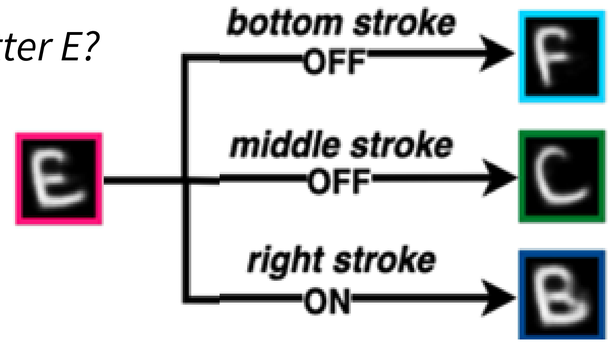
AI Security and Privacy Team

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

Unsupervised Causal Binary Concepts Discovery with VAE

Motivation
- Explain classification using high-level symbolic concepts
- "Data X is classified as class Y because X **have A, B** but **do not have C**"
- *Why this image is letter E?*
  Because…

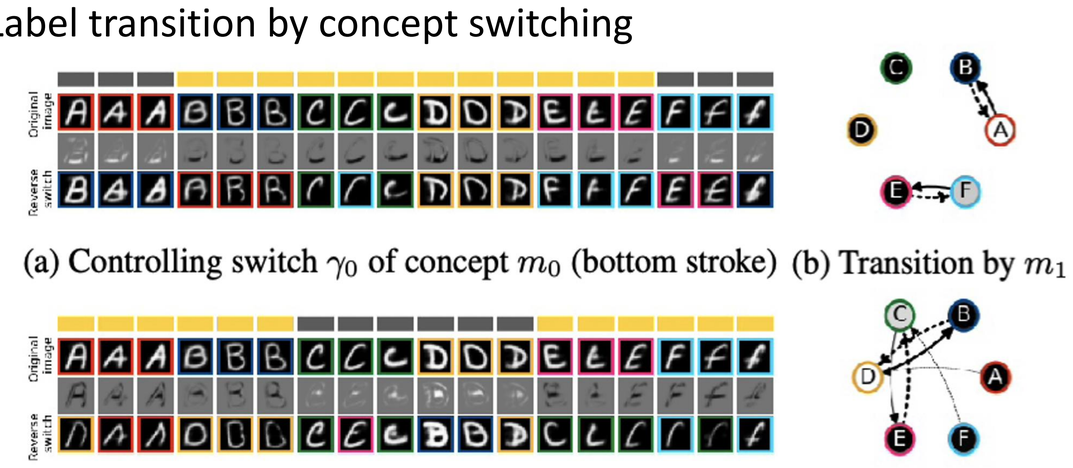


Objective function

$$\mathcal{L}(X) = \frac{1}{|X|}\sum_{x\in X}\mathcal{L}_{VAE}(x) + \lambda_{CE}\mathcal{L}_{CE}(X) + \lambda_R\mathcal{L}_R(x)$$

(1) VAE objective — **(2) Causal effect objective** — **(3)Interpretability regularizer**

Label transition by concept switching



(a) Controlling switch γ_0 of concept m_0 (bottom stroke) (b) Transition by m_1

(c) Controlling switch γ_1 of concept m_1 (middle stroke) (d) Transition by m_1

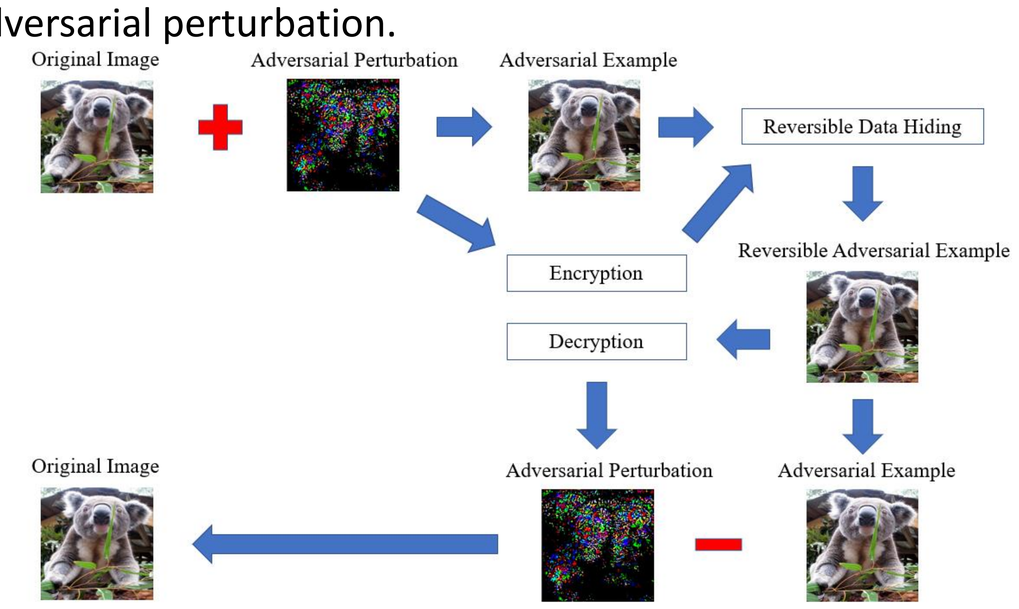
(Thien Q. Tran, K. Fukuchi, Y. Akimoto, J. Sakuma, AAAI2022)

Reversible Adversarial Example

Motivation
- Control how user's data is recognized and used by AI via exploiting the properties of adversarial examples

Objective
- The reversible adversarial example can be correctly recognized and used by the AI model specified by the user
- Other unauthorized AI models cannot recognize the reversible adversarial example correctly

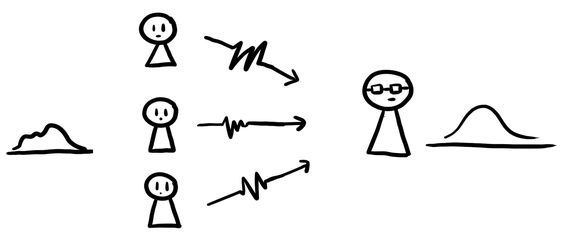
Realization of reversible adversarial example
- Combine three technologies, adversarial example, reversible data hiding for exact recovery of adversarial perturbation, and encryption for selective control of AIs who can remove adversarial perturbation.



(Jiayang Liu, Weiming Zhang, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, under review)

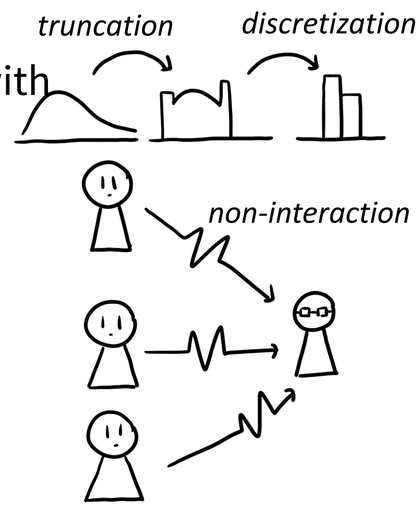
One-bit Submission for Locally Private Quasi-MLE

Motivation
- Find the model that most likely generated the data with LDP



- SGD-based algorithm is proposed [Bhowmick+,2018]
  - Pros: achieves the minimax optimal rate.
  - Cons: (i) High communication cost, (ii) Long waiting time, (iii) Requires bounded derivatives for loss function (data should be bounded)

Contribution : developed an algorithm with
- low communication cost (by *random discretization*)
- short waiting time (by *non-interactive algorithm*)
- applicable to unbounded and unknown data domain (by *truncation*)
- theoretical considerations:
  - we show *asymptotic normality* of the outputs
  ← can be used for statistical tests
  - Made some remarks and recommendations for the users of our algorithm



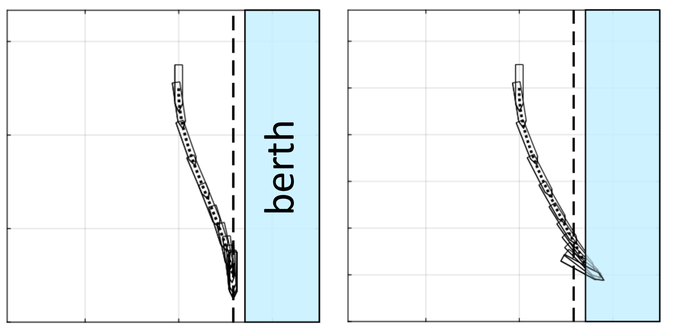
comparison of computational costs

Id	Scenario	Server	User	Wait
Bhowmick2018	$\mathbf{X}$ pub $\mathbf{X}$ pri	$32(k+d)$ $32d$	$32d$ $32d$	$O(n)$
Ours	$\mathbf{X}$ pub $\mathbf{X}$ pri	$32k$ 0	1 $k+1$	$O(1)$

(H. Ono, K. Minami, H. Hino, to appear in AISTATS2022)

Safe Berthing Control via Min-Max Optimization

Motivation
- Automatic berthing control task is a risk-critical task of obtaining a controller for berthing of a ship
- A controller is obtained on simulation, which includes uncertainties: estimated ship maneuvering model, weather conditions, etc.

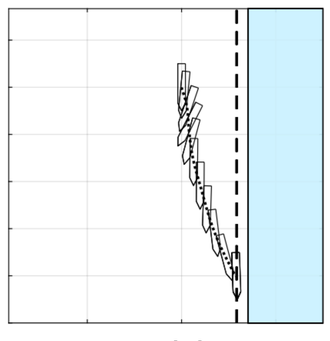


Controller optimized on a specific simulation environment tested on **(a)** the same environment and **(b)** a different environment. Collision to berth was observed when the test environment is different from the training environment.

Approach
- Solving min-max optimization $\min_{x\in X}\max_{y\in Y} f(x,y)$

i.e., optimizing controller x under the worst situation $y \in Y$
- A numerical solver relying on black-box local minimizers, CMA-ES, is proposed with its local convergence guarantee

Controller optimized by the proposed approach tested on **(c)** the worst environment $y_{worst} \in Y$. Collision was avoided successfully even on the worst case.



(Y. Akimoto, Y. Miyauchi, A. Maki, ACM TELO 2022)