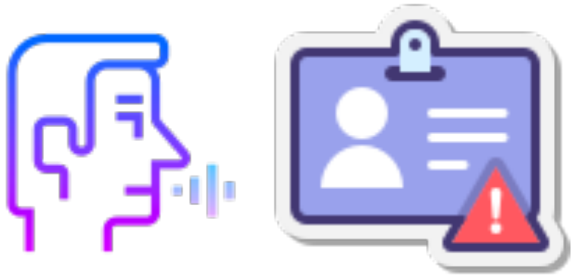
Distributed Training for Speech Recognition using Local Knowledge Aggregation and Knowledge Distillation in Heterogeneous Systems

Hongrui Shi
Joint work with Dr Valentin Radu and Dr Po Yang

The University Of Sheffield.

Training Speech Recognition Task on the Edge

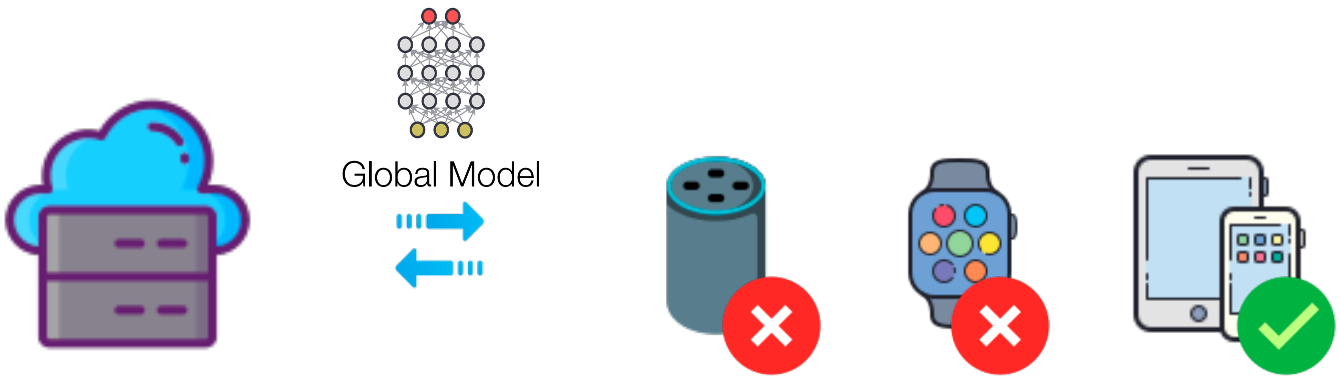
- Voice is biometric, spurring the privacy challenge for speech recognition training.
  - Cost: data communication and data protection
  - Risk: policies and legal restrictions



Solution: Distributed training on user devices without pooling the data to a central server.

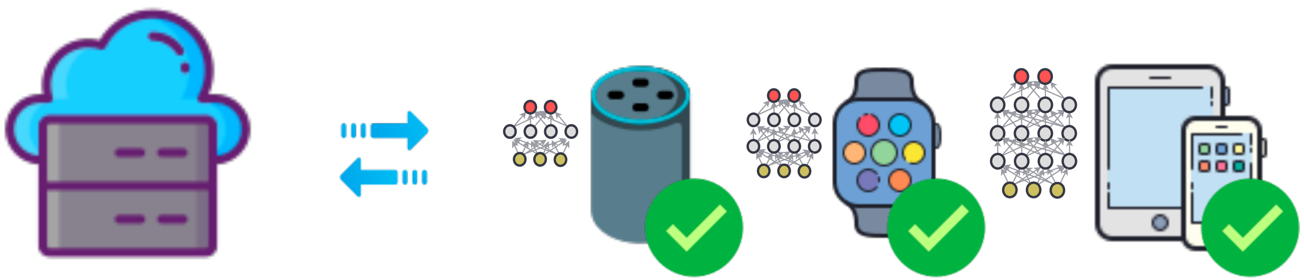
Uniform Model in FL

- In classical federated learning the server distributes a uniform model to all clients.
  - Pros: simple and effective model aggregation methods such as averaging parameters
  - Cons:
    - Stragglers: clients with lower computation resources, unable to complete model training in time
    - Effects of system heterogeneity: **performance degradation**, **slows training**, **unfairness**.



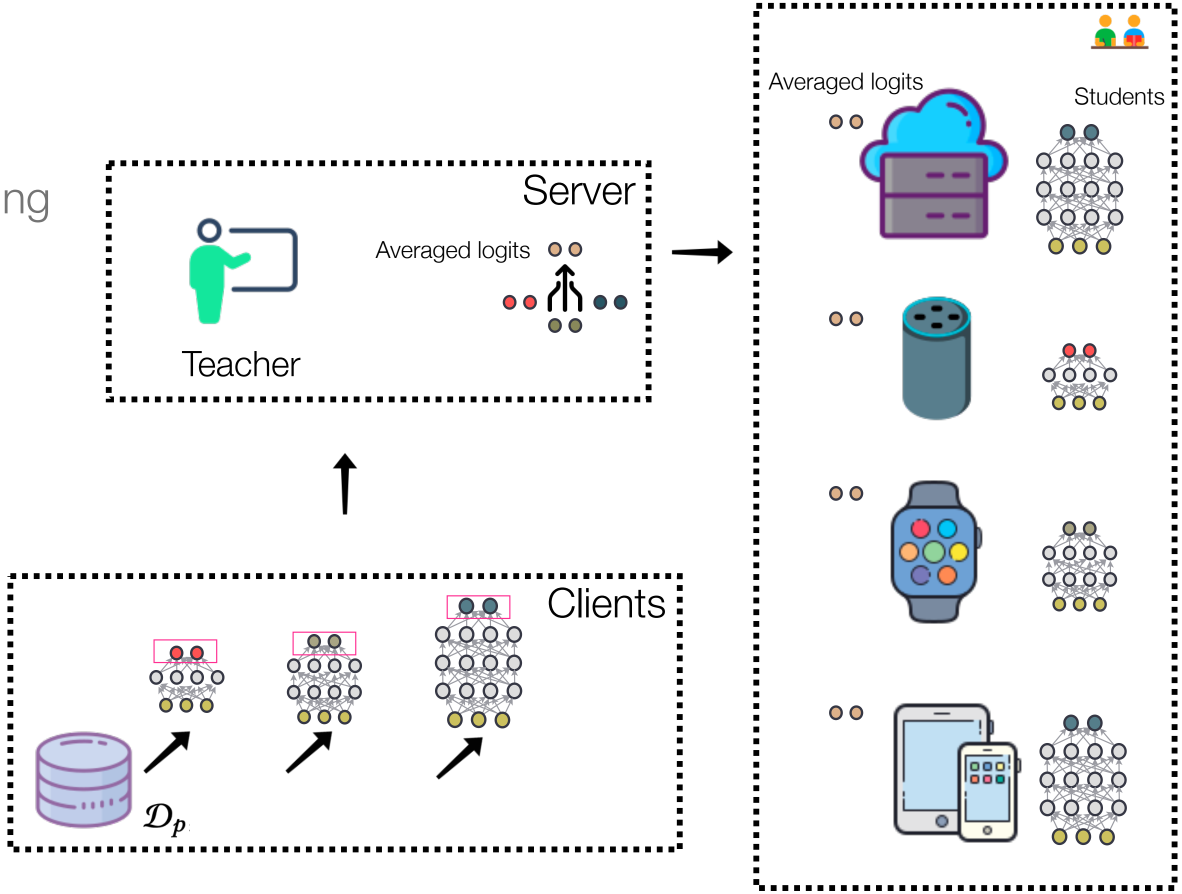
Heterogeneous Models in FL

- The server distributes custom size (heterogeneous) models to each device:
  - Pros: client model size is chosen based on the local system characteristics (personalised).
  - Cons: structural barriers for knowledge aggregation
- Solution: Student-Teacher learning with knowledge distillation



Proposed Method—FedKAD

- KD-based FL
  - A public dataset to support S-T learning
  - Teacher:
    - logits from clients
    - Aggregated on server
  - Students:
    - Client models/global model

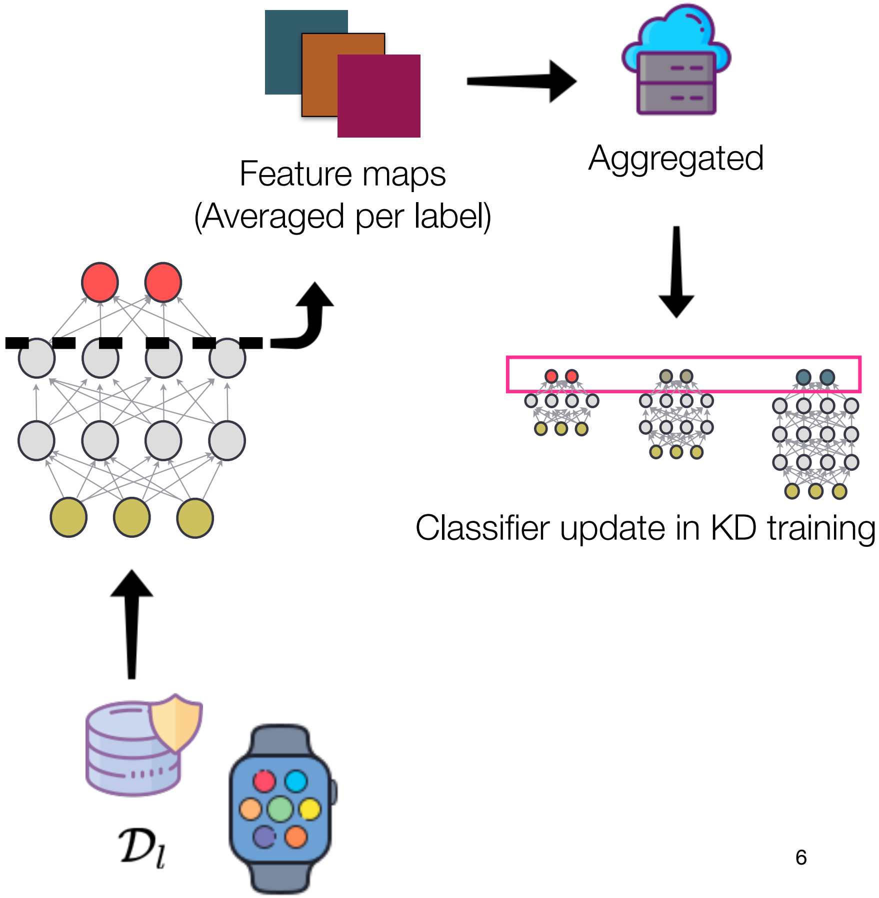


Proposed Method—FedKAD

- Feature maps on the client
  - To update model classifier in KD training
  - Locally averaged per label (prototype feature map)

Reduced communication cost with local aggregation based on our evaluation

FedKAD	FM Aggregation	GSC data
Comm. cost per round (MB)	w/o	1390.07
	w/	**11.47**

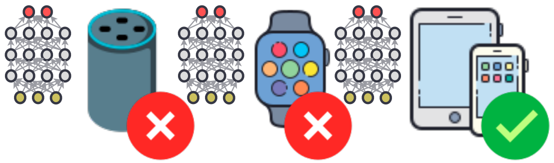


Evaluation—Experiment Setup

- **Dataset**: Google Speech Command transformed to Mel Spectrogram
- **Client size**: 20
- **Hetero. Models**: WideResnet with varying depths
- **Non-IID client data**: Dirichlet distribution (alpha to control data heterogeneity)

Evaluation — Baselines

- Uniform model (model fusion)
  - FedAvg
  - FedProx



Some clients are assumed constant stragglers

- Heterogenous models (KD training)
  - FedMD

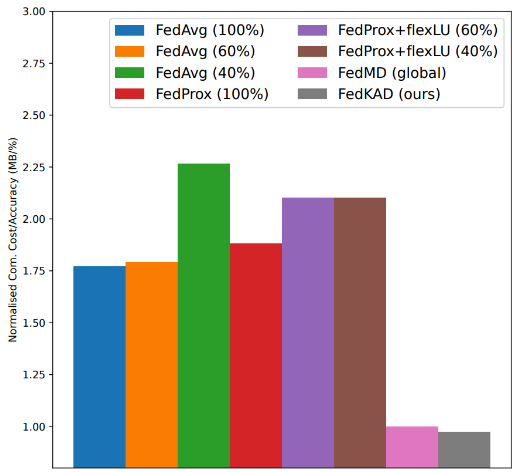
Assumed 100% client participation

Evaluation — Test Accuracy/Communication Cost

- Test accuracy (%)
  - FedKAD improves FedMD
  - FedKAD outperforms FedAvg/ FedProx with 40% client participation

Config.	Method	$\alpha = 0.1$	$\alpha = 0.5$
Uni. models	FedAvg (100%)	87.43%	91.08%
	FedAvg (60%)	81.63%	90.16%
	FedAvg (40%)	75.25%	87.22%
	FedProx (100%)	**87.57%**	**91.11%**
	FedProx (60%)	81.05%	89.60%
	FedProx (40%)	76.04%	86.55%
	FedProx+flexLU (60%)	87.34%	90.87%
	FedProx+flexLU (40%)	86.82%	90.48%
Hetero. models	FedMD (clients)	73.79%	77.38%
	FedMD (global)	76.68%	81.01%
	FedKAD (Ours)	**77.00%**	**81.13%**

- Communication cost (MB/per acc.%)
  - FedKDA reduces com. cost by half from FedAvg/FedProx
  - FedKAD and FedMD are on par



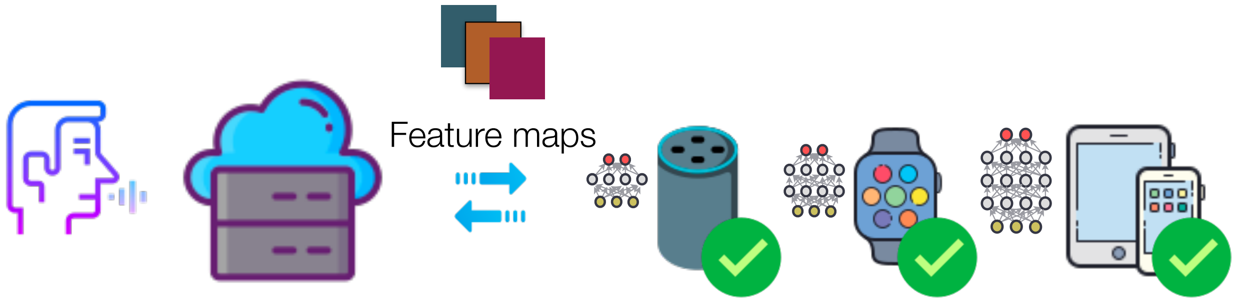
Problems of KD-based method

- Additional computation/memory cost:
  - A public dataset (or data generator).
  - KD training.

Summary/Takeaways

- We adapt client model size to ensure wider participation of clients to FL rounds.
- We exploit feature maps to boost KD-based heterogeneous FL.
- Uniform models methods (FedAvg, FedProx) need high client participation to outperform our FedKAD, which is not realistic for computing constrained devices.
- Our FedKAD surpasses another heterogeneous models method, FedMD, in both accuracy and communication efficiency.

FedKAD



Thank you!

hshi21@sheffield.ac.uk