G-3

XAI を用いた顔認識の公平性に関する検討

A Study on the Fairness of Face Recognition Using XAI

○陳澤舟[1]，関弘翔[2]，細野裕行[2]
*Zezhou Chen[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: In this study, we train models using racially balanced and unbalanced datasets separately and test the accuracy of gender detection for each race. we validate that the error rates on non-white people are usually much higher than white people and unbalanced dataset can amplify the bias. Furthermore, we try to identify the causes of bias by visualizing the classification basis using XAI.

１．まえがき

ディープラーニング技術の発展により，顔認識技術は世界中で活用されている．しかし，顔認識にはまだ懸念があり，様々な技術的及び倫理的な問題点が指摘されている．特に注目されるのは，顔認識システムに潜在している性別や人種に関するバイアスである[1].

2018 年，Buolamwini らの研究[2]では，当時既存の多くの顔データセットでは，黒人，特に黒人女性の割合が非常に低いと指摘した．その上，当時市販の性別分類システムを評価した結果として，黒人女性の誤分類率が一番高いことが報告されている．2019 年，Wang らの研究[3]では，バランスが取れたデータセットで学習させたモデルは非白人に対して良いパフォーマンスを示すことが指摘され，データセットにある人種の偏りが分類精度に反映されることが証明された．しかし，たとえ良いバランスのデータセットでの学習でも，非白人は白人に比べて精度が低いことも指摘された.

本研究では，人種バイアス問題を軽減する性別分類システムの構築を目的とする．そのため，データセットの人種と性別のバランスを調整しながら，各人種の精度を確認し，異なるモデルの精度を比較する．さらに，Explainable AI（XAI）で異なるモデルの分類根拠を可視化することでバイアスが生じる原因を検討する.

２．異なるモデルとデータセットの精度比較

既存の多くの顔画像データセットは，白人に強く偏っており，他の人種は著しく少ない．このようなデータセットで学習させたモデルは常に白人に対する精度が高く，バイアスが生じる[3]．本研究では，人種の偏りが少ないデータセット Fairface[4]を使用している．Fairface は，画像に人種，性別，年齢などのラベルが付いており，人種を黒人，東アジア人，インド人，ラテンアメリカ人，中東人，東南アジア人，白人の 7 つのグループに分類している．その上，各人種，性別，年齢層の画像枚数がバランスよく配置されている．人種ごとに男女それぞれおよそ 6000 枚の学習データとおよそ 800 枚の検証データがあり，合計 9 万枚以上の画像が含まれている.

性別分類モデルについて，ResNet34[5]と InsightFace[6]を採用した.

ResNet34 は 33 層の畳み込み層と残差接続，および 1 層の全結合層により構成される CNN である.

InsightFace は 2D・3D 顔解析のためのライブラリであり，顔認識，顔検出などの豊かなアルゴリズムが実装されている．本研究では，InsightFace の Gender＆Age という顔を検出して性別や年齢を予測する学習済みモデルを使用した.

本研究では，まず，事前学習済み ResNet34 を用いて転移学習を行った．次に，学習したモデルを用いて別々に ResNet34 と InsightFace でテストする．テストデータはカテゴリごとに 500 枚，合計 7000 枚である．それぞれの精度は Table 1.に示す.

Table 1. Accuracy of each category on ResNet34 model and InsightFace model

	ResNet34		InsightFace	
	Male	Female	Male	Female
Black	90.00%	87.80%	63.80%	78.00%
East Asian	92.40%	94.60%	53.40%	88.60%
Indian	95.60%	93.80%	74.00%	80.40%
Latino_ Hispanic	93.80%	93.80%	71.40%	84.60%
Middle Eastern	96.20%	93.00%	82.20%	80.80%
Southeast Asian	91.60%	91.40%	60.40%	82.60%
White	95.00%	93.80%	72.20%	88.00%

１：日大理工・院（前）・情報　２：日大理工・教員・情報

Table 1 から見ると，ResNet34 モデルの精度は全体的に Insightface モデルより高い．Insightface モデルでは，平均的に男性の方が女性よりも精度が低く，アジア人モデルに改善の余地がある．男性の精度が特に低い．ResNet34 モデルでは，全てのカテゴリの精度の差があまり大きくないが，黒人の精度はやや低い．

そして，特定の人種の学習データを減らして学習させて精度比較を行った．例として，黒人の学習画像を 25%に減らした ResNet34 モデルの精度は Table 2 に示す．Table 1.と比べると，黒人，特に女性の精度が低下した．これは，データセットの偏りは人種バイアスを広げることを示している．

Table 2. Accuracy of each category on ResNet34 model when training data of blacks reduced to 25%

	Male	Female
Black	88.60%	84.00%
East Asian	91.60%	95.20%
Indian	94.80%	94.20%
Latino_Hispanic	93.60%	93.00%
Middle Eastern	94.40%	93.20%
Southeast Asian	92.00%	92.20%
White	93.80%	94.20%

３．XAI を用いた分類根拠の可視化

XAI は，説明可能な AI を指す．本研究では，精度を高め，人種バイアスを減らす方法を探索するために，SHAP[7]という XAI を用いて分類根拠の可視化をした．

SHAP は元のモデルと同等の予測結果が得られる説明可能な代替モデルを構築し，モデルの予測結果に対する各特徴量の寄与を求める XAI である．本研究では，SHAP の GradientExplainer を用いてバランスが取れたデータセットと取れていないデータセットで学習した ResNet34 モデルの分類根拠の説明をそれぞれ行った．Figure 1.に説明例を示す．赤色が濃いほど正に大きく寄与し，青色が濃いほど負に大きく寄与することとなる．結果から見ると，ResNet34 モデルの精度は高いが，学習している特徴が非常にあいまいで，バイアスが生じる原因は把握できず，モデルに改善の余地がある．

４．まとめ

人種バイアスが低減する性別分類システムを構築するために，本研究では，データセットの人種偏りの有無が性別分類精度に与える影響について，ResNet34 モデルと Insightface の性別分類モデルで実験して精度比較を行った．さらに，人種バイアスの原因を探るために，SHAP を用いて分類根拠の可視化を行った．しかし，現段階での可視化の結果は，バイアスを引き起こす特徴を一般化するにはまだ不十分である．今後はより明確な分類根拠を示せるモデルの構築を検討する予定である．

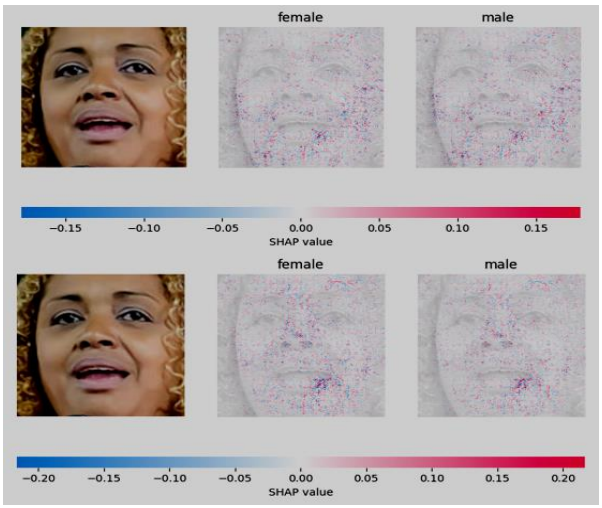

Figure 1. Example of visualizing the classification bais of different ResNet34 models with SHAP(top: trained with balanced dataset, bottom: trained with unbalanced dataset)

参考文献

[1] Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 145-151).

[2] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

[3] Wang, M., & Deng, W. (2020). Mitigating bias in face recognition using skewness-aware reinforcement learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9322-9331).

[4] Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.