Conditional Image Synthesis

Hongzhi Liu

July 8, 2018

Abstract

Characterizing the structure of natural images has been a rich research endeavor. Recent work has shown that GANs can produce convincing image samples on datasets with low variability and low resolution. And GANs struggle to generate globally coherent and high resolution samples, particularly from datasets with high variability. Today, I read a thesis written by Augustus Odena who is from Google Brain. His team introduce new methods for the improved training of generative adversarial networks (GANs) for image synthesis. They construct a variant of GANs employing label conditioning that results in 128×128 *resolution image samples exhibiting global coherence. Beside, they expand on previous work for image quality assessment to provide two new analyses for assessing the discriminability and diversity of samples from class-conditional image synthesis models.*

1. Overview of AC-GANs

Natural images obey intrinsic in variances and exhibit multi-scale statistical structures that have historically been difficult to quantify [7]. Recent advances in machine learning offer an opportunity to substantially improve the quality of image models. Improved image models advance the state-of-the-art in image denoising [1], compression [10], in-painting [5] and super-resolution [3]. Better models of natural images also improve performance in semi-supervised learning tasks [2] and reinforcement learning problems .

Generative adversarial networks (GANs) offer a distinct and promising approach that focuses on a game-theoretic formulation for training an image synthesis model. However, GANs struggle to generate globally coherent, high resolution samples-particularly from datasets with high variability. Moreover, a theoretical understanding of GANs is an on-going research topic.

In this paper, Augustus Odena and his team demonstrate that that adding more structure to the GAN latent space along with a specialized cost function results in higher quality samples [4]. They exhibit 128×128 pixel samples from all classes of the ImageNet dataset [6] with increased global coherence as shown in Fig. 1. Importantly, they demonstrate quantitatively that the high resolution samples are not just naive resizings of low resolution samples. In particular, downsampling their 128×128 samples to 32×32 leads to a 50% decrease in visual discriminability. The team also in-

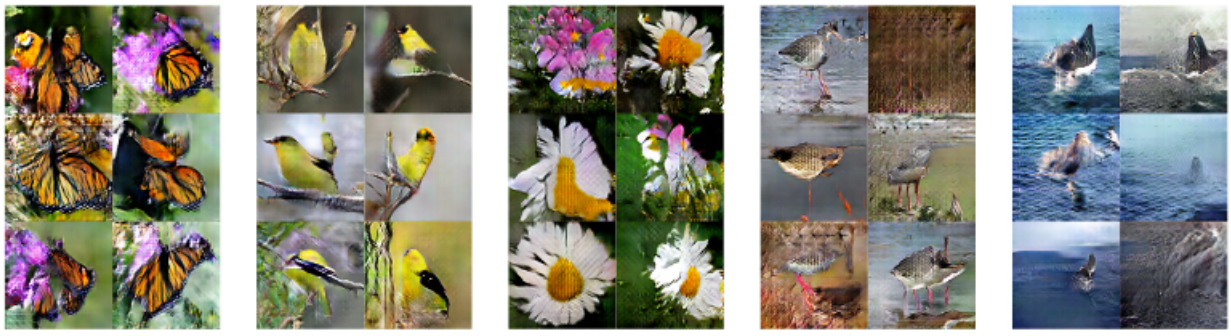

Figure 1. 128×128 resolution samples from 5 classes taken from an AC-GAN trained on the ImageNet dataset. Note that the classes shown have been selected to highlight the success of the model and are not representative. Samples from all ImageNet classes are linked later in the text.

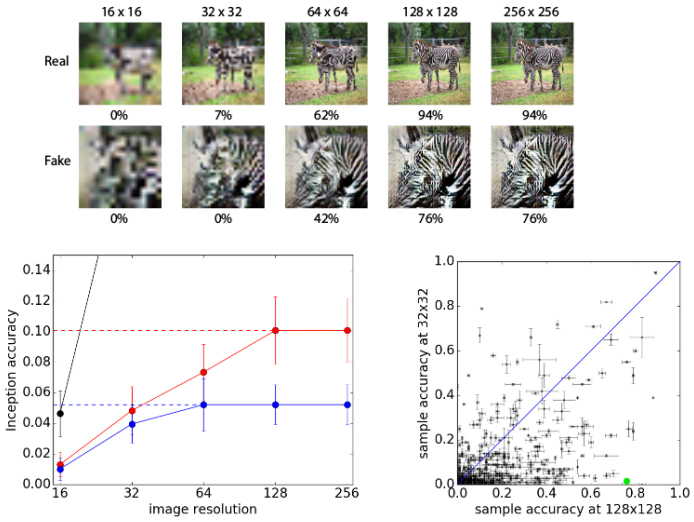

Figure 2. Generating high resolution images improves discriminability. Top: Training data and synthesized images from the zebra class resized to a lower spatial resolution and subsequently artificially resized to the original resolution. Bottom Left: Summary of accuracies across varying spatial resolutions for training data and image samples from 64×64 and 128×128 models. Bottom Right: Comparison of accuracy scores at 128×128 and 32×32 spatial resolutions

troduce a new metric for assessing the variability across image samples and employ this metric to demonstrate that their synthesized images exhibit diversity comparable to training data for a large fraction (84.7%) of ImageNet classes.

2. Architecture of AC-GANs

Augustus Odena and his team propose a variant of the GAN architecture which they call an auxiliary classifier GAN or AC-GAN. In the AC-GAN, every generated sample has a corresponding class label, $c \sim p_c$ in addition to the noise z. G uses both to generate images $X_{fake} = G(c, z)$. The discriminator gives both a probability distribution over sources and a probability distribution over the class labels, $P(S|X)$, $P(C|X) = D(X)$. The objective function has two parts that one is the log likelihood of the correct source, L_S and the log-likelihood of the correct class L_C as Eq. 1:

$$\begin{aligned} L_S &= E[\log P(S = real|X_{real})] \\ &\quad + E[\log P(S = fake|X_{fake})], \\ L_C &= E[\log P(C = c|X_{real})] \\ &\quad + E[\log P(C = c|X_{fake})]. \end{aligned} \tag{1}$$

where D is trained to maximize $L_S + L_C$ while G is trained to maximize $L_C - L_S$. AC-GANs learn a representation for z that is independent of class label.

Structurally, this model is not tremendously different from existing models. However, this modification to the standard GAN formulation produces excellent results and appears to stabilize training. Moreover, they consider the AC-GAN model to be only part of the technical contributions of the work, along with their proposed methods for measuring the extent to which a model makes use of its given output resolution, methods for measuring perceptual variability of samples from the model and a thorough experimental analysis of a generative model of images that creates 128×128 samples from all 1000 ImageNet classes.

Early experiments demonstrated that increasing the number of classes trained on while holding the model fixed decreased the quality of the model outputs. The structure of the AC-GAN model permits separating large datasets into subsets by class and training a generator and discriminator for each subset. All ImageNet experiments are conducted using an ensemble of 100 AC-GANs, each trained on a 10 class split.

3. Experiment Results of Models

Augustus Odena trains several AC-GAN models on the ImageNet data set [6]. Broadly speaking, the architecture of the generator G is a series of 'deconvolution' layers that transform the noise z and class c into an image. He trains two variants of the model architecture for generating images at 128×128 and 64×64 spatial resolutions. The discriminator D is a deep convolutional neural network with a Leaky ReLU nonlinearity.

Evaluating the quality of image synthesis models is challenging due to the variety of probabilistic criteria [9] and the lack of a perceptually meaningful image similarity metric. Nonetheless, they attempt to measure the quality of the AC-

GAN by building several ad-hoc measures for image sample discriminability and diversity.

To measure discriminability, Augustus Odena and his team feed synthesized images to a pre-trained Inception network [8] and report the fraction of the samples for which the Inception network assigned the correct label. They calculate this accuracy measure on a series of real and synthesized images which have had their spatial resolution artificially decreased by bilinear interpolation on top panels in Fig. 2. Note that as the spatial resolution is decreased, the accuracy decreases - indicating that resulting images contain less class information as shown on below top panels in Fig. 2. They summarized the finding across all 1000 ImageNet classes for the ImageNet training data (black), a 128×128 resolution AC-GAN (red) and a 64×64 resolution AC-GAN (blue) as shown on bottom left in Fig. 2. The black curve (clipped) provides an upper-bound on the discriminability of real images.

References

[1] J. Ballé, V. Laparra, and E. P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *ICLR*, 2016. 1

[2] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014. 1

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1

[4] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016. 1

[5] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. 1

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 1, 2

[7] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 2001. 1

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[9] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. *Computer Science*, 2015. 2

[10] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. In *CVPR*, 2017. 1