Conditional Generative Adversarial Nets

Mehdi Mirza
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7
mirzamom@iro.umontreal.ca

Simon Osindero
Flickr / Yahoo Inc.
San Francisco, CA 94103
osindero@yahoo-inc.com

Abstract

Generative Adversarial Nets [8] were recently introduced as a novel way to train generative models. In this work we introduce the conditional version of generative adversarial nets, which can be constructed by simply feeding the data, y, we wish to condition on to both the generator and discriminator. We show that this model can generate MNIST digits conditioned on class labels. We also illustrate how this model could be used to learn a multi-modal model, and provide preliminary examples of an application to image tagging in which we demonstrate how this approach can generate descriptive tags which are not part of training labels.

1 Introduction

Generative adversarial nets were recently introduced as an alternative framework for training generative models in order to sidestep the difficulty of approximating many intractable probabilistic computations.

Adversarial nets have the advantages that Markov chains are never needed, only backpropagation is used to obtain gradients, no inference is required during learning, and a wide variety of factors and interactions can easily be incorporated into the model.

Furthermore, as demonstrated in [8], it can produce state of the art log-likelihood estimates and realistic samples.

In an unconditioned generative model, there is no control on modes of the data being generated. However, by conditioning the model on additional information it is possible to direct the data generation process. Such conditioning could be based on class labels, on some part of data for inpainting like [5], or even on data from different modality.

In this work we show how can we construct the conditional adversarial net. And for empirical results we demonstrate two set of experiment. One on MNIST digit data set conditioned on class labels and one on MIR Flickr 25,000 dataset [10] for multi-modal learning.

2 Related Work

2.1 Multi-modal Learning For Image Labelling

Despite the many recent successes of supervised neural networks (and convolutional networks in particular) [13, 17], it remains challenging to scale such models to accommodate an extremely large number of predicted output categories. A second issue is that much of the work to date has focused on learning one-to-one mappings from input to output. However, many interesting problems are more naturally thought of as a probabilistic one-to-many mapping. For instance in the case of image labeling there may be many different tags that could appropriately applied to a given image, and different (human) annotators may use different (but typically synonymous or related) terms to describe the same image.

One way to help address the first issue is to leverage additional information from other modalities: for instance, by using natural language corpora to learn a vector representation for labels in which geometric relations are semantically meaningful. When making predictions in such spaces, we benefit from the fact that when prediction errors we are still often 'close' to the truth (e.g. predicting 'table' instead of 'chair'), and also from the fact that we can naturally make predictive generalizations to labels that were not seen during training time. Works such as [3] have shown that even a simple linear mapping from image feature-space to word-representation-space can yield improved classification performance.

One way to address the second problem is to use a conditional probabilistic generative model, the input is taken to be the conditioning variable and the one-to-many mapping is instantiated as a conditional predictive distribution.

[16] take a similar approach to this problem, and train a multi-modal Deep Boltzmann Machine on the MIR Flickr 25,000 dataset as we do in this work.

Additionally, in [12] the authors show how to train a supervised multi-modal neural language model, and they are able to generate descriptive sentence for images.

3 Conditional Adversarial Nets

3.1 Generative Adversarial Nets

Generative adversarial nets were recently introduced as a novel way to train a generative model. They consists of two 'adversarial' models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. Both G and D could be a non-linear mapping function, such as a multi-layer perceptron.

To learn a generator distribution p_g over data data $\boldsymbol{x}$, the generator builds a mapping function from a prior noise distribution $p_z(z)$ to data space as $G(z;\theta_g)$. And the discriminator, $D(x;\theta_d)$, outputs a single scalar representing the probability that $\boldsymbol{x}$ came form training data rather than p_g.

G and D are both trained simultaneously: we adjust parameters for G to minimize $log(1 - D(G(\boldsymbol{z}))$ and adjust parameters for D to minimize $logD(X)$, as if they are following the two-player min-max game with value function $V(G, D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{1}$$

3.2 Conditional Adversarial Nets

Generative adversarial nets can be extended to a conditional model if both the generator and discriminator are conditioned on some extra information $\boldsymbol{y}$. $\boldsymbol{y}$ could be any kind of auxiliary information, such as class labels or data from other modalities. We can perform the conditioning by feeding $\boldsymbol{y}$ into the both the discriminator and generator as additional input layer.