Inferencing Based on Unsupervised Learning of Disentangled Representations

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Tobias Hinz and Stefan Wermter

University of Hamburg, Department of Informatics, Knowledge Technology
Vogt-Kölln-Straße 30, D - 22527 Hamburg, Germany

Knowledge Technology

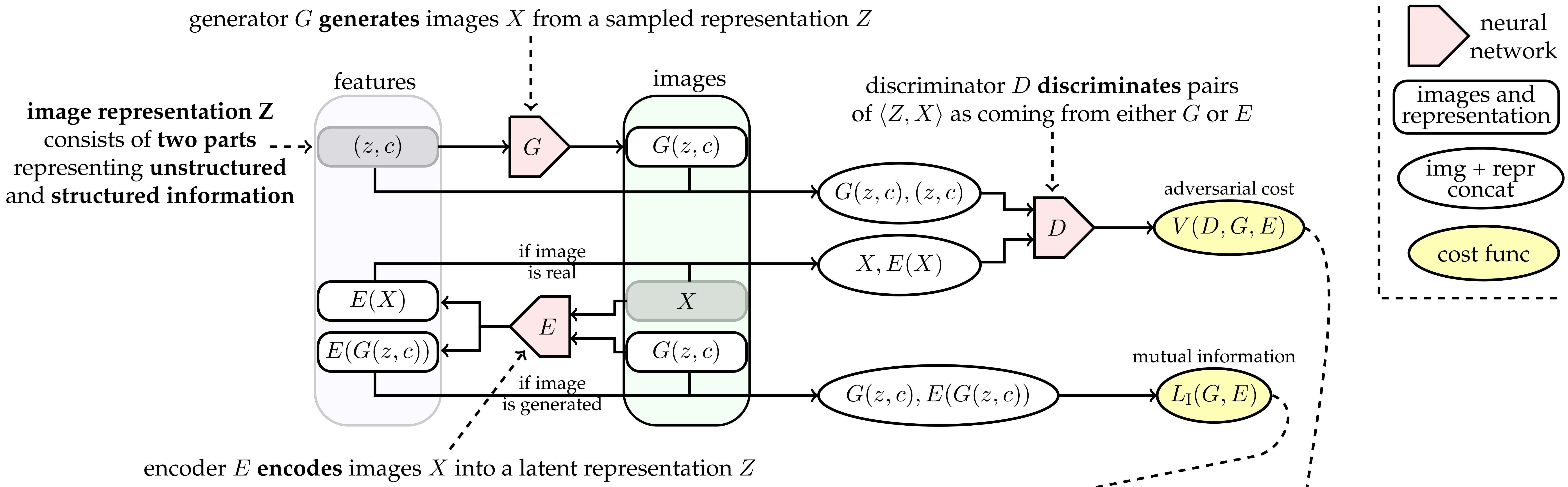
1. Motivation

- **Goal**: learn **meaningful information** about data in an **unsupervised** and **interpretable** way by encoding **data generating factors** into **disentangled representations**
- **Idea:** combine **Generative Adversarial Networks (GANs)** with an **encoder** that learns the **inverse of the generator**
- Use the generator and the encoder to **learn the underlying data generating factors** from the data without the need for explicit labels
- Use **disentangled representations** to make the data generating factors **explicitly accessible** within the learned representation

Additional information: `https://github.com/tohinz/Bidirectional-InfoGAN`

GitHub

2. Model: Bidirectional-InfoGAN



maximize the mutual information I between c and $G(z,c)$ to force the generator **to use the information provided in** c by maximizing the lower bound:

$$L_I(G,E) = \mathbb{E}_{c\sim P(c), z\sim P(z), X\sim G(z,c)}[log\ E(c|X)] + H(c) \leq I(c; G(z,c))$$

original **GAN minimax game** is extended and becomes:

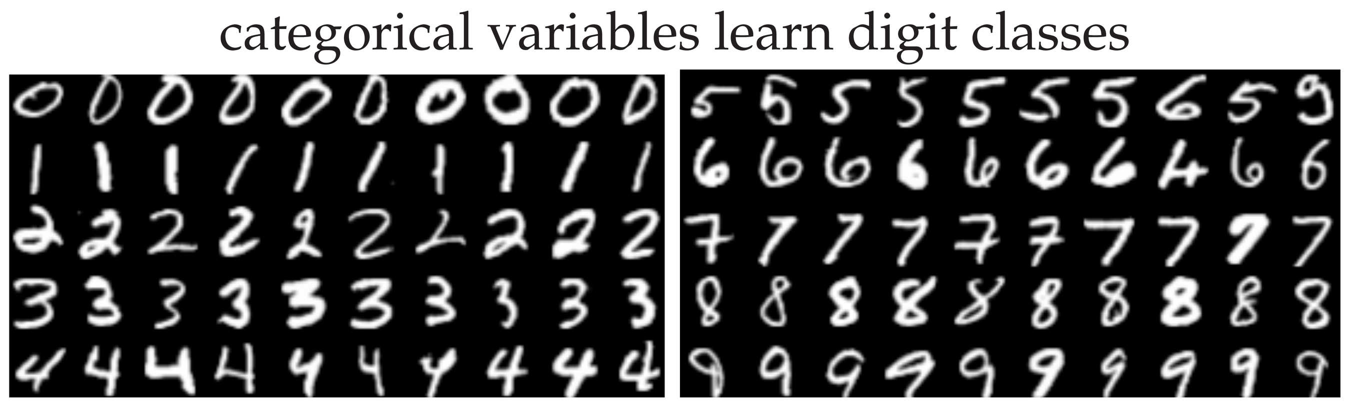
$$\min_{G,E}\max_{D} V(D,G,E) = \mathbb{E}_{X\sim P_{\text{data}}}[log D(X, E(X))] + \mathbb{E}_{Z\sim P_Z}[log(1 - D(G(Z), Z))]$$

final minimax game for the Bidirectional-InfoGAN (BInfoGAN) is:

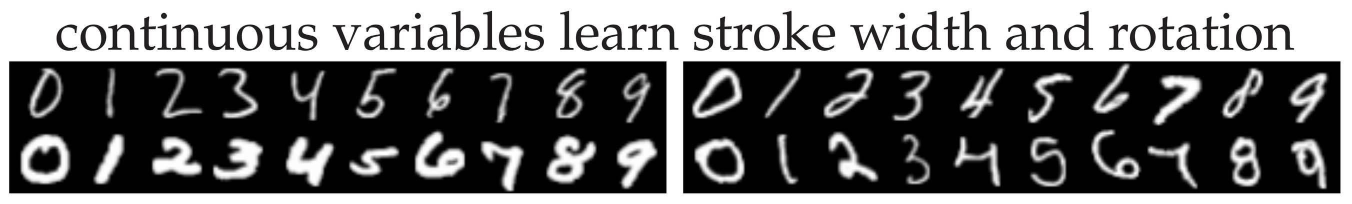
$$\min_{G,E}\max_{D} V_{\text{BInfoGAN}}(D,G,E) = V(D,G,E) - \lambda L_I(G,E)$$

3. Experiments and Results on the MNIST, SVHN, and CelebA data sets

categorical variables learn digit classes

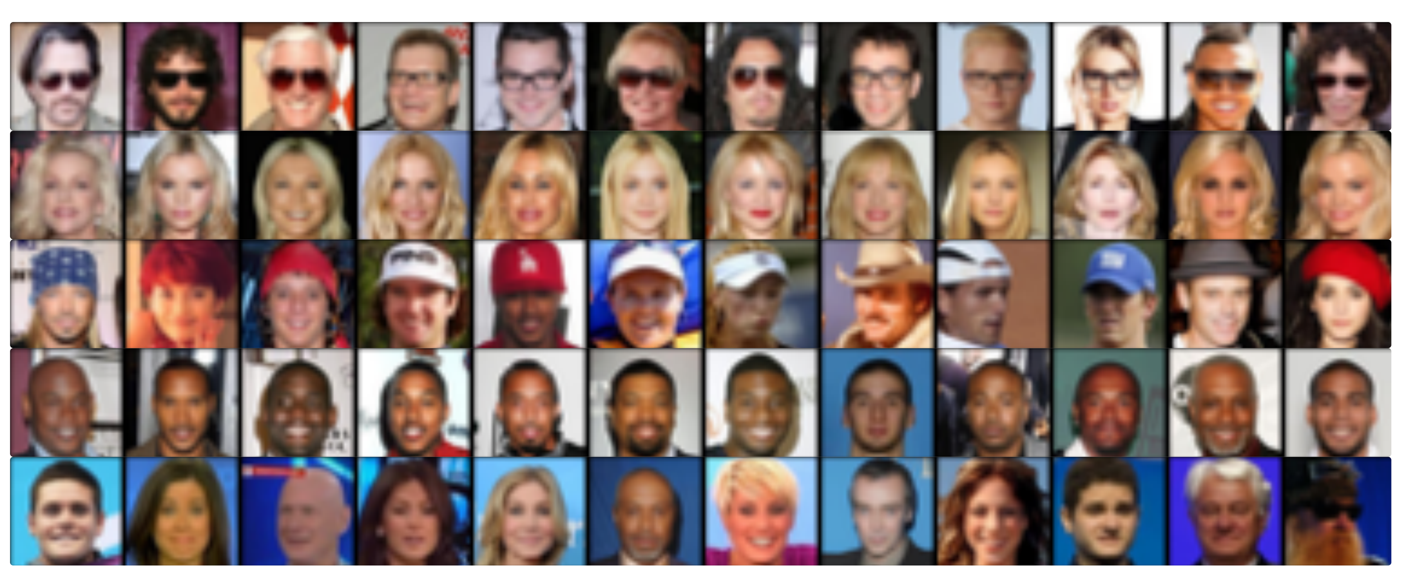


continuous variables learn stroke width and rotation



blue background
4 (dark background)
4 (light background)
1
0

glasses
blond hair
hats
skin tone
background



The Bidirectional-InfoGAN **learns to encode distinct visual characteristics** such as different digit classes, background, contrast, and facial characteristics **without the need for any labels during training**.

4. Outcome

- **Combination of an encoder and a generator** in a GAN can learn **disentangled representations of the data without any supervision**
- **Learned characteristics** are often **meaningful** and **interpretable**

Key References

[1] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proc. NIPS. pp. 2172–2180 (2016)

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. NIPS. pp. 2672–2680 (2014)