* The problem is that derivative of a loss function is so small that it is rounded to 0 (it is equal to around 10 \*\* -70). This is so called vanishing gradient problem. The other possible problem might be that derivative becomes very big (exploding gradient problem). Possible solutions might be:

- LSTM.

- gradient clipping: setting max and min threshold for all derivatives

- layer normalization

-

* Moge uzyc biblioteki caffe2 zeby przyspieszyc obliczenia