**SWS3009A Robotics and Deep Learning**

**Deep Learning Lab 1 Answer Book**

**SUBMISSION DEADLINE: Friday 4 July 2025, 11.59 pm**

Team Member 1 Name: Liu Yijia

Team Member 2 Name: Rui Yuhan

Team Member 3 Name (if any):

**Marks:** \_\_\_\_\_\_\_\_ / 3

Please save as PDF before submitting to Canvas.

**Question 1 Answer:**

(Fill answer here)

Q: What does an Embedding layer do?

A: Embedding layers can map discrete input tokens which are integers generated from the tokenizer into dense, continuous vector representations.

Q: Why can’t we feed raw integers directly into an LSTM?

A: LSTM expects continuous-valued vector inputs, not discrete integers. Feeding them raw token IDs would mislead the LSTM into interpreting them as ordinal or scalar values.

**Question 2 Answer:**

(Fill answer here)

Q: Explain why we don't need to chop up our tokens into groups of 5 tokens to predict the 6th for transformers, but must do so for LSTMs.

A: LSTM is a sequential model that only relies on a fixed number of previous tokens to predict the next token during training. Therefore, we need to split the text into multiple fixed-length segments, like using the first 5 tokens to predict the 6th. In contrast, the Transformer uses a self-attention mechanism, allowing it to see the entire input sequence at once and automatically learn the relationships between tokens. As a result, it does not require manually chop up the input into fixed-length segments like LSTM does.

**Question 3 Answer:**

(Fill answer here)

Q: Why can't we use a single output to predict the index of the next word?

A: Because it would become a regression task, but word prediction is classification problem. One-hot with softmax gives probabilities for all words, which is better for training.

Q: Why do we use softmax and categorical crossentropy?

A: Softmax gives a probability distribution; crossentropy compares it to the true one-hot label to optimize the model effectively.