Active Continual Learning with Metaplastic Binary Bayesian Neural Networks

Kellian Cottart[1], *Théo Ballet*[1], *Djohan Bonnet*[1,2], *Damien Querlioz*[1]

[1]Université Paris-Saclay, CNRS, Centre de Nanosciences et de Nanotechnologies, Palaiseau, France [2]Forschungszentrum Jülich, Germany

1. CONTEXT: AI AT THE EXTREME EDGE

Endless streams from sensors

At the edge, deep neural networks face a core tension: AI must run **inference continuously**

Edge devices

Autonomous vehicles

Medical

Monitoring

Compute

Energy

Storage

Binary Neural Networks

Ultra-**fast**, ultra-**efficient inference** (bitwise MACs)

Weights +1, -1

Activations -1, +1

BiNN

XNOR POPCNT

Why is training at the edge challenging?

Streaming and non stationary data

- Input data distribution changes over time → Catastrophic forgetting and catastrophic remembering
- Epistemic uncertainty collapses over time due to weight saturation → Vanishing uncertainties
- Imbalanced data with rare, pathological events → Heavy missclassification on potentially critical classes

Our contributions

- **Binary Metaplasticity from Uncertainty (BiMU):** Three forces for **stable**, **plastic** and **forgetful** learning in Binary Bayesian Neural Networks
- **Maintaining uncertainty via metaplasticity:** synapses learn as a **function of variance and curvature**
- **Continual active learning:** Retaining **uncertainty over time and using disagreement to select informative training examples**

2. MODELING: BINARIZED BAYESIAN NEURAL NEWORKS

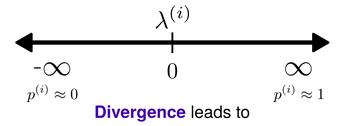
Each weight a Bernoulli random variable over {-1, 1}. We hold a mean-field posterior q over weights.

Bernoulli weights

$$\boldsymbol{\omega} \in \{-1,+1\}^s \quad | \quad q(\boldsymbol{\omega}|\boldsymbol{\lambda}) = \prod_{i=1}^{s} \frac{\exp\left(\lambda^{(i)}\omega^{(i)}\right)}{2\cosh\left(\lambda^{(i)}\right)}$$

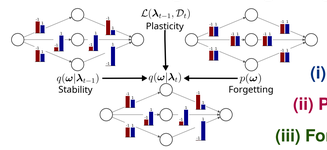
Natural Parameter

$$p^{(i)} = \text{sigmoid}\left(2\lambda^{(i)}\right)$$



Divergence leads to **deterministic behavior** (catastrophic forgetting)

Mean-field Variational Inference Continual learning and forgetting



(i) Stability

(ii) Plasticity

(iii) Forgetting

3. BiMU: LEARNING RULE FOR BINARY ONLINE LIFELONG LEARNING

Binary Metaplasticity from Uncertainty

$$\lambda_t^{(i)} = \lambda_{t-1}^{(i)} - \eta(\lambda_{t-1}^{(i)}) \left[\frac{\partial \mathcal{L}}{\partial \lambda^{(i)}}\Big|_{\lambda_{t-1}} + \frac{\lambda_{t-1}^{(i)} - \lambda_{\text{prior}}^{(i)}}{N \cosh^2(\lambda_{t-1}^{(i)})} \right]$$

BiMU updates the natural parameter of each weight in **online settings**

(i) Metaplastic learning rate (Stability)

- **Slows strong opposing gradient** to the most probable sign
- **Preserves non-consolidated** synapses

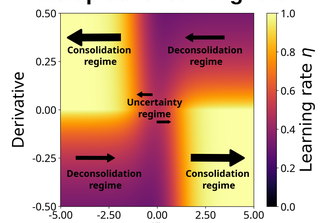
(ii) Gradient term (Plasticity)

- **Drives** the natural parameter **learning** (probability of sign)

(iii) Uncertainty-gated pull towards prior (Forgetting)

- **Sliding window N** controls the **forgetting**
- Synapses with **high variance** pull to the **prior**

Metaplastic learning rate



$$\frac{1}{\eta(\lambda_{t-1}^{(i)})} = \frac{1}{\cosh^2(\lambda_{t-1}^{(i)})} + 2\tanh(\lambda_{t-1}^{(i)}) \frac{\partial \mathcal{L}}{\partial \lambda^{(i)}}\Big|_{\lambda_{t-1}} + \frac{1}{\alpha_{\max}} + 2\left|\frac{\partial \mathcal{L}}{\partial \lambda^{(i)}}\Big|_{\lambda_{t-1}}\right|$$

4. METHODOLOGY FOR UNCERTAINTY QUANTIFICATION

Input data

$p(y|x,\boldsymbol{\omega}_1)$

$p(y|x,\boldsymbol{\omega}_2)$

K Monte Carlo samples

$p(y|x,\boldsymbol{\omega}_K)$

Epistemic = Predictive - Aleatoric

(Data & model uncertainty)

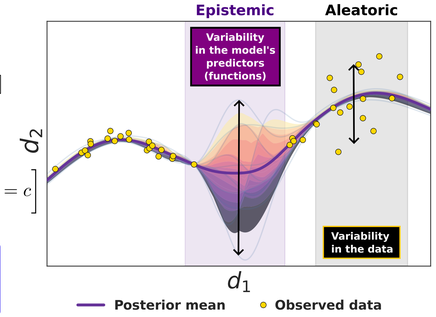
$$\mathcal{I}(\boldsymbol{\omega}, y|x) = H[p(y|x)] - \mathbb{E}_{q(\boldsymbol{\omega}|\boldsymbol{\lambda})}[H[p(y|x,\boldsymbol{\omega})]]$$

Variation ratio

(Disagreement between models)

$$\text{VR}(x) = 1 - \frac{1}{K} \max_{c \in \mathcal{C}} \sum_{k=1}^{K} \mathbb{I}\left[\arg\max_{c' \in \mathcal{C}} p(y = c'|x, \boldsymbol{\omega}_k) = c\right]$$

Forgetting preserves **uncertainties** in Binary Bayesian Neural Networks, allowing **variability detection**



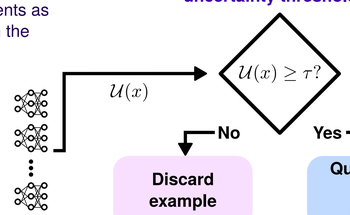
5. METHODOLOGY FOR ACTIVE LEARNING

Active learning at the edge

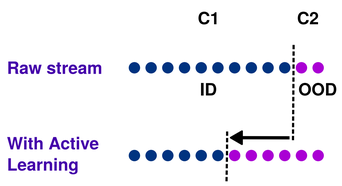
Intuition: uncertainty **catches rare** events as they deviate from the majority classes

Query labels from examples that exceed the **uncertainty threshold**

Input data



Why active learning helps with data imbalance?



C1 is in **majority** compared to C2

As training progresses, **C1** events are now **in distribution**

C2 events appear as **out of distribution**

Active learning queries more rare/uncertain examples, recovering informative events

6. EXPERIMENTS & RESULTS: 32x UPDATES REDUCTION ON LIFELONG LEARNING

A. Long-Horizon continual learning: mitigating forgetting and loss of plasticity (1000-tasks Permuted MNIST)

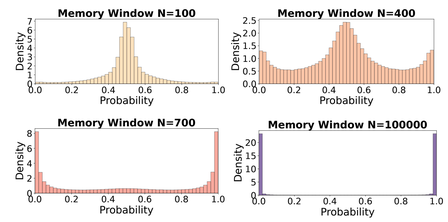
Method	Task bounds	MEAN ACC. 5 tasks (%)	OOD det. (AUC)	MMRR	Acc. 1 task (%)
Binary neural networks					
BiMU	NO	$\mathbf{90.30 \pm 0.38}$	$\mathbf{0.99 \pm 0.00}$	**139.47**	$\mathbf{94.67 \pm 0.11}$
BayesBiNN	YES	41.12 ± 1.62	0.57 ± 0.12	2.04	93.22 ± 0.09
Syn. Meta.	YES	10.27 ± 0.01	-	1.64	71.40 ± 1.48
STE	NO	29.35 ± 0.96	0.69 ± 0.04	9.32	77.56 ± 1.35
Real-valued neural networks					
MESU	NO	$\mathbf{91.69 \pm 0.58}$	$\mathbf{0.95 \pm 0.03}$	**261.10**	$\mathbf{96.10 \pm 0.18}$
EWC O.	YES	81.78 ± 0.82	0.66 ± 0.11	6.63	96.06 ± 0.11
SI	YES	74.41 ± 1.19	0.66 ± 0.17	5.11	95.55 ± 0.28
SGD	NO	66.64 ± 2.70	0.87 ± 0.05	43.50	96.03 ± 0.34

Protocol: 100 hidden units, fully online

- **BiMU** is **stable (high mean accuracy), plastic (high OOD detection), forgetful (high MMRR)**
- **Uncertainties** are maintained over **1000 tasks**
- **Memory window N** enforces **forgetting or stability**

In distribution

Out of distribution

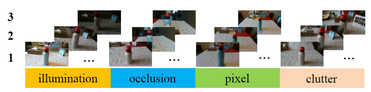


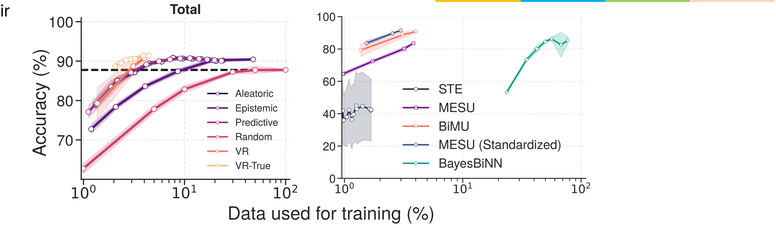
B. Imbalanced active continual learning: balancing classes, reducing annotations (OpenLORIS-Object)

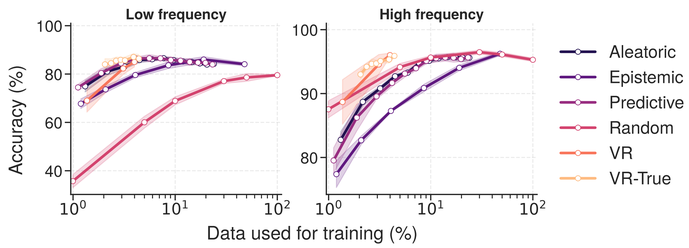
Protocol: Frozen features, linear classifier, fully online

Classes with less than 700 training examples have their training images removed up to 50-80%

- **Variation Ratio (VR)** active learning enables a **32x reduction in updates**
- **Outperforms full training** with only **4.0%** of the budget
- **BiMU (Binary)** performs **similarly** to **MESU (Real-valued)**







- **Low frequency accuracy drops significantly when using active learning with random examples**
- VR maintains the accuracy of **both high frequency classes** and **low frequency classes**
- At the 32x reduction, BiMU with VR recovers **28% accuracy** on the **low frequency classes**

Takeaway: BiMU allows efficient continual learning at the edge by preserving Bernoulli distribution uncertainties, using fast inference to compute disagreement, leading to selective updates

[1] Hubara, Itay, et al. "Binarized neural networks." Advances in neural information processing systems 29 (2016).
[2] Meng, Xiangming, et al. "Training binary neural networks using the Bayesian learning rule.Proceedings of the 37th International Conference on Machine Learning. 2020
[3] 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)
[4] Rajendran, Prajit T., et al. "Stream-Based Active Learning with Adaptive Uncertainty and Diversity Thresholds." 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2023.
[5] Bonnet, Djohan, et al. "Bayesian continual learning and forgetting in neural networks." Nature Communications 16.1 (2025): 9614.