Evaluation

Florence Regol

Contents

1 Testing

Most of the problems associated with evaluation methods of generative models stems from the fact that we don't have access to the ground truth distribution.

The problems:

1. Out-Of-distribution (OOD) log likelihood problem : High likelihood on a dataset doesn't guarantee OOD capability, which is an issue if we want to use the model for anomaly detection. In [1], they show that a model can have high likelihood on in-distribution data, generate good samples, and still assign higher likelihood to OOD data. → with ground truth distribution, we can directly evaluate the appropriate log likelihood that should be given to each element of the space without worrying about the ambiguity of knowing whether a sample is actually OOD or not. We can also explicitly look into the OOD space and see how the generative model behaves in those regions.

2. Sample log likelihood problem : Good likelihood doesn't guarantee good sample generation, as per the simple argument from Section 3.2 of [2]. → with ground truth distribution, the true log likelihood value that a point should have is known. Hence the goal becomes to approach that value, not to maximise the log likelihood on a testing set.

3. Degenerate ($p(\Omega) > 1$) log likelihood problem : Using the log likelihood on a test set as an evaluation metric makes sense if summing the likelihood over the whole support gives 1.

4. Repeating training samples problem: Most sampled-based metric can be worthless if the model repeats the training set. → We can control how much of the positive support is seen by the model during training.

The solution: Use an evaluation method that relies on knowing the ground truth distribution.

1.1 Given a ground truth p

The statistical distance between two distributions p, q, or the total variation, is defined as follow:

$$d_{TV}(p,q) \triangleq \frac{1}{2}||p-q||_1 = \frac{1}{2}\sum_{x\in\Omega}|p_x - q_x| \tag{1}$$

An other related metric is the Hellinger distance:

$$H(p,q) \triangleq \frac{1}{\sqrt{2}}||\sqrt{p}-\sqrt{q}||_2 = \frac{1}{\sqrt{2}}\sqrt{\sum_{\in\Omega}(\sqrt{p_x}-\sqrt{q_x})^2} \tag{2}$$

Given a ground truth distribution p, we can view the generative problem as minimizing those distances with the estimator $\hat{p}$ and p.

Density estimation If we have access to the probability weight given by the estimator to a point x $\hat{p}_x$, then we can directly evaluate those metrics. We can also get the KL distance as an additional metric:

$$KL(p||\hat{p}) = \sum_{x\in\Omega} p_x \log(\frac{p_x}{\hat{p}_x}) = H_p - \sum_{x\in\Omega} p_x \log(\hat{p}_x) \tag{3}$$

The d_{TV}-distances can be broken over positive and zero support regions ($\Omega^+ \triangleq \{x; p_x > 0\}, \Omega^0 \triangleq \{x; p_x = 0\}$), to give insight into how a model would perform in OOD detection:

$$\begin{aligned} d_{TV} &= \frac{1}{2}\sum_{x\in\Omega^+}|p_x-\hat{p}_x| + \sum_{x\in\Omega^0}\hat{p}_x && (4)\\ &= \hat{p}_{iod} + \hat{p}_{ood} && (5) \end{aligned}$$

Samples If we don't have access to the probability distribution of the generative model as it can be the case for machine learning gen. model (most notably GANs), then we have to rely on the generated samples. Throughout this section, the empirical distribution $\hat{p}^{emp}$ is obtained through "Poissonization" to remove any dependence between each $\hat{p}^{emp}_x$. That is, we first sample the number of samples from a Poisson distribution $m' \sim Poi(m)$, then sample $\tilde{\mathbf{x}}^{m'}$ from $\hat{p}$ and obtain

$$\hat{p}^{emp}_x = \frac{1}{m'}\sum_{i=1}^{m'} \mathbb{1}[\tilde{x}_i = x] \tag{6}$$

[Flo : Does that mean that we need to resample a set $\tilde{\mathbf{x}}^{m'}$ for each $\hat{p}^{emp}_x$? Or can we reuse the samples? If not, building the empirical pmf scales with the sampling space $O(|\Omega| * m)$?]

If we only have access to samples from $\hat{p}$, we can rely on estimates:

$$\hat{d}_{TV} = \frac{1}{2}||\hat{p}^{emp} - p||_1 \tag{7}$$

$$\hat{H} = \frac{1}{\sqrt{2}}||\sqrt{p} - \sqrt{\hat{p}^{emp}}||_2 \tag{8}$$

Identity Testing

1.1.1 Background

Based on [3]. The identity testing problem is : Given samples from an unknown distribution $\hat{p}$ over $|\Omega|$ elements, an explicitly given distribution p, and parameters $0 < \epsilon, \delta < 1$, we wish to distinguish, with probability at least $1-\delta$, whether the distributions are identical versus ϵ-far in total variation distance (d_{TV}).

The test goes as follow:

1. Transform the ground truth distribution p on sample space Ω into a uniform distribution Uni on sample space $\mathcal{U}$ using the m-grained trick from [4].

2. Draw $m = \Theta\Big(\frac{1}{\epsilon^2}(\sqrt{|\mathcal{U}|\log(\delta^{-1})} + \log(\delta^{-1})\Big)$ samples from $\hat{p}$.

3. Compute the empirical $\hat{p}^{emp}$

4. Compute the statistic

$$S = \frac{1}{2}\sum_{x\in\mathcal{U}}|\hat{p}^{emp}_x - \frac{1}{|\mathcal{U}|}| \tag{9}$$

5. Compute the threshold

$$t = \mathbb{E}_{Uni}[S] + C \begin{cases} \epsilon^2 \frac{m^2}{|\mathcal{U}|^2} & \text{for } m \leq |\mathcal{U}| \\ \epsilon^2 \sqrt{\frac{m}{|\mathcal{U}|}} & \text{for } |\mathcal{U}| < m \leq \frac{|\mathcal{U}|}{\epsilon^2} \\ \epsilon & o.w. \end{cases} \tag{10}$$

where

$$\mathbb{E}_{Uni}[S] = \frac{|\mathcal{U}|}{m} \sum_{k=\lceil m/|\mathcal{U}| \rceil}^{m} (k - \frac{m}{|\mathcal{U}|}) \frac{(1-\frac{1}{|\mathcal{U}|})^{m-k}}{|\mathcal{U}|^k} \tag{11}$$

6. if $S \geq t$ then we declare that $d_{TV}(p, \hat{p}) \geq \epsilon$.

The constant C can be derived from the requirement :

$$m \geq C \frac{1}{\epsilon^2} (\sqrt{|\mathcal{U}| \log(\delta^{-1})} + \log(\delta^{-1})) \tag{12}$$

1.1.2 Evaluation metric based on the test

The idea is to fix a significance level δ and a number of samples [Flo : or an error threshold], and to see at what error threshold ϵ the test can't differentiate between the true distribution and the proposal [Flo : or at how many samples].

Algorithm 1 S metric

```
**Input:** $m$, $C = 1, \delta = 0.05$
Compute the min error we can declare $\epsilon^* = \sqrt{\frac{\sqrt{|\mathcal{U}| \log(\delta^{-1})} + \log(\delta^{-1})}{m}}$
Sample m from $\hat{p}$ and build empirical $\hat{p}^{emp}$
Compute $S$ and threshold $t$
**if** $S \geq t$ **then**
    Find at which $\epsilon > \epsilon^*$ we can declare the test to be successful
    **return** $\epsilon$
**else**
    **return** $< \epsilon^*$
**end if**
```

We design a ground truth distribution to generate a synthetic dataset. We define the sample space $\mathbf{X} \in \{C_1, C_2, \ldots, C_K\}^K$ and only assign probability mass on permutations of $\{C_1, C_2, \ldots, C_K\}$ ($\mathcal{P}_K$ denotes the set of all permutations of K elements, i.e. $\mathcal{P}_K \triangleq \{x_{(i)} \neq x_{(j)} \forall i \neq j\}$). It is 3 times more likely to get a sequence with a "smaller" category at the start of the sequence than at the end.

$$p(\mathbf{X} = \mathbf{x}) = \begin{cases} \frac{1}{2K!} & \text{if } \mathbf{x} \in \mathcal{P}_K, x_{(1)} > x_{(K)} \text{ (rare)} \\ \frac{3}{2K!} & \text{if } \mathbf{x} \in \mathcal{P}_K, x_{(1)} < x_{(K)} \text{ (likely)} \\ 0 & o.w. \end{cases} \tag{13}$$

This synthetic dataset is designed to emulate believable characteristics of real world dataset. In practice, the distribution that we wish to model are believed to have a support on a small fraction of the totality of probability space. Whether we are trying to generate text, images or proteins, the likelihood of stumbling across a "valid" sample from a uniform distribution is extremely unlikely.

Research Questions/Outline for Anja

- We need to define a procedure to compare the gen. models. as a start, algorithm 1.
- Can we use this procedure to **efficiently** compare generative models? Meaning, given two sets of generated samples:
  - How many samples m are needed?
  - How large can $|\Omega|$ be?
  - What theoretical gar. can we have on this result?
- If the requirements on $m, |\Omega|$ are too restrictive, can we simplify the problem by reducing the induced space of the uniform distribution $|\mathcal{U}|$?
  - At what cost (do we still have the same theoretical guarantees, is the procedure still useful)
  - Can we still use this procedure?
- If the requirements on $m, |\Omega|$ are too restrictive, can we lower the power of the test and only try to differentiate between ok and terrible?

1.2 Given samples from the ground truth

One major capability of interest of a generative model is proper grasp of frequencies of second and higher order pattern of elements of the sequence. Those metrics are used in the generative proteins sequence modeling (GPSMs) literature as an evaluation baseline metric [5, 6]. They compare the empirical covariance of the element of such patterns in the generated samples to that of the ground truth samples.

The empirical positional frequency of a pattern of length $p < S$ $\hat{f}^{s_1,\dots,s_p}_{k_1,\dots,k_p}$ is the count of samples that contains the pattern of categories $k_1, \dots, k_p \quad k_i \in \mathcal{C}$ at positions $s_1, \dots, s_p, \quad s_i \in \{0, S-1\}$ in the sequence.

$$\hat{f}^{s_1,\dots,s_p}_{k_1,\dots,k_p}(\tilde{\mathbf{x}}^M) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[\hat{x}^i_{(s_1)} = k_1, \dots, \hat{x}^i_{(s_p)} = k_p] \tag{14}$$

A "pattern covariance" is obtained with $C^{s_1,\dots,s_p}_{k_1,\dots,k_p}(\tilde{\mathbf{x}}^M) = \hat{f}^{s_1,\dots,s_p}_{k_1,\dots,k_p}(\tilde{\mathbf{x}}^M) - \prod_{j=1}^{p} \hat{f}^{s_j}_{k_j}(\tilde{\mathbf{x}}^M)$. We compare two set of samples, one from our ground truth distribution $\mathbf{x}^N$ and one generated by our model $\tilde{\mathbf{x}}^M$ by taking the Pearson correlation coefficient between the "pattern covariance" of a set of patterns.

For patterns of length $p = 2$, the set of all "pattern covariance" $\{C^{s_1,s_2}_{k_1,k_2}(\tilde{\mathbf{x}}^M)\}; \forall s_1, s_2 \in \{0, S\}, k_1, k_2 \in \mathcal{C}\}$ is not too large ($S^2 \times K^2$), so we can compute them all.

For longer patterns $p > 2$, it gets too large. We select a subset of all patterns by following the procedure described in [6], which focus on the most likely pattern. The sampling procedure can be viewed in the Appendix **??**.

Hamming distance As a last metric of interest, we can compare the **sample variety** of $\tilde{\mathbf{x}}^M$ to the ground truth samples $\mathbf{x}^N$.

1- We first compute the hamming distance between every pair of sequence in the samples of both $\tilde{\mathbf{x}}^M$ and $\mathbf{x}^N$:

$$\mathbf{H}_{i,j} = hamming(x_i, x_j) \quad \forall x_i, x_j \in \mathbf{x}^N i \neq j, \quad \mathbf{H} \in \mathbb{N}^{N \times N} \tag{15}$$

$$\tilde{\mathbf{H}}_{i,j} = hamming(\tilde{x}_i, \tilde{x}_j) \quad \forall x_i, x_j \in \tilde{\mathbf{x}}^M i \neq j, \quad \tilde{\mathbf{H}} \in \mathbb{N}^{M \times M} \tag{16}$$

2- Then we obtain the empirical frequency of hamming distances within $\mathbf{H}$ and $\tilde{\mathbf{H}}$

$$\tilde{p}^{emp}_{dist}(d) = \frac{2}{M(M-1)} \sum_{i<j}^{M} \mathbb{1}[\tilde{\mathbf{H}}_{i,j} = d] \tag{17}$$

$$p^{emp}_{dist}(d) = \frac{2}{N(N-1)} \sum_{i<j}^{N} \mathbb{1}[\mathbf{H}_{i,j} = d] \tag{18}$$

3 - The final metric is to compute the total variation between those two distributions:

$$TV_{var} = TV(p^{emp}_{dist}, \tilde{p}^{emp}_{dist}) \tag{19}$$

2 Theory on generative capability issues

People realized that some generative models gave higher likelihood to images that were coming from other datasets [7].

- Understanding Failures in Out-of-Distribution Detection with Deep Generative Models `https://arxiv.org/pdf/2107.06908.pdf`. [Flo : This work investigate the typical OOD tests, and proposes that the failure of PGM is caused by model misestimation.] The results of our analysis suggest that it is the model that is at fault, not the method for OOD detection. We additionally highlight the importance of formalizing the out-distributions of interest for OOD detection in general, as well as the arbitrary choice of the typical set for OOD detection
- Why Normalizing Flows Fail to Detect Out-of-Distribution Data `https://arxiv.org/pdf/2006.08545.pdf`
- On Out-of-distribution Detection with Energy-based Models `https://arxiv.org/pdf/2107.08785.pdf`

2.1 Link diffusion model and gen. problems.

- TACKLING THE GENERATIVE LEARNING TRILEMMA WITH DENOISING DIFFUSION GANS `https://arxiv.org/pdf/2112.07804.pdf` [Flo : Also justifies dropping the normal assumption for small time step.]

3 Bibliography

References

[1] V. Nagarajan, A. Andreassen, and B. Neyshabur, “Understanding the failure modes of out-of-distribution generalization,” in *Proc. Int. Conf. Learning Representations ICLR*, 2021.

[2] L. Theis, A. van den Oord, and M. Bethge, “A note on the evaluation of generative models,” in *Proc. Int. Conf. Learning Representations ICLR*, 2016.

[3] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price, “Optimal testing of discrete distributions with high probability,” in *Proc. ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2021, New York, NY, USA, 2021, p. 542–555. [Online]. Available: https://arxiv.org/pdf/1708.02728.pdf

[4] O. Goldreich, *The Uniform Distribution Is Complete with Respect to Testing Identity to a Fixed Distribution.* Springer International Publishing, 2020, pp. 152–172.

[5] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, “Efficient generative modeling of protein sequences using simple autoregressive models,” *Nature Communications*, vol. 12, 2021.

[6] F. McGee, S. Hauri, Q. Novinger, S. Vucetic, R. Levy, V. Carnevale, and A. Haldane, “The generative capacity of probabilistic protein sequence models,” *Nature Communications*, vol. 12, 2021.

[7] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, “Do deep generative models know what they don’t know?” in *Proc. Int. Conf. Learning Representations ICLR*, 2019.