Robustifying Generative Models

Bui Binh An (Andrew) Pham, University of Melbourne

Introduction

Anomaly detection has a myriad of real-world applications such as intrusion dectection, fraud detection, or discovering unknowing diseases. With the ability to learn and approximate the density of input features, deep generative models are widely viewed to be able to detect outliers. However, recent works have demonstrated that certain deep generative models, such as flowed-based models, VAEs ([6]), and PixelCNNs ([13]) often assign a higher likelihood to dataset that differ from the one upon which the models were trained [11, 4]. For example, model trained on CIFAR-10 ([7]) and FashionMNIST ([14]) assign higher likelihood to out-of-distribution test datasets SVHN ([12]) and MNIST ([8]) respectively.

To investigate this peculiar phenomenon, we choose VAEs model specifically to replicate the experiment. Surprisingly, the phenomenon can only be observed in model trained on CIFAR-10 and tested on SVHN but not for the model trained on FashionMNIST and tested on MNIST.

We also study two proposed solution seeking to address the strange problem without having to provide a directed supervision such as giving the neural networks a means of assigning anomaly score to input [3, 9, 10].

Background

We begin by briefly reviewing the definition of VAEs models. We assume our training data x is generated from a latent representation z that follows distribution $p(z)$ (usually a standard Gaussian is used). Thus, our training data is sampled from distribution $p(x|z)$. We are interested in finding the parameter θ that make our model approximating to the true but unknown data distribution $p(x|z)$. This is equivalent to finding θ that maximise the likelihood of reconstructing training data:

$$p_\theta(x) = \int p(z)p_\theta(x|z)dz$$

However, computation of this equation is intractable, therefore, require the use of approximation techniques such as variational inference. In addition to decoder network $p_\theta(x|z)$, which decodes the representation vector back to the original space, we define additional encoder network $q_\phi(z|x)$, which maps the high-dimensional input data into a lower-dimensional representation vector, (often followed a multivariate Gaussian) that approximates $p_\theta(z|x)$. This enables us to derive a lower bound on the data likelihood that is tractable, so we can optimize:

$$logp_\theta(x) \geq E_{q_\phi(z|x)}[logp_\theta(x|z)] - KL[q_\phi(z|x)||p(z)]] = ELBO \quad (1)$$

where equality holds iff $q_\phi(z|x) \equiv p_\theta(z|x)$. Our objective is to train the encoder and decoder neural networks parameterized by ϕ and θ respectively to maximize the lower bound. By maximizing this lower bound function, the first term of the right-hand-side of Eq.(1) maximize the likelihood of original input being reconstructed, while the second term make the approximate posterior $q_\phi(z|x)$ close to prior $p_\theta(z)$, standard normal distribution. In addition, the reparametrization trick is used to reduce the variance of the gradient estimate.

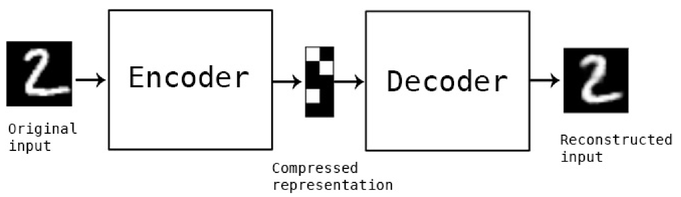
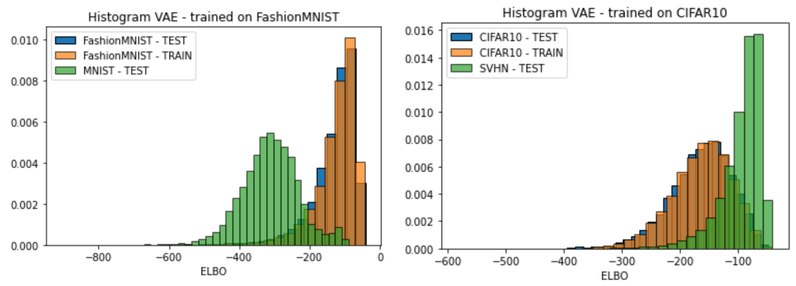


Fig. 1: Diagram of a VAE model, adopted from [1]

Experiment

We trained a VAE model architecture on FashionMNIST and CIFAR-10. We then caluculated the value of ELBO, which approximate the log-likelihood, of the two test datasets with the same dimensionality - MNIST and SVHN respectively. It is expected that the models assign a higher probablity to this test data, although they were not trained on it, as shown in [11]. However, our model successfully learn to assign lower probability to outlier test dataset MNIST on a FashionMNIST trained model (Figure 2a). The strange behaviour only persists for the CIFAR-10 vs SVHN case (Figure 2b). We believe one of the reasons leading to this phenomenon with CIFAR-10 and SVHN dataset is, and also pointed out in [11], that both these two datasets have roughly the same mean but SVHN has a smaller variance, which resulting in having a higher likelihood.



a) b)

Fig. 2: Histogram of VAE ELBO values for FashionMNIST vs MNIST (a), CIFAR-10 vs SVHN (b).

Proposed Solutions

The first paper claims that the counter-intuitive behavior of assigning higher likelihood to out-of-distribution test dataset in VAEs is due to the models' lack of capacity in learning the ground truth, true distribution [2]. Their proposed solution is, therefore, increasing the model capacity by increasing the number of channels in the convoluted neural networks. Although this idea is intriguing and ostensible, the original authors decided to retracted the paper due to an error in their empirical experiments. We independently run the proposed sufficient-capacity model but it could not produce the expected result, assigning lower likelihood to outlier test dataset SVHN on a CIFAR10 trained VAE model. Thus, we agree with the paper's authors decision to retract and delay the publish until the underlying problem can be better understood.

Secondly, we propose to look at forward Kullback-Leibler divergence (KL) instead of reverse KL as currently used for variational inference in VAEs models. Due to the difference in computation, forward KL can alleviate the issue of covariance underestimation and light tails in reverse KL [5]. In variational inference, we seek to find a candidate distribution q^* in a family of distributions Q according to some criterion. The typical criterion is to minimize $KL(q||p)$ (reverse KL), where p is the target distribution. This turns out to be equivalent to maximising the ELBO. Alternatively, we can seek a distribution q^* satisfied:

$$q^* = argmin_{q\in Q}\, KL(p||q) = argmin_{q\in Q}\, E_p[log_p] - E_p[log_q] \quad (2)$$

Since the expectation is under the target distribution p, we can not compute the forward KL divergence $KL(p||q)$ exactly. However, with simple rearrangement, we can make the expectation in Eq.2 under the approximated distribution q as follows:

$$KL(p||q) = E_p\,[log\frac{p}{q}] = E_q\,[\frac{p}{q}\,log\frac{p}{q}] \quad (3)$$

In addition, [5] pointed out that Eq.3 can be approximated, resulting in variational inference can be trained using forward KL. We, therefore, believe that using forward KL instead of reverse KL in VAEs may help mitigate the out-of-distribution phenomenon.

Conclusion

We have shown that with our proposed architecture, VAE model can recognize outlier test dataset MNIST when trained on FashionMNIST as opposed to its failure in [11]. However, we agree with [11] that the peculiar phenomenon can still be observed when testing out-of-distribution test data SVHN against a CIFAR-10 trained model. Because of this unstable performance, it is recommended that we should be cautious when using these models with potential out-of-distribution inputs. Moreover, we suggest trying to use forward KL instead of reverse KL for variational inference in VAEs models. Although we have not been able to implement this empirically due to time constraints, we believe the strange out-of-distribution phenomenon in VAEs might be alleviated by following this direction.

Acknowledgements

I would like to thank the Dept. of Math and Stats at the University of Melbourne for organizing this Vacation Scholarship program, helping me and other aspiring students getting start on the research journey. I would also like to thank my supervisor, Dr. Susan Wei, for her valuable insights, support, guidance, and encouragement in this project.

References

[1] Francois Chollet. *Building Autoencoders in Keras*. 2016. URL: https://blog.keras.io/building-autoencoders-in-keras.html.

[2] Bin Dai and David Wipf. "When Do Variational Autoencoders Know What They Don't Know?" 2020.

[3] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *International Conference on Learning Representations* (2017).

[4] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *International Conference on Learning Representations* (2019).

[5] Ghassen Jerfel et al. "Variational Refinement for Importance Sampling Using the Forward Kullback-Leibler Divergence". In: *Third Symposium on Advances in Approximate Bayesian Inference*. 2021. URL: https://openreview.net/forum?id=67p4Qb3fe4k.

[6] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

[7] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. "CIFAR-10 (Canadian Institute for Advanced Research)". In: (). URL: http://www.cs.toronto.edu/~kriz/cifar.html.

[8] Yann LeCun and Corinna Cortes. "MNIST handwritten digit database". In: (2010). URL: http://yann.lecun.com/exdb/mnist/.

[9] Kimin Lee et al. "Training confidence-calibrated classifiers for detecting out-of-distribution samples". In: *International Conference on Learning Representations* (2018).

[10] Si Liu et al. *Open Category Detection with PAC Guarantees*. 2018. arXiv: 1808.00529 [cs.LG].

[11] Eric Nalisnick et al. "Do Deep Generative Models Know What They Don't Know?" In: *International Conference on Learning Representations* (2019).

[12] Yuval Netzer et al. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS* (Jan. 2011).

[13] Aaron van den Oord et al. *Conditional Image Generation with PixelCNN Decoders*. 2016. arXiv: 1606.05328 [cs.CV].

[14] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. arXiv: 1708.07747 [cs.LG].