Variational Autoencoder

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

Motivation

Variational Autoencoders (VAEs) are used for

- Anomaly Detection
- Generating new data
- Dimensionality reduction for huge Datasets (e.g. in Finance or Bio-Processing)
- Maximum-Likelihood $p(\mathcal{D}|\theta)$ and Maximum a Posteriori estimation $p(\theta|\mathcal{D})$ where $\mathcal{D}$ is some Data and θ the parameters of some prior distribution $p(\theta)$
- Data representation tasks such as image denoising, image segmentation or super-resolution

Variational Bayesian

- Structure: Reduce dimension of input data (Encoder) e.g. by convolutional layers into some latent representation and try to reconstruct latent code back into initial space (Decoder), e.g. by transposed convolutional layers.
- Setting: Data generated by random process, involving an unobserved continuous random variable z (the latent variable)
- A Value $z^{(i)}$ is generated from the prior distribution $p_{\theta^*}(z)$ and $x^{(i)}$ is generated from conditional distribution $p_{\theta^*}(x|z)$ (likelihood) where θ^* are the true/ unknown parameters
- Intractability: 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.
- Define the probabilistic Decoder $p_\theta(x|z)$ and the probabilistic Encoder $q_\phi(z|x)$ as an approximation of the intractable true posterior $p_\theta(z|x)$.

The Evidence Lower Bound - ELBO

In general we want to maximize the (log-) marginal likelihood for which we can find the following bound

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

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

Sample from latent Space - Reparametrization

- Let the Encoder return two vectors $\mu,\sigma \in \mathbb{R}^n$ with latent dimension $n \in \mathbb{N}$
- Imagine μ,σ to be the mean and variance vectors, e.g. the parameters for a Normal Distribution

We then 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 reparametrize $z \sim \mathcal{N}(\mu,\sigma^2)$ as

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

where $\odot$ refers to the element-wise multiplication. We can now easily backpropagate through the Network w.r.t. μ and σ.

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

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

Then we can rewrite the ELBO 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

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

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

β-VAE

- Introduction of weight $\beta \in \mathbb{R}$ for KL-Divergence Term
- 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
- Keep in mind: The weight β of the KLD Loss is a hyperparameter which needs to be tuned
- The objective Loss function is

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

σ-VAE

- Let us now not assume Unit Variance for the likelihood, instead let $p_\theta(x|z) \sim \mathcal{N}(\mu(z), \sigma(z)^2)$
- 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

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

- The Reconstruction Term of the ELBO then satisfies

$$\mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))] = \frac{D}{2\sigma^2}\text{MSE}(\hat{x}, x) + D\log(\sigma) + C$$

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. We can just model quite simple gaussian posterior distributions.

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.
- Ladder VAE: Modelling more complex Priors though multiple stochastic layers which are shared between Encoder and Decoder
- Further readings: NVAE: A deep hierarchical VAE, Generating Diverse High-Fidelity Images with VQ-VAE-2

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