FY2023/2023年度

AI Security and Privacy Team

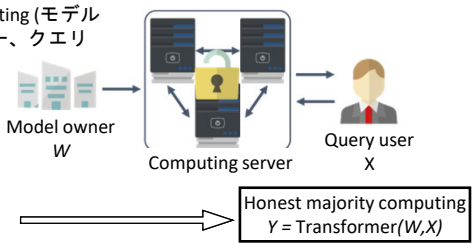
Jun Sakuma

人工知能セキュリティ・プライバシーチーム　佐久間 淳

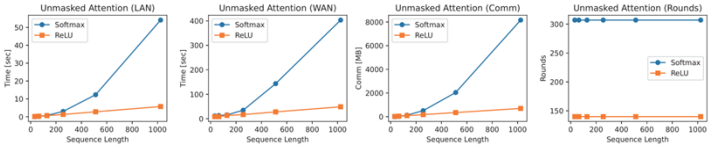
RIKEN Center for Advanced Intelligence Project

Privformer: Privacy-preserving Transformer with MPC, EuroS&P 2023

- APIモデルにおけるセキュアマルチパーティ計算(MPC)によるTransformer実装
- Honest 3-party setting (モデル所有者、サーバー、クエリユーザー)



- Performerバリアント(softmaxの代わりにReLU)とカーネルベースのアテンションマトリックスの近似を活用することで効率が向上 $O(S^2) \to O(S)$

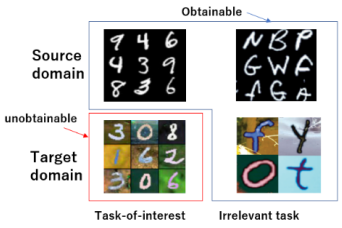


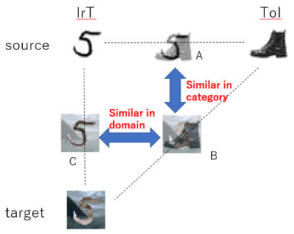
Yoshimasa Akimoto, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, Privformer: Privacy-preserving Transformer with MPC, 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)

Zero-shot Domain Adaptation Based on Dual-level Mix and Contrast

- **Zero-shot domain adaptation**

(1)トレーニングドメイン(ソース)とテストドメイン(ターゲット)
(2) ターゲットタスク(TgI)と関連タスク(IrY)
(3)ターゲットドメインにおけるTgIタスクへの適応





- Achieved 1st average classification accuracy

Domain shift	G→C	G→E	G→N	C→G	N→G
ZDDA	55.63	62.61	62.84	59.53	62.44
CoCoGAN	53.51	68.07	68.55	69.23	69.18
Wang2020	**66.90**	**71.51**	-	70.86	72.33
DF-ZSDA [3]	62.76	65.85	63.84	**88.45**	63.44
Ours	62.68	69.26	**82.4**	80.64	**83.08**

- **Proposal**

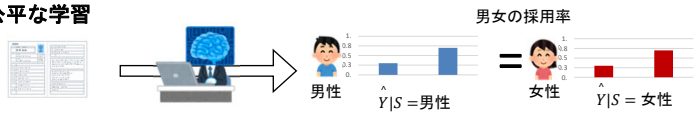
(1) 中間データの生成によるタスクバイアスとドメインバイアスの分離
(2) 敵対的学習によるドメインバイアスの軽減

Source	Pr			Rw			Ar			Cl		
Target	Ar	Cl	Rw	Ar	Cl	Pr	Cl	Pr	Rw	Ar	Pr	Rw
CoCoGAN	57.6	53.4	71.7	69.2	51.3	65.8	62.3	69.5	74.5	66.7	74.0	66.4
Wang2020	**70.3**	60.8	74.8	72.2	61.4	72.2	62.7	71.9	76.3	**72.6**	**75.1**	73.9
DF-ZSDA	64.4	**69.2**	**82.0**	**77.9**	**76.2**	**88.5**	71.0	76.5	**85.1**	62.1	68.7	**75.1**
Ours	67.5	65.1	78.9	74.3	69.0	75.8	**72.1**	**76.7**	83.8	69.8	73.0	71.5

under review

Demographic Parity Constrained Minimax Optimal Regression under Linear Model, NeurIPS2023

公平な学習



最も良い公平な回帰アルゴリズムは何か？

モデル

- Y:出力，X:非センシティブ属性，S:センシティブ属性（性別，人種）

$$Y = \langle \beta_S^*, X \rangle + \xi \qquad X \sim N(\mu_S, \sigma_X^2 I)$$

YのSへの依存：SがYの分散へ影響
XのSへの依存：SがXを通じて間接的にYへ影響

結果

Sが出力の分散へ与える影響の大きさ　（公平でない）通常の線形回帰と同じ

$$\frac{B^2\sigma_\xi^2 dM}{n} \lesssim \mathcal{E}_n \lesssim \frac{B^2\sigma_\xi^2 dM}{n} \vee \frac{\sigma_X^2 B^2 M}{n} \vee \frac{B^2 U^2}{n}$$

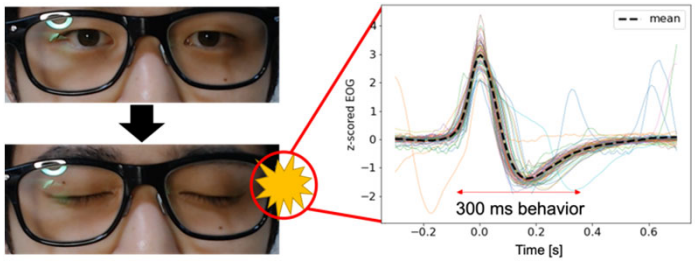
σ_ξ^2:ξの分散, d: Xの次元, M: Sの取りうる値の種類数, U : μ_Sのノルム上界

- Sの分散への影響が回帰の統計的難しさに影響する初めての結果
- 間接差別(SのXを通じたYへの影響)がある状況で初めての最適アルゴリズム

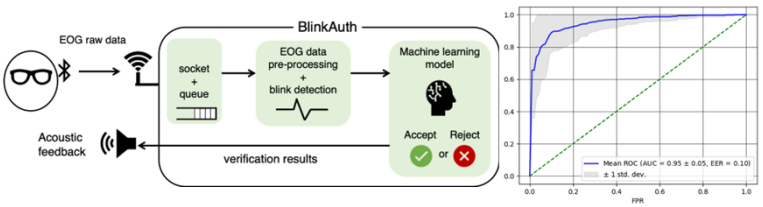
Kazuto Fukuchi and Jun Sakuma. Demographic Parity Constrained Minimax Optimal Regression under Linear Model. NeurIPS2023.

The Catcher in the Eye: Recognizing Users by their Blinks, AsiaCCS 2024

テーマ：人間の瞬きを利用した個人認証方式 BlinkAuthの提案



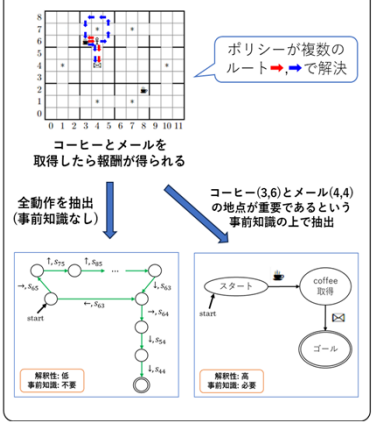
内容：市販のメガネ型のデバイス(JINS MEME)で測定可能な眼電位データに対して教師あり機械学習を適用し、リアルタイムで高精度な個人認証が可能であること、様々な環境（作業中、メイク有無、時間経過、攻撃）に対して頑健な認証が達成できること、ユーザスタディの結果高いユーザビリティを有することを明らかにした。



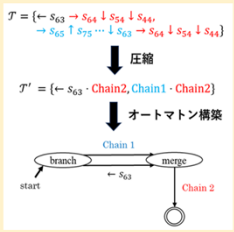
Ryo Iijima, Tetsuya Takehisa, Tetsushi Ohki, and Tatsuya Mori, The Catcher in the Eye: Recognizing Users by their Blinks, ACM ASIACCS 2024

Extracting Small Automata Representing Behavior of Reinforcement Learning Agent with Optimal Compression

強化学習ポリシーの動作を表すオートマトン生成:
事前知識なしで生のトレースを用いると、解釈性が低く、時間もかかる。



提案手法:
実行トレースサンプルから、系列のまとまり(Chain)を圧縮する。



結果:
事前知識なしで、解釈性の高いオートマトンを少ない実行時間で生成可能となった。

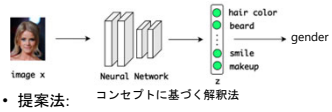
	Size of automata	Generating time (s)
Raw trace	29	1.92156
Compressed trace (proposed)	7	0.01028 (= 0.00522 [compression] + 0.00506 [automata generation])

オートマトンサイズと生成時間の比較。圧縮の時間を加えても全体の実行時間が短い

under review

Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs, IJCAI2023

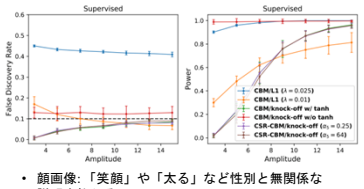
- コンセプトに基づく解釈法における問題点
  - どのコンセプトも多少予測に関係ある
  - 予測に関係のないコンセプトを重要だと取り間違える (偽陽性な解釈)



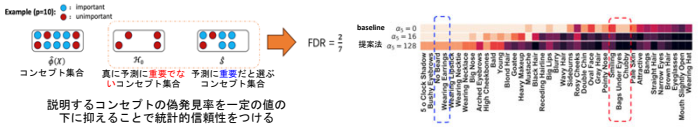
コンセプトに基づく解釈法

- 提案法:
  - 特徴抽出器でスパースな特徴を画像から学習する
  - Knock-offによる「不必要な要因」の偽発見率(False Discovery Rate)を理論的に抑える

- 結果
  - 人工データ: 指定のFDR (0.1)の下にコントロール



- 顔画像：「笑顔」や「太る」など性別と無関係な説明を抑える



説明するコンセプトの偽発見率を一定の値の下に抑えることで統計的信頼性をつける

Kaiwen Xu, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs, IJCAI 2023