Effects of Glow Data Augmentation on Face Recognition System based on Deep Learning

Jawad Rasheed
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0003-3761-1641

Erdal Alimovski
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0003-0909-2047

Ahmad Rasheed
Department of Electrical Engineering
Eastern Mediterranean University
Famagusta, Northern Cyprus
0000-0003-3730-1715

Yahya Sirin
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0001-5331-1804

Akhtar Jamil
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-2592-1039

Mirsat Yesiltepe
Department of Mathematical Engineering
Yildiz Technical University
Istanbul, Turkey
0000-0003-4433-5606

***Abstract*—Biometric artificial intelligence application depends on amount of material on which they are trained. In this paper, we integrated Glow data augmentation technique to diversify the facial images dataset to analyze its effects on face classification and identification system based on Convolutional Neural Network (CNN). In first phase, we trained our CNN with publicly available Labeled Faces in the Wild (LFW) database and evaluated the proposed system, which achieved accuracy of 92.2%. In second phase, we diversified LFW dataset with Glow method and then trained our CNN network. The experiment results shows that Glow data augmentation improved the accuracy of proposed network to 93.6%.**

Keywords—face recognition, glow, CNN

I. INTRODUCTION

In the world of advance technology, a sharp shift towards digital transformation is observed since last decade. Due to numerous amount of digital data production on daily basis, image processing has gained attention in research community. Like other digital transformation, image processing is widely adopted in different sectors having broad applications such as face recognition system for crime identification, object detection for video surveillance system, and text extraction for automatic image indexing.

Among vast domain of image processing, Biometric Artificial Intelligence applications are extensively used for security systems due to its non-invasive process as compared to iris recognition or finger-print systems. As stated by author in [1], the first facial recognition system in 1965, named as man-machine system, required manual marking of various features like nose, eye centers and others, due to limited machine computational capability. It is then fed to machine to automatically calculate the distance between the marked landmarks in images for person identification.

With the advancement in technology and development of high speed processors by tech companies, complex algorithms have been evolved in image processing and artificial intelligence domain. For object detection and classification various techniques have been emerged using machine learning and deep learning algorithms such as CNN, Support Vector Machine (SVM) and many others to accomplish better performance in minute amount of time. The performance of deep learning and machine learning models depend on the size of dataset on which they are trained. Typically, deep learning methods achieve high accuracy rate provided that there is significant amount of dataset to extract various features and learn different landmarks present in image while training.

To create large dataset from scratch is a bit hectic if done manually, therefore scientists developed various data enhancement and augmentation techniques such as Generative Adversarial Networks (GAN) [2], Variational Auto Encoders (VAE) [3], Deep Belief Networks (DBN), and Flow-based generative models [4] to cater the extremely challenging task of reproducing the data from original dataset.

The study in this paper is divided into two stages. As open source facial dataset available online is very limited, whereas as deep learning classifiers require huge amount of dataset for training, therefore; the existing dataset is enhanced in first stage using Glow data augmentation. In second stage, we developed a face recognition system based on deep learning technique, to analyze the effect of data enhancement on network's performance. To accomplish the goal, we downloaded dataset images containing people's faces, LFW from [5], and then applied Glow technique [6], a reversible generative model, to generate realistic face images with slight variation that resembles with the original face. After enhancing the dataset with flow-based generative model, we trained our deep learning mechanism that primarily includes CNN model. Later, its performance is evaluated on unseen data (original and augmented data).

The rest of paper is distributed in following regions. In next section, we outlined the facial recognition related work. Section 3 describes the proposed method for fake face generation and recognition, while experimental results are presented in Section 4. The paper ends with concluding remarks in last section.

II. RELATED WORK

Living creatures has a strong visual and brain system that can easily detect and identify an object in milliseconds, but machines don't have ability to recognize an object or face present in an image. Since last three decades, different methods and models have been introduced to train computers in order to have human like visual recognition system.

978-1-7281-9352-6/20/$31.00 ©2020 IEEE

Generally, the existing machine learning facial recognition techniques can be categorized into traditional and modern methods. Traditional methods such as Eigen faces algorithm outlined in [7] and Fisher vector in [8] are based on geometric approach to recognize the face from its features. The distances between facial features such as eye, eyebrows, and nose are calculated to recognize and identify the person's face.

The author in [9], formed vector containing 16 facial parameters to calculate Euclidean distance for face recognition that achieved an accuracy of 75% on small dataset of 40 images of two different persons. In [10], author used same technique but with geometric vector size of 35, which secured 90% accuracy. The authors in [11] formulated a bigger dataset of 658 images and developed face recognition system similar to previous work. They calculated the distance of manually extracted 35 features and acquired an accuracy of 95%. A comprehensive detail about traditional face recognition methods can be found in [12].

These conventional methods do not perform well in unrestricted environment as facial features were excerpted manually. Contrarily, modern approaches perform well by eliminating problems encountered in traditional methods. A major breakthrough in object classification was discovered in 1997 in which authors in [13] used CNN that automatically extracted the best features. Since its inception, CNN model has been extended in all aspects.

Later, researchers in Facebook lab introduced a DeepFace recognition technique, which was trained on huge dataset of 4.4 million tagged images of 4000 different individuals. They used CNN for feature extraction and accomplished an accuracy of 97.35%. Different variants were introduced afterward which achieved 99.47% accuracy as mentioned in [14]. In [15], Facenet architect was developed by Facebook researchers that mapped facial images directly to an Euclidean space followed by clustering. Face verification and recognition were performed using Facenet embedding as feature vector. The system used 200 million images of 8 million different individuals for training and outperformed all previous approaches by securing an accuracy of 99.63%.

Above studies show that established face identification systems based on deep learning architects required millions of images for training. In the beginning, each image in existing data was augmented manually using machine through color augmentation and position augmentation techniques. In color augmentation approach, contrast, brightness, hue and saturation were adjusted. While in position augmentation, the images are either flipped or rotated, whereas some pictures were scaled or cropped and padding is added to enrich the dataset.

With the passage of time, researchers introduced different data replication and regeneration techniques. As by author in [16], different data augmentation approaches were introduced by applying different hairstyles and glasses templates while adjusting light and exposure on facial images to enhance dataset. Authors in [17] obtained promising results by proposing a novel approach to generate facial images of various expression using Nonlinear Manifold Separator Neural Network (NMSNN).

A new era started when tech researchers from NVidia (graphics card manufacturing company) in [2] introduced generation of synthetic faces using Generative Adversarial Network (GAN). It revolutionized the field of data augmentation and the results were quite impressive as human eye was unable to distinguish between generated high quality synthetic faces and original faces.

Another mechanism emerged as VAE in [3] in which inferring the approximation of latent variables' values improved the lower-bound on log-likelihood of data, hence resulting in new dataset. In [6], authors proposed a flow-based generative model called Glow with an invertible 1x1 convolution. The proposed architecture is an extension of the NICE [4] and RealNVP [18] techniques that improves the log-likelihood to produce realistic looking synthetic facial images.

III. Material And Methodology

In deep learning model, network architecture such as number of layers and classifying parameters have great impact on classifier output. The classifier prediction is based on how the classifier is trained on its dataset. The amount of dataset plays a crucial role in training as model often requires larger dataset. The greater the training data is, the better the learning rate of architect will be.

However, with the increase of dataset, the success rate increases to a certain point but it stops if the diversity of dataset is low. In order to create larger dataset, we used Glow data enhancement technique to generate synthetic facial images as explained in this section below. After data augmentation, the 60% of the generated dataset along with the original dataset is fed to the system to train the classifier. Finally, the deep learning classifier is tested on unseen data.

A. Dataset

For testing and training purposes, we used LFW dataset. It has 13,233 facial images of 5,749 individual persons. Each images of three colored-channels (red, green, and blue) and size of 250 x 250 pixels is labeled with name of person pictured. The dataset is randomly divided into two sets; 60% for training while 40% for evaluation as shown in Table 1.

TABLE I. Dataset For Face Image Recognition

Dataset Type	Instances	Training Set	Testing Set
Original LFW	13233	7940	5293
Augmented (Glow Tech)	52932	31760	21172
Total	66165	39700	26465

B. Data Augmentation

In computer vision, scientists introduced various methods and techniques like DBN, GAN, VAE and flow-based generative models to generate new samples of images. Despite of above mentioned data availability, the data must be diverse enough in order to enhance the network's learning and generalizing capability while training and testing. To overcome the problem of narrow diversity and limited data, we choose Glow architect mentioned in [6] due to its reversible generative model approach which automatically optimizes the exact log-likelihood of data instead of approximation. More importantly, reversible models are memory saving as it takes fixed amount of memory.

In Glow [6], authors proposed 3 steps. In first step an activation normalization layer (Actnorm layer), like batch normalization, used scale and bias parameters for each channel to do affine transformation of activations. The Actnorm layer is then followed by an invertible 1x1

Fig. 1. Augmented samples obtained after applying Glow technique on one of instances of LFW dataset.

Fig. 2. Augmented samples of another sample with Glow technique

convolutional layer to generalize the fixed permutation operation. Finally, an affine coupling layer is introduced to perform reversible transformation to help the training of deep networks by first applying identity function and then splitting the input tensor along the channel dimension.

We diversified LFW dataset with Glow techniques by randomly generating 52932 facial images from the given dataset. We split the augmented dataset into 60%-40% for training and testing respectively as summarized in Table 1. Some random samples obtained after employing Glow on LFW are shown in Fig. 1 and Fig. 2.

C. *Classification Using CNN*

Among many deep learning architects and algorithms, CNN is most famous for complex image classification. The base of CNN is the chain of convolutional layers, pooling layers and dense layers connected together that takes the input image as feature maps and classify it to appropriate label by transforming it to meaningful representation, as depicted in Fig 3.

Each neuron in convolutional layer (see Fig. 4) forms a connection with some neurons of previous layer by assigning weights. The networks takes the image as input in form of matrix that represents pixel values. The model determines the activation of neuron, by calculating weighted sum of all activations 'a' of previous layer connected to this neuron and adding a reasonable bias 'b' for meaningful representation. It then uses sigmoid function to normalize these activations between 0 and 1. The activations of first convolutional layer is represented by function:

$$\sigma\left(\begin{bmatrix} w_{0,0} & \cdots & w_{0,n} \\ \vdots & \ddots & \vdots \\ w_{k,0} & \cdots & w_{k,n} \end{bmatrix}\begin{bmatrix} a_0^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ \vdots \\ b_n \end{bmatrix}\right) = \begin{bmatrix} a_0^{(1)} \\ \vdots \\ a_k^1 \end{bmatrix} \quad (1)$$

The network calculates weights and biases for all neurons in each layer, and generates activation maps for each convolutional layer. A kernel filter convolves across spatial dimensions of input feature map. Rectified Linear Unit performs the job of sigmoid function and resultant is passed to pooling layer for features' dimensionality reduction to boost the system performance.

Later, a dense layer, also known as fully connected layer, takes input representations of all preceding layers, combines it by sending the signal to each neuron in it. At the end classification layer, known as output layer, classify the input data based on probabilities computed by softmax layer.

For this experimental study, we proposed CNN architecture (see Fig. 3) to classify facial images. As facial images are colored (3 channel; red, green blue) pictures with height and width of 250 each, an input layer is defined in the Fig 3 as first layer with size of 250x250x3.

The model consists of 32 convolutional layers, each with ReLU that transforms the input by doing several convolutional operations on the given image for feature extraction. The size of input and spatial output is assured by padding. During scanning along the facial image, a 32x32 size of filter is used in each convolutional layer.

Training is performed using batch normalization layer between each ReLU layer and convolutional layer to speed-up the process. After each convolutional layer except the last convolutional layer, max-pooling layer reduces dimensionalities of feature representation thus increases system robustness.

Later, a fully connected layer in the end gathers all the features learned by previous layers to predict the face accordingly. Softmax activation function normalizes the probability of final output to one in fully connected layer. The

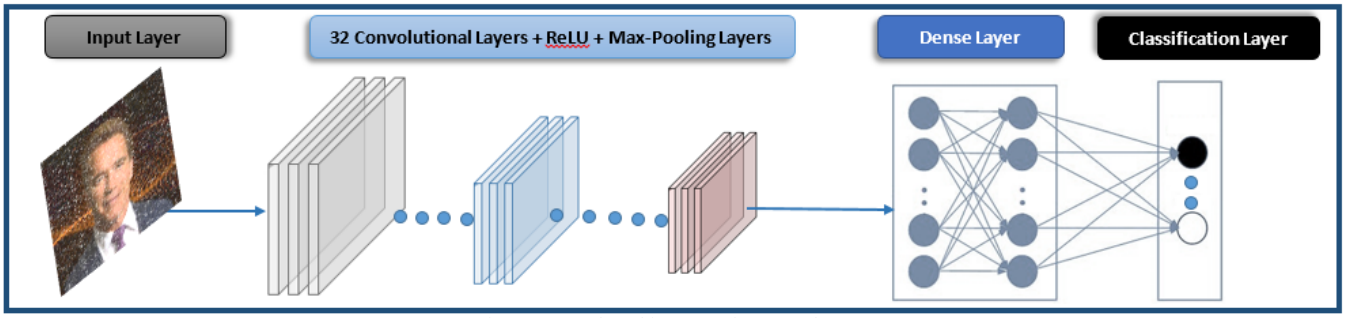

Fig. 3. Network architecture of proposed CNN.

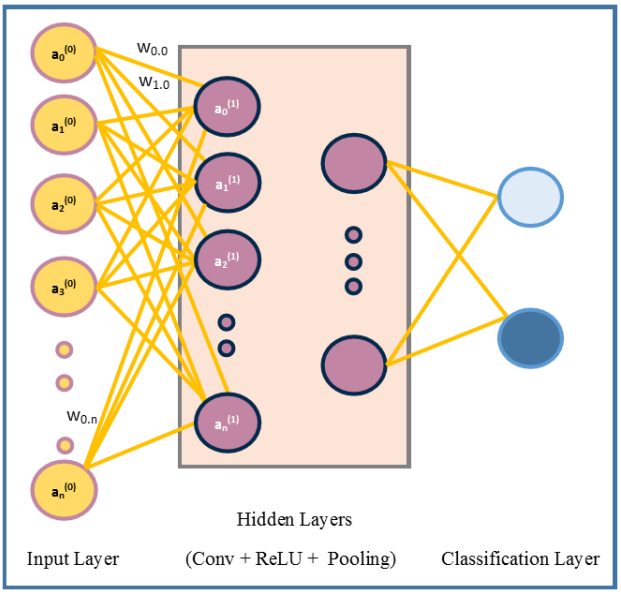

Fig. 4. Neurons working and connections in CNN.

output layer uses the probability outcomes of soft-max layer to identify the face.

IV. Experimental Results

The aim of the research is to analyze the effect of incorporating Glow augmentation technique on proposed CNN model. The LFW dataset of facial images is enhanced using Glow technique and later trained the proposed network to in two phases that can directly predict the input face to its corresponding label. In first phase, the network is trained with original LFW dataset only and then evaluated, while in second phase the network is trained with LFW dataset along with augmented dataset which is then tested accordingly.

The experiment is performed using Python 3.6 environment on Jupyter Notebook with Keras and Tensor flow GPU packages. The hardware constitutes of NVidia GeForce 410M (512MB RAM) for GPU acceleration with Intel® Core™ i5-2410M CPU @ 2.30GHz and 8GB RAM.

In first phase, to avoid overfitting due to lesser amount of data, LFW data is divided into 10-folds and the proposed CNN network is trained with this dataset. The model obtained an

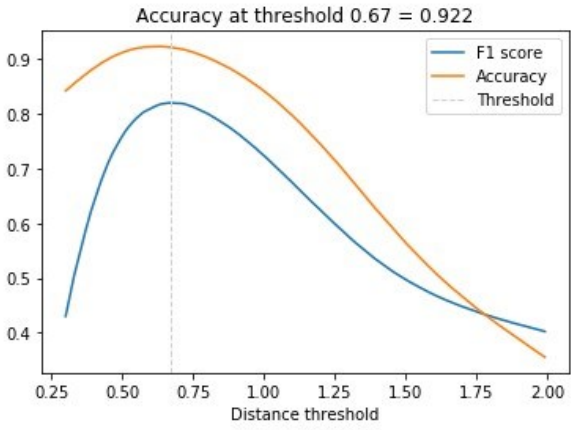

Fig. 5. CNN accuracy trained on LFW dataset only.

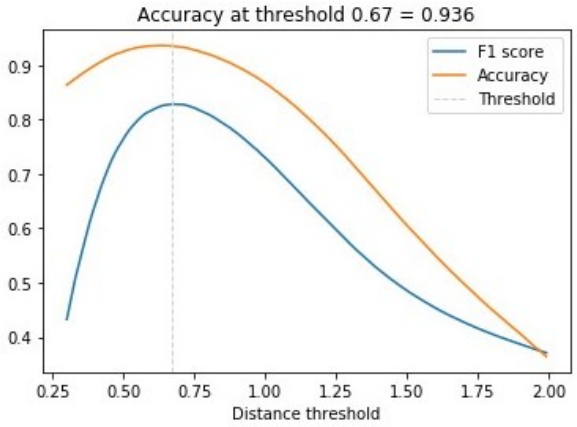

Fig. 6. CNN network accuracy, trained over augmented dataset.

accuracy of 92.2% (see Fig. 5). While in second phase, same folds of LFW and its augmented images are fed to CNN network for training and achieved an overall testing accuracy of 93.6% as outlined in detailed graph shown in Fig. 6. The experimental results shows that generative models can be utilized to improve generalizability through data augmentation.

V. Conclusion

This work builds upon idea of data augmentation using Glow methodology to diversify publicly available facial images dataset (LFW database) and analyzed its effects on our proposed CNN network. First, CNN network is trained with original LFW dataset in form of 10-folds and evaluated with test dataset. In second stage, we enhanced the same LFW 10-folds sets with Glow and trained our CNN with it. A total of 52932 images were generated, out of which 60% is used for training while 40% for testing. Although, the experimental result shows slight improvement in face identification using CNN, there is still capacity to improve the classification by testing with different number of convolutional layers in CNN.

References

[1] K. D. Leeuw and J. Bergstra, *The History of Information Security: A Comprehensive Handbook*. 2007.

[2] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," *Adv. neural Inf. Process. Syst.*, pp. 2672–2680, Jun. 2014, [Online]. Available: https://arxiv.org/abs/1406.2661v1.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.*, no. Ml, pp. 1–14, 2014.

[4] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc.*, vol. 1, no. 2, pp. 1–13, 2015.

[5] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "LFW Face Database : Main," 2007. http://vis-www.cs.umass.edu/lfw/index.html#download (accessed Dec. 20, 2019).

[6] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. 2, pp. 10215–10224, 2018.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul.

1997, doi: 10.1109/34.598228.

[8] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, vol. 4, no. 2, pp. 586–591, doi: 10.1109/CVPR.1991.139758.

[9] T. Kanade, “Picture Processing System by Computer Complex and Recognition of Human Faces,” Kyoto University, 1973.

[10] R. Brunelli and T. Poggio, “Face recognition through geometrical features,” in *Sandini G. (eds) Computer Vision — ECCV’92. ECCV 1992. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 1992, pp. 792–800.

[11] I. J. Cox, J. Ghosn, and P. N. Yianilos, “Feature-based face recognition using mixture-distance,” in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 209–216, doi: 10.1109/CVPR.1996.517076.

[12] R. Jafri and H. R. Arabnia, “A Survey of Face Recognition Techniques,” *J. Inf. Process. Syst.*, vol. 5, no. 2, pp. 41–68, 2009, doi: 10.3745/jips.2009.5.2.041.

[13] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back, “Face recognition: a convolutional neural-network approach,” *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997, doi: 10.1109/72.554195.

[14] G. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments,” *Tech. rep.*, 2008.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 815–823, 2015, doi: 10.1109/CVPR.2015.7298682.

[16] J. J. Lv, X. H. Shao, J. S. Huang, X. D. Zhou, and X. Zhou, “Data augmentation for face recognition,” *Neurocomputing*, vol. 230, pp. 184–196, 2017, doi: 10.1016/j.neucom.2016.12.025.

[17] S. Z. Seyyedsalehi and S. A. Seyyedsalehi, “Simultaneous Learning of Nonlinear Manifolds Based on the Bottleneck Neural Network,” *Neural Process. Lett.*, vol. 40, no. 2, pp. 191–209, 2014, doi: 10.1007/s11063-013-9322-9.

[18] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real NVP,” *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, 2017.