Image-to-Image Translation with Conditional Adversarial Networks

Yufeng Jiang

1. Optimization and inference

To optimize their networks, they follow the standard approach from [1]: they alternate between one gradient descent step on D, then one step on G. As suggested in the original GAN paper, rather than training G to minimize $\log(1 - D(x, G(x, z))$, they instead train to maximize $\log D(x, G(x, z))$ [1]. In addition, they divide the objective by 2while optimizing D, which slows down the rate at which D learns relative to G. They use minibatch SGD and apply the Adam slover, with learning rate 0.0002, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$.

At inference time, they run the generator net in exactly the same manner as during the training phase. This differs from the usual protocol in that they apply dropout at test time, and they apply batch normalization [2] using the statistics of the test batch, rather than aggregated statistics of the training batch. This approach to batch normalization, when the batch size is set to 1, has been termed "instance normalization" and has been demonstrated to be effective at image generation tasks [4]. In their experiments, they use batch sizes between 1 and 10 depending on the experiment.

2. Experiments

To explore the generality of conditional GANs, they test the method on a variety of tasks and datasets, including both graphics tasks, like photo generation, and vision tasks, like semantic segmentation.

2.1. Evaluation metrics

Evaluation the quality of synthesized images is an open and difficult problem. Traditional metrics such as perpixel mean-squared error do not assess joint statistics of the results, and therefore do not measure the very structure that structured losses aim to capture.

In order to more holistically evaluate the visual quality of their results, they employ two tactics. First, they run "real vs fake" perceptual studies on Amazon Mechanical Turk (AMT). For graphics problems like colorization and photo generation, plausibility to a human observer is often the ultimate goal. Therefore, they test their map generation, aerial photo generation, and image colorizaiton using this approach.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	**0.55**	**0.20**	**0.14**

Table 1. FCN-scores for different generator architectures (and objectives), evaluated on Cityscapes labels↔photos. (U-net (L1-cGAN) scores differ from those reported in other tables since batch size was 10 for this experiment and 1 for other tables, and random variation between training runs.)

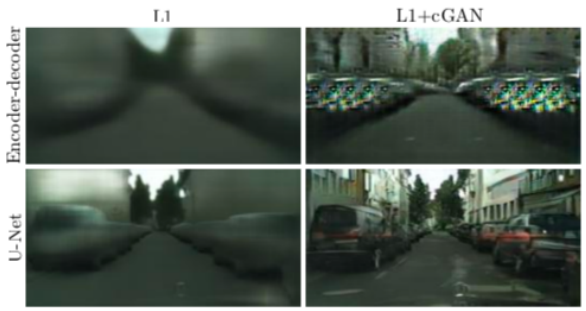

Figure 1. Adding skip connections to an encoder-decoder to create a "U-Net" results in much higher quality results.

Second, they measure whether or not their synthesized cityscapes are realistic enough that off-the-shelf recognition system can recognize the objects in them. This metric is similar to the "inception score" from [3], the object detection evaluation in [1], and the "semantic interpretabiity" measures in [5].

AMT perceptual studies For their AMT experiments, they followed the protocol from [5]: Turkers were presented with a series of trials that pitted a "real" image against a "fake" image generated by their algorithm. On each trial, each image appeared for 1 second, after which the images disappeared and Turkers were given unlimited time to respond as to which was fake. The first 10 iamges of each session were practice and Turkers were given feedback. No feedback was provided on the 40 trials of the main exeriment. Each session tested just one algorithm at a time, and Turkers were not allowed to complete more than one session.

"FCN-score" While quantitative evaluation of genera-

Figure 2. Patch size variations. Uncertainty in the output manifests itself differently for different loss functions. Uncertain regions become blurry and desaturated under L1. The 1×1 PixelGAN encourages greater color diversity but has no effect on spatial statistics. The 16×16 PatchGAN creates locally sharp results, but also leads to tiling artifacts beyond the scale it can observe. The 70×70 PatchGAN forces outputs that are sharp, even if incorrect, in both the spatial and spectral (colorfulness) dimensions. The full 286×286 ImageGAN produces results that are visually similar to the 70×70 PatchGAN, but somewat lower quality according to their FCN-score metric (Table 2). Please see https://phillipi.github.io/pix2pix/ for additional examples.

tive models is known to be challenging, recent works have tried using pre-trained semantic classifiers to measure the discriminability of the generated stimuli as a pseudo-metric. The intuition is that if the generatied images are realistic, classififers trained on real images will be able to classify the synthesized image correctlly as well.

2.2. Analysis of the generator architecture

A U-Net architecture allows low-level information to shortcut across the network. Does this lead to better results? Figure 1 and Table 1 compare the U-Net against an encoder-decoder on cityscape generation. The encoder-decoder is created simply by severing the skip connections in the U-Net. The encoder-decoder is unable to learn to generate realistic images in their experiments. The advantages of the U-Net appear not to be specific to conditional GANs: when both U-Net and encoder-decoder are trained with an L1 loss, the U-Net again achieves the superior results.

2.3. From PixelGANs to PatchGANs to ImageGANs

They test the effect of varying the patch size N of their discriminator receptive fields, from a 1×1 "PixelGAN" to a full 286×286 "ImageGAN"[1]. Figure 2 shows qualitative results of this analysis and Table 2 quantifies the effects using the FCN-score. Note that elsewhere in this paper, unless specified, all experiments use 70×70 PatchGANs, and for this section all experiments use an L1+cGAN loss.

Discriminator receptive field	**Per-pixel acc.**	**Per-class acc.**	**Class IOU**
1×1	0.39	0.15	0.10
16×16	0.65	0.23	**0.17**
70×70	**0.66**	**0.23**	**0.17**
286×286	0.42	0.16	0.11

Table 2. FCN-scores for different receptive field sizes of the discriminaotr, evaluated on Cityscapes lables→photos. Notes that input images are 256×256 pixels and larger rereptive fields are padded with zeros.

References

[1] I. 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] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 1

[4] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[5] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016. 1

[1] They achieve this variation in patch size by abjusting the depth of the GAN discriminator. Details of this process, and the discriminator architectures are provided in the supplemental materials online.