Large-scale neural electroencephalogram (EEG) activity analysis on big data platfrom

Final Project proposal

Ye Li

Introduction

The development of cutting-edge recording technology in neuroscience research increase the data size and complexity. For example, the new electroencephalogram (EEG) experiment can simultaneously record several neural activities from multiple brain regions, which induce huge data size data yield from one single experiment participants.

The problem in huge neural data analysis

- For large scale neural data, such as EEG or two-photon imaging data, even a simple calculation can spend hours to process on the single lab workstation.

- Since the complexity of neuron interaction in the brain, multiple differentiate analysis methods should be used in uncover the meaning of neural data. In this case, regular on-site workstation platform can not timely process multiple analysis when handling the big scale data.

Solution

- Apache Spark platform provides a potential solution to address this large-scale neural data problem. Using the resilient distributed data set (RDD) system, Spark cache a data set into memory across cluster node, which perform a fast computations and a possibility of interactive analysis. Also, the API of Spark can manipulate subsets of data, such as combine multiple data set.

- Thanks to the Spark platform, an open-source library analytical tools, name Thunder, built on the Spark platform is available to use. This tool contents multiple algorithms and featured by the capability of extendable library

Solution

Dataset:

- In this final project, I will use an EEG dataset from the OpenNeuro database. OpenNeuro database is an open-sharing database for neuroimaging research established from 2013. The datasets in OpenNeuro database can be accessed by public.
- This datasets include 17 experiment participants. The EEG data is recorded during the participants continuously play a video game.
- The datasets are storage in a specific format used in EEG recording software. Also, the datasets mixed with csv file for storing experiment setting information.
- These EGG data's link is here: https://openneuro.org/datasets/ds003517/versions/1.1.0.

Solution

Preliminary plan:

1. The EEG data are recorded into a EEG specific format used in EEG recording device, I may need to use the EEG recording software to convert the data into more user friendly format.

2. The second task is figure out how many brain region this EEG recorded simultaneously and how many different channel this EEG recorder used.

3. The third step is data manipulation. I want to convert the data into a time series form vs different neural channels of different brain region. In this way, the data can be represented as a key-value pair. The different time series can be used as a key, and paired signal reading can be used a value.

4. Next, I will try to convert the data into an RDD in Spark and process analysis for these EEG data.