Some Example Questions (DL course)

1.1 DL basics & convolutions

Q1 Suppose you have a single-channel image of size $M \times N$ and a filter of size $m \times n$. Assume M, N, m, and n are *odd* numbers. What is the output image size if stride=2 and zero padding is used?

Q2 If the input image has size $256 \times 256 \times 32$, how many parameters are there in a single 3x3 convolution filter, including bias?

Q3 List two reasons why max pooling operation is often used in CNNs.

Q4 In binary classification tasks, what is the difference between accuracy and precision?

1.2 Variational auto-encoders

Q5 Given the following factorisation structure of p and q, draw the corresponding graphical model.

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{z}_1, \boldsymbol{z}_2) = p(\boldsymbol{z}_1)p(\boldsymbol{z}_2)p(\boldsymbol{x}_1|\boldsymbol{z}_1, \boldsymbol{z}_2)p(\boldsymbol{x}_2|\boldsymbol{z}_2).$$
$$q(\boldsymbol{z}_1, \boldsymbol{z}_2|\boldsymbol{x}_1, \boldsymbol{x}_2) = q(\boldsymbol{z}_1|\boldsymbol{x}_1, \boldsymbol{x}_2)q(\boldsymbol{z}_2|\boldsymbol{x}_2).$$

Q6 Which of the following statements are TRUE for VAEs?

1. The optimal encoder minimises $\mathrm{KL}[p(\boldsymbol{z}|\boldsymbol{x})||q(\boldsymbol{z}|\boldsymbol{x})]$;
2. The latent variable $\boldsymbol{z}$ in a VAE needs to be continuous;
3. Reconstruction error is the best evaluation metric for a VAE's test performance;
4. To train the encoder with gradient ascent, the $q(\boldsymbol{z}|\boldsymbol{x})$ must be reparameterisable;
5. The VAE objective provides a lower-bound to the MLE objective for learning the decoder parameters.

1.3 Generative adversarial networks

Q7 Assume the data distribution is $p_{data}(\boldsymbol{x}, y)$ where $\boldsymbol{x}$ represents an image and y represent its label. Define the GAN generator distribution as $p_\theta(\boldsymbol{x}|y)p(y)$. How do you construct a discriminator and what is its optimal form?

Q8 Which of the following statements are TRUE?

1. With the original GAN loss, saturated gradients occur for the generator only when the fake images look similar to the real ones;
2. Putting gradient norm constraints for the discriminator can help for the saturated gradient problem of the generator;
3. The original GAN objective in Goodfellow et al. (2014) is equivalent to minimising the Jensen-Shannon divergence for the generator;
4. For GAN losses in general, we can swap the ordering of min-max to max-min;
5. Mode collapse is likely to happen when the data distribution contains isolated modes.

1.4 Recurrent neural networks

Q9 Consider the following RNN to process the input sequence $\boldsymbol{x}_{1:T}$:

$$\boldsymbol{h}_t = \sigma(W_h \boldsymbol{h}_{t-1} + W_x \boldsymbol{x}_t + \boldsymbol{b}_h), \quad t = 1, ..., T$$

with the convention that $\boldsymbol{h}_0 = \boldsymbol{0}$, and $\sigma(\cdot)$ is the identity activation function. The RNN is trained by minimising some loss function $\mathcal{L}(\boldsymbol{h}_T)$. Now assume the recurrent weight matrix is

$$W_h = \begin{pmatrix} -0.64 & -0.06 & -1.12 \\ -0.56 & 0.28 & -0.76 \\ 0.62 & -1.66 & -0.2 \end{pmatrix}.$$

State whether gradient vanishing/explosion would happen when $T \to +\infty$ and explain why.

Q10 Which of the following solutions can help address the gradient vanishing/explosion problem in practice?

1. Truncated back-propagation through time;
2. Use Softplus activations;
3. Use GRU network instead of simple RNN;
4. Use batch normalisation;
5. Use orthonormal matrix for W_h.

1.5 Attention & Transformers

Q11 Assume we want to build a Transformer-like neural network to process images. In this case an input image of size $H \times W \times C$ is viewed as a collection of small patches of size $h \times w \times C$. Assume no padding is used, and the small patches are overlapping with stride 1. What is the time complexity of computing the dot product for self-attention on these patches? We assume a single head and the projection matrices are identity matrices.

Q12 Which of the following statements are TRUE?

1. We need to use position encoding because dot-product attention is permutation invariant;
2. The expected maximum index of input informs the hyper-parameter choice of Sinusoidal position encoding;
3. Both time and memory complexity figures for self-attention are quadratic in the number of input datapoints;
4. The attention matrix in soft attention is symmetric;
5. The sinusoid encoding with fixed output dimensionality is a one-to-one mapping.

Example Answers

A1 The output size is $(\lfloor M/2 \rfloor + 1) \times (\lfloor N/2 \rfloor + 1)$.

A2 It should contain $3 \times 3 \times 32 = 288$ weight parameters and 1 bias parameters, so in total 289 parameters.

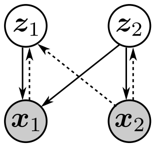
A3 First max pooling helps to reduce the feature map size (so subsequent layers have smaller dimensional inputs). Second, max pooling provides (approximate) local translation invariance (depending on the filter size).

A4 Every prediction falls into one of the following 4 categories: true positive (TP), false positive (FP), true negative (TF), false negative (FN). The definition of accuracy and precision are:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$
$$precision = \frac{TP}{TP + FP}.$$

Think about it in an information retrieval context where the system retrieves a document when the classifier predicts class "positive". Then accuracy concerns the overall prediction performance across all test inputs, while precision only cares about the percentage of correctly identified cases in all the retrieved documents.

A5 The p and q distributions correspond to the following graphical model (solid arrows for p and dashed arrows for q).



A6 Only choice 5 is correct.

The optimal encoder minimises $\mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{x})]$. VAEs allow the usage of discrete $\boldsymbol{z}$. In general VAEs can be trained without the reparameterisation trick (e.g., with the REINFORCE gradient, see lecture notes). It is possible that with poor tuning, a trained VAE achieves close to zero reconstruction error but still generates non-realistic fake data (refer to coursework practice).

A7 The discriminator takes a form like $D_\phi(\boldsymbol{x}, y) = p_\phi(\text{``}real\text{''}|\boldsymbol{x}, y)$. Assume for the GAN generator the label $y \sim p(y)$, then the optimal discriminator is

$$D^*(\boldsymbol{x}, y) = \frac{p_{data}(\boldsymbol{x}, y)}{p_{data}(\boldsymbol{x}, y) + p_\theta(\boldsymbol{x}|y)p(y)}.$$

Note here that it is not necessary that $p(y) = p_{data}(y)$.

A8 Choices 2, 3, 5 are correct.

For the original GAN loss, saturated gradients can occur for the generator at the very beginning when it is very easy for the discriminator to separate fake images from the real ones with high confidence. Gradient norm constraints for the discriminator help to avoid such sharp discriminator. In general the min-max ordering in an adversarial loss cannot be swapped. For mode collapse, refer to discussions in the lecture & coursework.

A9 The gradient will explode when $T \to +\infty$. The gradient $\nabla_{W_h}\mathcal{L}$ contains a $W_h^T \frac{d\boldsymbol{h}_1}{dW_h}$ term whose norm depends on the largest absolute eigenvalue of W_h. Since the largest absolute eigenvalue of W_h is approximately $1.15 > 1$, this means the gradient contains a term that will explode when $T \to +\infty$, therefore the gradient will explode.

A10 Choices 1, 3 are correct.

Changing activation in RNNs to Softplus alone does not address the issue of vanishing/exploding $||W_h^T||_2$. Making W_h orthonormal alone does not prevent vanishing gradient caused by activation function gradient. Batch-norm is not designed to help RNN convergence.

A11 The time complexity is $\mathcal{O}((H-h+1)^2(W-w+1)^2whC)$. There are $(H-h+1)\times(W-w+1)$ patches for computing self-attention, each patch has dimension $h \times w \times C$.

A12 Choices 2, 3 are correct.

Dot-product attention is permutation equivariant (notice the difference between invariance and equivariance). The sinusoid encoding with fixed output dimensionality is not one-to-one if the input position is in $\mathbb{R}$, so if we want to make it one-to-one (which might not be necessary), then we need to know the expected min & max input and design its frequency parameter accordingly. Also check the lecture notes for properties of dot-product attention.