Deep Learning and Practice — Final Exam

Date: Thursday, June 9, 2022
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

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

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

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

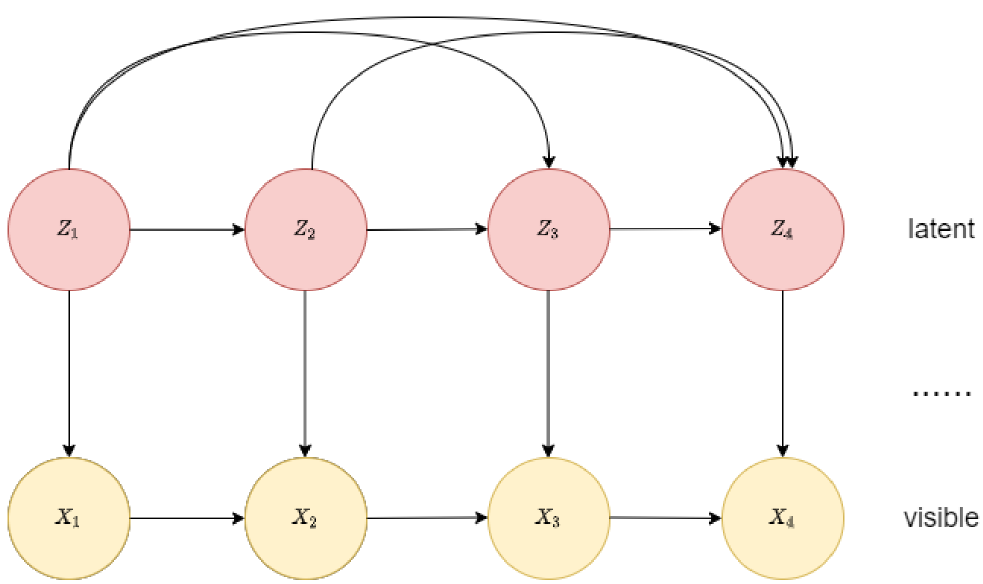
with the vectors $\boldsymbol{b}, \boldsymbol{c}$ and the matrix $\boldsymbol{W}$ denoting the model parameters.

(a) (5 pts) Show that $p(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{n} p(h_j|\boldsymbol{v})$ is factorial and $p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \boldsymbol{v}^T \boldsymbol{W}_{:,j})$, where $\boldsymbol{W}_{:,j}$ is the j-th column vector of $\boldsymbol{W}$.

(b) (5 pts) Show that $p(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{m} p(v_i|\boldsymbol{h})$ is factorial and $p(v_i = 1|\boldsymbol{h}) = \sigma(b_i + \boldsymbol{W}_{i,:} \boldsymbol{h})$, where $\boldsymbol{W}_{i;:}$ is the i-th row vector of $\boldsymbol{W}$.

(c) (5 pts) Assuming the model parameters are known, how can the $p(\boldsymbol{h}|\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h})$ be utilized to draw samples of $\boldsymbol{v}$ (and/or $\boldsymbol{h}$)? How would you draw independent samples of $\boldsymbol{v}$?

2. (32 pts) Consider the following latent factor model, where $Z_i, i = 1, 2, \ldots, T$ are latent variables and $X_i, i = 1, 2, \ldots, T$ 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 $\Pi_{t=1}^{T} p(X_t|" \cdot ")$. Make explicit the conditioning variables in " $\cdot$ " 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, dropout, activation function, and CNN.

   (a) (2 pts) Describe the difference between LeakyReLU, ELU, and ReLU.

   (b) (3 pts) Explain the idea of dropout.

   (c) (3 pts) Explain how the dropout works to evaluate multiple subnetworks during testing time.

   (d) (6 pts) What may cause gradient vanish problem and how to solve it? Explain your answer.

   (e) (6 pts) What is the size of the output feature map for an 256 x 256 input image after convolution with kernel (3,3), padding (2,2), and stride (2,2)?

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.