FMLU: Weighted Federated Mutual Learning with Uncertainty-aware Balancing Scheme

Kien T. Pham
The Hong Kong University of Science and Technology
COMP6311F - 20553388 - tkpham@connect.ust.hk

Abstract

Federated Learning (FL) is a technique that facilitates the collaborative training of deep learning models using decentralized data. However, the FL setting induces three types of unavoidable heterogeneities respectively residing on the data, training objective, and model architecture, presenting unique challenges to the conventional FL algorithms such as FedAvg and FedProx. Federated Mutual Learning (FML) is one of the first FL paradigms to address these problems. By introducing the intermediate meme model and leveraging Knowledge Distillation, FML enables clients to train a generalized model collaboratively and a personalized model independently on their own data distribution, and allows for flexibility in their private model designs and training objectives. In this work, we attempt to improve FML by quantifying the uncertainty levels of the meme and local models for each client and leverage it as the weighting scheme for both global model aggregation and client update stages. Extensive experiments demonstrate the effectiveness of our approach, achieving superior performance in heterogeneous scenario compared to FML and other alternatives in typical FL setting.

1. Introduction

In the era of big data, safeguarding data privacy has become increasingly crucial. It is not only a matter of public concern, but also a legal requirement enforced by regulations like the General Data Protection Regulation (GDPR) in the European Union. Consequently, the massive amount of data generated by devices (such as mobile phones, wearables, IoT devices) or within organizations (such as hospitals, companies, courts) cannot be centralized on a single server. This poses a significant challenge for deep learning. Federated Learning (FL) [23] offers a solution within the realm of deep learning. In this setting, clients can collaboratively train a shared model under the coordination of a central server, while ensuring that the data remains decentralized [9,19,32]. FL has emerged as a technique that overcomes the limitations of the ”data island” dilemma, finding extensive applications in areas such as mobile apps, autopilots, healthcare, and financial services. However, research on federated learning faces various challenges, with a particular focus in this paper on the heterogeneity problem. We categorize these challenges into three types of heterogeneities: data, objective, and model (DOM), as illustrated in Figure 1.

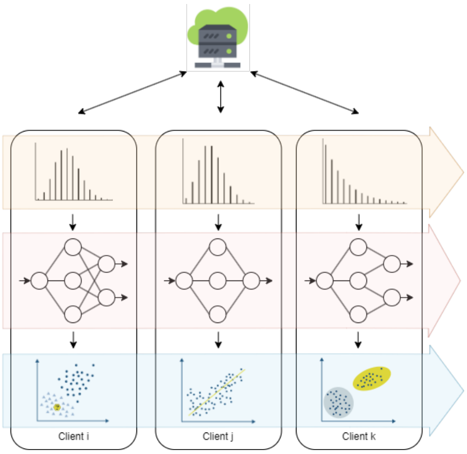

Figure 1. In typical FL, the concept of Non-IIDness of data often arises, highlighting the presence of data heterogeneity (DH). Additionally, the server and clients traditionally train a single model with the same architecture for a specific task. However, there can be a misalignment between the objectives of the server and the clients, and clients may be engaged in different tasks (OH). Consequently, clients may require the ability to train distinct models (MH) [27].

Data Heterogeneity (DH): In contrast to the independent and identically distributed (IID) data commonly found in centralized deep learning tasks, the data in a federated learning setting exhibits a Non-IID nature. This means that the data, denoted as $\{(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2), \ldots, (\mathcal{X}_n, \mathcal{Y}_n)\}$, distributed across different clients $\{C_1, C_2, \ldots, C_n\}$, may be generated from distinct distributions $(x, y) \sim \mathcal{P}_i(x, y) \neq$

$\mathcal{P}_j(x, y)$. The statistical heterogeneity resulting from this Non-IID data distribution introduces a significant decrease in accuracy compared to the IID data setting. This performance degradation can be attributed to the weight divergence [36, 37] that occurs during the stage of model weight averaging.

Objective Heterogeneity (OH): The objective of federated learning encompasses an inherent ambiguity, as the global model and local models may serve different purposes. Between the server and clients, the server aims to train a single generalized model that fits the joint distribution $\mathcal{P}_{joint}(x, y)$ for all clients and potential new participants, whereas clients aim to train personalized models that fit their distinct distributions $\mathcal{P}_i(x, y)$. Unfortunately, FedAvg [23] and FedProx [15] algorithms compromise the individuality of clients on reaching consensus due to DH. Besides, among clients each may possess similar features in their private data like visual features, yet they may have different tasks to accomplish, such as 10-category or 100-category classification. These hinder clients in benefiting from FL if they employ FedAvg or FedProx which are incapable of handling such variations.

Model Heterogeneity (MH): In FedAvg and FedProx, the global model aggregation is done by averaging weights of local models, i.e $\sum_{i=1}^{n} \frac{n_i}{n} w_i$, which is obviously incompatible for customizing local models architecture required for various scenes and tasks. Clients often possess varying hardware capabilities [5, 31], utilize different representations for local data [4, 18], and tackle different tasks [29]. Consequently, clients necessitate the ability to design their own models to cater to their specific requirements. Furthermore, the local model itself also involves privacy concerns as it constitutes private property that should be protected against theft or unauthorized access.

Federated Mutual Learning (FML) [27] is one of the first paradigms to leverage Knowledge Distillation (KD) in dealing with the three DOM heterogeneities. FML reinterprets FL as a learning method in which learned knowledge of data is transferred between global and local models, and introduces a KD method named Deep Mutual Learning (DML) [35] as the approach for local update of each client where a meme (forked from global) model and a local private model are trained together. Thanks to the intrinsic feature of KD, clients in FML setting can flexibly design and train their personalized model on their own data distribution and tasks. However, there are two drawbacks can be observed in FML. Firstly, each meme model is treated equally during global model aggregation ignoring the fact that they are trained on distinct local data. Secondly, it is troublesome to balance the training objectives of the employed DML. Inspired by [1, 2, 12, 25, 34] where uncertainty quantification has shown great success in improving performance of deep learning models, we attempt to leverage it as weighting scheme for both global model aggregation and local client update in FML, forming an extended version dubbed FMLU.

2. Related Works

Data Heterogeneity: One of the key distinctions between federated learning and distributed learning (typically referring to distributed training in data centers) lies in whether the clients' data remains fixed locally and inaccessible to others. This feature ensures data privacy but gives rise to challenges such as Non-IID and unbalanced data distributions, making the training process more difficult. Training on Non-IID data presents a difficulty in achieving high accuracy. It is explained in [36] that the accuracy reduction can be attributed to weight divergence, which introduces significant deviations from correct weight updates during the averaging stage. The authors propose a data-sharing strategy involving the creation of a small globally-shared subset of data. This strategy effectively improves accuracy, and for privacy preservation, the shared data can be extracted using distillation methods [30] or generated through generative adversarial networks (GANs) [3]. Several theoretical studies have also focused on FedAvg, specifically on convergence analysis and relaxing assumptions in the Non-IID setting [16, 17]. However, all these works primarily concentrate on training a single global model.

Objective Heterogeneity: The objective of traditional Federated Learning is to train a global model that can be utilized by all clients. However, in personalized scenarios, [33] demonstrates that some participants may not derive benefits from the global model if it is less accurate than their respective local models. This is particularly relevant for clients with small local datasets, as the global model may become overfitted to these limited data, thereby hindering its personalization capability. [8] emphasizes that optimizing solely for global accuracy can impede model personalization. Consequently, they proposes three objectives for personalized FL including developing personalized models that suit the majority of clients, creating an accurate global model that favors clients with limited private data for personalization, and achieving rapid model convergence within a small number of training rounds. In a related context, [20] puts forward a FL framework that enables the extraction of diverse types of image representations from different tasks, facilitating the integration of useful features from various vision-and-language grounding problems.

Model Heterogeneity: Smith et al. [29] introduce the MOCHA framework, which addresses challenges such as high communication costs, stragglers, and fault tolerance in multi-task FL. Khodak et al. [10] present the Average Regret-Upper-Bound Analysis (ARUBA) theoretical framework, which focuses on analyzing gradient-based meta-learning. These frameworks allow for separated model training, although the model architectures are still controlled by a central server. In contrast, Li and Wang [14] propose a decentralized framework based on knowledge distillation, enabling federated learning with independently designed models. However, this method requires a public dataset and lacks a global model for subsequent use. Moreover, it does not support new participant involvement as it may disrupt established models.

3. Preliminaries: Federated Mutual Learning

The objective of Federated Mutual Learning (FML) is to adjust the typical FL setting to mutually train a personalized and generalized models with flexibility in selected model architectures. Considering FL as knowledge transferring process between global model and client models, FML leverages Deep Mutual Learning (DML) as the distillation approach for local update of clients allowing each to train a personalized model for its own data and task. To deal with difference in model architecture, FML introduces the meme model that serves as the intermediate mean of knowledge distillation between global and local models.

Algorithm 1 Federated Mutual Learning (FML)

```
   Server Execution
1: global model $G_0$ initialization
2: for each global round $r$ do
3:     for each client $c$ do                        ▷ in parallel
4:         $M^c_{r+1} \leftarrow \text{ClientUpdate}(M^c_r)$
5:     end for
6:     Aggregate: $G_{r+1} \leftarrow \frac{1}{C}\sum_{c=1}^{C} M^c_{r+1}$
7: end for
   Client Update
1: local personalized model $L^c_0$ initialization
2: Fork: $M^c_r \leftarrow G_r$
3: for each local epoch $e$ do
4:     Conduct DML$(M^c_r, L^c_r)$ on private data $(\mathcal{X}_r, \mathcal{Y}_r)$
5: end for
```

As described in Algo. 1, the training process of FML begins with initializations of global model G_0 on the central server side, and personalized model L^c_0 on each client. Subsequently, all clients fork the global model as its medium meme model M^c_r, and conduct local update utilizing DML [35] to mutually train it and the corresponding personalized model for several epochs. The training loss functions of the two models are directly obtained from DML as:

$$\mathcal{L}_{local} = \alpha\mathcal{L}^{CE}_{local} + (1-\alpha)\mathcal{D}_{KL}(p_{meme}||p_{local}) \quad (1)$$

$$\mathcal{L}_{meme} = \beta\mathcal{L}^{CE}_{meme} + (1-\beta)\mathcal{D}_{KL}(p_{local}||p_{meme}) \quad (2)$$

where $\mathcal{L}^{CE}$ and $\mathcal{D}_{KL}$ are the Cross Entropy and Kullback Leibler (KL) Diverge, α and β are hyper-parameters to manipulate the proportion of knowledge from data or from the other model.

Benefitting from DML, knowledge transfer between meme and local models is bidirectional peer-to-peer. Intuitively, meme model retrieves global knowledge and transfers to personalized model and obtains feedback from it, and both are trained over private data. Eventually, each trained meme model is pushed to server and merged by averaging into the new generation of global model. The whole process is repeated until convgerence. From server perspective, the global model is learned by FedAvg with meme models. For clients side, the local models are trained over private data while simultaneously distilling knowledge from meme models during each communication round.

4. Methodology

FML demonstrates an adequate performance in dealing with the three DOM heterogeneities but we observe that there are two shortcomings in its settings. Firstly, the global model aggregation employed by FML directly inherits from FedAvg thus all meme models are equally weighted when updating the global model (Algo. 1 L6) despite each being trained on distinct local data with unique distribution. In another words, FML also inherits the drawbacks of FedAvg. Secondly, it is troublesome to tune for appropriate values for α and β hyper-parameters in DML at client update stage. Such problem is also briefly mentioned with a dubious suggestion in FML [27].

In attempt to deal with the two drawbacks, we propose to use a simple yet classical technique dubbed Shannon entropy [26] to quantify the uncertainty levels of the meme and local models as $\mathcal{H} = -\sum(p^k * \log(p^k))$, where p^k represents predictions on local private data and the higher the entropy the higher the uncertainty. The calculated entropy $\mathcal{H}$ are firstly leveraged to control the proportion of transferring knowledge between the two models in DML for client update, correspondingly replacing α and β. The training loss functions for the two models hence are reformulated as below:

$$\mathcal{L}_{local} = \mathcal{L}^{CE}_{local} + e^{-\mathcal{H}(p_{meme})}\mathcal{D}_{KL}(p_{meme}||p_{local}) \quad (3)$$

$$\mathcal{L}_{meme} = \mathcal{L}^{CE}_{meme} + e^{-\mathcal{H}(p_{local})}\mathcal{D}_{KL}(p_{local}||p_{meme}) \quad (4)$$

Such formulation is motivated by an intuition that the more uncertain a model is, the poorer the quality of the knowledge it possesses, thus the less it should contribute to the learning of other model in distillation process. We name this version DMLU shortened for the uncertainty-aware DML.

With the same observation, we additionally bring the calculated entropy $\mathcal{H}^c_r$ of the meme models to every global round. The lower the entropy, the more the corresponding meme model should contribute to the global model aggregation. Altogether, we compose a new version of FML dubbed Uncertainty-aware FML or FMLU that is described in Algo. 2. Note that due to limited time, we only experiment on Shannon entropy to measure the uncertainty of the meme and local models.

Algorithm 2 Uncertainty-aware FML (FMLU)

```
Server Execution
global model $G_0$ initialization
for each global round $r$ do
    for each client $c$ do                                    ▷ in parallel
        $M^c_{r+1} \leftarrow \text{ClientUpdate}(M^c_r)$
    end for
    Aggregate: $G_{r+1} \leftarrow \sum_c \left( \frac{e^{-\mathcal{H}^c_{r+1}}}{\sum_c e^{-\mathcal{H}^c_{r+1}}} M^c_{r+1} \right)$
end for
Client Update
local personalized model $L^c_0$ initialization
Fork: $M^c_r \leftarrow G_r$
for each local epoch $e$ do
    Conduct DMLU$(M^c_r, L^c_r)$ on private data $(\mathcal{X}_r, \mathcal{Y}_r)$
end for
```

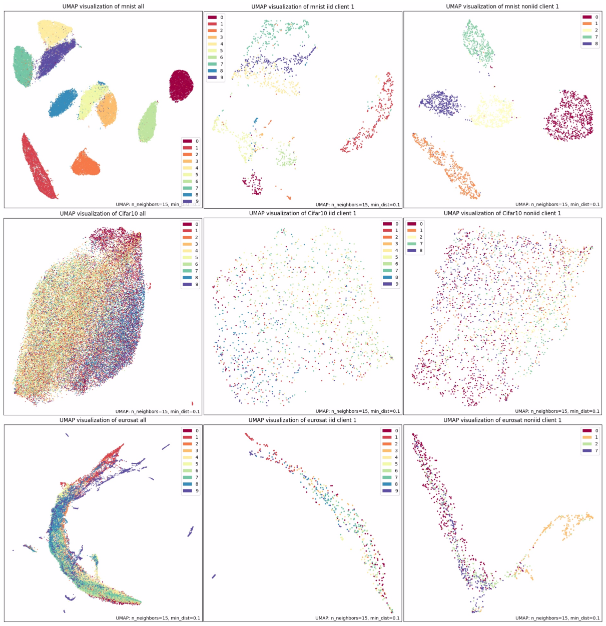

Figure 2. Visualization of data distribution for each client using UMAP [22]. From top to bottom respectively are for MNIST, CIFAR10, and EuroSAT datasets. From left to right corresponds to full distribution, IID setting, and Non-IID setting.

5. Experiments

Throughout this section, we validate the performance of FMLU over three image classification datasets under IID and Non-IID settings against standard FML as well as the two canonical FL methods FedAvg and FedProx.

5.1. Experiment Settings

Datasets: We conduct experiments on the two typical datasets MNIST and CIFAR10, and an uncommon EuroSAT. MNIST [13] is a dataset containing 60000 and 10000 28×28 gray-scale images of handwritten digits from 0 to 9 for training and testing, with 6000 and 1000 images per digit respectively. CIFAR10 [11] also consists of 50000 training and 10000 testing images categorized in 10 classes with 5000 and 1000 images per class yet all images are 3-channel colored with 32×32 resolution. The last one EuroSAT [7] provides 27000 3-channel land images of resolution 64×64 capturing from satellites and they are categorized into 10 classes with 2000 to 3000 images per class.

Federated Settings: Our experiments are configured and conducted under a simulated federated learning environment that includes 20 clients ($C = 20$) under the orchestration of a single central server. We set the total number of communication rounds $R = 50$ and the number of local epochs $E = 5$. All datasets are divided into 20 parts one for each client as its private data in IID and Non-IID settings. Specifically, for IID, each client will obtain a shuffled private data such that $\mathcal{P}_c(x, y) \approx \mathcal{P}_{join}(x, y)$. For Non-IID, we employ Dirichlet distribution at $\alpha = 0.1$ to practically allocate data among clients. The final data distribution for each client and the differences between the IID and Non-IID settings are illustrated in Fig. 2.

Training Settings: We select five different model architectures to conduct our experiments: a simple CNN with two 3×3 convolution layers (each followed with ReLu activation and 2×2 max pooling) and two FC layers, ResNet18 [6], VGG19 [28], ShuffleNetV2 [21], MobileNetV2 [24]. We slightly modify each model architecture according to each dataset image resolution. We use SGD optimizer with $lr = 0.01$ and set $batch\_size = 16$ to train each model on each dataset. Note that inheriting from FML, FMLU allows personalized model of each client to have different architecture from the meme and the global models. However, due to time limitation, we follow typical FL settings to employ the same architecture for initialization of all personalized and meme models of clients as well as global model for simplicity and convenience.

5.2. Main Results

We conduct all the training experiments and report the best Top-1 accuracies of global models in Tab. 1, 2, and 3 for MNIST, CIFAR10, and EuroSAT datasets as well as visualize them in Fig. 3 respectively.

General Observations: Firstly, we can observe from all tables that the performance of global models under Non-

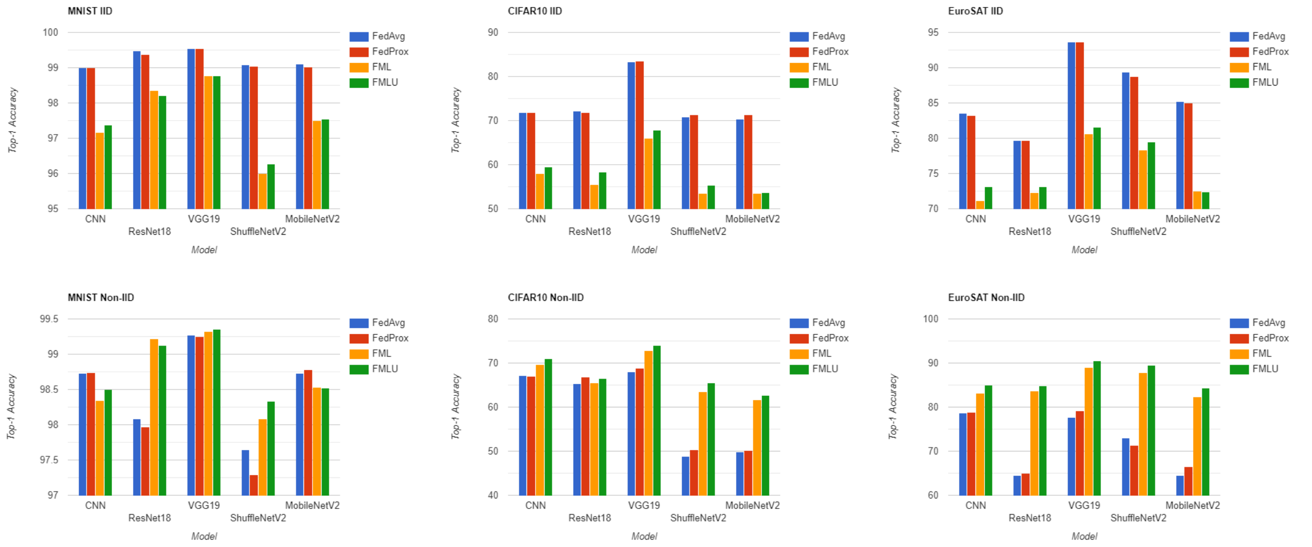

Figure 3. Visualization of every training experiments' results. From left to right respectively are for MNIST, CIFAR10, and EuroSAT dataset. First row represents IID settings while second row demonstrates Non-IID settings.

	Methods	FedAvg	FedProx	FML	FMLU
IID	CNN	99.01	99.00	97.17	97.37
	ResNet18	99.47	99.38	98.35	98.20
	VGG19	99.54	99.55	98.77	98.77
	ShuffleNetV2	99.09	99.04	96.00	96.27
	MobileNetV2	99.10	99.03	97.51	97.55
Non-IID	CNN	98.73	98.74	98.34	98.50
	ResNet18	98.08	97.97	99.22	99.12
	VGG19	99.27	99.25	99.32	99.35
	ShuffleNetV2	97.65	97.29	98.08	98.33
	MobileNetV2	98.73	98.78	98.53	98.52

Table 1. Top-1 accuracies (%) of global model in different FL settings on MNIST dataset.

	Methods	FedAvg	FedProx	FML	FMLU
IID	CNN	71.90	71.90	57.93	59.46
	ResNet18	72.25	71.89	55.49	58.31
	VGG19	83.32	83.58	66.08	67.78
	ShuffleNetV2	70.77	71.37	53.46	55.31
	MobileNetV2	70.40	71.35	53.54	53.66
Non-IID	CNN	67.10	67.06	69.64	70.96
	ResNet18	65.29	66.83	65.58	66.57
	VGG19	68.00	68.81	72.87	74.06
	ShuffleNetV2	48.78	50.32	63.52	65.47
	MobileNetV2	49.91	50.10	61.73	62.71

Table 2. Top-1 accuracies (%) of global model in different FL settings on CIFAR10 dataset.

IID setting are not as good as under IID counterpart for FedAvg and FedProx regardless of model architectures. For example, there is a significant performance degradation of

	Methods	FedAvg	FedProx	FML	FMLU
IID	CNN	83.52	83.24	71.11	73.09
	ResNet18	79.73	79.65	72.25	73.16
	VGG19	93.67	93.62	80.64	81.53
	ShuffleNetV2	89.38	88.78	78.31	79.52
	MobileNetV2	85.26	84.97	72.52	72.38
Non-IID	CNN	78.61	78.79	83.18	84.97
	ResNet18	64.44	64.97	83.59	84.83
	VGG19	77.74	79.12	89.07	90.57
	ShuffleNetV2	72.99	71.28	87.81	89.53
	MobileNetV2	64.55	66.47	82.27	84.37

Table 3. Top-1 accuracies (%) of global model in different FL settings on EuroSAT dataset.

21.99% for ShuffleNetV2 on CIFAR10 dataset in Tab. 2, or 20.71% for MobileNetV2 on EuroSat dataset in Tab. 3. This indicates that these canonical FL algorithms are susceptible to data heterogeneity. However, it is intriguing to see that it does not occur to FML and FMLU in which performances under Non-IID setting are unexpectedly better than IID one, requiring more investigation. Secondly, VGG19 significantly outperforms other architectures in every experimental settings. For instances, it attains the highest accuracies of 81.53% and 90.57% in IID and Non-IID scenarios respectively on EuroSAT dataset. Lastly, even a simple CNN can obtain an impressive performance in comparison to other more complex models despite being lightweight if FL is configured properly.

Comparison against FedAvg and FedProx: It is demonstrated in Tab. 1, 2, 3 and Fig. 3 that while FedAvg and FedProx perform better in IID scenario, FML and

FMLU dominate in every Non-IID experiments. Specifically, on one hand, FedAvg outperforms FML by 13.03% in IID-VGG19 setting. On the other hand, FMLU sets a gap of 11.45% on FedProx in Non-IID-VGG19 setting, highlighting their robustness against heterogeneous scenario. Considering the practicality of Non-IID in real world, FML and FMLU are arguably preferred over the other canonical FL approaches.

Comparison against FML: We can observe in Tab. 1, 2, 3 and Fig. 3 that FMLU achieves significantly better results compared to the baseline FML regardless of selection for model architecture and IID/Non-IID settings. For instance, in MNIST experiments, FMLU outperforms FML by 0.97% on IID and a larger gap of 1.67% on Non-IID setting avaraging on model architectures. Respectively, the improvements on CIFAR10 experiments are 2.02% and 1.17%. These advancements highlight the effectiveness of our proposed uncertainty-aware weighting scheme for global model aggregation and DMLU in local update stages.

5.3. Ablation Study

In addition to the main results, we also conduct a small ablation study to further highlight the effectiveness of the proposed uncertainty-aware weighting scheme in improving the baseline FML. We alternatively disable the weighting scheme in the local client update and the global model aggregation stage to examine its influence and show the results in 4. Note that we conduct the ablation study using simple CNN on EuroSAT dataset under both IID and Non-IID settings for simplicity. We can observe performance degradation occurred when either of the stage does not utilize the weighting scheme, and when neither of them enable the weighting scheme (FMLU degrades to FML), it produces the worst accuracy.

Client update	Server execution	IID	Non-IID
-	-	71.11	83.18
✓	-	72.83	84.95
-	✓	71.11	83.49
✓	✓	73.09	84.97

Table 4. Ablation study.

6. Conclusion

In this report, we propose Uncertainty-aware Federated Mutual Learning (FMLU), an improved version of FML leveraging uncertainty quantification to enhance the DML process of local and meme models in client update stage, and balance the weights of meme models in global model aggregation stage. Uncertainty measurement are done by calculating the Shannon entropy on the prediction logits, a simple and classical approach but it showcases great effectiveness. FMLU significantly outperforms not only FML but also other canonical FL algorithms including FedAvg and FedProx in almost every conducted experiments, especially under heterogeneous settings with Non-IIDness presented. In the future, we would like to adopt some more advanced techniques to quantify uncertainty level of involving models to further enhance the current settings.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. 2

[2] Marília Barandas, Lorenzo Famiglini, Andrea Campagner, Duarte Folgado, Raquel Simão, Federico Cabitza, and Hugo Gamboa. Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 101:101978, 2024. 2

[3] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3513–3521, 2019. 2

[4] Dashan Gao, Ce Ju, Xiguang Wei, Yang Liu, Tianjian Chen, and Qiang Yang. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *ArXiv*, abs/1909.05784, 2019. 2

[5] Chaoyang He, Murali Annavaram, and Amir Salman Avestimehr. Fednas: Federated deep learning via neural architecture search. *ArXiv*, abs/2004.08546, 2020. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 4

[8] Yihan Jiang, Jakub Konecný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *ArXiv*, abs/1909.12488, 2019. 2

[9] Peter Kairouz, H. B. McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Flo-

rian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2019. 1

[10] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Adaptive gradient-based meta-learning methods. In *Neural Information Processing Systems*, 2019. 2

[11] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 4

[12] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics Data Analysis*, 142:106816, 2020. 2

[13] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In *Neural Information Processing Systems*, 1989. 4

[14] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *ArXiv*, abs/1910.03581, 2019. 3

[15] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020. 2

[16] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *ArXiv*, abs/1907.02189, 2019. 2

[17] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Neural Information Processing Systems*, 2017. 2

[18] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *ArXiv*, abs/2001.01523, 2020. 2

[19] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Tao Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22:2031–2063, 2019. 1

[20] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Federated learning for vision-and-language grounding problems. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[21] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4

[22] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 4

[23] H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016. 1, 2

[24] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 4

[25] Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydogan, and John Reid. Misclassification risk and uncertainty quantification in deep classifiers. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2483–2491, 2021. 2

[26] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. 3

[27] Tao Shen, J. Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Fei Wu, and Chao Wu. Federated mutual learning. *ArXiv*, abs/2006.16765, 2020. 1, 2, 3

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 4

[29] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[30] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020. 2

[31] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Péter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10726–10734, 2018. 2

[32] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2019. 1

[33] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *ArXiv*, abs/2002.04758, 2020. 2

[34] Dell Zhang, Murat Sensoy, Masoud Makrehchi, Bilyana Taneva-Popova, Lin Gui, and Yulan He. Uncertainty quantification for text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3426–3429, New York, NY, USA, 2023. Association for Computing Machinery. 2

[35] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[36] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. 2018. 2

[37] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. 2