Student's-T Variational Auto-Encoder

Deep Generative Models Project

Ofek Ophir 207180191, Yaakov (Kobi) Rahimi 208675686

February 3, 2023

Abstract

We propose an alternative method for producing Variational Auto-Encoders in which the prior over the latent space is a centered isotropic multivariate Student's T-distribution, i.e. $St(0, I, \nu)$ with ν being the degree of freedom, and the posterior is multivariate Student's T-distribution with μ mean and σ variance, i.e. $St(\mu, \sigma, \nu)$. The objective of the Student's T Variational Auto-Encoder is therefore calculated by deriving the evidence lower bound of the model using the KL-divergence between the posterior and prior, and computing a suitable reconstruction error between the input data and the reconstructed data from the decoder.

1 Introduction

Variational Auto-Encoder (VAE) is a probabilistic graphical generative machine learning model introduced by [1]. It is a modification on auto-encoders which provides a statistic manner for describing the data in the latent space using an approximation to the posterior probability of the unobserved variables.

Auto-encoder is a bottleneck architecture which turns high-dimensional data into a latent low-dimensional encoding in an unsupervised fashion. It contains two neural networks, encoder and decoder:

- Encoder - Aims to learn an efficient encoding of the data in the bottleneck.
- Decoder - Uses the latent space in the bottleneck to regenerate the data.

The model is trained using a reconstruction loss between the input data and regenerated decoded data [2].

VAE is a generative model which uses variational inference with an auto-encoder architecture to learn the latent space encoding. Differently from an auto-encoder, VAE uses variational bayesian methods to learn the probability distribution, allowing for latent sampling, resulting in generative capability [3]. VAE uses Gaussian distribution for the latent variables because it has convenient properties such as analytical evaluation of the KL-divergence, and it allows for the reparameterization trick for efficient gradient computation, even though other distributions are also capable of the reparameterization trick [4], this includes the Student's-T distribution. We hypothesize that due to heavier tails (controlled by the degree of freedom), Student's t-distributions can reject outliers in a preferable fashion compared to Gaussian distributions when used as priors in modeling.

2 Method

The probability density function of the Student's t-distribution is denoted as:

$$f(z) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\left(1 + \frac{(z-\mu)^T\sigma^{-2}(z-\mu)}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-x}\,dt$ is the Gamma function, μ is the mean, σ is the scale, and ν is the degree of freedom.

The Student's t-distribution belongs to location-scale family distributions and it is possible to use the reparameterization trick as discussed in [1].

$$\begin{aligned} X &\sim St(\mu, \sigma, \nu), \\ X &= \mu + \sigma T, \\ T &\sim St(0, I, \nu) \end{aligned}$$

This is still not a differentiable transformation, However [4] found an implicit differentiation of the Gamma distribution, and with the reparameterization trick above, the Student's t-distribution is differentiable with respect to μ, σ, ν.

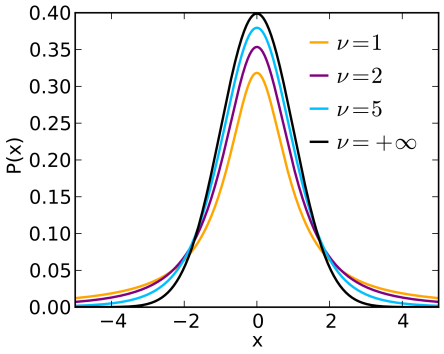

Figure 1: Probability density function of the Student's-T distribution with various degrees of freedom ν.

The objective of the VAE model is to maximize the Evidence Lower BOund (ELBO), defined as:

$$ELBO = \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log(p_\theta(\boldsymbol{x}|\boldsymbol{z}))] - \mathbb{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

Since $p_\theta(\boldsymbol{z}) \sim St(0, I, \nu)$ and $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \sim St(\mu, \sigma, \nu)$, the KL-divergence takes the shape of:

$$\mathbb{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})}\,d\boldsymbol{z}$$

Which equals:

$$= -\log|\sigma| - \left(\frac{\nu+1}{2}\right) \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log\left(1 + \frac{(\boldsymbol{z}-\mu)^T\sigma^{-2}(\boldsymbol{z}-\mu)}{\nu}\right)\right] + \left(\frac{\nu+1}{2}\right) \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log\left(1 + \frac{\boldsymbol{z}^2}{\nu}\right)\right]$$

The second expectation is further developed in [5].

We propose a similar architecture of the original VAE, yet different prior and posterior distribution sampling.

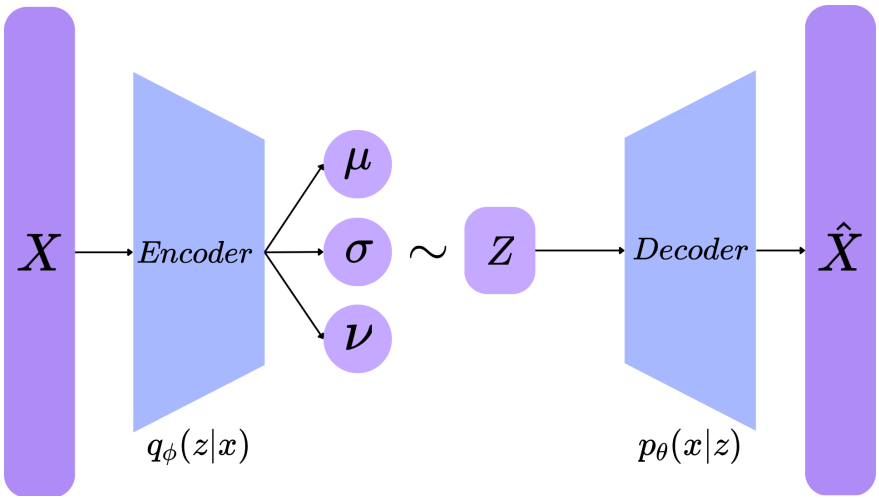

Figure 2: Student's-T VAE architecture. The posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is quantified by the encoder neural network, while $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is quantified by the decoder neural network.

3 Results

We trained two VAE models, a multivariate Gaussian VAE (the default model) and the proposed multivariate Student's-T VAE, we trained and tested the model on MNIST [6] and Fashion-MNIST [7]. We implemented the models using both fully connected and Convolutional Neural Network (CNN) architectures. The fully connected model includes 3 hidden layers with size [128, 64, 32] (for the encoder) and [32, 64, 128] (for the decoder), while the CNN model is comprised of 3 Conv2D layers, each followed by a batch normalization layer and LeakyRelu activation. We trained the models for 500 epochs with batch size = 512, and latent dim = 3. The learning rate for the fully connected architectures was set to 5e-4 and for the CNN architectures, it was set to 3e-4. The results from the CNN are sharper than those obtained from the fully connected architecture, however the difference was not significant, and all results attached to the GitHub repository below.

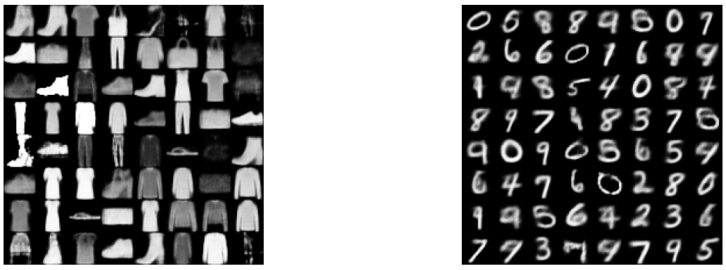

Figure 3: Generated images from the Student's-T VAE CNN model after 500 epochs for Fashion-MNIST (left) and MNIST (right).

Figure 4: Generated images from the Gaussian VAE CNN model after 500 epochs for Fashion-MNIST (left) and MNIST (right).

We also show the train and test loss function for the the two datasets for each of the models, the Student's-T VAE loss is shown in Figure 5 and the Gaussian VAE loss is shown in Figure 6.

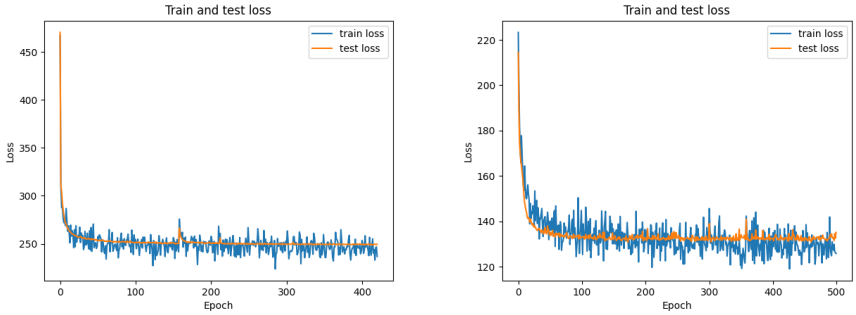

Figure 5: Fashion-MNIST loss (left) and MNIST loss (right) for the Student's-T VAE CNN model.

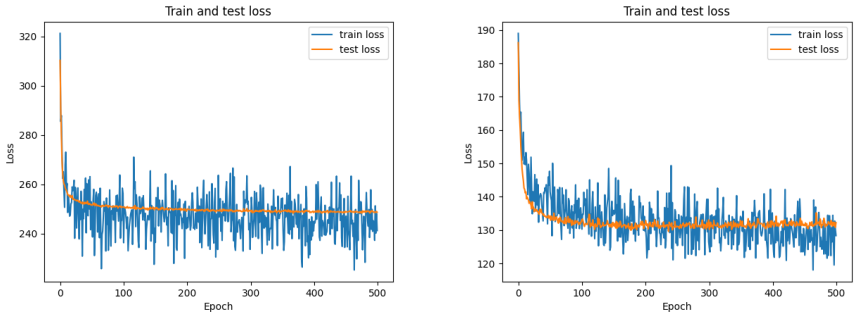

Figure 6: Fashion-MNIST loss (left) and MNIST loss (right) for the Gaussian VAE CNN model.

4 Conclusions

In this project for the course of deep generative models, we have implemented a variational auto-encoder with multivariate Student's-T distribution instead of a Gaussian distribution using both CNN and fully connected architectures. We have compared the results to a Gaussian distribution VAE on two datasets, MNIST and Fashion-MNIST. We optimized the models using grid search for hyperparameters tuning.

We show that Student's-T VAE is similar to Gaussian VAE in regards of generative capabilities on the experimented datasets, however Student's-T VAE allows more control to the researcher by that one can choose the ν (degree of freedom), we believe that using VAE with Student's-T distribution is beneficial for more complex data modeling.

5 Code

Code for the project can be found at: Student's-T VAE GitHub

References

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

[2] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[3] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[4] Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

[5] Konstantinos Zografos et al. On maximum entropy characterization of pearson's type ii and vii multivariate distributions. *Journal of Multivariate Analysis*, 71(1):67–75, 1999.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.