OpenMined

WOMEN of OM
Study Group

SEMI-SUPERVISED KNOWLEDGE TRANSFER FOR DEEP LEARNING FROM PRIVATE TRAINING DATA

Paper Session I

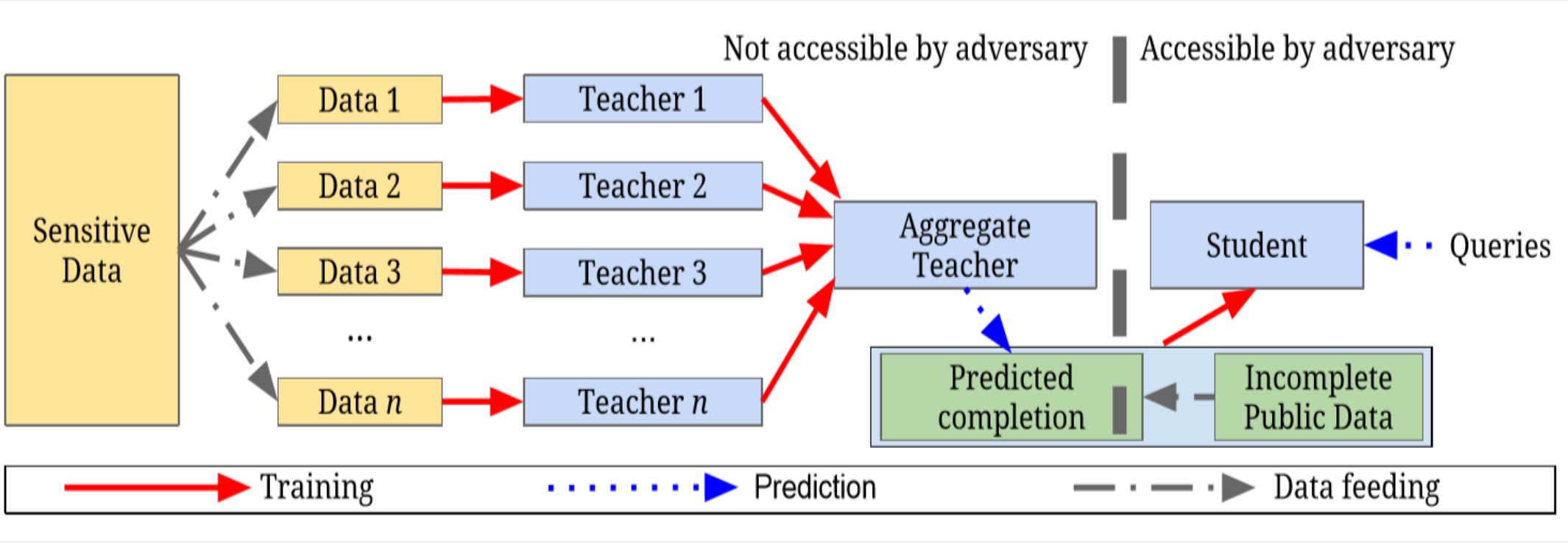
by Zumrut Muftuoglu

10/05/2020

- Some machine learning applications involve training data that is sensitive.
- Careful analysis of the model may reveal sensitive information, because a model may inadvertently and implicitly store some of its training data.
- PATE approach combines, in a black-box fashion, multiple models trained with disjoint datasets, such as records from different subsets of users.

OVERVIEW OF PATE:

- An ensemble of teachers is trained on disjoint subsets of the sensitive data,
- A student model is trained on public data labeled using the ensemble.


Not accessible by adversary
Accessible by adversary
Sensitive Data
Data 1
Data 2
Data 3
...
Data n
Teacher 1
Teacher 2
Teacher 3
...
Teacher n
Aggregate Teacher
Student
Queries
Predicted completion
Incomplete Public Data
Training
Prediction
Data feeding

TEACHER MODELS

- Each teacher is a model trained independently on a subset of the data whose privacy one wishes to protect.
- The data is partitioned to ensure no pair of teachers will have trained on overlapping data.
- Any learning technique suitable for the data can be used for any teacher.
- Training each teacher on a partition of the sensitive data produces n different models solving the same task. At inference, teachers independently predict labels.

AGGREGATION MECHANISM

- The aggregation step is a crucial component of PATE.
- It enables knowledge transfer from the teachers to the student while enforcing privacy.

STUDENT MODEL

- PATE's final step involves the training of a student model by knowledge transfer from the teacher ensemble using access to public—but unlabeled—data.

WHAT DO THEY WANT TO SHOW?

- They show how PATE can scale to learning tasks with large numbers of output classes and uncurated, imbalanced training data with errors

WHAT DID THEY?

- In the scope of this study, they improve the LNMax mechanism which adds Laplace noise to teacher votes and outputs the class with the highest vot

HOW DO THEY SHOW?

- They introduce new noisy aggregation mechanisms for teacher ensembles that are more selective and add less noise,and prove their tighter differential-privacy guarantees.
- They build the new mechanisms on two insights:
  - the chance of teacher consensus is increased by using more concentrated noise
  - lacking consensus, no answer need be given to a student.

STEPS OF STUDY/1

- They add Gaussian noise with an accompanying privacy analysis in the RDP framework.

STEPS OF STUDY/2

- the aggregation mechanism is now selective: teacher votes are analyzed to decide which student queries are worth answering. This takes into account both the privacy cost of each query and its payout in improving the student's utility.
- Their analysis shows that these two metrics are not at odds and in fact align with each other: **the privacy cost is the smallest when teachers agree, and when teachers agree, the label is more likely to be correct thus being more useful to the student.**

STEPS OF STUDY/3

- They propose and study an interactive mechanism that takes into account not only teacher votes on a queried example but possible student predictions on that query.
- Now, queries worth answering are those where the teachers agree on a class but the student is not confident in its prediction on that class.
- This third modification aligns the two metrics(mentioned in previous slayt) discussed above even further: queries where the student already agrees with the consensus of teachers are not worth expending their privacy budget on, but queries, where the student is less confident, are useful and answered at a small privacy cost.

TO WRAP UP

- During the study it is aimed that :
  - performing PATE on a task with a larger number of classes
  - showing the privacy-utility tradeoffs offered by PATE on data that is class imbalanced and partly mislabeled.

Thank You!