ρ-VAE:
Autoregressive parametrization of the VAE encoder

Sohrab Ferdowsi, **Shideh Rezaeifar**, **Maurits Diephuis**, and **Slava Voloshynovkiy**

Department of Computer Science, University of Geneva, Switzerland
{sohrab.ferdowsi, shideh.rezaeifar, maurits.diephuis, svolos}@unige.ch

Abstract

We make a simple, but very effective plug-and-play alteration to the standard VAE. This is about parametrization of the sample-dependent latent space as a Gaussian distribution whose covariance matrix is constructed as an AR(1) process. We argue that this is a more natural prior for images to consider, as compared to the standard Gaussian distribution with a diagonal covariance matrix. While the standard diagonal parametrization consists of learning the mean vector, as well as the vector of diagonal values for each sample, our parametrization consists of the mean and two scalar values s and ρ, where the first one is scaling the distribution and the second one is a measure of correlation among the latent dimensions. Therefore, the performance boost which we show in our experiments to be consistently noticeable cross different setups and variants of VAE models, comes at no cost, even reducing the number of covariance parameters from d, the dimensionality of the latent space to 2.

1 Introduction

Arguably, one of the most successful approaches to representation and generative learning is that of "Auto-encoding variational Bayes" [1], a parametrization of the standard variational Bayes in the form of an autoencoder neural network with a recipe for end-to-end learning of its parameters, while providing effective approximation of the intractable posterior. This has then given rise to the very popular Variational AutoEncoder (VAE), a set of models

bla bla

bla bla bla

2 The VAE models

Here we briefly review the standard VAE model, highlighting aspects relevant to our work, as well as some of its further developments.

In a typical probabilistic model where a latent variable $\mathbf{z} \in \Re^d$ is the underlying factor to generate the observable samples $\mathbf{x}$'s $\in \Re^n$, the standard variational Bayes [?] paradigm is concerned with finding an approximation $q(\mathbf{z})$ for the intractable posterior $p(\mathbf{z}|\mathbf{x})$. This is achieved by minimizing the Kullback-Leibler divergence between these two distributions , i.e., $D_{\mathrm{KL}}\Big[q(z)||p(z|x)\Big]$. Standard

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

treatments of this quantity, along with its non-negativity property will then amount to the following inequality:

$$\log(p(\mathbf{x})) \leqslant \mathbb{E}_{q(\mathbf{z})}\Big[\log(p(\mathbf{x}|\mathbf{z}))\Big] - D_{\mathrm{KL}}\Big[q(\mathbf{z})||p(\mathbf{z})\Big]. \tag{1}$$

Autoencoding variational Bayes is then constructing an explicit dependence of the latent variables to the i^{th} training sample by considering a parametrized distribution $q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$, whose construction resembles the encoder part of an autoencoder network with a set of learnable weights ϕ. Furthermore the training samples can be decoded with $p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})$, another network with parameters symbolized as θ.

Making this double-sided data dependence more explicit, and by summing over all N training samples results to the following inequality:

$$\frac{1}{N}\sum_{i=1}^{N}\log(p(\mathbf{x}^{(i)})) \leqslant \frac{1}{N}\sum_{i=1}^{N}\Big[\log(p(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})) - D_{\mathrm{KL}}\big[q(\mathbf{z}^{(i)}|\mathbf{z}^{(i)})||p(\mathbf{z})\big]\Big]. \tag{2}$$

This, in fact, is highly relevant for generative modeling as the log-likelihood of the training samples will be upper bounded by two terms, both of which amenable to mini-batch optimization with stochastic gradient descent.

However, this comes with a major caution, the fact that sampling $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$ is a non-differentiable operation. The work-around for this issue is the wise “reparametrization trick”, as proposed in [1].

The idea is to create randomness through a fixed distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$, where I_d is the unity matrix of dimension d. The samples of the appropriate distribution can then be generated by injecting the learnable moments, e.g., using $\mathbf{z}^{(i)} = \boldsymbol{\mu}^{(i)} + \tilde{\mathrm{C}}^{(i)}\boldsymbol{\epsilon}$, where $\boldsymbol{\mu}^{(i)}$ is the mean vector of the posterior learned for the i^{th} sample and $\tilde{\mathrm{C}}^{(i)}$ is the Choleskiy decomposition of the corresponding covariance matrix $\mathrm{C}^{(i)}$.

This then limits the choices for

3 The ρ-VAE

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.