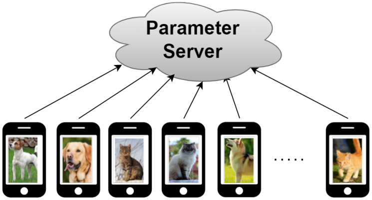
Compressed-VFL: Communication-Efficient Learning with Vertically Partitioned Data

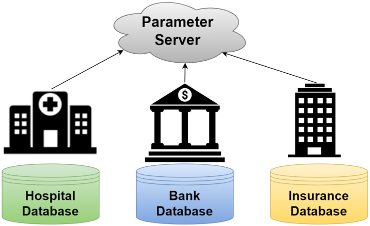
Timothy Castiglia[1], Anirban Das[1], Shiqiang Wang[2], Stacy Patterson[1]

[1]Department of Computer Science, Rensselaer Polytechnic Institute

[2]IBM T.J. Watson Research Center, IBM Research

7/20/2022

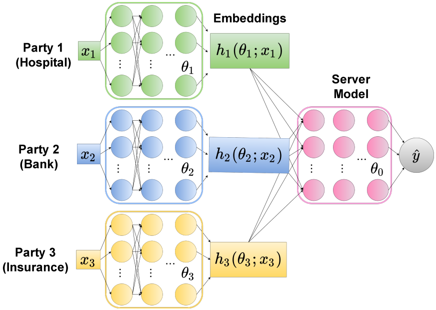




- Horizontal Federated Learning
- Feature space shared: images of animals
- Sample space not shared: each party stores different individual animals

- Vertical Federated Learning
- Sample space shared: individuals
- Feature space not shared: medical information, financial information, vehicle accident reports.

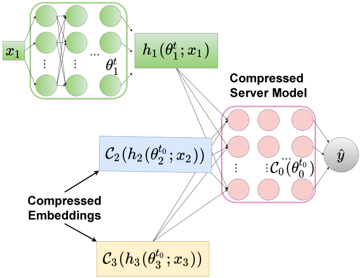
Message Passing in VFL



- Local feature extractors/models
- VFL shares embeddings: intermediate outputs from local models
- Large communication overhead
- Message compression

- HFL works with compression (Stich et al., 2018; Wen et al., 2017; Karimireddy et al., 2019).
- Several works in VFL (Liu et al., 2019; Hu et al., 2019; Chen et al., 2020)
- No work applies embedding compression in VFL

- Parties agree on mini-batch of samples
- Parties update local model parameters for Q local iterations using stochastic coordinate descent
- To calculate model updates, parties share:
    - Embeddings from each party for mini-batch $\mathcal{B}$
    - Server model parameters
- At the start of each round, parties share embeddings and server shares prediction model parameters
- Message compression applied to all messages



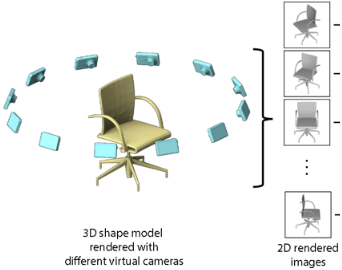
Convergence Results

Theorem

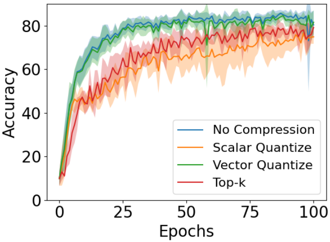
Under Assumptions 1-5, if $\eta^r = \eta$ for all iterations and satisfies $\eta^r \leq \frac{1}{16Q \max\{L, \max_m L_m\}}$, then the average squared gradient over R communication rounds of Algorithm 1 is bounded by:

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left[\|\nabla F(\Theta^r)\|^2\right] \leq \frac{4\left[F(\Theta^0) - \mathbb{E}\left[F(\Theta^T)\right]\right]}{\eta T} + 6\eta L \sum_{m=0}^{M} \frac{\sigma_m^2}{B} + \frac{92Q^2}{R}\sum_{m=0}^{M} H_m^2 G_m^2 \sum_{r=0}^{R-1} \sum_{j=0, j\neq m}^{M} \mathcal{E}_j^r.$$

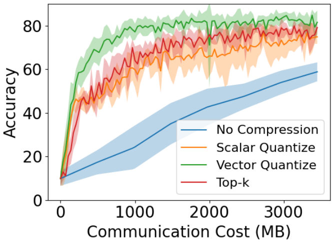
- C-VFL converges at a rate of $O\left(\frac{1}{\sqrt{T}}\right)$
  - $T = R \cdot Q$ is the total number of local iterations
- C-VFL can afford compression error without loss in convergence speed



- Compare no compression with three compression schemes:
  - Top-k Sparsification
  - Scalar Quantization
  - Vector Quantization
- ModelNet10: 3D models with 12 camera views
- 4 parties, 3 views each
- Parties train 2-layer CNN models
- Server with one fully-connected layer



(a) ModelNet10 by epochs



(b) ModelNet10 by cost

- 2 bits per embedding component
- Slight decrease in accuracy plotted by epochs: iterations/batches
- Significant improvement when plotted by communication cost

Thank you!

Contact me at castit@rpi.edu