Active Continual Learning with Metaplastic Binary Bayesian Neural Networks

Kellian Cottart, Théo Ballet, Djohan Bonnet, Damien Querlioz

Submitted to International Conference in Machine Learning (ICML)

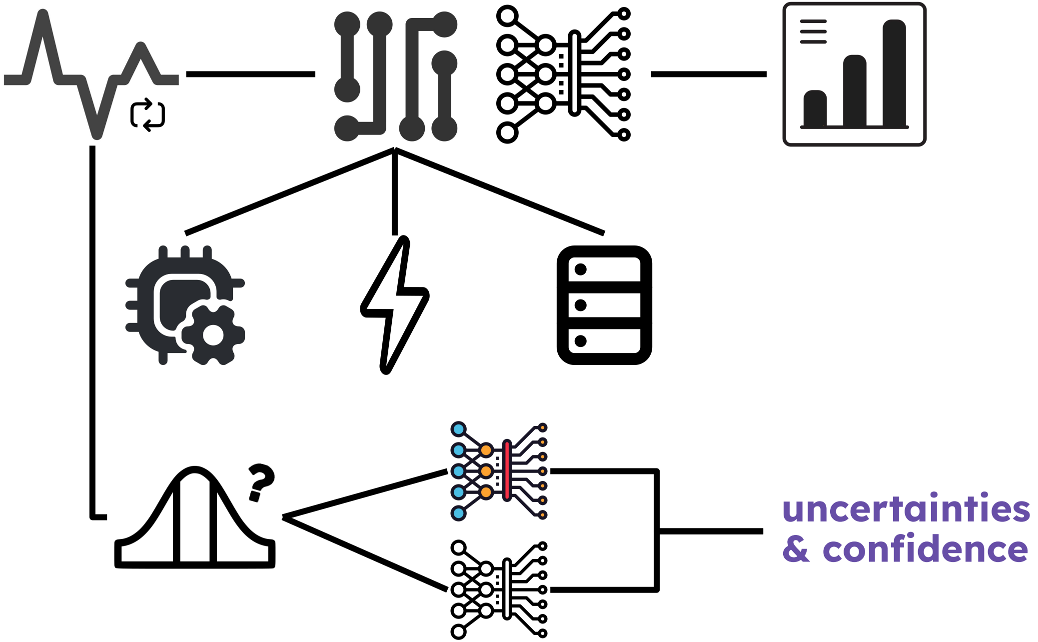
Introduction

AI systems at the Edge

Always on-edge systems must run inference **continuously:**
Data stream is **non-stationary** and potentially **unbalanced**

They face **computing power, energy and storage limitations:**
Weight updates are **expensive**

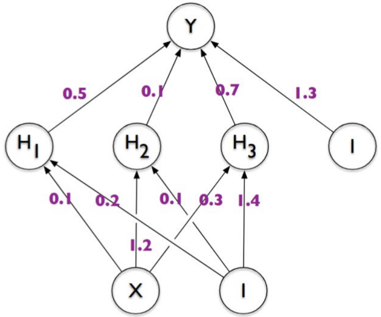
Safety critical systems need uncertainties
Bayesian Deep Learning offers ways to quantify **out-of-distribution data**

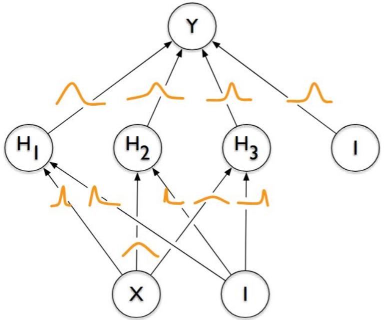


this talk: binary Bayesian neural networks at the edge are a solution for these limitations with **uncertainties and metaplasticity**

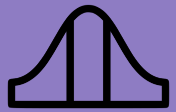
Bayesian neural networks

Equipping neural networks with a standard deviation

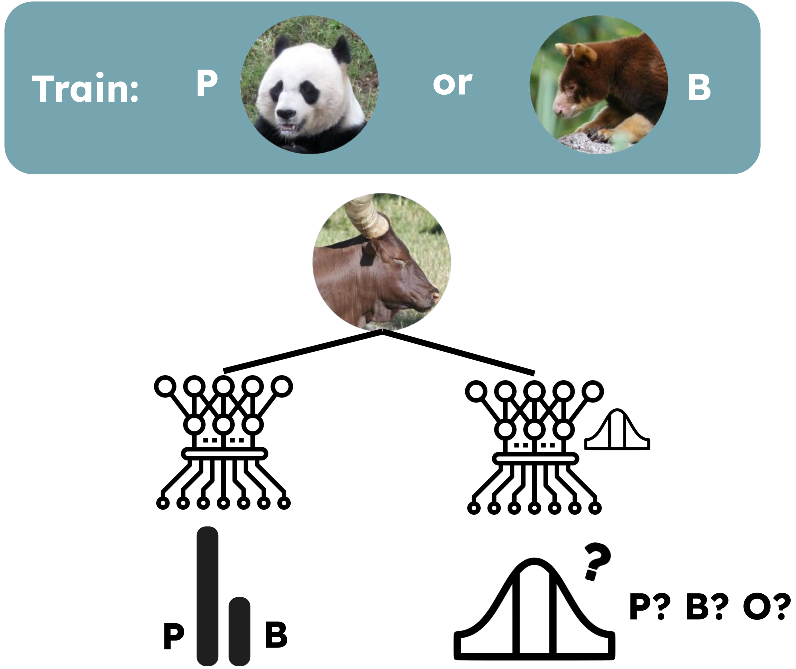




point estimate



random distribution



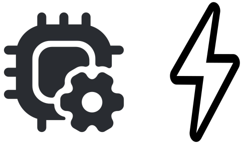
out-of-distribution detection

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

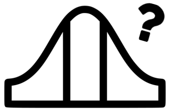
Disagreement-based Active Learning

Synapses carry previous activities, they know what they do not know

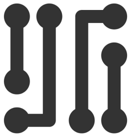
Labels or update steps are **expensive**
Device **endurance**, **energy** consumption, **computing** power

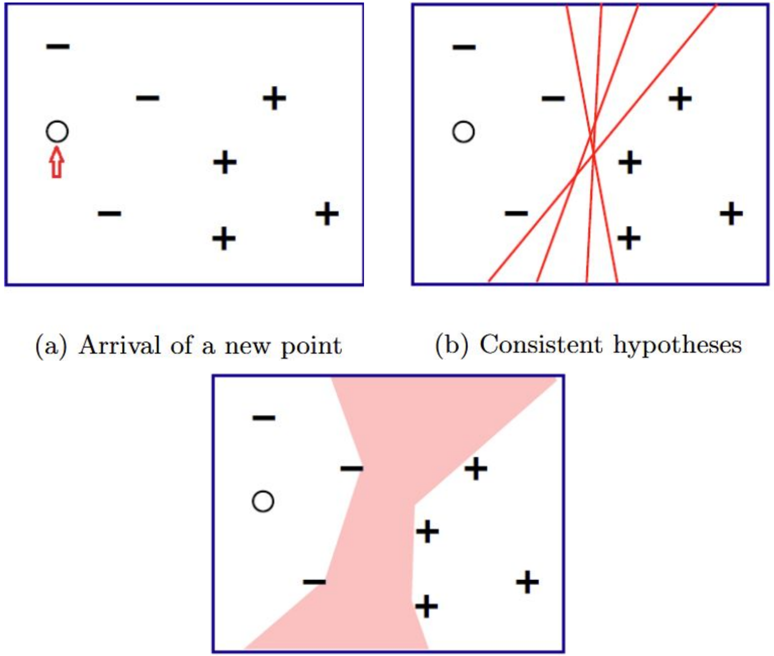


Learning efficiently **by selecting** highly relevant **examples**
Epistemic disagreement in the distribution



Above an uncertainty **threshold,** examples are labeled and used for training
Hardware-friendly examples selection





(a) Arrival of a new point

(b) Consistent hypotheses

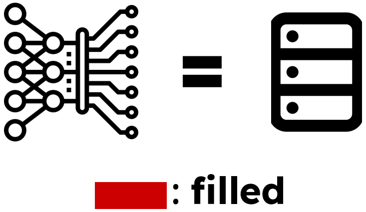
(c) Disagreement region

Epistemic disagreement learning

Catastrophic forgetting & remembering

How do you learn on endless data?

The system is **finite**. Weights are task **memories**. Learning is a **spectrum.**



Catastrophic Forgetting
(Forgetting everything / No continual learning abilities)

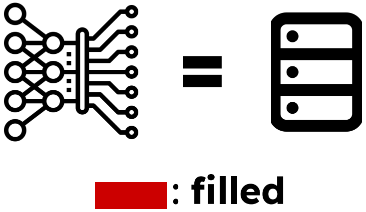
Gradual forgetting
(Gradually forgetting / Selecting relevant information)

Catastrophic Remembering
(Remember everything/ Inability to learn)

Catastrophic forgetting & remembering

How do you learn on endless data?

The system is **finite**. Weights are task **memories**. Learning is a **spectrum.**



Forgetting	Stability/Plasticity Trade-off	Stability

Catastrophic Forgetting
(Forgetting everything / No continual learning abilities)

Gradual forgetting
(Gradually forgetting / Selecting relevant information)

Catastrophic Remembering
(Remember everything/ Inability to learn)

Bayesian continual learning & forgetting

Uncertainties as a mean to know what to forget

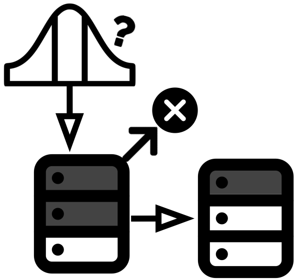
Bayes' Rule can be rewritten for **gradual forgetting**.

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

Updates steps are controlled with **uncertainty**. Uncertain information is **gradually forgotten**.

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

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



synapse's previous activities determines its current plasticity

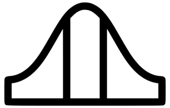
Metaplasticity

Bonnet, Djohan, et al. "Bayesian continual learning and forgetting in neural networks." *Nature Communications* 16.1 (2025): 9614.

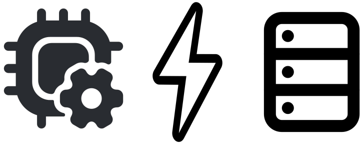
Binary Bayesian neural networks

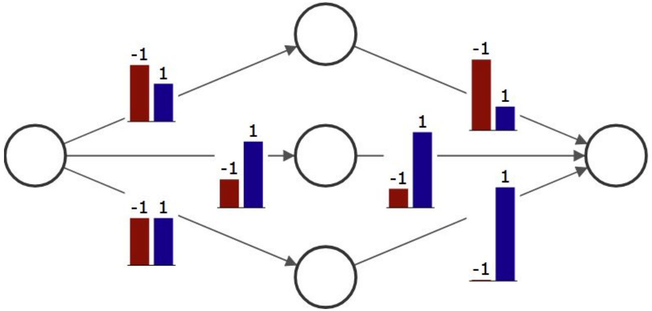
Efficient, work for CL, but prone to plasticity loss

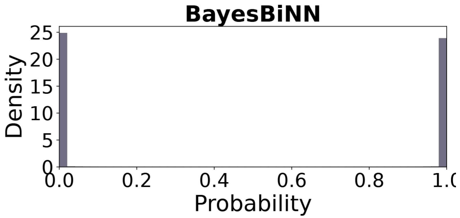
Each synapse models the probability p of a **Bernoulli distribution**

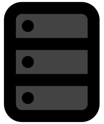


With **weights** on a one bit representation, **inference** is drastically **sped up**, **energy efficiency is improved**
Cheap Monte Carlo samples can be drawn









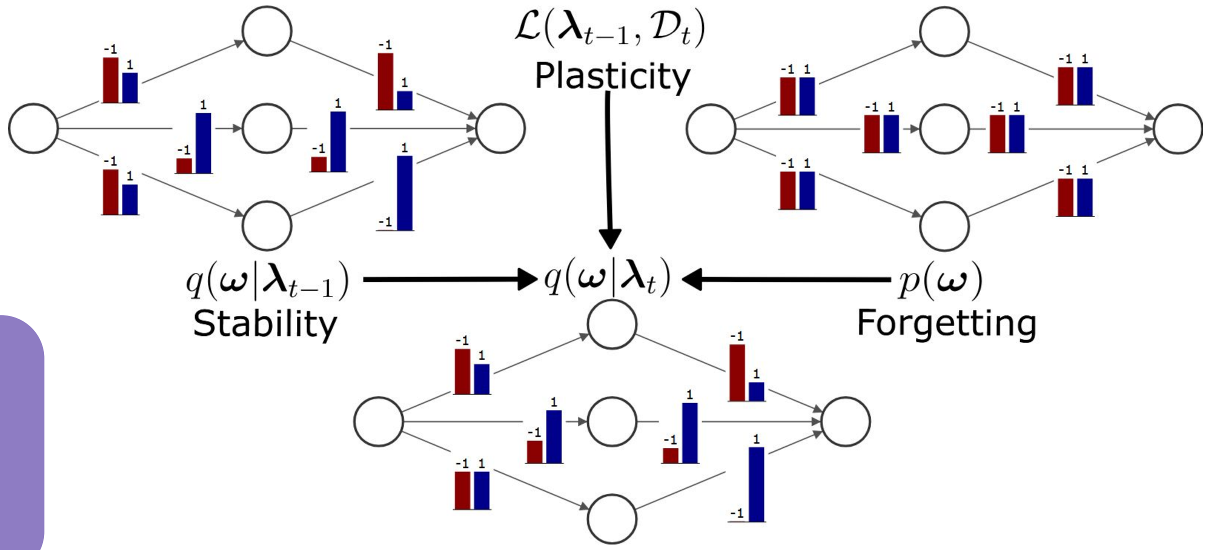
Network becomes **deterministic**, it cannot **learn** or **discriminate** continuously **anymore**

Synapse degeneracy occurs after training due to accumulated plasticity loss inducing inability to learn

Meng, Xiangming, Roman Bachmann, and Mohammad Emtiyaz Khan. "Training binary neural networks using the bayesian learning rule." *International conference on machine learning*. PMLR, 2020.

Binary Metaplasticity from Uncertainties

Gradually forgets to learn continually, uses its remembered knowledge to discriminate unseen data



Proposition 2. (Variational free-energy decomposition under controlled forgetting).

Implementing forgetting in binary setting

Uncertainties to guide learning of binary synapses

Forgetting is modulated by the **variance** of the Bernoulli, and both learning and forgetting are stabilized through a **metaplastic factor**

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

Theorem 1. (Memory-limited second-order asymmetric update for binary synapses)

Updates steps are controlled with **uncertainty**, and with a **surrogate curvature of the loss**

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

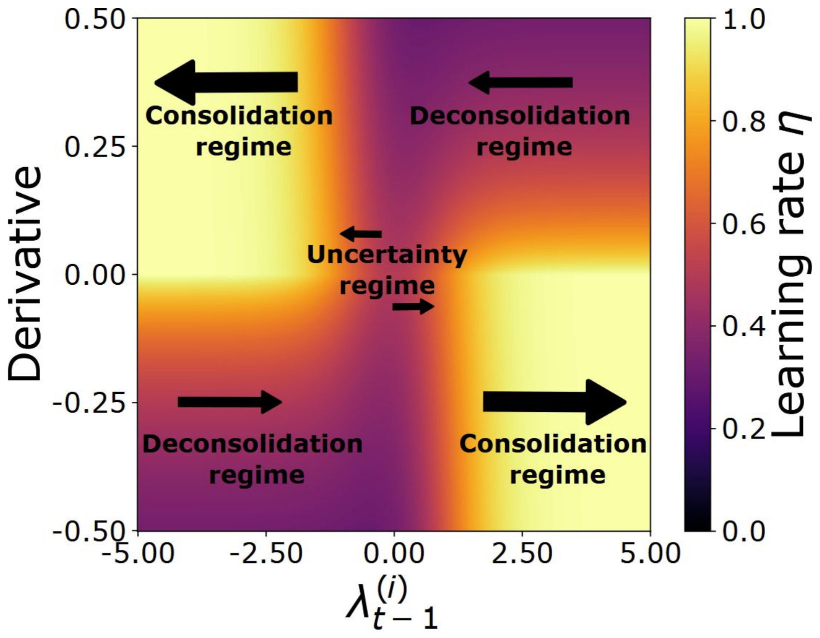
Proposition 4. (Bounded Learning Rate with a Curvature Surrogate).

Each synapse is modulated by its own uncertainty and contribution

Metaplasticity

Dynamics of the metaplastic synapse

Study of the learning rate modulated learning and forgetting



Depending on the value of the **derivative** and the current **synapse**:

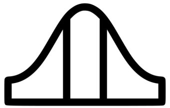
- **Uncertainty regime**: ***smallest*** learning rate, the weight is unsure about its sign.
- **Deconsolidation regime**: ***small*** learning rate to avoid flipping the sign unless decreasing certainty is repeatedly required.
- **Consolidation regime**: ***normal*** learning rate, consolidates the information and stores it long-term.

Information is retained through synaptic uncertainty: the more certain it is, the stronger it is kept

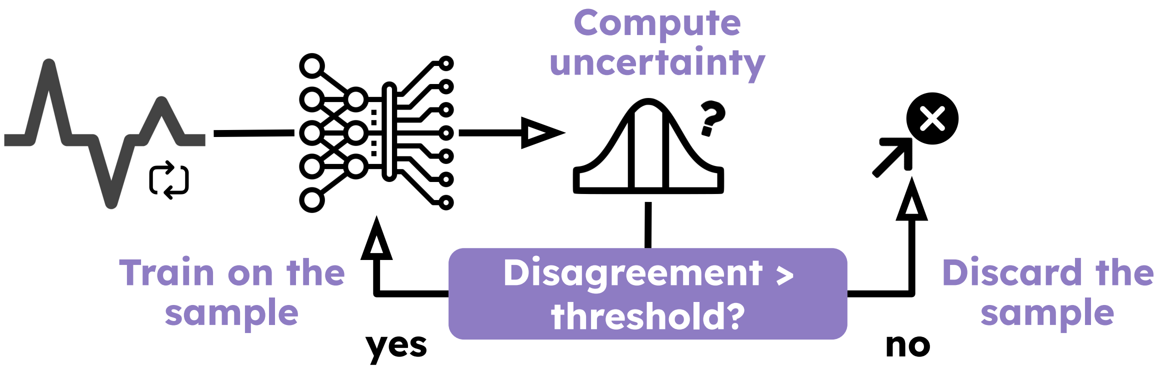
Uncertainty-based thresholding

Selecting informative examples to train

BiMU springs **epistemic disagreement**. We train if and only if the seen example exceed a given valued **threshold**.



Inference is cheap
Monte Carlo samples are used to compute **uncertainty**



With gradual forgetting, BiMU preserves uncertainties through time.

Huge savings in energy and computing, we only train on informative examples

Schwarz, Jonathan, et al. "Progress & compress: A scalable framework for continual learning." International conference on machine learning. PMLR, 2018.
Zenke, Friedemann, Ben Poole, and Surya Ganguli. "Continual learning through synaptic intelligence." International conference on machine learning. PMLR, 2017.
Laborieux, Axel, et al. "Synaptic metaplasticity in binarized neural networks." Nature communications 12.1 (2021): 2549.
Bengio, Yoshua, et al. "Estimating or propagating gradients through stochastic neurons for conditional computation." arXiv preprint arXiv:1308.3432 (2013).

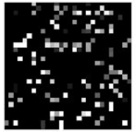
Empirical results on Continual Learning

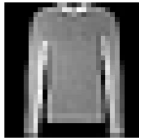
Maximum Memory Rigidity Resilience MMRR	Average accuracy (N tasks)	Out-of-distribution (OOD) ROC AUC
Measures the available plasticity	Measures the stability on the last N tasks	Measures data distinguishability

	Bayesian CL		CL Baselines	Non CL-Baselines
Binary	BiMU	BayesBiNN	Synaptic Metaplasticity	STE
Real	MESU	Online Elastic Weight Consolidation (EWC)	Synaptic Intelligence (SI)	SGD

Very long-horizon learning

Study of 1000-tasks Permuted MNIST

ID: 

OOD: 

Method	Task bounds	Mean Acc. 5 tasks (%)	OOD det. (AUC)	MMRR	Acc. 1 task (%)
Binary neural networks					
BiMU	no	**90.30 ± 0.38**	**0.99 ± 0.00**	**139.67**	**94.67 ± 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	**91.69 ± 0.58**	**0.95 ± 0.03**	**261.79**	**96.10 ± 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	43.52	96.03 ± 0.34

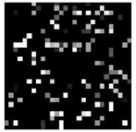
BiMU retains well information for very long term without losing plasticity
Mean ACC 5 tasks and **MMRR are very high** without task boundaries

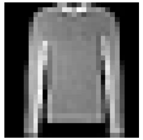
BiMU discriminates well OOD data
OOD detection is near-**perfect**

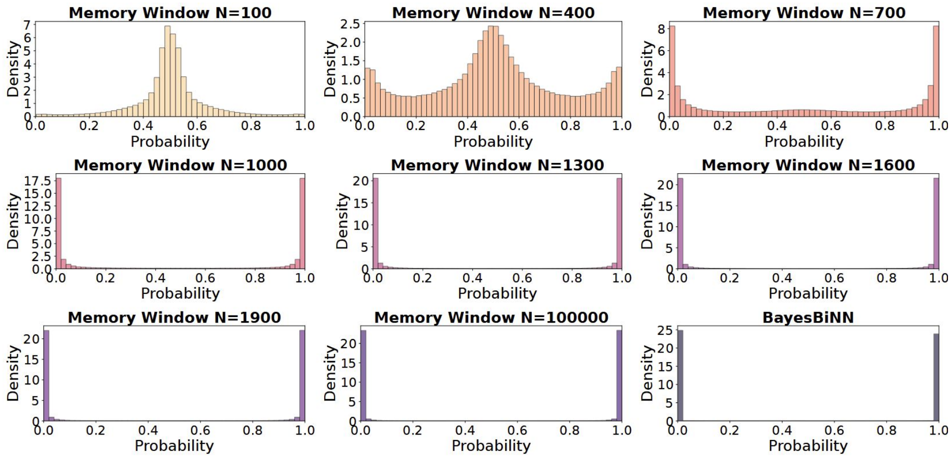
BiMU learns better than regular binary neural networks
ACC 1 task is the **highest** of all binary methods

Very long-horizon learning

Ablation study of N

ID: 

OOD: 



MEMORY WINDOW N	MEAN ACC. 5 TASKS (%)	OOD DET. (AUC)	MMRR
100	27.09 ± 0.57	0.99 ± 0.00	216.45
400	73.46 ± 2.59	$\mathbf{1.00 \pm 0.00}$	**574.72**
700	$\mathbf{90.29 \pm 0.24}$	0.99 ± 0.01	215.52
1000	89.07 ± 0.28	0.98 ± 0.01	46.30
1300	87.67 ± 0.33	0.97 ± 0.02	28.36
1600	86.94 ± 0.29	0.96 ± 0.02	24.91
1900	86.55 ± 0.24	0.96 ± 0.02	21.01
100000	83.70 ± 0.30	0.94 ± 0.05	13.13
BAYESBINN	67.91 ± 0.96	0.75 ± 0.20	5.03

BiMU conserves plasticity by conserving uncertain synapses

Robotics nuisance factors

Study of 12 factors OpenLORIS-Object

ID: 

OOD: 

Method	Features	Mean Acc. (%)	Aleatoric (AUC)	Epistemic (AUC)
		Binary neural networks		
BiMU	1,024	**73.61 ± 1.53**	**0.96 ± 0.01**	**1.00 ± 0.00**
	8,192	**89.19 ± 0.19**	**0.99 ± 0.00**	**1.00 ± 0.00**
	25,088	**90.62 ± 0.22**	**0.93 ± 0.00**	**0.90 ± 0.00**
BayesBiNN	1,024	72.01 ± 1.69	0.93 ± 0.01	**1.00 ± 0.00**
	8,192	86.93 ± 0.41	**0.99 ± 0.00**	**1.00 ± 0.00**
	25,088	89.37 ± 0.77	0.92 ± 0.00	**0.90 ± 0.01**
Syn. Meta.	1,024	62.82 ± 2.31	0.72 ± 0.04	–
	8,192	88.03 ± 0.38	0.63 ± 0.03	–
	25,088	86.72 ± 0.34	0.55 ± 0.00	–
STE	1,024	52.88 ± 3.39	0.73 ± 0.02	–
	8,192	79.12 ± 1.39	0.61 ± 0.05	–
	25,088	83.79 ± 1.13	0.55 ± 0.00	–
		Real-valued neural networks		
MESU	1,024	**80.82 ± 0.94**	0.90 ± 0.01	**0.99 ± 0.00**
	8,192	87.01 ± 0.69	1.00 ± 0.00	**0.98 ± 0.00**
	25,088	87.84 ± 0.11	**0.83 ± 0.00**	**0.86 ± 0.00**
EWC Online	1,024	75.72 ± 0.87	**0.97 ± 0.01**	–
	8,192	**87.18 ± 0.58**	**1.00 ± 0.00**	–
	25,088	**88.23 ± 0.05**	0.79 ± 0.00	–
SI	1,024	70.75 ± 0.57	0.95 ± 0.03	–
	8,192	86.44 ± 0.64	**1.00 ± 0.00**	–
	25,088	88.04 ± 0.03	**0.83 ± 0.00**	–
SGD	1,024	62.31 ± 1.80	0.50 ± 0.00	–
	8,192	86.27 ± 0.51	**1.00 ± 0.00**	–
	25,088	88.04 ± 0.03	**0.83 ± 0.00**	–

BiMU performs better than real-valued methods
After extracting features, data is **not normalized**, inducing performance discrepancies that BiMU **corrects**

BiMU behaves better than other method on data compression
x25 compression still allows **near-perfect OOD detection**
Mean ACC is highest

Shannon, Claude Elwood. "A mathematical theory of communication." The Bell system technical journal 27.3 (1948): 379-423.
Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." arXiv preprint arXiv:1112.5745 (2011).
Freeman, Linton C. "Elementary applied statistics: for students in behavioral science." (1965).
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).

Empirical results on Active Learning

Low frequency

Classes poorly represented

High frequency

Classes well represented

Data used for training

Percentage of data used to train out of the dataset

We compare uncertainty strategies to select examples to train on, from **epistemic disagreement** to **random acquisition**.

Aleatoric

Variability in the data

Epistemic

Variability in model entropies

Predictive

PU = EU + AU

Variation ratio

Variability of model predictions

Random

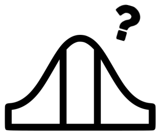
Selecting a random sample to train on

Active learning under data imbalance

Study of Animals dataset: hypothesis

Class	Number of train examples	Number of test examples
Butterfly	1997	50
Lizard	1412	50
Fish	1404	50
Monkey	1043	50
Spider	1015	50
Eagle	849	50
Frog	617	50
Jellyfish	501	50
Penguin	390	50
Whale	291	50
Zebra	164	50
Crocodile	136	50
Leopard	132	50
Sheep	125	50
Raccoon	106	50
Raven	91	50
Panda	62	50
Lynx	66	50
Bull	72	50
Scorpion	76	50

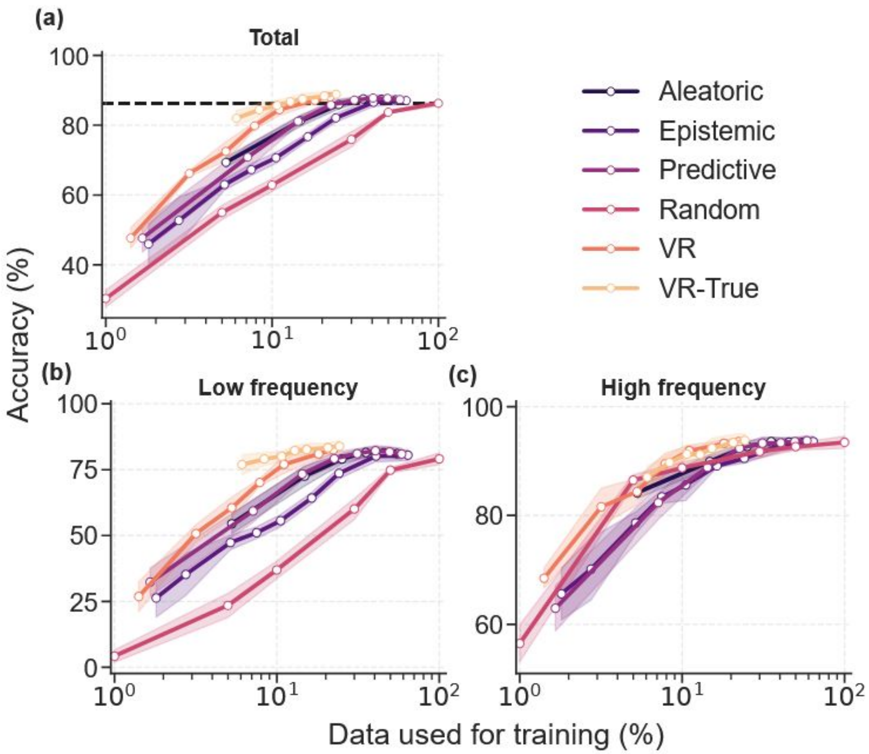
Intuition: if we can recognize **OOD, we** can recognize **imbalanced data, then train on it**

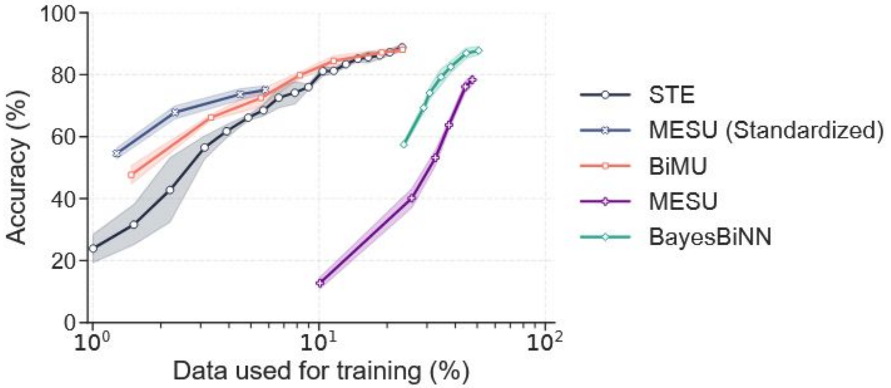


Imbalance train dataset,
balanced test dataset
26x data frequency disparity between **lowest** and **highest class**

Active learning under data imbalance

Study of Animals dataset



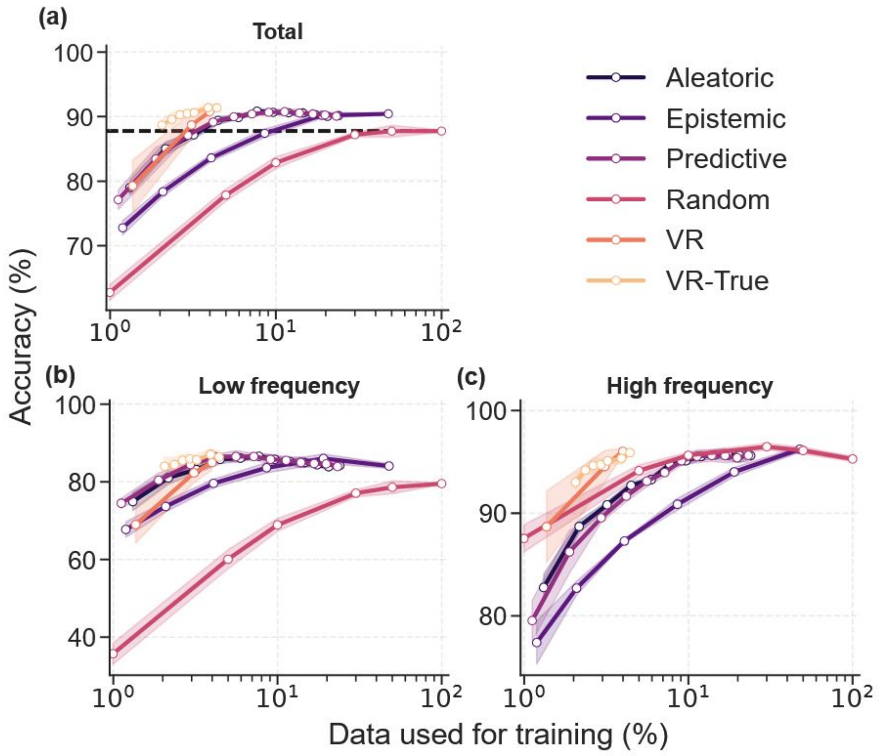


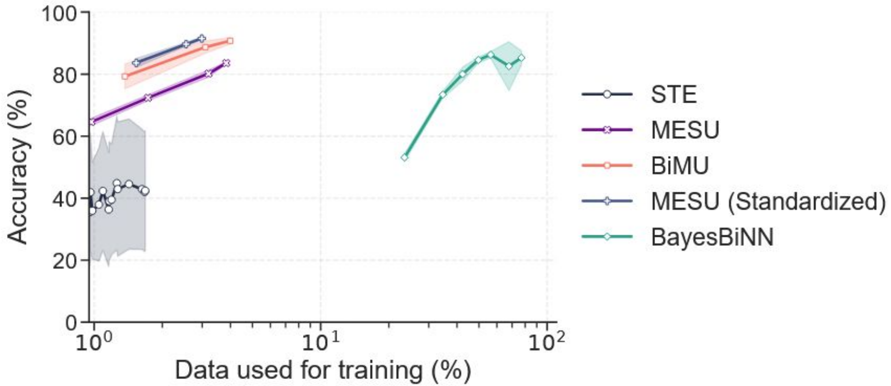
BiMU requires less data to be trained, using uncertainties
10x reduction in data usage using **variation ratio**

Low frequency classes are better taken into account

Active Continual Learning under data imbalance

Study of OpenLORIS-Object dataset





BiMU can do Active Continual Learning, maintaining uncertainties

32x reduction in data usage using **variation ratio**

BiMU actively helps with imbalance and saves updates

How many disagreement samples?

Cheap MC samples, yet how few?

Samples	Accuracy (%)	Data used (%)	Threshold
2	89.30 ± 0.88	3.30 ± 0.04	0.50
3	90.61 ± 0.53	3.87 ± 0.05	0.33
3	82.26 ± 1.15	1.62 ± 0.03	0.66
5	88.85 ± 0.92	2.91 ± 0.16	0.20
5	80.86 ± 1.90	1.47 ± 0.16	0.40 – 0.60
5	52.21 ± 23.71	0.34 ± 0.19	0.80
10	90.91 ± 0.98	3.97 ± 0.03	0.10
10	88.70 ± 1.94	3.10 ± 0.31	0.20 – 0.30
10	79.27 ± 4.00	1.37 ± 0.24	0.40 – 0.50
10	44.02 ± 32.02	0.42 ± 0.39	0.60
10	9.37 ± 8.22	0.02 ± 0.04	0.70
10	5.26 ± 0.00	0.00 ± 0.00	0.80 – 0.90
25	91.34 ± 0.50	5.63 ± 0.09	0.04
25	91.11 ± 0.44	4.40 ± 0.04	0.12
25	90.93 ± 0.70	4.03 ± 0.05	0.16
25	90.23 ± 0.68	3.65 ± 0.14	0.20
25	89.00 ± 0.64	3.27 ± 0.23	0.24
25	87.16 ± 1.88	2.93 ± 0.36	0.28
25	86.40 ± 2.68	2.61 ± 0.44	0.32
25	79.14 ± 13.18	1.96 ± 0.78	0.36
25	66.41 ± 30.90	1.48 ± 0.84	0.40
25	51.53 ± 37.93	1.03 ± 0.88	0.44
25	34.82 ± 36.25	0.52 ± 0.65	0.48
25	19.95 ± 29.37	0.23 ± 0.45	0.52
25	19.16 ± 27.79	0.17 ± 0.34	0.56
25	18.48 ± 26.44	0.13 ± 0.26	0.60
25	17.01 ± 23.49	0.10 ± 0.19	0.64
25	16.07 ± 21.61	0.07 ± 0.14	0.68
25	5.26 ± 0.00	0.00 ± 0.00	> 0.72
Baseline	87.76 ± 0.19	100.00 ± 0.26	-

Very **few inferences** are needed to perform well in **Active Continual Learning**

2 MC Samples are enough to cause a **30x** reduction in data usage using **variation ratio**

Conclusion

What to get out of this talk?

BiMU faces Continual Learning challenges
With Bayesian **forgetting,** it continuously learns with **metaplastic learning rate and curvature surrogate**

BiMU reduces the number of updates to be taken
Up to **35x update reduction** with Active Learning using **12.4x cheaper** inference

BiMU offers a hardware friendly solution
Binary representation on **one bit**,
Input compression up to **25x**,
Active Learning preserves **device endurance**

Uncertainty is a common denominator for lifelong embedded learning

Thank you for listening!

Kellian Cottart, Théo Ballet,
Djohan Bonnet, Damien Querlioz

Perspectives

Zero-order optimization schemes for BiMU using stochasticity and uncertainties

Exploit nanotechnologies and emerging memories (P-Bits) to perform AI in-memory computing with BiMU