MultimodalHD: Federated Learning Over Heterogeneous Sensor Modalities using Hyperdimensional Computing

Quanling Zhao Xiaofan Yu Shengfan Hu Tajana Rosing
Department of Computer Science and Engineering, University of California San Diego
{quzhao,x1yu,shh042,tajana}@ucsd.edu

***Abstract*—In recent years, Federated Learning (FL) has gained increasing interests as a distributed on-device learning paradigm while preserving users' privacy. Although previous works have addressed the data and system heterogeneities in FL, the *modality heterogeneity* where clients collect data from different types of sensors (accelerometer, gyroscope, etc) is less explored. Typical FL methods assume uni-modal sensor which is not applicable in MFL due to *modality heterogeneity*. State-of-the-art MFL methods use modality-specific blocks, usually recurrent neural networks (RNN), to process each modality but they are difficult and expensive to run on edge devices. A new MFL algorithm is desired to jointly learn from heterogeneous sensor modalities under limited resources and energy. In this paper, we propose a novel hybrid framework based on Hyperdimensional Computing (HD) and deep learning, named *MultimodalHD*, to learn effectively and efficiently from edge devices with different sensor modalities. MultimodalHD uses a static HD encoder to encode raw sensory data of different modalities into high-dimensional low-precision hypervectors, after which the multimodal hypervectors are fed to an attentive fusion module for learning richer representations via inter-modality attention. Moreover, we design a proximity-based aggregation strategy at the cloud to alleviate the modality interference between clients. MultimodalHD is designed to fully utilize the strengths of both worlds: the computing efficiency of HD, and the capability of deep learning techniques to learn complex patterns. We conduct experiments on multimodal human activity recognition (HAR) datasets. Our results show that MultimodalHD delivers comparable (if not better) accuracy performance compared to state-of-the-art MFL algorithms, while being 2x - 8x more efficient in terms of training time.**

I. Introduction

With recent advancements in machine learning and edge computing platforms, Federated learning (FL) has become a promising direction for distributed training and Internet-of-Things (IoT) deployments. While previous works have studied how to address data heterogeneity (e.g. non-iid data distribution on clients [1]), system heterogeneity (e.g. varied computational and communication delays [2]), and unexpected stragglers (e.g. client drops due to various types of failures [3]) in FL, very little has been done to address *Multimodal Federated Learning (MFL)*. In contrast to uni-modal FL which assumes a single sensor modality and an identical model architecture on all clients, Multimodal FL considers heterogeneous sensor modalities, which is a more realistic setting because not all edge devices always have exactly the same sensors. For example, gyroscopes, accelerometers and magnetometer can all be used to monitor human activity, but not all of them may be available on one client (Fig. 1). Traditional FL frameworks (such as FedAvg [4]) assume uni-modal sensors and uniform model architectures, hence struggle with heterogeneous sensor modalities among clients. State-of-the-art MFL designs use separate neural networks to process each modality, which are very expensive in terms of computation [5]–[7]. They use sophisticated and complex designs, such as deep canonically correlated autoencoders [5] and split nerual network design [7] to handle modality heterogeneity in MFL setting. Edge devices are usually much more limited in terms of computational capabilities, network connectivity and are often powered by batteries. Therefore, a new lightweight FL algorithm is needed to learn both *effectively* and *efficiently* under arbitrary modality combinations at the edge.

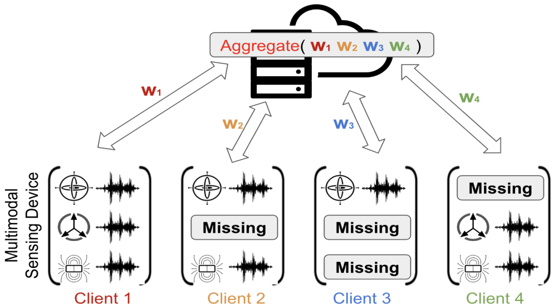

Fig. 1: An example of the **Multimodal Federated Learning** scenario with 3 different sensor modalities on 4 clients, and a cloud. Only model weights are exchanged and aggregated at the cloud.

Hyperdimensional Computing (HD) is a new brain-inspired computing paradigm where data are encoded into high-dimensional and often low-precision vectors called *hypervectors*. Cognitive tasks such as classification can be performed in HD space through a set of simple operations. Compared to traditional neural networks (NNs), HD-based designs have achieved similar accuracy in various applications while saving magnitudes of execution time and energy [8]–[12]. The efficiency of HD makes it a suitable candidate for MFL applications. However, previous works on HD multimodal fusion simply combine the hypervectors of different modalities into one and apply the common HD training [13], [14]. Such

[1] https://github.com/QuanlingZhao/DATE-24-MultimodalHD

approach failed to exploit the complex inter-modal dynamics, which has been shown to be important in many previous Deep Learning based approaches [6], [7], [15].

In this paper, we propose a novel hybrid framework named *MultimodalHD* which combines the efficiency of HD and capability of Deep Learning (DL). MultimodalHD utilizes a static HD encoder to encode the multimodal time-series data into hypervectors. We then design a novel attention module which fuses hypervectors with reinforced inter-modality correlations. Furthermore, we devise a proximity-based aggregation strategy in the cloud to alleviate interference between clients. Although our method is applicable to variety of MFL applications, in this paper we specifically focus on human activity recognition (HAR) tasks. HAR naturally comes with multimodal sensors (accelerometer, gyroscope, etc), and is often performed on small mobile devices. Thus HAR is an ideal use case that requires effective MFL while imposing strict resource constraint.

In summary, MultimodalHD is the first work that integrates HD and DL designs for effective and efficient MFL:

- MultimodalHD uses HD encoder to efficiently extract information from multimodal time-series sensor data, bypassing traditional recurrent neural network (RNN).
- MultimodalHD includes two novel DL components to improve multimodal representation learning and alleviate modality interference: attention-based fusion on local clients and proximity-based aggregation on the cloud.
- Our evaluation of three HAR datasets shows that MultimodalHD is 2x - 8x more time efficient to train compared to state-of-the-art multimodal FL baselines RNNs.

MultimodalHD is complementary to the other FL contributions, such as pruning [16] and client selection [2], thus can be easily combined with these techniques to achieve further improvements. As an example, in Sec. IV-B, we study the effects when combining MultimodalHD with various aggregation techniques.

II. Related Works

Multimodal Federated Learning (MFL). Learning from multimodal data in a federated setting has gained significant interests in recent years. In contrast to traditional FL which only focus on training uni-modal model, MFL is a more realistic and useful setting to consider especailly in IoT and HAR related tasks where clients are often multimodal and diverse. MFL adds complexity in model aggregation due to the presence of modality heterogeneity among clients and the fusing of different modalities.

Multimodal-FL [5] employs a split autoencoder on each client to learn multiple modalities without supervision. CreamFL [17] uses inter and intra-modal contrasts to complement information of the absent modality. However, both works do not have personalized models to account for client-specific pattern. MMFL [6] enables personalization with a metalearning-based approach, but its co-attention mechanism can only fuse between two modalities. Both FedMSplit [7] and Harmony [18] split the client models into modality-specific blocks to harness the modality heterogeneity. All these methods use separate Recurrent Neural Nets (RNN) as feature extractors for each modality, which are expensive to train and parallelize due to the sequential nature of RNN. In contrast, our design, MultimodalHD enables personalized multimodal federated learning while excelling in efficiency. The comparison is detailed in Table I.

TABLE I: Comparing MultimodalHD and state-of-the-art Multimodal Federated Learning works.

Method	**Modality Heterogeneity**	**Personal-ization**	**Hardware Efficiency**
[5], [17]	✓	×	×
[6]	Limited	✓	×
[7], [18]	✓	✓	×
MultimodalHD	✓	✓	✓

Hyperdimensional Computing. Although HD has been successfully applied in various scenarios [8]–[12], HD-based multimodal or federated learning are less visited. HDC-MER [13] and Schelegel *et al.* [14] bundle the encoded hypervectors from different modalities for fusion, and use the fused hypervectors for emotion recognition and driving style classification respectively. FHDnn [19] and FedHD [15] enable FL by sharing a fixed HD encoder among all clients, learning HD class hypervectors on clients from local data and aggregating the class hypervectors averagely at the cloud. To the best of the authors' knowledge, MultimodalHD is the first HD-based design for MFL.

III. Primitives

In this section, we provide background on the MFL problem definition, HD primitives and motivation for attention based multimodal fusion.

A. Multimodal Federated Learning: Problem Definition

We consider a supervised MFL problem with personalization. To model a realistic heterogeneous MFL setting, we pose no restriction on the number of modalities on a certain client. Let C_k denote a client for $k \in \{1, ..., N\}$. Specifically, $C_k = \{D_k, \theta, w_k\}$ where D_k is the labeled multimodal dataset on client k, θ is a HD encoder shared among all clients, and w_k denotes personalized model weights on client k. Suppose B is the set of all modalities in the system, B_k is the set of locally available modalities on client k with $B_k \subset B$. Assuming client k has n_k local data samples, let $D_k = \{(X_i, y_i)\}_{i=1}^{n_k}$ be the local multimodal dataset where $X_i = \{x_i^{(j)} | \forall j \in B_k\}$ and y_i are the raw multimodal sample and the label respectively. Each $x_i^{(j)}$ of a sample X_i represents time-aligned unimodal sensor readings in a sliding time window of length T. Following [7], we set the objective of our MFL problem as learning a set of different but correlated model weights $\{w_1 ... w_N\}, w_1 \neq w_2 \neq ... \neq w_N$.

$$\min_{w_1 ... w_N} \sum_{k=1}^{N} (\sum_{i=1}^{n_k} f(w_k; \theta(X_i), y_i)) + \mathcal{R}(w_k, w_1 ... w_N) \qquad (1)$$

where $f(w_k; \theta(X_i), y_i)$ is a loss function defined on model weights w_k, encoded hypervector $\theta(X_i)$ and true label y_i. $\mathcal{R}$

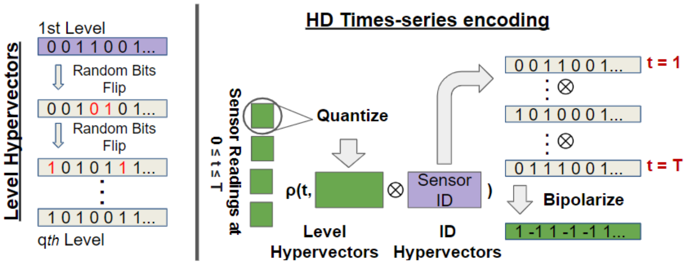

Fig. 2: Left: Generate random level hypervectors. Right: the complete HD encoding process for time-series data.

is a regularization term that forces certain level of similarity between w_k and the models from other clients, thus encouraging positive knowledge share among clients.

B. HD Primitives

Hyperdimensional computing (HD) is a lightweight computing paradigm that encodes data into hypervectors. Learning tasks can be performed through a set of simple arithmetic operations with excelling efficiency. Suppose the HD dimensionality is D. Associative learning is performed on hypervectors with well-defined operations: (1) *Bind*: $\otimes(\{0,1\}^D, \{0,1\}^D) \Rightarrow \{0,1\}^D$. Binding takes two hypervectors and returns a hypervector that is dissimilar to both operands. For binary hypervectors, binding is implemented via element-wise XOR. (2) *Bundle*: $\oplus(\mathbb{Z}^D, \mathbb{Z}^D) \Rightarrow \mathbb{Z}^D$. Bundling induce the notion of set in HD space as it returns a hypervector that is maximally similar to its constituting elements. Bundling is implemented via addition. (3) *Permute*: $\rho(t, \{0,1\}^D) \Rightarrow \{0,1\}^D$. Permutation is implemented using logical shift, t denoting the number of shifts.

Time-series HD Encoding. The first step in HD is to encode raw sample into hypervector. The goal of the HD encoder is to map high-precision, low-dimensional real valued sensor readings to low-precision, high-dimensional hypervector in HD space, while preserving the spatial and temporal patterns. In this paper, we use general encoding schemes for time-series data, i.e., spatial-temporal encoder [8], [20]. In order to represent numeric values, we generate *level hypervectors*. We begin by quantize the support of sensor reading in to q bins, and each bin is represented with a level hypervector. Starting with a random binary hypervector representing the 1st level, each subsequent level can be generated by randomly flipping $p \times D$ (p denoting flipping rate) bits from the previous level. In this way, we quantize sensor readings into hypervector while preserving the structure. Fig. 2 (left) details the generation of level hypervectors. *ID hypervectors* are randomly generated to represent different modality.

Fig. 2 (right) shows the full encoding process. Consider the encoding of $x_i^{(j)}$ which is a time series of length T. We begin by quantizing real-valued sensor readings; each quantized value is assigned to a level hypervector among $\overline{L}_1 ... \overline{L}_q$. Next, the level hypervectors are bound together with their corresponding ID hypervectors ID_j to encode information of modality j. To encode temporal information, we permute the bound hypervectors by their corresponding temporal order t in the time window. Lastly we bind all hypervectors across the temporal dimension and bipolarize it to produce the final hypervector. Formally, encoding $x_i^{(j)}$ can be expressed as:

$$\theta(x_i^{(j)}) = BP\big((\rho(1, \overline{L_1} \otimes ID_j) \otimes ... \otimes \rho(T, \overline{L_T} \otimes ID_j)\big) \quad (2)$$

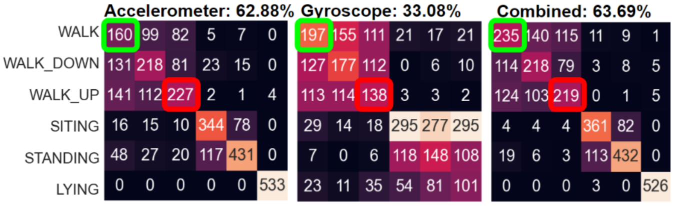

Fig. 3: The confusion matrices of the six activities in the HAR [21] dataset when using bundling as HD multimodal fusion. The green boxes highlight the case where fusion improves classification, while the red boxes indicate when fusion degrades performance.

C. Challenges of using HD in MFL

One main challenge for MFL is learning joint representation by fusing the information from different sensing modalities. Previous HD literature [13], [14] propose to simply bundle the hypervectors of different modalities to form class hypervectors and use cosine similarity metric for classification. We argue that such method does not fully utilize the multimodal data. Simply bundling hypervectors from different modalities makes the implicit assumption that all modalities are of the same importance for all situations. However, as many studies in the literature on multimodal learning have demonstrated, that is not the case [22]. We conduct a simple experiment on the HAR dataset [21] with standard HD classification pipeline [11] using bundling operation for multimodal fusion. Fig. 3 depicts the confusion matrices during classification. Only marginal performance improvement (sometimes even degradation) can be observed when incorporating new modalities. In order to capture the full inter-modality dynamics and complementary information, a more intelligent multimodal fusion method is needed, especially when modalities are many and diverse. Motivated by this observaton, we propose MultimodalHD.

IV. Proposed Framework: MultimodalHD

In this section, we present the proposed framework, MultimodalHD, whose overall structure is shown in Fig. 4. MultimodalHD uses a novel architecture that combines HD (static HD encoder) and deep learning (attentive fusion and aggregation). The design philosophy is to use HD encoder to map time-series data into feature space instead of RNN, use attention mechanism for multimodal fusion. As shown in Fig. 4, MultimodalHD first encodes the multimodal time-series data into hypervectors using a shared shared HD encoder. Afterwards, multimodal information is fused together with a novel attentive fusion design. Finally, the fused multimodal representations are passed to a multilayer perceptron (MLP) for classification. At federated level, we proposed an personalized aggregation strategy to alleviate modality interference due to modality heterogeneity at cloud. Notably, the model architecture of MultimodalHD is designed in a modality-invariant way, which means that all clients models share the

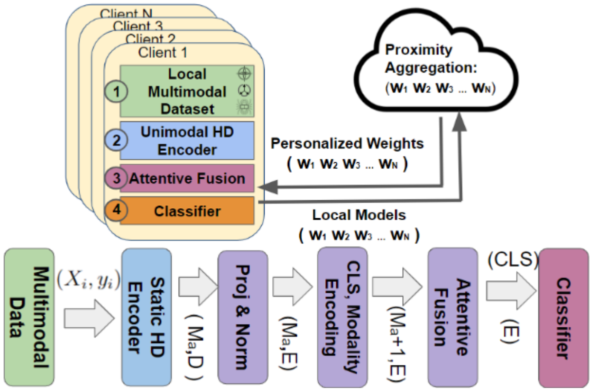

Fig. 4: Top: Overall structure of MultimodalHD. Bottom: An overview of MultimodalHD's local computation pipeline. Parentheses denoting tensor size.

same architecture and parameter space even in the presence of heterogeneous and unavailable modalities.

We detail the the two key designs in MultimodalHD: attentive multimodal fusion (Sec. IV-A) and aggregation (Sec. IV-B).

A. Attentive Multimodal Fusion

Inspired by the attention mechanism [23], we apply inter-modal attention to learn a fused representation from multi-modal hypervectors. The self-attention mechanism provides facilities for capturing dynamics in multimodal signals [24]. Inter-modality attention computation is shown in Fig. 5. By modeling modality interactions through attention score, it allows model to mix information from different modalities intelligently which previous HD-based methods failed to do.

As shown in Fig. 4, given M_a hypervector of dimension D, to ensure the the efficiency of attention computation, we first apply a trainable projection layer to reduce the dimensionality of the hypervectors to E. Gaining inspiration from the positional encoding in transformers [23], [25], we create modality encodings which are assigned to each sensing modalities and added with the corresponding projected hypervectors. Unlike the original positional encodings which are used to encode the position and order of inputs, the purpose here is to encourage the model to learn information associated with each modality itself rather than the data from that modality (e.g. same value from different modalities have different physical interpretation). Additionally, a classification (CLS) token, similar to the ViT [25] and BERT [26] architectures, are concatenated with the projected multimodal hypervectors before passing to attention computation. The output of CLS token can serve as a attentively aggregate representation from all modalities. Both modality encodings and CLS token are implemented as trainable parameters (as part of w_k on client k) and are aggregated across clients. After projecting and adding CLS/modality encodings, we have a matrix of size $((M_a + 1) \times E)$ denoted as $\mathbf{P}$. The computation of attention is shown in Fig. 5, formally as:

$$Q = W_{query} \cdot \mathbf{P},\ K = W_{key} \cdot \mathbf{P},\ V = W_{value} \cdot \mathbf{P} \quad (3)$$

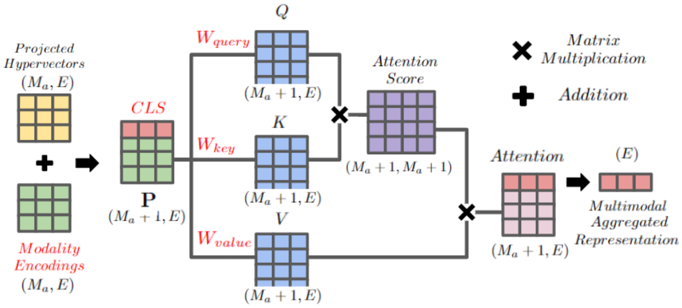

Fig. 5: The attention fusion module in MultimodalHD to fuse hypervectors from different sensing modalities. Parts in red are aggregated at cloud.

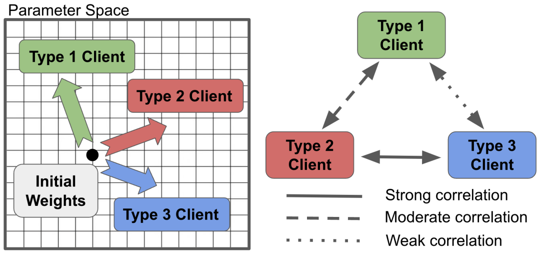

Fig. 6: An intuitive example of modality interference at the cloud that motivates the proximity-based aggregation.

$$Attention = softmax(\frac{Q \cdot K^T}{\sqrt{E}})V \quad (4)$$

Here $W_{query,key,value}$ are trainable attention weight matrices. The dimension of attention weights only depends on the embedding dimension E, while the dimension of $\hat{\mathbf{P}}$ only depends on E. Hence our attentive multimodal fusion module and classifier are invariant to the number of modalities on a client. We are able to use a uniform model architecture across all clients with heterogeneous modalities.

B. Proximity-based Cloud Aggregation

During aggregation phase, weights of attention module ($W^{(m)}_{query,key,value}$, projection layer, CLS, modality encodings) and the MLP classifiers are exchanged. At the cloud level, we propose a new proximity-based cloud aggregation strategy to mitigate interference between clients. Client models are trained on different combination of modalities, thus they are likely to be optimized towards different sub-regions in the parameter space. We call this *modality interference*.

Fig. 6 shows an intuitive example of modality interference between clients. It is more beneficial to emphasize the aggregation between strongly correlated clients (type 2 and type 3) for complementary information and information redundancy. Weakly correlated clients (type 1 and type 2) is likely to result in degraded performance due to modality mismatch. However, it is very difficult to measure modality interference between clients purely from their available modalities as we don't know the nature of the modality or what physical properties it is measuring. Hence we propose an adaptive aggregation strategy at the cloud using model weights similarity, to mitigate modality interference while allowing for personalization.

Let $\{w_1, ..., w_N\}$ denote model weights from clients $1, ..., N$ at the cloud. Let $S^{cos}_{ij} = cos(w_i, w_j)$ represents the pairwise cosine similarity between client i and j's model parameters. At the cloud, we adaptively adjust the aggregating weights based on the softmax of the cosine similarities between clients:

$$softmax(S_i^{cos})_k = \frac{\exp\left(\frac{S_{ik}^{cos}}{\tau}\right)}{\sum_{k'=1}^{N} \exp\left(\frac{S_{ik'}^{cos}}{\tau}\right)} \tag{5}$$

$$w_i^{new} = \sum_{k=1}^{N} softmax(S_i^{cos})_k \cdot w_k \tag{6}$$

Here τ is a temperature hyperparameter. Intuitively, our proximity-based aggregation strategy gives heavier weights to models from clients with similar modalities and suppresses modality interference between client pairs with dissimilar modalities with a small S_{ij}^{cos}.

V. Evaluation

A. Experimental Setup

Datasets. We use three commonly used multimodal human activity recognition datasets with continuous sensor readings, HAR [21], MHEALTH [27] and OPPORTUNITY(OPP) challenge dataset [28]. The HAR dataset is collected with smartphones contains time-series accelerometer and gyroscope readings of 30 subjects performing 6 common daily activities. Collected via wearable sensing devices, the MHEALTH dataset contains accelerometer, gyroscope and magenetometer data for 13 common activities. Following [5], we use accelerometer and gyroscope data from OPP dataset with 17 mid-level classes(Null class removed). The modality configuration in the MFL setting is reported in Table II. In all of our experiment, we use T=128 for all methods and split the datasets into individual multimodal time series samples with 75% overlap.

Baselines. In a MFL setting, we evaluate MultimodalHD in comparison to two representative state-of-the-art MFL methods. **Split-AE** [5] uses spilt-autoencoder to learn and extract correlated representations from different modalities. **FedMSplit** [7] uses separate blocks for available modalities on the clients, and update the global model based on a dynamically learned graph. We use 10 hidden units per LSTM block for both baselines. All methods are implemented using PyTorch. The important parameters in MultimodalHD are summarized in Table III. For all methods that require a classifier, we use a two-layer MLP with 25 hidden units in all experiments.

Metrics. We use weighed F1 score as our main metric for classification performance: $F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \times 100$. The F1 score for one client is the weighted average of all classes' F1 scores. For MFL, the overall performance is evaluated by the average client F1s. This metric accounts personalization on each client. For efficiency experiments, we measure and compare the time consumption per epoch of training on a Raspberry Pi model 4B.

B. Multimodal Federated Learning

We first experiment in the MFL setting where clients have different available sensory modalities as shown in Table II. The goal of personalized MFL is to achieve the best average F1 by utilizing information from different modalities and clients. Fig. 7(Top) shows the average weighted F1 score on all multimodal clients. It can be seen that, on all datasets, MultimodalHD achieves better/comparable final results compare to state of the art MFL baselines. On HAR, MultimodalHD also converges faster with regard to global communication rounds. Although FedMSplit ends up with slightly better result on HAR and MHEALTH datasets, we emphasize that the performance of MultimodalHD is close to optimal with huge efficiency advantage.

TABLE II: Sensor modality configurations in MFL on various datasets. Acc., Gyr, Mag. denote Accelerometer, Gyroscope, and Magetometer sensors in short respectively. #: Number of clients.

HAR [21]			**MHEALTH** [27]				**OPP** [28]		
Acc.	Gyr.	#	Acc.	Gyr.	Mag.	#	Acc.	Gyr.	#
✓	✓	10	✓	✓	✓	3	✓	✓	2
×	✓	10	✓	×	✓	3	✓	×	2
✓	×	10	✓	✓	×	4	-	-	-

TABLE III: Important parameters in MultimodalHD.

Param.	Description	Value (HAR, MHEALTH, OPP)
D	HD dimension	1000
E	Projected dimension	25
τ	Temperature in aggregation	$2e^{-4}$, $7e^{-4}$, $2e^{-3}$
q	Numner of quantization level	10, 100, 300
p	flipping rate	$1e^{-2}$, $2e^{-2}$, $1e^{-3}$

C. Effects of Different Federated Aggregation method

In this section we compare our proximity-based aggregation method with two commonly used aggragating methods: FedAvg [4] and FedPer [29]. FedAvg performs a weighted averaging and FedPer allows personalized weights for final MLP layers. Fig. 7(Bottom) shows the performance of FedAvg produces the least satisfactory results on HAR and MHEALTH dataset, which aligns with the modality interference issue discussed in Sec. IV-B as FedAvg equally aggregates models trained on different modalities. FedPer partially fixes this issue by allowing personalized weights, namely having the final layer not to be overwritten during federated aggregation. Our proximity-based aggregation in MultimodalHD further

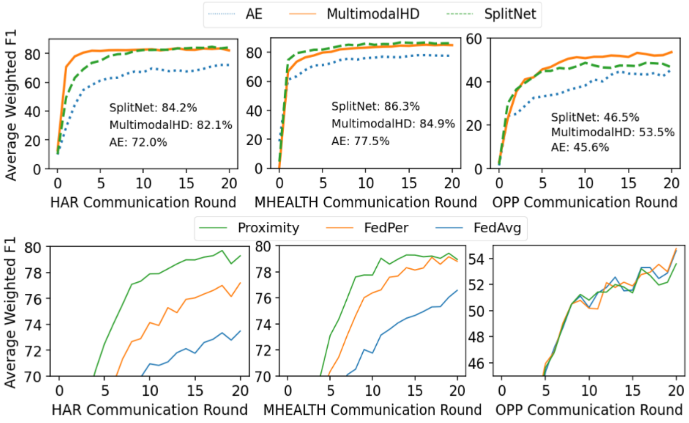

Fig. 7: Top: Under MFL setting, accuracy of MultimodalHD and baseline. Bottom: Effects of different aggregation methods in MultimodalHD.

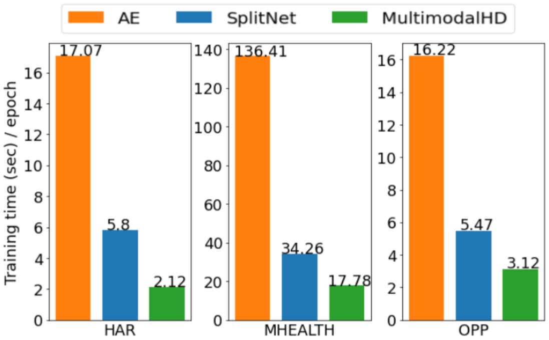

Fig. 8: Time (in seconds) consumption per epoch of training using different methods on Raspberry Pi 4B

improves on that by computing personalized model weights that takes modality interference into account when aggregating local models. The setup of OPP dataset is by far the most homogeneous one among the three, as a results, all three aggregation strategies shows similar final results.

D. *Efficiency*

Computation efficiency is a major bottleneck of edge computing. Fig. 8 summarizes the time saving capability of MultimodalHD. Compared to baselines, MultimodalHD takes significantly less time to train, achieving an improvement of 2x to 8x in terms of training cost on Raspberry Pi. The efficiency gain can be attributed to the lightweight nature of HD-based methods which avoids slow and expensive sequential computation of RNNs. We also observed simillar patterns in terms of energy consumption. Although MultimodalHD requires multimodal time-series data to be encoded first, this only needs to be done once at the start of the training, so the cost of encoding diminishes as the iterative training goes on. Since the encoder is static and highly parallelizable in hardware [11], [12], encoding can also be done while data are being collected with minimal cost.

VI. Conclusion

In this paper we propose MultimodalHD, a novel design for efficient and effective Multimodal Federated Learning (MFL) on clients with heterogeneous sensor modalities. Our hybrid model design combines the efficiency of HD and capability of Deep Learning. MultimodalHD uses a HD encoder to process multimodal sensor data efficiently and uses an attention mechanism to achieve multimodal fusion across different modalities. Additionally, we propose an aggregation method suitable for MultimodalHD to prevent modality interference. In our experiments, we systemically evaluate MultimodalHD in comparison with state-of-the-art MFL approaches on multimodal sensory datasets. We find MultimodalHD to be 2x - 8x more efficient in terms of time, while having better/comparable classification performance.

VII. Acknowledgement

This work was supported in part by National Science Foundation under Grants #2003279, #1826967, #2100237, #2112167, #1911095, #2112665, and in part by SRC under task #3021.001. This work was also supported in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA.

References

[1] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *ICML*. PMLR, 2020, pp. 5132–5143.

[2] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *OSDI*, 2021, pp. 19–35.

[3] A. Reisizadeh, I. Tziotis, H. Hassani, A. Mokhtari, and R. Pedarsani, "Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 197–205, 2022.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[5] Y. Zhao, P. Barnaghi, and H. Haddadi, "Multimodal federated learning on iot data," in *IoTDI*. IEEE, 2022, pp. 43–54.

[6] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.

[7] J. Chen and A. Zhang, "Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks," in *SIGKDD*, 2022, pp. 87–96.

[8] Y. Ni, N. Lesica, F.-G. Zeng, and M. Imani, "Neurally-inspired hyperdimensional classification for efficient and robust biosignal processing," in *ICCAD*, 2022, pp. 1–9.

[9] A. Rahimi, P. Kanerva, and J. M. Rabaey, "A robust and energy-efficient classifier using brain-inspired hyperdimensional computing," in *International symposium on low power electronics and design*, 2016, pp. 64–69.

[10] M. Imani, D. Kong, A. Rahimi, and T. Rosing, "Voicehd: Hyperdimensional computing for efficient speech recognition," in *ICRC*. IEEE, 2017, pp. 1–8.

[11] J. Morris, K. Ergun, B. Khaleghi, M. Imani, B. Aksanli, and T. Rosing, "Hydrea: Towards more robust and efficient machine learning systems with hyperdimensional computing," in *DATE*. IEEE, 2021, pp. 723–728.

[12] A. Dutta, S. Gupta, B. Khaleghi, R. Chandrasekaran, W. Xu, and T. Rosing, "Hdnn-pim: Efficient in memory design of hyperdimensional computing with feature extraction," in *the Great Lakes Symposium on VLSI 2022*.

[13] E.-J. Chang, A. Rahimi, L. Benini, and A.-Y. A. Wu, "Hyperdimensional computing-based multimodality emotion recognition with physiological signals," in *AICAS*, 2019, pp. 137–141.

[14] K. Schlegel, F. Mirus, P. Neubert, and P. Protzel, "Multivariate time series analysis for driving style classification using neural networks and hyperdimensional computing," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021.

[15] Q. Zhao, K. Lee, J. Liu, M. Huzaifa, X. Yu, and T. Rosing, "Fedhd: federated learning with hyperdimensional computing," in *Mobicom*, 2022, pp. 791–793.

[16] A. Li, J. Sun, P. Li, Y. Pu, H. Li, and Y. Chen, "Hermes: an efficient federated learning framework for heterogeneous mobile clients," in *MobiCom*, 2021, pp. 420–437.

[17] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," *arXiv preprint arXiv:2302.08888*, 2023.

[18] X. Ouyang, Z. Xie, H. Fu, S. Cheng, L. Pan, N. Ling, G. Xing, J. Zhou, and J. Huang, "Harmony: Heterogeneous multi-modal federated learning through disentangled model training," in *MobiSys*, 2023, pp. 530–543.

[19] R. Chandrasekaran, K. Ergun, J. Lee, D. Nanjunda, J. Kang, and T. Rosing, "Fhdnn: Communication efficient and robust federated learning for aiot networks," in *DAC*, 2022.

[20] I. Nunes, M. Heddes, T. Givargis, and A. Nicolau, "An extension to basis-hypervectors for learning from circular data in hyperdimensional computing," *arXiv preprint arXiv:2205.07920*, 2022.

[21] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz *et al.*, "A public domain dataset for human activity recognition using smartphones." in *Esann*, 2013.

[22] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang, "Cosmo: contrastive fusion learning with small data for multimodal human activity recognition," in *Mobicom*, 2022.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, 2017.

[24] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *NeurIPS*, vol. 34, pp. 14 200–14 213, 2021.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv*, 2020.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mhealthdroid: a novel framework for agile development of mobile health applications," in *IWAAL 2014*.

[28] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033–2042, 2013.

[29] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.