Classifying Genuine Face images from Disguised Face Images

Junyaup Kim
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, South Korea
yaup21c@g.skku.edu

Siho Han
Department of Applied Data Science
Sungkyunkwan University
Suwon, South Korea
siho.han@g.skku.edu

Simon S. Woo
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, South Korea
swoo@g.skku.edu

***Abstract*—Detecting fake or disguised face images become much more challenging due to the significant advancements made in machine learning, computer vision, and image processing techniques. In addition, due to the rise of various DeepFakes, fake images can be maliciously used to attack individuals and deter true information. Therefore, it is crucial to building a classifier that accurately distinguishes an individual from different or similar persons. In this preliminary work, we aim to detect a target person's face from different similar individuals, Doppelgängers, leveraging the dataset from Disguised Faces in the Wild (DFW) 2018. We use well-known off-the-shelf face detection classifiers, such as ShallowNet, VGG-16, and Xception to evaluate the classification performance. In order to further improve the detection performance, we apply data augmentation. Our preliminary result shows that the Xception model can classify one from different individuals with a 62% accuracy.**

***Index Terms*—DeepFakes, Fake Image Detection, Doppelgänger, Disguised Face in the Wild (DFW)**

Due to the significant advancements made in deep learning, creating and generating highly realistic new images have been much easier than before. At the same time, this can pose a significant threat because these algorithms and techniques can be misused to forge real persons and create fake information. The recent rise of various DeepFakes [1], [2], including DeepNude [3], clearly demonstrates how these deep learning techniques can be misused. Therefore, it is crucial to develop a classifier to detect fake or different individuals from the original person. As a first step toward detecting fake faces "in the wild", we aim to develop a classifier to distinguish the target individual from a set of other similar people (Doppelgängers). We leverage the dataset from DFW2018 [4], [5], which contains 2,403 normal face images, 4,814 disguised face images, and 4,440 similar imposters. As a preliminary study, we evaluate how easy and difficult it is for the best performing state-of-the-art off-the-shelf (OTS) image classifiers to distinguish genuine individuals and imposters.

In this preliminary study, we construct binary classifiers using the well-known ShallowNet, VGG-16, and Xception to directly distinguish the genuine images from the imposter images, which is a different task from that in the previous research [5]. We also use different training and test dataset ratios to compare the performance. Further, we apply data augmentation to improve the overall performance. The result for our preliminary work, shown in Table I, indicates that well-known OTS algorithms, including ShallowNet, which performed well on Generative Adversarial Networks(GAN)-generated images, are not good at classifying genuine and imposter images; the overall detection performance was low across different approaches. Meanwhile, Xception achieved the highest accuracy (62%) with data augmentation. Therefore, our preliminary work suggests that deeper OTS classifiers with additional optimizations can possibly improve the performance of detecting fake images.

This research was supported by Energy Cloud R&D Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (No. 2019M3F2A1072217). This work was also supported by the Basic Science Research Program through the NRF funded by the Ministry of Science, ICT (No. M2017R1C1B5076474).

I. Background and Dataset Description

In order to detect the subtle differences between an individual and his or her Doppelgängers, we define the following two classes for our binary classification task: the class 'Genuine', comprised of normal, validation, and disguised face images, and the class 'Imposter', comprised of impersonators' face images. This task is particularly challenging, given that unintentional disguises using disguise accessories might hinder the face recognition between the face images of the same subject, thus increasing the intra-class variability. On the other hand, imposters who impersonate other subjects intentionally further complicate the problem by lowering the inter-class variability. In that regards, the DFW dataset, proposed by Kushwaha et al. [4] and Singh et al. [5], are highly relevant for the problem we are tackling in this paper. The DFW dataset is comprised of 11,157 face images corresponding to 1,000 subjects: images for 925 subjects are from the Internet, and those for the remaining 75 are from the IIIT-Delhi Disguise Version 1 Face Database. Unlike existing disguised face datasets, which are mostly constructed in controlled settings, the novel DFW dataset is a better representation of the real-world scenario, because it also contains impersonators' images. Additionally, the images belonging to the majority of the dataset, i.e. 925 subjects, correspond to some of the most popular celebrities in the world and demonstrate unconstrained physical variations and disguise accessories such as hairstyle, facial hair, make-

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

Fig. 1: Example data of Owen Wilson from DFW 2018 [4]. The images in the black and yellow box are the normal and validation face images, respectively. The images in the green box are disguised face images, and those in the blue box are impersonators' face images. Note that all the images in the above row (normal, validation, and disguised face images) are of Owen Wilson, whereas those in the bottom row (imposters' face images) correspond to different individuals.

up, glasses, hats, masks, etc. Other variations include different lighting, poses, facial expressions, ages, genders, and image quality. The train and test sets from the DFW 2018 competition are provided as train and validation sets, respectively, for the DFW 2019 competition. Each set contains four types of images as shown in Figure 1: (1) normal, (2) validation, (3) disguised, and (4) impersonator face images. One set of normal, validation, and disguised face images pertains to the same person, whereas impersonators refer to other people who resemble the person regardless of their intention.

II. Proposed Approach

The overall descriptions of our approach is shown in Fig. 2. First, Fig. 2.(a) shows the genuine and imposter input face images of Oprah Winfrey; they are cropped using face coordinates as shown in Fig. 2.(b). Next, the cropped images are augmented via horizontal and vertical flipping as shown in Fig. 2.(c). A more detailed data augmentation process using Zooey Deschanel's face image is illustrated in Fig. 3. After this step, the face images are classified as either genuine or imposter using three different models, ShallowNet, Xception, and VGG-16 as shown in Fig. 2.(d). We defined two classes for our binary classification task, 0 and 1, to denote the genuine (normal, validation, and disguised face images) and imposter (impersonator face images) groups, respectively.

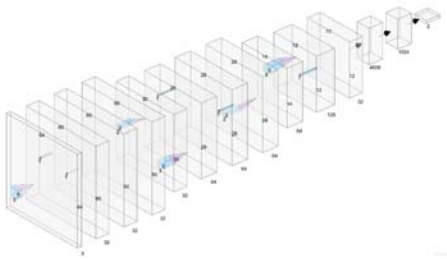
To compare the performance of classification for different dataset sizes, we manually constructed three types of data. All the images are cropped using the face coordinates from the DFW dataset with different training, validation, and test set ratios to observe the performance with respect to varying data size. We experiment with the following three datasets:

- **Dataset 1. Non-augmented data with train/test ratio = 1:1 (N,1,1)**: This dataset contains 3,386 train images, 3,885 validation images and 3,886 test images without any data augmentation.

(a) Original Dataset

(b) Cropped using Face Coordinates

(c) Data Augmentation



(d) Deep learning models used for classification.

Fig. 2: Our proposed approach.

- **Dataset 2. Augmented data with train/test ratio = 1:1 (A,1,1)**: Two types of data augmentation, horizontal and vertical flip, are applied to all images of the first dataset, resulting in 10,158 train images, 11,655 validation images and 11,658 test images.
- **Dataset 3. Augmented data with train/test ratio = 3:1 (A,3,1)**: 5,000 images that are randomly selected from the validation and test set in the second dataset are added to the train set, resulting in 20,158 train images, 6,655 validation images and 6,658 test images.

The above datasets are used to train three models, ShallowNet, Xception, and VGG-16, at a learning rate of 0.001 and for 15 epochs. The test accuracy is then calculated using the test set. We applied data augmentation to flip all original images horizontally and vertically, hence a three-fold increase of the dataset size (original image + horizontally flipped image + vertically flipped image for all images).

(a) Original image (b) Horizontal Flip (c) Vertical Flip

Fig. 3: Data augmentation example of Zooey Deschanel's face image.

III. Preliminary Results

With a train/validation/test set ratio of 1:1:1, the test accuracy increased for both ShallowNet and Xception through data augmentation (see Table I): Using Dataset 1, the test accuracy was 51% for ShallowNet and 55% for Xception. On the other hand, using Dataset 2, it increased by 2% for both ShallowNet (53%) and Xception (57%). Moreover, the train/validation/test set ratio adjustment led to different results for all three models: the test accuracy decreased by 2% for ShallowNet (51%), but increased by 5% for Xception (62%). However, the use of different datasets did not affect the performance of VGG-16: the test accuracy remained at 54% across all three datasets 1, 2, and 3. Some possible reasons for this are: (1) VGG-16 simply may not be suited for distinguishing genuine from imposter face images. (2) VGG-16 is a heavy model consisting of deep layers, so the size of the DFW dataset used for the task described in this paper may not have been sufficient and the model probably requires more data via other, carefully selected data augmentation techniques. Our result shows that Xception, with data augmentation and more training data, is able to achieve the highest performance due to its deeper layers, which can capture the subtle differences between genuine individuals and imposters.

IV. Discussion and Future Work

Using the DFW2018 dataset, our preliminary work shows that the classification of genuine and imposter face images is a challenging problem. However, the increase of the training dataset size via data augmentation can help improve the classification performance of Xception. Interestingly, ShallowNet, which performed well in detecting GAN-generated images, did not achieve high test accuracy. We hypothesize that ShallowNet is not well suited for detecting highly detailed face image features from the DFW dataset due to the small number of layers in the network. Also, the performance of VGG-16 was not enhanced after increasing the dataset size 3-fold. This suggests that we may need to apply more complex data augmentation techniques to train the heavy model.

TABLE I: Classification performance with different train/test set ratios and data augmentation.

Classifier (Dataset)	Accuracy
VGG (Dataset 1 (N:1:1))	54%
VGG (Dataset 2 (A:1:1))	54%
VGG (Dataset 3 (A:3:1))	54%
ShallowNet (Dataset 1 (N:1:1))	51%
ShallowNet (Dataset 2 (A:1:1))	53%
ShallowNet (Dataset 3 (A:3:1))	51%
Xception (Dataset 1 (N:1:1))	55%
Xception (Dataset 2 (A:1:1))	57%
Xception (Dataset 3 (A:3:1))	62%

We plan to further improve the performance with SiameseNet-based approaches or explore the pre-training and transferability between the genuine and disguised face images to better distinguish them from imposter face images. Future work will include collecting and experimenting with different disguised classifiers with more dataset.

References

[1] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[2] "Cable news network – "when seeing is no longer believing"," https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/, 2019, accessed: 2019-08-25.

[3] "Official deepnude algorithm source code," https://github.com/lwlodo/deep_nude, 2019, accessed: 2019-08-25.

[4] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa, "Disguised faces in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1–9.

[5] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa, "Recognizing disguised faces in the wild," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 97–108, 2019.