Isolating Sources of Disentanglement in Variational Autoencoders

Tian Qi Chen, Xuechen Li, Roger Grosse, David Duvenaud

University of Toronto. Vector Institute.

Contributions

Variational autoencoders naturally discover disentangled representations. To understand this behavior, we explore a **refined decomposition of the KL regularization term in VAEs**.

We can amplify the source of disentanglement in VAEs which results in an **improved algorithm with the same number of hyperparameters** as the β-VAE. We call it $\boldsymbol{\beta}$**-TCVAE**.

Quantifying disentanglement is hard, and existing approaches are mostly *ad hoc*. We design a **new measure rooted in information theory**.

Background

The penalized VAE objective can be written using the evidence lower bound (ELBO):

$$\frac{1}{N}\sum_{n=1}^{N}\Big(\mathbb{E}_q[\log p(x_n|z)] - \beta D_{\mathrm{KL}}\left(q(z|x_n)||p(z)\right)\Big)$$

- $\beta = 1 \longrightarrow$ Standard VAE objective.
- $\beta > 1 \longrightarrow \beta$-VAE [1] for disentangling. Reliable in practice but not explicitly analyzed.

Notion of Disentanglement

Each dimension of a disentangled representation should:

(1) Represent a different factor of variation in the data.

(2) Be able to be changed independently of the other dimensions.

It is conjectured the following may be important:

(1) Mutual information between the latent variables and the data.

(2) Independence between the latent variables.

$$p(n) = 1/N \qquad q(z) = \sum_{i=1}^{N} p(n)q(z|n)$$
$$q(z,n) = q(z|n)p(n)$$

ELBO TC-Decomposition

$$\mathbb{E}_{p(n)}\left[D_{\mathrm{KL}}\left(q(z|n)||p(z)\right)\right] = \underbrace{D_{\mathrm{KL}}\Big(q(z,n)||q(z)p(n)\Big)}_{\text{(i) Index-Code MI}} + \underbrace{D_{\mathrm{KL}}\Big(q(z)||\prod_j q(z_j)\Big)}_{\text{(ii) Total Correlation}} + \underbrace{\sum_j D_{\mathrm{KL}}\Big(q(z_j)||p(z_j)\Big)}_{\text{(iii) Dimension-wise KL}}$$

Decomposition Breakdown

The ELBO objective decreases all three terms:

(i) Mutual information between the training data and the latent variables [2].

(ii) **Total correlation** (TC) between the latent variables. A measure of statistical dependence.

(iii) **Dimension-wise KL.** Simple regularization acting on each dimension of the representation.

Minibatch-based Estimation

We can train with *arbitrary weights on each term* if we can stochastically estimate $\log q(z)$ and $\log q(z_j)$.

Problem. Evaluation of $q(z)$ depends on full data.

Solution. Estimate $q(z)$ based on the current minibatch, and *weight appropriately*. Inspired by importance sampling.

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

where $z(n_i)$ is a sample from $q(z|n_i)$

Special case: β-TCVAE

We designate a special case of the decomposition as a meaningful algorithm for learning disentangled representations, the $\boldsymbol{\beta}$**-TCVAE** objective:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}_{q(z|n)}[\log p(n|z)]\right) - \text{(i)} - \beta\,\text{(ii)} - \text{(iii)}$$

With $\beta > 1$, this should encourage the representation to more disentangled while preserving information about the data.

Preliminary experiments indicate that tuning the weights on either (i) or (iii) do not have as much of an effect for learning disentangled representations.

Measuring Disentanglement

If we have a set of latent variables $\{z_j\}$ and set of known factors $\{v_k\}$, then we can use the empirical mutual information $I_n(z_j; v_k)$ to quantize how well a latent variable z_j reflects a ground truth factor v_k. The full metric we call **mutual information gap** (MIG) is

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

where $j^{(k)} = \mathrm{argmax}_j\, I_n(z_j; v_k)$ and K is the number of known factors.

The gap encourages two important properties:

- Axis-alignment of the representation.
- Compactness of the representation.

Quantitative Comparisons

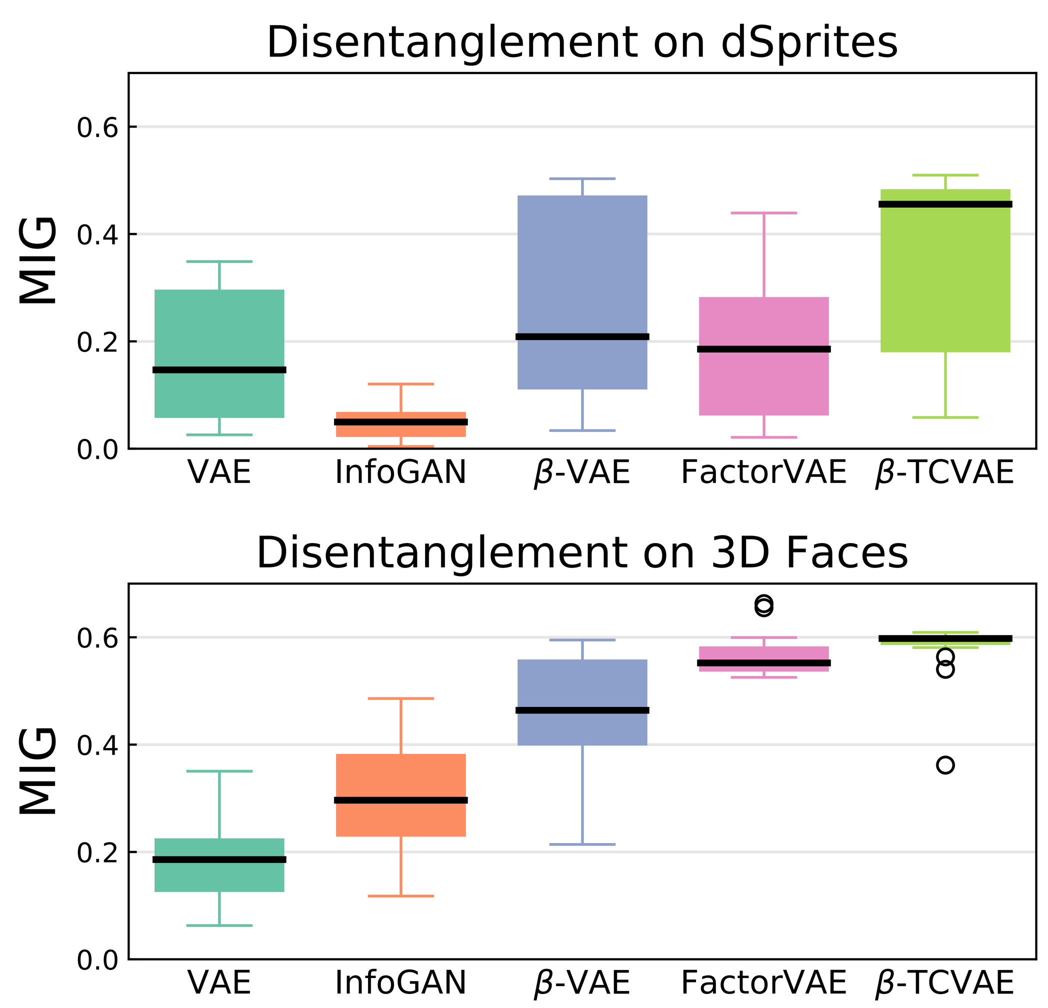


Figure: Distribution of disentanglement score (MIG) for representation learning algorithms.

Disentangled vs. Independent Representations

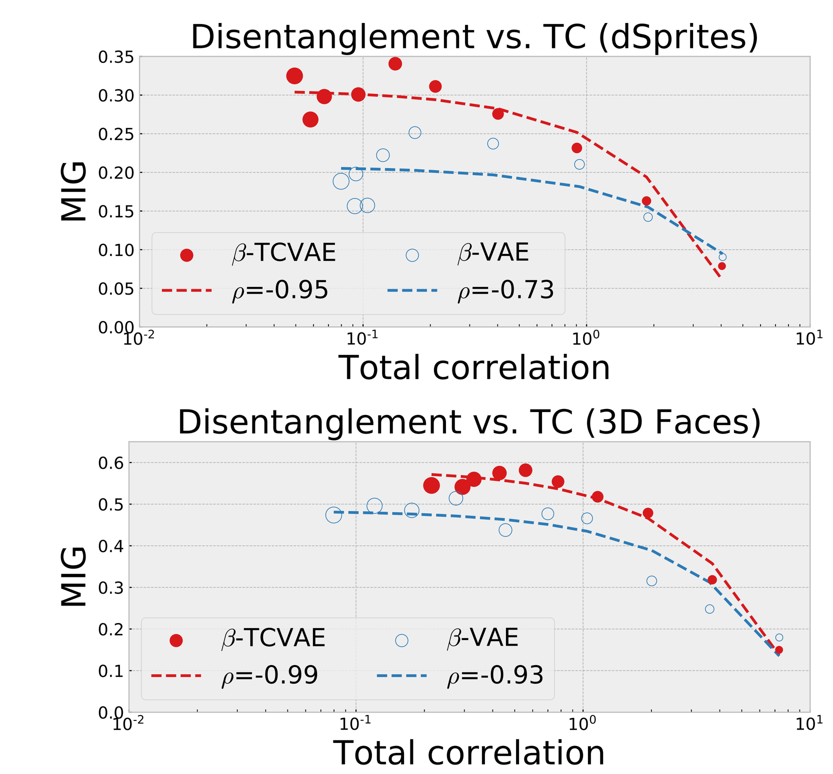
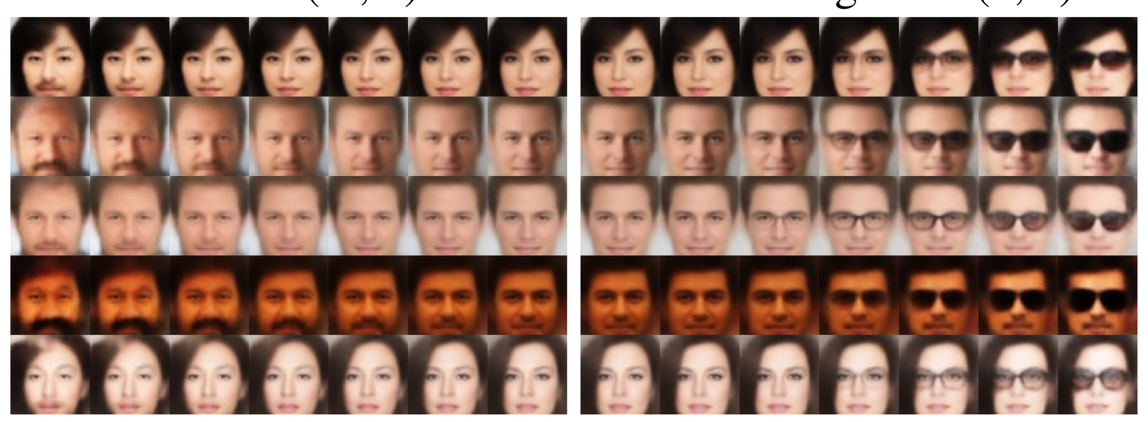


Figure: Scatter plots of the average MIG and TC per value of β. Larger circles indicate a higher β.

Qualitative Results

Gender (-6, 6) — Smoldering Gaze (0, 6)



Mustache (-6, 0) — Glasses (0, 6)

References

[1] Higgins *et al.* (2017). *Beta-VAE.*

[2] Hoffman & Johnson (2017). *ELBO Surgery.*

[3] Kim & Mnih (2018). *Disentangling by Factorising.*

[4] Achille & Soatto (2017). *Information Dropout.*