MSc in Data Science

Deep Learning

Academic Year: 2018-2019

Exercise 2: Create a Chatbox

Delivery Date: **07/01/2018**

Description

In this assignment you will build a “chatbox”. A “chatbox” is a conversational model, which in its simplest form predicts the next sentence given the previous sentence or sentences in a conversation.

You are provided with a dataset of movie dialogs: The Movie dialog corpus from Cornell University: https://www.kaggle.com/Cornell-University/movie-dialog-corpus#README.txt

The corpus contains:

- 220,579 conversational exchanges between 10,292 pairs of movie characters
- involves 9,035 characters from 617 movies
- in total 304,713 utterances

The utterances are contained in file “movie_lines.txt”.

The corpus can be downloaded from:
https://s3.amazonaws.com/pytorch-tutorial-assets/cornell_movie_dialogs_corpus.zip (original corpus)
https://www.kaggle.com/Cornell-University/movie-dialog-corpus/downloads/movie-dialog-corpus.zip/1 (pre-processed)

The goals of this exercise are:

1. Extract pairs of utterances from the provided dataset.
2. Select a neural network architecture that can adequately model this problem
3. Decide on the choice of activation and loss functions
4. Train you model, on pairs of utterances ([input, output])
5. Do some “discussions” with your chatbot (like the ones shown here: https://github.com/chiphuyen/stanford-tensorflow-tutorials/blob/master/assignments/chatbot/output_convo.txt)

There is no easy way for evaluating these kind of systems, so no evaluation with standard metrics is expected. You are free to use any code/tutorial you can find (there are plenty), and you can use any deep learning framework you want.

Deliverable

Write a 2-page report2 describing your approach and findings. The report should justify your choices when designing your neural network. It should also include same sample utterances of the learned model. You must also submit your source code, commented and structured as needed (Jupiter notebooks in python are preferred). Make sure you write your name both on the report and in the source code. Bundle and submit your report in PDF along with your source code, in a single zip or tar.gz file.

Hints

- Can the problem be seen as a "translation" problem?
    - https://github.com/tensorflow/nmt
    - https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html