Disentanglement

VAE 4

Dachun Sun

Department of Computer Science
University of Illinois at Urbana-Champaign

September 28, 2021

Variational Autoencoders (VAEs)

Given a dataset $\mathbf{x}$ characterized by $P(\mathbf{x})$ and a latent random vector $\mathbf{z}$, we model the data as a distribution $p_\theta(\mathbf{x})$, with θ being the parameter.

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})\, d\mathbf{z}$$

- Prior $p_\theta(\mathbf{z})$
- Likelihood (probabilistic decoder) $p_\theta(\mathbf{x}|\mathbf{z})$
- Posterior (probablistic encoder) $p_\theta(\mathbf{z}|\mathbf{x})$

$p_\theta(\mathbf{x})$ needs to compute high-d integral, so we need to approximate the posterior distribution as

$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$$

where ϕ is the parameter.

Variational Autoencoders (VAEs)

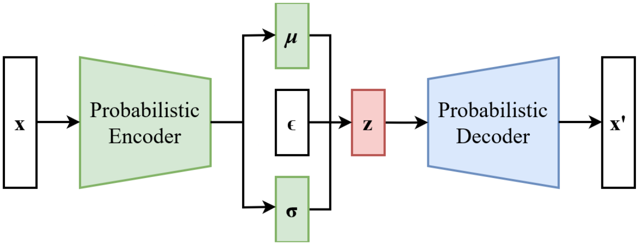


Figure 1: Model Architecture of VAEs.[a]

[a] Image credits to Wikipedia on Variational autoencoder.

VAE Loss

Evidence Lower Bound (ELBO)

We would like to maximize $p_\theta(\mathbf{x})$ through maximizing the lower bound of it.

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \log p_\theta(\mathbf{x}) - \overbrace{D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})\right)}^{\geq 0} \\ &= \underbrace{\underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\right)}_{\text{Regularization}}}_{\text{ELBO}}\end{aligned}$$

Therefore, during implementation, we have the VAE loss (negative ELBO):

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\right)$$

Disentanglement

Disentanglement = *Independence* + *Semantics*

- Unsupervised learning of a disentangled posterior distribution over the underlying generative factors of sensory data is a major challenge in AI research [BCV13] [LUTG17].
- Motivations include discovering independent components, controllable sample generation, and generalization/robustness.
- Facilitates **interpretable** decision making and controlled transfer.

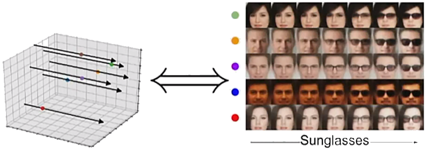


Figure 2: **Axis-aligned** traversal in the representation space and **Global interpretability** in data space.

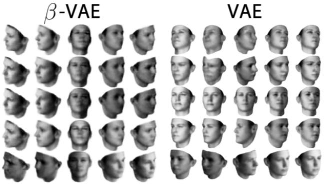


Figure 3: Traversal of the rotational latent dimension.

Credits to Ricky Chen's talk at NIPS 2018.

Datasets

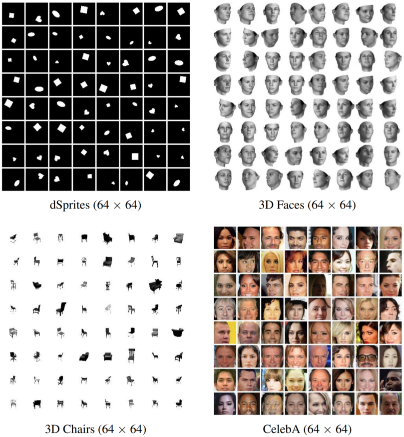


Figure 4: Real samples from the training datasets.

DC-IGN

- Deep Convolutional Inverse Graphics Network (DC-IGN) [KWKT15] has an architecture similar to VAE with special graphics code as the latent space.

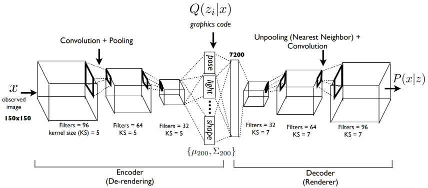


Figure 5: DC-IGN Architecture.

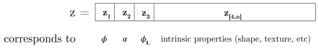


Figure 6: Structure of the representation vector.

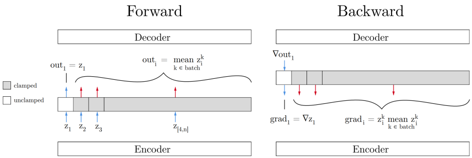


Figure 7: Training on a minibatch in which only ϕ, the azimuth angle of the face, changes.

InfoGAN

The GAN formulation uses a simple factored continuous input noise vector $\mathbf{z}$, but imposing no restrictions on how the generator may use it. So the generator may use it in a highly entangled way.
However, in InfoGAN [CDH+16],

- Uses a set of structured latent variables $\mathbf{c} = (c_1, \ldots, c_L)$, and assuming $p(\mathbf{c}) = \prod_{i=1}^{L} p(c_i)$.
- The generator becomes $G(\mathbf{z}, \mathbf{c})$.
- With no constraints, the generator could ignore $\mathbf{c}$, $p_G(\mathbf{x}|\mathbf{c}) = p_G(\mathbf{x})$.
- There should be high mutual information between latent code $\mathbf{c}$ and the generator distribution, meaning $I(\mathbf{c}; G(\mathbf{z}, \mathbf{c}))$ should be high.

A Note on Karush-Kuhn-Tucker (KKT) Conditions

Non-linear Programming

$$\begin{aligned} \text{Optimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \ldots, r \end{aligned}$$

Forming the Lagrangian function

$$L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\mu}^T [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})]^T + \boldsymbol{\lambda}^T [h_1(\mathbf{x}), \ldots, h_l(\mathbf{x})]^T$$

Karush-Kuhn-Tucker Conditions

1. **Stationarity**: $\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \mu_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^r \lambda_j \nabla h_j(\mathbf{x}^*) = 0$ for minimization.
2. **Primal Feasibility**: $g_i(\mathbf{x}^*) \geq 0, i = 1, \ldots, m$ and $h_j(\mathbf{x}^*) = 0, j = 1, \ldots, r$.
3. **Dual Feasibility**: $\mu_i \geq 0, i = 1, \ldots, m$.
4. **Complementary Slackness**: $\sum_{i=1}^m \mu_i g_i(\mathbf{x}^*) = 0$.

VAE Loss as an Optimization Problem

If we take a look at the VAE loss again

$$\theta, \phi = \underset{\theta,\phi}{\arg\min} \left\{ -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right) \right] \right\}$$

We can formulate it as a constrained optimization problem:

Optimization Problem from ELBO

$$\min_{\theta,\phi} -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \text{ subject to } D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right) \right] < \epsilon$$

Rewriting it as a Lagrangian under KKT conditions, we have

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + \beta \left(D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right) \right] - \epsilon \right)$$

Since $\beta, \epsilon \geq 0$ according to the complementary slackness.

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + \beta D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right) \right]$$

New Objective

β-VAE Loss

$$\mathcal{L}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + \beta D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right)$$

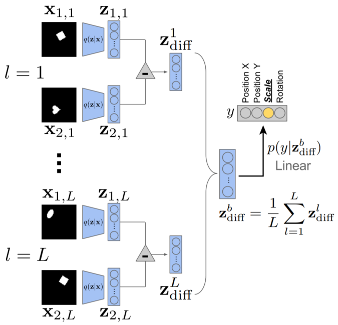
- Setting $\beta = 1$ corresponds to the original VAE formulation.
- Setting $\beta > 1$ puts a stronger constraint on the latent bottleneck
  - Limiting the capacity of $\mathbf{z}$ while trying to maximize the log-likelihood should encourage the model to learn a more efficient representation.
  - Higher value of β should encourage the conditional independence in $q_\phi(\mathbf{z}|\mathbf{x})$ because more weights are put on the D_{KL} term.
- Disentangled representation emerge when the right balance is found between reconstruction and latent capacity restriction.
  - Create a trade-off between reconstruction fidelity and the quality of the disentanglement.
- Note: In real implementations, β is usually a training-step dependent variable, from 0 to the set value. The intuition behind this warm-up is to first get the network to be able to learn reconstruction.

Measure Disentanglement

The Higgins' Metric

The accuracy that a low VC-dimension linear classifier can achieve at identifying a fixed ground truth factor [CLGD18].

1. Choose a factor $y \sim \mathcal{U}[1 \dots K]$.
2. For a batch of L samples:
   a. Sample two data points $\mathbf{x}_{1,l}, \mathbf{x}_{2,l}$ from the dataset where the chosen factor y has the same value.
   b. Obtain the latent representation $\mathbf{z}_{1,l}, \mathbf{z}_{2,l}$, and compute the difference $\mathbf{z}^l_{\text{diff}} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|$.
3. Use the average $\mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{z}^l_{\text{diff}}$ to predict $p(y|\mathbf{z}^b_{\text{diff}})$ and report the predictor's accuracy as **disentanglement metric score**.



Results

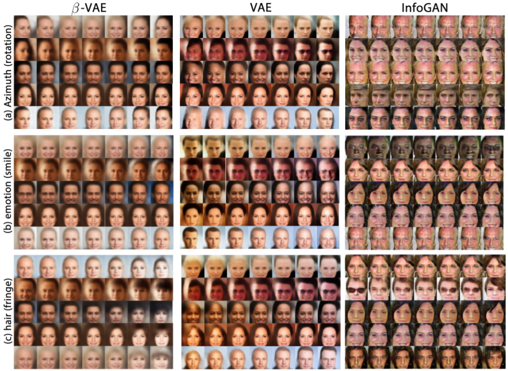


Figure 1: **Manipulating latent variables on celebA:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 250$), VAE (Kingma & Welling, 2014) ($\beta = 1$) and InfoGAN (Chen et al., 2016). *In all figures of latent code traversal each block corresponds to the traversal of a single latent variable while keeping others fixed to either their inferred (β-VAE, VAE and DC-IGN where applicable) or sampled (InfoGAN) values. Each row represents a different seed image used to infer the latent values in the VAE-based models, or a random sample of the noise variables in InfoGAN. β-VAE and VAE traversal is over the [-3, 3] range.* InfoGAN traversal is over ten dimensional categorical latent variables. Only β-VAE and InfoGAN learnt to disentangle factors like azimuth (a), emotion (b) and hair style (c), whereas VAE learnt an entangled representation (e.g. azimuth is entangled with emotion, presence of glasses and gender). InfoGAN images adapted from Chen et al. (2016). Reprinted with permission.

Results

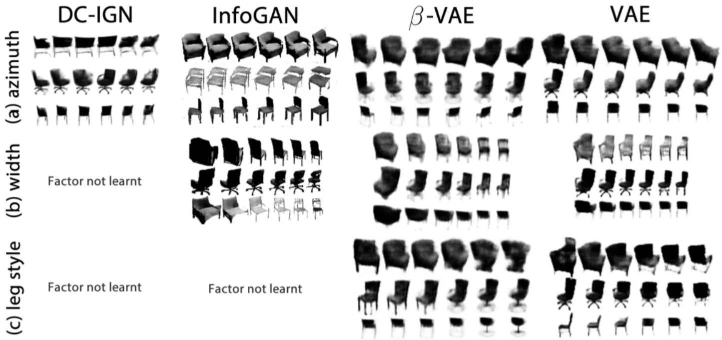


Figure 2: **Manipulating latent variables on 3D chairs:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 5$), VAE (Kingma & Welling, 2014) ($\beta = 1$), InfoGAN (Chen et al., 2016) and DC-IGN (Kulkarni et al., 2015). InfoGAN traversal is over the [-1, 1] range. VAE always learns an entangled representation (e.g. chair width is entangled with azimuth and leg style (b)). All models apart from VAE learnt to disentangle the labelled data generative factor, azimuth (a). InfoGAN and β-VAE were also able to discover unlabelled factors in the dataset, such as chair width (b). Only β-VAE, however, learnt about the unlabelled factor of chair leg style (c). InfoGAN and DC-IGN images adapted from Chen et al. (2016) and Kulkarni et al. (2015), respectively. Reprinted with permission.

Results

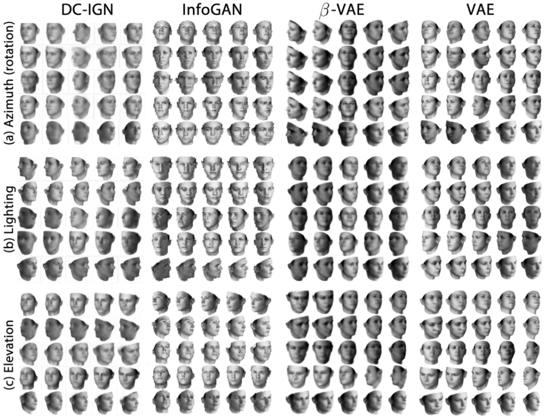


Figure 3: **Manipulating latent variables on 3D faces:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 20$), VAE (Kingma & Welling, 2014) ($\beta = 1$), InfoGAN (Chen et al., 2016) and DC-IGN (Kulkarni et al., 2015). InfoGAN traversal is over the [-1, 1] range. All models learnt to disentangle lighting (b) and elevation (c). DC-IGN and VAE struggled to continuously interpolate between different azimuth angles (a), unlike β-VAE, which additionally learnt to encode a wider range of azimuth angles than other models. InfoGAN and DC-IGN images adapted from Chen et al. (2016) and Kulkarni et al. (2015), respectively. Reprinted with permission.

Results

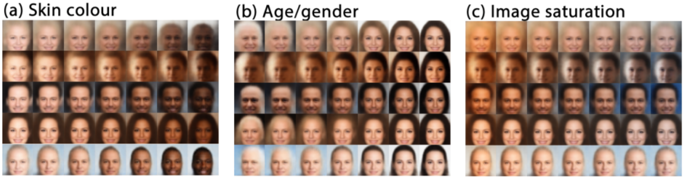


Figure 4: **Latent factors learnt by β-VAE on celebA:** traversal of individual latents demonstrates that β-VAE discovered in an unsupervised manner factors that encode skin colour, transition from an elderly male to younger female, and image saturation.

Results

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	$\mathbf{99.23 \pm 0.1\%}$

Figure 8: Disentanglement metric classification accuracy for 2D shapes dataset.

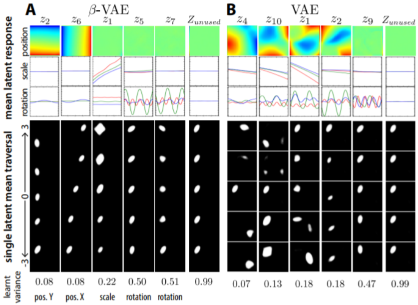


Figure 9: Representations learned by a β-VAE. Each column represents a latent z_i, ordered according to the learned Gaussian variance.

The Effect of β

- β is a mixing coefficient that weighs the gradients magnitudes between reconstruction and the prior-matching. So it is natural to consider normalized β in analysis by the latent space dimension M and input data dimension N, $\beta_{\text{norm}} = \frac{\beta M}{N}$.
- β being too low or too high, the model would learn a entangled representation due to either too much or too little capacity in the latent $\mathbf{z}$ bottleneck.
- Good disentanglement representations often lead to blurry reconstructions. However, in general, $\beta > 1$ is necessary to achieve good disentanglement.

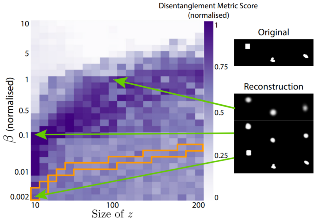


Figure 10: Positive correlation is present between the size of $\mathbf{z}$ and the optimal normalised values of β for disentangled factor learning for a fixed β-VAE architecture. Orange approximately corresponds to *unnormalized* $\beta = 1$.

The Problem with β-VAE

β-VAE Loss

$$\mathcal{L}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + \beta D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}) \right)]$$

Although tuning $\beta > 1$ showed promising results in disentanglement, β-VAE has several problems

- The trade-off between reconstruction and disentanglement.
- No mathematical explanation on the source of disentanglement by penalizing $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))$.
- The metric used lacks axis-alignment detection, tends to be ad-hoc, and sensitive to hyperparameters.

Decompose ELBO More

Mutual Information

Let (X, Y) be a pair of r.v.s over the space $\mathcal{X} \times \mathcal{Y}$. Then their mutual information is

1. $I(X; Y) = D_{\mathrm{KL}}(p(X, Y) \| p(X)p(Y))$
2. $I(X; Y) = \mathbb{E}_X \left[D_{\mathrm{KL}}(p(Y|X) \| p(Y))\right] = \mathbb{E}_Y \left[D_{\mathrm{KL}}(p(X|Y) \| p(X))\right]$

$I(X; Y)$ intuitively measures how much could you infer about the other random variable if you are given knowledge about one of them. $I(X; Y) = 0$ means independence because nothing can be inferred (not related at all).

ELBO TC-Decomposition

Define a uniform random variable on $\{1, 2, \ldots, N\}$ with which each data point relates. Denote $q(\mathbf{z}|n) = q(\mathbf{z}|x_n)$ and $q(\mathbf{z}, n) = q(\mathbf{z}|n)p(n) = q(\mathbf{z}|n)\frac{1}{N}$. $q(\mathbf{z}) = \sum_{n=1}^{N} q(\mathbf{z}|n)p(n)$ is the *aggregated posterior*. Then, we can decompose the regularization term in the ELBO as

$$\frac{1}{N}\sum_{n=1}^{N} D_{\mathrm{KL}}\left(q(\mathbf{z}|x_n) \| p(\mathbf{z})\right) = \mathbb{E}_{p(n)}\left[D_{\mathrm{KL}}\left(q(\mathbf{z}|n) \| p(\mathbf{z})\right)\right]$$

$$= \underbrace{D_{\mathrm{KL}}(q(\mathbf{z}|n) \| p(\mathbf{z}))}_{\text{Index-Code MI}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{z}) \| \prod_j q(z_j))}_{\text{Total Correlation}} + \underbrace{\sum_j D_{\mathrm{KL}}\left(q(z_j \| p(z_j))\right)}_{\text{Dimension-wise KL}}$$

Decompose ELBO More

ELBO TC-Decomposition

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

- The index-code MI is the mutual information $I_q(\mathbf{z}; n)$. It is argued that higher mutual information can lead to better disentanglement, but recent investigations also claim that a penalized one encourages compact and disentangled representations.
- The total correlation is one of many generalization of mutual information. It is a measure of dependency between the variables. This is claimed to be the main source of disentanglement.
- The dimension-wise KL divergence mainly prevents individual latent dimensions from deviating too far from priors. It acts like a complexity penalty.

β-TCVAE Loss

$$\mathcal{L} = -\mathbb{E}_{q(\mathbf{z}|n)p(n)}\left[\log p(n|\mathbf{z})\right] + \alpha I_q(\mathbf{z}; n) + \beta D_{\mathrm{KL}}(q(\mathbf{z})\|\prod_j q(z_j)) + \gamma \sum_j D_{\mathrm{KL}}\left(q(z_j\|p(z_j))\right)$$

Decompose ELBO More

β-TCVAE Loss

$$\mathcal{L} = -\mathbb{E}_{q(\mathbf{z}|n)||p(n)}\left[\log p(n|\mathbf{z})\right] + \alpha I_q(\mathbf{z}; n) + \beta D_{\mathrm{KL}}(q(\mathbf{z})\| \prod_j q(z_j)) + \gamma \sum_j D_{\mathrm{KL}}\left(q(z_j\|p(z_j))\right)$$

- With ablation studies, tuning β leads to the best results. The proposed model uses $\alpha = \gamma = 1$, which is the same object as in FactorVAE [KM18].
- Provides better trade-off between density estimation and disentanglement. Different from β-VAE, higher value of β would not penalize the mutual information term too much.

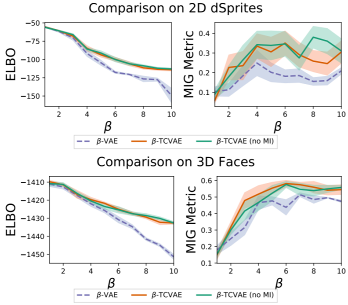


Figure S8: ELBO vs. Disentanglement plots showing β-TCVAE (4) but with α set to 0.

Decompose ELBO More

β-TCVAE Loss

$$\mathcal{L} = -\mathbb{E}_{q(\mathbf{z}|n)p(n)}\left[\log p(n|\mathbf{z})\right] + \alpha I_q(\mathbf{z}; n) + \beta D_{\mathrm{KL}}(q(\mathbf{z}) \| \prod_j q(z_j)) + \gamma \sum_j D_{\mathrm{KL}}\left(q(z_j \| p(z_j))\right)$$

- Decomposition expression requires the evaluation of the density $q(\mathbf{z}) = \mathbb{E}_{p(n)}\left[q(\mathbf{z}|n)\right]$, which depends on the entire dataset.
- Simple Monte Carlo approximation is not likely to work because if we view $q(\mathbf{z})$ as a mixture distribution where the data index n indicates the mixture component, a randomly sampled component $q(\mathbf{z}|n)$ is likely to be close to 0.
- $q(\mathbf{z}|n)$ would be large if n is the component that z comes from. So we should perform weighted sampling.
- Given a minibach of samples $\{n_1, \ldots, n_m\}$, we use the estimator

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

where $\mathbf{z}(n_i) \sim q(\mathbf{z}|n_i)$.

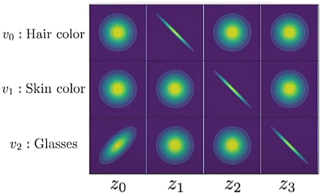
Measure Disentanglement

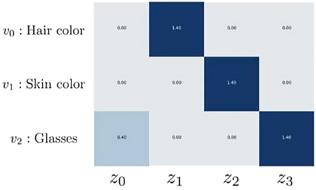
Mutual Information Gap (MIG)

- Estimate the mutual information between a latent variable z_i and a ground truth factor v_k by $q(z_j, v_k) = \sum_{n=1}^{N} p(v_k)p(n|v_k)q(z_j|n)$, and use it in some way.
- A higher mutual information implies that z_j contains a lot of information about v_k. MI is maximal if there exists a deterministic, invertible relationship between z_j and v_k.

1. For each v_k, take z_j, z_l that has the highest and the second highest mutual information with v_k.
2. $\text{MIG} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{H(v_k)} \left(I(z_j; v_k) - I(z_l; v_k)\right)$

- Averaging by K and normalizing by the entropy $H(v_k)$ provides a value between 0 and 1.
- MIG $\rightarrow$ 1 implies good disentanglement.



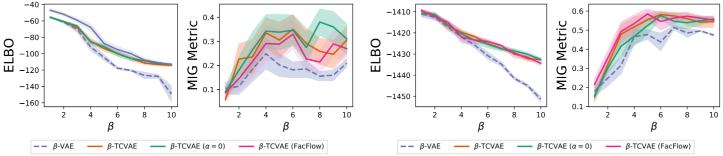


Results

(a) Baldness (-6, 6) (b) Face width (0, 6) (c) Gender (-6, 6) (d) Mustache (-6, 0)

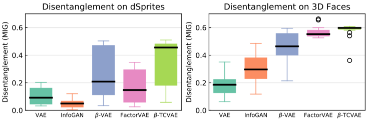
Figure 1: Qualitative comparisons on CelebA. Traversal ranges are shown in parentheses. Some attributes are only manifested in one direction of a latent variable, so we show a one-sided traversal. Most semantically similar variables from a β-VAE are shown for comparison.

Results



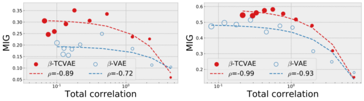
(a) dSprites (b) 3D Faces

Figure 2: Compared to β-VAE, β-TCVAE creates more disentangled representations while preserving a better generative model of the data with increasing β. Shaded regions show the 90% confidence intervals. Higher is better on both metrics.



(a) dSprites (b) 3D Faces

Figure 3: Distribution of disentanglement score (MIG) for different modeling algorithms.



(a) dSprites (b) 3D Faces

Figure 4: Scatter plots of the average MIG and TC per value of β. Larger circles indicate a higher β.

Results

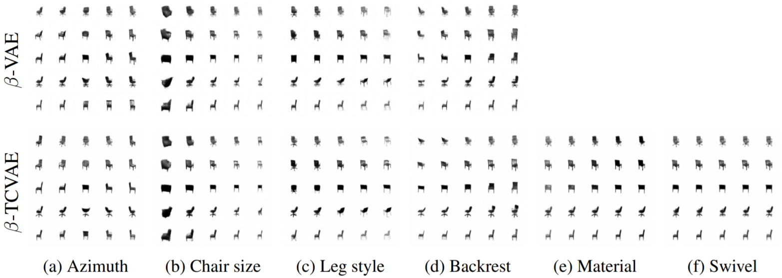


Figure 6: Learned latent variables using β-VAE and β-TCVAE are shown. Traversal range is (-2, 2).

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent.
Representation learning: A review and new perspectives.
IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[CDH+16] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel.
Infogan: Interpretable representation learning by information maximizing generative adversarial nets.
In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2180–2188, 2016.

[CLGD18] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud.
Isolating sources of disentanglement in variational autoencoders.
arXiv preprint arXiv:1802.04942, 2018.

[HMP+16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner.
beta-vae: Learning basic visual concepts with a constrained variational framework.
2016.

[KM18] Hyunjik Kim and Andriy Mnih.
Disentangling by factorising.
In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[KWKT15] Tejas D Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B Tenenbaum.
Deep convolutional inverse graphics network.
arXiv preprint arXiv:1503.03167, 2015.

[LUTG17] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman.
Building machines that learn and think like people.
Behavioral and brain sciences, 40, 2017.