Deep Learning for NLP 2020
Exercise 08

June 6, 2020

1 Pingo

Try to find the right answer(s) to each question on your own or in a group with your colleagues. The interactive survey will be conducted near the end of the practice class.

- What are components of GRUs?
  - ☐ Dropout Gate
  - ☐ Reset Gate
  - ☐ Update Gate
  - ☐ Gradient Clipping Gate
- How do GRUs differ from traditional RNNs?
  - ☐ They have better control over the impact of the input on the memory
  - ☐ They can overwrite the full memory if needed
  - ☐ They can, in theory, store dependencies to previous inputs which were fed arbitrarily far away in the past

2 RNN Extensions

1. What is the benefit of bidirectional RNNs over unidirectional RNNs? Explain in up to two sentences and give an example.
2. What is the benefit of adding "output connections" to RNNs? Explain in up to two sentences and give an example.

3 Theoretical Background of Vanishing Gradients

In the lecture on backpropagation, we derived the formula for the derivative of the loss function with respect to a neuron p_i in an MLP as:

$$\frac{\partial E}{\partial p_i} = \sum_j \frac{\partial E}{\partial y_j} \cdot \sigma'(z_j) \cdot w_{i,j}$$

If we refer to a neuron p_i in layer k as $p_i^{(k)}$, this becomes:

$$\frac{\partial E}{\partial p_i^{(k)}} = \sum_j \frac{\partial E}{\partial p_j^{(k+1)}} \cdot \sigma' \left(z_j^{(k+1)} \right) \cdot w_{i,j}^{(k+1)}$$

1. Unfold one iteration of backpropagation: Write $\frac{\partial E}{\partial p_r^{(k-1)}}$ as a function of $\frac{\partial E}{\partial p_j^{(k+1)}}$ by using the formula given above.
2. Without applying the formula another time: How would $\frac{\partial E}{\partial p_s^{(k-2)}}$ depend on σ'?
3. Based on your findings in a) and b): Which activation function would you recommend for an MLP with more than 3 hidden layers? What would happen if you chose a less suited activation function?