This is only in LUA. There is no documentation about it.

**Preprocess:**

we use data from deepsea dataset. We really appreciate that Jian Zhou, the author of Deepsea, replied to us about how to understand the dataset and motifs [1]. The data he used is from Encoder project for all the cell types [2]. We are still working on the tutorial he provided to see how he generated this dataset based on coordination for different cell types. Dataset is in .mat format. We need to read it into numpy then transpose and batch them to test our model.

Train data has shape: 4400000, 4, 1000

Train labels have shape: 4400000, 919

We batched those files for fast training sample, therefore we have 440 batched train data and labels. We name it reshaped\_data\_{start\_index}\_{end\_index}.

**Main Training:**

Our training now focuses on using only 1 batch from reshaped data. We print out our loss to see if our model is in the right way.

**Model:**

Our model is CNN\_beseline for only implementing the baseline function.

**Problems:**

We still have a hard time understanding our dataset. It makes it hard for us to seperate or add more cell types. In tutorial, he provided how they dealt with one cell type, we still have a problem understanding it. [3]

We still cannot run Lua environment, therefore, we switch to Selene model which is a Python library that helps train biological sequence data. When running its tutorial, there are many times we got different shape from the expected result. This makes us hard to understand or get this model running.

**References**

[1] <https://github.com/FunctionLab/selene>

[2] <http://deepsea.princeton.edu/help/>

[3] <https://github.com/FunctionLab/selene/tree/master/manuscript/case1/data>