Heterogeneous Zero-Shot Federated Learning with New Classes for Audio Classification

Gautham Krishna Gudur, Global AI Accelerator, Ericsson
Satheesh Kumar Perepu, Ericsson Research

Motivation

- On-device Federated Learning – characterization from multiple user devices for effective detection of audio frames.
- Address ***new class identification*** and ***statistical heterogeneities*** challenges from multiple local devices.
- Zero-shot FL framework tested on audio classification applications like **Keyword Spotting** and **Urban Sound Classification**.

Anonymized Data Impressions

- Construct anonymized data without transferring local sensitive data from user devices 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 parameter 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.

Proposed Framework

Algorithm 1 Our Proposed Framework

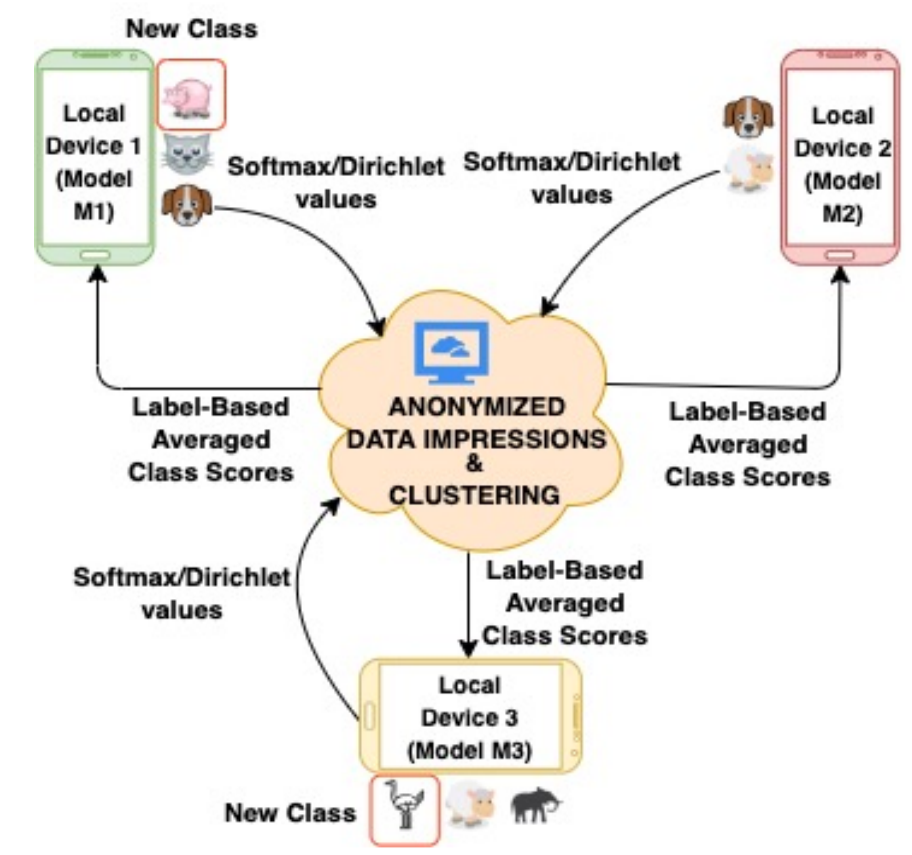
Input: Public Dataset $\mathcal{D}_0\{x_0, y_0\}$, Private Datasets $\mathcal{D}_m^i$, Total users M, Total iterations I, LabelSet l_m for each user, Overall Public LabelSet Y
Output: Trained Model scores f_G^I
Initialize $f_G^0 = \mathbf{0}$ (Global Model Scores)
for $i = 1$ **to** I **do**
  for $m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

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

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

  where $\text{acc}(f_{\mathcal{D}_m^{i+1}}(x_0))$ 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 $\mathbf{W}_m^i$. Average *DI* of all users with new classes, $\mathbf{X}^i = \sum_{m \in M_{S_k}} \mathbf{X}_m^i$, 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

Overall Block Diagram



Datasets and Preprocessing

- **Google Speech Commands (GKWS)**
  Total Classes – 10 keywords
  New Classes – {Stop, Go}
- **Urban Sound 8K (US8K)**
  Total Classes – 10 urban sounds
  New Classes – {Siren, Street Music}
- **Preprocessing**: Mel-frequency cepstral coefficients (**MFCC**), Window size – 50 ms

Experiments – Distribution of Models, Labels

	User 1	User 2	User 3	Global User (Public Dataset)
Model Arch.	2-Layer CNN {16, 32} Softmax Activation	3-Layer CNN {16, 6, 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

Heterogeneous Model Architectures, labels and Audio Frames per Iteration across all users

Iteration	New Model	New Class
User 1 Iteration 6	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 4 Iteration 3, 7	-	Go/Street Music
User 6 Iteration 5, 3	-	Stop/Siren
User 9 Iteration 4	-	Stop/Siren

Model Heterogeneities and New Classes across FL Iterations

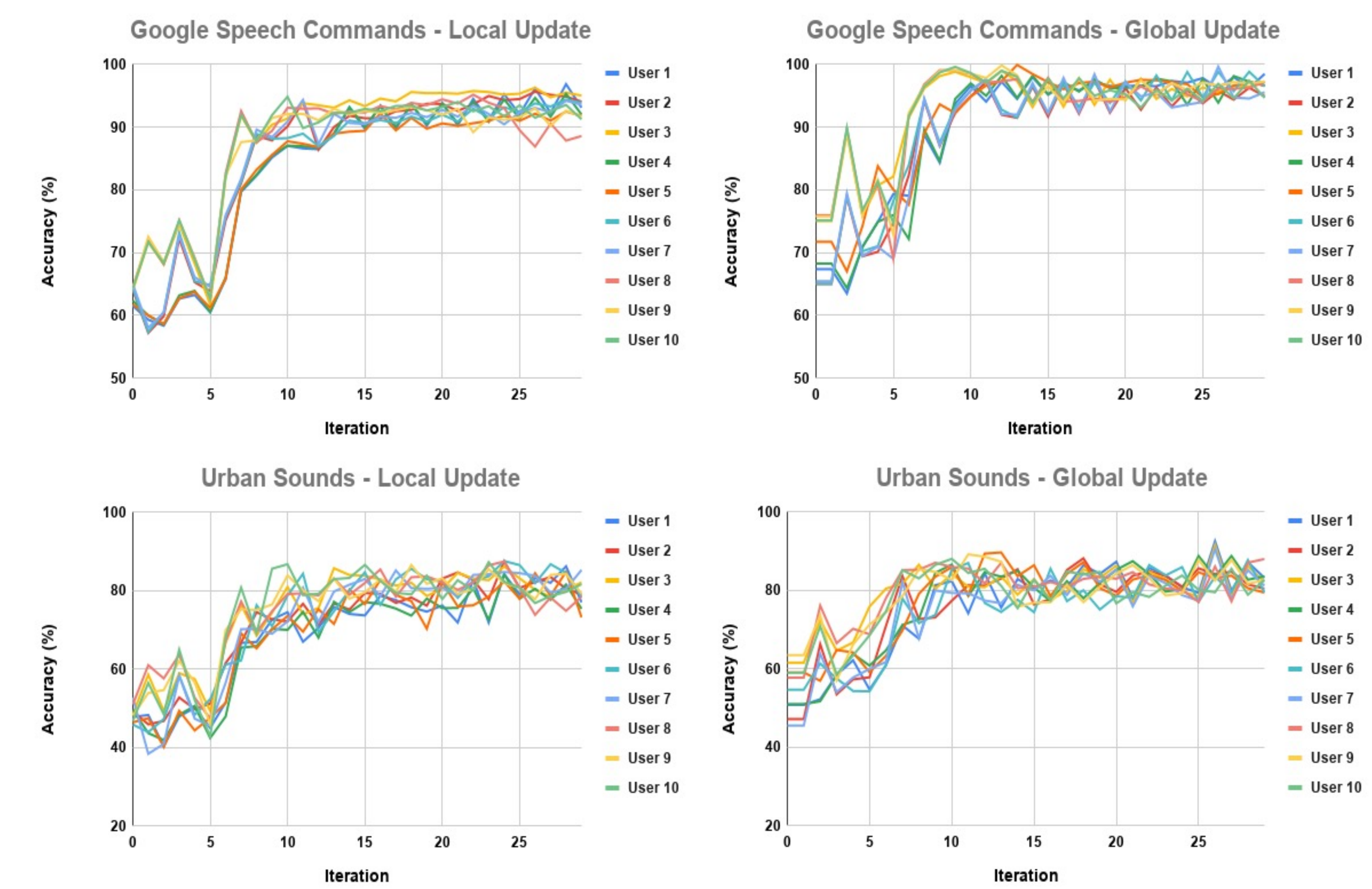
Results

	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**

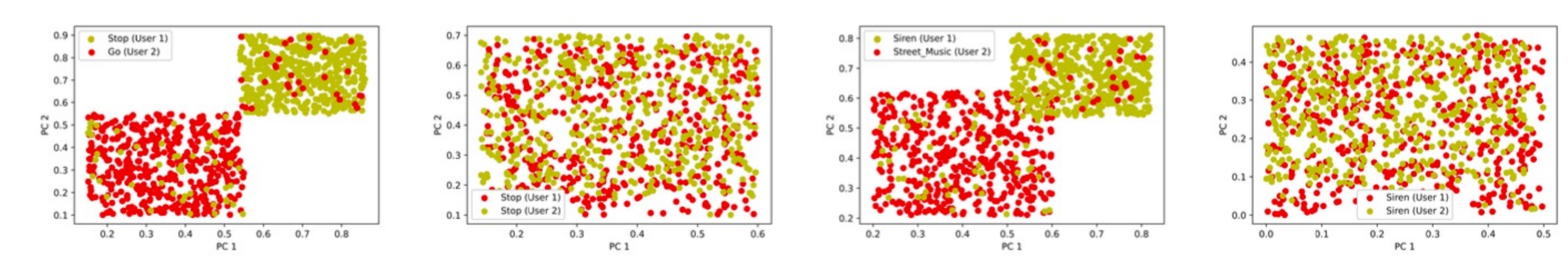
3 users and 10 FL iterations – Without heterogeneities

Update	GKWS	US8K
Local	92.5	78.24
Global	96.541	82.498
Increase	**4.041**	**4.258**

10 users and 30 FL iterations – With heterogeneities



Iterations vs Local Update and Global Update Accuracies across all 10 users and 30 FL iterations



(a) GKWS - Different Class (b) GKWS - Same Class (c) US8K - Different Class (d) US8K - Same Class

PCA (2 dimensions) with k-medoids Unsupervised Clustering of New Classes (Same/Different Classes)

On-Device Performance

- Raspberry Pi 2 used for evaluation of FL training and inference.
- The size of the models used are 520 kB, 350 kB, 270 kB respectively.

Process	Time
Training time per epoch in an FL iteration (i)	~1.2 sec
Inference time	~11 ms

On-Device Performance Metrics

References

[1] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, Anirban Chakraborty, (2019), "Zero-Shot Knowledge Distillation in Deep Networks" In: 36th International Conference on Machine Learning (ICML).

[2] Gautham Krishna Gudur, Bala Shyamala Balaji, Satheesh Kumar Perepu, (2020), "Resource-Constrained Federated Learning with Heterogeneous Labels and Models," In: The 3rd International Workshop on Artificial Intelligence of Things (AIoT), ACM SIGKDD.