Summary

Introduction

- VAE uses reparameterization tricks for capturing rich distributions from vast amounts of data (diversity).
- Prior plays a crucial role in mediating between the generative decoder and the variational encoder.
- Choosing a too simplistic prior like the standard normal distribution could lead to over regularization and, as a consequence, a very poor hidden representation.
- Re-formulating the variational lower bound gives two regularization terms :
    - The average entropy of the variational posterior, and
    - The cross-entropy between the averaged (over the training data) variational posterior and the prior.

The cross - entropy can be minimized by setting the prior = average of the variational posteriors over the training points. But, this is computationally expensive.

Hence, the paper proposes a new prior = mixture of variational posteriors priors, or VampPrior. A new two-level VAE that combined with our new prior that can learn a very powerful hidden representation.

The contributions are :

1. New flexible prior as a mixture of variational posteriors, conditioned on learnable pseudo-data. -> learn more powerful latent representation.
2. New two - layer generative VAE Model with two layers of stochastic latent variables. -> avoids the problems of unused latent dimensions.
3. VampPrior > normal prior and Hierarchical VAE = SOTA.

Variational Auto-Encoder

- Our goal is to maximize the average marginal log likelihood of the w.r.t parameters.
- However, when the model is parameterized by NN, the optimization could be difficult due to the intractability of the marginal likelihood.
- One possible way is to use VARIATIONAL INFERENCE and optimize the lower bound (ELBO method).

We have :
$q(x)$ = empirical distribution, $q(z|x)$ = variational posterior, $p(x|z)$ = generative model, $p(z)$ = prior

For continuous z, we use Monte-carlo estimate for L samples points, “ Reparameterization of $q(z|x)$ ”.

The first term of the objective function can be seen as the expectation of the negative reconstruction error that forces the hidden representation for each data case to be peaked at its specific MAP value.

The second and the third term are a kind of regularization that drives the encoder to match the prior.

Taking the gradients of L, we find prior plays a role of an “anchor” that keeps the posterior close to it.

Typically, the encoder is assumed to have a diagonal covariance matrix q(z|x), parameterized by the NN, and the prior is expressed using a standard normal distribution.

The decoder utilizes a suitable distribution for the data under consideration, e.g., Bernoulli for binary data or the normal distribution for continuous data, and it is parameterized by NN.

The Variational Mixture of Posterior Priors

The Idea :
The variational lower bound can be re-written as :

L = Neg. reconstruction error + entropy of variational posterior + Cross - entropy between the aggregated posterior and the prior.

The second term of the objective function encourages the encoder to have larger entropy of every data case and, the last term aims at matching the aggregated posterior and the prior.

Usually, prior is chosen in advanced i.e., standard normal prior, but can also be found by maximizing L with lagrange multiplier, giving the solution as the aggregated posterior.

But, this choice may potentially lead to overfitting, and optimization becomes costly due to sum over all training points. Also, having a simple prior like the standard normal distribution is known to result in over-regularised models with only few active latent dimensions.

In order to overcome issues like -

1. Overfitting,
2. Over-regularization,
3. High computational complexity.

The optimal sol. I.e., the aggregated posterior can be further approximated by a mixture of variational posteriors with pseudo inputs.

These pseudo inputs are learned through backpropagation and can be thought as the hyperparameters of the prior, alongside parameters of the posterior.

The resulting prior is a multimodal, thus it prevents the variational posterior to be over-regularised.
On the other hand, incorporating pseudo inputs prevents potential overfitting once we pick K << N, which also makes the model less expensive to train.

-> A comparison to a mixture of Gaussian priors

The process and, influence is the same but, VampPrior has two advantages :

1. By coupling prior with posterior, we entertain less params, and the prior and the posterior will at all times “cooperate” during training.
2. The coupling highly influences the gradient wrt a single weight of the encoder. The difference becomes close to zero, as long as $q(z|x) \sim q(z|u_k)$. The gradient points towards a solution where the variational posterior has high variance, whereas VAE lowers the variance.

-> A connection to the Empirical Bayes

Idea :
Empirical bayes (type - II maximum likelihood)
In standard Bayesian stats, you pick a prior (usually, gaussian), in Empirical Bayes, you look at the data first to decide which prior to work with.

VampPrior ~ Empirical Bayes since EB uses learnable hyperparameters to fit the prior until it fits the distribution of the actual data as closely as possible.
VampPrior does this with pseudo-inputs to shape the prior, and the models learn these by maximizing the ELBO.

They claims a new kind of Inference :

1. Variational Inference : using a nn (encoder) to approximate a complex, messy distribution.
2. Empirical Bayes : using the data itself to define the starting assumption (the prior).

By combining them, VampPrior VAE doesn’t just learn how to map the data, it learns the very structure of the world the data lives in.

-> A connection to the Information Bottleneck

Idea : The idea of any good encoder is to be a smart filter.
Keep the signal and throw away the noise.

In the Information bottleneck, the aggregated posterior is the bottleneck. It is the narrowest point through which information must flow. This is similar for VampPrior VAE, which also uses aggregated posterior as variation of posterior priors as the bottleneck.

This connection is interesting since,

1. The VAEs are usually seen through the lens of Generative Modeling (making new things).
2. IB is usually seen through the lens of compression/representation learning (understanding things).

Hierarchical VampPrior Variational Auto–Encoder

-> Hierarchical VAE and the inactive stochastic latent variable problem

In VAEs, the reconstruction term (which wants to use the latent variables z to store information), and the KL term (which wants to pull the latent variables towards the prior) are battling against each other.

- The lazy model : if the KL penalty is too strong or the prior is too simple, the model finds a cheating solution. It sets the encoder to just output the prior for every input.
- The result : the latents units become “inactive”. they carry zero information about the input x, and the decoder just ignores them, usually generating blurry, generic images.

This problem worse in deep (hierarchical) VAEs :
When you stack multiple layers of the latent variables, the information from the real data has to travel up a long ladder.

- Bottom - up (Encoding) : x-> z1->z2 by the time the signal reaches z2, it is very weak.
- Top-down (Generative) : z2-> z1->x the deeper layers are far from the actual image pixels.

The z2 feels so little pressure from the real data, the KL term easily wins, and z2 collapses into the prior and becomes useless.

The solution : two-layers VampPrior VAE

A. Generative path (Top-Down) :

   $p(x, z1, z2) = p(x|z1, z2)p(z1|z2)p(z2)$

   $p(z2) = vampprior$

B. The VampPrior as safety net :
   Since, the prior in vampprior is rich and multimodal, KL term doesn't kill the units.

-> Alternative Priors

1. Standard Gaussian tests necessity of complexity
2. Mixture of Gaussians tests benefits of coupling
3. VampPrior tests value of trainability

Experiments

Setup

MLPs with two hidden layers of 300 hidden units in the unsupervised permutation invariant setting.

Results

Quantitative results :
Test marginal log likelihood (LL) BETTER in all, more latent units active than previous VAEs and simple mixture of priors

Qualitative results :
VAE tends to generate blurry images, VampPrior VAE generates sharper images.

Take Home Message :

VampPrior is a simple, powerful idea that lifts VAE performance and illuminates the importance of prior–posterior alignment — but its stability and scalability deserve deeper future study.