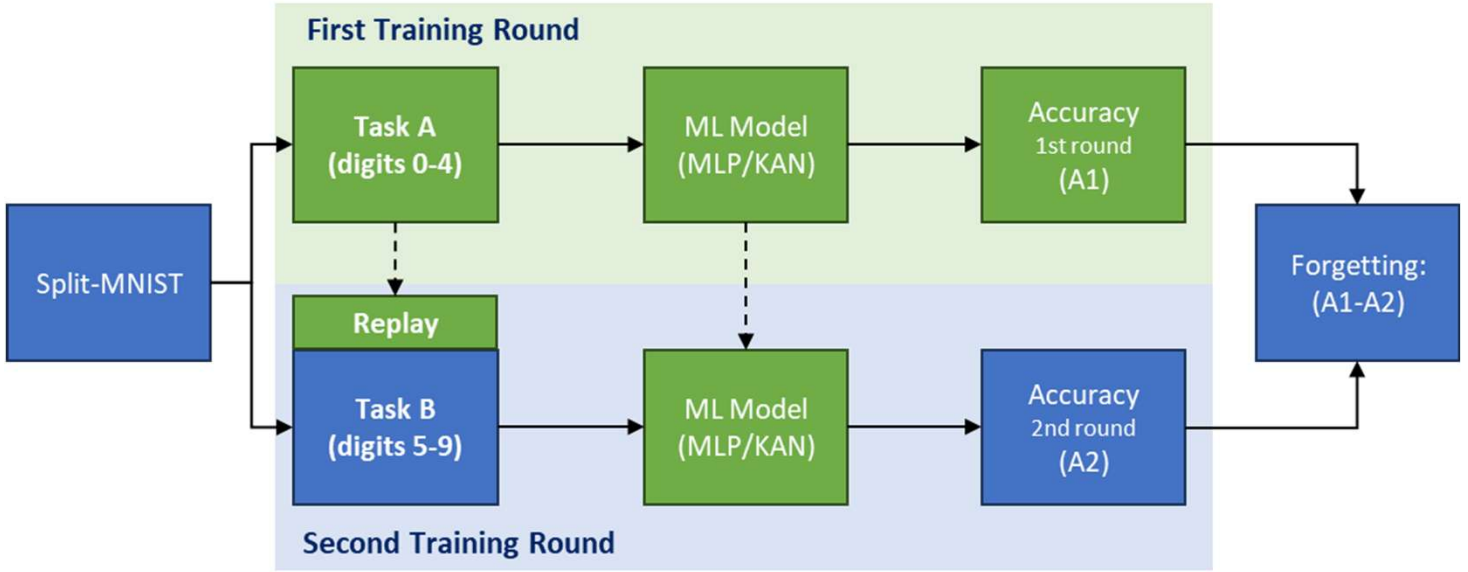
- To simulate continual learning under the Split-MNIST protocol, we split the dataset into Task A and Task B
- The model is first trained on Task A, then trained on Task B
- Forgetting is quantified as the drop in Task A accuracy after Task B training
  (how well the model can predict labels from test set A after two consecutive training rounds)



Toronto Metropolitan University

Replay and Stratified (balanced) Replay

- To mitigate forgetting, we implement two forms of experience replay, preserving a subset of Task A examples and mixing them into the Task B training set
- In replay scenario Task A examples are sampled randomly, In the balanced replay scenario, examples are sampled in a stratified fashion
- We tested replay buffer sizes of 50, 100, and 500, where the buffer size denotes the number of retained data samples

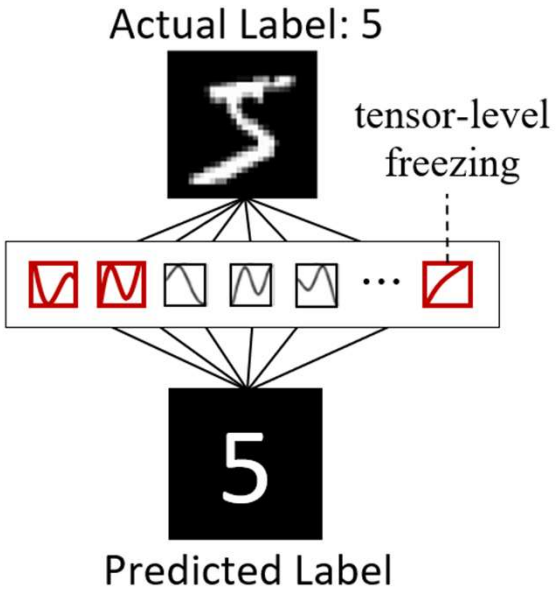


Toronto Metropolitan University

Spline Freezing Strategies

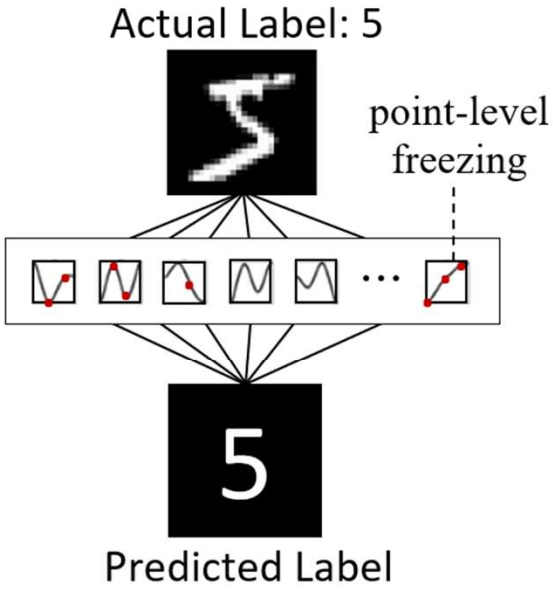
Tensor-level freezing

In this strategy, we compute a score for each spline row and freeze the top k% rows in the KAN architecture



Point-level freezing

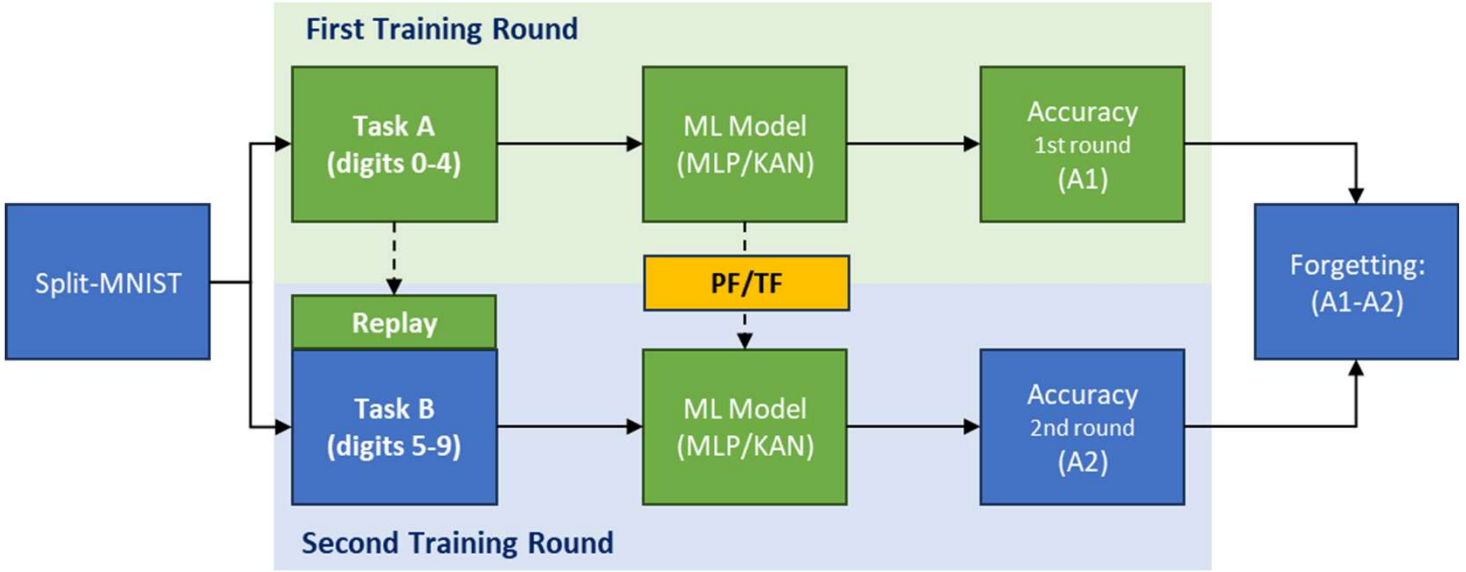
The top k% of individual control points in the spline weight matrix are frozen, regardless of their row or neuron association.



Toronto Metropolitan University

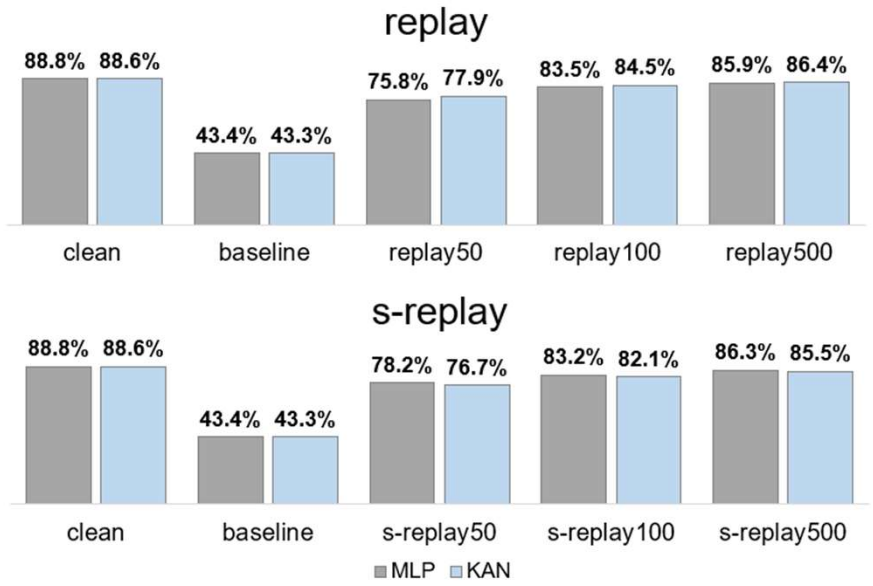
Spline Freezing Strategies

- For both strategies, we test k ∈ {0.05,0.1,0.25,0.5,0.75}, spanning from minimal to aggressive freezing
- Three scoring methods: **weight** - mean absolute value of the weights in each row; **grad** - mean absolute gradient magnitude per row; **weight grad** - a combination of both weight and grad
- Freezing is applied **after Task A** and **f**rozen parameters are excluded from optimization updates by masking their gradients before applying the optimizer steps



Results - MLP and KAN

- The clean setting refers to training on all MNIST classes simultaneously, serving as an upper-bound reference
- Both models achieve high accuracy on the full MNIST (88.8% and 88.6%, respectively)
- Both models but suffer from severe catastrophic forgetting when trained sequentially on separated tasks – baseline (Task A accuracy drops by nearly 96%, resulting in an overall accuracy of just 43.4% for MLP and 43.3% for KAN)
- Both replay and s-replay improve accuracy and retention. (with replay buffer size 100, MLP and KAN reach 83.5% and 84.5% accuracy, respectively, while s-replay achieves 83.2% for MLP and 82.1% for KAN)
- We selected buffer size of 100 for subsequent experiments, as it maintained measurable room for improvement while ensuring sufficient retention to validate the impact of freezing methods.

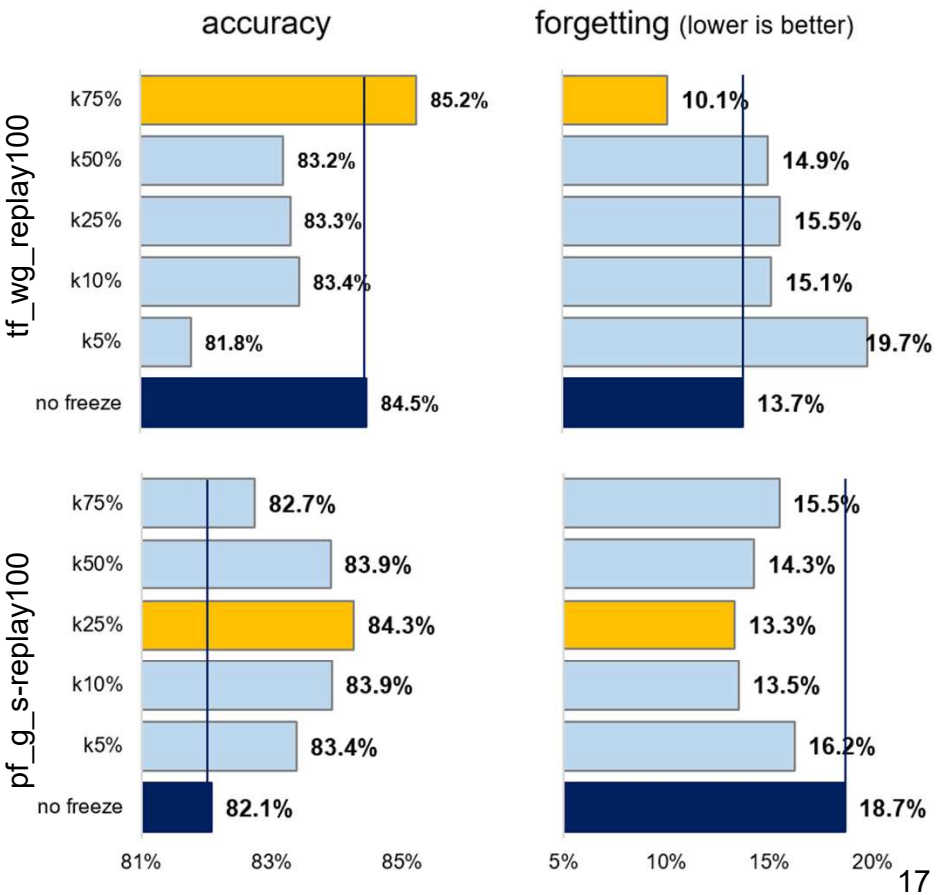


Scenario	MLP Acc.	KAN Acc.	MLP Forget.	KAN Forget.
Clean	0.888	0.886	-	-
Baseline	0.434	0.433	0.958	0.964
Replay 50	0.758	0.779	0.322	0.275
Replay 100	0.835	0.845	0.149	0.137
Replay 500	0.859	0.864	0.075	0.071
s-Replay 50	0.782	0.767	0.261	0.305
s-Replay 100	0.832	0.821	0.147	0.187
s-Replay 500	0.863	0.855	0.068	0.077

Toronto Metropolitan University

Results – Tensor(TF) and Point(PF) Freezing

- Both TF and PF improve continual learning when paired with replay, however, their effectiveness depends on the configuration.
- PF offers consistent gains, especially under s-replay. Most k values outperform the no-freeze baseline, with the best configuration pf_g_s-replay100 at k = 25%
- TF shows greater variability but also higher potential. The best configuration tf_wg_replay100 at k = 75% yielded the top accuracy overall



Results – Accuracy and Forgetting

- Tables indicate the top-k% of tensor rows or control points using heuristics based on weights (w), gradients (g), or a weighted average (wg)
- PF under s-replay remained effective across multiple k values, with most configurations improving over the no-freeze baseline.
- The top-performing setups for TF confirm that tensor-freezing can outperform point-freezing in certain cases when appropriately tuned
- The broader range of outcomes highlights that TF is more sensitive to the choice of k and scoring strategy, reinforcing the need for careful calibration.

Accuracy

Method	no freeze	k5%	k10%	k25%	k50%	k75%
pf_w_replay100	0.845	0.817	0.841	0.830	0.837	0.816
pf_g_replay100	0.845	0.833	0.835	0.822	0.845	0.824
pf_wg_replay100	0.845	0.838	0.832	0.842	0.824	0.844
pf_w_s-replay100	0.821	0.816	0.828	0.831	0.833	0.829
pf_g_s-replay100	0.821	0.834	0.839	0.843	0.839	0.827
pf_wg_s-replay100	0.821	0.825	0.838	0.840	0.838	0.829
tf_w_replay100	0.845	0.851	0.813	0.834	0.844	0.821
tf_g_replay100	0.845	0.840	0.846	0.830	0.834	0.843
tf_wg_replay100	0.845	0.818	0.834	0.833	0.832	0.852
tf_w_s-replay100	0.821	0.837	0.826	0.826	0.835	0.829
tf_g_s-replay100	0.821	0.831	0.832	0.834	0.842	0.843
tf_wg_s-replay100	0.821	0.851	0.841	0.841	0.841	0.837

Forgetting

Method	no freeze	k5%	k10%	k25%	k50%	k75%
pf_w_replay100	0.137	0.185	0.133	0.154	0.137	0.178
pf_g_replay100	0.137	0.166	0.153	0.178	0.123	0.157
pf_wg_replay100	0.137	0.141	0.144	0.133	0.166	0.112
pf_w_s-replay100	0.187	0.182	0.166	0.160	0.155	0.157
pf_g_s-replay100	0.187	0.162	0.135	0.133	0.143	0.155
pf_wg_s-replay100	0.187	0.173	0.135	0.148	0.130	0.152
tf_w_replay100	0.137	0.116	0.200	0.153	0.133	0.163
tf_g_replay100	0.137	0.121	0.133	0.165	0.158	0.116
tf_wg_replay100	0.137	0.197	0.151	0.155	0.149	0.101
tf_w_s-replay100	0.187	0.141	0.168	0.174	0.130	0.152
tf_g_s-replay100	0.187	0.143	0.164	0.152	0.137	0.127
tf_wg_s-replay100	0.187	0.123	0.134	0.132	0.137	0.136

Toronto Metropolitan University

Conclusions

- This paper investigated KANs in continual learning
- Both tensor-level and point-level spline freezing consistently improve retention in Split-MNIST when paired with simple replay, while the absolute improvements are moderate (up to +2.2 % overall accuracy and a 5.4 % reduction in forgetting)
- These KAN-specific freezing strategies leverage the spline structure to preserve prior task knowledge, opening a promising direction for more targeted retention strategies.

Toronto Metropolitan University

Future Work

- Exploring freezing in deeper KANs
- Integration with regularization and dynamic expansion methods
- Testing on more complex benchmarks beyond MNIST
- Developing adaptive freezing and unfreezing strategies

With these enhancements, we expect to achieve higher retention and greater robustness in continual learning tasks, further unlocking the potential of KANs for long-term knowledge consolidation

Toronto Metropolitan University

IARIA Congress 2025, July 06, 2025 to July 10, 2025 - Venice, Italy

Progressively Overcoming Catastrophic Forgetting in Kolmogorov–Arnold Networks

Q & A

eostanin@torontomu.ca