CS 674 Project 2

Rebecca Jones, Ethan Walker

March 27, 2021

1 Overview

Classifying images without labels is a difficult machine learning task. Last year, a new method, SimCLR, came out that substantially improved on previous state-of-the-art self-supervised learning.[2] Using small percentages of labeled data, the authors were able to achieve almost the same accuracy as supervised learning on some common benchmark datasets. An updated version, SimCLR2, was released in October 2020 that made some improvements to SimCLR. Notably, in increased accuracy by using deeper ResNets, scaling from ResNet-50 to ResNet-152 and making the projection head three layers instead of two. We apply this method to a combined dataset of facial images to classify the images by race and gender.

2 Method

SimCLR consists of several steps. The first is data augmentation. Each image is copied and both versions are modified in the following way; crop, flip, color jitter, grayscale, and gaussian blur. The images are compared using NT-Xent loss.

$$l_{ij} = -log\frac{exp(sim(z_i, z_j)/\tau}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq i} exp(sim(z_i, z_j)/\tau} \tag{1}$$

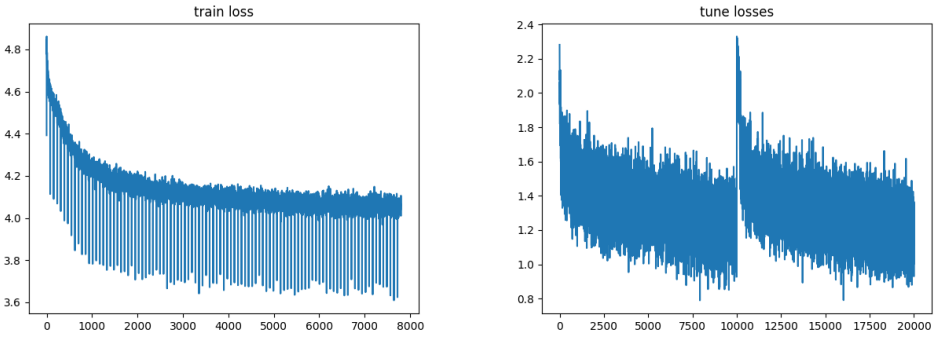
where $sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\|\|z_j\|}$, the cosine similarity.

The goal with this portion is to have a self-supervised portion of the training, where the model learns to create good representations of the images. Following this step is a fine-tuning step where the model will then learn to apply those representations to different problems, race and gender. The fine-tuning portion is performed using a standard cross entropy loss.

During the SimCLR training we used the Layer-wise Adaptive Rate Scaling (LARS) optimizer with SGD, so that we could increase the batch size [6]. In our CIFAR10 tests, we were able to get batch sizes up to 1024.

2.1 Architecture

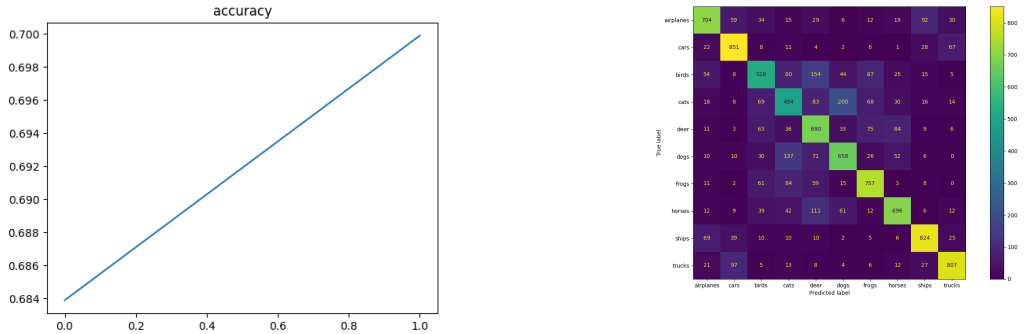
We used a 50-layer ResNet as our encoder, and a 3-layer feed forward network as a projection head. Although improvements to SimCLR suggest deeper networks, we did not have the GPU capacity to run a ResNet-152. The projection head that is learned during the NT-Xent loss portion is discarded during the fine tuning phase and a new projection head is created. This allows the projection head to be more easily trained for different tasks



(a) Loss of the SimCLR portion of the training

(b) Loss of the two fine tuning portions done

Figure 1: Losses over the training on CIFAR10



(a) Validation Accuracy

(b) Confusion Matrix for the SimCLR applied to CIFAR10

Figure 2: Validation results of CIFAR10

2.2 Model Testing

To test that our model was working, we validated it on CIFAR10. Sadly, because of memory limitations, we are not able to fully compare our testing with what is presented in the original SimCLR papers, as they were able to use much larger models than we were able to use.

We trained the model using SimCLR for 100 epochs and did fine-tuning for 20 epochs. In this process we stopped the SimCLR training after 50 epochs to do some validation, so you will see two fine tuning sessions in Figure 1b. We can see that additional contrastive learning using SimCLR improves our results in Figure 2a as our accuracy increases nicely after the 50 epochs of additional training.

3 Data

To create a dataset with enough representation, we combined three common facial recognition datasets: UTKFace[3], Flickr-Faces-HQ Dataset (FFHQ)[4], and CelebA[5]. The CelebA dataset consists of 202,599 images representing 10,177 people. To speed up training, we randomly selected one image per person to use in our final dataset. This gave us a total of 103,885 images for our training set. For each dataset, we used the thumbnail option if available, and each image was cropped to 128x128, the smallest image size.

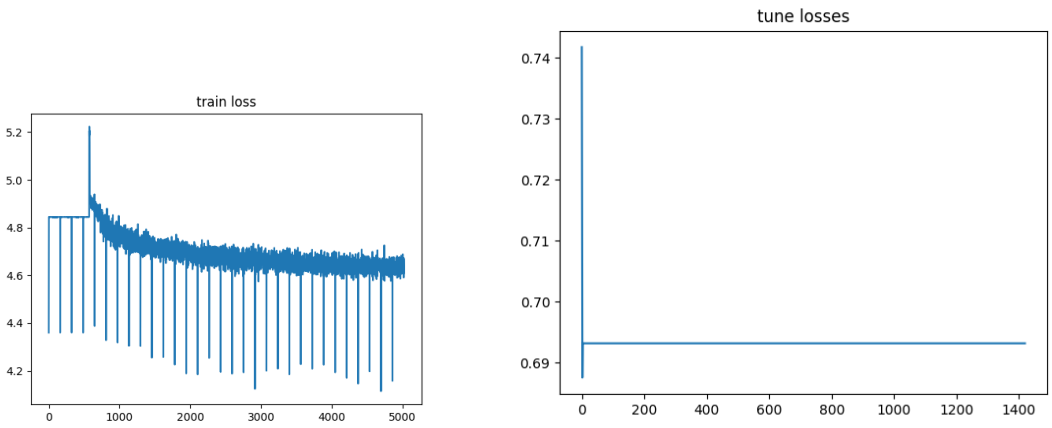
Of these datasets, only the UTKFace included classifications of race, gender, and age. This represented about 22.8% of the data. We used this dataset to do our fine tuning and validation, using 30% of the data to fine-tune and 70% to validate.

4 Results

Our results cover two different classification problems over the same custom faces dataset. The two problems we approached were classification based on gender and classification based on race. For these problems we used a 36-layer ResNet, because the face images are slightly larger than the CIFAR10 images.

4.1 Gender Results

We trained the SimCLR model on the images with two classes, male and female. The number of male and females in the data set can be seen in Figure 4. On our available machine with a 40 GB GPU, we were only able to complete 30 training epochs. Training seemed to go well as the loss is decreasing in Figure 3a. However, Figure 3b shows that the fine-tuning loss stayed constant. Looking at the confusion matrix, Figure 4, we can see that the model only predicted class 0, or male. Obviously something is not working, but since it takes about a day to train 30 epochs, we didn't have time to figure out why. We tried decreasing the learning rate, but the loss was still constant. Since we trained CIFAR10 on 100 epochs and the papers indicated training for longer, up to 800 epochs, was significant, we suspect that might contribute to the issues.



(a) Gender training loss

(b) Gender fine-tuning loss

Figure 3: SimCLR applied to gender classification problem

4.2 Race Results

The goal with this classification problem is to classify images by the race of the person in the image. The races in our labelled data were the following: White, Black, Asian, Indian, and other. The results of this problem are found in 5. The accuracy was 0.69, which is promising given that we did fine-tuning on only 6.8% of the total dataset. With a larger fine tuning set we hope to be able to achieve great results for classifying completely unlabelled images.

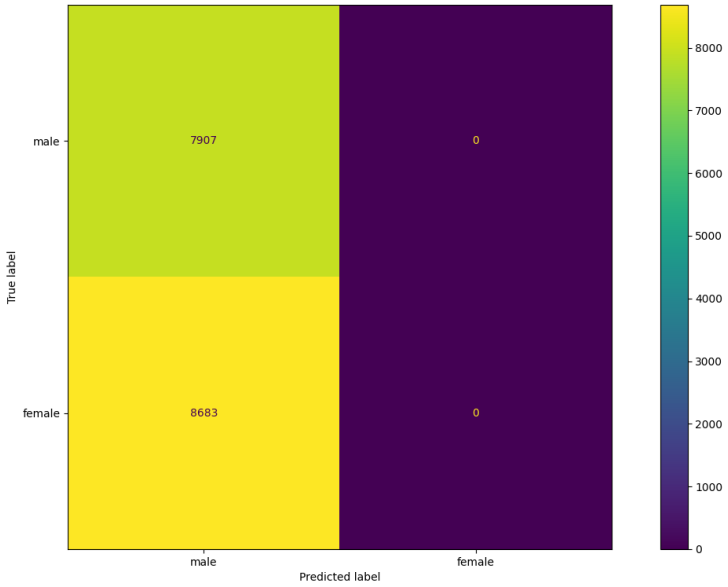
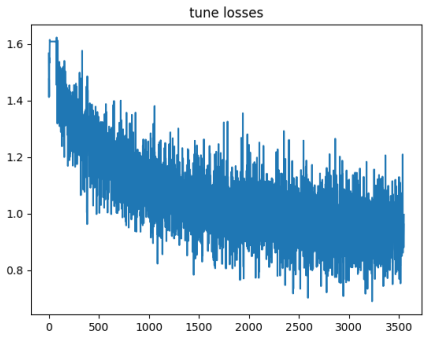
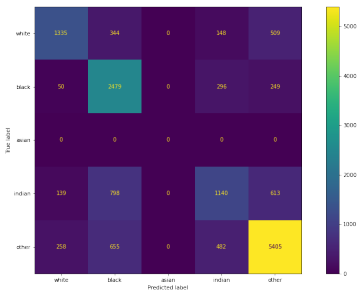


Figure 4: Confusion matrix for the gender model

Figure 5b reveals that, inexplicably, there were no Asian people in the validation set. This is simply caused by an error in how the data was organized and is an anomaly that needs to be addressed, but we don't believe that it invalidates our results.



(a) Fine tuning loss



(b) Confusion Matrix for the SimCLR applied to race classification problem

Figure 5: SimCLR applied to race classification problem

References

[1] Chen, Kornblit, Swersky, Norouzi, Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. October 2020. https://arxiv.org/abs/2006.10029

[2] Chen, Kornblit, Norouzi, Hinton. A Simple Framework for Contrastive Learning of Visual Representations. February 2020. http://proceedings.mlr.press/v119/chen20j/chen20j.pdf

[3] Zhang, Zhifei, Song, Yang. https://susanqq.github.io/UTKFace/

[4] Karras, Laine, Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. March, 2019. https://arxiv.org/abs/1812.04948

[5] Liu, Ziwei and Luo, Ping and Wang, Xiaogang and Tang, Xiaoou. Deep Learning Face Attributes in the Wild December, 2015 http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[6] You, Gitman, Ginsburg. Large Batch Training of Convolutional Networks. September 2017. https://arxiv.org/pdf/1708.03888.pdf