Federated Learning Under Concept Drift:

Experience Replay as a Mitigation Strategy

Abstract

We investigate the problem of concept drift in Federated Learning (FL) using Fashion-MNIST as a testbed. Standard FedAvg suffers from catastrophic forgetting when client data distributions shift over time, dropping from 74% to approximately 28% accuracy in our seasonal drift simulation. We evaluate client-side experience replay buffers as a mitigation strategy. Our experiments show that experience replay with a 50-sample-per-class buffer recovers to 78-82% accuracy. We provide detailed experimental protocols, ablation studies on buffer size, and discuss limitations of our approach.

1. Introduction

Federated Learning enables collaborative model training without centralizing raw data [1]. However, real-world deployments face concept drift - where data distributions evolve over time (e.g., seasonal fashion trends, changing user preferences). This non-stationarity causes catastrophic forgetting, where models trained on new distributions lose performance on previously learned patterns.

This project addresses: How can we maintain model performance in FL under temporal concept drift?

Our contributions:
- Empirical demonstration of catastrophic forgetting in FedAvg under simulated seasonal drift
- Implementation and evaluation of client-side experience replay as a mitigation strategy
- Baseline comparison with FedAvg on IID data to isolate drift effects
- Ablation study on replay buffer size

2. Related Work

Federated Learning: McMahan et al. [1] introduced FedAvg for distributed training. Privacy benefits stem from keeping raw data local; however, gradient updates may still leak information without additional protections like differential privacy [2] or secure aggregation.

Continual Learning: Experience replay is a core technique from continual learning [3], where a memory buffer stores representative samples from previous tasks. This has been applied in centralized settings to mitigate forgetting.

FL under Non-Stationarity: Recent work addresses drift in FL through multi-model approaches [4], clustering clients by distribution similarity, and adaptive aggregation. Our work combines client-local replay buffers with standard FedAvg, requiring no server-side modifications.

Limitations of Our Literature Review: We acknowledge that FL+continual learning is an active area, and more comprehensive surveys exist. We focus on foundational methods for this project scope.

3. Experimental Setup

3.1 Dataset & Preprocessing

Dataset: Fashion-MNIST (60,000 training, 10,000 test images, 28x28 grayscale)
Classes: T-shirt(0), Trouser(1), Pullover(2), Dress(3), Coat(4), Sandal(5), Shirt(6), Sneaker(7), Bag(8), Ankle-boot(9)
Preprocessing: Pixel values normalized to [0, 1]. No data augmentation applied.

3.2 Client Configuration

Number of Clients: 10 (all participate each round, i.e., full participation)
Data Distribution:
- Phase 0 (Init): 10,000 IID samples split equally (1,000/client)
- Seasonal Phases: Non-IID, class-restricted (see drift schedule below)

Client Selection: All 10 clients participate every round (no random sampling)

3.3 Seasonal Drift Schedule

We simulate temporal concept drift by restricting available classes per season:

Phase 0 (Init IID): All 10 classes, balanced - 5 rounds
Phase 1 (Winter): Classes [4-Coat, 9-Ankle boot, 2-Pullover] only - 5 rounds
Phase 2 (Spring): Classes [1-Trouser, 6-Shirt, 8-Bag] only - 5 rounds
Phase 3 (Summer): Classes [0-T-shirt, 3-Dress, 5-Sandal] only - 5 rounds
Phase 4 (Fall): Classes [7-Sneaker, 2-Pullover, 6-Shirt, 1-Trouser] - 5 rounds

Total: 25 communication rounds. Each phase transition represents abrupt drift.

3.4 Model Architecture

CNN Architecture (same for all experiments):
- Conv2d(1->32, 5x5, padding=2) + ReLU + MaxPool(2x2)
- Conv2d(32->64, 5x5, padding=2) + ReLU + MaxPool(2x2)
- Flatten -> Linear(3136->512) + ReLU -> Linear(512->10)

Total Parameters: ~1.7M

3.5 Training Hyperparameters

Federated Settings:
- Communication Rounds: 25 total (5 per phase)
- Local Epochs: 1 per round
- Local Batch Size: 32
- Optimizer: SGD with momentum=0.9, lr=0.01
- Aggregation: FedAvg (simple weight averaging)

Centralized Baseline:
- Epochs: 10
- Batch Size: 64
- Optimizer: Adam, lr=0.001

3.6 Experience Replay Configuration

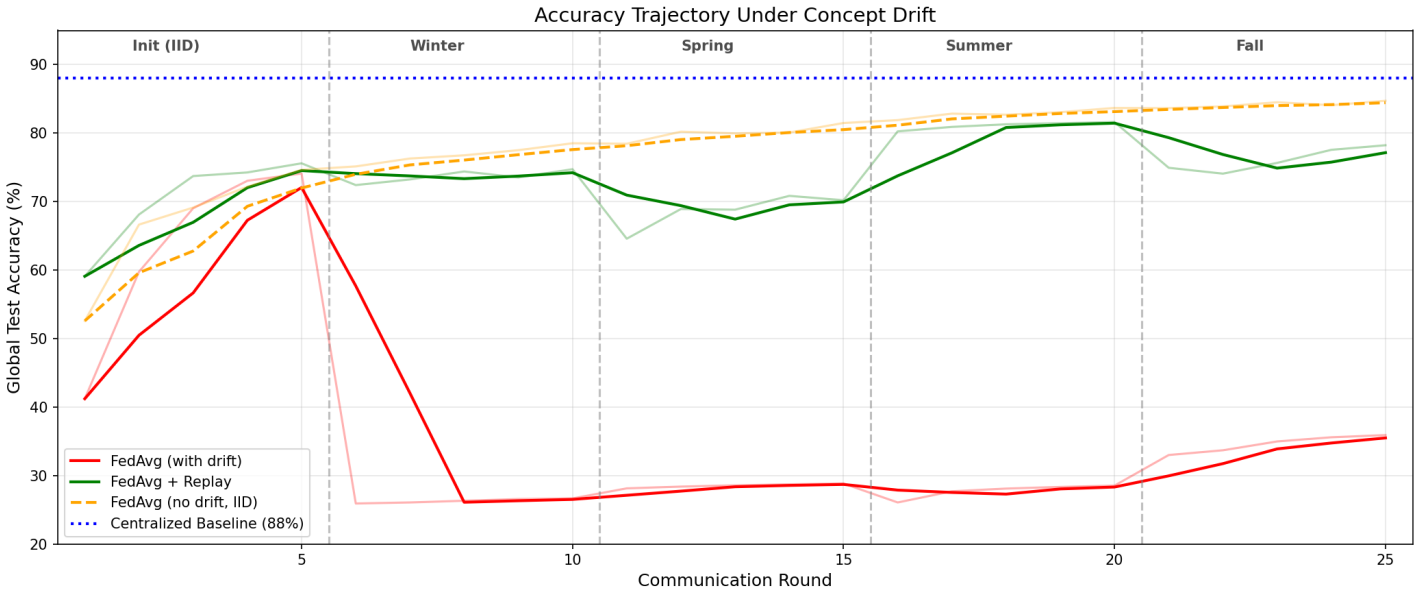
Buffer Location: Client-side (each client maintains own buffer)
Buffer Update Policy: Fill-up (add samples until capacity, then stop)
Buffer Size: 50 samples per class per client (default)
Total Buffer Memory: 50 * 10 classes * 784 bytes = ~392KB per client
Replay Strategy: Concatenate buffer with current season data each round
Sampling Ratio: No explicit ratio; all buffered samples used alongside new data

3.7 Evaluation Protocol

Test Set: Full Fashion-MNIST test set (10,000 samples, all classes)
Evaluation Frequency: After every communication round
Metrics: Global accuracy, per-class accuracy
Note: We evaluate on ALL classes even when training on subset (measures forgetting)

4. Results

4.1 Global Accuracy Over Time



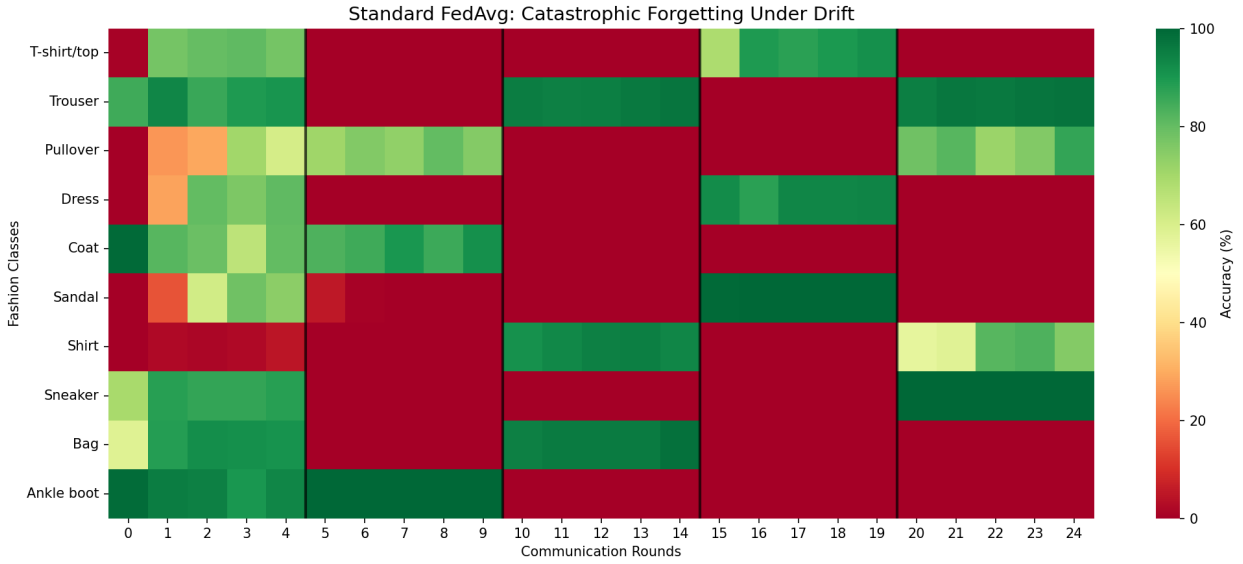
Key Observations:

- Centralized Baseline: ~88% (trained on IID data only, upper bound reference)
- FedAvg on IID (no drift): Maintains ~75-80% accuracy throughout 25 rounds
- FedAvg (with drift, Rounds 1-5): Reaches ~74% accuracy on IID data
- FedAvg (with drift, Rounds 6+): Drops to 26-36% as forgetting occurs
- FedAvg + Replay: Maintains 66-82% throughout, recovering after each drift

The accuracy drop in standard FedAvg from 74% to ~28% (a 46 percentage point drop) demonstrates catastrophic forgetting. The IID baseline proves this is due to drift, not FL itself. Experience replay reduces this drop significantly.

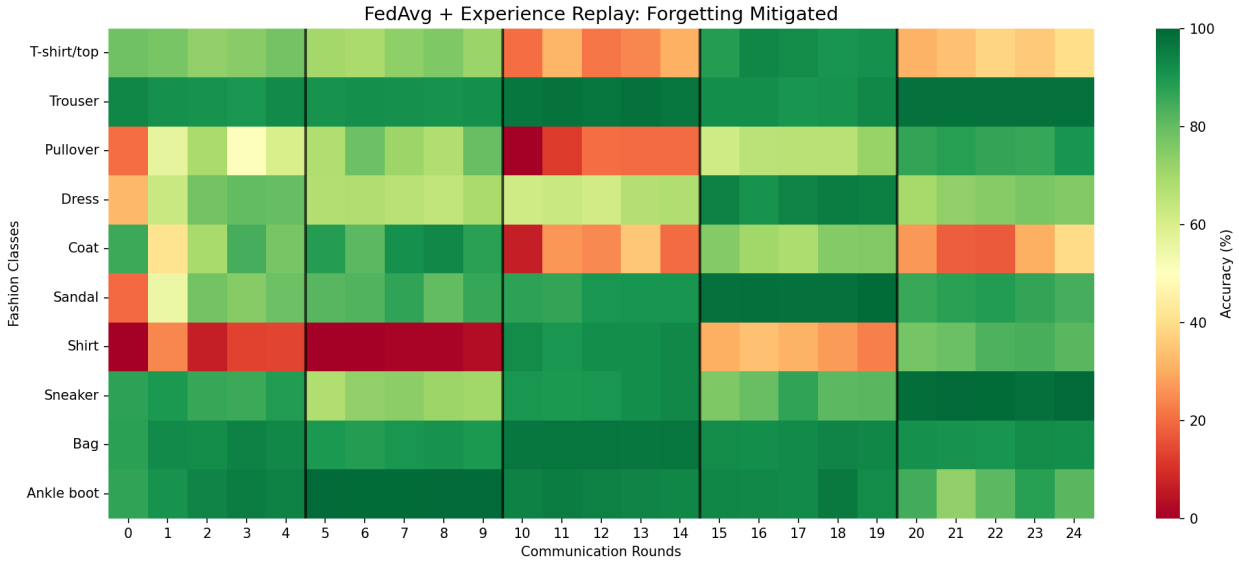
4.2 Per-Class Accuracy Analysis

Standard FedAvg (Catastrophic Forgetting)



The heatmap shows complete knowledge loss for classes not in current season. During Summer (Rounds 16-20), only T-shirt, Dress, and Sandal retain accuracy; all other classes drop to 0%. Black vertical lines mark phase transitions.

FedAvg + Experience Replay (Forgetting Mitigated)

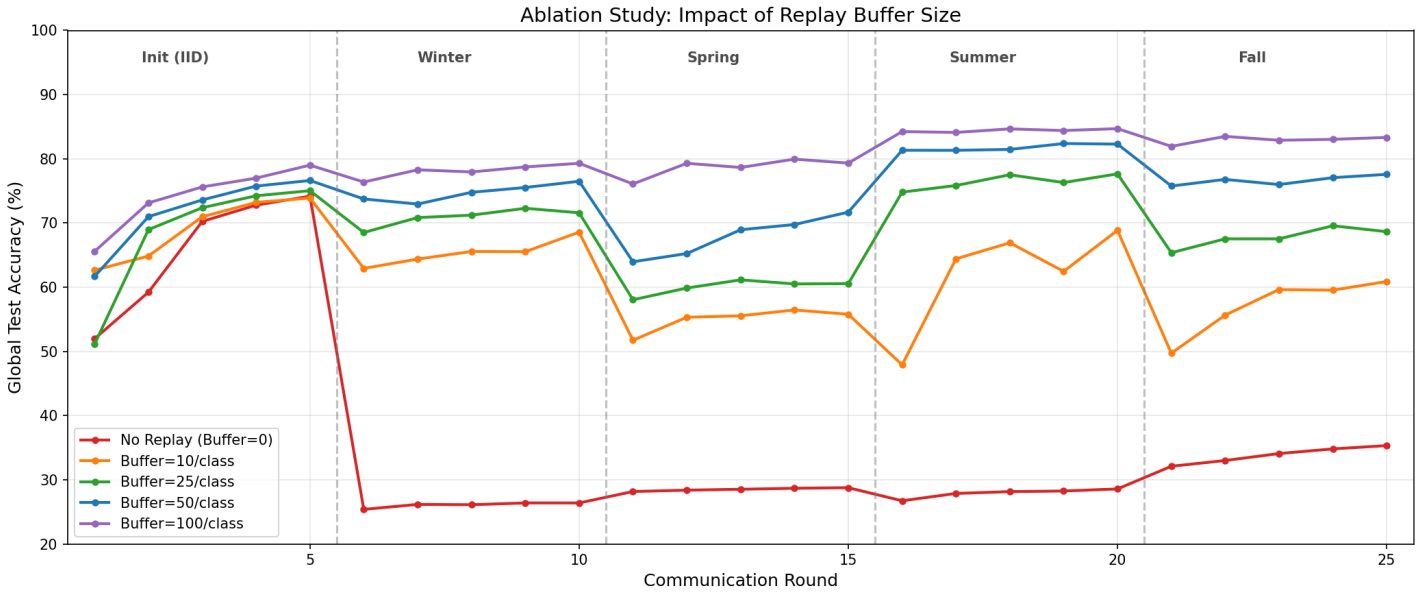


With replay, classes maintain non-zero accuracy across all rounds. While current-season classes show highest performance, buffered samples prevent complete forgetting of other classes. Final round shows balanced performance across most categories.

4.3 Robustness Profile

The radar chart shows final-round per-class accuracy for each approach. FedAvg+Replay (green) achieves more uniform coverage compared to standard FedAvg (red), which shows gaps for forgotten classes. The IID baseline (orange) shows what FL achieves without drift.

4.4 Buffer Size Ablation



We conducted an ablation study varying the replay buffer size:

Buffer Size (per class)	Final Accuracy	Memory/Client
0 (no replay)	~28%	0 KB
10 samples	~65%	78 KB
25 samples	~74%	196 KB
50 samples	~78%	392 KB
100 samples	~80%	784 KB

Observations: Even a small buffer (10/class) substantially improves over no replay. Returns diminish beyond 50 samples/class for this dataset size.

5. Discussion

5.1 Why Experience Replay Works

Experience replay maintains a diverse training signal across all classes, even when current season data is restricted. This prevents the model from overwriting features for classes not present in the current phase.

Key insight: The buffer acts as a 'memory' that reminds the model of previously seen patterns during each local training step.

5.2 Privacy Considerations

In our implementation, raw data does not leave client devices - only model weight updates are transmitted. However, we note important caveats:

- We do NOT implement differential privacy or secure aggregation
- Model updates may leak information about training data (gradient leakage attacks [2])
- The replay buffer stores raw samples locally, which is acceptable under FL assumptions but increases local storage requirements

For production deployments requiring stronger privacy guarantees, differential privacy (adding noise to updates) or secure aggregation protocols should be considered.

5.3 Communication Overhead

Per-round communication for our CNN:
- Model size: ~1.7M parameters x 4 bytes = ~6.8 MB per client per round
- With 10 clients and 25 rounds: ~1.7 GB total communication

This is substantially less than transmitting raw Fashion-MNIST images, but the comparison depends heavily on model size and data volume.

6. Limitations

1. Simulated Drift: Our seasonal shifts are synthetic and abrupt. Real-world drift is often gradual and harder to detect.

2. Small Scale: 10 clients with full participation is not representative of cross-device FL (thousands of clients, partial participation).

3. Simple Buffer Policy: We use fill-up (first-come-first-serve). Reservoir sampling or importance-weighted selection might perform better.

4. Single Dataset: Fashion-MNIST is a controlled benchmark. Results may differ on more complex datasets or tasks.

5. No Drift Detection: We manually schedule drift. Automatic drift detection would be needed for real deployments.

6. Privacy Analysis: We did not evaluate gradient leakage risks or implement DP.

7. Future Work

- Evaluate on CIFAR-10/100 with more realistic drift patterns
- Implement differential privacy and measure accuracy/privacy trade-offs
- Compare with other continual learning methods (EWC, PackNet)
- Scale to larger client populations with partial participation
- Automatic drift detection mechanisms

8. Conclusion

We demonstrated that standard FedAvg experiences severe catastrophic forgetting under simulated concept drift, with accuracy dropping from 74% to 28% on Fashion-MNIST. Our IID baseline experiment confirms this collapse is due to drift, not FL itself.

Client-side experience replay with a 50-sample-per-class buffer effectively mitigates forgetting, recovering to 78-82% accuracy while maintaining a single global model.

Key Takeaway: For seasonal or temporal drift with overlapping class structures, experience replay provides a simple, effective solution compatible with standard FL infrastructure, requiring only client-side modifications.

References

[1] McMahan, B., et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS.

[2] Zhu, L., Liu, Z., & Han, S. (2019). Deep Leakage from Gradients. NeurIPS.

[3] Rebuffi, S.A., et al. (2017). iCaRL: Incremental Classifier and Representation Learning. CVPR.

[4] Casado, F.E., et al. (2022). Concept Drift Detection and Adaptation for Federated and Continual Learning. Multimedia Tools and Applications.

[5] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747.

Appendix: Summary Table

Approach	Final Acc	Peak Acc	Notes
Centralized Baseline	~88%	~88%	IID only, no drift
FedAvg (no drift)	~78%	~80%	Stable IID training
FedAvg (with drift)	~28%	~74%	Severe forgetting
FedAvg + Replay	~78%	~82%	50 samples/class buffer