Disentangling by Factorising

Diego Aguado

October 25, 2019

Overview

1. Context: Disentanglement
2. Motivation: Analyzing β - VAEs objective.
3. Contributions
4. Experiments

Concept and Definition

Disentanglement:

- A representation where a change in a fixed dimension corresponds one factor of variation while being relatively invariant to changes in other factors. [1]

Factors could be know or unknown.

Relevance of Disentanglement

AI would benefit greatly from models that learn representations

- Of in high dimensional inputs through latent factors.
- With semantic meaning.
- That are disentangled, by approximating a prior that reflects it (factorization).

The previous would be particularly useful if obtained by an unsupervised methods.

Why ?

- Useful on supervised learning tasks.
- To train RL agents through latent representations instead of high dimensional input.
- Since humans in the loop is highly innefficient for the process
  - Generative models trained in an unsupervised fashion are apprealing.

Disentangling with β-Variational Autoencoders

β-VAE models are an extension of Variational Autoencoders that achieve disentangled representations by penalizing the KL divergence in the objective function with a parameter β.

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbb{E}_{q(z|x^i)}[\log(p(x^i|z)] - \beta \mathrm{KL}(q(z|x^i)||p(z))] \tag{1}$$

Motivation and Goals

Motivation

- $\beta - VAE$ models will sacrifice reconstruction quality to achieve disentanglement.
- To achieve disentanglement a factorized prior is imposed:

$$p(z) = \prod_j p(z_j) := \mathcal{N}(0, I). \tag{2}$$

Goals of the paper

- Achieve disentanglement without incurring into reconstruction error-disentanglement trade off (or reduce it as much as possible).
- To have a robust metric to evaluate said disentanglement.
- Provide quantitative (and qualitative) comparisson after evaluating different models that aim to get disentangled representations.

Reconstruction error - Disentanglement trade off

To understand the reconstruction error - disentanglemente trade off the authors analize the penalized term in the β-VAE objective.
Expanding on the penalized KL

$$\mathbb{E}_{p_{data}(x)}[KL(q(z|x)||p(z)] = \mathcal{I}(x;z) + KL(q(z)||p(z)) \tag{3}$$

The second term pushes q to the factorial p that'll produce disentanglement yet the first term penalizes mutual information on x and z.

A different approach to disentanglement: TC penalty and Factor VAE

To reach disentanglement without incurring in the trade off a different objective is proposed by the authors and the way to do it is to push the posterior to a factorized distribution.
The proposed objective penalizes the total correlation,

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbb{E}_{q(z|x^i)}[\log(p(x^i|z)] - \mathrm{KL}(q(z|x^i)||p(z))] - \gamma\mathrm{KL}(q(z)||\bar{q}(z)) \quad (4)$$

with $\bar{q}(z) = \prod_{i=1}^{d} q(z_i)$ and $q(z) = \frac{1}{N}\sum_{i=1}^{N} q(z|x^i)$ the aggregated posterior.

TC penalty and FactorVAE

By adding the last term, the objective minimizes when the total contribution of one component $q(z_j)$ to the total $q(z)$ is low.
This KL divergence is intractable, so the authors use the *density-ratio* trick to estimate it.

Train Strategy

- The VAE and classifier are trained jointly(almost adversarial fashion) with the objective proposed.
- The classifier is trained using divergence minimization.

Train Strategy

PseudoCode

Algorithm 2 FactorVAE

Input: observations $(x^{(i)})_{i=1}^{N}$, batch size m, latent dimension d, γ, VAE/Discriminator optimisers: g, g_D
Initialize VAE and discriminator parameters θ, ψ.
repeat
Randomly select batch $(x^{(i)})_{i\in\mathcal{B}}$ of size m
Sample $z_\theta^{(i)} \sim q_\theta(z|x^{(i)})\ \forall i \in \mathcal{B}$

$$\theta \leftarrow g(\nabla_\theta \frac{1}{m} \sum_{i\in\mathcal{B}} [\log \frac{p_\theta(x^{(i)}, z_\theta^{(i)})}{q_\theta(z_\theta^{(i)}|x^{(i)})} - \gamma \log \frac{D_\psi(z_\theta^{(i)})}{1-D_\psi(z_\theta^{(i)})}])$$

Randomly select batch $(x^{(i)})_{i\in\mathcal{B}'}$ of size m
Sample $z_\theta^{\prime(i)} \sim q_\theta(z|x^{(i)})$ for $i \in \mathcal{B}'$

$$(z_{perm}^{\prime(i)})_{i\in\mathcal{B}'} \leftarrow \texttt{permute\_dims}((z_\theta^{\prime(i)})_{i\in\mathcal{B}'})$$

$$\psi \leftarrow g_D(\nabla_\psi \frac{1}{2m} [\sum_{i\in\mathcal{B}} \log(D_\psi(z_\theta^{(i)})) + \sum_{i\in\mathcal{B}'} \log(1 - D_\psi(z_{perm}^{\prime(i)}))])$$

until convergence of objective.

Figure: Source: Kim et al. (2018)

Measuring disentanglement

One popular way to evaluate disentanglement is by doing latent traversal to see the efects but it's qualitative nature makes it not suitable for algorithm comparisson.

Measuring disentanglement

Higgins et al(2016)[2] proposed a metric to evaluate disentanglemente using a linear classifier:

1. Choose a factor k and generate data with said factor fixed and the rest varying randomly, obtain their representation.
2. Calculate the mean of the representation and take the absolute value of the difference pairwise.
3. Then the mean of these statistics across pairs give one training input for the classifier and the fixed factor k is the output.

$$(\frac{1}{L}\sum_{l=1}^{L}|z^{(l)} - z^{'(l)}|, k) \tag{5}$$

Measuring disentanglement

Considering the previous, the metric is the error rate of the classifier. Nevertheless, this is sensitive to hyperparameters and has a failure mode that involves learning $K - 1$ disentangled representations.

Measuring disentanglement

The metric proposed by the authors involves using the variance of the representation when fixing a factor k.

1. Choose a fixed factor k, generate data with this factor and the rest varying and obtain representations.
2. Normalize each dimension by the empirical σ
3. Use the normalized variance in each dimension to choose the index with the lowest variance and the target index k to provide a training input/ouput example.

The classifier is the majority voter classifier, again the metric is the error rate.

Measuring disentanglement

Old and new metrics:

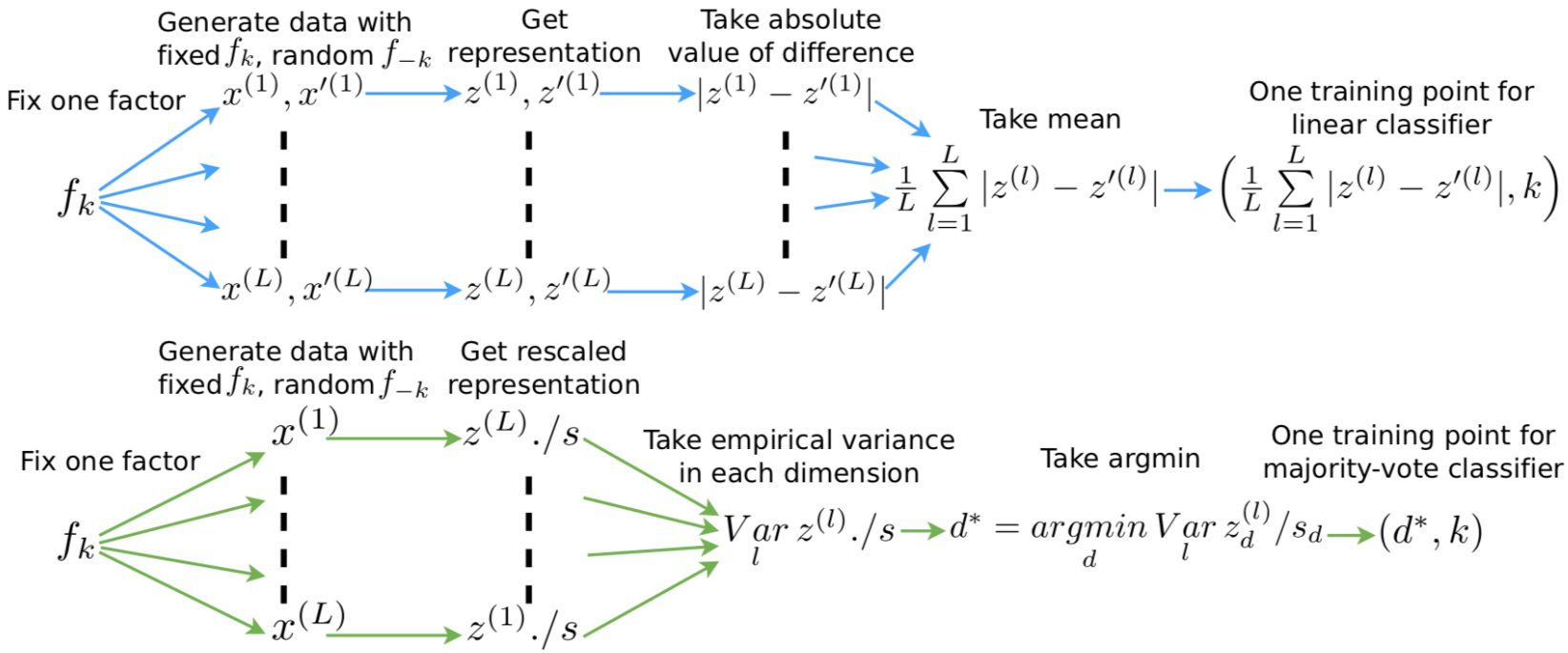

Figure: Diagrams for metrics. Source: Kim et al. (2018)

Types of Experiments

The authors perform two types of experiments:

- with known factors: 2D and 3D shapes
- with unknown factors: 3D faces, 3D Chairs, CelebA

Type I Experiments

In the first type of experiments, the authors are able to actually evaluate the old and new metrics of disentanglement.
Some observations on these experiments:

- For a given reconstruction error, FactorVAE reaches better disentanglement scores than VAEs.
- For 2D shapes, both Factor and β-VAEs are able to find x, y position and scale yet struggled to get orientation and shape. Neither robustly shape.
- For 3D shapes, both models struggled to disentangle shape and scale.
- Discrete factor, shape, is not being captured by any model.

Type I Experiments

Reconstruction error and disentanglement metrics.

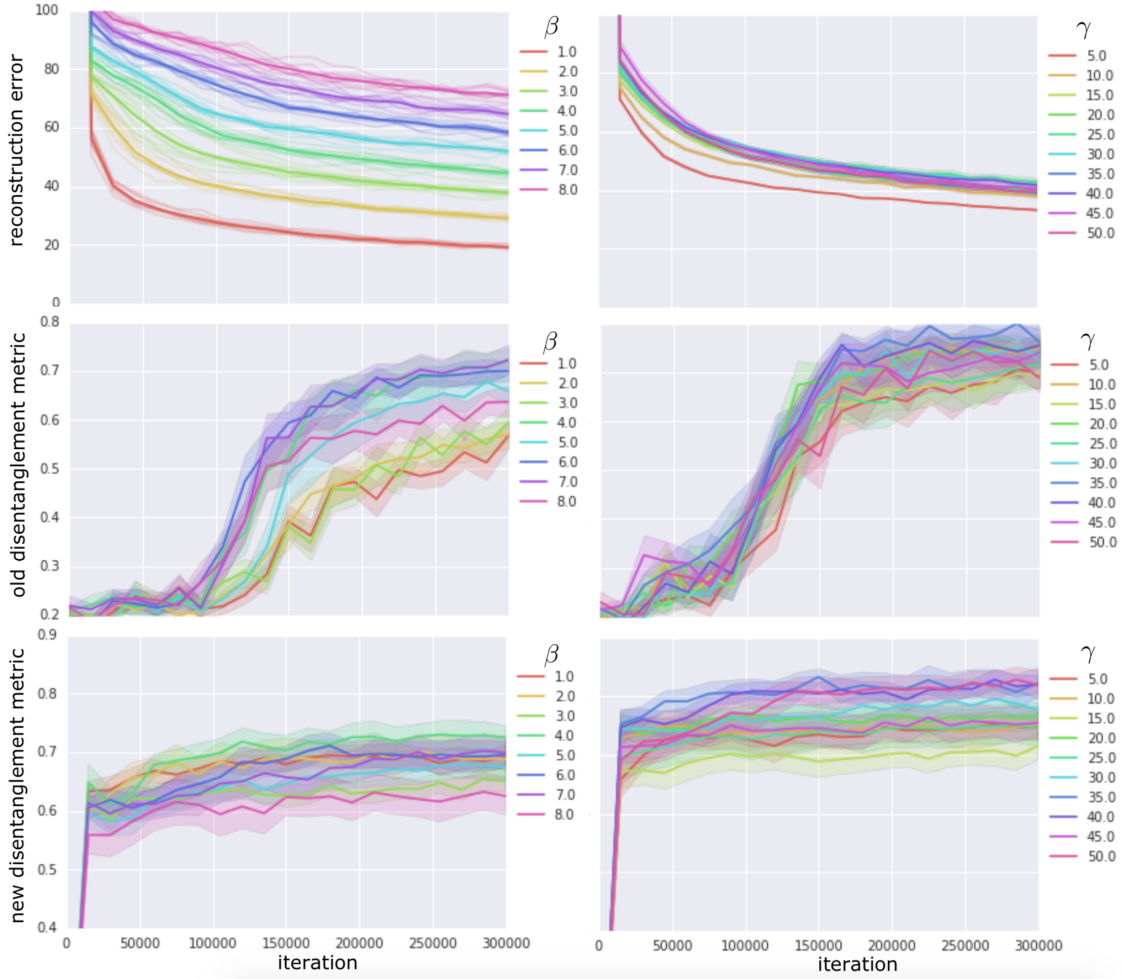

Figure: β-VAE (left), FactorVAE (right). Confidence intervals are over 10 random seeds. Source: Kim et al. (2018)

Disentanglement and reconstruction

FactorVAE presents less trade off between reconstruction and disentanglement than β-VAE.

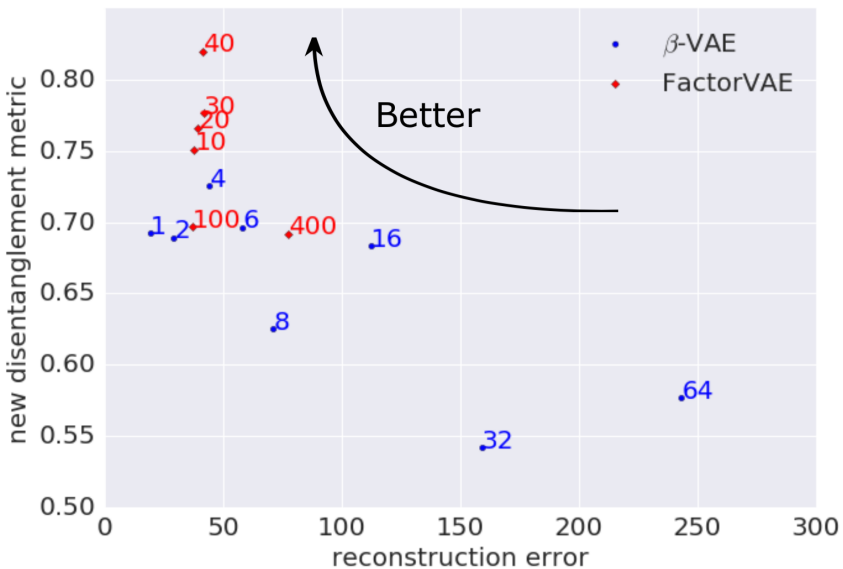

Figure: Reconstruction vs Disentanglement for FactorVAE and β-VAE. Source: Kim et al. (2018)

Latent Traversal

On 2D Shapes.

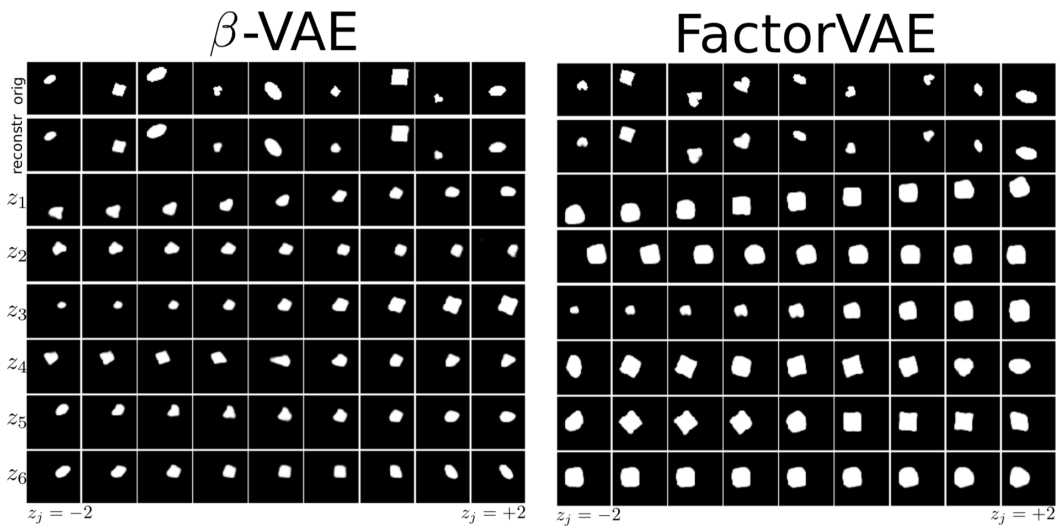

Figure 6. First row: originals. Second row: reconstructions. Remaining rows: reconstructions of latent traversals across each latent dimension sorted by $KL(q(z_j|x)||p(z_j))$, for the best scoring models on our disentanglement metric. Left: β-VAE, score: 0.814, $\beta = 4$. Right: FactorVAE, score: 0.889, $\gamma = 35$.

Figure: Source: Kim et al. (2018)

Latent Traversal

On 3D Shapes.

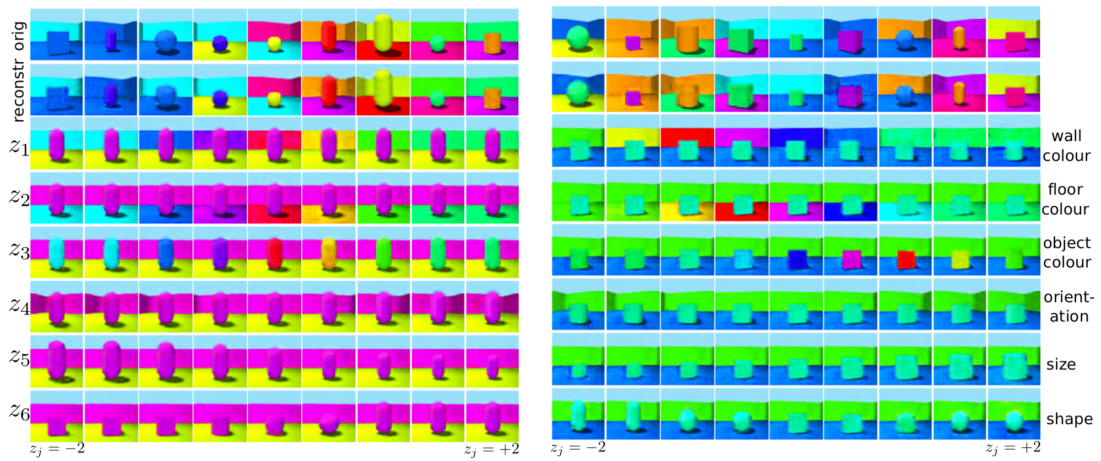

Figure 11. Same as Figure 6 but for 3D Shapes data. Left: β-VAE, score: 1.00, $\beta = 32$. Right: FactorVAE, score: 1.00, $\gamma = 7$.

Figure: Source: Kim et al. (2018)

Type I Experiments: InfoGANS

The authors compare the VAE-based models with InfoWGAN-GP, a particular case of InfoGan that uses the Wasserstein distance and gradient penalty.

Type I Experiments: InfoGANS

The theoretical advantage of InfoGANS is that their MC estimate of the objective is differentiable with respect to it's parameters even for discrete codes c.
In the experiments, the authors attempt to use this advantage to model properly discrete factors but the model struggles to learn the continuous factors. The previous reflected on low disentangling scores.

Type II Experiments

On datasets without known factors, latent traversals is the only evaluation possible. Yet FactorVAE presents smaller reconstruction error than β-VAEs.

Type II Experiments

Latent Traversal on Chairs.

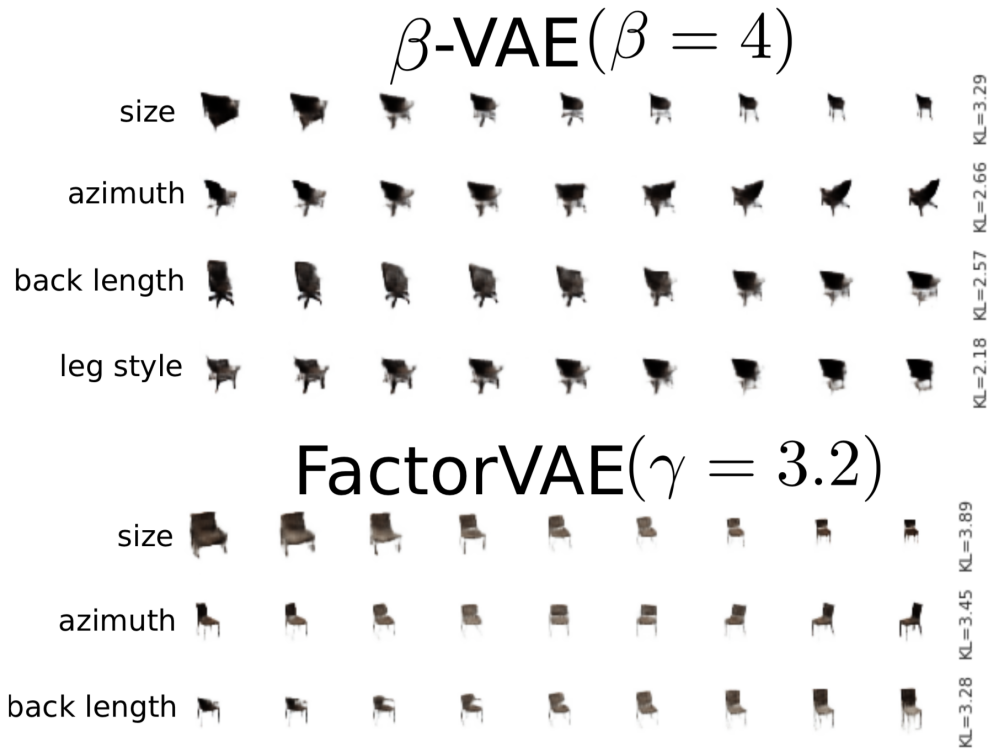

Figure 13. β-VAE and FactorVAE latent traversals across each latent dimension sorted by KL on 3D Chairs, with annotations of the factor of variation corresponding to each latent unit.

Type II Experiments

Latent Traversal on Faces.

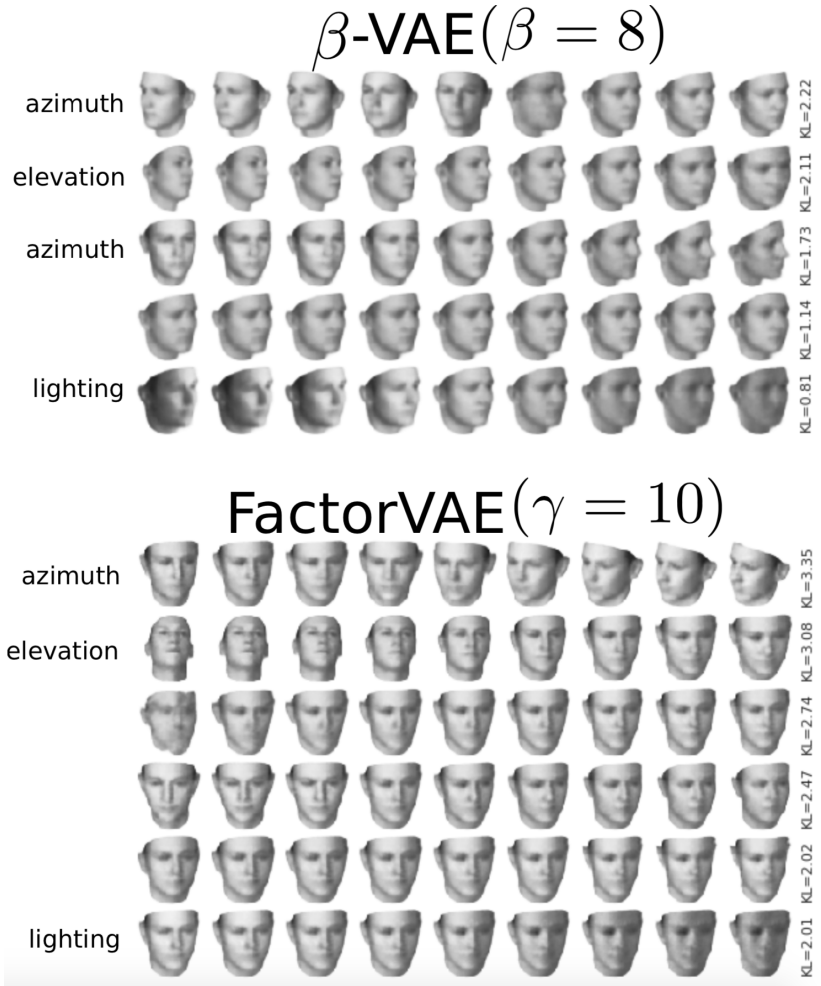

Figure: Source: Kim et al 2018[3]

[1] 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.

[2] 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. *ICLR*, 2(5):6, 2017.

[3] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.