

BiGAN

Julia Heine and Marta Sawko

January 23, 2020

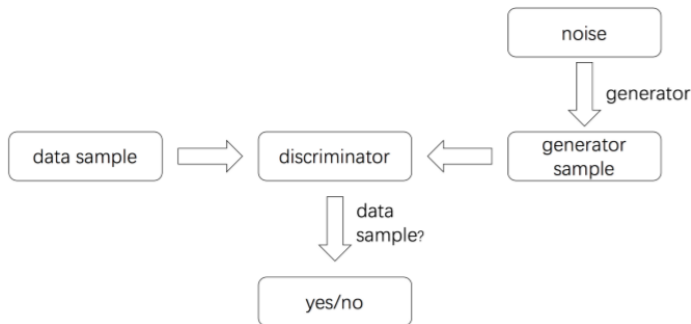
Overview

- 1 Introduction
- 2 Bidirectional GAN
- 3 Comparison
- 4 Examples
- 5 References

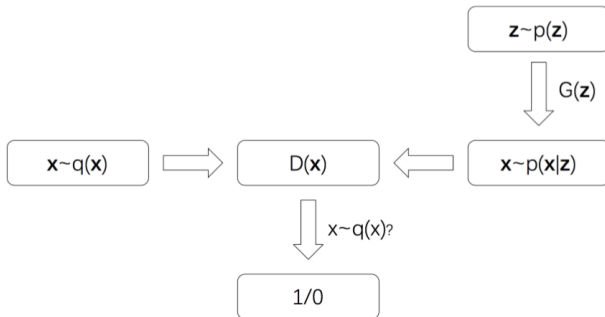
GAN

Strategy of GAN:

- generator tries to fool the discriminator by learning the conditional distribution $p(x) = \int p(z)p(x|z)$
- discriminator diminishes between real samples distribution $q(x)$ and generated ones $p(x)$



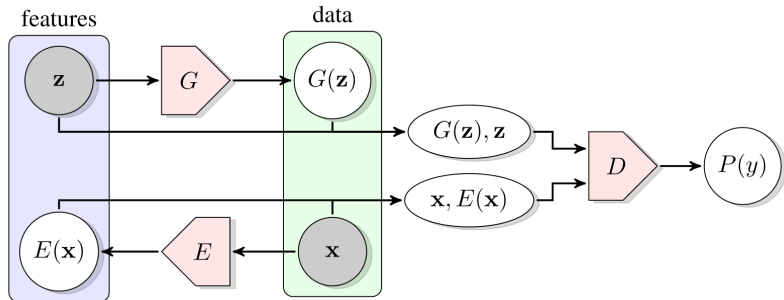
...or to make the diagram slightly more formal:



But what if we'd like to have an inverse mapping and project data back to latent space?

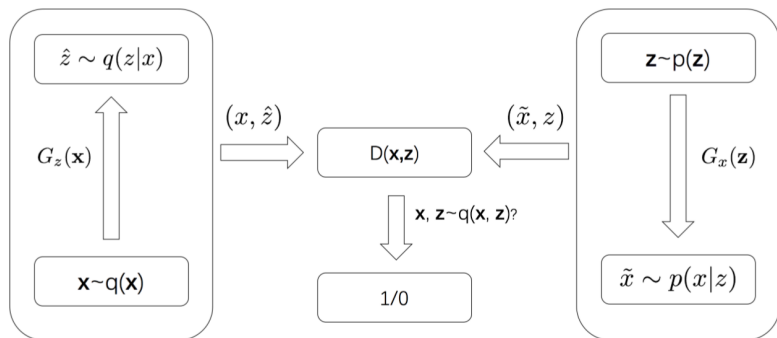
BiGAN: unsupervised feature learning framework

Let's introduce encoder E which maps data x to latent representations z .



D discriminates not only in data space (x versus $G(z)$), but jointly in data and latent space (tuples $(x, E(x))$ versus $(G(z), z)$), where the latent component is either an encoder output $E(x)$ or a generator input z .

BiGAN: the discriminator



The Discriminator is trained to diminish between joint samples (\mathbf{x}, \mathbf{z}) from:

- Encoder distribution $q(\mathbf{x}, \mathbf{z}) = q(\mathbf{x})q(\mathbf{z}|\mathbf{x})$
- Decoder distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$

BiGAN: labels for "free"

*"Because the BiGAN encoder learns to predict features z given data x , and prior work on GANs has demonstrated that these features capture semantic attributes of the data, we hypothesize that a trained BiGAN encoder **may serve as a useful feature representation** for related semantic tasks, in the same way that fully supervised visual models trained to predict semantic "labels" given images serve as powerful feature representations for related visual tasks. In this context, **a latent representation z may be thought of as a "label" for x , but one which came for "free," without the need for supervision.**"*

BiGAN: Encoder's learned feature representations

*"Interpolations in the latent space of the generator produce smooth and plausible semantic variations, and certain directions in this space correspond to particular semantic attributes along which the data distribution varies. For example, Radford et al. (2016) showed that a GAN trained on a database of human faces learns to **associate particular latent directions with gender and the presence of eyeglasses.**"*

BiGAN: Training objective

Minmax objective:

$$V(D, E, G) = \mathbb{E}_{x \sim p_x} [\mathbb{E}_{z \sim p_E(\cdot|x)} [\log D(x, z)]] \quad (1)$$

$$+ \mathbb{E}_{z \sim p_z} [\mathbb{E}_{x \sim p_G(\cdot|z)} [\log(1 - D(x, z))]], \quad (2)$$

or equivalently:

$$V(D, E, G) = \mathbb{E}_{x \sim p_x} [\log D(x, E(x))] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z), z))]. \quad (3)$$

Optimization is made with minimax objective using the same alternating gradient based optimization.

$$\min_{G, E} \max_D V(D, E, G) \quad (4)$$

Alternative methods: Discriminator

We may think about different tools serving as a feature representation learners.

Natural candidate is a GAN's **Discriminator**, taking intermediate representations of the real samples $x \sim p_x$.

Risk of this approach is that if generator generates well data with distribution $p_x(x)$, D may start ignoring input and predict:




$$P(Y = 1) = P(Y = 1|x) = \frac{1}{2} \quad (5)$$

Alternative methods: Latent Regressor
























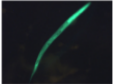

After normal GAN training we may think about the extra encoder that trains minimizing loss function $\mathcal{L}(z, E(G(z)))$. Problem of this approach is that, unlike the encoder in a BiGAN, the latent regressor encoder E is trained only on generated samples $G(z)$, and never “sees” real data. Let’s compare some methods, checking 1NN classification accuracy in MNIST:

BiGAN	D	LR	JLR	AE (ℓ_2)	AE (ℓ_1)
97.39	97.30	97.44	97.13	97.58	97.63

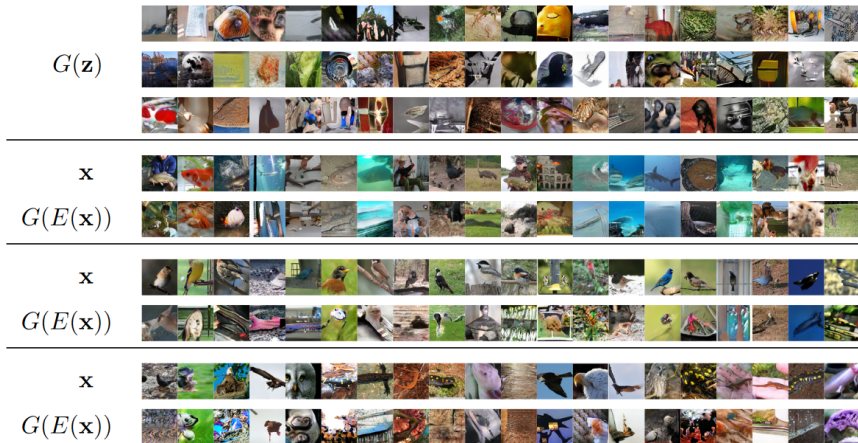
Table 1: One Nearest Neighbors (1NN) classification accuracy (%) on the permutation-invariant MNIST (LeCun et al., 1998) test set in the feature space learned by BiGAN, Latent Regressor (LR), Joint Latent Regressor (JLR), and an autoencoder (AE) using an ℓ_1 or ℓ_2 distance.

$G(z)$	
x	
$G(E(x))$	

ImageNet: nearest neighbour

Query	#1	#2	#3	#4
				
				
				
				
				

ImageNet: reconstruction



- [1] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Thank you!