InclusiveFaceNet:
Improving Face Attribute Detection with Race and Gender Diversity

Hee Jung Ryu 1 Hartwig Adam * 1 Margaret Mitchell * 1

Abstract
We demonstrate an approach to face attribute detection that retains or improves attribute detection accuracy across gender and race subgroups by learning demographic information prior to learning the attribute detection task. The system, which we call InclusiveFaceNet, detects face attributes by transferring race and gender representations learned from a held-out dataset of public race and gender identities. Leveraging learned demographic representations while withholding demographic inference from the downstream face attribute detection task preserves potential users’ demographic privacy while resulting in some of the best reported numbers to date on attribute detection in the Faces of the World and CelebA datasets.

1. Introduction
Detecting and recognizing different face attributes has become an increasingly feasible machine learning task due to the rise in deep learning techniques and large people-focused datasets (Niu et al., 2016; Liu et al., 2015). However, this rise has come at a cost: As public technology incorporates modern computer vision techniques, we are observing a troubling gap between how well some demographics are recognized compared with others (EOP, 2016). For example (Buolamwini & Gebru, 2018), the darker and the more feminine the face is the worse commercial gender classifiers performed.

In this paper, we examine the problem of unequal performance across different race and gender subgroups in the context of face attribute detection. We focus on inclusion, which we use in this paper to mean improving accuracy across minority subgroups; and demographic privacy, where our approach to embracing coarse-grained demographic characteristics that affect the visual characteristics of face attributes does not require demographic inference on users of the downstream face attribute task. Inclusion is related to but not identical to notions of fairness, where the goal is to obtain equal performance across all subgroups (even if that means lower accuracy). Our approach is similar in spirit to the recently introduced proposal for decoupled classifiers (Dwork et al., 2018), where the learning of sensitive attributes can be separated from a downstream task in order to maximize both fairness and accuracy. As in (Dwork et al., 2018), we use transfer learning. While (Dwork et al., 2018) uses its sub-type, domain adaptation, we use its different sub-type, task transfer learning. In our setting, both gender and race are learned as independent, coupled classifiers, and are not inferred at run time.

InclusiveFaceNet matches or improves over a baseline without transfer learning, defining a new state-of-the-art performance for face attribute detection across gender boundaries in both the CelebA (Liu et al., 2015) and FotW (Escalera et al., 2016) datasets.

2. Ethical Considerations
Intent. The intent of this work is to demonstrate the utility of reasoning about demographics – in this case, race and gender – in an ethical and legal setting.

Race and Gender categories. Race and gender categories are imperfect for many reasons. For example, there is no gold-
standard for ‘race’ categories, and it is unclear how many race and gender categories should be stipulated (or whether they should be treated as discrete categories at all). Throughout our experiments and in this paper, we obscure race and gender names in order to avoid introducing unconscious societal bias associated with existing labels, and anticipate more nuanced treatment of the sensitive categories in future work.

Race and gender identities in this work are learned from a held-out set of public identifications, and the learned representations are then transferred for learning the face attribute models. This permits face attribute detection to leverage race and gender representation without ever having to uniquely predict these characteristics on another individual.

3. Background & Related Work

Research in computer vision has included work on issues that have direct social impact, such as security and privacy, however, research on the related issue of diversity and inclusion in vision is surprisingly lacking (but see, e.g., (Buolamwini, 2016)). There are several other strands of research background relevant to the work in this paper: face attribute detection and machine learning fairness.

Face Attribute Detection. The ChaLearn “Looking at People” challenge from 2016 (Escalera et al., 2016) provides the Faces of the World (FotW) dataset, which annotates gender and the presence of smiling on faces. (Zhang et al., 2016) won first place in this challenge, utilizing multi-task learning (MTL) and fine-tuning on top of a model trained for face recognition (Parkhi et al., 2015). (Rajeev Ranjan & Chellappa, 2017) later published an out-performing result for the same task on FotW utilizing MTL and transfer learning from a face recognition model (Sankaranarayanan et al., 2016). (Wang et al., 2016) defined state-of-the-art performance on CelebA face attributes (Liu et al., 2015), employing both MTL and transfer learning from a face recognition model based on FaceNet (Schroff et al., 2015). This work is similar to the approach taken here, but includes additional processing and data, e.g., geo-location.

(Escalera et al., 2016) points out the importance of demographic diversity of the source material in a vision dataset, and postulates that analysis made on skewed datasets cannot be representative enough for benchmarking progress. The authors provide Faces of the World (FotW), collected with the aim to achieve a uniform distribution across two genders and four ethnic groups, and we use this dataset in the current paper (but note that we still find demographic skew).

Fairness in Machine Learning. (Dwork et al., 2012) demonstrates the importance of “Fairness through Awareness”, understanding sensitive characteristics like gender and race in order to build demographically inclusive models. Proposals for fairness have included parity, such as Demo-

4. Learning Diversity

Hypothesis. At the core of this work lies the idea that faces look different across different races and genders (S et al., 2014). If such a relationship holds, then face attribute detection could be improved by learning about race and gender characteristics.

Evaluation Metrics. In addition to overall accuracy, we also provide accuracy per demographic subgroup to examine the effect of our experiments on each race and gender subgroup. Inspired by recent work in machine learning fairness (see Section 3), we also evaluate with a metric that averages the false positive rate and false negative rate, Average False Rate (AFR), which is robust to label and subgroup imbalances in the test data.

4.1. Twofold Transfer Learning

Transfer learning is useful in the case where there is a feature space $X$ and a marginal distribution over the feature space $P(X_S)$, where $X_S = x_1, x_2, ..., x_n \in X$, as well as a marginal distribution over the related feature space $P(X_D)$, where $X_D \in X$. Here, $X_S$ and $X_D$ are completely disjoint sets, respectively representing face images for face attribute detection and face images for race classification. Formally, given:

- A face recognition domain $F_X$
- A face domain $F_R$ for the task of race classification $T_R$ with labels $Y_R$ and images $X_R$
- The target face domain $F_S$ for face attribute detection $T_S$ with labels $Y_S$ and images $X_S$

We optimize for $T_R$ by learning $P(Y_R|X_R)$ via transfer learning from a network trained in the face recognition space $F_X$ with the pre-trained network frozen, and then freeze the learned representations from $F_X$ and $F_R$ in the target domain $F_S$ when learning $P(Y_S|X_S)$. We refer to the use of transfer learning twice in this way as twofold transfer learning.
Table 1. Estimated Race & Groundtruth Gender Distribution across Datasets. Duplicated images in CelebA are counted only once in this table.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Estimated Race</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S1  S2  S3  S4  Other</td>
<td>G1  G2  Other</td>
</tr>
<tr>
<td>Train</td>
<td>6171</td>
<td>8% 48% 11% 15% 19%</td>
<td>54% 48% 2%</td>
</tr>
<tr>
<td>Valid</td>
<td>3086</td>
<td>8% 53% 11% 17% 11%</td>
<td>44% 55% 1%</td>
</tr>
<tr>
<td>Train</td>
<td>162687</td>
<td>7% 75% 7% 8% 4%</td>
<td>58% 42% -</td>
</tr>
<tr>
<td>Valid</td>
<td>19863</td>
<td>7% 76% 6% 6% 6%</td>
<td>58% 42% -</td>
</tr>
<tr>
<td>Test</td>
<td>19955</td>
<td>9% 69% 9% 9% 6%</td>
<td>61% 39% -</td>
</tr>
</tbody>
</table>

Table 1 details the racial inclusion in FotW and CelebA. For FotW, the race model estimates a high variance among the subgroups, with the plurality of the data for Subgroup 2. Subgroup 1 is the smallest of the subgroups, with only 523 training data instances. For CelebA, there is less variance across races, with Subgroup 2 the majority.

6. Experiments

6.1. Analyzing Inclusion

Table 1 details the racial inclusion in FotW and CelebA. For FotW, the race model estimates a high variance among the subgroups, with the plurality of the data for Subgroup 2. Subgroup 1 is the smallest of the subgroups, with only 523 training data instances. For CelebA, there is less variance across races, with Subgroup 2 the majority.

6.2. Evaluating

Results on Faces of the World are in Table 2. We achieve an accuracy higher than that of the prior art (Rajeev Ranjan & Chellappa, 2017), using both race and gender SMILING+R+G, and set a new state-of-the-art performance of 90.96% with the model that includes a learned race representation, SMILING+R.

On FotW, our smile model SMILING+R+G reaches an accuracy 1.62% (absolute) higher than the 2016 ChaLearn Looking at People Challenge winner SIAT_MMLAB (Zhang et al., 2016), and 0.13% (absolute) higher than the published state-of-the-art UMIACS (Rajeev Ranjan & Chellappa, 2017) on the FotW validation dataset. Our model outperforms prior arts (Li et al., 2016; Učić et al., 2016; Ehrlich et al., 2016) with an increased performance gap as large as 11.66% absolute. The only point in the evaluation where transferred knowledge hurts performance is from the addition of gender for Subgroup 4 in FotW, where accuracy is less than that of the baseline SMILING. The addition of race improves over the baseline.

The Faces of the World experiments suggest that learning race subgroups may be particularly useful for inclusive models. We next turn to evaluate the effect of learning race on all 40 face attributes found in the widely used dataset CelebA. We find that the proposed system sets new record accuracies across 23 face attributes, defining the state of the art on the 10 attributes in Table 5. Race representations also further improve accuracy for individual subgroups: For Gender 1, bangs, big lips, blurry, chubby, double chin, narrow eyes, pale skin, pointy nose, wearing necklace, and young improve with transfer learned race; For Gender 2, attractive, bangs, big lips, big nose, black hair, bushy eyebrows, goatee, wearing necklace, and young all improve with race representations (see Table 3). Improvements are marginal but consistent across the board.

Among the 10 attributes with accuracy improvements with race input, big lips improved by as large as .9% absolute for Gender 1. On the other hand, five out of 40 attributes have accuracy losses with the addition of race input: oval face, brown hair, heavy makeup, high cheekbones, and wavy hair. oval face has the largest accuracy loss delta of .4% absolute.

5. Face Attribute Detection Datasets

Faces of the World (FotW). FotW (Escalera et al., 2016) is originally introduced by the ChaLearn Looking at People Smile and Gender Challenge (Escalera et al., 2016) with a goal of demographic diversity. In the dataset, each image displays a single face labeled with smile, and some with gender. See Table 1 for details. We estimate Subgroup 2 to make up half the data, with the rest as minorities.

Large-scale CelebFaces Attributes (CelebA). CelebA (Liu et al., 2015) has each image labeled with presence of 40 face attributes (see Table 5), including smiling/not-smiling, and male/not-male. We estimate Subgroup 2 to be the major-
InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity

Table 2. Accuracy (Acc.) and Average False Rate (AFR) by race and gender subgroup for smiling detection with proposed systems, on FotW. +G denotes transferred representations from the gender head of the diversity classifier, +R from the race head of the diversity classifier. The last row serves as a baseline: A model with the same number of parameters as the +G, +R, but without the transferred representations. We see improvements across minority subgroups, as well as the plurality subgroups, for both race and gender. The transferred knowledge does not impact the race ‘Other’ category, which is a racially diverse set of people whose representative racial features our diversity classifier is not trained to capture and extract. We exclude the handful of unlabeled FotW gender instances in reporting the Gender MAFR.

Table 3. Example Improved Face Attributes in CelebA Evaluated by Gender. Examples of some of the most improved attributes, evaluated by gender. Accuracy (Acc.) and average false rate (AFR) are in percentage (%). VAN refers to vanilla. +R refers to our model with race input. We see an improvement up to .9% absolute.

Table 4. Example Face Attributes with Loss in CelebA Evaluated by Gender. Examples of some of the most lost attributes, evaluated by gender. Accuracy (Acc.) and average false rate (AFR) are in percentage (%). VAN refers to vanilla. +R refers to our model with race input. We see a loss up to .4% absolute.

Table 5. Example Improved Face Attributes in CelebA. Transfer learning with race representations improves or matches face attribute detection accuracy for 35 out of 40 attributes, with marginal but consistent gains improving 13 attributes over the baseline model, and setting new state-of-the-art on 10 face attributes (presented here). Prior art compared are LA=LeNet+ANet (Zhang et al., 2013) and WL=Walk&Learn (Wang et al., 2016).

for Gender 1. This result coincides with what we expected. (S et al., 2014) says that there are four racially salient facial features: mouth, eyes, lips, and nose. And, the top three face attributes in our experiments are closely related to one or more racially salient attributes. Big Lips is closely related to the racially salient attribute mouth. Big Nose is closely related to the racially salient attribute nose. Young is closely related to the racially salient attribute eyes and mouth according to (Porcheron et al., 2017). Meanwhile, the bottom three attributes are not closely related to the racially salient attributes.

7. Discussion

Building an inclusive model and building a state-of-the-art model can go hand-in-hand. This does not have to come at the expense of predicting characteristics that may be sensitive to predict at run-time.

Transferring race knowledge increases the accuracy in almost all cases in FotW and for 13 of the CelebA attributes. Notably, learning race improves performance for smiling across both genders in FotW. Transferring gender knowledge similarly leads to accuracy and AFR improvements, although it is not as impactful as race. In FotW, the addition of gender features increases the accuracy of the gender groups as well as race subgroups 2 and 3, in addition to improving accuracy overall.

Qualitatively, the addition of race and gender representations seem to improve smiling detection on faces with extreme head poses, black-and-white photos, and more diverse facial expressions such as open mouth. And, albeit incorrect groundtruth of marking neutral faces as smiling, the addition of race and gender representations seem to correctly categorize those neutral faces as not smiling. This work establishes the feasibility of using transfer learning with demographics to improve performance across demographic categories, which we refer to as algorithmic inclusion. Future work could examine the effects of joint training and decoupled classification.

8. Acknowledgements

Special thanks to Google, Blaise Agüera y Arcas and Corinna Cortes for their support and sponsorship. Thanks to Charina Chou and Google legal team for policy and legal advice. Thanks to Marco Andreetto, Timnit Gebru, Sergio Guadarrama, Moritz Hardt, Gautam Kamath, David Karam, Shrikanth Narayanan, Caroline Pantofaru, Florian
Schroff, Krishna Somandepalli, Rahul Sukthankar and Jason Yosinski for helpful conversations. Thanks to Ben Hutchinson, Yangsun Lee, Divya Tyam and Andrew Zaldivar for donating their face photos for our research.

References


