

Bayesian continual learning and forgetting in neural networks

Djohan Bonnet, Kellian Cottart, Tifenn Hirtzlin, Tarcisius Januel, Thomas Dalgaty, Elisa Vianello, Damien Querlioz

Nature Communications



Article

<https://doi.org/10.1038/s41467-025-64601-w>

Bayesian continual learning and forgetting in neural networks

Received: 4 April 2025

Accepted: 22 September 2025

Published online: 30 October 2025

Check for updates

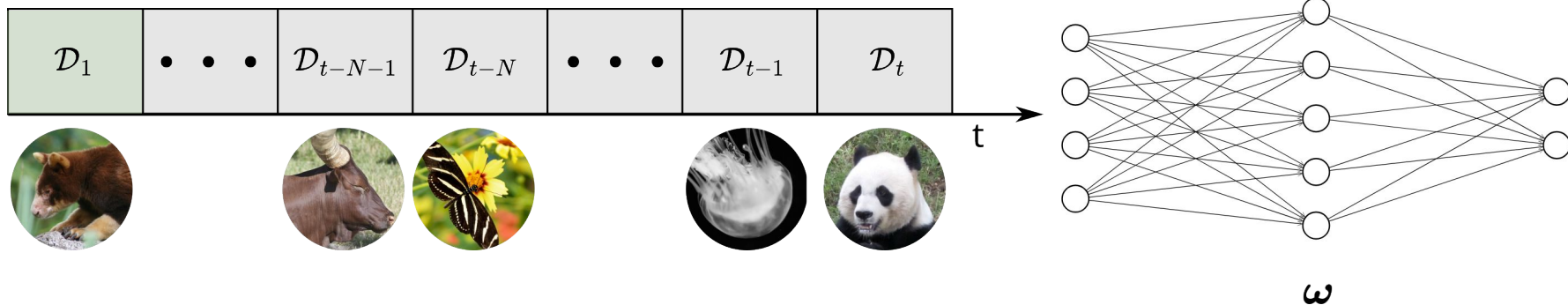
Djohan Bonnet¹, Kellian Cottart¹, Tifenn Hirtzlin², Tarcisius Januel², Thomas Dalgaty³, Elisa Vianello² & Damien Querlioz¹✉

Biological synapses effortlessly balance memory retention and flexibility, yet artificial neural networks still struggle with the extremes of catastrophic forgetting and catastrophic remembering. Here, we introduce Metaplasticity from Synaptic Uncertainty (MESU), a Bayesian update rule that scales each parameter's learning by its uncertainty, enabling a principled combination of learning and forgetting without explicit task boundaries. MESU also provides epistemic uncertainty estimates for robust out-of-distribution detection; the main computational cost is weight sampling to compute predictive statistics. Across image-classification benchmarks, MESU mitigates forgetting while maintaining plasticity. On 200 sequential Permuted-MNIST tasks, it surpasses established synaptic-consolidation methods in final accuracy, ability to learn late tasks, and out-of-distribution data detection. In task-incremental CIFAR-100, MESU consistently outperforms conventional training techniques due to its boundary-free streaming formulation. Theoretically, MESU connects metaplasticity, Bayesian inference, and Hessian-based regularization. Together, these results provide a biologically inspired route to robust, perpetual learning.

◆ Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Have you seen this animal yet?”

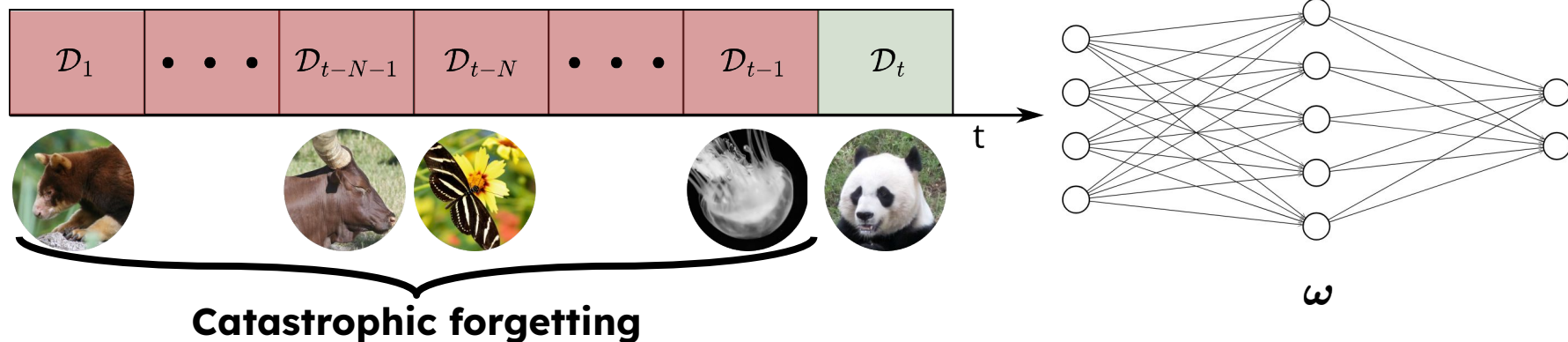


Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

◆ Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Have you seen this animal yet?”



Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." (2017)

Continual learning focuses on mitigating catastrophic forgetting

◆ Synaptic plasticity through variance

Reducing uncertainties is a biologically-plausible strategy for learning

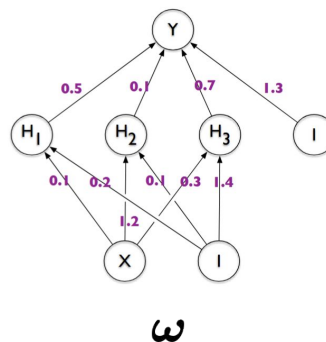


Synaptic plasticity as Bayesian inference

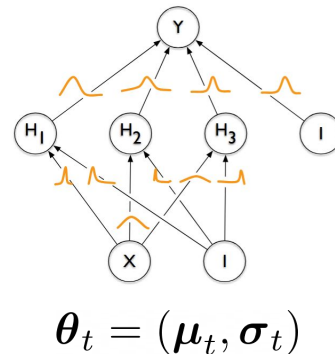
Laurence Aitchison^{1,2}✉, Jannes Jegminat^{3,4}, Jorge Aurelio Menendez^{1,5}, Jean-Pascal Pfister^{3,4}, Alexandre Pouget^{1,6,7} and Peter E. Latham^{1,7}

Aitchison L, et al. Synaptic plasticity as Bayesian inference. Nature Neuroscience 2021

Deterministic values



Random variables



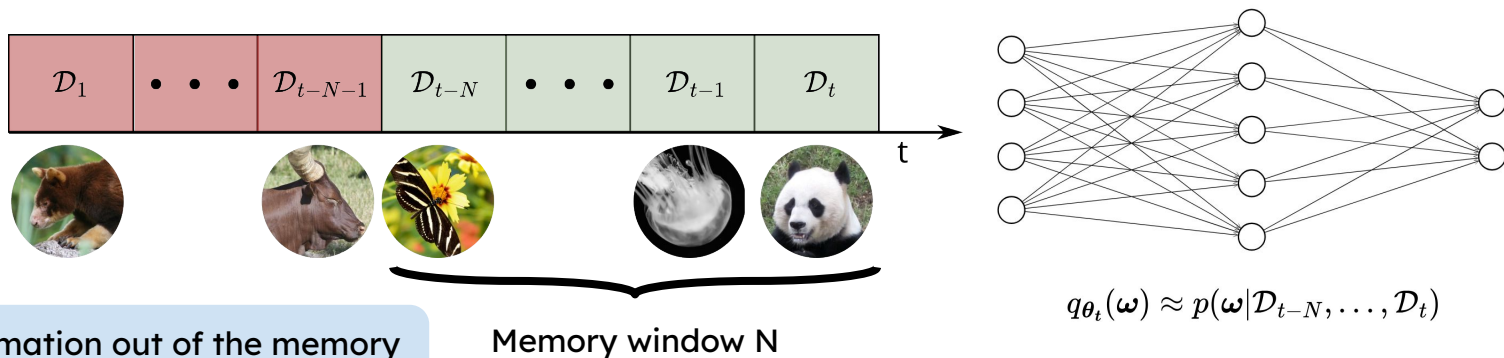
**Modelling synapses by their certainty helps
against catastrophic forgetting**

**Two parameters,
two update rules**

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

We consider a memory window of size N , corresponding to the maximum number of datasets presented to retain

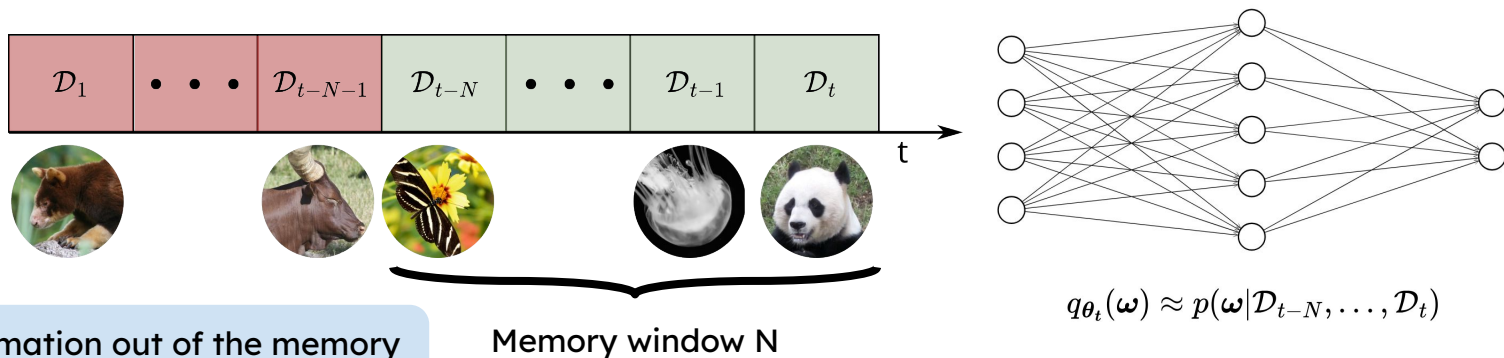


Information out of the memory window is gradually forgotten

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

We consider a memory window of size N , corresponding to the maximum number of datasets presented to retain



$$p(\omega | D_{t-N}, \dots, D_t) = \frac{p(D_t | \omega) \cdot p(\omega | D_{t-N}, \dots, D_{t-1})}{p(D_t)}$$

How to turn the memory window into a learning rule?

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

If we desire to learn only **N datasets**, Bayes' rule yields

$$\begin{array}{c}
 \text{Posterior} \\
 \underbrace{\hspace{10em}} \\
 p(\omega | \mathcal{D}_{t-N}, \dots, \mathcal{D}_t) = \underbrace{\frac{p(\mathcal{D}_t | \omega) \cdot p(\omega | \mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1})}{p(\mathcal{D}_t)}}_{\text{Learning}} \cdot \underbrace{\frac{p(\mathcal{D}_{t-N-1})}{p(\mathcal{D}_{t-N-1} | \omega)}}_{\text{Forgetting}}.
 \end{array}$$

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

If we desire to learn only **N datasets**, Bayes' rule yields

$$\begin{array}{c}
 \text{Posterior} \\
 \underbrace{\hspace{10em}} \\
 p(\omega | \mathcal{D}_{t-N}, \dots, \mathcal{D}_t) = \underbrace{\frac{p(\mathcal{D}_t | \omega) \cdot p(\omega | \mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1})}{p(\mathcal{D}_t)}}_{\text{Learning}} \cdot \underbrace{\frac{p(\mathcal{D}_{t-N-1})}{p(\mathcal{D}_{t-N-1} | \omega)}}_{\text{Forgetting}}.
 \end{array}$$

But we don't have access to that...

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

If we desire to learn only **N datasets**, Bayes' rule yields

$$\begin{array}{c}
 \text{Posterior} \\
 \underbrace{\hspace{10em}} \\
 p(\omega | \mathcal{D}_{t-N}, \dots, \mathcal{D}_t) = \underbrace{\frac{p(\mathcal{D}_t | \omega) \cdot p(\omega | \mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1})}{p(\mathcal{D}_t)}}_{\text{Learning}} \cdot \underbrace{\frac{p(\mathcal{D}_{t-N-1})}{p(\mathcal{D}_{t-N-1} | \omega)}}_{\text{Forgetting}}.
 \end{array}$$

But we don't have access to that...

Assuming each dataset has **equal marginal likelihood** and a **Gaussian prior, posterior and likelihood**, forgetting depends on the prior and the current state

$$p(\mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1} | \omega) = \prod_{i=t-N-1}^{t-1} p(\mathcal{D}_i | \omega) = [p(\mathcal{D}_{t-N-1} | \omega)]^N \propto \mathcal{N}(\omega; \mu_{L_{t-1}}, \text{diag}(\sigma_{L_{t-1}}^2))$$

◆ Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

If we desire to learn only **N datasets**, Bayes' rule yields

$$\underbrace{p(\omega | \mathcal{D}_{t-N}, \dots, \mathcal{D}_t)}_{\text{Posterior}} = \underbrace{\frac{p(\mathcal{D}_t | \omega) \cdot p(\omega | \mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1})}{p(\mathcal{D}_t)}}_{\text{Learning}} \cdot \underbrace{\frac{p(\mathcal{D}_{t-N-1})}{p(\mathcal{D}_{t-N-1} | \omega)}}_{\text{Forgetting}}.$$

But we don't have access to that...

Assuming each dataset has **equal marginal likelihood** and a **Gaussian prior, posterior and likelihood**, forgetting depends on the prior and the current state

$$p(\mathcal{D}_{t-N-1}, \dots, \mathcal{D}_{t-1} | \omega) = \prod_{i=t-N-1}^{t-1} p(\mathcal{D}_i | \omega) = [p(\mathcal{D}_{t-N-1} | \omega)]^N \propto \mathcal{N}(\omega; \mu_{L_{t-1}}, \text{diag}(\sigma_{L_{t-1}}^2))$$

**All likelihoods are supposed as probable,
leading to new update equations!**

◆ Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

Bayes By Backprop

$$\Delta\mu = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \mu_{t-1}}$$

$$\Delta\sigma = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \sigma_{t-1}}$$

◆ Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

Bayes By Backprop

$$\Delta\mu = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \mu_{t-1}}$$

$$\Delta\sigma = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \sigma_{t-1}}$$

Metaplasticity from Synaptic Uncertainty (MESU)

$$\Delta\mu = \underbrace{-\sigma_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \mu_{t-1}}}_{\text{Learning}} + \underbrace{\frac{\sigma_{t-1}^2}{N \sigma_{\text{prior}}^2} (\mu_{\text{prior}} - \mu_{t-1})}_{\text{Forgetting}}$$

$$\Delta\sigma = \underbrace{-\frac{\sigma_{t-1}^2}{2} \frac{\partial \mathcal{C}_t}{\partial \sigma_{t-1}}}_{\text{Learning}} + \underbrace{\frac{\sigma_{t-1}}{2 N \sigma_{\text{prior}}^2} (\sigma_{\text{prior}}^2 - \sigma_{t-1}^2)}_{\text{Forgetting}}$$

◆ Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

Bayes By Backprop

$$\Delta\mu = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \mu_{t-1}}$$

$$\Delta\sigma = -\alpha \frac{\partial \mathcal{C}'_t}{\partial \sigma_{t-1}}$$

Metaplasticity from Synaptic Uncertainty (MESU)

$$\Delta\mu = -\sigma_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \mu_{t-1}} + \frac{\sigma_{t-1}^2}{N \sigma_{\text{prior}}^2} (\mu_{\text{prior}} - \mu_{t-1})$$

$$\Delta\sigma = -\frac{\sigma_{t-1}^2}{2} \frac{\partial \mathcal{C}_t}{\partial \sigma_{t-1}} + \frac{\sigma_{t-1}}{2 N \sigma_{\text{prior}}^2} (\sigma_{\text{prior}}^2 - \sigma_{t-1}^2)$$

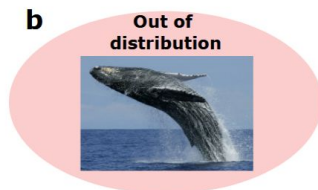
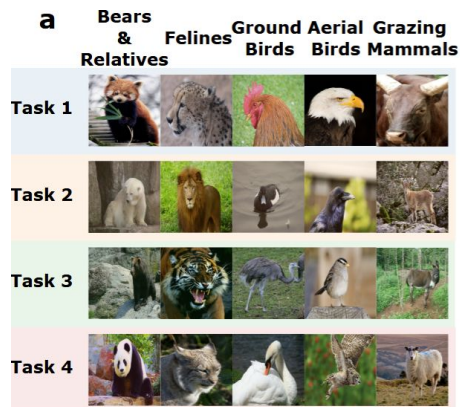
**In MESU, the variance scales the gradient.
Synapses with low variance do not update as quickly, retaining information**

MESU parameter updates are similar to biology-plausible updates!

Experimental results

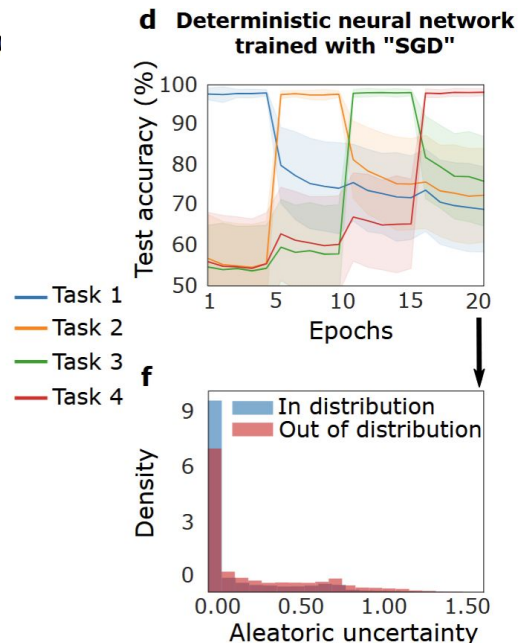
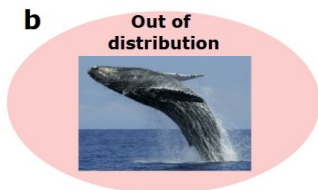
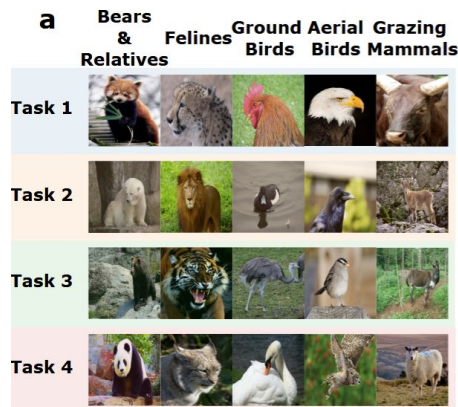
◆ Mitigating catastrophic forgetting

20 epochs of Animals dataset



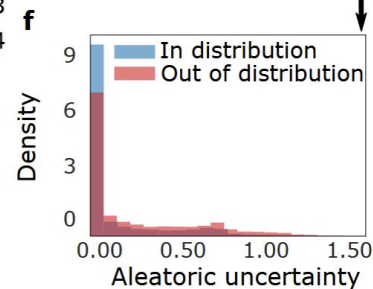
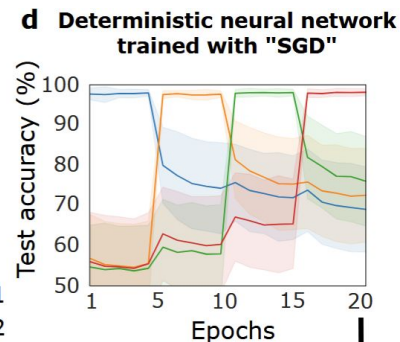
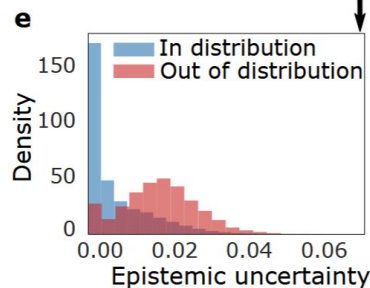
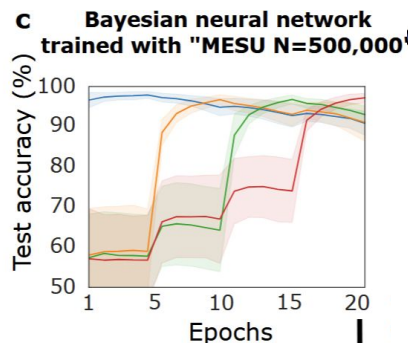
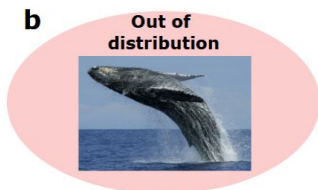
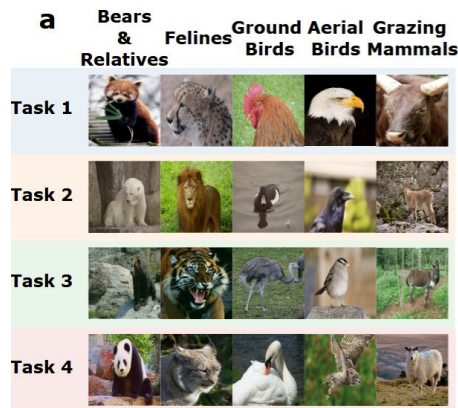
◆ Mitigating catastrophic forgetting

20 epochs of Animals dataset



◆ Mitigating catastrophic forgetting

20 epochs of Animals dataset

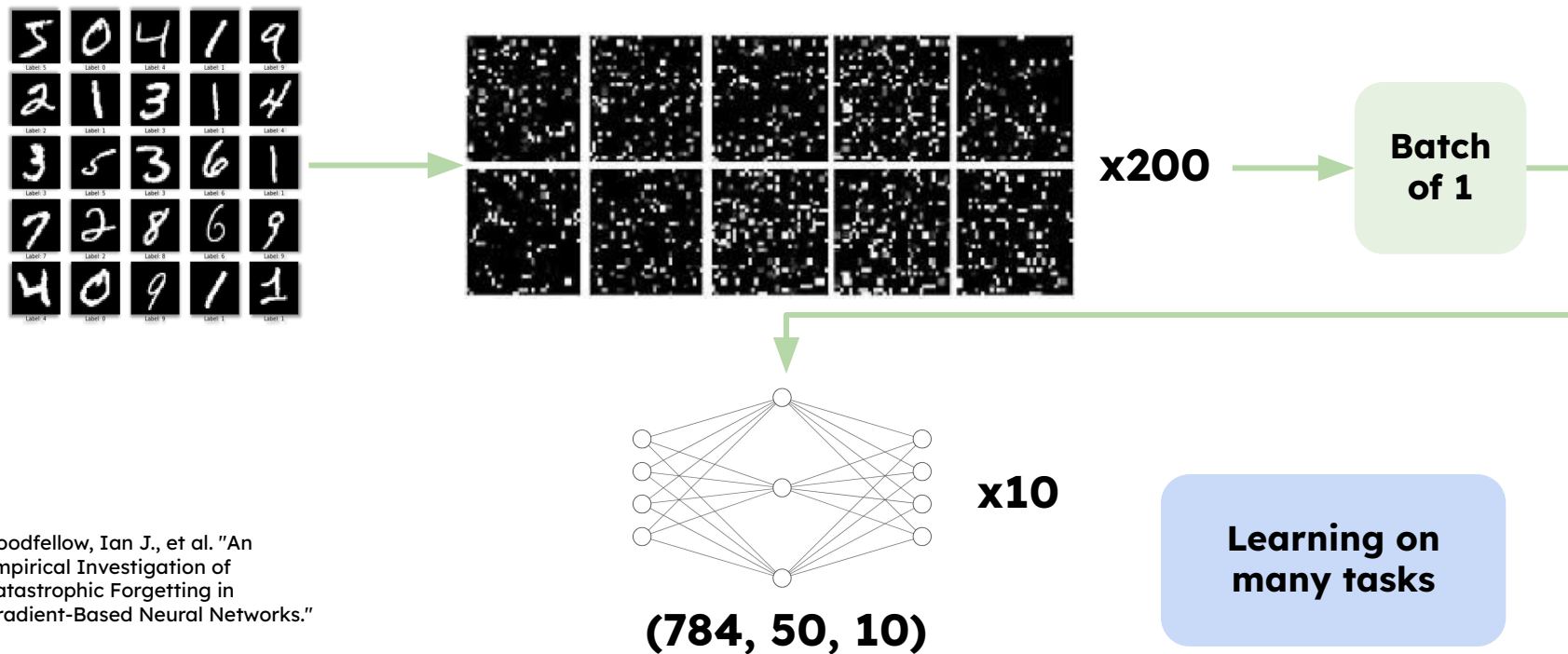


MESU mitigates catastrophic forgetting and allows to maintain multiple representations at the same time

◆ Permuted MNIST Benchmark

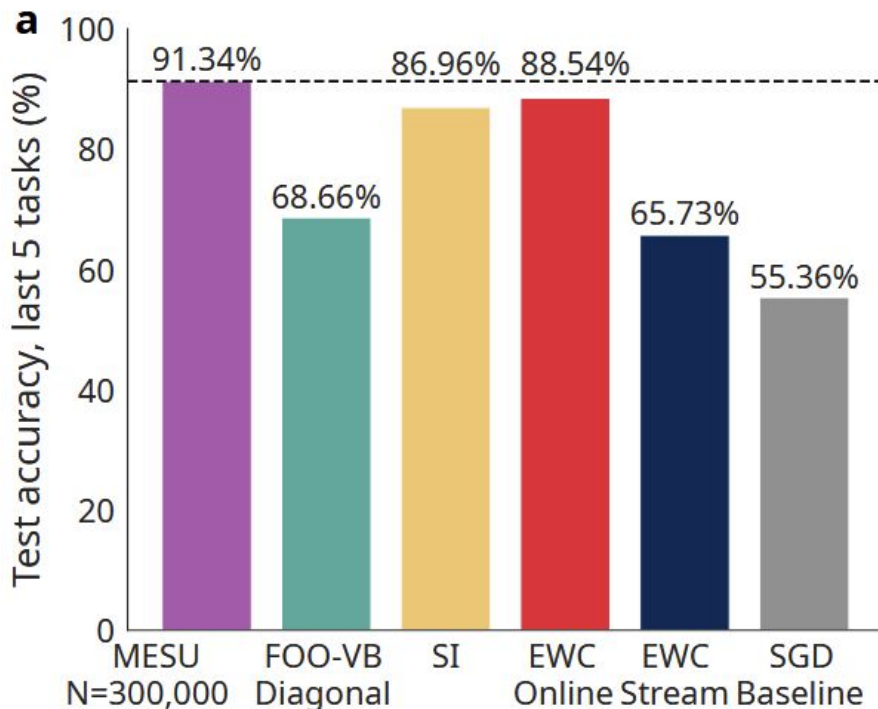
Permutation of pixels from MNIST

Simulation setup with a low amount of synapses in Online Continual Learning strategy



◆ Average accuracies over five tasks

Increased performance due to the memory window

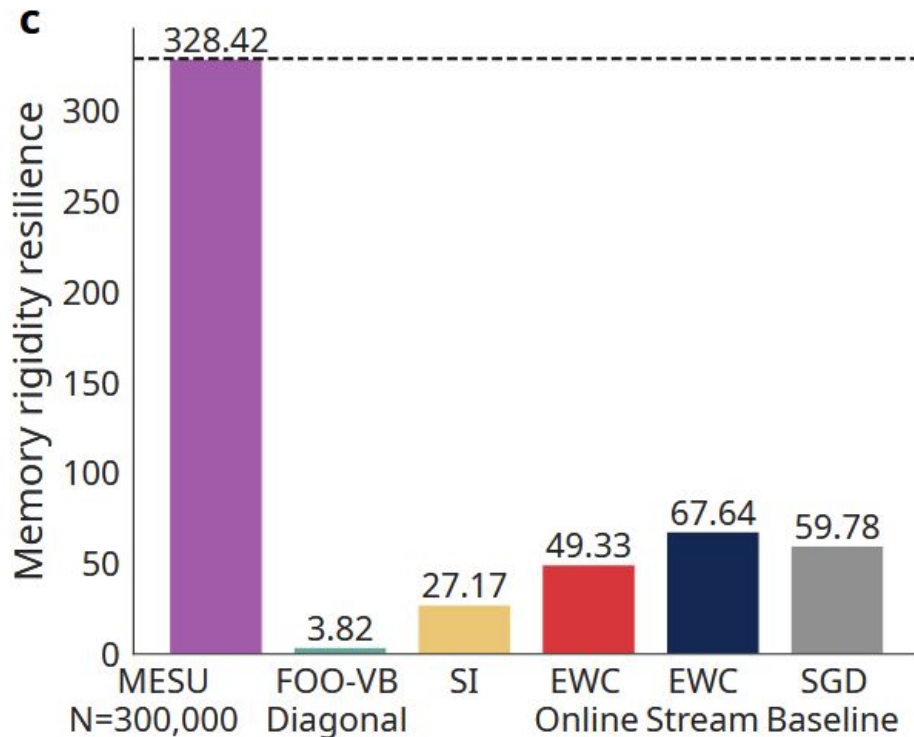


Stream: previous input
Online: previous task

Test accuracy on the
memory window
N = 300,000 = 5 tasks

◆ Resilience to loss of plasticity

Increased performance due to the memory window



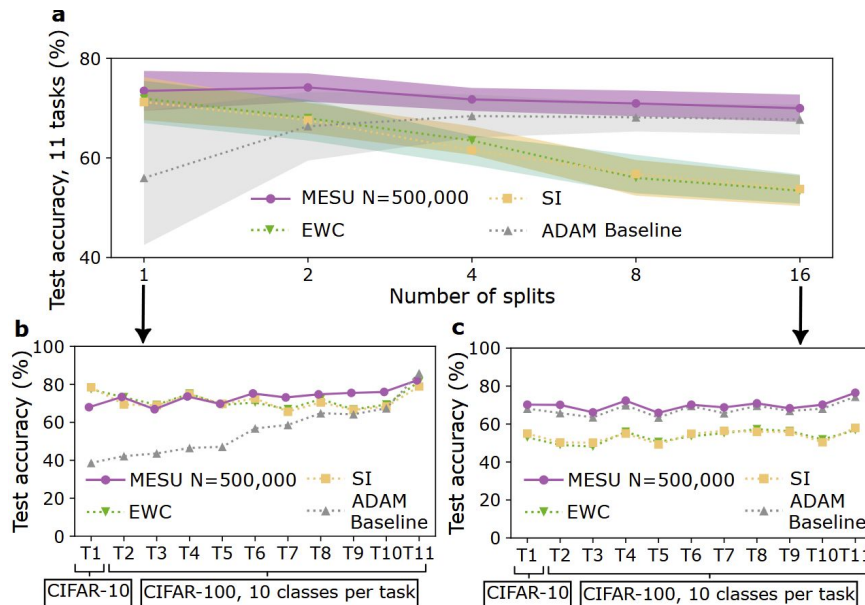
Memory rigidity
resilience between the
last and the first task

$$\mathcal{R}_t = \frac{1}{|\mathcal{A}_0 - \mathcal{A}_t|}$$

**Catastrophic
remembering** is
observed in all other
methods

◆ More complicated tasks

CIFAR-10 / CIFAR-100 splits



MESU outperforms methods with task-boundaries on end-to-end training of convolutional neural networks

Bayesian continual learning and forgetting in neural networks

Djohan Bonnet, Kellian Cottart, Tifenn Hirtzlin, Tarcisius Januel, Thomas Dalgaty, Elisa Vianello, Damien Querlioz

Nature Communications, Accepted

Thank you for listening!

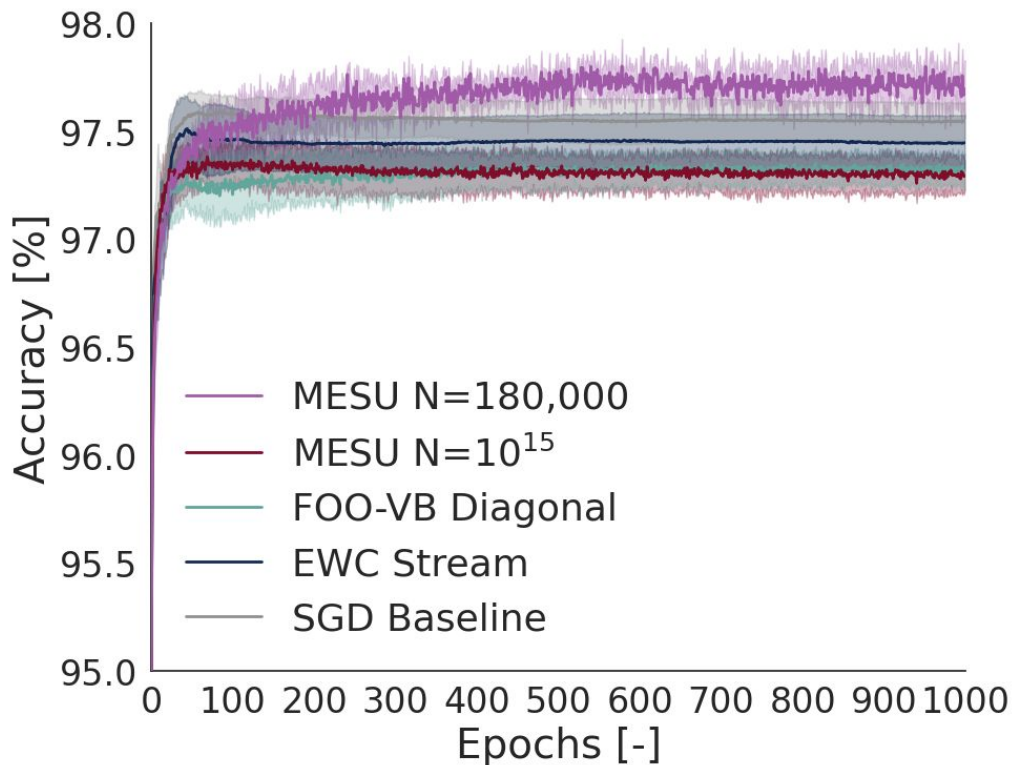
Q&A (5 minutes)



Appendix

◆ Rigidity and uncertainties - Accuracy

1000 epochs of MNIST (784 - 50 - 10) with a permutation of MNIST as OOD



◆ CNN architecture

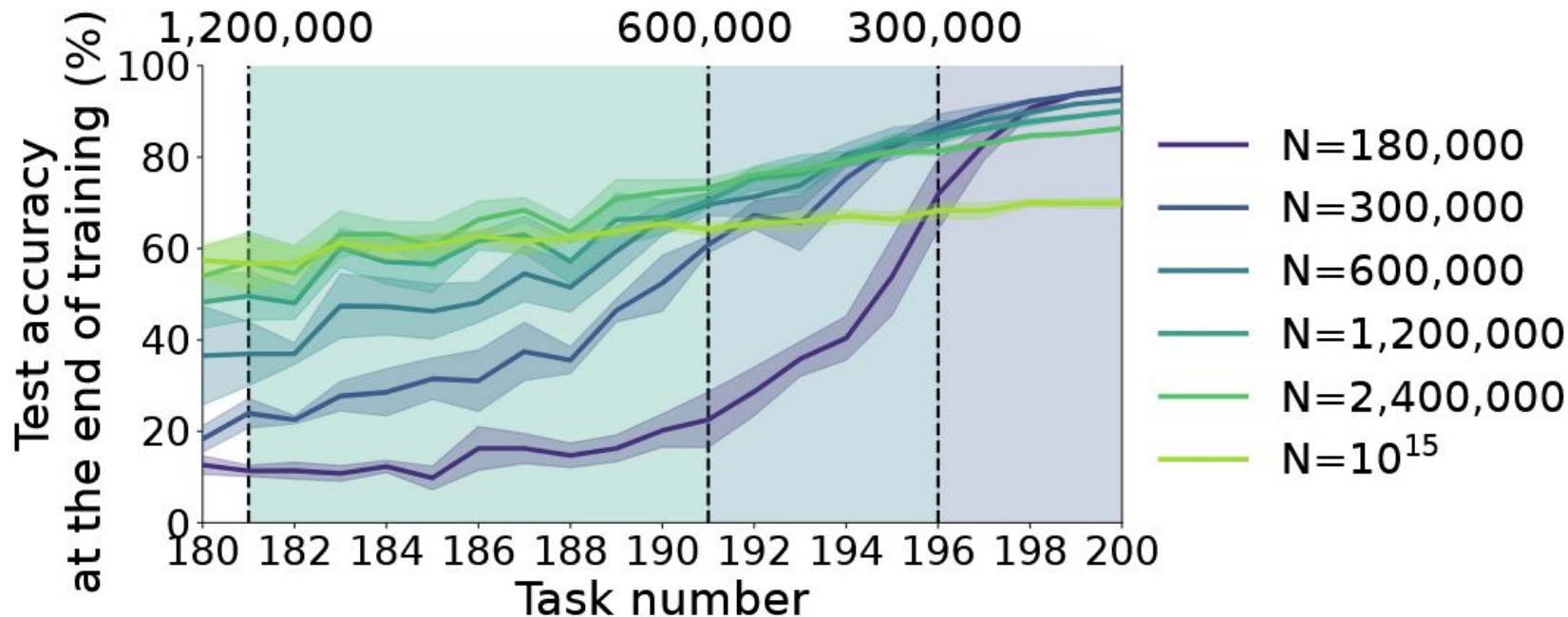
CIFAR10-100 CNN architecture

Operation	Kernel	Stride	Filters	Dropout	Nonlin.
Input	-	-	-	-	-
Convolution	3×3	1×1	32	-	ReLU
Convolution	3×3	1×1	32	-	ReLU
MaxPool	2×2	-	-	0.25	-
Convolution	3×3	1×1	64	-	ReLU
Convolution	3×3	1×1	64	-	ReLU
MaxPool	2×2	-	-	0.25	-
Fully connected	-	-	512	0.5	ReLU
Task 1: Fully connected	-	-	10	-	-
$\vdots$	-	-	$\vdots$	-	-
Task m : Fully connected	-	-	10	-	-

Table 2. CIFAR-10/100 model architecture, and dropout parameters. m : number of tasks.

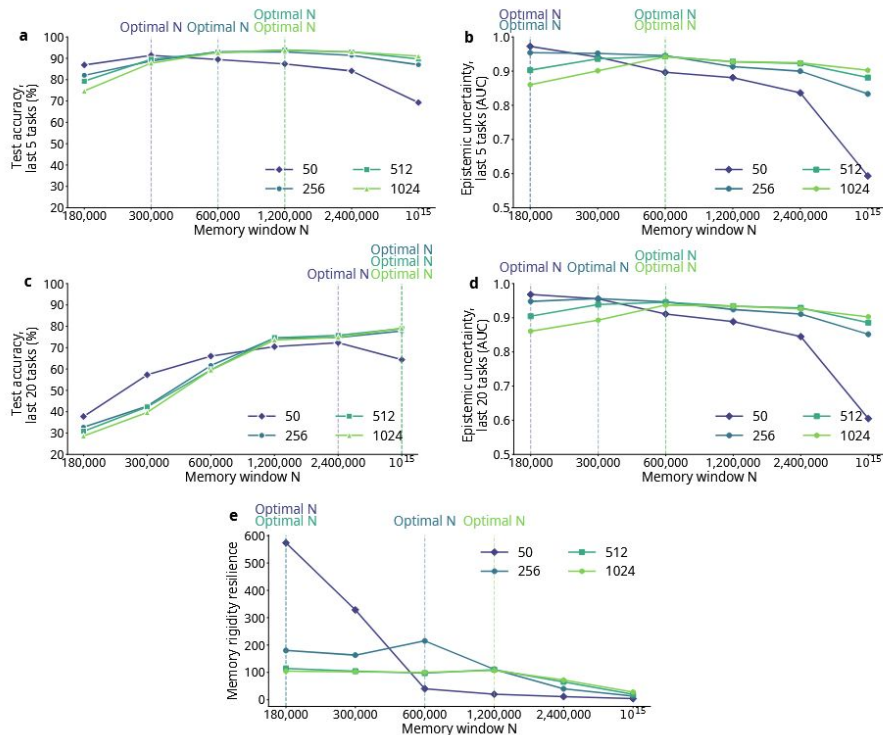
◆ N ablation study

Effect of the memory window on performance



◆ N ablation study

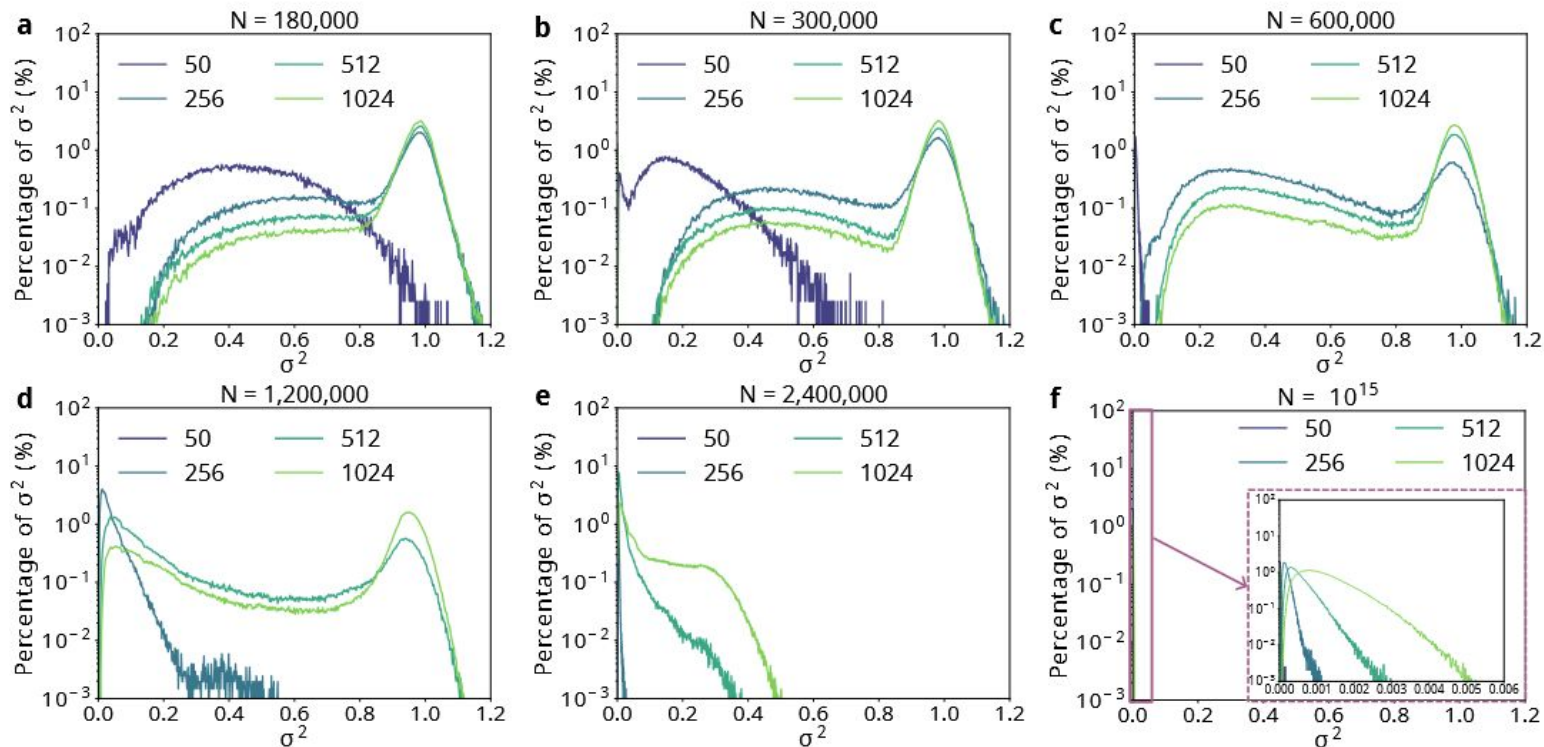
What is the optimal N value?



Optimal N value depends on what we are trying to maximize: amount of tasks, accuracy, OOD detection

◆ N ablation study

Effect of the memory window on variance



◆ Permuted MNIST detail

Effect of the memory window on variance

