Critical Review and Scalability Assessment of SplitNN

Jizhou Guo

1 Introduction

This is a paper aimed to solve the summer intern task: Vertical Federated Learning Challenge. This paper proposes a critical review of SplitNN[1] in the Context of VFL, report of Pseudo-Distributed SplitNN Implementation and scalability assessment of SplitNN.

2 Critical Review of SplitNN in the Context of VFL

SplitNN is a technique which can cover some shortages of FedAvg[2], such as lower computational cost on the client sides. However, it also faces several challenges(specifically as illustrated in Figure 2(c) of the cited paper):

- **Cost and Scalability**: The computational and communication cost of SplitNN on the server side is very high because:
  - For Split learning for vertically partitioned data(illustrated in Figure 2(c) of the cited paper), the outputs at the cut layer are concatenated, and when the number of clients is large, the concatenated result becomes very large and the computational cost would be unaffordable. It also may result in lower convergence speed.
  - For other types of SplitNN(illustrated in Figure 2(a) and Figure 2(b) of the cited paper), the server side can only handle request from only one client each time, and this harms the efficiency of SplitNN.

  Therefore, the SplitNN may face scalability challenges when dealing with a large number of clients or data sources.

  Potential Solution:

  - For split learning for vertically partitioned data, we can apply PCA or encoder models to reduce the dimensionality of the data.
  - For other types of SplitNN, we can handle the requests from different clients at the same time, then we average the gradients and send them back to the clients.

- **Dependence on a single modality**: For split learning for vertically partitioned data, as the server model may overly depend on a certain modality of patient data, this may drastically lower the performance when some client is unavailable.

  Potential Solution: We can use attention mechanism to better capture the relationship of different modalities and integrate them.

- **Data Privacy and Anti-attack**: Although the SplitNN does not need to share raw data, some sensitive information can still be revealed from the outputs at the cut layer. Plus, the SplitNN is quite vulnerable because it relies on a single centralized server. And if the server or the channel is unavailable the whole system can not work.

  Potential Solution:

  - We can add some noise on the outputs (such as dropout) for data privacy. This would not harm the training as adding noise is a common technique to alleviate overfitting in training.
  - We can use Homomorphic Encryption and Secure Multi-party Computation for data privacy.
  - We can use Blockchain to store the history updates of the model to prevent data from being tampered with.
  - We can use Knowledge Distillation so that the clients can make inferences alone.

3 Pseudo-Distributed SplitNN Implementation

We can use the **"torch.distributed"** API and deploy the model locally on different processes. The code and readme file are within the zip file in the attachment. Under the default hyperparameters in the code, the validation accuracy on Client $1, 2, 3$ in the last epoch are $94\%, 96\%, 96\%$ accordingly. The loss curve on the server is as follows:

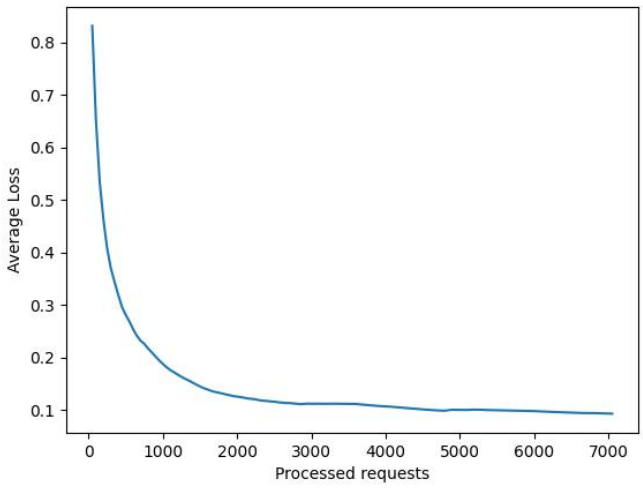

Figure 1: Average loss curve on the server

4 Scalability Assessment of SplitNN

As illustrated in section 2, the SplitNN may face scalability challenges, and making it inappropriate to support a large number of parties in VFL due to the high computation cost and high communication cost. Our experimental implementation further illustrates the conclusion as our code shows that when the amount of data for each client is constant, the server running time and communication cost increase linearly with the number of clients.

References

[1] Vepakomma et al. "Split learning for health: Distributed deep learning without sharing raw patient data" *arXiv preprint arXiv:1812.00564* (2018).

[2] Konečný et al. "Federated Learning: Strategies for Improving Communication Efficiency" *arXiv preprint arXiv:1610.05492* (2016)