InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

By Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel

Explainable Machine Learning
Peter Huegel
Heidelberg 12th July 2018

Heidelberg University

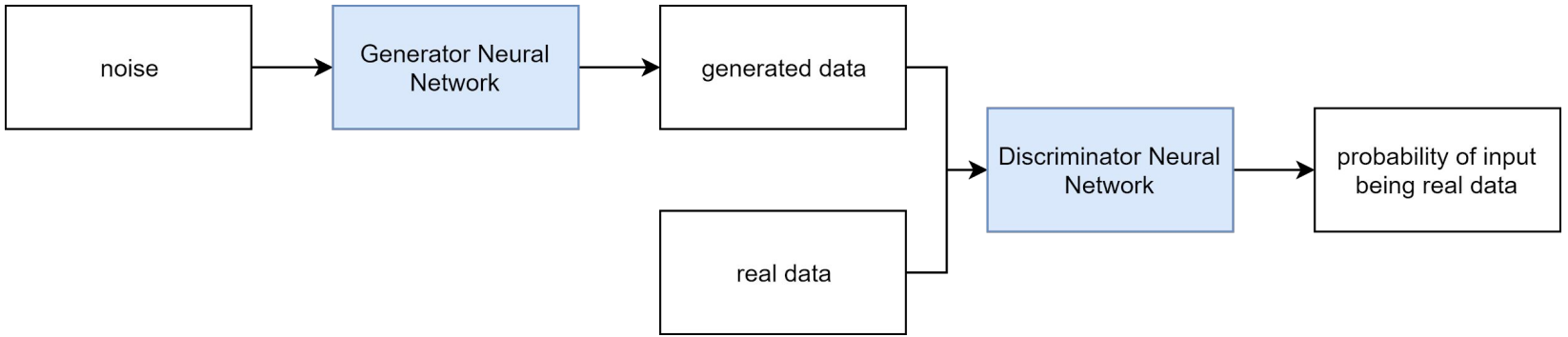
Table of Contents

A generative model trained through the competition between two neural nets:

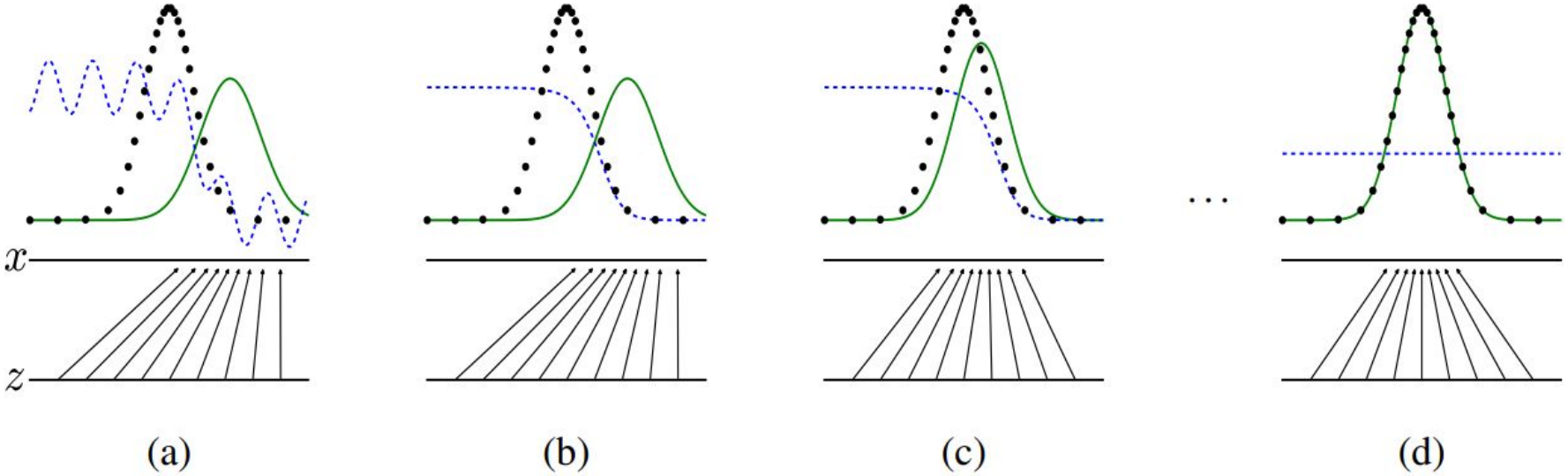
- Generator: $G(z), z \sim P_{noise}(z)$
  $\hat{x} = G(z)$: Sample generated by the generator
  $p_{noise}(z)$: arbitrary noise distribution
- Discriminator: $D(x) \in [0, 1]$
  probability of x being from the true data distribution P_{data} rather than a generated sample from the generator's distribution P_G.
- Optimization of both through the following minimax game:

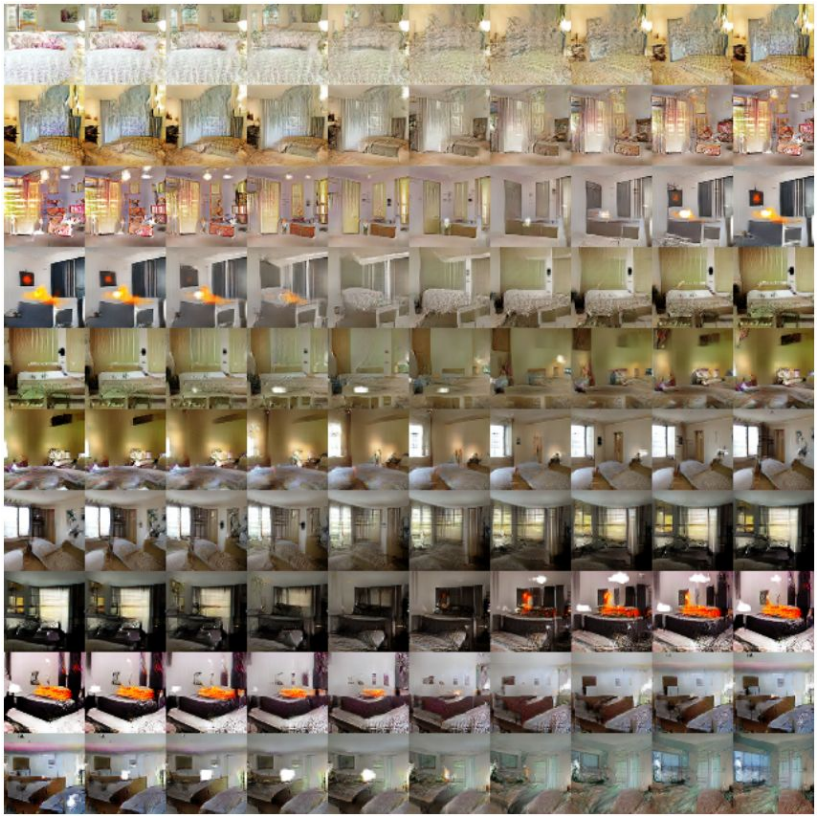
$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim \text{noise}}[\log (1 - D(G(z)))]$$

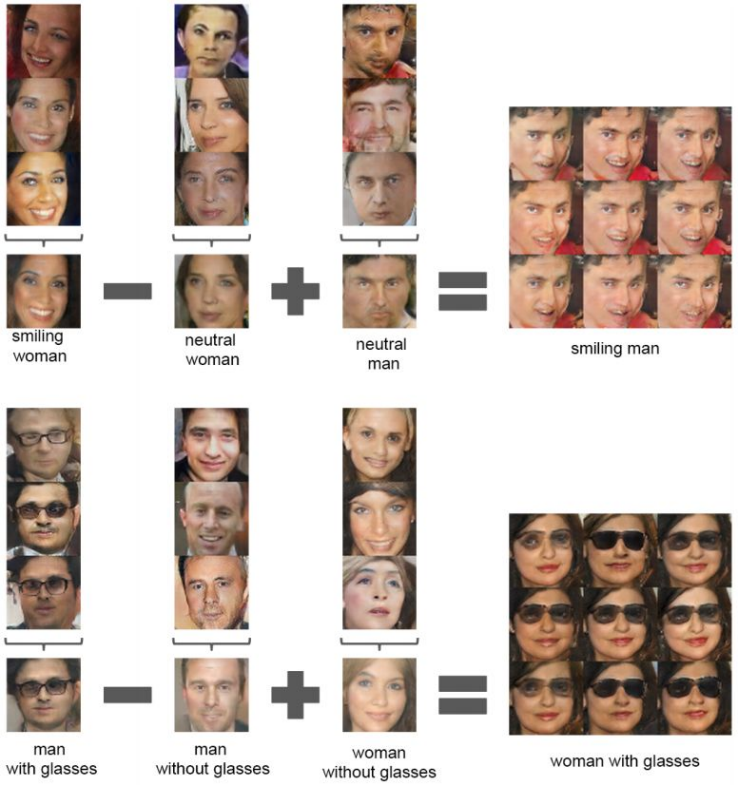
Structure of a Generative Adversarial Net:



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim \text{noise}}[\log(1 - D(G(z)))]$$



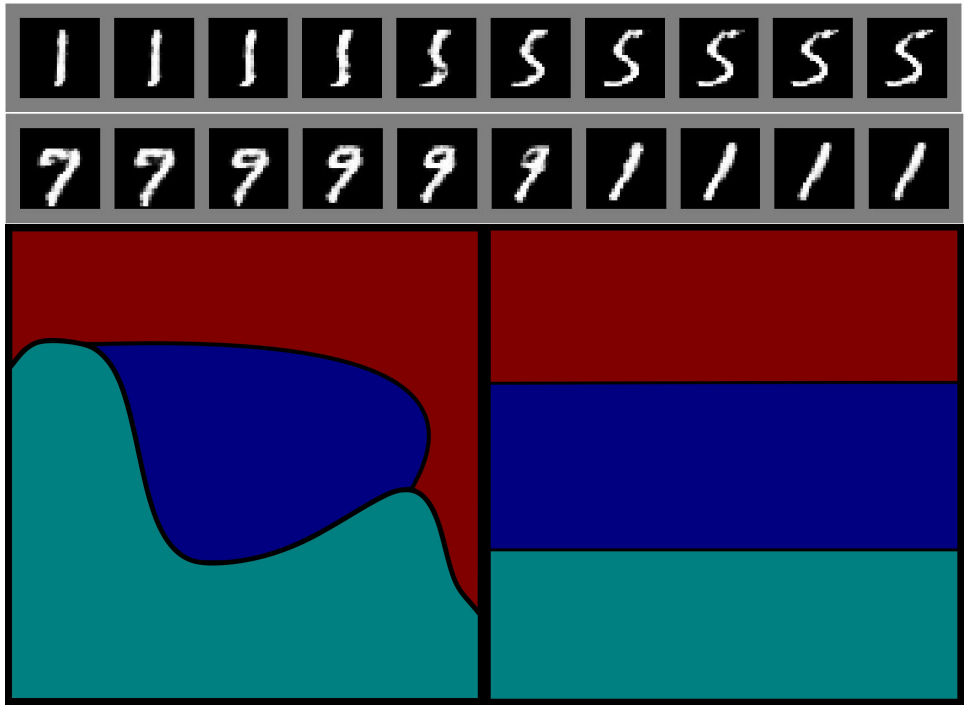



smiling woman
neutral woman
neutral man
smiling man
man with glasses
man without glasses
woman without glasses
woman with glasses

In general, GANs produce entangled representations.

In order to be more interpretable and easier to apply to tasks, a disentangled representation is desired.

A disentangled representation would allow for one parameter to specify an important feature of the generated data. In the case of MNIST, this could be the identity of the generated number.



Desired feature dimensions for MNIST dataset:

- Numerical identity (0-9)
- Angle of rotation
- Stroke thickness

Desired feature dimensions for a dataset of faces:

- Facial expression
- Eye color
- Hairstyle
- Glasses

InfoGAN

- Learns a disentangled representation of the dataset.
- Instead of just using the noise z for the generator, the noise is split into noise z and latent code c:

$$G(z)$$
$$G(z, c)$$

- We know that the MNIST dataset contains ten different types of digits.
- The latent code can be modeled to have a categorical code that represents the type of digit.

$$c_1 \sim \mathrm{Cat}(K = 10, p = 0.1)$$

- GAN is not required to use the specified latent code and is free to choose to ignore it:

$$P_G(x|c) = P_G(x)$$

- To force the network to make use of the latent code, mutual information is measured needs to be measured.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- The value function is adjusted to achieve maximal mutual information between the latent code and the image generated from noise and the latent code:

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

- To compute the mutual information $I(c; G(z, c))$, the posterior probability $P(c|x)$ is required.
- Calculating the posterior probability is difficult.
- Instead, an auxiliary distribution $Q(c|x)$, that approximates the posterior probability, is calculated with a neural network.
- With this approximation of the posterior probability, a lower boundary can be calculated:

$$I(c; G(z, c)) \geq L_I(G, Q)$$
$$= E_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c)$$

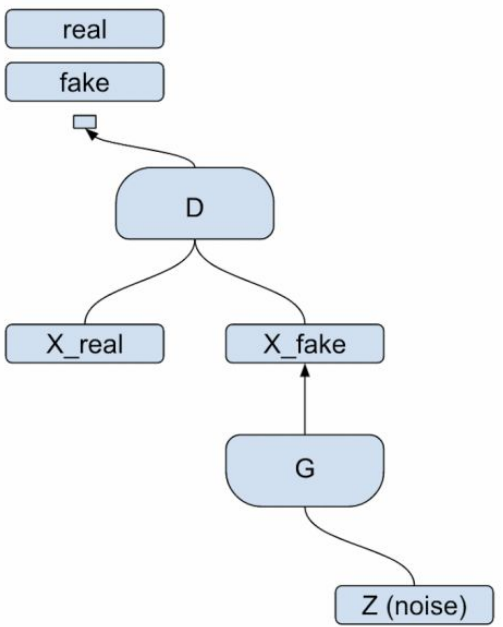
Final form of the new value function:

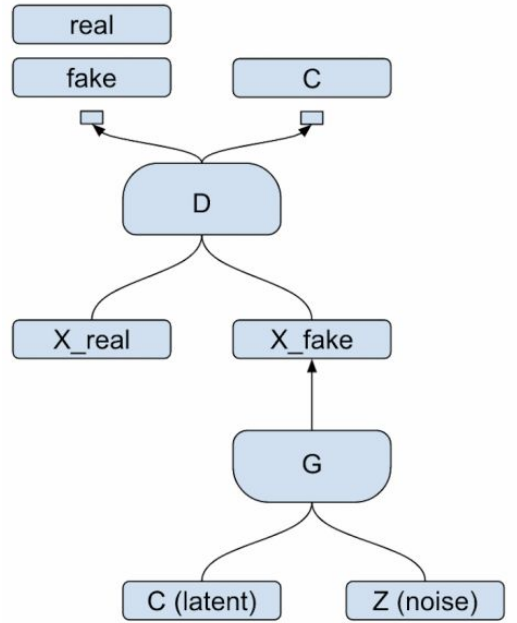
$$\min_{G,Q} \max_{D} V_{\text{InfoGAN}}(D, G, Q) = V(D, G) - \lambda L_I(G, Q)$$

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim \text{noise}}[\log (1 - D(G(z)))]$$

$$L_I(G, Q) = E_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c)$$

- In practice, the neural network approximating the posterior probability is one fully connected layer that is attached to the final layer of the discriminator.





Reflection

- InfoGAN is unsupervised. No labels are required for the training data.
- For MNIST, the following latent codes were specified:

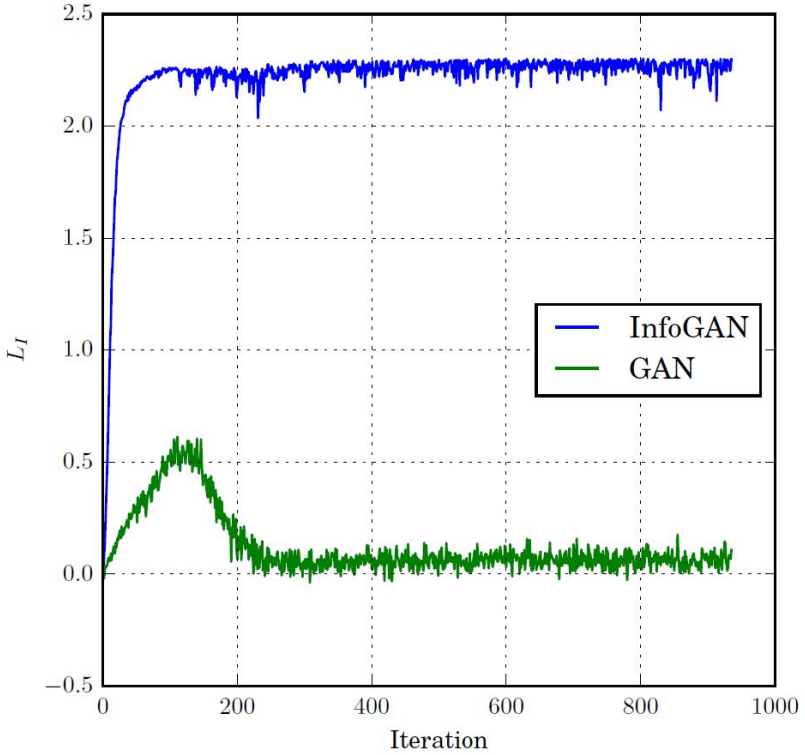
$$c_1 \sim \mathrm{Cat}(K = 10, p = 0.1)$$
$$c_2, c_3 \sim \mathrm{Unif}(-1, 1)$$

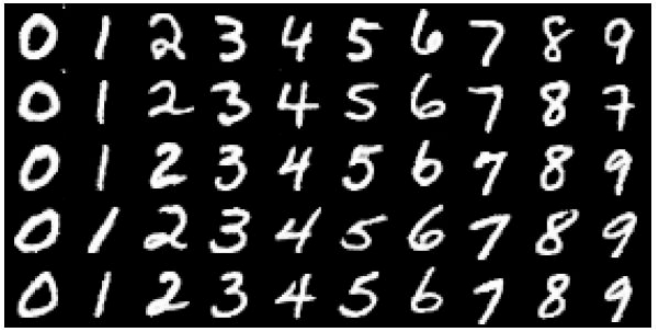
- Without using the labels of MNIST, a 5% error rate was achieved when matching each digit type to one category of the latent code corresponding to the numerical identity.

Comparison of $L_I(G, Q)$ with and without explicitly encouraging maximal mutual information. The latent code is a uniform categorical distribution $c \sim \text{Cat}(K = 10, p = 0.1)$.

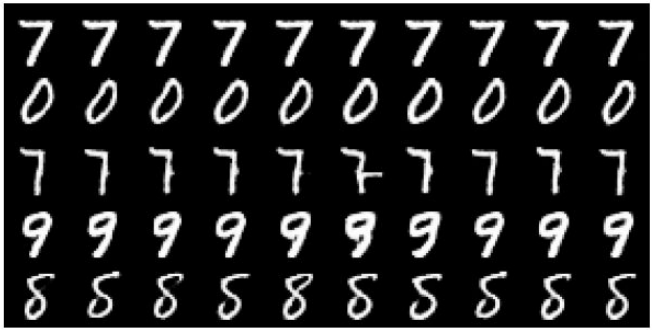
$L_I(G, Q)$ is quickly maximized to $H(c) \approx 2.30$, as it's first term becomes zero.

$$L_I(G, Q) = E_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c)$$

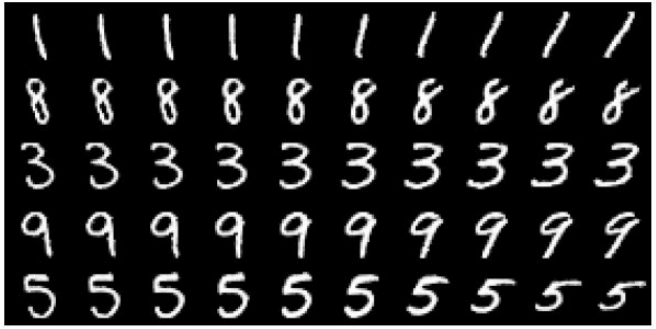




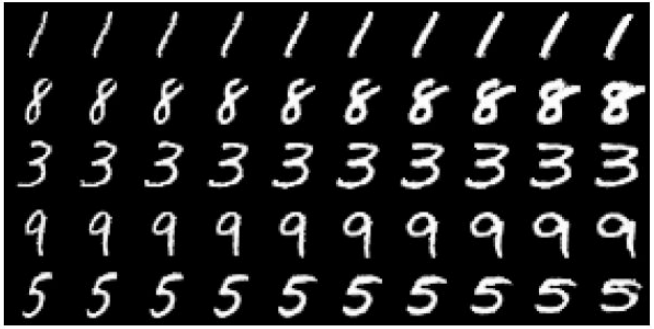
(a) Varying c_1 on InfoGAN (Digit type)



(b) Varying c_1 on regular GAN (No clear meaning)



(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)



(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

(a) Azimuth (pose)

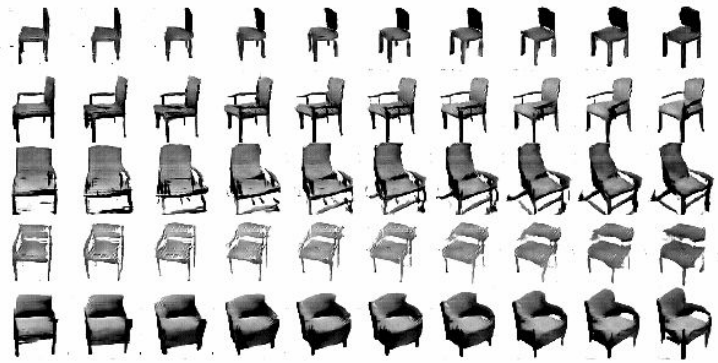
(b) Elevation

(c) Lighting

(d) Wide or Narrow

(a) Continuous variation: Lighting

(b) Discrete variation: Plate Context



(a) Rotation

(b) Width

(a) Azimuth (pose)

(b) Presence or absence of glasses

(c) Hair style

(d) Emotion

Sources

- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P., 2016. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets”. In *Advances in neural information processing systems*(pp. 2172-2180).
- Evtimova, K. and Drozdov, A., “Understanding Mutual Information and its Use in InfoGAN”.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. “Generative adversarial nets”. In *Advances in neural information processing systems* (pp. 2672-2680).
- Radford, A., Metz, L. and Chintala, S., 2015. “Unsupervised representation learning with deep convolutional generative adversarial networks”. *arXiv preprint arXiv:1511.06434*.