Machine Learning and Physics
Winter Semester 2024/2025

Exercise Sheet #7

Due: 09.12.24 14h00

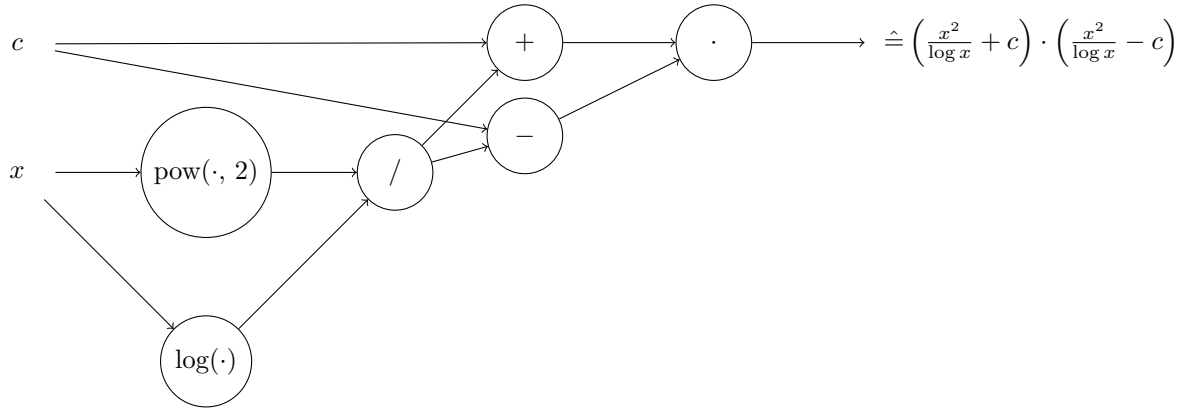
Peter Lippmann

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group). Submissions by a single person alone will not be corrected.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes please make sure that they are legible and not too blurry.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/sciai-lab/mlph_w24`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in your notebook (`.ipynb`), as well as an exported pdf-version of it.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of two. Specify all names of your group in the submission.

1 Reverse Mode Automatic Differentiation

Consider the following (admittedly contrived) computational graph. It was created as a small example with a central node that has a fan-out both in the upstream and downstream direction.



(a) Write down the analytical expression of the derivative of the output w.r.t. x via the chain rule, treating each node as a separate function (no need to simplify the result, e.g. keep terms like $\frac{\partial(\frac{x^2}{\log(x)})}{\partial x^2}$ as they are). (2 pts)

(b) Give the backprop forward trace for $x = 3,\ c = 5$ (by hand). (2 pts)

(c) Give the backprop backward/reverse trace for $x = 3,\ c = 5$ (by hand). Which is simpler, backprop or symbolic differentiation? (3 pts)

(d) Implement the computation in pytorch and use the autograd framework (via `.backward()`) to compute the derivatives w.r.t. x and c. (1 pt)

2 ADAM optimizer

(a) Write down the formulae for the Adaptive Moment Estimation (ADAM) optimizers and write a short sentence describing what each line does. (2 pts)

(b) Show that in the very first iteration the components of the gradient g are reduced to $\text{sign}(g)$. (2 pts)

(c) Bonus: How about the second iteration? (2 pts)

(d) Do you have a proposal how to deal with this problem? (1 pt)

(e) Consider an MLP with a set of weights w trained with Adam and L2-regularization. Does it make a difference whether the L2-penalty $\|w\|_2^2$ is included in the loss or whether weight decay is applied to the weights directly? Can you argue why one may be better than the other? *Hint: AdamW.* (3 pts)

3 Receptive Field of VGG16

(a) Using pen and paper, compute the field of view of the famous VGG16 convolutional network (https://arxiv.org/abs/1409.1556) for perceptrons just before the fully connected layers. That is, compute the set of input pixels that the prediction in a certain output pixel depend upon. All filters have size 3×3 and the max pooling is over 2×2 pixels. (3 pts)

(b) Compute the number of parameters of the full VGG16 architecture. What is the ratio of the parameters in the convolutional layers and the parameters in the fully connected layers? (3pts)

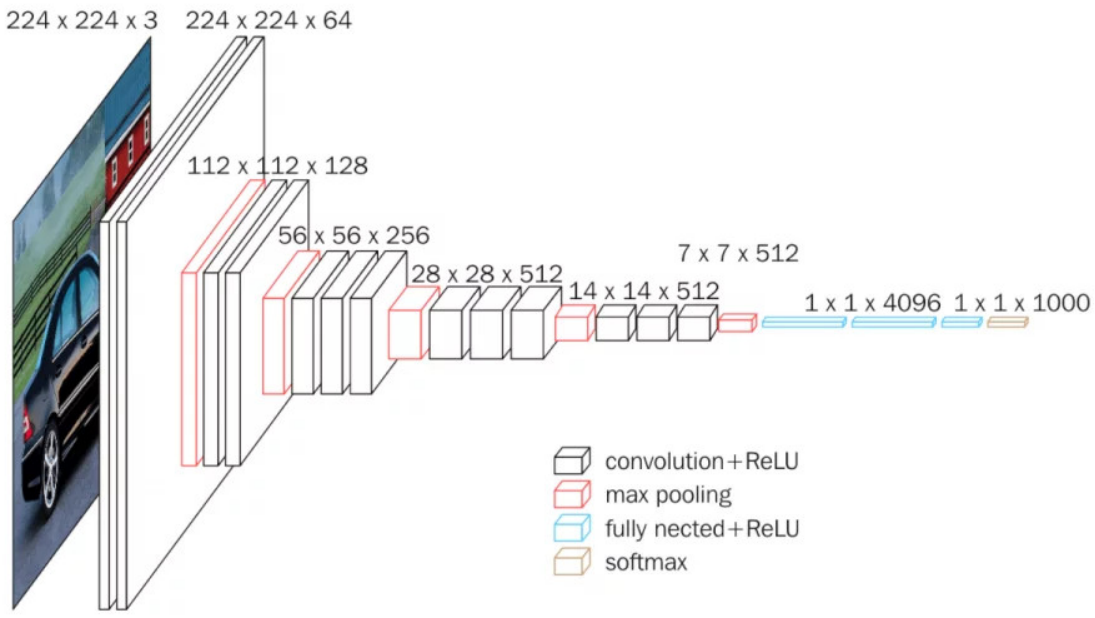


Figure 1: Schematic VGG16 architecture, taken from here.