＜年次報告書＞
第 1 年度（2025 年度）
日 ASEAN 科学技術・イノベーション協働連携事業（NEXUS）
日本－シンガポール「AI」研究領域

研究課題名（和文）	マルチモーダル対応の切断：より安全で公平なマルチモーダル AIGC の構築
研究課題名（英文）	Breaking Multimodal Correspondence: Crafting Safer and Fairer Multimodal AIGC
研究期間	2025 年 4 月 1 日～2028 年 3 月 31 日

氏名	所属 部署　役職
佐藤 真一	国立情報学研究所 コンテンツ科学研究系　教授

1. 研究進捗状況と成果

まず、画像識別器の公平性の解析並びに不公平さ(バイアス)の解消について取り組んだ。より具体的には、顔画像に基づく性別識別の画像解析モデルが広く使われているが、これらが人種や年齢により著しく性能が落ちることを指摘した。特に Kaggle コンペで広く公開されているモデルは、白人には高性能を示すが、アジア人では著しく性能が落ちることを示した。この問題は、教師ありファインチューニングで解決することをまず示し、かつ Pseudo Labeling による教師なしファインチューニングでもある程度解消できることを示した。さらに、教師なしファインチューニングで用いる顔画像自体にバイアスがある場合に対応するため、新たに Pseudo Balancing という手法を考案し、その効果を確かめた。
次いで、ターゲットの識別モデルが外部 API 等でしか公開されておらず、わずかな実験しかできず、かつ公平性検証のためのデータについては、データ流出を防ぐため外部 API に投入することができない場合に、効果的に当該モデルの公平性を検証する手法について検討した。基本アイデアは、まず検証用のデータを用いて様々なバイアスを有するリファレンスモデルを多数用意する。その上で、評価用データ(検証用データとは別)を構築したすべてのリファレンスモデルとターゲットモデルとに投入し、その振る舞いを比較し、最も振る舞いの似ているリファレンスモデルのバイアスをターゲットモデルの推定バイアスとするというものである。検証の結果、数十程度のわずかな評価データのみで十分な推定が可能であることが実証できている。

2. 代表的な成果事例

1. Ana Manzano Rodriguez, Camille Guinaudeau, and Shin'ichi Satoh, ``Uncovering Gender Biases in Gender Identification Models for Japanese Data Analysis,'' Workshop on Demographic Diversity in Computer Vision, in conjunction with CVPR, Nashville, US, June 11 (2025).
   概要:顔画像に基づく性別識別の画像解析モデルが人種や年齢により著しく性能が落ちることを指摘し、教師ありファインチューニングで解決すること、また Pseudo Labeling による教師なしファインチューニングでもある程度解消できることを示した。
2. Haohua Dong, Ana Manzano Rodriguez, Camille Guinaudeau, and Shin'ichi Satoh, ``Fairness Without Labels: Pseudo-Balancing for Bias Mitigation in Face Gender Classification,'' Second workshop on Fairness and ethics towards transparent AI: facing the challEnge through model Debiasing (FAILED), in conjunction with ICCV, Honolulu, Hawaii, USA, Oct 20 morning, (2025).
   概要:顔画像に基づく性別識別の画像解析モデルが人種や年齢により著しく性能が落ちるという不公平さの解消において、教師なしファインチューニングで用いる顔画像自体にバイアスがある場合に対応するため、新たに Pseudo Balancing という手法を考案し、その効果を確かめた
3. Zhaohui Zhu, Marc A. Kastner, and Shin'ichi Satoh, ``Fairness Auditing of Limited-Access Models Using Protected Datasets,'' IEEE Access, Vol.13, pp.208511-208532, (2025).
   概要:ターゲットの識別モデルが外部 API 等でしか公開されておらず、わずかな実験しかできず、かつ公平性検証のためのデータについては、データ流出を防ぐため外部 API に投入することができないという、厳しいアクセス制限がある場合でも、数十程度のわずかな評価データのみで十分な推定が可能であることを示した。