Improving Smiling Detection with Race and Gender Diversity

Hee Jung Ryu
imiheej@gmail.com
Google Inc.

Margaret Mitchell
margarmitchell@gmail.com
Google Inc.

Hartwig Adam
hadam@google.com
Google Inc.

Abstract

Recent progress in deep learning has been accompanied by a growing concern for whether models are fair for users, with equally good performance across different demographics [26, 17]. In computer vision research, such questions are relevant to face detection and the related task of face attribute detection, among others [15].

We measure race and gender inclusion in the context of smiling detection, and introduce a method for improving smiling detection across demographic groups. Our method introduces several modifications over existing detection methods, leveraging twofold transfer learning to better model facial diversity. Results show that this technique improves accuracy against strong baselines for most demographic groups as well as overall.

Our best-performing model defines a new state-of-the-art for smiling detection, reaching 91% on the Faces of the World dataset. The accompanying multi-head diversity classifier also defines a new state-of-the-art for gender classification, reaching 93.87% on the Faces of the World dataset. This research demonstrates the utility of modeling race and gender to improve a face attribute detection task, using a twofold transfer learning framework that allows for privacy towards individuals in a target dataset.

1. Introduction

Detecting and recognizing people and faces has become an increasingly feasible machine learning task due to the rise in deep learning techniques and large people-focused datasets [3, 20, 36, 27, 24, 19, 44]. The closely related task of face attribute detection has seen similar gains, with work on everything from age estimation [25] to arched eyebrow detection [22] demonstrating significant strides in what modern machine learning can do.

However, this rise has come at a cost: As modern and public technology has incorporated modern computer vision techniques, users have noticed a troubling gap between how well some demographics are recognized compared with others [26]. For example, a woman's face with short hair and no makeup may be classified as male [31]. A black face may not be recognized at all in a face tracking task, while a face with a white mask will be [9].

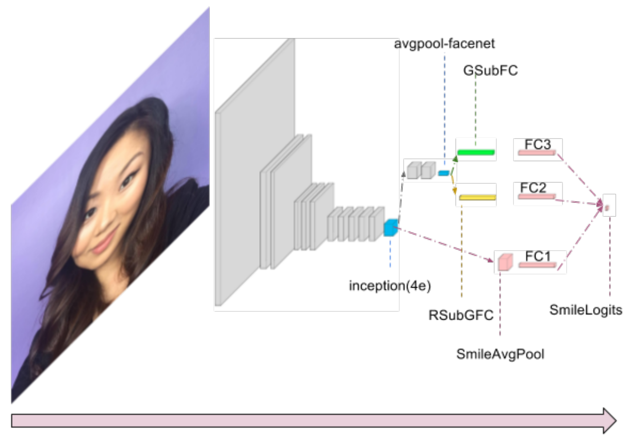

Figure 1. **End-to-end Architecture of our Smile Detector.** Gray and blue layers are part of the pretrained face recognition model, modeled after FaceNet v8 [36]. Blue layers are transferred to the diversity and smile classifiers. Green denotes the layer pretrained for gender classification and from which we transfer learn our smile classifier. Yellow denotes the layer pretrained for race classification and from which we transfer learn our smile classifier. Pink denotes layers trained from scratch for the smile classification task.

In this paper, we examine the problem of unequal performance across different races and genders in the context of smiling detection and the ChaLearn: Looking at People challenge [15]. We document disparities between demographic subgroups, working with four race subgroups and two genders, which serve as an initial set of subgroups, with room to grow.

We detail a method (see Figure 1 for details) to improve accuracy across the demographic subgroups and overall while preserving the demographic privacy of unique individuals. Our approach utilizes transfer learning, using some of the last hidden layers of a face recognition model to train a *race* classifier and a *gender* classifier (as shown in Figure 4) on held-out public data, then combining the last hidden layers from these classifiers into a third model trained for the task of smiling detection. This results in overall improved smiling detection accuracy across the board, including improvements for minority demographic subgroups. It also reduces the errors captured in false negative rate (FNR) and false positive rate (FPR) calculations, decreasing the average false rate between the two. Our contributions include:

1. We analyze the diversity across existing smiling detection datasets, focusing on the ChaLearn: Looking at People challenge [15], and documenting demographic differences.
2. We propose an approach for face attribute detection that is sensitive to race and gender underrepresentation in training data, race and gender disparities in system predictions, and user privacy.
3. We demonstrate improved performance for the *smiling* task, reaching state-of-the-art on Faces of the World.
4. We present quantitative and qualitative results on system performance across race and gender subgroups.

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 order to do better across those demographics, as well as overall, on a downstream task – in this case, smiling detection. What this research does not do is motivate race and gender identification as an end-goal, nor motivate inferring race and gender for individuals without their consent. Use of ideas in this work towards those goals are not supported by the authors. Although this work concerns race and gender, it does not provide evidence or support for inferring race and gender directly.

Race and Gender categories. The categories that we choose within race and gender are imperfect for a number of reasons, including the fact that there is no gold-standard for 'race' categories; race is often a collection of a variety of demographic information; and it is debatable whether gender should be treated as discrete categories or a spectrum. A deep-dive into the subgroup categories is intentionally absent from this work in order not to promote the idea that the subgroup categories we use are a good set to use generally.

3. Related Work

Research in computer vision has included work on issues that have direct social impact, such as security and privacy [37, 30], however, research on the related issue of diversity in computer vision is surprisingly lacking (but see, e.g., [9] for details about some of the pressing issues).

However, there are several other strands of research relevant to the work in this paper: 1) Face Attribute detection, including smile detection and gender/race classification from face images, 2) Dataset Bias, and 3) Machine learning fairness.

Face Attribute Detection. The ChaLearn "Looking at People" challenge from 2016 [15], piqued the interests and accelerated the research progress in detecting face attributes such as smile and gender. The challenge provides the Faces of the World (FotW) dataset for training and validation. [46] won the first place of the challenge, utilizing multi-task learning (MTL) and fine-tuning on top of a model trained for face recognition [27]. [28] later published an out-performing result for the same task on FotW utilizing transfer learning from a face recognition model [35] as well as MTL. The authors explore frozen parameter values rather than fine-tuning all the parameters. [42] reached a state-of-the-art performance on the smiling task on the CelebA test dataset [22]. The authors employ both transfer learning from a face recognition model similar to [36] and MTL in addition to providing additional features as inputs besides the features extracted from a face recognition model. The authors utilize the geo-location and weather information inferred from a given face as two additional input features to their MTL model.

This work explored the latent relationship between smaller attributes of the face and attributes related to identity, such as gender, finding that performance on the smaller attribute tasks could be improved by also inferring identity attributes.

Several lines of research on face detection and recognition utilize skin tone as a proxy for race [18, 43, 32, 23], however, skin tone is a variable visual feature within any race group (and across lighting conditions), and has been shown to be one of the least helpful features for distinguishing between races [34, 7, 45, 4, 39].

[42] explores inferring race from geo-location features extracted from a face. [16] explores inferring racial population geo location distribution from car maker geo location distribution. However, the application of such race classification models are limited and have not been adopted as tools to measure and improve fairness performance of a computer vision model. [38] has open-sourced the first race classification model with five races, including a subset of the US Census and Hispanic.

Bias and Diversity [40] is one of the first papers to present the dataset bias problem, documenting the issue in popular object recognition datasets and exploring its implications and effects on computer vision models. [15] points out the importance of the 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 of achieving a uniform distribution across different gender (female and male) and ethnic groups (Asian, Black, Hispanic, White).

Recent work has also applied computer vision techniques to bring into light gender bias present in movies [2]. Using a tool they introduced called GDIQ, the research team detects faces and measures the scene time by gender. The team discovered that actresses are seen 60% fewer times than actors.

Fairness in Machine Learning. The area of machine learning and fairness has recently surged, with newly uncovered theoretical guarantees and demonstrated results using modern deep learning techniques. [13] advocates for "Fairness through Awareness", demonstrating the importance of being aware of sensitive characteristics such as gender and race in order to work towards models that work well regardless of demographics. Proposals for fairness have included *parity*, such as Demographic Parity [13, 17, 5], which requires that the proportion of individuals from each demographic group matches for different decisions of the algorithm. However, as pointed out by [13], this provides no guarantee that the quality of the algorithmic decisions are equally good for each group. Building from this idea, [17] discusses measures for *equality*, such as Equality of Opportunity, which requires equal False Negative Rates across subgroups, and Equality of Odds, which additionally requires equal False Positive Rates across subgroups.

[5] applies Equality of Opportunity within a multi-task paradigm, where the main head predicts a task such as income bracket, and the auxiliary head predicts a sensitive attribute: The gradient from that prediction is then negated and passed back into the model, effectively minimizing accuracy on the sensitive attribute task while maximizing attribute on the main task. In this work, we take a somewhat opposite approach, learning race and gender, and then transferring that learning into a new model.

4. Learning Diversity

4.1. Demographic Accuracy

At the core of this work lies the idea that faces look different across different races and genders [34, 6], and that it is equally important to do well on each demographic group. To the best of our knowledge, it has not been studied whether there exists a direct relationship between smile detection and race and gender subgroup classification tasks. If such a relationship holds, then smile detection might be improved across races and genders by the model learning about them. As we will demonstrate, learning race and gender does indeed improve performance.

A classifier trained to maximize the likelihood of a given dataset will learn representations within the bounds of that dataset [40]: If the dataset exhibits racial disparity, then models trained on it will learn to encode similar disparities, with poor performance on underrepresented races and better performance on the more represented races in the dataset.

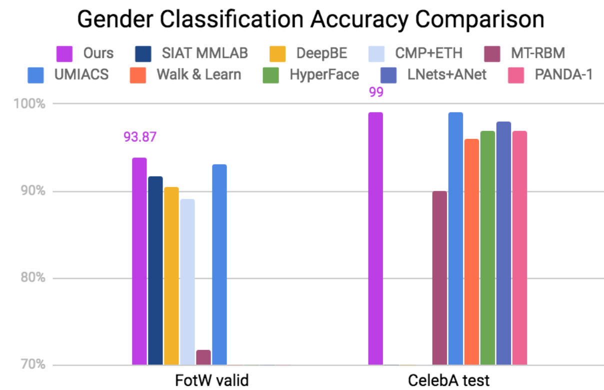

Figure 2. Our multihead diversity classifier's gender head sets a new state-of-the-art accuracy on FotW and achieves the state-of-the-art accuracy on CelebA. UMIACS [28] fine-tuned their gender model on FotW train dataset before evaluating against FotW validation dataset. For accurate performance comparison to their model, we adopted their method in this analysis. The models compared are [46, 21, 41, 14, 28, 42, 29, 22, 47] from left to right in the legend.

Poor performance may take the form of low accuracy on some subgroups, and frequent false positives or false negatives on some subgroups. These disparities may be missed when the training and validation datasets reflect a similarly skewed demographic distributions.

To evaluate systems in terms of their performance across subgroups, we report values for accuracy per subgroup. Inspired by recent work in *fairness* in machine learning (see Section 3), we also introduce a metric that evaluates false positive rates and false negatives rates in a way that is robust to label and subgroup imbalances, the *average false rate* (AFR). Given:

FP = False Positives; FN = False Negatives
TP = True Positives; TN = True Negatives
FPR = False Positive Rate = $\frac{\text{FP}}{\text{FP+TN}}$
FNR = False Negative Rate = $\frac{\text{FN}}{\text{FN+TP}}$

The Average False Rate is equal to the weighted average between the False Positive Rate and False Negative Rate:

$$\text{AFR} = \frac{(\text{FNR}n + \text{FPR}p)}{n+p} \tag{1}$$

where n and p are weights on FNR and FPR. Weighting provides the opportunity to capture different relative importance between the two kinds of error rates depending on the task the metric is used on. AFR is somewhat comparable to Error Rate, however, balances positive and negative labels so that there is no effect of label imbalance in the error calculation. In this work, we weight FPR and FNR equally ($n = p = 1$). When we calculate performance for the full dataset, in addition to reporting accuracy, we report the mean of the Average False Rate across subgroups to additionally address issues with both label and class imbalance.

4.2. Twofold Transfer Learning

Transfer learning is useful in the case where there is a feature space $\mathcal{X}$ and a marginal distribution over the feature space $P(X_S)$, where $X_S = x_1, x_2...x_n \in \mathcal{X}$, as well as a marginal distribution over the related feature space $P(X_D)$, where $X_D \in \mathcal{X}$. In this setting, $\mathcal{X}$ is the space of all face features, X_S are face images for the smiling task, and X_D are face images for the demographic classification task, race or gender.

Formally, given:

- A face domain F_R for the task of race classification T_R with labels Y_R and images X_R
- A face domain F_G for the task of gender classification T_G with labels Y_G and images X_G
- The target face domain F_S for the task of smiling detection T_S with labels Y_S and images X_S

We optimize for T_R and T_G by learning $P(Y_R|X_R)$, $P(Y_G|X_G)$ and then include those learned representations in the target domain F_S when learning $P(Y_S|X_S)$, hence the name twofold transfer learning. This has the added benefit of protecting the privacy of individuals in the smiling detection task, without directly inferring the gender and race of those individuals.

4.3. Sampling

One concern in training classifiers is dataset class imbalance, where some classes or subgroups are more prevalent than others in the training data. This negatively impacts how well a trained classifier can generalize beyond the frequent groups in the training data to groups that are in the minority. Working with object recognition and CNNs, [8] demonstrates that oversampling outperforms other popular sampling methods such as undersampling in most types of class imbalance. However, when we oversampled more than 10 folds on our face dataset with a high class imbalance in the context of race classification, we did not see any performance improvements. In order to balance our dataset across different classes for training, we instead list the eight (race $\times$ gender) combinations and undersample our dataset in a stratified sampling fashion, with up to about 25K face images per race and gender intersection.

A few minority subgroups in our dataset could not even meet the under-sample threshold. For these, we employ an off-line oversampling method in order to make sure each training batch contains faces across all race $\times$ gender. Our resulting diversity classifier sets a new state-of-the-art gender classification accuracy on FotW dataset with 93.87% and reaches the current state-of-the-art accuracy on the CelebA datset. See Figure 2 for more comparisons. Our multihead diversity classifier's gender head also performs

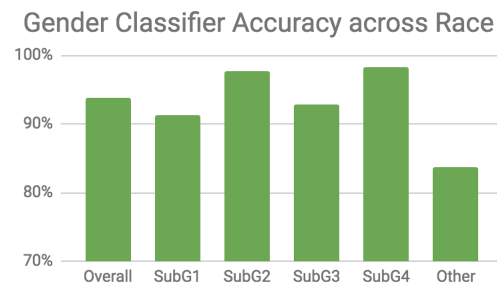

Figure 3. To understand how inclusive our multi-head diversity classifier is on gender classification, we measured its accuracy by race subgroups against FotW validation set.

relatively equal across different race subgroups in the FotW dataset, as reported in Figure 3.

4.4. Race and Gender Architectures

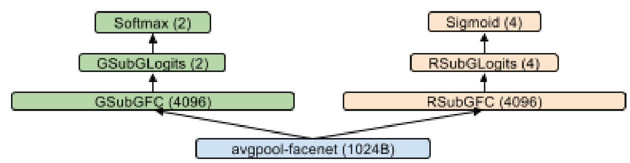

Figure 4. Multihead (gender and race) diversity classifier is used to provide features that improve the task of smiling detection across different races.

Figure 4 provides a high-level overview of the architecture used for learning in the demographic domains, F_R and F_G. First, a face detector (similar to Picasa [10]) is run to crop and align faces. Using these face crops, we train a face recognition system modeled after the FaceNet v8 architecture [36]. Following [11], we extract features from the layer called *avgpool*, which we refer to as *avgpool-facenet* in the rest of this paper. We add fully connected layers and then fine-tune the model for the race and gender estimation tasks.

To train the multihead (race and gender) diversity classifier, we use data collected from publicly available resources of celebrities appearing in [1, 33, 38], and further curate the data so that train and validation sets consist of disjoint sets of identities.

Unlike the traditional approach to labeling race and gender, we obscure the race and gender subgroup names whenever possible[1] in order to avoid introducing unconscious, societal bias associated with existing labels. We call gender *Gender 1* and *Gender 2*, with an eye towards adding more gender subgroups in future work. We call the races *Subgroup 1, Subgroup 2, Subgroup 3, Subgroup 4*, and so on, and similarly anticipate using this naming scheme to extend to further race subgroups.

[1]FotW is annotated with 'Female', 'Male', and 'Not Sure'; CelebA has 'Male' and 'Non-Male'; we adopt these labels for direct comparisons with previous work on these datasets.

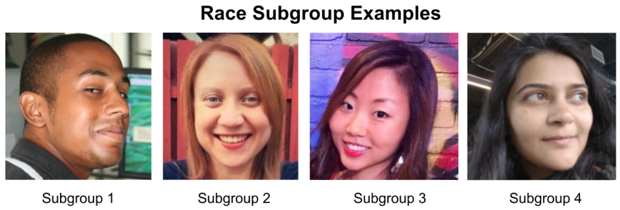

Figure 5. Example subgroups used in race classification.

Gender Classification. Our model's performance on gender classification in the FotW [15] and CelebA [22] datasets are 93.87% and 99% respectively, setting the new state-of-the-art gender classification accuracy. Figure 2 illustrates the model's performance compared to other existing models. Additional experiments demonstrate that gender classification works surprisingly well with only the top part of the face, also known as the "attention region" (eyes and nose) [34]. The trained model achieves accuracy that only slightly underperforms the experiments with full faces.

Race Classification. We explore four race subgroups for which we were able to scrape over 100,000 web images of famous identities listed in [1, 33, 38]. This allows us to train a reasonably performing race head of a multihead (race and gender) diversity classifier, however, this is a limited set of all the races that might be considered.

With the initial classifier, we can demonstrate the potential usefulness of learning race when detecting face attributes. We train our model to reach 98% or greater AUC across the four race subgroups.

4.5. Smile Detector Architecture

Figure 1 provides a high-level overview of the architecture used for learning to detect smiles in the smiling domain F_G, building from the learned representations F_R and F_G. Figure 6 is a close-up on the part of the model architecture trained from scratch and outputs smiling detection results.

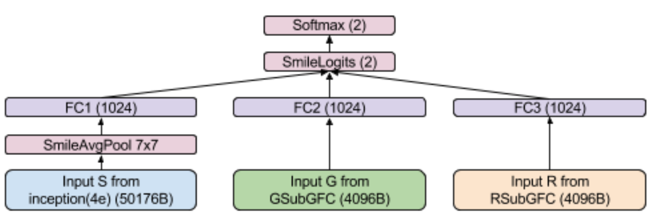

Figure 6. Smile Detector Architecture (Close-up).

Input Features and Twofold Transfer Learning. We transfer learn our final smile detection model with three sets of features extracted from two different pretrained models: a face recognition model and a multihead (race/gender) diversity classifier.

Bottleneck Layers. We use AvgPool of dimension 7x7 filter and stride 1 with 'valid' padding and smaller FCs as our bottleneck layers to generalize our rich input features of total size 58,368 bytes (= S 50,176B + R 4,096B + G 4,096B) down to total 3,072 bytes (= FC1 1,024B + FC2 1,024B + FC3 1,024B). Details of all architectures explored are presented in Table 2.

	S1	S2	S3	S4	Other	Total
avgpool-facenet	85.7	88.9	89.4	92.0	88.4	89.04
inception(4e)	86.1	90.4	90.9	92.4	87.6	**89.82**

Table 1. Accuracy (in %) by different input features according to the layers detailed for FaceNet v8 [36], Subgroups (S) 1-4. (Train: FotW, Eval: FotW, 1FC(4096))

Learning Rate. Learning rate plays an important role in our small neural net for it to avoid getting stuck in a local minima before having enough exposure to learn from and fine-tune on its teacher model(s). It is also not ideal for the model to overfit to the training dataset due to the learning rate being under a certain threshold. In our experiments, we adjusted the initial learning rate through brute force search, with an aim to reduce the oscillation of the continuous training accuracy- and loss- curves. We found that initial learning rates between the range of 0.005 and 0.001 (inclusive) work the best for our experiments, with an early stopping between 60K and 100K training steps when trained with a batch size of 64.

Number of Parameters Trained. $1024 \times 1024 + 4096 \times 1024 \times 2 + 1024 \times 3 \times 2 = 944{,}332$ weight parameters are trained from scratch. Comparison experiments with the same number of parameters but without the transfer learning demonstrate that the transfer learning adds additional benefit to the final model.

Optimization and Loss. We use softmax along with a cross-entropy loss for smiling, following the state-of-the-art smiling model [28]. For optimization, we use 'AdaGrad'[12].

	Accuracy
(a) 4096	89.82%
(b) 4096, 4096	89.72%
(c) 1024, 1024	89.85%
(d) MaxPool, Conv, 2048, 512	90.01%
(e) MaxPool7x7x1(V), 1024, 1024	90.47%
(f) AvgPool7x7x1(V), 1024, 1024	**90.57%**

Table 2. Architecture experiment (Train: FotW, Eval: FotW, Input: Inception(4e))

5. Smile Detection Datasets

5.1. Looking at People, Smile and Gender Challenge

The ChaLearn Looking at People[2] Smile and Gender Challenge [15] requires classifying images according to

[2] http://chalearnlap.cvc.uab.es/challenge/13/track/20/description/

		Total	Estimated Race					Gender		
			S1	**S2**	**S3**	**S4**	**Other**	**F**	**M**	**Other**
FotW	**Train**	6171	8%	48%	11%	15%	19%	52%	46%	2%
	Valid	3086	8%	53%	11%	17%	11%	55%	44%	1%
CelebA	**Train**	162687	7%	75%	7%	8%	4%	0*	42%	58%
	Valid	19863	7%	76%	6%	6%	6%	0*	42%	58%
	Test	19955	9%	69%	9%	9%	6%	0*	39%	61%

Table 3. **Estimated Race and Groundtruth Gender Distribution across Datasets.** Duplicated images in CelebA are counted only once in this dataset analysis. **FotW comes with 'Female' (F) and 'Male' (M) labels. CelebA comes with 'Male' (M) and 'Non-Male' (Other).*

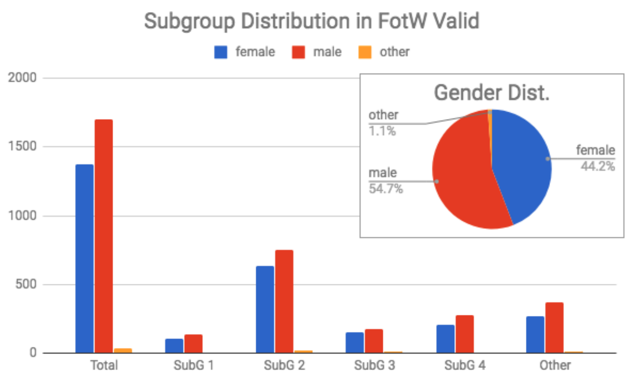

Figure 7. Diversity in FotW validation set measured by image count.

gender and smiling. Models may be trained on the provided FotW and CelebA datasets [22], as well as pretraining from other sources.

In the Faces of the World (FotW) dataset, each image displays a single face labeled with smile, and some with gender. The train set has 2,234 smile and 3,937 no smile, and the validation set has 1,969 smile and 1,117 no smile. See Table 3 for more details across genders and races. We estimate that Subgroup 2 makes up about half the data, with the rest as minorities.

The Large-scale CelebFaces Attributes (CelebA) data contains 162,770 images with duplicates and 162,687 unique images in the train set; 19,867 images with duplicates and 19,863 unique images in the validation set; and 19,962 images with duplicates and 19,863 unique images in the test set. Each image is labeled with smiling/not-smiling and male/not-male. We estimate Subgroup 2 to be the vast majority here, with relatively little data for other race subgroups. See Table 3 for more details.

6. Experiments

6.1. Analyzing Inclusion

Figure 7 and Table 3 detail the racial inclusion in the Faces of the World and CelebA datasets used in the LAP challenge. For Faces of the World, 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. This is contrary to the reported racial diversity of FotW in [15], and may be due to the fact that these are estimated labels from our model, as well as due to differences in how different parties will categorize racial subgroups. As we will demonstrate, agreement on race categories is not needed to improve performance.

For CelebA, there is less variance across subgroups, with Subgroup 2 being the biggest.

6.2. Learning Smiling from Race and Gender

To improve our models for the smiling task, we first examine the effect of different layers as input to the model. We compare *avgpool-facenet* and *inception(4e)* in Table 1, where the 'Other' label is applied when the confidence score is < 0.5 for all subgroups.

Inception(4e) yields the best results across all racial subgroups, boosting the overall accuracy by +0.78% and the accuracy of the minority Subgroup 1 by +0.41% on a simple fully connected layer of output size 4096. In the rest of the experiments, we use features extracted from the inception(4e) layer as input to our smile detection system. Considering how avgpool-facenet helps with our multihead (gender and race) diversity classifier and how the layer comes after various bottleneck layers, we conjecture that avgpool-facenet has a richer set of features, more representative of facial identities than inception(4e). Meanwhile, inception(4e) may have a richer set of features representative of facial muscle movements like smiling. This layer may encode other features (possibly lighting or head poses) irrelevant to smiling such that we see an accuracy improvement of +0.72% after introducing a bottleneck layer similar to *avgpool-facenet* into our system (See (c) and (f) in Table 2 for details).

We then compare MaxPool, Conv, AvgPool with two fully connected (FC) layers, comparing different FC layers with the inception(4e) layer. Moving from one fully connected layer to two does not improve performance, and changing the size of the layer does not lead to significant gains. However, adding a pooling layer leads to additional gains. In Table 2, (d) was inspired by [28] and (f), by [42]. Since (f) with AvgPool worked the best, we experiment with another popular pooling layer MaxPool. The final baseline architecture that we adopt is (f) in Table 2, achieving 90.47% accuracy on the smiling detection task. From here

Train: FotW Eval: FotW	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		Other		Total	
	ACC.%	AFR%	ACC.%	AFR%	ACC.%	AFR%	ACC.%	AFR%	ACC.%	AFR%	ACC.%	MAFR%
SMILING	88.5	12.2	90.9	10.7	90.0	12.2	94.3	6.2	88.2	18.4	90.57	12.0
+G	88.5	**12.1**	91.2	10.1	90.9	10.9	94.1	6.1	88.2	18.4	90.76	11.5
+R	88.5	**12.1**	**91.6**	**9.7**	90.6	11.3	**94.5**	5.9	88.2	18.4	**90.96**	**11.5**
+R +G	88.5	**12.1**	91.4	9.9	90.3	11.8	**94.5**	5.9	88.2	18.4	90.86	11.6
on 1FC(2048)	88.1	12.5	91.3	10.2	90.3	11.8	**94.5**	5.9	88.2	18.7	90.80	11.8
on 1FC(3072)	87.7	13.0	91.4	10.3	89.8	12.4	94.1	6.5	87.9	19.6	90.67	12.3
on 2FCs(930,4608)	87.7	12.8	90.8	10.1	91.2	10.2	94.3	**5.8**	88.1	18.4	90.57	**11.5**

Table 4. Accuracy (ACC.), Average False Rate (AFR), and mean AFR (MAFR) by race subgroup for smiling detection with proposed systems, on FotW dataset. +G denotes transferred representations from the gender head of the diversity classifier, +R from the race head of the diversity classifier, and +R+G from both. The last three FC rows represent models with the same number of parameters as the +G, +R, +R +G models, but without the transferred representations.

Train: FotW Eval: FotW	Gender 1		Gender 2	
	ACC.%	AFR%	ACC.%	AFR%
SMILING	90.2	10.6	91.0	12.5
+G	90.3	10.4	91.3	11.6
+R	**90.7**	**10.0**	91.3	11.8
+R +G	90.4	10.5	**91.4**	**11.5**

Table 5. Accuracy (Acc.) and Average False Rate (AFR) by gender for smiling detection with proposed systems, FotW dataset.

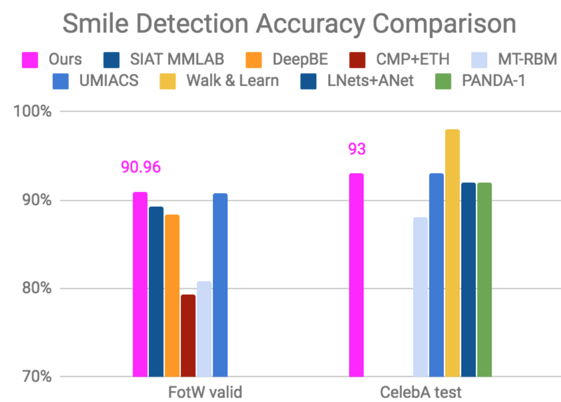

Figure 8. Our smile classifier sets a new state-of-the-art accuracy on FotW. On CelebA, it is only behind by [42]. The models compared are [46, 21, 41, 14, 28, 42, 22, 47] from left to right in the legend.

on, we call this baseline model S or SMILING (See Table 4).

As discussed earlier in Section 3, studies show a latent relationship between physical appearance of lips and sexual and/or racial identity. From here, we postulate that the appearance of smiling may likely differ by race and gender and affects a smiling model's performance. However, we are often faced with face dataset lack of sexual and racial diversity. To mitigate this issue, we train a multi-head diversity classifier that classifies gender and race (detailed in Section 4.4) and augment our baseline model S with additional input features transfer learned from each head (race and gender) of the diversity classifier (architecture found at Figure 4). From here, we denote input features extracted from GSubGFC in Figure 4, G and those extracted from RSubGFC in Figure 4, R. We experiment with a few combinations of the input features on each of the datasets, included in the results below.

6.3. Evaluating

First, the face detector is run on the provided CelebA or FotW thumbnails. If it fails to align the face (this happens for 0.65% of CelebA test set faces and 12.57% of FotW validation set faces), the provided CelebA or FotW face crops and alignments are used.

In addition to accuracy for smiling detection, we measure the Average False Rate (AFR) across race subgroups (see Section 4.1). Intuitively, this measures how much the model fails to predict smiling across smiling instances for each race subgroup, and how much the model over-predicts smiling across non-smiling instances for each race subgroup.

Results are shown in Table 4. We achieve an accuracy higher than that of the previous state-of-the-art smile classifier [28], 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.

To test whether the transfer learning has transferred representations that help for this task, we compare to a models with the same number of parameters but without the transferred learning. See rows 1FC(2048), 1FC(3072), 2FCs(930,4608) in the lower half of Table 4. These correspond to 1 fully-connected layer with 2048 dimensions, 3072 dimensions, and 2 fully-connected layers with 930 and 4608 dimensions. We find that the models with the transfer achieve higher overall accuracy, although, interestingly, the same number of parameters lead to reduced AFR in some cases, and the highest accuracy for Subgroup 3, outperform-

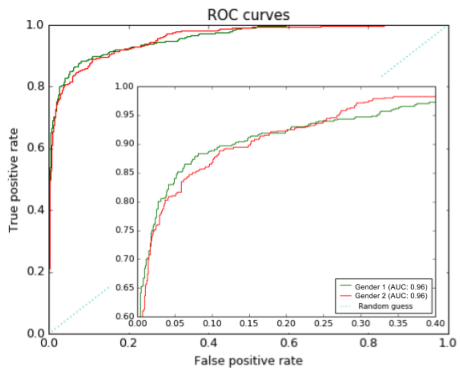

Figure 9. Smile Detector Performance by Gender. For both genders, AUCs are equally 0.96.

ing the next best performance achieved by adding gender. The reason for this is not entirely clear, and we look towards ethical approaches for a deeper dive into differences between subgroups in future work.

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 [46], and 0.13% (absolute) higher than the published state-of-the-art UMIACS [28] on the FotW validation dataset. Both of the models used transfer learning from a face recognition model and MTL. SIAT_MMLAB exploited fine-tuning the entire face recognition model with a couple of iterations of additional pre-training/fine-tuning using CelebA dataset and then finally fine-tuning using FotW. UMIACS takes a similar approach to ours in that the model is transfer learned from features extracted from a multiple mid-/low-layer of a face recognition model. They also pretrained with CelebA and fine-tuned on FotW. Our best performing model is trained on FotW train dataset on top of a face recognition model, skipping the intermediate step of fine-tuning CelebA dataset. Our model outperforms other models published [21, 41, 14] with an increased performance gap as large as 11.66% absolute. Figure 8 presents an accuracy comparison graph.

We also plot ROCs with respect to smiling by gender in Figure 9. AUCs of our model per gender are equally 0.96, demonstrating Equality of AUC similar to the fairness metrics of [17].

7. Discussion

This paper is one of the first to demonstrate that building an inclusive vision model and building a state-of-the-art model can harmoniously go hand-in-hand. Moreover, this vision diversity research result gives us more insight into understanding our vision models, the interaction between our vision models, and visually distinctive features on our faces. We present a more in-depth analysis of race/gender intersectionality in the Appendix.

Transferring race knowledge increases the accuracy for subgroups 2, 3, and 4, and improves total accuracy to the state of the art. It also decreases the average false rate for the four race subgroups and two gender subgroups, as well as overall. Notably, learning race improves performance across both genders.

Transferring gender knowledge similarly leads to accuracy and false rate improvements, although it is not as impactful as race. 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. The only point in the evaluation where transferred knowledge *hurts* performance is from the addition of gender for Subgroup 4, where accuracy is less than that of the baseline SMILING.

Adding both race and gender does not significantly improve over adding race alone for all race subgroups and overall, however, it does improve performance on Gender 2 beyond improvements provided by race or gender alone. The limited improvements provided by the model +R+G in Tables 4 and 5 may be due in part to the fact that with this model comes with more parameters, which may be more difficult to tune with the relatively small amount of data provided by the FotW dataset.

We did not see any accuracy change for subgroups with a dataset size much smaller than a hypothetical threshold. For example, we did not see any performance changes to race Subgroup 1 (denoted as SubG1). We attribute this to the data for that subgroup being too small compared to the rest, as shown in Figure 7. This is similar to the class imbalance problem discussed in Section 4.3. From this, we make a soft conclusion that our methodology's effectiveness is limited when the size of a minority demographic subgroup dataset does not meet some minimum threshold. This gives a strong motivation for collecting a more diverse dataset in order to achieve an even greater model performance boost.

Despite our diversity classifier's state-of-the-art performance on gender classification (see Figure 3 and 2), the model S+G fine-tuned on the gender-head of the diversity classifier shows some accuracy drop in *Subgroup 4* in Table 4. Looking at Figure 7, we see a good gender parity in the dataset. This may explain why the additional two-gender input features do not help much: Two genders are already well represented in the training data.

It is worth noting that Subgroup 4 has the highest accuracy compared to the other subgroups, and is the only subgroup that results in a decrease in accuracy when gender representations are added. Perhaps the information transferred for gender is already better captured in the baseline smiling model for this subgroup, although it's not clear why the model performs particularly well on this group. Also noteworthy is the fact that Subgroup 3 achieves the best performance when two fully connected layers are added –

Figure 10. **Baseline vs. Models with Extra Input Features.** Left column shows examples the baseline model got wrong that the transfer learned models corrected. Right column shows examples the baseline model got right, and the transfer learned models got wrong.

without the transferred knowledge. We note these differences for future work that dives into details of each race subgroup.

Throughout our experiments, the transferred knowledge does not impact the 'Other' category, which is a diverse set of people from many different backgrounds. This is interesting because while we do see a positive effect of transfer learning for specific race and gender subgroups, this knowledge does not transfer to the other potential subgroups outside of the four represented. Simply adding more parameters in 2 fully connected layers decreases the average false rate very slightly, but also decreases accuracy significantly. We think this could be attributed to the fact that our diversity classifier's race head was not trained to capture and extract representative features of the Other group here.

Finally, Figure 10 shows examples of the images in which additional gender and/or race feature inputs made differences. The faces in the examples tend to have a thin smile, a neutral face, or a mouth open. With the additional input features, we see more correct smile detections in black-and-white photos and in faces turned away from the camera.

In the future, we intend to explore methods to boost the lowest demographic accuracy and overall accuracy even when the disparity between demographic subgroups, like gender, is as small as that of this dataset. In addition, we also see a possibility that the performance boost by gender features could have been better measured if we measured the performances by the estimated gender slices rather than the groundtruth gender slices.

We acknowledge that our efforts taking an unbiased look at diversity and fairness in our models has its limitations. In future work, we aim to add more race and gender groups, and obscure specific race and gender groundtruth labels from the start. For example, by using unsupervised or weakly supervised learning, such as clustering, we may be able to create coherent groups that pattern along demographic lines, without the need for explicit labels.

8. Conclusion

We have demonstrated the problem of race and gender differences in smiling detection, and have proposed a method for improving accuracy across races and genders. The model utilizes twofold transfer learning for better representations of the race and gender subgroups that are present in faces, and reaches a new state-of-the-art accuracy for smiling detection on the Faces of the World dataset. In this process, we also reach a new state-of-the-art accuracy on gender classification using sampling to balance demographic labels and classes (subgroups).

When we use only FotW and transferred racial representations, the smiling detection model is improved to a new state-of-the-art, with 90.96% accuracy. This is 1.62% absolute improvement over the previous ChaLearn winner [46], and 0.13% absolute improvement over the previous state of the art [28]. By incorporating the learned representations of both gender and race, Gender 2 further improves.

In future work, we plan to address the fact that our gender and race subgroup categories are limited and do not represent the real world to its entirety.

9. Acknowledgements

Special thanks to Blaise Agüera y Arcas and Corinna Cortes for their supports and sponsorship. Thanks to Caroline Pantofaru, Sergio Guadarrama, Rahul Sukthankar, Florian Schroff, Marco Andreetto, David Karam, Krishna Somandepalli, Shrikanth Narayanan, and Moritz Hardt for helpful conversations. Thanks to Divya Tyam, Andrew Zaldivar, and Ben Hutchinson for donating their face photos for our research.

References

[1] Wikidata. `http://www.wikidata.org/`.

[2] H. Adam, I. Breckinridge-Jackson, R. Cooper, T. Guha, I. Hall, C. wei Huang, N. Kumar, N. Malandrakis, M. Nasir, A. Ramakrishna, K. Somandepalli, and Y. Zhu. The reel truth: Women arent seen or heard: An automated analysis of gender representation in popular films. `https://seejane.org/wp-content/uploads/gdiq-reel-truth-women-arent-seen-or-heard-automated-analysis.pdf`, 2016.

[3] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] G. Anzures, O. Pascalis, P. C. Quinn, A. M. Slater, and K. Lee. Minimizing skin color differences does not eliminate the own- race recognition advantage in infants. *Infancy*, page 640654, 2011.

[5] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[6] U. Bindal, P. G. Bindal, and N. A. bt Raml. Labial impressions: A tool for identification. *Journal of Forensic Research*, 2014.

[7] K. R. Brooks and O. S. Gwinn. No role for lightness in the per- ception of black and white? simultaneous contrast affects per- ceived skin tone, but not perceived race. *Perception*, 39:11421145, 2010.

[8] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, abs/1710.05381, 2017.

[9] J. Buolamwini. How i'm fighting bias in algorithms. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-63664.

[10] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, 2014.

[11] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Face synthesis from facial identity features. *CoRR*, abs/1701.04851, 2017.

[12] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

[14] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[15] S. Escalera, X. Baro, H. J. Escalante, and I. Guyon. Chalearn looking at people: A review of events and resources. In *Proceedings of the The 2017 International Joint Conference on Neural Networks (IJCNN 2017)*. IEEE, 2017.

[16] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. A. Lieberman, and L. Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of the US. *CoRR*, abs/1702.06683, 2017.

[17] M. Hardt, E. Price, , and N. Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

[18] H. Hill, V. Bruce, and S. Akamatsu. Perceiving the sex and race of faces: The role of shape and color. In *the Royal Society B: Biological Sciences*, volume 261, page 367373, 1995.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[20] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[21] C. Li, Q. Kang, and Q. Ge. Deepbe: Learning deep binary encoding for multi-label classification. In *Computer Vision and Pattern Recognition Workshops CVPRW, 2016 IEEE Conference on*. IEEE.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[23] C. R. Malskies, E. Eibenberger, and E. Angelopoulou. The recognition of ethnic groups based on histological skin properties. In *the Vision, Modeling, and Visualization Workshop*, page 353360, 2011.

[24] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] E. O. of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[28] C. D. C. Rajeev Ranjan, Swami Sankaranarayanan and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition*. IEEE.

[29] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *CoRR*, abs/1603.01249, 2016.

[30] N. Raval, A. Machanavajjhala, and L. P. Cox. Protecting visual secrets using adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017*, pages 1329–1332, 2017.

[31] K. Rodden. Is that a boy or a girl? exploring a neural networks construction of gender. https://medium.com/\spacefactor\@m{}kerryrodden/is-that-a-boy-or-a-girl-cb93abbae6da.

[32] S. M. M. Roomi, S. Virasundarii, S. Selvamegala, S. Jeevanand-ham, and D. Hariharasudhan. Race classification based on facial features. In *NCVPRIPG*, page 5457, 2011.

[33] R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[34] F. S, H. H, and H. ZG. Learning race from face: A survey. *IEEE Trans Pattern Anal Mach Intell.*, 36:2483–509, Dec 2014.

[35] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. *CoRR*, abs/1604.05417, 2016.

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[37] J. Sokolic, Q. Qiu, M. R. D. Rodrigues, and G. Sapiro. Learning to identify while failing to discriminate. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[38] K. Somandepalli. Prediction race from face for movie data. `https://github.com/usc-sail/mica-race-from-face/wiki`, 2017.

[39] M. A. Strom, L. A. Zebrowitz, S. Zhang, P. M. Bronstad, and H. K. Lee. Skin and bones: The contribution of skin tone and facial structure to racial prototypicality ratings. *PLoS ONE*, 7:18, 2012.

[40] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[41] M. Uřičář, R. Timofte, R. Rothe, J. Matas, and L. V. Gool. Structured output SVM prediction of apparent age, gender and smile from deep features. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition Workshops*, Las Vegas, USA, 2016.

[42] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, 2016.

[43] Y. Xie, K. Luu, and M. Savvides. A robust approach to facial ethnicity classification on large scale face databases. In *BTAS*, page 143 149, 2012.

[44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[45] L. Yin, J. Loi, and J. Jia. Towards race-related face identification: Research on skin color transfer. In *IEEE Int. Conf. Autom. Face Gesture Recog*, page 362368, 2004.

[46] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–38, 2016.

[47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. D. Bourdev. PANDA: pose aligned networks for deep attribute modeling. *CoRR*, abs/1311.5591, 2013.

10. Appendix

Evaluation Analysis at Gender and Race Subgroup Intersections, Accuracy (See Figure 11). Adding Race (+R) increases accuracy over the baseline in all evaluations. Adding gender on top of race (+R +G) can lead to further gains. In isolation, adding the gender representation (+G) helps for two of the four tested subgroups. For subgroup 4, accuracy is decreased.

Statistical Significance (See Table 6) Differences between the baseline S and the twofold transfer learned models S+R, S+G, and S+R+G are statistically significant. Difference between S+R and S+R+G is not significant. We set the significance level α with Bonferroni correction for α=.05. With correction for the 6 tests we run, the Bonferroni-corrected α is .008. At this level, S+G and S+R are not significantly different, although they are significant at the uncorrected .05 significance level. Results demonstrate that transfer learned race and gender representations help for improving smiling detection across races, with race leading to the highest overall accuracy.

	S		**S+G**		**S+R**		**S+R+G**	
	p-value	χ^2	*p-value*	χ^2	*p-value*	χ^2	*p-value*	χ^2
S	-	-	$5.7e^{-7}$	25.0	$1.0e^{-4}$	15.1	$4.0e^{-4}$	12.4
S+G	$5.7e^{-7}$	25.0	-	-	$3.9e^{-2}$*	2.0*	$9.8e^{-4}$*	0.0*
S+R	$1.0e^{-4}$	15.1	$3.9e^{-2}$*	2.0*	-	-	0.38*	1.0*
S+R+G	$5.7e^{-7}$	25.0	$9.8e^{-4}$*	0.0*	0.38*	1.0*	-	-

Table 6. **McNemar Test.** (*) indicates P-value from an exact binomial test and χ^2 from McNemar's test with continuity correction.

	S	S+G	S+R	S+R+G
	Accuracy	Accuracy	Accuracy	Accuracy
SubG1 & gender unknown	75.0	75.0	75.0	75.0
SubG1 & female	88.8	88.8	88.8	88.8
SubG1 & male	88.7	88.7	88.7	88.7
SubG2 & gender unknown	92.9	92.9	92.9	92.9
SubG2 & female	91.1	91.2	**91.8**	91.4
SubG2 & male	90.8	91.3	91.3	**91.4**
SubG3 & gender unknown	83.3	83.3	83.3	83.3
SubG3 & female	90.7	**91.4**	**91.4**	90.7
SubG3 & male	89.6	**90.8**	90.2	90.2
SubG4 & gender unknown	80.0	80.0	80.0	80.0
SubG4 & female	94.2	93.7	94.2	94.2
SubG4 & male	94.6	94.6	**95.0**	**95.0**
Other & gender unknown	83.3	83.3	83.3	83.3
Other & female	85.9	85.9	85.9	85.9
Other & male	90.0	90.0	90.0	90.0
Overall	90.57	90.76	**90.96**	90.86
SubG1	88.5	88.5	88.5	88.5
SubG2	90.9	91.2	**91.6**	91.4
SubG3	90.0	**90.9**	90.6	90.3
SubG4	94.3	94.1	**94.5**	**94.5**
Other	88.2	88.2	88.2	88.2
gender unknown	85.3	82.4	85.3	85.3
female	90.2	90.3	**90.7**	90.4
male	91.0	91.3	91.3	**91.4**

Figure 11. Green rows indicate accuracy with an upward boost from the baseline, **S**, using the transfer learned representations. The best-performing model is in bold.