JOURNAL OF
EMERGING INVESTIGATORS

Article

Propagation of representation bias in machine learning

Upamanyu Dass-Vattam and Elizabeth MacGregor
Westford Academy, Westford, MA

SUMMARY

Transfer learning is an emerging paradigm in machine learning that involves reusing existing pretrained models to develop new machine learning applications. As machine learning grows in importance, keeping new models and applications clear of biases is of paramount importance. While previous works focus on the development of bias free models, they fail to address the mechanism of bias propagation in transfer learning. Using facial recognition as a use-case scenario, we attempt to identify sources of bias in a model developed using transfer learning. To achieve this task, we developed a model based on a pre-trained facial recognition model, and scrutinized the accuracy of the model's image classification against factors such as age, gender, and race to observe whether or not the model performed better on some demographic groups than others. By identifying the bias and finding potential sources of bias, his work contributes a unique technical perspective from the view of a small scale developer to emerging discussions of accountability and transparency in AI.

INTRODUCTION

Our society is adopting artificial intelligence (AI) at an unprecedented rate, especially machine learning technology. Across both the public and private sectors, organizations use machine learning to aid decision making on high-stakes tasks. For example, government organizations use machine learning for predictive policing or determining a person's eligibility for pension payments, housing assistance, or unemployment benefits. In the private sector, companies use machine learning to select job applicants, and banks use them to determine the creditworthiness of loan applicants or set interest rates. Machine learning systems are versatile and can perform high-stakes in technology pipelines without being explicitly designed for them. For example, face recognition models can be used to identify suspects by authorities, although such models might not have originated in the context of law enforcement. Therefore, machine learning systems have to be developed and deployed with extreme care, else bias infects such crucial decisions. Data bias in machine learning is a type of error in which elements of a dataset are more heavily weighted or represented. In this study, we examine representation bias and racial bias.

Deep learning (1) has emerged as the state-of-the-art in machine learning. Over the past decade, deep learning models have achieved remarkable success in various research areas. Evolved from previous research on artificial neural networks, large-scale deep learning models with billions of parameters have shown superior performance compared to other machine learning algorithms in areas such as image and voice recognition and natural language processing, among others.

Transfer learning (2) is an emerging archetype within deep learning which involves reusing and repurposing existing pre-trained models to develop new machine learning systems and is rapidly becoming commonplace in AI development. However, pre-trained models can harbor latent biases that, unbeknownst to the developer, are spread to the deployed applications through transfer learning. Therefore, studying pre-trained models as first-class objects and examining the impact of transfer learning on bias propagation is crucial.
Through this work, we aim to identify and clarify the mechanism of bias propagation in machine learning, with the hope of contributing a unique technical perspective to emerging discussions on fairness, accountability, and transparency of AI systems.

In a typical use case scenario, the input to a facial recognition algorithm is a single image. The output is a label, which could be the identity of that face or a trait associated with the face, such as age or gender. In machine learning, researchers treat facial recognition as supervised learning, and it remains an active area of research in computer vision. Since the breakthrough of AlexNet (3) for general image classification, there has been a flurry of research on applying deep learning methods to face recognition. Deep learning approaches have not only achieved, but exceeded human-level performance on standard facial recognition datasets within a few years of wider adoption of this approach. Wang and Deng (4) provide a helpful summary of the state of face recognition research, highlighting the broad trends from earlier simpler learning methods to the state-of-the-art deep learning methods.

Training high-performance deep learning models like the ones used for facial recognition requires enormous computational resources, well beyond the vast majority of organizations' reach. As such, researchers in large institutions and for-profit giants such as Google or Microsoft are largely responsible for the development of these models. These institutions then release the models as pre-trained models for the users in the rest of the AI community to use. The release of the pre-trained models allows the common

user from the AI community to reuse, re-purpose, fine-tune, and transfer them for use in a variety of real world machine learning applications.

Following the discovery of unintended bias in many machine learning systems that use deep learning approaches, research into fair and transparent AI is gaining significant attention. Bolukbasi et al. (5) exposed gender bias in a commonly used text analysis technique involving a well-known pre-trained word embedding model. Unfortunately, this issue goes beyond text and encompasses other modalities as well, including images. Buolamwini et al. (6) analyzed the accuracy of commercial face recognition products across light- and dark-skinned males and females. Their research considered products sold by Microsoft, Face++, and IBM and found them to perform far better on males and light-skinned people. The table below shows each product's error rates in predicting a binary classification of male or female from an image (Table 1). These numbers are concerning given that these products are being used by governments and businesses in decision making.

One source of this problem originates with the use of pre-trained models that have been trained on large publicly-available image datasets, sourced from popular online sites such as IMDB and Wikipedia. Since the datasets comprise famous people and celebrities, the training data tend to have a higher representation of males and light-skinned people. Thus, these pre-trained models likely carry an inherent representation bias on account of sampling of the training data.

Representation bias is not the only form of bias that can afflict machine learning systems. Suresh and Guttag (7) provide a taxonomy of biases which includes historical bias (which can arise during data collection or generation), measurement bias (which can arise when choosing and measuring particular features), evaluation bias (which can occur during model interpretation and evaluation), and aggregation bias (which can occur as a result of flawed assumptions about model's population influence).

We restrict the scope of this work to strictly representation bias. The following sections provide a deep dive into the occurrence of representation bias in the context of a specific task (face recognition), a specific type of deep learning model (CNN), and a specific mechanism (transfer learning).

Age and gender are two key facial attributes that play a foundational role in many real-world applications. For instance, Quividi(8) detects the age and gender of users who pass by digital signage and provides targeted advertising, and AgeBot is an Android app that determines age from stored photos. Therefore, age and gender estimation from a single facial image is a task of significant importance in many domains such as human-computer interaction, law enforcement, surveillance, or marketing.

	Microsoft	Face++	IBM
dark skin female	20.8%	34.5%	34.7%
light skin female	1.7%	6.0%	7.1%
dark skin male	6.0%	0.7%	12.0%
light skin male	0.0%	0.8%	0.3%

Table 1: Accuracy of state of the art facial recognition models on images categorized by skin tone and gender.

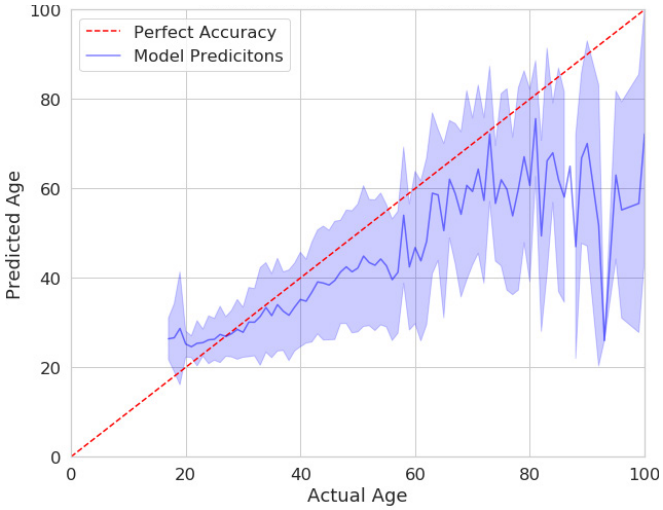

Figure 1: Overall accuracy of the model in terms of actual age vs predicted age. The red dotted line represents the perfect prediction line, where the predicted age meets the actual age.

To analyze the problem, we developed a convolutional neural network (CNN) model, a popular form of deep learning model. This type of network takes advantage of the fact that pixels that are close together are related by reducing the number of pixels and correspondent weights by filtering n x n squares of input pixels into a single destination pixel (3). A CNN was used because studies show that they consistently outperform other models on image recognition tasks and have become the industry standard for such tasks.

The pre-trained model used to develop the current solution is called VGG-Face (9). It is a popular model for implementing face recognition tasks and was originally trained on approximately 2.6 million images of 2000+ celebrity faces. The distribution of age, race, and gender of these 2000+ personalities was not found during a literature search, but the odds are high that these faces are disproportionately white and in the age range of 20 to 50. According to the Hollywood Diversity report, 77% of all film roles were played by white actors, and these datasets are based on celebrities from film and TV (10). We chose to use VGG-Face because of its excellent benchmark performance, extensive documentation, and ease of implementing transfer learning.

The task of age estimation is the focus of the model studied in this work. The objective of examining this model is to not only determine whether or not biases exist, but whether or not those biases match the aforementioned issues with a lack of representation in the data. Since the model used

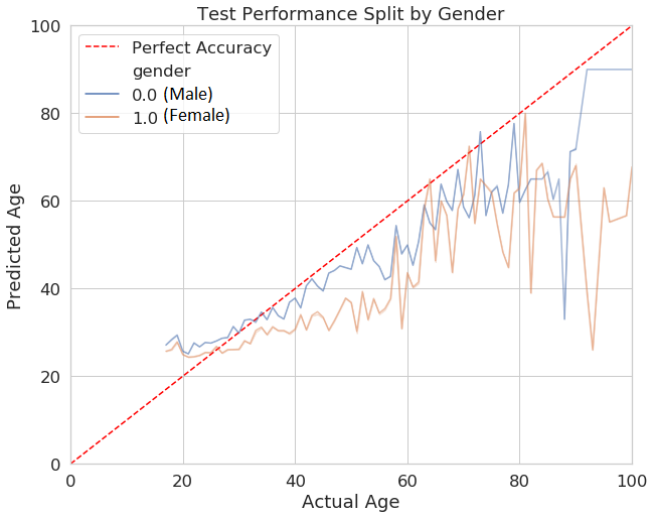

Figure 2: Accuracy of the model in terms of actual age vs predicted age when split by gender. The dataset labels "male" images with a 0 and "female" images with a 1, which was how they were separated.

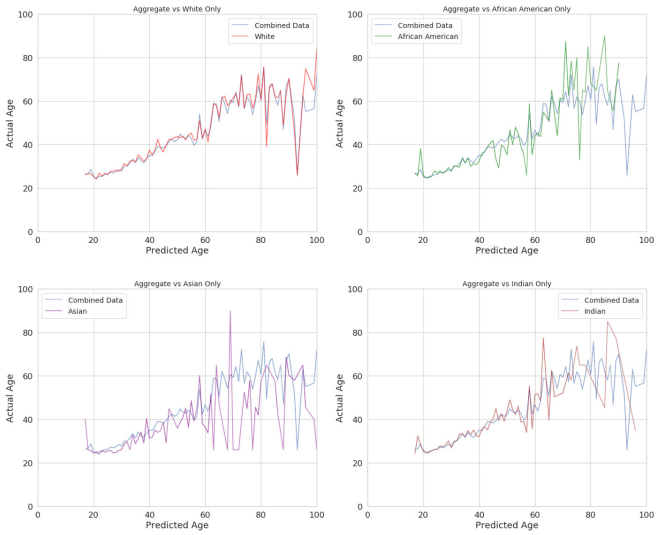

Figure 3: Accuracy of the model on different ethnicities in terms of actual age vs predicted age. Included on each of the plots is the plot overall accuracy of the model (in blue) for comparison.

is a direct offshoot from the VGG-Face model, finding the representation bias from the data used to train VGG-Face demonstrates that those biases were propagated to the new model. In addition to identifying how biases can infect a new model's performance, this work would provide valuable insight to small scale developers reliant on transfer learning. Without the resources to train entirely new massive models from scratch, they must instead be vigilant against the spread of biases.

RESULTS

To test the predictive power of the trained model, we subjected it to samples from the test split consisting of 5915 facial images. We compared the age predictions obtained from the model against the ground truth age labels that came with the dataset, and used mean absolute error (MAE) as a performance measure as it is more resistant to outliers and is considered the industry standard metric for age prediction tasks. The lower the MAE score, the better the performance. On the full test set, we obtained an MAE of 8.704 (compared to a train MAE of 7.793), showing that the model generalizes well to the test set. In table 2, the yellow row was not part of the original paper, and shows our results in comparison to state-of-the-art methods. However, the models in Dehghan et al. (11) trained on the data using a roughly 75/25 train/test split, while the trained model trained on a 60/10/20 train/val/test split.

A prediction curve that closely hugs the perfect accuracy line indicates better performance by the model. The model's estimation is relatively good in the age range of 20 to 50. However, the rapid performance degradation after age 60 is striking (Figure 3).

We examined the performance of the model across different age groups, using three age groups – 20 to 40, 40 to 60, and 60 plus. The MAEs were 4.685 years, 12.889 years, and 19.390 years respectively. The mean average error varies across different age groups, and the degradation of the model performance was higher in upper age groups.

Method	MAE
Sighthound	5.76
Rothe et al.	7.34
Microsoft	7.62
My model	8.70
Kairos	10.57
Face++	11.04

Table 2: A Comparison of MAE for several state-of-the-art methods taken from Deghan et al. (11). The MAE is calculated by taking the average difference between the prediction and actual age of a face in an image.

We also examined the performance of the model across two different gender groups. The error rate for the male group was 12.837 and for the female group was 14.1. Even in the age group of 20 - 40, where the model performance was superior, the age estimation for males was better than the age estimation for the females. This is a clear indicator of possible gender, as the only time that the male and female accuracies are comparable are at the higher variability ages (Figure 2).

We also examined performance of the model on test data across different ethnic groups. The UTKFace dataset labels all the facial images as one of five ethnic categories: White, African American, Asian, Indian, and Other(We ignored the Other category for this analysis). The White performance curve closely tracked the aggregate performance compared to the performance curves of the other ethnic groups, demonstrating the weight that ethinc group carried on the model's performance. Further, the non-White performance

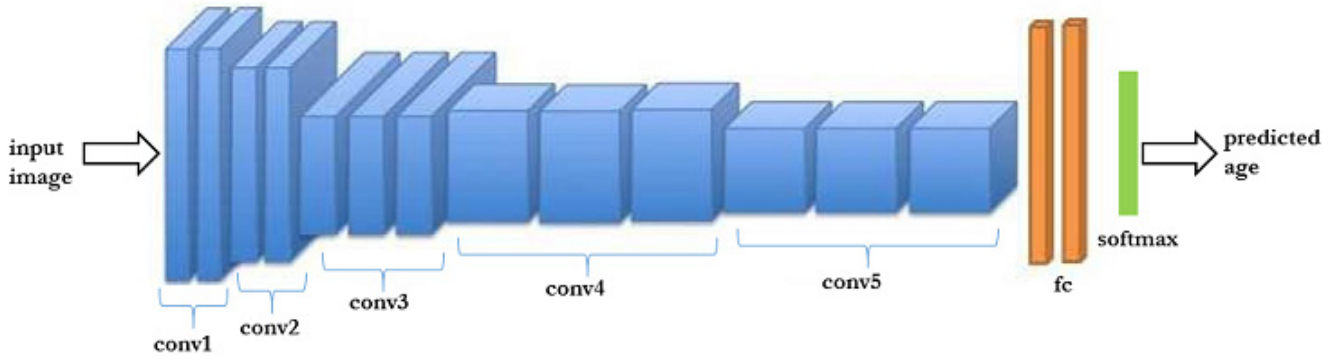

Figure 4: The structure of the CNN. The model is composed of five convolutional layers of different filter size, two fully connected layers, and a softmax layer that returns the predicted age.

Loss function	Categorical Cross-entropy
Optimizer	Adam optimizer
Learning rate	0.001
Decay rate	0.00001
Momentum	0.9
Batch size	512
Epochs	100

Table 3: Parameters used for training the age estimation CNN.

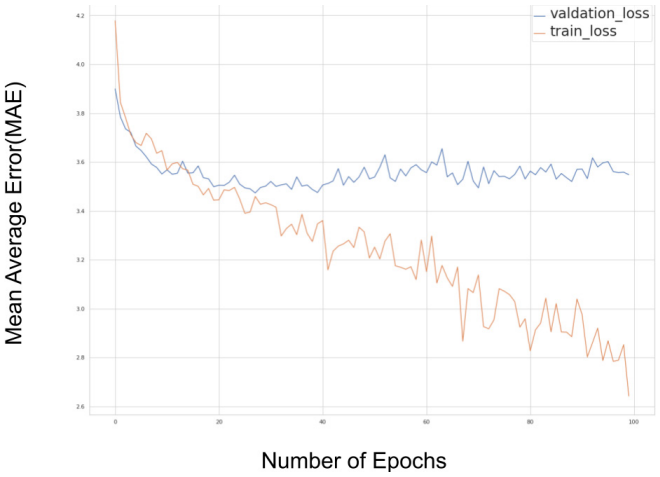

Figure 5: How the MAE decreased as the number of epochs increased. The validation MAE plateaued around 60 epochs even as the training accuracy contained to decrease.

curves were visibly noisier after 60 plus years of age compared to the other ethinc groups (Figure 3).

The MAEs for each of the four ethnic groups were computed. The MAE for whites was 7.819 years, while the MAEs for African Americans and Asians was higher at 8.174 and 10.33 years, respectively. Surprisingly, the MAE for Indians was 7.397 years.

The model's performance generalized well to the test data. The difference between the train and test MAE was about one year, and its performance was comparable to benchmarked results in the literature, as it placed fourth among the six cited systems.

MATERIALS AND METHODS

The age estimator model was developed by deriving a new CNN from VGG-Face. Once instantiated, all the convolutional layers remained the same, but the last fully-connected layer was replaced with a new one containing 101 nodes (one for each age in the range [0,100]. Finally, a Softmax layer with 101 class nodes was added at the end. The input to the model is a single facial image, and the output is a number in the range [0, 100], representing the image of the age.

UTKFace (12) is an image dataset containing faces. It consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover a long age span (from 0 to 116 years old). It also covers a large variation in pose, facial expression, illumination, resolution, and other features. This dataset can be used for a variety of tasks like face detection, age estimation, or gender recognition. It provides two versions, "in the wild" faces and "aligned and cropped" faces. The former is used for the experiments presented here.

The data was split into 60/10/30 percent Train/Validation/Test splits. The model was trained on Google Cloud GPU platform with 4 GPUs. It took approximately 3 hours to fully train the model. The training yielded a test loss of 2.6422 and a validation loss of 3.5481, respectively.

DISCUSSION

When tested for performance across different age groups, the model showed strong tendencies for age bias. The best performing age group was 20-40. The percentage difference in MAE between age groups 20-40 and 40-60 was around 93%, and between 20-40 and 60-plus was around 122%. When tested across male and female gender groups, the percentage difference in MAE between the male (best performing group) and female groups was around 9%. These results suggested that gender bias was less pronounced compared to age. When tested for performance across four different ethnic groups, the demonstrated tendency for racial

bias was mixed. The approximate percentage differences between the best performing group (White) and other groups (African American, Asian, and Indian) were respectively 4.5%, 27%, and -5%. It is curious to note that the MAE for the Indian group was the lowest, although the performance curve tells a different story. One possible explanation is that the small sample size leaves MAE more vulnerable to outliers and deviations observed in the post-60, non-White groups. Perhaps, using a different metric such as mean absolute scaled error or mean absolute deviation might bring the differences into sharper contrast. To sum up, the overall model performance was satisfactory, but it exhibited strong signs of age bias, moderate signs of racial bias, and low levels of gender bias.

There are many benefits and risks associated with transfer learning. It brings high-performance deep learning models within reach of individuals and smaller institutions. Even high school students can work with state-of-the-art deep learning models without the need for access to massive computer resources that only larger institutions can support. We trained the model on just 4 GPUs within a few hours because we were able to use a pre-trained model. It levels the playing field to some extent and avoids massive concentration of AI power in the hands of a few institutions. It democratizes the development of machine learning applications and promotes open-source culture. However, latent biases tend to propagate from pre-trained models to derived models, as demonstrated in the current work. The technical-knowledge barrier to using the pre-trained models within developer-friendly tools such as TensorFlow and PyTorch is quite low. Case in point - we trained a sophisticated CNN without only a surface level understanding of how CNNs work, increasing the risk of developing models with many unknowns and embedded assumptions. Deep learning models are black-box models that lead to the interpretability problem, where although models may be highly accurate, humans cannot understand the causes of the decision. This problem is exasperated in transfer learning due to many levels of indirection. We believe that the benefits of transfer learning outweigh the risks and that transfer learning is here to stay. However, we need to take sufficient measures to mitigate the risks outlined above.

We propose a few steps that we can take to mitigate the risks of transfer learning. Firstly, we need better education: Consumers of pre-trained models should develop an in-depth understanding of these models' workings. It is not enough to simply use them, but one has to keep in mind that these models carry with them inherent biases depending on the training data. Ensuring a similarity between the data the model was trained on and the data one will use to fine-tune the model should help mitigate the problem. Secondly, we should test on different datasets: It is important to test the derived model on a diversity of datasets. Deploying a machine learning model by training and testing on a single dataset is reckless, as the model will carry with it the tendencies of the dataset. If that model then spreads widely, any biases of the dataset would be propagated. Lastly, we recommend peer reviews: Sharing models with the larger community and allowing the community to give feedback can reduce some of the risks associated with transfer learning. Reproducing results can provide a good check of a pre-trained model, especially if an intentionally different dataset is used.

Transfer learning is an emerging paradigm in machine learning. One of the risks of this approach is propagating biases from producers to consumers of pre-trained models. In this work, the mechanism of representation bias propagation was examined in the context of the facial recognition task. The results confirmed the risks of bias propagation and allowed quantification and fine-grained analysis of these risks.

Any claims from this work must be taken with the caveat that this study was limited to one type of model on one type of data using one data set. As such, generalization claims about the racist, sexist, and ageist nature of this class of models cannot be strongly supported. However, even this narrowly-scoped study highlights the need for critical consumption of transfer learning. Those who use pre-trained models cannot be content to accept what is given to them as infallible. The methods of data sampling alone risk the introduction of many kinds of biases. However, the layers upon layers of modeling bundled together based on the assumption of "correctness" should alert the user to potential biases.

At least two research directions can follow this work. One direction is to expand the scope of the study by including additional types of models and a larger set of datasets. Another direction is to explore model correction methods, which can be applied post hoc to account for and correct biases during transfer learning.

REFERENCES

1. LeCun, Y., Bengio, Y., & Hinton, G. "Deep learning." Nature. Volume 521, Issue 7553, 2015. pp. 436.
2. Weiss, K., Khoshgoftaar, T. M., & Wang, D. "A survey of transfer learning." Journal of Big Data, Volume 3, Issue 1, 2016. Pp. 9.
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. pp. 1097-1105.
4. Wang, M., & Deng, W. "Deep face recognition: A survey." arXiv preprint arXiv:1804.06655. 2018.
5. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 2016. pp. 4349-4357.
6. Buolamwini, J., & Gebru, T. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. January 2018. pp. 77-91.
7. Suresh, H., & Guttag, J. V. "A Framework for

Understanding Unintended Consequences of Machine Learning." arXiv preprint arXiv:1901.10002. 2019.
8. Quividi. quividi.com/.
9. Parkhi, O. M., Vedaldi, A., & Zisserman, A. "Deep face recognition." bmvc. Vol. 1, No. 3, 2015. p. 6.
10. Hunt, D., Ramon, A., Tran, M. Hollywood Diversity Report. UCLA College of Social Sciences, 2019.
11. Dehghan, A., Ortix, E., Shu, G., & Masood, S. “DAGER: Deep Age, Gender, and Emotion Recognition Using Convolutional Neural Networks.” arXiv preprint arXiv: 1702.042280, 2017
12. Zhang, Z., Song, Y., & Qi, H. "Age progression/regression by conditional adversarial autoencoder." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017. pp. 5810-5818.

Article submitted: June 9, 2020
Article accepted: Septermber 3, 2020
Article published: February 20, 2021