國立中央大學通識教育中心
National Central University General Education Center

「人工智慧跨域應用」學分學程

「平權之路：數位時代的社會風險與治理」課群

人臉辨識的性別偏差與修正：從演算法偏見到公眾參與的多元資料庫

國立中央大學　大氣三 王冠程 / 大氣三 陳棋恩 / 中文四 金秀珍 / 太空二 王瑀釩

作品連結

專案簡介

本計畫旨在探討人臉辨識系統中的性別偏差問題，並提出具體的技術與社會解方。我們首先利用機器學習模型，實作「偏差模型」與「公正模型」，以模擬並驗證資料集性別比例失衡對AI辨識能力的影響 。針對實驗證實的偏誤，我們提出建構一個「公眾參與的多元人臉資料庫網站」，期望透過群眾協力的方式收集不同性別與族群的臉部特徵，從源頭解決訓練資料單一化的問題，以提升AI的泛用性與公平性。

而我們也製作了一個網站，讓對我們主題感興趣的人們能夠明白我們為甚麼會想開啟這個計畫，我們是怎麼證實這個問題存在，並且決定用架設人臉資料庫的方式來解決目前技術限制的窘境。透過視覺化的呈現方式，讓大家能意識到人工智慧並非全能。AI仍需要人性作為其韁繩，避免在科技進步的過程中忽略了弱勢群體的權益。

網站封面

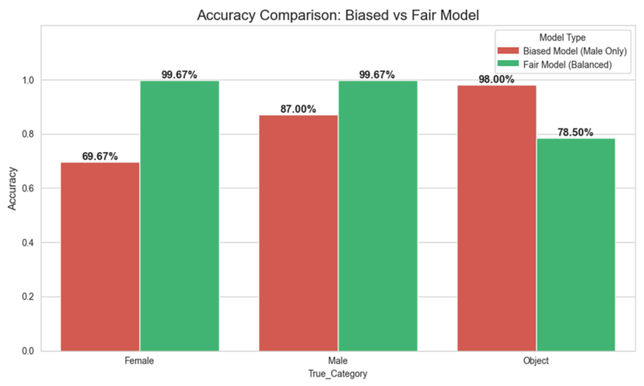
近年來，深度學習在人臉分析領域有了顯著進展，但同時也引發了對「公平」與「偏差」的關注。許多研究指出臉部偵測與辨識模型在不同族群、性別或膚色之間可能存在性能差異，導致某些群體受到不公平的影響。Mittal 等人(2022) 在 Are Face Detection Models Biased? 文中指出，現有的人臉偵測模型在性別與膚色等屬性上存在顯著差異，並且這些偏差往往源自於訓練資料分布不均或缺乏多樣性。我們期望能透過親自驗證，將偏差訓練集對臉部偵測模型的影響以統計的方式可視化，並規劃方法來解決資料集不公平的問題。

研究成果

為了探討訓練資料的公正性是否會對人臉偵測模型造成顯著影響，我們基於ResNet18架構設計了兩個人臉偵測模型，並以男性臉部、女性臉部和各種物件的圖片作為訓練:

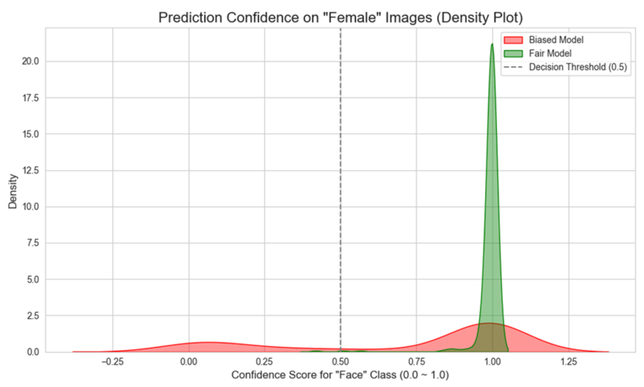
1. 公正(Fair)模型:使用等量的男、女與物件圖片進行訓練
2. 偏誤(Biased)模型:僅使用等量的男性與物件圖片進行訓練

實驗結果顯示，偏誤模型在對女性臉部的辨識準確率為69.67%，對男性為87%；而公正模型則在男女人臉偵測的準確率均達99.67%。然而，公正模型在物件辨識的準確率僅為78.5%。這表明，公正模型在提升人臉偵測的泛化能力上優於偏誤模型，但同時也容易將物件誤識別為人臉。



模型偵測準確率比較圖

接著我們分析兩個模型在女性圖片上的人臉偵測信心分數。由核密度估計圖可觀察到，公正模型在面對女性圖片時呈現較高的機率密度，顯示其對女性人臉的判定具有更高的信心。



模型對女性圖片的人臉偵測信心分數之核密度估計圖

心得與反思

科技並不總是客觀的，它反映了我們的社會。我們常以為AI是絕對理性的，但實驗證明，如果教它學習的資料本身就充滿資料偏誤(例如男生的照片較多)，它的認知就會出現偏見。演算法的偏見，其實就是將人類社會不平等的數位放大。

而我們希望從「被動接受」到「主動改變」。希望透過我們架設的網站，每個人都可以上傳自己的照片來修正AI的認知。這不只是一個收集資料的平台，更是一種「科技正義」的實踐--讓我們不再只是冷冰冰的數據，而是能主動參與、一起修正科技偏見的推手。

透過「性別、科技與社會」課程接觸到AI偏誤相關議題後，我們對於科技背後潛藏著對不同膚色、不同性別的不公感到驚訝。為了解決偏差的問題，我們決定貢獻我們的一份力量，架設一個網站提供大眾上傳臉部資料，以收集多元樣本。在專題製作的過程中，我們除了學會訓練ResNet18來驗證假設，更在架設「公眾參與的多元人臉資料庫網站」時，學到了如何將平權的概念實體化。這次的專題讓我們知道，即便我們只是學生，我們也有能力實踐「科技正義」，成為修正AI偏誤的關鍵一員。

課程名稱：性別、科技與社會　授課教師：姜貞吟老師、楊燕枝老師
特別感謝：教育學習網STEAM 張宗彥創辦人、劉百耕助教、林樺霖助教、王儀涵助教

教育部數位人文跨域人才智慧領航計畫
Digital Humanities Smart Leadership for Interdisciplinary Talent Program