12: Convolutional Neural Networks for Attribute Classification

Sergio Steven Leal Cuellar* and Mateo Rueda Molano†,
Department of Biomedical Engineering, University of the Andes, Bogota, Colombia
Email: *ss.leal10@uniandes.edu.co, †ms.rueda10@uniandes.edu.co,

***Abstract*—Due to the recent relevance of the neural networks methods, several classification tasks have been published, one of them is the multi-label classification task. For that purpose the CelebFaces Attributes (CelebA) Dataset is a suitable experimental frame in order to develop suitable methods to recognize various attributes (characteristics) on human faces. Because of that, a pair of methods based on convolutional neural networks were developed, one of them based on a modified architecture of ResNet and the other one is a simpler implementation of convolutional neural networks inspired from VGG. The results obtained suggest that the algorithms that were developed present an acceptable performance on the 10 categories multi-label classification task for the CelebA dataset. Nevertheless, applying some pre-processing and data augmentation might help to improve the performance obtained for the ResNet and VGG networks. Finally, *the CNN architecture that was used for the challenge was our smaller implementation of VGG (ourVGG)*.**

I. Introduction

Attributes represent semantic features allowing for mid-level representation between the low level features and the labels, with many applications such as recognition of people, activities, face recognition and verification., etc. Attributes can include biometrics such as age and gender but also particular specific categories such as hair color, hair style, accessories (glasses, hat), makeup and even mood and facial expressions. Accurate prediction of different kind of facial attributes is important of success of HCI applications and robotics where the machine needs to scan and understand the human by facial expressions to interact in a proper way. However, face attribute prediction is a challenging problem for biometric research over decades and the vast majority of the research threats each attribute individually as a prediction from the face image domain to the attribute label [1].

The first approximations in the literature of this problem had a great work of the scientific community some years ago. At that time, the direction of the gradient and the local descriptors were identified as first steps to face the lighting in the images [2] [3]. Methods close to the 2000s used SVM for the classification of ethnicity, pose, expression [8], with a similar characteristic and it is the use of face detectors by the Viola Jones algorithm, as well as position and detection algorithms. form for specific objects such as nose, eyes, mouth and train supervised classifiers taking into account the differences of these objects [4]. For the shape descriptors, BOW has been used in addition to histograms, taking into account variations in lighting, pose, among others [5]. Finally, near 2010 new approaches emerge such as handcrafted features (hc) which three typical: LBP, SIFT and Color histograms with SVM classifiers [1].

However, Neural Networks have helped us perform some machine learning tasks much better than we ever could, but that is not all. Convolutional Neural Networks (CNN) a specialized version of Deep Neural Networks, have revolutionized both machine learning and computer vision. They have shown great promise in tasks of recognition. Networks such as LNet and ANet, had promising results in facial attribute recognition. ANet, for example, extracts high-level face representation making attribute recognition from the entire face region possible [6]. Later, to investigate the effect of networkdepth on accuracy, VGG was released by Karen Simonyan [7]. Rasmus Rothe [8] finetuned VGG for apparent age estimation and won ChaLearn LAP competition in 2015. After the success of VGG, deeper network architectures such as Inception and ResNet [9] were developed. This networks had great success and impressive results in attribute recognition in huge datasets such as CelebA dataset which is a build up from 202599 images from celebrities with 40 attributes for recognition [10][11].

In this order of ideas, this paper will explore the performance of neural networks such as Resnet and VGG and their own derivations from the same networks, with the aim of observing performance in the task of classifying 10 of the 40 attributes.

II. materials and methods

A. Dataset description

The CelebFaces Attributes (CelebA) Dataset is a build up from 202,599 images of various celebrities with 10,177 unique identities. Each image haves annotations of 40 binary attribute possibilities. For the training of both convolutional neural networks, just 10 classes/attributes were used; 'Eyeglasses', 'Bangs', 'BlackHair', 'BlondHair', 'BrownHair', 'GrayHair', 'Male', 'PaleSkin', 'Smiling' and 'Young'. The images may vary severely on their resolution from one another.

(a) CelebA Dataset image example A.

(b) CelebA Dataset image example B.

Figure 1. Sample images and brigthness changes from the CelebA Dataset.

(a) CelebA Dataset image example A.

(b) CelebA Dataset image example B.

Figure 2. Sample images for old people from the CelebA Dataset.

The annotations are given in csv file with the imageid and all the 40 attributes set on columns with 1 if that image has that attribute and -1 if it doesn't have it which ease the multi-label task using a one-hot encoding.

B. Methods description

1) Modified Resnet Architecture.: We try ResNet architecture because networks such as AlexNet and LeNet when they start converging, a degradation problem is exposed. With the network increasing the accuracy gets saturated and then degrades rapidly. ResNet address the degradation problem in this way: instead of hoping each few stacked layers directly fit a desired underlying mapping, it explicitly let these layers fit a residual mapping. Formally, denoting the desired underlying mapping as $H(x)$, we let the stacked nonlinear layers fit another mapping of $F(x) := H(x)x$. The original mapping is recast into $F(x)+x$. Furthermore, Resnet hypothesizes it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers. The basic architecture is shown in fig 3

For this task we use a network resnet 18 which has 4 blocks of convolutional layers of 2 layers per block. However, due to the computational complexity and the high number of images we change 2 blocks to 1 convolutional layer. The architecture can be seen in the figure 4

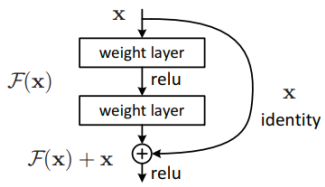

Figure 3. Basic structure of a residual neural network (ResNet)

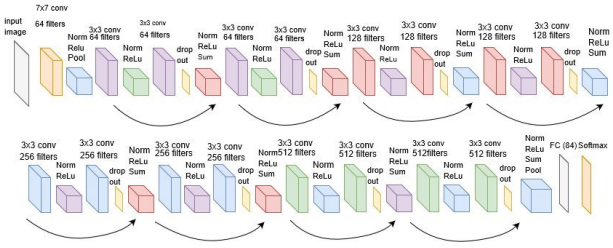

Figure 4. ResNet18 Architecture. However, in our algorithm due to the big data of training set we drop 2 blocks (block of 128 convolutional filters and block of 512 convolutional filters. Image retrieved from [12])

The network architecture consists of an input layer of 64 3x3 convolutional filters. Subsequently, blocks of 2 convolutional layers start and after the first convolutional layer of each block, a normalization is carried out by the Relu activation function that takes into account the output of the last block. Likewise, inside each block, after the second filter (convolutional layer) a 50% dropout is made to prevent overfitting of training co-adaptations. The filters (convolutional layers) are of 64,128,256 and 512 filters and at the end a maximal pooling and a multilabel softmax classifier is carried out. The used function of Loss was BCE because it is independent for each vector component (class), meaning that the computed for every CNN output vector component is not affected by other component values. That's why it is used for multi-label classification, were the insight of an element belonging to a certain class should not influence the decision for another class. Finally, the learning rate was of 0.003, momentum of 0.9 and a weight decay of 0.0001.

2) Our Architecture: The architecture that was established in order to face this task was inspired from the VGG architecture. In fact, our architecture is highly influenced by the ConvLayer+ConvLayer+Pooling sequency that VGG has for many of it's variants. In contrast our architecture uses just a single convolutional layer after every max pooling operation. In addition, our CNN uses is 3 layers smaller that the 11-Layer implementation of VGG and also reduces the ammount of parameters to a 5,791,818 model for simplicity. The activation function used for the data, due to the problem, was a sigmoid with a loss function of binary cross entropy loss function. For every convolutional layer 64, 128, 192, 256 and 256 number of filters were set each of them with a kernel size of 3 and setting 1 to the padding and the stride. Also, after every convolutional layer a batch normalization and and max pooling operation is done. Then, in order to get the 10-category classification, a set

of 3 fully conected layers were established. For the training parameters a static learning rate of 0.003 was set, with a momentum of 0.9 and a weight decay of 0.00001. Finally, to avoid over-fitting, a dropout of 50% was set between each of the FC layers. Then, a ablation study was done by removing the second an third layers (the hidden layers) of the network and its performance was analyzed. **This CNN architecture was the one that we used for the challenge.**

III. Results

Class	Accuracy	Precision	Recall	F-measure
Eyeglasses	1146/1289	0.999	0.889	0.941
Bangs	2424/3109	0.983	0.780	0.870
Black Hair	4354/5422	0.926	0.803	0.860
Blond Hair	2288/2660	0.972	0.860	0.9130
Brown Hair	2421/3587	0.919	0.675	0.778
Gray Hair	440/636	0.988	0.692	0.814
Male	7264/7715	0.995	0.942	0.968
Pale Skin	424/840	0.984	0.505	0.667
Smiling	9089/9987	0.943	0.910	0.926
Young	13992/15114	0.775	0.926	0.844
AVG	79.8%	0.948	0.798	0.867

Table I
Results obtained for the test using our VGG modification.

IV. Discussion

Taking in to account the results obtained using the modified smaller version of VGG, it can be seen that the proposed neural network presents an acceptable performance for most of the categories of the 10-class task. Nevertheless, there are several drawbacks in categories like "Pale Skin", "Brown Hair" and "Gray Hair", as can be seen in table I,the algorithm tends to have troubles in finding those characteristics on the images. Because of that, it can be said that these limitations on the classes may be a result of the illumination variation along the images, as can be seen in figure 1, some skin color may seem different due to the brightness in the image. For that reason, these categories tend to look similar to other, a suitable way to overcome this limitation might be to apply a histogram equalization to all the images. Additionally, the category "young" presented a vast amount of miss-matches, this may be caused by the lack of clear visual cues that help to distinguish young people from old one, as can be seen in figure 2 some images might not be evident. Theoretically, adding some noise to the input data of a neural network may cause an increase in the variation across the different images which may be useful because the model will face more unrelated data and so it will be forced to realize better generalizations on the dataset. Because of that jitter can serve as regularization method to avoid severe over fitting on the training set. Nevertheless, the results obtained without jitter were acceptable, but augmenting the amount of samples and diversifying the features of the samples may help to increase the overall performance of the method. After developing the abblation study, removing second and third layer, it came up that the smallest value of loss that the ablated model got was 0.163 meanwhile the whole model got a minimum loss of 0.124, which represents how the models differ from each and the relative importance of this layers because less high level information is being obtained due to the less amount of convolutions that are done. The results of ResNet was not as expected since the network was very time consuming and we cannot submit the results to the challenge. However, it will be important to check on these results to analyze how our algorithm will perform.

V. Conclusions

A face detector may will improve both methods because we are just introducing the entire image and this can cause noise in the results, algorithms such as Viola-Jones to preprocess the image may have many advantages for increasing the performance.

On the other hand, our method based on the VGG architecture presents a vast amount of difficulties due to it's reduced amount of parameters compared to the whole architecture that is used for VGG, lack of significant pre-processing and data augmentation. A way to improve this limitations may be to implement bigger networks with a big amount of filters of every convolutional layer. Also, adding a ResNet-similar mechanism of adding residual connections between the several layers. Additionally, since the learning rate was constant for every epoch of the training process it might be suitable to decrease the learning rate after some epochs or re scaling the learning rate if the loss is not decreasing after a specific number of epochs. Another possible option might be to generate a secondary "head" for the network, this secondary head might serve as a detector for every one of the attributes, thanks to the detector some high probability of feature regions can be obtained and those regions can serve the CNN to focus on classifying the regions extracted from the detector.

References

[1] R. Torfason, E. Agustsson, R. Rothe, and R. Timofte., "From face images and attributes to attributes," *Computer Vision Laboratory, ETH Zurich, Switzerland*, 2016.

[2] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs, "In search of illumination invariants," *CVPR*, 2000.

[3] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," *CVPR*, 1994.

[4] S. Srivastava and K. Asawa, "Real Time Facial Expression Recognition: Using a Novel Method," *The International Journal of Multimedia Its Applications (IJMA)*, vol. 4, 2012.

[5] N. Kumar, A. Berg, P. Belhumeur, and S.Nayar, "Attribute and Simile Classifiers for Face Verification," *Columbia University ICCV*, 2009.

[6] L. Ziwei, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," *Department of Information Engineering, The Chinese University of Hong Kong*.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[8] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of apparent age from a single image," *ICCV*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[10] Y. Zhong, J-Sullivan, and H. Li, "Face Attribute Prediction Using Off-the-Shelf CNN Feature," *Computer Science and Communication KTH Royal Institute of Technology*, 2016.

[11] R. Jahandideh, A. Tavakoli, and M. Tahmasb, "Physical Attribute Prediction Using Deep Residual Neural Networks," *Computer Science Shahid Beheshti University*, 2018.

[12] M. Al Rabbani Alif, S. Ahmed, and M. Hasan, "Isolated bangla handwritten character recognition with convolutional neural network," pp. 1–6, 12 2017.