FY2022/2022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

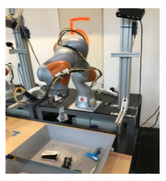
Max-Min Off-Policy Actor-Critic Method Focusing on Worst-Case Robustness to Model Misspecification, NeurIPS2022

Max-Min Off-Policy Actor-Critic Method Focusing on Worst-Case Robustness to Model Misspecification, NeurIPS2022

- 異なる環境へのポリシーの転移
  - 不確実性はパラメータ化されている
  - ただし、転移後の環境が未知

学習時：シミュレータ

Policy 転移



適用時：実環境

- 提案法:最悪ケース環境に頑健な強化学習
  - 獲得報酬の加減を最大化

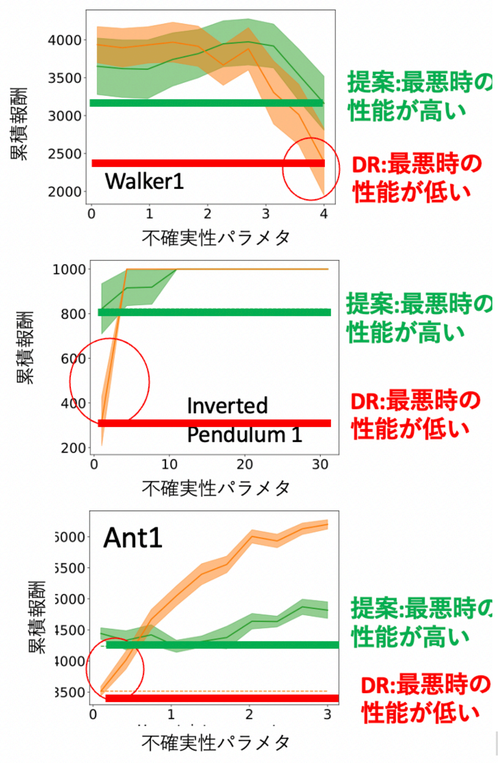
→未知環境転移後でも優れた予測性能

$$\max_{\mu \in \mathrm{M}} J(\mu, \omega^*) \geq \max_{\mu \in \mathrm{M}} \min_{\omega \in \Omega} J(\mu, \omega)$$

転移後　　転移前

転移前
不確実性パラメータに対して
最悪の環境で報酬を最大化

- 結果　Green：M2TD3 (提案法)　Orange：DR (比較手法)



Takumi Tanabe, Kazuto Fukuchi, Jun Sakuma, Yohei Akimoto, Max-Min Off-Policy Actor-Critic Method Focusing on Worst-Case Robustness to Model Misspecification, NeurIPS2022

Domain Generalization via Adversarially Learned Novel Domains, IEEE Access (2022)

Domain Generalization via Adversarially Learned Novel Domains, IEEE Access (2022)

- ドメイン汎化問題
  - 訓練時と別ドメインで予測したい
  - 訓練時にはターゲットドメインが未知

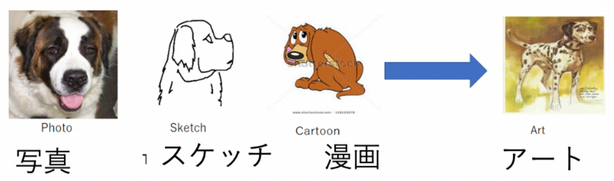


写真　スケッチ　漫画　アート

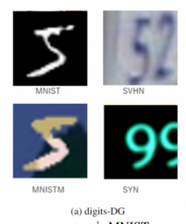
訓練時に利用できるドメインの画像　目標ドメイン

- 提案法
  - 訓練ドメインから多様なドメイン画像を生成
  - 最も訓練が困難なドメインでデータ増強

→未知ドメインでも優れた予測性能

3ドメインからの画像生成例

- 結果
  - 4ドメイン文字画像におけるドメイン汎化でSoTA



(a) digits-DG

	MNIST	SVHN	MNIST-M	SYN	Avg
ERM	96.0 ± 0.2	62.7 ± 0.6	59.6 ± 0.5	78.6 ± 0.5	74.2
CrossGrad [17]	96.7	65.3	61.1	80.2	75.8
JiGen [22]	96.5	63.7	61.4	74.0	73.9
L2A-OT [11]	96.7	68.6	63.9	83.2	78.1
Mix-style [16]	96.5 ± 0.3	64.7 ± 0.7	63.5 ± 0.8	81.2 ± 0.8	76.5
GUD [10]	97.1 ± 0.2	66.7 ± 0.3	63.9 ± 0.3	81.5 ± 0.2	77.3
DDAIG [18]	96.6 ± 0.2	68.6 ± 0.6	64.1 ± 0.4	80.2 ± 0.2	77.6
DLOW [5]	97.6 ± 0.1	68.3 ± 0.5	65.1 ± 0.6	82.2 ± 0.3	78.3
GADA-D	98.1 ± 0.3	**72.2** ± 0.5	**67.1** ± 0.5	**84.9** ± 0.3	**80.6**
GADA-NN	**98.2** ± 0.5	70.2 ± 0.4	66.5 ± 0.5	84.5 ± 0.6	79.9

- 4ドメイン画像分類ではNo. 2 (左上の4ドメイン)

	Art	Cartoon	Photo	Sketch	Avg
ERM	77.6 ± 0.5	76.4 ± 0.3	96.3 ± 0.1	69.8 ± 0.6	80.0
CrossGrad	79.8	76.8	93.6	66.8	79.3
Jigen	79.4	75.3	96.0	71.6	80.6
L2A-OT	83.3	78.2	96.2	73.6	82.8
Mix-style	84.1 ± 0.4	78.8 ± 0.4	96.3±0.3	75.9 ± 0.9	83.8
GUD	79.5 ± 0.4	77.2 ± 0.5	94.9 ± 0.2	71.1 ± 0.3	80.7
DDAIG	84.2 ± 0.3	78.1 ± 0.6	95.3 ± 0.4	74.7 ± 0.8	83.1
DLOW	80.7 ± 0.6	76.1 ± 0.5	94.1 ± 0.3	76.7 ± 0.8	81.9
Wang2021 [19]	81.4	79.6	95.5	80.6	84.3
Yang2021 [20]	**85.8** ±0.6	80.7 ± 0.5	**97.3**±0.3	77.3 ± 0.5	85.3
MBDG [12]	80.6 ± 1.1	79.3 ± 0.2	97.0 ± 0.4	**85.2**±0.2	**85.6**
GADA-D	84.3 ± 0.7	**82.7** ± 1.1	96.2 ± 0.4	76.8 ± 0.8	85.1
GADA-NN	84.3 ± 1.2	81.4 ± 0.8	96.0 ± 0.8	78.6 ± 0.6	85.1

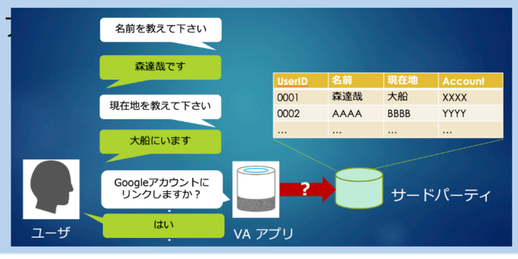
Yu Zhe, Kazuto Fukuchi , Youhei Akimoto , Jun Sakuma: Domain Generalization via Adversarially Learned Novel Domains. IEEE Access 10: 101855-101868 (022)

Understanding the Behavior Transparency of Voice Assistant Applications Using the ChatterBox Framework. RAID 2022

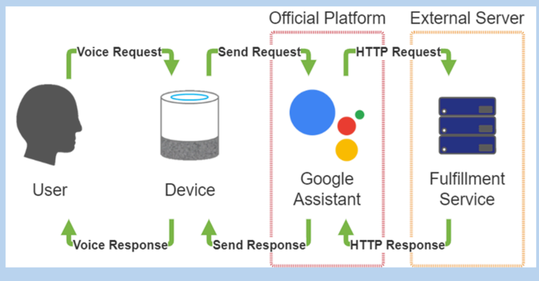
RQ: 音声アシスタント(VA)はどの程度個人情報を収集しているか？

VAアプリに固有な課題

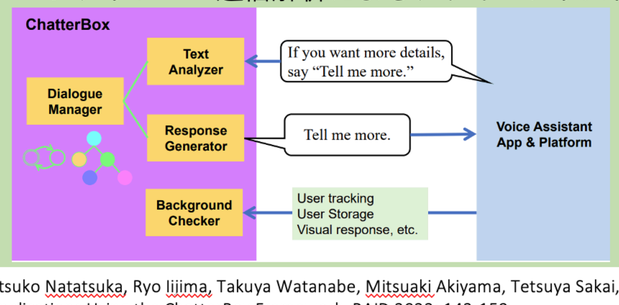
ダウンロード実行できない(クラウドのみ)
　ローカル実行できない
→対話を通じてしか機能が明らかにできない



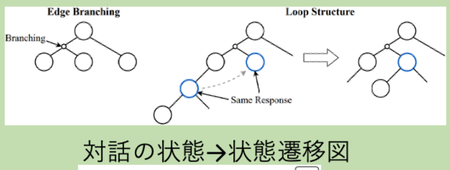
クラウド（見えないところ）で実行される
→ユーザに対する**透明性に欠ける**



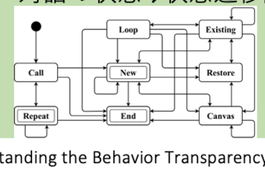
- **アプローチ**: 自然言語処理を利用した対話生成・解析＋アプリレベル通信解析によるプライバシリスク評価



過去の対話履歴→ツリー構造データ



対話の状態→状態遷移図



Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, Tatsuya Mori: Understanding the Behavior Transparency of Voice Assistant Applications Using the ChatterBox Framework. RAID 2022: 143-159

- 結果 (対象VAアプリ EN: 9,732, JP: 931)

対話で個人情報を収集するVAアプリ

個人情報	アプリ数（英語）	アプリ数（日本語）
名前	9	2
メールアドレス	6	0
電話番号	7	0
居住住所	3	12
現在位置	0	3
年齢	2	3
性別	0	1
血液型	1	0
誕生日	1	8
勤務地	1	1
通勤経路（駅）	0	2
いずれか1つ以上当てはまる	22 (3.0%)	28 (6.0%)

全VAアプリの3-6% は対話を通じて何らかの個人情報を収集

ユーザストレージを利用するVAアプリ

Data	#apps (EN)	#apps (JA)
ユーザID	133 (18.2%)	42 (8.99%)
最終利用時刻	4 (0.546%)	7 (1.50%)
アプリ利用回数	6 (0.820%)	10 (2.14%)
ユーザの現在位置	1 (0.137%)	0 (0%)
その他	59 (8.06%)	45 (9.64%)
上記のいずれか1つ	160 (21.9%)	79 (16.9%)

これらのVAアプリは、ユーザに対して秘密裏に情報収集が可能（潜在リスク）

プライバシポリシ分析

Class	#apps (EN)	#apps (JA)
妥当	324 (44.3%)	18 (3.85%)
不足あり	367 (50.1%)	431 (92.3%)
掲示なし	41 (5.60%)	18 (3.85%)

全VAアプリの4-6%はプライバシポリシなし
プライバシポリシを持つVAアプリの76-94%は記述不十分

■研究の貢献

VAアプリの解析フレームワークを提案

自然言語処理による対話＋アプリレベル通信解析
日本語＋英語で動作することを実証
自然言語を用いた対話インタフェースを持つシステムのテストに適用可能
※チューニングなしのGPT-3では対話は成立せず

VAアプリの実態を明らかにした

スマートフォンと比較して、個人情報の扱いがまだ発展途上
適切なユーザIFの開発が必要であることを示唆

Team Achievement

2022年度の主な業績

- 査読付き国際会議論文･英文ジャーナル論文合わせて24編（重要な業績のみ）
  - 機械学習･人工知能に関する論文20編
  - セキュリティ･プライバシーに関する論文4編
  - うちトップ国際会議/ジャーナルはNeurIPS'21 (1編, core rank A*), AAAI'22 (1編, core rank A*),NDSS'22 (1編, core rank A*), Applied Soft Computing (1編, IF 8.263) の4編
- 主な受賞は以下の四件
  - セキュリティ分野の中堅国際会議であるECML/PKDD2022にて機械学習公平性に関する論文でTest of Time Award
  - 日本統計学会にてデータサイエンスに関する著作で日本統計学会出版賞
  - 情報論的学習理論と機械学習 (IBISML) 研究会にて研究会賞
  - コンピュータセキュリティシンポジウムにおいて優秀論文賞(1件)および学生論文賞(2件)