# Example-Reweighting Meta-Learning Model for Debiasing Race Imbalances in Face Datasets

**Seoyoung Lee**
**Purdue Undergrad ECE**

## Abstract

Biases and class imbalances often lead machine learning models (e.g. classification and detection) to have poorer performances for the minority dataset. Facial image datasets often suffer from the imbalances in the representations of races. These datasets often tend to be skewed toward Caucasian images. The biases in data would lead the machine learning models to have poorer performances of recognizing the people of color. The underrepresentation in the datasets and poor performances of the models would lead to further discrepancies in future models and social representations. I plan to re-implement a meta-learning model that reweights training examples for face datasets. I reproduce the reweighting results using face datasets and evaluate the performances of the binary classifications of Caucasians and people of color. Debiasing classification models builds upon providing balanced datasets by ensuring the people of minority are still well represented when using previous imbalanced datasets or when there are insufficient data of the minority. This initiative to debias classification models would bring broader social implications in fostering fairness and increasing the representations of the people of color.

## 1. Introduction

Machine Learning models should strive to have high accuracy and be robust. When training the models, the distribution of the datasets may impact the learning of the models, possibly introducing biases. Class imbalances in datasets is one source of biases, leading the model to be skewed toward the majority class and have poor performance for the minority dataset.

Many previous face datasets suffer from the skew towards Caucasian images. There has been research about providing a new face dataset with balances among races. My research builds upon the provision of a balanced and clean dataset and studies whether additional reweighting would be necessary.

I re-implement a reweighting model that strives to maximize the model performance and robustness even with an imbalance in the input data. I plan to reproduce the results and evaluate the reweighting performance for classifying different races in face datasets. I hope to ensure the face classification models can still have high accuracy for minority races, people of color and other races. I experiment the models' performances for binary classifications of whether the face image represents Caucasians or people of color. I compare the performance of different reweighting models for a range of imbalances.

I believe it is critical to improve the dataset balance at the input stage and increase the representations of the people of minority. However, the reweighting model has validity because there may be countries or areas with a natural distribution with an imbalance of races. It is critical to apply the reweighted models in these situations. I study how much the reweighting model improves the classification performance beyond the base standard Le-Net models.

## 2. Literature Review

There is previous research that addresses the machine learning models' struggles with imbalance and noisy-label datasets. The meta-learning model that reweights training examples (Ren et al., 2018) (1) presents the the problem that deep neural networks tend to overfit to training set biases and label noises. The research proposes a meta-learning algorithm that teaches the model to reassign weights to training examples based on their gradient directions. The model takes the mini batches of training examples and the clean unbiased validation set and computes the gradient descents. It aims to minimize the loss on the validation set by giving higher weights to training examples that have similar gradient directions as the validation set.

My research focuses on the re-implementations of the the class imbalance experiment. The meta-learning model uses a small unbiased validation set that is consistent with the evaluation procedure. An online reweighing method is used in order to perform validation at every training iteration, instead of performing validation at the end of training. This method ensures that that the model is robust for general forms of training set biases.

To improve the robustness against training set biases, the model implements a weighted loss and instantiated meta-learning. Meta-learning object to online approximation: Given a training set of input-label pairs and a small unbiased and clean validation set, the model minimizes the expected loss for the training set. Each input example is first weighted equally. Then, the inputs are reweighted while minimizing the weighted loss. The weight is treated as training hyperparameters and it is optimized based on the validation performance. The weights should be nonnegative since minimizing the negative training loss tends to result in unstable behavior.

The weight is adapted online through a single optimization loop. The descent direction of some training examples are inspected locally on the training loss surface and reweighted according to their similarity to the descent direction of the validation loss surface. A mini-batch of training examples is sampled, which then the parameters are adjusted according to their descent direction of the expected loss on the mini-batch. It then takes a single gradient descent step on a mini-batch of validation samples with respect to eta and corrects the output to get a non-negative weighting. A batch-normalization is conducted to match the original training step size.

Learning to reweight examples in a multi-layer perceptron (MLP): The gradients of the validation loss is computed with respect to the local perturbation. The meta-gradient on the perturbation is composed of the sum of the similarity between the training and validation inputs and the similarity between the training and validation gradient directions.

Implementation using automatic differentiation: The unnormalized weights are calculated based on the sum of the correlations of layerwise activation gradients and input activations. Firstly, the gradient graph of the training batch is unrolled. Secondly, the backward-on-backward automatic differentiation is used to take a second order gradient.

Convergence of the reweighted training: The method converges to the critical point of the validation loss function under certain conditions. Using a larger training set and converging to an approximation from the clean and balanced validation dataset improves both generalization and robustness.

The research about the meta-learning reweighting model used the MNIST dataset for the experiment. The research analyzed the test error of different reweighting mechanisms and different imbalance proportions. The research has limitations in that it did not evaluate the debiasing performance for other datasets with class imbalances. I raise the question of whether the reweighting model would still show promising performances for face datasets. Face datasets tend to be more complicated than handwriting (MNIST) datasets because the images have more complex features. I plan to re-implement the reweighting model and conduct the debiasing experiment using face datasets of Caucasians and People of Color.

Many face datasets have significant underrepresentation of the people of color. The imbalance of races may lead to inconsistency in classifying the people of color, impacting the performance of further machine learning models. These poor machine learning models would lead to further translations of biases, including the marginalization and misreporting of the people of color. These implications raise ethical concerns and questions about the fairness of models.

Fairface dataset (Karkkainen et al., 2021) (2) is a novel face image dataset with 7 race groups, collected primarily from the YFCC-100M Flickr dataset. The racial groups are as follows: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern and Latino. These groups represent the commonly accepted race classification from the US Census Bureau. During dataset construction, the authors addressed the biases in limited photography situation (quality bias) or search result filtering bias. The faces were detected from the images without any preselection. The race, gender and age group were annotated and these annotations were further refined by training a model from the initial ground truth annotations.

The research of Fairface delivered a rebalanced face image dataset and has improved the accuracy of classification models. The Fairface research primarily focused on reconfiguring the dataset. I want to evaluate whether the reweighting of training inputs or the creation of balanced datasets would lead to better improvement of classification models. Due to insufficient data, the Fairface dataset still suffers from the inability of representing more marginalized people such as Native Americans, Hawaiian and Pacific Islanders, etc. Different countries have different proportions of races and I want to answer the question of whether it is enough to provide a balanced dataset. My research continues from the Fairface research's balanced dataset by adding the reweighting of training examples.

## 3. Method

I have obtained about 10000 images from the Fairface dataset and organized the dataset by the racial groups. The Fairface dataset has seven racial groups, but to simplify the classification experiment, I have created binary groups of the Caucasians and the people of color.There are about 2000 Caucasian images and about 8000 images of people of color. Many datasets have biases of favoring Caucasians so I redistributed the dataset to create balanced and imbalanced datasets. The ratio of Caucasians and the other races is 50:50 in the balanced dataset. For the imbalanced dataset, I implemented datasets that are majorly Caucasian, the proportions of the majority ranging from 60 % to 90 %.

I am re-implementing the Reweighting Meta Learning research's method of testing the performances of the baseline model and different reweighting models with a ranging proportion of the majority class. When comparing the models, a standard LeNet is trained as the baseline model and then the research compares the baseline model and the following reweighting models:

Proportion - weight each example by the inverse frequency

Resample - sample a class-balanced mini-batch for each iteration

Hard mining - highest loss example from the majority class

Random - random example weight

Autodiff - take the batches of the data, reweight the training example by computing the gradients of the training and validation examples

The Reweighting Meta Learning research proposed the autodiff reweighting as a new method to reweight the training examples. Their research proved that the autodiff reweighting showed significantly less test error for imbalanced datasets compared to the baseline model and other previous reweighting methods. I plan to use this method for binarily classifying the face dataset.

I have used the code that the research published for conducting the imbalance MNIST dataset experiment. Each reweighting model had different experiment configurations and reweighting functions defined. Some of the key differences were the number of gradient descent steps, batch size and the validation set size. The base model was a standard LeNet written in Tensorflow.

When obtaining the imbalanced dataset, I passed the 10,000 Fairface dataset and first divided them into the Caucasian and the people of color datasets. Then, I split each dataset into the training and test examples with the ratio of 9:1. I shuffled each dataset and created training and testing subsets: in the final subsets, there were 3600 training samples and 360 test samples. Some reweighting models and the baseline model did not use the validation set. For reweighting models that did use the validation set, I took a very small balanced and clean validation set of 10 images from the training subset. The validation set is split directly from the training set so that I do not introduce additional information outside the original training set. I finally shuffled all of these datasets.

After subsampling the dataset, I have performed a binary classification of the race: Caucasian or people of color. For the balanced dataset, I have used the base LeNet model and for the imbalanced dataset, I have used both the base LeNet model and the reweighting models. For all experiments, the learning rate is reduced by half for the second half of training. When running the experiment, I run it over 10 random seeds and take the average to ensure that the performance accuracy is consistent for multiple executions. I compare the training and test accuracies of the base LeNet model, the past reweighting models and the autodiff reweighting model for a range of the proportion of the majority class.

## 4. Results

Using the face dataset, I re-implement the meta-learning model that reweights training examples. I modified the algorithm used for the MNIST dataset class imbalance experiment. I parsed the code into data import, dataset configurations, training, reweighting, evaluation and execution of the experiment. Due to its usage of Tensorflow version 1, I had to do debug the code so that it was compatible with my environment. I modified other hyperparameters and simplified the dataset configurations.

I used 3600 training samples and 360 test samples. As the model was taking gradient steps, I tracked the losses and the training and test accuracies. I ran the experiment with 10 trials with different random seeds to ensure that there was consistency in the performances with random executions. For the proportions of the majority class ranging from 60% to 90%, I evaluated the final training and test accuracies of the base standard Le-Net model and the hardmining, ratio, random and autodiff reweighting models.

The following graphs show the training and test accuracies of the base model and different reweighting models for the proportions of the majority class (Caucasians) ranging from 60% to 90%.

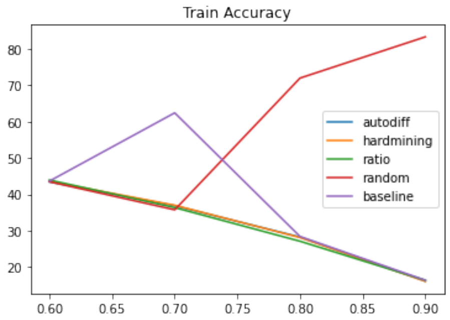

*Figure 1.* Training Accuracy (%) of models on imbalanced datasets of 60% to 90% majority

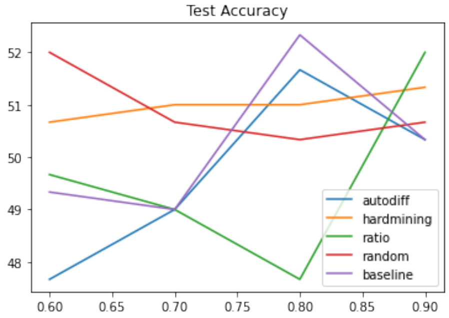

*Figure 2.* Test Accuracy (%) of models on imbalanced datasets of 60% to 90% majority

For the training accuracy, I have found out that the baseline model and the hardmining, ratio and autodiff reweighting models generally have decreasing training accuracy with an increase in the imbalance. The baseline model had a spike in the training accuracy in the 70% majority proportion, but then experienced a decline in the train accuracy afterward. The random reweighting model slightly dipped in the training accuracy at 70% majority, but actually improved the training accuracy afterward.

The graph shows that during the training phase, the reweighting models have not shown any significant difference from the baseline model. The training accuracy decreased more steadily so this may indicate the reweighting models may help stabilize the performance of the model, but does not improve the training accuracy performance. The random reweighting models have a high training accuracy (70% to 80%) in big imbalances (80% to 90% majority), but the evaluations of the test accuracy is needed to determine whether the improved performance comes from fundamentally improving the model or plainly overfitting to the training set.

All of the models have had test accuracy of about 50%, which is not a significant improvement from random binary classification. The autodiff reweighting model generally had even lower performance than the baseline model. The models with the highest general test accuracies are hardmining and random reweighting. The result may indicate the combination of hardmining and random reweighting models may lead to improved test accuracy.

I computed the runtime to conduct the experiment with majority class proportions ranging from 60% to 90% (with 10 % increments) and with 5 experiment configurations (hardmining, ratio, random, baseline, autodiff). With the original number of steps of 2880, it took more than 5 hours. The runtime was especially high with the hardmining reweighting model. As I tracked the losses and the accuracies, there were very small changes in the loss and accuracy so I instead evaluated with the number of steps of 72. Since there are 3600 training samples and the batch size is 100, I concluded 72 steps would cover the full training dataset. The high runtime result may indicate that the reweighting model may lead to improved performances and higher accuracies, but there are areas of improvement for efficiency and runtime speeds.

I have additionally studied the performance results with extreme imbalances of 91% to 99% majority proportions. The following graphs show their training and test accuracy results.

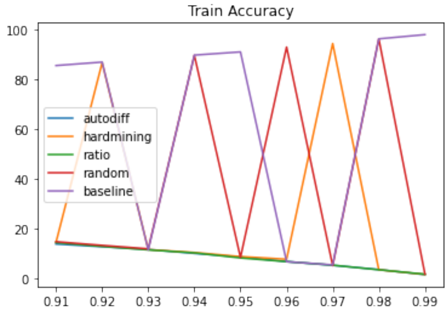

*Figure 3.* Training Accuracy (%) of models on imbalanced datasets of 91% to 99% majority

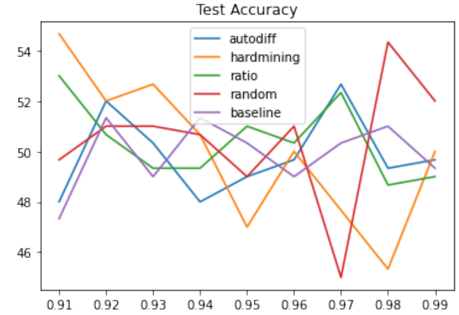

*Figure 4.* Test Accuracy (%) of models on imbalanced datasets of 91% to 99% majority

During the training phase, the autodiff and ratio reweighting models had steady declines in the accuracy. This may indicate these two models are more stable than the other models who experiences sudden spikes in accuracies. The spikes are pretty extreme, going from about 10% to 80% accuracy.

During the test phase, all models still had about 50% accuracy. Some models such as autodiff and ratio reweighting models and the baseline model were more stable. In extreme imbalances, it may be beneficial to prioritize more stable models.

The analyses of the training and test accuracies of different models with a range of majority class proportions describe the debiasing performances of reweighting models. Some models such as hardmining and random reweighting had a slightly better accuracy, showing that these models are candidates for the face data classification task.

## 5. Conclusion

The reweighting model for face datasets serves the purpose of improving classification performance with imbalances in races. I re-implemented the meta-learning reweighting model for the binary classification of Caucasians and people of color.

For the experiment on majority class proportions of 60% to 90%, the training accuracy results show that many reweighting models have similar performance with the baseline model during training. The random reweighting model had a lot higher training accuracy than the other models, which may show promising performances but also concerns for overfitting. The test accuracy results show that all models have about 50% accuracy, which is not a big improvement from random binary classifications. However, the accuracy still did not decline for bigger imbalances (80% to 90%) so this result may indicate that the reweighting models may have prevented the decline in test accuracy from the data imbalance. Random and hardmining reweighting models have had slightly better performance than the baseline model so these models may be studied further to further improve classifications for face datasets. Unlike the MNIST dataset, the autodiff model did not show much improvement. This result may show that different datasets may benefit from different reweighting models.

There may be cautions needed not only for accuracies but also for runtime speeds and efficiency. Some reweighting models such as the hardmining reweighting model had a high runtime so their efficiency may be further improved.

For the experiment on majority class proportions of 91% to 99%, the autodiff and ratio models were a lot more stable than the other reweighting models and the baseline model. In extreme imbalances, there should be concern for not only improving the accuracy but ensuring the stability of the model performance.

Further areas of research is performing multiclass classifications. Many research shows that face datasets have major representations of Caucasians, but don't provide sufficient information about imbalances within people of color. Another area is studying social groups outside the racial groups in the Fairface dataset: for example, Hawaiian and Pacific Islanders and Native Americans. I hope to test how the model would perform for social groups of extremely small populations. This further research would further reveal the debiasing performance of marginalized people.

Imbalances in datasets do not stop at limiting the data. Underrepresentations and marginalizations of racial groups would lead further machine learning models to poorly reflect these groups and deepen the biases and discriminations. It is important that machine learning models are robust for and reflective of all racial groups.

## 6. Code

Source code: https://github.com/uber-research/learning-to-reweight-examples

My re-implementation: https://colab.research.google.com/drive/1zI28exHWm9gUrtYGqwObl4recAYY2tKz?usp=sharing

## 7. Citations and References

### References

[1] Mengye Ren and Wenyuan Zeng and Bin Yang and Raquel Urtasun, Learning to Reweight Examples for Robust Deep Learning ICML, 2018

[2] Karkkainen, Kimmo and Joo, Jungseock, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, p. 1548–1558