FY2024/2024年度

AI Security and Privacy Team

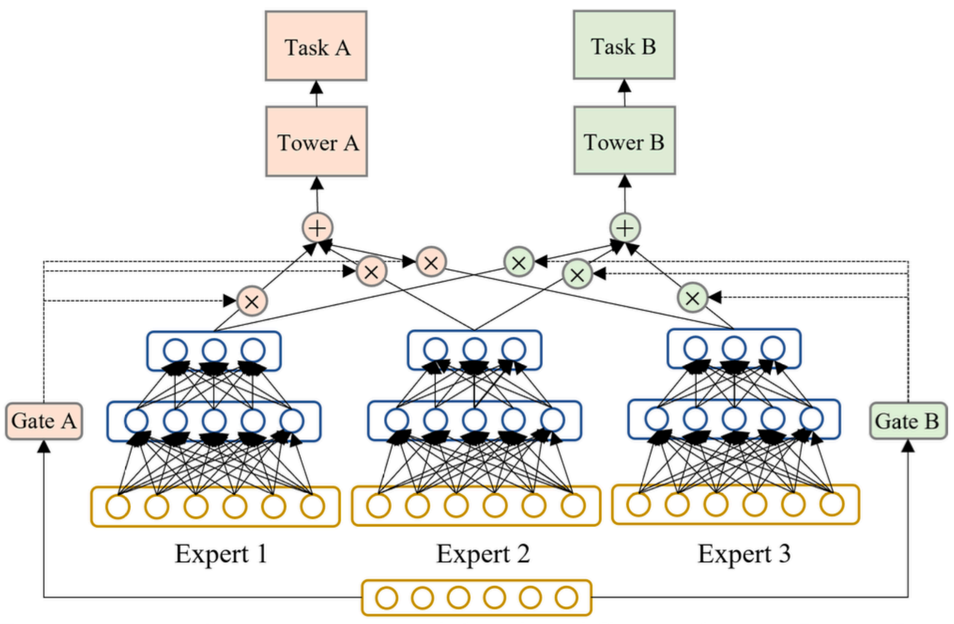
Jun Sakuma
人工知能セキュリティ・プライバシーチーム　佐久間 淳

RIKEN Center for Advanced Intelligence Project

MoME: Mixture-of-Masked-Experts for Efficient Multi-Task Recommendation (SIGIR'24)

Motivation: deep neural network suffer from:
- High computational consumption
- Poor scaling-up ability

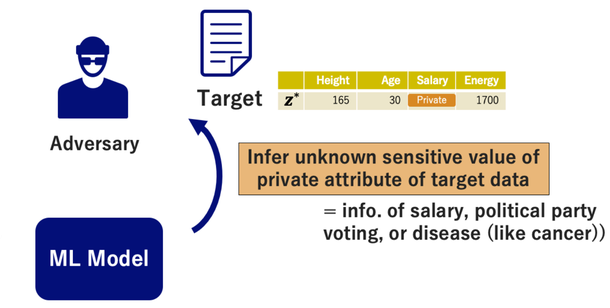
Our solution -- MoME
1. **Mixture of Experts (MoE):** Activate a subset of parameters (experts) for each input
2. 2. One overparameterized base network and a mixture of binary masks!



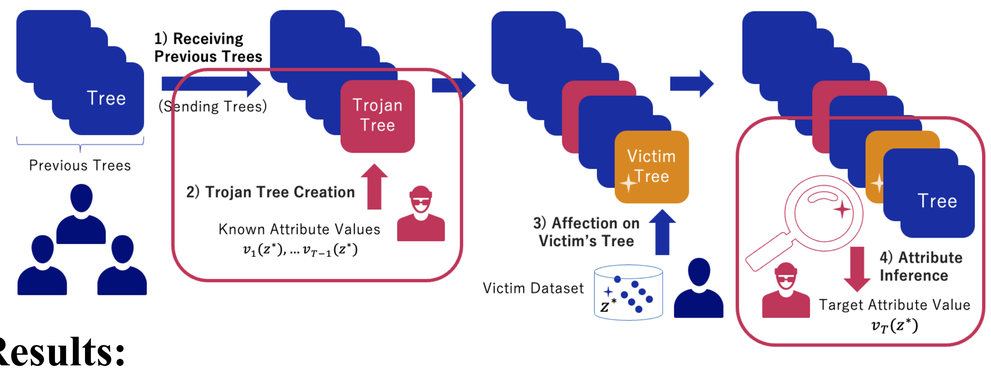
Better performance with significantly **reduced model size**.

Trojan attribute inference attack on gradient boosting decision trees (IEEE Euro S&P'24)

Motivation:
- Attribute Inference attack (AIA) has been investigated for DNN intensively, but not for boosting trees
- Establish AIA for boosting trees in federated learning setting



Key Idea:
Insert a trojan tree that cause attribute lealage



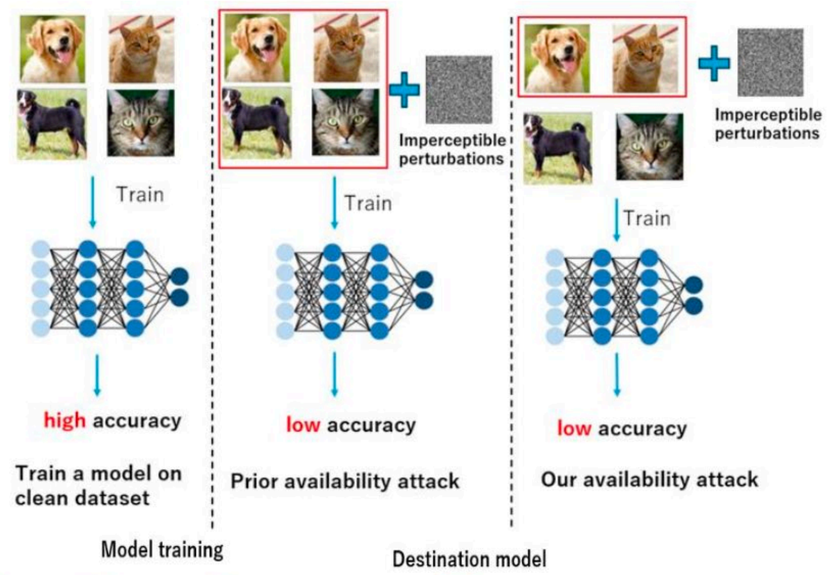
Results:
Our AIA always predicts the ground-truth att. value

Theorem 1. *Fix n_{max} and r_{max}. For $c = 1, \ldots, C$, define w_c and $(\tilde{w}_{c,min}, \tilde{w}_{c,max})$ as (5) and (6). Suppose that f_{vic} is trained after the next round that f_{Trojan}, generated by **MakeTrojan** with the width ε, is trained. Suppose the following conditions hold:*
(A) for all $z \in D^{(vic)} \setminus \{z^*\}$, there exists $t \in \{1, \ldots, T-1\}$ such that $\|x_t(z) - x_t(z^*)\| \geq \varepsilon$;
(B) $|D_{L^*}| \leq n_{max}$; and,
(C) for all $z \in D^{(vic)} \setminus \{z^*\}$, $|r_{k+1}(z)| \leq r_{max}$.
*Then, the output of **Infer** (Algorithm 5) coincides with the ground-truth value of the target attribute.*

	HSKC		Census	
	Basic	SimFL	Basic	SimFL
Predict Most Common	0.53	0.51	0.56	0.59
Whitebox AIA [17]	0.53	0.51	0.56	0.59
Imputation [27]	0.78	0.74	0.75	**0.80**
Trojan AIA (Ours)	**0.99**	**0.99**	**0.89**	**0.80**

Parameter Matching Attack: Enhancing Practical Applicability of Availability Attacks (CSS'24 symposium best paper award)

Motivation:
data owners want their data to be used by humans and not by ML model.

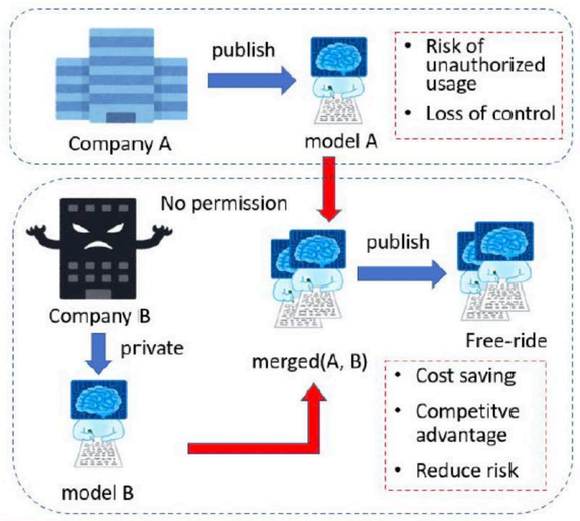


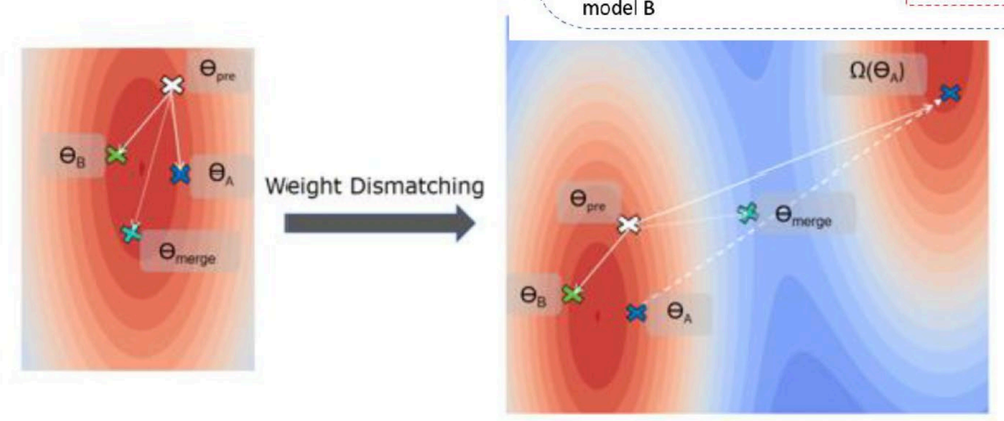
High-level idea:
- Optimize the perturbation so that the resulting model performs similarly to a low-performance model.
- Resulting model is trained on a mixture of perturbed dataset and clean dataset.

Results:
Show better results than previous methods when poison ratio is smaller than 100%

Model Merging Disruption via Weight Dispatching (on going)

- **Motivation:**
  - Privacy issue in model merging
- **Proposed method**
  - Transfer weight basin
  - Equivalent transformation for consistent performance
- **Results:**
  - Model merging disrupted from A to B





Members 2024

PI: Jun Sakuma
Researcher: Yu Zhe, Sun Lu
Visiting Researcher: Hideitu Hino, Kazuto Fukuchi, Takao Murakami, Tatsuya Mori, Youhei Akimoto, Yuki Koike, Yuwei Sun
Part-timers: Daiki Nishiyama, Oiso Hideyuki, Kudo Mikoto, Ragib, Nihal, Shiwen An, Takaaki Toda, Junhao Wei, Zling He

Achievement in 2024

Journals:
Transactions on Machine Learning Research 1
Proceedings of Machine Learning Research 1
International Journal of Information Security 2
Information Geometry 1
ACM Transactions on Evolutionary Learning and Optimization 1
Evolutionary Computation 1他

Conference papers:
NeurIPS'24 2
AISTAS'24 1
Euro S&P 2024 2
IEEE BigData 1
GECCO '24 1
ASIACCS 2024 1 他