Appendix for One-Class Collaborative Filtering with the Queryable Variational Autoencoder (SIGIR-19)

Ga Wu[1,2], Mohamed Reda Bouadjenek[1], and Scott Sanner[1,2]

[1]Department of Mechanical and Industrial Engineering, University of Toronto
[2]Vector Institute of Artificial Intelligence
{wuga,mrb,ssanner}@mie.utoronto.ca

1 Background

In this section, we briefly review the background knowledge needed for this project. We start by summarizing Variational Auto-encoder and several project related VAE variants. We then describe the pros and cons of conditional inference methods that applicable on VAE structure.

1.1 Variational Auto-encoder and Its Variants

This project based on three important work in Variational Auto-encoder literature: VAE, CVAE and BCDE. While there exist numerous variants of VAE, the following discussion would stand in most of the cases. Thus, we would only concentrate the canonical versions.

1.1.1 Variational Auto-encoder [Kingma and Welling, 2013]

Original Variational Autoencoder optimize the Lower-bound of log probability of the full observations:

$$\begin{aligned}\log p(v) &= \log \int_z q(z|v)\frac{p(v|z)p(z)}{q(z|v)}dz\\ &\geq \int_z q(z|v)\log\frac{p(v|z)p(z)}{q(z|v)}dz\\ &= \int_z q(z|v)\log p(v|z)dz + \int_z q(z|v)\frac{p(z)}{q(z|v)}dz\\ &= E_{q(z|v)}[\log p(v|z)dz] - KL[q(z|v)||p(z)],\end{aligned} \tag{1}$$

where the prior latent distribution $p(z)$ was assumed to be standard Normal distribution. While there exists multiple works on changing this relaxation with complex latent priors [Kingma et al., 2016], in practice, most of the VAE implementation is still in its original form with expoential family due to simplicity. Alternatively, we notice that if the latent dimension is large enough, the Gaussian assumption is still valid due to no complex latent distribution needed for the representation.

1.1.2 Conditional Variational Auto-encoder [Sohn et al., 2015]

Variational Auto-encoder as a generative model cannot effectively generate observations that fit certain requirements. For example, requiring the model generate MNIST image that represents 7 is hard. Thus, Conditional Variational Auto-encoder extends the VAE family by allowing the network carrying evidence when generating observations.

More specifically, CVAE optimize the lower-bound of log conditional probability $p(y|x)$:

$$\begin{aligned}
\log p(y|x) &= \log \int_z q(z|x,y)\frac{p(y|z,x)p(z|x)}{q(z|x,y)}dz \\
&\geq \int_z q(z|x,y)\log\frac{p(y|z,x)p(z|x)}{q(z|x,y)}dz \\
&= \int_z q(z|x,y)\log p(y|z,x)dz + \int_z q(z|x,y)\frac{p(z|x)}{q(z|x,y)}dz \\
&= E_{q(z|x,y)}[\log p(y|z,x)dz] - KL[q(z|x,y)||p(z|x)]
\end{aligned} \tag{2}$$

Note that, in CVAE, $p(z|x)$ is a prior network that either pre-trained or assumed to be relaxed as p(z). The generative network $p(y|z,x)$ always take evidence x as input.

One major problem of CVAE is that people have to train the CVAE network for each particular split of evidence variables X and query variables Y, which loss the ability to do arbitrary conditional inference.

1.1.3 Bottleneck Conditional Density Estimation [Shu et al., 2017]

Bottleneck Conditional Density Estimation(BCDE) tries to conquer completely different problems with CVAE and VAE. The goal of BCDE is to estimate the conditional probability $p(y|x)$ instead of generating new observations given evidence. It is a refined version of CVAE particular for density estimation, it combines generative models and discriminative model with mutual parameter regularization. Generative models are used to improve the training effectiveness as multi-objective training.

Interestingly, for the discrimiative part of the BCDE model, it re-derived the log conditional likelihood as:

$$\begin{aligned}
\log p(y|x) &= \log \int_z q(z|x,y)\frac{p(y|z)p(z|x)}{q(z|x,y)}dz \\
&\geq \int_z q(z|x,y)\log\frac{p(y|z)p(z|x)}{q(z|x,y)}dz \\
&= \int_z q(z|x,y)\log p(y|z)dz + \int_z q(z|x,y)\frac{p(z|x)}{q(z|x,y)}dz \\
&= E_{q(z|x,y)}[\log p(y|z)dz] - KL[q(z|x,y)||p(z|x)],
\end{aligned} \tag{3}$$

where the decoder network $p(y|z)$ no longer need evidence x as input(which also suggesting it cannot to conditional observation generation as CVAE).

1.2 Conditional Inference on Variational Auto-Encoders

In this section, we review the methods that can do arbitrary conditional inference on the VAE structure. This is different with CVAE or BCDE whose evidence and query variables are pre-partitioned before training the network. Such fixed partitioning does not support arbitrary inference where evidence could be any subset of the observation variables (instead of fixed).

1.2.1 Hamilton Monte Carlo [Neal et al., 2011]

Given pre-trained VAE, the HMC does conditional inference by exploiting VAE decoder structure. More specifically, we model the conditional inference as:

$$p_\theta(y|x) = \int_z p_\theta(z,y|x) = \int_z p_\theta(z|x)p_\theta(y|z)dz. \tag{4}$$

HMC exploit this factorization by sample z along from $p(z|x)$ instead of y, z together, and then it draw exact samples from $p(y|z)$ just by forward propagation in decoder network for each sample of z.

While this method works in some simple cases, when the latent space of z is multi-mode, it fails to exploit all modes and lead into biased inference due to travel difficulty. Unfortunately, latent distribution given partial observation $p(z|x)$ is usually complex distribution with multiple near zero density area. Note that, even though VAE latent for full observations is assumed to be gaussian, for the partial observations, it is not the case. In another word, the original VAE is not friendly for conditional inference.

One may argue the HMC is able to exploit multi-mode distribution. Our claim is it is very hard to tune HMC with proper hyper-parameters that allows HMC move smoothly in arbitrary conditional inference cases, since each query evidence may dramatically change the latent distribution, in which pre-set hyper-parameter may not effective.

1.2.2 Rezende Approach [Rezende et al., 2014]

The Rezende approach constructs a transition function:

$$T(\hat{y}|y,x) = \int_{\hat{x}}\int_{z} p(\hat{y},\hat{x}|z)q(z|x,y)dzd\hat{x}, \tag{5}$$

where $q(z|x,y)$ and $q(\hat{x},\hat{y}|z)$ are encoder and decoder of VAE. It assumes $q(z|x,y)$ is a good approximation of $p(z|x,y)$. By fixing the input evidence x and iterative update $\hat{y}$, this approach approximate the blocked Gibbs sampling and is able to find stationary distribution of $p(y|x)$ for arbitrary evidence variable set X and query variable set Y.

As pointed out in paper [Wu et al., 2018], Rezende approach needs more evidence variables ratio over all variables to trigger the effective transition. General observation is that, when less than 40% variables are observed, the Rezende approach may not find the stationary distribution due to the transition is potentially not aperiodic and irreducible everywhere in the latent space.

1.2.3 Cross-coding Inference [Wu et al., 2018]

Cross-coding is a variational inference method which approximate $p(y|x)$ by minimize the upper-bound of KL divergence

$$\begin{aligned} KL[q_\psi(y)||p_\theta(y|x)] &\leq KL[q_\psi(y,z)||p_\theta(y,z|x)] \\ &= KL[q_\psi(z)||p_\theta(z|x)] + KL[p_\theta(y|z)||p_\theta(y|z,x)] \\ &= KL[q_\psi(z)||p_\theta(z|x)] + KL[p_\theta(y|z)||p_\theta(y|z)] \\ &= KL[q_\psi(z)||p_\theta(z|x)], \end{aligned} \tag{6}$$

where $q_\psi(z)$ is assumed to be an prior function that project standard Normal distribution into complex distribution that close to $p_\theta(z|x)$.

While, from conditional generation quality perspective, Cross-coding inference achieves desirable performance, it suffers from low efficiency due to it requires to train the prior network for each instance of the conditional inference.

2 Queryable Variational Auto-encoder

We notice that arbitrary conditional inference has not been solved yet due to the limitation from either efficiency from inference methods or from the training of VAE itself.

Leveraging on previous VAE, CVAE and BCDE works, we propose Queryable Variational Auto-encoder. The proposed model could directly deploy on lots of VAE variants without changing basic structure.

Assume the full variable list is V. $X, Y \subseteq V$ and $X \cap Y = \emptyset$. Note, there may exist free variables so that $X \cup Y \neq V$.

For any random partition of variables X and Y, the joint probability of any assignment $p(x,y)$ is

$$\begin{aligned} &\log p(x,y) \\ &= \log p(y|x) + logp(x) \\ &\geq E_{q(z|x,y)}[\log p(y|z)dz] - KL[q(z|x,y)||p(z|x)] + E_{q(z|x)}[\log p(x|z)dz] - KL[q(z|x)||p(z)] \\ &\geq E_{q(z|x,y)}[\log p(y|z)dz] - KL[q(z|x,y)||q(z|x)] + E_{q(z|x)}[\log p(x|z)dz] - KL[q(z|x)||p(z)] \end{aligned} \tag{7}$$

where we relax $p(z|x)$ by its lower-bound $q(z|x)$ instead of $q(z)$, and we do not need additional prior network.

By learning a model that maximize the probability of any subset of V, it approximate an Bayesian model averaging where each prediction output is given by multiple trained subset models. Mathmatically, if we assume the random partition probability is $p(s)$, the joint objective function could be represented as

$$\sum_{i}^{|D|}\sum_{s} p(s) \log p(x^{(i)}, y^{(i)}) \tag{8}$$

2.1 Property 1: Bounded Estimation

Comparing to original Variational Auto-encoder with conditional independent assumption (generative probability $p(x|z)$ and $p(y|z)$ are independent):

$$\log p(x, y) \geq \int_z q(z|x, y) \log \frac{p(x, y|z)p(z)}{q(z|x, y)} dz \tag{9}$$

Use (7) minus (9), we have:

$$\begin{aligned}
&\int_z q(z|x, y) \log \frac{p(y|z)p(z|x)}{q(z|x, y)} dz + \int_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} dz - \int_z q(z|x, y) \log \frac{p(x, y|z)p(z)}{q(z|x, y)} dz \\
&= -logp(x) + \int_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} dz \\
&= KL(q(z|x)||p(z|x)),
\end{aligned} \tag{10}$$

which indicate the gap between the different types of lower-bounds are simply $KL(q(z|x)||p(z|x))$.

Thus, the lower bound of p(x,y) in Q-VAE is: $KL(q(z|x)||p(z|x)) + KL(q(z|x, y)||p(z|x, y))$.

2.2 Property 2: Robust Latent Representation with More or Less Observations

Look at the KL terms for Normal distribution in the objective function in equation 7

$$\begin{aligned}
&KL[q(z|x, y)||q(z|x)] \\
&= \log \frac{\sigma_x}{\sigma_{x,y}} + \frac{\sigma_{x,y}^2 + (\mu_{x,y} - \mu_x)^2}{2\sigma_x^2} - \frac{1}{2},
\end{aligned} \tag{11}$$

which suggests the mean of more observations should be close to any mean of less observations, where overlapped observations are identical. This interpretation is corresponding to the experiment result as shown in section 3.2.

2.3 Property 3: Support Arbitrary Combination of Evidence and Query Variables

From the conditional inference perspective, in stead of fixing the split of variables X and Y as in CVAE and BCDE, Q-VAE randomly split variables during training through cascade random dropouts as described in section 2. Such random dropout training allows model being able to do arbitrary conditional inference with different set of evidence or query variables without retrain a new model.

For example, we can obtain random $X \cup Y$, X and Y through

$$\begin{aligned}
&X \cup Y = \mathit{Dropout}(V, \rho) \\
&X = \mathit{Dropout}(X \cup Y, \rho) \\
&Y = X \cup Y - X,
\end{aligned} \tag{12}$$

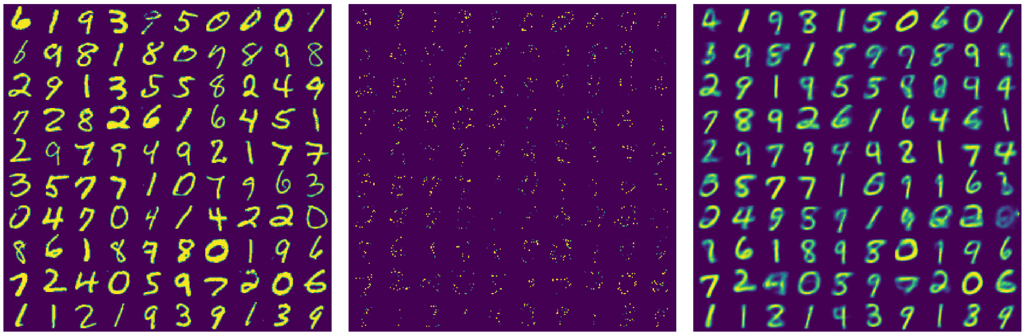
where ρ indicates dropout ratio. Clearly, the partition probability $p(s)$ is a function of the dropout rate ρ. Note: if the evidence feeding into network is not full, we have to re-scale the input to guarantee the statistic in-variance.

3 Demo with MNIST

3.1 Ability of Conditional Inference

To demonstrate the ability of conditional inference of Q-VAE, we do image inpainting on corrupted MNIST images. In this demo, the evidence varaibles are pixels that observed by the network and query variables are everything corrupted.

Note, Variational auto-encoder without modification can also do image impainting by approximating blocked gibbs sampling through repeating forward propagation as shown in Rezende et al. [2014]. However, our previous experiments show it does not work when the evidence percentage is less than 40%.



(a) Original

(b) Corrupted

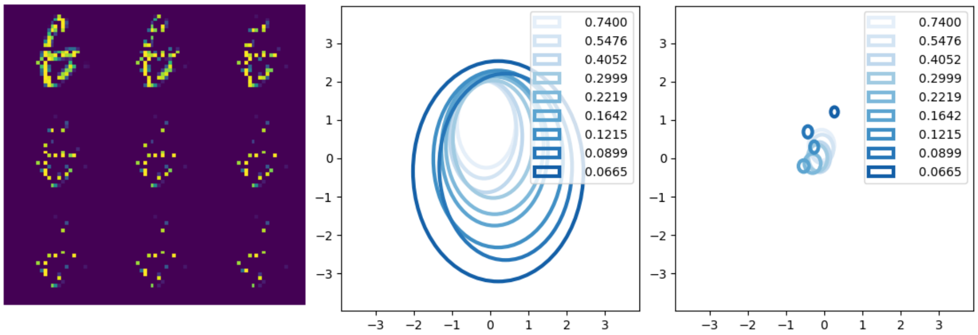
(c) Mean of Prediction

Figure 1: Image in-painting example. We feed corrupted images as input of the trained Q-VAE network. The network produce the full filled prediction that matches evidence with single forward propagation. The evidence percentage is around 10%

3.2 Uncertainty with Partial Observation

We show the Q-VAE is able to correctly represent the uncertainty with different level partial observations. Intuitively, we want the latent representation distribution has the following properties:

- The less we observe, the larger variance we should obtain in the latent space.
- With more observations, the latent distribution should be inside of the distribution of that with fewer observations.



(a) Corrupted

(b) Q-VAE Latent Distributions

(c) VAE Latent Distributions

Figure 2: Uncertainty of different observation percentage. (a)Shows different level of input corruptions from less to more. (b) shows the corresponding latent distributions of Q-VAE.(first two dimensions) (c) shows the latent distributions of original VAE. Note the circles cover 99% latent distribution for each partial input.

References

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

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 1278–1286, 2014.

Rui Shu, Hung H Bui, and Mohammad Ghavamzadeh. Bottleneck conditional density estimation. *International COnference on Machine Learning*, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

Ga Wu, Justin Domke, and Scott Sanner. Conditional inference in pre-trained variational autoencoders via cross-coding. *arXiv preprint arXiv:1805.07785*, 2018.

Appendix

3.3 Algorithm

The following algorithm shows the rough step of the coding, where we assume encoder and decoder are multi-layer neural networks with bottleneck latent dimension.

Algorithm 1 Queryable Variational Auto-encoder(Q-VAE)

```
procedure TRAIN($D$, $V$)
    for $i \in range(1, |D|)$ do                                        ▷ Loop over data
        $x, y, \rho \leftarrow GetRandomSplit(d_i, V)$
        $\mu_x, \sigma_x \leftarrow Encoder_\psi(x, \rho)$
        $\mu_{x,y}, \sigma_{x,y} \leftarrow Encoder_\psi((x, y), \rho^2)$
        $z_x \leftarrow \mu_x + \sigma_x * Random(0 \cdots 1)$                    ▷ Re-parameterization Trick
        $z_{x,y} \leftarrow \mu_{x,y} + \sigma_{x,y} * Random(0 \cdots 1)$
        $\hat{x} \leftarrow Decoder_\theta(z_x)$
        $\hat{y} \leftarrow Decoder_\theta(z_{x,y})$
        $L \leftarrow (1-\rho)[KL(\mu_{x,y}, \sigma_{x,y}, \mu_x, \sigma_x) + KL(\mu_x, \sigma_x, 0, 1) + \ell(x, \hat{x}) + \ell(y, \hat{y})]$   ▷ Loss
        $\psi \leftarrow \psi + \frac{\partial L}{\partial \psi}$                                   ▷ Parameter Update
        $\theta \leftarrow \theta + \frac{\partial L}{\partial \theta}$

procedure GETRANDOMSPLIT($d_i$, $V$)                ▷ Create Random Split for Pairwise Training
    $\rho \leftarrow Random(0.1 \cdots 0.9)$                                 ▷ Random Dropout Rate
    $X \cup Y \leftarrow Dropout(V, \rho)$
    $X \leftarrow Dropout(X \cup Y, \rho)$
    $x \leftarrow d_i[X]$
    $y \leftarrow d_i[X \cup Y - X]$
    return $x, y, \rho$

procedure KL( $\mu_{x,y}$, $\sigma_{x,y}$, $\mu_x$, $\sigma_x$)                   ▷ Gaussian Close Form KL Divergence
    return $\log \frac{\sigma_x}{\sigma_{x,y}} + \frac{\sigma_{x,y}^2 + (\mu_{x,y} - \mu_x)^2}{2\sigma_x^2} - \frac{1}{2}$
```