Bayesian continual learning and forgetting in neural networks

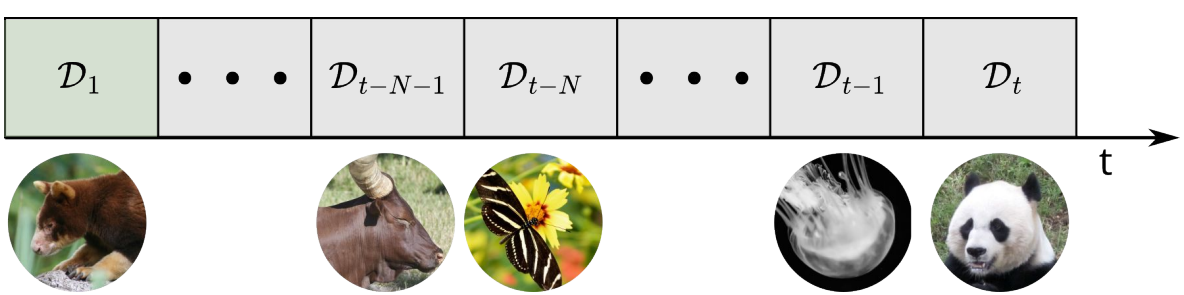
Djohan Bonnet, **Kellian Cottart**, Tifenn Hirtzlin, Tarcisius Januel, Thomas Dalgaty, Elisa Vianello, Damien Querlioz

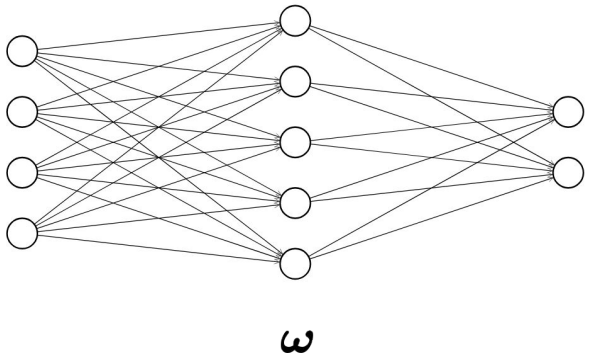
Nature Communications

Catastrophic forgetting

Forgetting about previously seen data

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



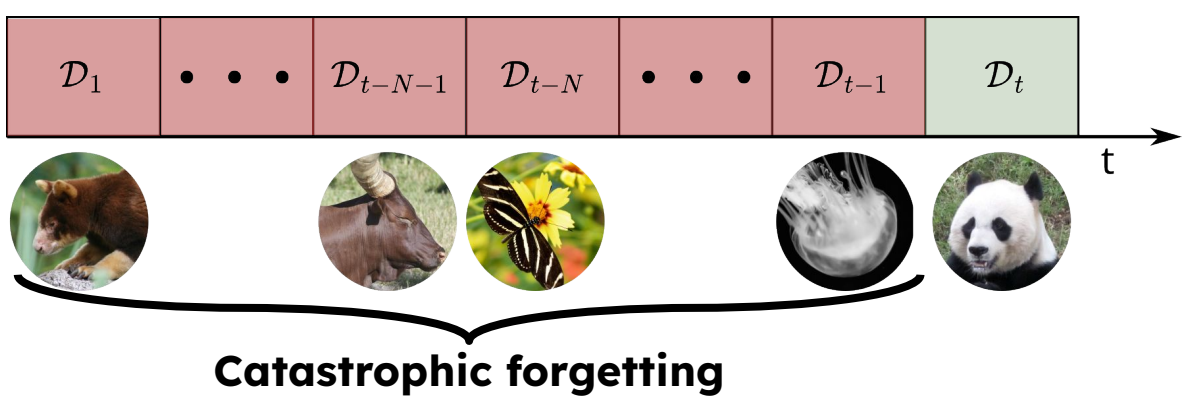


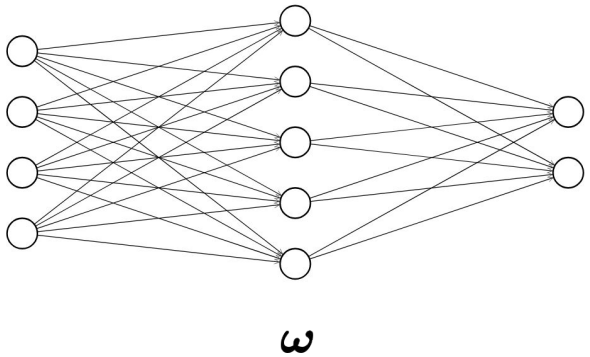
Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

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

nature neuroscience

ARTICLES

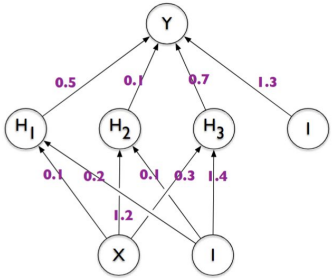
https://doi.org/10.1038/s41593-021-00809-5

Check for updates

Synaptic plasticity as Bayesian inference

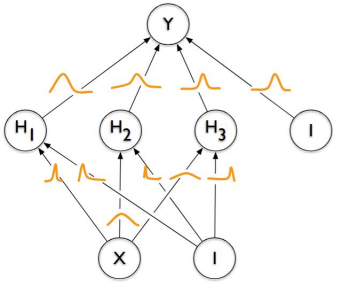
Laurence Aitchison[1,2], Jannes Jegminat[3,4], Jorge Aurelio Menendez[1,5], Jean-Pascal Pfister[3,4], Alexandre Pouget[1,6,7] and Peter E. Latham[1,7]

Deterministic values



ω

Random variables



$\theta_t = (\mu_t, \sigma_t)$

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

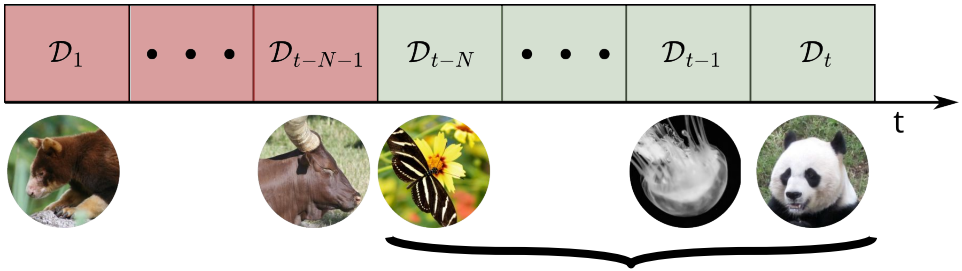
Two parameters, two update rules

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

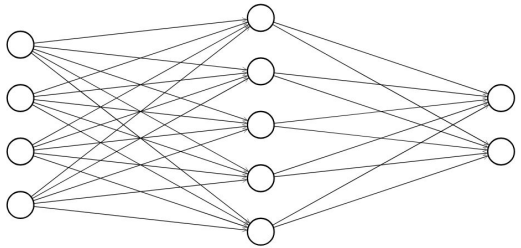
Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

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



Memory window N



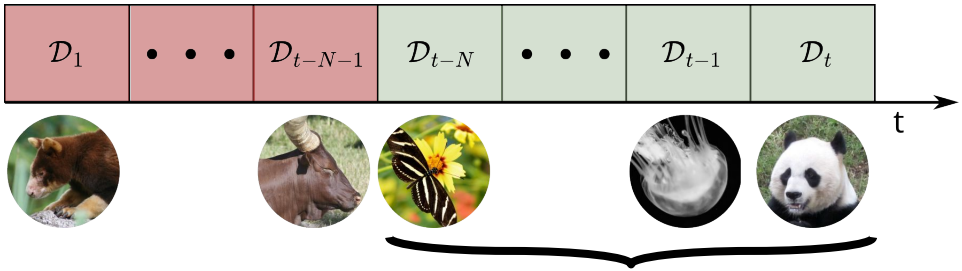
$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

Information out of the memory window is gradually forgotten

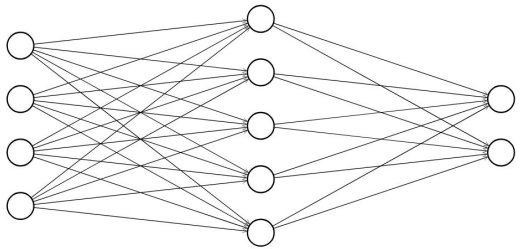
Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

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



Memory window N



$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

Information out of the memory window is gradually forgotten

$$p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, ..., \mathcal{D}_t) = \frac{p(\mathcal{D}_t|\boldsymbol{\omega}) \cdot p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, ..., \mathcal{D}_{t-1})}{p(\mathcal{D}_t)}$$

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

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

Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

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

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

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

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

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

$$p\left(\mathcal{D}_{t-N-1},\ldots,\mathcal{D}_{t-1}\mid\boldsymbol{\omega}\right) = \prod_{i=t-N-1}^{t-1} p(\mathcal{D}_i\mid\boldsymbol{\omega}) = \left[p(\mathcal{D}_{t-N-1}\mid\boldsymbol{\omega})\right]^N \propto \mathcal{N}\left(\boldsymbol{\omega};\boldsymbol{\mu}_{L_{t-1}},\mathrm{diag}(\boldsymbol{\sigma}^2_{L_{t-1}})\right)$$

Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

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

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

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

$$p(\mathcal{D}_{t-N-1},\ldots,\mathcal{D}_{t-1}\mid\boldsymbol{\omega}) = \prod_{i=t-N-1}^{t-1} p(\mathcal{D}_i\mid\boldsymbol{\omega}) = \left[p(\mathcal{D}_{t-N-1}\mid\boldsymbol{\omega})\right]^N \propto \mathcal{N}\left(\boldsymbol{\omega};\boldsymbol{\mu}_{L_{t-1}},\mathrm{diag}(\boldsymbol{\sigma}^2_{L_{t-1}})\right)$$

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

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

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

Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

Bayes By Backprop

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

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

Metaplasticity from Synaptic Uncertainty (MESU)

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

$$\Delta\boldsymbol{\sigma} = \overbrace{-\frac{\boldsymbol{\sigma}^2_{t-1}}{2} \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\sigma}_{t-1}}}^{\text{Learning}} + \overbrace{\frac{\boldsymbol{\sigma}_{t-1}}{2\,N\,\boldsymbol{\sigma}^2_{\text{prior}}} \left(\boldsymbol{\sigma}^2_{\text{prior}} - \boldsymbol{\sigma}^2_{t-1}\right)}^{\text{Forgetting}}$$

Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

Bayes By Backprop

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

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

Metaplasticity from Synaptic Uncertainty (MESU)

$$\Delta\boldsymbol{\mu} = -\boldsymbol{\sigma}_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}_{t-1}} + \frac{\boldsymbol{\sigma}_{t-1}^2}{N\,\boldsymbol{\sigma}_{\text{prior}}^2}\left(\boldsymbol{\mu}_{\text{prior}} - \boldsymbol{\mu}_{t-1}\right)$$

$$\Delta\boldsymbol{\sigma} = -\frac{\boldsymbol{\sigma}_{t-1}^2}{2} \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\sigma}_{t-1}} + \frac{\boldsymbol{\sigma}_{t-1}}{2\,N\,\boldsymbol{\sigma}_{\text{prior}}^2}\left(\boldsymbol{\sigma}_{\text{prior}}^2 - \boldsymbol{\sigma}_{t-1}^2\right)$$

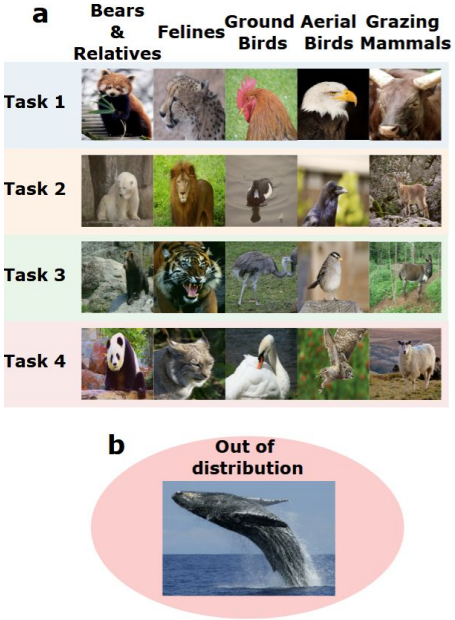
In MESU, the variance scales the gradient. Synapses with low variance do not update as quickly, retaining information

MESU parameter updates are similar to biology-plausible updates!

Experimental results

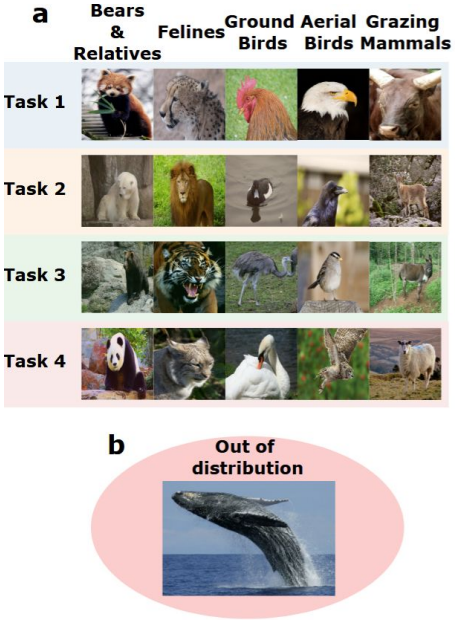
Mitigating catastrophic forgetting

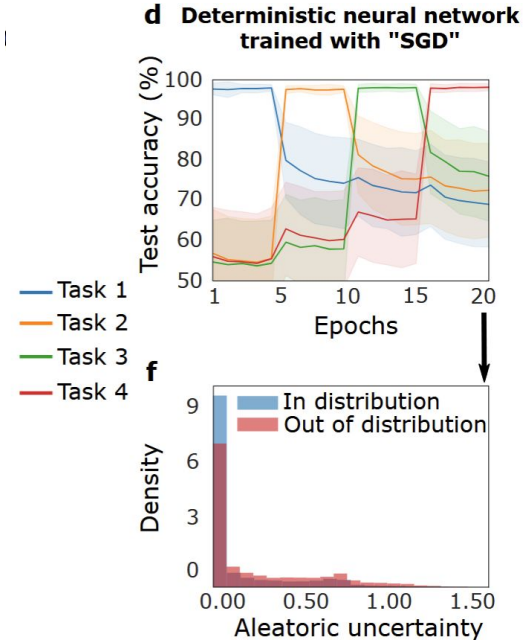
20 epochs of Animals dataset



Mitigating catastrophic forgetting

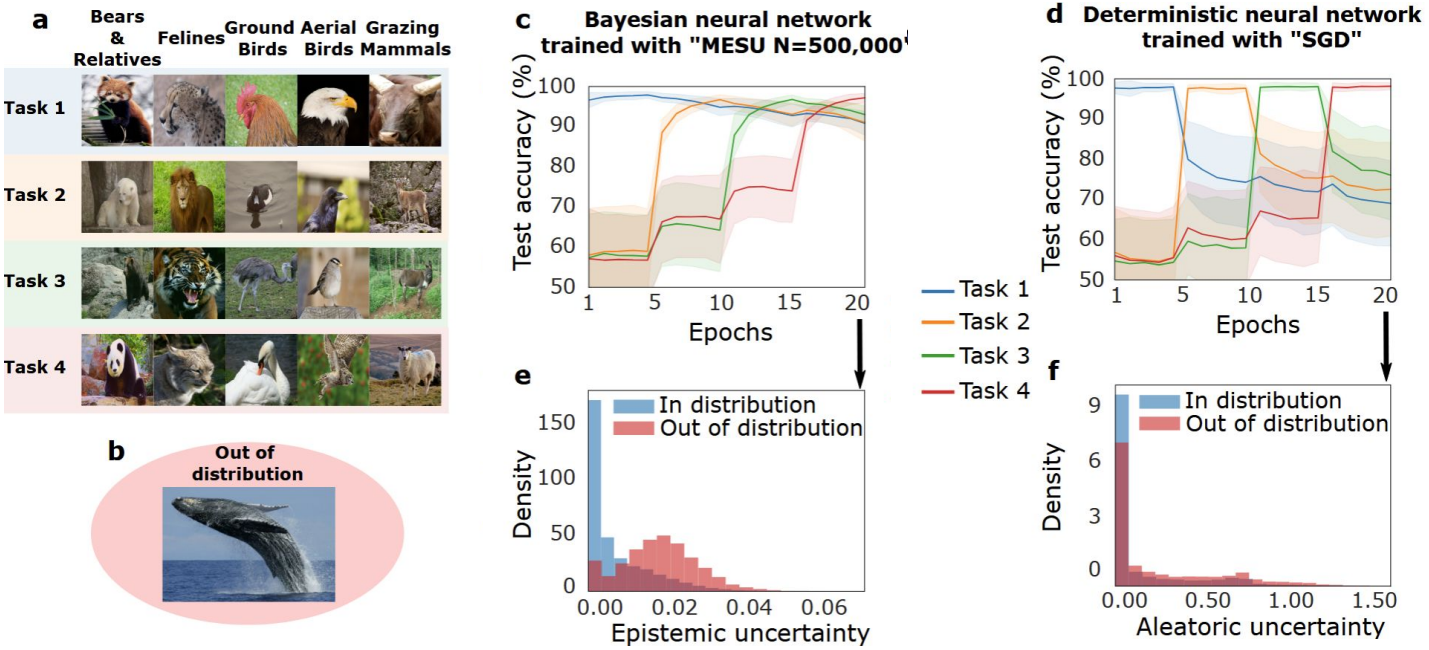
20 epochs of Animals dataset





Mitigating catastrophic forgetting

20 epochs of Animals dataset

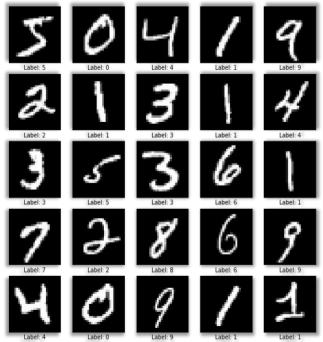


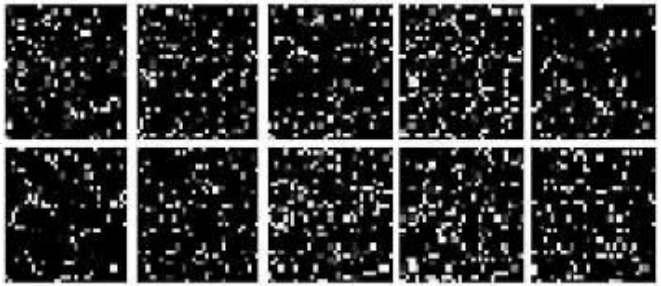
MESU mitigates catastrophic forgetting and allows to **maintain multiple representations at the same time**

Permuted MNIST Benchmark

Permutation of pixels from MNIST

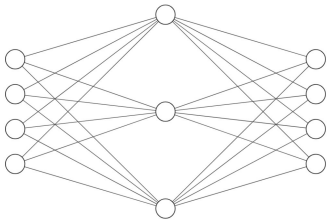
Simulation setup with a low amount of synapses in Online Continual Learning strategy





x200

Batch of 1



x10

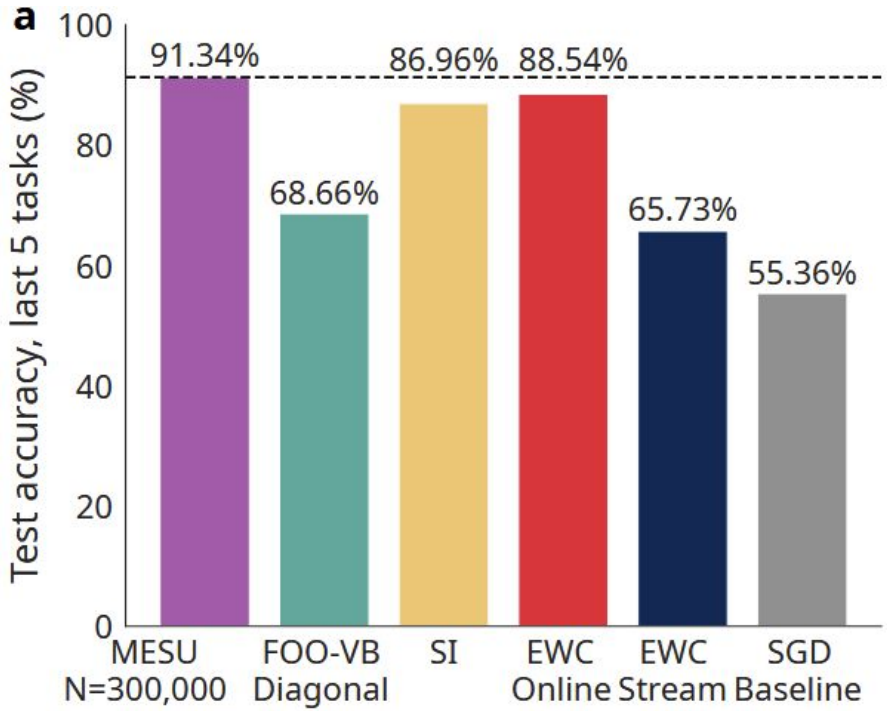
(784, 50, 10)

Learning on many tasks

Goodfellow, Ian J., et al. "An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks."

Average accuracies over five tasks

Increased performance due to the memory window

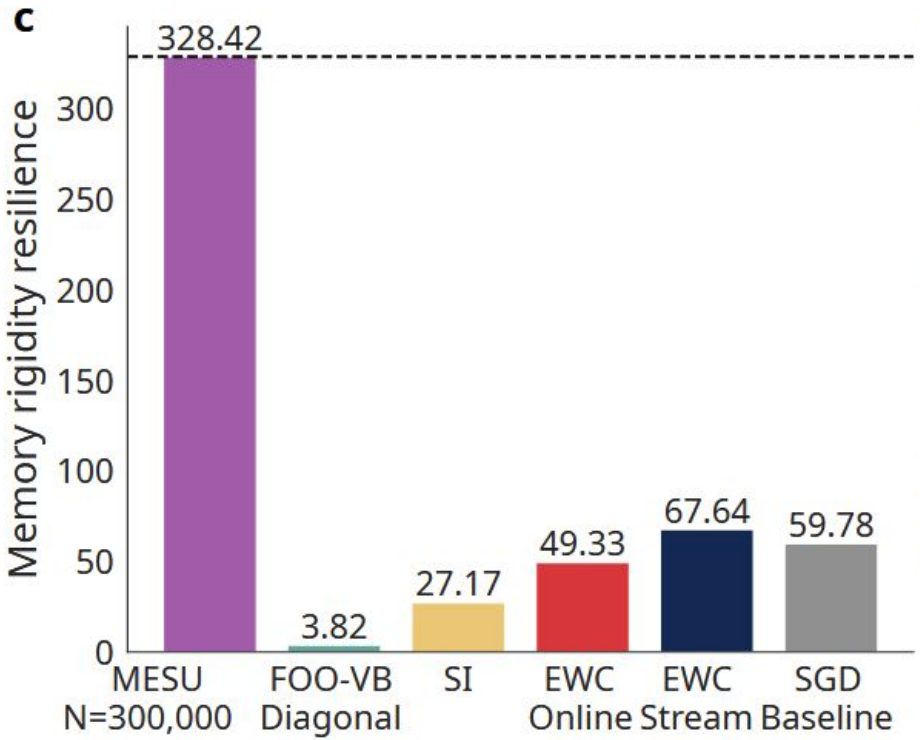


Stream: previous input
Online: previous task

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

Resilience to loss of plasticity

Increased performance due to the memory window



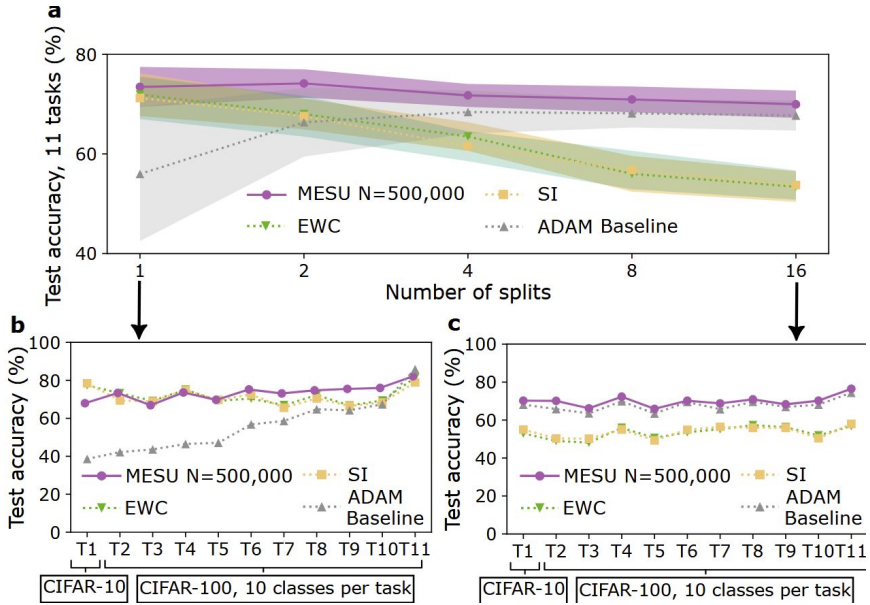
Memory rigidity resilience between the last and the first task

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

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

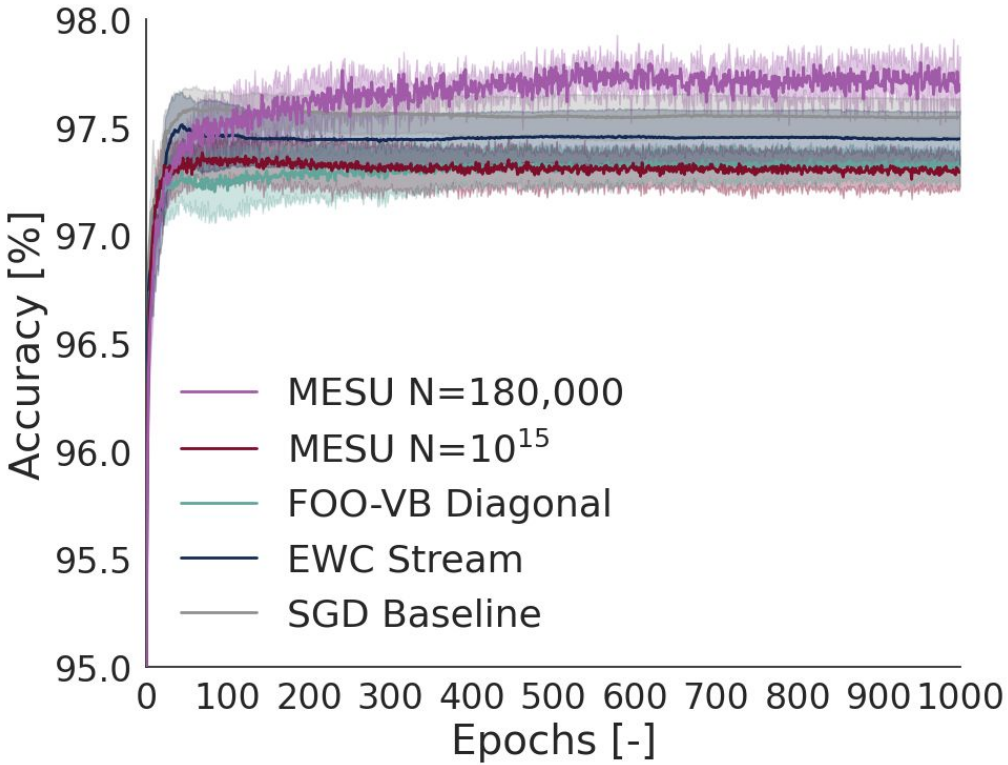
Thank you for listening!

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

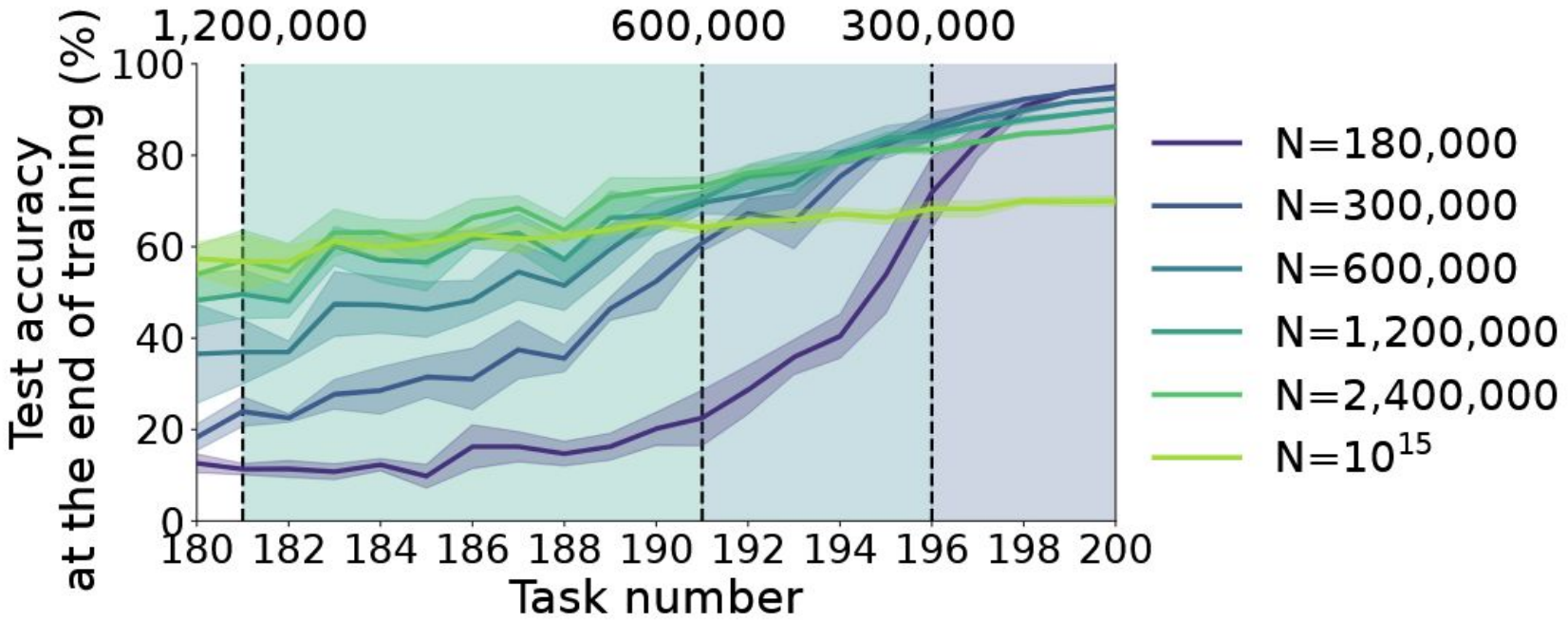
CIFAR10-100 CNN architecture

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	-	-
⋮	-	-	⋮	-	-
Task m : Fully connected	-	-	10	-	-

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

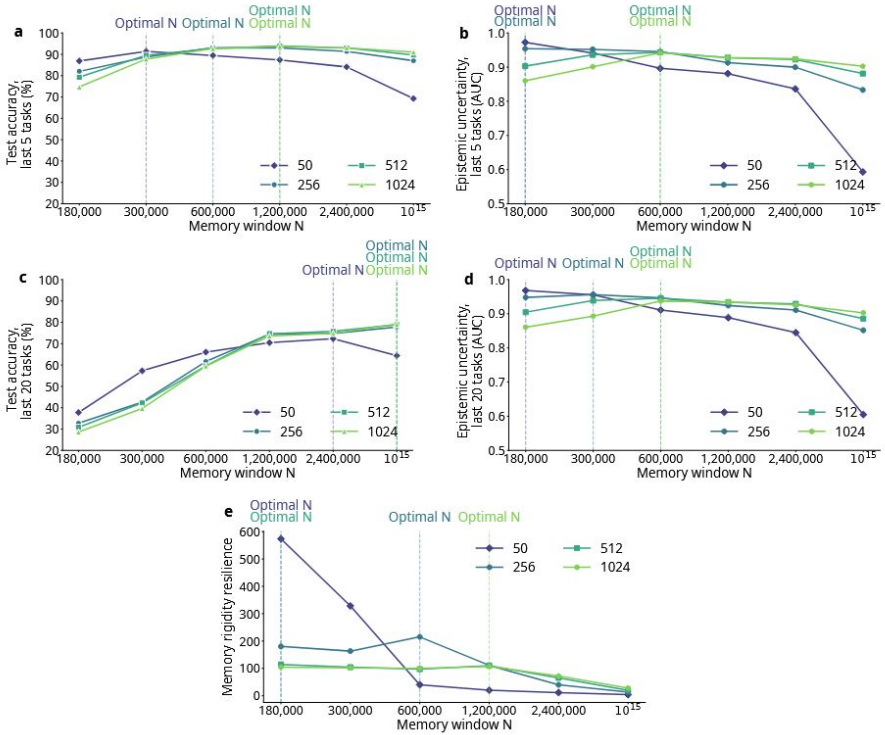
N ablation study

Effect of the memory window on performance



N ablation study

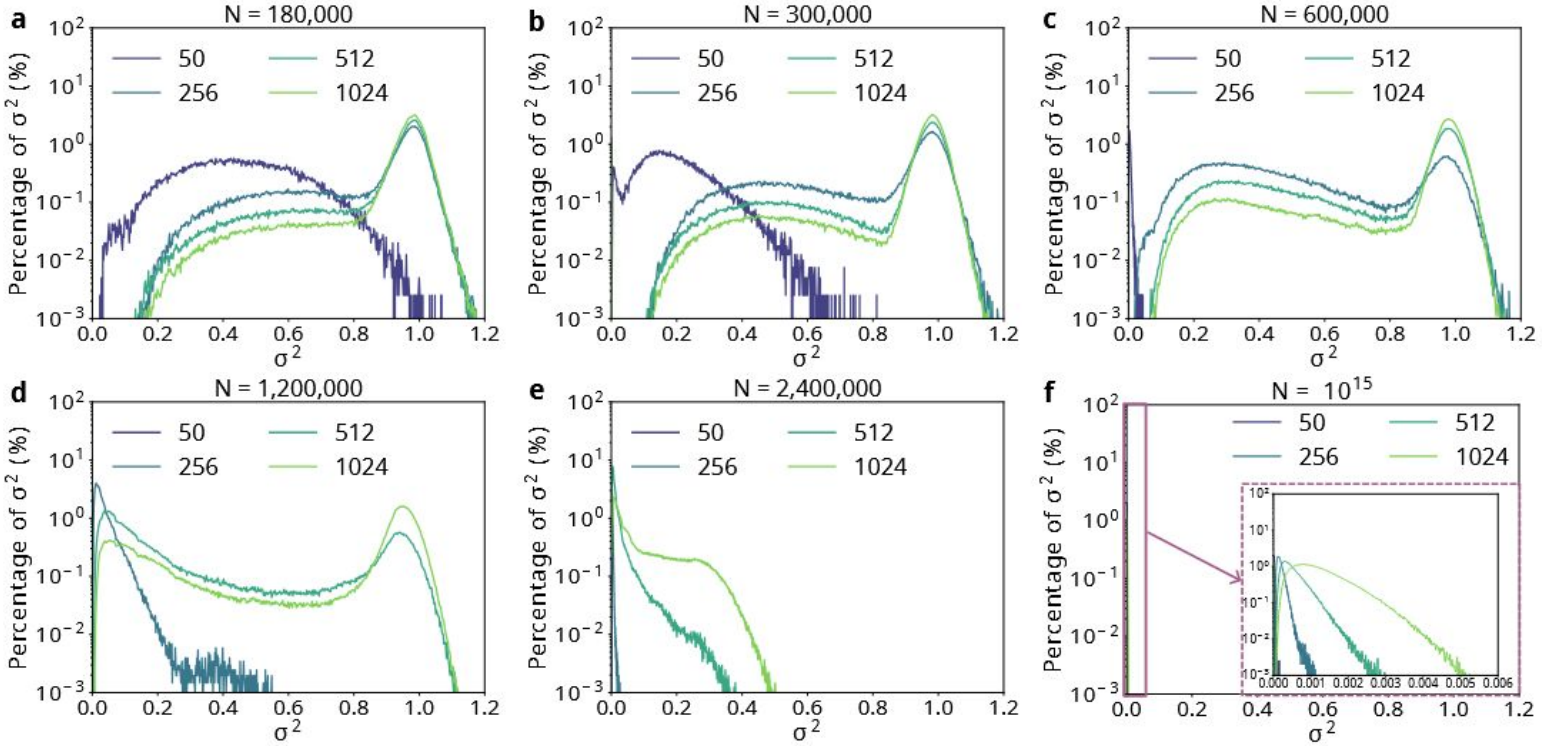
What is the optimal N value?



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