

Project 3:

Run the Transformer Model

I. Source Code Download

After downloaded from bb and unzipped, The directory will be like the following:

名称	修改日期	类型	大小
📁 _pycache_	2023/11/28 19:08	文件夹	
📁 models	2023/2/23 10:57	文件夹	
📁 THUCNews	2023/2/23 10:57	文件夹	
🐍 predict.py	2022/2/20 14:27	Python 源文件	2 kB
🐍 run.py	2023/11/28 18:21	Python 源文件	4 kB
🐍 train.py	2023/11/28 19:12	Python 源文件	6 kB
🐍 utils.py	2022/2/20 14:05	Python 源文件	7 kB

II. The Transformer Model

1. Training Task

The task in this project is to classify sentences of news in Chinese into 10 categories, namely 财经, 房产, 股票, 教育, 科技, 社会, 时政, 体育, 游戏, and 娱乐. Each of the categories is represented by an integer from 0 to 9 in the given order.

2. Dataset

The transformer model will be trained, evaluated, and tested on a truncated version of the THUCNews dataset with 50000 sentences of news and their categories, where 90% are in the training set and 5% are in the evaluation set (also known as the validation or development set) and test set respectively. To ensure the validity of evaluation and test results, these three sets should not have any shared content.

3. Running the Model

To run the transformer model, simply run the run.py file in the above directory. The program will automatically train the model for 6 epochs and evaluate its performance (loss and accuracy) on the evaluation set at the end of each epoch. The evaluation can provide a reference for when to terminate training. After the training terminates, the model will run on the test set for a final evaluation. There are also some classification cases at the end of the program to provide an intuitive perception of the model's performance. You should include a screenshot of the whole running outputs and the model's loss and total accuracy on the test set in your report.