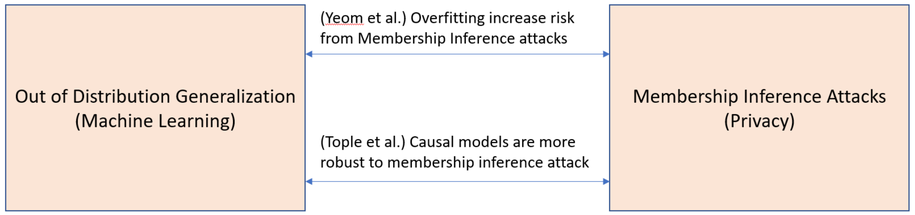
Does Domain Generalization Provide Inherent Membership Privacy?

Divyat Mahajan [1] Shruti Tople [2] Amit Sharma [1]

[1]Microsoft Research, India [2]Microsoft Research, UK

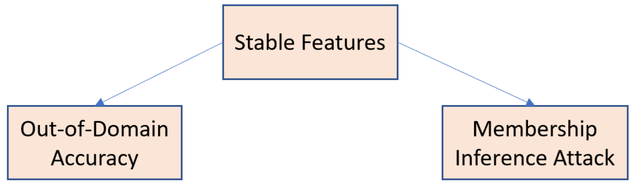
https://github.com/microsoft/robustdg

Machine Learning and Privacy Attacks



- **Yeom et al:** ML models with poor generalization leads to worse membership privacy
- **Tople et al:** ML models that learn stable features (causal models) are more robust to MI attacks
- **Our Contribution**: We propose a connection between Domain Generalization (DG) and MI Attacks under stable features

Domain Generalization and Membership Privacy



- Stable Features: Features X_c such that their relationship with the outcome $p(Y|X_c)$ is invariant to changes in data distributions
- DG algorithms aim to learn stable features using heuristics given varying data distributions at training time
- DG algorithms are evaluated using accuracy on unseen data distributions or out-of-distribution (OOD) accuracy
- OOD accuracy evaluated using restricted test domains, hence does not always imply learning stable features

Contributions:

- Higher OOD accuracy on restricted number of test domains does imply imporved robustness against MI attacks
- MI attacks can be used to evaluate DG algorithms and obtain better understanding about their generalization

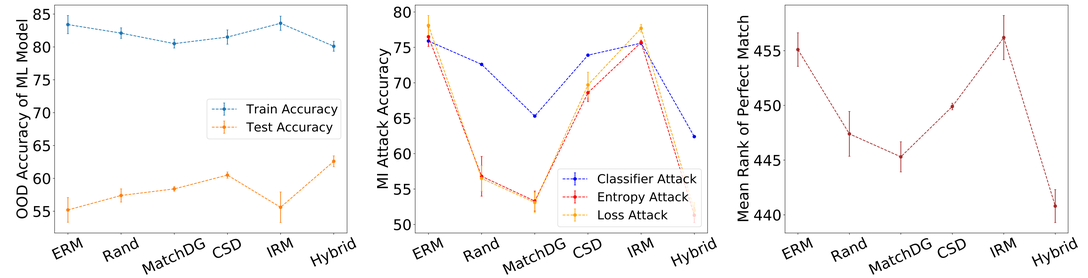
References:

- Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S. (2018, July). Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (pp. 268-282). IEEE.
- Tople, S., Sharma, A., Nori, A. (2020, November). Alleviating Privacy Attacks via Causal Learning. In International Conference on Machine Learning (pp. 9537-9547). PMLR.

Membership Inference Attacks Results

Table: Connection between learning stable representation, high OOD accuracy and better membership privacy.● denotes the training method satisfies the metric, ○ means it does not.

Metrics	Dataset	ERM	Random-Match	IRM	CSD	MatchDG	Perfect-Match/ Hybrid
Stable Features (Theory)		○	○	●	●	●	●
Stable Features (Practice)	Rotated-MNIST	●	○	●	●	●	●
	Fashion-MNIST	●	○	●	●	●	●
	ChestXray	○	○	○	○	●	●
Out-of-Domain Accuracy	Rotated-MNIST	○	●	○	○	●	●
	Fashion-MNIST	○	○	○	○	●	●
	ChestXray	○	○	○	●	●	●
MI Attack Accuray	Rotated-MNIST	●	○	●	●	●	●
	Fashion-MNIST	●	●	●	●	●	●
	ChestXray	○	○	○	○	●	●



(a)

Figure: OOD generalization and MI privacy attacks on Chest X-Ray dataset. left: OOD Accuracy (higher is better), center: MI Attack Accuracy and right: Mean Rank (lower is better)

Attribute Inference Attacks Results

Table: Attribute Attack Accuracy for different datasets with domains as an attribute.

Dataset	Random Guess	ERM	Random-Match	IRM	CSD	MatchDG	Perfect-Match/ Hybrid
Rotated-MNIST	14.3%	27.6 (0.84)	20.4 (0.46)	26.6 (0.71)	25.0 (0.40)	19.3 (0.62)	**16.6 (0.55)**
Fashion-MNIST	14.3%	26.6 (0.59)	21.8 (0.77)	23.8 (0.86)	26.9 (0.93)	21.9 (0.29)	**21.5 (0.97)**
ChestXray	33.3%	61.8 (1.02)	58.4 (0.58)	63.4 (2.28)	67.8 (3.05)	57.9 (0.58)	**57.1 (0.21)**