Exercise Sheet 5: Sequence Modeling in Vision

Solution

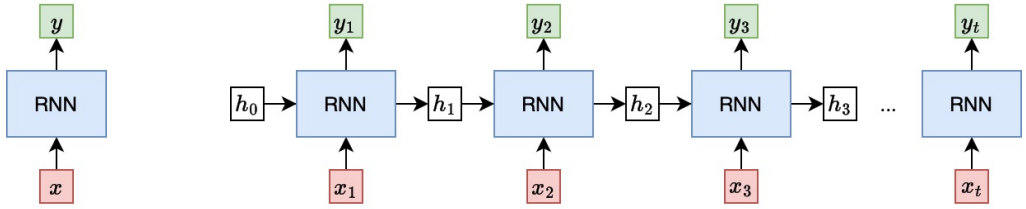
Important Notes:

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)
   [Solution: The hidden state can contain more information than is visible in the output, e.g. long-term dependencies which are not visible in the output at t.]

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)
   [Solution:
   - **The expressiveness (dimensionality) of h_t is fixed which limits the amount of information that can be stored/remembered.**
   - **Backpropagation through time can lead to vanishing/exploding gradients during training which is exacerbated by long sequences.**

   ]

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)

[Solution: Frame-by-frame captioning cannot directly capture transitions between the frames whereas sequence modeling enables video captioners to aggregate information over the entire video.]

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)
[Solution:

- **Live video captioning: harder problem because of less context for immediate prediction (no content from unseen frames). As a consequence, accuracy can be expected to suffer compared to offline captioning.**
- **Offline video captioning: can use and often requires more resources (memory, runtime) to maximize accuracy. This also prevents applicability to real-time settings like e.g. livestreams.**

]

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)
[Solution:
Formally, $Q \in \mathbb{R}^{N_Q \times d_k}$, $K \in \mathbb{R}^{N_{KV} \times d_k}$ and $V \in \mathbb{R}^{N_{KV} \times d_v}$.
In order to obtain the attention matrix $QK^T \in \mathbb{R}^{N_Q \times N_{KV}}$, Q and K need to be eligible for a matrix multiplication (share the common dimension d_k).
The attention matrix then needs to be multiplied with K which means they need to share a common dimension N_{KV} which is the number of key-value pairs.]

2. Which 2 factors determine the shape of the attention output? (1P)
[Solution: The number of queries (N_Q) and the dimensionality of the values (d_v).]

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)
[Solution: The softmax operation normalizes the attention scores across N_{KV} such that the scores for every key sum up to 1 per query.]

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)
[Solution:

- **Matrix multiplication of Q with K^T.**

$$QK^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

- **Normalization of the attention matrix.**

$$\frac{QK^T}{\sqrt{d_k}} = \frac{QK^T}{\sqrt{4}} = \frac{QK^T}{2} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 \\ 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

- **Matrix multiplication of the normalized attention matrix with**

V.

$$\frac{QK^T}{\sqrt{d_k}}V = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 \\ 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 6 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 11.0 & 15.0 & 18.0 \\ 5.0 & 6.5 & 7.5 \\ 2.0 & 2.5 & 3.0 \\ 0.5 & 1.0 & 1.5 \\ 3.5 & 5.0 & 6.0 \end{pmatrix}$$

]

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)
[Solution:

- **Possible if V is multiplied with the identity matrix I.**
- **Recall the definition of softmax:**

$$\textbf{softmax}(x)_i = \frac{e^{x_i}}{\sum_j^D e^{x_j}}$$

where $i, j \in [1, D]$ and $x \in \mathbb{R}^D$. Note that this means if x is a one-hot vector (only one element, e.g. $x_o = 1$ where $o \in [1, D]$, the rest $x_r = 0$ for $\forall_{r \in [1,D]} : r \neq o$), then the softmax returns $\textbf{softmax}(x)_o = \frac{e^1}{\sum_j^D e^{x_j}} = \frac{e}{e+D-1}$ and $\textbf{softmax}(x)_r = \frac{e^0}{\sum_j^D e^{x_j}} = \frac{1}{e+D-1}$. For high D, the difference between $\textbf{softmax}(x)_o$ and $\textbf{softmax}(x)_r$ diminishes $\rightarrow$ if given a one-hot input, the softmax output is closer to a uniform distribution than a one-hot vector for high D.

- **The softmax can distribute weights such that one key (o) is weighted close to 1 and all other keys (r) are weighted close to 0 if the logit for that one key is substantially higher than for all other keys, such that $x_o >> x_r$. This is due to the fact that the exponential functions in the softmax will increase the deltas between scores.**
- **The attention matrix QK^T should be square ($N_Q = N_{KV}$) and diagonal (zeros everywhere except on the main diagonal). For $\lim_{c \to +\infty}$ where c are the scalars on the main diagonal, the softmax output will approximate the identity matrix.**
- **In order for QK^T to be diagonal, (A) the rows of Q and K should be (ideally) orthonormal and (B) Q and K should be**

inverses of another (if we neglect the scale c) such that $QK^T = QQ^{-1} = I$.

- **As such a possible solution is $K = I$ and $Q = c \cdot I$ (for large $c >> 1$).**

]

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)
[Solution:

- **Attention: the matrix multiplications (QK^T and attention matrix with V) and the softmax operation each have $\mathcal{O}(N^2)$ complexity (due to N queries and N key-values**
- **Convolution: for each pixel, convolution has $\mathcal{O}(k^2)$ complexity and thus complexity is $\mathcal{O}(N \times k^2)$**

]

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. iii