

Deep Learning and Practice — Final Exam

Date: Tuesday, June 17, 2021

Time: 12:20pm – 15:20pm (180 minutes)

Format: Open book

Instructions:

- 1) You may give your answers in Chinese or English.
- 2) Please give your answers in succinct phrases or point form.
- 3) Please write your answers clearly (with explicit denotation of labels and symbols used).

1. (15 pts) Consider an energy-based model with the following probability distribution

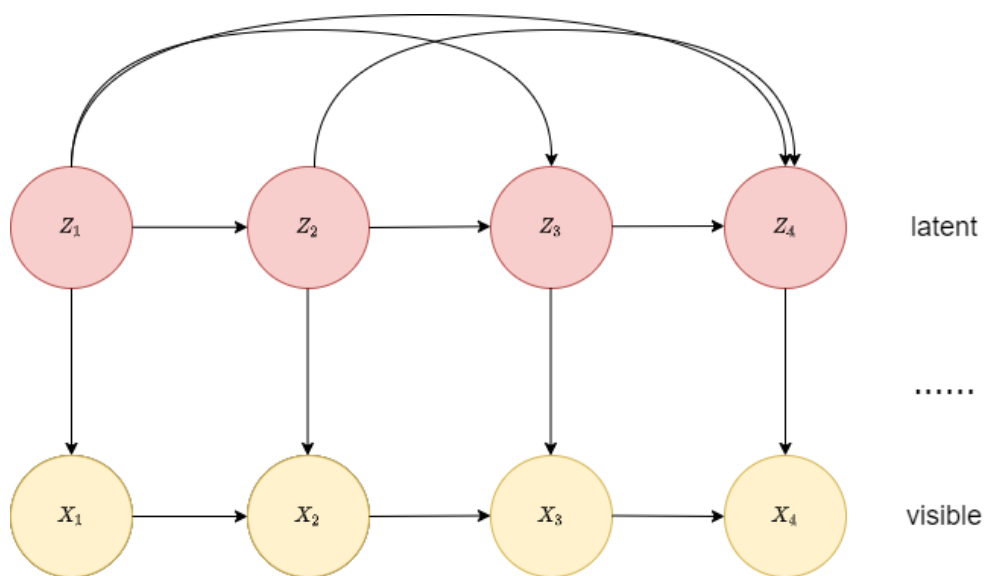
$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h}))$$

where $\mathbf{v} = (v_1, v_2, \dots, v_m)$ are *binary visible units*; $\mathbf{h} = (h_1, h_2, \dots, h_n)$ are *binary hidden units*; $Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$ is the partition function; and $E(\mathbf{v}, \mathbf{h})$ is the energy function defined as

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h} - \mathbf{v}^T \mathbf{W} \mathbf{h},$$

with the vectors \mathbf{b}, \mathbf{c} and the matrix \mathbf{W} denoting the model parameters.

- (a) (5 pts) Show that $p(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^n p(h_j|\mathbf{v})$ is factorial and $p(h_j = 1|\mathbf{v}) = \sigma(c_j + \mathbf{v}^T \mathbf{W}_{:,j})$, where $\mathbf{W}_{:,j}$ is the j -th column vector of \mathbf{W} .
 - (b) (5 pts) Show that $p(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^m p(v_i|\mathbf{h})$ is factorial and $p(v_i = 1|\mathbf{h}) = \sigma(b_i + \mathbf{W}_{i,:} \mathbf{h})$, where $\mathbf{W}_{i,:}$ is the i -th row vector of \mathbf{W} .
 - (c) (5 pts) Assuming the model parameters are known, how can the $p(\mathbf{h}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{h})$ be utilized to draw samples of \mathbf{v} (and/or \mathbf{h})? How would you draw independent samples of \mathbf{v} ?
2. (32 pts) Consider the following latent factor model, where $Z_i, i = 1, 2, \dots, T$ are latent variables and $X_i, i = 1, 2, \dots, T$ are visible variables.



- (a) (5 pts) Factorize $p(X_{1-T}|Z_{1-T})$. That is to express $p(X_{1-T}|Z_{1-T})$ as the product $\prod_{t=1}^T p(X_t | \dots)$. Make explicit the conditioning variables in \dots using the d-separation rule.
 - (b) (5 pts) Follow (a) and factorize $p(Z_{1-T}|X_{1-T})$.
 - (c) (5 pts) Following (a) and factorize $p(Z_{1-T})$.
 - (d) (5 pts) Design an encoding distribution $q(Z_{1-T}|X_{1-T})$ to approximate the true posterior distribution $p(Z_{1-T}|X_{1-T})$, providing that the generation process of Z_t is based on causal information only, i.e. $X_{\leq t}, Z_{< t}$. What would be the factorization of $q(Z_{1-T}|X_{1-T})$?
 - (e) (6 pts) Train this latent factor model by maximizing the evidence/variational lower bound. Describe the network architecture (CNN, RNN, etc.) for the encoding, decoding and prior distribution, and the training objective function.
 - (f) (6 pts) Consider all X_i 's and Z_i 's to be visible. Convert the graphical model into a flow model. Use $T = 3$ as an example.
3. (20 pts) Convolution, Pooling, Activation function, CNN.
- (a) (6 pts) What is the output feature map size for an 128 x 128 input image after convolution with kernel (3,3), padding (2,2), and stride (2,2) ?
 - (b) (6 pts) Which activation function may cause serious gradient vanish problem? Why?
 - (c) (5 pts) Describe the function of pooling layer.
 - (d) (3 pts) What are the major reasons that contribute to the success of convolutional neural networks?
4. (15 pts) Training the VAE.
- (a) (3 pts) In training the VAE, we try to maximize a variational lower bound on the data log-likelihood. Explain the main idea and provide the exact objective function to be maximized.
 - (b) (3 pts) What distribution does the approximate posterior $q(z|x)$ take for training VAE? Is this an assumption?
 - (c) (3 pts) Explain the notion of the re-parameterization trick.
 - (d) (3 pts) True or False: In maximizing the variational lower bound, the approximate posterior $q(z|x)$ should ideally be identical to the prior $p(z)$ when the variational lower bound is maximized. Explain your answer.
 - (e) (3 pts) How would you evaluate the KL divergence $KL(q(z|x)||p(z))$ if the prior $p(z)$ is replaced with a Gaussian Mixture distribution?
5. (5 pts) In evaluating the KL divergence between the ground-truth distribution $p(z)$ and the learned distribution $q(z)$, explain how $q(z)$ may turn out to be if the objective is to minimize $KL(p(z)||q(z))$ and $KL(q(z)||p(z))$. Here $q(z)$ is assumed to be an uni-modal distribution while $p(z)$ has two peaks.