顔認識における人種バイアス軽減のためのデータセット・学習モデルの考察と改善法の検討

論文番号 M11
テクノロジーデザイン講座
上西研究室 松本 大樹

人種バイアスとは、顔認識アルゴリズムがもたらすバイアスで、白人と比べ非白人における認識性能が低下する問題である。よって、この問題を回避するための方法を探索することは、社会科学だけでなく、より良い利便性を持った社会を構築するために必要な非常に重要な課題である。そのため、顔認識における人種バイアスを低減するために、さまざまな研究がなされている。既存研究の多くは，偏りの主要な原因であるデータセットに直接言及するのではなく，学習中にモデルに現れる偏りを是正しようとするものである。例外として、Mei らは、人種間で等しいデータセットを作っており、従来の人種間で偏りのあるデータセットと比較し、認識性能の向上を達成した。そのため、人種バイアスに対抗する方法として、人種間で等しいデータセットを作ることが重要だという仮説を示している。しかし、人種間で等しいデータセットを用いた学習でも、全人種で完全にバランスのとれた性能をあげることはできていない。

本研究では、この要因に深く言及するために、学習モデルおよび学習データセットと人種バイアスの関連性を理解することを目的とする。そのため、画像認識分野で広く用いられている Resnet、Densenet、VGG をベースとしたネットワークを用いて学習モデルを作り、認識結果を考察することで上記問題に言及する。学習データセットに関しては、K-means、t-SNE を用いた分析を行うことで、人種間で等しいデータセットが持つ問題点を分析し、Mei らの仮説が必ずしも成り立たないことを示す。学習モデルに関しては、Attention Map を用いて学習モデルの注目領域を可視化することで、人種バイアスと学習モデルの関連性を考察する。また、人種バイアスを改善するための 1 つの手法として Data Augmentation に着目し、人種バイアスへの影響を考察する。Data Augmentation には、Random Erasing、Cutout を応用した 3 つの手法を学習データセットに適用し、学習モデルごとに結果を比較する。ここでも、学習モデルの挙動を分析するために Attention Map を用いて注目領域を可視化し考察する。そして、Data Augmentation を適用していない同じネットワーク同士を比較することで、本手法が人種バイアス改善にどのような影響を与えるか考察した。

実験の結果、学習モデルおよび学習データセットと人種バイアスの関連性を理解することができた。学習モデルに関しては、モデルごとに認識性能や注目領域が大きく異なることが示された。学習データセットに関しては、Mei らの仮説が必ずしも成り立たず、クラスタに着目したデータセット構成の重要性が示された。また、本実験の Data Augmentation が学習モデルの認識性能を向上させ、人種バイアス改善に効果を及ぼす可能性が示された。本研究では、学習済みモデルを用いた転移学習および Fine-tuning によって実験・考察している。そのため、実データとしては 1 万枚ほどの画像しか用意していない。実社会の顔認識では、顔画像のデータ収集にかかるコストが高いことから、少ない画像枚数で高い認識精度を上げる仕組みが求められる。したがって、本研究で得られる知見は、人種や個人の認識など、さまざまな場面で応用することができる可能性がある。