Paper review for "Kingma & Dhariwal, Glow: Generative flow with invertible 1 × 1 convolutions"

EECS 598 Paper Review - Week 7 - Changyuan Qiu

This paper explores the advance of flow-based generative models, which have gained little attention in the academic community compared with VAEs, GANs and autoregressive models. The flow-based generative models are promising due to tractability of the exact latent-variable inference and log-likelihood evaluation (while VAE only provides a lower bound for log-likelihood and only gives an approximation of the latent variable and GAN does not even have an encoder), and parallelizability of both training and synthesis (while synthesis for autoregressive models is difficult to parallelize and thus inefficient).

The core approach of this paper, referred to as *Glow: Generative Flow with Invertible 1 × 1 convolution*, consists of a series of 3 steps of flow: (1) actnorm (a scale and bias layer with data independent initialization which performs similarly as batch normalization), (2) invertible 1 × 1 convolution over channels, and (3) an affine coupling layer introduced by Dinh et al [1] This flow is combined with a multi-scale architecture due to space constraints.

There are limited prior works focusing on flow-based generative models, and the most inspiring prior works are NICE and RealNVP proposed by Dinh et al [1, 2]. Compared with the flow proposed in NICE and RealNVP that utilizes the equivalent of a permutation that reverses the ordering of the channels, this work replace and generalize this fixed permutation with a learnable invertible 1 × 1 convolution. **This core replacement brings about significant improvement in negative log-likelihood results (bits per dimension) on CIFAR-10, ImageNet and LSUN compared against RealNVP.** In addition, the authors trained the model on CelebA-HQ dataset (which consists of high resolutions (256 × 256) face images) and generated realistic random samples from the model. They also performed linear interpolation in the latent space between real images and the obtained manifold is smooth with realistic intermediate samples. They also tried manipulating attributes of images with the generated latent space and got good results on manipulation tasks like changing the color of hair. **These experiments demonstrated that the glow model could also scale to high resolutions, generate realistic high-resolution images, and produce a meaningful latent space that can be easily used for downstream tasks like manipulation of data.** Regarding computational efficiency, they reported that sampling 256 × 256 images with their largest model takes less than one second on current hardware.

Regarding limitations of this paper, I discovered that the largest face generation model uses ~200M parameters and ~600 convolution layers, which is quite expensive to train. Maybe future work could consider improving the architecture (for example, use self-attention architectures, or perform progressive training to scale to high resolutions) to make it computationally cheaper to train.

Reference

[1] Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: non-linear independent components estimation. *arXiv preprint arXiv:1410.8516.*

[2] Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803.*