Waste Not, Want Not: Service Migration-Assisted Federated Intelligence for Multi-Modality Mobile Edge Computing

Hansong Zhou[1], Shaoying Wang[1], Chutian Jiang[1], Xiaonan Zhang[1], Linke Guo[2], and Yukun Yuan[3]
[1]Department of Computer Science, Florida State University, USA
[2]Department of Electrical and Computer Engineering, Clemson University, USA
[3]Department of Computer Science & Engineering, University of Tennessee at Chattanooga, USA
{hz21e, sw22bh, cj20cn}@fsu.edu, xzhang@cs.fsu.edu, linkeg@clemson.edu, yukun-yuan@utc.edu

ABSTRACT

Future mobile edge computing (MEC) is envisioned to provide federated intelligence to delay-sensitive learning tasks with multi-modal data. Conventional horizontal federated learning (FL) suffers from high resource demand in response to complicated multi-modal models. Multi-modal FL (MFL), on the other hand, offers a more efficient approach for learning from multi-modal data. In MFL, the entire multi-modal model is split into several sub-models with each tailored to a specific data modality and trained on a designated edge. As sub-models are considerably smaller than the multi-modal model, MFL requires fewer computation resources and reduces communication time. Nevertheless, deploying MFL over MEC faces the challenges of device mobility and edge heterogeneity, which, if not addressed, could negatively impact MFL performance. In this paper, we investigate an Service Migration-assisted Mobile Multi-modal Federated Learning (SM3FL) framework, where the service migration for sub-models between edges is enabled. To effectively utilize both communication and computation resources without extravagance in SM3FL, we develop the optimal strategies of service migration and data sample collection to minimize the wall-clock time, defined as the required training time to reach the learning target. Our experiment results show that the proposed SM3FL framework demonstrates remarkable performance, surpassing other state-of-art FL frameworks via substantially reducing the computing demand by 17.5% and dramatically decreasing the wall-clock time by 25.3%.

CCS CONCEPTS

• **Computing methodologies** → **Distributed artificial intelligence**; **Multi-agent systems**.

KEYWORDS

Mobile edge computing, Multi-modality data, Federated learning

ACM Reference Format:

Hansong Zhou[1], Shaoying Wang[1], Chutian Jiang[1], Xiaonan Zhang[1], Linke Guo[2], and Yukun Yuan[3]. 2023. Waste Not, Want Not: Service Migration-Assisted Federated Intelligence for Multi-Modality Mobile Edge Computing. In *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3565287.3610277

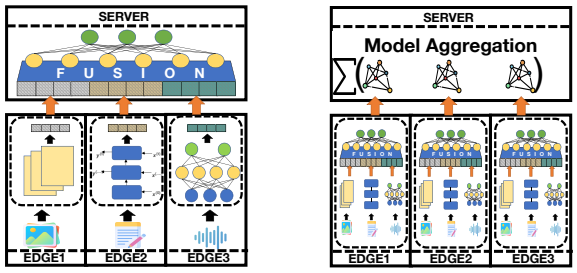
MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3610277

1 INTRODUCTION

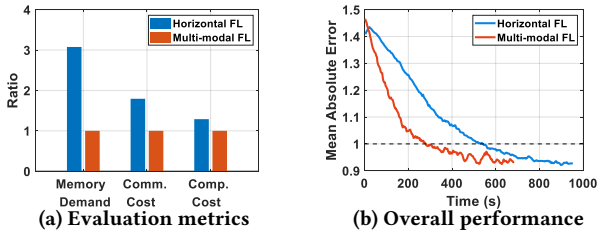
Mobile edge computing (MEC) has emerged as a promising paradigm for next-generation computing systems, which brings computation and storage resources to the network edge in proximity to mobile devices. Driven by its salient features of low latency and bandwidth saving, MEC enables a diverse range of applications such as advanced manufacturing [12], intelligent health care [27], and smart cities [29]. Future MEC is anticipated to support increasingly complex applications with high reliability and robustness, which necessitates the use of data with multiple modalities [23]. Object tracking in autonomous vehicles, for instance, requires the integration of data generated by multiple sources, such as cameras, radars, and LiDAR [4]. Similarly, in traffic surveillance tasks, time-series data, including location, altitude, and velocity, as well as vision data from cameras, are collected and further analyzed for pattern recognition [3]. Typically, the volume of multi-modal data is higher than that of single-modal data, requiring significantly more computation and storage resources [11].

Federated learning (FL) [17] is surging as a key enabler to learn from the massive collected data to provide useful insights in MEC, where edges collaboratively train a powerful learning model under the coordination of a edge server. When the conventional horizontal FL (HFL) meets multi-modal data, as shown in Fig. 1b, each edge builds a complex multi-modal model with a huge amount of parameters and further communicates it with the server back and forth. However, the limited resources at the edges significantly compromise the HFL efficiency. Even worse, inadequate edge resources pose a substantial risk of HFL failure in the time requirement. Multi-modal FL is a better FL paradigm to learn from the multi-modal data. As depicted in Fig. 1a, the entire multi-modal model in MFL is divided into several sub-models with each corresponding to a single-modal data. Each edge trains a sub-model and then outputs an intermediate result. The server deploys a top fusion model to aggregate intermediate results of all modalities and generate new gradients for training. This process is iterated until reaching the target loss. MFL shifts the training of the complicated fusion layer to the edge server, and only a sub-model is trained on the edge. Thus, MFL greatly reduces the edge training load. Meanwhile, compared to exchanging complex models with the edge server in HFL, the sub-model in MFL significantly reduces the volume of parameters to be transferred and thus saves transmission time. The comparison



(a) Multi-modal FL (b) Horizontal FL

Figure 1: Illustration of different FL frameworks



(a) Evaluation metrics (b) Overall performance

Figure 2: Performance comparison of different FL frameworks

between MFL and HFL in Fig. 2 highlights that MFL utilizes only one-third of the memory and half the communication resources required by HFL, as well as reduces half of the wall-clock time, which is defined as the required training time to achieve a target loss. Please refer to Section 6 for the settings in this experiment.

Despite its potential benefits, device mobility and edge heterogeneity pose a great challenge to efficient MFL. The movement of mobile devices, such as autonomous vehicles [13] and smartphones [2], causes various latencies in uploading data to edges across rounds. As a mobile device moves farther away, the edge may experience prolonged waiting times for data to perform local training, leading to a significant increase in wall-clock time. Moreover, edges in an area typically undertake multiple computing tasks that serve both public and government purposes, such as real-time transcription services [18] and spectrum management [20]. The sharing of edge resources is dynamic in nature and may result in resource depletion, thereby disrupting MFL operation. Edge resource sharing also exacerbates edge heterogeneity, thereby rising to the challenge of learning from multi-modal data that require diverse resources. Obviously, it is not the best option to train LSTM on resource-exhausted edges while performing simple CNN on edges with abundant resources, since it would not only waste edge resources but also increase the local training time divergence among edges. In the worst case, the whole MFL would fail caused by insufficient edge resources for training LSTM. Therefore, how to utilize both the communication and computation resources without extravagance becomes a critical issue in MFL.

In this paper, we propose a Service Migration-assisted Mobile Multi-modal Federated Learning (SM3FL) framework as shown in Fig. 3. The modality-associated sub-models, taken as the service, will be moved from one edge to another more proper one to balance the learning performance and the available resources in each round, which is the idea of "waste not, want not". To develop an optimal service migration strategy for efficient MFL with convergence guarantee, we mainly focus on these two problems: *Whether to move the sub-model between edges?* and *Which is the pair between the target edge and the sub-model for a specific data modality?* To get answers, we first provide the convergence analysis and reveal the relationship between the convergence round and data sample size. An offline wall-clock time minimization problem is then formulated taking service migration decisions and the sample collection ratio as variables. Solving this problem requires the information of mobile devices and edges during the whole training process, which is not always available. Therefore, we reformulate it as an online problem, but the above variables are coupled. To tackle this issue, we first obtain the service migration strategy by coverting the online problem into a Makespan minimization problem, which can be solved by a variant of Longest-Processing-Time-first (LPT) algorithm [8]. The optimal sample collection ratio is then determined based on its slope feature.

In light of the above discussion, we summarize our key attributions in this paper as follows:

- We propose a novel service migration-assisted multi-modal federated learning (SM3FL) framework, which is highly efficient and applicable for multi-modal learning tasks in MEC.
- We provide the convergence analysis to SM3FL under the assumption of the non-convex loss function and get the maximum estimated number of rounds to achieve a target loss.
- We formulate a wall-clock time minimization problem in SM3FL. We solve it by determining the optimal and service migration strategies and sample collection ratio in each round.
- We conduct extensive experiments to reveal how SM3FL works. We further demonstrate its advantages in reducing wall-clock time, communication cost, as well as computation demand.

2 RELATED WORK

Federated Learning over Wireless Network. The deployment of FL over wireless networks faces the challenge caused by the communication and computation resource constraints [19]. Quantization algorithms are developed to minimize the local model transmission cost between clients and the server in resource-constrained IoT networks [24]. Assisted by deep reinforcement learning (DRL) approaches, various resource allocation strategies are designed for efficient FL over wireless networks [10, 31, 33]. However, most DRL-based approaches take the offline policy and regard to only the single-modal data. By contrast, we focus on improving FL efficiency in an online manner for multi-modal learning tasks.

Edge Computing with Multi-modal Data. Multi-modal data widely exists in various applications in MEC, including but not limited to mobile crowd sensing [34], objective trajectory [4], the smart home [32], the smart city [22], and the smart health [1, 15]. For instance, Zhou *et al.* in [34] map multi-modal data into the same feature space and fuse the representations through the bi-linear pooling technique for the classification tasks. Their experiments prove that the classification with multi-modal data is more accurate than the single-modal one. In [26], Mohammad-Parsa *et al.* explore a cloud-edge framework for real-time health diagnosis. They apply an unsupervised feature extraction model for the identification of Interictal epileptic discharge (IED) and nonIED time intervals using

EEG and rs-fMRI data, respectively. However, these works place the model for processing multi-modal data on a single edge, without considering the high cost of multi-modal learning tasks and the limited edge resources in the MEC network.

Service Migration in Wireless Network. The device mobility and the edge server's limited coverage could lead to significant network performance degradation [28]. Service migration ensures high service quality by deciding when or where to migrate the service. The optimal decision of service migration in the 2D area is investigated in [25]. They form the migration as a Discrete Time Markov Decision Process and gain an optimal decision in each round with Q-matrix. Recently, a few works come out on migrating local models in FL. Aergia in [7] speeds up the FL process by migrating part of the CNN model to faster clients for efficient training. They either require devices' trajectories, contradicting our assumption of unpredictable device locations, or pre-training, unsuitable for our scenarios demanding fast service deployment.

3 SYSTEM MODEL

Fig. 3 presents a 3-layer MEC network including M mobile devices for multi-modal data collection, e.g., the smartphone and autonomous vehicles, a set of trusted edges $\mathcal{N} = \{1, 2, \cdots, N\}$ for local training, e.g., the small base stations (SBS) and roadside units (RSU), and a powerful third-party edge server, such as the remote cloud of Intelligent Transportation Systems [16]. They collaborately perform SM3FL over R rounds.

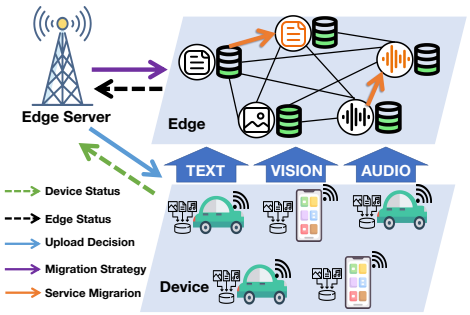


Figure 3: System overview

3.1 System Overview

In each round $r = 1, 2, \cdots, R$, mobile device $m = 1, 2, \cdots, M$ deploys multi-modal sensors for collecting the same number of K-modality data samples, written as $S_m^r = \rho^r S$, where S is the number of total samples in the environment and ρ^r is the sample collection ratio. For MEC in large areas, the number of modalities is far less than the number of deployed edges, for which we assume $K \leq N$. MFL works in the single-modality single-edge manner, that is each modality is uniquely processed by only one edge. Note that single-modality multi-edge will lead to huge communication costs of model duplication and transfer, whereas multi-modality single-edge will result in resource exhaustion on some edges but under-utilization on others.

The edge server requests mobile device locations and edge available computing resources, based on which it makes decisions on service migration and sample collection for all mobile devices. Particularly, at the beginning of each round, the edge server decides K out of N target edges to learn from K modalities, respectively, denoted as $\{n_k^r, k \in K\} \subseteq \mathcal{N}$, for which we use (n_k^r, k) to denote each edge-modality pair. Once model migration is completed, the device uploads each modality to the associated edge for local training based on the collection ratio ρ^r provided by the edge server.

3.2 Learning Model

Let $\mathcal{X} = \{X_1, \ldots, X_K\} \in \mathbb{R}^{S \times D}$ represent all available multi-modal data in the environment, where D denotes the dimension of feature space. $X_k \in \mathbb{R}^{S \times D_k}$ indicates the data of modality k and $\sum_{k=1}^{K} D_k = D$. In round r, device m collects a subset $\mathcal{X}_m^r \in \mathbb{R}^{S_m^r \times D} \subseteq \mathcal{X}$ consisting of S_m^r samples. Device m then sends each modality data to the associated target edge n_k^r. After contaminating data of modality k from all devices $X_k^r = \{X_{m,k}^r\}_{m=1}^{M} \in \mathbb{R}^{S^r \times D_k}$, the edge forwards each data sample $x_k^{r,i} \in \mathbb{R}^{D_k}$ through its neural network $\boldsymbol{\theta}_k \in \mathbb{R}^{V_k}$ to generate the embedding $h_k(\boldsymbol{\theta}_k; x_k^{r,i})$, which is then sent to the edge server. The edge server maintains a fusion model parameterized by $\boldsymbol{\theta}_f \in \mathbb{R}^{V_f}$, which is a function of K embeddings. We represent the entire model as $\boldsymbol{\Theta} = [\boldsymbol{\theta}_f, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \in \mathbb{R}^{V}$, where $V = V_f + \sum_k V_k$. The above learning model structure can be referred to Fig. 1a. The long-term object of MFL is

$$\min_{\boldsymbol{\Theta}} F(\boldsymbol{\Theta}, \mathcal{S}; \boldsymbol{y}) = \frac{1}{S} \sum_{i=1}^{S} f(\boldsymbol{\theta}_f; h_1(\boldsymbol{\theta}_1; x_1^i); \ldots; h_K(\boldsymbol{\theta}_K; x_K^i)), \quad (1)$$

where $F(\cdot)$ is the overall loss of all samples; $f(\cdot)$ denotes the loss function on the edge server; and $\boldsymbol{y} \in \mathbb{R}^S$ represents the label set.

The derivative of modality k evaluated by $\boldsymbol{\Theta}$ is given as

$$\nabla_k F(\boldsymbol{\Theta}, \mathcal{S}; \boldsymbol{y}) = \frac{1}{S} \sum_{i=1}^{S} \nabla_k f(\boldsymbol{\theta}_f; h_1(\boldsymbol{\theta}_1; x_1^i); \ldots; h_K(\boldsymbol{\theta}_K; x_K^i)). \quad (2)$$

3.3 Communication Model

The communication time per round is comprised of service migration and data uploading. We ignore the time spent in embedding transmission between edges and server due to its tiny size.

Service Migration. Denote the edge for modality k in last round as n_k^{r-1}. $n_k^{r-1} = n_k^r$ indicates no migration performed for modality k in round r. Otherwise, the local model for modality k is migrated from n_k^{r-1} to n_k^r. Given the local model size v_k in bytes, the reference distance d_0, the distance $d_{n_k^{r-1}, n_k^r}$ between n_k^{r-1} and n_k^r, and the migration speed π_n in bytes/second, the service migration time is,

$$T_{sm}^r(n_k^r, k) = \frac{v_k \log_2(1 + d_{n_k^{r-1}, n_k^r}/d_0)}{\pi_n}. \quad (3)$$

Data Uploading. In round r, mobile device m uploads samples of each modality $X_{m,k}^r$ to the associated target edge, respectively. We denote size of each sample in $X_{m,k}^r, k = 1, 2, \cdots, K$ as q_k. The overall size of $X_{m,k}^r$ is $\rho_m^r S q_k$. Each device deploys MIMO antennas for sample transmission. Given the transmission bandwidth B_m and power P_m, the time in sending $X_{m,k}^r$ is calculated as

$$T_{up,m}^r(n_k^r, k) = \frac{\rho_r S q_k}{B_m \log_2(1 + \frac{P_m h(d_{m,n_k^r}^r)}{N_0})}, \quad (4)$$

where d^r_{m,n^r_k} is the distance between device m and target edge n^r_k in round r; $h(d^r_{m,n^r_k})$ denotes the corresponding channel gain, which is positively related to the distance, such as the Rayleigh fading channel; and N_0 denotes additive white Gaussian noise (AWGN).

The communication time of edge-modality pair (n^r_k, k) is

$$T^r_{cm}(n^r_k, k) = T^r_{sm}(n^r_k, k) + \max_m T^r_{up,m}(n^r_k, k), \tag{5}$$

where max function indicates the longest time in sending modality k to target edge n^r_k among all devices. It is because the target edge cannot perform training until getting the modality from all devices.

3.4 Computation Model

Due to the resource sharing between other edge computing tasks and MFL tasks, the available computing resource on each edge varies across rounds, which is depicted by computation capability ξ^r_n in cycles/second. We assume that each edge performs E epochs of local training to update Θ every round. With GPU computation density ϑ_k cycles/sample for model $\boldsymbol{\theta_k}$, we calculate the computation demand of each epoch as $S^r\vartheta_k$ in cycles, where $S^r = \sum_{m=1}^{M} .S^r_m = \rho^r MS$ represents the total samples from all devices. The computation time on edge n^r_k in round r is calculated as

$$T^r_{cp}(n^r_k, k) = E(\frac{S^r\vartheta_k}{\xi^r_n}) = \rho^r EM(\frac{\vartheta_k}{\xi^r_n}). \tag{6}$$

We ignore the computation time spent by the edge server due to its sufficient computation resources. Another reason is that the fusion model on the server is much simpler than the learning models on edges. For example, the RNN model processing time series data on edges is much more complicated than the MLP-based fusion model on the edge server.

4 THEORETICAL ANALYSIS

In this section, we analyze the convergence rate of SM3FL. For simplicity, we use k to refer to the target edge n^r_k that trains the local model for modality k.

As in Eq. (2), edge k needs the embedding set from all target edges, including itself and the model from server $\boldsymbol{\theta}^r_f$, to calculate the partial gradient $g_k(\Theta^r, \mathcal{S}^r)$. We denote the set of required components as

$$\Phi^r = \{(\boldsymbol{\theta}^r_f; h_1(\boldsymbol{\theta}^r_1; x^i_1); \ldots; h_K(\boldsymbol{\theta}^r_K; x^i_K))\}. \tag{7}$$

We use Φ^r_{-k} to represent the subset of Φ^r without the embeddings from edge k. Denote the local gradient as $g_k(\{\boldsymbol{\theta}^r_k|\Phi^r_{-k}\}, \mathcal{S}^r)$, the expectation over $\mathcal{S}^r$ on edge k is expressed as

$$\nabla_k F(\{\boldsymbol{\theta}^r_k|\Phi^r_{-k}\}, \mathcal{S}^r) = \frac{1}{S^r}\sum_{i=1}^{S^r} g_k(\{\boldsymbol{\theta}^r_k|\Phi^r_{-k}\}, \mathcal{S}^r), \tag{8}$$

where $\{\boldsymbol{\theta}^r_k|\Phi^r_{-k}\}$ indicates that embeddings Φ^r_{-k} remain static when edge k updates its local gradient θ^r_k in round r.

The update of the global model Θ^r can be decomposed to the collaboration of the updates on the local model $\boldsymbol{\theta}^r_k$. Here we introduce a global gradient $\mathbf{G}$ represented by the set of local gradients

$$\mathbf{G} = [g_1(\{\boldsymbol{\theta}^r_1|\Phi^r_{-1}\}, \mathcal{S}^r); \ldots; g_K(\boldsymbol{\theta}^r_K|\Phi^r_{-K}\}, \mathcal{S}^r)]. \tag{9}$$

With Eq. (9), the global update becomes

$$\Theta^{r+1} = \Theta^r - \hat{\eta}^T \mathbf{G}^\mathbf{r}, \tag{10}$$

where $\hat{\eta} = [\eta_1; \ldots; \eta_K]$ is a set of learning rate w.r.t sub-models on all edges. The local update for edge k with a single step is

$$\boldsymbol{\theta}^{r+1}_k = \boldsymbol{\theta}^r_k - \eta_k g_k(\{\boldsymbol{\theta}^r_k|\Phi^r_{-k}\}, \mathcal{S}^r). \tag{11}$$

According to [5, 6, 14], we make the following common assumptions on the model and simply denote the local gradient as $g_k(\boldsymbol{\theta}^r_k)$.

Assumption 1 (*L-Smoothness*). The global model is L-smooth with positive constants L and $L_k, k \in K$, which is described as:

$$\|\nabla F(\Theta_1) - \nabla F(\Theta_1)\| \le L\|\Theta_1 - \Theta_2\|, \forall \Theta_1, \Theta_2, \tag{12}$$

$$\|\nabla_k F(\Theta_1) - \nabla_k F(\Theta_1)\| \le L_k\|\Theta_1 - \Theta_2\|, \forall \Theta_1, \Theta_2. \tag{13}$$

Assumption 2.1 (*Uniform Sample Distribution*). We assume that after sufficient rounds $\bar{r}$, the sampling distribution of each sample x^i approaches the uniform distribution as $p^r(x^i|r = \bar{r}) \approx 1/S$.

Assumption 2.2 (*Unbiased Gradient*). The expectation of the stochastic gradient is presented as

$$\mathbb{E}_{S^r}[g_k(\boldsymbol{\theta}^r_k, \mathcal{S}^r)] = \frac{1}{\sum_r S^r}\sum\nolimits_r \sum\nolimits_{\mathcal{S}_r} \nabla_k F(\boldsymbol{\theta}^r_k, \mathcal{S}). \tag{14}$$

When Assumption 2.1 holds, the expectation can be simplified as

$$\mathbb{E}_{S^r}[g_k(\boldsymbol{\theta}^r_k, \mathcal{S}^r)] = \frac{1}{S}\sum\nolimits_{\mathcal{S}_r} f(\Theta^r, \mathcal{S}) = \nabla_k F(\boldsymbol{\theta}^r_k, \mathcal{S}). \tag{15}$$

Assumption 2.3 (*Bounded Variance*). Given the unbiased gradient, we further obtain a bounded variance as

$$\mathbb{E}_{S^r}\|g_k(\boldsymbol{\theta}^r_k, \mathcal{S}) - \nabla_k F(\boldsymbol{\theta}^r_k, \mathcal{S})\|^2 \le \frac{\sigma^2}{S}. \tag{16}$$

σ is a small constant regarding variances of the sampling process.

Lemma 1. The local gradient evaluated by global model Θ^r in Eq. (2) and by local model $\boldsymbol{\theta}^r_k$ cannot be regarded as unbiased. This is because the evaluation of $g_k(\boldsymbol{\theta}^r_k)$ contains stale embeddings from other edges, which introduces bias from the local gradient on the global view. Following the proof in [14], such bias is bounded as

$$\mathbb{E}\sum_{k=1}^{K}\eta_k\|\nabla_k F(\Theta^r) - g_k(\boldsymbol{\theta}^r_k)\|^2 \le 2E^2(K+3)\sum_{k=1}^{K}\eta_k^3 L_k^2\frac{\sigma^2}{S} + 2E^2\sum_{k=1}^{K}\eta_k^3(A+3L_k^2)\mathbb{E}\|\nabla_k F(\boldsymbol{\theta}^r_k)\|^2 + 2K\frac{\sigma^2}{S}, \tag{17}$$

where $A = \sum_{k=1}^{K} L_k$ is a constant determined by local smoothness.

Theorem 1. With the above assumptions, when the minimum learning rate among all edges satisfies $\eta^{min}_k \le \frac{\sqrt{L^2+4(C+3L_k^2)}-L}{2C+6L_k^2}$, we have the following bound on the gradient of global loss

$$\frac{1}{R}\sum\nolimits_{r=0}^{R-1}\mathbb{E}\|\nabla F(\Theta^r)\|^2 \le \frac{2K}{\gamma R}\mathbb{E}[F(\Theta^0) - F(\Theta^R)] + (2+4E^2(K+3))\sum\nolimits_{k=1}^{K}\eta_k^3 L_k^2)\frac{\sigma^2}{\gamma S}. \tag{18}$$

γ is the sum of learning rates over all edges. On the right-hand side, the first item indicates the convergence rate and the second item is the residual error. Please refer to Appendix A.1 for the proof.

Remark 1 (*Impact of modalities*). The number of modalities K is positively related to both the convergence rate and the residual error. The model with more modalities will result in a faster convergence but a higher residual error, making SM3FL more challenging.

Corollary 1 (*Maximum rounds for target loss*). For the multi-modal learning model, set ε as the convergence loss. When $S > \alpha\varepsilon^{-1}$ is satisfied, the number of rounds to achieve the target loss is

$$R(S|\varepsilon) = \frac{S\beta\varepsilon^{-1}}{S - \alpha\varepsilon^{-1}}. \tag{19}$$

Here, we regard $(2+4E^2(K+3)\sum_{k=1}^{K}\eta_k^3 L_k^2)\frac{\sigma^2}{\gamma}$ and $\frac{2K}{\gamma}\mathbb{E}[F(\mathbf{\Theta}^0) - F(\mathbf{\Theta}^R)]$ as two positive constant α and β, respectively. The proof is given in Appendix A.2.

5 PROBLEM STATEMENT AND SOLUTION

In this section, we aim to minimize the wall-clock time in MFL to facilitate delay-sensitive computing tasks. Specifically, Corollary 1 suggests that a larger sample size reduces the number of rounds required for a target loss. However, increasing the number of samples collected in each round causes a higher communication and computation time cost, as illustrated in Eq. (3) and (4). This may not be feasible for MEC networks with limited resources. Hence, a sample collection strategy for efficient MFL needs careful consideration. Additionally, Eq. (3) to (6) demonstrate that the selection of the target edge for each modality significantly impacts the time spent on service migration, data uploading, and edge computation, which necessities the design of an optimal service migration strategy.

5.1 Problem Formulation

We minimize the wall-clock time w.r.t the target edge n_k^r for all modalities and the sample collection ratio ρ^r in each round by solving the following optimization problem

$$\mathbf{P1:} \min_{\rho^r, \{n_k^r\}_{k=1}^K} \sum_{r=1}^{R}\left(T_{cm}^r(n_k^r,k) + T_{cp}^r(n_k^r,k)\right), \tag{20a}$$

$$\text{s.t.} \quad n_k^r \neq n_{k'}^r, \forall k \neq k', \tag{20b}$$

$$\beta > 0, \alpha > 0, \tag{20c}$$

$$\eta_k^{min} \leq \frac{\sqrt{L^2 + 4(C + 3L_k^2)} - L}{2C + 6L_k^2}, \tag{20d}$$

$$S \geq S^r = \sum_{m=1}^{M} S_m^r \geq \alpha\varepsilon^{-1}, \tag{20e}$$

where Eq. (20b) denotes that each edge learns from a single modality; Eq. (20c) and Eq. (20d) set the range for α, β, and η_k^{min}, respectively; and Eq. (20e) limits the total number of samples every round for learning convergence, which is found in Eq. (18) and Eq. (19).

The optimal solution to **P1** requires a known round R for target loss as well as the exact dynamic mobile device locations and edge computing resources in all rounds, which, however, is agnostic in real-world scenarios. Instead, we obtain its sub-optimal solution with the optimal decision for each round, for which we reformulate **P1** into an online optimization problem

$$\min_{\rho^r, \{n_k^r\}_{k=1}^K} T = R\left(T_{cm}^r(n_k^r,k) + T_{cp}^r(n_k^r,k)\right), \tag{21a}$$

$$\text{s.t.} \quad Eq.(20b) - (20e). \tag{21b}$$

Eq. (21a) represents the estimated wall-clock time, which consists of the estimated convergence round from Corollary 1 and the duration of the current round given device locations and edge resources. The edge server estimates the wall-clock time T by supposing mobile devices stay in the same location. Thus, the optimal n_k^r and ρ^r do not change across rounds. In round r, in total $S^r = MS\rho^r$ data samples will be collected for training. We estimate the expected $R(S^r)$ by replacing S in Eq. (19) with $MS\rho^r$. After substituting $R(S^r)$ into Eq. (20a), we rewrite the above optimization problem as

$$\mathbf{P2:} \min_{\rho^r, \{n_k^r\}_{k=1}^K} \quad T = \frac{\rho^r\beta\left(T_{cm}^r(n_k^r,k) + T_{cp}^r(n_k^r,k)\right)}{\rho^r - \alpha}, \tag{22}$$

$$\text{s.t.} \quad Eq.(20b) - (20e). \tag{23}$$

We let $\beta = \beta\varepsilon^{-1}$ and $\alpha = \alpha(MS\varepsilon)^{-1}$ for simplicity.

Obviously, the joint optimization of float value ρ^r and $\{n_k^r\}_{k=1}^K$ is a Mix Integer Non-Linear Problem (MINLP). We tackle this issue by first decoupling the above two variables and solving the optimal migration strategy followed by the optimal sample collection ratio. Specifically, we enumerate edges for each modality to generate an edge-modality matrix. Each element in the matrix records the time cost for the corresponding edge-modality pair, which is a non-linear function of ρ^r. By proving the convexity w.r.t ρ^r, we can get an optimal ρ and the corresponding edge-modality preference matrix, based on which the edge-mobility pairs minimizing the wall-clock time are determined. With the given optimal $\{n_k^{r*}\}_{k=1}^K$, we further obtain the optimal ρ^{r*} from $\rho_{n,k}^{r*}, \forall n,k$ based on its feature of curvature. The details are discribed in the following.

5.2 Edge-Modality Preference Matrix

As demonstrated in Eq. (22), once the optimal sample collection ratio ρ^{r*} is determined, the optimal migration strategy $\{n_k^{r*}\}_{k=1}^K$ is deterministic. However, the coupled relationship make it impossible to directly obtain the first variable. Therefore, instead of straightforward representing the migration strategy w.r.t the globally optimal ρ^{r*}, we introduce an edge-modality preference matrix $\mathbb{T}_{NxK}(\rho_{n,k}^{r*}) = [T(\rho_{n,k}^{r*})]_{N\times K}$ to denote the estimated minimal wall-clock time for every potential edge-modality pair w.r.t its own optimal sample collection ratio $\rho_{n,k}^{r*}$ in round r. In other words, the $\rho_{n,k}^{r*}$ is their preferred sample collection ratio. Consequently, we convert the MINLP problem into a assignment problem.

From communication model described by Eq. (4) and Eq. (5) as well as computation models by Eq. (6), the initial edge-modality matrix $\mathbb{T}_{NxK}(\rho_{n,k}^r)$ is represented as

$$\mathbb{T}_{NxK}(\rho_{n,k}^r) = [T(\rho_{n,k}^r)]_{N\times K} = \frac{\rho_{n,k}^r\beta}{\rho_{n,k}^r - \alpha}[T^r(\rho_{n,k}^r)]_{N\times K}$$

$$= \frac{\rho_{n,k}^r\beta}{\rho_{n,k}^r - \alpha}[T_{sm}^r + \max_m T_{up,m}^r(\rho_{n,k}^r) + T_{cp}^r(\rho_{n,k}^r)]_{N\times K}, \tag{24}$$

where $T^r(\rho_{n,k}^r)$ is the estimated time cost of each edge-modality pair in round r. As each edge-modality pair is enumerated, we replace n_k^r and ρ^r in **P2** with n and $\rho_{n,k}^r$, respectively. Since T_{up}^r and T_{cp}^r rely on $\rho_{n,k}^r$, we integrate them by putting aside $\rho_{n,k}^r$ as follows

$$\max_m T_{up,m}^r(n,k) = \rho_{n,k}^r \max_m h(m,n,k) = \rho_{n,k}^r H(n,k), \tag{25}$$

$$T_{cp}^r(n,k) = EM(\frac{\vartheta_k}{\xi_n^r}) = \rho_{n,k}^r L(n,k), \tag{26}$$

where $h(m,n,k) = Sq_k/(B_m \log_2(1+\frac{P_m h(d^r_{m,n})}{N_o}))$ is the uploading time for device m. With Eq. (25) and Eq. (26), $T^r(\rho^r_{n,k})$ is written as

$$T^r(\rho^r_{n,k}) = \rho^r_{n,k}(H^r(n,k)+L^r(n,k)) + T^r_{sm}(n,k). \tag{27}$$

For simplicity, we denote $\Psi^{v,r}_{n,k} = H^r(n,k)+L^r(n,k)$ and $\Psi^{s,r}_{n,k} = T^{r,sm}(n,k)$. We then are able to obtain the edge-modality preference matrix by solving the following problem **P3**

$$\textbf{P3:} \min_{\rho^r_{n,k}} T(\rho^r_{n,k}) = \frac{(\rho^r_{n,k})^2\beta}{\rho^r_{n,k}-\alpha}(\Psi^{v,r}_{n,k}+\Psi^{s,r}_{n,k}), \forall n,k \tag{28a}$$

$$\text{s.t.} \quad Eq.(20b)-(20e). \tag{28b}$$

Now, we attempt to prove that there exists a optimal and unique $\rho^{r*}_{n,k}$ for each edge-modality pair. The first and second derivative of $T(\rho^r_{n,k})$ w.r.t $\rho^r_{n,k}$ in Eq. (28b) are

$$\frac{dT(n,k)}{d\rho^r} = \beta\frac{\Psi^{v,r}_{n,k}(\rho^r)^2 - 2\alpha\Psi^{v,r}_{n,k}\rho^r - \alpha\Psi^{s,r}_{n,k}}{(\rho^r-\alpha)^2}. \tag{29}$$

$$\frac{d^2T(n,k)}{d^2\rho^r} = \beta\frac{\Psi^{v,r}_{n,k}\alpha^2 + \Psi^{s,r}_{n,k}\alpha}{(\rho^r-\alpha)^3} > 0. \tag{30}$$

Since $\frac{d^2T(n,k)}{d^2\rho^r} > 0$, $T(n,k)$ is a convex function of ρ^r. Hence, by setting Eq. (29) equal to 0, we are able to obtain the optimal $\rho^{r*}_{n,k}$ for each (n,k) in $\mathbb{T}_{N\times K}$ as

$$\rho^{r*}_{n,k} = \alpha + \sqrt{\frac{\alpha^2\Psi^{v,r}_{n,k} + \alpha\Psi^{s,r}_{n,k}}{\Psi^{v,r}_{n,k}}}, \tag{31}$$

which is the solution to **P3**. By substituting $\rho^{r*}_{n,k}$ into Eq. (24), the edge-modality preference matrix obtained in round r becomes

$$\mathbb{T}_{N\times K}(\rho^{r*}_{n,k}) = \left[\left(\sqrt{\frac{\alpha\Psi^{v,r}_{n,k}}{\alpha\Psi^{v,r}_{n,k}+\Psi^{s,r}_{n,k}}} + \sqrt{\frac{\alpha^2\Psi^{v,r}_{n,k}+\alpha\Psi^{s,r}_{n,k}}{\Psi^{v,r}_{n,k}}} + 2\alpha\right)(\beta\Psi^{v,r}_{n,k}+\Psi^{s,r}_{n,k})\right]_{N\times K}, \tag{32}$$

where each element $T(\rho^{r*}_{n,k})$ represents the minimal time cost of the potential edge-modality pair (n,k).

5.3 Service Migration and Sample Collection

We now explore the optimal service migration strategy through the given edge-modality preference matrix. Since each modality is parallelly processed by edges, the overall wall-clock time is determined by the slowest modality, who has the longest estimated wall-clock time. Thus, the optimization problem **P2** can be regarded as minimizing the maximum estimated wall-clock time over a set of modalities, which drops into a Makespan minimization problem.

In accordance with the idea of LPT algorithm, which provides a polynomial time-complexity solution to the Makespan problem, we propose the following modality assignment procedure. First, we denote the set of modality existing in the preference matrix $\mathbb{T}_{N\times K}$ as $\mathcal{K}_T$. For each modality $k \in \mathcal{K}_T$, we select out a set of edges with minimum estimated wall-clock time, where

$$\{n^r_k\} = \{\arg\min_n T(n,k), k \in \mathcal{K}_T\}. \tag{33}$$

Algorithm 1 Sample collection and Server migration strategy

Input: Edge resource ξ^r_n, device distance to edges $\mathbb{D}^r_{m,n}$, target loss ε
Output: Number of samples to be collected S^r, edges to work for MFL $\mathcal{N}^r$ in this round

Initialize the selected edge set $\mathcal{N}^r = \emptyset$
Calculate the H^r, L^r, and T^r_{sm} for all potential edge-modality pairs with Eq. (25), Eq. (26) and Eq. (3).
Find the optimal $\rho^{r*}_{n,k}$ for each edge-modality pair (n,k) with Eq. (31)
Generate the preference matrix $\mathbb{T}_{N\times K}$ by Eq. (32)
for j **do** = 1, 2, …, K
  Obtain edge-modality candidates for remaining modalities by Eq. (33)
  Sort the candidates by estimated time in descending order.
  Select out the first pair (n^r_j, j) with the longest wall-clock time.
  Add the edge n^r_j into final edge list $\mathcal{N}^r$.
  Remove the corresponding modality and edge from $\mathbb{T}_{N\times K}$.
end for
Set the sample collection ratio as $\rho^{r*} = \min\{\text{argmax}_{\rho^{r*}}\{\Psi^{v,r}_{n,k}, n \in \mathcal{N}^r\}, \frac{1}{M}\}$
Obtain the optimal number of samples $S^r = \rho^r MS$

Return: S^r, $\mathcal{N}^r$

Algorithm 2 Service Migration-assisted Mobile MFL (SM3FL)

Input: Server fusion model $\boldsymbol{\theta_f}$, edge models $\{\boldsymbol{\theta_k}\}^K_{k=1}$ and learning rate set $\{\eta_k\}^K_{k=1}$, for K modalities, target loss ε, epochs per round E.
Output: Trained global model Θ, wall-clock time T_{wc}.

Training round $r = 0$. Wall-clock time $T_{wc} = 0$
while $F(\boldsymbol{\Theta^r}) > \varepsilon$ **do**
  Edges report available computing resource ξ^r_n to the edge server
  Devices report current locations to the edge server
  Edge Server determines the sample collection threshold S^r and the target edges set $\mathcal{N}^r$ through **Algorithm** 1
  Edges perform service migration based on $\mathcal{N}^r$
  Devices collect and upload data to target edges
  After receiving S^r samples from devices, the edge server and edges perform training by (9) for E epochs
  Update the wall-clock time by $T_{wc} = T_{wc} + T^r_{cm} + T^r_{cp}$
  $r = r + 1$
end while

We then sort them in a descending order as $\{n^r_k(i), i = 1, \ldots, |\mathcal{K}_T|\}$, where i is the order. Since the wall-clock time is decided by the slowest modality, the edge-modality pair represented by $\{n^r_k(0)\}$ is first settled. After that, we remove the corresponding edge and modality from the preference matrix and we have $\mathcal{K}_T = \mathcal{K}_T/k$. The above steps are repeated until all modalities are assigned to edges.

Given the optimal target edge for each modality, we now decide the optimal ρ^{r*} from $[\rho^{r*}_{n,k}]_{N\times K}$. The key challenge lies in: no matter which $\rho^{r*}_{n,k}$ is selected, it is not the optimal one for other edge-modality pairs. Hence, ρ^{r*} will always increase the wall-clock time for those edge-modality pairs. To address this issue, we minimize the total increment of wall-clock time T in all optimal edge-modality pairs. Since $\rho^r > 2\alpha$, the first derivative w.r.t ρ^r in Eq. (29) is a linear function of $\Psi^{v,r}_{n,k}$ with the positive slope, demonstrating that a higher cost of $\Psi^{v,r}_{n,k}$ leads to the faster changes of T under different $\rho^{r*}_{n,k}$. Therefore, we select the $\rho^{r*}_{n,k}$ associated with the modality that results in the highest $\Psi^{v,r}_{n,k}$. By doing this, we mitigate the increment

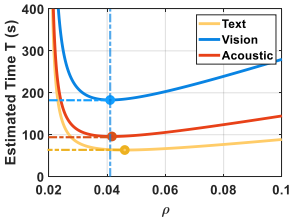


Figure 4: Sample collection

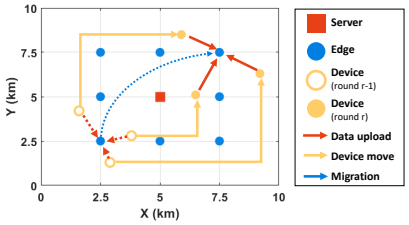


Figure 5: Experiment map

Modality	Text	Acoustic	Vision
Size per sample (bits)	500K	1M	2M
Local learning model	GloVe	Bi-LSTM	Bi-LSTM
Embedding shape	128	256	256
Number of params	2,560,000	596,480	792, 320
Density ϑ (Cycles/sample)	2000	1000	1000
Learning rate η	5e-5	0.001	0.001

Table 1: Experiment setting for CMU-MOSI dataset

of T for other edge-modality pairs. An example is shown in Fig. 4, vision is the modality that causes the highest $\Psi_{n,k}^{v,r}$ brought by the largest sample size. While selecting its corresponding ρ^{r*} as the optimal value might increase the overall wall-clock time, ρ^{r*} guarantees the minimum increment of T.

5.4 Complexity analysis

The complexity of Algorithm 1 is mainly derived from three parts, which are obtaining the upload time for each edge-modality pair in Eq. (4), determining the service migration as in Line 5-11 of Algorithm 1, and setting the sample collection ratio as in Line 12 of Algorithm 1. The max operation in Eq. (4) requires going through all devices for all edge-modality pairs, resulting a complexity of $O(NKM)$. As for the service migration strategy, Eq. (33) has a complexity of $O(NK)$ and selecting out the pair with longest estimated wall-clock time has that of $O(K)$. Since it is repeated for all K modalities, the complexity of service migration is $O(NK^2)$. Following with it, the complexity of deciding the sample collection ratio is $O(N)$. Therefore, the complexity of Algorithm 1 becomes $O(NK(M+K))$. When the $M \gg K$ as followed by the real-world scenarios, the overall complexity approximates $O(NKM)$.

6 EVALUATION

In this section, we evaluate the performance of the SM3FL framework on a desktop with the GeForce RTX 3060 graphic card.

6.1 Experiment Settings

System setting. We deploy a MEC system over a 4G cellular network in an area of $10km \times 10km$. We involve 50 mobile devices, 8 edges, and an edge server for MFL. The transmission power of each device P_m is limited to 23dBm and the bandwidth is $B_m = 20$MHz located at the center frequency of 2100MHz. The power spectral density of AWGN is set to $10^{-14.7} mW/Hz$. The WINNER II model is adopted to estimate the urban wireless channel.

As shown in Fig. 5, the coordinates of all edges are from $(2.5, 2.5)km$ to $(7.5, 7.5)km$ by step of $2.5km$ on each axis, excluding the center point. The computation capability of each edge varies every round, following the normal distribution with $f_m \sim N(10^6, 3 \times 10^6)$ cycles/seconds. We set the reference distance as $3km$ and the model migration speed as $100Mb/s$. The constant α and β in Eq. (31) is set as 0.02 and 100 in SM3FL, respectively. The settings related to the data modality are given in the followings.

Dataset and models. We evaluate the proposed SM3FL using a real-world multi-modal dataset CMU-MOSI [30] on both regression and binary classification tasks. The CMU-MOSI dataset contains language, vision, and acoustic data from 2199 videos. There are in total 1284 samples for training and 686 samples for testing. The vision and acoustic are pre-embedded to vectors of size 35 and 74, respectively. The specification of each modality is listed in Table 1. In particular, GloVe [21] is a pre-trained word-embedding model but can be fine-tuned with our dataset. Bi-LSTM is the bi-directional LSTM model. The fusion model refers to the setting in [9]. SGD optimizer and MSE loss are applied. The epoch for training all local models and fusion model is set to 5 in each round. We use Mean Absolute Error (MAE) and accuracy as evaluation metrics for regression and classification tasks, respectively.

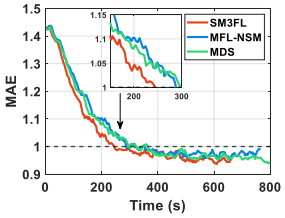
Benchmark schemes. We compare the performance of SM3FL with the following benchmarks

- FedAvg: The base HFL proposed by [17]. Since edges own the entire multi-modal models, service migration is not necessary.
- FedBCD: A communication-efficient VFL framework that allows parallel local iterations [14]. However, neither service migration nor adaptive sample collection is performed by this framework. The edges for each modality are randomly distributed at the beginning and fixed during training, denoted as $\{n_k^0\}_{k=1}^K$.
- MFL-NSM: The variant of SM3FL where no service migration is enabled. The edge-modality pairs are fixed as $\{n_k^0\}_{k=1}^K$ as well.
- MFL-FR: Another variant applying the service migration strategy. In MFL-FR, we directly obtain the edge-modality matrix from Eq. (24) with a fixed collection ratio without calculating Eq. (31).

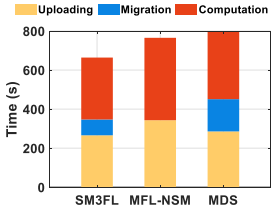
6.2 Service Migration Strategy

We give an ablation study of service migration in comparison with MFL-NSM and MDS, respectively. MDS aims at minimizing the average communication cost without considering the migration cost. In detail, the square area is divided into quarters; each edge group contains the edge at the corner and another two in the same quarter. The edge group which is the closest to the average coordinates of all devices will be selected.

In Fig. 6a, we set the target MAE as 1. The overall time in SM3FL is 243s, around 16.5% less than that in MFL-NSM. We then analyze the portion of data uploading, service migration, and computation stages in total time. Here, we take the log function to the time spent on each stage for clarity. The results are shown in Fig. 6b, where SM3FL outperforms all other benchmarks in terms of the overall time. MFL-NSM spends longer time than others in all three stages because it cannot adapt to the device mobility and thus will only select edges with sufficient computation resources. Note that the blind migration carried by MDS severely corrupts the performance. It is true that MDS reduces the data uploading time compared to NFL-NSM, but it doesn't consider the migration cost. This results in high migration costs and the longest overall time.

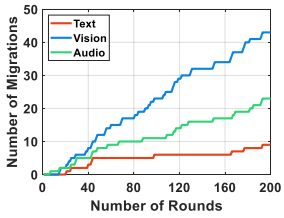


(a) MAE by time

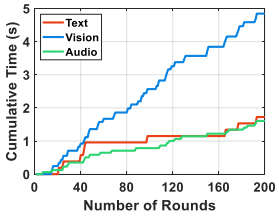


(b) Portion of different stages

Figure 6: Comparison of different migration strategies

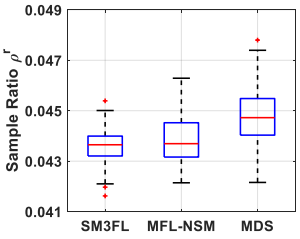


(a) Cumulative # of migrations

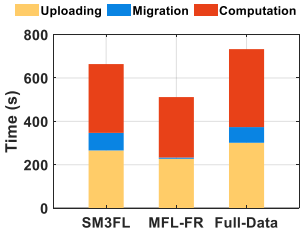


(b) Cumulative migration costs

Figure 7: Service migration of different modalities

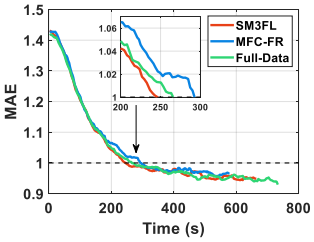


(a) Imapct of migration on ρ^r

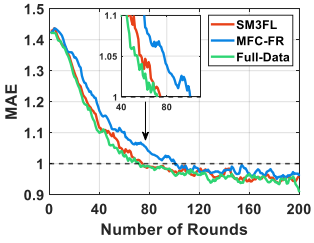


(b) Portion of different stages

Figure 8: The relationship between migration and sample collection



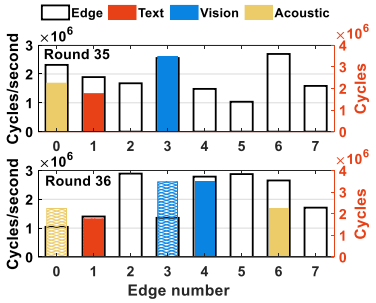
(a) MAE by time



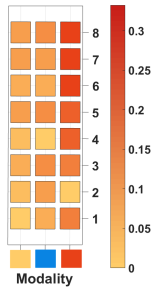
(b) MAE by rounds

Figure 9: Comparison of different sample collection decisions

We then explore how the service migration strategy exploits the available computation resources on edges. In Fig. 10, the edge distribution fits well with the computation demand in round 35. When the available edge computation resource changes in the next round, the previous edge-modality pairs may bring stragglers to the training, such as the vision and acoustic modality. SM3FL migrates their models to edge 6 and 4, respectively, to prevent performance deterioration. It is worthwhile noting that text is not migrated to the idle edge 5 with rich resources. The reason is that its migration cost is higher than the potential benefits as shown in Fig. 10b.



(a) Service migration at round 36



(b) Migration time

Figure 10: Effectiveness of service migration strategy in SM3FL

We also record the migration behaviors of each modality in SM3FL in Fig. 7. The vision modality is migrated 42 times, which is four times of text and double of acoustic because vision data has the largest size per sample among all 3 modalities and thus is more likely to become the slowest modality in each round. Hence, the benefit brought by migrating the sub-model for vision modality is higher than other choices. This makes the service migration strategy prone to migrate the sub-model for this modality. As for other modalities, the text modality is migrated the least times due to its high migration costs as shown in Fig. 10b. This explains that in Fig. 7b, although text migrates much less frequently than audio, its cumulative time cost is quite close to that of audio modality.

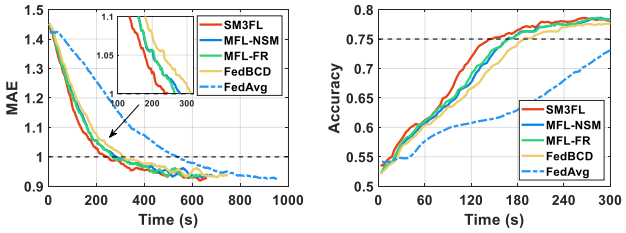
6.3 Adaptive Sample Collection Decision

As mentioned in Section 5, the decisions of sample collection ratio and edge-modality pairs are coupled. This relationship is depicted in Fig. 8. From Fig. 8a, ρ^r ranges from 0.041 to 0.046 with an average value of around 0.044. However, MDS indicates a higher demand for sample collection as shown in the third bar. It is due to the high communication cost for migration, which follows the positive relationship demonstrated by Eq. (31). An explanation for this positive relationship is that: when the estimated time cost is high, more samples are asked to be collected to improve the training quality of the current round, thus the overall time can be minimized.

We then compare SM3FL with MFL-FR and the Full-Data condition. In this comparison, $\rho^r = 0.35$ is set for MFL-FR, slightly lower than the range of SM3FL shown in Fig. 8a; Full-Data indicates that training will not start until all data are collected. From Fig. 9b, SM3FL demonstrates a similar performance with the Full-Data condition at the scale of the round but reduces 10.4% overall time. As for the wall-clock time, SM3FL is 16.9% and 10.1% less than MFC-FR and Full-Data, respectively. Note that MFC-FR has the shortest overall time but the longest wall-clock time, because its training quality of each round suffers from insufficient sample collection. Thus, we can say, our dynamic sample collection decision minimizes the data demand without sacrificing the training quality.

6.4 Performance Comparison

We compare SM3FL with benchmarks on MAE for regression and accuracy for classification. For fairness, we set ρ^r in both FedBCD and MFL-FR equal to the average of SM3FL as 0.044. As shown in Fig. 11a, the wall-clock time of FedAvg is double of the SM3FL. Compared to FedBCD, either MFL-NSM or MFL-FR can reduce the wall-clock time by around 10% with the benefits from either



(a) MAE by time **(b) Accuracy by time**

Figure 11: Comparison of different FL frameworks

Framework	Comm. C	Comp. C	Wall-Clock T	MAE
FedAvg	241.152	721.05s	593.28s	0.9253
FedBCD	178.44s	570.10s	311.45s	0.9319
MFL-NSM	**162.54s**	521.68s	284.01s	0.9324
MFL-FR	**163.83s**	475.13s	270.38s	0.9307
SM3FL	189.35s	**470.53s**	**247.57s**	**0.9185**

Table 2: Comparison on cost, wall-clock time and final MAE

adaptive sample collection or service migration strategies. On top of that, SM3FL integrates these techniques and thus surpasses all benchmark frameworks. As depicted in Table. 2, it reduces 20.8% wall-clock time from 311s in FedBCD to 247s. Moreover, benefiting from service migration strategy, SM3FL also lowers 17.5% computation cost. For the classification task with target accuracy 75%, SM3FL is still the best among all frameworks. It reduces around 25.3% wall-clock time of FedBCD. It is worth noting that FedAvg cannot achieve the target accuracy even after the double wall-clock time of SM3FL due to the high cost of the model exchange.

7 CONCLUSION

In this paper, we present a service migration-assisted multi-modal FL framework SM3FL to support delay-sensitive learning tasks with multi-modal data, which are anticipated in future MEC. In SM3FL, the entire learning model is spitted into several sub-models, each trained from single-modal data on a specific edge. The edge server then fusions those sub-models to fit multi-modal learning tasks. As learning from data with different modalities require various communication and computation resources, SM3FL increases the resource utilization efficiency by assigning each single-modal data to a proper edge. Moreover, SM3FL enables the service migration to overcome the learning efficiency degradation brought by device mobility and edge heterogeneity. Particularly, online strategies of both service migration and data sample collection are proposed to minimize the wall-clock time, with which we fully leverage the available resources for efficient training without causing failures. Extensive simulations have shown that our proposed SM3FL framework can dramatically reduce both the computation cost and the wall-clock time compared to other benchmark frameworks.

8 ACKNOWLEDGEMENTS

The work of L. Guo is supported by National Science Foundation under grant IIS-1949640, CNS-2008049, and CCF-2312616. The work of X. Zhang is supported by National Science Foundation under grant CCF-2312617.

REFERENCES

[1] Alaa Awad Abdellatif, Amr Mohamed, Carla Fabiana Chiasserini, Mounira Tlili, and Aiman Erbad. 2019. Edge Computing for Smart Health: Context-Aware Approaches, Opportunities, and Challenges. *IEEE Network* 33, 3 (2019), 196–203.

[2] Sergio Barbarossa, Stefania Sardellitti, and Paolo Di Lorenzo. 2014. Communicating While Computing: Distributed mobile cloud computing over 5G heterogeneous networks. *IEEE Signal Processing Magazine* 31, 6 (2014), 45–55.

[3] Ilker Bozcan and Erdal Kayacan. 2020. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Virtual, 8504–8510.

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Virtual, 11618–11628.

[5] Timothy Castiglia, Shiqiang Wang, and Stacy Patterson. 2022. Flexible Vertical Federated Learning with Heterogeneous Parties. https://arxiv.org/abs/2208.12672

[6] Tianyi Chen, Xiao Jin, Yuejiao Sun, and Wotao Yin. 2020. VAFL: a Method of Vertical Asynchronous Federated Learning. https://arxiv.org/abs/2007.06081

[7] Bart Cox, Lydia Y. Chen, and Jérémie Decouchant. 2022. Aergia: Leveraging Heterogeneity in Federated Learning Systems. In *Proceedings of the 23rd ACM/IFIP International Middleware Conference* (Quebec, QC, Canada) *(Middleware '22)*. Association for Computing Machinery, New York, NY, USA, 107–120.

[8] R.L Graham. 1969. Bounds on Multiprocessing Timing Anomalies. *SIAM J. Appl. Math.* 17, 2 (1969), 416–429.

[9] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 9180–9192.

[10] Mounssif Krouka, Anis Elgabli, Chaouki Ben Issaid, and Mehdi Bennis. 2022. Communication-Efficient and Federated Multi-Agent Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking* 8, 1 (2022), 311–320.

[11] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.

[12] Liangzhi Li, Kaoru Ota, and Mianxiong Dong. 2018. Deep Learning for Smart Industry: Efficient Manufacture Inspection System With Fog Computing. *IEEE Transactions on Industrial Informatics* 14, 10 (2018), 4665–4673.

[13] Shaoshan Liu, Liangkai Liu, Jie Tang, Bo Yu, Yifan Wang, and Weisong Shi. 2019. Edge Computing for Autonomous Driving: Opportunities and Challenges. *Proc. IEEE* 107, 8 (2019), 1697–1716.

[14] Yang Liu, Xinwei Zhang, Yan Kang, Liping Li, Tianjian Chen, Mingyi Hong, and Qiang Yang. 2022. FedBCD: A Communication-Efficient Collaborative Learning Framework for Distributed Features. *IEEE Transactions on Signal Processing* 70 (2022), 4277–4290.

[15] Qianxia Ma, Yongfang Nie, Jingyan Song, and Tao Zhang. 2020. Multimodal Data Processing Framework for Smart City: A Positional-Attention Based Deep Learning Approach. *IEEE Access* 8 (2020), 215505–215515.

[16] Pavel Mach and Zdenek Becvar. 2017. Mobile Edge Computing: A Survey on Architecture and Computation Offloading. *IEEE Communications Surveys and Tutorials* 19, 3 (2017), 1628–1656. https://doi.org/10.1109/COMST.2017.2682318

[17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. JMLR, Fort Lauderdale, Florida, USA, 1273–1282.

[18] Stefan Nastic, Thomas Rausch, Ognjen Scekic, Schahram Dustdar, Marjan Gusev, Bojana Koteska, Magdalena Kostoska, Boro Jakimovski, Sasko Ristov, and Radu Prodan. 2017. A Serverless Real-Time Data Analytics Platform for Edge Computing. *IEEE Internet Computing* 21, 4 (2017), 64–71.

[19] Solmaz Niknam, Harpreet S. Dhillon, and Jeffrey H. Reed. 2020. Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges. *IEEE Communications Magazine* 58, 6 (2020), 46–51.

[20] Haixia Peng, Qiang Ye, and Xuemin Shen. 2020. Spectrum Management for Multi-Access Edge Computing in Autonomous Vehicular Networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 7 (2020), 3001–3012.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543.

[22] Nikos Piperigkos, Aris S. Lalos, and Kostas Berberidis. 2021. Multi-modal cooperative awareness of connected and automated vehicles in smart cities. In *2021 IEEE International Conference on Smart Internet of Things (SmartIoT)*. IEEE, Jeju, Korea, 377–382.

[23] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*

34, 6 (2017), 96–108.
[24] Dian Shi, Liang Li, Maoqiang Wu, Minglei Shu, Rong Yu, Miao Pan, and Zhu Han. 2022. To Talk or to Work: Dynamic Batch Sizes Assisted Time Efficient Federated Learning Over Future Mobile Edge Devices. *IEEE Transactions on Wireless Communications* 21, 12 (2022), 11038–11050.
[25] Tarik Taleb, Adlen Ksentini, and Pantelis A. Frangoudis. 2019. Follow-Me Cloud: When Cloud Services Follow Mobile Users. *IEEE Transactions on Cloud Computing* 7, 2 (2019), 369–382.
[26] Md. Zia Uddin. 2019. A wearable sensor-based activity prediction system to facilitate edge computing in smart healthcare system. *J. Parallel and Distrib. Comput.* 123 (2019), 46–53.
[27] Prabal Verma and Sandeep K. Sood. 2018. Fog Assisted-IoT Enabled Patient Health Monitoring in Smart Homes. *IEEE Internet of Things Journal* 5, 3 (2018), 1789–1796.
[28] Shangguang Wang, Jinliang Xu, Ning Zhang, and Yujiong Liu. 2018. A survey on service migration in mobile edge computing. *IEEE Access* 6 (2018), 23511–23528.
[29] Wei Yu, Fan Liang, Xiaofei He, William Grant Hatcher, Chao Lu, Jie Lin, and Xinyu Yang. 2018. A Survey on the Edge Computing for the Internet of Things. *IEEE Access* 6 (2018), 6900–6919.
[30] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. https://arxiv.org/abs/1606.06259
[31] Qixun Zhang, Hao Wen, Ying Liu, Shuo Chang, and Zhu Han. 2022. Federated-Reinforcement-Learning-Enabled Joint Communication, Sensing, and Computing Resources Allocation in Connected Automated Vehicles Networks. *IEEE Internet of Things Journal* 9, 22 (2022), 23224–23240.
[32] Xiaonan Zhang, Sihan Yu, Hansong Zhou, Pei Huang, Linke Guo, and Ming Li. 2023. Signal Emulation Attack and Defense for Smart Home IoT. *IEEE Transactions on Dependable and Secure Computing* 20, 3 (2023), 2040–2057.
[33] Hansong Zhou, Sihan Yu, Xiaonan Zhang, Linke Guo, and Beatriz Lorenzo. 2022. DQN-based QoE Enhancement for Data Collection in Heterogeneous IoT Network. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, Denver, CO, 188–194.
[34] Jiancun Zhou, Tao Xu, Sheng Ren, and Kehua Guo. 2020. Two-Stage Spatial Mapping for Multimodal Data Fusion in Mobile Crowd Sensing. *IEEE Access* 8 (2020), 96727–96737.

A PROOFS OF MAIN RESULTS

A.1 Proof of Theorem 1

Applying the Smoothness assumption to the global loss F, we have

$$F(\Theta^{r+1}) - F(\Theta^r) \leq \langle \nabla F(\Theta^r), \Theta^{r+1} - \Theta^r \rangle + L/2 \|\Theta^{r+1} - \Theta^r\|^2.$$

By substituting the global update with the global gradient and taking the expectation on both sides, we have

$$\begin{aligned}
&\mathbb{E}[F(\Theta^{r+1}) - F(\Theta^r)] \leq -\mathbb{E}\langle \nabla F(\Theta^r), \hat{\eta}^T \mathbf{G}^r \rangle + L/2 \mathbb{E}\|\hat{\eta}^T \mathbf{G}^r\|^2 \\
&\leq -\sum_{k=1}^K \frac{\eta_k}{2k} (\mathbb{E}\|\nabla F(\Theta^r)\|^2 + \mathbb{E}\|\mathbf{G}^r\|^2 - \mathbb{E}\|\nabla F(\Theta^r) - \mathbf{G}^r\|^2) + \frac{L}{2}\mathbb{E}\|\hat{\eta}^T \mathbf{G}^r\|^2 \\
&\leq \mathbb{E}\sum_{k=1}^K \frac{\eta_k}{2K} \|\nabla_k F(\Theta^r) - g_k^r(\theta_k^r)\|^2 \\
&- \sum_{k=1}^K \frac{\eta_k}{2K} \mathbb{E}\|\nabla F(\Theta^r)\|^2 - \mathbb{E}\sum_{k=1}^K \frac{\eta_k - L\eta_k^2}{2K} \|\mathbf{G}^r\|^2. \qquad (34)
\end{aligned}$$

According to the definition of global gradient $\mathbf{G}$ in Eq. (9) and Assumption 2.2, we have $\mathbb{E}\|\mathbf{G}^r\|^2 \geq \mathbb{E}\sum_{k=1}^K \|\nabla_k F(\theta_k^r)\|^2$. Replace the global gradient in Eq. (34), we then obtain

$$\begin{aligned}
&\mathbb{E}[F(\Theta^{r+1}) - F(\Theta^r)] \leq \mathbb{E}\sum_{k=1}^K \frac{\eta_k}{2K} \|\nabla_k F(\Theta^r) - g_k^r(\theta_k^r)\|^2 \\
&- \sum_{k=1}^K \frac{\eta_k}{2K} \mathbb{E}\|\nabla F(\Theta^r)\|^2 - \mathbb{E}\sum_{k=1}^K \frac{\eta_k - L\eta_k^2}{2K} \|\nabla_k F(\theta_k^r)\|^2. \qquad (35)
\end{aligned}$$

By replacing the first item in the above equation with the result obtained from Lemma 1, we get

$$\begin{aligned}
&\mathbb{E}[F(\Theta^{r+1}) - F(\Theta^r)] \leq 2E^2(K+3) \sum_{k=1}^K \frac{\eta_k^3 L_k^2}{K} \frac{\sigma^2}{S} \\
&+ 2E^2 \sum_{k=1}^K \frac{\eta_k^3 (C + 3L_k^2)}{K} \mathbb{E}\|\nabla_k F(\theta_k^r)\|^2 + 2\frac{\sigma^2}{S} \\
&- \sum_{k=1}^K \frac{\eta_k}{2K} \mathbb{E}\|\nabla F(\Theta^r)\|^2 - \mathbb{E}\sum_{k=1}^K \frac{\eta_k - L\eta_k^2}{2K} \|\nabla_k F(\theta_k^r)\|^2.
\end{aligned}$$

Switch $\mathbb{E}\|\nabla F(\Theta^r)\|^2$ and $\mathbb{E}[F(\Theta^{r+1}) - F(\Theta^r)]$. Let $\gamma = \sum_{k=1}^K \eta_k$, the sum of learning rates over all edges, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\Theta^r)\|^2 \leq \frac{2K}{\gamma} \mathbb{E}[F(\Theta^r) - F(\Theta^{r+1})] + 4E^2(K+3) \sum_{k=1}^K \eta_k^3 L_k^2 \frac{\sigma^2}{\gamma} \\
&+ \frac{4E^2}{\gamma} \sum_{k=1}^K (\eta_k^3 (C + 3L_k^2) + L\eta_k^2 - \eta_k) \mathbb{E}\|\nabla_k F(\theta_k^r)\|^2 + 2\frac{K\sigma^2}{\gamma}.
\end{aligned}$$

When $\eta_k^3(C + 3L_k^2) + L\eta_k^2 - \eta_k < 0$ holds for $\forall k$, we obtain

$$\eta_k \leq \frac{\sqrt{L^2 + 4(C + 3L_k^2)} - L}{2C + 6L_k^2}, \forall k. \qquad (36)$$

The $\mathbb{E}\|\nabla F(\Theta^r)\|^2$ is bounded by

$$\begin{aligned}
\mathbb{E}\|\nabla F(\Theta^r)\|^2 &\leq \mathbb{E}[F(\Theta^r) - F(\Theta^{r+1})] \\
&+ (2 + 4E^2(K+3) \sum_{k=1}^K \eta_k^3 L_k^2) \frac{\sigma^2}{S}. \qquad (37)
\end{aligned}$$

By averaging over all global rounds $r = 0, 1, \ldots, R-1$, we get

$$\begin{aligned}
\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\|\nabla F(\Theta^r)\|^2 &\leq \frac{2K}{\gamma R} \mathbb{E}[F(\Theta^0) - F(\Theta^R)] \\
&+ (2 + 4E^2(K+3) \sum_{k=1}^K \eta_k^3 L_k^2) \frac{\sigma^2}{\gamma S}. \qquad (38)
\end{aligned}$$

The proof of Theorem 1 is finished.

A.2 Proof of Corollary 1

Referring to the work in [24], achieving target loss ε within R round can be represented with the equation given in Theorem 1 as

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\|\nabla F(\Theta^r)\|^2 \leq \varepsilon. \qquad (39)$$

We then let the right-hand side of Eq. (18) satisfy

$$\varepsilon = \frac{2K}{\gamma R} \mathbb{E}[F(\Theta^0) - F(\Theta^R)] + (2 + 4E^2(K+3) \sum_{k=1}^K \eta_k^3 L_k^2) \frac{\sigma^2}{\gamma S}. \qquad (40)$$

Regarding the number of sample S as variable, we can replace $(2 + 4E^2(K+3) \sum_{k=1}^K \eta_k^3 L_k^2) \frac{\sigma^2}{\gamma}$ with constant α and $\frac{2K}{\gamma} \mathbb{E}[F(\Theta^0) - F(\Theta^R)]$ as β. The requirement is rewritten as

$$\varepsilon = \frac{\beta}{R} + \frac{\alpha}{S}. \qquad (41)$$

We then finish the proof by reorganizing ε, R and S as

$$R = \frac{S\beta\varepsilon^{-1}}{S - \alpha\varepsilon^{-1}}. \qquad (42)$$