Homework Report

Ting-Wei Lu（盧庭偉）

Problem Description:

Given the browsing history of the last few months, predict whether the user will use the app in different time slot of the next two weeks.

Model:

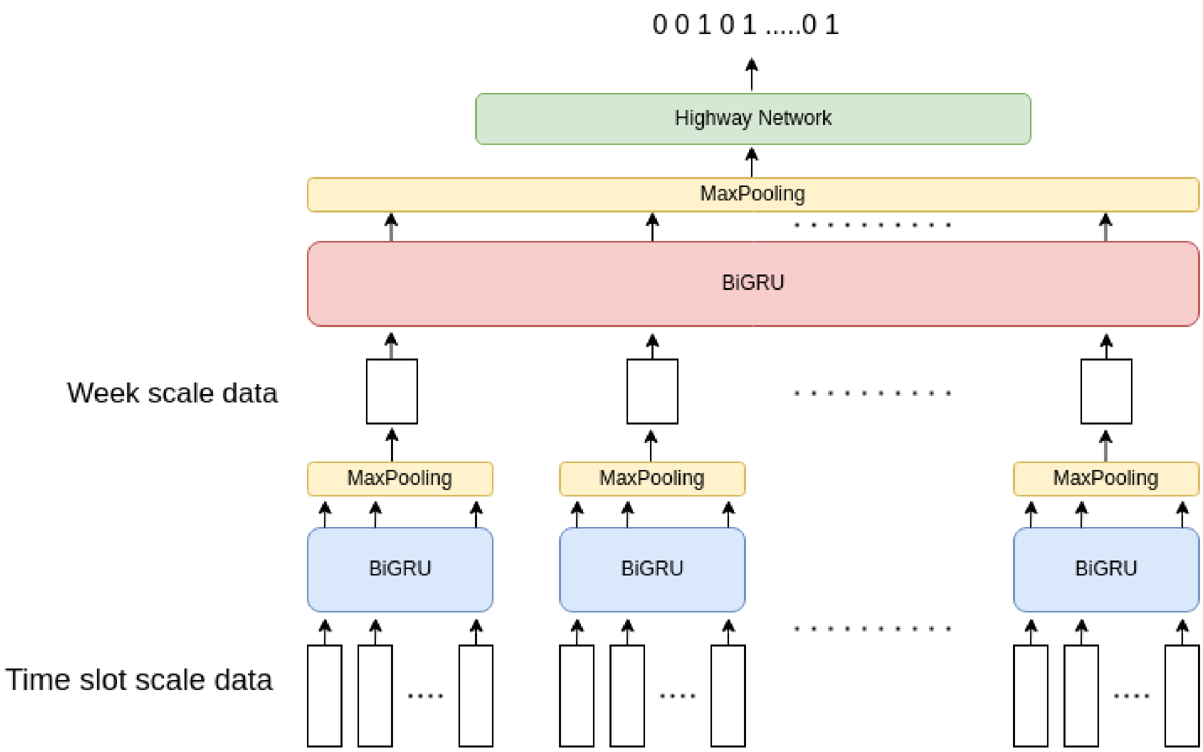

Figure 1: Model

Data:

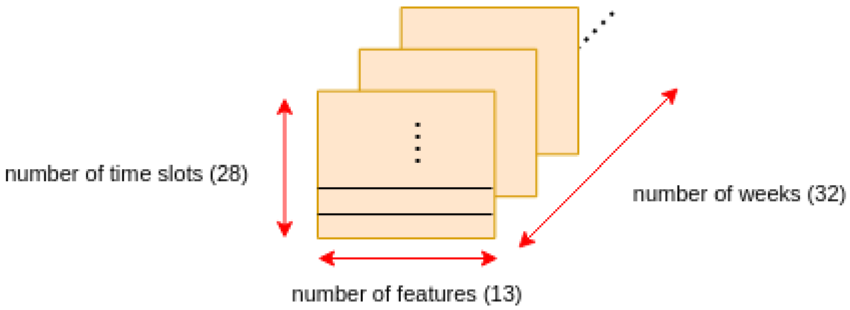

Figure 2: Data

(Detail of our model and data will be discussed in the next part)

Selected Feature:

1. Platform:
   One hot encoding for three different platform

2. Connection type:
   One hot encoding for different connection types which are categorized as wifi, cellular, offline or others

3. Watched ratio:

$$\frac{episode\_number}{series\_total\_episodes\_count}$$

4. Total number of episodes:

$$log(1 + \frac{series\_total\_episodes\_count}{210})$$

   where 210 *is the maximum count of total episodes*

5. Limit playzone countdown:
   If action_trigger is ‘limit playzone countdown' -> 1
   otherwise -> 0

6. Error:
   If action_trigger is 'error' -> 1
   otheriwise -> 0

7. Video ended:
   If action_trigger is either ‘video ended’, ‘program stopped or enlarged-reduced’ or program stopped -> 1
   otherwise -> 0

8. Played duration:

$$log(1 + \frac{T}{9550675})$$

   where T is the total watch time of a time slot, 9550675 *is the maximum total watch time of a time slot*

9. Popular titile:

$$log(1 + \frac{T}{5224})$$

   where T is the total watch time of a time slot that watches the top k most viewed title, 5224 *is the maximum total watch time of a time slot*

Challenges:

A. **Incomplete time series data:**

We categorize each set of data of a week into 28 tim slots and calculate the average values of each feature of that time slot (except for played duration which we sums the total time), which gives us 32 weeks of data, each includes 28 time slot data(as described in figure 2). After that, we apply RNN on the 28 time slot and the 32 weeks.

Since the data is composed of two levels of ordered data, it is quite similar to character based language modeling / character word embedding problem. Previous works (Kim et al., 2016) and (Kim et al., 2018) on relative field show that utilizing highway network can gives us a better result, so we also implement a highway network after the RNN output.

B. **Imbalanced label** (90% labeled 0, 10% labeled 1):

(a) Solution 1: Focal Loss (Lin et al., 2017):

Focal is a modified version of cross entropy loss proposed by Facebook in 2017 aimed to address imbalanced data problem, , which is define as:

$$\mathrm{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t)$$

where gamma decides how much emphasis we put on rare class, which is suggested by the author to set to 2.

(b) Solution 2: F1 Loss:

A common criterion for imbalanced data class classification isf1 score, which is define as:

$$F_1 = \left(\frac{recall^{-1} + precision^{-1}}{2}\right)^{-1}$$

So intuitively, we can try to maximize f1 score in order to have a better result on the classification of the rare class. There is just one problem, it' is non-differentiable. Fortunately, we can make it differentiable with a slight modification. Instead of ground truth, we now use probability to calculate recall and precision instead, i.e

$$Truth\ Positive\ =\ y\_true\ \times\ y\_pred$$
$$Truth\ Negative\ =\ (1 - y\_true)\ \times\ (1 - y\_pred)$$
$$False\ Positive\ =\ (1 - y\_true)\ \times\ y\_pred$$
$$False\ Negative\ =\ y\_true\ \times\ (1 - y\_pred)$$

then our objective function became minimizing $1\ -\ F1\ Score$

Validation Score	AUROC	Discrete-AUROC	F1	Accuracy
BCE Loss	0.8536	0.5011	0.4622	**0.9113**
Focal Loss	**0.8881**	0.5844	0.5235	0.8984
BCE Loss + F1 Loss	0.8001	**0.7882**	**0.6811**	0.8451

AUROC: Predict Probability, Discrete-AUROC: Predict 1 or 0 (with 0.5 as threshold)

Conclusion:

Since my research topic of undergraduate research program(there are two of them, actually) is deep learning for NLP, my work here focuses mainly on the model construction and training optimization. In a nutshell, I leveraged some knowledge in NLP to build a reasonable to solve time series problem and model, and utilized some special loss function to deal with imbalancing problem, which might be what differentiate this work from others.

Feature work on this topic can be refine the feature engineering of the preprocessing stage, we can apply some stats-based methods to evaluate the relativity of a certain feature and the outcome.

Reference:

[1] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In Proceedings of the Association for the Advancement of Artificial Intelligence, pages 2741–2749.

[2] Kim, Y., Kim, K.-M., Lee, J.-M., Lee, S. 2018. Learning to Generate Word Representations using Subword Information Proceedings of the 27th International Conference on Computational Linguistics, pages 2551–2561

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017.