Image-to-Image Translation with Conditional Adversarial Networks

Liangjie Cao

July 2, 2018

1. Method

To test the importance of conditioning the discriminator, the authors also compare to an unconditional variant in which the discriminator does not observe x:

$$L_GAN(G,D) = E_{y-p_{data}(y)}[logD(y)]+ \\ E_{x-p_{data}(x),z-p_z(z)}[log(1-D(G(x,z)))] \quad (1)$$

Previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as $L2$ distance. The discriminator's job remains unchanged, but the generator is tasked to not only fool the discriminator but also to be near the ground truth output in an $L2$ sense. They also explore this option, using $L1$ distance rather than $L2$ as $L1$ encourages less blurring:

$$L_{l1}(G) = E_{x,y,z}[\| \ y - G(x,z) \ \|_1] \quad (2)$$

Their final objective is:

$$G^* = \arg\min_G \max_D L_{cGAN}(G,D) + \lambda_{L1}(G) \quad (3)$$

Without z, the net could still learn a mapping from x to y, but would produce deterministic outputs, and therefore fail to match any distribution other than a delta function. Past conditional GANs have acknowledged this and pro- vided Gaussian noise z as an input to the generator, in addi- tion to x (*e.g.*, [3]). In initial experiments, they did not find this strategy effective the generator simply learned to ig- nore the noise which is consistent with Mathieu *et al.* [1]. Instead, for our final models, we provide noise only in the form of dropout, applied on several layers of our generator at both training and test time. Despite the dropout noise, the authors observe very minor stochasticity in the output of their nets. Designing conditional GANs that produce stochastic out- put, and thereby capture the full entropy of the conditional distributions they model, is an important question left open by the present work.

2. Generator with skips

A defining feature of image-to-image translation problems is that they map a high resolution input grid to a high resolu- tion output grid. In addition, for the problems they consider, the input and output differ in surface appearance, but both are renderings of the same underlying structure. Therefore, structure in the input is roughly aligned with structure in the output. They design the generator architecture around these considerations.

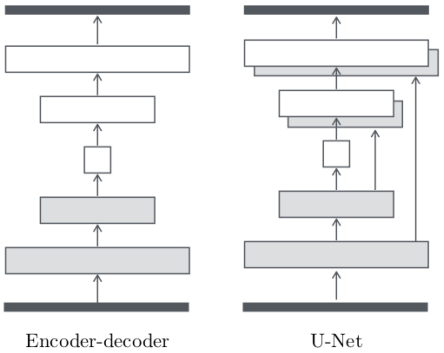

Figure 1. Two choices for the architecture of the generator. The "U-Net" [2] is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

In such a network, the input is passed through a series of lay- ers that progressively downsample, until a bottleneck layer, at which point the process is reversed (Figure 1). Such a network requires that all information flow pass through all the layers, including the bottleneck. For many image translation problems, there is a great deal of low-level information shared between the input and output, and it would be desirable to shuttle this information directly across the net. For example, in the case of image colorizaton, the input and output share the location of prominent edges.

To give the generator a means to circumvent the bot- tleneck for information like this, they add skip connections, following the general shape of a "U-Net" [2] (Figure 1). Specifically, they add skip connections between each layer i and layer $n-i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n-i$.

References

[1] M. Mathieu, C. Couprie, and Y. Lecun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2015. 1

[2] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[3] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 1