EPI RQ4 Research Update: Privacy Preserving Distributed Machine Learning

Saba Amiri

s.amiri@uva.nl

Supervisor: Adam Belloum, Advisor: Eric Nalisnick

Promoters: Sander Klous, Leon Gommans

Multiscale Networked Systems Group

17 January 2022

Recap

- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Different federated learning fusion schemes
  - Different non-i.i.d data distribution schemes
  - Impact of differential privacy
- Differentially private synthetic data generation
  - Distributed datasets
  - Privacy preserving
  - Non-i.i.d data distribution among nodes
  - Skewed/imbalanced dataset
- Modeling tail behavior of long-tailed distribution in generative models
  - Synthetic data generation probabilistic machine learning models
  - Verify their ability to capture the tail behavior of different data distributions, especially when the tail is long
  - The long tail – as will be shown at the end of this presentation – can be masked during the privacy preserving distributed machine learning process
- Distributed learning pipeline
  - Research on using Vantage6[1] as the distributed machine learning infrastructure (as opposed to more generic solutions, e.g. managing the distributed pipeline through use of Pytorch distributed)

[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Update

- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Paper submitted to IJCAI 2022 this month
  - Work in progress to extend the results (ETA: March 2022)
  - Next phase already planned (ETA: July 2022)
- Differentially private synthetic data generation
  - Distributed datasets
  - Privacy preserving
  - Non-i.i.d data distribution among nodes
  - Skewed/imbalanced dataset
- Modeling tail behavior of long-tailed distribution in generative models
  - Synthetic data generation probabilistic machine learning models
  - Verify their ability to capture the tail behavior of different data distributions, especially when the tail is long
  - The long tail – as will be shown at the end of this presentation – can be masked during the privacy preserving distributed machine learning process
- Distributed learning pipeline
  - Research on using Vantage6[1] as the distributed machine learning infrastructure (as opposed to more generic solutions, e.g. managing the distributed pipeline through use of Pytorch distributed)

[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Update

- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Paper submitted to IJCAI 2022 this month
  - Work in progress to extend the results
  - Next phase already planned for summer
- Differentially private synthetic data generation
  - Ablation study finished (Planning on submitting results, ETA: March 2022)
  - New model already beating SOTA benchmarks
  - Early results on semantic integrity improvements (ETA: March 2022)
- Modeling tail behavior of long-tailed distribution in generative models
  - Synthetic data generation probabilistic machine learning models
  - Verify their ability to capture the tail behavior of different data distributions, especially when the tail is long
  - The long tail – as will be shown at the end of this presentation – can be masked during the privacy preserving distributed machine learning process
- Distributed learning pipeline
  - Research on using Vantage6[1] as the distributed machine learning infrastructure (as opposed to more generic solutions, e.g. managing the distributed pipeline through use of Pytorch distributed)

[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Update

- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Paper submitted to IJCAI 2022 this month
  - Work in progress to extend the results
  - Next phase already planned for summer
- Differentially private synthetic data generation
  - Ablation study finished
  - New model already beating SOTA benchmarks
  - Early results on semantic integrity improvements
- Modeling tail behavior of long-tailed distribution in generative models
  - Early results in, clear improvements observed on artificial data
  - Working on formalizing the method, extend results (ETA: June 2022)
- Distributed learning pipeline
  - Research on using Vantage6[1] as the distributed machine learning infrastructure (as opposed to more generic solutions, e.g. managing the distributed pipeline through use of Pytorch distributed)

[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Update

- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Paper submitted to IJCAI 2022 this month
  - Work in progress to extend the results
  - Next phase already planned for summer
- Differentially private synthetic data generation
  - Ablation study finished
  - New model already beating SOTA benchmarks
  - Early results on semantic integrity improvements
- Modeling tail behavior of long-tailed distribution in generative models
  - Early results in, clear improvements observed on artificial data
  - Working on formalizing the method, extend results
- Distributed learning pipeline
  - Finished, will be submitted to a FL workshop (ETA: August 2022)

[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Thank you!

My direct collaborators in chronological order

- Serge van Haag (AI)
- Boris Egelie (AI)
- Tidi Stamatiou (AI)
- Carlijn Nijhuis (Computer Science)
- Mike Schouw (Computer Science)
- Jetske Beks (Computer Science)
- Willemijn Beks (Computer Science)
- Yu Wang (Computer Science)
- Simon Tokloth (Data Science)