Capture the Present Before Time Moves On

Reducing Catastrophic Forgetting via Contrastive Learning and Dataset Distillation

Nikola Vasić, Nevena Denić

Faculty of Science and Mathematics, University of Niš, Serbia

Reproducing 2 significant online class-incremental learning papers:

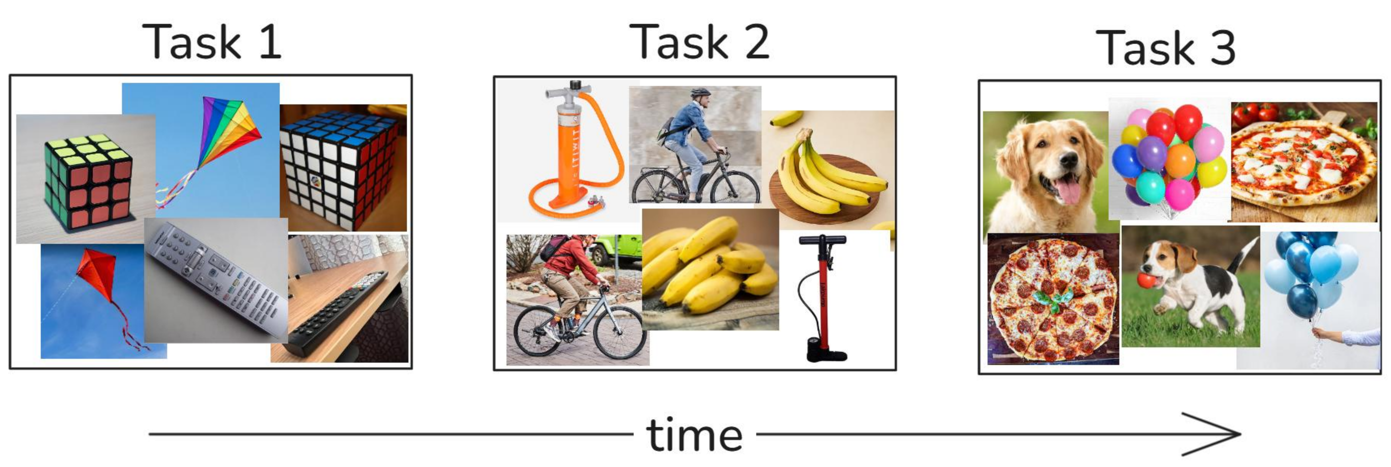
SCR - Supervised Contrastive Replay: Revisiting the Nearest Class Mean Classifier in Online Class-Incremental Continual Learning (CVPR 2021)

- Leverages **contrastive learning** in replay-based methods
- Outperforms all previous SOTA by a large margin

SSD - Summarizing Stream Data for Memory-Constrained Online Continual Learning (AAAI 2024)

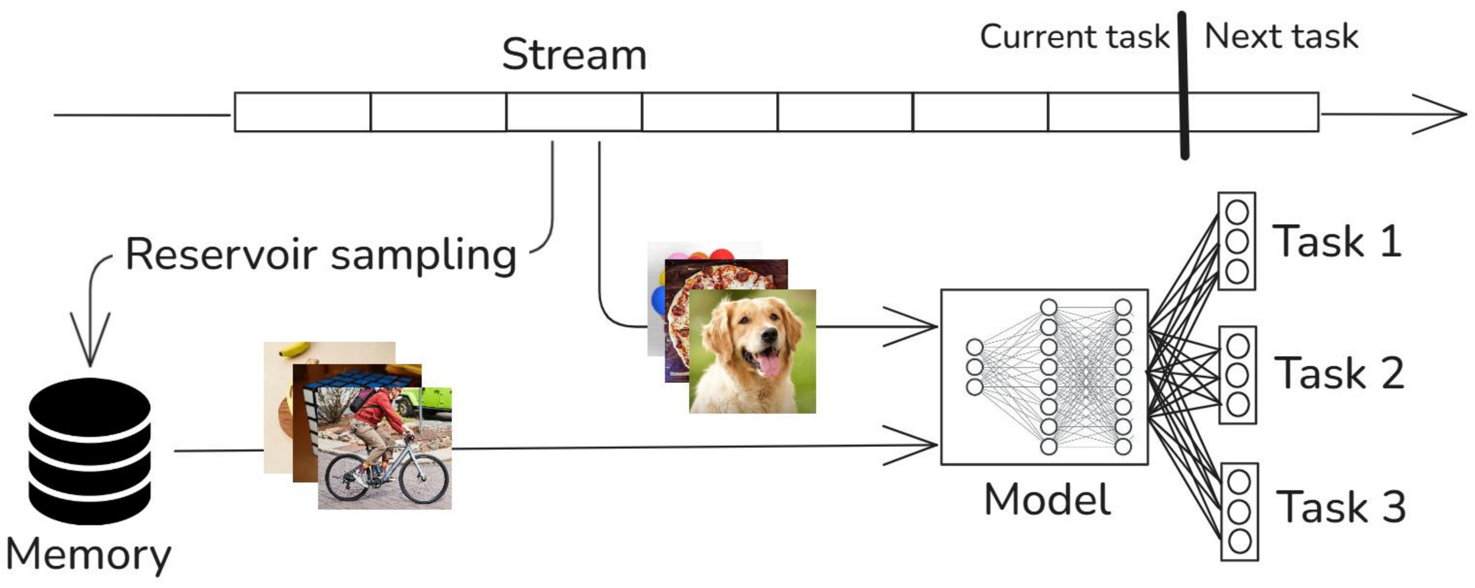
- Uses **dataset distillation** techniques to generate compact memory representations
- Provides significant accuracy boost under restricted memory size

ONLINE CLASS-INCREMENTAL LEARNING



Setup	One task is N classes	Task-ID available	Data comes from stream	Info about future tasks
Task-incremental	✅	✅	❌	❌
Class-incremental	✅	❌	❌	❌
Online class-incremental	✅	❌	✅	❌

While learning new tasks, models forget previous tasks – **catastrophic forgetting**.
Simple fix: store previous inputs and replay them later – **replay-based methods**.



ENHANCING THE MODEL (SCR)

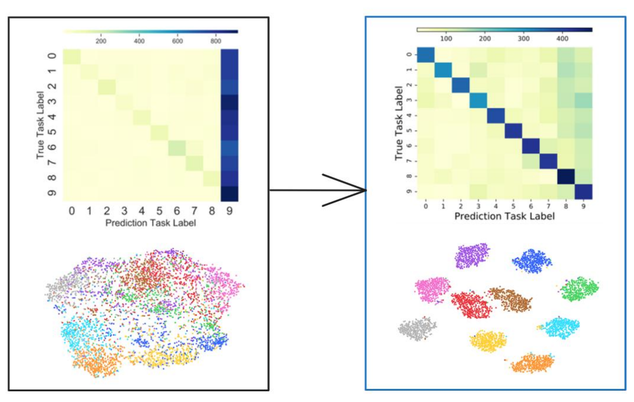
Softmax classifier and **cross-entropy loss** introduce problems:

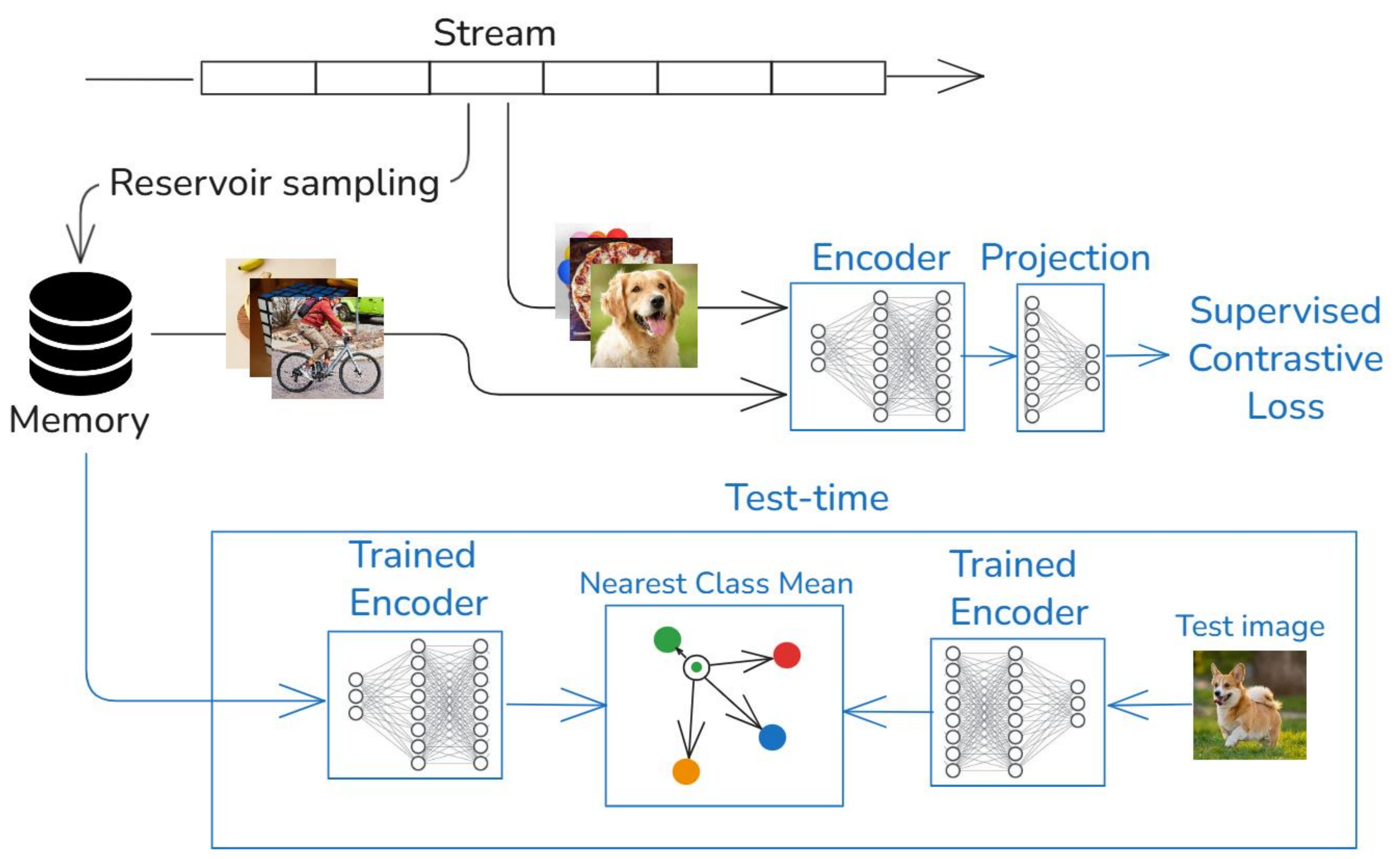
- A strong task-recency bias – most predictions are recently seen classes
- Architectural changes are needed when new classes arrive

Alternative: dismissing the classification layer and using the **Supervised Contrastive Loss** and the **Nearest Class Mean classifier:**

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\left(\frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_p/\tau\right)}{\sum_{a\in A(i)}\exp\left(\mathbf{z}_i^\top \mathbf{z}_a/\tau\right)}\right)$$

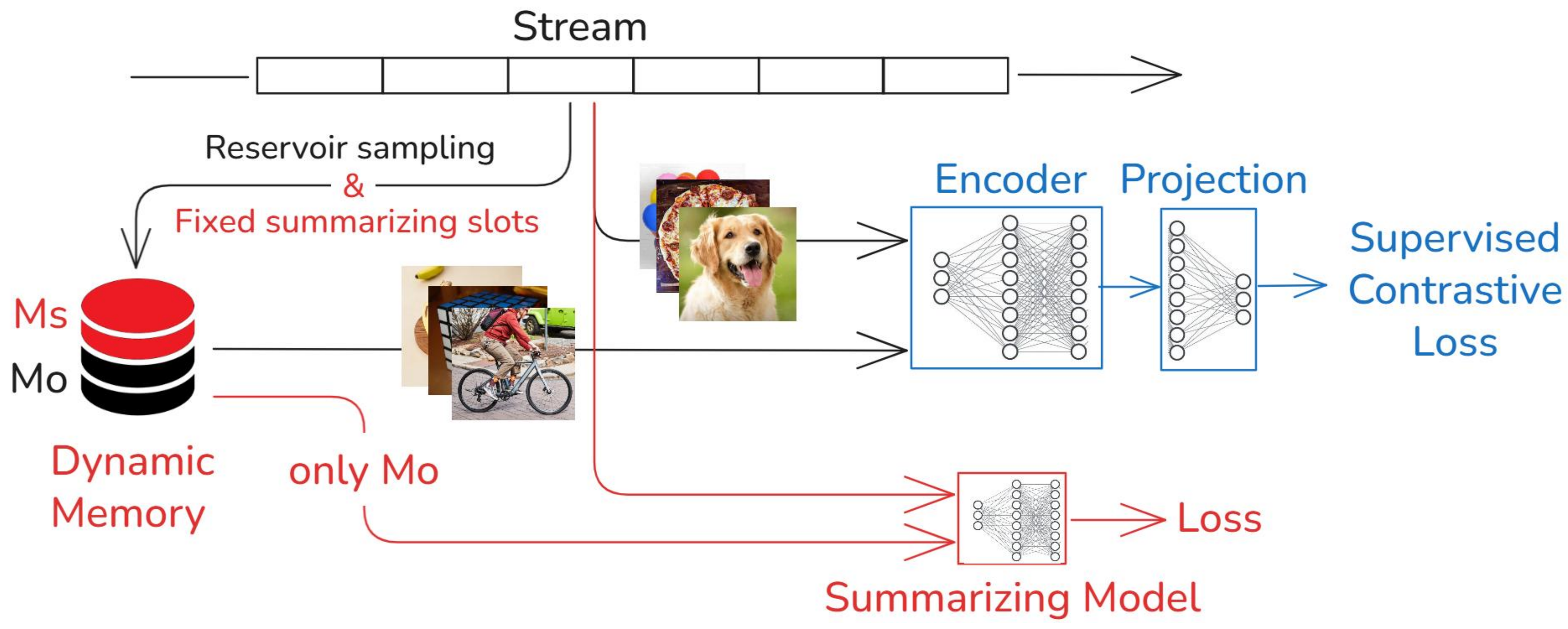
$$\mathbf{z}_i = \mathrm{Proj}(\mathrm{Encoder}(x_i))$$
$$P(i) = \{p \in \{1,\dots,N\}\setminus\{i\} \mid y_p = y_i\}$$
$$A(i) = \{a \in \{1,\dots,N\}\setminus\{i\}\}$$





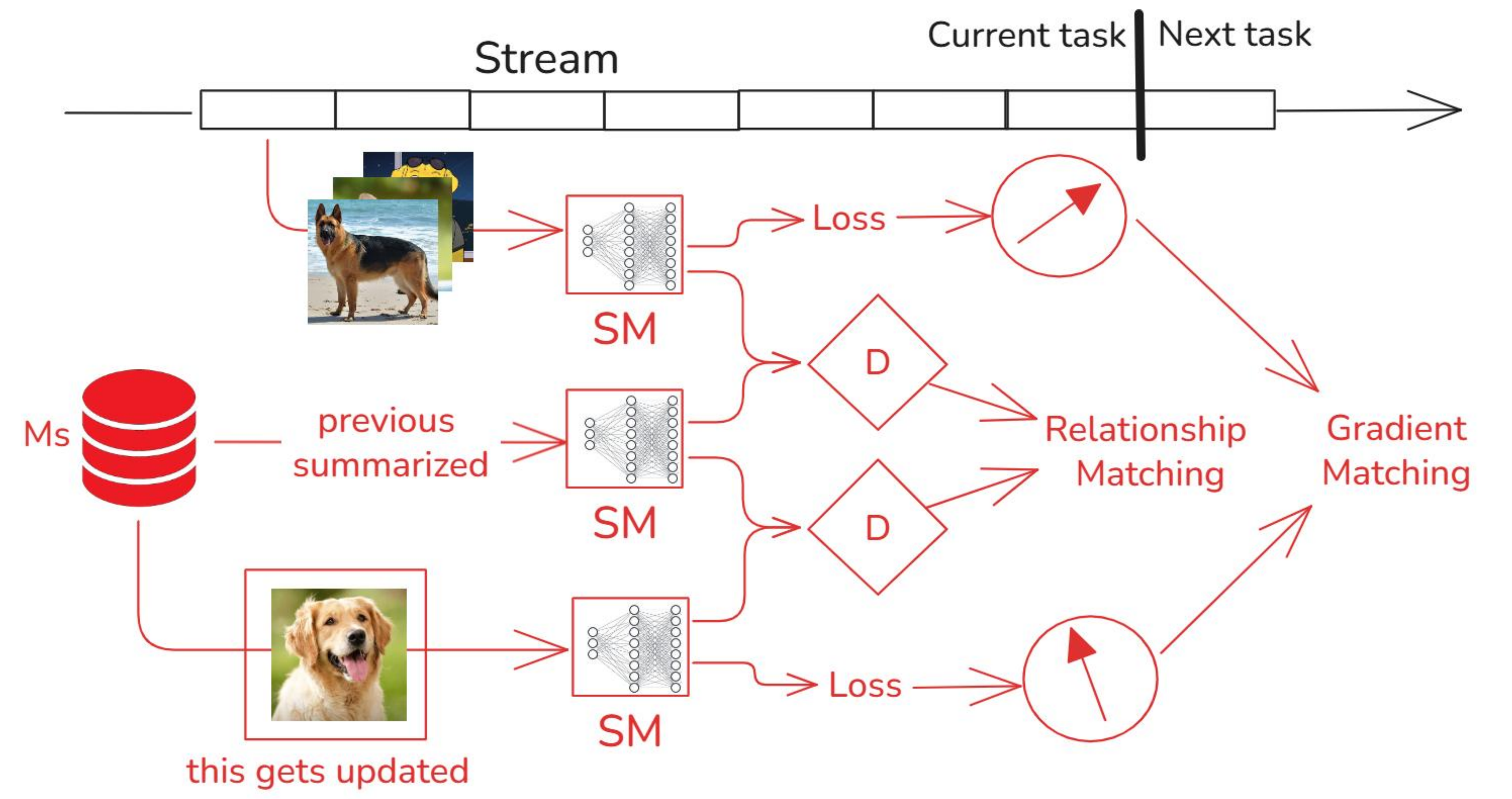
ENHANCING THE MEMORY (SSD)

- After task t is finished, the memory subset that belongs to task t carries some fixed amount of information about task t to be used for replay
- We want a more **informative** memory, without increasing memory size
- Particularly makes a difference in **low memory** settings



Main components:

- **Summarizing model**
- **Dynamic memory**
- **Stream data summarization**



- Information from the stream gets distilled into the fixed summarizing memory slots through Relationship Matching + Gradient Matching
- Calculating gradients in the forward pass - second-order optimization
- Distilling the **present** into the **past** for better **future** remembering

EXPERIMENTS

Mainly evaluated using **Average End Accuracy** on test datasets of all tasks.

The main model's encoder is a ResNet-18, while a 2-layer MLP with output size 128 is used as the projection network. A smaller 3-layer CNN is used as the summarizing model.

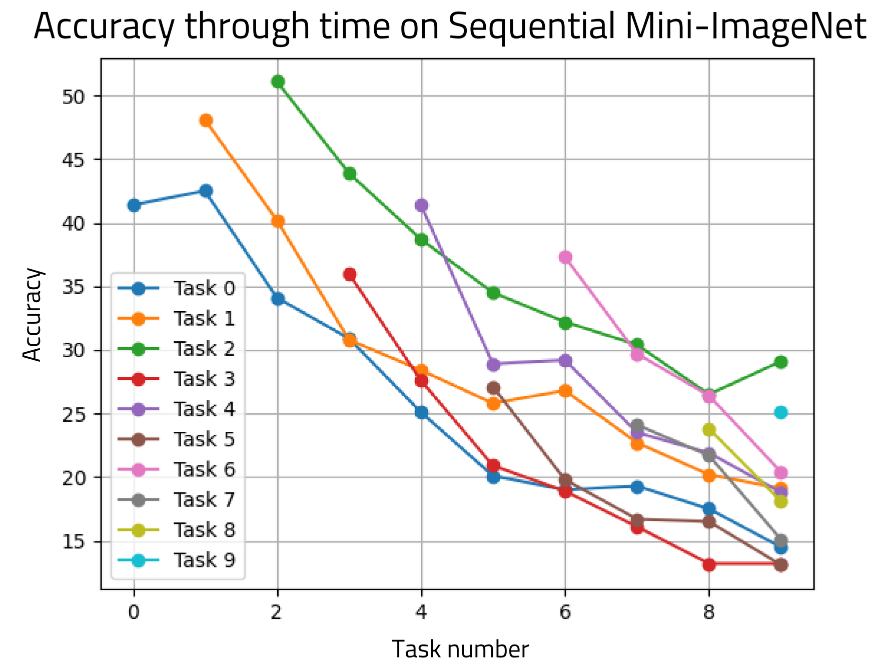
Datasets:

- **Sequential CIFAR-100** (10 tasks of 10 classes)
- **Sequential Mini-ImageNet** (10 tasks of 10 classes)
- **Sequential Tiny-ImageNet** (20 tasks of 10 classes)

AEA on Sequential CIFAR-100

SSD paper results		Our current results			
SCR	SSD	SCR	SSD	SSD+IS	Mem size
9.0%	+3.1%	**9.1%**	**+3.3%**	**+3.7%**	100
20.6%	+2.4%	19.7%	**+2.5%**	**+3.1%**	500
26.6%	+2.2%	**26.4%**	+1.4%	**+2.0%**	1000

Accuracy through time on Sequential Mini-ImageNet



- We **implement** both methods from scratch
- We *mostly* **reproduce** the benchmark results
- We suggest a couple of **practical improvements**