MyThisYourThat: interpretable bias identification in federated learning for biomedical images

Klavdiia Naumova[1,*], Arnout Devos[2], Sai Praneeth Karimireddy[3,4], Martin Jaggi[1], Mary-Anne Hartley[5]

[1]EPFL, Switzerland; [2]ETH Zurich, Switzerland; [3]UC Berkeley, CA, USA; [4]USC, CA, USA [5]Yale University, CT, USA

*incoming ELLIS PhD student at DKFZ, Heidelberg, Germany

1 BACKGROUND

Federated learning (FL) allows multiple data owners (clients) to jointly build predictive models without sharing any original data. This method is actively used in privacy-sensitive domains such as medicine and finance.

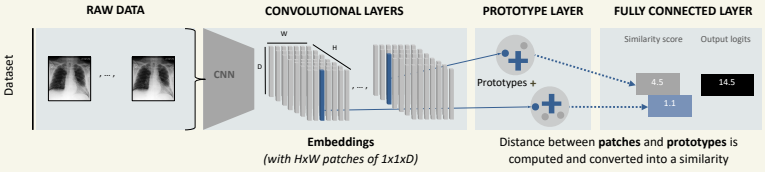
Challenges of FL:

- Low interpretability induced by *black-box data*
- Low robustness to systematic bias between datasets.

These problems are particularly important for **images** since the deep learning models they require are also poorly interpretable (a.k.a. *black-box models*).

2 OUR SOLUTION

MyTH (MyThisYourThat) adapts an inherently interpretable **prototypical part learning** network ProtoPNet to an FL setting, where each client learns parts of its images most important for the task.



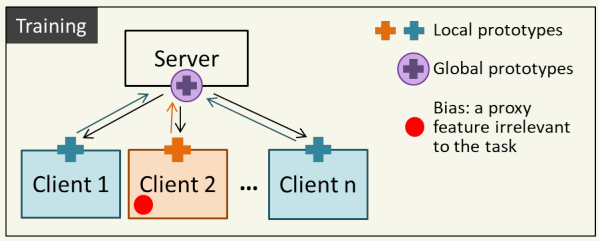
ProtoPNet by Chen, et al. (2018)

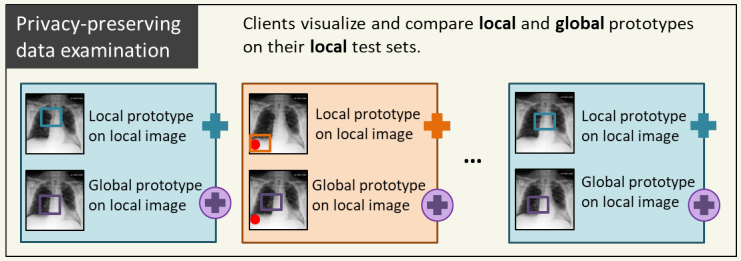
In every **communication round of MyTH**:

1. Each client learns **local** prototypes on its dataset and sends them to the server.
2. The server aggregates local prototypes to obtain the **global** ones and sends them back to clients to continue training.

Image patch most activated by a prototype

After training, each client visualizes and compares its locally and globally learned prototypes on its local dataset: comparing its own *This* to others' *That*. The difference between **local** and **global** prototypes indicates the presence of bias among the clients.





Naumova et al. (2024), accepted to *npj Digital Medicine*

3 EXPERIMENTS

Data: benchmark chest X-rays CheXpert
Tasks: binary classification of cardiomegaly and pleural effusion
FL setting: 4 IID clients (3 unbiased and 1 biased client)
Models:

- Local (LM) trained on each client data without communication
- Global (GM) trained collaboratively sharing all ProtoPNet parameters
- Personalized (PM) trained collaboratively sharing only prototypes

Baselines:

- Centralized model trained on the whole unbiased data

Examples of biased data

Synthetic bias added to the cardiomegaly class

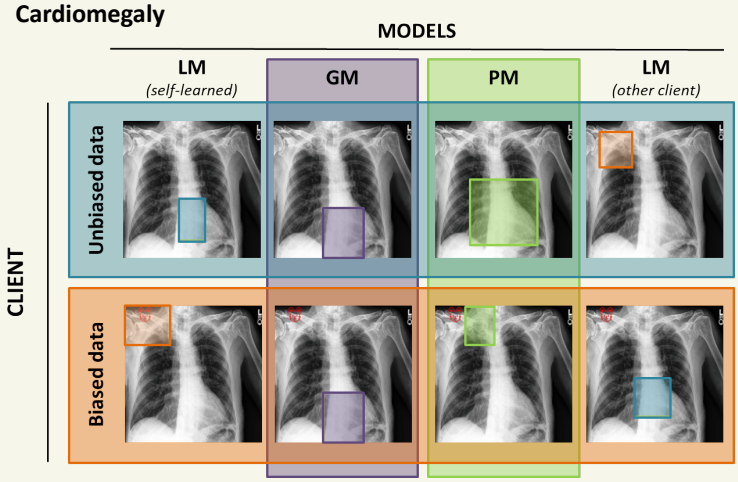
Real-world bias: chest drains in the pleural effusion class

4 RESULTS

Examples of prototypes learned by local, global and personalized models and visualized on unbiased and biased test images.

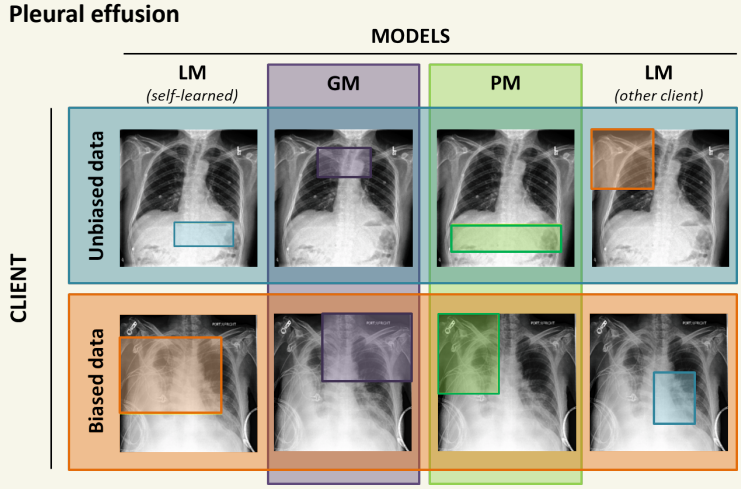
Difference between the prototypes => bias in FL setting

Cardiomegaly



Large difference between LM and GM on biased data

Pleural effusion



Large difference between LM and GM on unbiased data

5 CONCLUSION AND FUTURE WORK

MyTH allows visually interpretable and privacy-preserving identification of data bias in federated learning for imaging data.

Future work:

- Investigating privacy-transparency trade-off
- Introducing debiasing option
- Implementing MyTH in a web tool for distributed learning: discolab.ai