Exercise Sheet 5: Sequence Modeling in Vision

Due on 13.06.2025, 10:00

Timy Phan (timy.phan@lmu.de)

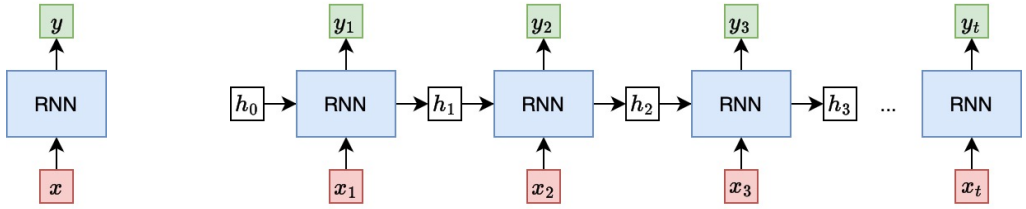
Important Notes:

1. **Email**: Frequently check your email address registered for Moodle. All notifications regarding the course will be sent via Moodle.
2. **Due date**: The exercise sheets will usually be uploaded 1 week in advance of the due date, which is indicated at the top of the exercise sheet.
3. **Submission**: Put your report (a single PDF file) inside a single ZIP file. Both PDF file and ZIP file should contain your surname and your matriculation number (e.g. *Surname-MatriculationNumber.zip*). You may use Jupyter[1] for exporting your PDF. **Submissions that fail to follow the naming convention will not be graded.**

General Information:

If you have any problems or questions about the exercise, you are welcome to use the student forum on the lecture Moodle page, as most of the time other students might have similar question. For technical issues about the course (for example, in case you cannot upload the solution to Moodle) you can write an email to the person responsible for the exercise (indicated at the top of the exercise sheet).

[1] https://jupyter.org/



(a) Vanilla RNN. (b) Vanilla RNN unrolled over time, indicated by the subscript t.

Figure 1: Example of a Vanilla Recurrent Neural Network.

Task 1: Recurrent Neural Networks (RNNs) (3P)

In this task, you will take a closer look at the workings and limits of RNNs (s. Figure 1).

1. Why is the hidden state h_t not equivalent to the model output y_t? (1P)
2. Given long sequences, name and briefly explain 2 factors which can limit the RNN's ability to handle long-term relations/dependencies in the data. (2P)

Task 2: Video captioning (3P)

In this task, you will review the implications of sequence modeling for video captioning.

1. Compared to performing image captioning for each individual frame, what is the advantage of modeling videos as a sequence of frames for video captioning? (1P)
2. Compare challenges of live video captioning (producing a caption while the model is processing the video) vs. offline video captioning (making a caption after the model has processed the entire video already)? (2P)

Task 3: Basics of the Attention mechanism (14P)

In this task, you will dive into the details of basic single-head (self-)attention. As in [Vaswani et al., 2017], Attention is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

The product QK^T is also known as the attention matrix.

1. Give an algebraic definition of the Q, K and V matrices. What are the relationships between the shapes of these matrices? (3P)

2. Which 2 factors determine the shape of the attention output? (1P)

3. The softmax operation is meant to normalize the scores in the attention matrix such that they can be interpreted as a probability distribution. Which dimension is this normalization meant for, i.e. which values in the normalized matrix are meant to sum up to 1? (1P)

4. Given

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, K = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, V = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 6 & 8 & 9 \end{pmatrix} \tag{2}$$

   calculate the attention matrix and the final output of the attention operation. For simplicity, *omit the softmax normalization*, such that the final output is $\frac{QK^T}{\sqrt{d_k}}V$. (3P)

5. In the general case for Equation (1), can we choose the Q and K matrices such that the attention mechanism approximates an identity function for V? (e.g. by setting $Q = K$)? (4P)

6. Compare the (self-)attention mechanism to the convolutional layer in terms of runtime complexity w.r.t. image size. Assume an image with height H and width W is flattened to a sequence of length $N = H \cdot W$ (the number of pixels), such that we have N queries and N key-value pairs for self-attention. For the convolutional layer, assume kernel size k. (2P)

Important Note: Submit exactly one ZIP file via Moodle before the deadline. The ZIP file should contain your report in PDF format. The PDF file should contain all figures, explanations and answers to questions. Make sure that plots (if included) have informative axis labels, legends, and captions.

References

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. ii