Trustworthy ML Reading Group

31 Jan 2024

About Us

- Focus on technical and non-technical works
  - Work can be your own or others
- Modeled loosely on ML Collective groups
  - Focus on informal and interdisciplinary work

- Meetings are meant to be informal and accessible for different experience levels
- Ask “stupid questions”
  - Explore your own ideas and get critical feedback

About Us

- Go deep!
  - Since we have similar interests, it is very helpful to be as technical as needed to explain an idea
- We are fully collaborative. Papers discussions are meant to be shared among members

Conduct

- **From Ethos of MLC…**
  - https://mlcollective.org/wiki/code-of-conduct/
- **Highlights**
  - Expectation of Confidentiality
  - Reporting -> send me a direct message

Happening Today

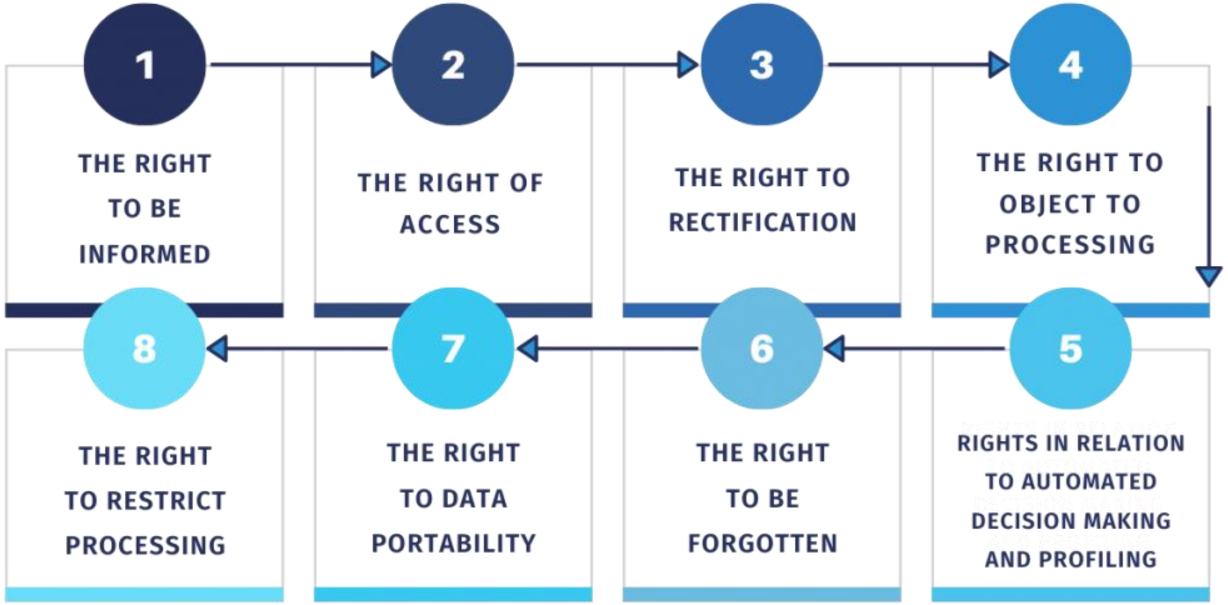
Machine Unlearning

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2019, December 9). ***Machine unlearning***. arXiv.org. https://arxiv.org/abs/1912.03817. IEEE S&P

Motivations

- Aspects of Data Governance
  - Security, availability, integrity
  - Data should remain consistent
  - Must comply with Laws (e.g., GDPR, CA Consumer Privacy Act, etc.)

- GDPR's provides data subject rights provide legal basis and motivation for Trustworthy AI systems



Source: https://dataprivacymanager.net/what-are-data-subject-rights-according-to-the-gdpr/

Motivations

- Naïve Approaches to “the Right to Forget”:
  - Retrain model from scratch
  - Train model in increments, use saved parameters as “check points”
- Both approaches may result in a long time to unlearn data points
- “The Right to Forget” produces large time overhead
  - ^a key critique from people (e.g., Google)

Problem

- Online Learning
  - A machine learning model is continually updated
  - Individuals reserve the right to have data deleted (EU GDPR, CCPA)
  - ML models make are complicated due to:
    - Memorization
    - Black-box nature

Problem

- *…Unlearning guarantees that training on a point and unlearning it afterwards will produce the same distribution of models that not training on the point at all, in the first place, would have produced*

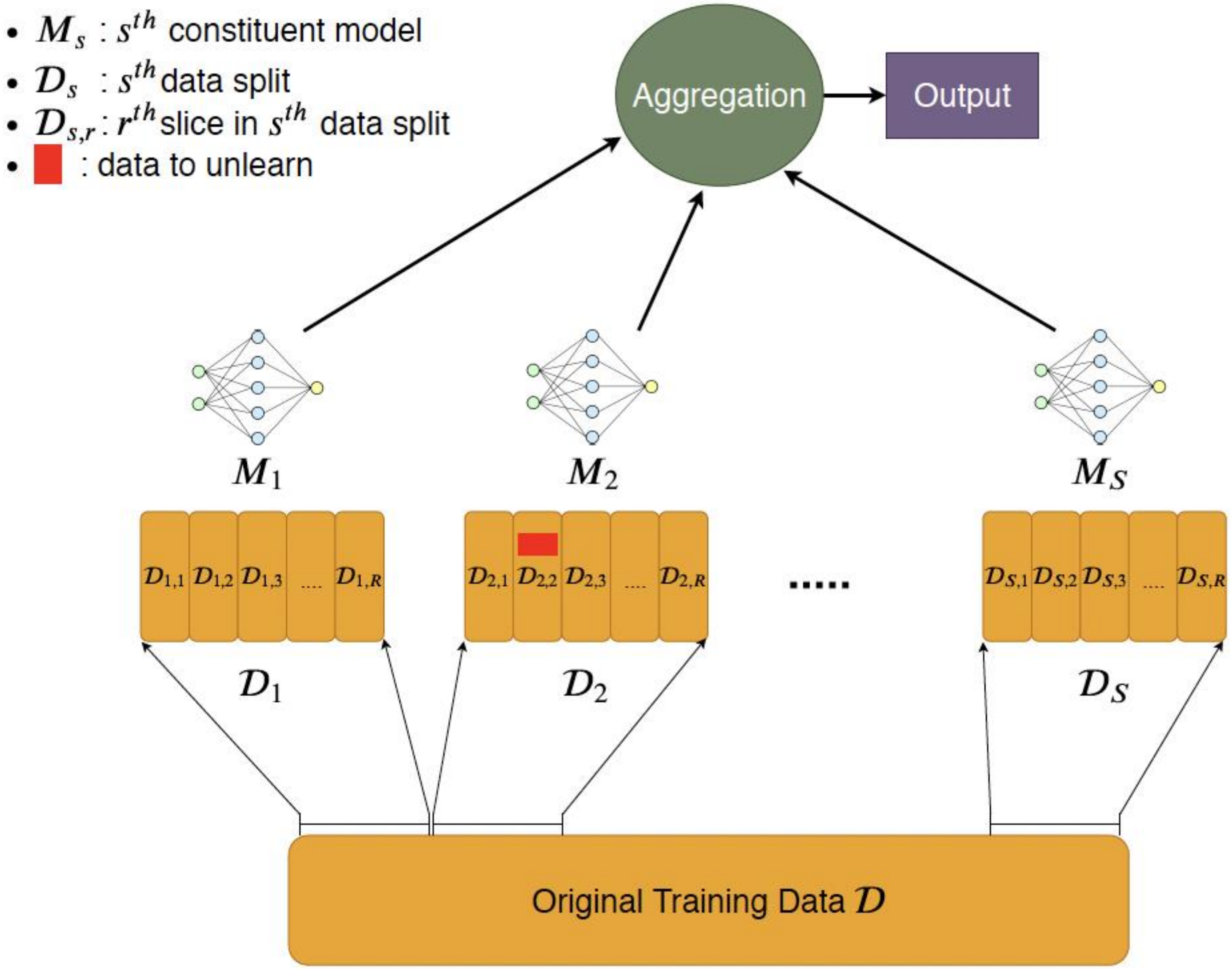
Approach

- **S**harded **I**solated **S**liced **A**ggregated training

  Key point: Limit the influence of individual data points. Train model in increments
- Models are trained on separate shards. An aggregation mechanism outputs the popular prediction

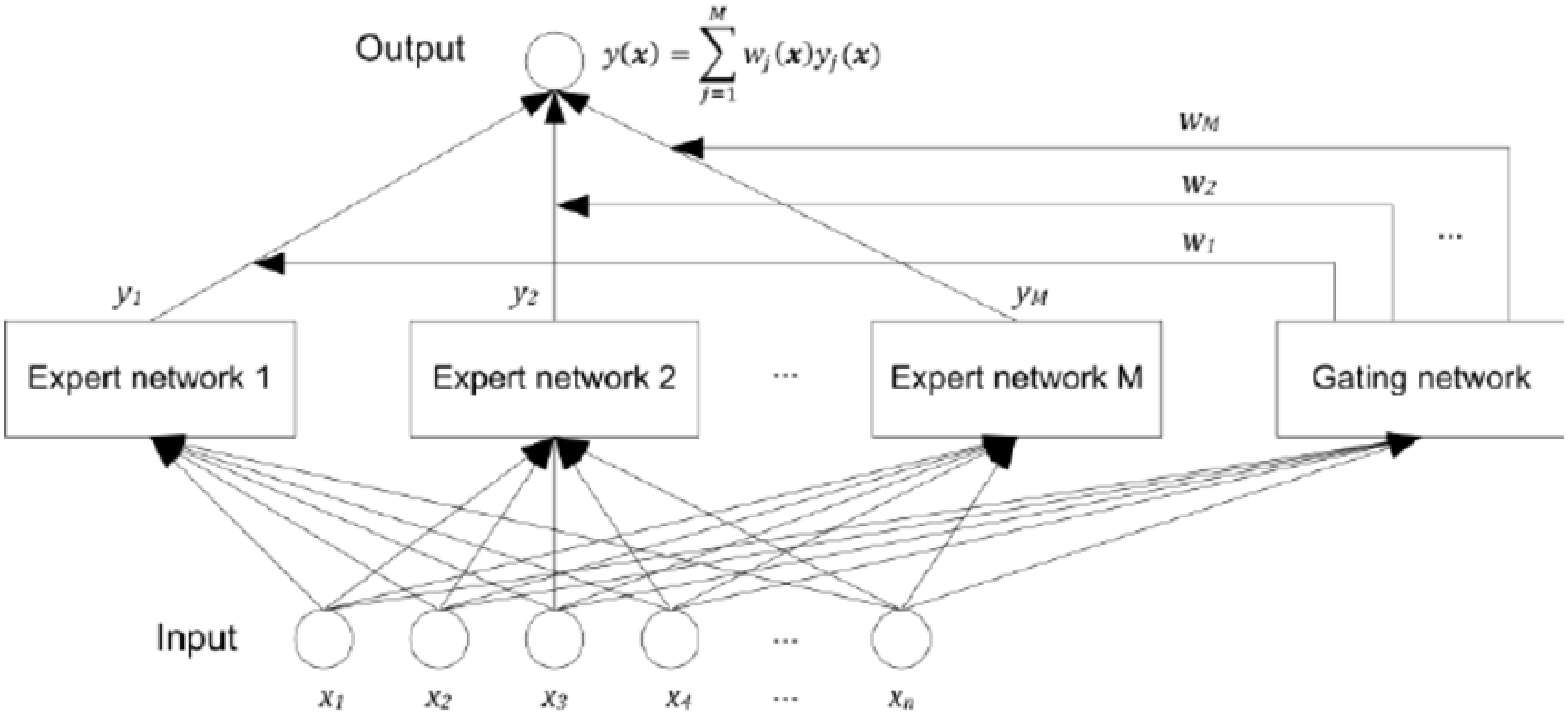
Approach

- Data is split into disjoint groups (shards).
- Models are trained on Disjoint shards. Information is not shared among shards. Shards can further divided into slices (batch like)
- **Key Idea:** multiple checkpoints may be saved within a shard. We can start from closest checkpoint not including the data to unlearn



Interesting

- Approach is similar to mixture of experts model



Problems to Consider

- “Weak Learners”
  - Training on small datasets hurts complex tasks
- Generalization ability of model
- Tradeoffs
  - Accuracy vs Time (to retrain)
  - Small Shards and Complex Learning Tasks
  - How to verify unlearning to end-users
    - External auditing
    - Core Question: How much do individual points influence a model?

Personal Takeaways

- Affirms that privacy should not be “one size fits all”
  - E.g. Reiterates that Differential privacy can’t solve all privacy problems
- I like that they consider solution scalability (e.g., simple and complex tasks)
- They are vocal about using an iterative design. They are very transparent on their research approach
- Practicality vs Novelty

Read more

- Y. Cao and J. Yang, “**Towards making systems forget with machine unlearning**,” in 2015 IEEE Symposium on Security and Privacy. IEEE, 2015, pp. 463–480. [Online]. Available: https://ieeexplore.ieee.org/document/7163042/
- Papernot, Nicolas et al. “**Scalable Private Learning with PATE**.” *ArXiv* abs/1802.08908 (2018): n. pag.
- A. Ginart, M. Y. Guan, G. Valiant, and J. Zou, “**Making AI forget you: Data deletion in machine learning**,” CoRR, vol. abs/1907.05012, 2019. [Online]. Available: http://arxiv.org/abs/1907.05012