EECS 6322 Week 9, Paper 1
Fixing a Broken ELBO

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes an alternative framework for the ELBO (evidence lower bound) with different quantitative and qualitative characteristics. The proposed framework provides an upper and lower bound on the mutual information between the latent variable and the input variable. It characterises the tradeoff between the reconstruction accuracy and compression by deriving a rate-distortion curve using the upper and lower bounds. It also suggests a new method to prove that the latent variable models with powerful stochastic decoders take better consideration of their latent code.

A common method to learn a useful representation of data in an unsupervised approach is to fit a latent variable model to the data by maximizing the likelihood training, or maximizing the lower bound on the likelihood training, such as the evidence lower bound (ELBO), as used in the variational autoencoders (VAEs). Alternatively, other divergent methods such as the reverse KL is used to fit certain kinds of generative adversarial networks (GANs). The main problem with these loss functions is that they do not measure the quality of the representation. This is because the loss functions primarily depend on $p(x|\theta)$, and not on $p(x,z|\theta)$. Thus, if we use a stochastic decoder, we may obtain a good ELBO score or rather a good marginal likelihood, but we still may not obtain a good representation. Thus, to resolve this problem the paper proposes mutual information I between the latent variable Z and the input variable X to assess the value of representational learning. Since this mutual information can be intractable, an upper and lower bounds are computed which give the value of the representational learning. The proposed framework also shows that VAEs with powerful autoregressive decoders can target certain points in the rate-distortion curve to pay more attention to their latent code and that it is possible to recover the "true generative process" of a simple model only with the knowledge of the mutual information I.

The proposed framework defines a joint density for unsupervised representation learning given by $p_e(x,z) = p(x)\ e(z|x)$ where $e()$ is an encoder, z is the latent variable, and x is the input. Based on the joint density, the mutual information is given by: $I_e(X;Z) = \lim\lim dxdzp_e(x,z)log\frac{p_e(x,z)}{p(x)p_e(z)}$. Since it is difficult to compute the marginal density and the true data density $p(x)$, the authors leverage tractable variational bounds on mutual information to compute the variational lower and upper bounds given by: $HDI_e(X;Z)R$ where $H \equiv \lim dxp(x)logp(x)$, $D \equiv \lim dxp(x)Zdz\ \ e(z|x)logd(x|z)$, and $R \equiv \lim dxp(x)\lim dze(z|x)log\frac{e(z|x)}{m(z)}$. Here in the given equations, $m(z)$ is a variational approximation to $p_e(z)$ and $d(x|z)$ is a variational approximation to $p_e(x|z)$. The variables H or the data entropy measures the complexity of the dataset, D or the distortion is equal to the reconstruction negative log likelihood, and R or the rate is the average relative KL divergence between the encoding distribution and the learned marginal approximation. All variables H, D, and R are non-negative in nature.

Thus, the upper and lower bounds of the mutual information divide the rate-distortion plane into feasible and infeasible regions, where distortion is given on the y-axis and rate is given on the x-axis. Hence, on the x-axis the distortion is equal to zero which signifies that the data can be perfectly encoded or decoded. This is known as the auto-encoding limit. The lowest possible rate in the feasible region with distortion equal to zero is given by H, which is the entropy of the data. This point H corresponds to the point $(R = H, D = 0)$. On the contrary, the y-axis refers to the state when the rate is equal to zero which implies $e(z|x) = m(z)$. Thus, the encoding distribution $e(z|x)$ is independent of x which means that the latent representation is not encoding any information about the input. To resolve this, the authors use a suitably powerful decoder, $d(x|z)$, that is able to capture correlations between the components of x and hence reduce the distortion to the lower bound of H given by the point $(R = 0, D = H)$ which is also known as the auto-decoding limit. Thus, the solutions along the diagonal line, from H on the y-axis and H on the x-axis, satisfy the equation $D = HR$ which mean both the bounds are tight and hence $m(z) = p_e(z)$ and $d(x|z) = p_e(x|z)$.

The authors also propose Legendre transformation, as opposed to tracing out the optimal distortion as a function of the rate $D(R)$, that can find the optimal rate and distortion for a fixed $\beta = \frac{\partial D}{\partial R}$ by minimizing $min_{e(z|x),m(z),d(x|z)}D + \beta R$. When β is set to 1, the Legendre transformation objective matches the ELBO objective used when training a VAE with the distortion term matching the reconstruction loss, and the rate term matching the "KL term". However, the ELBO objective alone cannot distinguish between models that make use of the latent variable and learn useful representations for reconstruction versus models that make no use of the latent variable, in the infinite model family. In the finite model family, the ELBO targets a single point along the rate distortion curve, which is the point with slope 1. This point depends on the model architecture and its respective encoder, decoder and *marginal. Thus, for a general $\beta \geq 0$, the β-VAE objective is obtained which smoothly interpolates between auto-encoding behavior $(\beta \ll 1)$ to auto-decoding behavior $(\beta \gg 1)$. Thus, the model interpolates between a point where the distortion is low but the rate is high to a point where the distortion is high but the rate is low, all without having to change the model architecture.*

The framework is evaluated on the binary MNIST dataset to show that comparing models in terms of rate and distortion separately is more useful than simply observing marginal log likelihoods. The framework is applied on different simple and complex variants for the encoder and decoder, and three different types of marginal, which totals to 12 models. All 12 models are trained to minimize the β-VAE objective and the results include the converged rate-distortion location for a total of 209 distinct runs across the 12 architectures, with different initializations and βs on the binary MNIST dataset. The rate-distortion curve was evaluated for the dataset and the best ELBO achieved was $\widehat{H} = 80.2$ nats, at $R = 0$, which set an upper bound on the true data entropy H for the static MNIST dataset. The 12 model families considered in the results performed worse in the auto-encoding limit of the rate-distortion plane due to a lack of power in the marginal approximations.

The framework was quantitatively evaluated by sampling reconstructions and generations from some of the runs, which were grouped into categories such as autoencoders, syntactic encoders, semantic encoders, and autodecoders. When $\beta = 1.10$, the obtained R, D, and ELBO values classify the model as an autodecoder. The R value is very small and hence indicates that the decoder ignores its latent code, and the reconstructions are independent of the input x. However, the images sampled from the decoder are generated well. When $\beta = 0.1$, the obtained R, D, and ELBO values classify the model as an autoencoder, which generated nearly pixel-perfect reconstructions. When $\beta = 1.0$, the obtained R, D, and ELBO values classify the model as semantic encoder, which generated highly compressed representations that retained the semantic features of the data. When $\beta = 0.15$, the obtained R, D, and ELBO values classify the model as syntactic encoder, that retained both semantic and syntactic information of the representations, but visually degraded the generated samples. Thus, the framework shows that by using $\beta < 1$, the model can be forced to do well at reconstruction.

In conclusion, the authors propose a theoretical framework for understanding representation learning using latent variable models in terms of the rate-distortion tradeoff. The constrained optimization problem allows the models to fit the data by targeting a specific point on the RD curve, which cannot be achieved using the β-VAE framework.

II. Strengths

The paper proposes a framework that considers the quality of the representation learning of the models as compared to the ELBO method. The framework computes the optimal values of β for which the images are reconstructed well and also explains the concept of the rate-distortion curve that can be utilized to compute the optimal point on the curve for different models. The authors also proposed a solution to the expressive decoders that ignore the latent code by reducing the KL penalty term to $\beta < 1$.

III. Weaknesses

The model does not compare the different versions of the existing methodology in the results obtained in the paper.

IV. Correctness

The claims and empirical methodology of the paper are corrected based on the theoretical grounding and the results as compared to the existing prior work. The framework was proved to evaluate the representation learning of the model better than the existing methods and has a strong theoretical grounding for the framework proposed and the results obtained.

V. Clarity

The paper is well written in general with clear illustrations and comparisons between the models while evaluating the results. The paper also explained the theoretical concepts along the implementation of the framework which was later elaborated in the appendix. This helped get a clear understanding of the model and its implementation details.

VI. Relation to prior work

Prior work to improve the VAE representations include methods to reduce the problem of unused latent variables in VAEs by annealing the weight of the KL term of the ELBO from 0 to 1 over the course of training. This method however, did not consider the ending weights that differed from 1 as opposed to the framework proposed by the authors. Other works to improve the VAE representation also included the β-VAE for unsupervised learning, which is a generalization of the original VAE in which the KL term is scaled by β. This differs from the proposed framework as it focuses mainly on disentangling and does not discuss rate-distortion trade offs across model families.

Prior work in information theory and representation learning include information bottleneck frameworks that leverage information theory to learn robust representations. These frameworks allow a model to smoothly trade off the minimality of the learned representation (Z) from data (X) by minimizing their mutual information, $I(X;Z)$, against the informativeness of the representation for the task at hand (Y) by maximizing their mutual information, $I(Z;Y)$ but they differ from the proposed framework as they work only under supervised settings. Under unsupervised representational learning, several works have been proposed that reinterpreting the ELBO and present information maximization as a solution to the problem of VAEs ignoring the latent code. However, as compared to the proposed framework, the objectives in the prior work leverage techniques from implicit variational inference as the aggregated posterior is intractable to evaluate and some also *require to sample the generative model (which is slow for autoregressive models) and compute gradients through model samples (which is challenging for discrete input spaces).*

Prior work in generative models and compression include methods that utilize rate-distortion theory in compression to trade off the size of compressed data with the fidelity of the reconstruction. Some methods also leveraged deep latent-variable generative models for images, and explored tradeoffs in the RD plane. These differ from the proposed framework as the prior methods focus on a restricted set of architectures with simple posteriors and decoders and do not study the impact that architecture choices have on the marginal likelihood and structure of the representation.

VII. Reproducibility

There are enough details included in the paper in terms of elaborated theoretical grounding and code implementation in the github links that would enable us to reproduce the major components of the paper. The framework has been applied on top of several models whose details are also mentioned in the paper that can be utilized to replicate the model as tested by the authors.

VIII. Additional comments

An additional comparison or discussion on the other VAE approaches that incorporate sparsity into the latents would supplement the work proposed by the authors.