Understanding Disentangling in VAE

Yann Dubois, Alexandros Kastanos, Dave Lines and Bart Melman

Department of Engineering, University of Cambridge

Introduction

Learning meaningful representations is crucial to improve generalisation, robustness and interpretability of machine learning models. Deep generative models such as variational autoencoders (VAEs) are a promising approach for learning such representations without supervision [1]. Various losses have been proposed to enforce *disentangling* [2, 3, 4, 5], such that the representation factorises into latent units that are sensitive to changes in single generative factors. We explore the benefits and differences of these losses by using a fixed architecture and various datasets. We propose a new metric to provide quantitative insights into the mechanisms at play.

Model

- Parametrised Gaussian posterior $q_\phi(\mathbf{z}|\mathbf{x})$ (encoder) and likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ (decoder), with a standard normal prior $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$
- Trained using reparameterisation trick: $\mathbf{z}(\mathbf{x}) = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

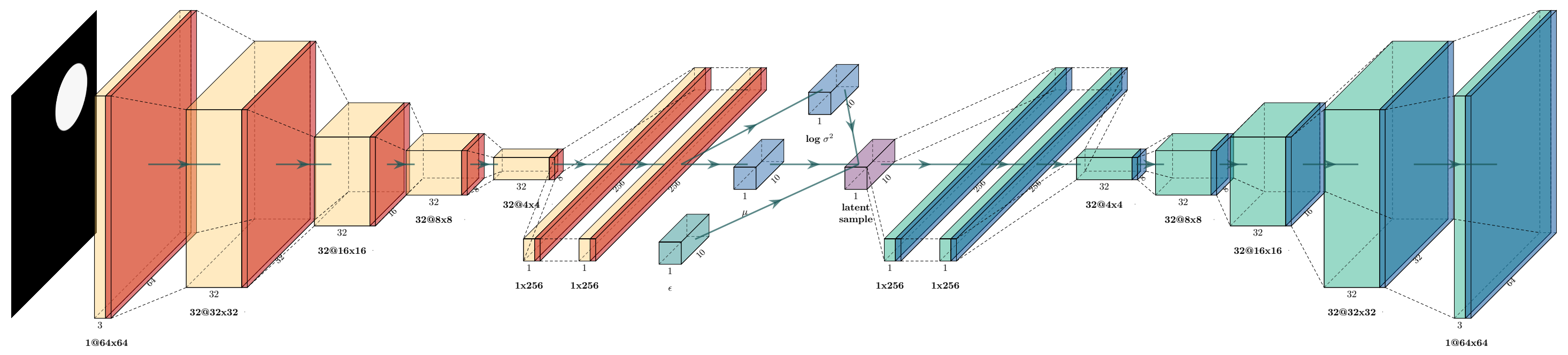


Fig. 1: VAE architecture

VAE's are trained to estimate the evidence $p(\mathbf{x})$ by minimising its (log) upper bound. Different losses balance the tightness of the bound and the amount of disentangling:

$$\underbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction Error}} + \underbrace{\alpha I_q[\mathbf{z};\mathbf{x}]}_{\textcircled{i}\text{ Index-Code MI}} + \underbrace{\beta \mathrm{KL}\left[q(\mathbf{z}) \| \prod_j q(z_j)\right]}_{\textcircled{ii}\text{ Total Correlation}} + \underbrace{\gamma \sum_j \mathrm{KL}\left[q(z_j) \| p(z_j)\right]}_{\textcircled{iii}\text{ Dimension-wise KL}} \quad (1)$$

- **VAE** [1]: $\beta = \gamma = \alpha = 1$ the tightest evidence lower bound for α, β, γ.
- β-**VAE**$_1$[2]: $\beta = \gamma = \alpha > 1$. Further penalising $\textcircled{i}$, $\textcircled{ii}$ and $\textcircled{iii}$, forces compression of $\mathbf{z}$ at the expense of reconstruction. β-**VAE**$_2$ [3], builds upon this, by only penalising the sum of $\textcircled{i}$, $\textcircled{ii}$ and $\textcircled{iii}$ once they deviate from a *capacity*, C.
- **Factor-VAE** [4]: $\gamma = \alpha = 1$, $\beta > 1$. Further penalising $\textcircled{ii}$ forces factorised $\mathbf{z}$.

References

[1] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[2] Irina Higgins et al. "beta-vae: Learning basic visual concepts with a constrained variational framework". In: *International Conference on Learning Representations*. 2017.

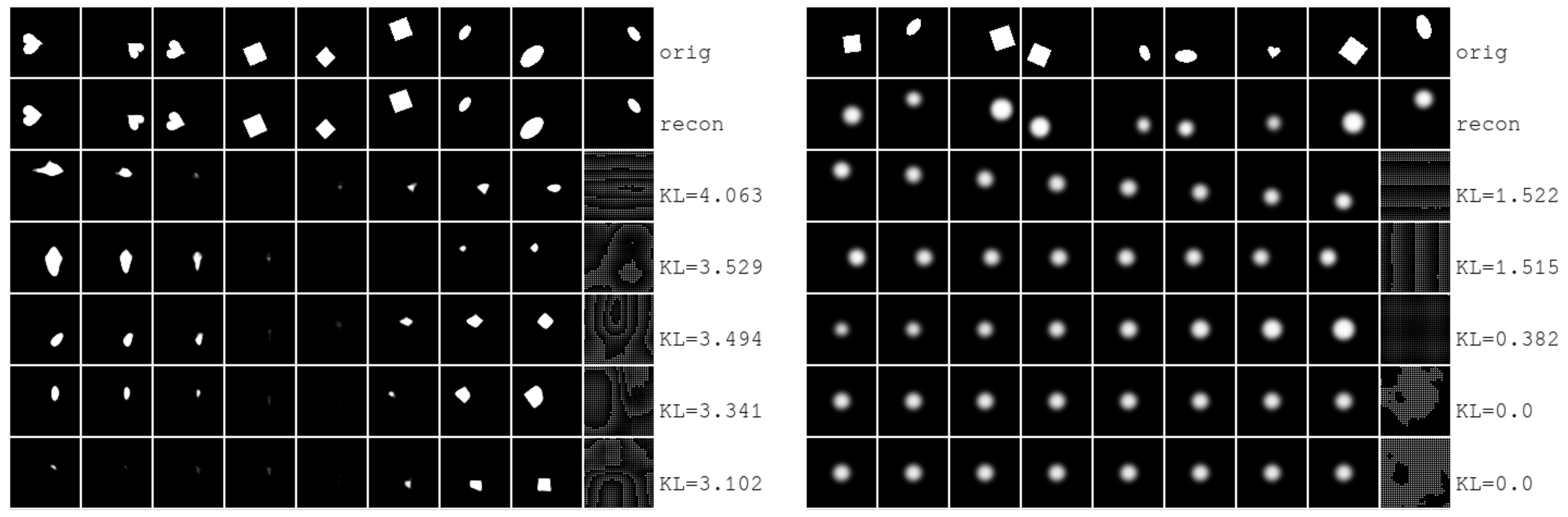
[3] Christopher P Burgess et al. "Understanding disentangling in β-VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[4] Hyunjik Kim and Andriy Mnih. "Disentangling by factorising". In: *arXiv preprint arXiv:1802.05983* (2018).

[5] Tian Qi Chen et al. "Isolating sources of disentanglement in variational autoencoders". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2615–2625.

Results

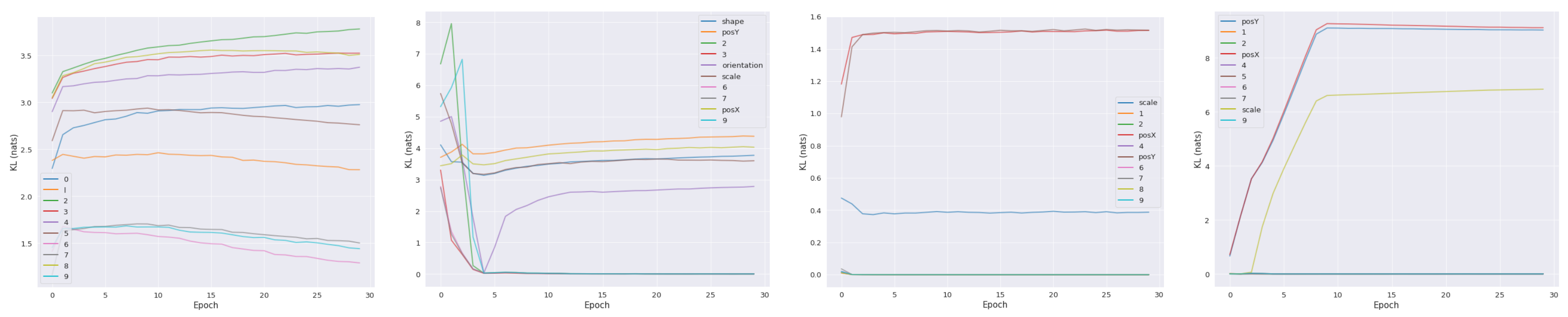
Figure 2 illustrates the β trade-off. VAE (Fig.2a) reconstructs nearly perfectly and generates sharp images from the prior, but does not disentangle factors of variation. $\beta-$VAE$_1$ (figure 2b) has poor reconstruction and generates blurry images, but uses an interpretable and factorised latent space. The heat maps indicate that the coordinates of the shape are disentangled for the entire dataset.



(a) VAE ($\beta = 1$) (b) β-VAE$_1$ ($\beta = 50$)

Fig. 2: Latent traversals from the posterior for dSprites. Top row: original images. Second row: corresponding reconstructions. Remaining rows: latent traversals sorted by their KL divergence to the prior. Right column: heat map showing the value of $\mathbf{z}$ for different positions of the shape.

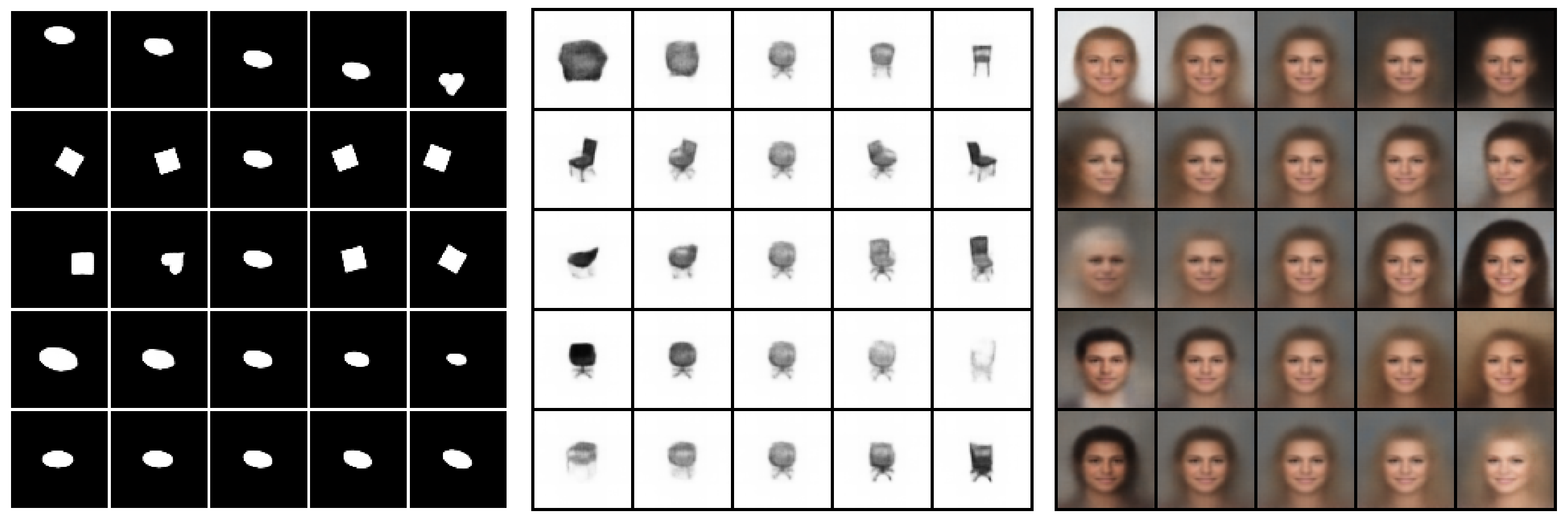
Figure 3 shows that VAE uses all latent dimensions without disentangling, while $\beta-$VAE only focuses on a few factor of variations. Factor-VAE is able to only use the correct number of latent dimensions.



(a) VAE (b) Factor-VAE (c) β-VAE$_1$ (d) β-VAE$_2$

Fig. 3: Dimension-wise KL divergence. The legend is manually set to the true factor of variation in case of strong qualitative evidence from the latent traversals.

Figure 4 shows that Factor-VAE learns disentangled representations with higher reconstruction accuracy by only increasing the regularisation of $\textcircled{iii}$, forcing factorised representations .



(a) Factor-VAE dSprites (b) Factor-VAE Chairs (c) Factor-VAE CelebA

Fig. 4: Latent traversals from the prior, dimensions are sorted by their KL and truncated to 5.

Metric

Performance of deep generative models is often measured by qualitative inspection of reconstructed images and their latent traversals (Fig.4). Following [5, 2, 4] we use quantitative measures of disentanglement based on the ground truth factors of variation $\{v_k\}_{i=1}^K$ and the latent dimensions $\{z_j\}_{j=1}^D$ to understand and compare models reliably. Specifically, we propose a metric to quantify axis alignment (Eq.2), and our modified version of mutual information gap [5] (Eq.3):

$$\mathrm{AAM}[\mathbf{v};\mathbf{z}] = \frac{1}{K}\sum_{k=1}^{K} \frac{\max\left(\max_j I_{x_n}\left[z_j; v_k\right] - \sum_{j'=1}^{D-1} I_{x_n}\left[z_j; v_k\right]_{(j')}, 0\right)}{\max_j I_{x_n}\left[z_j; v_k\right]} \quad (2)$$

$$\mathrm{MMIG}[\mathbf{v};\mathbf{z}] = \frac{1}{K}\sum_{k=1}^{K} \frac{I_{x_n}\left[z_j; v_k\right]_{(D)} - I_{x_n}\left[z_j; v_k\right]_{(D-1)}}{H\left[v_k\right] - \max_{k'} I[v_k; v_{k'}]} \quad (3)$$

Where the subscript (D) denotes the D^{th} order statistic and $I_{x_n}\left[z_j; v_k\right] = \mathbb{E}_{q(z_j, v_k)}\left[\log \sum_{x_n} q\left(z_j|x_n\right) p(x_n|v_k)\right] + H\left[z_j\right]$ is estimated using empirical distributions and stratified sampling over $p(v_k)$ and $p(x_n)$.

Table 1: Quantitative model comparison on dSprites.

Model	AAM	MMIG	KL	Log Like.	Total Loss	Var. Dim. KL
VAE	0.20	7.0e-4	0.025	-0.001	0.034	1.5e-5
$\beta -$ **VAE$_1$** ($\beta = 4$)	0.65	3.3e-2	0.025	-0.015	0.086	1.1e-6
$\beta -$ **VAE$_1$** ($\beta = 50$)	0.08	4.0e-4	0.003	-0.140	0.316	3.6e-7
$\beta -$ **VAE$_2$**	0.12	2.0e-4	0.025	-0.096	0.108	1.5e-5
FactorVAE	0.87	3.3e-2	0.028	-0.036	0.064	5.7e-6

A small increase from $\beta = 1$ to $\beta = 4$ increases AAM (Table 1) due to the regularisation of $\textcircled{ii}$. A large increase from $\beta = 4$ to $\beta = 50$ decreases AAM due to $\textcircled{iii}$ which penalises $\left(\mathbb{E}[q(z_j)] - \mathbb{E}[p(z_j)]\right)^2$. As a result, the model is forced to use multiple dimensions to encode a single factor of variation to obtain a smaller variance of dimension-wise KL. FactorVAE only increases the regularisation of $\textcircled{ii}$, enabling it to have both a large likelihood and AAM.

Conclusion

A variety of VAE objectives were explored under a fixed model architecture. Increased regularisation of terms $\textcircled{i}$, $\textcircled{ii}$ and $\textcircled{iii}$ for β-VAE encouraged greater disentangled representations of the generative factors compared to that of standard VAE but at the cost of reconstruction accuracy. We confirm the Factor-VAE postulate that total correlation $\textcircled{ii}$ is responsible for independence in the latent distribution and that penalising all three terms reduces the amount of data information stored in the latent representation. Indeed for Factor-VAE, we achieved both higher levels of disentanglement and higher reconstruction accuracy compared to β-VAE. Our proposed metric, AAM provides a quantitative evaluation of axis alignment between the latent representation and the generative factors of the data.