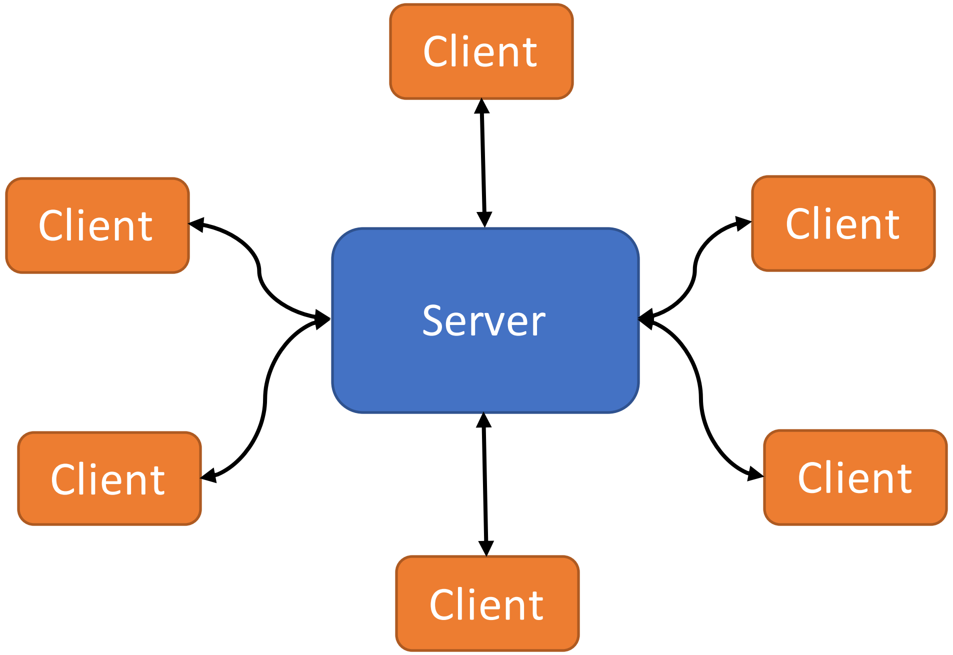
Causality and Privacy

Sandipan Sikdar

Privacy

- We consider a distributed setting rather than a centralized one



- Exchange of information between the server and client
- Client data is private and should not be revealed to the adversary
- Such settings are vulnerable to adversarial attacks

Federated Learning

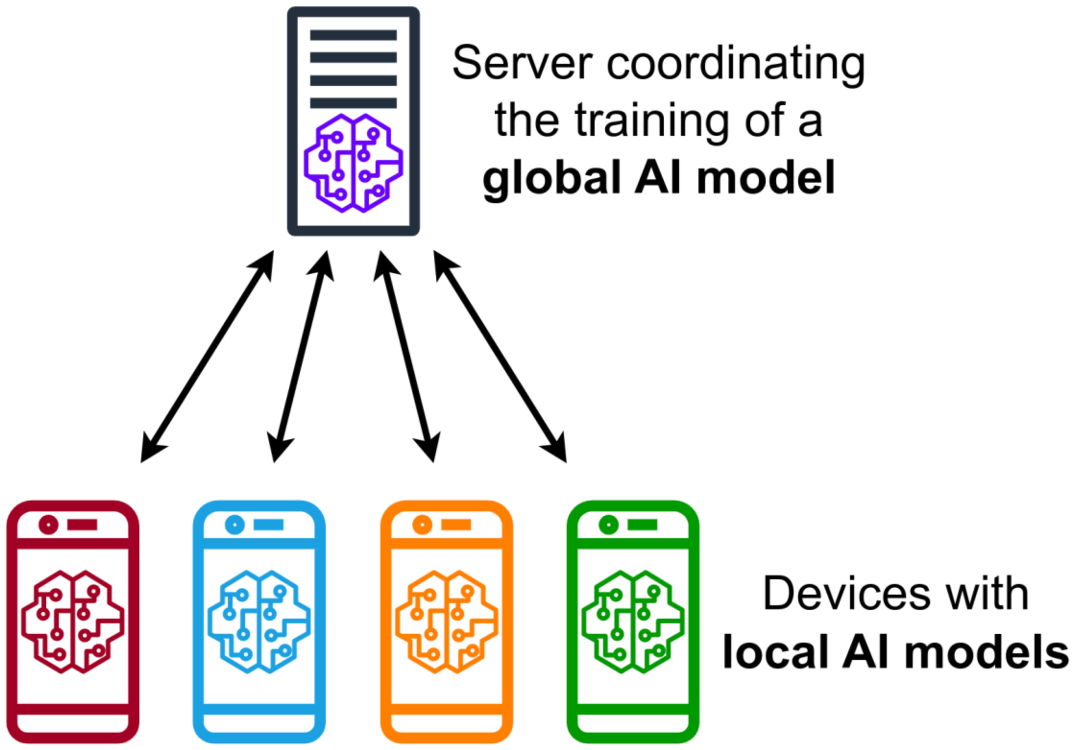


Image source: Wikipedia

- Learn a shared prediction model while keeping all the training data on device
- Training data need not be stored in the cloud

- A client downloads the current model
- Improves it by learning from local data
- Summarize the changes as a small focused update
- Only this update to the model is sent to the server
- Server collects updates from all the clients and then updates itself.

FederatedAveraging

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \ldots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$

ClientUpdate(k, w): // *Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

Initialize the server weights/parameters

Randomly select a set of clients

Update each client on local data

Compute a weighted sum of the weights

Update weights following SGD and send the new weights to server

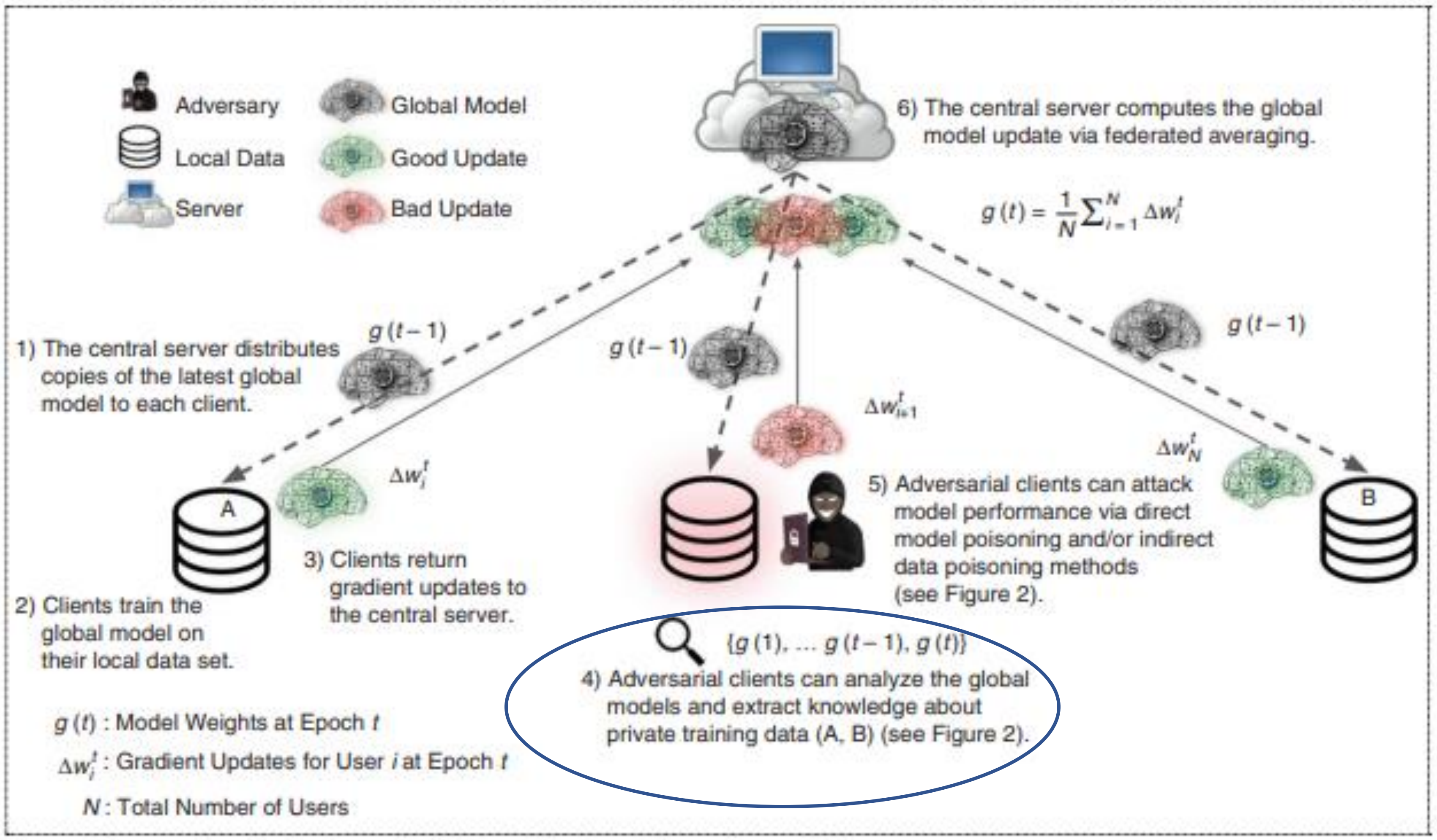
FedSGD: Send the gradient updates to the server instead of weights

McMahan et. al., AISTATS 2017.

Federated Learning

- The model needs to generalize across multiple distributions
  - The data at each client need not follow the same distribution
- Ensure client data is not leaked to adversaries

Federated Learning: Attacks



Jere, M. S., Farnan, T., & Koushanfar, F. IEEE S & P, 2020.

Membership Inference Attacks

- Given a data point $d = (X, y)$ determine whether d was used for training
- Can reveal sensitive information
  - Multiple hospitals train a model on COVID-19 diagnosis
  - Membership inference attacks can reveal if an individual tested for COVID-19
- Blackbox vs. whitebox
- Active vs. passive

Membership Inference Attacks

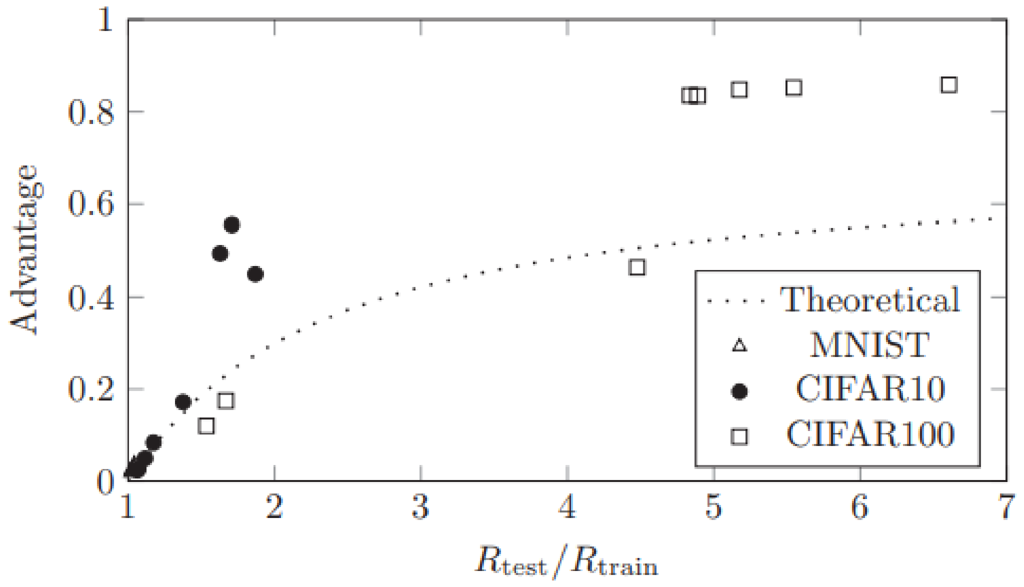
- SGD algorithm minimizes the empirical expectation of the loss function over a dataset D

$$\min_{\mathbf{W}} \mathbb{E}_{(\mathbf{x},y)\sim D}\big[L(f(\mathbf{x};\mathbf{W}),y)\big]$$

- Over the training steps SGD repeatedly updates $\mathbf{W}$ towards reducing the loss
- For any data point in the training set the gradient $\frac{\partial L}{\partial \mathbf{W}}$ is pushed to 0
- Distribution of the model‘s gradients for the data used for training would significantly differ for the data not used for training

Model Generalizability

- Distribution of the model's gradients for the data used for training would significantly differ for the data not used for training
- More pronounced for models that overfit/do not generalize

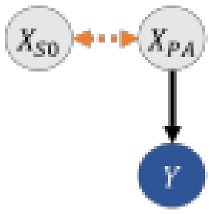


Advantage: Difference between true positive rate (TPR) and false (FPR) on the task of detecting membership

Yeom et. al. CSF 2018

Model Generalizability

- Two important results (Tople et.al. ICML 2020)
- The worst case generalization error for a causal model is less than or equal to an association model
- The worst case membership advantage of a causal model is less than or equal to an association model



Marked in blue

Colored MNIST: $\{0,1,2,3,4\} \rightarrow 1$

$\{5,6,7,8,9\} \rightarrow 0$

- Association models can reach zero error by correlating with the color
- Causal models will look into the causal features
- Association models won’t generalize to distribution shift (e.g., digits in red)

Causality and Privacy

Federated learning

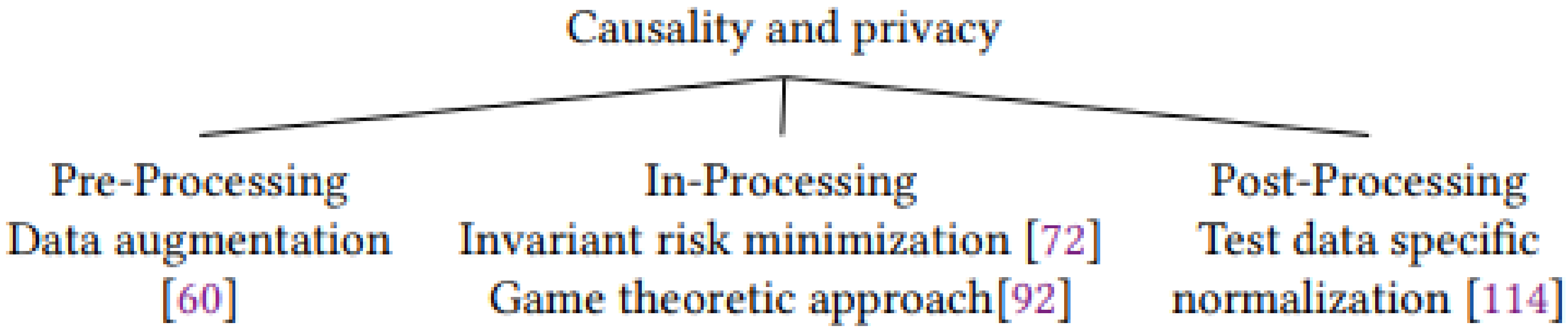
- Distribution shifts
- Overfitting -> Vulnerability to attacks

Causal models

- Ability to deal with distribution shifts

Causal models -> Improved generalizability -> Less vulnerable to attacks

Causal Models for Privacy

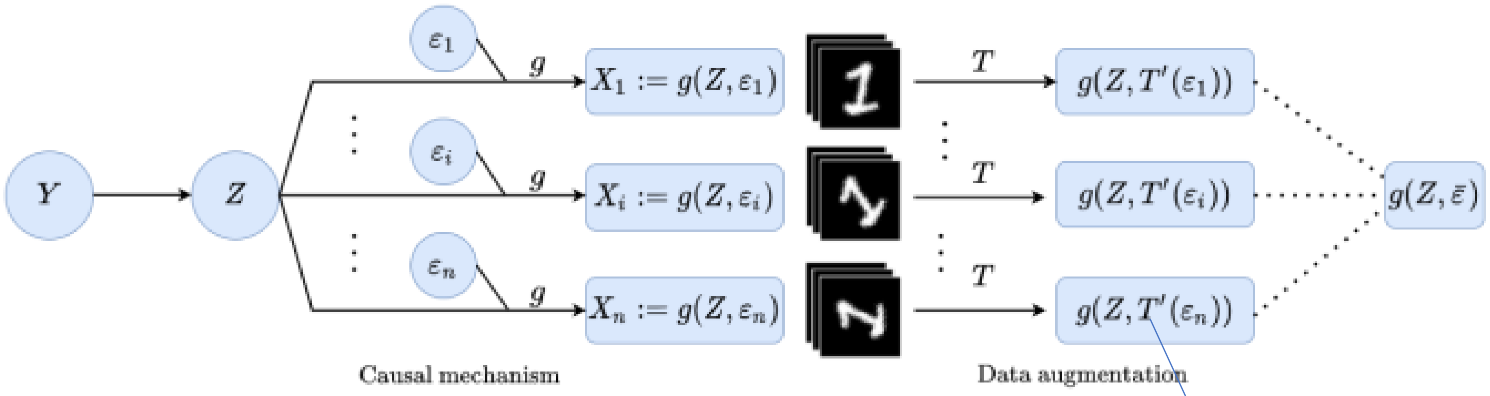


Pre-processing Methods

- Manipulate the training data at each client, which would result in better generalization.
- Data augmentation
  - Strategy for increasing diversity of samples
  - Helps obtaining robust models
  - Better performance on OOD data samples

Causal Data Augmentation

$$X_i := g(Z, \epsilon_i)$$



Transformation over the style feature, e.g., rotate the image with a random amount

de Luca, A.B., Zhang, G., Chen, X. and Yu, Y., 2022. Mitigating Data Heterogeneity in Federated Learning with Data Augmentation. *arXiv preprint arXiv:2206.09979*.

In-processing Methods

- Learn domain invariant (causal) features while training
- Invariant risk minimization

Find ϕ such that each environment has the same classifier i.e., $W_A = W_B$

Env A $\longrightarrow \phi \longrightarrow W_A \longrightarrow \widehat{Y_A}$

Env B $\longrightarrow \phi \longrightarrow W_B \longrightarrow \widehat{Y_B}$

ϕ has to extract invariant features that are useful for all the environments

$$min_{\phi} \sum_{e \in \epsilon} \mathcal{L}(w \cdot \phi(x), y)) + \lambda \left\| \nabla \mathcal{L}(w \cdot \phi(x), y)) \right\|$$

Causal Fed

ServerCausalUpdate:

Initialize $\mathbf{W}_0^s$
for each server epoch, t = 1,2,..k **do**
  Select random set of S clients
  Share initial model with the selected clients
  for each client $k \in S$ **do**
    $(\phi(x_t^k), \mathbf{Y}^k) \leftarrow ClientRepresentation(k, \mathbf{W}_t^k)$
    Evaluate loss $\mathcal{L}_k$
  end for
  $\boxed{\mathcal{L}_s = \sum_k^S \mathcal{L}_k + \lambda \sum_k^S \left\| \nabla \mathcal{L}_k \right\|^2}$
  $\mathbf{W}_{t+1}^s \leftarrow \mathbf{W}_t^s - \eta \nabla \mathcal{L}_s$
end for
$\mathbf{W}_t^k \leftarrow ClientUpdate(\nabla \mathcal{L}_s)$

Minimize

ClientRepresentation($\mathbf{W}_t^k$):

if k is first client to start training **then**
  $\mathbf{W}_t^k \leftarrow$ initial weights from server
else
  $\mathbf{W}_t^k \leftarrow \mathbf{W}_{t-1}^{k-1}$ from the previous $ClientUpdate(\nabla \mathcal{L}_s)$
end if
for each local client epoch, i=1,2,..k **do**
  Calculate hidden representation $\phi(x_t^k)$
end for
return $\phi(x_t^k)$ and $\mathbf{Y^k}$ to server

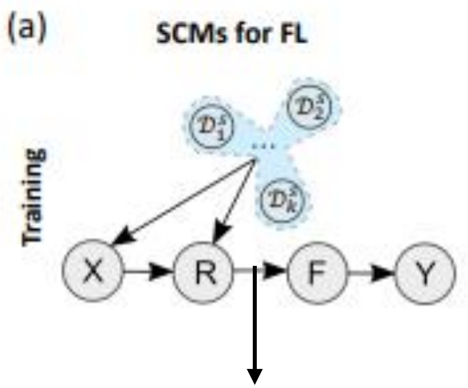
ClientUpdate:

for each client $k \in S$ **do**
  $\mathbf{W}_{t+1}^k \leftarrow \mathbf{W}_t^k - \eta \nabla \mathcal{L}_s$
end for
return $\mathbf{W}_{t+1}^k$ to server

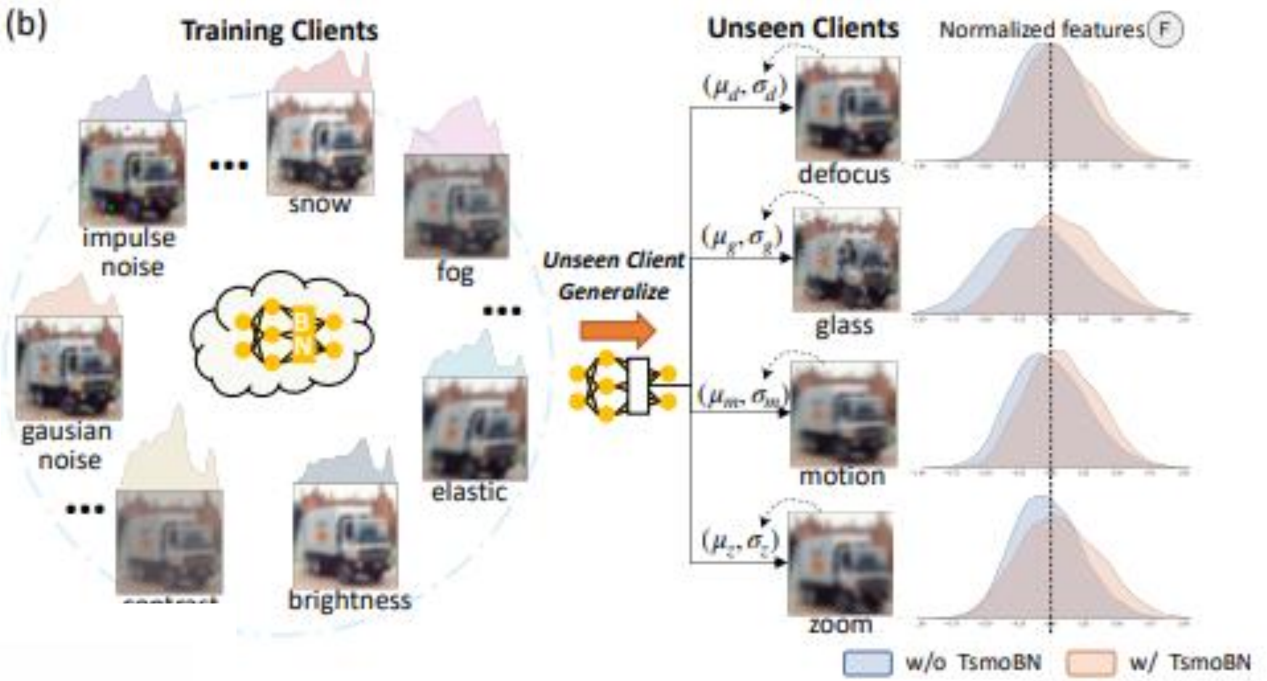
Improves generalization and is effective against membership inference attacks

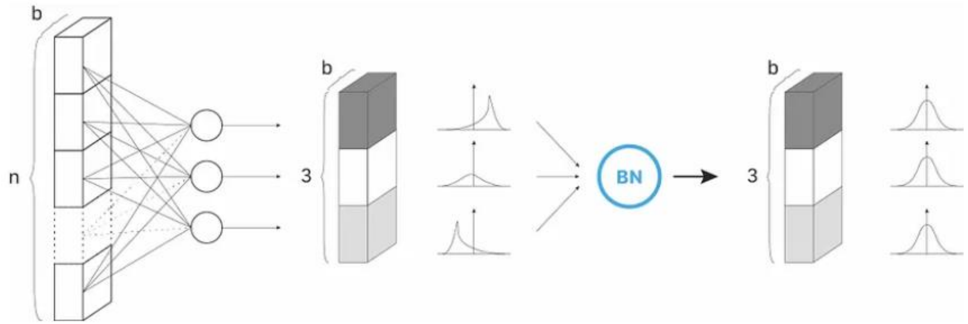
Francis, S., Tenison, I. and Rish, I., 2021. Towards causal federated learning for enhanced robustness and privacy. *arXiv preprint arXiv:2104.06557*.

Post-processing Methods



Normalization step: Batch normalization



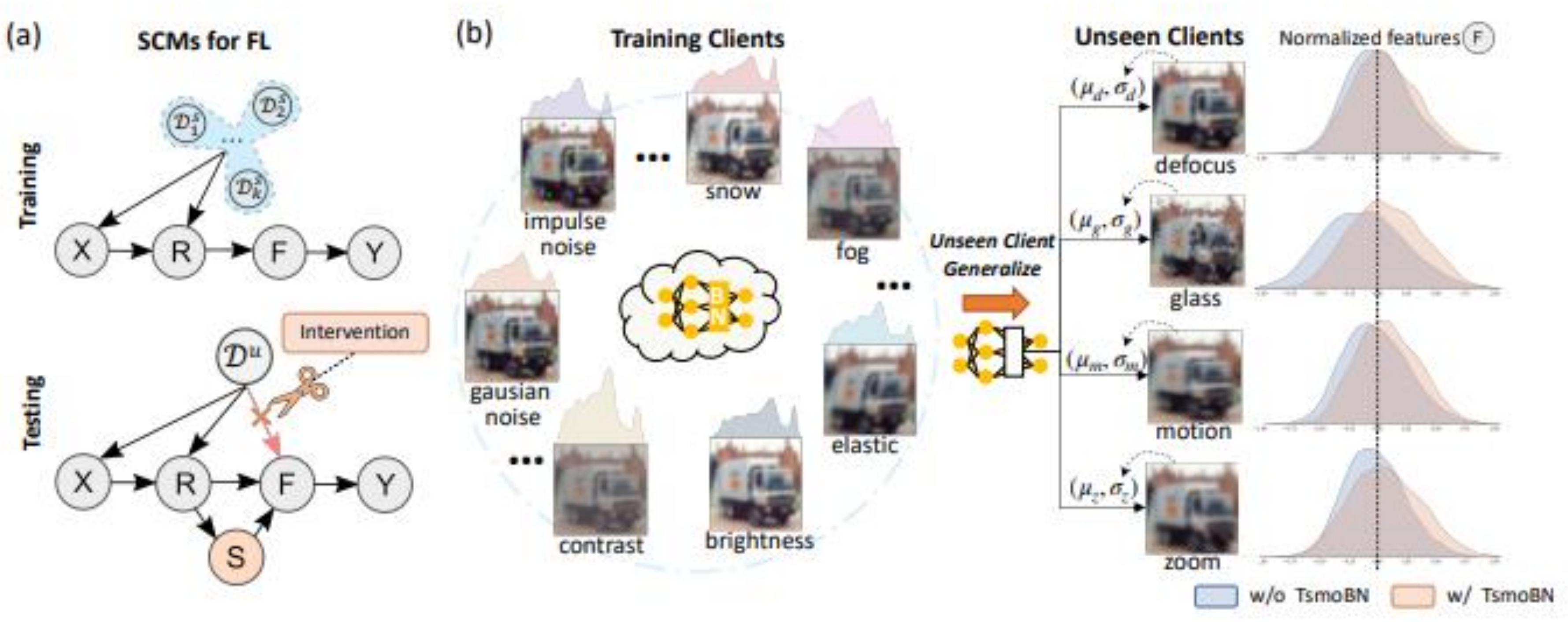


$$(1)\ \mu = \frac{1}{n}\sum_i Z^{(i)} \qquad (2)\ \sigma^2 = \frac{1}{n}\sum_i (Z^{(i)} - \mu)^2$$

$$(3)\ Z^{(i)}_{norm} = \frac{Z^{(i)} - \mu}{\sqrt{\sigma^2 - \epsilon}} \qquad (4)\ \breve{Z} = \gamma * Z^{(i)}_{norm} + \beta$$

Jiang, M., Zhang, X., Kamp, M., Li, X. and Dou, Q., 2021. TsmoBN: Interventional Generalization for Unseen Clients in Federated Learning. *arXiv preprint arXiv:2110.09974*.

Post-processing Methods



Calculate the mean and variance pair at test time in BN to normalize features

Summary

- Only generalization aspect of causal models have been explored
  - Defenses against membership attacks
  - What about other types of attacks?
- Causal models for Differential privacy?
- How to deal with scalability?
- Benchmark datasets?
- Check out our paper - https://arxiv.org/pdf/2302.06975.pdf