Machine Learning Engineer Nanodegree

Capstone Proposal

Sora
(kotori4c@gmail.com)
Sep 22th, 2017

Domain Background

Deep learning is one of the most promising fields in machine learning where already much development has taken place and is currently used in real world application.TV and movies are popular among the young, but it is not easy to generate a good TV script and it takes a lot of time. In order to generate TV scripts more efficiently, machine learning helps.
My personal motivation is that we can watch more interesting TV or movies if TV scripts could be generated with machine learning.

Problem Statement

Recently I have watched Simpsons and want to watch more about it, however, there are no more stories.
The problem is to generate a new TV script for a scene at Moe's Tavern. As input, we are provided the Simpsons dataset of scripts from 27 seasons. Obviously, we can handle the problem with deep learning model.

Datasets and Inputs

The dataset contains the Simpsons dataset of scripts from 27 seasons.
Path: data/ simpsons /moes_tavern_lines.txt (4520 lines)

Solution Statement

We want to generate TV scripts from the original scripts. We have to find the relationship between words. We can train the network on the provided scripts, and then use it to generate an original piece of writing with a RNN model.

Benchmark Model

To build to RNN model, we should pick up some of the provided scripts instead of all of them to make sure that the script content will be generated as expected. In details, we use a subset of the original dataset including only the scenes in Moe's Tavern. This doesn't include other versions of the tavern, like "Moe's Cavern", "Flaming Moe's", "Uncle Moe's Family Feed-Bag", etc. And then we will preprocess the data, build RNN cell, build word embedding, build a RNN model step by step. Finally we will train the neural network on the preprocessed data to generate the scripts.

Evaluation Metrics

We can read the generated scripts to see whether it follows the grammar and whether it is related to the original scripts by checking the character names.

Project Design

We will pick the sub-dataset and do data preprocessing at first. And then train a RNN model using the testing dataset, and try to make the training loss as low as possible by adjusting several hyperparameters. When the training model is reliable enough, we can begin to write our own TV scripts.