DL4NLP 2022 — Exercise 5

Jonas Belouadi, Steffen Eger
Natural Language Learning Group (NLLG),
University of Bielefeld,
Summer Semester 2022

Task 1 (10min): Pingo

To give your answers to the following multiple-choice questions please head over to `https://pingo.coactum.de/432849`.

(a) Which of the following statements about CNNs are correct?

- ☐ Filter sizes usually attend to an uneven number of words.
- ☐ Padding is used so that all sentences in one minibatch have the same length.
- ☐ A stride of 2 means that every convolution operation folds two neighboring input values.
- ☐ When applying CNNs on NLP tasks, the features learned by every filter matrix can be easily visualized and interpreted.

(b) How do CNNs differ from traditional MLPs?

- ☐ Instead of dense connections, within CNNs neurons are only sparsely connected.
- ☐ CNNs do not require an activation function.
- ☐ The computational complexity of CNNs is quadratic with regards to input length, while MLPs are linear.
- ☐ CNNs are less prone to overfitting.

(c) What statements about pooling are correct?

- ☐ In CNNs, one usually applies a pooling layer first, followed by a convolutional layer.
- ☐ To get a multi-dimensional output tensor multiple pooling operations must be applied.
- ☐ A pooling layer does not have any trainable parameters.
- ☐ The dimensions of the fixed size output tensor of a pooling layer depend on the stride and the window size of the previous convolution layer.

Task 2 (10min): Theoretical Background of CNNs

(a) Describe the difference between wide and narrow convolution in up to two sentences.

(b) Is a convolution over individual words (window size $k = 0$, i.e., one word per convolution window) useful? Explain in up to three sentences.

(c) Why is it important to use nonlinear activation functions after a convolution operation?

Task 3 (15min): Dimensions and Parameters

We use a convolutional neural network for sentence classification. Each input sentence consists of 197 words. Each word is represented by a vector from a 300-dimensional embedding space. There are 5581 unique words. The network consists of a trainable embedding layer followed by a convolutional layer, a global max-pooling layer and another fully-connected hidden layer. 111 filters, each with a window size of $k = 2$ convolve over the input. The stride is 1. The convolution is narrow. The hidden layer has 42 neurons. The output layer is a single neuron with sigmoid activation function. Using pen and paper…

(a) Compute the output shape of each layer.

(b) Compute the number of trainable parameters of each layer. Don't forget the bias for the filters and the hidden layers!