xDISCO: eXplainable DIStributed COllaborative learning for images

Klavdiia Naumova[1], Mary-Anne Hartley[1], Arnout Devos[1], Sai Praneeth Karimireddy[2], Martin Jaggi[1]

[1]Intelligent Global Health Research Group, Machine Learning and Optimization Laboratory, EPFL

[2]Berkeley AI Research Laboratory, UC Berkeley

1 BACKGROUND

Federated learning (**FL**) is a method of building collaborative predictive models between clients without sharing any original data. FL is actively used in privacy-sensitive domains such as medicine and finance.

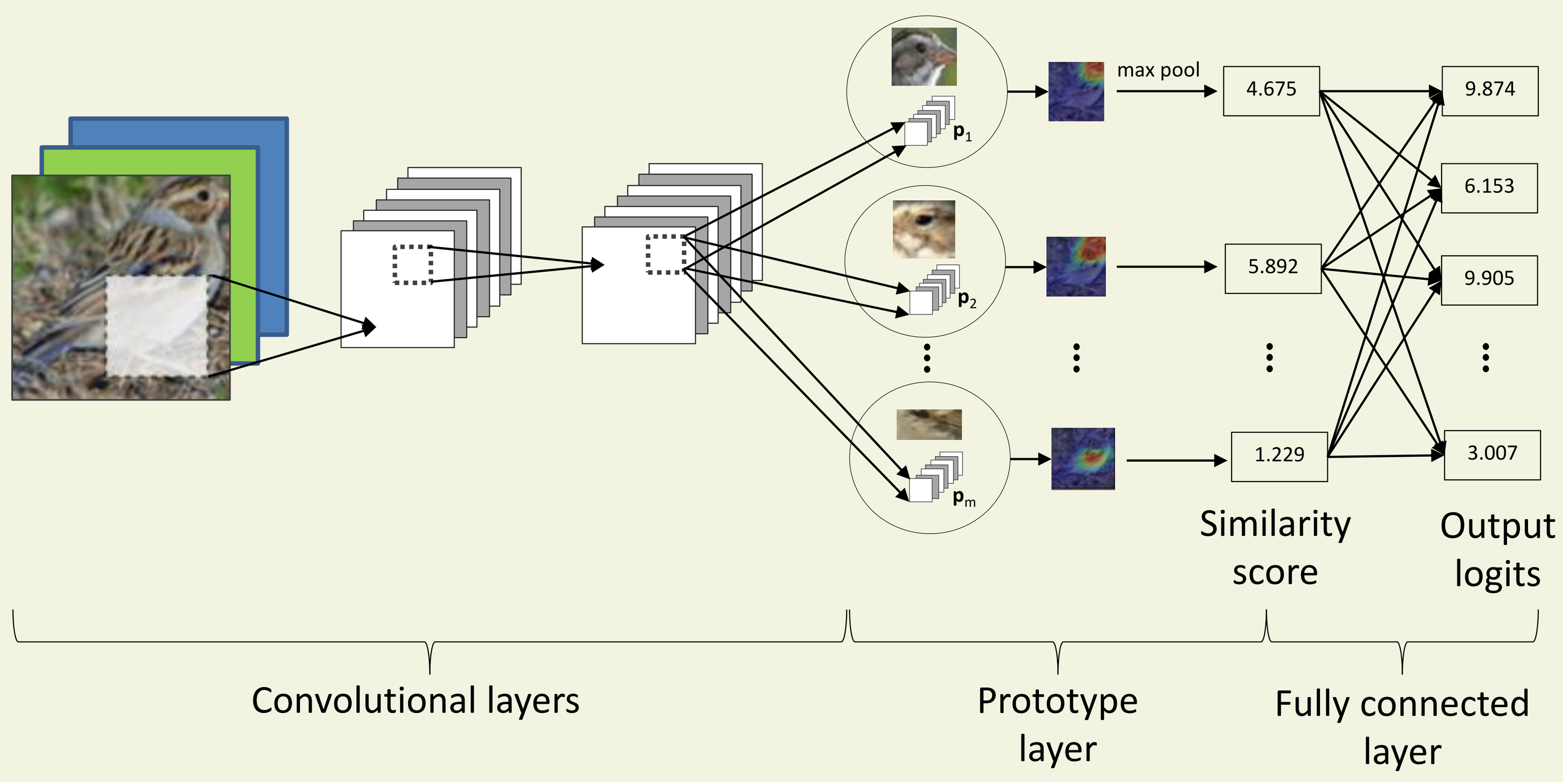
Challenges of FL:

- Low interpretability
- Low robustness to systemic bias between datasets.

These problems are particularly important for **images** since the deep learning models they require are also poorly interpretable.

2 OUR SOLUTION

xDISCO adapts interpretable **"prototypical part learning"** to an FL setting, where each client learns which parts of its images are most important for the task.

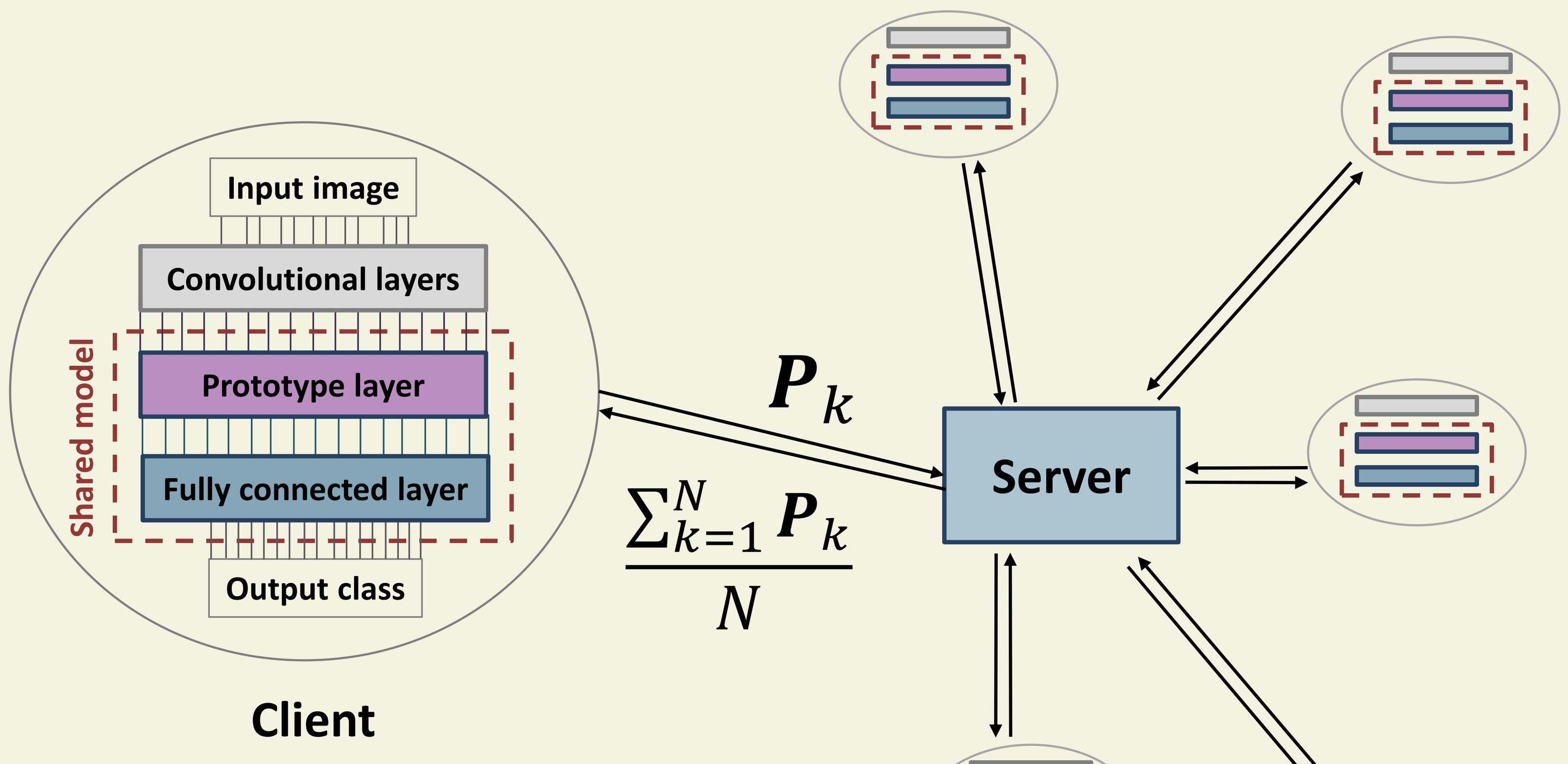


ProtoPNet by Chen, et al. (2018)

to FL

In every **communication round**:

1. Each client k learns m **local** prototypes $\boldsymbol{P}_k = \{p_{kj}\}_{j=1}^{m}$ on its dataset and sends them to the server.
2. The server aggregates and averages local prototypes to obtain the **global** ones and sends them back to N clients.



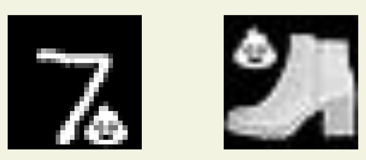
After training we can visualize global and local prototypes on each client's dataset and compare.

3 EXPERIMENTS

Datasets

Name	Number of train/test images	Color/size	Example
MNIST	50,000/10,000	Grayscale/ 28×28	
Fashion MNIST	50,000/10,000	Grayscale/ 28×28	

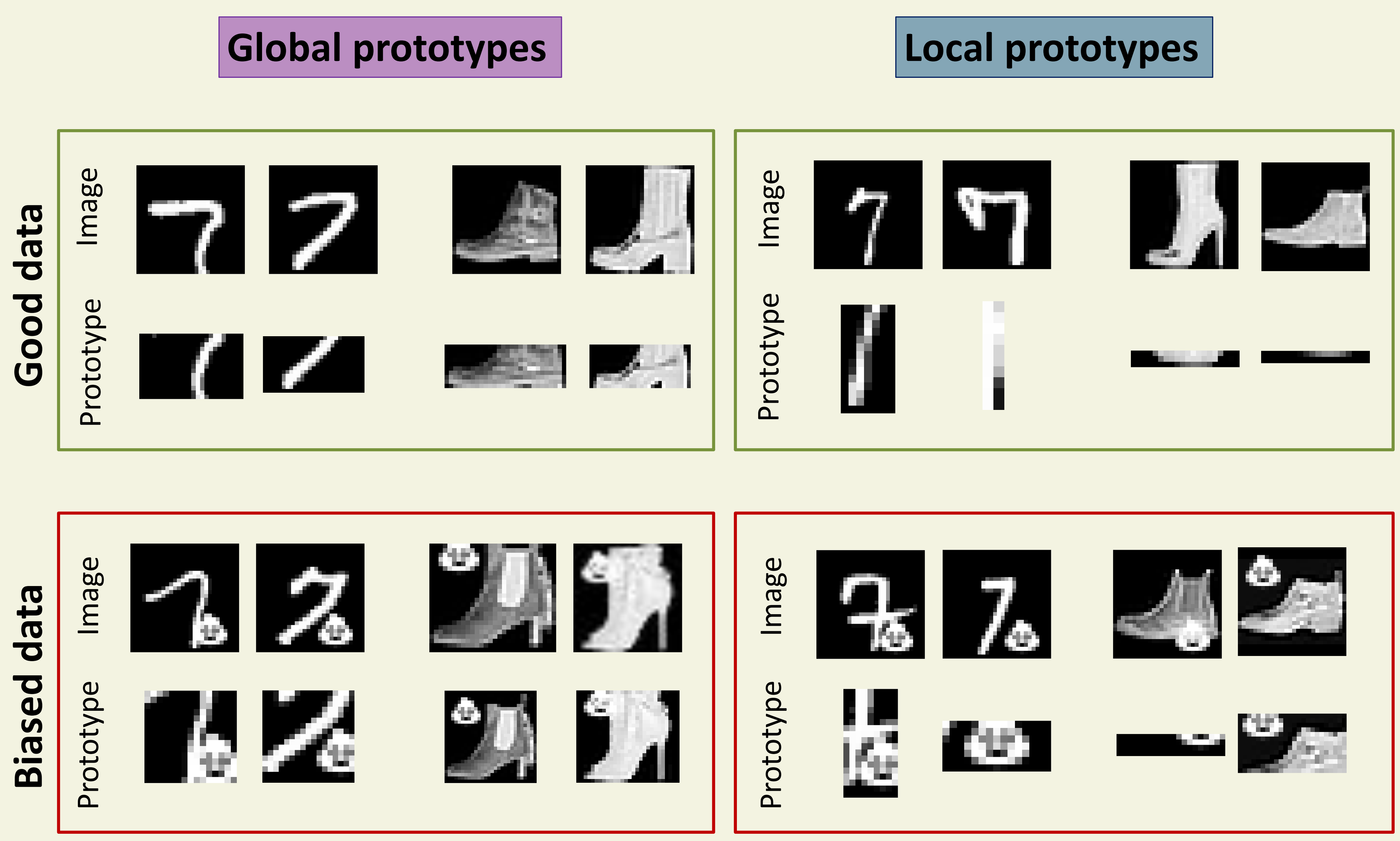
Examples of biased data



We added bias (an emoji) to one client's images of a particular class.

Results

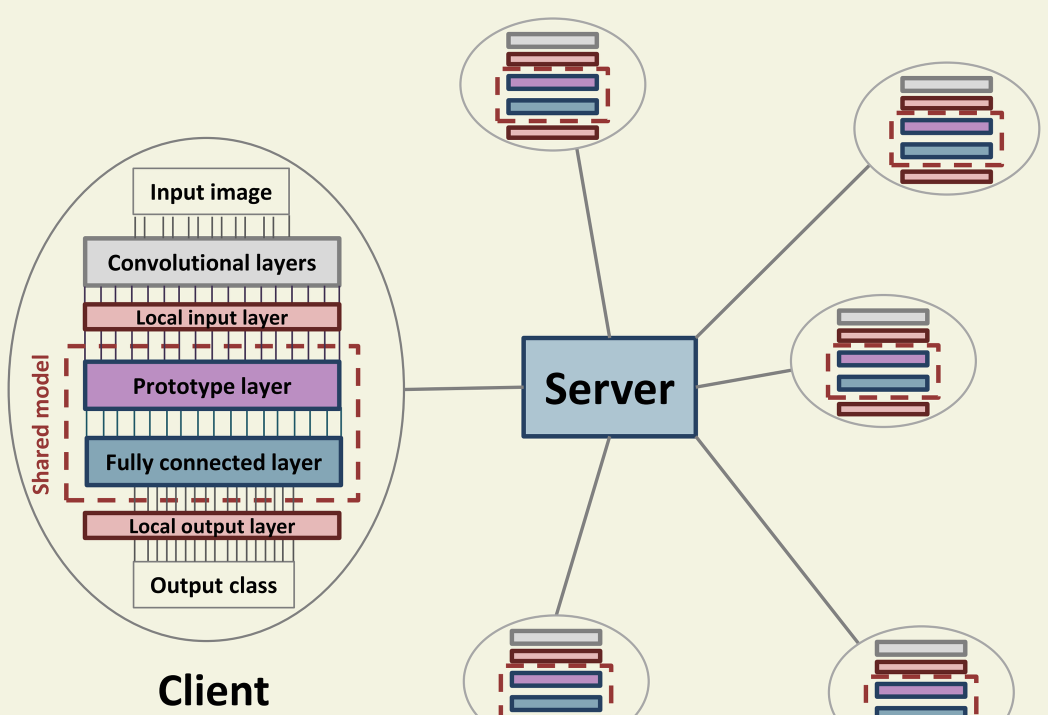
Model	Accuracy, %	
	MNIST	Fashion MNIST
Baseline*	98.0	89.2
xDISCO (ours) good data	91.7	81.5
xDISCO (ours) biased data	91.7	81.9



*The baseline model is a ProtoPNet trained on good data in a centralized setting

4 FUTURE WORK

- Adding personalization layers suggested by Roschewitz, et al. (2021) around a shared part of the model to identify and correct local bias by learning a shift from local to global prototypes;
- Quantifying the privacy risk of sharing prototypes.



5 CONCLUSIONS

- A prototypical part learning model can be used in an FL setting on good and systematically biased data to provide interpretability.
- Learned prototypes activate a part of an image at which the network looks to base its prediction and this activated region changes in presence of data bias.
- We hypothesize that with personalization layers, it would be possible to identify and correct bias in federated learning in a privacy-preserving way.