Correlated Variational Auto-Encoders

Da Tang * Columbia University

Dawen Liang Netflix

Tony Jebara Columbia University & Netflix

Abstract

Variational Auto-Encoders (VAEs) are capable of learning latent representations for high dimensional data. However, due to the i.i.d. assumption, VAEs only optimize the singleton variational distributions and fail to account for the correlations between data points, which might be crucial for learning latent representations from dataset where *a priori* we know correlations exist. We propose correlated VAEs that can take the correlation structure into consideration when learning latent representations with VAEs. Experimental result on learning matchings on a public benchmark movie rating dataset shows the effectiveness of the proposed method over several baseline algorithms.

1 Introduction

Variational Auto-Encoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) are a family of powerful deep generative models that learns stochastic latent embeddings for input data. By applying variational inference on deep generative models, VAEs are able to successfully identify the latent structures of the data and learn latent distributions that can potentially extract more compact information which are not easily and directly obtained from the original data.

VAEs assume each data point is i.i.d. generated, which means we do not consider any correlations between the data points. This is a reasonable assumption under many settings. However, sometimes we known *a priori* that data points are structured and correlated (e.g. networked data with useful side-information (Shi et al., 2014)). For example, in a recommender system, it is reasonable to assume a user's social network can impact what items they would interact with. When we fit a regular VAE on user's click data as in Liang et al. (2018), we will lose such information. It is more reasonable to assume the latent representation for each user is also correlated following the same network structure.

In this paper, we extend the regular VAEs by encouraging the latent representations to take the correlation structure into consideration. Instead of only learning singleton latent space mappings, we also learn pairwise mappings to capture the correlations from the graph structure. More specifically, we define the prior distribution of the latent variables according to the pairwise correlations. To make the computation tractable, we modify the standard VAE objective to contain a KL-divergence term with both edge and non-edge information from the graph. The experiment results show that our method can outperform regular VAE and some other baseline methods on matching dual user pairs using the movie rating records on a movie recommendation dataset.

2 Correlated VAEs

2.1 Variational Auto-Encodings

Assume that we have input data $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^D$. VAEs assume that the data point $\boldsymbol{x}_i$ is generated i.i.d. from the following process: First, generate the latent embeddings $\boldsymbol{z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \subseteq \mathbb{R}^d$ (usually $d \ll D$) by drawing i.i.d. $\boldsymbol{z}_i \sim p_0(\boldsymbol{z}_i)$ from the prior distribution p_0

*Work done while DT was an intern at Netflix.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

(parameter-free, usually a standard Gaussian distribution) for each $i \in \{1, \ldots, n\}$. Then generate the data points $\boldsymbol{x}_i \sim p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\theta})$ from the model conditional distribution p, for $i \in \{1, \ldots, n\}$ independently. The joint likelihood of $(\boldsymbol{z}, \boldsymbol{x})$ is $p(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p_0(\boldsymbol{z}_i) \cdot \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\theta})$.

We are interested in optimizing $\boldsymbol{\theta}$ to maximize the likelihood $p(\boldsymbol{x}; \boldsymbol{\theta})$, which requires computing the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$. For most models, this is usually intractable. VAEs sidestep the intractability and resort to variational inference by approximating this posterior distribution as $q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda}) = \prod_{i=1}^{n} q(\boldsymbol{z}_i|\boldsymbol{x}_i; \boldsymbol{\lambda})$ and maximize the *evidence lower bound* (ELBO):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta}) &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda})} \left[\log p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})\right] - \beta \cdot \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}; \lambda)||p_0(\boldsymbol{z})) \\ &= \sum_{i=1}^{n} (\mathbb{E}_{q(\boldsymbol{z}_i|\boldsymbol{x}_i;\boldsymbol{\lambda})} \left[\log p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\theta})\right] - \beta \cdot \mathrm{KL}(q(\boldsymbol{z}_i|\boldsymbol{x}_i; \boldsymbol{\lambda})||p_0(\boldsymbol{z}_i))) \end{aligned} \tag{1}$$

Here $\beta > 0$ is a constant that we can control. When $\beta = 1$, ELBO is a lower bound of the log-likelihood $\log p(\boldsymbol{x}; \boldsymbol{\theta})$ and maximizing this lower bound is equivalent to minimizing the KL divergence between the variational distribution $q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda})$ and the true posterior $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})$. Recent work (Higgins et al., 2017; Alemi et al., 2018) suggests that setting with $\beta \neq 1$ can lead to other desirable behaviors for representation learning.

The KL-divergence term in the ELBO can be viewed as a regularization that pulls the variational distribution $q(\boldsymbol{z}|\boldsymbol{x})$ towards the prior distribution $p_0(\boldsymbol{z})$. Since the approximation family $q(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda})$ factorizes over data points, the KL-divergence in the ELBO is simply a sum over the per-data-point KL-divergence terms, which means that we do not consider any correlations of latent representation between data points.

2.2 Inference with correlation

As motivated earlier, sometimes we know *a priori* there exists correlations between data points. If we have access to such information, we can incorporate it into the generative process of VAEs, which we term Correlated VAEs.

Formally, Assume that we have n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In addition, we know correlation structure of these data points through an undirected graph $G = (V, E)$, where $V = \{v_1, ..., v_n\}$ is the set of vertices corresponding to all data points (i.e., v_i corresponds to $\boldsymbol{x}_i$) and $(v_i, v_j) \in E$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are correlated. Making use of the correlation information, we change the prior distribution p_0 of the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ to take the form of a distribution over $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathbb{R}^d \times \ldots \times \mathbb{R}^d$ whose marginal distribution on any two variables $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ satisfies

$$p(\boldsymbol{z}_i, \boldsymbol{z}_j) = \begin{cases} p_0^p(\boldsymbol{z}_i, \boldsymbol{z}_j) & \text{if} (v_i, v_j) \in E \\ p_0^n(\boldsymbol{z}_i, \boldsymbol{z}_j) & \text{if} (v_i, v_j) \notin E. \end{cases} \tag{2}$$

Here $p_0^p(\cdot, \cdot)$ and $p_0^n(\cdot, \cdot)$ are two parameter-free distributions that capture the correlated and uncorrelated relationships between each pair of variables. For example, we can set p_0^p to have high density when the two latent variables have closer value while set p_0^n in the opposite way. This change will help the model take the correlation information into consideration since we have a KL-divergence regularization term in ELBO that will regularize the variational distribution towards the prior distribution. With the latent representation $\boldsymbol{z}$ sampled from this new prior $p_0(\boldsymbol{z})$, we assume the each data point $\boldsymbol{x}_i$ is again conditionally independently generated from $\boldsymbol{z}$, similar to a regular VAE.

With this choice of prior, ideally it would make sense to use a fully joint q distribution to approximate the joint posterior and optimize $\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda})} \left[\log p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})\right] - \beta \cdot \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}; \lambda)||p_0(\boldsymbol{z}))$.

However, this is intractable. To address this issue, we relax the exact ELBO in the following way:

$$\begin{aligned}\mathcal{L}^G(\boldsymbol{\lambda}, \boldsymbol{\theta}) = &\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{z}_i|\boldsymbol{x}_i;\boldsymbol{\lambda})} \left[\log p(\boldsymbol{x}_i|\boldsymbol{z}_i;\boldsymbol{\theta})\right] \\ &- \beta \cdot \sum_{(v_i,v_j)\in E} \mathrm{KL}(q(\boldsymbol{z}_i, \boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{x}_j;\boldsymbol{\lambda})||p_0^p(\boldsymbol{z}_i, \boldsymbol{z}_j)) \\ &- \gamma \cdot \sum_{(v_i,v_j)\notin E} \mathrm{KL}(q(\boldsymbol{z}_i, \boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{x}_j;\boldsymbol{\lambda})||p_0^n(\boldsymbol{z}_i, \boldsymbol{z}_j)).\end{aligned} \tag{3}$$

Here $\gamma > 0$ is a constant that controls the contribution from uncorrelated pair of nodes. We use a variational distribution that only considers singleton $q(\boldsymbol{z}_i|\boldsymbol{x}_i;\boldsymbol{\lambda})$ and pairwise relations $q(\boldsymbol{z}_i, \boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{x}_j;\boldsymbol{\lambda})$. As one example for such variational distribution, we can set the $q(\boldsymbol{z}|\boldsymbol{x})$ as a multivariate Gaussian distribution where the marginal distribution of $\boldsymbol{z}_i$ is a function of $\boldsymbol{x}_i$ while the covariance $\mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{z}_j)$ is a function of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. By applying this approximation, we only need to learn singleton and pairwise parametric distributions. On the other hand, this approximation has the advantage over regular VAE in that we take the pairwise correlations into consideration by directly learning the correlations in the variational distribution q.

In most of the real-world data, the graph G is highly sparse. In this case, we can approximate the non-edge KL-divergence (the last term in Equation (3)) by sampling edges from the complement graph $\bar{G} = (V, \bar{E})$. This is akin to negative sampling commonly used in language modeling.

2.3 Related work

Shaw et al. (2011) incorporated graph structure to metric learning. The major difference with correlated VAEs is that the metric learned from (Shaw et al., 2011) is inherently linear while correlated VAEs is capable of capturing more complex non-linear relations in the feature space.

There has been some recent work on incorporating structures in VAEs. Johnson et al. (2016) proposed structured VAEs which enable the prior to take a more complex form (e.g., a Gaussian mixture model, or a hidden Markov model). Similarly, Ainsworth et al. (2018) proposed output-interpretable VAEs which combine a structured VAE comprised of group-specific generators with a sparsity-inducing prior. However, both structured VAEs and output-interpretable VAEs are designed to model the structures between dimensions *within* each data point, while correlated VAEs considers structures *between* data points.

Another related line of work is the recent advances in graph convolutional networks (Kipf and Welling, 2016; Hamilton et al., 2017). We leave the comparison with GCN for future work.

3 Experiments

In this simple experiment, we evaluate the correlated VAEs with a bipartite correlation graph. We use the **MovieLens 20M** dataset (Harper and Konstan, 2016). This is a public movie rating dataset that contains ≈ 138K users and ≈ 27K movies. We binarize the rating data and only consider whether a user has watched a movie or not, i.e., the feature vector for each user is a binary bag-of-word vector. For all the experiments, we did a stochastic train/test split over users with a $90/10$ ratio.

Task. For each user u_i, we randomly split the movies that this user has watched into two halves and construct two synthetic users u_i^A and u_i^B. This creates a bipartite graph where we know the synthetic users which were generated from the same real user should be more related than two random synthetic users. The goal of the evaluation is that, when given the watch history of a synthetic user u_i^A from a held-out set, we try to identify its dual user u_i^B. This can be potentially helpful with identifying close neighbors when using matching to estimate causal effect, which is generally a difficult task especially in high-dimensional feature space (Imbens and Rubin, 2015).

Method. We learn a correlated VAE from training synthetic user pairs. Since we know the correct dual user matchings, we can build an undirected graph $G = (V, E)$ to capture this correlation

information. For each pair of synthetic users (u_i^A, u_i^B), we have an edge $(v_i^A, v_i^B) \in E$. For the edge relation prior distribution p_0^p, we use:

$$p_0^p(\boldsymbol{z}_i^A, \boldsymbol{z}_i^B) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{0}_{2d}, \Sigma = \begin{pmatrix} \mathbf{1}_d & \tau \cdot \mathbf{1}_d \\ \tau \cdot \mathbf{1}_d & \mathbf{1}_d \end{pmatrix}\right). \tag{4}$$

Here $0 \ll \tau < 1$ is a parameter that controls the correlation between the variables. Since the watch history of u_i^A and u_i^B are correlated, we set τ to be close to 1, as we hope the corresponding latent embeddings are close to each other. On the other hand, we use a standard Gaussian distribution for the non-edge relation prior distribution p_0^n since we do not expect the two variables to have some correlations.

For the variational distribution q, we set both $q(\boldsymbol{z}_i^A)$ and $q(\boldsymbol{z}_i^B)$ as diagonal Gaussian distribution on $\mathbb{R}^d$ and $q(\boldsymbol{z}_i^A, \boldsymbol{z}_j^B)$ as a Gaussian distribution such that its marginal on both $\boldsymbol{z}_i^A$ and $\boldsymbol{z}_j^B$ are diagonal Gaussian distributions and the covariance matrix between $\boldsymbol{z}_i^A$ and $\boldsymbol{z}_j^B$ is also a diagonal matrix. This approximation helps make the inference tractable and efficient while still take the correlation information into considerations. For the observational model $p(\boldsymbol{x}_i|\boldsymbol{z}_i)$, we use a multinomial likelihood similar to Liang et al. (2018).

Baselines. We compare correlated VAEs with both regular and some variations of VAEs:

- VAE: Simply train on user's watch history without using the correlation structure.
- VAE_{BI}: On top of a regular VAE, we add a bi-directional transformation between the latent embeddings $\boldsymbol{z}_i^A$, $\boldsymbol{z}_i^B$ for each synthetic user pair (u_i^A, u_i^B). More specifically, we learn a bi-directional mapping of the mean and standard deviation between the diagonal normal approximation $q(\boldsymbol{z}_i^A)$ and $q(\boldsymbol{z}_i^B)$.
- $\text{VAE}_{\text{EdgePrior}}$: Learn a regular VAE on pairs of synthetic users with a correlated prior in Equation (4), instead of standard Gaussian distributions. This is effectively setting γ in Equation (3) to 0 and only learning the singleton parametric distributions $q(\boldsymbol{z}_i|\boldsymbol{x}_i)$ (not learning the pairwise distributions).

Evaluation. We train all the methods on all the synthetic user pairs from the training set. To evaluate, we select a fixed number of $N^{\text{eval}} = 1000$ pairs of synthetic user from the test sets. For each synthetic user u_i^A (or u_i^B), we find the ranking of u_i^B (or u_i^A) among all candidates in the set of u_j^B (or u_j^A) in terms of the latent distribution distance to u_i^A (or u_i^B). This latent distribution distance can vary from methods to methods, as different methods provide different information to compute the distance. For correlated VAEs, we use

$$\text{dis}(u_i^A, u_j^B) = \mathbb{E}_{q(\boldsymbol{z}_i^A, \boldsymbol{z}_j^B)}\left[\left\|z_i^A - z_j^B\right\|^2\right]$$

as the distance measure which will make use of the correlation between pairs of users. For the baseline methods, since they do not compute the correlations between data points directly, we use the variation of this distance, which is the squared L_2-Wasserstein distance, to compute the distance:

$$\text{dis}(u_i^A, u_j^B) = \min_{q'(\boldsymbol{z}_i^A)=q(\boldsymbol{z}_i^A),\ q'(\boldsymbol{z}_j^B)=q(\boldsymbol{z}_j^B)} \mathbb{E}_{q'(\boldsymbol{z}_i^A, \boldsymbol{z}_j^B)}\left[\left\|z_i^A - z_j^B\right\|^2\right].$$

Results. We report normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002), mean reciprocal rank (MRR), and mean rank in Table 1. The two numbers in each parenthesis represent metrics for finding dual users of the u_i^A's among all u_j^B candidates and finding dual users of the u_i^B's among all u_j^A candaidates, respectively. We can see that correlated VAEs successfully outperform baseline methods. Notably, correlated VAE achieves a consistently smaller mean rank, which indicates that it is more stable as mean rank is sensitive to bad ranking.

4 Conclusion and future work

We introduce correlated VAEs to account for correlation between data points. It extends the regular VAEs by adding pairwise variational distribution approximations. This method successfully outperforms some other baseline methods based on VAEs on a matching task on a public movie-rating

Table 1: Metrics for dual user matching

Method	NDCG	MRR	Mean rank
VAE	(0.617, 0.613)	(0.519, 0.515)	(19.5, 20.3)
VAE_{BI}	(0.746, 0.752)	(0.673, 0.681)	(**5.76**, 6.25)
$\text{VAE}_{\text{EdgePrior}}$	(0.721, 0.721)	(0.643, 0.643)	(7.13, 6.82)
Correlated VAE	(**0.761**, **0.768**)	(**0.691**, **0.701**)	(**5.77**, **5.53**)

dataset. For future work, we will perform more thorough experiments on a more general type of undirected graphs, as well as with related baselines from the literature.

References

Samuel Ainsworth, Nicholas Foti, Adrian KC Lee, and Emily Fox. Interpretable VAEs for nonlinear group factor analysis. *arXiv preprint arXiv:1802.06765*, 2018.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pages 159–168, 2018.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1025–1035, 2017.

F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

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

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the International Conference on World Wide Web*, pages 689–698, 2018.

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

Blake Shaw, Bert Huang, and Tony Jebara. Learning a distance metric from a network. In *Advances in Neural Information Processing Systems 24*, pages 1899–1907. 2011.

Tianlin Shi, Da Tang, Liwen Xu, and Thomas Moscibroda. Correlated compressive sensing for networked data. In *UAI*, pages 722–731, 2014.