Modeling assumptions and evaluation schemes: On the assessment of deep latent variable models

Judith Bütepage Petra Poklukar Danica Kragic
Robotics, Perception and Learning
KTH Royal Institute of Technology
butepage@kth.se, poklukar@kth.se, dani@kth.se

Abstract

Recent findings indicate that deep generative models can assign unreasonably high likelihoods to out-of-distribution data points. Especially in applications such as autonomous driving, medicine and robotics, these overconfident ratings can have detrimental effects. In this work, we argue that two points contribute to these findings: 1) modeling assumptions such as the choice of the likelihood, and 2) the evaluation under local posterior distributions vs global prior distributions. We demonstrate experimentally how these mechanisms can bias the likelihood estimates of variational autoencoders.

1. Background

In recent years, we have witnessed an increasing interest in density modeling using deep generative models (DGMs). Their striking ability to generate e.g. natural looking images has gained attention both in the artificial intelligence community as well as the public media. A recent paper [6] raised an important question: ”Do Deep Generative Models Know What They Don’t Know?”. The authors show in extensive experiments that likelihood-based methods, such as variational autoencoders (VAEs) [4, 9] and flow-based methods [2, 3], can assign higher likelihoods to out-of-distribution (OOD) samples than to in-distribution (ID) samples, which are drawn from the same distribution as the training data. For example, a DGM trained on the FashionMNIST dataset will assign higher likelihoods to images from the MNIST dataset compared to testing images from the FashionMNIST dataset.

In this work we take a Bayesian approach to the problem and discuss why VAEs can assign high likelihoods to OOD examples. We believe that this problem is caused by a) modeling assumptions, and b) evaluation schemes. Due to the lack of space, we refer the reader to [4, 9] for the specifics of VAEs.

Modeling assumptions: A modeling assumption that contributes to the performance of VAEs is the choice of the likelihood. For images for example, we often define a rather simplistic likelihood function, such as independent and identically distributed (iid) pixels. Humans however would rate the likelihood of an image in a certain dataset based on its semantic content and spatial decomposition. In fact, the true likelihood function of a data distribution is often unknown.

Local vs global evaluation: For a test point $\hat{\mathbf{x}}$, a VAE can be evaluated by marginalizing over the latent variable $\mathbf{z}$ either globally, using the prior, or locally, under a local, point-estimated probability distribution that depends on $\hat{\mathbf{x}}$. To-date, there exists no protocol evaluation scheme that is used by all papers on VAEs.

2. Modeling assumptions

From a Bayesian view point, we would like to find a model M which explains the observed data $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_N}\}$ in the best manner. We assume that the data was generated from the following distribution $p(\mathbf{X}, \mathbf{Z}, \theta_x, \theta_z | M^*) = p(\mathbf{X}|\mathbf{Z}, \theta_x) p(\mathbf{Z}|\theta_z) p(\theta_x, \theta_z | M^*)$, where $\mathbf{Z}$ are local latent variables, θ_x and θ_z are global parameters and M^* is the true, data generating model. Here, local variables are data point specific, such as the cluster assignment in a Gaussian Mixture Model, while global parameters generalize over all data points, such as the means and variances of the mixtures. Thus, we would like to find the model that maximizes

$$M = \arg\max_M \int p(\mathbf{X}, \mathbf{Z}, \theta_x, \theta_z | M) \partial\theta_x \partial\theta_z.$$

Most VAE literature to-date discusses modeling assumptions related to the structure and choice of distribution over the latent variables $p(\mathbf{z}|\theta_z)$. A common assumption is that a more flexible posterior distribution $p(\mathbf{z}|\mathbf{x}, \theta_z)$ will lead to a better model performance, e.g. [8].

Few papers, e.g. [1, 7], discuss the structure of the likelihood term itself. We are interested in how this choice influences the model performance not only on ID test data but also OOD testing points. In most cases, the likelihood function is chosen to describe only local statistics in the data, such as iid pixels, instead of the global structure of the training data distribution. Note that the VAE model itself does capture global information about the data point, especially in the latent variable, whereas the likelihood function only focuses on local errors. While it appears unintuitive that a likelihood function of a VAE, trained on Cifar-10, will assign a higher value to an image from the The Street View House Numbers (SVHN) dataset than to a test image of a car from Cifar-10 [6], we need to keep in mind that it judges pixels, not semantic content. An image from the SVHN dataset usually consists of a background with a single color and some darker regions that constitute a number. The image of a car on the other hand consists of a varied background with parts of the street, grass and the sky as well as the vehicle itself. This variance in color intensity moves the Cifar-10 image further away from the average Cifar-10 image than the SVHN test image and can therefore be judged as less likely.

Even without including semantic knowledge, the modeling choice of the likelihood function can play a major role. For instance, compared to MNIST, the intensity values in FashionMNIST vary within and between images, e.g. due to shading. If we assume an iid Bernoulli likelihood and binarize the images, low intensity areas will disappear. Thus, the Bernoulli likelihood might not be the correct choice for FashionMNIST. In the experiments (Section 4), we demonstrate how the choice of either an iid Bernoulli or an iid Gaussian likelihood changes the behavior of a VAE with respect to OOD testing data.

3. Local vs global evaluations

In the traditional variational inference (VI) setting, we approximate the often intractable posterior distributions over latent variables $\mathbf{Z}$ and parameters θ_x, θ_z with simpler, approximate distributions q:

$$p(\mathbf{Z}, \theta_x, \theta_z|\mathbf{X}, M) \approx q(\mathbf{Z}|\theta_z)q(\theta_x, \theta_z|\mathbf{X}, M).$$

After having inferred the parameters $(\mathbf{Z}, \theta_x, \theta_z)$, a model can be evaluated on a testing point $\hat{\mathbf{x}}$ using

$$\begin{aligned}
p_{VI}^{PR}(\hat{\mathbf{x}}|M) &= \int_{\hat{\mathbf{z}}, \theta_x, \theta_z} p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \theta_x)p(\hat{\mathbf{z}}|\theta_z)p(\theta_x, \theta_z|M) \\
p_{VI}^{PO}(\hat{\mathbf{x}}|\mathbf{X}, M) &= \int_{\hat{\mathbf{z}}, \theta_x, \theta_z} p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \theta_x)p(\hat{\mathbf{z}}|\theta_z)p(\theta_x, \theta_z|\mathbf{X}, M) \\
p_{VI}^{APO}(\hat{\mathbf{x}}|\mathbf{X}, M) &= \int_{\hat{\mathbf{z}}, \theta_x, \theta_z} p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \theta_x)q(\hat{\mathbf{z}}|\theta_z)q(\theta_x, \theta_z).
\end{aligned}$$

Here p_{VI}^{PR} is the prior predictive (PR), p_{VI}^{PO} is the posterior predictive (PO) and p_{VI}^{APO} is the approximate posterior predictive (APO) distribution. Note that after the inference, the parameters (θ_x, θ_z) are fixed which implies that e.g. $p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \theta_x)$ has the same structural form for all testing points.

VAEs on the other hand assume the generative model

$$\begin{aligned}
p(\mathbf{X}, \mathbf{Z}, \theta_z, M) &= p(\mathbf{X}|\phi_x(\mathbf{Z}))p(\mathbf{Z}|\theta_z, M), \\
p(\mathbf{Z}|\mathbf{X}, M) &\approx q(\mathbf{Z}|\phi_z(\mathbf{X}), M),
\end{aligned}$$

where a parameterized neural network $\phi_z(\mathbf{X})$ determines the parameters of the approximate posterior distribution q and the neural network $\phi_x(\mathbf{Z})$ determines the structure of the likelihood. θ_z is interpreted as prior parameters. Following the structure above, the model can be evaluated using

$$\begin{aligned}
p_{VAE}^{PR}(\hat{\mathbf{x}}|\mathbf{X}, M) &= \int_{\hat{\mathbf{z}}} p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}), M)p(\hat{\mathbf{z}}|\theta_z, M) \\
p_{VAE}^{PO}(\hat{\mathbf{x}}|\mathbf{X}, M) &= \int_{\hat{\mathbf{z}}} p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}), M)p(\hat{\mathbf{z}}|\phi_z(\hat{\mathbf{x}}), M) \\
p_{VAE}^{APO}(\hat{\mathbf{x}}|\mathbf{X}, M) &= \int_{\hat{\mathbf{z}}} p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}), M)q(\hat{\mathbf{z}}|\phi_z(\hat{\mathbf{x}}), M),
\end{aligned}$$

where the influence of the training data $\mathbf{X}$ manifests itself in form of the fixed parameters of the neural networks ϕ_x and ϕ_z, which is why we condition even the prior predictive distribution p_{VAE}^{PR} on $\mathbf{X}$. The original VAE papers [4] and [9] use importance weighted versions of the posterior predictive p_{VAE}^{PO} and the approximate posterior predictive p_{VAE}^{APO} respectively to compute the marginal likelihood of the testing data. Compared to the traditional VI setting, these two evaluation schemes depend on the testing point $\hat{\mathbf{x}}$, which influences both the structure of the (approximate) posterior and the form of the data likelihood. Assume that a VAE was trained on the FashionMNIST dataset and $\hat{\mathbf{x}}$ to be an image of a pair of trousers. Then p_{VAE}^{PO} and p_{VAE}^{APO} will only evaluate this image under other images of trousers. It will not be evaluated under images of e.g. shoes or sweaters. Therefore, the image is evaluated locally and not under the global training data distribution. If the test point $\hat{\mathbf{x}}$ on the other hand was an image from the MNIST dataset, e.g. the digit 1, it would also be evaluated locally, under images that resemble the digit 1 but are close to the FashionMNIST dataset, i.e. thin and long pieces of clothes. While this behavior is not necessarily unwanted, it can skew the likelihood estimates of ID and OOD data points. We demonstrate this on a simple 3D Gaussian example and compare it to the traditional VI in the next section.

4. Experiments

We present experimental evidence that points towards the problems that arise due to modeling assumptions and evaluation schemes of VAEs.

Local vs global: a 3D Gaussian example We generate toy data by sampling iid samples from a 3D Gaussian $p_{\mathbf{X}}(x)$ and train a VAE with a Gaussian likelihood function on these samples. In Figure 1 **a)** we show samples from the

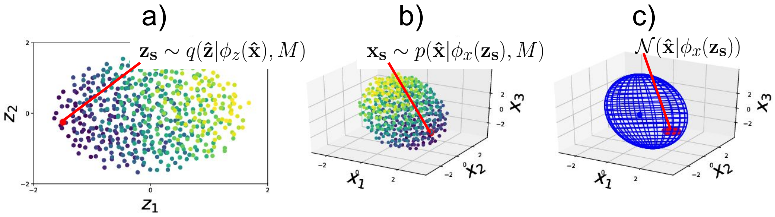

Figure 1: 3D Gaussian: The color coding maps the data points from the latent space (**a)**) to the data space (**b)**). The red point is a test point. **c)** visualizes the Standard Deviational Ellipse (SDE) of the data distribution (blue) and the VAE approximate posterior predictive (red).

prior in the latent space that are then decoded into the output space shown in Figure 1 **b)**. The red point in **a)** is a sample from an encoded testing point and the red point in **b)** is a sample from its decoded likelihood. The likelihood $p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}),M)$ is a Gaussian that is centered close to the position of the testing point and has a low variance. When comparing the difference between the Standard Deviational Ellipse (SDE) of the inferred approximate posterior predictive $p^{APO}_{VAE}(\hat{\mathbf{x}}|\mathbf{X},M)$ (red) compared to the data likelihood $p_{\mathbf{X}}(\hat{\mathbf{x}})$ (blue) in Figure 1 **c)** we see that the testing point is more likely under the local $p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}),M)$ compared to the global $p_{\mathbf{X}}(\hat{\mathbf{x}})$. Effectively, a test set is no longer evaluated under a single Gaussian $p_{\mathbf{X}}(\hat{\mathbf{x}})$ but under an infinite mixture of Gaussians [5] where each testing point is assigned its own mixture component $p(\hat{\mathbf{x}}|\phi_x(\hat{\mathbf{z}}),M)$. This becomes apparent when we compare the log likelihoods (LL) of testing points under a) the true distribution, b) the VAE PR, c) the VAE APO and, for comparison, d) the VI APO as shown in Figure 2. The VAE APO assigns local, overconfident likelihoods compared to the VAE PR that samples globally from the prior while evaluating under sample-dependent likelihoods. The traditional VI approach in which we fitted a Gaussian Mixture Model with ten mixtures to the data outperforms both VAE approaches. This is caused by the fact that it models a global likelihood function instead of point-estimates of distributions.

Modeling assumptions To demonstrate the difference between modeling assumptions, we trained two VAEs, M_1 and M_2, on the FashionMNIST dataset. The only difference between these two models is that they assume an iid Bernoulli and an iid Gaussian likelihood respectively. For M_1 we binarize the images and for M_2 we scale the pixel values to lie between zero and one. We evaluate the log likelihood of these models under the prior and under the approximate posterior. In the latter case, we evaluate with importance weights as in [9]. We test both images on the FashionMNIST and the MNIST dataset and visualize the results in Figure 3. It becomes apparent that the M_2 model is better at detecting the OOD data points. We also see that the VAE APO produces more reasonable estimates than the VAE PR. However, this might be caused by overconfidence as discussed in the previous section. Note that only the PR evaluation with a Bernoulli likelihood reproduces the results reported in [6] while the APO evaluation with a Gaussian likelihood reverses them.

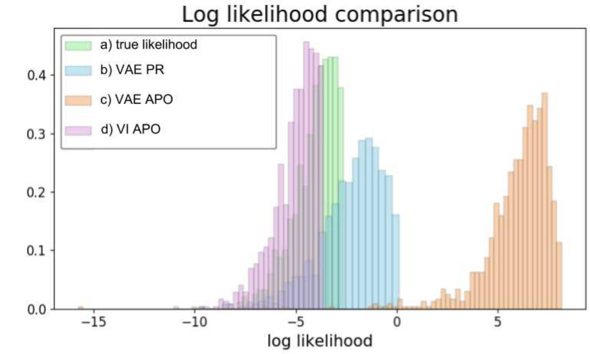

Figure 2: Test data log likelihoods under a) the true LL, b) the prior predictive LL $p^{PR}_{VAE}(\hat{\mathbf{x}}|\mathbf{X},M)$, c) the approximate posterior predictive LL $p^{APO}_{VAE}(\hat{\mathbf{x}}|\mathbf{X},M)$ and d) the VI approximate posterior predictive LL $p^{APO}_{VI}(\hat{\mathbf{x}}|\mathbf{X},M)$.

5. Conclusion

We discussed two problems of VAEs that might contribute to inappropriate likelihoods of OOD samples. Firstly, we show how the modeling assumption on the likelihood function can impact the judgment of the model. We show this by reversing the OOD phenomenon discussed in [6] by simply changing the likelihood assumption. Secondly, we demonstrate that the local evaluation under the approximated posterior leads to overconfidence in case of the toy data. This phenomenon does not appear in the higher dimensional image data, supposedly because of the curse of dimensionality and sample complexity.

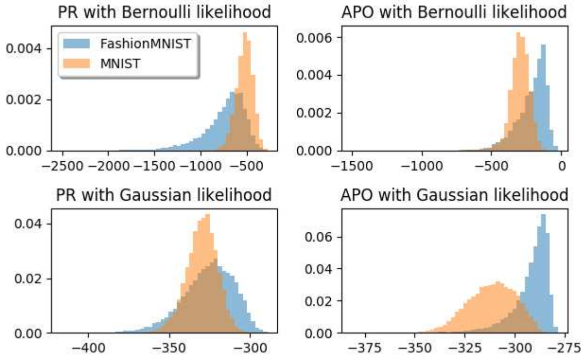

Figure 3: The log likelihood under the prior (left) and an importance weighted approximate posterior (right) using model M_1 with an iid Bernoulli likelihood function (top) and model M_2 with an iid Gaussian likelihood function (bottom).

References

[1] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *ICLR*, 2017.

[2] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *ICLR Workshop Track*, 2015.

[3] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2015.

[5] P.-A. Mattei and J. Frellsen. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, pages 3855–3866, 2018.

[6] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? *ICLR*, 2019.

[7] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016.

[8] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

[9] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

This work was supported by the EU through the project socSMCs (H2020-FETPROACT-2014), the Swedish Foundation for Strategic Research and the Knut and Alice Wallenberg Foundation.