ENGR 2900 Project 1 Report

Javier Farach and Tobby Zhu

March 2024

1 Abstract

A data wrangling and machine learning pipeline is implemented for a binary face recognition task. Using Inception ResNet v1 model pre-trained on the VGGFace 2 dataset and training it on manually labeled face image data, we were able to achieve 87% accuracy on the validation dataset when identifying the face of Bill Gates. The trained model is able to correctly identify most instances of Bill Gate's face when tested on two separate video clips.

2 Introduction

In this project we implemented a pipeline for real time facial recognition in videos. We performed transfer learning using Inception ResNet v1, which has been pre-trained on the VGGFace 2 dataset. We train the model on a dataset collected from two separate public datasets, CelebA and the Flickr-Faces-HQ Dataset, and manually labeled with two classes "bill_gates" and "not_bill_gates". We use stochastic gradient descent to train a logistical regression model with a sigmoid activation function.

The model achieves 87% accuracy on validation set. We then use the model to perform facial recognition in videos. We use the DNN face detection model to output a bounding box around detected faces, and then feed the cropped face image to our facial recognition model to output either "bill_gates" or "not_bill_gates". Using two video clips involving both Bill Gates and other people, we observe that our model is able to successfully detect Bill Gate's face when it is not occluded. The model's performance deteriorates for side views of faces and when the face is occluded or motion blurred.

3 Related Works

We heavily referenced literatures discussed in class for this project. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning (2016)" by Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi provides the groundwork for our model used in our work. We also studied the original Resnet paper, "Deep Residual Learning for Image Recognition (2015)" by Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun to better understand the model's performance.

4 Methods

We implement CNN network with transfer learning to achieve face recognition and classification. Transfer learning is a technique where we can use a model pre-trained for one task, and repurpose it on a new and related task to achieve higher performance than training a model from scratch.

In this project we choose Inception ResNet v1, which is a hybrid network inspired both by Inception and the performance of ResNet. We use a version of the model pre-trained on VGGFace 2, a large face database, as the input to our training. All layers of the model are kept unchanged except the last fully connected layer, which is trained on a dataset of approximately 1000 images labeled either "bill_gates" or "not_bill_gates".

Two different loss functions, the standard PyTorch cross entropy loss for classification, and BCEWithLogitsLoss are used. BCEWithLogitsLoss is logistical loss layers on top of a binary cross entropy loss, enhanced with better numerical computational stability. The model trained with BCEWithLogitsLoss is chosen for video face recognition tasks due to its superior validation accuracy.

Then, we use the DNN face detection model to detect faces in a video. We extract detected faces as images, and input the images to our trained face classification model. Finally, we output the bounding boxes for faces and the prediction of our face classification model in a video.

5 Experiments

We experimented with two different loss functions and different training dataset. We first tried using 600 images of 5 different celebrities from the CelebA dataset as training set, split between 20% "bill_gates" images and 80% "not_bill_gates" images. Another 300 images of similar composition are used as the validation set. After many epochs of training, we were only able to achieve roughly 65% validation accuracy.

We then increased the number of images to about 1000 and achieved 75% validation accuracy. However, when tested on real world videos, the model almost always failed to recognize Bill Gates. We believe this is due to the fact that we used cropped face images from CelebA, and all five celebrities we chose are white males, which lead to biases in the model.

Therefore, we use a different dataset, Flickr-Faces-HQ, to expand our training data. We include an extra 500 images of faces from people of different ages, races, and genders for the "not_bill_gates" class. With the new training set, we achieved 79% best validation accuracy after 27 epochs. Despite relatively small improvement in accuracy, the model performed significantly better in videos, likely due to the better diversity of the Flickr-Faces-HQ dataset.

Finally, we experimented with a different loss function. Because we are only concerned about a binary classification task, we changed the cross entropy loss to PyTorch's BCEWithLogitsLoss, which is optimized for binary classification. This helped us increase the validation accuracy to 89%.

Figure 1: Semi-successful example of our pipeline working in a frame of a video. Note that the side face of Bill Gates is not detected as a face.

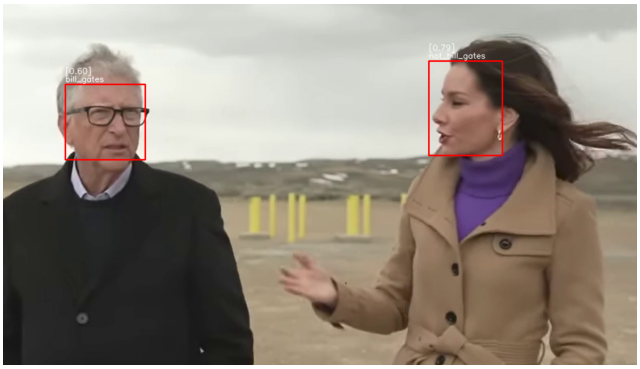

Figure 2: Successful example of our pipeline working in a frame of a video.

We compare our results with other facial recognition models used to detect the face of a particular individual. These include CNN based models similar to ours, but more models are based on feature extraction and similarity matching. As shown in the table in Figure 4, most models achieve better accuracy performance than our model, likely due to their specific find tuning on facial features.

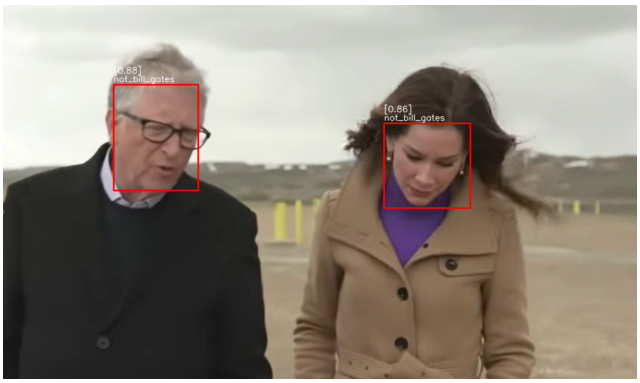

Figure 3: A failed case of our algorithm. The failure happens likely due to Bill Gate's face showing only its side.

Author/Technique Used		Database	Matching	Limitation	Advantage	Result
Local Appearance-Based Techniques						
Khoi et al. [20]	LBP	TDF CF1999 LFW	MAP	Skewness in face image	Robust feature in fontal face	5% 13.03% 90.95%
Xi et al. [15]	LBPNet	FERET LFW	Cosine similarity	Complexities of CNN	High recognition accuracy	97.80% 94.04%
Khoi et al. [20]	PLBP	TDF CF LFW	MAP	Skewness in face image	Robust feature in fontal face	5.50% 9.70% 91.97%
Laure et al. [40]	LBP and KNN	LFW CMU-PIE	KNN	Illumination conditions	Robust	85.71% 99.26%
Bonnen et al. [42]	MRF and MLBP	AR (Scream) FERET (Wearing sunglasses)	Cosine similarity	Landmark extraction fails or is not ideal	Robust to changes in facial expression	86.10% 95%
Ren et al. [43]	Relaxed LTP	CMU-PIE Yale B	Chisquare distance	Noise level	Superior performance compared with LBP, LTP	95.75% 98.71%
Hussain et al. [60]	LPQ	FERET/ LFW	Cosine similarity	Lot of discriminative information	Robust to illumination variations	99.20% 75.30%
Karaaba et al. [44]	HOG and MMD	FERET LFW	MMD/MLPD	Low recognition accuracy	Aligning difficulties	68.59% 23.49%
Arigbabu et al. [46]	PHOG and SVM	LFW	SVM	Complexity and time of computation	Head pose variation	88.50%
Leonard et al. [50]	VLC correlator	PHPID	ASPOF	The low number of the reference image used	Robustness to noise	92%
Napoléon et al. [38]	LBP and VLC	YaleB YaleB Extended	POF	Illumination	Rotation + Translation	98.40% 95.80%
Heflin et al. [54]	correlation filter	LFW/PHPID	PSR	Some pre-processing steps	More effort on the eye localization stage	39.48%

Figure 4: Synopsis of accuracy performance of various facial recognition models. Source: Kortli Y, Jridi M, Falou AA, Atri M. Face Recognition Systems: A Survey. Sensors (Basel). 2020

6 Analysis and Potential for Future Work

We implemented a binary face recognition model with relatively high accuracy that can be directly applied to videos. Our model has a validation accuracy of 89%, and is able to recognize most instances of the face of Bill Gates in our test video clips. We believe the data pipeline and model we implemented can aid the labeling of facial data in future machine learning projects, where the labeling of a particular person in a video dataset is helpful. Our pipeline can also be used for low-risk smart home applications, such as a facial wakeup call for AI assistants (similar to the audio "hey Siri"). However, the current accuracy of our model makes it less capable in security related tasks that require an ultra-low false positive rate.

While our particular work used publicly available datasets, we acknowledge that deploying our pipeline to specific use cases require further training on private facial image data. Social considerations including privacy, consent, and fairness are important when using our pipeline for non-public individuals. Therefore, transparent data practices and clear contract between the machine learning algorithm provider and the end user are pivotal in ensuring an ethical and legally compliant application of our pipeline.

We note that our model has the following weaknesses:
1. When face is blurred or occluded, the recognition accuracy significantly decreases.
2. The model's performance is worse among similar faces of white males – it often misclassified white male faces, while mostly correctly classifies female faces and faces of other races.

We believe we should continue this work in the following aspects:
1. Further increase the diversity of our training data to increase accuracy and decrease bias.
2. Use other metrics, such as FPR, FNR, to evaluate the performance of the model.
3. Use asymmetric loss functions to penalize false positives more than false negatives, due to the requirement of most face recognition tasks.
4. Experiment with multi-class classification to recognize the faces of multiple known individuals.