Out of Distribution Detection Using Bayesian Variational Autoencoder

Kamran Chitsaz Alireza Razaghi

1 Introduction

The recent success of deep learning methods plays an essential role in the evolution at an unprecedented rate of current technology. Despite their remarkable performance in a wide range of complex tasks including image recognition [1], object detection [2], and speech recognition [3], deep learning models can make confident, but meaningless, predictions on unobserved data [4]. To perform reliable decisions, Deep learning models need to be able to determine whether the input data is an anomaly or outlier and significantly different from the training data. Such data are called out-of-distribution (OoD) data. Several approaches have been proposed to detect OoD data such as spectral anomaly detection techniques which try to find lower dimensional embeddings of the input data where OoD and normal data are expected to be separated from each other. The Principal components analysis (PCA) [5] and Autoencoders [6] are among the most popular dimension reduction methods for detecting outliers.

Likelihood estimators are one of the most promising OoD detection approach which work for probabilistic generative models. Such models can evaluate the likelihood of input data, and if a generative model fits the training data distribution well enough, it should assign high likelihood to samples from the similar distribution and low likelihood to OoD samples. So to detect whether a data is OoD or not, by this approach first trains a density estimator parameterized by θ, and then classifies an input $\mathbf{x}^*$ as OoD based on a threshold on the density of $\mathbf{x}^*$, i.e., if $p\left(\mathbf{x}^* \mid \theta\right) < \tau$ [7]. In this project we used variational autoencoder to calculate log likelihood and detect OoD.

2 Background

One of the most popular models for density estimation is the Variational Autoencoder [8] which successfully can approximate the distribution of training data. By this generative model we can learn a joint model $p(x, z)$ of the data x and some latent variables z. The idea is that each data X has a corresponding latent variable Z that data X is caused by this Z, so we can marginalize out Z.

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

The conditional distribution $p(x|z)$ is a Gaussian. So, we have a kind of mixture of infinitely many Gaussians, for each value of Z, there's one Gaussian and we mix them with weights. In this approach we assumed that prior is normal and likelihood is a Gaussian which it's parameters are depend on latent variable z. One way to define parameters is using neural

networks. So the parameters of our model depends on leading variable z which is generated by a neural network with parameter θ. So we can build the observation model as follow:

$$\begin{aligned} &p(x) = \int p(x|z)p(z)dz \\ &p(z) = \mathcal{N}(0, I) \\ &p(x \mid z) = \mathcal{N}\left(x \mid \mu_\theta(z), diag\left(\sigma_\theta^2(z)\right)\right) \end{aligned}$$

Since computing $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$ is intractable in high dimention, variational inference is used to derive a lower bound on the log likelihood of x. So we can approximate posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ using amortized variational inference via another neural network with parameter ϕ. Similarly, we can build the inference model as

$$q(z \mid x) = \mathcal{N}\left(z \mid \mu_\phi(x), \mathrm{diag}\left(\sigma_\phi^2(x)\right)\right)$$

Using Jensen's inequality we have

$$\begin{aligned} \log p(x) &= \log \int \frac{q(z \mid x)}{q(z \mid x)} p(x, z)dz \geq \int q(z \mid x) \log \frac{p(x, z)}{q(z \mid x)} dz \\ &= E_{q(z|x)} \log \frac{p(x, z)}{q(z \mid x)} = E_{q(z|x)} \log \frac{p(x \mid z)p(z)}{q(z \mid x)} \end{aligned}$$

So the objective we're trying to maximize is

$$\begin{array}{ll} \underset{w,\phi}{\text{maximize}} & \sum_i E_{q_i} \log \frac{p(x_i|z_i,w)p(z_i)}{q_i(z_i)} \\ \text{subject to} & q_i\left(t_i\right) = \mathcal{N}\left(m\left(x_i, \phi\right), \mathrm{diag}\left(s^2\left(x_i, \phi\right)\right)\right) \end{array} \tag{1}$$

We can rewrite equation 1 as below:

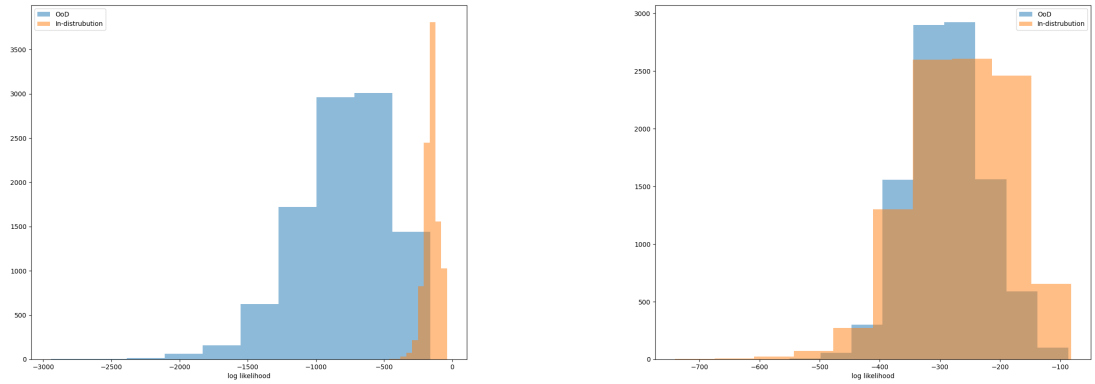
$$\begin{aligned} &\sum_i E_{q_i} \log p\left(x_i \mid t_i, w\right) + E_{q_i} \log \frac{p(t_i)}{q_i(t_i)} \\ &= \sum_i E_{q_i} \log p\left(x_i \mid t_i, w\right) - \mathcal{KL}\left(q_i\left(t_i\right) \| p\left(t_i\right)\right) \end{aligned} \tag{2}$$

When we maximize this minus KL we are actually trying to minimize KL so we are trying to push the variational distribution q_i as close to the prior as possible which is standard normal. Also if for simplicity we set all the output variances to be 1, the first term, log likelihood of x_i given t_i, can be interpreted as $-\left\|x_i - \mu\left(t_i\right)\right\|^2 + \text{const}$ which is actually a reconstruction loss. It tries to push x_i as close to the reconstruction as possible.

Finally, one can use importance sampling w.r.t. the variational posterior $q(z \mid x, \phi)$ to get an estimator $\hat{p}(\mathbf{x} \mid \theta, \phi)$ of the probability $p(x \mid \theta)$ of an input x under the generative model, i.e.,

$$p(\mathbf{x} \mid \theta) \simeq \hat{p}(\mathbf{x} \mid \theta, \phi) = \frac{1}{K} \sum_{k=1}^{K} \frac{p\left(\mathbf{x} \mid \mathbf{z}_k, \theta\right) p\left(\mathbf{z}_k\right)}{q\left(\mathbf{z}_k \mid \mathbf{x}, \phi\right)}$$

where $\mathbf{z}_k \sim q(\mathbf{z} \mid \mathbf{x}, \phi)$ and where the estimator $\hat{p}(\mathbf{x} \mid \theta, \phi)$ is conditioned on both θ and ϕ to make explicit the dependence on the parameters ϕ of the proposal distribution $q(z \mid x, \phi)$.



(a) (b)

Figure 1: Histogram of the log likelihood of test samples for (a) VAE traind on Mnist and Fashion-Mnist as OoD, and (b) VAE traind on Fashion-Mnist and Mnist as OoD

3 Results

We have trained our VAEs with samples only from training data from in distribution dataset and evaluate our method by test sample from different dataset to measure OoD performances. In first experiment we have set Mnist dataset as in distribution and Fashion-Mnist as OoD which you the results are illustrated in Fig. 1a. In this case we observed that our method assign higher probability to in distribution data and can detect OoD with accurcy 89.90%. But as another experiment we have set Fashion-Mnist dataset as in distribution and Mnist as OoD which you the results are illustrated in Fig. 1b. As you can see our method failed. Also, reference [9] shows that almost all major types of probabilistic generative models, including VAE, can assign spuriously high likelihood to OoD samples which make them unreliable metrics for OoD detection.

3.1 Metrics

To classifie whether a data is OoD or not we compute maximum likelihood and then compare with a threshold. The choice of a threshold depends on the particular application. We have tune the threshold, as same as other model hyper parameters, by validation set.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[3] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. Recent advances in deep learning for speech research at microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8604–8608. IEEE, 2013.

[4] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[6] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, 2014.

[7] Christopher M Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

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

[9] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.