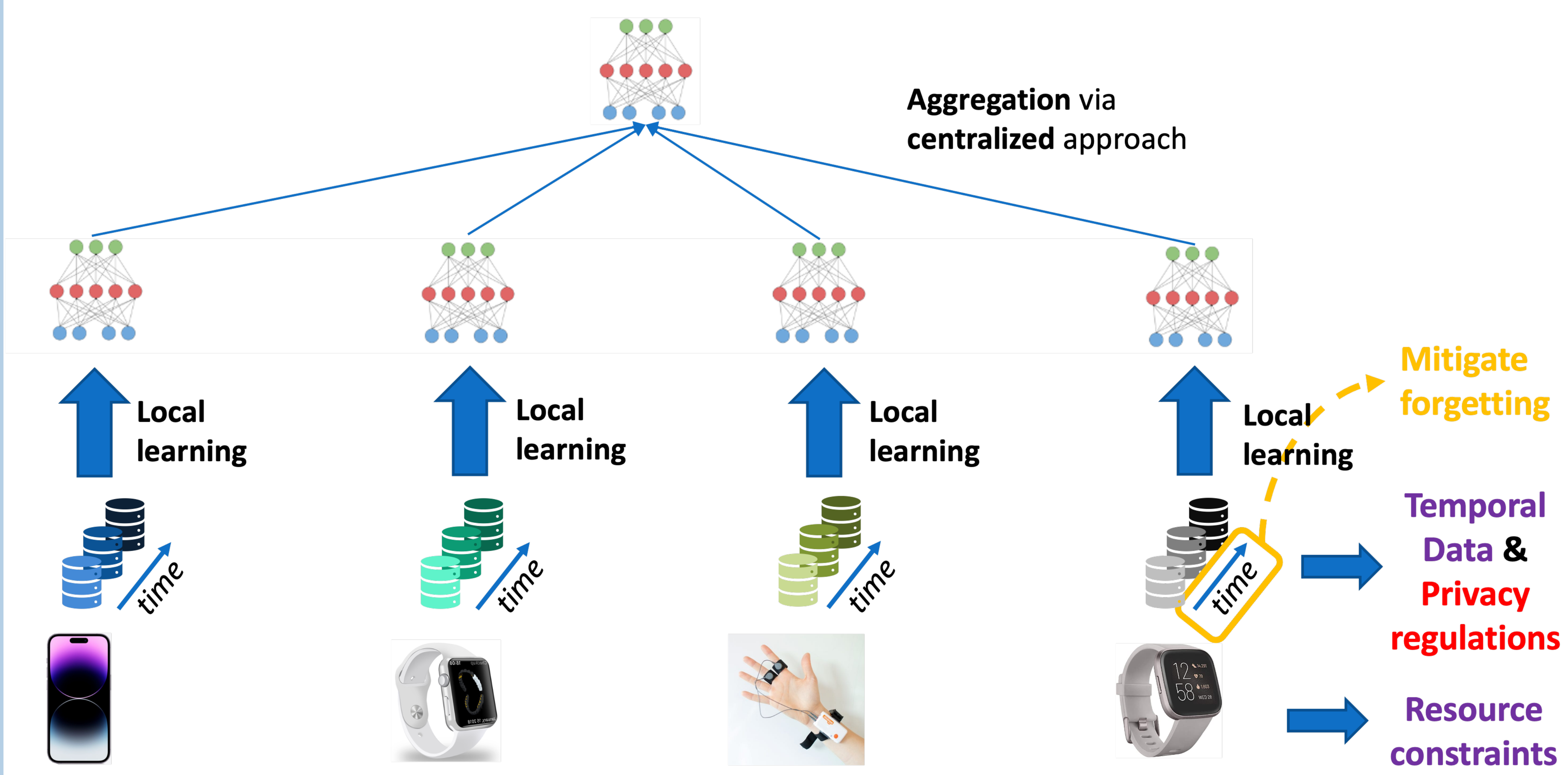
Decentralized Plasticity in Reservoir Dynamical Networks for Pervasive Environments

Valerio De Caro Davide Bacciu Claudio Gallicchio

Department of Computer Science, University of Pisa, Italy

Learning in Pervasive Environments: Summary and Objective



Pervasive environments → massive distribution of resource-constrained devices collecting private temporal data over time

Learn a global model which *generalizes over space and time* by *localizing the learning process on these dimensions*

Echo State Networks with Intrinsic Plasticity

Federated Averaging to deal with local private datasets

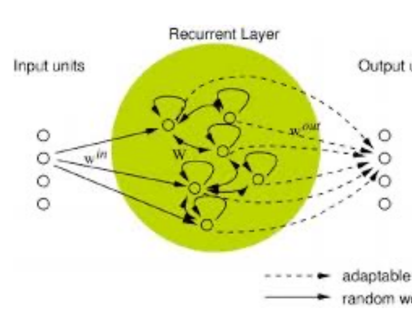
Replay strategies to deal with streaming data

Combine the tools in a unified framework which accounts for uncertainty from distribution over space and time

Learning from Temporal Data, Efficiently

Echo State Networks

$\mathbf{W}_{in}$ and $\widehat{\mathbf{W}}$ are *randomized* and ***fixed***



$$\mathbf{x}_{net}(t) = \mathbf{W}_{in}\mathbf{u}(t) + \mathbf{b}_{rec} + \hat{\mathbf{W}}\mathbf{x}(t-1)$$
$$\mathbf{x}(t) = (1-a)\,\mathbf{x}(t-1) + af\left(\mathbf{g}\mathbf{x}_{net}(t) + \mathbf{b}\right)$$
$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) + \mathbf{b}_{out}$$

Reservoir Adaptation: Intrinsic Plasticity

Maximize the information gain via Maximum Entropy

- *tanh* activation → Gaussian as ME distribution

$$\mathcal{L}(\theta;\mu,\sigma) = D_{KL}(\tilde{q}\,||\,\mathcal{N}_{\mu,\sigma}) \quad \text{with} \quad \theta = \{\mathbf{g}, \mathbf{b}\}$$

$$\Delta b = -\eta\left(-\frac{\mu}{\sigma^2} + \frac{\tilde{x}}{\sigma^2} + 1 - \tilde{x}^2 + \mu\tilde{x}\right)$$
$$\Delta g = \frac{\eta}{g} + \Delta b x_{net}$$

Readout Learning: Ridge Regression

$$\mathbf{W} = \mathbf{Y}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \lambda\mathbf{I})^{-1}$$

- Extended for learning over space and time with an ***exact*** approach

$$\mathbf{W} = (\textstyle\sum_i \mathbf{A}_i)^T(\sum_i \mathbf{B}_i + \lambda\mathbf{I})^{-1}$$

where i indexes a slice

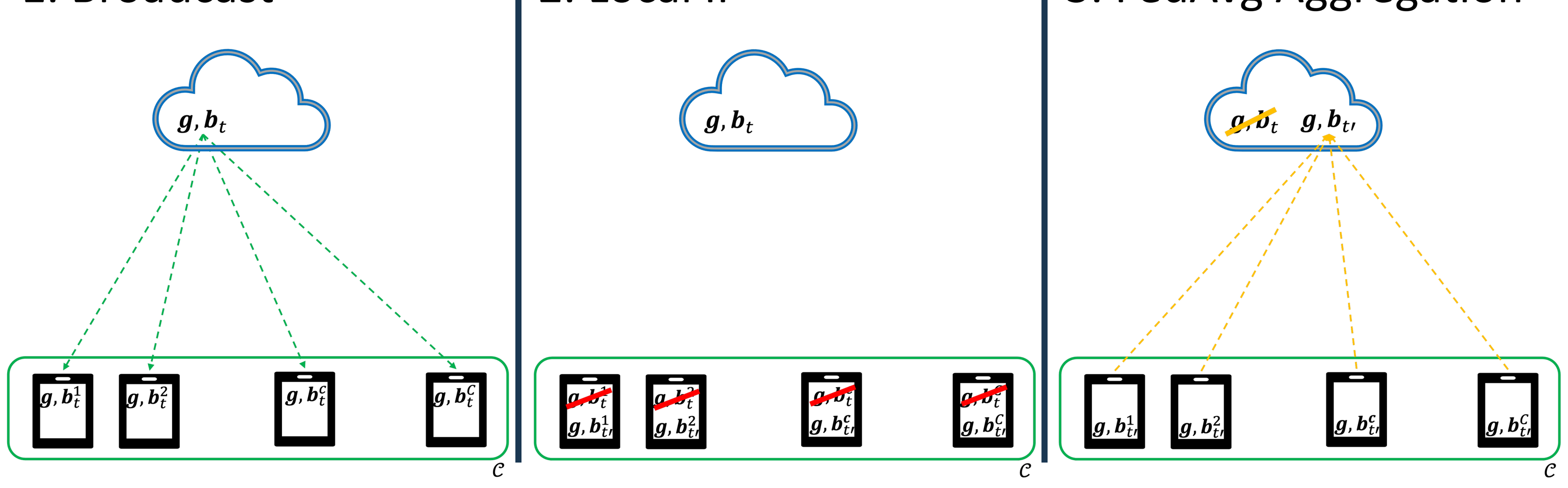
Localizing Plasticity on Space...

$$\mathcal{L}(\theta;\mu,\sigma) \le \sum_{c\in\mathcal{C}} p_c D_{KL}(\tilde{q}_c\,||\,\mathcal{N}_{\mu,\sigma})$$
$$= \sum_{c\in\mathcal{C}} p_c \mathcal{L}_c(\theta_c;\mu,\sigma) = \mathcal{L}_F(\theta;\mu,\sigma)$$

Convex combination of clients models, accounts for heterogeneity

Federated Intrinsic Plasticity

1. Broadcast 2. Local IP 3. FedAvg Aggregation



Only 2N parameters exchanged per round (N = number of reservoir units)

... and Time

$$\mathcal{L}(\theta;\mu,\sigma) \le \sum_{t=1}^{T}\sum_{k\in\mathcal{K}} p_{t,k} D_{KL}(\tilde{q}_k\,||\,\mathcal{N}_{\mu,\sigma})$$
$$\simeq \sum_{t=1}^{T} D_{KL}(\tilde{q}_t\,||\,\mathcal{N}_{\mu,\sigma}) \quad \Longrightarrow \quad \mathcal{L}_{FC}(\theta;\sigma,\mu) = \sum_{t=1}^{T}\sum_{c\in\mathcal{C}} p_{t,c} D_{KL}(\tilde{q}_{t,c}\,||\,\mathcal{N}_{\mu,\sigma}),$$

Accounts for *evolving* heterogeneity

Federated Continual Intrinsic Plasticity

Algorithm 3 Federated Continual Intrinsic Plasticity (`FedCLIP`)

```
Input: clients C, learning rate η, local epochs E, batch size B
R ← {W_in, Ŵ, α}
θ_0^0 ← {g_0^0 ← 1, b_0^0 ← 0}
Send R to all clients c ∈ C
for each experience t ∈ {0, 1, ..., T−1} do
  for each round r ∈ {0, 1, ..., R−1} do
    for each client c ∈ C in parallel do
      Send θ_t^r to client c
      θ_{t,c}^r ← ContinualIPUpdate_c(t, g_t^r, b_t^r, η, E, r == R−1)
    end for
    θ_t^{r+1} ← { Σ_{c∈C} (n_{t,c}/n_t) g_{t,c}^r , Σ_{c∈[C]} (n_{t,c}/n_t) b_{t,c}^r }
  end for
end for
```

Synchronized experiences

Algorithm 4 ContinualIPUpdate (on client c)

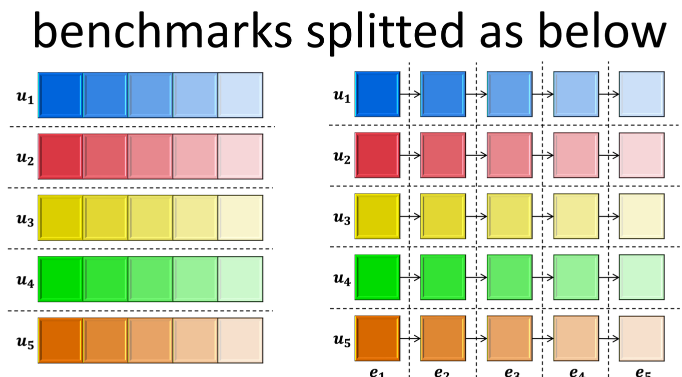
```
Env: stream = [D_0, D_1, ..., D_{T−1}], M_0 = {}
Input: experience t, global gain g_t^r, global bias b_t^r, learning rate η, epochs E, boolean update
B_t ← split data D_t ∪ M_t into a set of batches of size B
for epoch e ∈ {0, 1, ..., E−1} do
  for batch b ∈ B do
    Compute the average Δg^b, Δb^b over b
    g_t, b_t ← g_t + Δg^b, b_t + Δb^b
  end for
end for
if update then
  M_{t+1} ← UpdateWithStrategy(D_t, M_t)
end if
return {g_t, b_t}
```

Retain memory of previous experiences via replay strategy

Experiments

Setup

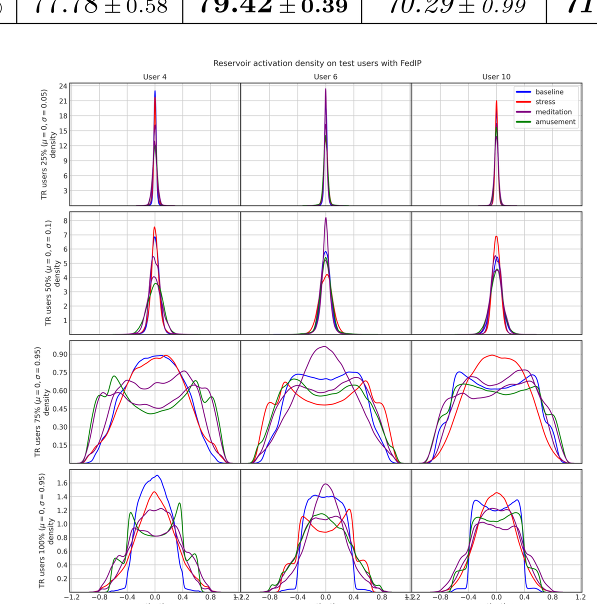
- Assessed on two HAR benchmarks splitted as below



- *Baselines*:
  - No FedIP on the stationary setting;
  - Naïve and Joint strategies on the non-stationary setting
- Repeated with different percentages of training clients
- Measured accuracy and activation density

Federated and Stationary

%TR	WESAD		HHAR	
	w/o FedIP	w/ FedIP	w/o FedIP	w/ FedIP
25%	72.09 ± 0.59	**78.68 ± 0.12**	57.08 ± 3.11	**69.83 ± 0.64**
50%	72.04 ± 1.03	**77.43 ± 0.19**	***63.88 ± 6.02***	***57.74 ± 0.19***
75%	76.53 ± 1.08	**77.97 ± 0.41**	***71.09 ± 0.56***	***71.08 ± 0.69***
100%	77.78 ± 0.58	**79.42 ± 0.39**	*70.29 ± 0.99*	***71.38 ± 0.43***



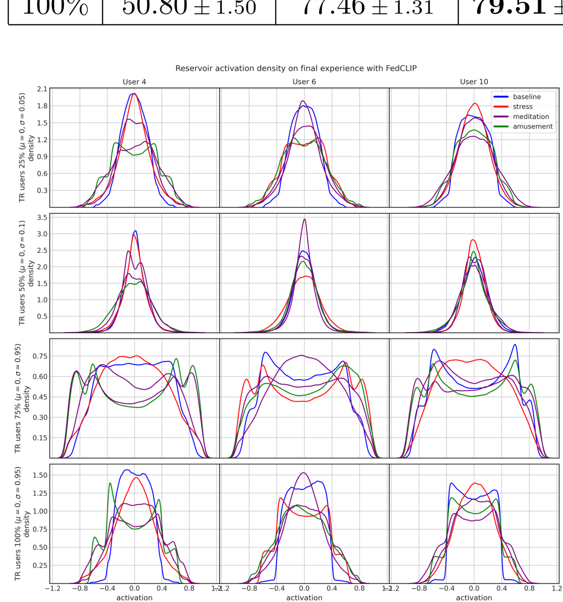
✓ FedIP outperforms baseline

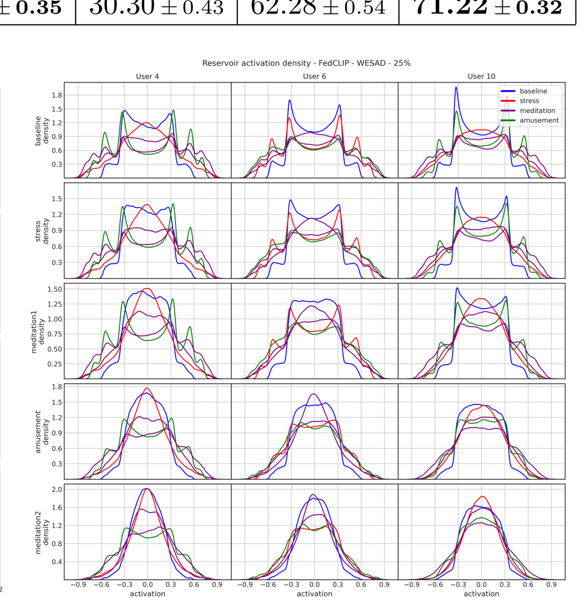
✓ Regularizes with ≤ 50% and improves information gain with 75-100%

✓ Does not suffer from averaging approximation

Federated and Non-Stationary

%TR	WESAD			HHAR		
	Naïve	Replay	Joint	Naïve	Replay	Joint
25%	27.32 ± 10.86	***79.23 ± 0.44***	*78.75 ± 0.67*	34.85 ± 3.08	51.16 ± 5.88	**69.44 ± 0.38**
50%	30.60 ± 7.51	**77.49 ± 0.89**	75.95 ± 1.07	30.16 ± 2.10	43.77 ± 1.50	**60.85 ± 4.37**
75%	51.50 ± 4.10	77.04 ± 0.89	**78.17 ± 0.54**	28.62 ± 0.93	59.83 ± 0.88	**71.14 ± 0.84**
100%	50.80 ± 1.50	77.46 ± 1.31	**79.51 ± 0.35**	30.30 ± 0.43	62.28 ± 0.54	**71.22 ± 0.32**





✓ Results consistent with the stationary setting

✓ Replay strategy consistently **mitigates forgetting**

✓ **Good forward transfer**

Conclusions

- Extended ESNs' **Intrinsic Plasticity** towards a learning approach **localized on space and time**
- Proposed **two communication-efficient algorithms for stationary and non-stationary data**
- Both approaches **do not suffer from the approximation** and lead to good generalization

Acknowledgements

This work is supported by the EC H2020 program under project TEACHING (grant n. 871385).