Deep-Learning-TAU-0510-7255-Spring-2020

Homework 2

Ophir Shurany 304867716

Tim Mironov 321297111

**Theory: Question 1**

1. To handle variable length **input** sequences, we can do the following:
   1. **Padding**: Padding can be used, whereby you would have to fix the length of each sample (either to the length of the longest sample, or to a fixed length — longer samples would be trimmed or filtered somehow to fit into that length)
   2. **Tokenize** **(encoder**): create a state vector of the input by process the inputs one by one by the encoder RNN, while disregarding the output. The RNN should stop by an agreed EOS sign.
   3. **Length Information**: set the *sequence\_length* argument when calling the *dynamic\_rnn()* (or *static\_rnn()*) function; it must be a 1D tensor indicating the length of the input sequence for each instance
2. To handle variable length **output** sequences, we can do the following:
   1. **Tokenize** **(decoder)**: Define a special output called an end-of-sequence token (EOS token). Any output past the EOS should be ignored
   2. **Length Information**: If you know in advance what length each sequence will have (for example if you know that it will be the same length as the input sequence), then you can set the *sequence\_length* parameter as described above (a.II)
   3. **Last Relevant Output**: For sequence classification, we want to feed the last output of the recurrent network into a predictor, e.g. a softmax layer. While taking the last frame worked well for fixed-sized sequences, we do not have to select the last relevant frame.

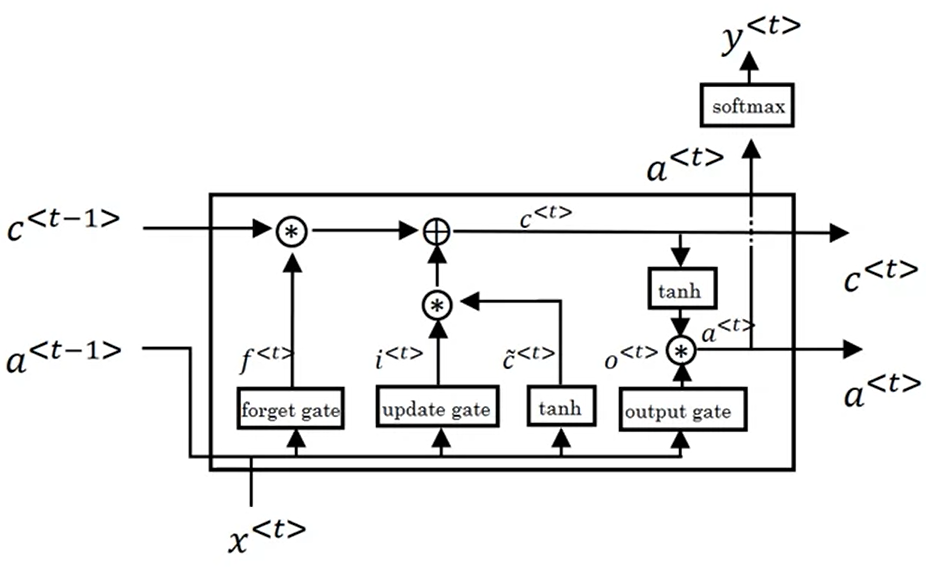
**Theory: Question 2**

the advantages of the GRU:

1. **Architecture**: GRU is a simpler model and so it is easier to build a much bigger network
2. **Computations**: GRU has only has two gates (LSTM has three gates), so computationally, it runs a bit faster than LSTM

**Theory: Question 3**

LSTM structure:



LSTM equations:

Where is a compressed notation =

Total parameters calculating for 3 gates and 1 cell:

**Theory: Question 4**

The GRU equations are defined as follows:

Given is a two iteration GRU network with a defined loss function:

The gradient is given. The BP gradients of the second time stamp are calculated as follows: