REPORT FOR DATA CHALLENGE 2

Student Name: Jinhe Zhang
Student Number: 21068423
Kaggle Name: j423zhan
Leaderboard Score: 0.70588

1. STRUCTURE

```
├── info.txt            # a txt file contains student number,
kaggle name and leader board score
├── src
│   ├── data            # data folder
│       ├── test.csv    # test data
│       └── train.csv   # train data
├── utils
│   ├── load_data.py    # load ground truth lables from
`train.csv`
│   ├── logger.py       # customized logger
│   └── dataset.py      # customized dataset
├── requirements.txt    # required packages
└── readme.md           # current file
```

2. SETUP

The code is based on:

- Python: 3.10
- CUDA: 12.1
- GPU: 4 `*` Nvidia RTX-4090
- PyTorch: 2.2.0
- Torch vision: 0.17.0
- Other packages can be found in requirements.txt
- Torch seed is mannually set to 42
- Pre-Trained Model: `XLNet-Large`

3. HOW TO RUN

```
python -m torch.distributed.launch --nproc_per_node=4 --nnodes=1
train_xlnet.py --batch-size 8 --epochs 20 --lr 0.000002 --val
```

4. BRIEF INTRODUCTION OF THE CODE

The source code can be divided to such few steps:

1. Load model

    1. The model used in this challenge is `XLNet-Large-Cased` on hugging face, only the body, the classification head is initialized with random weights.

2. Load dataset

3. Train

    1. Load inputs data and labels to gpu

    2. Compute the outputs

    3. Compute loss

    4. Backwords

4. Tricks:

    1. ADAM optimizer

    2. Linear learning rate scheduler