Isolating Sources of Disentanglement in VAEs

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, David Duvenaud

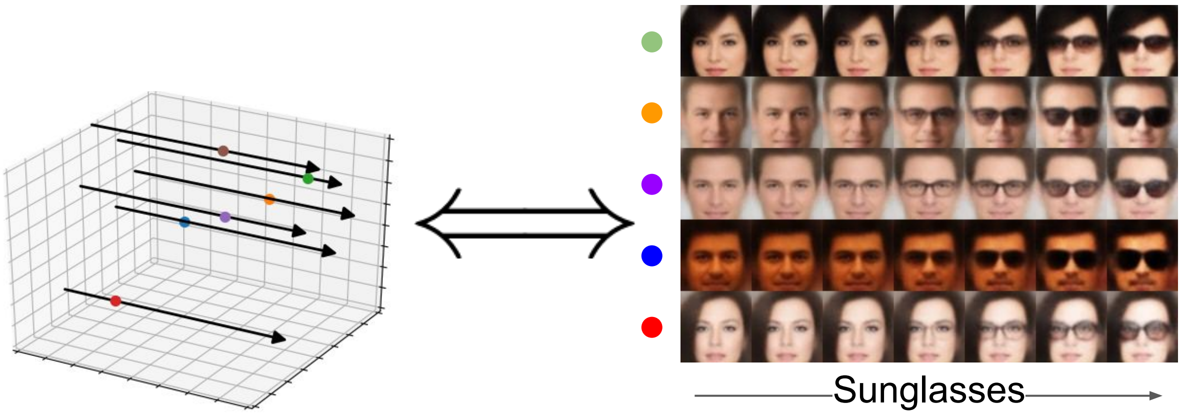
University of Toronto, Vector Institute

Presenter: Ricky

Disentanglement = Independence + Semantics

<u>Axis-aligned</u> Traversal in the Representation:

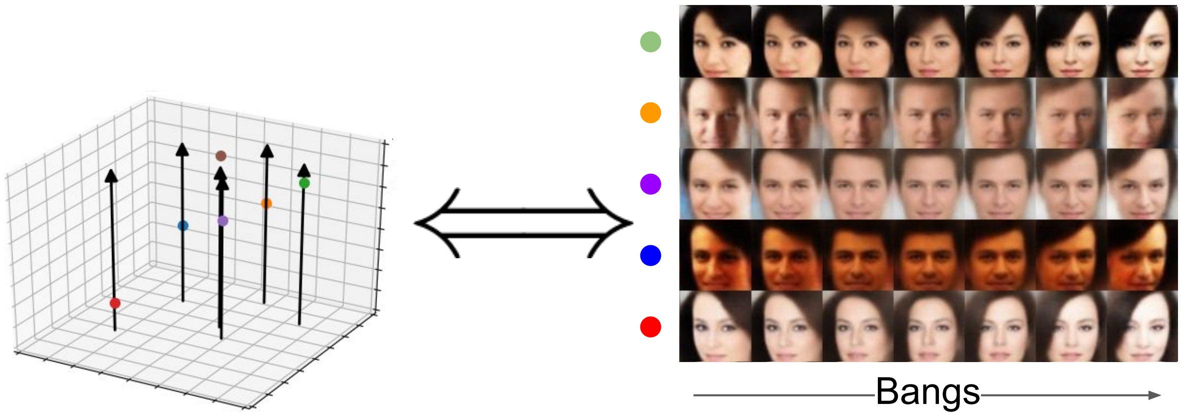
<u>Global</u> Interpretability in Data Space:



Disentanglement = Independence + Semantics

<u>Axis-aligned</u> Traversal in the Representation:

<u>Global</u> Interpretability in Data Space:



Disentanglement = Independence + Semantics

<u>Motivations:</u>

- Independent Components

<u>Downstream Tasks:</u>

- Interpretable Decision Making

Disentanglement = Independence + Semantics

<u>Motivations:</u>

- Independent Components
- Controllable Sample Generation

<u>Downstream Tasks:</u>

- Interpretable Decision Making
- Semantic Inpainting

Disentanglement = Independence + Semantics

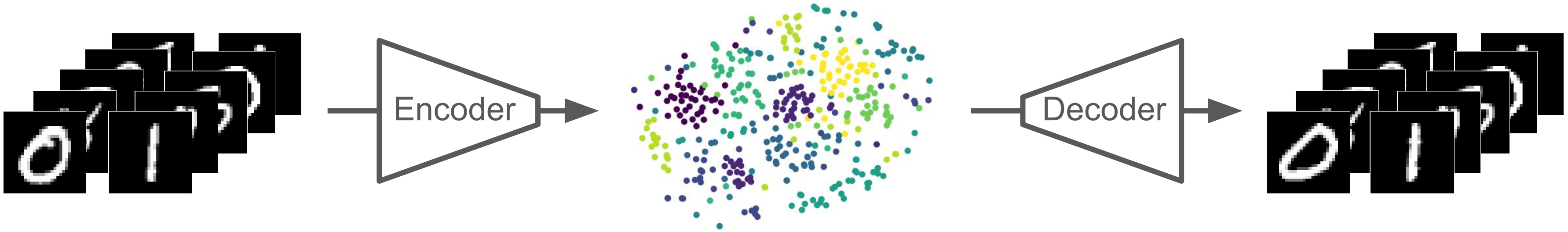
Motivations:

- Independent Components
- Controllable Sample Generation
- Generalization and Robustness

Downstream Tasks:

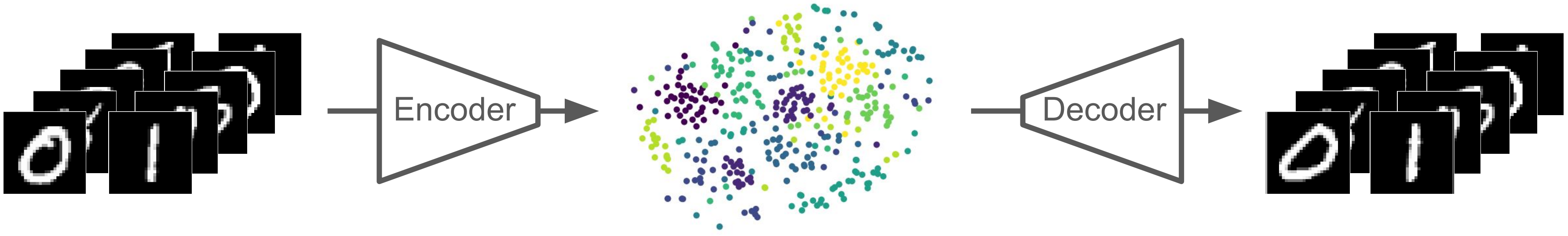
- Interpretable Decision Making
- Semantic Inpainting
- Controlled Transfer

Regularization in VAEs



$x_n \sim dataset$

Regularization in VAEs

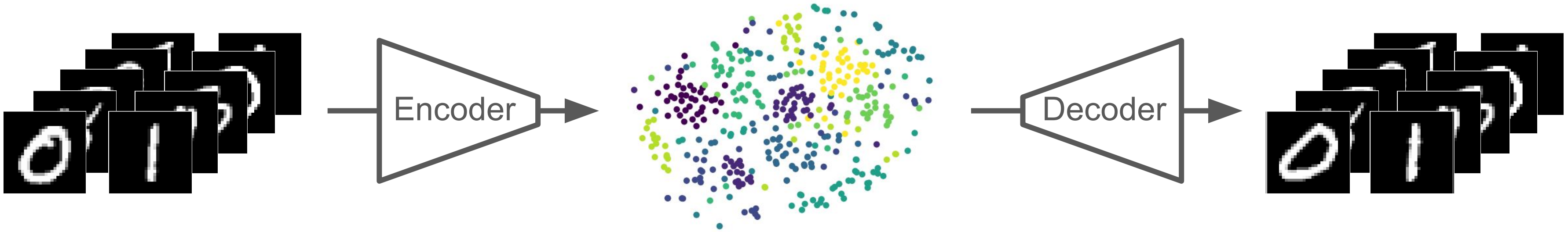


$x_n \sim dataset$ $\qquad$ $z \sim q(z|x_n)$ $\qquad$ $x \sim p(x|z)$

Evidence Lower Bound (ELBO):

$$\log p(x_n) \geq \underbrace{\mathbb{E}_{q(z|x_n)}\left[\frac{p(x_n, z)}{q(z|x_n)}\right]}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q(z|x_n)}[\log p(x_n|z)]}_{\text{reconstruction}} - \beta \underbrace{\mathrm{KL}[q(z|x_n)||p(z|x_n)]}_{\text{regularization}}$$

Regularization in VAEs



$x_n \sim dataset$ $\qquad$ $z \sim q(z|x_n)$ $\qquad$ $x \sim p(x|z)$

Evidence Lower Bound (ELBO):

$$\log p(x_n) \geq \underbrace{\mathbb{E}_{q(z|x_n)}\left[\frac{p(x_n, z)}{q(z|x_n)}\right]}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q(z|x_n)}[\log p(x_n|z)]}_{\text{reconstruction}} - \beta \underbrace{\mathrm{KL}[q(z|x_n)||p(z|x_n)]}_{\text{regularization}}$$

How does this affect disentanglement?

Different Forms of Regularization in the ELBO

Define joint distribution $q(z,n) = p(n)q(z|n)$ where $p(n) = \dfrac{1}{N}$.

Different Forms of Regularization in the ELBO

Define joint distribution $q(z,n) = p(n)q(z|n)$ where $p(n) = \frac{1}{N}$.

The marginal distribution $q(z) = \mathbb{E}_{p(n)}[q(z|n)]$.

Isolating Different Forms of Regularization

Define joint distribution $q(z, n) = p(n)q(z|n)$ where $p(n) = \dfrac{1}{N}$.

The marginal distribution $q(z) = \mathbb{E}_{p(n)}[q(z|n)]$.

ELBO TC-DECOMPOSITION

$$\mathbb{E}_{p(n)}\Big[\underbrace{D_{\mathrm{KL}}\big[q(z|n)\big|\big|p(z)\big]}_{\text{Regularization}}\Big] = \underbrace{D_{\mathrm{KL}}\big[q(z,n)\big|\big|q(z)p(n)\big]}_{\text{(1) Index-code Mutual Info.}} + \underbrace{D_{\mathrm{KL}}\big[q(z)\big|\big|\textstyle\prod_j q(z_j)\big]}_{\text{(2) Total Correlation}} + \underbrace{\textstyle\sum_j D_{\mathrm{KL}}\big[q(z_j)\big|\big|p(z_j)\big]}_{\text{(3) Dim-wise KL}}$$

Index-code MI: Hoffman & Johnson “ELBO Surgery.” (2016)

Isolating Different Forms of Regularization

Define joint distribution $q(z,n) = p(n)q(z|n)$ where $p(n) = \dfrac{1}{N}$.

The marginal distribution $q(z) = \mathbb{E}_{p(n)}[q(z|n)]$.

ELBO TC-DECOMPOSITION

$$\mathbb{E}_{p(n)}\Big[\underbrace{D_{\mathrm{KL}}\big[q(z|n)\big|\big|p(z)\big]}_{\text{Regularization}}\Big] = \underbrace{D_{\mathrm{KL}}\big[q(z,n)\big|\big|q(z)p(n)\big]}_{\text{(1) Index-code Mutual Info.}} + \underbrace{D_{\mathrm{KL}}\big[q(z)\big|\big|\textstyle\prod_j q(z_j)\big]}_{\text{(2) Total Correlation}} + \underbrace{\textstyle\sum_j D_{\mathrm{KL}}\big[q(z_j)\big|\big|p(z_j)\big]}_{\text{(3) Dim-wise KL}}$$

The beta-VAE (*Higgins et al., 2017*) penalizes all three terms evenly.

Isolating Different Forms of Regularization

Define joint distribution $q(z, n) = p(n)q(z|n)$ where $p(n) = \frac{1}{N}$.

The marginal distribution $q(z) = \mathbb{E}_{p(n)}[q(z|n)]$.

ELBO TC-DECOMPOSITION

$$\mathbb{E}_{p(n)}\Big[\underbrace{D_{\mathrm{KL}}\big[q(z|n)||p(z)\big]}_{\text{Regularization}}\Big] = \underbrace{D_{\mathrm{KL}}\big[q(z,n)||q(z)p(n)\big]}_{\text{(1) Index-code Mutual Info.}} + \underbrace{D_{\mathrm{KL}}\big[q(z)||\textstyle\prod_j q(z_j)\big]}_{\text{(2) Total Correlation}} + \underbrace{\textstyle\sum_j D_{\mathrm{KL}}\big[q(z_j)||p(z_j)\big]}_{\text{(3) Dim-wise KL}}$$

The beta-VAE (*Higgins et al., 2017*) penalizes all three terms evenly.

We should amplify the independence regularization in isolation!

Stochastic Estimation of $\log q(\cdot)$

$\mathrm{q}(\cdot)$ is a *mixture distribution*.

- Evaluating $\mathrm{q}(\cdot)$ requires the full dataset.
- Stochastic estimate $\mathrm{q}(\cdot)$ based on a minibatch?
    - Randomly chosen n will give q(z|n) close to zero.

$$q(z) = \mathbb{E}_{p(n)}[q(z|n)]$$
$$= \frac{1}{N}\sum_{n=1}^{N} q(z|n)$$

Mixture of q(z|n)

Stochastic Estimation of $\log q(\cdot)$

$q(\cdot)$ is a *mixture distribution*.

- Evaluating $q(\cdot)$ requires the full dataset.
- Stochastic estimate $q(\cdot)$ based on a minibatch?
  - Randomly chosen n will give q(z|n) close to zero.
- We can reuse the **same minibatch**.

$$\mathbb{E}_q(z)[\log q(z)] \approx \frac{1}{M}\sum_{i=1}^{M}\left[\log \frac{1}{NM}\sum_{j=1}^{M} q(z(n_i)|n_j)\right]$$

- Better minibatch estimators since our work:
  - Esmaeili et al. "*Structured Disentangled Representations.*"

$$q(z) = \mathbb{E}_{p(n)}[q(z|n)]$$
$$= \frac{1}{N}\sum_{n=1}^{N} q(z|n)$$

Mixture of q(z|n)

Isolating Different Forms of Regularization & TCVAE

$$\text{Modified ELBO} = \underbrace{\mathbb{E}_{q(z,n)}[\log p(n|z)]}_{\text{reconstruction}} - \boldsymbol{\alpha}\underbrace{\mathrm{KL}[q(z|n)||q(z)]}_{\text{index-code MI}} - \boldsymbol{\beta}\underbrace{\mathrm{KL}[q(z)||\prod_j q(z_j)]}_{\text{Total Correlation}} - \boldsymbol{\gamma}\underbrace{\sum_{j=1}\mathrm{KL}[q(z_j)||p(z_j)]}_{\text{Dim-wise KL}}$$

Pseudo-code

```
def ELBO(x):
  z = encode(x)
  x_recon = decode(z)
  logpx_z = logp_likelihood(x_recon)
  logqz_x = logp_approximate_posterior(z, x)
  logpz = logp_prior(z)
  elbo = logpx_z + logp - logqz_x
```

Isolating Different Forms of Regularization & TCVAE

$$\text{Modified ELBO} = \underbrace{\mathbb{E}_{q(z,n)}[\log p(n|z)]}_{\text{reconstruction}} - \boldsymbol{\alpha}\underbrace{\mathrm{KL}[q(z|n)||q(z)]}_{\text{index-code MI}} - \boldsymbol{\beta}\underbrace{\mathrm{KL}[q(z)||\prod_j q(z_j)]}_{\text{Total Correlation}} - \boldsymbol{\gamma}\underbrace{\sum_{j=1}\mathrm{KL}[q(z_j)||p(z_j)]}_{\text{Dim-wise KL}}$$

Pseudo-code

```
def ELBO_decomp(x):
  z = encode(x)
  x_recon = decode(z)
  logpx_z = logp_likelihood(x_recon)
  logqz_x = logp_approximate_posterior(z, x)
  logpz = logp_prior(z)
  logqz = logsumexp(logqz_x.sum(axis=2), axis=1) - log(M * N)
  logqz_factorized = (logsumexp(logqz_x, axis=1) - log(M * N)).sum(axis=2)
  elbo = logpx_z - (logqz_x - logqz) - (logqz - logqz_factorized) - (logqz_factorized - logpz)
```

Isolating Different Forms of Regularization & TCVAE

$$\text{Modified ELBO} = \underbrace{\mathbb{E}_{q(z,n)}[\log p(n|z)]}_{\text{reconstruction}} - \boldsymbol{\alpha}\underbrace{\mathrm{KL}[q(z|n)||q(z)]}_{\text{index-code MI}} - \boldsymbol{\beta}\underbrace{\mathrm{KL}[q(z)||\prod_j q(z_j)]}_{\text{Total Correlation}} - \boldsymbol{\gamma}\underbrace{\sum_{j=1}\mathrm{KL}[q(z_j)||p(z_j)]}_{\text{Dim-wise KL}}$$

Pseudo-code

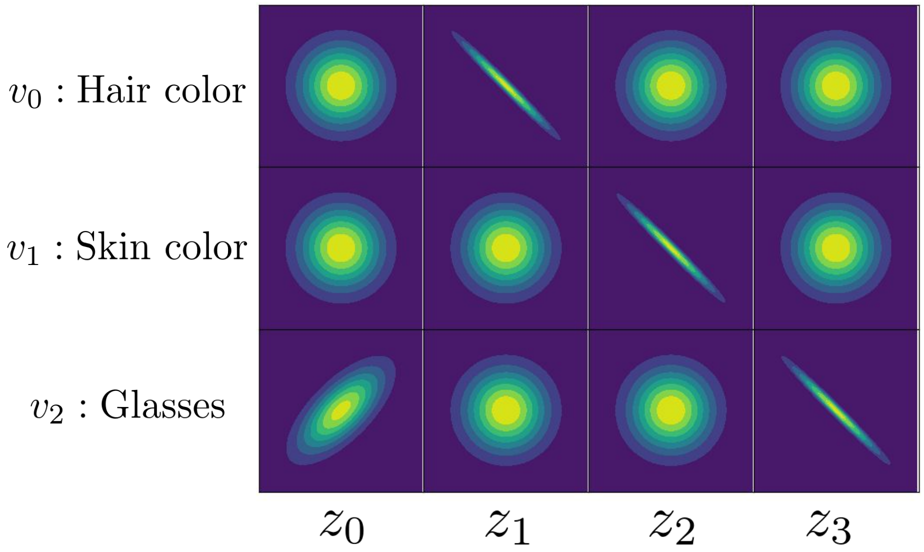
```
def ELBO_decomp(x):
  z = encode(x)
  x_recon = decode(z)
  logpx_z = logp_likelihood(x_recon)
  logqz_x = logp_approximate_posterior(z, x)
  logpz = logp_prior(z)
  logqz = logsumexp(logqz_x.sum(axis=2), axis=1) - log(M * N)
  logqz_factorized = (logsumexp(logqz_x, axis=1) - log(M * N)).sum(axis=2)
  elbo = logpx_z - (logqz_x - logqz) - (logqz - logqz_factorized) - (logqz_factorized - logpz)
```

The case $\alpha = \gamma = 1$ results in an equivalent objective to FactorVAE (*Kim & Mnih. 2017.*), though they use a discriminator to estimate KL.

Evaluating Disentanglement

Suppose we have some ground truth factors $\{v_k\}_{k=1}^{K}$.

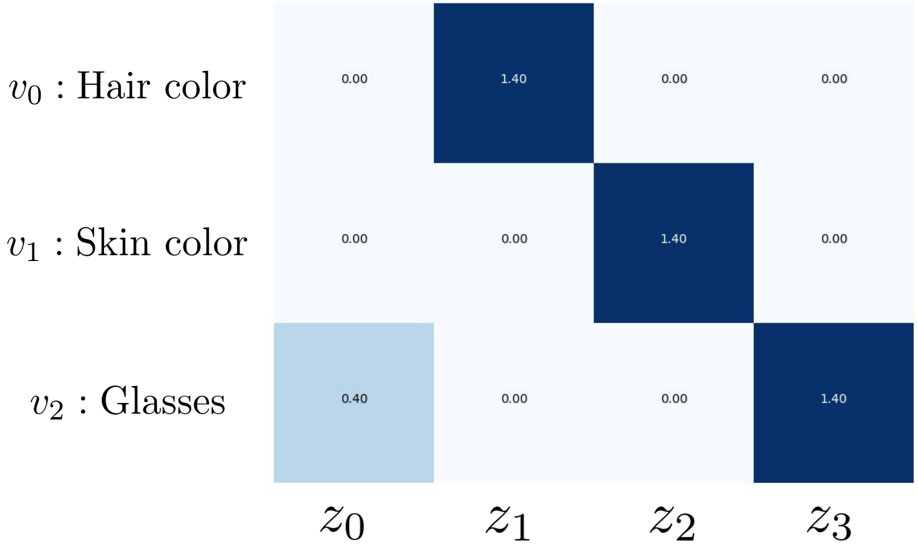
We can define a joint distribution $q(z_j, v_k) = \sum_{n=1}^{N} p(v_k)p(n|v_k)q(z_j|n)$.



Evaluating Disentanglement

Suppose we have some ground truth factors $\{v_k\}_{k=1}^{K}$.

We can define a joint distribution $q(z_j, v_k) = \sum_{n=1}^{N} p(v_k)p(n|v_k)q(z_j|n)$.

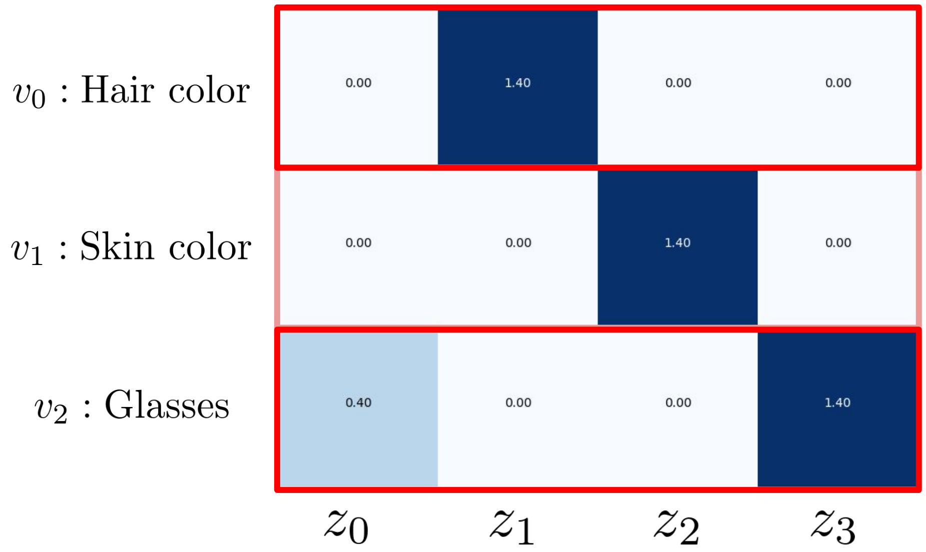


Ideally, one axis for each factor.

Evaluating Disentanglement

Suppose we have some ground truth factors $\{v_k\}_{k=1}^{K}$.

We can define a joint distribution $q(z_j, v_k) = \sum_{n=1}^{N} p(v_k) p(n|v_k) q(z_j|n)$.



One Factor == One Dimension

Mutual Information Gap (MIG):

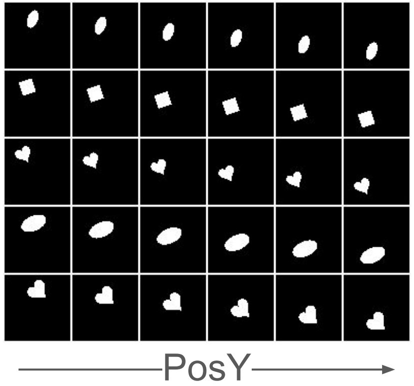
$$\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H(v_k)} \left(I(z_{j^{(k)}}; v_k) - \max_{j \neq j^{(k)}} I(z_j; v_k) \right)$$

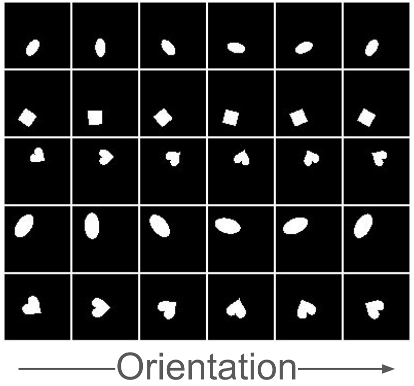
where $\quad j^{(k)} = \arg\max_j I(z_j; v_k)$

Datasets Used for Quantitative Experiments

dSprites:

- Scale
- Orientation
- PosX
- PosY

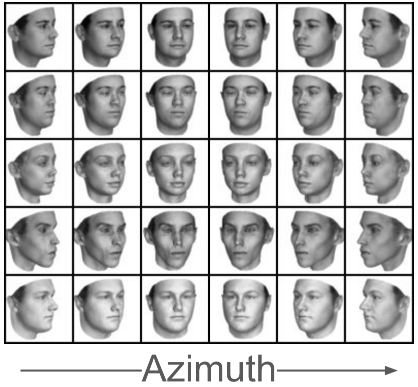


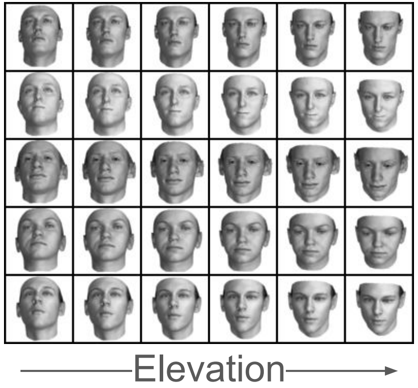


Matthey et al. (2017)

3D Faces:

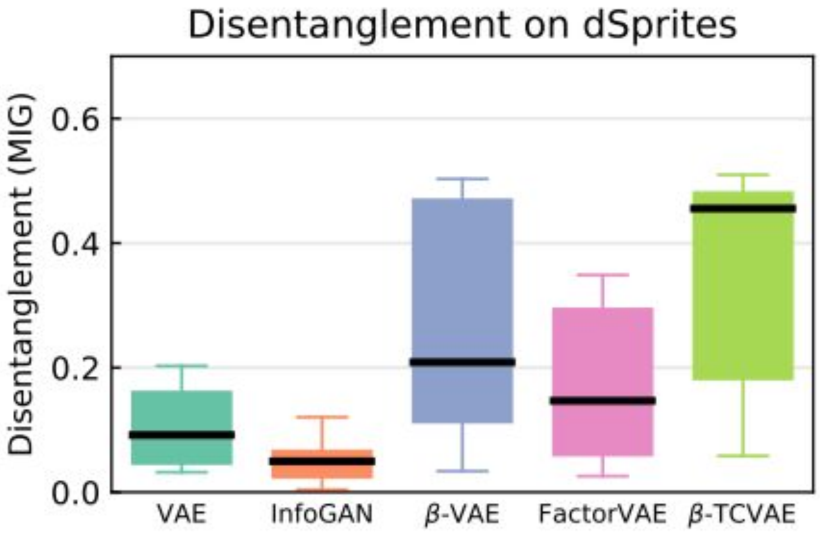
- Azimuth
- Elevation
- Lighting



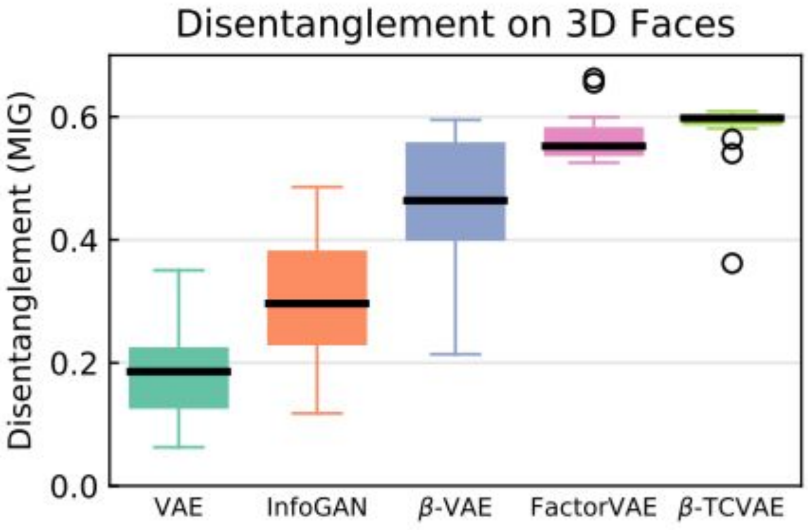


Paysan et al. (2009)

Penalizing Only Total Correlation Works Better



(Fully connected)

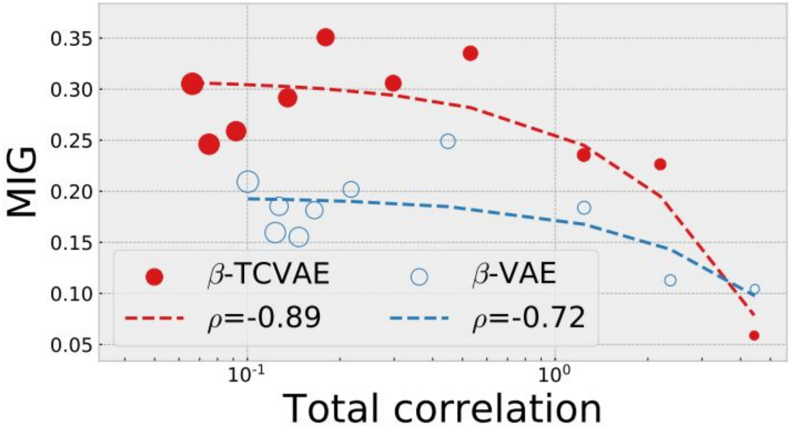


(Convolutional)

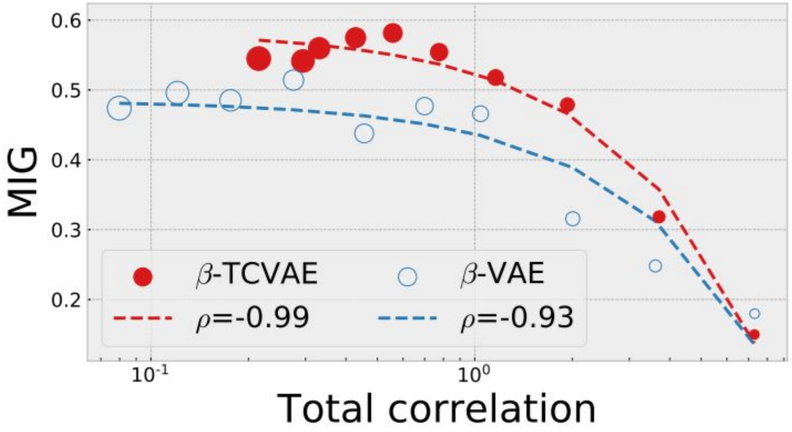
How is Independence related to Disentanglement?

Empirically, they seem to be correlated in both beta-VAE and TCVAE.

Slightly stronger correlation using TCVAE.



$$\mathrm{KL}[q(z)||\prod_j q(z_j)]$$

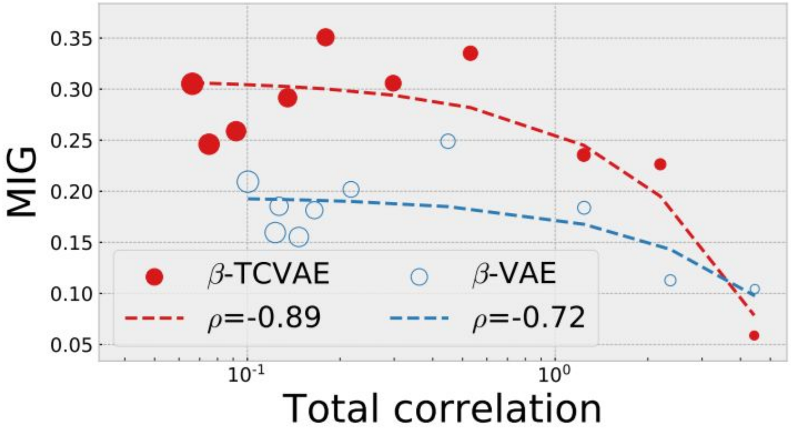


$$\mathrm{KL}[q(z)||\prod_j q(z_j)]$$

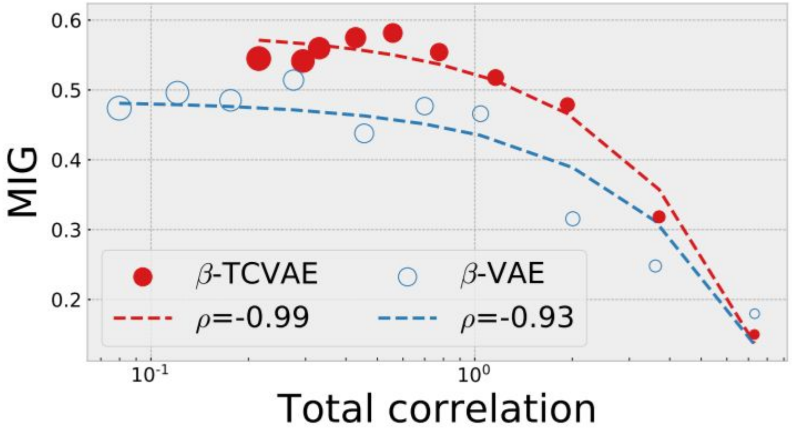
How is Independence related to Disentanglement?

Empirically, they seem to be correlated in both beta-VAE and TCVAE.

Slightly stronger correlation using TCVAE.



$\mathrm{KL}[q(z)|| \prod_j q(z_j)]$



$\mathrm{KL}[q(z)|| \prod_j q(z_j)]$

Hyvärinen & Oja. "Independent component analysis: algorithms and applications." (2000)

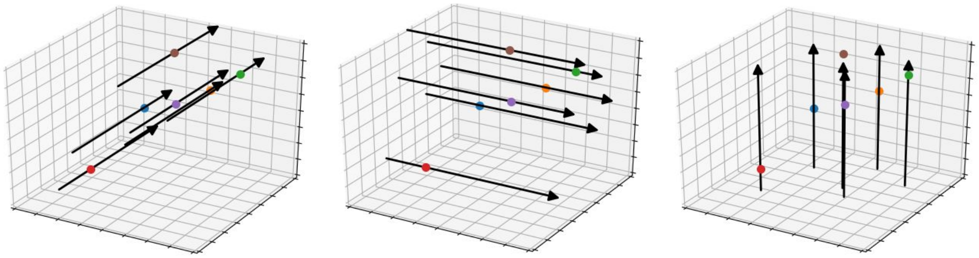
Qualitative Results

Bangs

Azimuth

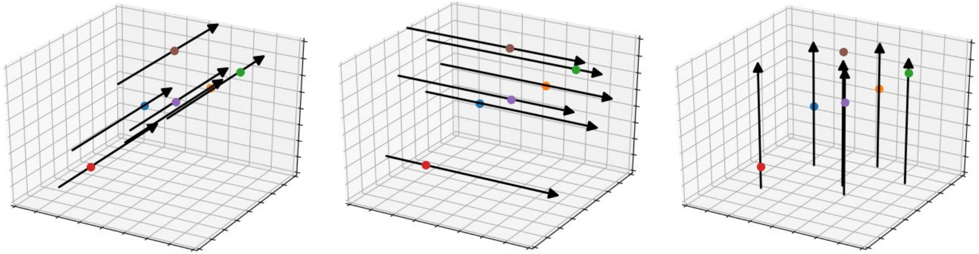
Glasses

CelebA: 15 interpretable dimensions.
Extrapolations are 6 standard deviations.



Qualitative Results

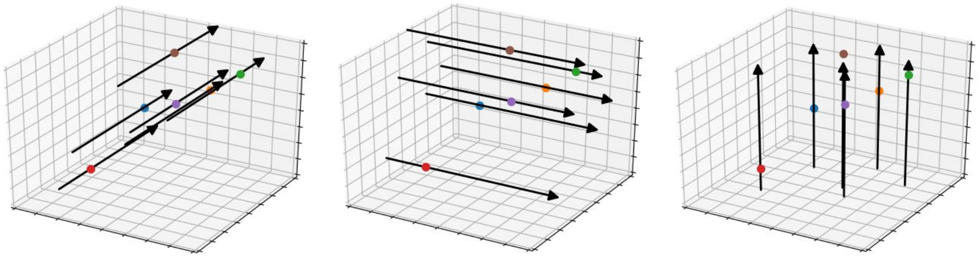
Smile Shadow Gender



CelebA: 15 interpretable dimensions.
Extrapolations are 6 standard deviations.

Qualitative Results

CelebA: 15 interpretable dimensions.
Extrapolations are 6 standard deviations.



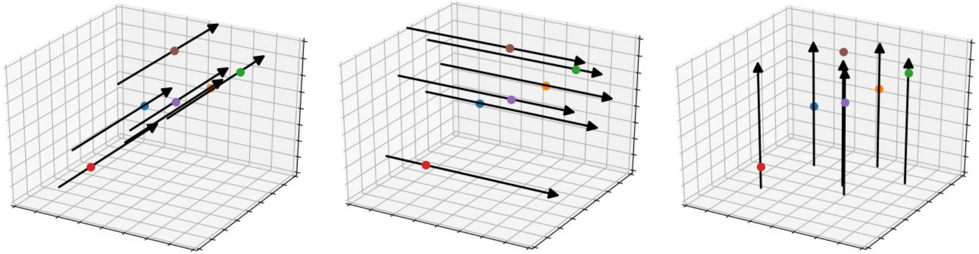
Qualitative Results

Baldness

Face Width

Eye shadow

CelebA: 15 interpretable dimensions.
Extrapolations are 6 standard deviations.



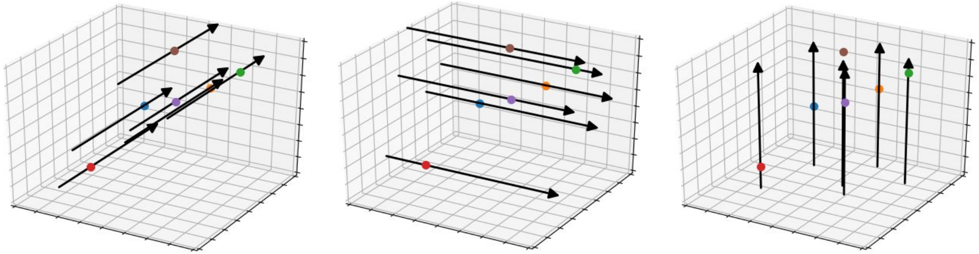
Qualitative Results

Hue

Smoldering Look

Mustache

CelebA: 15 interpretable dimensions.
Extrapolations are 6 standard deviations.



Future Directions

- Specific inductive biases for recovering specific factors.
- Better stochastic estimators of information theoretic quantities.
- Generalized notions of disentanglement.

Future Directions

- Specific inductive biases for recovering specific factors.
- Better stochastic estimators of information theoretic quantities.
- Generalized notions of disentanglement.

You may also be interested in…

- Kim & Mnih. “Disentangling by Factorising.”
- CP Burgess et al. “Understanding disentangling in beta-VAE”.
- Eastwood and Williams. “A Framework for the Quantitative Evaluation of Disentangled Representations.”
- Mathieu et al. “Disentangling Disentanglement.”
- Locatello et al. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations.”

Collaborators

UNIVERSITY OF TORONTO COMPUTER SCIENCE

Xuechen Li

Roger Grosse

David Duvenaud

VECTOR INSTITUTE

You may also be interested in…

- Kim & Mnih. “Disentangling by Factorising.”
- CP Burgess et al. “Understanding disentangling in beta-VAE”.
- Eastwood and Williams. “A Framework for the Quantitative Evaluation of Disentangled Representations.”
- Mathieu et al. “Disentangling Disentanglement.”
- Locatello et al. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations.”