A multi-label Classification of Attributes on Face Images

Giang H. Le, * Yeejin Lee
Seoul National University of Science and Technology
gianghle@seoultech.ac.kr, *yeejinlee@seoultech.ac.kr

Abstract

Generative adversarial networks (GANs) have reached a great result at creating the synthesis image, especially in the face generation task. Unlike other deep learning tasks, the input of GANs is usually the random vector sampled by a probability distribution, which leads to unstable training and unpredictable output. One way to solve those problems is to employ the label condition in both the generator and discriminator. CelebA and FFHQ are the two most famous datasets for face image generation. While CelebA contains attribute annotations for more than 200,000 images, FFHQ does not have attribute annotations. Thus, in this work, we introduce a method to learn the attributes from CelebA then predict both soft and hard labels for FFHQ. The evaluated result from our model achieves 0.7611 points of the metric is the area under the receiver operating characteristic curve.

1. Introduction

Generative Adversarial Networks (GANs) [1] have opened a new era for the numerous types of tasks in the computer vision field. Specifically, the quality of synthesis images produced by GANs in face generation tasks has been a significant increase [2] [3].

The performance of GANs model is evaluated by two factors: fidelity and diversity [4]. The fidelity accesses the degree to how much the generated images resemble real images. The diversity measures whether the generated samples cover the entire variability of the real samples. Currently, GANs are good at fidelity owing to the truncate tricks and the novel architecture [5] [6]. However, the outputs from GANs still are strongly dependent on the noise from latent space, which indicates that the generated images are entangling and uncontrollable.

The study in [7] implies that the labels of the face image are necessary for controlling the output image, leads to the increasing of diversity, and extends the implementation field of GANs. CelebFaces Attributes (CelebA) [8] is a large-scale face attributes dataset with more than 200K celebrity images,

Figure 1. Each image contains a 40 binary labels, which leads to multi-label tasks. For example, the left image positive attributes are big lips, big nose, chubby, male, wearing hat, young, and the right are attractive, blond hair, heavy makeup, young.

each with 40 attribute annotations. The CelebA dataset has been employed as the training and test sets for many computer vision tasks thanks to its attribute annotations and landmark locations. Despite its contribution, the resolution of images in the CelebA dataset 178 x 229 is not sufficiently large by the fact the state-of-the-art GANs model is able to generate pictures that reach 512 x 512 [9] or even 1024 x 1024

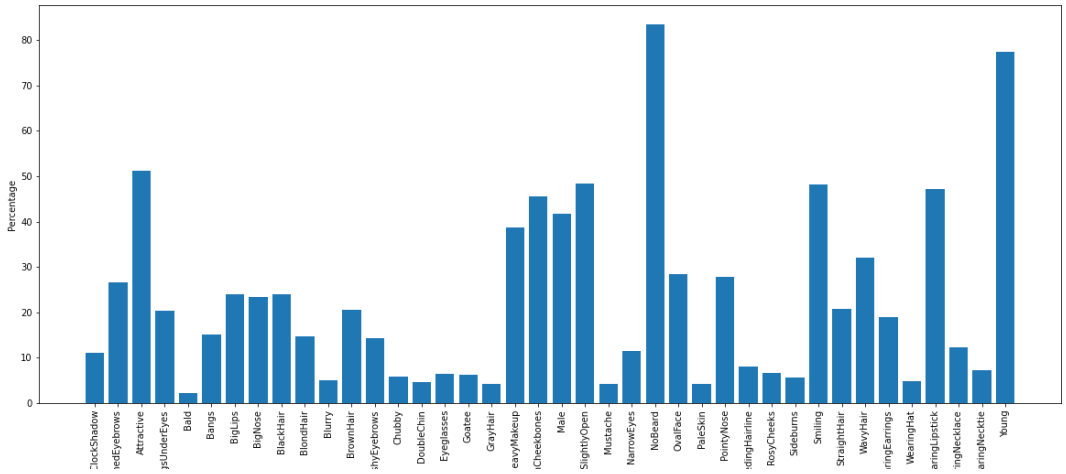

Figure 2. The density distribution in the percentage out of the whole dataset

[5] [6] [10].

FFHQ [10] is a more recent dataset of human faces, containing 70k images with 1024 x 1024 resolution and higher variation in terms of ages, ethnicity, and viewpoints. However, it does have neither attribute annotations nor landmark labels, which makes its usage in limited tasks due to less controllability. Thus, we want to add such missing attributes to the FFHQ dataset and to have more tractability on it. To do this, in this work, we will train the multi-label classification model by using CelebA then transfer the result to determine the attributes of FFHQ. However, the imbalance of CelebA dataset induces some issues in the training and the inference. Specifically, the imbalance problem causes more difficulty in selecting model architecture and evaluation metrics. While in inference mode, to predict the hard label such as the binary attributes in CelebA, we need to define the specific threshold for each class since it is hard to set the general threshold for all classes in CelebA.

2. Proposed Solution

2.1. CelebA Dataset

CelebA dataset contains 202,599 face images of the size 178x229 from 10,177 celebrities, each annotated with 40 binary labels indicating facial attributes like hair color, gender, and age, as illustrated in the examples of Figure 1. However, the dataset is imbalanced, as demonstrated in Figure 2. The imbalance issue makes a model poorly recognize the low-density labels like Chubby or Blurry occupied lower than 10% of the whole dataset. We will solve this problem by adding the class weight into the loss function in Section 2.3.

2.2. Model Selection

We formulate the attribute annotation on the high-resolution FFHQ dataset as multi-label classification. Although this multi-label classification problem can be solved by applying any state-of-the-art classification deep learning model, the capacity and the complexity of the model should be compatible with the difficulty of the dataset.

As observed in Figure 1, most attributes of the CelebA are easy to classify since they are definitely identifiable features, for example, young, male, mustache, wearing a hat, etc. On the other hand, some attributes, such as attractiveness, pale skin, and blurry, are not easy to classify. It is hard to find common features that represent these kinds of attributes, even for humans.

We empirically find that SE-ResNeXt-101(32 x 4d) [11] is suitable for this multi-label classification problem. The 101 layers with residual block architecture are light enough to solve the easy labels. On the other side, the combination between squeeze-and-excitation and the inception module of ResNext should be able to handle the confusing ones.

2.3. Loss function

The goal of training the proposed network is to predict the corresponding label y of an input image X to match with the correct class label $\boldsymbol{c}$. Recalling the discussion in Section 2.1., as the CelebA dataset is imbalanced, we force the model to focus more on the fewer density labels than others. This is achieved by giving different weights to the majority and

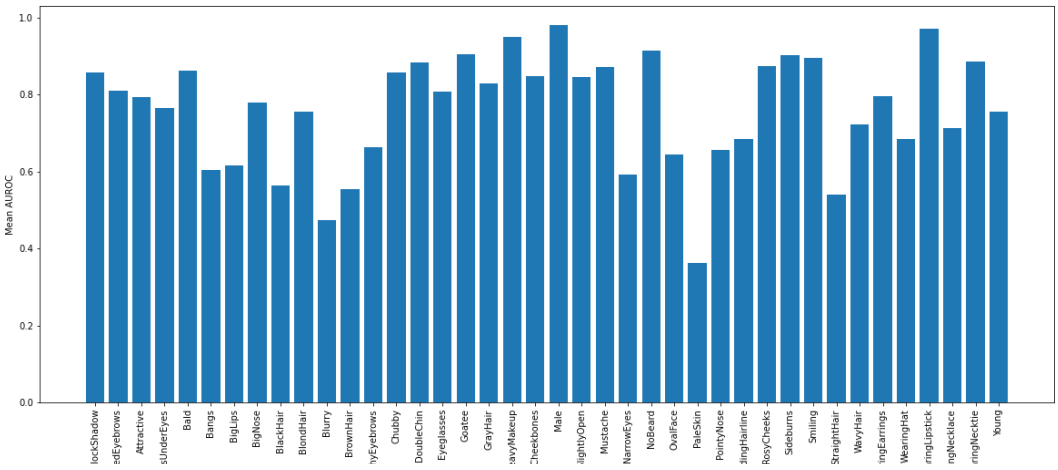

Figure 3. The AUROC of each attribute predicted by the trained model.

minority classes in the loss function, as follows:

$$loss(X|y = \boldsymbol{c}) = weight(c) \times criterion(X, y). \quad (1)$$

The criterion function of the model output H and the true label y is computed as the average of the loss $l_{i,j}$ over the entire classes and the examples:

$$criterion(H, y) = \frac{1}{m \times n} \sum_{i=1}^{n} \sum_{j=1}^{m} l_{i,j}\left(H_{i,j}, y_{i,j}\right). \quad (2)$$

In Eq. (2), m is the total number of classes, and n is the total number of examples. The loss $l_{i,j}$ is defined as the binary cross-entropy with the sigmoid activation σ:

$$l_{i,j}\left(H_{i,j}, y_{i,j}\right) = y_{i,j} \log \sigma\left(H_{i,j}\right) + \left(1 - y_{i,j}\right)\left(1 - \log \sigma\left(H_{i,j}\right)\right), \quad (3)$$

where the sigmoid activation function σ normalizes the outputs int the probability range [0,1].

The $weight(\boldsymbol{c})$ for each class is computed as following [12], even though their work applies to the multiclass classification, we modify it to adapt to the multi-label classification by adding a weighting factor for each class. The weighting factor for class t in multiclass classification is defined as:

$$w_t = \frac{N_1 + N_2 + \cdots + N_t}{M \times N_t} = \frac{N}{M \times N_t}, \quad (4)$$

where N is the total number of samples, M is the total number of classes and N_m represents the number of samples of class m. In multilabel classification, an example can have multiple class, so in this work we consider the N and N_m as:

$$N = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{i,j}, \quad (5)$$

$$N_t = \sum_{i=1}^{n} y_{i,t}, \quad (6)$$

and we set $M = m = 40$. We also apply random augmentation [13] to avoid overfitting and diversifying the training.

3. Experiments

3.1 Experimental Setup

Datasets: We trained the proposed multi-label classification model on CelebA then used it to predict the face attributes of FFHQ. We resized the CelebA images to 224 x 224 and used them as an input of the model. We randomly selected 20% out of more than 200k images of CelebA to be test set. Finally, we reduced the resolution of FFHQ from 1024 x 1024 to inference the model.

Evaluation metrics: Since the task is multi-label classification, it is hard to measure the model's performance by accuracy, precision, or recall. Thus, we controlled the training process by hamming loss in Eq. (7) and the test set result by the area under the receiver operating characteristic curve (AUROC). The hamming distance is defined as:

$$hamming\ loss = \frac{1}{n \times m} \sum_{i=0}^{n} \sum_{j=0}^{m} XOR\left(y_{i,j}, y'_{i,j}\right), \quad (7)$$

where $\boldsymbol{y}'_{i,j}$ represents the $\boldsymbol{j}$-th class of the $\boldsymbol{i}$-th example in our dataset, and XOR is a digital logic gate that returns 0 if $\boldsymbol{y}_{i,j}$ and $\boldsymbol{y}'_{i,j}$ is similar and return 1 vice versa.

3.2. Result

Figure 3 shows the AUROC of each attribute predicted by our model. As discussed in Section 2.1, the hard and low-frequency classes will return the worse result. Although we

Figure 4. The images from FFHQ dataset. Our model predicts the left image with the positive attributes are male, no beard, young, and the positive attributes of the right positive attributes are attractive, bangs, heavy makeup, high cheek bones, mouth slightly open, no beard, pointy nose, smiling, wearing lip stick and young.

get the lowest AUROC attribute is Pale skin (<40), the mean AUROC of all classes is 0.7611.

While predicting FFHQ, we aim to predict 2 sets of labels, soft and hard. The soft label is simple since we just keep the output from the model then get through the sigmoid activation to get the result. On the other hand, the hard label is difficult to decide the threshold of each class. By using the test set of CelebA, we do the small grid search to find the best threshold for each attribute. Specifically, we arrange a list of thresholds from 0.01 to 1 with step 0.01, we then apply each threshold to the output from our model and apply the hamming loss to find the threshold with the lowest loss for each class. Figure 4 presents the example of our hard labels on FFHQ images.

4. Conclusion

This paper presents a method to find the labels for the FFHQ dataset based on the attributes from the CelebA dataset, dealing with the imbalanced problem on the multi-label classification task. The proposed method is to combine the class weight into loss function and the data augmentation in the training process. While in the inference step, the specific threshold for each class is important to find the hard label. In experiments, our AUROC achieves 0.7611 on average, and this implies that the proposed model succeeded to classify the sets of hard and soft labels from FFHQ.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00994, Development of autonomous VR and AR content generation technology reflecting usage environment).

References

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. "Generative Adversarial Nets", in Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 2014.

[2] J. Lin, R. Zhang, F. Ganz, S. Han and J.-Y. Zhu, "Anycost GANs for Interactive Image Synthesis and Editing", in Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[3] E. Schönfeld, B. Schiele and A. Khoreva, "A U-Net Based Discriminator for Generative Adversarial Networks", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[4] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi and J. Yoo, "Reliable Fidelity and Diversity Metrics for Generative Models," in Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

[5] Karras, T., Laine, S., & Aila, T. " A style-based generator architecture for generative adversarial networks.", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401-4410.

[6] Karras, T., Laine, S., & Aila, T. " A style-based generator architecture for generative adversarial networks.", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401-4410.

[7] A. Odena, C. Olah and J. Shlens, "Conditional Image Synthesis with Auxiliary Classifier GANs," in Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 2017, pp. 2642-2651.

[8] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (ICCV), Santiago, Chile, 2015.

[9] Brock, A., Donahue, J., & Simonyan, K. "Large Scale GAN Training for High Fidelity Natural Image Synthesis", in International Conference on Learning Representations (ICLR), Vancouver, Canada, September, 2018.

[10] Karras, T., Aila, T., Laine, S., & Lehtinen, J. "Progressive Growing of GANs for Improved Quality, Stability, and Variation", in International Conference on Learning Representations (ICLR), February, 2018.

[11] Hu, J., Shen, L., & Sun, G. "Squeeze-and-excitation networks", in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Utah, USA, 2018, pp. 7132-7141.

[12] G. King and L. Zeng, "Logistic Regression in Rare Events Data," Political Analysis, vol. 9, p. 137-163, 2001.

[13] Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. "Randaugment: Practical automated data augmentation with a reduced search space", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR), 2020, pp. 702-703.