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

Ilya Sutskever

Research Director of OpenAI
Verified email at openai.com - Homepage
Machine Learning   Neural Networks

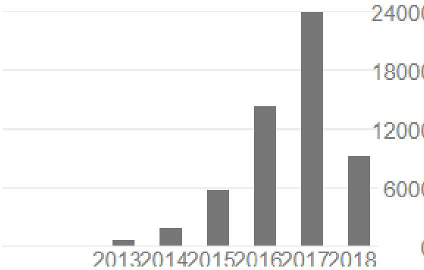
FOLLOW

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
Imagenet classification with deep convolutional neural networks A Krizhevsky, I Sutskever, GE Hinton Advances in neural information processing systems, 1097-1105	23698	2012
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Advances in neural information processing systems, 3111-3119	7649	2013
Dropout: A simple way to prevent neural networks from overfitting N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov The Journal of Machine Learning Research 15 (1), 1929-1958	6225	2014
Sequence to sequence learning with neural networks I Sutskever, O Vinyals, QV Le Advances in neural information processing systems, 3104-3112	3552	2014

Cited by

	All	Since 2013
Citations	56520	56032
h-index	41	41
i10-index	61	60



Co-founder and Research Director of OpenAI.

I spent three wonderful years as a Research Scientist at the Google Brain Team.

Before that, I was a co-founder of DNNresearch.

And before that, I was a postdoc in Stanford with Andrew Ng's group.

And in the beginning, I was a student in the Machine Learning group of Toronto, working with Geoffrey Hinton.

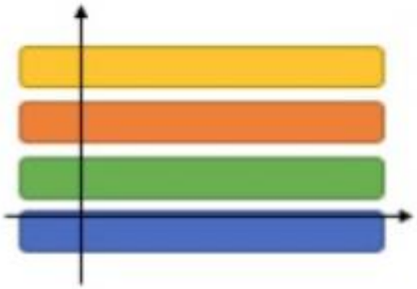
My email address is ilyasu@openai.com.

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

- Unsupervised learning of representation
  (no mechanism to force disentanglement)
  - ✓ Stacked (often denoising) auto-encoder, RBM
  - ✓ Many others, including semi-supervised approach
- Supervised learning of disentangled representation
  - ✓ Bilinear models, multi-view perceptron
  - ✓ VAEs, adversarial auto-encoders
- 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 networks:

- ✓ Generator $x = G(z), z \sim p_z(Z)$

$p_z(Z)$: an arbitrary noise distribution

- ✓ Discriminator $\mathrm{D}(x) \in [0,1]$:

Probability that x is sampled from the real data $p_{data}(X)$
Rather than generated by the generator $\mathrm{G}(Z)$

Optimization problem to solve:

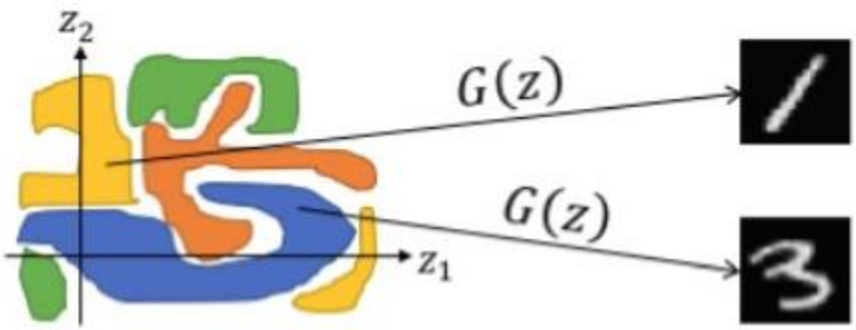
$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim \text{noise}}[\log (1 - D(G(z)))]$$

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

Learning And Mining from DatA
http://lamda.nju.edu.cn

Observation in conventional GANs:

a generated data 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 $\mathrm{I}(\mathrm{C}|\mathrm{X})$ between the latent code C and the generated data X

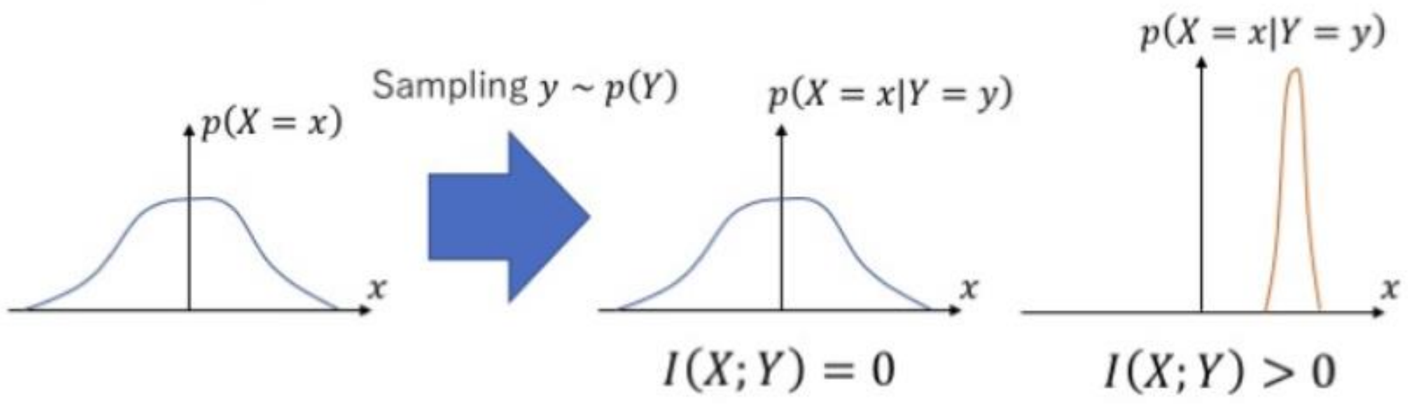
$$\min_G \max_D \{V_{\mathrm{GAN}}(G, D) - \lambda I(C|X = G(Z, C))\}$$

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 $\mathrm{Q}(C|X) = p(C = c|X = x)$
- ✓ Sharing the neural net
  between $\mathrm{Q}(C|X)$ and the discriminator $D(x)$

Variational Maximization of MI

Lower bounding mutual information

$$
\begin{aligned}
I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\
&= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)
\end{aligned}
$$

Variational Maximization of MI

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

$$
\begin{aligned}
L_I(G,Q) &= E_{c\sim P(c),x\sim G(z,c)}[\log Q(c|x)] + H(c) \\
&= E_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\leq I(c;G(z,c))
\end{aligned}
$$

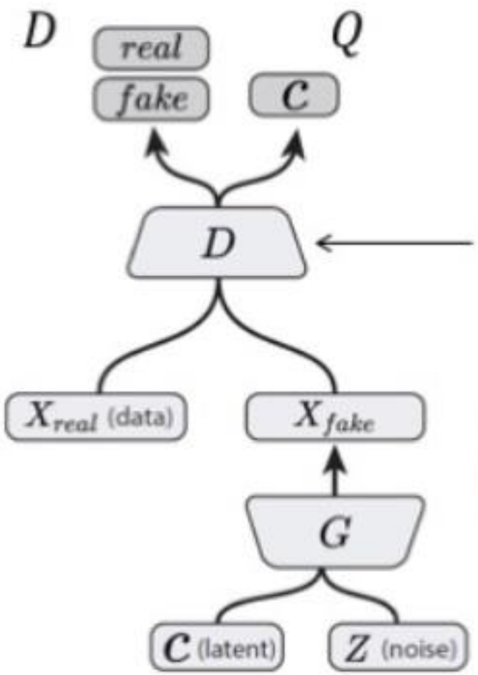
Optimization problem to solve in InfoGAN:

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

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

Learning And Mining from DatA
http://lamda.nju.edu.cn

- InfoGAN on MNIST dataset
- Latent code $c \sim Cat(K = 10, p = 0.1)$
  =10-class categorical code

The lower bound $L_I(G, Q)$ is quickly Maximized to $H(c) = 2.3$

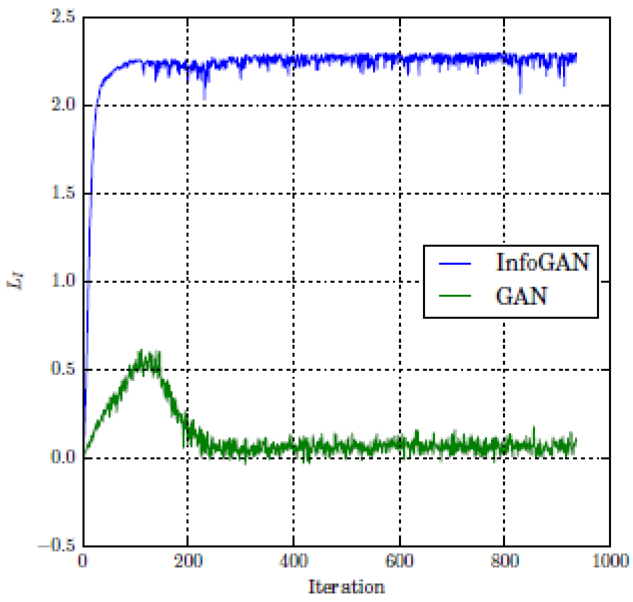
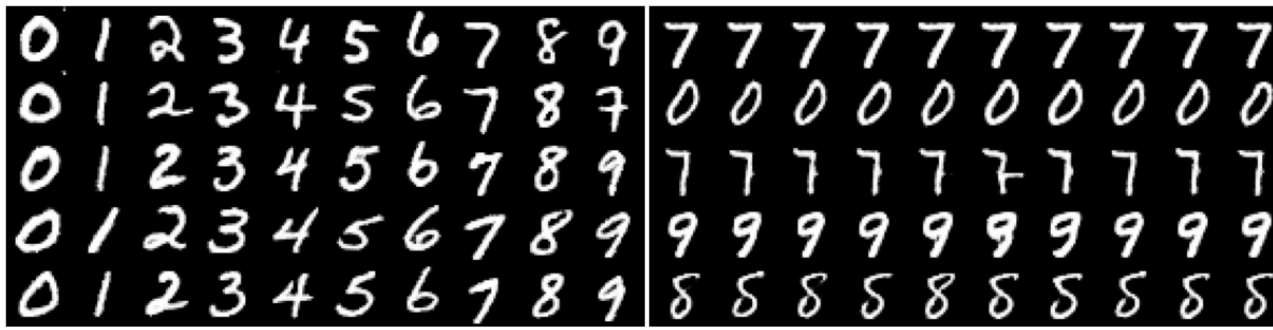


Figure 1: Lower bound L_I over training iterations

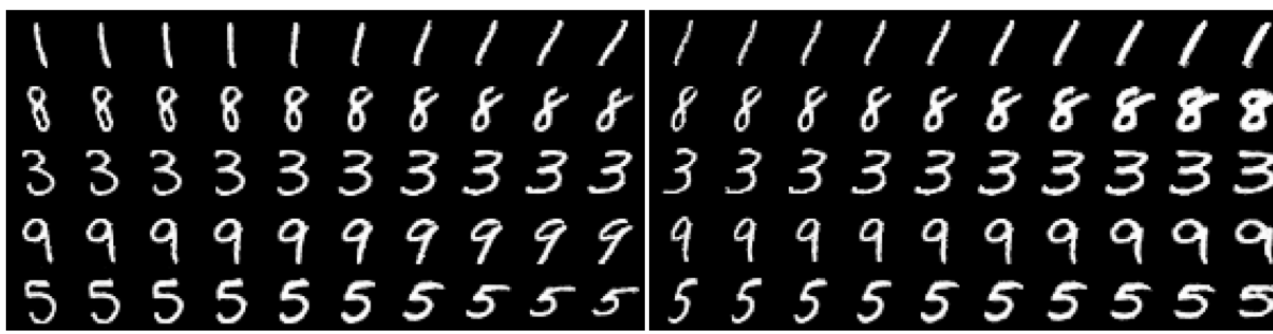
Experiment – Disentangled Representation

- InfoGAN on MNIST dataset
- Latent codes
  - c_1:10-class categorical code $c_1 \sim \mathrm{Cat}(\mathrm{K}=10, \mathrm{p}=0.1)$
  - c_2, c_3:continuous code,

  $c_2, c_3 \sim Unif(-2,2)$



(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)



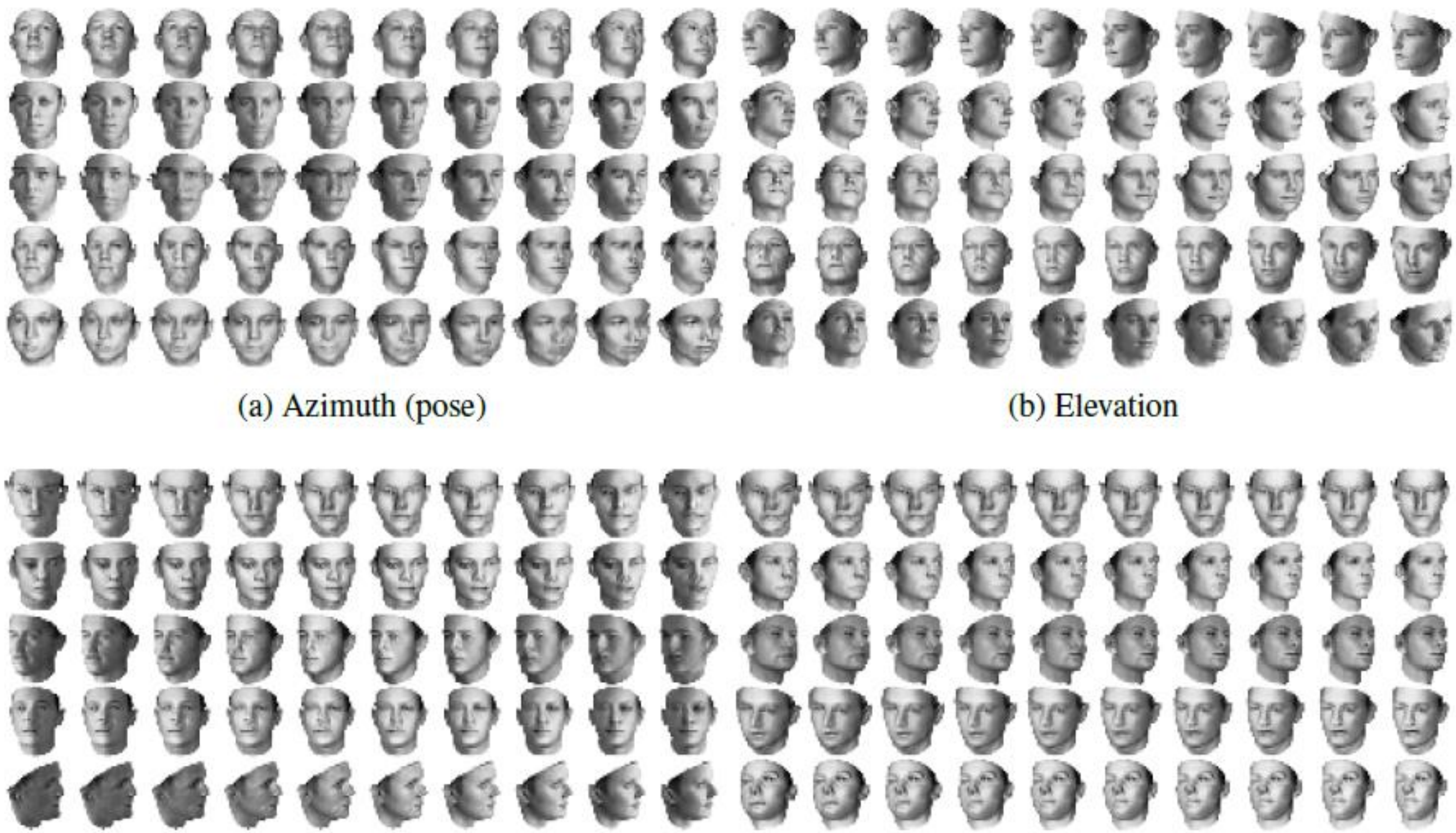
(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation) (d) Varying c_3 from -2 to 2 on InfoGAN (Width)

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

✓ c_2 and c_3 captured the rotation and width respectively

Experiment – Disentangled Representation

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



(a) Azimuth (pose) (b) Elevation

(c) Lighting (d) Wide or Narrow

Experiment – Disentangled Representation

Learning And Mining from DatA
http://lamda.nju.edu.cn

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

(a) Rotation (b) Width

InfoGAN learned salient features without supervision

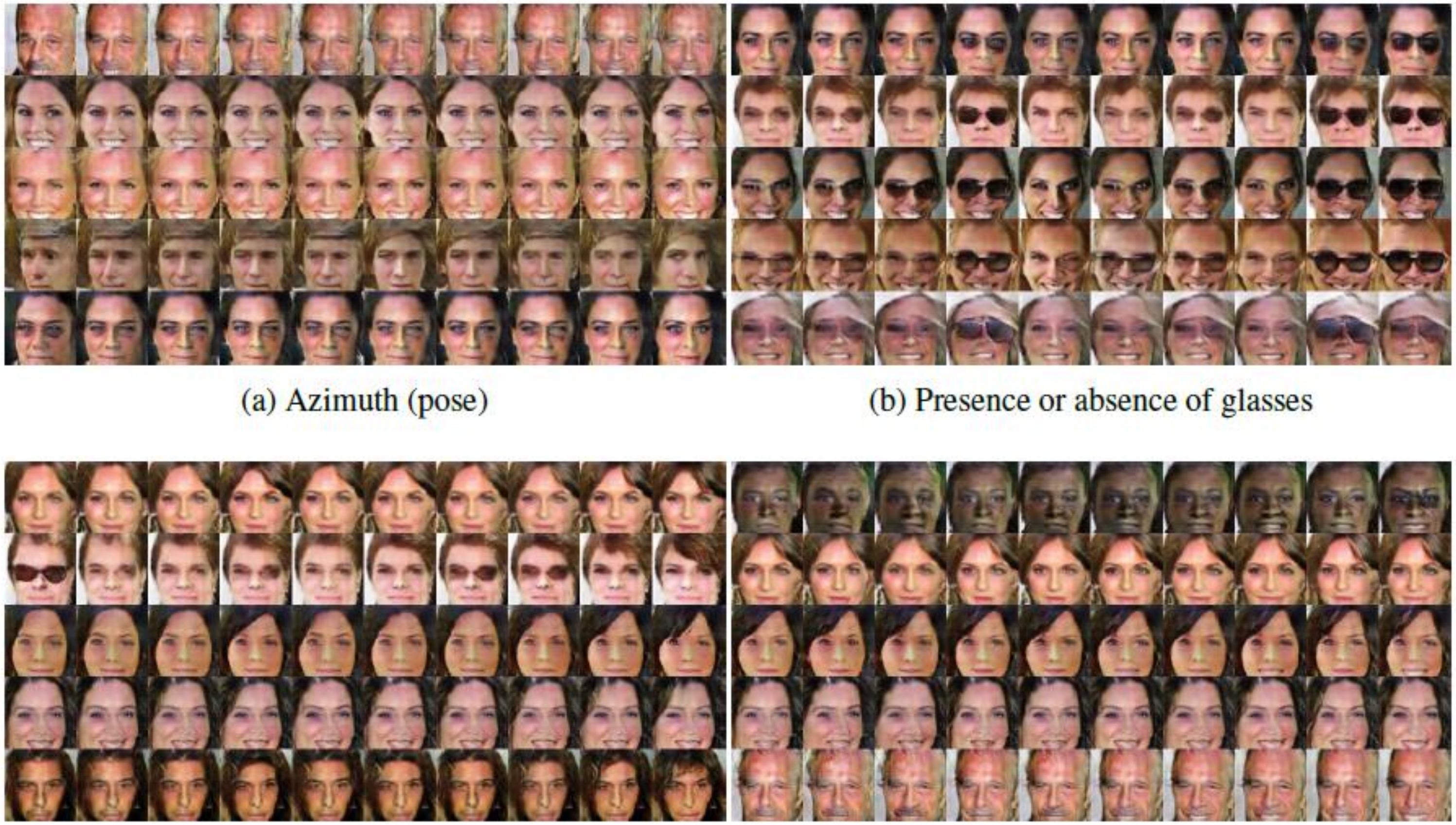
Experiment – Disentangled Representation

Dataset: Street View House Number

(a) Continuous variation: Lighting (b) Discrete variation: Plate Context

Experiment – Disentangled Representation

Dataset: CelebA



(a) Azimuth (pose) (b) Presence or absence of glasses

(c) Hair style (d) Emotion

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

Conclusion

- 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

Future Prospect

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