Propagation of Representation Bias in Machine Learning

RP485

Abstract

Transfer learning is an emerging paradigm in machine learning that involves reusing existing pretrained models to develop new machine learning applications. Transfer learning is quite common these days. But pretrained models can harbor biases that are unintentionally propagated to the deployed applications through transfer learning. It is important to study pretrained models as first-class objects and examine the mechanism of bias propagation in transfer learning. Using the task of face recognition as a use case, this work explores the issue of representation bias propagation in the context of transfer learning. A Convolutional Neural Network (CNN) model was developed using the (pretrained) VGG-Face model for the proposed task. The predictions of the model were investigated for age, gender and racial bias, revealing that the trained model indeed demonstrated tendencies for representation bias. By making the mechanism of bias propagation explicit, this work contributes a unique technical perspective to emerging discussions on fairness, accountability, and transparency of AI.

Propagation of Representation Bias in Machine Learning

Introduction

Motivation

Our society is adopting AI at an unprecedented rate, especially machine learning technology. Across both the public and private sectors, organizations are using machine learning to aid decision making on high-stakes tasks. For example, government organizations are using machine learning for predictive policing, or determining a person's eligibility for pension payments, housing assistance or unemployment benefits. In the private sector, companies are using machine learning to select job applicants, or banks use them for determining the creditworthiness of loan applicants or setting interest rates. Even machine learning systems that are not specifically designed to perform high-stakes tasks can be used in technology pipelines to such effect. 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.

Deep learning (LeCun, Bengio & Hinton, 2015) 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, natural language processing, among others.

Transfer learning (Weiss et al., 2016) is an emerging archetype within deep learning which involves reusing and re-purposing existing pretrained models to develop new machine learning systems, and is rapidly becoming commonplace in AI development. However, pretrained models can harbor latent biases that, unbeknownst to the developer, are spread to the deployed applications through transfer learning. Therefore, studying pretrained models as first-class objects and examining the impact of transfer learning on bias propagation is crucial.

Goal

The objective of this work is to make the mechanism of bias propagation in machine learning more clear, with the hope of contributing a unique technical perspective to emerging discussions on fairness, accountability, and transparency of AI systems.

Problem

To analyze the problem, this work will:

- Focus on the task of trait prediction from facial images.
- Develop a convolutional neural network (CNN) model as a solution.
- Transfer a pretrained model known as *VGG-Face* to develop the said CNN
- Use a dataset known as *UTKFace* consisting of 20000+ facial images to train and evaluate the model

The problem here is to show that the CNN model derived from VGG-Face, which potentially harbors representation bias, also exhibits tendencies of racial, age, and gender bias.

Related Work: Face Recognition, Deep Learning, Representation Bias

In a typical use case scenario, the input to a facial recognition algorithm is a single image and the output is a label, which could be the identity of that face or a trait associated with the face such as age, gender, etc. In machine learning, this task is generally treated as supervised learning. Facial recognition remains an active area of research in computer vision. Since the breakthrough of AlexNet (Krizhevsky et al., 2012) 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 this approach. Wang and Deng (2018) provides 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 require enormous computational resources, well beyond the reach of the vast majority of organizations. As such, the development of these models has largely been carried out by researchers in large research institutions and for-profit giants such as Google, Microsoft, and Facebook. These pretrained models are then released as *pretrained models* for the users in the rest of the AI community to use. This allows the community to reuse, re-purpose, fine-tune, and transfer them for use in a variety of machine learning applications which are then deployed in many real world settings. *Transfer learning* (Weiss et al., 2016) refers to this paradigm of deriving new deep learning models from existing pretrained models.

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. (2016) exposed gender bias in a commonly used text analysis technique involving a well-known pretrained word embedding model. Unfortunately, this issue goes beyond text and encompasses other modalities as well, including images. Buolamwini et al. (2018) 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. These numbers are concerning given that these products are being used by governments and businesses in decision making.

One source of this problem can be traced to the use of pretrained models that have been trained on large publicly-available image datasets, sourced from popular online sites such as IMDB and Wikipedia. Being datasets of famous people and celebrities, the training data mostly contains images of light skinned people and tend to have a higher representation of males. So these pretrained models likely carry an inherent representation bias on account of sampling of the training data.

	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%

Representation bias is not the only form of bias that can afflict machine learning systems. Suresh and Guttag (2019) 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).

The scope of this work will be restricted to representation bias alone. 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).

The Task: Age Prediction from Facial Image

Age and gender are two key facial attributes that play a foundational role in many real world applications. For instance, Quividi[1] detects age and gender of users who pass by a 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

[1] https://quividi.com/

facial image is a task of significant importance in many domains such as human-computer interaction, law enforcement, surveillance, or marketing.

The task of age estimation is the focus of this work. The input to the model is a single facial image and the output is a number in the range [0, 100].

The Model: Convolutional Neural Network

A solution to the proposed task is developed here using *Convolutional Neural Networks (CNN)* (Krizhevsky et al., 2012). CNNs are a popular form of deep learning models. A CNN consists of an input and an output layer, and typically multiple hidden layers. The hidden layers of a CNN consist of a series of convolutional layers. The activation function commonly used in convolution layers is Rectified Linear Unit (RELU). Convolution layers are followed by additional pooling, fully-connected, and normalization layers. In case of classification tasks, the final convolution layer is followed by a Softmax layer.

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. In the past few years, the ImageNet Competition has provided a proving ground for the new architectures and developments in CNNs. Their superior performance in this annual competition vis-à-vis other machine learning approaches has led to large-scale acceptance of CNNs as the go-to models for image classification tasks. The ImageNet Competition is a 1000 category classification task with over 1million images in the training set. Within a span of few years, the classification error in the competition was reduced from around 28% to around 2% due to the advancement in CNN architectures.

The pretrained model used to develop the current solution is called *VGG-Face* (Parkhi et al., 2015). 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 literature search, but the odds are high that these faces are disproportionately white and in the age range of 20 to 50. So, it is assumed here that VGG-Face harbors latent representational bias.

VGG-Face is a 16 layer CNN with 13 convolution layers (some with down sampling), 2 fully-connected layers, and a softmax output layer. VGG-Face was selected because of its excellent benchmark performance, extensive documentation, and the ease of implementing transfer learning.

The age estimator model was developed by deriving a new CNN from VGG-Face using the following steps:

1. Instantiate the VGG-Face model
2. Keep all the convolutional layers of the VGG-Face leading up to the last fully-connected layer unchanged
3. Replace the last fully-connected layer with a new fully-connected layer containing 101 nodes (one for each age in the range [0,100])
4. Add an Softmax layer with 101 class nodes at the end

The final architecture is presented in Figure 1.

The Dataset: UTKFace

UTKFace[2] (Zhang et al., 2017) is an image dataset containing faces. It consists of over 20000 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 variety of tasks like face detection, age estimation, gender recognition, etc. It provides two versions, “in the wild” faces and “aligned and cropped” faces. The former is used for the experiments presented here. Samples from UTKFace dataset are presented in Figure 2.

[2] https://susanqq.github.io/UTKFace/

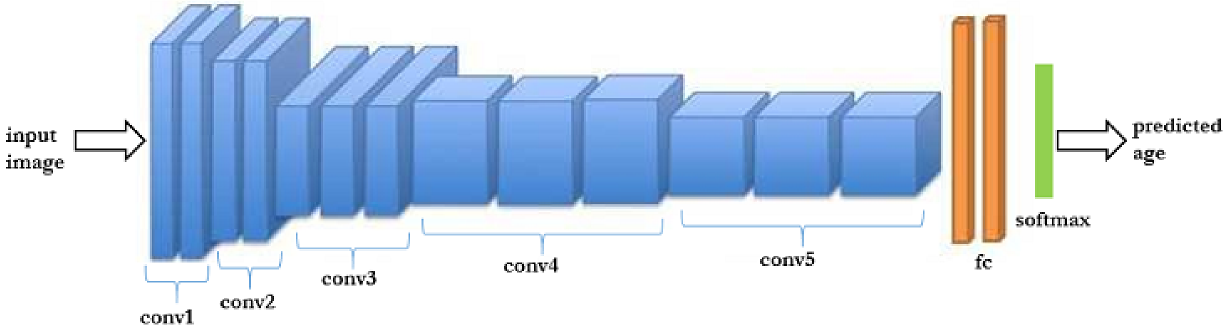

Figure 1. A CNN model for age prediction derived from VGG-Face

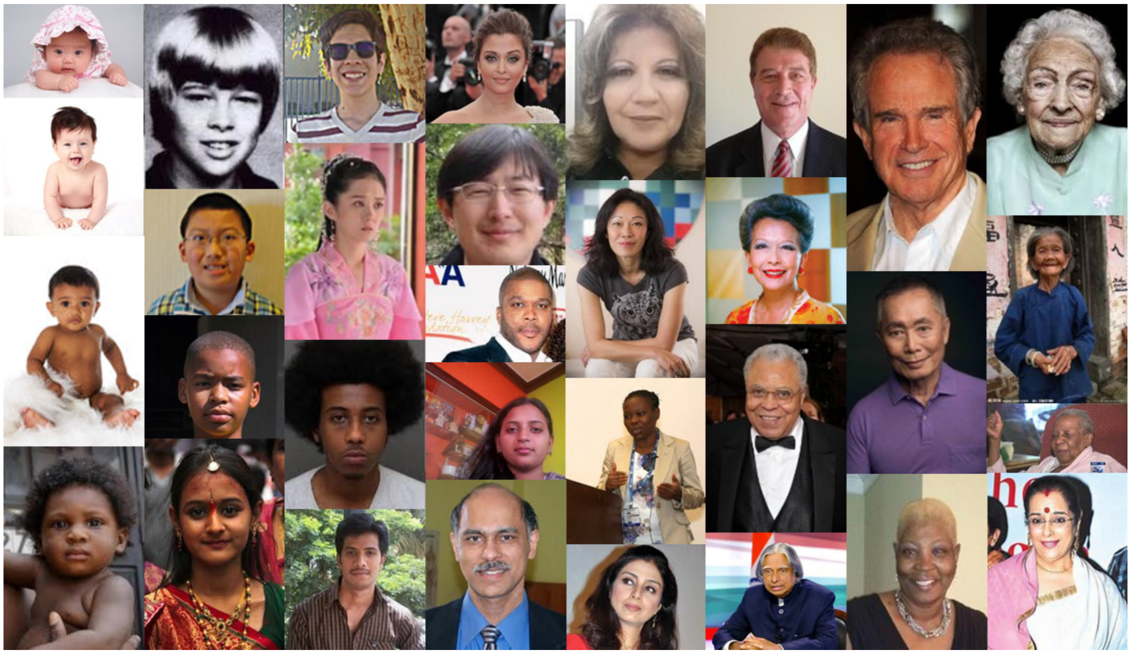

Figure 2. Sample images from UTKFace dataset

(source: https://susanqq.github.io/UTKFace/)

The Training

The CNN for age estimation discussed above was trained with the parameters listed below. 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. The learning curves are shown in Figure 3.

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

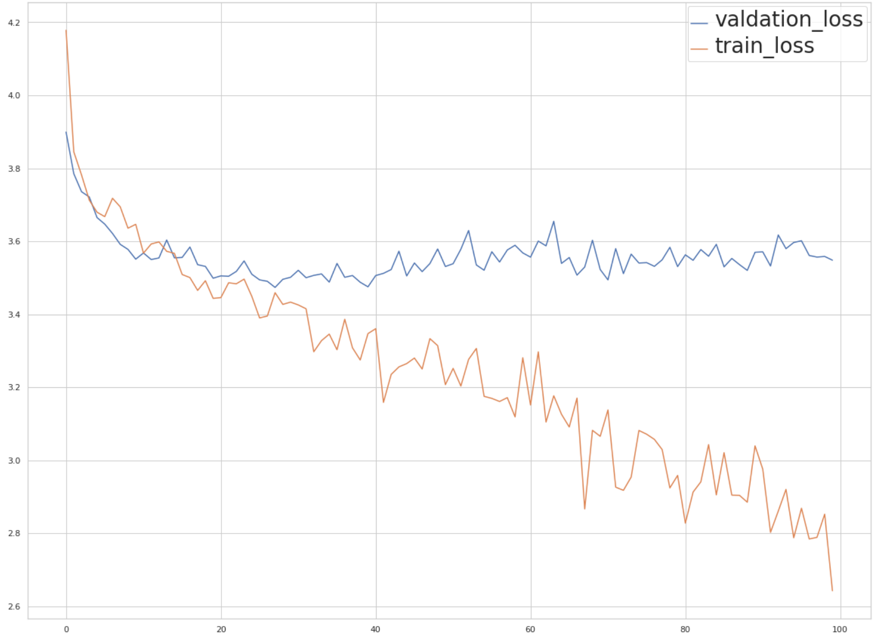

Figure 3. Learning curves (epochs vs. loss) at training time

Evaluation and Results

To test the predictive power of the trained model, it was subjected to samples from the test split consisting of 5915 facial images. The age predictions obtained from the model were compared against the ground truth age labels that came with the dataset. Mean Absolute Error (MAE) was used as a performance measure as it is more resistant to outliers and is considered the industry standard metric for age prediction tasks. Lower the MAE score, the better the performance.

Overall results

On the full test set, a MAE of 8.704 was obtained (compared to a train MAE of 7.793) showing that the model generalizes well to the test set. For comparison, the table below was cited by Dehghan et al. (2017) comparing the MAE of several state-of-the-art methods for age prediction (the yellow row was added here and was not part of the original paper).

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

Figure 4 below plots the performance curve of the model on the full test data. The closer the prediction curve hugs the perfect accuracy line, the better the performance of the model. This plot shows that the model's estimation is relatively good in the age range 20 to 50.However, the rapid performance degradation after age 60 is striking.

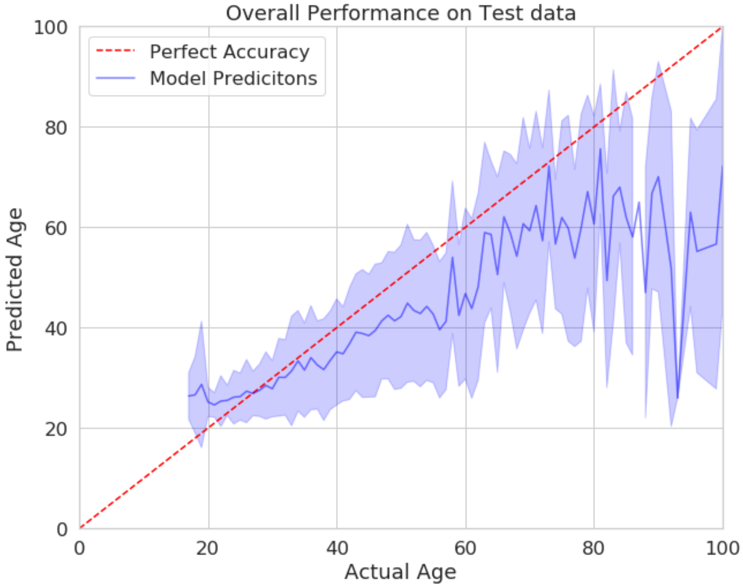

Figure 4. Performance curve on full test data

Results by age groups

The performance of the model across different age groups was examined. Three age groups were created: 20 to 40, 40 to 60, and 60 plus. The following shows MAE per age group. The error rates clearly vary across different age groups and the degradation of the model performance is higher in upper age groups.

Age group	MAE
20 - 40	4.685
40 - 60	12.889
60 plus	19.390

Results by gender

The performance of the model across two different gender groups was also examined. The error rates for the Male group was 12.837 and for Female group was 14.1. The plot in Figure 5 captures the performance of the model across the two gender groups. The plot shows that even in the age group of 20 - 40 where the model performance was superior, the age estimation of males are better than the age estimation of the females.

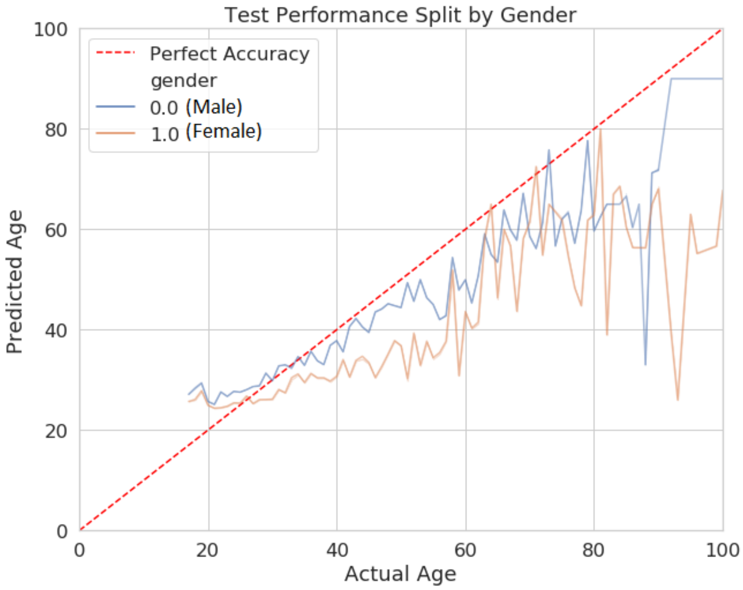

Figure 5. Gender-wise performance curves on test data

Results by ethnic group

Performance of the model on test data across different ethnic groups was also examined. The UTKFace dataset contains ground truth race labels for all the facial images. There are a total of 5 ethnic categories: White, African American, Asian, Indian, and Other. The Other category was ignored for this analysis.

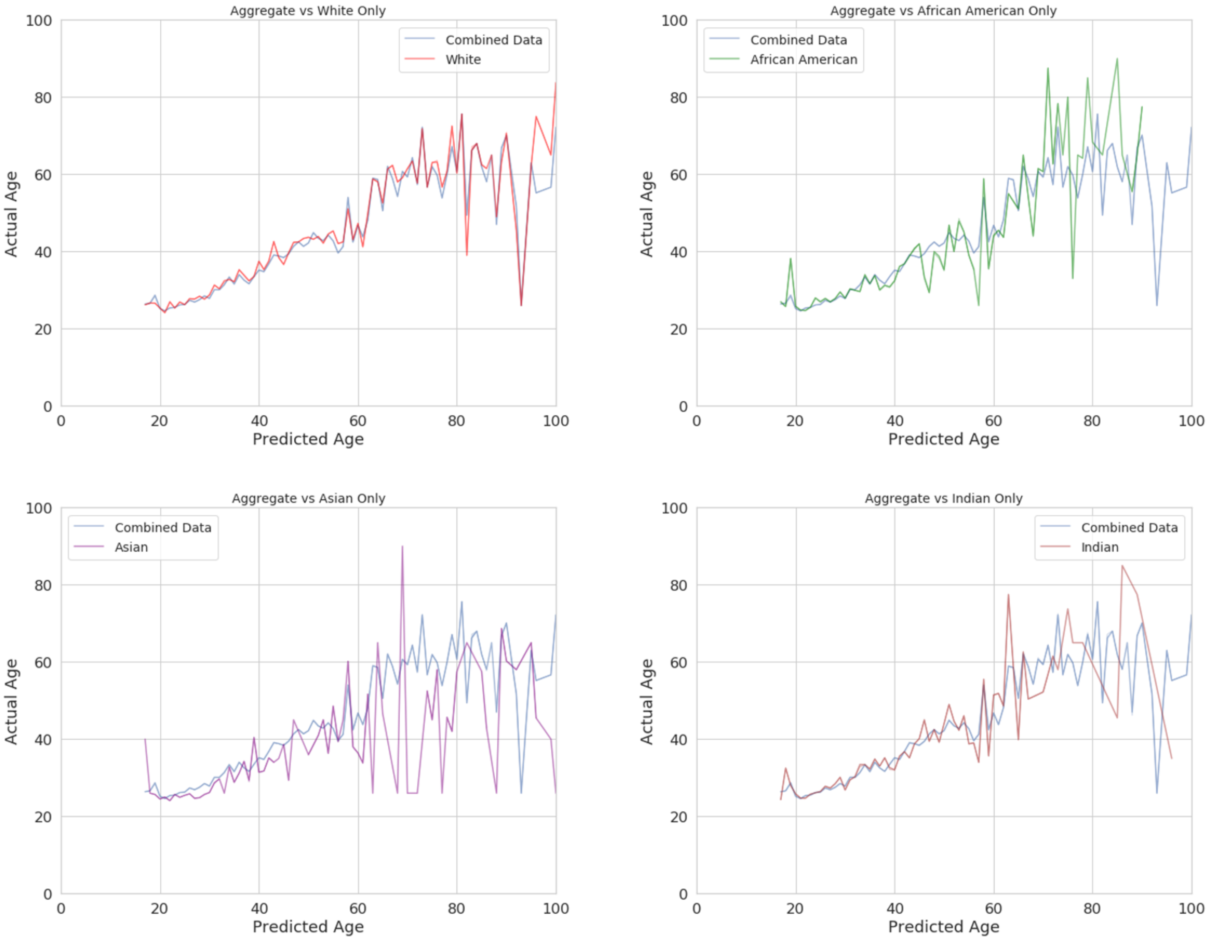

Figure 6. Model performance across different ethnic groups

The plots in Figure 6 show the performance curves across these ethnic groups. Notice how closely the White performance curve tracks the aggregate performance compared to the performance curves of the other ethnic groups. Further, the non-White performance curves are significantly more noisy after 60 plus years of age.

The MAEs for each of the four ethnic groups was computed. The following table presents the MAE values obtained across different groups.

Ethnic group	MAE
White	7.819
African American	8.174
Asian	10.33
Indian	7.397

Analysis of results

Results from the evaluation reveal the following insights about the current model.

- The model's performance generalized well. The difference between train MAE and test MAE was about 1 year where the range was 0 to 100 years.
- The model's performance was comparable to benchmarked results in the literature. It was placed fourth among the six cited systems.
- 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%. This suggests that the 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 lowest although the performance curve tells a different story.

One possible explanation is that MAE is not robust to the level of 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.

Discussion

Transfer learning: A boon or bane?

There are many benefits and risks associated with transfer learning. Some of the benefits are:

- It brings high-performance deep learning models within reach of individuals and smaller institutions. Even high school students like me can work with state-of-the-art deep learning models without the need for having access to massive compute resources that only larger institutions can support. I trained my model on just 4 GPUs within a few hours because I was able to use a pretrained 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.

The risks associated with transfer learning are:

- Latent biases tend to propagate from pretrained models to derived models as demonstrated in the current work.
- The technical-knowledge barrier to using the pretrained models within developer-friendly tools such as TensorFlow and PyTorch is quite low. Case in point -

I trained a pretty sophisticated CNN without having a deep understanding of how CNNs work. This increases the risk of developing models with a lot of unknowns and embedded assumptions.

- Deep learning models are black box models which lead to the *interpretability problem*. The decisions of these models are not easily interpretable by humans. This problem is exasperated in transfer learning due to many levels of indirection.

I 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.

What steps can be taken to mitigate the risks of transfer learning?

In my opinion there are at least three steps that we can take to mitigate the risks of transfer learning.

- Better education: Consumers of pretrained models should develop an in depth understanding of the workings of these models. It is not enough to simply use them, but one has to use them with proper understanding.

- Testing 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.

- Peer reviews: Sharing models with larger community and allowing the community to give feedback can reduce some of the risks associated with transfer learning.

- Model correction: Several techniques are being researched that corrects a trained model to account for biases and assumptions in pretrained models. Applying those might also be a good idea.

Conclusion and Next Steps

Transfer learning is an emerging paradigm in machine learning. One of the risks of this approach is the propagation of biases from producers to consumers of pretrained models. In this work, the mechanism of representation bias propagation was examined in the context of the facial recognition task. The results not only confirmed the risks of bias propagation, but also 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 the class of models used here cannot be strongly supported. However, even this narrowly-scoped study highlights the need for critical consumption of transfer learning. Those who use pretrained models cannot be content to accept what is given to them as infallible. The methods of data sampling alone risks the introduction of many kinds of biases, but the layers upon layers of modeling bundled together based on the assumption of "correctness" should signal danger signs.

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

References

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Buolamwini, J., Gebru, T. (2018, January). *Gender shades: Intersectional accuracy disparities in commercial gender classification.* In Conference on fairness, accountability and transparency (pp. 77-91).

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Parkhi, O. M., Vedaldi, A., Zisserman, A. (2015). Deep face recognition. In *bmvc* (Vol. 1, No. 3, p. 6).

Suresh, H., Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002.

Wang, M., Deng, W. (2018). *Deep face recognition: A survey.* arXiv preprint arXiv:1804.06655.

Weiss, K., Khoshgoftaar, T. M., Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 9.

Zhang, Z., Song, Y., Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE Conference on *Computer Vision and Pattern Recognition* (pp. 5810-5818).