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

September 16, 2025

Introduction

Motivations and context

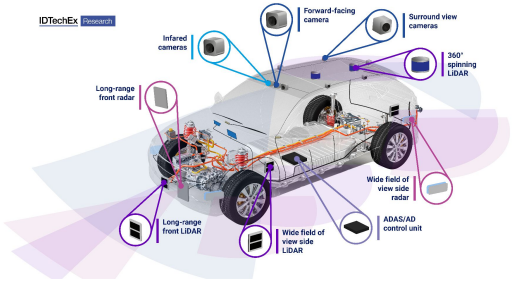
Continual learning

Real-time applications (Autonomous Vehicles, Healthcare, IoT) receive **unbatched streams of data**

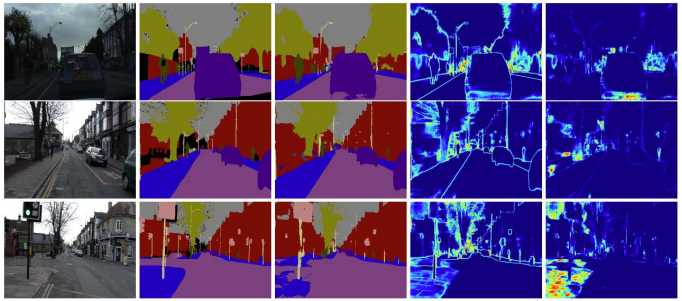
Uncertainties

Safety-critical contexts require **uncertainties** to **trust** the **predictions** that are made by the model

This talk: Uncertainties help with continual learning



IDTechEx



Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).

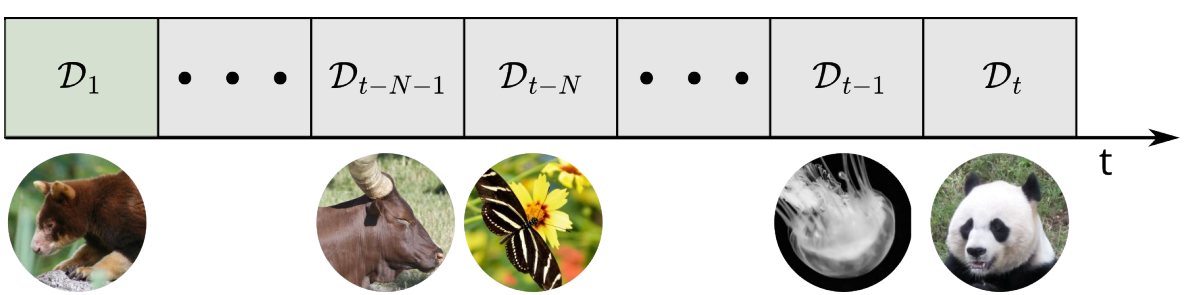
Table of contents

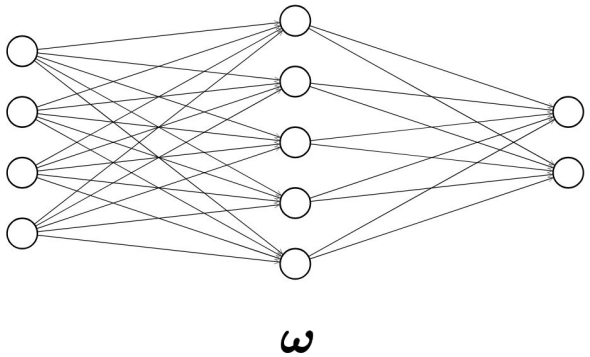
Catastrophic forgetting & remembering

Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Is this animal the one I’m looking for?”



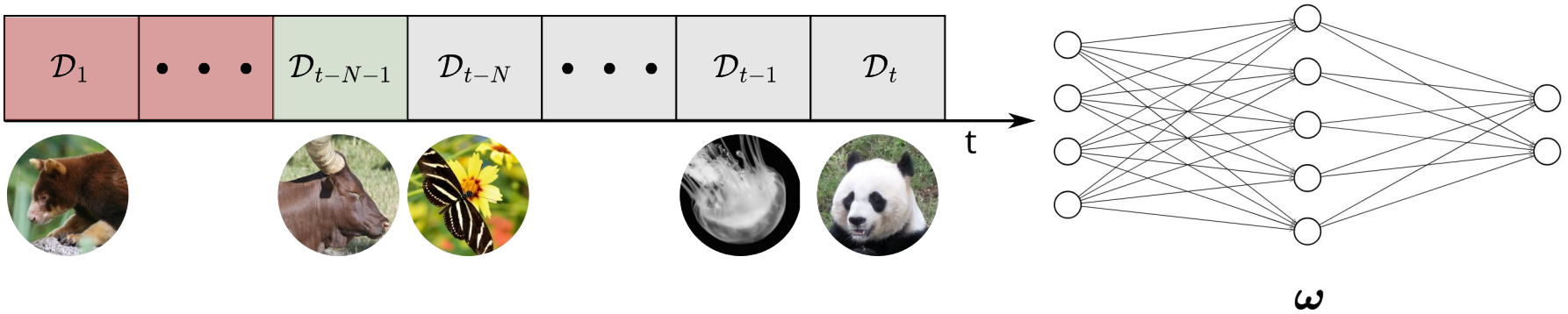


Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Is this animal the one I’m looking for?”

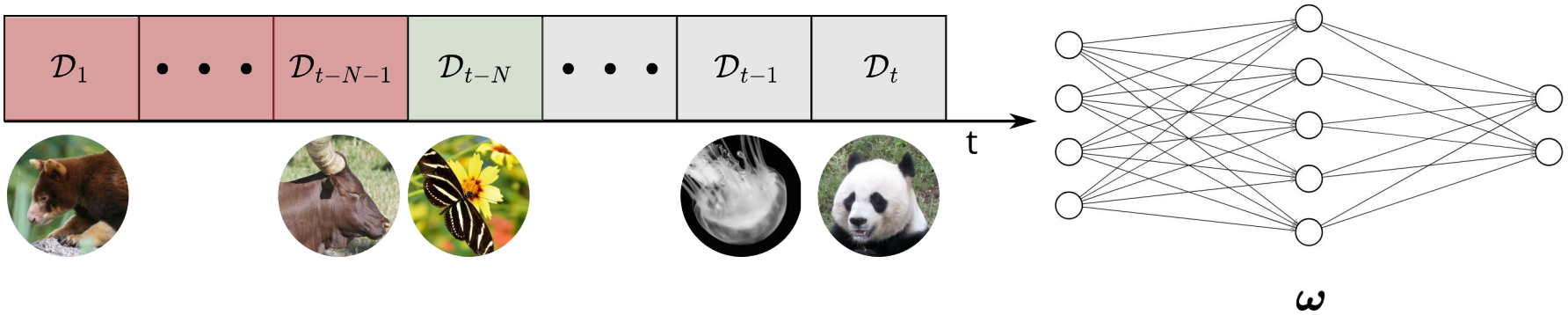


Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Is this animal the one I’m looking for?”

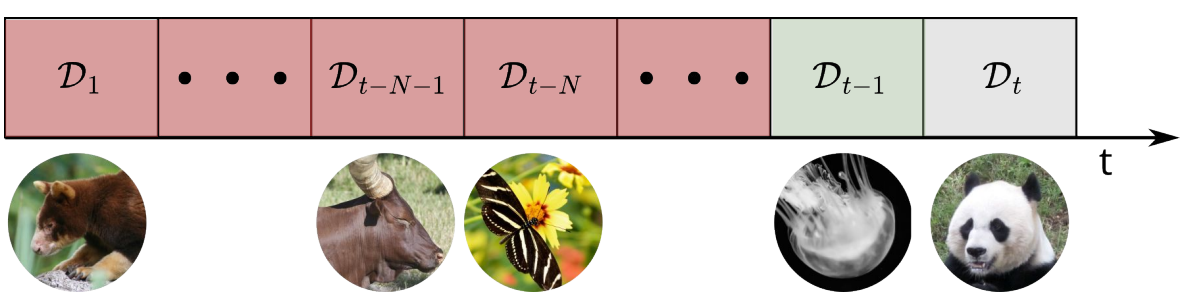


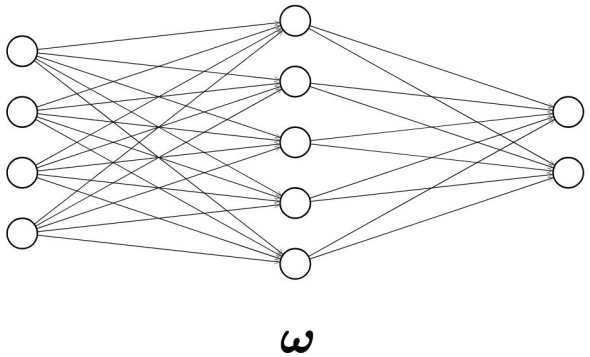
Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Is this animal the one I’m looking for?”

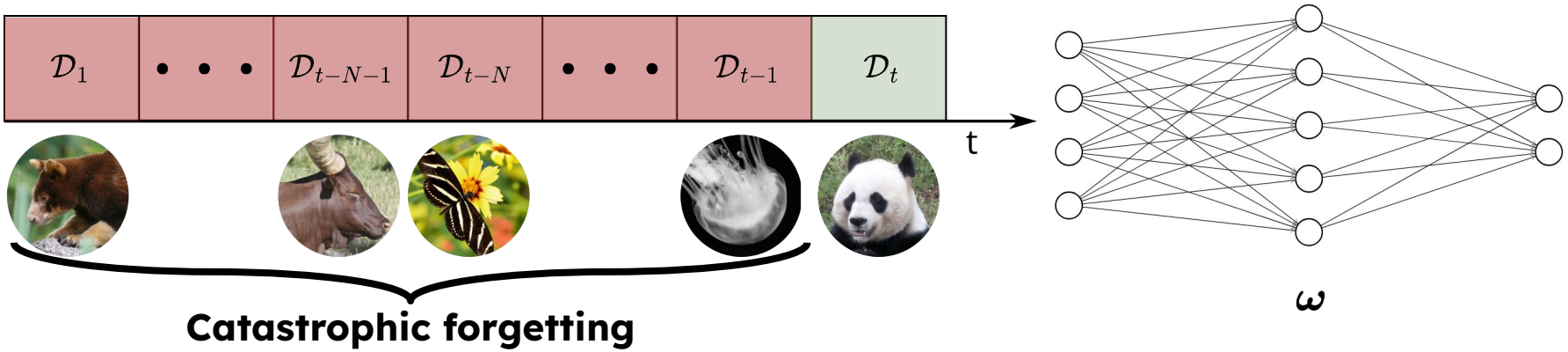




Catastrophic forgetting

Forgetting about previously seen data

When deployed in an environment, data distributions may evolve through time.
“Is this animal the one I’m looking for?”



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

Deep neural network are weak to evolving data distributions and forget previous representations

Mitigating catastrophic forgetting

Literature review for continual learning

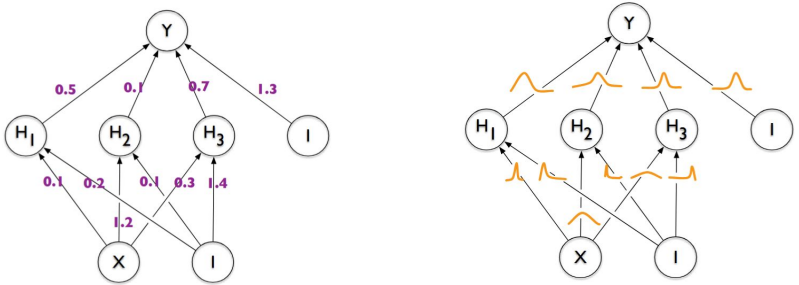
Regularization | Optimization | Replay | Representation | Architecture

Mitigating catastrophic forgetting

Literature review for continual learning

Regularization | Optimization | Replay | Representation | Architecture

Equipping each weight with a standard deviation allows to track uncertainty of each synapse

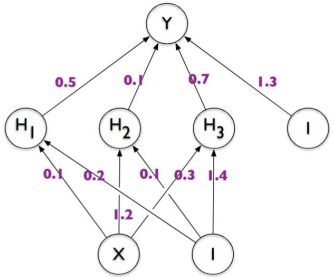


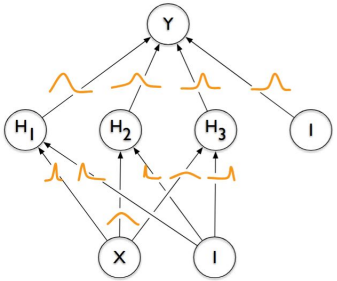
Mitigating catastrophic forgetting

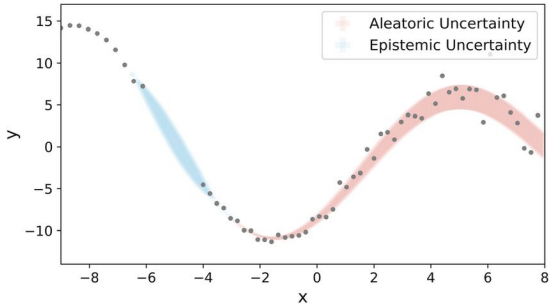
Literature review for continual learning

Regularization | Optimization | Replay | Representation | Architecture

Equipping each weight with a standard deviation allows to track uncertainty of each synapse





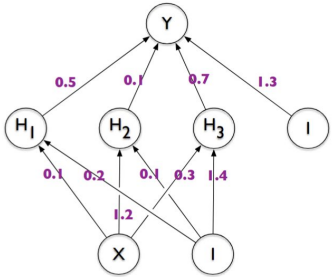


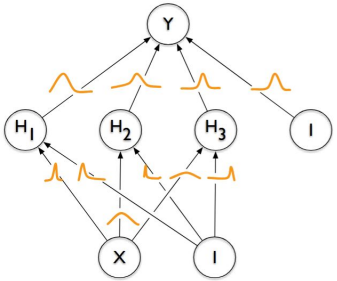
Mitigating catastrophic forgetting

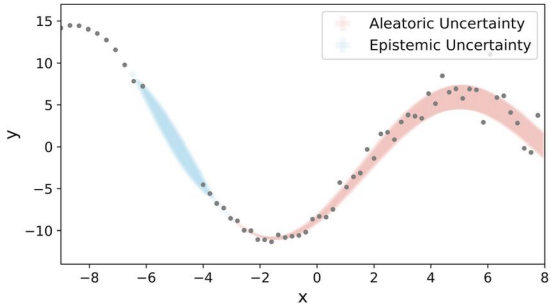
Literature review for continual learning

Regularization | Optimization | Replay | Representation | Architecture

Equipping each weight with a standard deviation allows to track uncertainty of each synapse







Blundell, Charles, et al. "Weight uncertainty in neural network." 2015.

Conditioning the model on Bayesian statistics allows to express the degree of belief in an event and face lifelong learning

Bayes' Rule against catastrophic forgetting

Updating the neural network based on prior knowledge

$$\overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1, ..., \mathcal{D}_t)}^{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}_t|\boldsymbol{\omega})}^{\text{Likelihood}} \cdot \overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1, ..., \mathcal{D}_{t-1})}^{\text{Prior}}}{p(\mathcal{D}_t)}.$$

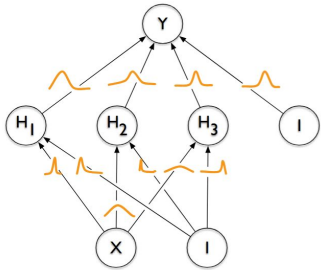
Bayes' Rule against catastrophic forgetting

Updating the neural network based on prior knowledge

$$\overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_t)}^{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}_t|\boldsymbol{\omega})}^{\text{Likelihood}} \cdot \overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_{t-1})}^{\text{Prior}}}{p(\mathcal{D}_t)}.$$

Variational Mean-field Gaussian

$$\overbrace{q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega})}^{\text{Variational Mean-field Gaussian}} = \prod_i \mathcal{N}(\omega_i; \mu_{t,i}, \sigma_{t,i}^2)$$



$$\boldsymbol{\theta}_t = (\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$$

Bayes' Rule against catastrophic forgetting

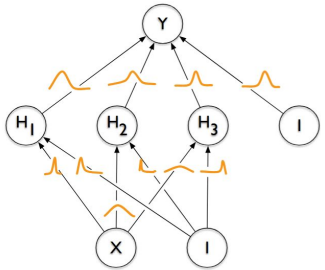
Updating the neural network based on prior knowledge

$$\overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_t)}^{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}_t|\boldsymbol{\omega})}^{\text{Likelihood}} \cdot \overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_{t-1})}^{\text{Prior}}}{p(\mathcal{D}_t)}.$$

Minimize using Kullback-Leibler Divergence

Variational Mean-field Gaussian

$$\overbrace{q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega})}^{\text{Variational Mean-field Gaussian}} = \prod_i \mathcal{N}(\omega_i; \mu_{t,i}, \sigma^2_{t,i})$$



$$\boldsymbol{\theta}_t = (\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$$

Bayes' Rule against catastrophic forgetting

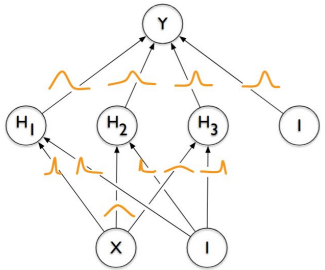
Updating the neural network based on prior knowledge

$$\overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_t)}^{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}_t|\boldsymbol{\omega})}^{\text{Likelihood}} \cdot \overbrace{p(\boldsymbol{\omega}|\mathcal{D}_1,...,\mathcal{D}_{t-1})}^{\text{Prior}}}{p(\mathcal{D}_t)}.$$

Minimize using Kullback-Leibler Divergence

Variational Mean-field Gaussian

$$\overbrace{q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega})}^{} = \prod_i \mathcal{N}(\omega_i; \mu_{t,i}, \sigma^2_{t,i})$$



$$\boldsymbol{\theta}_t = (\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$$

Zeno, Chen, et al. "Task Agnostic Continual Learning Using Online Variational Bayes." (2022)

Neural networks are mitigating catastrophic forgetting using Bayes' Rule

Synaptic metaplasticity through variance

Bayesian neural networks for lifelong learning

FOO-VB Diagonal parameter updates

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

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \right]$$

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\sigma}} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \times \boldsymbol{\epsilon} \right]$$

Synaptic metaplasticity through variance

Bayesian neural networks for lifelong learning

FOO-VB Diagonal parameter updates

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

Metaplasticity: each synapse adjusts its own learning rate

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \right]$$

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\sigma}} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \times \boldsymbol{\epsilon} \right]$$

Synaptic metaplasticity through variance

Bayesian neural networks for lifelong learning

FOO-VB Diagonal parameter updates

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

Metaplasticity: each synapse adjusts its own learning rate

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right]$$

$$\frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\sigma}} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{\partial \mathcal{L}_t(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \times \boldsymbol{\epsilon}\right]$$

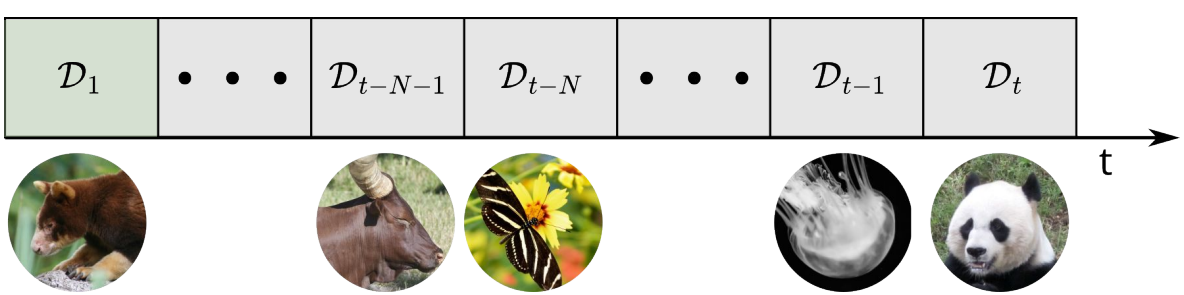
- Variance tracks **uncertainty:** the **higher** the variance, the **larger** the learning rate
- When synapses are **very certain**, the mean cannot change: **information is retained**

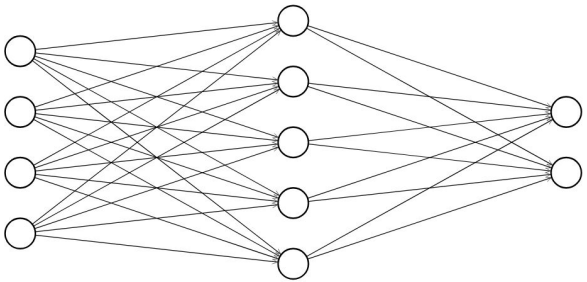
Bayes' Rule applied to multiple datasets leads **inherently metaplastic** updates

Catastrophic remembering

Inability to learn induced by too much data

Using Bayesian neural networks adapted to lifelong learning, catastrophic forgetting is prevented.





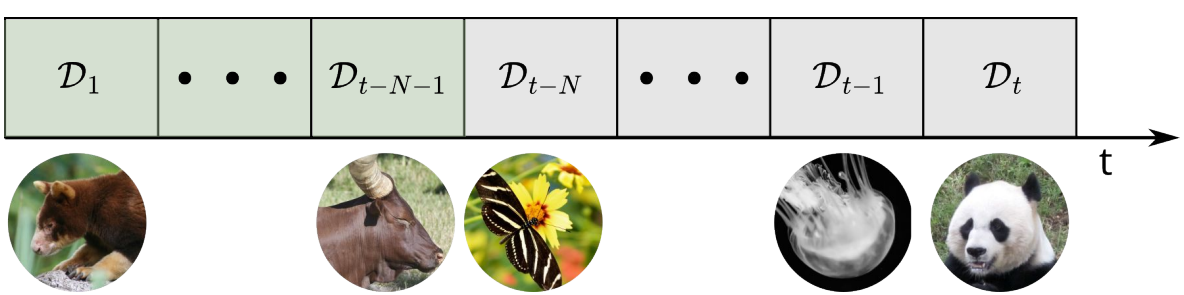
$$q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_1, \ldots, \mathcal{D}_t)$$

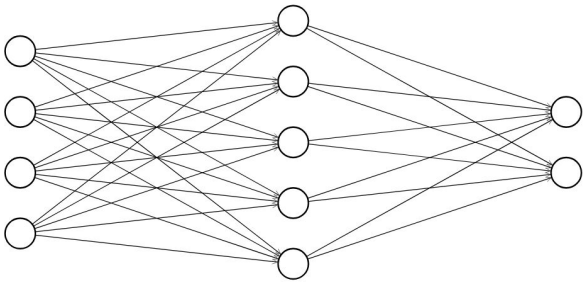
Kaushik, Prakhar, et al. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping."

Catastrophic remembering

Inability to learn induced by too much data

Using Bayesian neural networks adapted to lifelong learning, catastrophic forgetting is prevented.





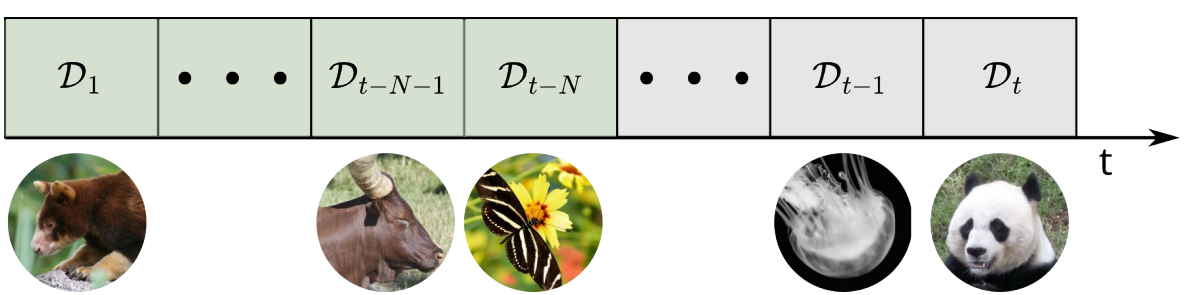
$$q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_1, \ldots, \mathcal{D}_t)$$

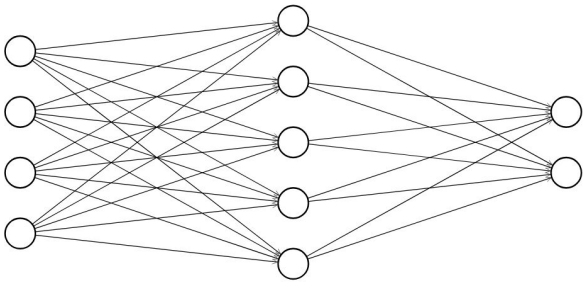
Kaushik, Prakhar, et al. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping."

Catastrophic remembering

Inability to learn induced by too much data

Using Bayesian neural networks adapted to lifelong learning, catastrophic forgetting is prevented.





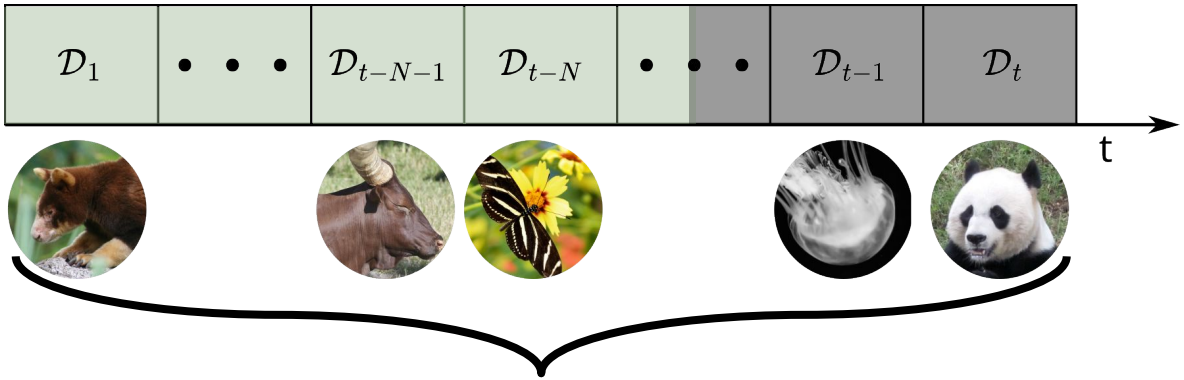
$$q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_1, \ldots, \mathcal{D}_t)$$

Kaushik, Prakhar, et al. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping."

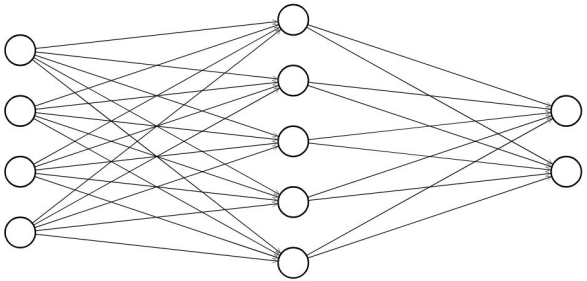
Catastrophic remembering

Inability to learn induced by too much data

Using Bayesian neural networks adapted to lifelong learning, catastrophic forgetting is prevented.



Catastrophic remembering



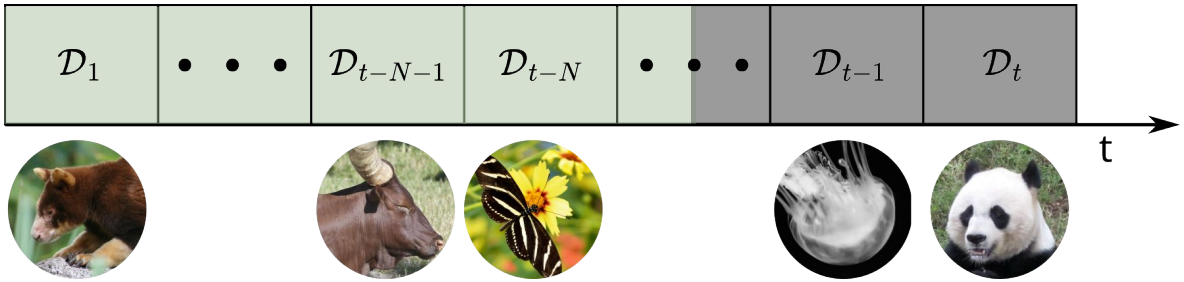
$$q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_1, \ldots, \mathcal{D}_t)$$

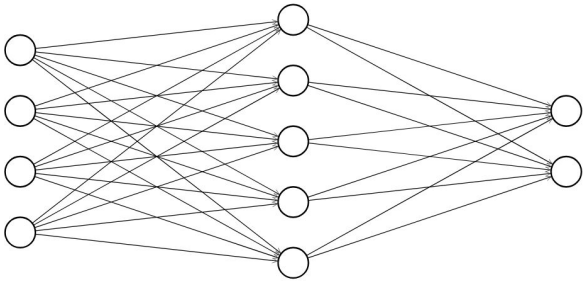
Kaushik, Prakhar, et al. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping."

Catastrophic remembering

Inability to learn induced by too much data

Using Bayesian neural networks adapted to lifelong learning, catastrophic forgetting is prevented.





$$q_{\boldsymbol{\theta}_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_1, \ldots, \mathcal{D}_t)$$

Kaushik, Prakhar, et al. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping."

However, remembering all datasets prevents learning

Research direction

Improving and furthering Bayesian neural networks

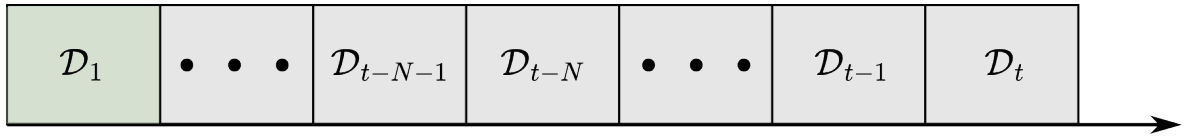
- **What triggers catastrophic remembering?**
- **How to maintain optimal discriminative capabilities through time?**
- **Is it possible to avoid both catastrophic remembering and catastrophic forgetting?**

Regulate forgetting and remembering through Metaplasticity from Synaptic Uncertainty

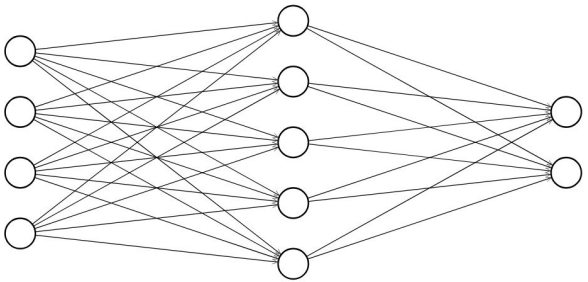
Bayesian continual learning and forgetting

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

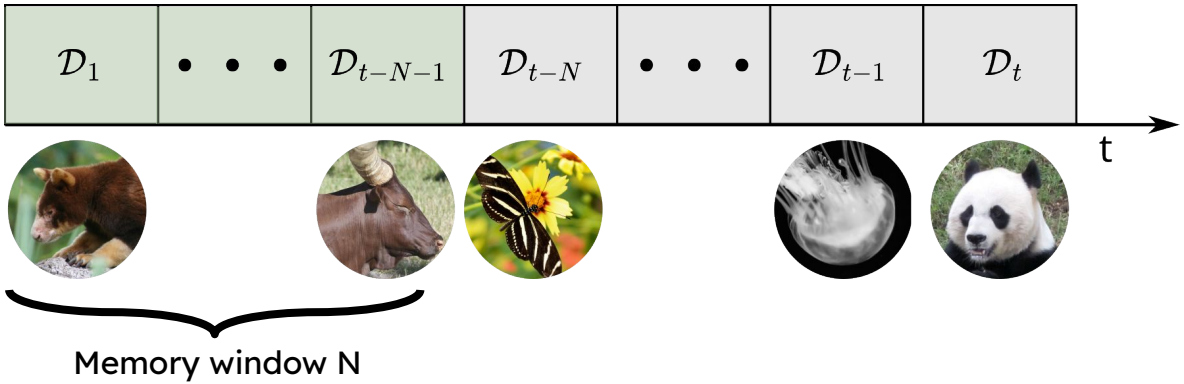


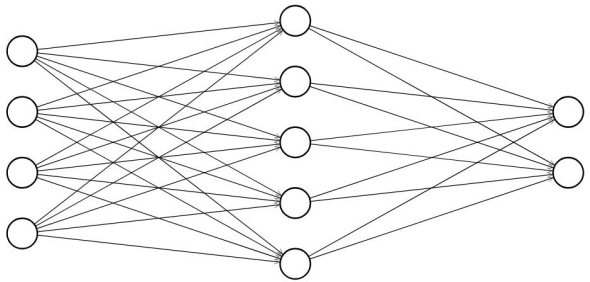
$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

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



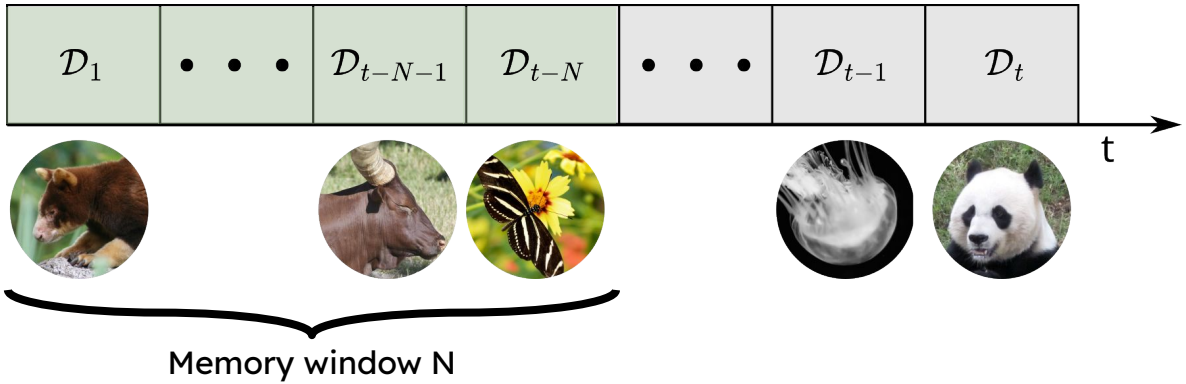


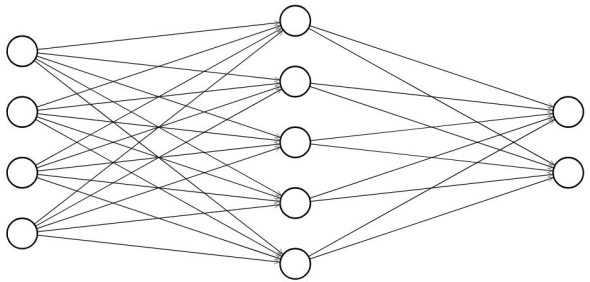
$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

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



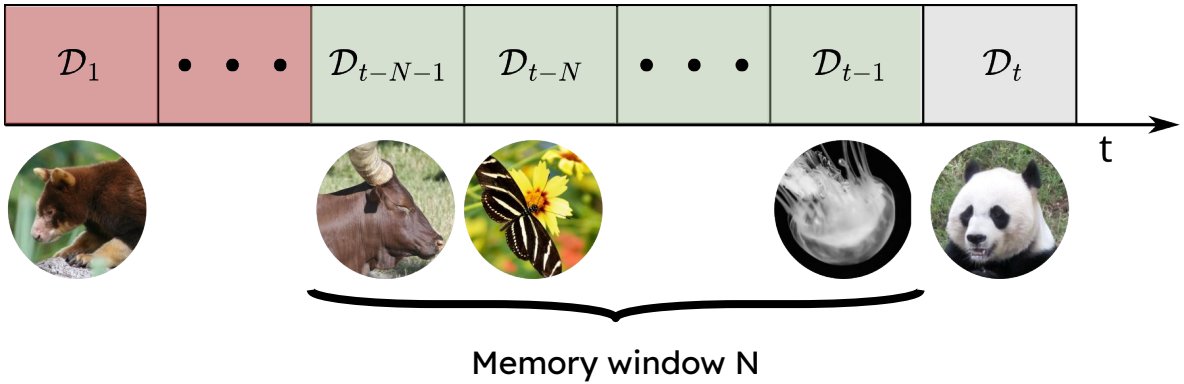


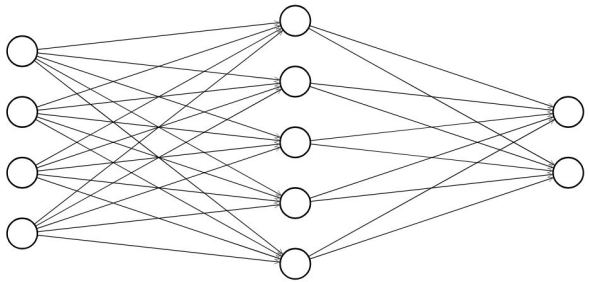
$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

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



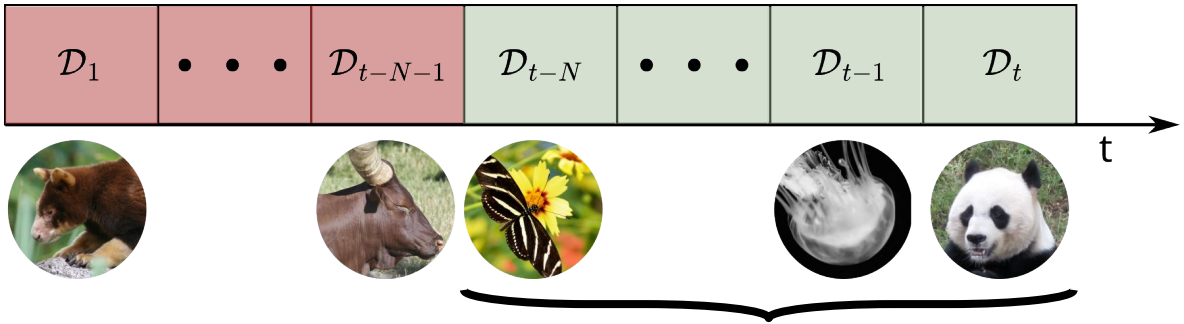


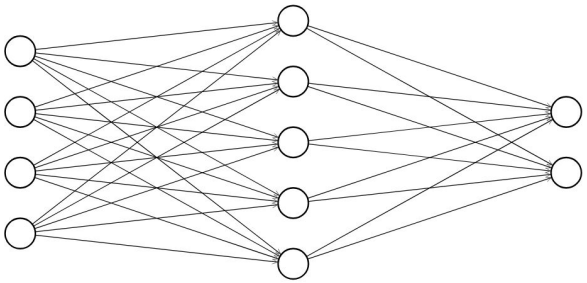
$$q_{\theta_t}(\boldsymbol{\omega}) \approx p(\boldsymbol{\omega}|\mathcal{D}_{t-N}, \ldots, \mathcal{D}_t)$$

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





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

Information out of the memory window is gradually forgotten

Bayesian continual learning and forgetting

Avoiding catastrophic remembering and vanishing uncertainties

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

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

Learning and forgetting is formalized through a **truncated posterior distribution**

Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

FOO-VB Diagonal

$$\Delta\boldsymbol{\mu} = -\boldsymbol{\sigma}_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}_{t-1}}$$

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

MESU

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

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

Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

FOO-VB Diagonal

$$\Delta\boldsymbol{\mu} = -\boldsymbol{\sigma}_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}_{t-1}} \longrightarrow \mathbf{0}$$

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

MESU

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

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

Metaplasticity from Synaptic Uncertainty

Synapse-wise regularization based on standard deviation

FOO-VB Diagonal

$$\Delta\boldsymbol{\mu} = -\boldsymbol{\sigma}_{t-1}^2 \frac{\partial \mathcal{C}_t}{\partial \boldsymbol{\mu}_{t-1}} \longrightarrow \mathbf{0}$$

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

MESU

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

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

Forgetting gradually bring synapses that have the highest standard deviation back to their prior

Main claims

What is MESU capable of doing?

- **MESU prevents catastrophic remembering and rigidity by gradually forgetting information**
- **MESU conserves OOD detection abilities through time, expressing high uncertainty**
- **MESU mitigates catastrophic forgetting within the memory window**

Experimental results

Estimating uncertainties

The neural network is a probability distribution

$$\begin{aligned} \mathcal{I}(\boldsymbol{\omega}, y \mid \mathcal{D}, x) &= H\big[p(y \mid x, \mathcal{D})\big] - \mathbb{E}_{p(\boldsymbol{\omega}\mid\mathcal{D})}\, H\big[p(y \mid x, \boldsymbol{\omega})\big], \\ \mathrm{EU} &= \mathrm{TU} - \mathrm{AU}. \end{aligned}$$

Estimating uncertainties

The neural network is a probability distribution

$$\begin{aligned} \mathcal{I}(\boldsymbol{\omega}, y \mid \mathcal{D}, x) &= H\big[p(y \mid x, \mathcal{D})\big] - \mathbb{E}_{p(\boldsymbol{\omega}\mid\mathcal{D})}\, H\big[p(y \mid x, \boldsymbol{\omega})\big], \\ \mathrm{EU} &= \mathrm{TU} - \mathrm{AU}. \end{aligned}$$

Epistemic

- Computes the mutual information, the **uncertainty of the model given the whole data**

Estimating uncertainties

The neural network is a probability distribution

$$\begin{aligned}\mathcal{I}(\boldsymbol{\omega}, y \mid \mathcal{D}, x) &= H\big[p(y \mid x, \mathcal{D})\big] - \mathbb{E}_{p(\boldsymbol{\omega}\mid\mathcal{D})}\, H\big[p(y \mid x, \boldsymbol{\omega})\big], \\ \text{EU} &= \text{TU} - \text{AU}.\end{aligned}$$

Epistemic

- Computes the mutual information, the **uncertainty of the model given the whole data**

Aleatoric

- Computes the expected **uncertainty of the output given the model and the input**

Estimating uncertainties

The neural network is a probability distribution

$$\begin{aligned}\mathcal{I}(\boldsymbol{\omega}, y \mid \mathcal{D}, x) &= H\big[p(y \mid x, \mathcal{D})\big] - \mathbb{E}_{p(\boldsymbol{\omega}\mid\mathcal{D})}\, H\big[p(y \mid x, \boldsymbol{\omega})\big], \\ \text{EU} &= \text{TU} - \text{AU}.\end{aligned}$$

Epistemic

- Computes the mutual information, the **uncertainty of the model given the whole data**

Aleatoric

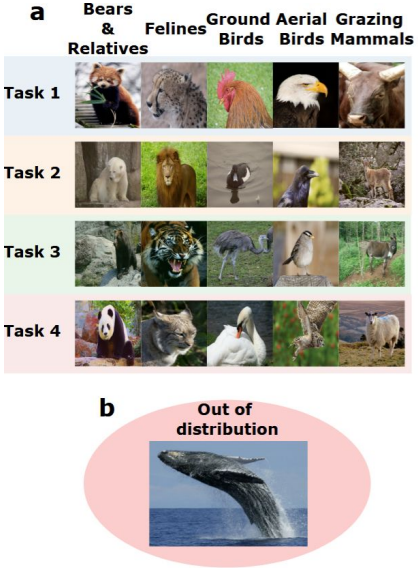
- Computes the expected **uncertainty of the output given the model and the input**

Smith, L. & Gal, Y. Understanding measures of uncertainty for adversarial example detection. (2018)

Bayesian neural networks generate both **aleatoric** and **epistemic uncertainties** to evaluate **out of distribution (OOD)** data

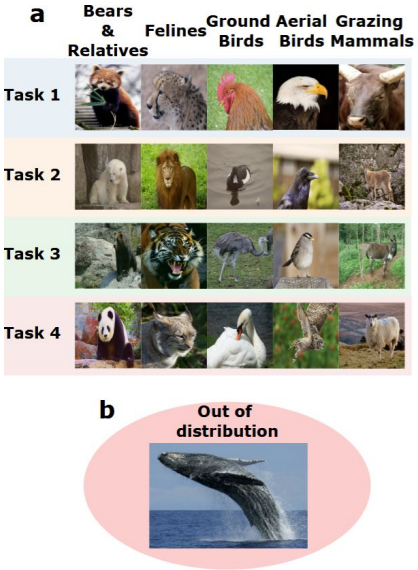
Mitigating catastrophic forgetting

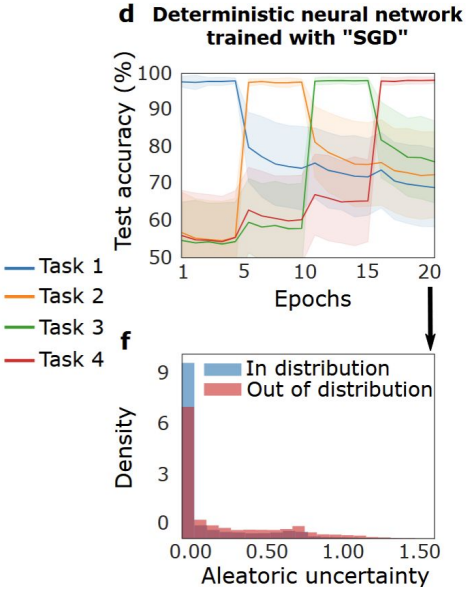
20 epochs of Animals dataset



Mitigating catastrophic forgetting

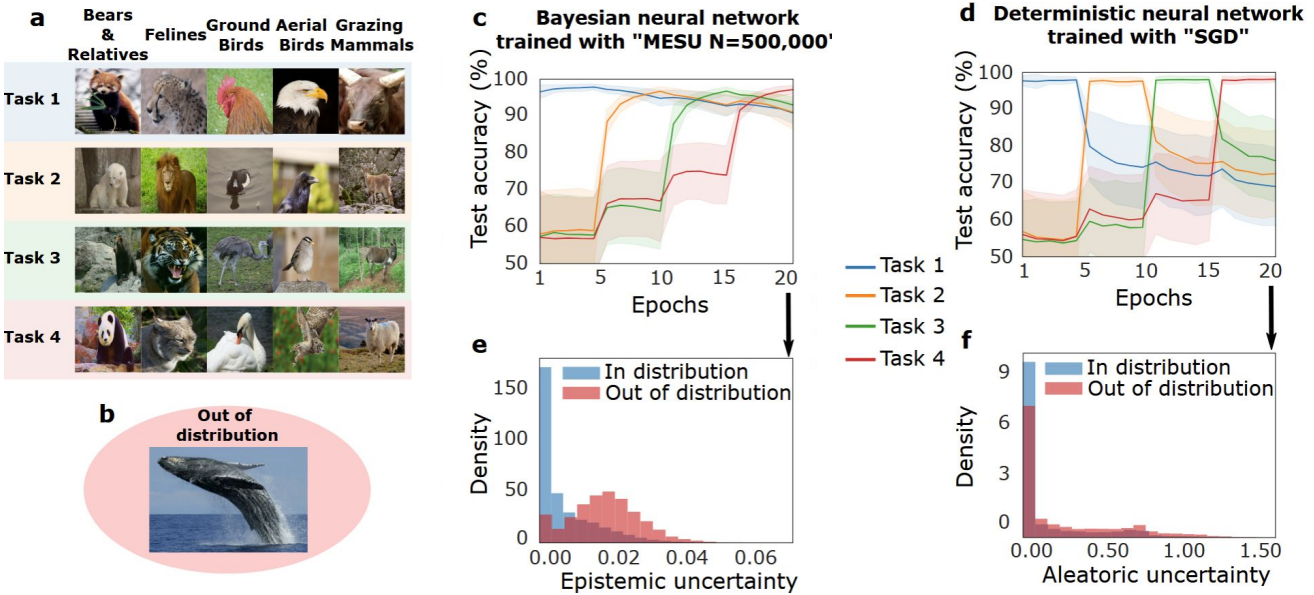
20 epochs of Animals dataset





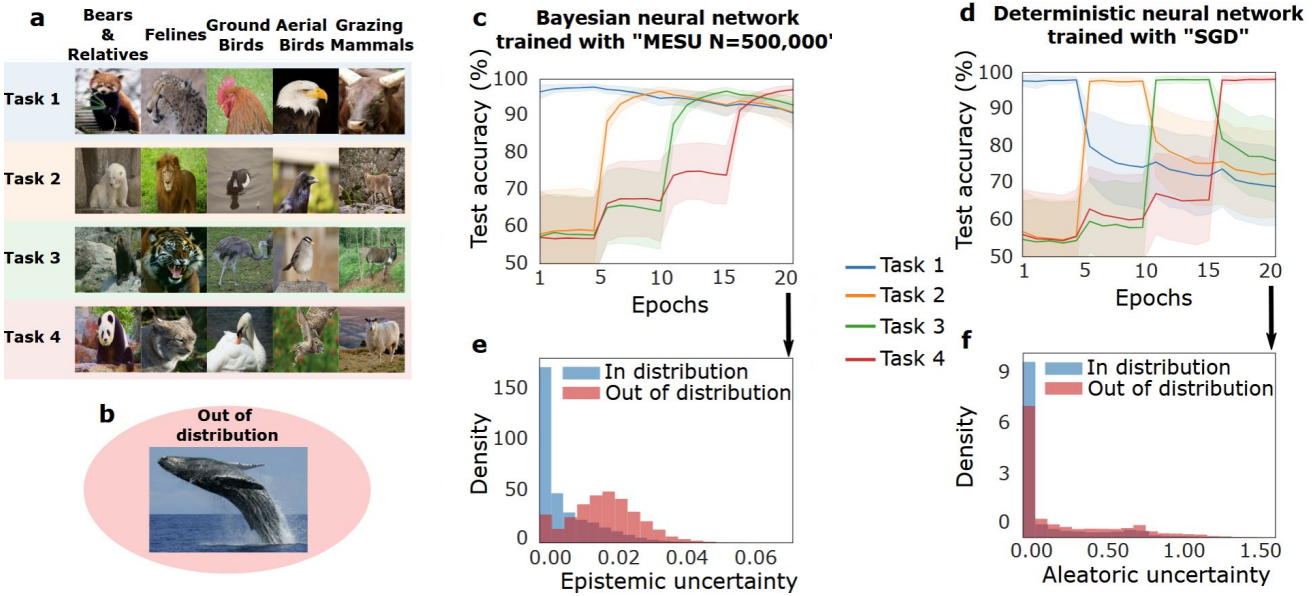
Mitigating catastrophic forgetting

20 epochs of Animals dataset



Mitigating catastrophic forgetting

20 epochs of Animals dataset



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

State-of-the-art

About task boundaries

Regularization | Optimization | Replay | Representation | Architecture

- Synaptic Intelligence

$$\tilde{L}_\mu = L_\mu + c\sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2$$

State-of-the-art

About task boundaries

Regularization | Optimization | Replay | Representation | Architecture

- Synaptic Intelligence
- Elastic Weight Consolidation

$$\tilde{L}_\mu = L_\mu + c\sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2$$

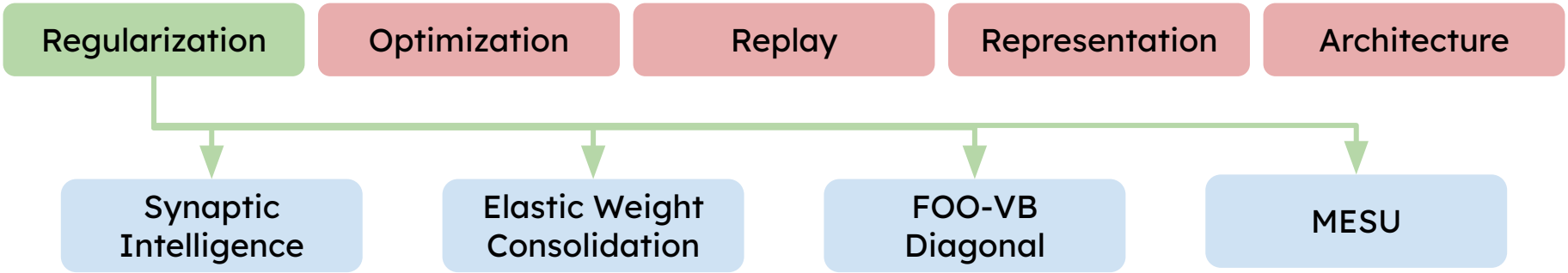
$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

Both SI and EWC methods require tasks boundaries to work, whereas FOO-VB Diagonal and MESU do not

Task boundaries: the algorithm knows ahead of time when the task changes, and **requires the previous parameters** in memory

State-of-the-art

About task boundaries



$$\tilde{L}_\mu = L_\mu + c\sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2 \qquad \mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

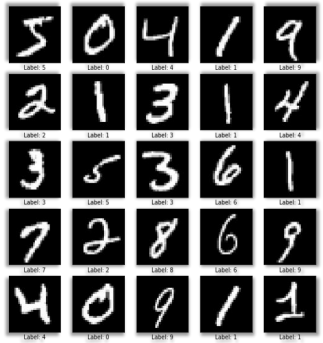
Both SI and EWC methods require tasks boundaries to work, whereas FOO-VB Diagonal and MESU do not

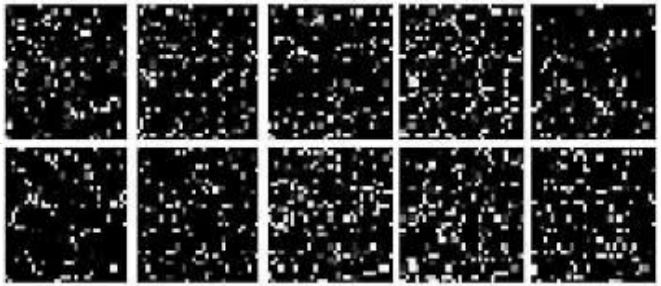
Task boundaries: the algorithm knows ahead of time when the task changes, and **requires the previous parameters** in memory

Permuted MNIST Benchmark

Permutation of pixels from MNIST

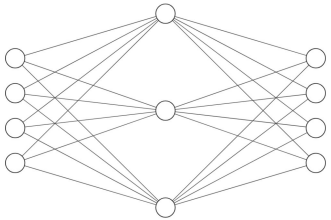
Simulation setup with a low amount of synapses in Online Continual Learning strategy





x200

Batch of 1



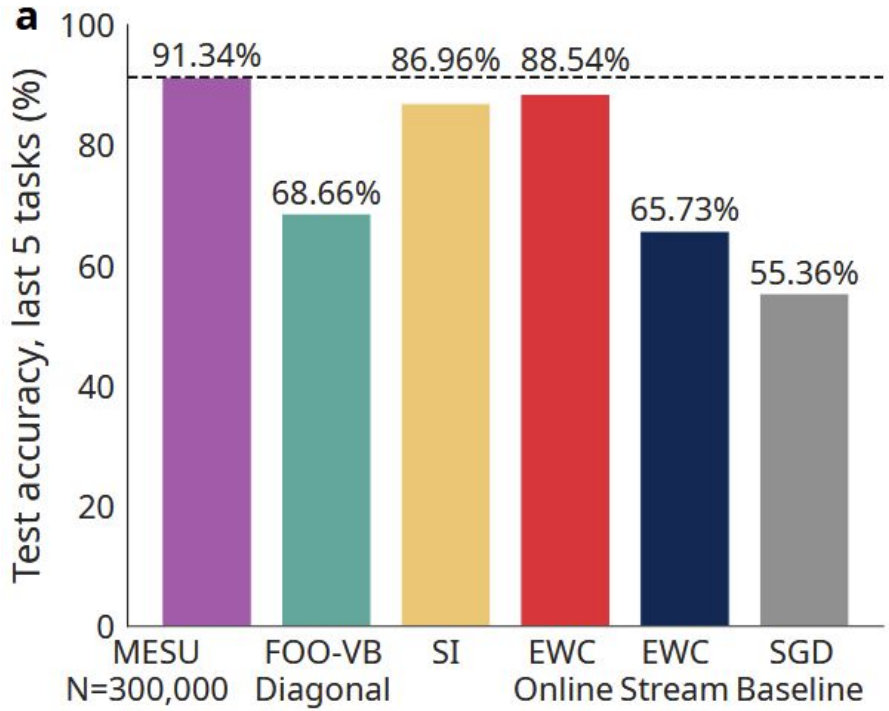
x10

(784, 50, 10)

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

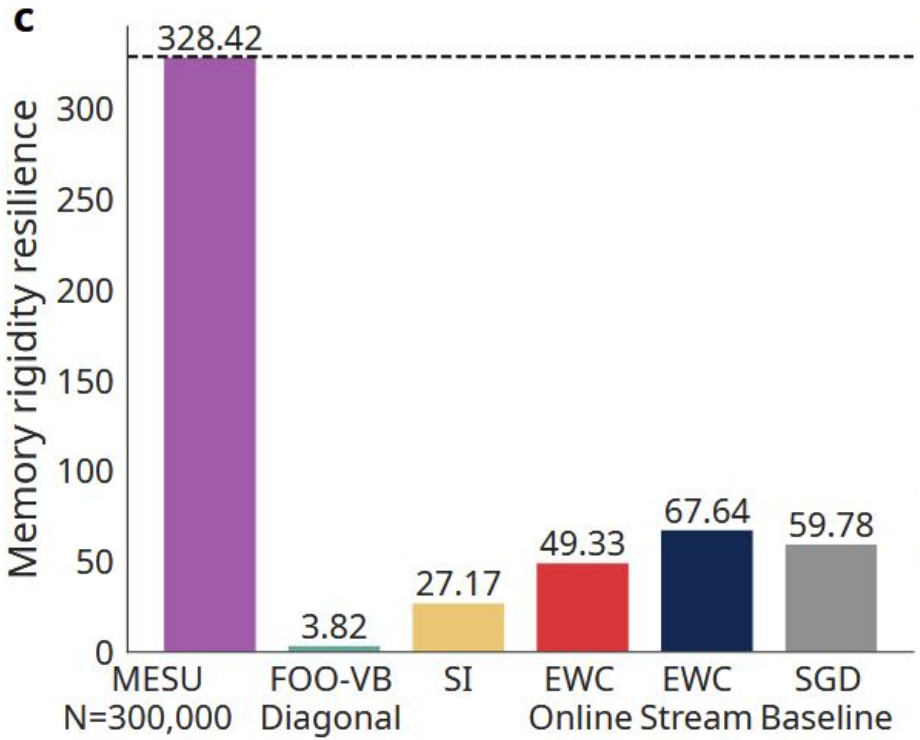


Stream: previous input
Online: previous task

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

Resilience to loss of plasticity

Increased performance due to the memory window



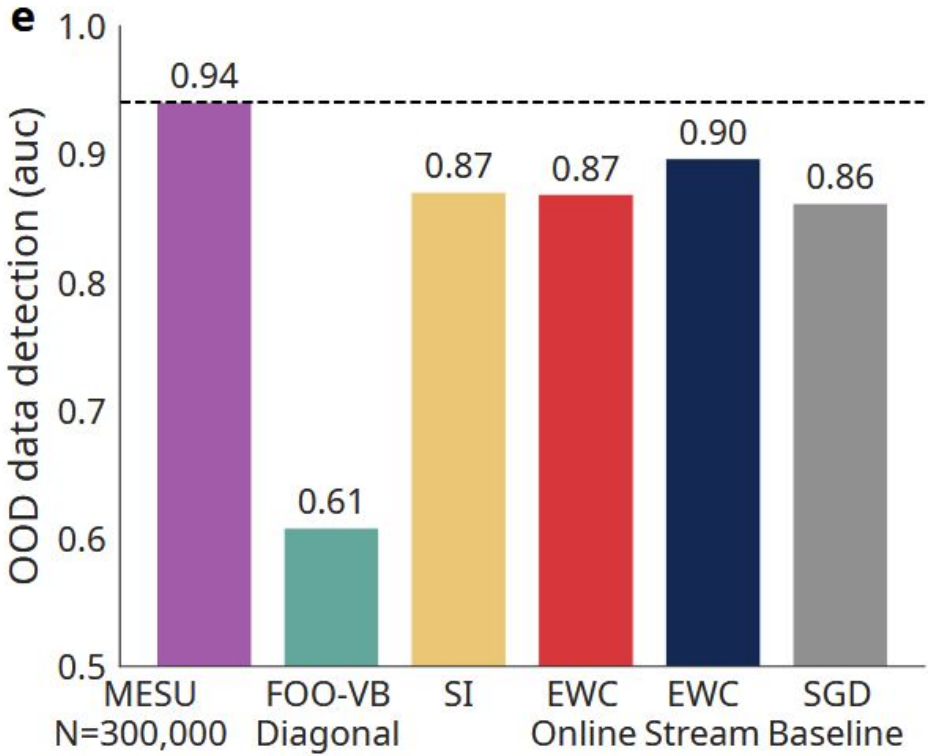
Stream: previous input
Online: previous task

Memory rigidity resilience between the last and the first task

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

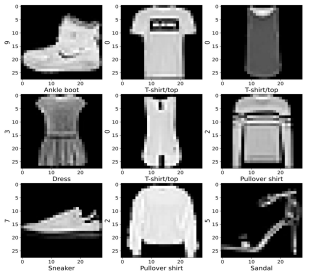
Out-of-distribution data detection

Increased performance due to the memory window



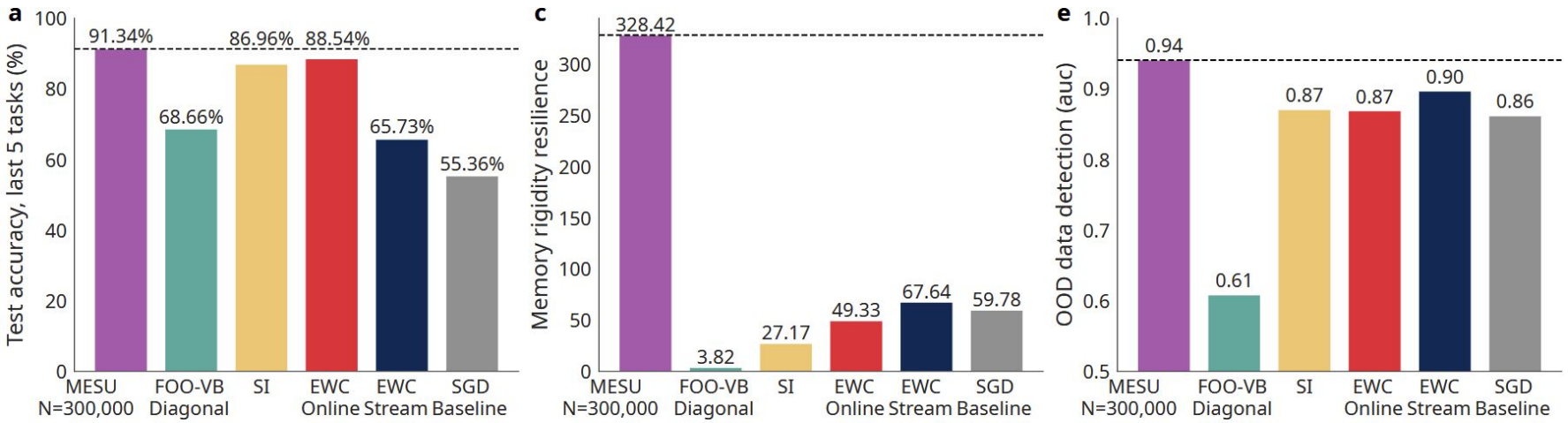
Stream: previous input
Online: previous task

OOD detection between ID distribution Permuted MNIST and OOD distribution Fashion-MNIST



Summary

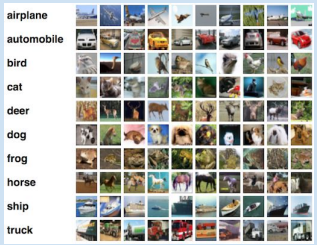
Increased performance due to the memory window



MESU is resilient against loss of plasticity, even after more than 200 tasks and outperforms methods with task-boundaries

CIFAR-10 / CIFAR-100 Benchmark

Domain-incremental learning



CIFAR-10

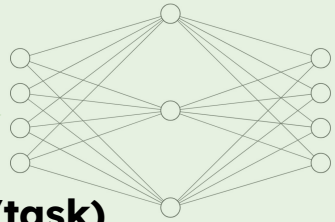
Task 1

CIFAR-100
(10 classes/task)

Task 2-11

Batch
of 200
60
epochs

CNN



(512, 10+10/task)

x8

1	2	3	4	5	6	7	8	9	10	11

2				3			
2-1	2-2	2-3	2-4	3-1	3-2	3-3	3-4

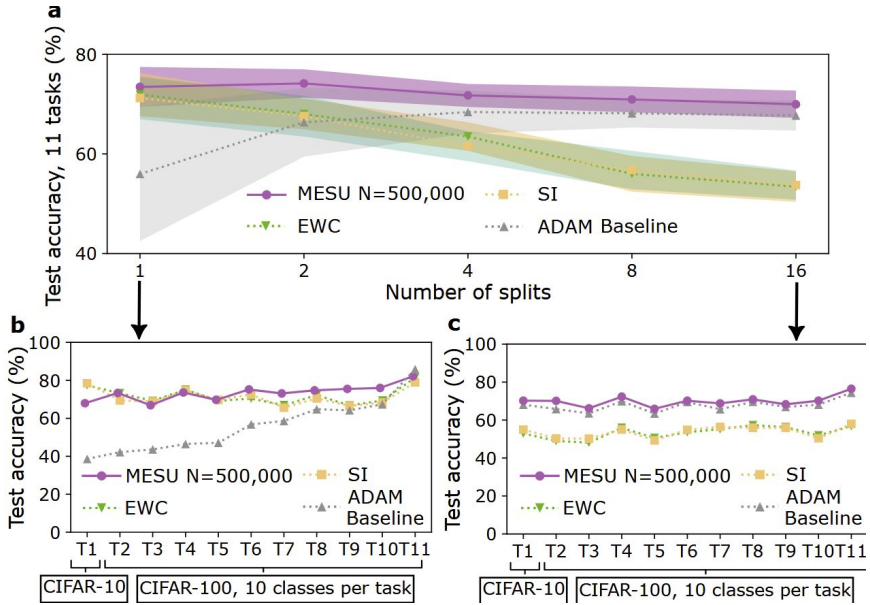
Interleaved

9-2	4-2	2-1	5-3	6-1	1-4	2-2	8-3

Tasks are then subdivided into 1, 2, 4, 8,
16 splits and interleaved

Beyond Multi-layer perceptrons

CIFAR-10 / CIFAR-100 splits



MESU outperforms methods with task-boundaries on end-to-end training with domain-incremental tasks

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

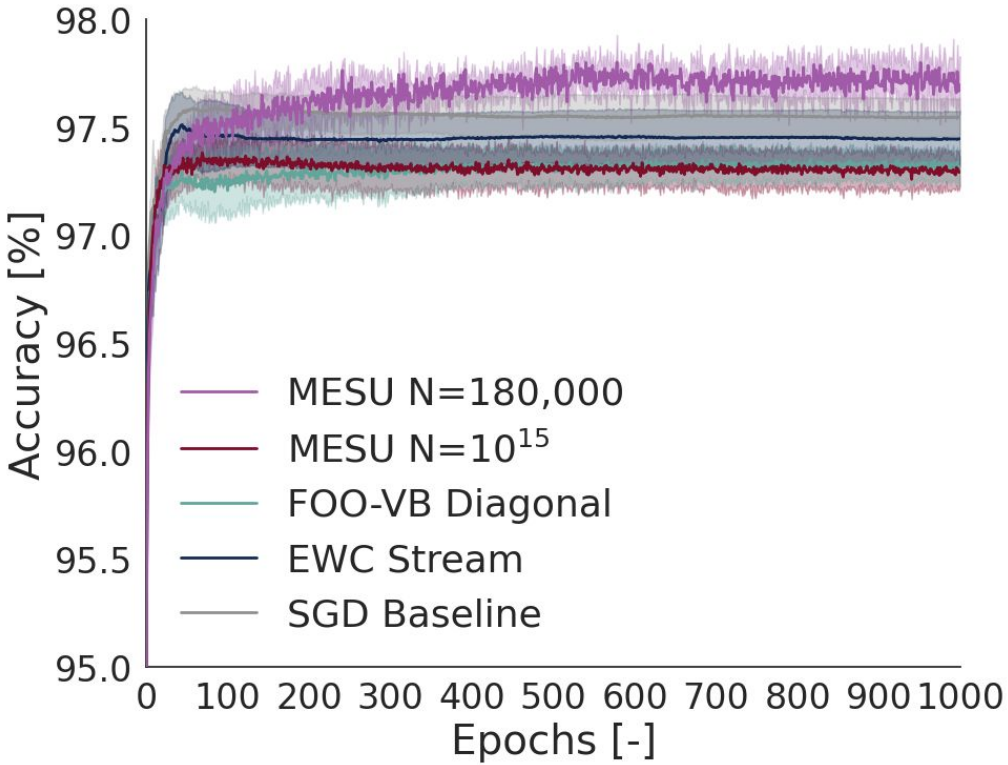
Thank you for listening!

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

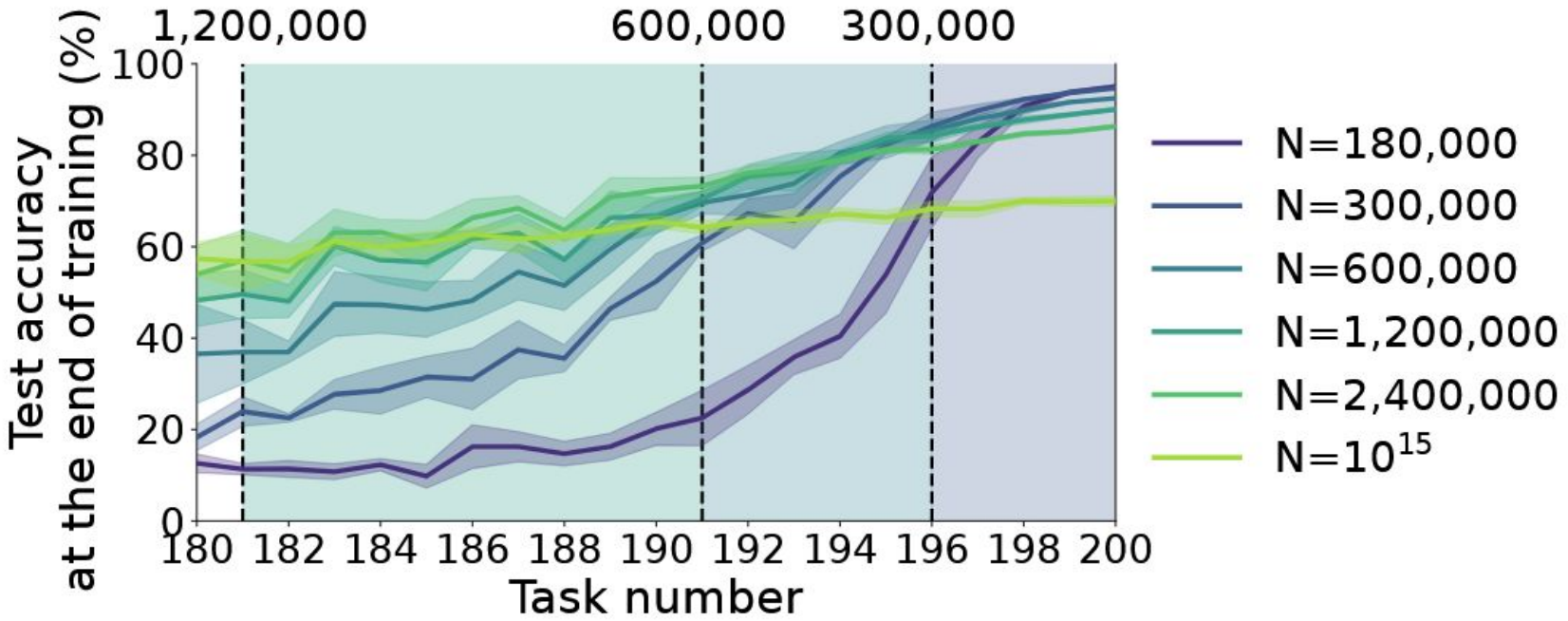
CIFAR10-100 CNN architecture

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.

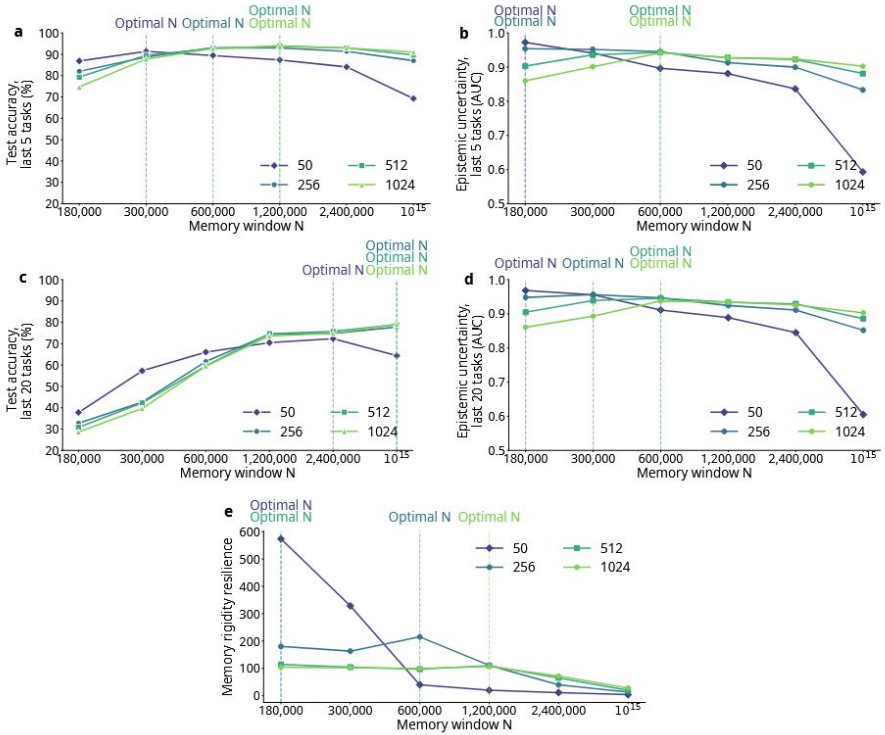
N ablation study

Effect of the memory window on performance



N ablation study

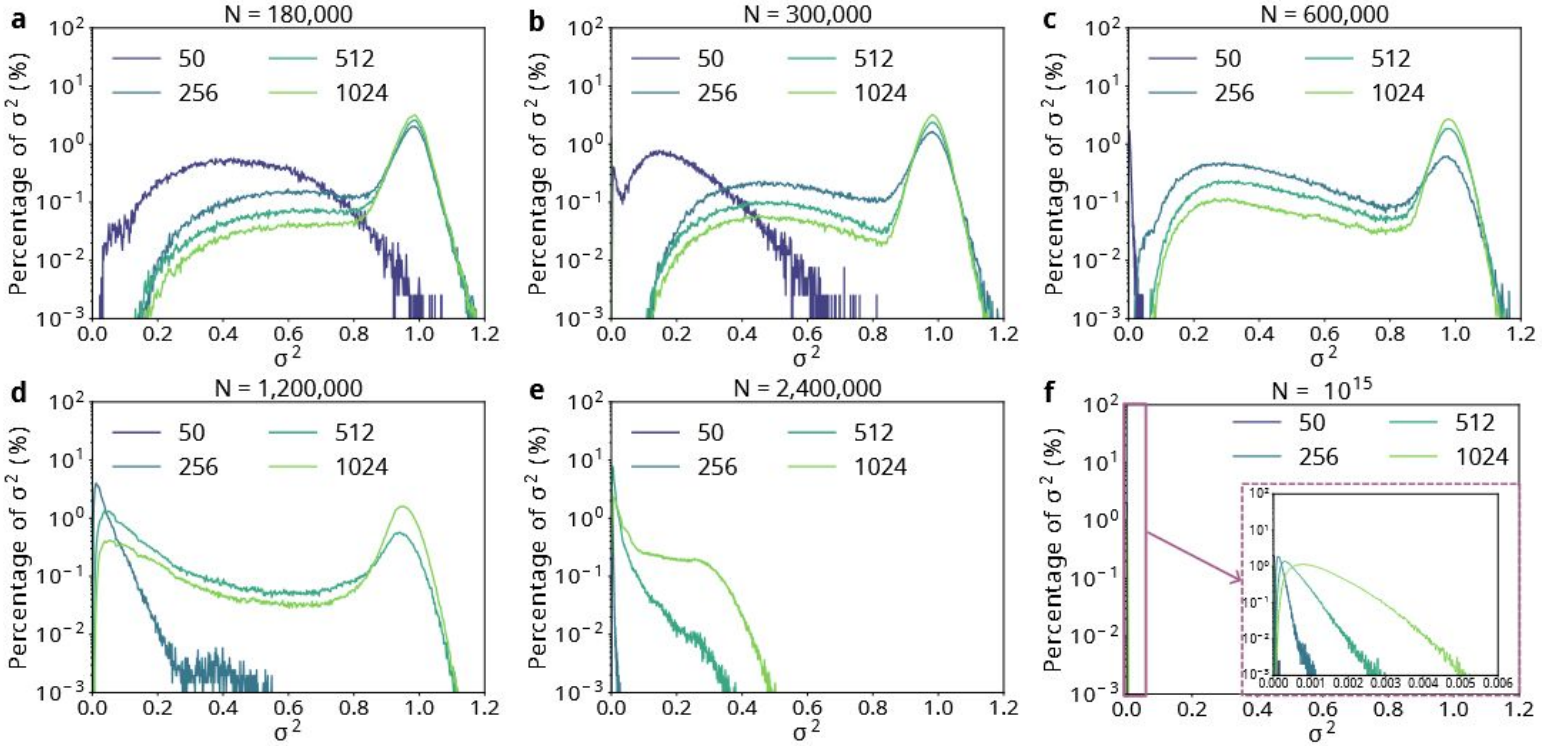
What is the optimal N value?



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