Variational Autoencoder

The structure of a Variational Autoencoder (VAE) consists of an Encoder and Decoder, where the Encoder reduces the dimensionality of the input data from the initial, high dimensional space, e.g. $\mathbb{R}^n$, into a lower dimensional latent space $\mathbb{R}^d$ with $d << n$. This can be done by convolutional or linear layers. Given the latent representation the Decoder tries to reconstruct the datapoint, e.g. with transposed convolutional layers. In comparison to Standard Autoencoders, VAEs regularize the latent space, enforcing the latent codes to look e.g. standard gaussian, from which we can sample from to generate new data.

Motivation

Variational Autoencoders (VAEs) are used for

- Anomaly Detection
- Generating new data
- Dimensionality reduction for huge Datasets (e.g. in Finance or Bio-Processing)
- Data representation tasks such as image denoising, image segmentation or super-resolution

Variational Bayesian

- Assumptions: Dataset $\{x_i\}_{i=0}^N$ of N i.i.d. samples is generated by a random process, involving an unobserved continuous random variable z, the latent variable. [1]
- The process consists of two steps: First a value $z^{(i)} \sim p_{\theta^*}(z)$ is sampled from the prior and then a value $x^{(i)} \sim p_{\theta^*}(x|z)$ is generated from the conditional distribution. θ^* are the true and unknown parameters. [1]
- The true posterior $p_\theta(z|x)$ and thus the Integral of the marginal likelihood

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

  is intractable. This is very common in Neural Networks due to None-Linearities. [1]
- In order to approximate the unknown parameters, we define the probabilistic Encoder $q_\phi(z|x)$ as an approximation of the intractable true posterior and the probabilistic Decoder $p_\theta(x|z)$, since given a code z it produces a distribution over the possible values of x, e.g. a Normal/ Bernoulli Distribution for continuous/ binary Data. [1]

The Evidence Lower Bound - ELBO

- In general we want to maximize the (log-) marginal likelihood. As the integral is intractable we can find a lower bound which we can then optimize to approximate our parameters ϕ, θ. The ELBO is defined as [1]

$$\log(p_\theta(x)) \geqslant \mathcal{L}(\theta, \phi|x) = -\mathcal{D}_{KL}[p_\theta(z) \parallel q_\phi(z|x)] \\ + \mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))]$$

- The first term is the Kullback-Leibler divergence between the approximate posterior $q_\phi(z|x)$ and the prior $p_\theta(z)$, which we in most cases fix to be a Standard Normal $\mathcal{N}(0, I)$. It encourages the codes z to look Gaussian and acts as a regularization term.
- The second term $\mathbb{E}_{q_\phi(z|x)[}\log(p_\theta(x|z))]$ is the expected log-likelihood of the observed x given the code z that we have sampled. One can think of it as a Reconstruction Term.

Sample from latent Space - Reparametrization

- Let the Encoder return two vectors $\mu, \sigma \in \mathbb{R}^n$ with latent dimension $n \in \mathbb{N}$, which we define as the mean and variance parameters for our posterior $q_\phi(z|x) = \mathcal{N}(\mu, \sigma^2)$. [1]
- Instead of directly sampling from this posterior, we use a reparametrization trick in order to be able to backpropagate through the network w.r.t. μ, σ
- For that we sample $\varepsilon \sim \mathcal{N}(0, I)$, where $0 \in \mathbb{R}^n$ and $I \in \mathbb{R}^{n \times n}$ the identity matrix and then reparametrize z as

$$z = \mu + \sigma \odot \varepsilon,$$

  where $\odot$ refers to the element-wise multiplication. [1]

Robin Mittas
Technische Universität München
Computation, Information and Technology (MA)
Applied Mathematics

Common Approaches

1. Classic Gaussian VAE

- Let the prior $p_\theta(z) = \mathcal{N}(0, I)$ be an n-dimensional Multivariate Standard Gaussian and $n \in \mathbb{N}$ the latent dimension
- Assume the likelihood $p_\theta(x|z) = \mathcal{N}(\hat{x}, I)$ is Gaussian with Unit-Variance, where $\hat{x} \in \mathbb{R}^D$ is the output of the VAE for some input x

The ELBO Term can be rewritten for some constant $C \in \mathbb{R}$ and $D \in \mathbb{N}$ the dimension of the data with the help of Monte-Carlo Estimates as

$$\frac{1}{2}\mathrm{MSE}(\hat{x}, x) - \frac{1}{2}(1 + \log(\sigma) - \mu^2 - \sigma) + DC$$

which we can use as the Loss Function to optimize. A potential limitation is that the regularization and the reconstruction terms could be on very different scales and that we assumed Unit Variance for the likelihood.

2. β-VAE

- Introduction of weight $\beta \in \mathbb{R}$ for KL-Divergence Term, which results in following Loss function [3]

$$\mathcal{L}_\beta(\hat{x}, x) = \mathrm{Reconstruction}(\hat{x}, x) - \beta \mathcal{D}_{KL}[q(z|x) \parallel p(z)].$$

- The goal is for some $\beta >> 1$ to have a disentangled representation, such that each latent dimension learns its own features as we force the posterior to be closer to the prior. citedisentangled-vae
- The weight β of the KLD Loss is a hyperparameter which needs to be tuned.

3. σ-VAE

- In this type of VAE $p_\theta(x|z) \sim \mathcal{N}(\mu(z), \sigma(z)^2)$ is assumed, instead of a unit variance. This parameterization of the decoding distribution outputs one variance value per each pixel and channel. [4]
- While powerful, it has been observed, that this approach attains suboptimal performance, and is moreover prone to numerical instability. [4]
- The optimal results, also due to efficiency could be achieved with a covariance matrix with a single shared parameter $\sigma \in \mathbb{R}$ defined as [4]

$$\sigma^* = \log\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(\mathrm{MSE}(\hat{x}_i, x_i, reduction = mean))}\right).$$

- The Loss function for the specified σ^* is $\mathcal{L}(\hat{x}, x) = \frac{D}{2}\frac{\mathrm{MSE}(\hat{x}, x)}{exp((\sigma^*)^2)} + D\sigma^* - \mathcal{D}_{KL}[q(z|x) \parallel p(z)].$

Limitations and Challenges

- Dimension of the Latent Space is a hyperparameter which needs to be tuned carefully: Wrong choices can lead to over-/ underfitting.
- Tendency to generate blurry, unrealistic images
- Assumption that the variational posterior distribution $q_\theta(z|x)$ follows an isotropic Gaussian. VAE cannot guarantee that the inference algorithm will converge onto the standard Gaussian prior if the k latent neurons $z = (z_1, ..., z_k)$ are in fact correlated.
- Assumption that prior is standard gaussian: We try to approximate the posterior to also be standard gaussian. This leads to potential limitations in the flexibility of our posterior approximation.

State of the Art Approaches

- Inverse Autoregressive Flow VAEs: Building flexible posterior distributions $q_\theta(z|x)$ through an iterative procedure by transforming the initial simple posterior to a more complex one. [2]
- Ladder VAE: Modelling more complex Priors though multiple hierarchical stochastic layers shared between Encoder and Decoder. [5]
- NVAE: A deep hierarchical VAE, Generating Diverse High-Fidelity Images with VQ-VAE-2.

References

[1] Diederik P Kingma und Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[2] Diederik P. Kingma, Tim Salimans und Max Welling. “Improving Variational Inference with Inverse Autoregressive Flow”. In: *CoRR* abs/1606.04934 (2016). arXiv: 1606.04934. URL: http://arxiv.org/abs/1606.04934.

[3] Emile Mathieu u. a. *Disentangling Disentanglement in Variational Autoencoders*. 2018. DOI: 10.48550/ARXIV.1812.02833. URL: https://arxiv.org/abs/1812.02833.

[4] Oleh Rybkin, Kostas Daniilidis und Sergey Levine. “Simple and Effective VAE Training with Calibrated Decoders”. In: *CoRR* abs/2006.13202 (2020). arXiv: 2006.13202. URL: https://arxiv.org/abs/2006.13202.

[5] Casper Kaae Sønderby u. a. *Ladder Variational Autoencoders*. 2016. DOI: 10.48550/ARXIV.1602.02282. URL: https://arxiv.org/abs/1602.02282.

For some own Coding examples refer to github.com/robinmittas/variational-autoencoder.

Robin Mittas
Technische Universität München
Computation, Information and Technology (MA)
Applied Mathematics