Is Continuous Bottleneck Pressure on Variational Autoencoder Critical for Disentanglement?

2019122076 이상규

I. Introduction

Variational Autoencoder[1](VAE)는 Generative Model로서 성공적인 approach 중 하나이다. VAE는 intractable한 data $\mathbf{x}$에 대한 log likelihood $\log p(\mathbf{x})$를 Autoencoder(AE) architecture를 응용하여 approximate한다. 이는 $\log p(\mathbf{x})$를 maximize하는 것을 variational inference의 evidence lower bound(ELBO) technique와 reparameterization trick을 통해서 tractable하게 다룰 수 있음을 활용한다. VAE는 다른 Generative Model보다 stable한 convergence를 보이면서 latent variable이 상대적으로 control하기 쉽다는 점에서 장점을 보인다.

$$\mathcal{L}(\phi, \theta; \mathbf{x}, \mathbf{z}) = -E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

prior research로서 VAE의 ELBO object를 constrained optimization problem으로 해석하는 시도가 있다. 이는 VAE가 AE의 architecture가 기본적으로 수행하는 reconstruction loss에 posterior와 prior의 KL divergence가 constraint로 적용되는 것이라는 접근이다. 단, 이렇게 얻은 object는 Lagrangian function으로 나타나기 때문에 multiplier β에 따라서 더 이상 ELBO를 maximize한다는 theoretical evidence가 적용될 수 없다.

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

$$\rightarrow \mathcal{L}(\phi, \theta, \beta; \mathbf{x}, \mathbf{z}) = -E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

그러나 β라는 새로운 hyperparameter를 조정하는 과정에서 추가적인 merits를 얻을 수 있다는 것이 관찰되었다. 대표적으로 [2], [3]은 VAE의 Encoder, Decoder에 sequential model을 사용하였을 때 발생하는 posterior collapse을 이러한 approach로 개선했다. 해당 problem은 sequential model이 data property를 memorize하여 latent variable의 역할이 상실되는 현상을 말한다. 해당 연구들은 KL divergence term의 multiplier를 0에서 1 사이에서 적절히 scheduling 하는 방법을 통해서 이러한 problem을 개선할 수 있음을 보여준다.

또 다른 prior research로서 disentanglement를 달성하기 위해 이러한 approach를 택하는 경우가 있다. β-VAE[4]는 high value β를 통해서 disentanglement를 달성할 수 있음을 보여준 대표적인 research이다. 해당 approach는 Information Bottleneck[5]를 theoretical evidence로 사용한다. 이는 task Y를 input X로 표현하기 위해 최대한으로 압축된 signal Z를 찾는 방법이다. 이에 근거해 [4]는 KL divergence term을 regularize하는 것이 X에 대한 Z의 효율적인 encoding을 학습하도록 유도할 수 있다고 보였다.

$$\max\left[(Z;Y) - \beta I(X;Z)\right]$$

해당 research는 object의 property를 결정하는 factor가 reconstruction에 미치는 영향이 다르다는 hypothesis가 사용되었다. 예를 들어 object의 position은 shape과 같은 factor에 비해 상대적으로 많은 pixel의 변화를 가져온다. 이는 KL divergence라는 bottleneck capacity가 제한되는 상황에서 reconstruction loss를 최소화하기 위해서는 최대한 position이라는 factor의 information이 먼저 posterior에 담겨야 함을 의미한다.

이러한 capacity와 disentanglement factor 측면에서 β-VAE는 empirical하게 좋은 결과를 보여주었지만 동시에 좋지 못한 reconstruction을 보여주었다. 이를 개선하기 위해서 해당 KL divergence term의 capacity를 training 진행 시에 증가시키는 방법이 제안되었다[6]. 그러나 이에 대한 해석에서 theoretical evidence와 완전히 일관적이지 않은 사항이 있다. 본 project에서는 bottleneck capacity의 증가와 disentanglement와의 관계를 조금 더 깊이 이해하기 위하여 β-VAE variant를 활용하여 experiment를 진행하였다.

II. Hypothesis

[6]은 다음과 같은 이유로 KL divergence term에 강한 pressure를 주는 것이 disentanglement를 달성하는 데 도움을 준다고 주장한다. KL divergence에 high pressure가 발생하면 posterior는 각 data $\mathbf{x}$마다 large variance 또는 localized mean을 갖도록 유도된다. 동시에 reconstruction을 위해서 data $\mathbf{x}$마다 Decoder가 구분하기 쉬운 posterior를 갖도록 유도된다. 그러나 VAE는 unit Gaussian prior와 reparameterization trick을 사용하기 때문에 posterior가 diagonal covariance matrix를 갖는다. 이로 인해 latent channel의 combination으로 표현되는 axis에 대해서는 factor encoding이 수행될 수 없다. 따라서 각 latent channel에 align하도록 factor는 담기게 되는데 각 factor를 다른 channel에 encoding하는 것이 가장 KL divergence term을 최소화 할 수 있다.

따라서 [6]은 각 factor를 차례대로 담기에 충분한 KL divergence capacity를 보장하기 위해서 새로운 parameter C를 도입한다. 충분히 하나의 factor가 encoding으로서 담겼다면 조금 더 reconstruction에도 최적화되도록 KL divergence term에 대한 bottleneck pressure를 완화시키는 방법이다. 이는 주어지는 C의 maximal value까지 일정한 step으로 linearly increasing하는 방법으로 구현되었다.

$$\mathcal{L}(\phi, \theta, \beta, C;\ \mathbf{x}, \mathbf{z}) = -E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta\ (\mathbf{x}|\mathbf{z})] + \beta|D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|$$

이렇게 bottleneck capacity를 조절하는 방법을 통해서 reconstruction과 disentanglement를 동시에 달성할 수 있다는 것을 [6]은 보였다. 그러나 FactorVAE[7]은 이렇게 KL divergence term을 regularize하여 disentanglement를 달성하는 것은 부적절한 approach라고 주장한다. 따라서 FactorVAE는 marginal posterior를 approximate하는 neural network를 추가하여 posterior가 factorize되도록 유도한다. 이는 VAE object의 KL divergence term에 대해서 다음과 같이 decomposition이 가능하다는 observation을 활용한다.

$$E_{\mathbf{x}\sim\mathcal{D}}[D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] = I_q(\mathbf{x};\mathbf{z}) + D_{KL}(q(\mathbf{z})||p(\mathbf{z}))$$

이 때 $D_{KL}(q(\mathbf{z})||p(\mathbf{z}))$ term은 prior $p(\mathbf{z})$가 factorized prior이기 때문에 marginal posterior $q(\mathbf{z})$가 factorized된 형태일 때 최적화된다. 따라서 해당 term을 regularize하는 것은 disentanglement에 긍정적인 영향을 줄 것이다. 그러나 mutual information $I_q(\mathbf{x};\mathbf{z})$ term은 그렇지 않다. 이를 regularize하는 것은 결과적으로 $\mathbf{z}$가 $\mathbf{x}$의 information을 잘 encoding하지 못하는 결과를 낳을 수 있다. 실제로 [8]과 같은 approach는 이러한 mutual information을 maximize하는 것으로 disentanglement를 달성하였다. 따라서 두 term을 동시에 regularize 하는 것은 disentanglement를 달성하는데 효과적인 approach가 아닐 수 있다는 것이다.

해당 approach로 reconstruction과 disentanglement이 동시에 달성된 현상을 통해 parameter C가 수행하는 역할에 대해 이해할 수 있다. increasing하는 C는 $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ term에서 mutual information $I_q(\mathbf{x};\mathbf{z})$ term에 대한 보상의 역할을 하는 것이다. model은 training을 하는 과정에서 $\mathbf{z}$에 $\mathbf{x}$에 대한 information을 담을 수 밖에 없기 때문에 반드시 $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$은 증가하게 된다. 따라서 실제 latent z가 유의미한 information을 담은 만큼의 capacity는 C를 통해 regularize하는 quantity에서 보상하여 reconstruction에도 효율적인 encoding을 찾을 수 있게 유도하는 것이다.

이러한 observation을 바탕으로 다시 C를 보면 단순히 linearly increasing하는 것은 적절치 못할 수 있다. C는 현재 model이 $\mathbf{z}$로서 표현하고 있는 $\mathbf{x}$에 대한 information 만큼을 regularize하는 것에서 제외해주어야 하는 것이다. 때문에 C를 설정할 때 현재 model의 $I_q(\mathbf{x};\mathbf{z})$의 quantity가 반영되는 것이 이상적이다. 그러나 $I_q(\mathbf{x};\mathbf{z})$는 기본적으로 intractable하고 [7]에서 Monte Carlo Sampling으로 approximate 되기 어려웠다는 것이 언급되어 있다. 따라서 FactorVAE는 이를 완전히 회피하는 approach를 제시했던 것이다.

그러나 여전히 FactorVAE가 제시한 방법이 아니라 기존의 parameter β를 well-scheduling하는 방법으로

더 좋은 reconstruction과 안정적인 disentanglement를 달성한 ControlVAE[9]가 있다. ControlVAE는 training step마다 발생하는 KL divergence term의 변화를 PID controller[10]를 통해 step마다 변하는 적절한 $\beta(t)$를 선택하는 approach이다. KL divergence term이 목표로 하는 capacity보다 높다면 높게, 낮다면 낮게 $\beta(t)$를 설정하는 것을 통해서 적절한 regularization을 수행한다.

이렇게 time step마다 변화하는 $\beta(t)$는 FactorVAE의 approach보다 안정적으로 disentanglement를 달성하면서 높은 reconstruction을 보여주었다. 그러나 여전히 목표로 하는 capacity C는 β-VAE에서 사용된 방법과 동일하게 linearly increasing 되는 방식으로 수행하였다. 즉, FactorVAE에서의 observation에서 얻을 수 있었던 현재 model의 mutual information $I_q(\mathbf{x};\mathbf{z})$의 state를 반영하지 않는다.

그러나 [4]의 hypothesis를 생각해보면 이렇게 linear increasing하는 capacity C는 Information Bottleneck 측면에서 부적절하다. 만약 현재의 mutual information $I_q(\mathbf{x};\mathbf{z})$이 목표로 하는 capacity C에 비해 낮은 상황에서 더 많은 capacity를 보상하게 된다면 model은 disentanglement를 위한 효율적인 encoding을 찾도록 유도되지 않는다. capacity를 만족할 수 있도록 최대한 많이 factor에 대한 information을 담으려고 유도될 뿐 latent channel과 align되는 encoding을 수행하려는 pressure는 완화된다.

하지만 PID controller를 통해서 KL divergence term을 tracking하는 ControlVAE는 이러한 문제에서 조금 더 자유롭게 접근할 수 있다. capacity를 increasing 시킬 때 단순히 step마다 linear하게 증가시키는 것이 아니라 PID controller의 결과를 통해 capacity를 increasing 시킬 것인지 결정할 수 있다. 만약 현재의 capacity로 충분한 pressure가 발생했다고 추론할 수 있으면 capacity를 증가시킨다. 그렇지 않다면 현재의 capacity를 유지하여 model이 Information Bottleneck에 따라 효율적인 encoding을 하도록 유도한다. 만약 정말 disentanglement가 달성되는 evidence가 Information Bottleneck이라면 이렇게 time step 별로 model의 state를 반영하는 capacity bottleneck이 더 좋은 disentanglement를 달성할 수 있을 것이다.

III. Experiments

앞서 언급된 hypothesis를 검증하기 위해서 ControlVAE를 활용하여 capacity를 조절하는 방법에 따라 disentanglement와 reconstruction의 변화를 비교하는 experiment를 진행하였다. 이를 위해 disentanglement를 학습하기 위해 주로 사용되는 dSprites[11] dataset을 이용하였다. 이는 [4]에서 직접 disentanglement를 검증하기 위한 dataset이다 보니 많은 research가 disentanglement에 대한 reference로 사용하기 때문에 선정하였다. 특히 disentanglement를 수치적으로 계산할 수 있도록 Mutual Information Gap[12](MIG)라는 metric이 존재한다는 점에서 disentanglement quality의 비교를 위해 적절한 dataset이라고 할 수 있다.

model의 state를 반영하는 capacity를 위해서는 다음과 같은 방법을 사용하였다. initial value $C(0)$로 시작해서 매 step마다 PID controller가 반환하는 $\beta(t)$의 변화를 바탕으로 $C(t)$를 결정한다. 만약 $\beta(t)$가 이전 step에 비해서 증가했다면 아직 capacity를 달성하지 못했다는 것을 의미하므로 $C(t)$를 유지한다. 그렇지 않다면 남은 iteration을 바탕으로 linearly increasing 됐을 때 목표로 하는 maximal capacity $C_{\max}$가 달성될 수 있는 step size만큼을 반영하여 $C(t)$를 증가시킨다. 이렇게 capacity를 regularize 할 경우 linearly increasing 하는 capacity보다 주어지는 capacity level에 대해 지속적인 pressure가 발생할 수 있다.

$$C(t) = \begin{cases} C(t-1) & \text{if } \beta(t) > \beta(t-1) \\ C(t-1) + \dfrac{C_{\max} - C(t-1)}{t_{\max} - (t-1)} & \text{otherwise} \end{cases}$$

이렇게 hypothesis를 검증하기 위해 제안된 capacity와 비교를 위한 linearly increasing capacity는 두 경우에서 모두 $C(0) = 0.5, C_{\max} = 16$ 으로 설정되어 학습되었다. 마찬가지로 사용한 model architecture는 두

경우에서 모두 [4]와 동일하며 optimizer 또한 동일하게 Adam[13]을 사용하였다. 또한 PID controller에 대한 hyperparameter와 linear increasing capacity에 대한 rule은 [9]에서 제시된 것과 동일하다. 정확한 비교를 위해서 두 경우에서 같은 random seed로 64 mini-batch size로 1,500,000 iteration 동안 학습되었다.

다음의 Figure들은 두가지 case에서 학습이 진행되는 동안의 trace plot, 최종적인 model에 대한 reconstruction과 latent traversal result, MIG score와 final reconstruction loss에 대한 table이다. trace plot은 reconstruction loss, KL divergence loss, $\beta(t)$, $C(t)$ 에 대해서 두 model의 iteration의 증가에 따른 변화이다. reconstruction의 비교는 주어지는 original image와 이에 대한 reconstruction의 pair이다. latent traversal은 model의 latent dimension인 10개의 channel에 대해 -2부터 2까지 일정하게 변화시킨 output이다.

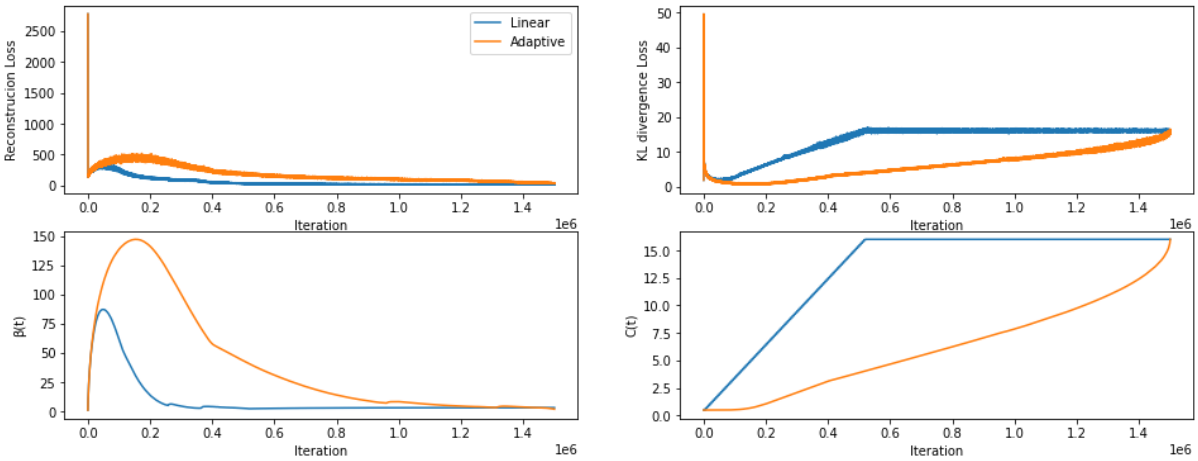

Figure 1: trace plot of reconstruction loss, KL divergence loss, $\beta(t)$, $C(t)$

Figure 2: reconstruction(lower) from given input(upper); Left = Linear, Right = Adaptive

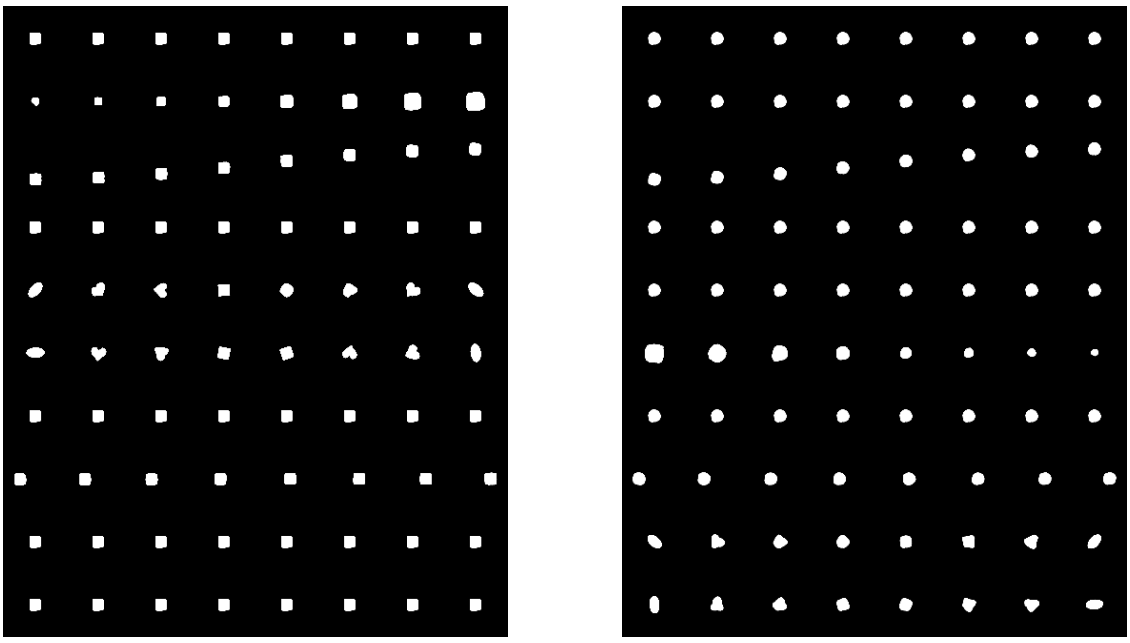

Figure 3: latent traversal for each latent channel; Left = Linear, Right = Adaptive

	MIG Score	Reconstruction Loss
Linear	0.5279	**18.2191**
Adaptive	0.5273	37.1294

Figure 4: MIG Score and Reconstruction Loss from each final model

IV. Discussion

먼저 hypothesis의 검증을 위해 제시한 $C(t)$의 scheduling은 어느 정도 의도한 방향으로 이루어졌음을 Figure 1을 통해서 확인할 수 있다. linearly increasing하는 경우와 달리 $\beta(t)$의 state에 따라 $C(t)$가 결정 될 경우 상대적으로 특정한 capacity level에 대해서 유지되려는 현상을 확인할 수 있다. 이는 최대한 적은 capacity를 통해 data $\mathbf{x}$를 encoding하는 latent variable $\mathbf{z}$를 찾도록 pressure를 가하는 환경이 유도됨을 의미한다. 이에 따라 $\beta(t)$ 또한 linearly increasing하는 경우보다 더 큰 값을 평균적으로 갖게 된다는 것도 확인할 수 있으며, 이는 더 많은 KL divergence term에 대한 pressure가 발생함을 의미한다.

따라서 Information Bottleneck과 [6]의 hypothesis에 따르면 더 지속적인 pressure가 가해지는 경우 더 많은 disentanglement가 달성되어야 한다. 그러나 Figure 3 및 Figure 4의 경우를 살펴보면 disentanglement에 긍정적인 영향을 주지 못했음을 알 수 있다. MIG score 상으로도 linearly increasing하는 경우와 차이가 거의 없었으며, latent traversal에서도 linearly increasing하는 경우보다 나은 모습을 보여주지 못했다. 예를 들어 capacity를 linearly increasing 시키는 경우 shape와 rotation이 entangling하는 모습을 보인다(channel 4, 5). 이는 adaptive increasing하는 경우 또한 관찰되는 현상이다(channel 8, 9). 오히려 reconstruction에서 adaptive한 경우는 더 떨어지는 현상이 관찰되는 것을 보면 theoretical evidence에 대해 재고가 필요해 보인다.

이러한 현상에 대해 practical한 경우에는 [4]에서 제시한 hypothesis가 Information Bottleneck이라는 theoretical evidence보다 disentanglement에 critical 하다는 것으로 해석 할 수 있다. 두 경우 모두 position, size와 같은 factor에 대해서는 disentanglement가 잘 달성되었지만, shape나 rotation과 같은 factor는 그렇지 못했다. 이 때 shape와 rotation과 같은 factor는 상대적으로 다른 factor에 비해 적은 pixel의 변화를 가져오므로 reconstruction에 주는 영향이 적다. 따라서 강한 bottleneck pressure가 가해져도 latent channel에 align하는 최적의 encoding을 찾도록 유도되기 어려울 수 있다.

이를 통해 [6]의 hypothesis 중 reconstruction을 위해서 data $\mathbf{x}$마다 Decoder가 구분하기 쉬운 posterior를 갖도록 유도되는 것이 disentanglement에 critical하다고 할 수 있다. 그러나 특정 factor의 reconstruction loss를 결정 짓는 magnitude가 강하지 못하다면 Encoder에게 entangling이 발생한다는 signal을 전달하기 어렵다. 따라서 bottleneck pressure가 지속적으로 가해지더라도 factor의 고유한 property에 따라 KL divergence term에 대한 regularization의 effectiveness가 다를 수 있는 것이다. 이는 본질적으로 KL divergence loss라는 frame만으로는 general domain의 disentanglement를 보장할 수 없다는 것을 의미한다.

이러한 결론은 VAE에서 더 좋은 disentanglement을 위한 follow-up study의 방향성을 제시해준다. KL divergence term에 존재하는 mutual information $I_q(\mathbf{x};\mathbf{z})$의 보다 좋은 approximation이 필요하다. 이를 통해서 marginal posterior가 factorize 되도록 유도하는 $D_{KL}(q(\mathbf{z})||p(\mathbf{z}))$ term이 직접적으로 regularize되도록 유도 해야 한다. FactorVAE와 같이 neural network를 통해 marginal posterior를 approximate하는 것은 그 오차를 피하기 어렵다. 또한 ControlVAE와 같이 marginal posterior의 factorization을 고려하지 않고 KL divergence term을 well-scheduling하는 것은 reconstruction에 critical하지 않은 factor에 대해서 disentanglement가 달성되지 않을 수 있다. 따라서 Monte Carlo Sampling이 잘 작동하지 않았던 mutual information $I_q(\mathbf{x};\mathbf{z})$에 대해서 다른 approach를 통해 좋은 approximation을 얻을 수 있도록 research가 필요하다고 판단된다.

V. Reference

[1] Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes”, arXiv, 1312.6114, 2013.

[2] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz and Samy Bengio, “Generating Sentences from a Continuous Space”, arXiv, 1511.06349, 2015.

[3] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz and Lawrence, “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing”, arXiv, 1903.10145, 2019.

[4] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner, “β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”, In International Conference on Learning Representations (ICLR), 2017.

[5] Naftali Tishby, Fernando C. Pereira and William Bialek, “The information bottleneck method”, arXiv, 2000.

[6] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins and Alexander Lerchner, “Understanding disentangling in β-VAE”, arXiv, 1804.03599, 2018.

[7] Hyunjik Kim and Andriy Mnih, “Disentangling by Factorising”, In International Conference on Machine Learning (ICML), 2018.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever and Pieter Abbeel, “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”, In Neural Information Processing Systems (NeurIPS), 2016.

[9] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang and Tarek Abdelzaher, “ControlVAE: Controllable Variational Autoencoder”, In International Conference on Machine Learning (ICML), 2020.

[10] Karl Johan Åström and Tore Hägglund, “Advanced PID control”, volume 461, International Society of Automation (ISA), 2006.

[11] Loic Matthey, Irina Higgins, Demis Hassabis and Alexander Lerchner, “dSprites - Disentanglement testing Sprites dataset”, https://github.com/deepmind/dsprites-dataset, 2017.

[12] Ricky T. Q. Chen, Xuechen Li, Roger Grosse and David Duvenaud, “Isolating Sources of Disentanglement in Variational Autoencoders”, In Neural Information Processing Systems (NeurIPS), 2018.

[13] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization”, In International Conference on Learning Representations (ICLR), 2015.