NNDL2024: Exercices & Assignments
Session 4 (18 Apr), Deadline 22 Apr

Mathematical exercises

1. (20pts) Positional encoding in transformers. Consider a transformer with N inputs $\mathbf{x} \in \mathbb{R}^D$ and positional embeddings $\mathbf{p} \in \mathbb{R}^D$. Assume the positional embeddings are orthogonal, so that $\mathbf{p}_i^T \mathbf{p}_j = 0$ if $i \neq j$.

   (a) Is it always possible to construct such positional embeddings? If not, what are the necessary requirements?

   (b) The core building block of self-attention is the inner product $(\mathbf{W}_q \mathbf{x}_i)^T (\mathbf{W}_k \mathbf{x}_j)$ between the queries and keys. Write down and simplify the corresponding inner product for two alternative ways of encoding the positional information: Addition $\mathbf{x}_i + \mathbf{p}_i$ and concatenation $[\mathbf{x}_i; \mathbf{p}_i]$ of the elements. Explain how the two alternatives differ.

   (c) Would one-hot-encoding (each $\mathbf{p}_i$ is a vector of zeroes except for the ith entry that is one) be a good positional embedding? If not, explain what kinds of problems it would have.

2. (20 points) Find one scientific article that uses transformers for modelling some other type of data than language, images or videos. Tell which paper you read, giving the paper title, the author names, and a url. Read the paper and describe how they use it, covering aspects like:

   - What is the task? For instance, what do they try to predict.
   - How is the input data processed? Is is a set or a sequence? Explain key preprocessing steps needed to convert the data into a suitable format (e.g. tokenization in language models).
   - What kind of architecture was used? How many layers and parameters? Are there some interesting non-standard components?
   - How was the training done? Supervised? Self-supervised? Is some form of transfer learning used? How much data was used?
   - Did it work well? What was the method compared against and how did it fare?
   - Your subjective feeling about the paper: Is the model good? Was the paper clear in communicating these aspects?

3. (20pts) Suppose we observe a bivariate Gaussian data $(x, y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (0, 0), \boldsymbol{\Sigma} = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$$

   for some constant c. Now, we want to estimate c. Let's devise a very simple self-supervised algorithm: Learn a linear regression by least-squares to predict y from x. Show (mathematically, not empirically) that this learns the parameter c. Discuss how you can interpret this as self-supervised learning.

Computer Assignment

1. (40pts) Self-attention. Your task is to implement the multi-head self-attention operation, taking in a set of N vectors of D dimensions and outputting a matrix of the same size. Do this without relying on neural network libraries, but rather write directly the required operations in `NumPy`.

   (a) Implement a function that performs the standard self-attention operation. Include the dot product scaling, but do not add any extra components (positional encodings etc).

   (b) Implement the multi-head version, so that use the function you implemented above as a component. You can assume that the number of heads H is chosen so that D/H is an integer. Remember to include the final transformation of the concatenated output.

   (c) Apply your codes for the case provided in the notebook (with $N = 5$ and $D = 6$) for both $H = 1$ and $H = 3$, using the identity matrix as the final transformation in the multi-head case. For $H = 3$ you should use the same weight matrices but interpret them so that the first H rows correspond to the first head, the next H rows to the 2nd head etc.
   Report: For both cases print out (a) the output of the operation and (b) the attention weights the operation uses. We use the exact same weight matrices in both cases but the results are not identical, even when using identity transformation in the end. Why?

   (d) What happens for the output if you change the order of the first two inputs?