Institut for Theoretical Physics
University Leipzig

Prof. Dr. B. Rosenow
M. Kühn

Statistical Mechanics of Deep Learning - Problem set 4

Winter Term 2024/25

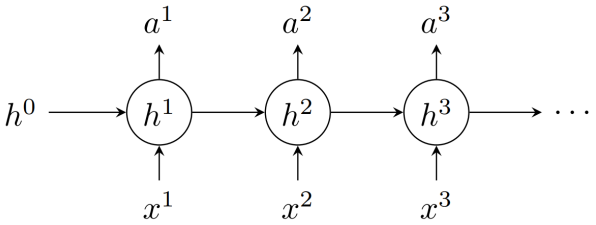
Hand in Python code: Before **Monday 11.11.2024, 9:15**, only submit the Python code you have written. Share a Google Colab Notebook with your code and send the link via email to itpleipzig@gmail.com.

8. Recurrent Neural Network: Manual Construction *2+1+3 Points*

In this exercise, we consider a straightforward Recurrent Neural Network (RNN) with ReLU nonlinearity, i.e.,

$$h^t = \mathrm{ReLU}(Wh^{t-1} + Ux^t + b),$$
$$a^t = Vh^t$$

for $t = 1, 2, \ldots$, where ReLU is evaluated component-wise. The RNN operates on the characters 'h', 'a', 'l', 'o', 'e', and 'g' as input and output, encoded by 'h' $= e_1$, 'a' $= e_2$, 'l' $= e_3$, 'o' $= e_4$, 'e' $= e_5$, and 'g' $= e_6$ (with $e_i \in \mathbb{R}^6$ representing the i-th unit vector).



(a) First, let $U = 0$ and $b = 0$, meaning the input x^t does not play a role. Choose a suitable dimension H for the hidden state h^t (thus $h^t \in \mathbb{R}^H$ for all t), and find an initial state h^0 as well as matrices $W \in \mathbb{R}^{H\times H}$ and $V \in \mathbb{R}^{6\times H}$ so that the output sequence "hallo" is generated, i.e., $a^1 = e_1$, $a^2 = e_2$, $a^3 = e_3$, $a^4 = e_3$, $a^5 = e_4$.

Hint: Distinguish between the first and second occurrence of 'l' in "hallo" within the hidden state.

(b) Show that for $u, v \in \{0, 1\}$ and a suitable choice of the bias value b, the expression $\max(u + v + b, 0)$ is equal to the product uv, i.e., it computes the logical AND $u \wedge v$.

(c) Now, using an appropriate choice of W, U, b, and V with initial state $h^0 = 0$, implement a simple version of word completion: the network should output the sequence "hallo" when given the input $x^1 = e_1$, $x^2 = e_2$, and $x^t = 0$ for all $t \geq 3$; and the sequence "helga" when given $x^1 = e_1$, $x^2 = e_5$, and $x^t = 0$ for all $t \geq 3$. The correct entry of the first two letters 'h', 'a' or 'h', 'e' is mandatory: for example, the input $x^1 = e_2$, $x^t = 0$ for all $t \geq 2$ should not be completed to "allo".

9. Residual connections *6 Points*

(a) Train fully connected neural networks on the dataset MNIST for different numbers of hidden layers, try at least 2, 6, and 8 hidden layers. Use the ReLU activation function for all hidden layers and cross entropy as the loss function. Train for 5 epochs and use a hidden layer size of 30 for all hidden layers. Plot the final test accuracy achieved against the number of hidden layers. What do you observe?

(b) Repeat part a, but alter the network architecture by adding residual connections every two hidden layers (https://arxiv.org/abs/1512.03385). This means that you should add the output of the first hidden layer to the preactivations of the third hidden layer, then add the new output of the third hidden layer to the preactivations of the fifth hidden layer (if there is one), and so on. Again, plot the final test accuracy achieved against the number of hidden layers. Do you observe any differences as compared to part a?

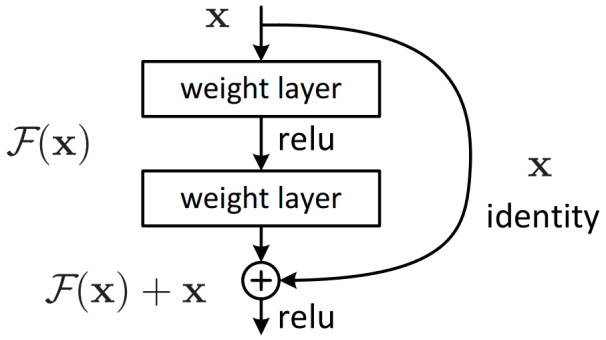


Figure 1: Residual building block. Taken from https://arxiv.org/abs/1512.03385