HETEROGENEOUS ZERO-SHOT FEDERATED LEARNING WITH NEW CLASSES FOR AUDIO CLASSIFICATION

Gautham Krishna Gudur
Ericsson R&D

Satheesh Kumar Perepu
Ericsson Research

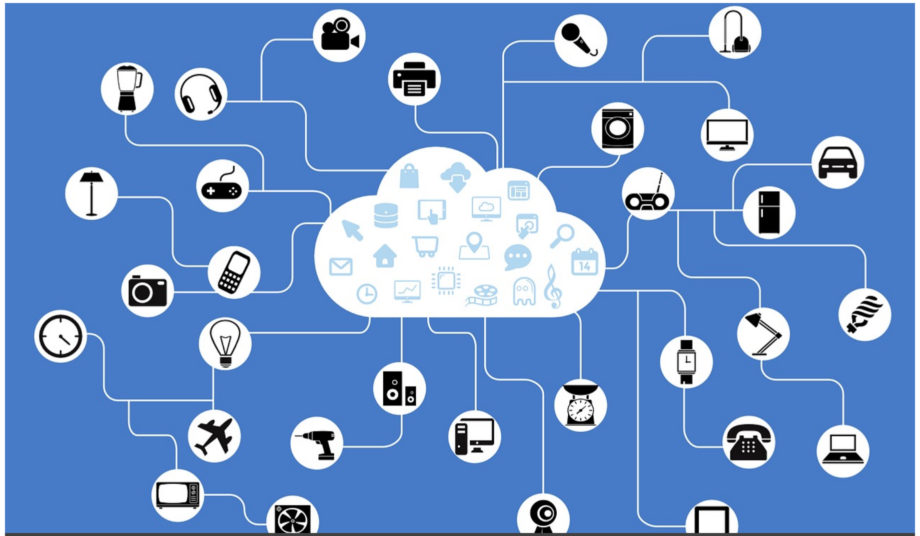
Algorithms are trained across a federation of multiple decentralized devices.

Effectively train a global/centralized model without compromising on sensitive data of various users.

Transfer of model weights and updates from local devices to cloud, rather than conventional sharing of sensitive data.

Privacy Preserving; Minimal Latency; More Personalization

ON-DEVICE FEDERATED LEARNING FOR AUDIO CLASSIFICATION



- Expansive growth in usage of IoT devices.
- Significant research on ML/DL on-device for audio sensing.
- Applications of importance:
  - **Keyword Spotting**
  - **Urban Sound Classification**

PROMINENT CHALLENGES IN FEDERATED LEARNING

Privacy Concerns about sharing sensitive data to the cloud from local user devices

Low Latency between cloud and local devices

System Heterogeneities - HW/SW, Network, Power (Resource Constraints)

New Class Identification across devices

Statistical Heterogeneities

- **Label Heterogeneities**
- **Model Heterogeneities**

ANONYMIZED DATA IMPRESSIONS

- Construct anonymized data without transferring local sensitive data in a zero-shot manner [1].
- **Sample Softmax values:**
  - Create ***Class Similarity Matrix*** – similar weights between connections of penultimate layer to the nodes of the classes.

$$\mathbf{C}(i,j) = \frac{\mathbf{w}_i^T \mathbf{w}_j}{||\mathbf{w}_i|| \, ||\mathbf{w}_j||}$$

  - From Dirichlet distribution (*K* classes, Concentration param *C),* sample the softmax values,

$$\text{Softmax} = Dir(K, C)$$

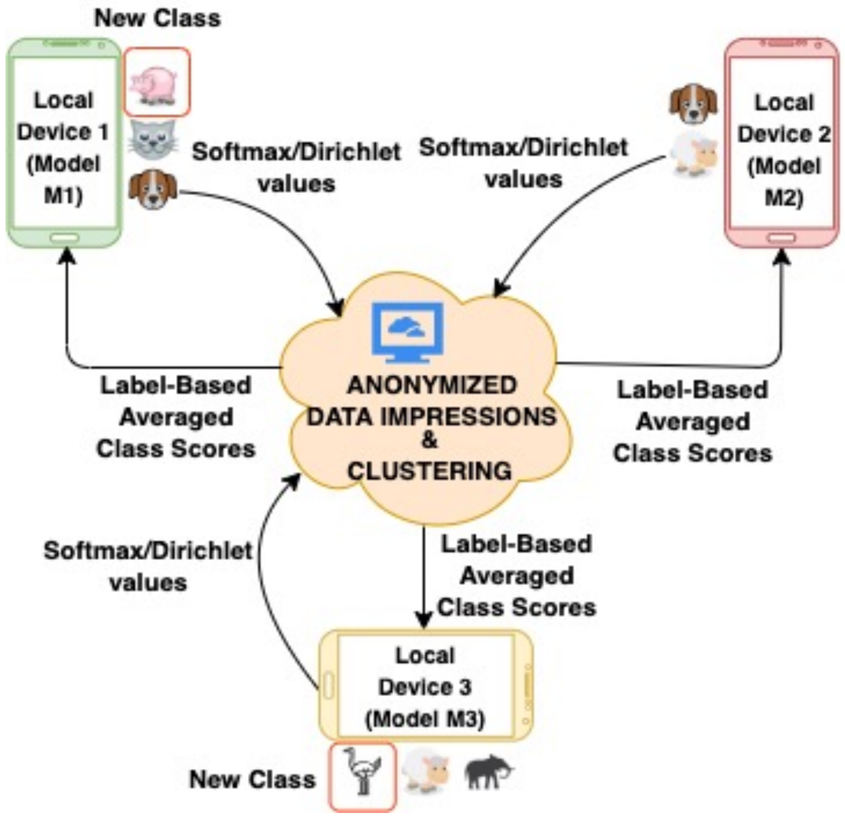
- Synthesize Data Impressions (*DI*),

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x}} L_{CE}(\mathbf{y}_i^k, \mathcal{M}(\mathbf{x}))$$

by minimizing cross-entropy loss (L_{CE}), where M is the model with random initialization and y_i^k are the softmax values sampled.

[1] Zero-Shot Knowledge Distillation in Deep Networks, ICML '19

PROPOSED SYSTEM/ ARCHITECTURE



PROPOSED FRAMEWORK

- ***Build:*** We build the model on the incoming data pertaining to each local user.
- ***Local Update:*** To obtain scores across different iterations on a single user.
  - ➢ When new classes are not reported, perform typical federated learning workflow with weighted α-update.
  - ➢ When new classes are reported, train the new model with public and newly acquired data.
- ***Global Update:*** Weighted average of scores across all users in same iteration.
  - ➢ When new classes are not reported, perform typical federated learning workflow with parameter β.
  - ➢ When new classes are reported, create Anonymized Data Impressions followed by k-medoids clustering.

Algorithm 1 Our Proposed Framework

Input: Public Dataset $\mathcal{D}_0\{x_0, y_0\}$, Private Datasets $\mathcal{D}^i_m$, Total users M, Total iterations I, LabelSet l_m for each user, Overall Public LabelSet Y,
Output: Trained Model scores f^I_G
Initialize $f^0_G = \mathbf{0}$ (Global Model Scores)
for $i = 1$ **to** I **do**
  for $m = 1$ **to** M **do**
    Build: Model $\mathcal{D}^i_m$ and predict $f_{\mathcal{D}^i_m}(x_0)$
    Local Update:
    Choice 1: New classes are not reported
    $f_{\mathcal{D}^i_m}(x_0) = f^I_G(x_0^{l_m}) + \alpha f_{\mathcal{D}^i_m}(x_0)$, where $f^I_G(x_0^{l_m})$ are global scores of l_m with m^{th} user, $\alpha = \frac{len(\mathcal{D}^i_m)}{len(\mathcal{D}_0)}$
    Choice 2: New classes are reported
    Train a new model with $\mathcal{D}_0$ and $\mathcal{D}^i_m$ (new data) together, and send weights of the last layer $(\mathbf{W}^i_m)$ to global user.
  end for
  Global Update:
  Choice 1: No user reports new classes
  Update label wise

$$f_G^{i+1} = \sum_{m=1}^{M} \beta_m f_{\mathcal{D}^i_m}(x_0), \text{ where}$$

$$\beta = \begin{cases} 1 & \text{If labels are unique} \\ \mathrm{acc}(f_{\mathcal{D}^{i+1}_m}(x_0)) & \text{if labels are not unique} \end{cases}$$

  where $\mathrm{acc}(f_{\mathcal{D}^{i+1}_m}(x_0))$ is the accuracy metric, defined by the ratio of correctly classified samples to total samples for a given local model.
  Choice 2: Any user reports new classes
  Create *Data Impressions (DI)* for each user m with weights $\mathbf{W}^i_m$ (Section 2.2). Average *DI* of all users with new classes, $\mathbf{X}^i = \sum_{m \in M_{S_k}} \mathbf{X}^i_m$, where M_{S_k} is set of users with new label k.
  Perform *k-medoids clustering* on $\mathbf{X}^i$ across M_{S_k}. Number of clusters = Number of new labels (l_{new}).
  Update public dataset with new DI $(\mathbf{X}^i)$, $\mathcal{D}_{new} = \mathcal{D}_0 \bigcup \mathbf{X}^i$, add l_{new} to l_m and Y.
end for

EXPERIMENTAL SETUP

Datasets used:

- **Google Speech Commands** (*GKWS*)
  Total: 10 keywords
  New Classes – {Stop, Go}
- **Urban Sound 8K** (*US8K*)
  Total: 10 urban sounds
  New Classes – {Siren, Street Music}

Preprocessing: Mel-frequency cepstral coefficients (MFCC) with windowing.

	User 1	User 2	User 3	Global User (Public Dataset)
Model Arch.	2-Layer CNN {16, 32} Softmax Activation	3-Layer CNN {16, 16, 32} ReLU Activation	3-Layer ANN {16, 16, 32} ReLU Activation	—
Keywords	{Yes, No, Up, Down}	{Up, Down, Left, Right}	{Left, Right, On, Off}	{Yes, No, Up, Down, Left, Right, On, Off}
Keyword Frames per Iteration	{200-300, 200-300, 200-300, 200-300}	{200-300, 200-300, 200-300, 200-300}	{200-300, 200-300, 200-300, 200-300}	{300 * 8} = 2400
Urban Sounds	{air conditioner, car horn, children playing}	{children playing, dog bark, drilling}	{drilling, engine idling, gun shot, jackhammer}	{air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer}
Sound Frames per Iteration	{40-50, 40-50, 40-50}	{40-50, 40-50, 40-50}	{40-50, 40-50, 40-50, 40-50}	{50 * 8} = 400

AVERAGE ACCURACIES ACROSS USERS

3 users, 10 FL Iterations

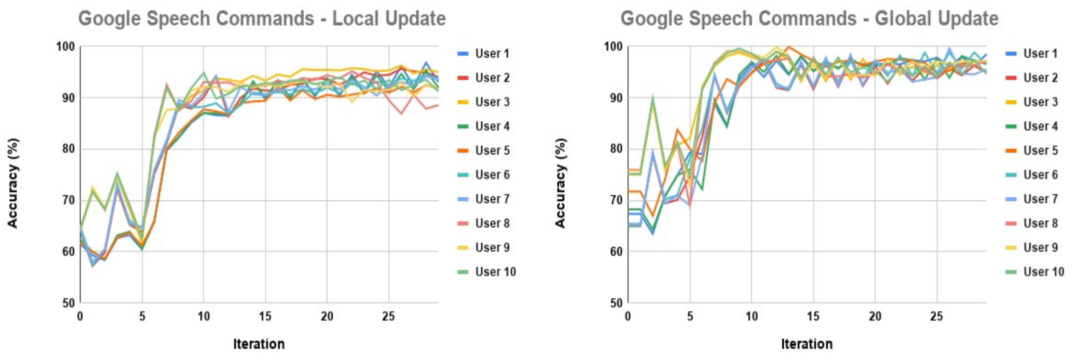
(only new classes without heterogeneities)

	GKWS			US8K		
User	**Local**	**Global**	**Increase**	**Local**	**Global**	**Increase**
User 1	89.684	93.166	3.482	76.526	80.214	3.688
User 2	91.888	95.28	3.391	75.272	77.944	2.672
User 3	91.517	94.727	3.211	77.61	81.838	4.228
Average	**91.03**	**94.391**	**3.361**	**76.469**	**80**	**3.529**

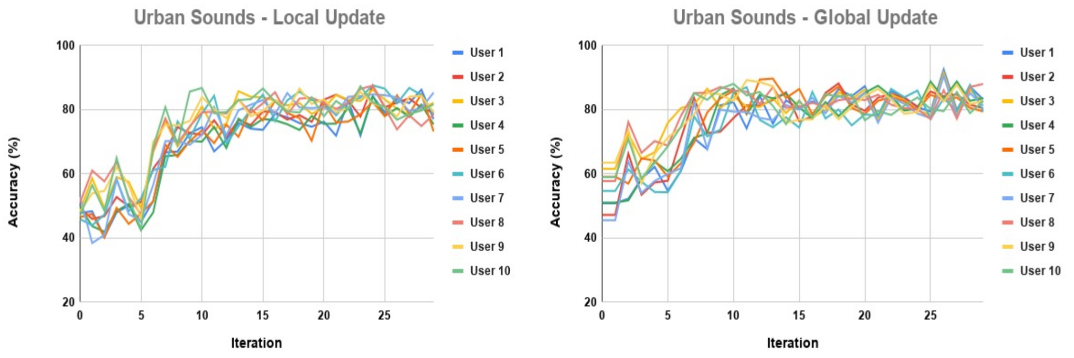
Accuracies of all global updates higher than their respective local update accuracies.

HETEROGENEITIES IN MODEL ARCHITECTURES & NEW CLASS DISTRIBUTIONS ACROSS FL USER ITERATIONS

User FL Iteration	New Model	New Class
User 1 Iteration 16	3-Layer ANN (16, 16, 32) ReLU Activation	–
User 1 Iteration 8	1-Layer CNN (16) Softmax Activation	–
User 2 Iteration 4, 6	3-Layer CNN (16, 16, 32) Softmax Activation	Stop / Siren
User 3 Iteration 5	4-Layer CNN (8, 16, 16, 32) Softmax Activation	–
User 2 Iteration 3, 7	–	Go / Street Music
User 6 Iteration 3, 5	–	Stop / Siren
User 9 Iteration 4	–	Stop / Siren



Google Speech Commands –
Accuracy vs Iterations



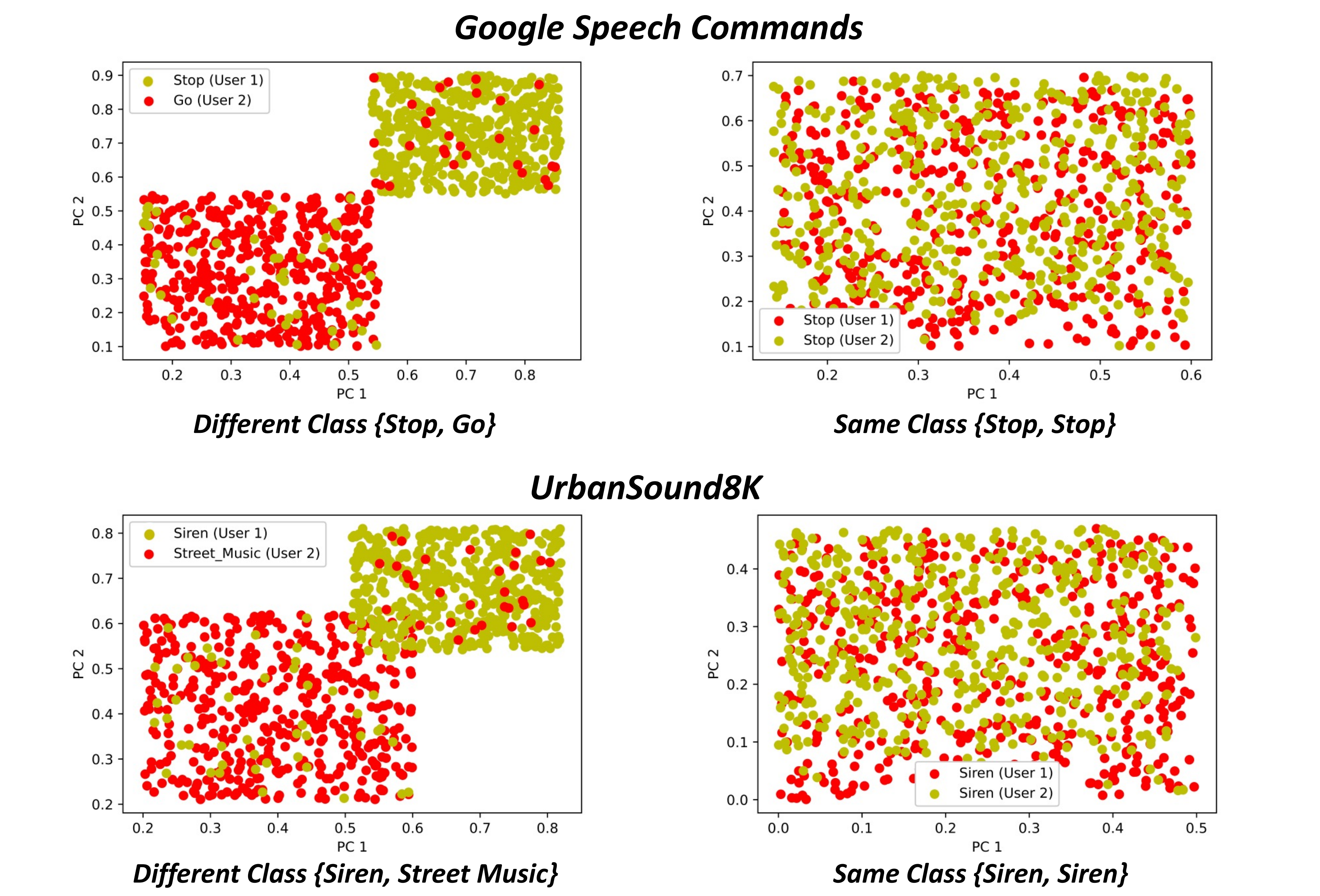
UrbanSound8K –
Accuracy vs Iterations

WITH NEW CLASSES & HETEROGENEITIES – LOCAL & GLOBAL UPDATES

10 users, 30 FL Iterations

Update	Google Speech Commands	UrbanSound8K
Local	92.5	78.24
Global	96.541	82.498
Accuracy Increase	**4.041**	**4.258**

PCA (2-dim) – UNSUPERVISED CLUSTERING WITH K-MEDOIDS



ON-DEVICE PERFORMANCE

- On-device performance of our proposed framework is experimented on a Raspberry Pi 2.
- Similar (HW/SW) specifications with that of predominant contemporary IoT/edge/mobile devices.
- The size of the models used are 520 kB, 350 kB, 270 kB for the three users.
- Clearly feasible.

Process	Computational Time
Training time per epoch in a FL iteration	1.2 sec
Inference time	11 ms

Contact

Gautham Krishna Gudur

Satheesh Kumar Perepu

Let’s chat!

THANK YOU!

QUESTIONS?