Study Attention using in LSTM2LSTM (for example for Machine Translation problem).

* Theory presentation:

Sequence-to-sequence tasks involve input sequences and output sequences of variable length. Examples of such tasks include machine translation, where the input is a sentence in one language and the output is the same sentence in another language, and question answering, where the input is a question and the output is the answer. These tasks can be solved using an LSTM2LSTM architecture.

LSTM2LSTM is a neural network architecture that consists of two stacked layers of Long Short-Term Memory (LSTM) networks, where the output of the first LSTM layer is fed as input to the second LSTM layer. The first LSTM layer is used to encode the input sequence into a fixed-length vector representation, which is then used as input to the second LSTM layer to generate the output sequence. This architecture is commonly used in sequence-to-sequence tasks, as it allows the model to better capture the long-term dependencies and nuances of the input sequence.

However, a limitation of the LSTM2LSTM architecture is that it treats all parts of the input sequence equally, without considering which parts of the input are more important for generating each part of the output sequence. This can lead to suboptimal performance, particularly for longer input sequences.

To address this limitation, attention mechanisms can be used in LSTM2LSTM. Attention allows the model to focus on the parts of the input sequence that are most relevant for generating each part of the output sequence. This is done by computing a weight for each part of the input sequence, based on how relevant it is for generating the current output sequence. These weights are then used to compute a weighted average of the input sequence, which is used as the input to the second LSTM layer.

There are several types of attention mechanisms that can be used in LSTM2LSTM. One common type is additive attention, which involves computing a score for each part of the input sequence based on a learned parameter matrix, and then computing a softmax over these scores to obtain the weights. Another type is multiplicative attention, which involves computing a score for each part of the input sequence based on a learned parameter matrix, and then computing a dot product between these scores and the hidden state of the second LSTM layer to obtain the weights.

In summary, attention mechanisms can be used in LSTM2LSTM to improve the performance of sequence-to-sequence tasks, by allowing the model to focus on the parts of the input sequence that are most relevant for generating each part of the output sequence. Different types of attention mechanisms can be used, such as additive attention and multiplicative attention, depending on the specific requirements of the task.