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

# Exercise Sheet **Attention**

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

The Transformer model [1] 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 (QKV-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:

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

$$p_{ij} = \frac{\exp(q_{ij})}{\sum_{i'} \exp(q_{i'j})}, \tag{2}$$

with embedded sequences $\{\boldsymbol{x}, \boldsymbol{y}\} \in \mathbb{R}^{L \times D}$ of length $L$ and hidden dimension $D$, and $p_{ij}$ the scalar attention weights computed from raw query-key dot-products $q_{ij}$ (that depend on embedding the input $\boldsymbol{x}$ using the $W_K$ weight matrix and $\boldsymbol{x'}$ using $W_Q$ respectively). Inputs $\boldsymbol{x}$ and $\boldsymbol{x'}$ contain the same values but get assigned different functions in the self-attention module.

Hint: Consider the multivariate chain rule in (c) and (d). You do not need to solve for the full analytical solution, only write down the structure of the required gradients. For simplification, you may assume a hidden dimension of $D = 1$.

(a) State the formula for how $q_{ij}$ is computed from the inputs $x_i$ and $x'_j$ using the key and query weight matrices.

(b) Draw a schematic diagram how the block inputs $x_i$ and $x'_j$ interact with the attention weights $p_{ij}$ to produce the layer output $y_j$ and finally $f$ the output of the neural network (e.g. a class probability score). The computation graph should consider $\{f,\ \boldsymbol{p},\ \boldsymbol{y},\ \boldsymbol{x} \text{ and } \boldsymbol{x'}\}$.

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

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

## Exercise 2: Programming (50 P)

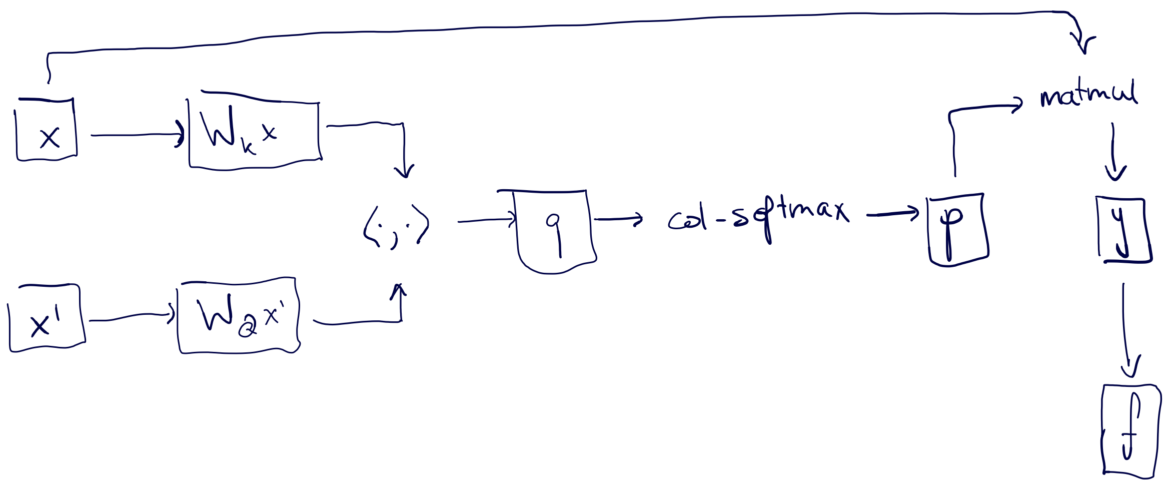
Download the programming files on ISIS and follow the instructions.

# References


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


1 a) $$q_{ij} = (W_K x_i)^T (W_Q x'_j) = x_i^T W_K^T W_Q x'_j$$

b) Alternative to the solution in the PDF:




c)

$f(y_1, \dots, y_L)$

$y_i(x_1, \dots, x_L, p_{1i}, \dots, p_{Li})$

$p_{ij}(q_{1j}, \dots, q_{Lj})$

$q_{ij}(x_i, x'_j)$

Multivariable chain rule:

Simple case: $f(x,y)$ ; $x(t)$ ; $y(t)$

$$\frac{d f(x,y)}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}$$

more general: $f(x_1, \dots, x_L)$ ; $x_i(t)$, $i = 1, \dots, L$

$$\frac{df}{dt} = \sum_{i=1}^{L} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}$$

$$\frac{df}{dx'_j} = \sum_k \frac{\partial f}{\partial y_k} \frac{dy_k}{dx'_j} = \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{i=1}^{L} \left[ \overbrace{\frac{\partial y_k}{\partial x_i} \overbrace{\frac{dx_i}{dx'_j}}^{=0}}^{=0} + \frac{\partial y_k}{\partial p_{ik}} \frac{dp_{ik}}{dx'_j} \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{i=1}^{L} \frac{\partial y_k}{\partial p_{ik}} \frac{d p_{ik}}{d x'_j} = \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{i=1}^{L} \frac{\partial y_k}{\partial p_{ik}} \sum_{m=1}^{L} \frac{\partial p_{ik}}{\partial q_{mk}} \frac{d q_{mk}}{d x'_j}$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{i=1}^{L} \frac{\partial y_k}{\partial p_{ik}} \sum_{m=1}^{L} \frac{\partial p_{ik}}{\partial q_{mk}} \left( \underbrace{\frac{\partial q_{mk}}{\partial x_m} \frac{d x_m}{d x'_j}}_{=0} + \frac{\partial q_{mk}}{\partial x'_k} \frac{d x'_k}{d x'_j} \right)$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{i=1}^{L} \frac{\partial y_k}{\partial p_{ik}} \sum_{m=1}^{L} \frac{\partial p_{ik}}{\partial q_{mk}} \frac{\partial q_{mk}}{\partial x'_k} \underbrace{\frac{d x'_k}{d x'_j}}_{= \begin{cases} 0 & \text{if } k \neq j \\ 1 & \text{if } k = j \end{cases}}$$

$$= \frac{\partial f}{\partial y_j} \sum_{i=1}^{L} \frac{\partial y_j}{\partial p_{ij}} \sum_{m=1}^{L} \frac{\partial p_{ij}}{\partial q_{mj}} \frac{\partial q_{mj}}{\partial x'_j} = \sum_{i=1}^{L} \frac{\partial f}{\partial y_j} \frac{\partial y_j}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial x'_j}$$

* at the bottom

Recall:

$f(y_1, \dots, y_L)$

$y_i(x_1, \dots, x_L, p_{1i}, \dots, p_{Li})$

$p_{ij}(q_{1j}, \dots, q_{Lj})$

$q_{ij}(x_i, x'_j)$

d)

$$\frac{df}{dx_i} = \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \frac{d y_k}{d x_i}$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \sum_{j=1}^{L} \left[ \frac{\partial y_k}{\partial x_j} \underbrace{\frac{d x_j}{d x_i}}_{= \begin{cases} 0 & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases}} + \frac{\partial y_k}{\partial p_{jk}} \frac{d p_{jk}}{d x_i} \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \left[ \frac{\partial y_k}{\partial x_i} + \sum_{j=1}^{L} \frac{\partial y_k}{\partial p_{jk}} \sum_{m=1}^{L} \frac{\partial p_{jk}}{\partial q_{mk}} \frac{d q_{mk}}{d x_i} \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \left[ \frac{\partial y_k}{\partial x_i} + \sum_{j=1}^{L} \frac{\partial y_k}{\partial p_{jk}} \sum_{m=1}^{L} \frac{\partial p_{jk}}{\partial q_{mk}} \left( \frac{\partial q_{mk}}{\partial x_m} \overbrace{\frac{d x_m}{d x_i}}^{= \begin{cases} 0 & \text{if } m \neq i \\ 1 & \text{if } m = i \end{cases}} + \overbrace{\frac{\partial q_{mk}}{\partial x'_k} \frac{d x'_k}{d x_i}}^{=0} \right) \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \left[ \frac{\partial y_k}{\partial x_i} + \sum_{j=1}^{L} \frac{\partial y_k}{\partial p_{jk}} \sum_{m=1}^{L} \frac{\partial p_{jk}}{\partial q_{mk}} \frac{\partial q_{mk}}{\partial x_m} \frac{d x_m}{d x_i} \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \left[ \frac{\partial y_k}{\partial x_i} + \sum_{j=1}^{L} \frac{\partial y_k}{\partial p_{jk}} \frac{\partial p_{jk}}{\partial q_{ik}} \frac{\partial q_{mk}}{\partial x_i} \right]$$

$$= \sum_{k=1}^{L} \frac{\partial f}{\partial y_k} \left( \frac{\partial y_k}{\partial x_i} + \sum_{j=1}^{L} \frac{\partial y_k}{\partial p_{jk}} \cdot \frac{\partial p_{jk}}{\partial x_i} \right)$$

(*) the solution pdf states the following formula, which looks different from what is written above:

$$\frac{df}{dx_j'} = \sum_{k=1}^{L} \sum_{i=1}^{L} \frac{\partial f}{\partial y_k} \cdot \frac{\partial y_k}{\partial p_{ik}} \cdot \frac{\partial p_{ik}}{\partial x_j'}$$

The difference are the two sums vs. the single sum, as written above. But, if we look at the formula for $p_{ik}$,

$$p_{ik} = \frac{\exp(q_{ik})}{\sum_{i'} \exp(q_{i'k})} = \frac{\exp(x_i W_k W_Q x_k')}{\sum_{i'} \exp(x_i W_k W_Q x_k')}$$

we see that, for $k \neq j$, $p_{ik}$ does not depend on $x_j'$. Thus, $\frac{\partial p_{ik}}{\partial x_j'}$ is zero for every $k \neq j$. That in turn leaves us with the formulation used above.