Learnt Prior VAE
Previous Papers

Presenter: Arshdeep Sekhon
https://qdata.github.io/deep2Read

Nonparametric Variational Auto-encoders for Hierarchical Representation Learning

- Prasoon Goyal, Zhiting Hu, Xiaodan Liang Chenyu Wang, Eric P. Xing
- ICCV 2017

Motivation

- VAE Loss = Reconstruction Error + Regularization

$$E_{\boldsymbol{z} \sim q_\phi(z|\boldsymbol{x}_m)}[log_{p_\theta}(\boldsymbol{x}_m|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_m)||p_\theta(\boldsymbol{z})) \tag{1}$$

- Bayesian regularization over the latent space, which enforces the posterior of the hidden code vector matches a prior distribution.
- this prior is a standard single mode normal distribution that enables convenient inference and learning
- overly simplified representations which lose rich semantics present in the data
- For example, a large video corpus can encode rich human activity with underlying intricate temporal dependencies and hierarchical relationships.
- desirable to develop new representation learning approaches with great modeling flexibility and structured interpretability

hierarchical nonparametric variational autoencoders

- **hierarchical nonparametric bayesian prior** with VAE
- unsupervised hierarchical representation learning of sequential data
- As opposed to fixed prior distributions learn both the VAE parameters and the nonparametric priors jointly from the data

Method

- $\boldsymbol{x}^m = (x_{mn})_{n=1}^{N_m}$: sequence m of length N_m
- capture a representation of each data sample $\boldsymbol{x}^m$ and learn a structured representation of the entire corpus

Dirichlet Process

- Intuition: A stick of unit length, break at random location, left part is π_1 and right part keep doing this
- $\sum_{i=1}^{N} \pi_i = 1$

$$v_i \sim Beta(1, \gamma), \qquad \pi_i = v_i \prod_{j=1}^{i-1} (1 - v_j)$$

$$w_i \sim G_0, \qquad G = \sum_{i=1}^{\infty} \pi_i \delta_{w_i}$$

Nested Chinese Restaurant Process Prior

- Extend Stick Breaking Process to tree structure
- Start at the root node (level 0), and obtain probabilities over its child nodes (level 1) using a DP. Then recursively run a DP on each level 1 node to get probabilities over level 2 nodes, and so on.
- This defines a probability distribution over paths of an infinitely wide and infinitely deep tree
- Root Node $\pi_1 = 1$
- ith node at level 1 $\pi_{1i} = \pi_1 v_{1i} \prod_{j=1}^{i}(1 - v_{1j})$
- For *jth* child at level 2 of *ith* level node $\pi_{1ij} = \pi_1 \pi_{1i} v_{1ij} \prod_{k=1}^{j}(1 - v_{1ik})$. This process is repeated to infinity

nested Chinese Restaurant Prior

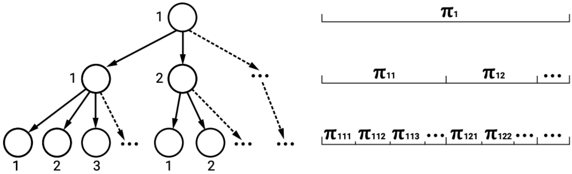


Figure 2. **Left**: a sample tree structure draw from nCRP. **Right**: The respective tree-based stick-breaking construction. The stick length of the root node is $\pi_1 = 1$. Each node performs a stick-breaking process on its stick segment to construct its children.

Generative Model

- The generative model assumes a tree with infinite depth and branches
- generates data sequences through root-to-leaf random walks along the paths of the tree.
- Each node p has a parameter α_p depends on the parameter vector of the parent node to encode the hierarchical relation.
- for every node p of the tree, draw a D-dimensional parameter vector α_p, according to $\alpha_p \sim N(\alpha_{par(p)}, \sigma^2 I)$, For root node define $\alpha_{par(p)} = \alpha^*$
- Each data sequence $\boldsymbol{x}_m$ is modeled as a mixture of the paths down the tree, and each element x_{mn} is attached to one path sampled from the mixture.

Sampling from the generative model

- For each edge e of the tree, sample $v_{me} \sim Beta(1, \gamma^*)$
- denote the collection of all v_{me} for sequence m as $\boldsymbol{V}_m$
- $\pi(\boldsymbol{V_m})$ denotes the probabilities of the leaf nodes
- For each element x_{mn} in $\boldsymbol{x}_m$, draw a path c_{mn} according to the multinomial distribution $Mult(\pi(\boldsymbol{V}_m))$.
- the final latent representaion z_{mn} according to $N(c_{\alpha_{mn}}, \sigma_D^2 I)$. which is the emission distribution defined by the parameter associated in the leaf node of path c_{mn}.
- parameters to estimate: α_p, V_m, path assignments c, θ, ϕ

Optimization: Variational Inference

- For each node p of the tree, the parameter vector α_p is distributed as $\alpha_p \sim N(\mu_p, \sigma_p^2 I)$, where μ_p is a D-dimensional vector and σ_p is a scalar.
- For sequence m, the DP variable at edge e, v_{me} is distributed as $v_{me} \sim Beta(\gamma_{me,0}, \gamma_{me,1})$, where $\gamma_{me,0}$ and $\gamma_{me,1}$ are scalars.
- For data x_{mn}, the path assignment variable c_{mn} is distributed as $c_{mn} Mult(\phi_{mn})$, where the dimension of ϕ_{mn} is equal to the number of paths in the tree.
- We want to find optimal variational parameters that maximize the variational lower bound

$$\mathcal{L} = E_q[\log p(W, X|\Theta)] - E_q[\log q_\nu(W)] \tag{2}$$

where W denotes the collection of latent variables, $X = \{z_{mn}\}$ are the latent vector representations of observations, Θ are the hyperparameters, and $\nu = \{\boldsymbol{\mu}_p, \sigma_p, \gamma_{me,0}, \gamma_{me,1}, \phi_{mn}\}$ are variational parameters.

$$
\begin{aligned}
& p(W, X|\Theta) \\
& = \sum_p \log p(\boldsymbol{\alpha}_p|\boldsymbol{\alpha}_{par(p)}, \sigma_N) + \sum_{m,e} \log p(v_{me}|\gamma^*) \\
& + \sum_{m,n} \log p(c_{mn}|\boldsymbol{V_m}) + \log p(\boldsymbol{z}_{mn}|\boldsymbol{\alpha}, c_{mn}, \sigma_D)
\end{aligned} \tag{5}
$$

$$
\mu_p = \sigma_p^2 \cdot \left(\frac{\boldsymbol{\mu}_{par(p)} + \sum_{r \in ch(p)} \boldsymbol{\mu}_r}{\sigma_N^2} \right)
$$

$$q^*(v_{me}|\gamma_{me,0},\gamma_{me,1}) \sim Beta(\gamma_{me,0},\gamma_{me,1}) \tag{11}$$

where

$$\gamma_{me,0} = 1 + \sum_{n=1}^{N_m} \sum_{p:e\in p} \phi_{mnp} \tag{12}$$

$$\gamma_{me,1} = \gamma^* + \sum_{n=1}^{N_m} \sum_{p:e<p} \phi_{mnp} \tag{13}$$

$$q^*(c_{mn}|\boldsymbol{\phi}_{mn}) \sim Mult(\boldsymbol{\phi}_{mn}) \tag{14}$$

where

$$\begin{aligned}\phi_{mnp} \propto \exp \Bigg\{ & \sum_{e:e\in p} [\Psi(\gamma_{me,0}) - \Psi(\gamma_{me,0}+\gamma_{me,1})] \\ & + \sum_{e:e<p} [\Psi(\gamma_{me,1}) - \Psi(\gamma_{me,0}+\gamma_{me,1})] \\ & - \frac{1}{2\sigma_D^2} \left[(\boldsymbol{z}_{mn} - \boldsymbol{\mu}_p)^T (\boldsymbol{z}_{mn} - \boldsymbol{\mu}_p) + \sigma_p^2 \right] \Bigg\}\end{aligned} \tag{15}$$

Optimization

- Keep the nCRP parameters fixed and then optimize NN parameters
- Alternating optimization
- Heuristic Ways to dynamically update tree structure

Experiments: Test Set Reconstruction

Algorithm	**Mean test log-likelihood**
VAE-StdNormal	-28886.90
VAE-nCRP	**-28438.32**

Experiments: Video Classification

Category	**K-Means**	**VAE-GMM**	**VAE-nCRP**
Board_trick	44.6	**47.2**	31.3
Feeding_an_animal	**57.0**	42.5	53.8
Fishing	33.7	39.0	**48.9**
Woodworking	38.9	40.5	**60.8**
Wedding_ceremony	59.8	54.3	**63.6**
Birthday_party	6.5	7.4	**27.8**
Changing_a_vehicle_tire	31.9	39.7	**45.3**
Flash_mob_gathering	**43.4**	40.1	38.2
Getting_a_vehicle_unstuck	52.9	50.6	**65.9**
Grooming_an_animal	2.9	14.5	**17.3**
Making_a_sandwich	47.1	**54.7**	49.3
Parade	28.4	**33.8**	19.8
Parkour	4.5	19.8	**27.7**
Repairing_an_appliance	42.3	**58.6**	47.4
Sewing_project	1.6	**24.3**	18.4
Aggregate over all classes	34.9	39.1	**42.4**

Table 3. Classification Accuracy (%) on TRECVID MED 2011.

The Loracs Prior for VAEs: Letting the trees speak for the data

- Sharad Vikram, Matthew D. Hoffman, Matthew J. Johnson
- AISTATS 2019

Motivation

- unimodal simple distribution too simple
- more opinionated prior on the VAE's latent vectors: the time marginalized coalescent(TMC)
- TMC: interpretable Bayesian nonparametric hierarchical clustering model that can encode rich discrete and continuous structure

Background: Bayesian priors for hierarchical clustering

- incorporates uncertainty over tree structure $r(\tau)$
- a likelihood model for data $r(z_{1:N}|\tau)$, with the goal of sampling the posterior distribution $r(\tau|z_{1:N})$.
- Phylogeny: rooted binary trees with N labeled leaves adorned with branch lengths

Time-marginalized coalescent

- TMC defines a prior disrtibution over phylogenies
- A phylogeny is (V, E, T) a directed fully binary tree with vertex set V , and edges E with time labels $T : V \rightarrow [0, 1]$. where $t_v = T(V)$
- $V_{leaf} = \{1, \cdots, N\}$
- Separate nodes into V_{leaf} and V_{int}
- $V = V_{int} \bigcup V_{leaf}$
- directed edges of the tree are encoded in the edge set $E \subset V_{int} \times V$, where we denote the root vertex as v_{root} and for $v \in V$ v_{root} we denote the parent of v as $\pi(v) = w$ where $(w, v) \in E$

TMC

- The TMC samples a random tree structure (V, E) by a stochastic process in which the N leaves are recursively merged uniformly at random until only one vertex is left.
- This process yields the probability mass function on valid (V, E) pairs given by

$$r(V, E) = \frac{(N-1)!}{\prod_{v \in V_{\text{int}}} c(v)} \prod_{i=1}^{N-1} \binom{i+1}{2}^{-1},$$

TMC

Given the tree structure, time labels are generated via the stick-breaking process

$$t_v = \begin{cases} 0 & v = v_{\text{root}}, \\ 1 & v \in V_{\text{leaf}}, \\ t_{\pi(v)} - \beta_v(1 - t_{\pi(v)}) & v \in V_{\text{int}} \setminus \{v_{\text{root}}\}, \end{cases}$$

where $\beta_v \sim Beta(a, b)$ for $v \in V$. These time labels encode a branch length $t_v - t_{\pi(v)}$ for each edge $e = (\pi(v), v) \in E$. We denote the overall density on phylogenies with N leaves as $TMC_N(\tau; a, b)$

Prior

- Data points are leaf nodes
- Define a likelihood model $r(z_{1:N}|\tau)$
- z_n corresponds to leaf vertex $n \in V_{leaf}$
- $z_{v_{root}} = N(0, I)$
- parent of $v = \pi(v)$
- $z_v|z_{\pi_v} = N(z_{\pi(v)}, (t_v - t_{\pi(v)})I)$
- $z_{v_{root}} = N(0, I)$
- Why? : Use GGM structure to marginalize out internal vertices $v \in V_{int}$ to get marginal density $r(z_{1:N}|\tau)$
- Final overall prior density $r(z_{1:N}|\tau) = TMC_N(\tau; a, b)r(z_{1:N}|\tau)$

TMC Posterior Predictive Density

- The above with N leaves and a GRW likelihood model can be a prior on a set of N hierarchically structured data : nodes closer to each other in terms of tree dsitance have similar location values
- Location values
- Sampling new data: $z_{N+1} : r(Z_{N+1}|z_{1:N}, \tau)$
- Need to select a branch to add a new leaf node and a time label
- $r(e_{N+1}|V, E)$ and $r(t_{N+1}|e_{N+1}, V, E)$

Connecting to VAE

- this phylogeny based prior is interpetable discrete structure in the latent space
- Step 1: generate $z_{1:N}$ according to the TMC prior
- Step 2: generate $x_{1:N}$

$$\tau \sim TMC_N(\tau; a, b) \tag{3}$$

$$z_{1:N}|\tau \sim r(z_{1:N}|\tau) \tag{4}$$

$$x_n|z_n \sim p_\theta(x_n|z_n) \tag{5}$$

TMC Inference

- posterior distribution is intractable $r(\tau|z_{1:N})$ is analytically intractable due to $r(z_{1:N})$ (sum over all possible structures)
- Use MCMC + Metropolis Hastings + subtree prune and regraft

$$\mathcal{L}[q] = \mathbb{E}_q \left[\log \frac{p(\tau, z_{1:N}, x_{1:N})}{q(\tau) \prod_n q(z_n \mid x_n)} \right] \tag{8}$$

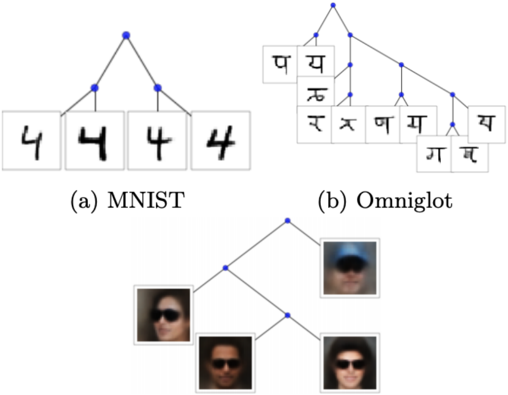
For fixed $q_\phi(z_n \mid x_n)$, we can sample the optimal $q^*(\tau)$,

$$q^*(\tau) \propto \exp\{\mathbb{E}_q [\log p(\tau, z_{1:N}, x_{1:N})]\} \tag{9}$$

Experiments

- Visualizing inducing points
- Hierarchical Clustering
- Generating Samples

Experiments: Hierarchical Clustering



(a) MNIST (b) Omniglot

(c) CelebA

Figure 4: An example learned subtree from a sample of $q(\tau; s_{1:M})$ for each dataset. Leaves are visualized by passing inducing points throught the decoder.

Experiments:Held-out log-likelihood

Prior	MNIST	Omniglot
Normal	-83.789	-89.722
MAF	**-80.121**	**-86.298**
Vamp	-83.0135	-87.604
LORACs	-83.401	-87.105

Table 2: MNIST/Omniglot test log-likelihoods

Faithful inversion of generative models for effective amortized inference

- Stefan Webb, Adam Golinski, Robert Zinkov, N. Siddharth, Tom Rainforth, Yee Whye Teh, Frank Wood

Motivation

- $q(z|x)$ is the inference network: deficiencies in inference network are propagated to the generative model
- VAE: Coarse Grain Structure(Bayesian Netwrok encodes a dependency) and a Fine Grain Structure (Neural nets)
- Amortized Inference: Graphical Mode Inversion (Invert the generative model to give a GM approximating the posterior)

Choosing the best structure for the ivnerse graphical model?

- invert structure of generative model
- introduces conditional independencies not present in true distribution
- Consequently, they cannot represent the true posterior even in the limit of infinite neural network capacities.
- Fully connected Bayesian Network for inverse graphical model
- Though such a model is expressive enough to correctly represent the data given infinite capacity and training time, it ignores substantial available information from the forward model
- leading to reduced performance for finite training budgets and/or network capaciti

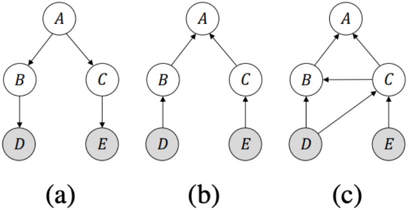


Figure 1: (a) Generative model BN: