Nonparametric Variational Auto-encoders for Hierarchical Representation Learning

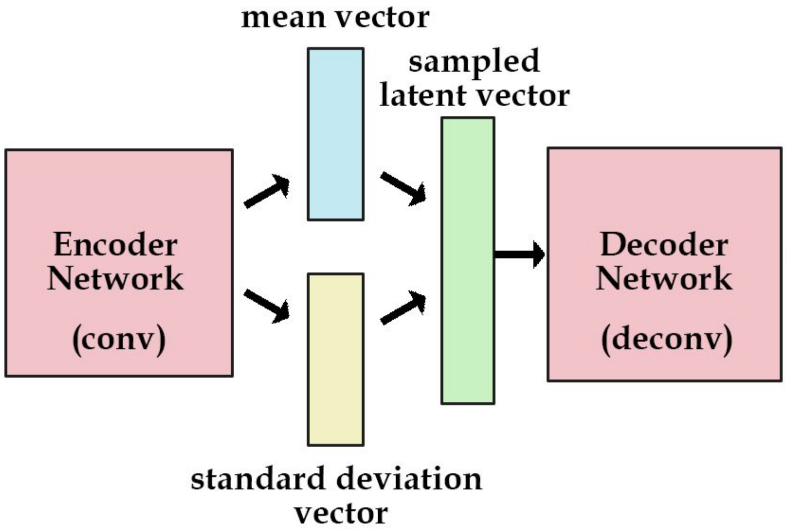
Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, Eric P. Xing

Presented by: Zhi Li

Motivation

- Variational Autoencoders can be used for unsupervised representation learning
    - However, most of these approaches employ a simple prior over the latent space

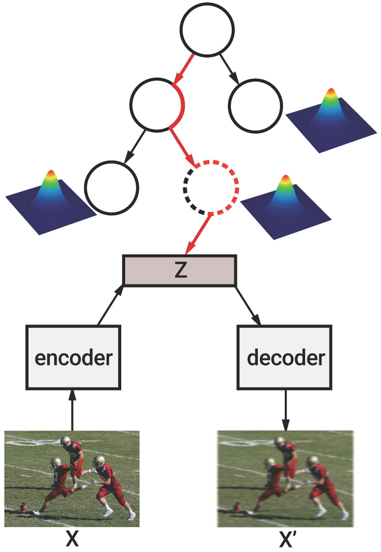
- It's desirable to develop a new approach with great modeling flexibility and structured interpretability



Variational Autoencoder Explained. Retrieved from http://kvfrans.com/content/images/2016/08/vae.jpg

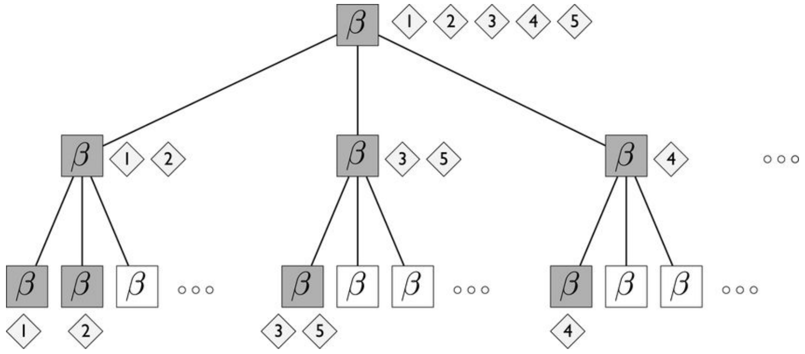
Hierarchical nonparametric variational autoencoders

- Bayesian nonparametric prior
  - Allows infinite information capacity to capture rich internal structure of the data
- Hierarchical structure
  - serves as an aggregated structured representation
  - coarse-grained concepts are higher up in the hierarchy, and fine-grained concepts form their children



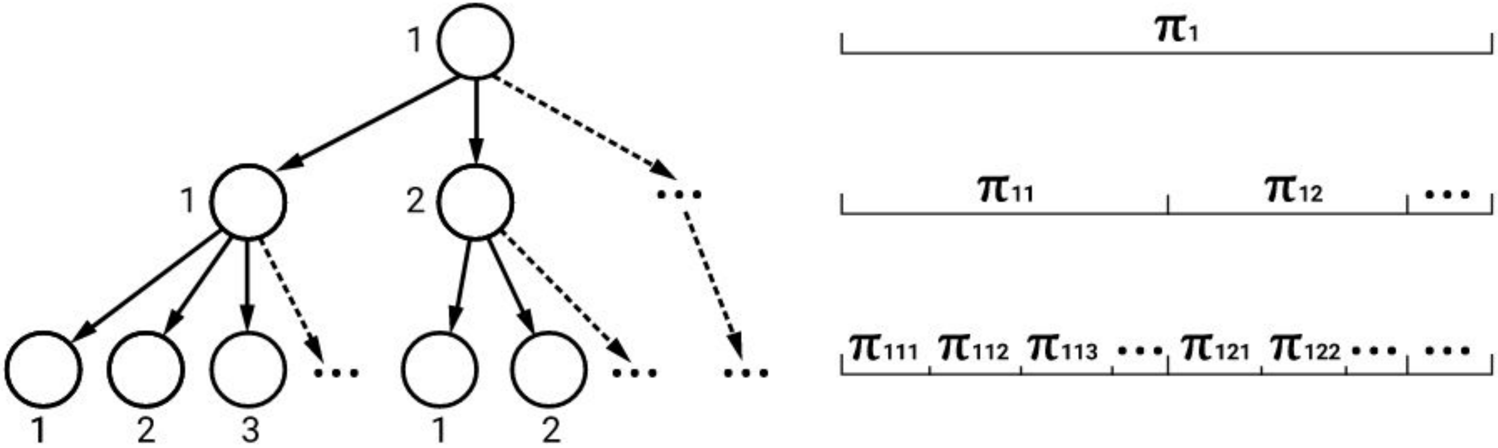
Nested Chinese Restaurant Process

- Similar to CRP, but imagine now that tables are organized in a hierarchy
  - Each table has infinite number of tables associated with it at the next level
  - Moving from a table to its sub-tables at next level, the customer draws following the CRP
- nCRP defines a probability distribution over paths of an infinitely wide and infinitely deep tree



Nested Chinese Restaurant Process. From: DAVID M. BLEI “The nested Chinese restaurant process and Bayesian inference of topic hierarchies”

Stick-breaking Interpretation



nCRP as the Prior

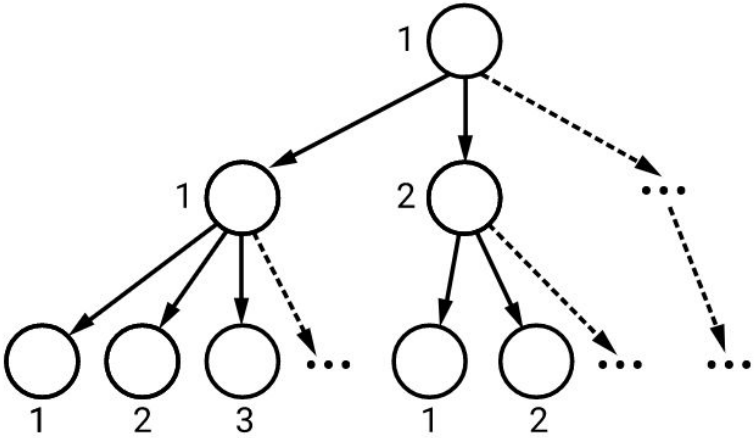
Image we have an infinitely wide and infinitely deep tree

For every node *p* of the tree, it has a parameter vector α_p, which depends on the parameter vector of its parent node to encode the hierarchical relation

$$\alpha_p \sim \mathcal{N}(\alpha_{par(p)}, \sigma_N^2 I)$$

For the root node, we set

$$\alpha_{par(p)} = \alpha^*$$



nCRP as the Prior

We can now use this tree to sample a sequence of latent representation through root-to-leaf random walks along the paths of the tree.

- For each sequence x_m, draws a distribution, V_m, over the paths of the tree based on nCRP
- For each element x_{mn} of the sequence, a path c_{mn} is sampled according to the distribution V_m
- Draw the latent representation z_{mn}, according to the parameter associated in the leaf node of the path c_{mn}

$$\mathcal{N}(\alpha_{c_{mn}}, \sigma_D^2 I)$$

Parameters Learning

The paper uses ***alternating optimization*** to jointly optimize for the neural network parameters (the encoder and the decoder) and the parameters of the nCRP prior.

- first fix the nCRP parameters and perform backpropagation steps to optimize for the neural network parameters
- then fix the neural network, and perform variational inference updates to optimize for the nCRP parameters

Variational Inference for nCRP

The trick is to derive ***variational inference*** on a ***truncated*** tree

- Variational inference - - to approximate a conditional density of latent variables
    - It uses a family of densities over the latent variables, parameterized by free “variational parameters”, and then find the settings of the parameters that is closest in KL divergence to the density of interest
    - The fitted variational density then serves as a proxy for the exact conditional density

- Truncated tree -- instead of an infinitely wide and deep tree
    - If the truncation is too large, the algorithm will still isolate only a subset of components
    - If the truncation is too small, the algorithm will dynamically expand the tree based on heuristic during training

Encoder and Decoder Learning

Similar to standard VAE, the object is to maximize the lowerbound of data log-likelihood

$$\mathcal{L} = \mathbb{E}_{z \sim q_\phi(z|x^m)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) || p_\theta(z))$$

Parameters are updated using gradient descent

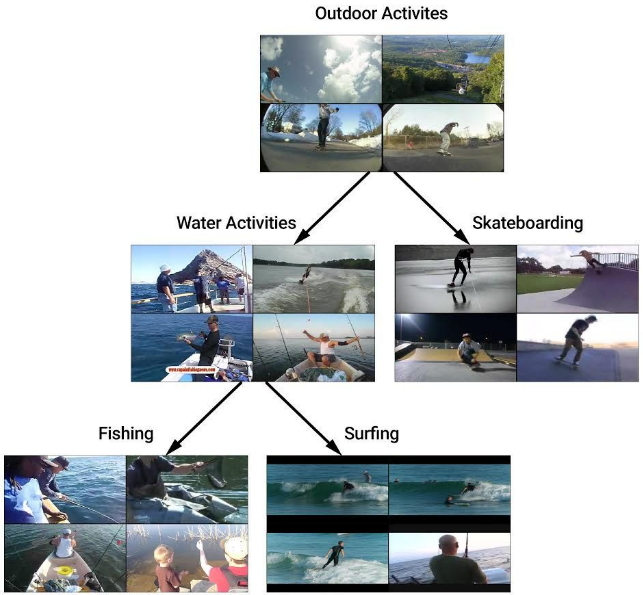
Experiment

- Evaluate the models on TRECVID Multimedia Event Detection 2011 dataset.
    - Only use the labeled videos: 1241 videos for training, 138 for validation and 1169 for testing
    - Each video is considered as a data sequence of frames
    - One frame is sample from the video every 5 seconds, and each frame is used as the elements within the sequence

- A CNN (VGG16) is used to extract features and the feature dimension is 4096

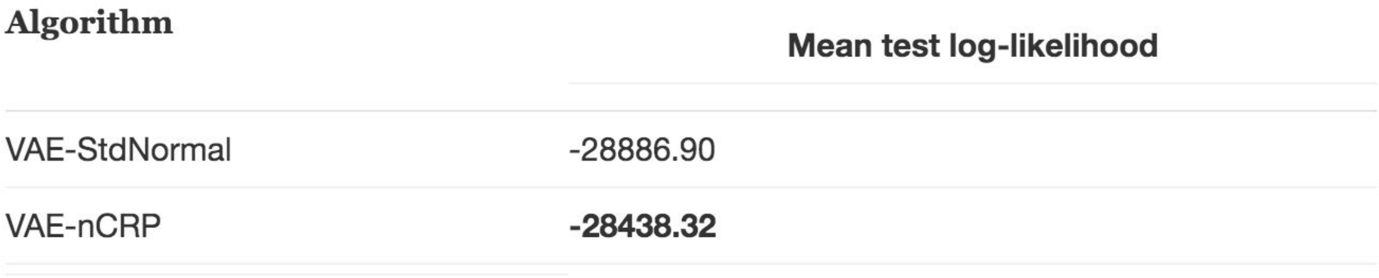
Video Hierarchical Representation Learning

- For each frame, the model obtain the latent representation ***z*** of the frame feature
- then find the node to which it's assigned by minimizing the distance between the latent representation ***z*** and the node parameters α_p



Likelihood Analysis

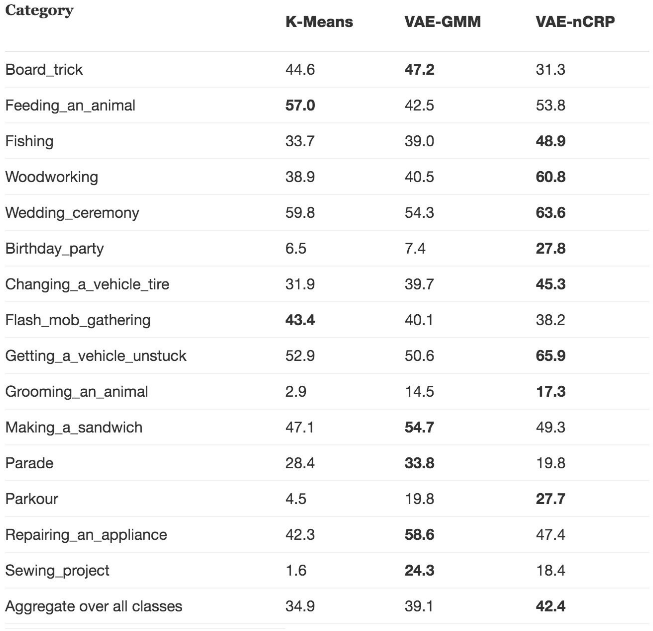
Test-set log-likelihoods comparison between the model “VAE-nCRP” and traditional variational autoencoder “VAE-StdNormal”



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

Classification

- Assigns the label to each node (either leaf nodes or internal nodes) by taking a majority vote of the labels assigned to the node
- Predicts label of a frame is then given by the label assigned to the closest node to the frame



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**

Takeaway

The paper presented a new unsupervised learning framework to combine rich nCRP prior with VAEs

This approach embeds the data into a latent space with rich hierarchical structure, which has

- more abstract concepts higher up in the hierarchy
- less abstract concepts lower in the hierarchy