NIPS2016読み会

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen, Yan Duan, Rein Houthooft, John Schulman,
Ilya Sutskever, Pieter Abbeel (UC Berkeley, Open AI)

Presenter: Shuhei M. Yoshida (Dept. of Physics, UTokyo)

Goal

Unsupervised learning of disentangled representations

Approach

GANs + Maximizing Mutual Information
between generated images and input codes

Benefit

Interpretable representation obtained
without supervision and substantial additional costs

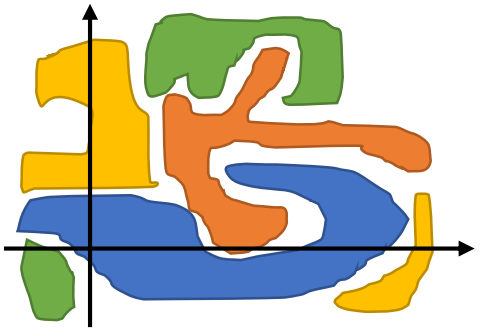
Reference
https://arxiv.org/abs/1606.03657 (with Appendix sections)
Implementations
https://github.com/openai/InfoGAN (by the authors, with TensorFlow)
https://github.com/yoshum/InfoGAN (by the presenter, with Chainer)

Motivation

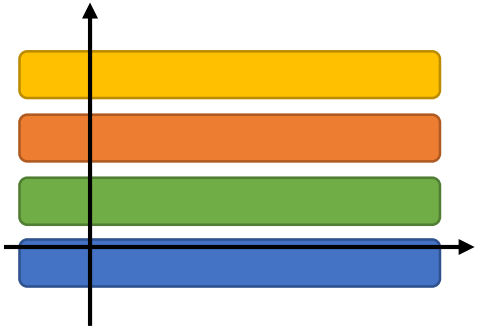
How can we achieve

unsupervised learning of disentangled representation?

In general, learned representation is entangled,
i.e. encoded in a data space in a complicated manner



When a representation is **disentangled**, it would be
more interpretable and easier to apply to tasks



Related works which the presenter has almost no knowledge about.

- Unsupervised learning of representation (no mechanism to force disentanglement)
  - ✓Stacked (often denoising) autoencoder, RBM
  - ✓Many others, including semi-supervised approach
- Supervised learning of disentangled representation
  - ✓Bilinear models, multi-view perceptron
  - ✓VAEs, adversarial autoencoders
- Weakly supervised learning of disentangled representation
  - ✓disBM, DC-IGN
- Unsupervised learning of disentangled representation
  - ✓hossRBM, applicable only to discrete latent factors

This work:
Unsupervised learning of disentangled representation
applicable to both continuous and discrete latent factors

Generative Adversarial Nets(GANs)

Generative model trained by competition between two neural nets:

- ✓Generator $x = G(z),\ z \sim p_z(Z)$
  $p_z(Z)$: an arbitrary noise distribution
- ✓Discriminator $D(x) \in [0,1]$:
  probability that x is sampled from the data dist. $p_{\text{data}}(X)$ rather than generated by the generator $G(z)$

Optimization problem to solve:

$$\min_G \max_D V_{\text{GAN}}(G,D)\text{, where}$$

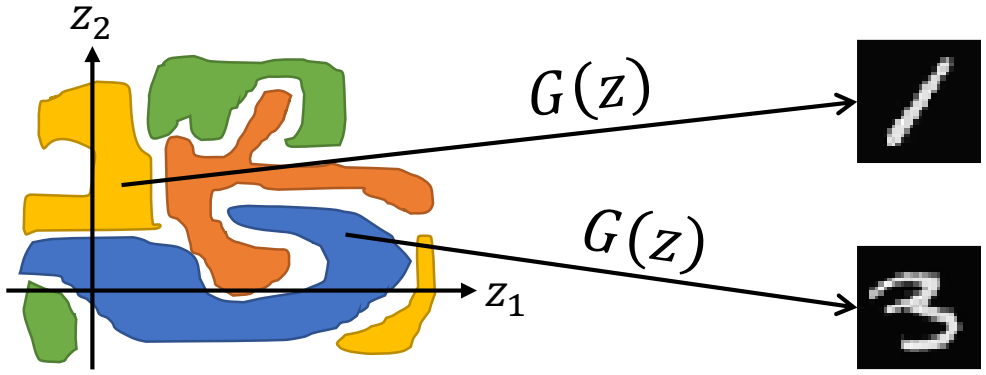
$$V_{\text{GAN}}(G,D) \equiv E_{x\sim p_{\text{data}}(X)}[\ln D(x)] + E_{z\sim p_z(Z)}\left[\ln\left(1 - D\left(G(z)\right)\right)\right]$$

Problems with GANs

From the perspective of representation learning:

✓No restrictions on how $G(z)$ uses z

- z can be used in a highly entangled way
- Each dimension of z does not represent any salient feature of the training data



Proposed Resolution: InfoGAN -Maximizing Mutual Information -

Observation in conventional GANs:

a generated date x does not have much information on the noise z from which x is generated

because of heavily entangled use of z

Proposed resolution = InfoGAN:

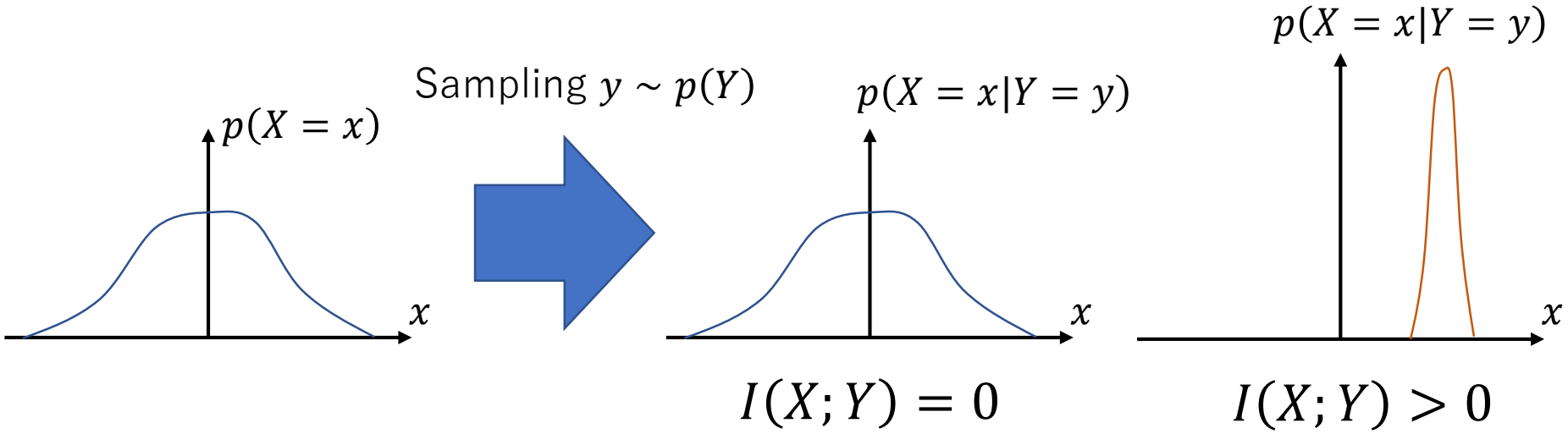
the generator $G(z, c)$ trained so that it maximize the mutual information $I(C|X)$ between the latent code C and the generated data X

$$\min_G \max_D \{V_{\mathrm{GAN}}(G, D) - \lambda I(C|X = G(Z, C))\}$$

Mutual Information

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

- $H(X) = E_{x\sim p(X)}[-\ln p(X = x)]$:
  Entropy of the prior distribution
- $H(X|Y) = E_{y\sim p(Y),x\sim p(X|Y=y)}[-\ln p(X = x|Y = y)]$:
  Entropy of the posterior distribution



Avoiding increase of calculation costs

Major difficulty:

Evaluation of $I(C|X)$ based on
evaluation and sampling from the posterior $p(C|X)$

Two strategies:

- ✓Variational maximization of mutual information
  - ✓Use an approximate function $Q(c|x) = p(C = c|X = x)$
- ✓Sharing the neural net
  between $Q(c|x)$ and the discriminator $D(x)$

Variational Maximization of MI

For an arbitrary function $Q(c, x)$,

$$E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln p(C=c|X=x)]$$

$$= E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)] + E_{x\sim p_G(X),c\sim p(C|X=x)}\left[\ln \frac{p(C=c|X=x)}{Q(c,x)}\right]$$

$$= E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)] + E_{x\sim p_G(X)}\left[D_{\mathrm{KL}}\left(p(C|X=x)||Q(C,x)\right)\right]$$

$$\geq E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)] \quad (\because \text{positivity of KL divergence})$$

Variational Maximization of MI

✓ With $Q(c,x)$ approximating $p(C=c|X=x)$, we obtain an variational estimate of the mutual information:

$$L_I(G,Q) \equiv E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)] + H(C)$$
$$\lesssim I(C|X=G(Z,C))$$

✓ Maximizing $L_I(G,Q)$ w.r.t. G and Q

$\Leftrightarrow$

- Achieving the equality by setting $Q(c,x) = p(C=c|X=x)$
- Maximizing the mutual information

Optimization problem to solve in InfoGAN:

$$\min_{G,Q}\max_{D}\{V_{\mathrm{GAN}}(G,D) - \lambda L_I(G,Q)\}$$

Eliminate sampling from posterior

Lemma

$$E_{x\sim p(X),y\sim p(Y|X=x)}[f(x,y)] = E_{x\sim p(X),y\sim p(Y|X=x),x'\sim p(X'|Y=y)}[f(x',y)].$$

By using this lemma and noting that

$$E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)]$$
$$= E_{c\sim p(C),z\sim p_Z(Z),x=G(z,c),c\sim p(C|X=x)}[\ln Q(c,x)],$$

we can eliminate the sampling from $p(C|X=x)$:

$$E_{x\sim p_G(X),c\sim p(C|X=x)}[\ln Q(c,x)] = \boldsymbol{E_{c\sim p(C),z\sim p_Z(Z),x=G(z,c)}\,[\ln Q(c,x)]}$$

Easy to estimate!

Proof of lemma

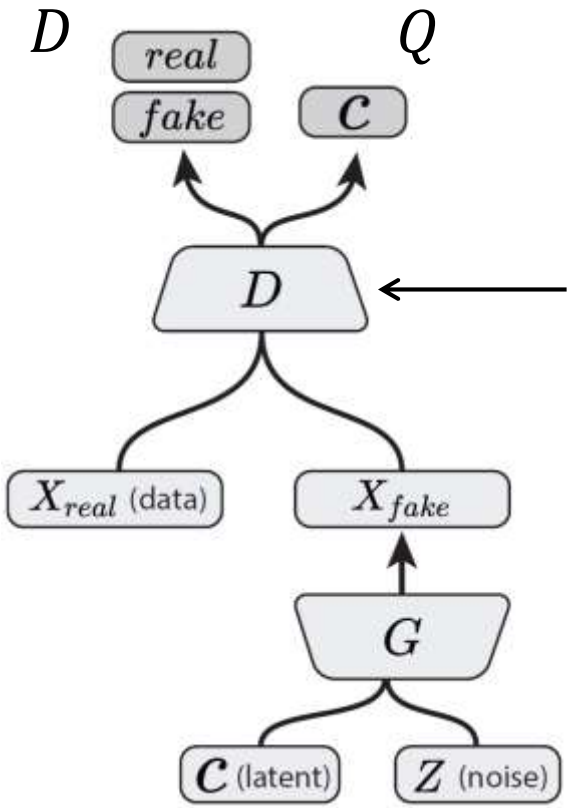
Lemma

$$E_{x\sim p(X),y\sim p(Y|X=x)}[f(x,y)] = E_{x\sim p(X),y\sim p(Y|X=x),x'\sim p(X'|Y=y)}[f(x',y)].$$

$$\begin{aligned}
\because \text{l.h.s.} &= \int_x \int_y p(X=x)p(Y=y|X=x)f(x,y) \\
&= \int_x \int_y p(Y=y)p(X=x|Y=y)f(x,y) \quad \because \text{Bayes' theorem} \\
&= \int_x \int_y \int_{x'} p(X=x',Y=y)\, p(X=x|Y=y)f(x,y) \\
&= \int_x \int_y \int_{x'} p(X=x')p(Y=y|X=x')p(X=x|Y=y)f(x,y) \\
&= \text{r.h.s.}
\end{aligned}$$

Sharing layers between D and Q

- ✓Model $Q(c, x)$ using neural network
- ✓Reduce the calculation costs by
  sharing all the convolution layers with D



Convolution layers of the discriminator

Given DCGANs,
InfoGAN comes for negligible additional costs!

Image from Odena, *et al.*, arXiv:1610.09585.

Experiment – MI Maximization

- InfoGAN on MNIST dataset
- Latent code c
  $=$ 10-class categorical code

L_I quickly saturates to
$H(c) = \ln 10 \sim 2.3$ in InfoGAN

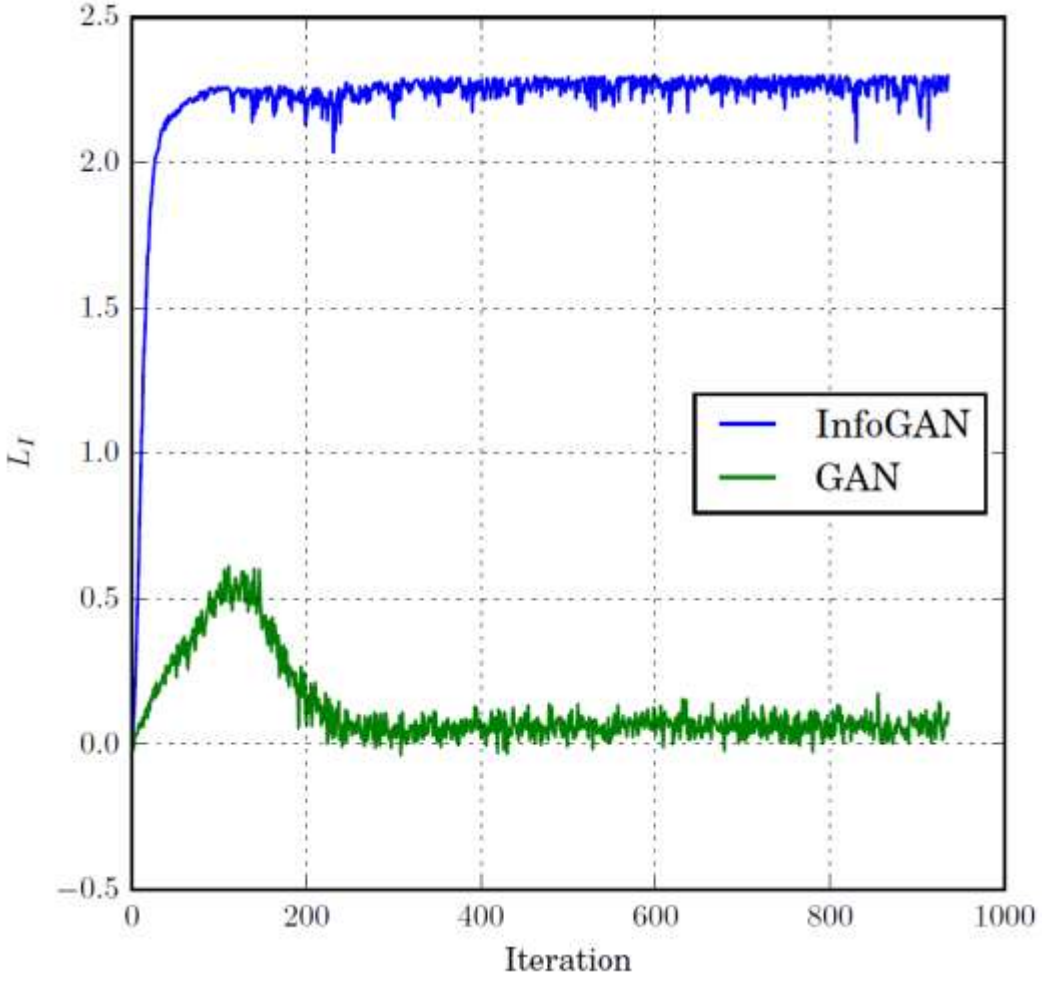


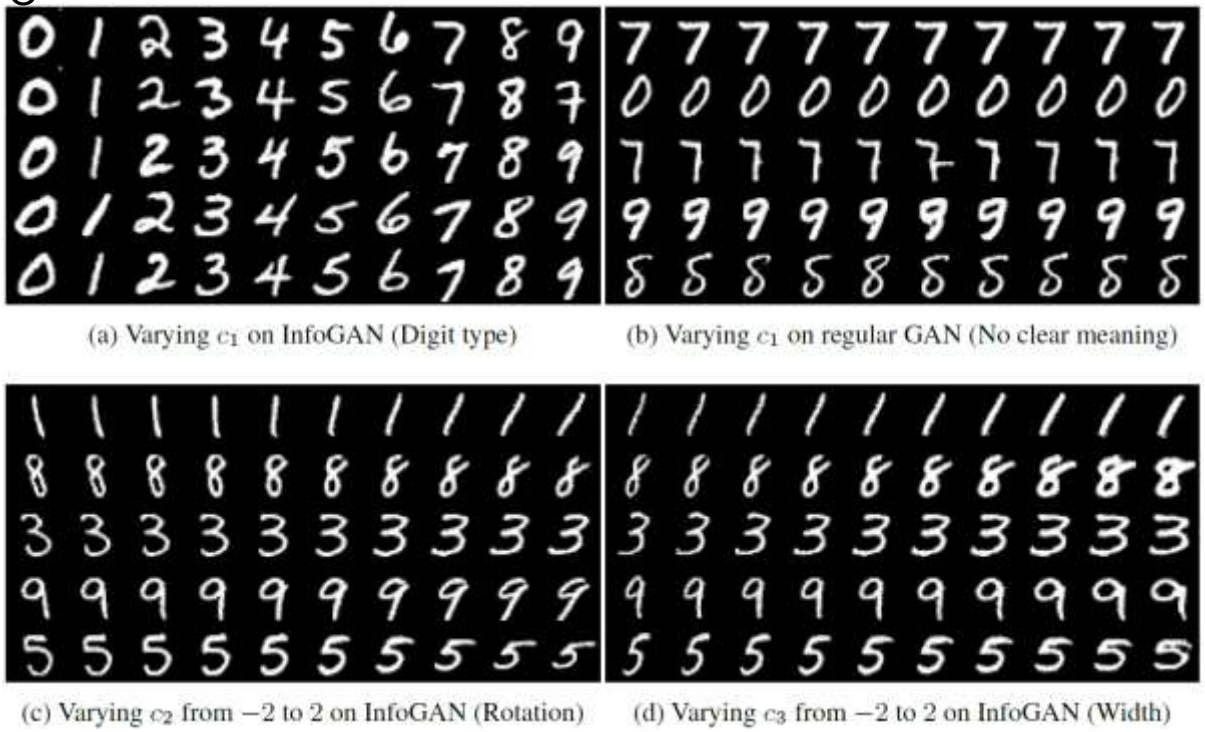
Figure 1 in the original paper

Experiment
– Disentangled Representation –

- InfoGAN on MNIST dataset
- Latent codes
  - ✓ c_1: 10-class categorical code
  - ✓ c_2, c_3: continuous code

✓ c_1 can be used as a classifier with 5% error rate.

✓ c_2 and c_3 captured the rotation and width, respectively



(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)

Figure 2 in the original paper

Experiment
– Disentangled Representation –

Dataset: P. Paysan, *et al.*, AVSS, 2009, pp. 296–301.

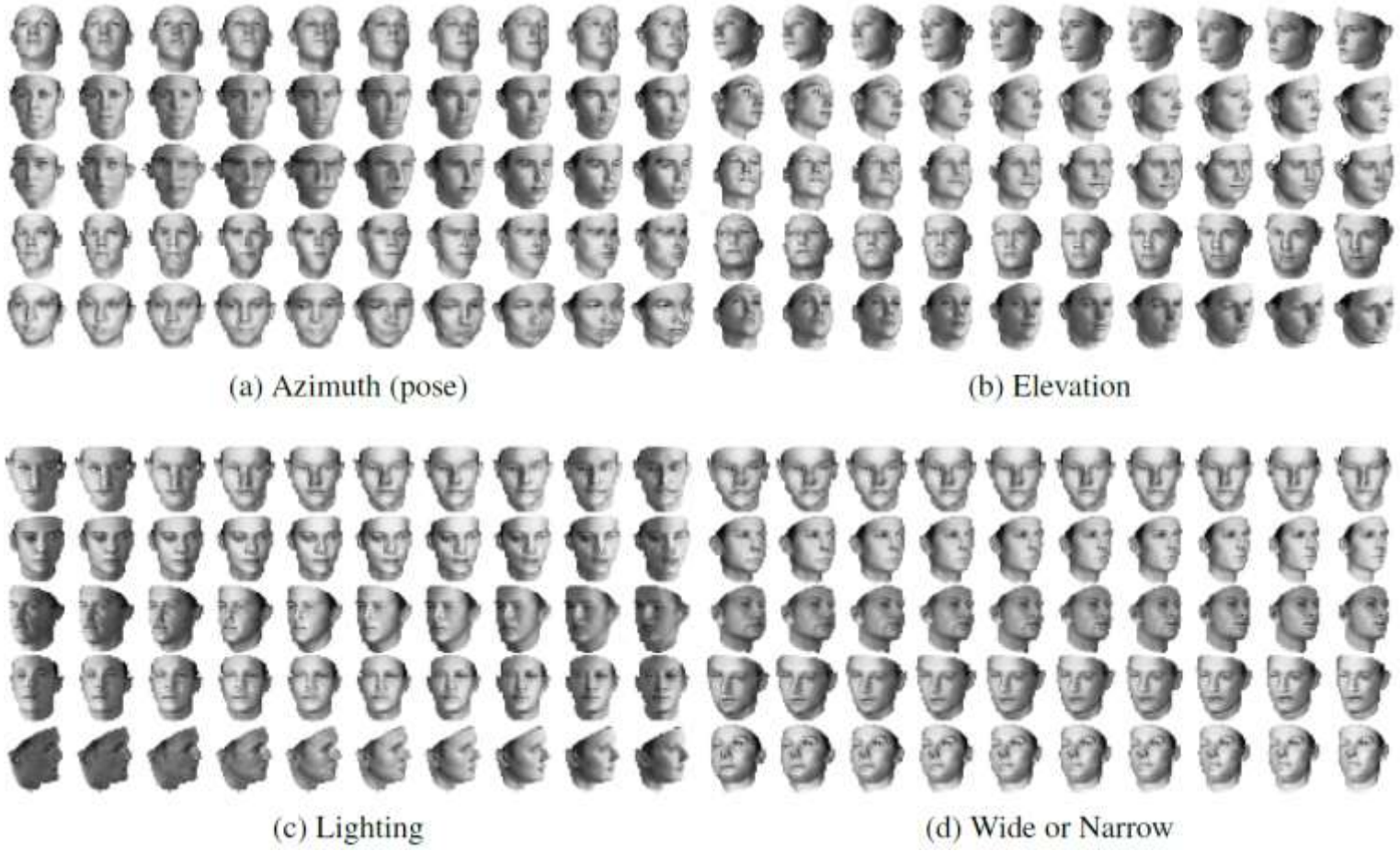


Figure 3 in the original paper

Experiment
– Disentangled Representation –

Dataset: M. Aubry, *et al.*, CVPR, 2014, pp. 3762–3769.

(a) Rotation (b) Width

Figure 4 in the original paper

InfoGAN learned salient features without supervision

Experiment
– Disentangled Representation –

Dataset: Street View House Number

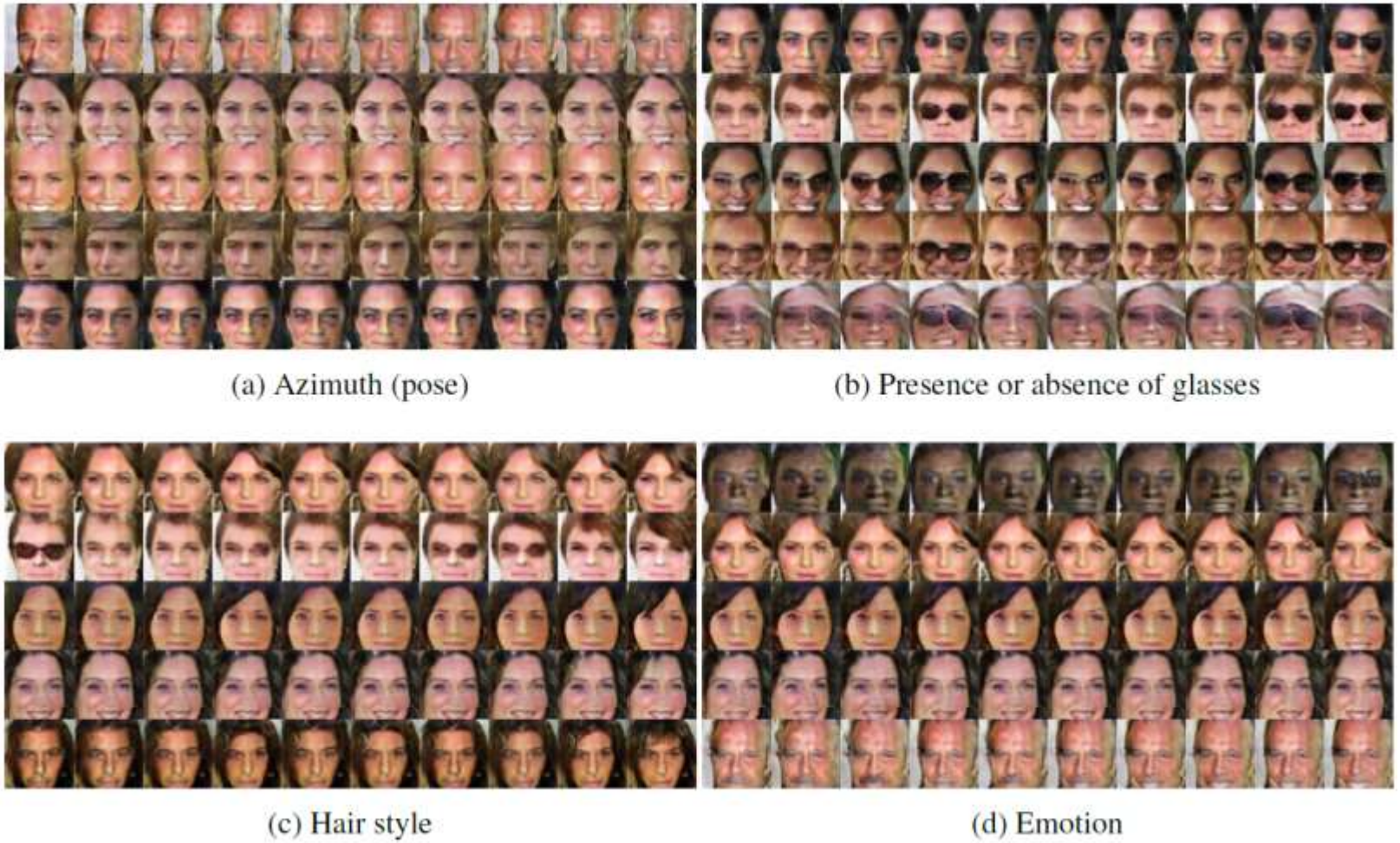
(a) Continuous variation: Lighting

(b) Discrete variation: Plate Context

Figure 5 in the original paper

Experiment
– Disentangled Representation –

Dataset: CelebA



(a) Azimuth (pose)

(b) Presence or absence of glasses

(c) Hair style

(d) Emotion

Figure 6 in the original paper

Future Prospect and Conclusion

- ✓Mutual information maximization can be applied to other methods, e.g. VAE
- ✓Learning hierarchical latent representation
- ✓Improving semi-supervised learning
- ✓High-dimentional data discovery

Goal
Unsupervised learning of disentangled representations

Approach
GANs + Maximizing Mutual Information
between generated images and input codes

Benefit
Interpretable representation obtained
without supervision and substantial additional costs