We extend the visualization techniques of 2D loss surface and optimization trajectories to understand federated learning for both global and local scope. We visually demonstrate the phenomenon of model drifting, the effect of data heterogeneity and model initialization. With proper initialization, the trajectories under different non-IID degrees would enter the same loss basin, which provides an explanation of why pre-training could largely improve FL.

**Background**

- **Federated Learning (FL)**
  - M clients, each client holds a data set \( D_m = \{(x_i, y_i)\}_{i=1}^{D_m} \)
  - The optimization problem formulated as:
    \[
    \min_{\theta} \mathcal{L}(\theta) = \sum_{m=1}^{M} \frac{|D_m|}{|\mathcal{D}|} \mathcal{L}_m(\theta),
    \]
    where \( \mathcal{L}_m(\theta) = \sum_{i} f(x_i, y_i; \theta) \)

- **Visualization of loss landscape**
  - A powerful technique to analyze the high-dimensional learning behavior and generalization of neural networks [2]
  - Visualizing the loss in a 2D subspace at \( \alpha \) a center point \( \beta \) with two direction vectors, \( \delta \) and \( \eta \)
    \[
    f(\alpha, \beta) = L(\theta^* + \alpha \delta + \beta \eta)
    \]

**Loss Landscape Visualization for FL**

- **Data preparation**
  - CIFAR-10 image classification
  - Number of clients: \( M = 10 \)
  - Coefficient \( \alpha \) indicates the non-IID degree
    - \( \alpha = 0.3 \) (middle non-IID)

- **Global scope**
  - Trajectory of the aggregated global models over 100 rounds
  - Global model initialized at \( \star \)
  - Loss values are computed using global training data

- **Training details**
  - Types of neural networks: ConvNet and ResNet20
  - FedAvg with 100 rounds
  - 5 epochs for local training
  - Random initialization
  - Full client participation

- **Local scope**
  - Trajectories of 10 local models at round 9
  - Local training starts from \( \Delta \)
  - The final aggregated global model \( \delta \) moves closer to the minima

**Highlights**

- We extend the visualization techniques of 2D loss surface and optimization trajectories to understand federated learning for both global and local scope.
- We visually demonstrate the phenomenon of model drifting, the effect of data heterogeneity and model initialization.
- With proper initialization, the trajectories under different non-IID degrees would enter the same loss basin, which provides an explanation of why pre-training could largely improve FL.

**Effect of Data Heterogeneity**

- **Random weight initialization**
  - Global scope
    - Visualizations of the final global models
  - With the same initialization, global models under different non-IID soon diverge into different loss basins (fig: a, b)
  - Global loss along the interpolation between the IID global model and each of the non-IID global model (fig: c)
    - A higher intermediate loss indicates existence of a larger barrier between two models

- **Pre-trained weight initialization**
  - Global scope
    - Initialization with a well pre-trained model leads to a much smoother loss landscape
    - All the final global models enter the same loss basin (fig: a, b)

**Detailed Study**

- **Local scope**
  - Different non-IID conditions, at round 30
  - Severer non-IID cases have denser loss contour, meaning that the local models quickly move away from the global minima

- **Visualizations on training trajectories**
  - FedAvg initialized with the ImageNet pre-trained weights
  - When the data become more non-IID, the trajectory gradually deviates away from the ideal trajectory (purple one, under IID data) and ends at a higher loss value (fig: a, b)
  - The degree of deviation is governed by the non-IID degree
    - Four levels of data heterogeneity: IID to \( \alpha = 1.0, \alpha = 0.5, \) and \( \alpha = 0.1 \)

**Partial client participation**

- **Global scope**
  - Under non-IID condition, if we only use a small portion of clients, the overall performance of FedAvg drops significantly
  - Performance of non-IID is affected more seriously than IID data
    - At each round, randomly sample a portion of 50 clients
    - Three client sampling rates: 0.2, 0.5, 0.8

This research is supported in part by grants from the National Science Foundation (IIS-2107077, OAC-2118240, and OAC-2112606) and Cisco Systems, Inc.

---

[2] Li et al., Visualizing the Loss Landscape of Neural Nets, 2018