Machine Learning Summer Training

Homework 3 Text Generating

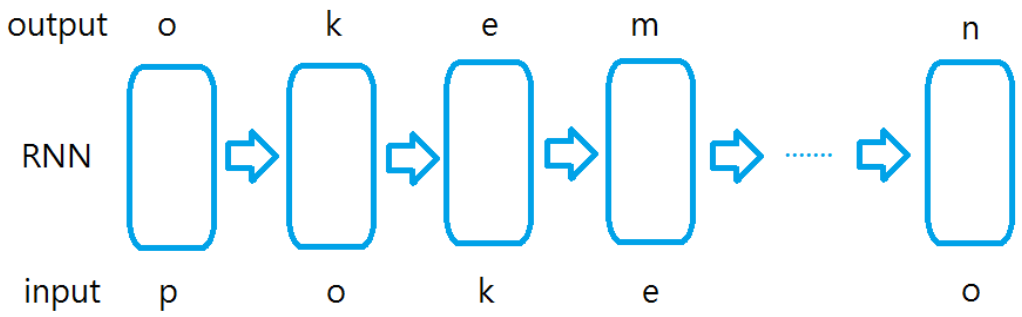
TA: Chih-Wei Lee

1. Goal

In this homework, we are going to teach machine to learn how to generate meaningful sentences by RNN or LSTM. The machine will learn the probability distribution of the next character given a fixed length sentence, and then generate the whole sentence recurrently.

2. Recurrent neuron network

For example, if we only have six characters "pokemo" and we want machine to learn to generate the word "pokemon", we can give the RNN(or LSTM) an input 'p' and set its training target as 'o', give NN the input 'o' and set the training target as 'k', and so on… Because of RNNs' property, it can somehow realize the sequential information in each time step, so RNNs is useful in text generating. The brief concept is shown below:



But in this homework, I recommend you to let the input be fixed length, for example: In "Happy birthday to you", input can be "Happy birthday to yo" and the training target is then 'u'. How many characters should it be in the input depends on you, you can try several number and observe the outcome.

3. Hint

For encoding the characters, one-hot encoding is commonly used. For example, in English, 'c' should be encoded to [0,0,1,0,….,0], 'f' should be encoded to [0,0,0,0,0,1,….,0], and so on… but if your corpus has other symbols, such as '\n' or some punctuation marks, they should be encoded as well.

4. TODO

First, you can try to generate **English sentence**. After you know how RNN works, you can then try to generate **Chinese sentence**. Note that in Chinese text generating, encoding is a problem. Chinese characters should be decoded using **utf-8**.

In early epoch, machine generates the sentences which are meaningless, but as training epoch increases, it will be more and more meaningful. You can observe the difference between English and Chinese and try to explain why they're different.

There's no baseline in this homework, but you should train your model until it generates meaningful sentences in both English and Chinese

5. Notice

Notice that the sample code still has some simple **TODOs** for you to fill in. After realizing the whole concept, you can build your own neuron network and try different architecture, enjoy it!