beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Minjae Kim

Disentangled

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

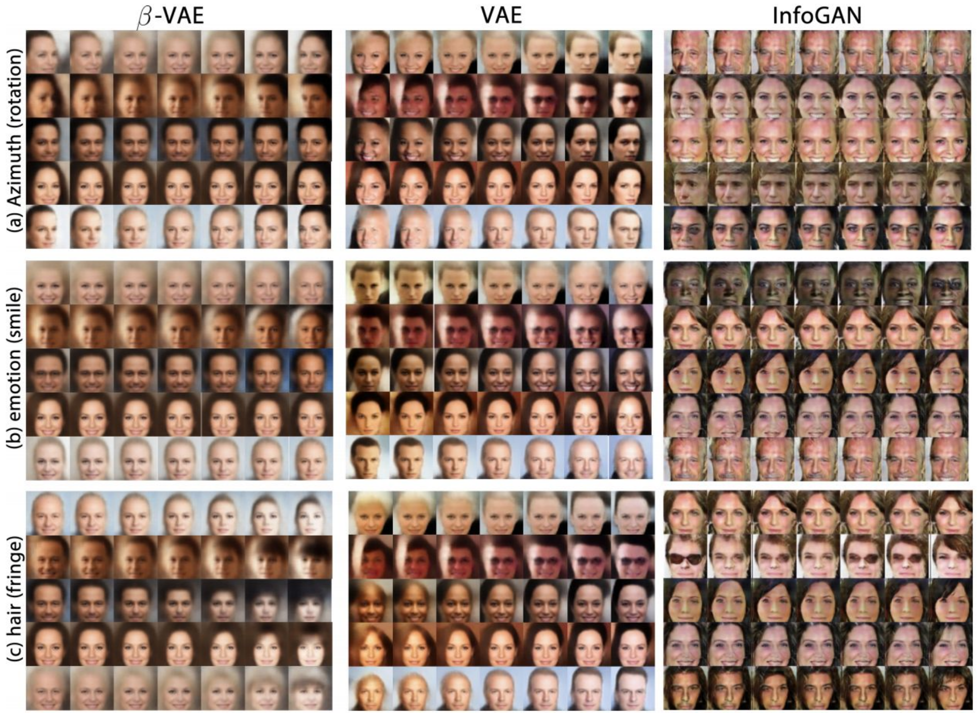
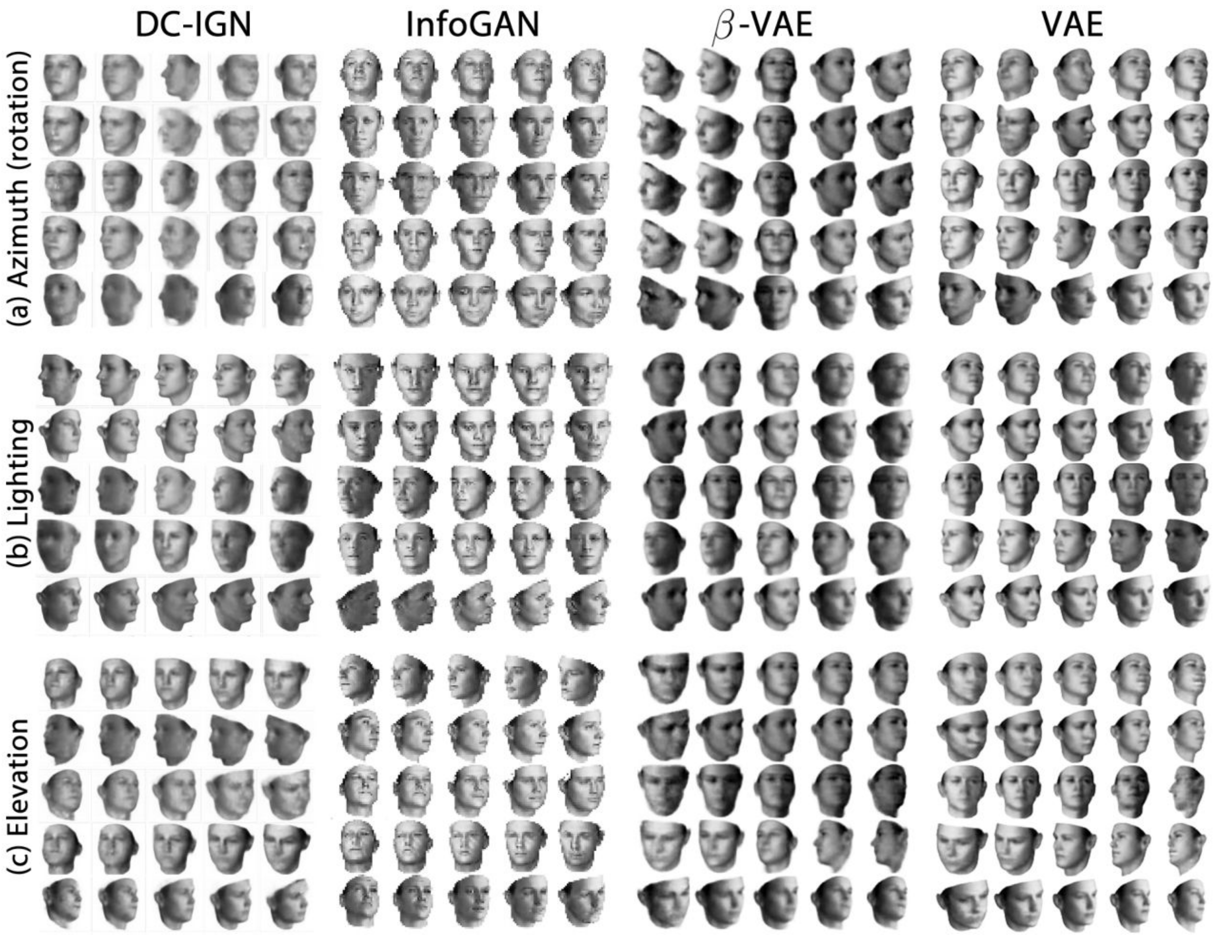


Figure 1: **Manipulating latent variables on celebA:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 250$), VAE (Kingma & Welling, 2014) ($\beta = 1$) and InfoGAN (Chen et al., 2016). *In all figures of latent code traversal each block corresponds to the traversal of a single 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.

Disentangled



Contribution

1) we propose beta-VAE, a new unsupervised approach for learning disentangled representations of independent visual data generative factors
2) we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models
3) we demonstrate both qualitatively and quantitatively that our beta-VAE approach achieves state-of-the-art disentanglement performance compared to various baselines on a variety of complex datasets.

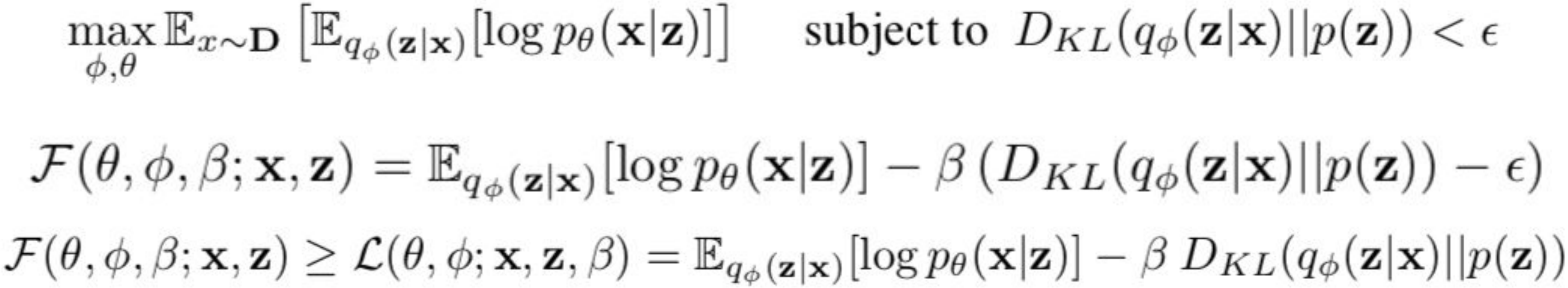
Variational AutoEncoder

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$$

$$\log p_\theta(\mathbf{x}|\mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big)$$

$$z_i = \mu_i + \sigma_i \epsilon$$

beta-VAE



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

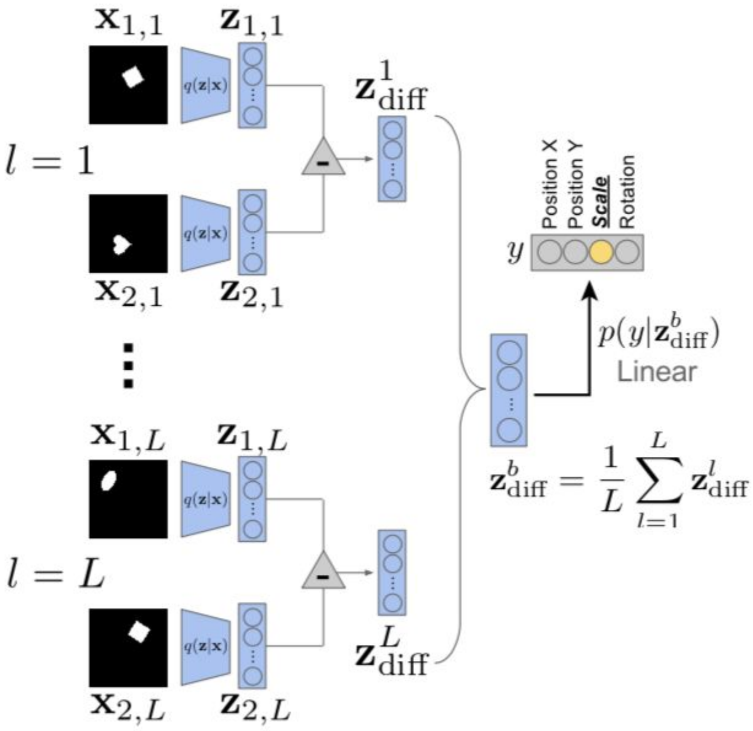
$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\,(D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \epsilon)$$

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

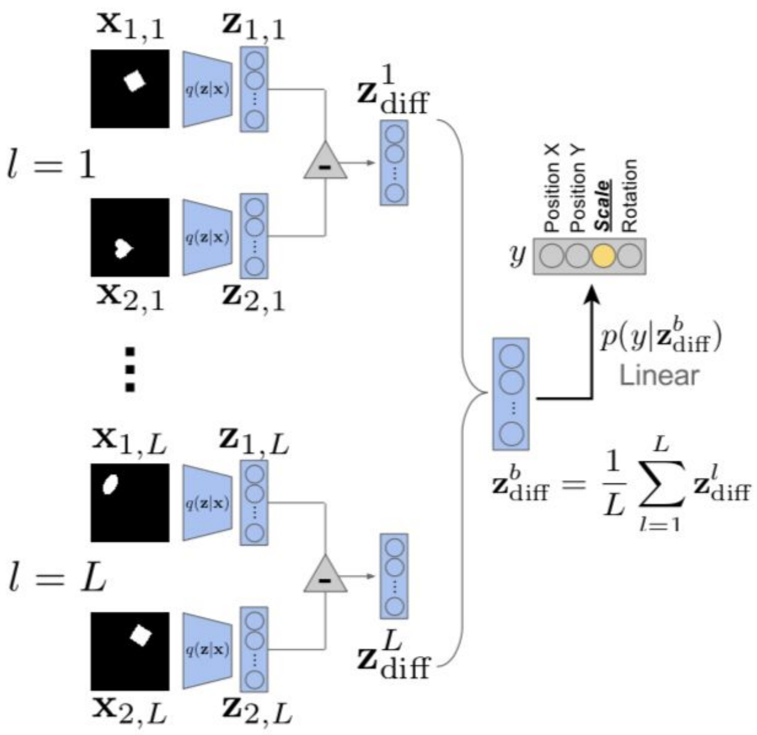
The extra pressures coming from high beta values may create a trade-off between reconstruction fidelity and the quality of disentanglement within the learnt latent representations.

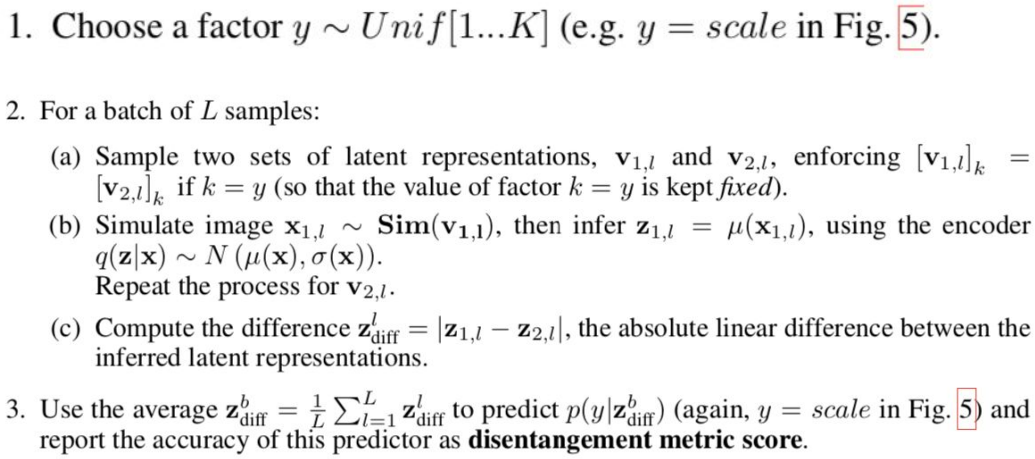
Disentanglement Metric

A representation of the data that is disentangled with respect to these generative factors, i.e. which encodes them in separate latents, would enable robust classification even using very simple linear classifiers.


$\mathbf{x}_{1,1}$
$\mathbf{z}_{1,1}$
$q(\mathbf{z}|\mathbf{x})$
$l = 1$
-
$\mathbf{z}_{\text{diff}}^{1}$
$\mathbf{x}_{2,1}$
$\mathbf{z}_{2,1}$
$\mathbf{x}_{1,L}$
$\mathbf{z}_{1,L}$
$l = L$
$\mathbf{z}_{\text{diff}}^{L}$
$\mathbf{x}_{2,L}$
$\mathbf{z}_{2,L}$
Position X
Position Y
Scale
Rotation
y
$p(y|\mathbf{z}_{\text{diff}}^{b})$
Linear
$\mathbf{z}_{\text{diff}}^{b} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{z}_{\text{diff}}^{l}$

Disentanglement Metric


$\mathbf{x}_{1,1}$
$\mathbf{z}_{1,1}$
$q(\mathbf{z}|\mathbf{x})$
$\mathbf{z}_{\text{diff}}^{1}$
$l = 1$
-
$\mathbf{x}_{2,1}$
$\mathbf{z}_{2,1}$
Position X
Position Y
Scale
Rotation
y
$p(y|\mathbf{z}_{\text{diff}}^{b})$
Linear
$\mathbf{z}_{\text{diff}}^{b} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{z}_{\text{diff}}^{l}$
$\mathbf{x}_{1,L}$
$\mathbf{z}_{1,L}$
$l = L$
$\mathbf{z}_{\text{diff}}^{L}$
$\mathbf{x}_{2,L}$
$\mathbf{z}_{2,L}$



1. Choose a factor $y \sim Unif[1...K]$ (e.g. $y = scale$ in Fig. 5).

2. For a batch of L samples:

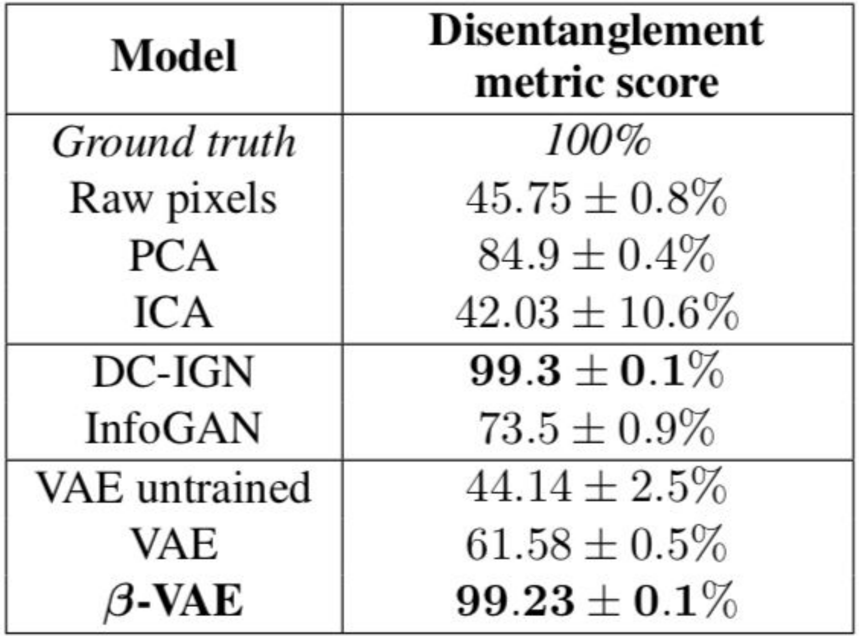
   (a) Sample two sets of latent representations, $\mathbf{v}_{1,l}$ and $\mathbf{v}_{2,l}$, enforcing $[\mathbf{v}_{1,l}]_k = [\mathbf{v}_{2,l}]_k$ if $k = y$ (so that the value of factor $k = y$ is kept *fixed*).

   (b) Simulate image $\mathbf{x}_{1,l} \sim \mathbf{Sim}(\mathbf{v_{1,l}})$, then infer $\mathbf{z}_{1,l} = \mu(\mathbf{x}_{1,l})$, using the encoder $q(\mathbf{z}|\mathbf{x}) \sim N(\mu(\mathbf{x}), \sigma(\mathbf{x}))$.
   Repeat the process for $\mathbf{v}_{2,l}$.

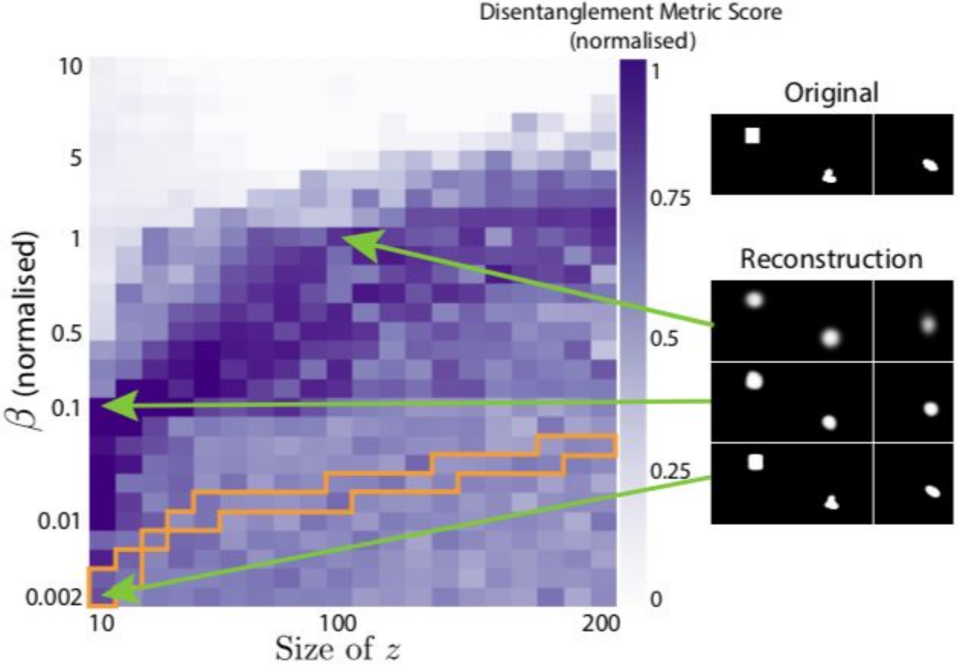
   (c) Compute the difference $\mathbf{z}_{\text{diff}}^{l} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|$, the absolute linear difference between the inferred latent representations.

3. Use the average $\mathbf{z}_{\text{diff}}^{b} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{z}_{\text{diff}}^{l}$ to predict $p(y|\mathbf{z}_{\text{diff}}^{b})$ (again, $y = scale$ in Fig. 5) and report the accuracy of this predictor as **disentangement metric score**.

Result



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



Disentanglement Metric Score (normalised)
β (normalised)
Size of z
Original
Reconstruction