G-15

深層距離学習を用いた顔認識の公平性に関する検討

A Study on Fairness of Face Recognition Using Deep Metric Learning

○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, to enhance fairness, we compared the effects of uniformizing data for all races and reweighting based on the quantity of data for each race. We also introduced deep metric learning into the model. We assessed fairness of models by existing fairness index and visualizing the feature space through t-SNE.

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，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．さらに，たとえ良いバランスのデータセットでの学習でも，白人以外の人種は白人に比べて精度が低いことも指摘された.

よって本研究では，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．

本稿では，公平性を向上させるため，まずデータセットのアンバランス解消を検討する．さらに深層距離学習をモデルに導入することを検討し，それらの評価を行う．モデルの公平性評価には，公平性の指標および次元削減アルゴリズムで特徴空間の可視化を用いる．

２．公平性評価

公平性を評価する指標として，Equal Opportunity Difference(EOD)[4]を使用した．式(1)に EOD の定義を示す.

$$EOD = \left|P\left(Y=1\middle|\hat{Y}=1,S=1\right) - P\left(Y=1\middle|\hat{Y}=1,S=0\right)\right| \quad (1)$$

Y をラベルの確率変数，$\hat{Y}$ を分類器の出力ラベル，S をセンシティブ属性の確率変数とする．つまり，センシティブ属性における真陽性率の差の絶対値である．0 に近いほど公平性が高いといえる．ラベルとして，0 を男性，1 を女性とする．また，人種をセンシティブ属性とし，0 を黒人以外の人種，1 を黒人とする.

３．データセットにある不公平の解消

本研究では，既存の多人種データセット FairFace[5]から黒人，東アジア人，インド人，白人という 4 つの人種のデータを抽出した．各人種のデータ量は大体においてバランスが取られているが，最大 2605 枚の差異がある．これらの差異がもたらす不公平を可能な限り解消するため，2 つの方法を採用して比較した．1 つ目は，各人種から男性と女性それぞれ 5000 枚の画像を取り，合計 40000 枚の画像でモデルを学習させる方法である．2 つ目は，データをモデルに入力する際，それが所属する人種の総データ量の逆数を重みとして付け，モデルにより観察される機会を変えた合計 49361 枚の画像で学習させる方法である．性別分類のモデルとして，Vision Transformer(ViT)モデル[6]を使用した．ViT は Transformer が単語を扱うように，入力画像から分割されたパッチをベクトル化して学習する．Table 1 に元の ViT モデルの精度および EOD を示す．Table 2 に 2 つの方法で学習したモデルの精度および EOD を示す.

精度をみると，女性に対する分類精度は，単純にデータの枚数を均一にする場合大幅に低下した．これはデータ量の減少によるものだと考えられる．また，EOD を見ると，2 つの手法は元のモデルより公平であり，ほぼ同じ公平性を持っていることが分かる.

Table 1. Evaluation of accuracy and EOD on the original ViT model

	Female	Male
Black	0.834	0.856
East Asian	0.944	0.892
Indian	0.948	0.926
White	0.944	0.914
EOD	0.111	

１：日大理工・院（前）・情報　２：日大理工・教員・情報

Table 2. Evaluation of accuracy and EOD on ViT models, after uniformizing data for all races or reweighting based on the quantity of data for each race

	Uniformizing		Reweighting	
	Female	Male	Female	Male
Black	0.774	0.892	0.854	0.818
East Asian	0.876	0.920	0.946	0.856
Indian	0.858	0.942	0.946	0.914
White	0.880	0.938	0.956	0.872
EOD	0.0973		0.0953	

４．深層距離学習の導入

公平性を改善するための手法として，ArcFace[7]を導入した．ArcFace とは，同一クラス内の特徴ベクトルの距離を小さくし，異なるクラス間の距離を大きくする深層距離学習手法である．再重み付けをしたうえで，ViT モデルの出力層を ArcFace に入れ替えて学習した．Table 3 に ArcFace を適用した ViT モデルの精度を示す．

結果を Table 2 の再重み付けのモデルと比較してみると，ArcFace を適用したモデルは，ほぼすべてのカテゴリ，特に黒人の分類精度が向上した．さらに，EOD も 0.0313 低下したので，公平性も改善されたことがわかる．

Table 3. Evaluation of accuracy and EOD on the ViT+ArcFace model trained with reweighting

	Female	Male
Black	0.882	0.892
East Asian	0.946	0.908
Indian	0.948	0.946
White	0.944	0.918
EOD	0.0640	

５．特徴空間の可視化

本研究では，t-SNE[8]を用いてモデルの特徴空間を可視化することで公平性を分析する．t-SNE は高次元データを 2 次元に落とし込める次元削減アルゴリズムである．Fig. 1 に再重み付けで学習した ViT モデルと ViT+ArcFace モデルの特徴空間を示す．グラフをみると，少数のインド人女性を除き，特定の人種のかたまりはほとんど見られない．これは，モデルが人種に関して基本的に公平であることを示している．

６．まとめ

本研究では，データセットのアンバランスによる不公平を解消する，人種ごとの枚数の均一化とデータ量による再重み付けの２つの手法を検討した．結果から，再重み付けの精度はデータ量の均一化より優れていた．また，深層距離学習の ArcFace を導入することで，モデルの精度および公平性が向上することが明らかとなった．t-SNE による特徴空間の可視化で，モデルが人種に関して公平であることを検証した．

今後の課題として，さらに公平性を向上させつつ精度を高めることがあげられる．

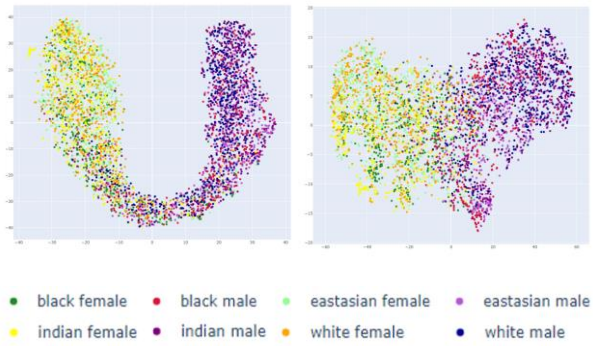

Figure 1. the feature space of the ViT model (left) and the ViT+ArcFace model (right) trained with reweighting visualized by t-SNE

参考文献

[1] Raji, I. D., et al. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conf. 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 Conf. 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] 綿岡晃輝, 他. (2020). 公平性により生じる敵対的攻撃に対する脆弱性. In 人工知能学会全国大会論文集 第 34 回 (2020) (p. 4N2OS26a01). 一般社団法人 人工知能学会.

[5] 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).

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[7] Deng, J., Guo, J., et al. (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).

[8] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).