Does Domain Generalization Provide Inherent Membership Privacy

Divyat Mahajan, Shruti Tople, Amit Sharma

Microsoft Research

PRIML/PPML'20

Microsoft

Machine Learning and Privacy Attacks

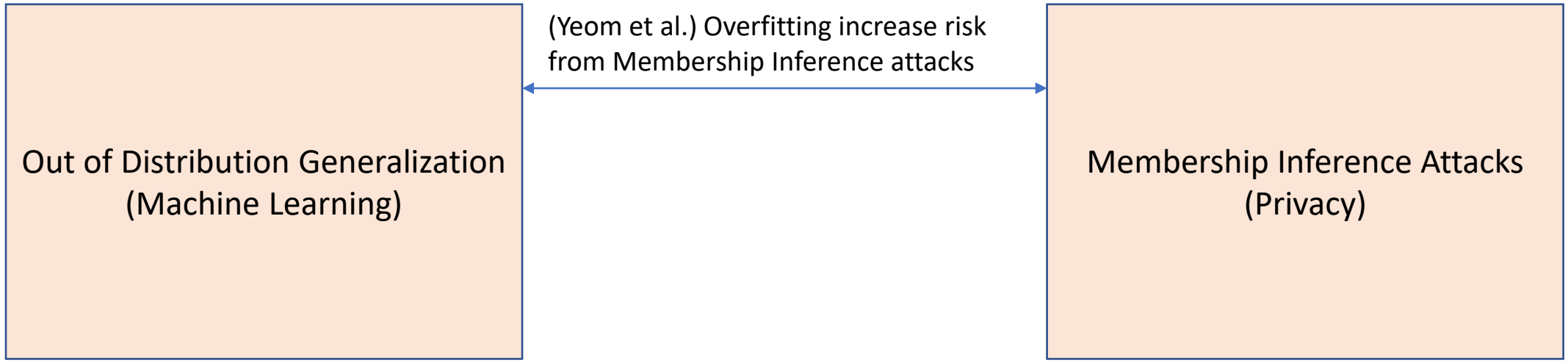
Out of Distribution Generalization
(Machine Learning)

Model trained on some hospital is used in another hospital (Distribution Shift)

Membership Inference Attacks
(Privacy)

Model should not leak information about the sensitive training data of patients

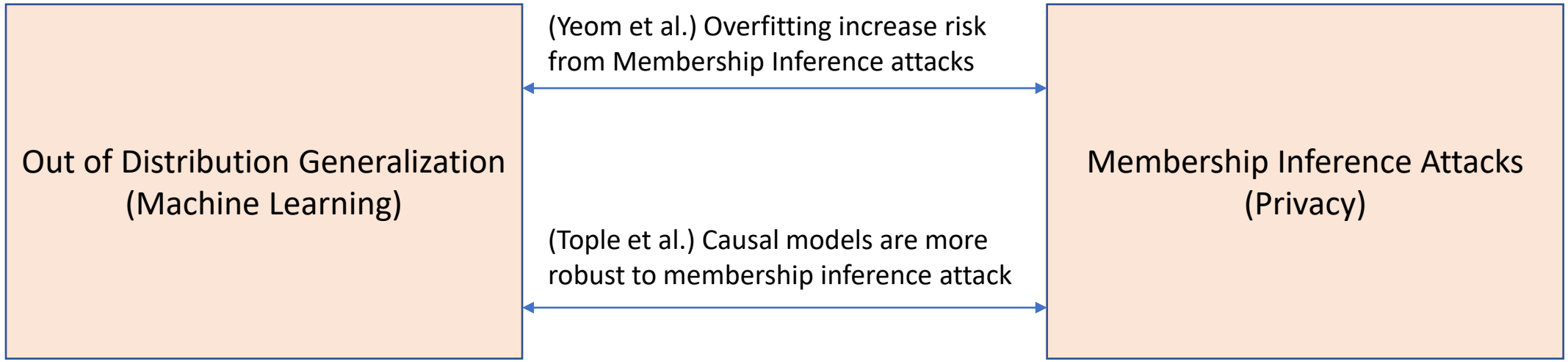
Machine Learning and Privacy Attacks



Model trained on some hospital is used in another hospital (Distribution Shift)

Model should not leak information about the sensitive training data of patients

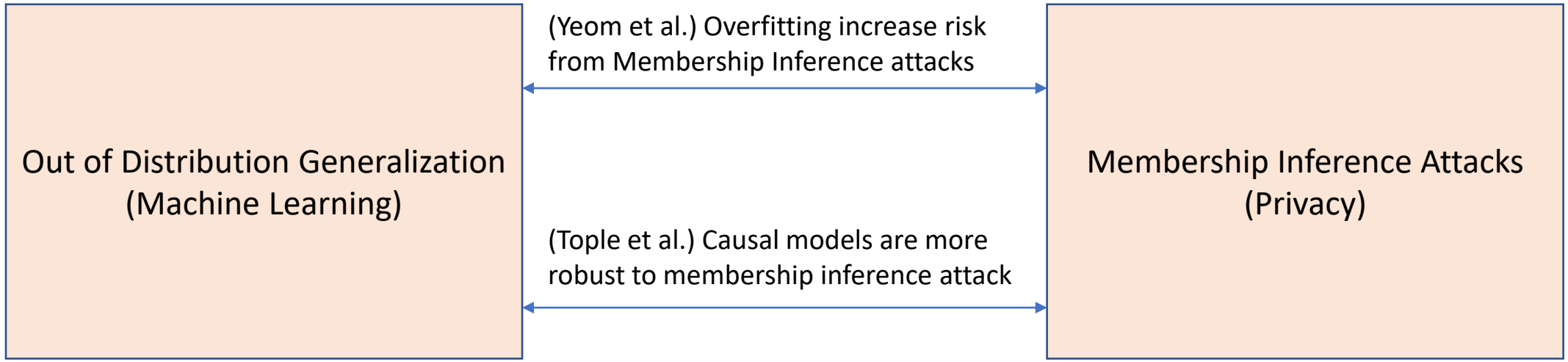
Machine Learning and Privacy Attacks



Model trained on some hospital is used in another hospital (Distribution Shift)

Model should not leak information about the sensitive training data of patients

Machine Learning and Privacy Attacks



Model trained on some hospital is used in another hospital (Distribution Shift)

Model should not leak information about the sensitive training data of patients

Our Contribution

We show the connection between Domain Generalization and Membership Inference Attacks

Domain Generalization -> Membership Inference Attacks

- Domain Generalization setup exposes ML models to domain shifts during training and expects them to generalize to unseen domains at test time

Domain Generalization -> Membership Inference Attacks

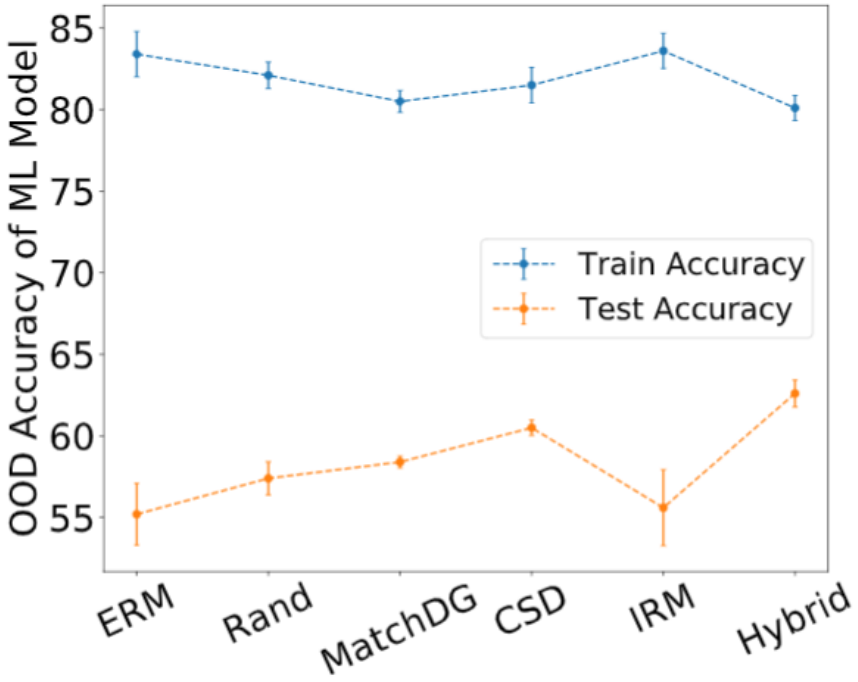
- Domain Generalization setup exposes ML models to domain shifts during training and expects them to generalize to unseen domains at test time
- Domain Generalization methods improve generalization on unseen data distributions and lead to robustness against Membership Inference Attacks

Domain Generalization -> Membership Inference Attacks

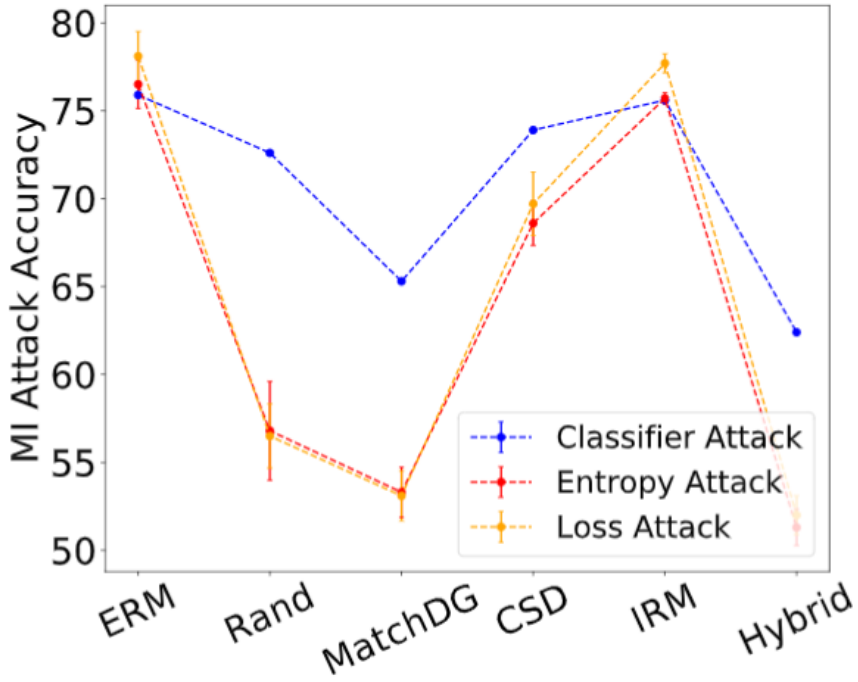
- Domain Generalization setup exposes ML models to domain shifts during training and expects them to generalize to unseen domains at test time
- Domain Generalization methods improve generalization on unseen data distributions and lead to robustness against Membership Inference Attacks

Chest X-Ray Dataset

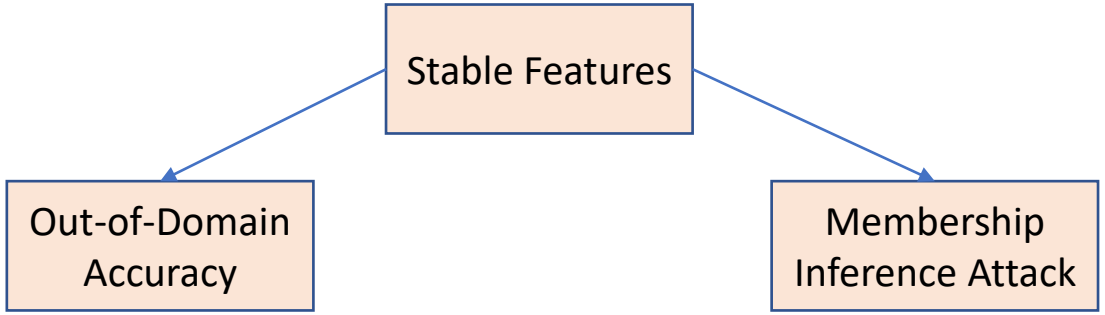
Out-of-Domain Accuracy



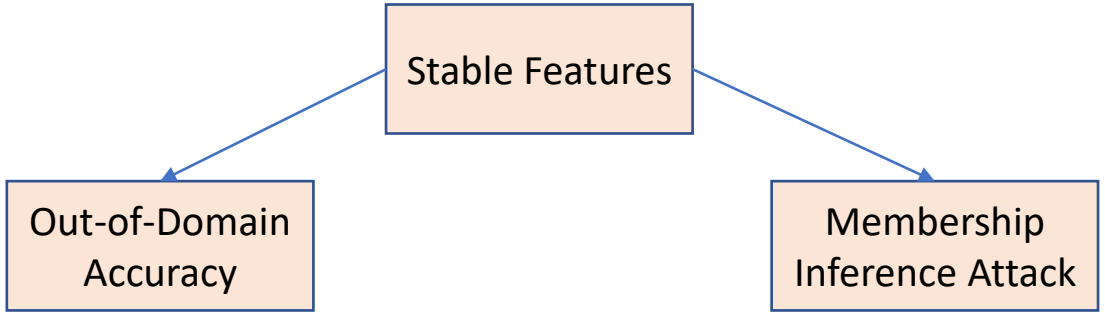
Membership Inference Attack Accuracy
(Lower is Better)



Membership Inference Attacks->Domain Generalization

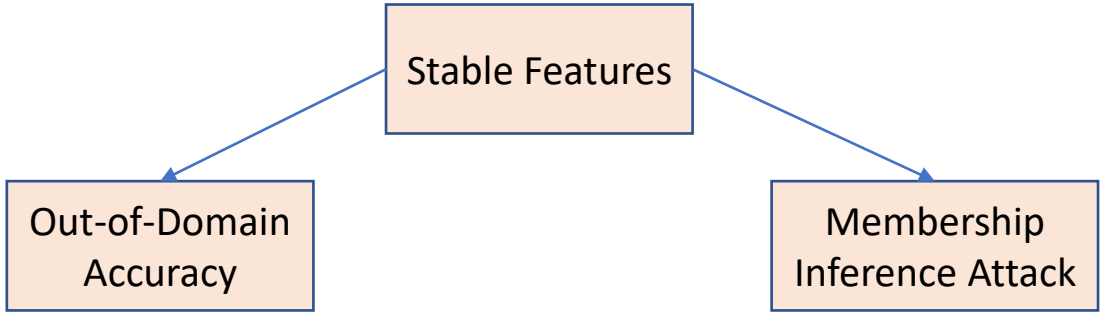


Membership Inference Attacks->Domain Generalization



- Better Out-of-Domain Accuracy is not a sufficient metric to evaluate Domain Generalization algorithms
  - Stable Features determine the true generalization performance but they are hard to determine

Membership Inference Attacks->Domain Generalization



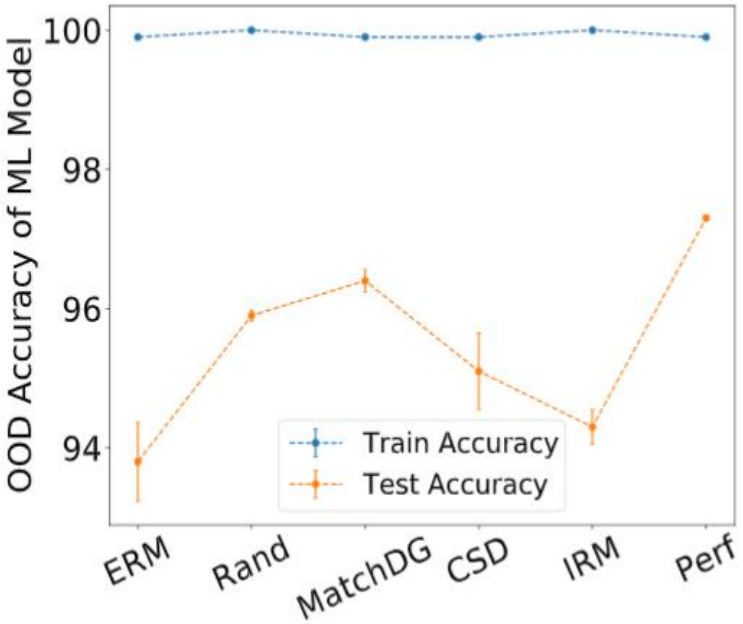
- Better Out-of-Domain Accuracy is not a sufficient metric to evaluate Domain Generalization algorithms
  - Stable Features determine the true generalization performance but they are hard to determine
- Membership Inference Attacks can be used to evaluate Domain Generalization as they capture the extent to which stable feature were learnt

Membership Inference Attacks->Domain Generalization

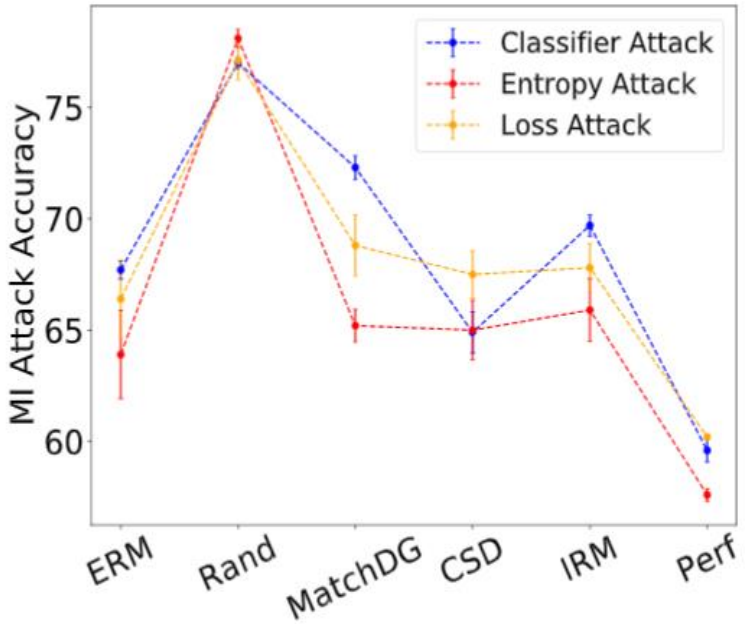
- High Out-of-Domain accuracy does not always imply more robustness against Membership Inference Attacks
- Membership Attack Accuracy correlates with metrics to compute stable features (Mean Rank)

Rotated MNIST Dataset

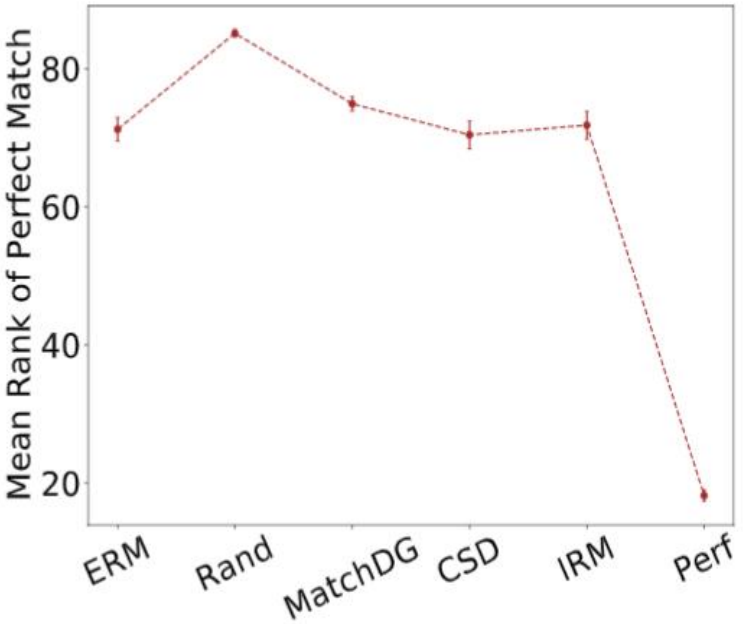
Out-of-Domain Accuracy



Membership Inference Attack Accuracy
(Lower is Better)



Mean Rank: Stable Features
(Lower is Better)



Conclusion

- Please visit RobustDG (https://github.com/microsoft/robustdg) for further details

Toolkit for Building Robust ML models that generalize to unseen domains (RobustDG)

Divyat Mahajan, Shruti Tople, Amit Sharma

ICML 2020 Paper | MatchDG paper | Privacy & DG Connection paper

For machine learning models to be reliable, they need to generalize to data beyond the train distribution. In addition, ML models should be robust to privacy attacks like membership inference and domain knowledge-based attacks like adversarial attacks.

To advance research in building robust and generalizable models, we are releasing a toolkit for building and evaluating ML models, *RobustDG*. RobustDG contains implementations of domain generalization algorithms and includes evaluation benchmarks based on out-of-distribution accuracy and robustness to membership privacy attacks. We will be adding evaluation for adversarial attacks and more privacy attacks soon.

It is easily extendable. Add your own DG algorithms and evaluate them on different benchmarks.

Thank You

Chat with us during the poster session!