Brandon Polk
CSC 578
Final Survey Paper

Introduction/Overview

Humans are diverse creatures. Whether it's the way we look, the way we talk, or how we see the world both figuratively and literally, we all are different, unique. Of course, there are strong through-lines connecting us. We are still human after all. But our diversity is one of those through-lines, and how we as humans and societies have understood it, identified with it, related to it, and acted upon it in the past and how we choose to do so in the present and the future has an impact on every aspect of our lives. Machine learning, artificial intelligence, neural networks, and the like are no exception. As the prevalence and uses of these data methods broaden, it becomes increasingly important to understand how they interact with the diversity of our world and make intentional efforts to mitigate their potential to exacerbate existing problems or even create new harm regarding demographic and non-demographic biases.

The primary focus of this survey is demographic bias pertaining to facial recognition algorithms. The first article surveyed addresses some of the harms due to biased facial recognition. The second article evaluates a method to decrease bias in facial recognition. Facial recognition is often the most recognized and discussed form of bias in machine learning applications, however it is important to note that it is not the only area where bias appears. Issues with demographic bias affects natural language processing applications, recommender systems, healthcare applications, and more. All these areas have their own unique challenges in terms of dealing with bias, suggesting that this is an area of research requiring a lot of attention, especially given its potential impact. For example, in NLP, biases inherent in a training corpus tends to find its way into a model's word embeddings. This may lead to several issues including gender bias, biases for certain accents or dialects, political bias, and many forms of stereotyping in real-world applications like movie sentiment analysis or messaging applications (Sweeney). How issues of bias apply to various machine learning applications may be different in some respects, but the real-world effects can cause harm all the same.

Description/Summary of Articles (at least 2)

The Harms of Demographic Bias in Deep Face Recognition Research

The first research article speaks to the varying reliability and predictive ability of several popular facial recognition algorithms related to certain physical features. The authors take note of how this kind of bias is still persistent in present day deep face representation models and how there is a current lack of research and analysis focusing on the real-world implications (Garcia). They suggest that this lack of references is evidence of a larger problem of awareness of this issue within the field. Their work takes on two tasks. The first is to demonstrate the existence of demographical bias in several of the most prevalent recognition models, and the second is to address a real-world consequence of this kind of bias, specifically forgery attacks that target automated border control systems. By doing this, the authors hope to inspire more research and development teams to improve benchmark datasets and procedures to mitigate the issue bias and improve the overall quality and usability of facial recognition systems.

The paper considers 3 deep learning facial representation models, VGGFace, DLib, and FaceNet. Each of these models are recognized for attaining high accuracy results on the LFW benchmark dataset. They are all different convolutional neural network learning approaches. The performance of these models is compared on multiple benchmark facial recognition datasets, LFW, Multi-PIE, UTKFace, UniTexas, and ColorFeret. To measure the ability of these models to distinguish faces within different demographic groups, the average euclidean distance between the face embeddings of the different samples within that demographic is used. The lower this average distance, the more similar the model presumes the individual faces within that group to be. A low average distance means that the model has a difficult time discerning variation in the facial features of a particular group of people and is therefore susceptible to bias and inaccurate identification of these faces. Additionally, a linear SVM classifier is trained and tested using 5-fold cross-validation on the embeddings produced by each model-dataset combination. This was done to measure how well each model can recognize traits of different ethnicities. A higher accuracy indicates that the model can sufficiently pick up traits distinguishing different ethnicities.

In their evaluation of the different face representation models, they discovered that while the different models generally performed well recognizing different ethnicities, they failed to distinguish between faces of the same ethnicity equally between the different groups. Regarding identifying ethnicity, DLib and FaceNet models were highly accurate while VGG-Face had less success. Across the board, all the models performed significantly better using datasets with standardized images (even lighting, neutral expression, neutral pose). Regarding a model's ability to differentiate between faces within an ethnic group, each model performed significantly better on male, White/European faces than other groups. While the performance varied across different models and ethnic groups, performance on Asian, particularly female faces performed the worse. These variations in performance by ethnic group indicates the presence of significant bias.

After demonstrating the bias present in the facial representation models, the authors then discuss why this bias is a problem in practical applications using automated border patrol as an example. To demonstrate this vulnerability, a morphing attack on face recognition models is simulated. A morphing attack involves a forged travel document that mixes the face images of two individuals (an accomplice and an imposter) with a morphing technique. The document is then used by the imposter to bypass automatic face verification by tricking the system into positively verifying the imposter as both the accomplice and as themself. The results of their simulation demonstrated that morphing attacks were significantly more successful when the forged images were of Asian or Indian ethnicity. Asian women were significantly more effective in this kind of attack than other groups.

With the evidence of their study in hand, the authors conclude that demographic bias in current face representation models not only exists but have real consequences in their application. These issues if not dealt with, may raise significant problems including security and civil rights. To address this bias problem, the authors suggest the development and use of more representative datasets and demographic-based evaluation measurements. They also emphasize the need for better data gathering methods and the adoption for stronger face recognition benchmarks.

Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer

The first article surveyed, established the existence of bias in commonly used facial representation models and datasets and the potential harms of this bias. This article evaluates a new

method to mitigate that bias. The approach that these researchers take addresses the lack of diversity in publicly available by implementing a data augmentation method to improve training image diversity. With greater diversity in the training data, much of the demographic bias in the resulting models may be avoided. The paper breaks down its method for data augmentation and image mixing, evaluates their method's ability to improve dataset diversity, and then evaluates how bias in datasets may affect model performance.

The method developed in this paper to increase dataset diversity is a generating adversarial network (GAN) that can apply multiple demographic feature transformations to a facial image simultaneously. What makes their method different from other models that also attempt to apply multiple demographic transformations to augment image data, however, is that instead of using categorical labels, it merges attribute-specific representations of desired demographic changes (Georgopoulos). What this means is that instead of applying a rigid representation of a given category, this method has the flexibility to create intra-class variants, so its class representations adapt based on the face images that they are extracted from. The process allows for face images to have age, gender, and skin tone modified simultaneously in a photorealistic way and for the synthesis of training image data that is more demographically representative than the original training set.

In order validate their method, it was applied to several common annotated image datasets and its results were compared to several other multi-attribute image translation methods as a benchmark. The databases used, MORPH, CACD, KANFace, and UTKFace each have some level of significant deficiency in representation of age, gender, or race/ethnicity. A qualitative examination of the results of the transformation methods revealed that the proposed method succeeded in creating clear, photorealistic transformations of generally superior quality to the other methods. Several methods, AttGAN, IPCGAN, and CAAE produced images that were barely distinguishable from each other, while StarGAN produced images that were not photorealistic (Georgopoulos). A quantitative evaluation of the method involved measuring the increase in the diversity of a test set of face images using diversity metrics. The metrics used are the Shannon H (ShH) and Simpson D (SiD) diversity indices and the Shannon E (ShE) and Simpson E (SiE) evenness indices. The diversity metrics of the augmented image sets are compared along with those of the original test images. The model proposed in this paper achieved the high score in every metric.

With the ability of the proposed model to improve the diversity of datasets established, the authors next explore how the diversity or lack thereof in a dataset effects the performance of recognition models. To do this a gender recognition model was trained on the MORPH and KANFace datasets and the model performance measured using the equality of opportunity metric, defined by the difference in true positive rate between subpopulations in the data and true positive rate. The results of the model used on the non-augmented image set serves as a baseline. A gender recognition model is also trained on augmented image sets created from StarGAN, AttGAN, and the method proposed in the paper. The results are all compared, and the proposed method resulted in considerable improvement over all other models in the classification of young males (a notably underrepresented group within the original training set). The efficacy of the proposed model was further evidenced by the comparing its performance to several other de-biasing methods. Equality of opportunity was again used as the measurement of fairness in the model. Here the proposed method produced competitive results achieving the best or near-best results in evaluations of the age and skin tone attributes.

The data augmentation method demonstrated in this paper showed considerable success synthesizing image data to diversify training data. It improved on many popular methods used for this task, and data augmented using this method was demonstrated to improve the performance and fairness of recognition models that had been trained on it. Despite the success, the authors do note some limitations with this method. First, its ability to mitigate biased data depends on having a large enough set of data to build from. When the size of the training set is too small its ability to reduce bias may fail or even exacerbate the problem. Second, this method of data augmentation is dependent on an external training set. If the biases inherent in the training set are too severe, the method will again fail to mitigate bias or make it worse. This will happen because the method cannot synthesize images from attributes it does not observe. With these limitations in mind, it is important to consider the quality of the training set before applying this method of data augmentation.

Discussion/Comments

The Harms of Demographic Bias in Deep Face Recognition Research and *Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer* both establish that while there has been progress made in addressing bias in deep learning, specifically in the case of facial recognition, the problem still exists. Discourse around the consequence of this bias in real-world applications often focuses on issues of civil rights, but another potential problem, discussed in the first article, is around the security of these applications. Modern society is dependent on computing technology and increasingly machine learning applications. As with nearly all forms of progress, it comes with new problems. Deep learning applications such as facial recognition and biometrics are often touted as smart security and law enforcement solutions that allow for cost-effective and efficient decision making and enforcement. However, as the first article demonstrated, the failure to address the bias within these systems leaves them vulnerable to circumvention and has serious implications for security. An interesting part of this problem is how that bias reveals itself, an important aspect to understand if it is to be addressed. As mentioned in the second article, “While such systems have proven to be *accurate* by standard evaluation metrics and benchmarks, a surge of work has recently exposed the demographic bias that such algorithms exhibit–highlighting that *accuracy* does not entail *fairness*.” (Georgopoulos). These deep learning applications are often sold on their accuracy, but clearly, that is not a sufficient metric on its own. New metrics measuring the bias and fairness in learning models need more research, development, attention, and commercial marketing if this problem is to be addressed.

New evaluation metrics, improved standards and methods of data collection, data augmentation, and other data de-biasing techniques are all useful approaches to mitigate bias in databases and deep learning models. However, the effectiveness of these techniques in combating bias is ultimately limited by human awareness and behavior. It does not matter how good an algorithm is if the people who wield it are either unaware or simply do not care about the issue. It does not matter how equitable an algorithm is if it is applied inappropriately. A recent NPR report (Allyn) details the story of one of the first documented cases of someone being wrongfully arrested as a result of a false positive produced by a facial recognition application. Robert Williams, a Michigan man, was wrongfully accused and arrested for theft by Detroit police after facial recognition software falsely matched store surveillance video of the thief with a driver’s license photo of Williams. This false identification was the only basis for the accusation against Williams whom they arrested in front of his children and detained for 30 hours. Luckily for Mr.Williams, his case was dropped due to lack of evidence and since then, the Detroit police department has modified some of its rules regulating the use of facial recognition, but this

case is a great example of the harm that can happen when biased models meet human error. In this case, the police were not aware or educated on the potential shortcomings or biases of the technology they were using. In effect, they were handed a tool that they were not ready to use appropriately and that gave them the excuse of “the computer made me do it”, possibly allowing them to avoid examining their own failure to investigate this case properly and thoroughly. As this case demonstrates, reducing bias in deep learning models is an important endeavor, but if it is to succeed, it requires significant effort to raise awareness and education among everyone as well.

Conclusions

Bias in machine learning applications is a deep subject that deserves attention and dedicated research. The failure to do so and implement the resulting work will render these applications less usable, less effective, less deserving of public trust, and may even bring significant, unintentional harm to the very people they are employed to help. The two papers surveyed are a small part of the effort to address this, but more must be done. It is not enough to only improve databases, benchmarks, standards, and models. An effort must also be made to raise the awareness of these issues and educate more people about the subject.

References

Sweeney, Chris, and Maryam Najafian. "A transparent framework for evaluating unintended demographic bias in word embeddings." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

Garcia, Raul Vicente, et al. "The harms of demographic bias in deep face recognition research." *2019 International Conference on Biometrics (ICB)*. IEEE, 2019.

Georgopoulos, Markos, et al. "Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer." *International Journal of Computer Vision* (2021): 1-20.

Allyn, Bobby. “'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man.” *NPR*, NPR, 24 June 2020, www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig.