CIS 522 Final Report

Mingyung Kim Sophie Trotto Qingyun Zeng Yiliang Zhang

May 2020

1 Introduction

Deep learning has infiltrated into various fields with countless applications. However, many of these tasks are expected to be operated with awareness of fairness — lawfully and without discrimination. Consider an example: an HR employee is training a deep neural network to decide whether or not an applicant should be hired. They need be cautious to make sure the decision should not be related to gender and race. As a result, the training process of the neural network involves a trade-off between fairness (non-discrimination) and quality of decision.

In general, having class balance across sensitive groups is critical in the creation of a fair algorithm. In the case of deep learning, one answer to this "trade-off" is to train a neural network while preserving group fairness is by adding a penalty term in loss function, which regularizes the fairness of the algorithms. A more extensive list of different metrics and bias mitigation algorithms has been compiled by IBM in their AI Fairness 360 toolkit [1].

Automatic facial recognition algorithms, which are used by developers including social media companies and federal law enforcement, have been shown to be biased with respect to race/ethnicity, gender, and age. A NIST study evaluated 189 software algorithms from 99 developers and found 10-100 times higher rates of false classification for (East) Asian and Black faces relative to Caucasian ones. Examining only algorithms created in the U.S., the highest rate of misclassification of ethnicity was of Native Americans. Women were also less likely to be correctly identified than men, and older adults were misclassified up to 10 times more than the middle-aged [2]. Clearly, there is significant room for improvement when it comes to eliminating bias from automatic facial recognition algorithms.

2 Related Work

2.1 FairFace paper [3]

Findings: Shows that most of the current and commonly-used face datasets are biased toward white faces and against faces of color. In order to address this racial bias and to provide more consistent training accuracy across racial/ethnic groups, the authors construct a novel large-scale face dataset with class balance among seven ethnic groups, called FairFace.

Relevance: Our project utilizes this FairFace dataset for training and validation data.

2.2 Review on The Effects of Age, Gender, and Race Demographics on Automatic Face Recognition [4]

Findings: Recognition accuracy of automatic face recognition models differs across gender, race, and age, and more strongly across interactions of these demographics. In particular, recognition accuracies of males and older people are greater than those of females and younger people.

Relevance: This paper identifies the current biases with respect to face recognition models, and what the sensitive groups of interest are.

2.3 Deep Learning for Face Recognition: A Critical Analysis [5]

Findings: Contains a review of different methodologies (models, loss function, number of neural networks) for face recognition, as well as available databases. Notes, importantly, that the best models have extremely high computational cost and require large databases to provide good results.

Relevance: This paper identifies which deep learning face recognition models perform the best, which will be helpful for our group when it comes to designing our own model. It also had information on the effectiveness of the IMDB-Faces dataset, as well as the UTKFaces dataset, which overlaps significantly with FairFace.

2.4 Fairness Criteria for Face Recognition Applications [6]

Findings: Found that CNN classifier was biased against non-white individuals for gender classification despite passing standard fairness metrics. Adding confidence criteria and identifying a minimal test sample size can improve fairness.

Relevance: This paper is extremely relevant to our project, as it deals with the exact same subject: model fairness with respect to face recognition. It discusses some ways to improve fairness, which we can potentially use in our model design.

2.5 Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification [7]

Findings: While commercially-used gender-classification deep learning tools (i.e. IBM, Microsoft, and Face++) provided high classification performance on average, their error rates are significantly different across groups. Notably all of these tools had the worst performance when it came to classifying images of darker-skinned females over other combinations of skin color and ethnicity.

Relevance: This is one of the key papers that study group fairness in computer vision. Our project may contribute to this research, by proposing a way to improve group fairness in facial recognition and classification.

2.6 Fairness-aware Learning through Regularization Approach [8]

Findings: Identifies three causes of unfairness in machine learning (specifically, regression and classification): prejudice, underestimation, and negative legacy. All of these causes are studied by building corresponding models on a class Y, a sensitive feature X, and a non-sensitive feature S. The author also proposed a method to remove unfairness caused by indirect prejudice, which uses regularization.

Relevance: This is an theory-inclined paper which identifies causes of unfairness and the tools to identified them. We can adapt its method directly into group fairness problem into deep learning, though it is not clear whether this might be helpful for individual fairness problems we proposed such as facial recognition. Also, the idea that how to convert definition of unfairness into statistical model and the construction of metric on quantifying the unfairness is helpful.

2.7 Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations [9]

Findings: In training classifiers with data that are full of ambiguities, we often use regularization to narrow down choices and encourage robustness in order to get high accuracy. We usually also employ domain-specific knowledge to assist our model. However, these methods may not generalize well due to confounding in the data set, leading to a "right model for the wrong reasons." The author built methods which can effectively explain and regularize differentiable models by using techniques that selectively penalize input gradient.

Relevance: This is a more application-inclined paper, aiming for a "right model for the right reasons." This paper is useful for us since its technique of regularizing on specific attributes is aligned with one of our novel methods.

3 Data

3.1 FairFace

The FairFace dataset contains 108,501 face images collected primarily from the YFCC-100M Flickr dataset. Attributes include age, gender, and race, and is notably balanced in distribution across the seven racial/ethnic groups represented in the dataset (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino) [3]. There are no names or ids in the dataset, so it cannot be used to judge how accurately a facial recognition model can identify a person based on an image.

Five images from the FairFace training dataset.

Table 1: Attributes of the above images.

file	age	gender	race
train/1.jpg	50-59	Male	East Asian
train/2.jpg	30-39	Female	Indian
train/3.jpg	3-9	Female	Black
train/4.jpg	20-29	Female	Indian
train/5.jpg	20-29	Female	Indian

3.2 IMDb-Faces

The IMDb-Faces dataset contains 460,723 face images from 20,284 celebrities from IMDb. Attributes include gender, birthdate, name, and year when the photo was taken [10]. Notably, there is no information on race or ethnicity, which poses an issue for algorithm fairness: many face recognition models are unfair with respect to race/ethnicity because there is significant class imbalance in the dataset. More specifically, there are often many more white subjects in the dataset than subjects of other races/ethnicities — which is true of the IMDb-Faces dataset [5]. Because we do not have label information for race or ethnicity, we cannot correct this imbalance. Furthermore, since we do not have ethnicity information, we cannot use it as a feature in our training or testing.

As a solution, we webscraped ethnicity information for celebrities from ethniccelebs.com, and cleaned these longer ethnic data into one of the seven ethnic categories from FairFace. Our cleaned IMDb dataset, with ethnicity, contains 347,701 face images from 6,224 celebrities.

Five images from the IMDb-Faces dataset.

Table 2: Cleaned attributes of the above images.

celeb_id	name	gender	ethnicity	full_path
2	Weird Al Yankovic	Male	White	Weird_Al_Yankovic_0.jpg
4	50 Cent	Male	Black	50_Cent_0.jpg
5	A Martinez	Male	Latino	A_Martinez_0.jpg
8	A.J. Cook	Female	White	A.J._Cook_0.jpg
11	A.J. McLean	Male	Latino	A.J._McLean_0.jpg

Interestingly, two of the faces (1 and 2) clearly do not match the celebrity name label. Likely, these are costars of the labelled celebrity whose faces were also tagged in their IMDb images. However, it raises the issue of inaccuracies in the dataset.

4 Definitions and Methods

4.1 Definition of fairness

There are in general two types of fairness in related literatures: the first is individual fairness, which is based on the belief that similar individuals should be treated similarly. The other is group fairness, which states members from different groups should have similar chance to achieve target outcomes. In this project, we narrow our scope to group fairness. Following the idea of group fairness [11, 12, 13], suppose for the i^{th} protected group, the classification accuracy is α_i, then we define the unfairness in classification tasks as

$$UF := \text{Var}(\alpha_i) \tag{1}$$

It's straightforward to interpret UF as the divergence of accuracies among all the protected groups. The closer their accuracies are, the smaller UF is.

4.2 Non-deep learning baseline - Logistic Regression

The hyperparameters of our logistic regression model are as follows:

- Loss function: CrossEntropyLoss
- Optimizer: SGD
- Learning rate: 0.001
- Number of epochs: 30

4.3 Deep learning baseline (1) - FeedForward Neural Network

The hyperparameters of our FeedForward model are as follows:

- Structure: 2 fully connected linear layers, followed by ReLU and BatchNorm1d
- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Learning rate: 0.001
- Number of epochs: 20

4.4 Deep learning baseline (2) - Convolutional Neural Network

The hyperparameters of our CNN model are as follows:

- Structure: 3 convolutional layers followed by a fully connected linear layer (See Figure 1)

- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Learning rate: 0.001
- Number of epochs: 10

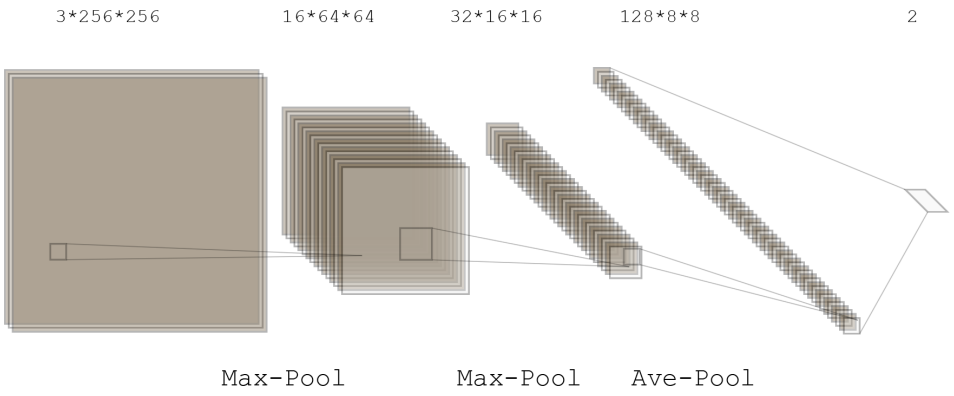

Figure 1: CNN baseline structure

4.5 Deep learning baseline (3) - Residual Network

The hyperparameters of our ResNet model are as follows:

- Structure: ResNet-34
- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Learning rate: 0.0001
- Number of epochs: 30

4.6 Novel approach (1) - Regularized ResNet

We propose a regularized ResNet. Our basic idea is to update the loss function by adding a UF-based regularization (penalty) term:

$$\text{New Loss} = \text{CrossEntropyLoss} + w * \text{Penalty} \tag{2}$$

Here, w is a coefficient of the penalty term. In the process of tuning a model, we find a critical issue when we assign a value to w that is too large. That is, while UF becomes smaller than the baseline UFs, the validation accuracy soon drops and converges to around 50%, or random chance. Hence, we use a step decay, i.e. $w = \delta(\text{step})$, to mitigate the decrease in loss as training occurs. The formula for step decay is as follows

$$\delta(\text{step}) = w_{start} - \frac{\text{step}}{N_{steps}}(w_{start} - w_{end}), \tag{3}$$

where N_{steps} is the total number of steps and w_{start} and w_{end} are the starting (maximum) and ending (minimum) weights for the penalty term, respectively. We set $(w_{start}, w_{end}) = (1, 0.001)$.

We also find that "MiddleEastern" has an exceptionally higher accuracy compared to the accuracies of other ethnic groups. Hence, we apply a weight 10 to "MiddleEastern" the penalty function, i.e.

$$\text{Penalty} = 10 * (acc_{\text{MiddleEastern}} - acc_{\text{average}})^2 + \sum_{i \in \{\text{other ethnicities}\}} (acc_{\text{i}} - acc_{\text{average}})^2$$

,

where $acc_{\text{MiddleEastern}}$ and acc_{average} are the validation accuracy of gender among Middle Eastern faces and average validation accuracy amongst all ethnic groups, respectively.

The hyperparameters of our regularized ResNet model are as follows:

- Structure: ResNet-34
- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Learning rate: 0.0001
- Number of epochs: 30

4.7 Novel approach (2) - Fair Masking

We also propose fair masking, which is another novel approach to mitigate the bias of neural networks. The general idea is that we focus on one hidden layer of the network and regularize the values of all neurons strongly dependent with the protected feature.

Our model is based on convolutional neural network, in which we specify a hidden layer $\mathcal{H}$. Suppose there are K neurons in $\mathcal{H}$ and the value passed from previous layer is $V = \{v_1, \ldots, v_K\}$. We add a mask $M = \{m_1, \ldots, m_K\}$ on previous output at $\mathcal{H}$, then pass them to the next layer. That is, the value restored in k^{th} neuron is

$$m_k \odot v_k.$$

Initially, all the m_k are set to be 1. This implies that the training process of this fair-masked CNN is the same as a normal CNN. After the training process, we record all the Vs, since for each image X_i, we can have a corresponding V_i by simply inputting image into the network. We select those v_k ($k \in \{1, \ldots, K\}$) that are strongly dependent with the protected features and set the corresponding $m_k = 0$. After this masking process, we obtain the final fair-masked CNN.

However, the correct dependence test can be hard to find. The two protected groups we chose are gender and ethnicity, both of which are categorical, whereas the hidden features in the network are all continuous . There is no widely used test for the dependence between a non-binary categorical variable (ethnicity in our case) and a continuous variable. Hence, we proposed two novel approaches to tackle this problem: a correlation-based test and a logistic regression-based test.

Correlation-based dependence test

In this test, we need an additional validation set. When training is complete, we calculate the classification accuracy of the network on this validation set. Then, we rank all the protected features in a ascending order according to classification accuracy.

For example, in the gender classification task, we specify the ethnicity as protected group, in which there are seven classes. For each class, we can calculate the classification accuracy of gender and rank all of the accuracies in an ascending order. Then, we assign a value R_c to each class which equals to its ranking (e.g. if the class "Latino" has the highest classification accuracy among the seven classes, it would have a value $R_c = 7$, and if "White" has the lowest accuracy, it would have value $R_c = 1$). Then, we can calculate

the correlation between R_c with each hidden feature. As a result, we would apply the threshold T_c on the correlation and update the value of masks by:

$$m_k = \mathbf{1}\{|\text{Cor}(v_k, R_c)| < T_c\}. \tag{4}$$

Logistic regression-based dependence test

As an alternative approach, we propose to fit a logistic regression to test the dependence between hidden features and protected ones. Recall that for multi-class logistic regression, if we have $\mathcal{S}$ classes in total, then we have

$$\mathcal{P}(Y_i = s) \propto \exp\left(-X_i \beta_s\right),$$

where $X_i = \{X_{i,1}, \ldots, X_{i,K}\}$, $\beta_s = \{\beta_{s,1}, \ldots, \beta_{s,K}\}^\top$ and $\beta_0 = \{0, \ldots, 0\}^\top$. This implies that for the k^{th} feature, there will be $\mathcal{S} - 1$ associated coefficients, namely $\{\beta_{1,k}, \ldots, \beta_{\mathcal{S},k}\}$. Once we fit a logistic regression, we assign a value $R_l := \max_i \beta_{i,k} - \min_i \beta_{i,k}$. We sort all K features using R_l and select top-T_l features to mask out.

In the context of gender prediction with ethnicity as protected group, psuedocode for a fair-masked CNN is shown in Algorithm 1.

Algorithm 1 `Fair-Masked CNN`

Data: Facial images
`Network`: A Convolutional Neural Network with a pre-specified hidden layer $\mathcal{H}$ for fair masking
Result: Classification result
Set $m_i = 1$ for $i \in \{1, \ldots, K\}$
while *training* **do**
  normally train the network
end
Conduct dependent test to identify relevant features $\{s_1, \ldots, s_k\} \subseteq \{1, \ldots, K\}$ that are strongly associated with the protected features.
Set $m_i = 0$ for $i \in \{s_1, \ldots, s_k\}$.
while *prediction* **do**
  calculate $V = (v_1, \cdots, v_K)$, the values passed to layer $\mathcal{H}$
  use $\{m_k \odot v_k\}_{k=1}^K$ as input and forward the network from $\mathcal{H}$
end

The hyperparameters of our fair-masked CNN model are as follows:

- Structure: see Figure 2 (number of hidden features changes for value of T_c)
- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Learning rate: 0.001
- Number of epochs: 10

Correlation-based test: We set T_c at values between 0.3 and 0, to select 1 to total 32 hidden features.

Logistic regression-based test: We set T_l at values between 1 to 32, to select 1 to total 32 hidden features.

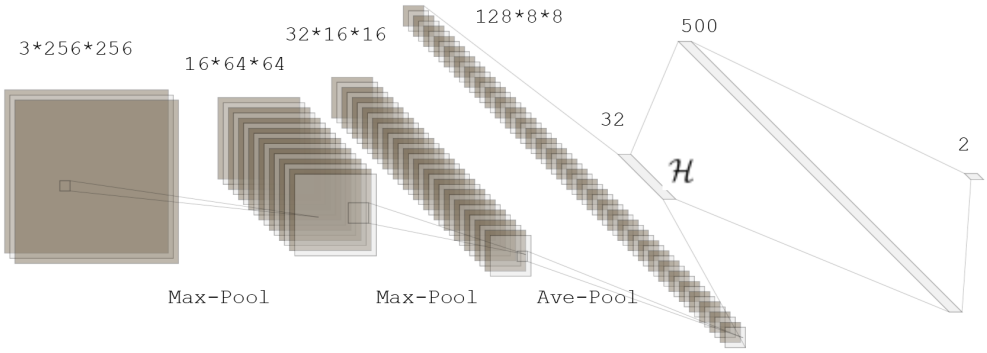

Figure 2: Fair Masked CNN

4.8 Novel approach (3) - Transfer Learning

Finally, we propose transfer learning from the ResNet-34 baseline model trained on FairFace onto the cleaned IMDb-Faces dataset. Simply put, the IMDb data is tested on the pretrained ResNet model. The IMDb dataset differs from FairFace in that it relatively larger and much more imbalanced with respect to gender and ethnicity classes.

Given a source domain D_S and a corresponding source task T_S, as well as a target domain D_T and a target task T_T, the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in D_T with the information gained from D_S and T_S, where $D_S \neq \mathrm{D}_T$ and $T_S \neq \mathrm{T}_T$.

Knowing that domain $D = \{X, P(X)\}$ and target $T = \{Y, P(Y)\}$, in our transfer learning scenario, we have $P(Y_S|X_S) \neq \mathrm{P}(\mathrm{Y}_T|X_T)$. This means that the conditional probability distributions of the source and target tasks (where source refers to FairFace and target refers to IMDb) are different, due to the differences in class balance [14].

5 Analysis

The results shown in this section are for gender prediction with ethnicity as protected group, and are summarized for all models below. The results for fair masking have $T_c = 0.1$.

Model	Training accuracy (%)	Validation/Testing accuracy (%)	Training loss	Validation /Testing loss	Unfairness
Logistic regression baseline	73.51	72.97	0.0044	0.0045	0.0010
FeedForward baseline	79.32	77.20	0.0066	0.0421	0.0015
CNN baseline	95.22	85.29	0.0009	0.0025	0.0012
ResNet baseline	97.93	86.71	0.0003	0.0035	0.0016
Regularized ResNet	98.26	88.03	0.0007	0.0074	0.0008
Fair masking	99.35	85.31	0.0001	0.0028	0.0011
Transfer learning	N/A	62.81	N/A	0.0103	0.0002

Table 3: Training and validation accuracy and loss and unfairness measures of gender classification by model

Note that, for all models aside from transfer learning, the evaluation accuracies are the lowest for the label "Black," which shows that the models are still somewhat unfair.

Model	Black	E. Asian	Indian	Latino	Mid. Eastern	S.E. Asian	White
Logistic regression baseline	67.35	72.26	74.74	75.97	77.58	71.52	72.37
FeedForward baseline	74.17	78.18	81.84	81.97	83.70	77.39	81.61
CNN baseline	79.58	83.74	86.28	87.92	90.57	87.49	84.80
ResNet baseline	79.95	86.00	86.87	89.03	92.64	86.22	86.28
Regularized ResNet	81.61	87.93	88.32	90.81	91.06	87.49	88.05
Fair masking	80.14	82.54	87.60	88.97	88.17	84.81	85.95
Transfer learning	63.14	60.91	62.49	65.54	64.83	62.85	62.60

Table 4: Validation accuracy (%) amongst ethnic groups for gender classification by model

5.1 Non-deep learning baseline - Logistic Regression

We use logistic regression for gender classification.

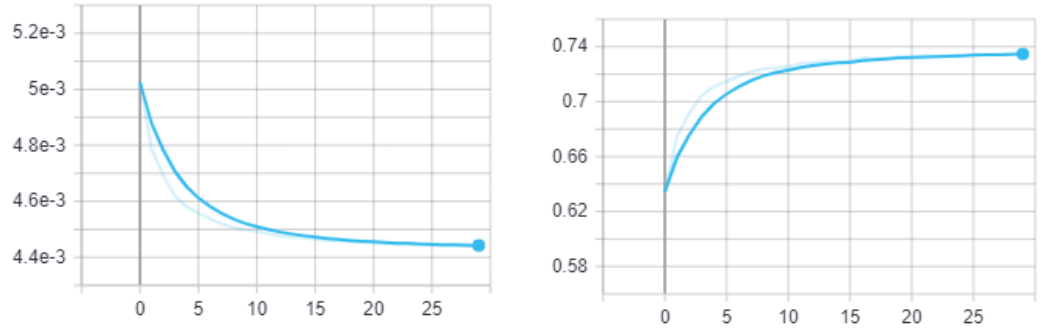

Figure 3: Training loss and accuracy in gender classification for logistic baseline

- Training: final training accuracy and loss are 73.51% and 0.0044, respectively.
- Validation: final validation accuracy and loss are 72.97% and 0.0045, respectively.
- UF for the validation set: unfairness parameter UF in equation (1) on the validation set is 0.0010.

5.2 Deep learning baseline (1) - FeedForward Neural Network

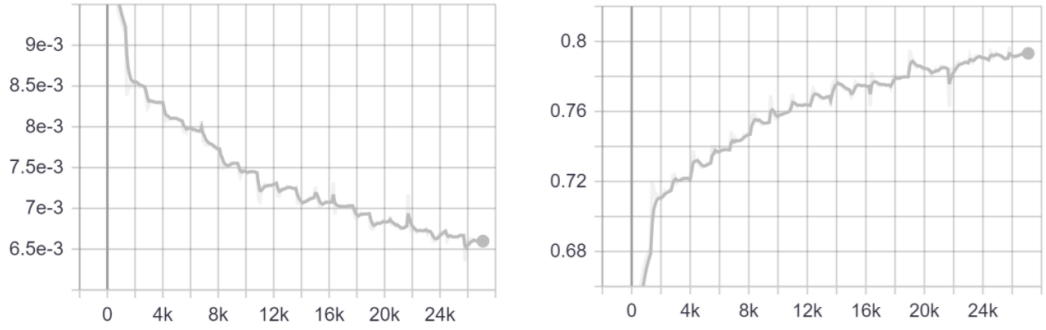

Figure 4: Training loss and accuracy in gender classification for CNN baseline.

- Training: final training accuracy and loss are 79.32% and 0.0066, respectively.
- Validation: final validation accuracy and loss are 77.20% and 0.0421, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0015.

5.3 Deep learning baseline (2) - Convolutional Neural Network

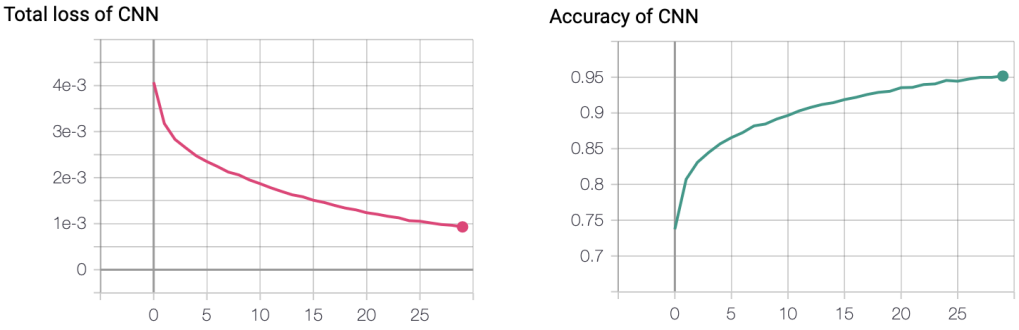

Figure 5: Training loss and accuracy in gender classification for CNN baseline.

- Training: final training accuracy and loss are 95.22% and 0.0009, respectively.
- Validation: final validation accuracy and loss are 85.29% and 0.0025, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0012.

5.4 Deep learning baseline (3) - Residual Network

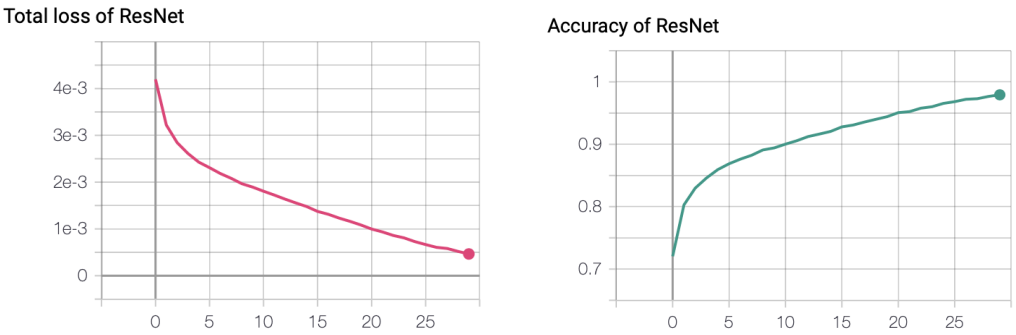

Figure 6: Training loss and accuracy in gender classification for ResNet baseline.

- Training: final training accuracy and loss are 97.93% and 0.0003, respectively.
- Validation: final validation accuracy and loss are 86.71% and 0.0035, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0014.

5.5 Novel approach (1) - Regularized ResNet

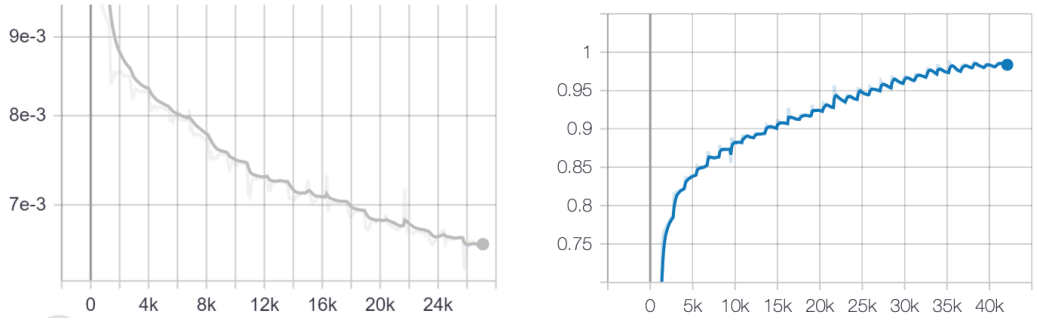

Figure 7: Training loss and accuracy in gender classification for regularized ResNet.

- Training: final training accuracy and loss are 98.26% and 0.0007, respectively.
- Validation: final validation accuracy and loss are 88.03% and 0.0074, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0008.

Note that the penalty is clearly applied, as the loss on the regularized ResNet is greater than that of the baseline ResNet. The UF value is significantly lower than those of the baseline models, confirming that our approach is effective with respect to improving fairness. However, we can see the accuracy of "Black" is still significantly lower than other ethnicity, which implies that we may need to use a lower weight in our penalty function, e.g. take $\frac{1}{10} * (acc_{black} - acc_{average})^2$. Due to time constraints, we were not able to test this, but it seems to be a promising direction for future work.

5.6 Novel approach (2) - Fair Masking

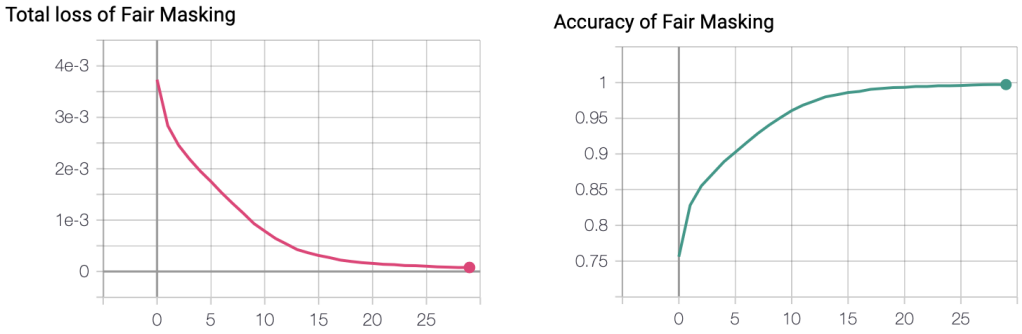

Figure 8: Training loss and accuracy in gender classification for Fair Masked CNN.

- Training: final training accuracy and loss are 99.35% and 0.0001, respectively.
- Validation: final validation accuracy and loss are 85.42% and 0.0028, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0012.

Performance with correlation-based dependence test

Below are the plots of the unfairness and accuracy of fair-masked CNN on the validation set with different values of T_c.

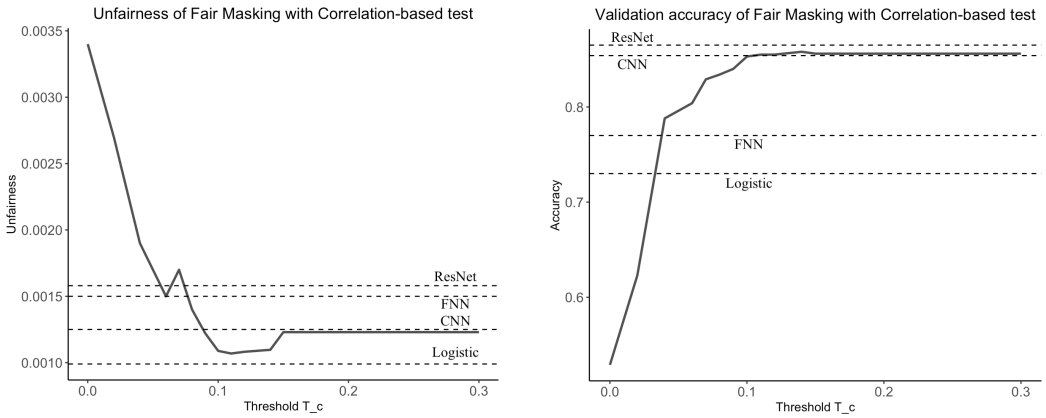

Figure 9: Unfairness and accuracy of Fair Masked CNN in gender prediction

The ranking of unfairness, from fairest to least fair, among all the baseline models is: logistic regression, CNN and FNN, and ResNet. Since our fair-masked CNN also adopts a CNN structure, we observe that when $T_c \to \infty$, in which none of the hidden features are masked out, the unfairness of the algorithm is very similar to that of CNN baseline.

We observe that the unfairness of the network decreases from around $T_c = 0.15$, where we select the first hidden feature from $\mathcal{H}$ to mask out, to $T_c = 0.10$, where we select 6 out of 32 features. Afterwards, the unfairness increases as we select more hidden features. Notice that at the minimum of the curve, in which we select 6 features to mask out, the prediction accuracy is 85.31%, which is similar to that of the CNN and ResNet baselines and significantly better than that of the logistic regression. However, the logistic regression model is a bit fairer than fair-masked CNN at $T_c = 0.1$, implying a fairness-performance trade-off involved in fair masking.

The result implies that, for fair masking with correlation test, a carefully chosen threshold will effectively mitigate the unfairness while doing little harm to the overall performance of the CNN in gender prediction.

Performance with logistic-regression-based dependence test

Below are the plots of the unfairness and accuracy of fair-masked CNN on the validation set with different values of T_l.

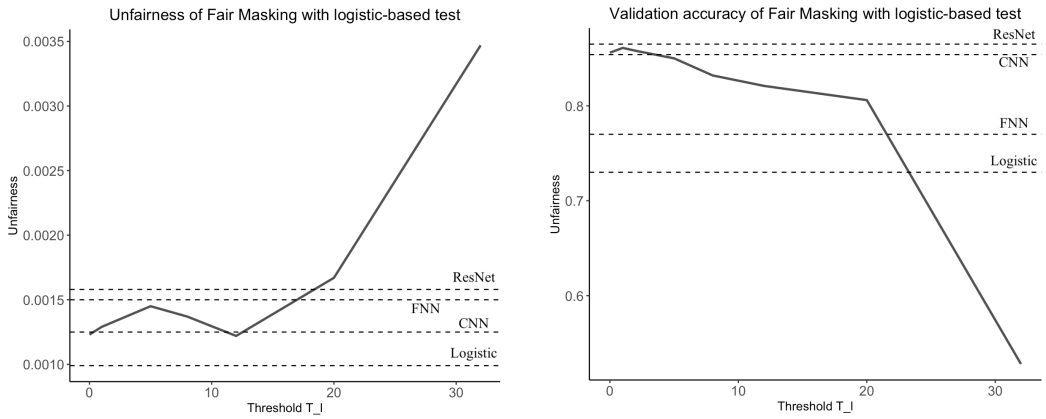

Figure 10: Unfairness and accuracy of Fair Masked CNN in gender prediction

Unfortunately, we observe that masking out several hidden features from $\mathcal{H}$ (no more than five) doesn't significantly affect the prediction accuracy on validation set, and, in fact, amplifies the algorithmic bias. We believe one potential reason for this failure is that the dependence test based on logistic regression is less effective than the correlation-based test, failing to select those that truly affect the bias of algorithm. We believe it is still possible to propose other novel dependence tests to select those features that affect algorithm bias.

5.7 Novel approach (3) - Transfer Learning

- Test: final test accuracy and loss are 62.81% and 0.0103, respectively.
- UF for the validation set: unfairness parameter on the validation set is 0.0002.

The overall test accuracy is significantly lower than the baselines' results, but above random chance.

While it inherently makes sense that accuracies for transfer learning would be below the baseline, these drops in accuracy are significant. We suggest some reasons why this may have occurred.

- The class imbalance of IMDb is extreme (see Figure 11 below), whereas the data that the model was trained on was fairly balanced.
- There is significant mislabelling of data in IMDb. This is noted in 3.2, and certainly would have a negative effect on accuracy.
- The FairFace images are cropped much more closely than the images in IMDb, which could impede classification accuracy. This can also be visually observed in the Data section.

We can observe the distributions for gender and ethnicity across FairFace training, validation, and IMDB datasets below.

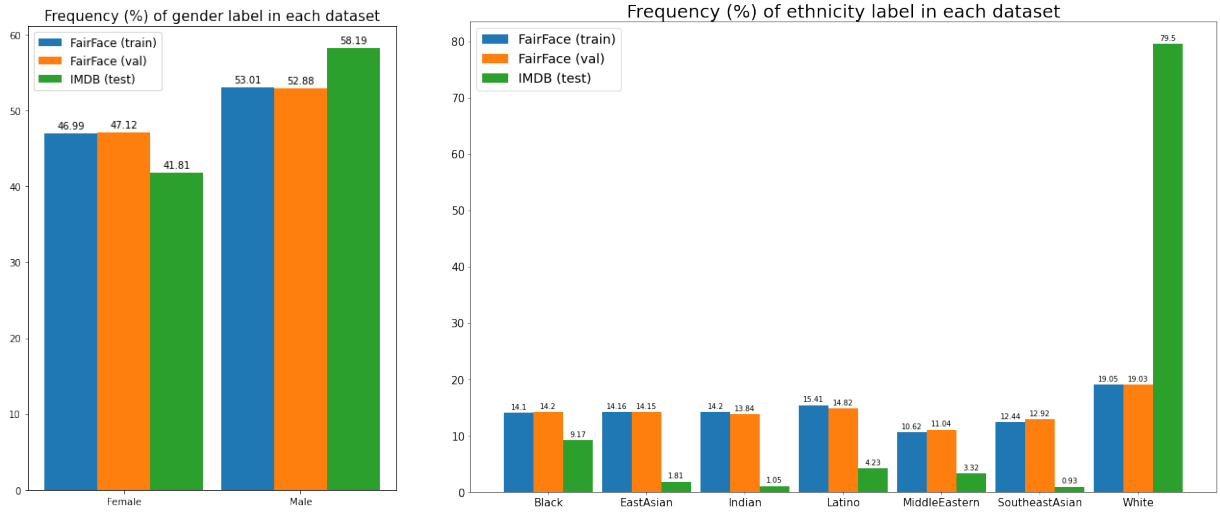

Figure 11: Distributions of gender and ethnicity across all datasets.

Interestingly, the unfairness measure for transfer learning was 0.0002, the lowest of all models. This means that the model was the most equally accurate (or inaccurate) across all ethnic groups. This may have to do with the fairness-performance trade-off discussed in the fair masking analysis.

6 Discussion

In this project, we have examined the algorithmic fairness of deep neural networks in facial classification tasks. We first set up non-deep learning and deep learning baseline algorithms and observe non-negligible unfairness in gender prediction. In particular, the validation accuracy for the Middle Eastern ethnicity group is consistently at least 10% higher than that of the Black ethnicity group. By proposing two novel approaches, fair regularization and fair masking, we illustrate that it is possible to reduce the unfairness while maintaining (or even improving upon) the high performance of the baselines. Compared to the baselines, these two novel fair algorithms show improvement of fairness without losing much validation accuracy. We further investigate the performance of the algorithms in transfer learning.

In the IMDb-Faces dataset, we observe that a ResNet pretrained in FairFace shows significant reduced performance (62.81%) compared to the FairFace validation set (86.71%). We believe potential reasons include significantly imbalanced distribution of ethnicity and gender groups, mislabelling, and the fact that faces in IMDb-Faces are not cropped as closely as those in FairFace dataset. Notably, the UF measure for the transfer learning on IMDb was by far the lowest, compared to all other models. This may have to do with the fairness-performance trade-off, where lower accuracy is associated with lower unfairness. In order to increase the accuracy of IMDb, we suggest using IMDb data to tune the hyperparameters of a model, instead of FairFace.

We struggled to find a dataset that contained interesting protected features as well as names of subjects, so that identity classification based on facial recognition could be performed. While FairFace has gender and ethnicity well-balanced amongst its dataset, it does not have IDs or names for the individuals in the photograph, or multiple photographs per individual. IMDb-Faces has those names and photographs, but a highly imbalanced and mislabelled dataset, and we had to webscrape its ethnicity data.

Due to the limited time, there are still ideas haven't been investigated throughout the project study. We are curious about the performance and unfairness of DenseNet, another popular network structure in computer vision. We also wish to learn more about fairness in transfer learning: whether a fair model appears to be fairer than baselines after domain transfer can also be an exciting question in this area of research. We also would like to test more protected features, such as age, and explore areas of intersectionality between multiple features (e.g. examining female vs. male classification accuracy across ethnicities).

Regarding the topic of fairness in deep learning, much room is remained for future investigation and study as a whole. First, the effect of hyperparameters in algorithmic fairness hasn't been well-studied or documented yet. We believe it would be interesting and useful to investigate and provide insight into that how to design a network structure to mitigate potential bias. For example, how does the depth and width of the network influence fairness? How about optimizers, learning rate as well as batch size? Furthermore, existing works only consider the case when only single feature is protected. How to control unfairness of the algorithms regarding multiple protected features can be even more challenging but meaningful.

References

[1] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[2] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. *NIST Interagency/Internal Report (NISTIR)*, (8280), December 2019.

[3] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[4] Salem Abdurrahim, Salina Samad, and Aqilah Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34, 08 2017.

[5] Andrew Jason Shepley. Deep learning for face recognition: A critical analysis, 2019.

[6] Filip Michalsky. Fairness criteria for face recognition applications. *Association for Computing Machinery*, pages 527–528, 01 2019.

[7] Joy Buolamwini and Gebru Timnit. Gender shades - intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*, pages 1–15, 2018.

[8] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, Dec 2011.

[9] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670, 2017.

[10] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[14] Sebastian Ruder. Transfer learning - machine learning's next frontier, Apr 2019.