Prof. Dr. Andreas Hotho, Albin Zehe, Konstantin Kobs
Lehrstuhl für Data Science

29./30.06.2021

7. Assignment in "Machine Learning for Natural Language Processing"

Summer Term 2021

1 General Questions

1. What are the differences between a recurrent neural network that processes inputs of length $l = 3$ and a multilayer perceptron (fully-connected network) that has three layers?
2. Describe two tasks that are better suited for RNN models than for MLP models and provide a short explanation.

? Something to think about

3. Two problems in Machine Translation that still remain with biRNN Seq2Seq models are a) unknown words (e.g., names) and b) forgetting/duplicating parts of the input. Think about possible solutions to these problems!

? Something to think about

4. Propose a neural network architecture that would be suited for generating image descriptions, that is, given an image as input, creates a sentence describing the content of the image! Provide reasoning why your architecture is a good choice for this task.

2 Beam Search

In the lecture, you learned that Beam Search can sometimes lead to better translations than simple Greedy Search, where the decoder always chooses the locally most likely token.

To show that this can indeed happen, look at the following situation: Given an Encoder-Decoder network trained to translate from English to German. The encoder has read

the sentence "I won't tell you nothing!" and produced an internal representation h that encodes this sentence. This representation is now fed to the decoder and a translation is generated. The decoder will return the following probabilities for output tokens:

$$P(\text{Ich}|\text{<s>}) = 0.9$$
$$P(\text{Er}|\text{<s>}) = 0.01$$
$$P(\text{verrate}|\text{<s>,Ich}) = 0.3$$
$$P(\text{verrate}|\text{<s>,Er}) = 0.05$$
$$P(\text{sage}|\text{<s>,Ich}) = 0.5$$
$$P(\text{sage}|\text{<s>,Er}) = 0.08$$
$$P(\text{euch}|...,\text{verrate}) = P(\text{euch}|...,\text{sage}) = 0.3$$
$$P(\text{dir}|...,\text{verrate}) = P(\text{dir}|...,\text{sage}) = 0.45$$
$$P(\text{nichts}|...,\text{euch}) = P(\text{nichts}|...,\text{dir}) = 0.35$$
$$P(\text{nicht}|...,\text{euch}) = P(\text{nicht}|...,\text{dir}) = 0.4$$
$$P(!|...,\text{nichts}) = 0.8$$
$$P(!|...,\text{nicht}) = 0.1$$
$$P(\text{nichts}|...,\text{nichts}) = 0.01$$
$$P(\text{nichts}|...,\text{nicht}) = 0.2$$

1. Perform a greedy search to get the output of the decoder network!
2. Perform a beam search with $B = 2$ to get the output of the decoder network!
3. Compare the two outputs.

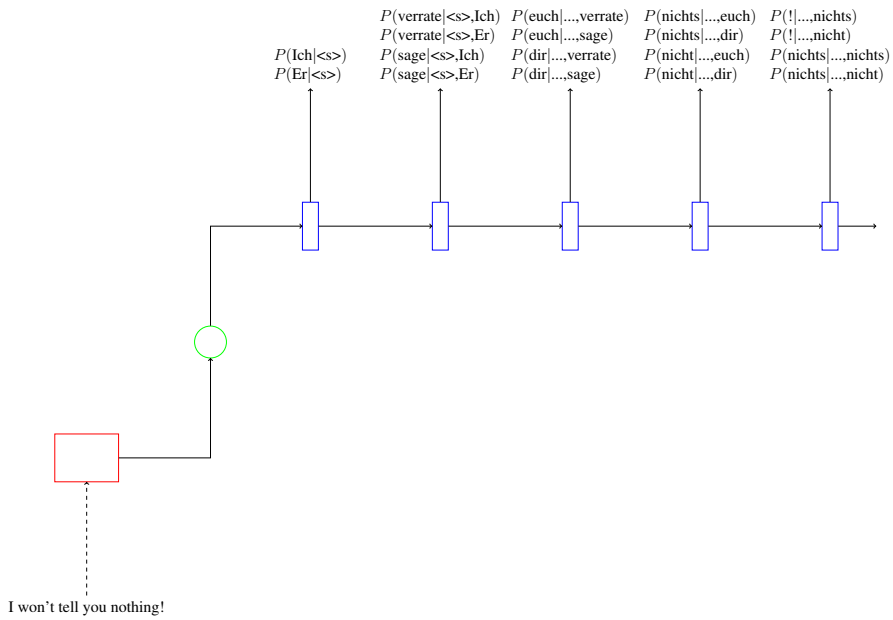

Figure 1: Visualisation of the Encoder-Decoder network used in this task