Image-to-Image Translation with Conditional Adversarial Networks

Liangjie Cao

Jun 30, 2018

Abstract

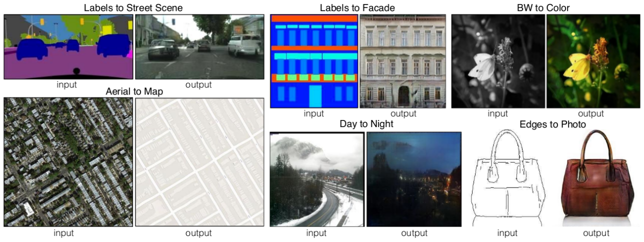

Figure 1. Many problems in image processing, graphics, and vision involve translating an input image into a corresponding output image. These problems are often treated with application-specific algorithms, even though the setting is always the same: map pixels to pixels. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Here we show results of the method on several. In each case we use the same architecture and objective, and simply train on different data

The authors investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. So they demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. As a commu- nity, they no longer hand-engineer our mapping functions, and this work suggests they can achieve reasonable results without hand-engineering our loss functions either.

*Many problems in image processing, computer graphics, and computer vision can be posed as translating an input image into a corresponding output image. Just as a concept may be expressed in either English or French, a scene may be rendered as an RGB image, a gradient field, an edge map, a semantic label map, etc. In analogy to automatic language translation, we define automatic image-to-image translation as the problem of translating one possible representation of a scene into another, given sufficient training data (see Fig- ure 1). One reason language translation is difficult is be- cause the mapping between languages is rarely one-to-one any given concept is easier to express in one language than another. Similarly, most image-to-image translation problems are either many-to-one (computer vision) map- ping photographs to edges, segments, or semantic labels, or one-to-many (computer graphics) mapping labels or sparse user inputs to realistic images. Traditionally, each of these tasks has been tackled with separate, special-purpose machinery (*e.g.,* [3, 7, 5, 1, 2, 10, 8, 9, 4, 11, 12]), despite the fact that the setting is always the same: predict pixels from pixels. Our goal in this paper is to develop a common framework for all these problems.*

The community has already taken significant steps in this direction, with convolutional neural nets (CNNs) becoming the common workhorse behind a wide variety of image prediction problems. CNNs learn to minimize a loss function an objective that scores the quality of results and although the learning process is automatic, a lot of manual effort still goes into designing effective losses. In other words, we still have to tell the CNN what we wish it to minimize. But, just like Midas, we must be careful what we wish for! If we take a naive approach, and ask the CNN to minimize Euclidean distance between predicted and ground truth pix- els, it will tend to produce blurry results [12]. This is because Euclidean distance is minimized by averaging all plausible outputs, which causes blurring. Coming up with loss functions that force the CNN to do what we really want e.g., *output sharp, realistic images is an open problem and generally requires expert knowledge.*

1. Method

GANs are generative models that learn a mapping from random noise vector z to output image y: $G : z \to y$ [6]. In contrast, conditional GANs learn a mapping from observed image x and random noise vector z, to y: $G : \{x, z\} \to y$. The generator G is trained to produce outputs that cannot be distinguished from "real" images by an ad- versarially trained discrimintor, D, which is trained to do as well as

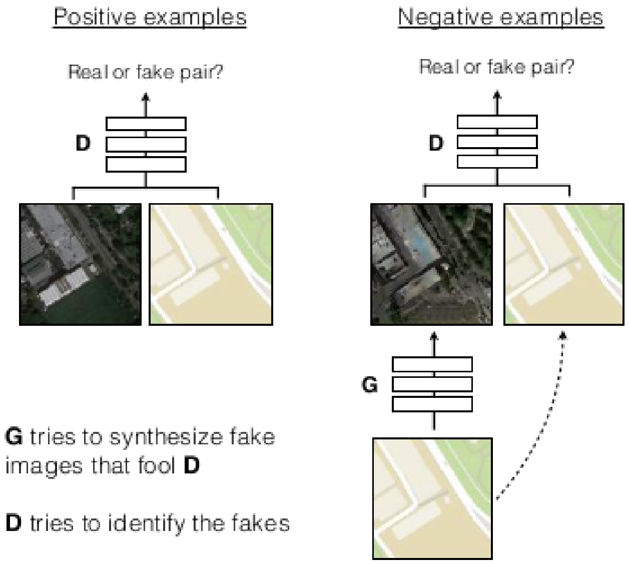
Positive examples
Negative examples
Real or fake pair?
Real or fake pair?
D
D
G
G tries to synthesize fake images that fool **D**
D tries to identify the fakes

Figure 2. Training a conditional GAN to predict aerial photos from maps. The discriminator, D, learns to classify between real and synthesized pairs. The generator learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe an input image

possible at detecting the generators "fakes". This training procedure is diagrammed in Figure 2.

The objective of a conditional GAN can be expressed as:

$$\begin{aligned} L_{cGAN}(G, D) = E_{x,y\ Pdata}[logD(x, y)]+ \\ E_{x\ Pdata(x),z\ p_z(z)}[log(1 - D(x, G(x, z)))] \end{aligned} \quad (1)$$

where G tries to minimize this objective against an ad- versarial D that tries to maximize it, i.e. $G = argmin_G \max_D L_{cGAN}(G, D)$.

References

[1] A. Buades, B. Coll, and J. M. Morel. A non-local algorithm for image denoising. In CVPR*, pages 60–65, 2005. 1*

[2] T. Chen, M. M. Cheng, P. Tan, A. Shamir, and S. M. Hu. Sketch2photo: internet image montage. TOG*, 2009. 1*

[3] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In SIGGRAPH*, pages 341–346, 2001. 1*

[4] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV*, 2014. 1*

[5] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. TOG*, 2006. 1*

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

[7] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In SIGGRAPH*, 2001. 1*

[8] P. Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. TOG*, 2014. 1*

[9] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR*, 2015. 1*

[10] Y. Shih, S. Paris, and W. T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. TOG*, 2013. 1*

[11] S. Xie and Z. Tu. Holistically-nested edge detection. In ICCV*, 2015. 1*

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