sentimental analysis of IMDB 50k dataset with GPU optimization

Zhuopu Wang zw3743 Deze Zhu dz2372

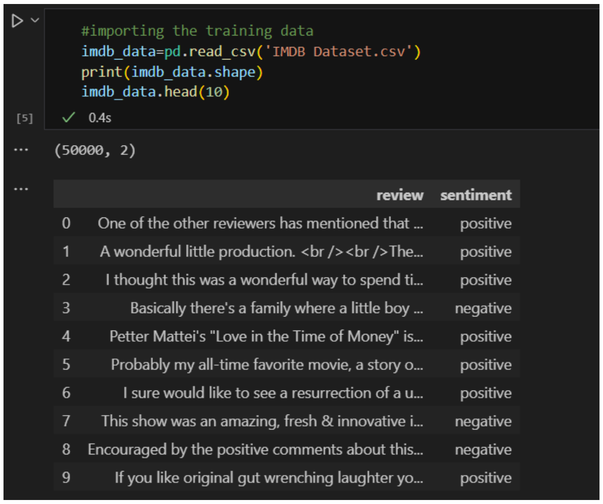
Project Milestones

- ☑ **Literature Review and Model Selection** :Review relevant literature on sentiment analysis and LSTM models.
- ☑ **Dataset Collection and Preprocessing** :Acquire and preprocess the IMDB movie review dataset.
- ☑ **Model selection and Training** :Choose the proper models for the task, and do the training process with the IMDB dataset
- ☑ **Model Accuracy Optimization** :Pick up the best model that optimal the final result
- □ GPU-accelerated Training Optimization
- □ Inference Speed Optimization
- □ Evaluation and Analysis
- □ Report Writing

Current Achievement

We have gone over some papers about sentiment analysis with different machine learning techniques. Some most mentioned models include: LSTM, BagOfWords, Transformers and so on.

The data we used in the project is from IMDB dataset, which is a dataset of people's reviews on movies. We will use these data to analysis people's attitude towards the movie. They are labeled with sentiment and well balanced. After cleaning, we can see there is no missing rows or data.

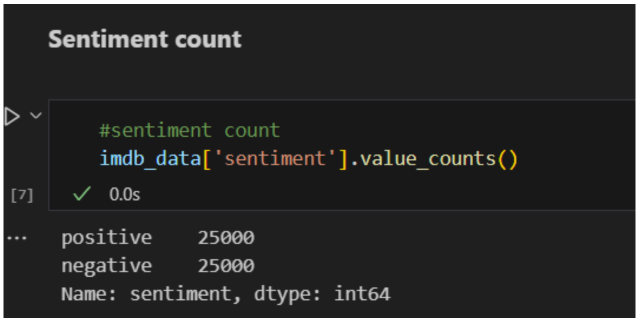


```
#importing the training data
imdb_data=pd.read_csv('IMDB Dataset.csv')
print(imdb_data.shape)
imdb_data.head(10)
```

[5] ✓ 0.4s

```
(50000, 2)
```

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive



Sentiment count

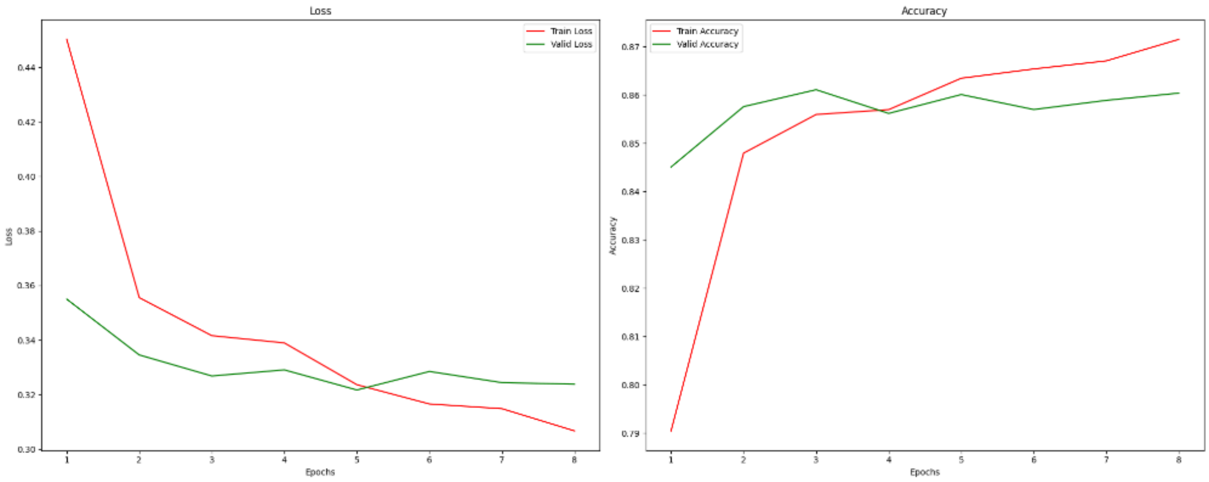
```
#sentiment count
imdb_data['sentiment'].value_counts()
```

[7] ✓ 0.0s

```
positive    25000
negative    25000
Name: sentiment, dtype: int64
```

Also, there are different cleaning methods we could choose. We can keep all the letters while ignoring the others for simple understanding cause symbols may be very hard to learn with. Also some stopwords like “the” “is”, should not be used by AI because they are grammar words, we also need to delete them from the reviews.

After properly choosing the hyper parameters, a relatively good model with LSTM could reach a 80+ accuracy after 8 epochs of tranning.



Remaining Bottlenecks

For the remaining work, there are mainly the following potential difficulties:

- Deploying the model to HPC environments and debugging and completing the training process in the distributed environment.
- In data parallelism, each GPU needs to frequently exchange gradients to keep the model synchronized. For the LSTM model, the gradient tensors can be quite large, which may cause communication to become a bottleneck.
- The data is unevenly distributed across GPUs, causing some GPUs to be idle and waiting.
- Balance the relationship between weight pruning ratio and model accuracy loss.
- Designing the distillation objectives in Knowledge Distillation.

Work Distribution

Zhuopu Wang

- Parameter tuning for non-distributed models(original models)
- GPU acceleration based on data parallelism
- Optimization of GPU acceleration-related issues
- Inference speed of optimization based on model pruning
- Analysis and evaluation of model
- Writing related reports

Deze Zhu

- Preliminary literature review and model selection
- Data collection and cleaning
- Building non-distributed model
- Model building in the HPC environment
- Inference optimization based on knowledge distillation
- Writing related report