Autoencoding Variational Bayes

Mohamed Amine GRINI, Marine VIEILLARD

Introduction

Context:

- Article from 2012, in the development of autoencoders that date back to 1986
- Before its publication, approaches such as classical variational inference and MCMC methods were used, but these were often too expensive and unstable to be applied with large neural networks
-

$\longrightarrow$ Founding article of Variational Auto Encoder (VAE)

Introduction

Objective: resolve three major limitations

- difficulty in calculating or differentiating marginal likelihood
- the necessity of using restrictive analytical approximations in classical variational inference
- the impossibility of performing efficient gradient learning in deep probabilistic models

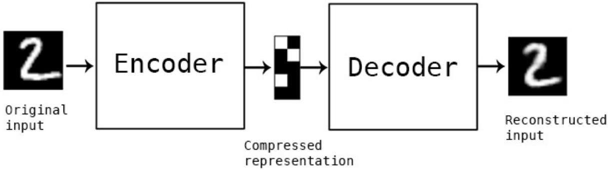
Introduction of key contributions :

- Reparametrization trick
- Stochastic Gradient Variational Bayes estimator (SGVB)

$\longrightarrow$ Efficient alternative to the Monte Carlo EM which cannot be used when the posterior density is intractable, and would be too slow applied on large datasets

What is an Autoencoder ?

First appeared in 1986 *[1]*, an autoencoder is a neural network that learns to compress data into a lower-dimensional representation (the latent space) and reconstruct it.



AEs limitations

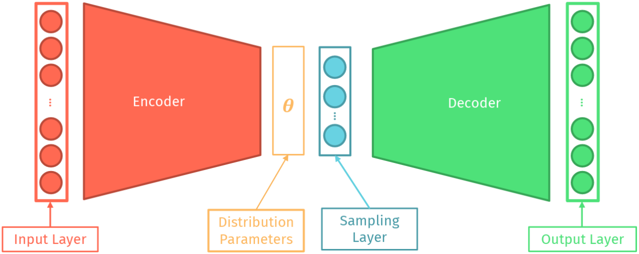
Deterministic Encoding

Autoencoders provide a deterministic mapping from input to latent space, meaning each input is mapped to a fixed point in the latent space.

This can limit their ability to generate diverse and novel samples.

Variational Autoencoders

Appeared in 2013 (*Kingma et al.[2]*), learns the distribution parameters to generate diverse samples. Instead of a single latent representation z, VAEs model a probability distribution over z



What changes from regular AE ?

- We force a structured latent space using variational inference.
- Instead of encoding into a single point z, we encode into the probability distribution $p_\theta(z|x)$

Variational Inference

Bayes' Theorem:

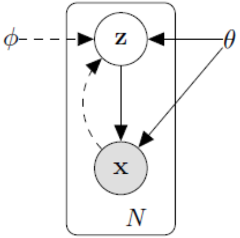
$$p_\theta(z|x) = \frac{p_\theta(z,x)}{p_\theta(x)} = \frac{p_\theta(z,x)}{\int p_\theta(x|z)p_\theta(z)\,dz}$$

Integral $\int p_\theta(x|z)p_\theta(z)\,dz$ *is intractable (very high computational cost) due to integrating over all possible values of* z

Approximation:

$$q_\phi(z|x) \approx p_\theta(z|x)$$

q (parameterized by ϕ, usually $q_\phi \sim \mathcal{N}(\mu, log\sigma^2)$) is an approximate posterior to p (parameterized by θ)



Variational Bound

Kullback-Leibler Divergence:

$$KL(P||Q) = \int P(x) log(\frac{P(x)}{Q(x)})\, dx = \mathbb{E}_P \left[log \left(\frac{P(x)}{Q(x)} \right) \right]$$

Properties:

- Distance metric not symmetric
- Always ≥ 0
- Equal to 0 if and only if $P = Q$

Goal: Maximizing the marginal likelihood $log \ p_\theta(x)$

$$\Rightarrow logp_\theta(x) \geq \mathcal{L}(\theta, \phi, x)$$

Variational Lower Bound/ELBO:

$$\mathcal{L}(\theta, \phi, x) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[log(p_\theta(x|z))\right]}_{\text{Maximize the reconstruction quality}} - \underbrace{KL(q_\phi(z|x)||p_\theta(z))}_{\text{Minimize the distance}}$$

We want to differentiate and optimize the lower bound $\mathcal{L}(\theta, \phi, x)$ w.r.t. both the variational parameters ϕ and generative parameters θ. The KL-Divergence can be calculated analytically:

$$-KL(q_\phi(z|x)||p_\theta(z|x)) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j^2)\right)$$

The expected reconstruction error $\mathbb{E}_{q_\phi(z|x)}\left[log(p_\theta(x|z))\right]$ requires estimation by sampling.

Monte Carlo Sampling: $\mathbb{E}_{q_\phi(z|x)}[f(z)] \approx \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})$, where:

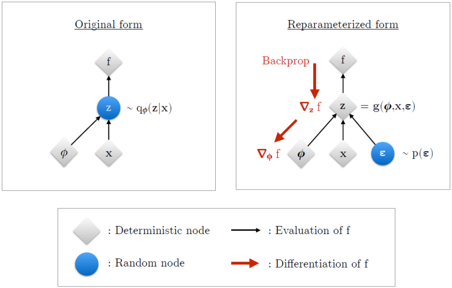
- $z^{(l)}$ are samples drawn from variational distribution
- L the number of Monte Carlo Samples

$$\Rightarrow \tilde{\mathcal{L}}^A(\theta, \phi, x^{(i)}) = -KL(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)}) + \frac{1}{L}\sum_{l=1}^{L}(\log(p_\theta(x^{(i)}|z^{(i,l)})))$$

This is called **Stochastic Gradient Variational Bayes (SGVB)** estimator $(\mathcal{L}(\theta, \phi, x) \simeq \tilde{\mathcal{L}}^A(\theta, \phi, x^{(i)}))$

Optimizing the lower bound

(Naïve) Monte Carlo Gradient:
$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] \simeq \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z})\nabla_{q_\phi(\mathbf{z}^{(l)}|x)} \log q_\phi(\mathbf{z}^{(l)}|x)$
This gradient estimator exhibits very high variance (*Jordan et al.[3]*), cannot backpropagate through $q_\phi(z|x)$ because it's a **stochastic function** of ϕ
Solution: Reparameterization Trick



Reparametrization Trick

Express random variable z as a deterministic variable:

$$z = g_\phi(\epsilon, x)$$

with g_ϕ a differentiable transformation and ϵ an auxiliary variable

Three approaches:

- "location-scale" family of distributions, choose the standard distribution as ϵ and $g(.) = location + scale * \epsilon$
- tractable inverse CDF, let $\epsilon \sim \mathcal{U}[0, I]$ and let $g_\phi(\epsilon, x)$ the inverse CDF of $q_\phi(z|x)$
- composition, compose transformations used in the previous points

$$\tilde{\mathcal{L}}^B(\theta, \phi, x^{(i)}) = -KL(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)})) + \frac{1}{L}\sum_{l=1}^{L}(\log(p_\theta(x^{(i)}|z^{(i,l)})))$$

where $\epsilon \sim p(\epsilon)$ and $z^{(i,l)} = g_\phi(\epsilon^{(i,l)}, x^{(i)})$

Second version of **SGVB** estimator with less variance :
Instead of using the whole dataset X (N datapoints), using a minibatch X^M ($M < N$ datapoints) is computationally more efficient:

$$\mathcal{L}(\theta, \phi, X) \simeq \tilde{\mathcal{L}}^M(\theta, \phi, X^M) \simeq \frac{N}{M}\sum_{i=1}^{M}\tilde{\mathcal{L}}^B(\theta, \phi, X^M)$$

AEVB Algorithm

Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators can be used.

Initialize: θ, ϕ

Repeat:

- $X^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
- $\epsilon \leftarrow$ Random samples from noise distribution $p(\epsilon)$
- $g \leftarrow \nabla_{\theta,\phi} \hat{\mathcal{L}}^M(\theta, \phi; X^M, \epsilon)$
- $\theta, \phi \leftarrow$ Update parameters using gradients g (e.g., SGD or Adagrad)

Until: Convergence of parameters (θ, ϕ)

Return: θ, ϕ

Experiments

We trained generative models of images from the **MNIST** and **CIFAR-10** datasets (both are continuous data) and compared our results with those of the article.

Reconstruction

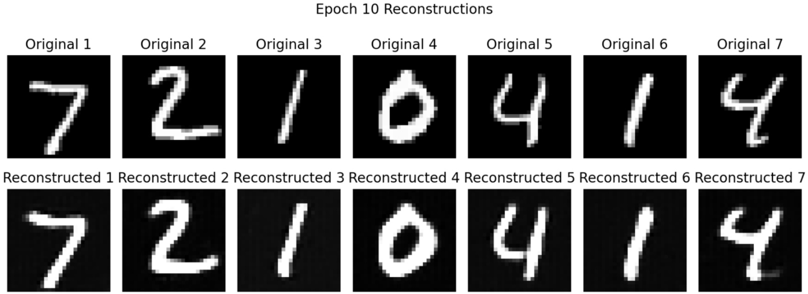
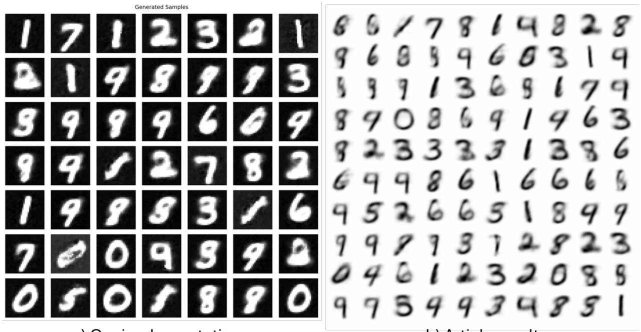


Figure: Reconstruction at the end of epoch 10

Generation :



a) Our implementation   b) Article result

Figure: Generation result for a latent space with 2 dimensions

Generation :

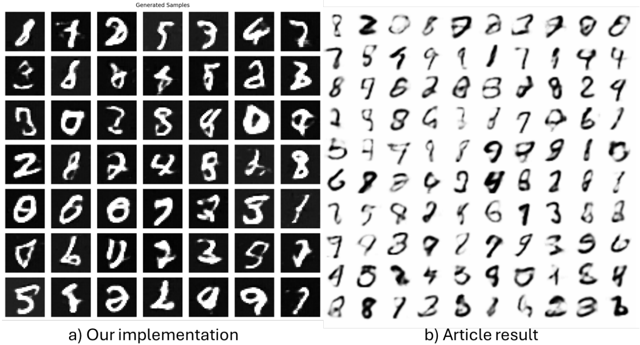


Figure: Generation result for a latent space with 20 dimensions

References

[1] Ronald J. Williams David E. Rumelhart Geoffrey E. Hinton. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536. DOI: https://doi.org/10.1038/323533a0.

[2] Max Welling Diederik P. Kingma. "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114* (2013). DOI: https://doi.org/10.48550/arXiv.1312.6114.

[3] Michael Jordan John Paisley David Blei. "Variational Bayesian Inference with Stochastic Search". In: *ICML 2012* (2012). DOI: https://doi.org/10.48550/arXiv.1206.6430.