Understanding Disentangled Representation in Variational Autoencoders

Viet Nhat Nguyen, Hao Qiu
DISI, University of Trento
Via Sommarive, 9, 38123 Povo, Trento TN, Italy
{vietnhat.nguyen, hao.qiu}@studenti.unitn.it

Abstract

In this project we study about disentangled representation of images by using Variation Autoencoders framework with different settings. We compare and analyse different variations with different settings including our ideas. Finally we assess the disentanglement both qualitatively and quantitatively. Our results show that this framework is able to disentangle objects having simple structures.

1. Introduction

Disentangling underlying generative factors is one of the desired characteristics of feature learning that Machine Learning community have been wanted to solve for a long time. In this problem, we assume that our data is generated and controlled by some independent factors and we want a neural network that can represent each single factor by one neuron. For example, we wish to have a representation in which one dimension corresponds to one property of the object in image such that angle, color, thickness, background, style, and so forth.

There are some works trying to address this learning problem. One of the first papers is DC-IGN [7] which tried to learn graphic codes of object in images but it was semi-supervised and required special dataset.

King ma et al. [6] introduced Variational Autoencoder (**VAE**) framework for generative model at first but then it is surprisingly suitable for disentangling task because it allows the smoothness and independence in latent space. β-VAE [3] is a popular modified version of VAE with a modification on loss function to archive better disentanglement. One drawback of β-VAE is that reconstruction quality (compared to VAE) must be sacrificed in order to obtain better disentangling. Following β-VAE, [3] found that an annealing training manner could lead to higher quality reconstruction images.

InfoGAN [2] is another popular approach using GANs framework. However, this model is difficult to converge as a common problem of GANs.

In this project, we do the following works: 1) We study and implement VAE and β-VAE. 2) We also implement the proposed method to promotes robust learning of disentangled representation combined with better reconstruction fidelity. 3) We try our ideas about training with conditional VAE and slight change in loss function. 4) We give quantitative and qualitative comparisons of VAE, β-VAE for disentanglement.

The rest of this report is organized as follows: in section 2 we will briefly present VAE and further improvement, respectively original VAE, β-VAE, annealing way to VAE train network, and our propose using conditional VAE. Experimental validation will be showed in section 3, including both qualitative and quantitative results. Finally we conclude what we have studied in section 4.

2. Proposed Methods

2.1. Variational Autoencoder (VAE)

Suppose we have a sample $\mathbf{x}$ of samples from a distribution parameterised by ground truth generative factors $\mathbf{z}$. VAE aims to learn the marginal likelihood of the data in such a generative process:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \tag{1}$$

where ϕ, θ parameterise the distributions of the VAE encoder and the decoder respectively. This can be re-written as:

$$\log p_\theta(\mathbf{x}|\mathbf{z}) = D_{KL}\big(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big) + \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) \tag{2}$$

where $D_{KL}(\cdot\|\cdot)$ stands for the non-negative Kullback-Leibler divergence between the true and the approximate posterior. Hence, maximising $\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z})$ is equivalent to maximising the lower bound to the true objective in Eq. 1:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big) \tag{3}$$

2.2. β-VAE

β-VAE is a modification of the VAE framework, which introduces an adjustable hyperparameter β to the original

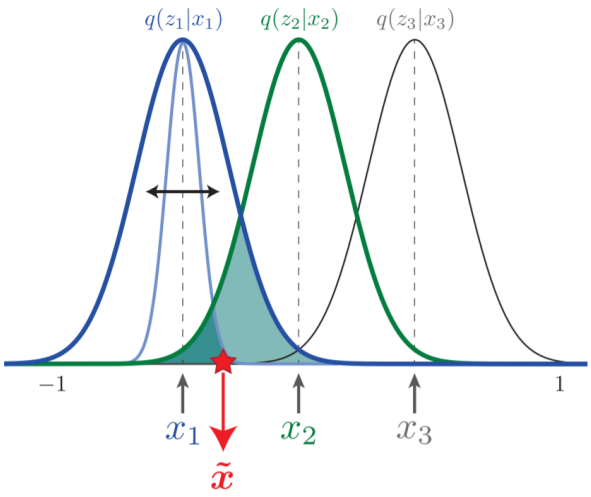

Figure 1. An illustration for the situation of increasing value of β. The sample $\tilde{x}$ is sampled from $q(z_2|x_2)$. Initially β is small, there is little relationship between this sample and the posterior $q(z_2|x_2)$ of sample x_1. But when we increase β, the variance of $q(z_2|x_2)$ is widen. Now the sample $\tilde{x}$ is likely drawn from both posterior distributions of x_1 and x_2. Consequently, to lower the reconstruction error, the network is forced to arrange nearby points in data space close together in the latent space. This figure is taken from Burgess et al. [1].

VAE objective:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \quad (4)$$

with β is usually set to be ≥ 1.

We can see β-VAE as a general version of standard VAE. When $\beta = 1$, it corresponds to the original framework. When $\beta > 1$, this constraint imposes a limit on the capacity of the latent information channel and control the emphasis on learning statistically independent latent factors. The larger β is, the more posterior $q\left(z_i|\mathbf{x}\right)$ is encouraged to match the unit Gaussian prior $p(z_i) = \mathcal{N}(0, 1)$. In order to reduce the KL divergence, the network has to broaden the posterior distributions and move their means close together (and closer to 0). As a result of this effects, the overlaps between these posteriors are increased, which leads to more smoothness in the latent space. Figure 1 illustrates this situation. Since we have assumed that our data is generated by independent factors, the best way for the network to reconstruct dataset while maintaining high smoothness is disentangling representation in latent neurons.

2.3. An annealing method to train VAE

Although β-VAE is able to learn a disentangled representation, the reconstructed images are usually blurry due to the increase of regularization term. We implement the annealing method proposed in [1] to train VAE, that explicitly forces the KL divergence between posterior and marginal distribution equal to $C \geq 0$ in which C is linearly increased from 0 to C_{max} during training progress. From information bottleneck perspective, C controls how much information are allowed to pass throughout the network and is measured by natural unit of information (nat). When C is small, only factors with large variation to be learn, which means reconstructions are also blurry. By gradually adding more latent encoding capacity, the network is enable to learn progressively more factors of variation while retaining the disentangling from previously learned factors. This phenomenon allows adding more details to reconstruction and predictably enhances reconstruction quality as well as disentanglement. Specifically, the Variational Lower Bound on marginal log-likelihood of data now turns to be:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, C) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \gamma\, |D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) - C| \quad (5)$$

Since the absolute value function in Eq. 5 is quite strict, we replace it by a ReLU function in our implementation. The intuition is that we should only penalize the KL divergence when this term excesses C. We found that this modification is beneficial for the sake of optimization:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, C) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \gamma\, \text{ReLU}\left((D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) - C)\right) \quad (6)$$

2.4. Conditional VAE

In this problem, one assumes that the dataset is generated by independent factors that could be represented as continuous numbers. However, in practice, except datasets which are artificially created, the underlying factors could be dependent or discrete. For this reason, we make our generator be conditioned on categorical information. In other words, we concatenate the label of classes with the latent vector then feed it into generator. By doing this, the network is encouraged to focus on continuous factors therefore gets more chance to learn a disentangled representation.

3. Experiments

3.1. Setting

Datasets We trained VAE and β-VAE on the **MNIST** [1] and **dSprites** [8] datasets. The dSprites is a dataset of 2D shapes procedurally generated from 6 ground truth independent latent factors. These factors are color, shape, scale, rotation, x and y positions of a sprite. Each factor controls one feature of image independently.

Model Architecture. VAE and β-VAE both utilised same basic architecture for experiments. The encoder for the VAEs consists of 4 convolutional layers, using batch normalisation after each convolution layer. The size of latent vectors is 10. The decoder architecture is simply symmetric to the encoder. All the activation functions in hidden layers are ReLU while it is Sigmoid for the output layer.

[1] Available at http://yann.lecun.com/exdb/mnist/

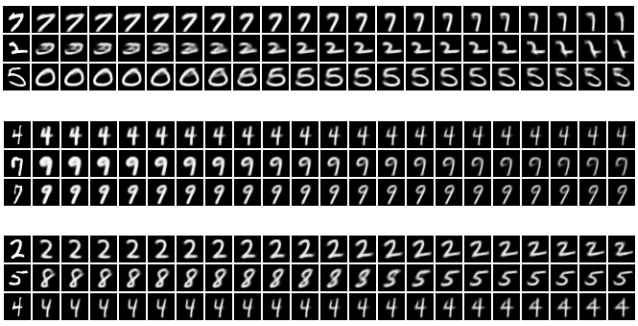

Figure 2. Latent travesal from -3 to +3 in β-VAE model (zoom in for better observation). The first column contains original images. From top to bottom: azimuth, thickness and height of digits are changed, corresponding to latent traversal on 3 neurons. This is also applied for the other figures in this section.

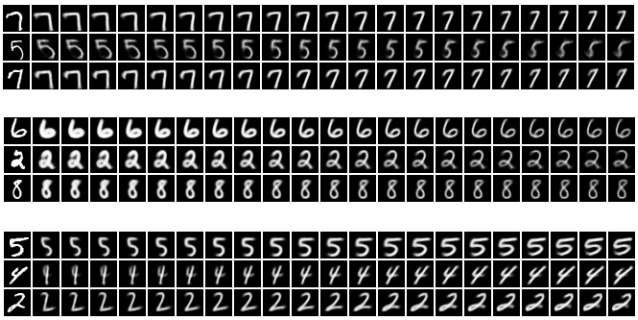

Figure 3. Latent travesal from -3 to +3 in β-VAE model with label information.

Training We used Adam [4] as the optimiser with a fixed learning rate of 5e-4. The β value was set to 4 for β-VAE and C was gradually increased to 20 (nats) in 20 epochs when using the annealing method. The reconstruction loss was binary cross entropy and summed up over all pixels in images. All the networks were trained in 100 epochs with batch size 64.

3.2. Qualitative results.

One of the popular way to assess the disentanglement is varying the value of one dimension while fixing all the others in the latent representation. Figures 2, 3, 4 respectively provide reconstructed images after doing latent traversal on neurons which have significant KL divergence with $\mathcal{N}(0, 1)$ in β-VAE model, β-VAE with label information and annealing VAE model with label information. Generally, all the models could learn some independent factors of digits. β-VAE model was still entangled and becomes disentangled when we fed the information about class label. However, the model trained with annealing method gave the best qualitative result with better reconstruction and discovers more factors, namely azimuth, thickness, width and height of digits.

In term of dSprites dataset, our model trained with annealing method could learn four out of five underlying factors. Figure 5 shows the latent traversal. Position and size of object were learned successfully while azimuth was still be entangled with shape of object. This is due to the fact that azimuth is the factor that has smallest variance therefore it is not sensitive to small change of input.

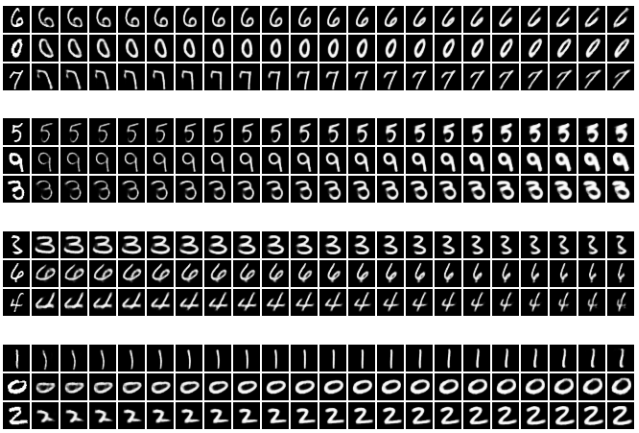

Figure 4. Latent travesal from -3 to +3 in VAE model trained with annealing method with label information.

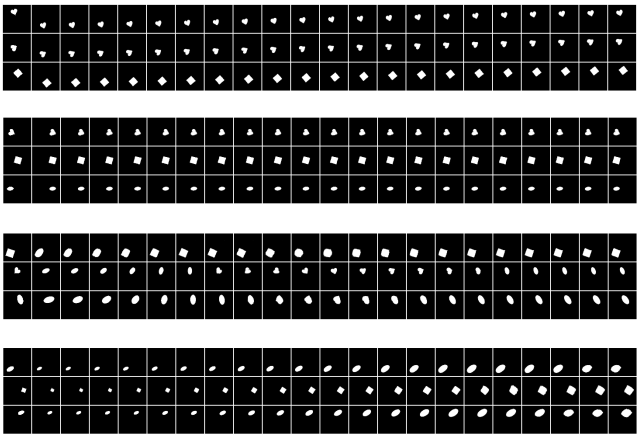

Figure 5. Latent travesal from -3 to +3 in VAE model trained with annealing method with label information.

3.3. Quantitative results

3.3.1 Quality of reconstruction images

Let us first investigate quantitatively the quality of the reconstruction images of different methods. We use the Laplacian metric proposed in [9], which is widely used for auto focus assessment. This method simply convolves the image with a Laplacian operator to measure the second derivative of an image, then compute the variance as a sharpness score. If this score is low, then there is a tiny spread of responses, indicating there are very little edges in the image. As we know, the more an image is blurred, the less edges there are. We report this score and the recon-

Table 1. The reconstruction loss and sharpness score on MNIST dataset with different methods. The lower the sharpness score is, the blurrier the reconstructions are.

Methods	Reconstruction loss	Sharpness score
Standard VAE	76.72	2159
$\beta-$VAE ($\beta = 4$)	99.04	1231
VAE (anneal) - abs	74.32	2318
VAE (anneal) - ReLU	**73.94**	**2343**

Table 2. The reconstruction loss and disentanglement score on dSprites dataset with different methods.

Methods	Reconstruction loss	Disentanglement score
Standard VAE	25.98	0.5476
$\beta-$VAE ($\beta = 4$)	51.54	0.7502
VAE (anneal) - abs	12.97	**0.8232**
VAE (anneal) - ReLU	**12.07**	0.8227

struction loss as well for MNIST dataset in table 1.

3.3.2 Disentanglement metric

Up to this point, we have assessed the quality of disentanglement merely by looking at latent traversal of reconstruction images. It is important to be able to compare the level of disentanglement achieved by different models. There are some metrics proposed to address this problem. In this project we followed the metric described in [3], which attempts to measures both the independence and interpretability of the inferred latents. This metric is based on an intuition that if two images share one property then the corresponding neuron in latent of them should be identical. Formally, this metric is calculated as following steps:

- Choose a random factor k uniformly.
- Sample two sets of latent representations $\mathbf{u}_{1..L}$ and $\mathbf{v}_{1..L}$ such as $[u_l]_k = [v_l]_k$.
- Simulate images $\mathbf{x}_u$ and $\mathbf{x}_v$ from $\mathbf{u}$ and $\mathbf{v}$. Use encoder to infer $\mathbf{z}_v$ and $\mathbf{z}_u$ respectively.
- Compute the difference $z^l_{diff} = |z_{u,l} - z_{v,l}|$ then use $\mathbf{z}_{diff} = \sum_{l=1}^{L} z^l_{diff}$ to predict factor k and report accuracy as disentanglement score.

Figure 6 demonstrates how to compute the metric. Because we use the true latent representation to simulate images, this metric could only be used to measure on dSprite dataset. In our setting, we generated 10000 samples then chose a SVM as the classifier for stability. The scores obtained from our models are reported in table 2. We can see that standard VAE has better reconstruction but not good at disentangling as β-VAE.

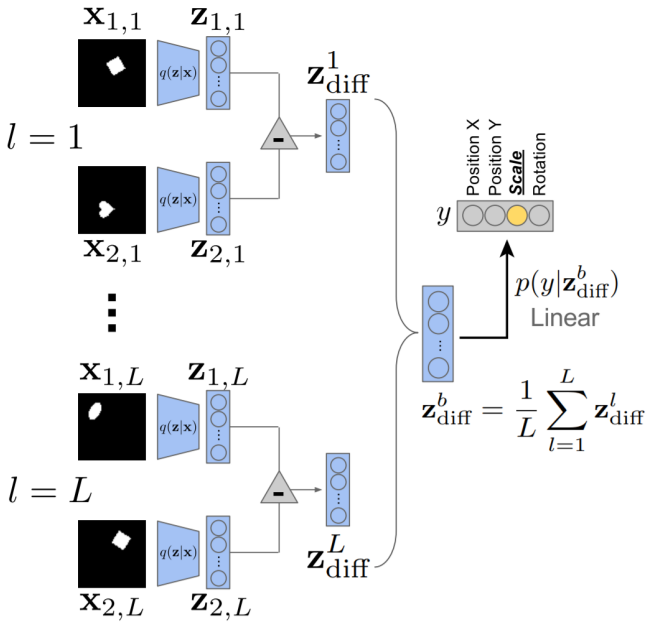

Figure 6. Illustration of procedure to calculate disentangling metric. Image is taken from [3].

4. Conclusions

In this project, we have studied and implemented the original VAE [5] and β-VAE [3]. In particular, we have compared the disentangling performance of different models qualitatively and quantitatively on the MNIST and dSprites. We also used two methods to improve disentangled representation in VAE: one is an annealing method; the other one is adding label information inside latent variables. Our results showed that both methods can promote the performance of image quality and disentangling.

References

[1] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[2] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016.

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[7] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

[8] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[9] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *ICPR*, 2000.