Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

LH Neural Computation
LM Neural Computation extended

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Turn Over

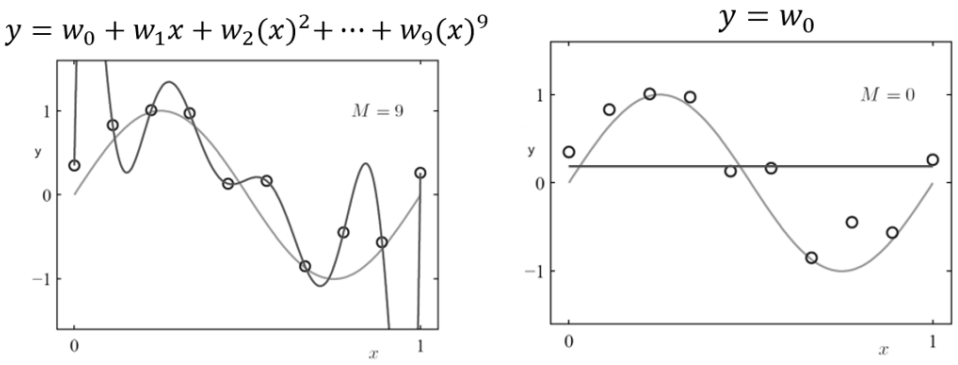
Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1

Please answer the following questions about machine learning fundamentals and diffusion models.

(a) Consider polynomial regression model and the plots below where x denotes a 1-dimensional independent variable (depicted on the horizontal axis), w denotes the model weight, y denotes the dependent response variable (depicted on the vertical axis), and M denotes the model complexity. The left-hand plot shows the model with $M = 9$ and the right-hand plot shows the model with $M = 0$. For both plots, the hollow circle marker depicts the original data points, the sinusoidal curve depicts the underlying function which was used to generate these points, and the polynomial curve depicts the learned model prediction.



With reference to the above context:

(i) Briefly describe underfitting and overfitting. (roughly 3-4 sentences)

(ii) Briefly describe how we can find a suitable model, which neither underfits nor overfits. (roughly 3-4 sentences)

[6 marks]

(b) Consider that you are required to train a neural network on a diverse dataset that includes data from different modalities (e.g., text, images, and numerical features). Propose an outline of a plan (in no more than 6 bullet points, of 1-2 sentences each) for mini-batch selection in the context of mini-batch stochastic gradient descent (SGD). Your proposed plan should: (i) handle the heterogeneity of the dataset, (ii) describe the rationale of your plan, and (iii) describe how it balances the computational efficiency of mini-batch training with the challenges posed by diverse data

types. Briefly state any hypotheses that you make to propose your outline of the plan. **[8 marks]**

(c) Microscopic imaging plays a crucial role in various scientific disciplines, providing insights into cellular structures and processes. Assume you are a researcher in the field of biology, and you have acquired a dataset of low-resolution microscopic images capturing cellular details. Your task is to employ diffusion models for image super-resolution, to enhance the image resolution. Provide a brief outline of your proposed plan to achieve image super-resolution utilising diffusion model, particularly highlighting:

- dataset preparation (roughly 2-3 sentences)
- changes needed to a conventional diffusion model (roughly 2-3 sentences)
- training strategy (roughly 2-3 sentences)

[6 marks]

Question 2

Given the weights (w_1, w_2, w_3) and the biases (b_2, b_3), we have the following recurrent neural network (RNN) which takes in an input vector x_t and a hidden state vector h_{t-1} and returns an output vector y_t:

$$y_t = \mathbf{g}(w_3\mathbf{f}(w_1 x_t + w_2 h_{t-1} + b_2) + b_3), \tag{1}$$

where $\mathbf{g}$ and $\mathbf{f}$ are activation functions. The following computational graph depicts such a RNN.

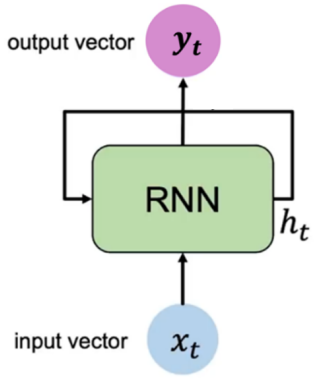


Figure 1: RNN Computational Graph

(a) Write down clearly which part of Equation (1) defines the current (updated) hidden state vector h_t shown in Figure 1. **[3 marks]**

(b) When $t = 3$ (starting from 1), please show how information is propagated through time by drawing an unfolded feedforward neural network that corresponds to the RNN in Figure 1. Please make sure that hidden states, inputs and outputs as well as network weights and biases are annotated on your network. **[4 marks]**

(c) Assume x_t, h_{t-1}, h_t and y_t are all scalars in Equation (1), and the activation functions are a linear unit and a binary threshold unit, respectively defined as:

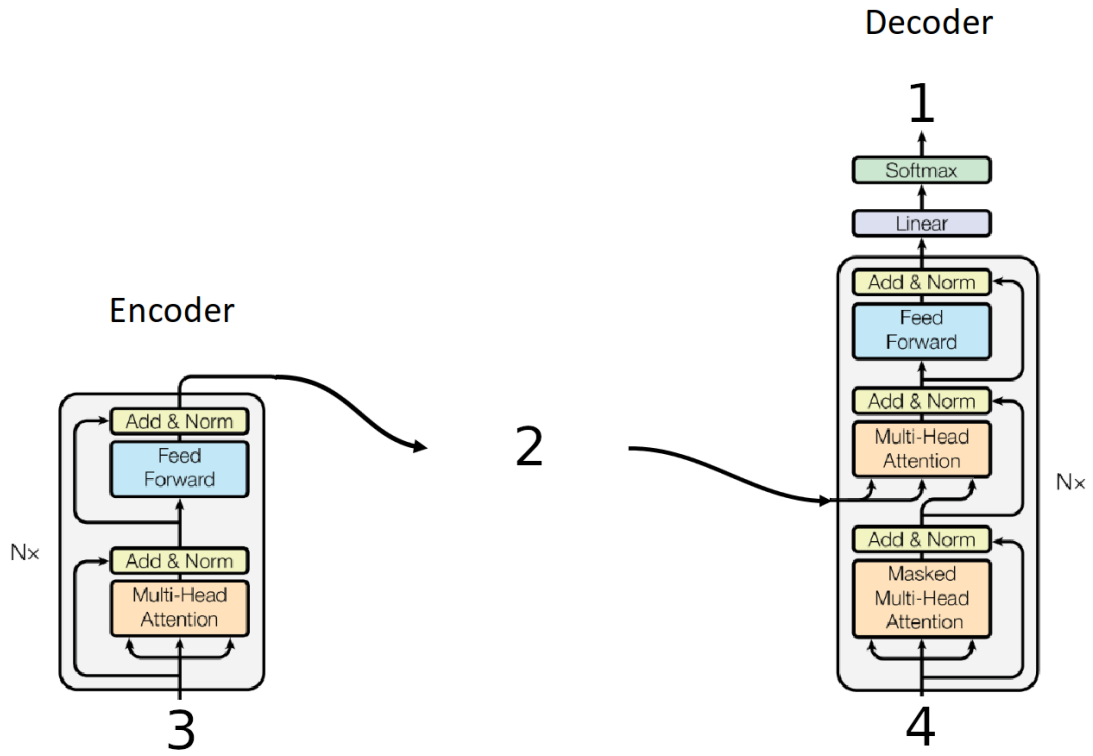
$$\mathbf{g}(x) = x,$$
$$\mathbf{f}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}.$$

When $t = 3$ (starting from 1), please calculate the values of the outputs (y_1, y_2, y_3) given $(w_1 = 1, w_2 = -3, w_3 = 5)$, $(b_2 = 1, b_3 = 3)$, $(x_1 = 5, x_2 = 3, x_3 = 1)$ and $h_0 = 0$. Please show your calculations in detail. **[3 marks]**

(d) Again let us assume x_t, h_{t-1}, h_t and y_t are all scalars with $h_0 = 0$ and the activation functions the same as above. Compute (w_1, w_2, w_3) and (b_2, b_3) such that the network outputs 0 initially, but when it receives an input of 1, it outputs 1 for all subsequent time steps. For example, if the input is 00001000100, the output will be 00001111111. Please justify your answer.

Note: Here we want a solution that satisfies (1) the hidden state h_t is zero until the input x_t becomes 1, at which point the hidden state changes to 1 forever, and (2) the output always predicts the same as the hidden state, i.e. $y_t = h_t$. Also note that to get full marks all possible values for (w_1, w_2, w_3) and (b_2, b_3) should be given. **[10 marks]**

Question 3



(a) A transformer network is being trained to translate German sentences into English sentences. Before processing, each word is mapped to a separate token. During training, the network is shown a training pair consisting of the German sentence *"Ich studiere Informatik"* and the English version *"I study computer science"*. Considering the German and English sentences and the network shown in the image above, please explain what the missing elements (1,2,3,4) are when the network is being trained (roughly 1-2 sentences each). **[8 marks]**

(b) Consider a single-head self-attention layer in the encoder part of the network, while the network is translating the sentence from (a). The layer uses a key length of 64 and the token embedding vector and value length 512. The data at every step of the self-attention calculation can be described as matrices. Use short bullet points to describe the calculation step by step. In particular answer: Which matrices hold the learnable parameters? What is the size of the involved matrices (num. rows×num. columns), including input and output matrices? **[8 marks]**

(c) How does the multi-head self-attention layer differ from the single-head attention layer discussed in (b)? (roughly 2 sentences) **[1 mark]**

(d) How does the masked self-attention layer differ from the basic unmasked layer? Why is it important for training? (roughly 2 sentences) **[2 marks]**

(e) Could training be possible without using masked self-attention? If so, how? (roughly 2 sentences) **[1 mark]**

End of Paper

LH Neural Computation
LM Neural Computation extended **Non-alpha only**

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.