Exercises for the course
Deep Learning 2
Summer Semester 2023

Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 2

Exercise 1: Gradient computation in attention heads (50 P)

The Transformer model [?] uses a specific form of attention, namely self-attention layers, to extract the task-relevant information from the available features. It herein uses the query, key and value projections of the layer inputs (QKVa-attention). In this analytical exercise 1 of this week's sheet, we will focus on the structure of the gradient computation of the attention head module. In the programming exercise 2, you will implement the QKV-attention module used in Transformer models.

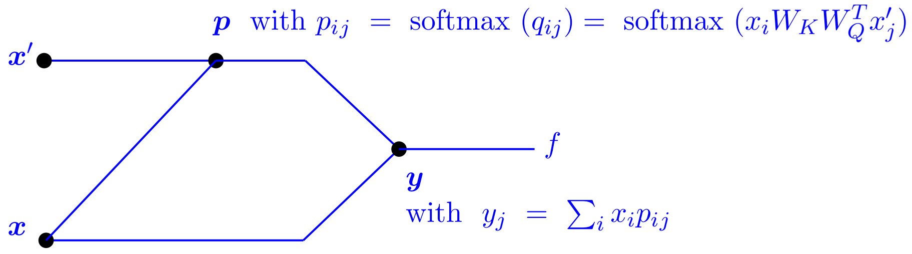
For this, recall the following equations:

$$y_j = \sum_i x_i p_{ij} \tag{1}$$

$$p_{ij} = \frac{\exp(q_{ij})}{\sum_{i'} \exp(q_{i'j})} \quad \text{with} \quad q_{ij} = x_i^T W_K W_Q^T x'_j \tag{2}$$

Hint: Consider the multivariate chain rule. You do not need to solve for the full analytical solution, but write down the correct structure of the required gradients.

(a) Draw a schematic diagram how the block input vectors x_i and x'_j interact with the attention weights p_{ij} to produce the layer output y_j and finally the attention block output f. Nodes of the computation graph should consider $\{f,\ \boldsymbol{p},\ \boldsymbol{y},\ \boldsymbol{x}$ and $\ \boldsymbol{x'}\}$.


$\boldsymbol{p}$ with $p_{ij} = \text{softmax}\ (q_{ij}) = \text{softmax}\ (x_i W_K W_Q^T x'_j)$
$\boldsymbol{x'}$
f
$\boldsymbol{y}$
with $y_j = \sum_i x_i p_{ij}$
$\boldsymbol{x}$

(b) Write down the gradient to compute $\partial f / \partial x'_j$ using the relevant local gradients of the involved variables.

$$\frac{\partial f}{\partial x'_j} = \sum_k \sum_i \frac{\partial f}{\partial y_k} \cdot \frac{\partial y_k}{\partial p_{ik}} \cdot \frac{\partial p_{ik}}{\partial x'_j} \tag{3}$$

(c) Write down the gradient to compute $\partial f / \partial x_i$using the relevant local gradients of the involved variables.

$$\frac{\partial f}{\partial x_i} = \sum_j \frac{\partial f}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i} + \sum_l \sum_j \frac{\partial f}{\partial y_j} \cdot \frac{\partial y_j}{\partial p_{lj}} \cdot \frac{\partial p_{lj}}{\partial x_i} \tag{4}$$

$$\tag{5}$$

$$= \sum_j \frac{\partial f}{\partial y_j} \left(\frac{\partial y_j}{\partial x_i} + \sum_l \frac{\partial y_j}{\partial p_{lj}} \cdot \frac{\partial p_{lj}}{\partial x_i} \right) \tag{6}$$

Exercise 2: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

References

[Unke et al.(2021)Unke, Chmiela, Sauceda, Gastegger, Poltavsky, Schütt, Tkatchenko, and Müller] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.