High-Resolution Deep Convolutional Generative Adversarial Networks

Joachim D. Curto*[,1,2,3], Irene C. Zarza*[,1,2,3], Fernando De La Torre[2],
Irwin King[1], and Michael R. Lyu[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]The Robotics Institute, Carnegie Mellon University
[3]Department of Electronic Engineering, City University of Hong Kong

{curto,zarza,king,lyu}@cse.cuhk.edu.hk , ftorre@cs.cmu.edu

*Both authors contributed equally

Abstract

Generative Adversarial Networks (GANs) [2] convergence in a high-resolution setting with a computational constrain of GPU memory capacity (from 12GB to 24 GB) has been beset with difficulty due to the known lack of convergence rate stability. In order to boost network convergence of DCGAN (Deep Convolutional Generative Adversarial Networks) [6] and achieve good-looking high-resolution results we propose a new layered network structure, HR-DCGAN, that incorporates current state-of-the-art techniques for this effect. A novel dataset, CZ Faces (CZF)[1], containing human faces from different ethnical groups in a wide variety of illumination conditions and image resolutions is introduced. We conduct extensive experiments on CelebA [5] and CZF.

1. Introduction

Developing a Generative Adversarial Network (GAN) structure [2] able to produce good quality high-resolution samples from images has important applications including image inpainting, 3D data, localization and semi-supervised learning.

In this paper, we focus on the task of face generation, as it gives GANs a huge space of learning attributes. In this context, we introduce CZ Faces (CZF), a well-balanced dataset containing 14.248 human faces from different ethnical groups and rich in a wide range of learnable attributes, such as gender and age diversity, hair-style and pose variation or presence of smile, glasses, hats and fashion items. We also ensure the presence of changes in illumination and image resolution. We propose to use CZF as de facto approach to empirically test the distribution learned by a GAN, as it offers a challenging problem to solve, while keeping the number of samples, and therefore training time, bounded.

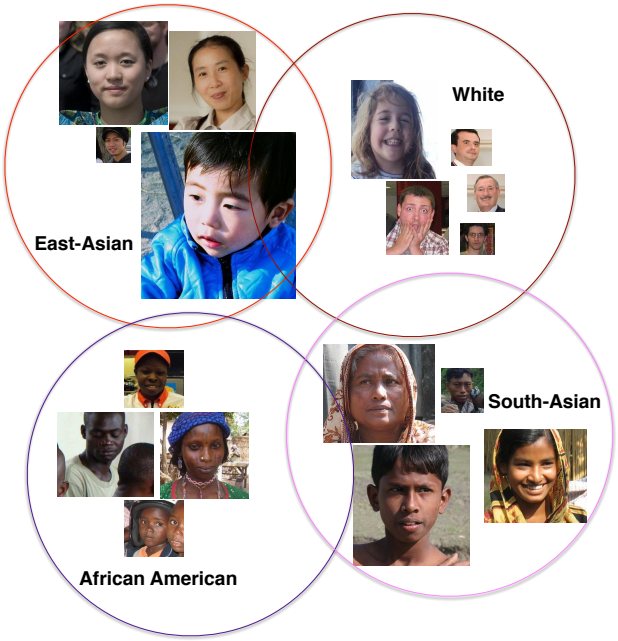

Figure 1. **CZF image samples**. A set of random samples for each ethnicity category.

Despite improvements in GANs training stability [7] and specific-task design during the last years, it is still challenging to train GANs to generate high-resolution images. Mainly due to the possible disjunction in the high dimensional pixel space between supports of the real image and implied model distributions [1, 8].

[1]CZF is available at http://www.cse.cuhk.edu.hk/∼curto/

Our goal is to be able to generate indistinguishable sample instances using face data to push the boundaries of GAN image generation that scale well to high-resolution images (*i.e.* 512×512) and where context information is maintained.

In this sense, Deep Learning has a tremendous appetite for data. The question that arises instantly is, what if we were able to generate additional realistic data to aid the learning process using the same techniques that are later used to train the system. The first step would then be to have an image generation tool able to sample from a very precise distribution (*e.g.* faces from celebrities) which instances resemble or highly correlate with real sample images of the underlying true distribution. Once this is achieved, what is desirable and comes next is that these generated image points not only fit well into the original distribution set of images but also add additional useful information such as redundancy, different poses or even generate highly-probable scenarios that would be possible to see in the original dataset but are actually not present.

To achieve the former goal this work contributes in the following:

- Network structure that achieves compelling results and scales well to the high-resolution setting where to the best of our knowledge other variant network architectures are unable to continue learning or fall into mode collapse.
- New dataset targeted for GAN training, CZF, that introduces a wide space of learning attributes. It aims to provide a well-posed difficult task while keeping training time and resources tightly bounded to spearhead research in the area.

2. Related Work

Generative image generation is a key problem in Computer Vision. Remarkable advances have been made with the renaissance of Deep Learning. Variational Autoencoders (VAE) [3] formulates the problem with a probabilistic graphical model approach, where the lower bound of data likelihood is maximized. Autoregressive models (*i.e.* PixelRNN [9]), based on modeling the conditional distribution of the pixel space, have also presented relative success generating synthetic images. Lately, Generative Adversarial Networks (GAN) [2] have shown strong performance in image generation. However, training instability makes it very hard to scale to high-resolution (256× 256 or 512×512) samples. Some current works on the topic pinpoint this specific problem [12], where conditional image generation is also tackled while other recent techniques [7] try to stabilize the training process.

3. CZF Dataset

CZF contains a total of 14.248 face images balanced in terms of ethnicity: african american, east-asian, south-asian and white. Some sample images can be seen in Figure 1. We crawled Flickr to download face images from several countries that contain different hair-style variations and style attributes. These images were then processed to extract 49 facial landmark points using [10]. We ensure using Amazon Mechanical Turk that the detected faces are correct in terms of ethnicity and face detection. Cropped faces are then extracted to generate multiple resolution sources. Mirror augmentation is performed to further enhance pose variation.

CZF introduces a difficult learning paradigm, where different ethnical groups are present, with very varied fashion and hair styles. The fact that the photos are taken using non-professional cameras in a non-controlled environment, gives us multiple face poses, illumination conditions and camera quality.

4. High-Resolution Structure

Generative Adversarial Networks (GANs) proposed by [2] are based on two dueling networks; Generator G and Discriminator D. In essence, the learning process consists on a two-player game where D tries to distinguish between the prediction of G and the ground truth, while at the same time G tries to fool D by producing fake instance samples as closer to the real ones as possible.

The min-max game entails the following objective function.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] + \\ + \mathbb{E}_{z\sim p_z}\left[\log(1 - D(G(x)))\right]. \quad (1)$$

where x is a ground truth image sampled from the true distribution p_{data}, and z is a noise vector sampled from p_z (*i.e.* uniform or normal distribution). G and D are parametric functions where $G : p_z \rightarrow p_{data}$ maps samples from noise distribution p_z to data distribution p_{data}.

As an extension to this framework, DCGAN [6] proposes an architectural topology based on Convolutional Neural Networks (CNNs) to stabilize training and re-use state-of-the-art network topologies from image classification tasks. This direction has recently received lots of attention due to its compelling results in several supervised and unsupervised learning problems. We build on this to propose a novel DCGAN architecture to address the problem of high-resolution image generation. We name this approach HR-DCGAN.

4.1. HR-DCGAN

Despite the undoubtable success, GANs are still arduous to train, particularly when we use big images (*e.g.* 512×512). It is very common to see D beating G in the learning process, or the reverse. Ending in unrecognizable imagery, also known as mode collapse. Just when stable learning is achieved, the GAN structure is able to succeed in getting better and better results with time.

Is this issue what drives us to carefully derive a simple yet powerful structure that leverages common problems and gets a stable and steady training mechanism.

Self-normalizing Neural Networks (SNNs) were introduced in [4]. We consider a neural network with activation function f, connected to the next layer by a weight matrix $\mathbf{W}$, and whose inputs are the activations from the preceding layer x, *i.e.* $y = f(\mathbf{W}x)$.

We can define a mapping g that maps mean and variance from one layer to mean and variance of the following layer

$$\begin{pmatrix}\mu\\ \nu\end{pmatrix} \longmapsto \begin{pmatrix}\tilde{\mu}\\ \tilde{\nu}\end{pmatrix} : \begin{pmatrix}\tilde{\mu}\\ \tilde{\nu}\end{pmatrix} = g\begin{pmatrix}\mu\\ \nu\end{pmatrix}. \quad (2)$$

Common normalization tactics such as batch normalization ensure a mapping g that keeps (μ, ν) and $(\tilde{\mu}, \tilde{\nu})$ close to a desired value, normally (0, 1).

SNNs go beyond this assumption and require the existence of a mapping $g : \Omega \longmapsto \Omega$ that for each activation y maps mean and variance from one layer to the next layer and at the same time have a stable and attracting fixed point depending on (ω, τ) in Ω. Moreover, the mean and variance remain in the domain Ω and when iteratively applying the mapping g, each point within Ω converges to this fixed point. Therefore, SNNs keep activations normalized when propagating them through the layers of the network.

Scaled Exponential Linear Units (SELU) [4] is introduced as the choice of activation function in Feed-forward Neural Networks (FNNs) to construct a mapping g with properties that lead to SNNs.

$$selu(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha \exp^x - \alpha & \text{if } x \leq 0 \end{cases} \quad (3)$$

Empirical observation leads us to say that the use of SELU activation function greatly improves the convergence speed on the DCGAN structure, however, after some iterations mode collapse and gradient explosion completely destroy training when using high-resolution images. We conclude that although SELU gives theoretical guarantees as the optimal activation function in FNNs, numerical errors in the GPU computation degrade its performance in the overall min-max game of DCGAN. To alleviate this problem, we propose to use always both SELU and BatchNorm together. The motivation is that when numerical errors move $(\tilde{\mu}, \tilde{\nu})$ away from the attracting point that depends on $(\omega, \tau) \in \Omega$, BatchNorm will ensure it is close to a desired value and therefore maintain the convergence rate. Experiments show that this technique stabilizes training and allows us to use fewer GPU resources, having a steady diminishing G and D learning loss. It also accelerates convergence speed by a great factor, as can be seen after just some few epochs of CelebA training in Figure 4.

We observe that when having difficulty in training DCGAN, it is always better to use a fixed learning rate and instead increase the batch size. This is because having more diversity in training, gives a steady diminishing loss and better generalization results. To aid the learning process, additive noise is added to both the inputs of D and G. We see that this helps overcome mode saturation and collapse.

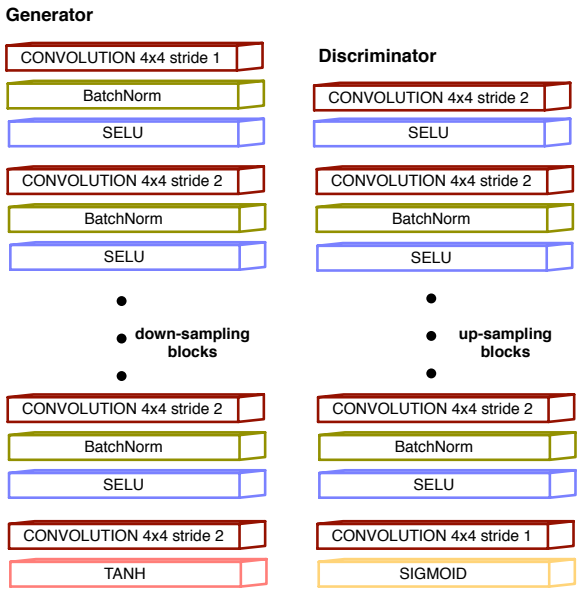

Figure 2. **HR-DCGAN.** Discriminator and Generator architectures.

5. Experiments and Results

We build on DCGAN and extend the framework to train with high-resolution images using Pytorch. Our experiments are conducted using a fixed learning rate of 0.0002 and ADAM solver [26] with batch size 32 and 512×512 training images with the number of filters of G

Figure 3. **Example results by HR-DCGAN on CZF dataset**. 150 epochs of training. Image size 512×512.

and D equal to 64.

In order to test generalization capability, we train HR-DCGAN in the newly introduced CZF dataset, CelebA and Street View Research dataset. We use 2×NVIDIA Titan X and 12 CPU cores.

5.1. CZF

The results after 150 epochs are shown in Figure 3. We can see that HR-DCGAN captures the underlying image features that represent faces and not only memorizes training examples. We retrieve nearest neighbors to the generated images in Figure 5 to illustrate this effect.

5.2. CelebA

CelebA is a large-scale dataset with 202.599 celebrity faces. It mainly contains frontal faces and is particularly biased towards white ethnical groups. The fact that it presents very controlled illumination settings and good photo resolution, makes it a considerably easier problem than CZF. The results after just 12 epochs of training are shown in Figure 4.

5.3. Street View

The Street View dataset [11] contains 62.058 high-quality images from Pittsburgh, Manhattan and Orlando. We use this dataset to illustrate the case of 1024×1024 image size learning. We lower the filter resolution of D and G to 16 so that it fits into 24GB memory. The results after 249 epochs of training can be seen in Figure 6.

6. Conclusions

In this paper, we propose High-Resolution Generative Adversarial Networks (HR-DCGAN) by stacking blocks of SELU and BatchNorm. The proposed method generates high resolution images (*e.g.* 512×512 and 1024×1024) in circumstances where other methods fail. It exhibits a steady and smooth training mechanism.

Figure 4. **Example results by HR-DCGAN on CelebA**. 12 epochs of training. Image size 512×512.

Furthermore, we present a face dataset containing well-balanced ethnical groups for GAN training, CZ Faces (CZF), that poses a very difficult challenge and is rich on learning attributes to sample from.

Figure 5. **Nearest Neighbors**. Generated images in the first row and retrieving their five nearest neighbors in the training images (rows 2-6).

Figure 6. **Example results by HR-DCGAN on StreetView dataset**. 249 epochs of training. Image size 1024×1024.

References

[1] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial neworks. *NIPS*, 2014.

[3] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[4] G. Klarbauer, T. Unterthiner, and A. Mayr. Self-normalizing neural networks. 2017.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *ICCV*, 2015.

[6] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial network. *ICLR*, 2016.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Redford, and X. Chen. Improved techniques for training gans. *NIPS*, 2016.

[8] C. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Hussar. Amortised map inference for image super-resolution. *ICLR*, 2017.

[9] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016.

[10] X. Xiong and F. De La Torre. Supervised descent method and its application to face alignment. *CVPR*, 2013.

[11] A. R. Zamir and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *TPAMI*, 2014.

[12] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CVPR*, 2017.

Supplementary Materials

More Results on CelebA Dataset

Additional results on CelebA dataset after 19 and 39 training epochs. The experiments are conducted using a fixed learning rate of 0.0002 and ADAM solver with batch size 32 and 512×512 training images. Number of filters of G and D equal to 64.

Figure 7. **Example results by HR-DCGAN on CelebA dataset**. 19 epochs of training. Image size 512×512.

Technical Specifications: 2 × NVIDIA Titan X, Intel Core i7-5820k@3.30GHz.

Figure 8. **Example result by HR-DCGAN on CelebA dataset**. 19 epochs of training. Image size 512×512. Full scale image.

Figure 9. **Example results by HR-DCGAN on CelebA dataset**. 19 epochs of training. Image size 512×512. Failure cases.

Figure 10. **Example results by HR-DCGAN on CelebA dataset**. 39 epochs of training. Image size 512×512.

Technical Specifications: 2 × NVIDIA Titan X, Intel Core i7-5820k@3.30GHz.

Figure 11. **Example result by HR-DCGAN on CelebA dataset**. 39 epochs of training. Image size 512×512. Full scale image.

Figure 12. **Example results by HR-DCGAN on CelebA dataset**. 39 epochs of training. Image size 512×512. Failure cases.