Deep Generative Image Models

Hongzhi Liu

July 2, 2018

Abstract

Building a good generative model of natural images has been a fundamental problem within computer vision. However, images are complex and high dimensional, making them hard to model well despite extensive efforts. Today, I read a thesis written by Emily Denton who is in Department of Computer Science Courant Institute from New York University. His team introduce a generative parametric model capable of producing high quality samples of natural images. Their approach uses a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. In a quantitative assessment by human evaluators, their CIFAR10 samples were mistaken for real images around 40% of the time, compared to 10% for samples drawn from a GAN baseline model. They also show samples from models trained on the higher resolution images of the LSUN scene dataset.

1. Overview of Deep Generative Image Models

Generative image models are well studied, falling into two main approaches that is non-parametric and parametric. The former copy patches from training images to perform, for example, texture synthesis [3] or super-resolution [4]. More ambitiously, entire portions of an image can be in-painted, given a sufficiently large training dataset. Early parametric models addressed the easier problem of tex-ture synthesis, with Portilla and Simoncelli [7] making use of a steerable pyramid wavelet representation, similar to Denton's use of a Laplacian pyramid. For image processing tasks, models based on marginal distributions of image gradients are effective [8], but are only designed for image restoration rather than being true density models. Very large Gaussian mixture models [10] and sparse coding models of image patches can also be used but suffer the same problem.

Several recent papers have proposed novel generative models. Dosovitskiy *et al.* [2] showed how a convnet can draw chairs with different shapes and viewpoints. While their model also makes use of convnets, it is able to sample general scenes and objects. The DRAW model of Gregor *et al.* [6] used an attentional mechanism with an RNN to generate images via a trajectory of patches, showing samples of MNIST and CIFAR10 images. Sohl-Dickstein *et al.* [9] use a diffusion-based process for deep unsupervised learning and the resulting model is able to produce reasonable CIFAR10 samples.

In this paper, Emily Denton and his team exploit the multiscale structure of natural images, building a series of generative models, each of which captures image structure at a particular scale of a Laplacian pyramid [1]. This strategy breaks the original problem into a sequence of more manageable stages. At each scale they train a convolutional networkbased generative model using the Generative Adversarial Networks (GAN) approach of Goodfellow *et al.* [5].

2. LAPGAN model

The basic building block of Denton's approach is the generative adversarial network (GAN) of Goodfellow *et al.* [5]. They introduce the LAPGAN model which integrates a conditional form of GAN model into the framework of a Laplacian pyramid.

2.1. Generative Adversarial Networks

The GAN approach [5] is a framework for training generative models, which the team briefly explain in the context of image data. The method pits two networks against one another. One is a generative model G that captures the data distribution, and the other one is a discriminative model D that distinguishes between samples drawn from G and images drawn from the training data. In their approach, both G and D are convolutional networks. The former takes as input a noise vector z drawn from a distribution $p_{Noise}(z)$ and outputs an image $\tilde{h}$. The discriminative network D takes an image as input stochastically chosen to be either $\tilde{h}$ – as generated from G, or h – a real image drawn from the training data $p_{Data}(h)$. D outputs a scalar probability, which is trained to be high if the input was real and low if generated from G. A minimax objective is used to train both models

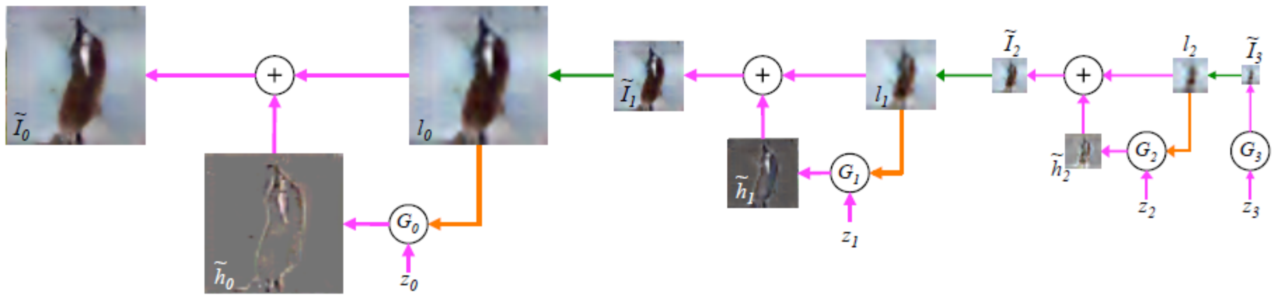

Figure 1. The sampling procedure for their LAPGAN model. They start with a noise sample z_3 and use a generative model G_3 to generate $\tilde{I}_3$. This is upsampled and then used as the conditioning variable l_2 for the generative model at the next level, G_2. Together with another noise sample z_2, G_2 generates a difference image $\tilde{h}_2$ which is added to l_2 to create $\tilde{I}_2$. This process repeats across two subsequent levels to yield a final full resolution sample I_0.

Table 1. Log-likelihood estimates for a standard GAN and their proposed LAPGAN model on CIFAR10 and STL10 datasets. The mean and std. dev. are given in units of nats/image. Rows 1 and 2 use a Parzen-window approach at full-resolution, while row 3 uses their multi-scale Parzen-window estimator.

Model	CIFAR10	STL10
GAN [5](Parzen window estimate)	-3617 ± 353	-3661 ± 347
LAPGAN(Parzen window estimate)	-3572 ± 345	-3563 ± 311
LAPGAN (multi-scale Parzen window estimate)	-1799 ± 826	-2906 ± 728

together:

$$\min_G \max_D \mathbb{E}_{h \sim p_{Data}(h)}[\log |D(h)] + \mathbb{E}_{z \sim p_{Noise}(z)}[\log(1 - D(G(z)))]. \quad (1)$$

This encourages G to fit $p_{Data}(h)$ so as to fool D with its generated samples $\tilde{h}$. Both G and D are trained by backpropagating the loss in Eq. 1 through both models to update the parameters.

2.2. Laplacian Pyramid

The Laplacian pyramid is a linear invertible image representation consisting of a set of band-pass images, spaced an octave apart, plus a low-frequency residual. Formally, let d(.) be a downsampling operation which blurs and decimates a $j \times j$ image I, so that d(I) is a new image of size $j/2 \times j/2$. Also, let u(.) be an upsampling operator which smooths and expands I to be twice the size, so u(I) is a new image of size $2j \times 2j$. They first build a Gaussian pyramid $G(I) = [I_0, I_1, \ldots, I_K]$, where $I_0 = I$ and I_k is k repeated applications of d(.) to I, *i.e.* $I_2 = d(d(I))$. K is the number of levels in the pyramid, selected so that the final level has very small spatial extent.

2.3. Laplacian Generative Adversarial Networks (LAPGAN)

Denton proposed approach combines the conditional GAN model with a Laplacian pyramid representation. The model is best explained by first considering the sampling procedure. Following training (explained below), they have a set of generative convnet models $G_0, \ldots, G_K$, each of which captures the distribution of coefficients hk for natural images at a different level of the Laplacian pyramid. Sampling an image is akin to the reconstruction procedure in equation above, except that the generative models are used to produce the $\tilde{h}_k$ as Eq. 2:

$$\tilde{I}_k = u(\tilde{I}_{k+1}) + \tilde{h}_k = u(\tilde{I}_{k+1}) + G_k(z_k, u(\tilde{I}_{k+1})). \quad (2)$$

The recurrence starts by setting $\tilde{I}_{k+1} = 0$ and using the model at the final level G_K to generate a residual image $\tilde{I}_k$ using noise vector $z_K : \tilde{I}_k = G_K(z_K)$. Note that models at all levels except the final are conditional generative models that take an upsampled version of the current image $\tilde{I}_{k+1}$ as a conditioning variable, in addition to the noise vector z_k. Fig. 1 shows this procedure in action for a pyramid with $K = 3$ using 4 generative models to sample a 64×64 image.

3. Evaluation of the model

Like Goodfellow *et al.* [5], Denton and his team are compelled to use a Gaussian Parzen window estimator to compute log-likelihood, since there no direct way of computing it using their model. Tab. 1 compares the log-likelihood on a validation set for their LAPGAN model and a standard GAN using 50k samples for each model. Their approach shows a marginal gain over a GAN. However, they can improve the underlying estimation technique by leveraging the multi-scale structure of the LAPGAN model. This new approach computes a probability at each scale of the Laplacian pyramid and combines them to give an overall image probability. The multi-scale Parzen estimate, shown in Tab. 1, produces a big gain over the traditional estimator.

References

[1] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 1

[2] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015. 1

[3] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999. 1

[4] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE CG&A*, 2013. 1

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[6] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. *Computer Science*, 2015. 1

[7] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 2000. 1

[8] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, 2005. 1

[9] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1

[10] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, 2011. 1