Augment Artificial Data to Aid Face Recognition

Sarantos Tzortzis (11863331) **Juan Buhagiar** (11576014) **Tomas Fabry** (11868414)
Yiangos Georgiou (11807288)[1]
https://github.com/sarantinio/Face-Recognition.git

June 25th, 2018

Abstract

In this paper, we explore the possibilities of increasing the robustness, reliability and most importantly, accuracy, of neural networks when applied to areas with limited data. Nowadays, to introduce variations in image datasets, simple flipping and rotating of images is done. However, we wanted to go beyond that and find out whether augmenting the dataset by generating new images can improve the accuracy of a face recognition model after re-training. We use a state of the art face recognition model and two data augmentation methods to augment a limited sub-sampled dataset. We discuss their effectiveness and how they can be used to solve the problem of algorithm reliability and limited data.

1. Introduction & Related work

Deep learning methods have become very popular in the recent years, namely, architectures such as Convolutional Neural Networks(CNNs) have become popular in the computer vision domain. CNNs have been applied in many computer vision sub-domains such as object detection and classification (Ren et al., 2015), (Krizhevsky et al., 2012), scene classification (Zhou et al., 2014), semantic segmentation (Long et al., 2015), pose estimation (Toshev & Szegedy, 2014) and face recognition (LeCun et al., 2015). The success of CNNs can be attributed to multiple factors, such as the increasing amount of annotated datasets, GPU implementations of libraries used for CNNs and improvements in the architectures of CNNs.

Initial work in face recognition was done with handcrafted features in constrained environments which achieved promising results (Sim et al., 2002). These methods failed drastically when introduced to face recognition datasets in uncontrolled environments (Huang et al., 2007). An initial improvement to these handcrafted features was metric learning, which tries to learn a transformation such that objects with the same labels are close together while objects with different labels are further away (Guillaumin et al., 2009).

Although metric learning showed improvements to face recognition in uncontrolled environments, it was still hard to generalize for changes in the pose, illumination and facial expression. Like many other computer vision tasks, researches investigated the possibility of changing the problem to a data-driven supervised learning task using CNNs. These methods achieved state-of-the-art results on the LFW (Labelled Faces in the Wild) dataset (Huang et al., 2007). These methods, namely DeepFace (Parkhi et al., 2015) and FaceNet (Schoff et al., 2015), were trained on 4 and 200 million images respectively.

Unfortunately, it is not always feasible, or even possible, to obtain such large amounts of data. This can be because of practical reasons such as costs, time, privacy regulations or very specific domains such as night vision faces. This problem has become evident to many researchers in the field and that is why alternative methods to solve this have been investigated.

Two solutions to the data scarcity problem that have been used in the past are the generation of synthetic data and data augmentation. In this paper, we will focus on the data scarcity problem starting from an annotated dataset of 25K images of 9K+ identities which has 3 images per identity. This dataset will be used to train a baseline model, then we will enlarge the dataset by augmenting the original images to produce new, artificial ones, with the hopes of training more robust models. Thus, our primary goal is to verify whether artificially augmented datasets can be used to improve the accuracy of the model.

[1] Supervisor: dr. Sezer Karaoglu.
Sarantos Tzortzis <sarantos_tzortzis@icloud.com>
Juan Buhagiar <juanbuhagiar@gmail.com>
Tomas Fabry <tomasfabry1@gmail.com>
Yiangos Georgiou <ygeorgiou12@gmail.com>.

Project AI 2018, MSc Artificial Intelligence, Univ. of Amsterdam

2. Addressing Algorithm Reliability

The secondary goal of this paper is to address the problem of reliability and robustness of face detection algorithms. More data generally leads to better results. More varied data generally leads to even better and more generalized results. Current state-of-the-art algorithms such as facenet are trained on images of a certain type, more specifically lacking occlusions such as glasses, beards, hats or variations in skin tone, hair color, and others, within 1 person. It is common sense that the first method to hide a person's identity from automatic face detection systems is to put on a hat, glasses and perhaps grow a beard if possible. By artificially augmenting the original dataset with data that introduces variations and occlusions, the algorithm can be trained to be more robust and to learn the features that constitute face similarities, while also taking into account occlusions, color variations etc.

3. Datasets

In this project we have used two publicly available datasets for studying the problem of unconstrained face recognition, the VGGFace2 (Cao et al., 2017) and LFW(Labelled Faces in the Wild)(Huang et al., 2007).

The VGGFace2 contains 3.31 million images of 9131 identities containing an average of 362.6 images per identity. The images were downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity, and profession. Since we are investigating the relationship between data augmentation and face recognition when using limited data, we sub-sampled the VGGFace2 dataset to 3 images per identity, leading to 27,393 images. We have used this dataset for the baseline experiments and to generate the augmented datasets discussed later.

The LFW dataset is used for evaluation and contains 13233 images of 5749 identities where 1680 identities have two or more images. The only constraint on the images of faces is that they were detected by the Viola-Jones face detector. This dataset serves as a set of example images that are taken under imperfect conditions in real life and is used for benchmarking modern face detection algorithms.

4. Data Augmentation

We use a pre-processing method called "One-to-many augmentation" (Wang & Deng, 2018). More specifically, we generate many images of the pose or features variability from a single image to enable deep networks to learn pose-invariant and noise-invariant representations.

As described in the introduction, it is extremely time-consuming and expensive to create a large dataset. Surveillance cameras, for example, may only have spot 2 to 3 images of a person which is not enough to train a network to identify him. 'One-to-many augmentation' can mitigate this time-consuming and expensive procedure.

We use two methods of augmentation: a) using the StarGAN neural network model(Choi et al., 2017) to introduce variations in hair color, skin color, age and arched eyebrows, and b) generating a 3d model from the original 2d image in order to be able to rotate it and put a 3d model of glasses on the 3d model of the face, in the end projecting the augmented 3d model back to 2d to get the new image.

4.1. Augmentation Method 1: StarGAN

Generative Adversarial Networks(GANs) have been introduced by Ian Goodfellow (Goodfellow et al., 2014). They have since found their main use to be changing or generating new images of some kind. They use two networks: a generator which generates new images, given an original image and noise or attributes, and a discriminator, which tries to predict whether the produced image belongs to the original dataset or not (i.e evaluates them for authenticity). Authors of (Choi et al., 2017) have been able to architecture a network to generate faces with different attributes, given an input face. We use the best performing pre-trained model to generate multiple images from each image of a face.

4.1.1. StarGAN: Attribute Selection

We want as many variations in appearance or expression as possible. However, certain attributes of StarGAN seem to generate images that are not realistic enough or change the inherent facial features such that it is hardly recognizable whether the new image is of the same person, which is the goal. In other cases, the images are distorted or blurry. For example, attributes such as *wearing_necklace* or *wearing_lipstick* do not even try to transform the image to include a necklace or lipstick, but instead, feminize the image. This is due to the inherent imbalance in the training dataset for StarGAN, where almost every image labeled as including a necklace or lipstick, is an image of a woman. The network has therefore learned to transform an image to look more feminine when asked to augment with a necklace. Furthermore, the results are very blurry and distorted, as can be seen in Figure 1

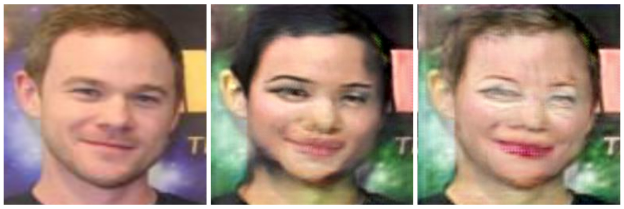

Figure 1. Original - Necklace - Lipstick.

The original dataset *MS-Celeb-1M* that StarGan was trained on consists of around 80% women. This gender imbalance leads to feminine results.

We have found attributes that perform reasonably well for both genders. The attributes *Arched_Eyebrows*, *Blond_Hair*, *Brown_Hair*,*Pale_Skin*, *Age* appeared to produce more consistent results, albeit still not perfect.

The Images 2 and 3 indicate the results using these on one male and one female image. In both cases, the first image is the original.

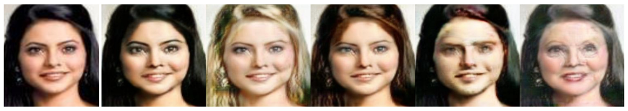

Figure 2. Original - rched Eyebrows - Blond Hair - Brown Hair - Pale Skin - Age.

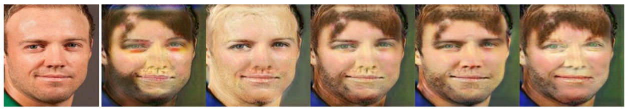

Figure 3. Original - rched Eyebrows - Blond Hair - Brown Hair - Pale Skin - Age.

The results are still better for the female images, as the images with males are always more distorted. Both images though could be useful in terms of train our network to be more robust to noise. It is possible that even though the images are distorted and unclear, the features that are important for the network to distinguish between people might be preserved. We would, therefore, like to find out whether the augmented dataset with these selected attributes is going to improve in terms of accuracy.

4.2. Augmentation Method 2: 3D face reconstruction

The second augmentation method is 2D to 3D face reconstruction. We enrich our benchmark data-set by generating per-subject appearance variations. Using an application interface by 3duniversum(URL: http://yourface.3duniversum.com), we create a 3D estimation from each image and use it to create poses in different angles with additional face accessories. This 3D model is then projected back to 2D and used as new data.

We generate 5 images for each real image in the dataset. More precisely, we use 5 different poses (up, down, left, right and center) with 15 degrees deviation from the original pose. We add glasses with 1 black material as additional accessories. An example can be seen in Figure 4

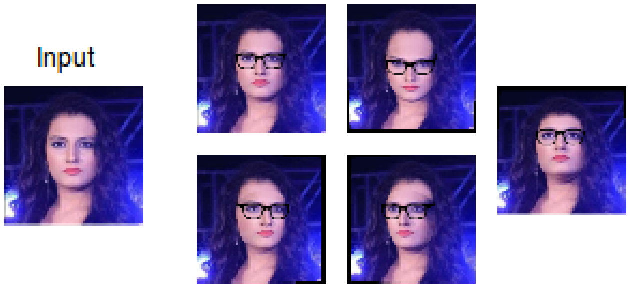

Figure 4. 3D sample result. Left is the input image, on the right is the generated images from different poses.

5. Pipeline

We use a pipeline that follows the one that presented in (Schoff et al., 2015). We give an image of a person as an input. The image is then aligned so that the face is centered and most of the background is cut off. Then we use our two data augmentation methods to generate new images, as illustrated in Figure 5. Once we have the new datasets, we can re-train the facenet network. Facenet first extracts features from each image to create an embedding. This is a high-level feature representation of the image. Using these embeddings, we can either train a linear SVM classifier to retrieve the ID of the person or we can feed in 2 images, get 2 embeddings and decide if it is the same person or not, depending on the distance between them.

5.1. Evaluation

Face Recognition can be categorized as face verification and face identification. In either scenario, a set of known subjects(training set) is initially enrolled, and during testing, new subjects (test set) is presented. In our experiments, we monitor the face verification accuracy for the purposes of comparing our methods.

Face verification computes the one-to-one similarity between the two input face images to determine whether the two images are of the same subject. The procedure that we follow is to generate image embeddings from each image using a pre-trained model. Then, given the two embeddings, the pair is annotated as the same or different person based on the euclidean distance between the two embeddings and a threshold.

As our test set, we use solely Labelled Faces in the Wild dataset, as described in the dataset section.

6. Experiments

For our experiments, we used two already implemented state of the art models for facenet and StarGan written in python

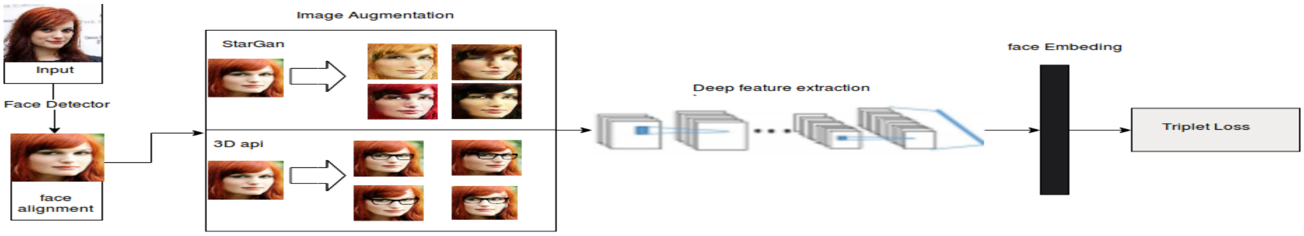

Figure 5. Deep face recognition pipeline.

and we used DAS4 to run the experiments on TitanX GPU. Each training took from 20 to 30 hours, saving checkpoints and monitoring the progress and convergence.

We have sub-sampled our original dataset *vggface2* to contain only 3 images per person, with roughly 9000 identities. Following the process described in the pipeline section, we have generated 2 more datasets, one with each augmentation method. The ratio of real to fake images is 1:5 for the first two experiments and 1:1 for the last experiment, where only the original pose with added glasses was used for each image. We have then re-trained the facenet model on each of these datasets and evaluated on LFW.

7. Results

We develop three different data-sets using the augmentation methods that were described previously in the paper. Our results are illustrated in table 1. As we can see, the benchmark (dataset that contains only 3 real images) performs very well and reaches a verification score of ***84.5%***. On the other hand, StarGan dataset(3 real and 15 generated images) gives the lowest accuracy with ***74.3%***, while 3D dataset performance is also high and reaches ***83.8%***, which is very close to benchmark's accuracy. The last dataset (containing the same number of real and fake images) outperforms all our previous experiments and results in an accuracy of ***88.7%***.

Table 1. Face Verification accuracies on various data sets.

DATA SET	VERIFICATION
BENCHMARK 3	84.5± 0.021
STARGAN DATA	74.3± 0.014
3D DATA	83.8± 0.023
3D DATA NO POSE (1:1 RATIO)	88.7± 0.015

8. Discussion

In terms of the augmentation methods, we can observe that in many cases the results were not what we expected, especially with the StarGan approach.

StarGan performance was very poor in many cases, especially for male face images, as we can see in Figures 1 and 3. StarGan outputs are more blurry than the real images and because of that, it also alters facial characteristics such as nose, mouth, and eyes. Moreover, some results are non-realistic. That makes it difficult for our neural network to find strong correlations on facial characteristics between images of the same person. Another issue with the pre-trained model of StarGan is that it was trained with a not diverse data-set(80% females) and it seems to be biased on feminine face images, which makes it perform badly on images with male faces.

Regarding 3D data augmentation, it seems to sustain and alter the image in a more natural way. As it can be seen in Figure 4, the images are very realistic and not blurry. However, we can observe that in some images the geometry of the face was altered despite not rotating the image too much in order to avoid revealing the ”unseen” part of the face.

In terms of different data-sets, StarGan and 3D data performed worst among the others, despite the fact that these contained more images than Benchmark and 3D data-set(1:1 ratio). To further investigate where our model fails we track the pair of images where only one model fails to annotate it with the right label.

In figure 4.1, we observe that StarGan model considers mostly the geometry of the face to decide whether the two face images depict the same identity. As we can see, people with the same pose and similar face geometry are annotated as the same identity (false positive), while the same person with a different pose is annotated as different. An explanation is that our StarGan data contains many fake images with the same pose but different facial characteristics, which consequently forces the deep network to focus mostly on the pose in order to cluster all these images together.

For 3D data model, we have not observed the same issue as for the StarGan model. However, a pattern also arose for this case. This model considers both the geometry of the face and the facial characteristics of a person and provides decent results, but as we can see sometimes if the facial expression is very similar it annotates the two images as the

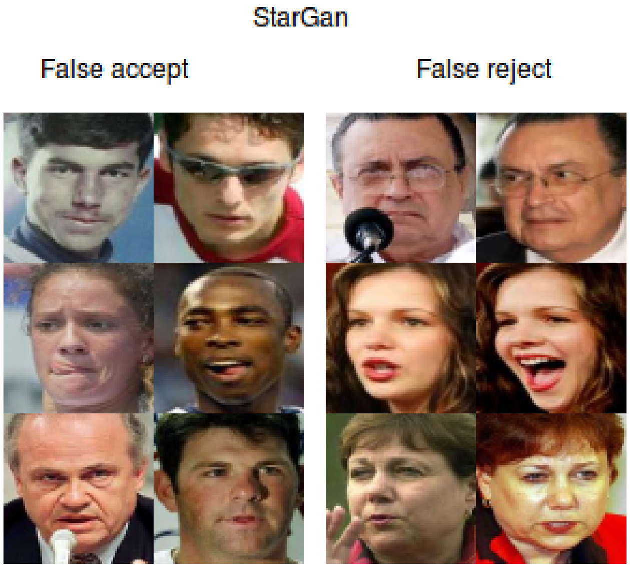

Figure 6. False positives(right) and False negatives(left) for StarGan Data model.

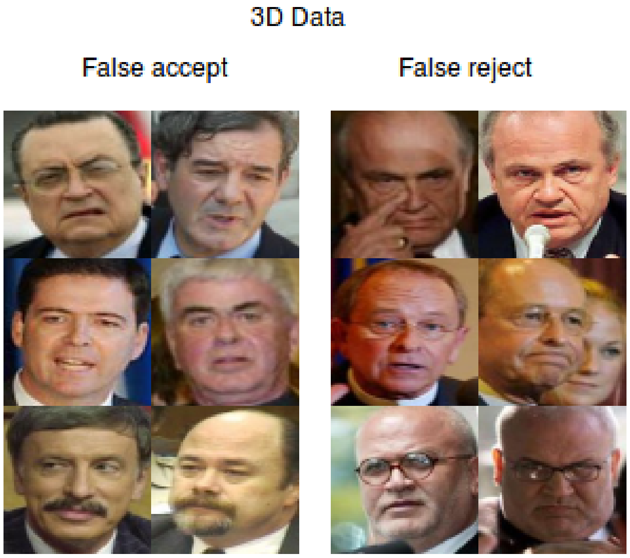

Figure 7. False positives(right) and False negatives(left) for 3D Data model.

same identity while if the facial expression is too different, it leads to the opposite result. Thus, as is illustrated in Figure 7, many false positive arise that share similar expressions. This might be also explained through the dataset because it contains many images with the same facial characteristics with different poses. This makes the dataset biased towards the facial characteristics.

Another factor that seems to affect the accuracy is the ratio between real and fake images. Using too many fake images leads to worse results. As we can observe in table 1, the dataset with the ratio 1:1 of real to fake images provides the best accuracy. Perhaps by using more variations such as accessories or slight and realistic facial feature changes, the number of fake images can be higher.

9. Future Work

There is a variety of future experiments to be done that could improve the results. Using meta-data such as gender in our original dataset could be used to pick image-specific attributes for the StarGan augmentation so that we can generate a more consistent and realistic dataset.

Furthermore, using a 1:1 ratio with the StarGAN augmentation again (as we did with the last experiment with the glasses) could be tested. For this experiment the *blond_hair* is the recommended attribute as it was the most consistent one.

At last, expanding the 3D reconstruction to include more facial occlusions and items such as has, beards, mustaches etc. could potentially improve the accuracy even further. Moving a black rectangle over the image to occlude the face could also be used to test and see what the minimal surface area of the face is to reach decent accuracy when it comes to recognizing the identity of a person.

10. Conclusion

To conclude, in this work we illustrate the importance of a diverse dataset in terms of poses, facial characteristic, and expressions for face recognition tasks. Moreover, it is important to say that more data does not always result in accuracy increase, for face recognition tasks, diversity in poses and features is very important.

Secondly, we compare StarGan and 3D reconstruction approach for data augmentation both for output quality and also for Face Recognition accuracy. The 3D reconstruction approach seems to be more promising as it generates clearer and more realistic images and also is more beneficial for training instead of the StarGan data as we can observe in 1.

We have managed to slightly out-perform the baseline by using a 1:1 ratio of real to generated images using 2D to 3D reconstruction by generating different poses and putting on glasses as the occlusion.

And finally, we have found that even the current state of the art GAN models do not produce sufficient enough results to be used to augment datasets(different from what the original GAN was trained on), to be used for face recognition. Given the positive result of the 2D to 3D reconstruction method, it is possible that further improvements in GANs might eventually make for a very generalizable method for artificially enlarging datasets.

References

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017. URL http://arxiv.org/abs/1710.08092.

Choi, Y., aa, and aa. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. Korea University, Clova AI Research,The College of New Jersey, 2017.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. *ArXiv e-prints*, June 2014.

Guillaumin, M., Verbeek, J., and Schmid, C. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, pp. 498–505. IEEE, 2009.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Parkhi, M., Vedaldi, A., and Zisserman, A. Deep face recognition. Department of Engineering Science, Oxford University, 2015.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Schoff, F., Kalenichenko, D., and Phibin, J. Facenet: A unified embedding for face recognition and clustering. Google Inc., 2015.

Sim, T., Baker, S., and Bsat, M. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 53–58. IEEE, 2002.

Toshev, A. and Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.

Wang, M. and Deng, W. Deep face recognition: A survey. Department of Engineering Science, Oxford University, 2018.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pp. 487–495, 2014.