Deep Generative Models
Lecture 11

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Disentangled representations

Goal

Learning an interpretable factorised representation of the independent data generative factors of the world without supervision.

Informal definition

A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors.

Example

Model trained on a dataset of 3D objects might learn independent latent units sensitive to single independent data generative factors, such as object identity, position, scale, lighting or colour.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017

Generative process

- $p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \text{Sim}(\mathbf{v}, \mathbf{w})$ – true world simulator;
- $\mathbf{v}$ – conditionally independent factors: $p(\mathbf{v}|\mathbf{x}) = \prod_{k=1}^{K} p(v_k|\mathbf{x})$;
- $\mathbf{w}$ – conditionally dependent factors.

Goal

Develop an unsupervised deep generative model

$$p(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x}|\mathbf{v}, \mathbf{w}).$$

- Ensure that the inferred latent factors $q(\mathbf{z}|\mathbf{x})$ capture the factors $\mathbf{v}$ in a disentangled manner.
- The conditionally dependent factors $\mathbf{w}$ can remain entangled in a separate subset of $\mathbf{z}$ that is not used for representing $\mathbf{v}$.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

InfoGAN

GAN objective

$$\min_G \max_D V(G, D)$$

$$V(G, D) = \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z})))$$

Latent vector $\mathbf{z}$ is not imposed to be disentangled.
InfoGAN decomposes input vector:

- $\mathbf{z}$ – incompressible noise;
- $\mathbf{c}$ – structured latent code.

Information-theoretic regularization

$$\max I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

Information in the latent code $\mathbf{c}$ should not be lost in the generation process.

https://arxiv.org/abs/1606.03657

InfoGAN

Objective

$$\min_G \max_D V(G, D) - \lambda I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

Variational Information Maximization

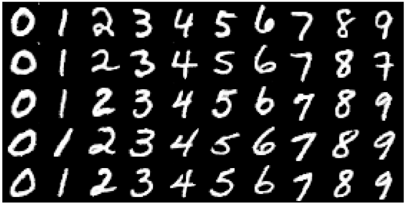
$$\begin{aligned}
I(\mathbf{c}, G(\mathbf{z}, \mathbf{c})) = H(\mathbf{c}) - H(\mathbf{c}|G(\mathbf{z}, \mathbf{c})) = \\
= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})} \left[\mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}|\mathbf{x})} \log p(\mathbf{c}'|\mathbf{x})\right] = \\
= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})} KL(p(\mathbf{c}'|\mathbf{x})||q(\mathbf{z}'|\mathbf{x})) + \\
+ \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) \geq \\
\geq H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) = \\
H(\mathbf{c}) + \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})} \log q(\mathbf{c}'|\mathbf{x})
\end{aligned}$$

---

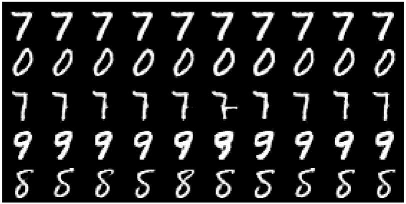
https://arxiv.org/abs/1606.03657

InfoGAN

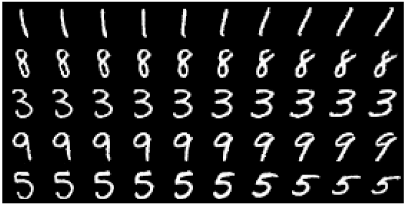
Latent codes on MNIST



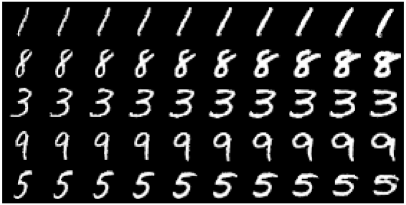
(a) Varying c_1 on InfoGAN (Digit type)



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



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



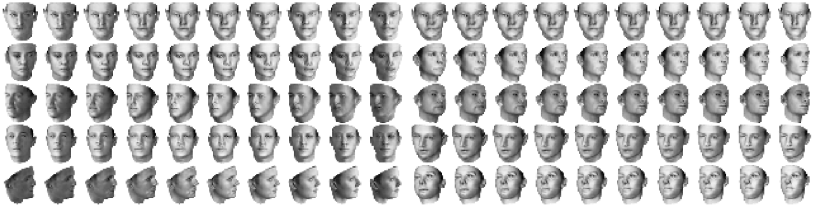
(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

https://arxiv.org/abs/1606.03657

InfoGAN

Latent codes on 3D Faces

(a) Azimuth (pose) (b) Elevation



(c) Lighting (d) Wide or Narrow

https://arxiv.org/abs/1606.03657

β-VAE

Constrained optimization

$$\max_{q,\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}), \quad \text{subject to } KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon.$$

Objective

$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - \beta \cdot KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})).$$

What do we get at $\beta = 1$?

Hypothesis

To learn disentangled representations of the conditionally independent factors $\mathbf{v}$, it is important to set stronger constraint on the latent bottleneck: $\beta > 1$.

Note: It could lead to poorer reconstructions due to the loss of high frequency details when passing through a constrained latent bottleneck.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE

Disentangling metric

Accuracy of classifier $p(y|\mathbf{z}_{\text{diff}})$ with a low VC-dimension in order to ensure that it has no capacity to perform nonlinear disentangling itself.

$$\mathbf{x}_{li} \sim \text{Sim}(\mathbf{v}_{li}, \mathbf{w}_{li}); \quad \mathbf{x}_{lj} \sim \text{Sim}(\mathbf{v}_{lj}, \mathbf{w}_{lj}); \quad y \sim U[1, K].$$

$$\mathbf{v}_{li} \sim p(\mathbf{v}); \quad \mathbf{w}_{li} \sim p(\mathbf{w}); \quad \mathbf{v}_{lj} \sim p(\mathbf{v}) \left([v_{li}]_y = [v_{lj}]_y\right); \quad \mathbf{w}_{lj} \sim p(\mathbf{w}).$$

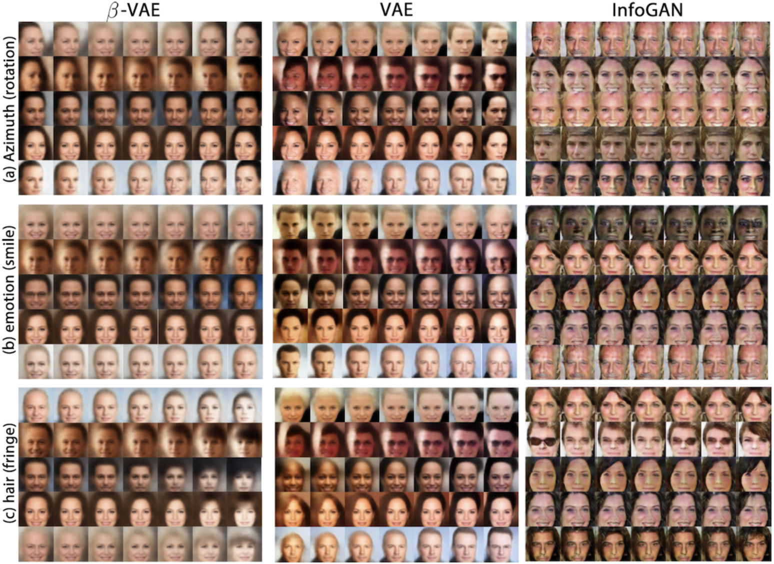
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x})|\sigma^2(\mathbf{x})\right); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

$$\mathbf{z}_{\text{diff}} = \frac{1}{L} \sum_{l=1}^{L} |\mathbf{z}_{li} - \mathbf{z}_{lj}|.$$

---

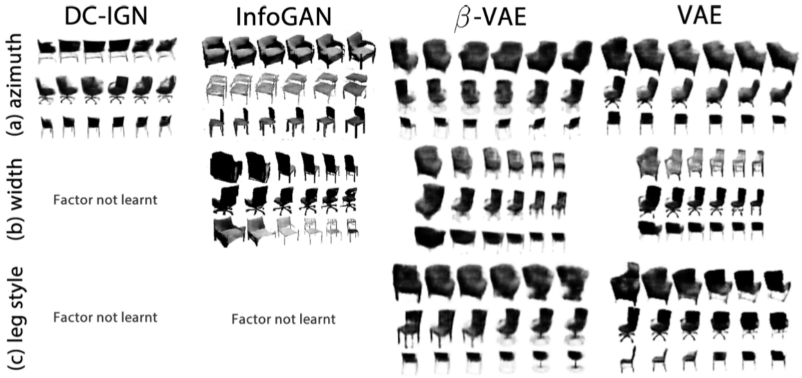
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE



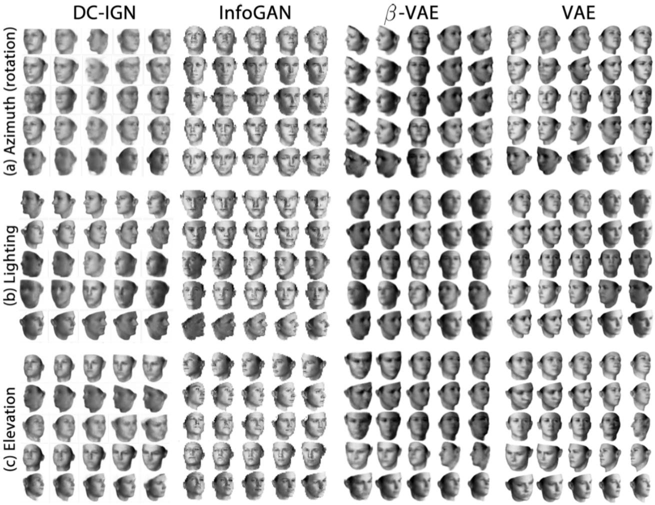
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE



https://openreview.net/references/pdf?id=Sy2fzU9gl

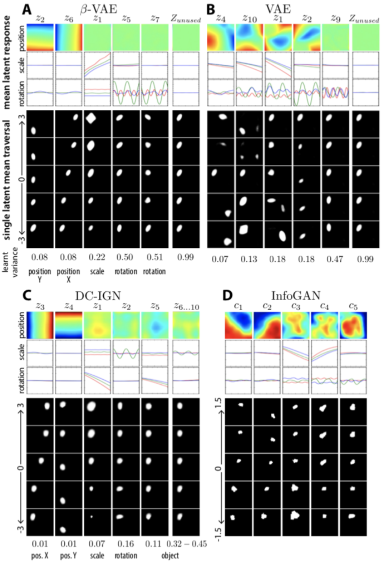
β-VAE



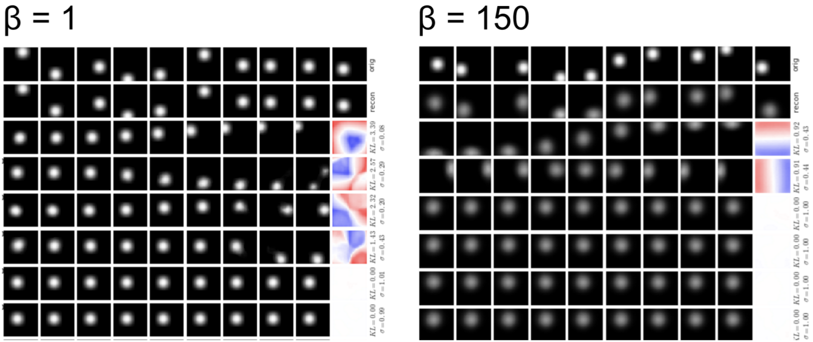
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE

Model	**Disentanglement metric score**
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%\%}$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$



β-VAE

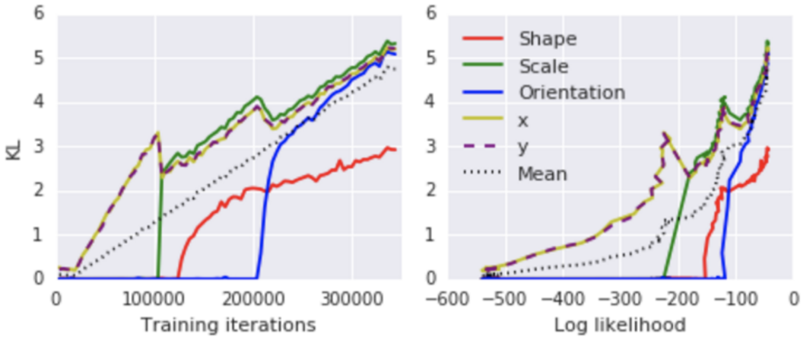


https://arxiv.org/pdf/1804.03599.pdf

β-VAE

Controlled encoding capacity

$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$

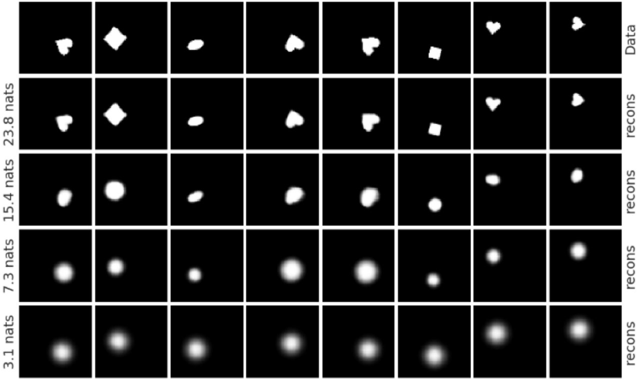


https://arxiv.org/pdf/1804.03599.pdf

β-VAE

Controlled encoding capacity

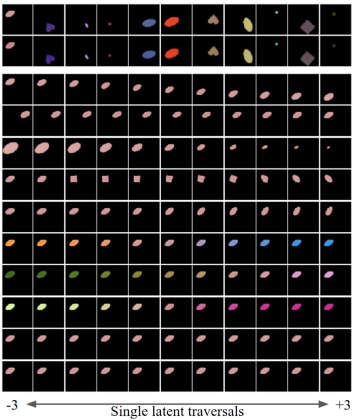
$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



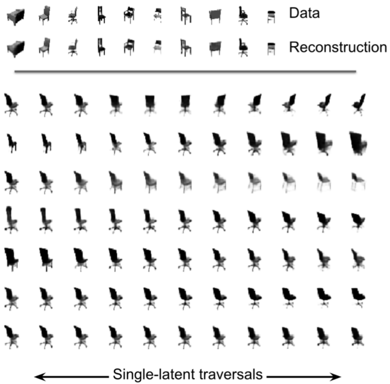
https://arxiv.org/pdf/1804.03599.pdf

β-VAE

(a) Coloured dSprites



(b) 3D Chairs



https://arxiv.org/pdf/1804.03599.pdf

β-VAE

ELBO

$$\mathcal{L}(q, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) - \beta \cdot KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) \right].$$

ELBO surgery

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})}_{\text{Reconstruction loss}} - \beta \cdot \underbrace{\mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}]}_{\text{Mutual info}} - \beta \cdot \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

Minimization of MI

- It is not necessary and not desirable for disentanglement.
- It hurts reconstruction.

---

https://arxiv.org/pdf/1804.03599.pdf

DIP-VAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} q(\mathbf{z}|\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x})\pi(\mathbf{x}) d\mathbf{x} = \prod_{j=1}^{k} q(z_j)$$

Variational inference with disentangled prior encourages inferring factors that are close to being disentangled:

$$KL(q(\mathbf{z}) || \mathbb{E}_{\pi(\mathbf{x})} p(\mathbf{z}|\mathbf{x})) \leq \mathbb{E}_{\pi(\mathbf{x})} KL(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}))$$

DIP-VAE Objective

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z}) || p(\mathbf{z})) =$$

$$= \underbrace{\mathbb{E}_{\pi(\mathbf{x})} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) \right]}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}]}_{\text{Mutual info}} - (1 + \lambda) \cdot \underbrace{KL(q(\mathbf{z}) || p(\mathbf{z}))}_{\text{Marginal KL}}$$

https://arxiv.org/abs/1711.00848

DIP-VAE

DIP-VAE Objective

$$\mathcal{L}(q,\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z}))$$

- $KL(q(\mathbf{z})||p(\mathbf{z}))$ is intractable.
- Let match the moments of $q(\mathbf{z})$ and $p(\mathbf{z})$.

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})}\text{cov}_{q(\mathbf{z}|\mathbf{x})}(\mathbf{z}) + \text{cov}_{\pi(\mathbf{x})}\left(\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\mathbf{z}\right).$$

For most common case $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x}))$:

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})}\boldsymbol{\Sigma}(\mathbf{x}) + \text{cov}_{\pi(\mathbf{x})}\boldsymbol{\mu}(\mathbf{x})$$

DIP-VAE regularizes $\text{cov}_{q(\mathbf{z})}(\mathbf{z})$ to be close to the identity matrix.

https://arxiv.org/abs/1711.00848

DIP-VAE

DIP-VAE Objective

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z}) || p(\mathbf{z}))$$

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} \boldsymbol{\Sigma}(\mathbf{x}) + \text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x})$$

DIP-VAE-I

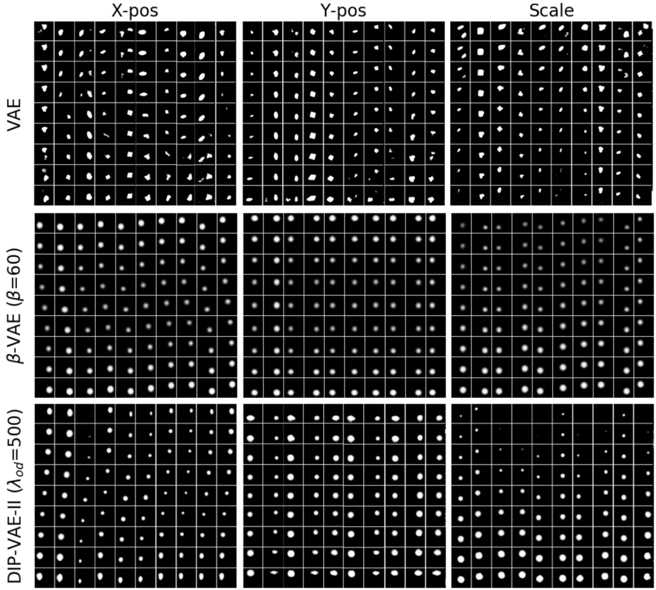
$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \lambda_1 \sum_{i \neq j} \left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ij}^2 - \lambda_2 \sum_i \left(\left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ii} - 1 \right)^2$$

DIP-VAE-II

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \lambda_1 \sum_{i \neq j} \left[\text{cov}_{q(\mathbf{z})}(\mathbf{z}) \right]_{ij}^2 - \lambda_2 \sum_i \left(\left[\text{cov}_{q(\mathbf{z})}(\mathbf{z}) \right]_{ii} - 1 \right)^2$$

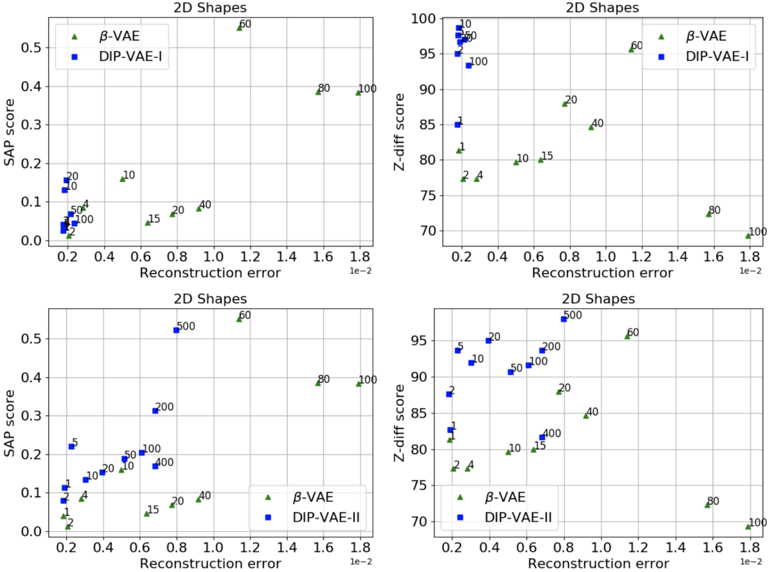
https://arxiv.org/abs/1711.00848

DIP-VAE



https://arxiv.org/abs/1711.00848

DIP-VAE



https://arxiv.org/abs/1711.00848

FactorVAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} q(\mathbf{z}|\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \prod_{j=1}^{k} q(z_j)$$

Total correlation regularizer

$$\min KL(q(\mathbf{z}) || \prod_{j=1}^{k} q(z_j))$$

FactorVAE objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \gamma \cdot KL(q(\mathbf{z}) || \prod_{j=1}^{k} q(z_j))$$

- The last term is intractable.
- FactorVAE uses density ratio trick for estimation.

---

https://arxiv.org/abs/1802.05983

FactorVAE

Consider two distributions $q_1(\mathbf{x})$, $q_2(\mathbf{x})$ and probilistic model

$$p(\mathbf{x}|y) = \begin{cases} q_1(\mathbf{x}), & \text{if } y = 1, \\ q_2(\mathbf{x}), & \text{if } y = 0, \end{cases} \qquad y \sim \text{Bern}(0.5).$$

Density ratio trick

$$\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Bigg/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} =$$

$$= \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \frac{p(y=1|\mathbf{x})}{1 - p(y=1|\mathbf{x})} = \frac{D(\mathbf{x})}{1 - D(\mathbf{x})}$$

Here $D(\mathbf{x})$ could be treated as a discriminator model, which outputs the probability that $\mathbf{x}$ is a sample from $q_1(\mathbf{x})$ rather than from $q_2(\mathbf{x})$.

https://arxiv.org/abs/1802.05983

FactorVAE

FactorVAE objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta},\boldsymbol{\phi}) - \gamma \cdot KL(q(\mathbf{z})||\prod_{j=1}^{k} q(z_j))$$

Total correlation regularizer

$$KL(q(\mathbf{z})||\prod_{j=1}^{k} q(z_j)) = KL(q(\mathbf{z})||\bar{q}(\mathbf{z})) =$$

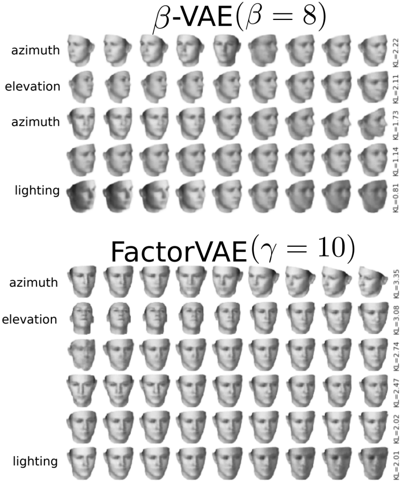
$$= \mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{\bar{q}(\mathbf{z})} \approx \mathbb{E}_{q(\mathbf{z})} \log \frac{D(\mathbf{z})}{1 - D(\mathbf{z})}$$

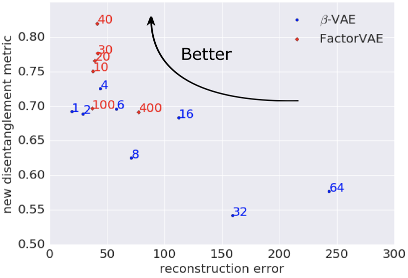
VAE and GAN are trained simultaneously.

---

https://arxiv.org/abs/1802.05983

FactorVAE





https://arxiv.org/abs/1802.05983

Challenging Disentanglement Assumptions

Whether unsupervised disentanglement learning is even possible for arbitrary generative models?

Theorem

For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that

- $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled);
- and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).

Theorem claims that unsupervised disentanglement learning is impossible for arbitrary generative models with a factorized prior.

---

https://arxiv.org/abs/1811.12359

Challenging Disentanglement Assumptions

Assume we have $p(\mathbf{z})$ and some $p(\mathbf{x}|\mathbf{z})$ defining a generative model. Consider any unsupervised disentanglement method and assume that it finds a representation that is perfectly disentangled with respect to $\mathbf{z}$ in the generative model.

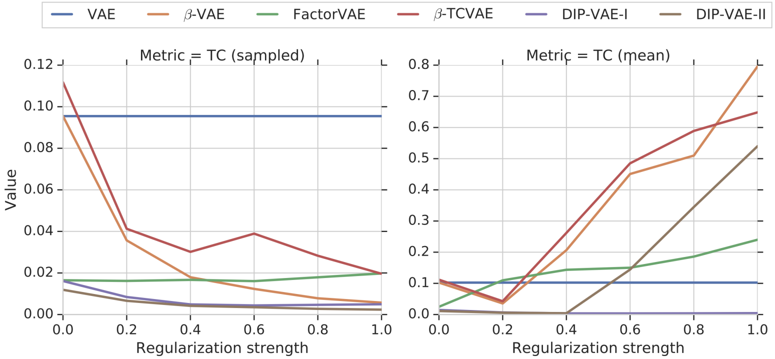
- Theorem claims that $\exists\ \hat{\mathbf{z}} = f(\mathbf{z})$ where $\bar{\mathbf{z}}$ is completely entangled with respect to $\mathbf{z}$.
- Since the (unsupervised) disentanglement method only has access to observations $\mathbf{x}$, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

https://arxiv.org/abs/1811.12359

Challenging Disentanglement Assumptions

- **Training:** Factorizing **samples** from aggregated posterior $q(\mathbf{z}) = \prod_{i=1}^{d} q(z_i)$.
- **Inference:** Use the **mean** vector (usually mean of Gaussian encoder) as the representation.

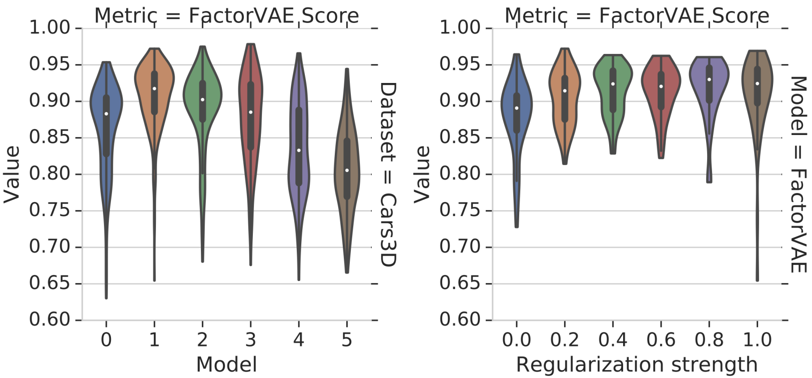


---

https://arxiv.org/abs/1811.12359

Challenging Disentanglement Assumptions

Importance of different models and hyperparameters for disentanglement



https://arxiv.org/abs/1811.12359

Challenging Disentanglement Assumptions

Agreement of different disentanglement metrics

Dataset = Noisy-dSprites

	(A)	(B)	(C)	(D)	(E)	(F)
BetaVAE Score (A)	100	80	44	41	46	37
FactorVAE Score (B)	80	100	49	52	25	38
MIG (C)	44	49	100	76	6	42
DCI Disentanglement (D)	41	52	76	100	-8	38
Modularity (E)	46	25	6	-8	100	13
SAP (F)	37	38	42	38	13	100

https://arxiv.org/abs/1811.12359

References

- **InfoGAN:** Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets
  https://arxiv.org/abs/1606.03657
  Summary: An information-theoretic extension to the GANs that disentangles representations in an unsupervised manner. InfoGAN maximizes the MI between a small subset of the latent variables and the observation. Lower bound for MI objective is derived that can be optimized efficiently.
- **beta-VAE:** Learning Basic Visual Concepts with a Constrained Variational Framework
  https://openreview.net/references/pdf?id=Sy2fzU9gl
  Summary: Modifications of VAE objective. The task is represented as constrained optimization. Increasing the weight of KL divergence term in ELBO allows to disentangle latent space factors and makes model more interpretable. The assessment of disentanglement is provided by constructing the classifier.
- Understanding disentangling in β-**VAE**
  https://arxiv.org/pdf/1804.03599.pdf
  Summary: Consider beta-VAE from the position of the rate-distortion theory (information bottleneck). Propose the modified ELBO with controlled latent capacity.
- **DIP-VAE:** Variational Inference of Disentangled Latent Concepts from Unlabeled Observations
  https://arxiv.org/abs/1711.00848
  Summary: Introduce a regularizer on the expectation of the approximate posterior over observed data that encourages the disentanglement. Penalize the mismatch between the aggregated posterior and a factorized prior. Comparison with beta-VAE.
- **FactorVAE:** Disentangling by Factorising
  https://arxiv.org/abs/1802.05983
  Summary: Penalizes the total correlation with density-ratio estimation. Comparison with beta-VAE. Does not degrate the reconstructions.
- **FactorVAE:** Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations
  https://arxiv.org/abs/1811.12359
  Summary: There infinite family of equivalent generative models with entangled factors. Uncorrelated samples vs Correlated means. Agreement of all disentanglement metrics. Hyperparameter choice is crucial.