Ethics of Data Science – Part III

Week 4: Federated Learning

Governance

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

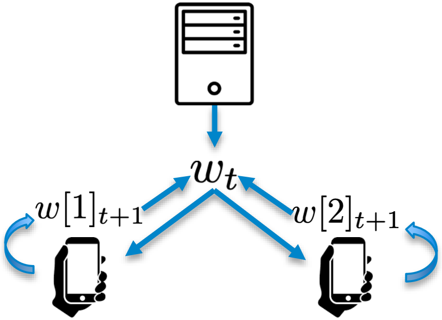
Co-training involves optimizing a federated loss

$$\min_{w} \sum_{k=1}^{m} p_k F_k(w)$$

$$\text{where } F_k(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} f(w; x[k]_j)$$

Co-training via federated averaging (FedAvg)

$$
\begin{aligned}
w[k]_t &= w_t, \quad \text{(broadcast)} \\
d[k]_t &= -\eta \nabla_{w=w_t} f(w; x[k]_t), \quad \text{(local gradient)} \\
w[k]_{t+1} &= w[k]_t + d[k]_t, \quad \text{(local update)} \\
w_{t+1} &= \sum_{k=1}^{m} q_k w[k]_{t+1}, \quad \text{(aggregation)}
\end{aligned}
$$



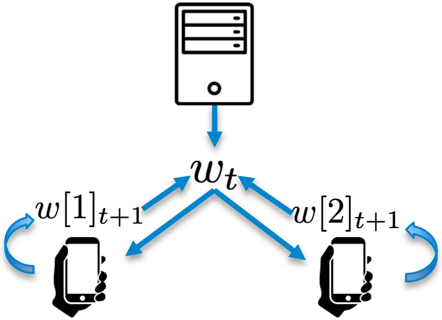
- Is it safe to communicate $w[k]_t$?
- Is averaging the estimated parameters a good way to perform model averaging?
- What if we have dataset shift in each client?

McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

Co-training via federated SGD (FedSGD)

$$d[k]_{t+1} = \nabla_{w=w_t} f(w; x[k]_t), \quad \text{(local gradient)}$$

$$w_{t+1} = w_t - \eta \sum_{k=1}^{m} p_k \nabla d[k]_{t+1}, \quad \text{(global learning)}$$



- Is it safer to communicate $d[k]_t$ than $w[k]_t$?
- Is averaging the gradients a better way to perform model averaging than averaging parameters?
- What if we have dataset shift in each client?

McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

Distributed versus federated learning

$$\bar{w}[k]_t = \sum_j d_{k,j} w[j]_t, \quad \text{(broadcast and local aggregation)}$$

$$w[k]_{t+1} = w[k]_t + \eta \nabla_{w=\bar{w}[k]_t} f(w; x[k]_t), \quad \text{(local learning)}$$

- Larger communication overhead
- No central model available in the end
- Dataset shift in each client can be handled through neighborhood-based kernels/weighting

Many flavors of federated learning

- $$w_{t+1} = \frac{m}{L} \sum_{k \in S_t} p_k w[k]_t, \text{ where } |S_t| = L$$
  - S_t = first L clients to respond in that cycle [1]
  - S_t = set that excludes 10% of clients with extreme values [2]
  - S_t = set that includes L randomly selected clients [3]
- $$w_{t+1} = w_t + \frac{\alpha}{m} \sum_{k=1}^{m} p_k \left(w[k]_t - w_t \right)$$ Convergence forced globally
- $$w[k]_{t+1} = w[k]_t + \eta \nabla_w f(\bar{w}[k]_t; \ldots) + \alpha(w[k]_t - w_t)$$ Convergence forced locally (FedProx, [4], or clipping)
- $$d[k]_{t+1} = -\eta \nabla_{w=w_t} f(w; x[k]_t) + \epsilon, \epsilon\ N(0, \sigma^2)$$ Differentially private federated SGD (DP-FedSGD)

1: similar to Li, Xiang, et al. "On the convergence of FedAvg on non-iid data." *arXiv preprint arXiv:1907.02189* (2019).
2: similar to Ghosh, Avishek, et al. "Robust federated learning in a heterogeneous environment." *arXiv preprint arXiv:1906.06629* (2019).
3: Abadi, Martin, et al. "Deep learning with differential privacy." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.
4: similar to Sahu, Anit Kumar, et al. "On the convergence of federated optimization in heterogeneous networks." *arXiv preprint arXiv:1812.06127* 3 (2018): 3.
5: McMahan, H. Brendan, et al. "Learning differentially private recurrent language models." *arXiv preprint arXiv:1710.06963* (2017).

TensorFlow Federated