Final Project Proposal

107061212 劉亦傑、107061218 謝霖泳
107061234 張博閔、107062217 鍾凱恩

1. Title

Using Random Forest, LSTM, Transformer and CatBoost to Perform Behavior Classification of Exposition Visitors

2. Methods

● Data Preprocessing

Intuitively, the route is only related to the visiting order. Therefore, we only consider the order of locations visited for a certain person (mac_hash) and ignore the absolute visit time. However, the number of locations that a person visited is not identical in the dataset, which varies from 1 to 14. Therefore, the preprocess phase of the dataset is needed. We will attempt the following two methods.

1. With grouping method: Distribute the people (mac_hash) into several groups by the number of locations that he/she visited. For example, group the people who visited 1~5 locations together, and people who visited 6~10 locations for another group. To maintain the length of samples in each group, we will perform padding up to the limits of the group. For instance, for the group with people who visited 1~5 locations, we will perform padding to all the samples in this group to 5 locations. For each group, we will train a model and the corresponding model will be applied in the inference stage according to the length of the test sample.

2. Without grouping method: Treat all the samples as a same group, and padding all the samples to 14 locations is needed.

● Classification Methods

After observing the dataset, we found that the data has sequential properties. So, directly applying a traditional CNN model with 2D convolution may not be so helpful for this task since 2D convolution generally focuses on spatial relation, which usually captures the neighboring features when training.

Therefore, we surveyed some papers and proposed 4 different model architectures corresponding to this problem, including **Transformer**, **LSTM**, **Random Forest** and **CatBoost**. Transformer is a seq2seq architecture with self-attention mechanism, which could be useful to fit our dataset. LSTM is a classical architecture for prediction on sequential data, and it can greatly handle the sequential data in the history. Also, since it’s a multi-class classification problem, we will try the tree-based method with ensemble learning, which is random forest. Finally, we also want to use the CatBoost package to build our model, which is a gradient boosting on decision trees.

We will compare the result by training the above four kinds of models with two different data preprocessing methods after our experiments. Refer to our complete system design flow chart at the following page (Fig. 1).

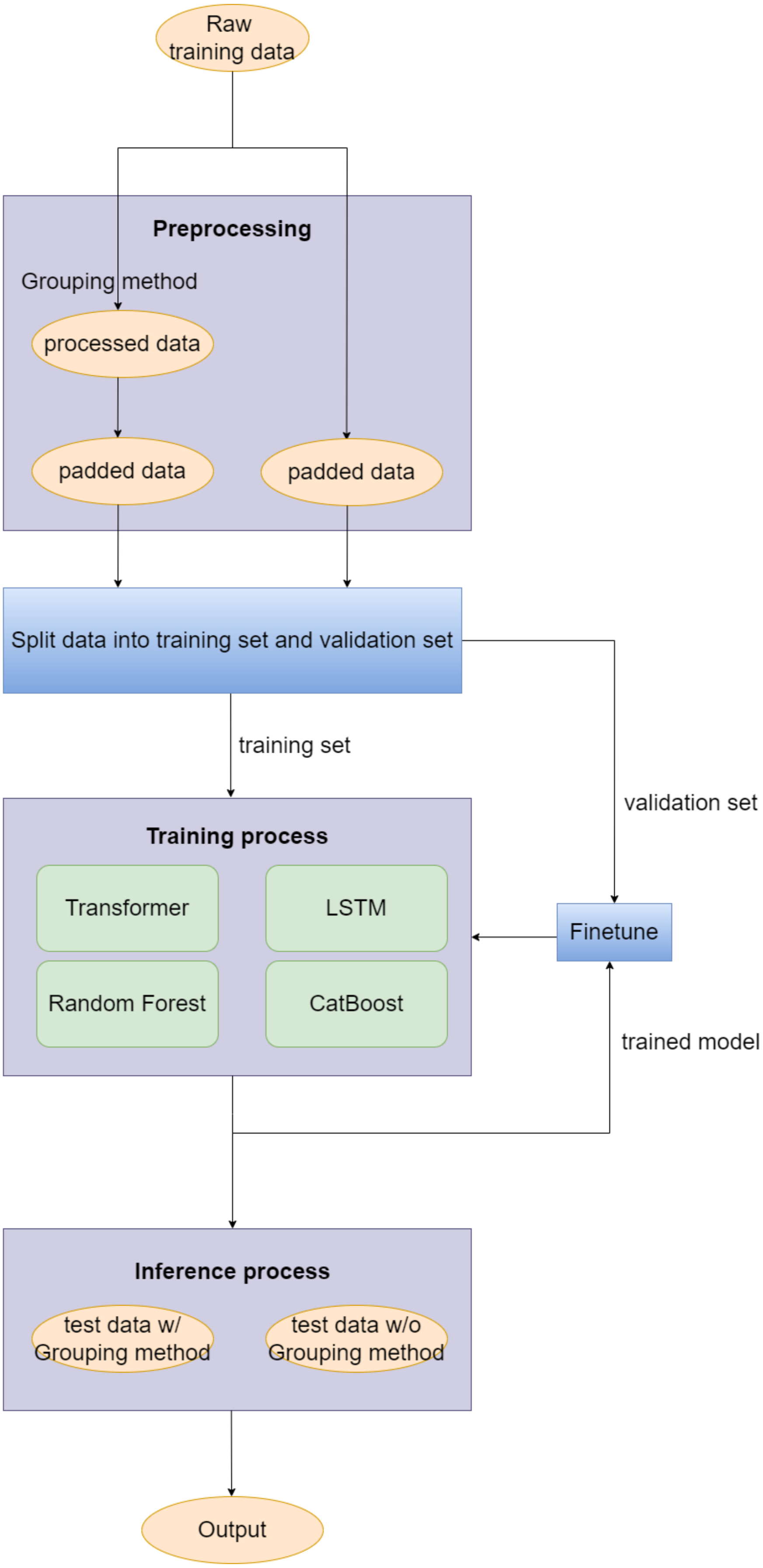

Figure 1. Flow Chart

3. Reference

[1] Leo Breiman. Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[2] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. CatBoost: unbiased boosting with categorical features, 2017.

[3] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-term Memory. In *Neural Computation*, 1997.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. In *Neural Information Processing Systems (NIPS)*, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Google Research*, 2018.