FedCTTA: A Collaborative Approach to Continual Test-Time Adaptation in Federated Learning

Rakibul Hasan Rajib[1][0009−0007−3705−9400], Md Akil Raihan Iftee[1][0009−0001−1459−6365], Mir Sazzat Hossain[1][0000−0001−6999−6879], A K M Mahbubur Rahman[1][0000−0001−9941−4817], Sajib Mistry[2][0000−0001−7513−3789], Md. Ashraful Amin[1][0000−0003−2330−9775], and Amin Ahsan Ali[1][0000−0002−0129−8705]

[1] Center for Computational & Data Sciences
Independent University, Bangladesh
[2] Curtin University

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Deep learning models experience performance degradation when a distribution shift occurs between the training and testing data. For instance, a self-driving car system trained on images captured in clear weather conditions may perform well in detecting pedestrians, traffic signs, and obstacles under sunny skies. However, during deployment, adverse weather conditions such as fog, rain, or, snow can significantly impair its ability to recognize objects, if these scenarios were not part of the training data. To address distribution shifts, researchers have explored approaches such as Domain Generalization (DG)[2] and Domain Adaptation (DA)[cite]. DG aims to improve model robustness to unseen domains by training on diverse source domains, while DA focuses on adapting a model trained on one domain to perform well on a related target domain. Despite their success, these methods have inherent limitations: DG depends on sufficient domain diversity during training, and DA typically requires access to source domain data, which is not always practical due to privacy or storage constraints.

Test-Time Adaptation (TTA) has emerged as a promising alternative, enabling models to adapt during deployment using only incoming test samples. Unlike Domain Adaptation (DA) and Domain Generalization (DG), TTA operates without requiring access to the original training data, making it particularly suitable for scenarios where sharing source data is restricted due to privacy or logistical constraints. For example, TTA can allow a self-driving car to adapt dynamically to foggy conditions in real time, utilizing only test data from the foggy environment without needing to revisit the original dataset.

In recent years, data privacy has become a critical concern across industries, driven by the growing awareness of data misuse and stringent privacy regulations such as GDPR[cite] and CCPA[cite]. This shift towards privacy-conscious practices poses significant challenges for traditional machine learning approaches, which often rely on centralized data access. Federated Learning (FL) [cite] has emerged as a prominent decentralized approach for collaborative model training. By enabling multiple clients to jointly learn a global model without exchanging raw data, FL effectively addresses privacy concerns while accommodating a wide range of applications across domains such as healthcare, finance, and autonomous systems.

In such decentralized settings, performing Test-Time Adaptation (TTA) is a challenging task due to the heterogeneous and continually changing data distributions across clients. While clients can adapt locally based on their own test-time data, this approach may not fully capture the broader environmental shifts. Collaborative adaptation, where clients share insights from their diverse data distributions, has the potential to improve model performance in dynamic environments. However, collaboration is difficult due to privacy concerns and the challenges of aligning adaptation strategies across clients with varying data distributions. This makes achieving effective collaboration while maintaining privacy and ensuring robust performance a non-trivial problem.

Recent work has explored Test-Time Adaptation (TTA) in federated learning settings. For example, FedICON[**?**] uses contrastive learning to fine-tune models across heterogeneous client environments, while ATP[**?**] introduces client-specific adaptation by adjusting module-specific adaptation rates. However, these approaches face significant challenges: they assume static test-time distributions, fail to fully utilize inter-client collaborations. Methods like FedTSA[**?**], which leverage temporal-spatial correlations based on local feature means for personalized model aggregation, raise privacy concerns. The use of local feature statistics can potentially reveal sensitive information through the reconstruction process. Additionally, transmitting full model parameters from clients to the central server increases communication costs, further complicating the implementation of such methods in resource-constrained environments.

To address these limitations, we propose Federated Continual Test-Time Adaptation (FCTTA), a novel approach that balances privacy, efficiency, and performance. Our method updates only the batch normalization (BN) layers of the model, reducing bandwidth overhead while ensuring adaptability. By leveraging similarity between BN layers, FCTTA enables dynamic client collaboration without sharing sensitive information such as local feature means. Our contributions are as follows:

- ✧ Eliminates the need for sharing local feature means by utilizing BN layer similarity for client collaboration preserving privacy and security
- ✧ Limits parameter sharing to BN layers, significantly reducing bandwidth usage.
- ✧ Enhances model accuracy in dynamic environments through continual test-time adaptation and client collaboration.

2 Related Works

2.1 Federated Learning

Federated learning is a decentralized approach to training machine learning models while keeping data localized, thereby addressing privacy and security concerns. McMahan et al. [5] introduced the FederatedAveraging algorithm, enabling efficient model training on unbalanced and non-IID data with significantly reduced communication costs. Li et al. [4] proposed using a globally shared dataset to mitigate performance degradation in non-IID data settings, improving model accuracy by up to 30% on skewed datasets like CIFAR-10. Zhao et al. [13] introduced FedProx, an extension of FedAvg, to handle statistical and system heterogeneity, ensuring robust convergence and improving accuracy by 22% in highly heterogeneous settings

2.2 Test Time Adaptation

Test-Time Adaptation (TTA) methods aim to enable models to adapt dynamically to distribution shifts without access to source data. TENT [9] introduces a lightweight adaptation approach by minimizing entropy through updates to BatchNorm parameters, achieving state-of-the-art results on corrupted datasets like ImageNet-C while being efficient for online updates. DUA [6] extends this idea by dynamically adapting BatchNorm statistics with minimal unlabeled test data, showcasing robust performance gains in real-time scenarios like autonomous driving. EATA [7] addresses challenges of catastrophic forgetting and noisy updates by introducing an entropy-based sample selection strategy and a Fisher regularizer to constrain significant model parameters. CoTTA [10] further pushes the boundaries by tackling non-stationary environments with weight-averaged and augmentation-averaged pseudo-labeling, alongside stochastic restoration of source weights to preserve long-term knowledge, achieving superior results on continually changing domains. Together, these methods highlight diverse strategies to enhance TTA efficiency and robustness across various applications.

2.3 Federated Test Time Adaptation

Adaptive Test-Time Personalization (ATP) [1] adaptively learns module-specific adaptation rates based on inter-client distribution shifts. Clients simulate unsupervised adaptation during training, refining adaptation rates to enhance performance on unseen, unlabeled data. FedTHE+ [3] ensembles global and local classifiers for robust test-time personalization and performs unsupervised fine-tuning, significantly improving accuracy across in-domain (ID) and out-of-domain (OOD) distributions. FedICON [8] employs contrastive learning to capture invariant knowledge from inter-client heterogeneity during training and uses self-supervision for smooth test-time adaptation, tackling intra-client heterogeneity. While leveraging inter-client heterogeneity to address test-time shifts, FedICON requires extensive contrastive learning, which may be computationally intensive for resource-constrained clients. Zhang et al. [12] developed Temporal-Spatial Aggregation (TSA), a server-side module that captures temporal and spatial correlations among clients during dynamic test-time adaptation. TSA is self-supervised and robust to temporal-spatial heterogeneity, enabling collaborative adaptation in multi-device settings. **TSA assumes synchronized client activity, limiting its applicability in asynchronous or sporadic communication settings.** Xu et al. [11] proposed FedCal, a lightweight framework that performs test-time classifier calibration using estimated label priors from global model predictions. FedCal handles label shifts efficiently without extra labeled data, ensuring flexibility for unseen clients.

3 Methodology

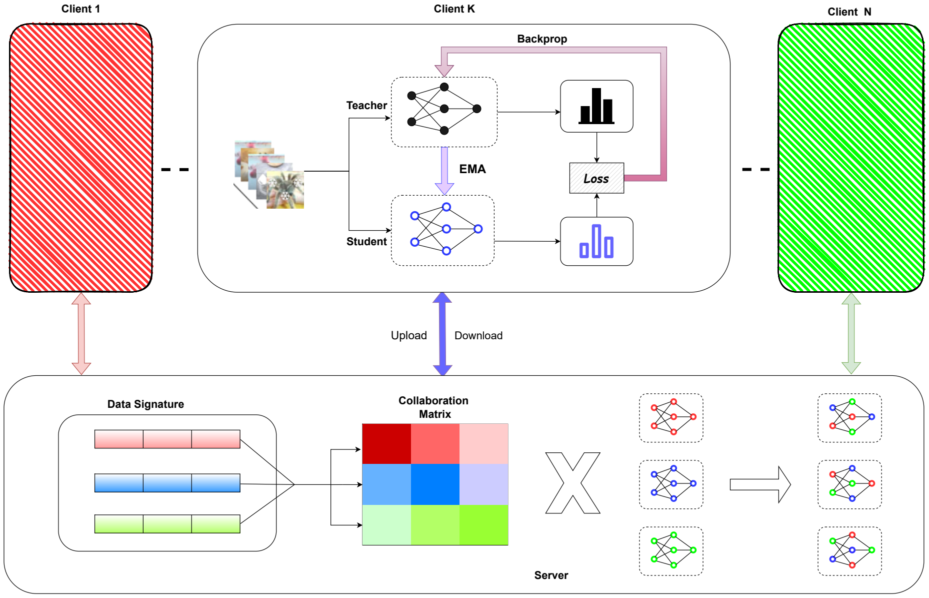

Fig. 1. a more descriptive caption needed, do it now. Plus this doesn't show that the aggregate models are passed back to clients

3.1 Problem Definition

Continual test-time adaptation (TTA) aims to address the challenge of adapting machine learning models to sequentially arriving data from non-stationary target domains. This is particularly challenging in federated settings where each client observes different data from similar or distinct domains, and these domains continually change over time. Unlike traditional domain adaptation methods, continual TTA does not assume access to the source domain data during deployment.

In this setting, we consider a federated system with N clients, denoted $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$. Each client encounters a stream of data $\mathcal{D}_t^{(i)}$ over time t, originating from dynamically evolving target domains. The goal is to adapt the model θ at each client to maintain performance on the incoming data while facilitating knowledge sharing across clients, without compromising data privacy. This problem is further complicated by the need to prevent catastrophic forgetting of earlier knowledge while reducing error accumulation over time.

3.2 Local Test-time Adaptation: A Knowledge Distillation Approach

At each client, we employ a mean teacher framework to perform local test-time adaptation. The framework consists of a student model f_s and a teacher model f_t and works based on a self-supervised training. The teacher model generates pseudo-labels that guide the student model's adaptation to the evolving target data.

The student model updates its parameters by the consistency loss between the student and the teacher predictions. The adaptation is guided by minimizing the consistency loss as symmetric cross-entropy loss in student model:

$$\mathcal{L}_{\text{SCE}} = \frac{1}{2}\left(H(y_t, f_s(x)) + H(f_s(x), y_t)\right),$$

where H represents the cross-entropy loss, y_t is the teacher's prediction, and x denotes the input data. This loss function ensures that the student aligns closely with the teacher's predictions.

To reduce error accumulation challenges of continual adaptation, the teacher model is updated using an exponential moving average (EMA) of the student model's weights which aggregates knowledge from previous iterations and is therefore more robust to noise:

$$\theta_t^{(f_t)} = \alpha\theta_{t-1}^{(f_t)} + (1-\alpha)\theta_t^{(f_s)},$$

where α is the decay factor for EMA. This update mechanism ensures that the teacher model retains a memory of past states, reducing the likelihood of catastrophic forgetting.

3.3 Similarity-Aware Aggregation

In a federated setting, knowledge sharing among clients is crucial for improving performance across the system. To achieve this while preserving data privacy, we propose a similarity-aware aggregation strategy. This approach aggregates the parameters of student models from different clients based on their similarity, enabling collaborative adaptation without requiring data exchange.

The aggregation mechanism starts with computing the similarity between different clients. For each client C_i, let $\theta_s^{(i)}$ represent the parameters of its student model. The aggregated parameters θ_s^{agg} are calculated as:

$$\theta_s^{\text{agg}} = \sum_{i=1}^{N} w_i \theta_s^{(i)},$$

where w_i is the similarity-based weight assigned to the client C_i. These weights are calculated using one of two similarity measures.

give names to each measure and, if possible, add equations that describe these. You must add a line or two describing the intuition behind each measure. For example, discuss why you are looking at the student model similarity instead of the teacher model.

The first measure calculates the similarity between the parameters of the student models for all clients. Specifically, the similarity between two clients' models is measured using a function such as cosine similarity. The similarity scores are then normalized using a softmax function to produce the weights.

The second measure considers the output tendency of each client's student model. Each client maintains an exponential moving average (EMA) of its current output to calculate output tendencies, denoted as $\bar{y}_t^{(i)}$. The similarity is computed based on the output tendencies between clients. This approach leverages the slow-moving nature of the teacher model and the fast updates of the student model to incorporate temporal dynamics into the similarity calculation. The similarity scores are normalized using a softmax function to determine the aggregation weights.

The aggregated model θ_s^{agg} is distributed back to the clients, serving as the updated student model for the next time step. By aligning the models of clients with similar domains, this process facilitates knowledge sharing between clients while ensuring not to access of individual clients' data is preserved.

Table 1. Performance comparison for all corruptions

Method	NIID				IID			
	CIFAR10-C		CIFAR100-C		CIFAR10-C		CIFAR100-C	
	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn
No-Adapt	35.19±0.27	35.19±0.27	30.22±0.12	30.22±0.12	35.19±0.27	35.19±0.27	30.22±0.12	30.22±0.12
Local	49.37±0.32	56.93±0.26	52.85±0.32	55.99±0.34	49.37±0.32	56.93±0.26	52.85±0.32	55.99±0.34
FedAvg	54.40±0.36	56.52±0.21	51.63±0.17	57.13±0.43	57.28±0.19	61.29±0.04	62.60±0.31	63.96±0.31
FedProx	53.86±0.62	56.52±0.21	53.00±0.38	57.13±0.43	55.03±0.29	61.29±0.04	62.27±0.67	63.96±0.31
FedAMP	55.88±0.24	57.27±0.25	55.57±0.46	58.62±0.39	56.43±0.77	57.26±0.07	61.70±0.63	58.24±0.44
pFedGraph	55.68±0.74	57.24±0.24	57.01±0.38	58.73±0.38	56.79±0.53	57.44±0.09	62.52±0.30	58.73±0.63
FedTSA	57.32±0.36	60.27±0.23	58.03±0.38	62.93±0.29	57.41±0.12	61.56±0.52	62.63±0.36	63.72±0.34
Ours(V1)	65.5		63.14		65.12		63.69	
Ours(V2)	64.59		60.08		64.49			
Ours(V3)	65.49		62.54		65.12		63.61	
Ours(V3)	64.58		60.08		64.59		59.97	

Table 2. Performance comparison under spatial IID and temporal heterogeneity

	Time	t ⟶															
Datasets	Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*Jpeg*	*Mean*
CIFAR10-C	Source	37.30	38.44	26.08	28.99	33.92	27.44	30.21	34.53	32.89	10.63	36.46	23.53	37.51	41.43	43.70	30.54
	Local	61.11	61.94	54.13	61.65	55.47	57.93	60.58	59.27	59.13	45.76	63.18	30.63	59.53	62.96	63.78	57.14
	FedAvg	65.34	66.04	59.22	65.97	59.74	62.17	64.95	63.32	63.02	48.99	67.70	33.04	64.41	67.71	68.90	61.37
	FedAMP	61.22	62.15	54.39	61.81	55.63	58.10	60.81	59.53	59.39	45.91	63.44	30.80	59.83	63.24	64.02	57.35
	pFedGraph	61.26	62.21	54.44	61.82	55.75	58.09	60.81	59.55	59.33	45.90	63.46	30.83	59.80	63.28	64.03	57.37
	FedTSA	65.28	66.06	59.21	66.64	59.93	62.49	65.25	63.11	63.57	49.55	67.88	32.97	64.56	67.70	69.28	61.57
	Ours																
CIFAR100-C	Source	14.05	16.64	34.76	41.60	19.35	38.15	43.32	36.48	27.63	20.96	54.91	17.24	35.02	11.45	41.73	30.22
	Local	51.59	53.02	50.26	65.13	50.77	63.23	65.07	58.10	58.17	51.10	66.70	61.25	56.67	59.98	51.57	57.51
	FedAvg	57.33	58.60	56.74	69.07	57.51	68.94	70.92	64.29	64.19	57.44	72.40	67.36	63.70	66.33	57.56	63.50
	FedAMP	51.96	53.39	50.47	65.52	51.12	63.66	65.49	58.36	58.56	51.39	67.10	61.66	57.11	60.39	52.01	57.88
	pFedGraph	52.05	53.42	50.48	65.60	51.19	63.61	65.49	58.39	58.64	51.39	67.08	61.64	57.14	60.46	52.06	57.91
	FedTSA	57.56	58.75	57.23	69.73	56.27	69.18	71.05	64.33	64.60	56.44	73.10	67.77	63.30	66.58	58.21	63.61
	Ours																

Table 3. Performance comparison under spatial heterogeneity and temporal IID.

Datasets	Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*Jpeg*	*Mean*
CIFAR10-C	Source	37.30	38.44	26.08	28.99	33.92	27.44	30.21	34.53	32.89	10.63	36.46	23.53	37.51	41.43	43.70	30.54
	Local	55.27	52.70	56.46	48.25	55.30	50.87	58.25	46.40	55.58	52.45	58.48	45.80	51.75	56.12	49.00	52.85
	FedAvg	61.49	56.33	60.15	50.26	61.14	53.02	63.64	50.17	60.02	53.88	63.53	50.95	55.86	59.55	52.36	56.82
	FedAMP	62.04	57.12	61.29	52.32	61.83	55.28	63.87	51.69	60.83	55.89	63.94	52.16	56.70	61.32	53.65	58.00
	pFedGraph	61.66	56.41	61.01	51.87	61.18	54.44	63.70	51.60	60.23	55.17	63.44	52.42	56.46	60.22	53.56	57.56
	FedTSA	62.16	57.59	61.72	52.58	61.91	55.36	63.96	50.87	61.33	56.67	64.36	51.37	57.15	61.78	53.23	58.14
	Ours-v1(FM)	65.12															
CIFAR100-C	Source	14.05	16.64	34.76	41.60	19.35	38.15	43.32	36.48	27.63	20.96	54.91	17.24	35.02	11.45	41.73	30.22
	Local	50.68	54.84	56.29	55.56	51.77	54.06	59.04	51.24	54.80	54.43	57.43	51.30	53.15	57.56	54.87	54.47
	FedAvg	52.14	43.19	63.30	48.91	53.90	49.31	65.61	46.71	54.31	49.38	61.93	45.47	51.57	57.00	56.34	53.27
	FedAMP	62.04	57.12	61.29	52.32	61.83	55.28	63.87	51.69	60.83	55.89	63.94	52.16	56.70	61.32	53.65	58.00
	pFedGraph	61.66	56.41	61.01	51.87	61.18	54.44	63.70	51.60	60.23	55.17	63.44	52.42	56.46	60.22	53.56	57.56
	FedTSA	57.33	58.60	56.74	71.07	57.51	68.94	70.92	64.29	64.19	57.44	72.40	67.36	63.70	66.33	57.56	63.63
	Ours																

Table 4. The experimental scenario and performance comparison of the case study.

Time	t ⟶															
Client 1-4	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Mean
Client 5-7	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Gaussian	Mean
Client 8-10	Jpeg	Pixelate	Elastic	Contrast	Brightness	Fog	Frost	Snow	Zoom	Motion	Glass	Defocus	Impulse	Shot	Gaussian	Mean
No-Adapt	20.85	19.55	34.65	26.70	33.40	33.45	35.45	31.80	27.15	36.30	31.75	27.70	25.80	20.40	23.75	28.58
FedAMP	48.00	54.75	61.35	53.30	60.05	61.60	62.30	59.05	58.60	58.30	52.15	54.75	54.45	51.65	54.80	56.34
pFedGraph	47.00	51.90	61.05	51.20	58.45	62.50	63.25	59.75	55.95	58.55	50.30	54.00	52.50	50.80	53.65	55.39
FedTSA	48.95	61.15	62.25	54.90	60.15	63.75	63.75	63.05	58.35	59.90	54.00	58.15	58.15	57.45	56.40	58.69
Ours																

4 Experimental Results

4.1 Implementation Details

4.2 Baselines

4.3 Performance Analysis

4.4 Decoding Collaboration Relationship

5 Ablation Study

6 Conclusion

References

1. Bao, W., Wei, T., Wang, H., He, J.: Adaptive test-time personalization for federated learning. Advances in Neural Information Processing Systems **36** (2024)
2. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
3. Jiang, L., Lin, T.: Test-time robust personalization for federated learning. arXiv preprint arXiv:2205.10920 (2022)
4. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
5. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)

6. Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14765–14775 (2022)
7. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
8. Tan, Y., Chen, C., Zhuang, W., Dong, X., Lyu, L., Long, G.: Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. Advances in Neural Information Processing Systems **36** (2024)
9. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
10. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
11. Xu, J., Huang, S.L.: A joint training-calibration framework for test-time personalization with label shift in federated learning. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 4370–4374 (2023)
12. Zhang, J., Liu, X., Zhang, Y., Zhu, G., Niu, J., Tang, S.: Enabling collaborative test-time adaptation in dynamic environment via federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 4191–4202. KDD '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3637528.3671908
13. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)