Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework

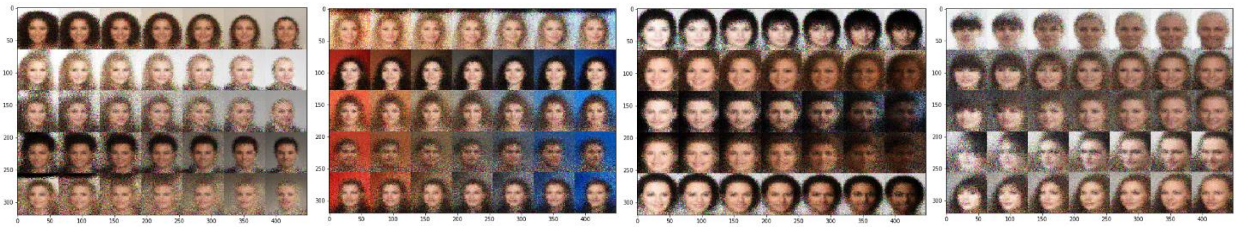
1. Experiment details

Augmenting Variational Autoencoder, the author introduced an additional hyper-parameter $\beta(>1)$ to encourage disentanglement between features of latent variable. (Note that $\beta = 1$ in Variational Autoencoder.) This is derived from the Lagrange theory on constrained optimization problem where the constraint is bounding the KL divergence between the posterior and the prior of latent variable. This further weights on regularization term so that the learned posterior cannot deviate far from the Gaussian prior with 0 mean and identity covariance matrix. Another main contribution of this paper is that they devised a new metric for quantifying the degree of disentanglement given the underlying factors of variations in data. Still, the definition of 'disentanglement' is open to debate, but referring to current researches, most of the newly devised disentanglement metrics are revised, improved, and evaluated based on the metric in this paper. Both of the experiments are focused on comparing the performance of VAE and β-VAE.

- For the first experiment, I decided to plot the latent traverse with CelebA dataset just as the Figure 1 in the paper. This is the qualitative analysis on disentanglement to verify the effectiveness of the hyper-parameter β which is set to 250. (The value is way bigger than the vanilla Variational Autoencoder where $\beta = 1$.) First, 5 latent variables are sampled and then certain dimension is selected to vary the value from -3 to 3 with interval 6/7 while other dimensions are stay fixed. Finally, these samples are feed to the trained decoder and then the generated outputs are plotted in a single canvas.

- For the second experiment, I decided to calculate the disentanglement metric classification accuracy with 2D shapes dataset just as in the Figure 6 (left) in the paper. This is the quantitative analysis on disentanglement where β is set to 4 this time. First, construct a dataset where each sample in particular batch shares a certain factor of variations which is sampled from the uniform distribution. Then, for each batch, all the instances are feed to the encoder to get a latent representation. By computing absolute difference between the first half and the second half and taking the mean along the instances, one training instance is generated to train the classifier. The classifier is set to be linear and with soft-max output nonlinearity and a cross-entropy loss function was used.

2. Reproduction results



Hair **Background** **Skin color (cropped image)** **Bald (cropped image)**

The latent traverse plot of β−VAE verifies that the additional hyperparameter β encourages disentanglement between features. Comparing to the generated images from VAE, those from β-VAE were more blurred. However, since it is more robust to the backgrounds, the generated images can catch facial features better, while VAE are kind of overfitted to background. Therefore, here I cropped the center of images to allow the model to focus more on the facial features. As a result, the situation gets better in the sense that the generated images become clearer and more robust. Furthermore, it was further observed that the image gets cleaner just as the paper if the decoder output is estimated as its mean rather than sampled by the estimated normal distribution. You can check all the details in the 'plots' folder.

Model	VAE	β−VAE
Disentanglement score	62%	79%

The disentanglement metric further verifies the validity of β. While VAE achieves only 62% accuracy, β−VAE achieves 79% which significantly achieves better score. Without doubt, this indicates that β−VAE results in better disentanglement between features. However, even if the score of VAE was exactly in the confidence bound referred in the paper, there was still a gap in β−VAE scores. It implies that β−VAE model parameters are not converged yet so that the biased encoder prevents the linear classifier from predicting the factor of variation well. This critique can be further supported by the fact that the train loss was yet higher that the test loss. Since β−VAE essentially possesses the constraint, optimization was bit tricky where the loss curve fluctuates a lot. (I found that warming up the learning rate in the early stage empirically helps)

3. Discussion

To begin with, as the regularization gets stronger, the performance on reconstruction deteriorates, which results in blurred images. How can this problem be resolved? Many researches such as β-TCVAE, Factor-VAE, DIP-VAE are suggested as variants of VAE which also deal with disentanglement by modifying the ELBO through information theoretic approach.