Enhancing the Dependency Mechanism of RoBERTa

指導教授 ： 張炎清 教授

專題組員 ： 李昱廷、郭為軒、曹仲辰、吳岳霖、林裕峰

Abstract

Our proposed model, EDM-RoBERTa, uses the Boom Layer to improve the multi-headed attention mechanism in the encoder of Transformer model. Compared with the original Transformer and SHA-RNN models, the reorganization of Boom Layer and RoBERTa can meet both long and short input text sequences, and keep the original long-term dependency of Transformer. In the calculation process, it can also reduce the amount of calculation, thereby improving the accuracy and performance on text classification.

The research source are Transformer-based models (BERT, RoBERTa, XLNet, DistilBERT), SHA-RNN, and the self-attention mechanism in the Transformer as the main structure. The Boom Layer of SHA-RNN is transformed to realize the attention mechanism for high-dimensional vector convertion and then improve the multi-headed attention mechanism in the encoder of Transformer.

Result

➢ Fine-tuning Transformer-based Models with IMDb Dataset

Fine-tuning Transformer-based Models with IMDb Dataset

	Epoch	Accuracy	train loss	valid loss	error rates
BERTLARGE	6	92.6	0.35	0.55	0.29
RoBERTaLARGE	6	93.17	0.22	0.53	0.26
XLNet	6	89.53	0.28	0.69	0.37
DistilBERT	6	86.48	0.32	0.74	0.35
EDM-RoBERTa	6	94.76	0.27	0.49	0.2

➢ Fine-tuning Transformer-based Models with Rotten Tomatoes Dataset

Fine-tuning Transformer-based Models with Rotten Tomatoes Dataset

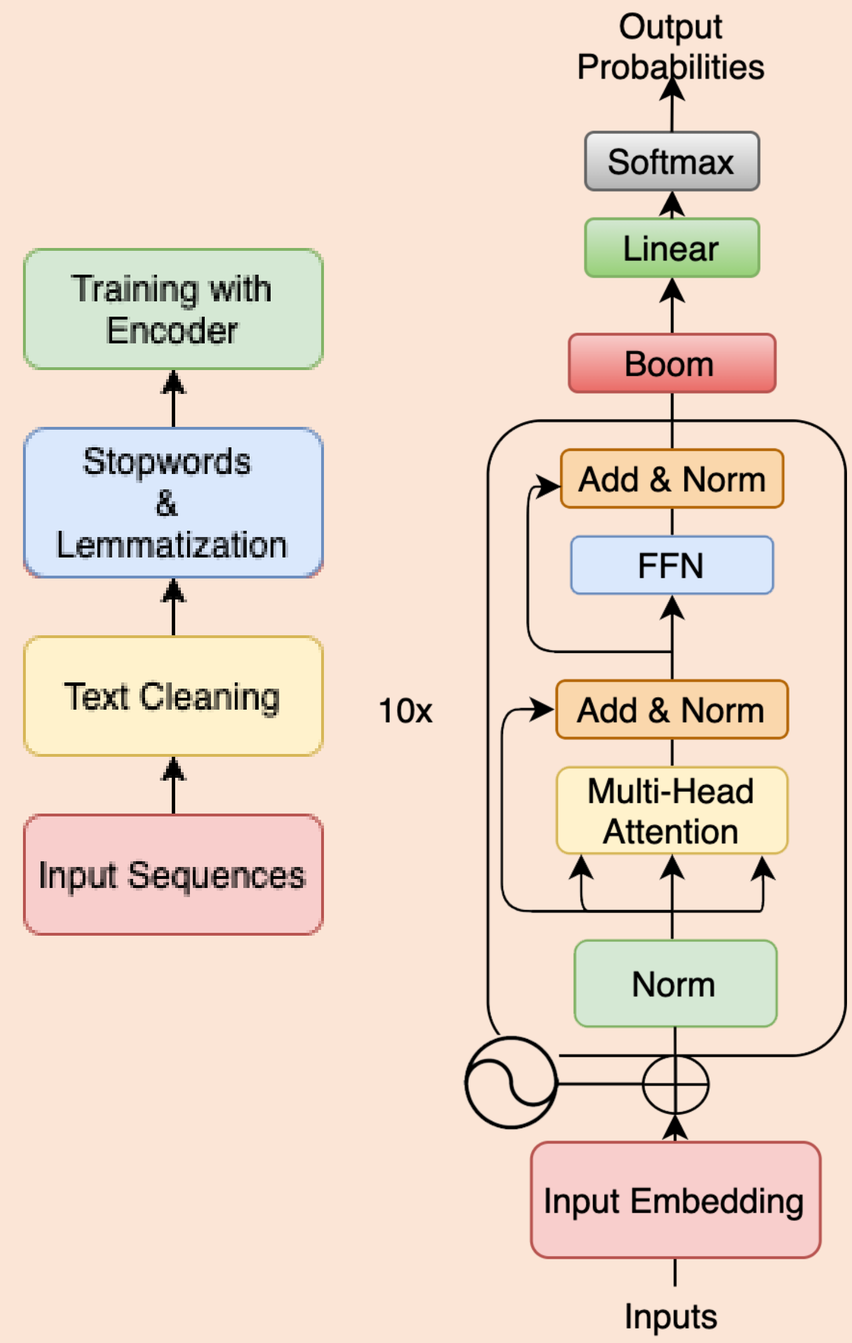
	Epoch	Accuracy	train loss	valid loss	error rates
BERTLARGE	5	66.21	0.64	0.68	0.3
RoBERTaLARGE	5	68.91	0.67	0.7	0.29
XLNet	5	62.83	0.73	0.79	0.38
DistilBERT	5	54.65	0.8	0.77	0.44
EDM-RoBERTa	5	76.18	0.64	0.62	0.26

- EDM-RoBERTa
  (Enhance the Dependency Mechanism of RoBERTa)

EDM-RoBERTa (Enhance the Dependency Mechanism of RoBERTa)

bsz	steps	lr	ppl	SST-2	SST-5
256	1M	1.00E-05	3.83	92.6	74.57
2K	**125K**	**2.00E-04**	**3.61**	**94.76**	**76.18**
8K	31K	1.00E-03	3.72	92.1	74.31

Architecture



Conclusions

The results shows that our EDM-RoBERTa model obtains more accurate prediction results than the original Transformer model on word meaning analysis output, and thereby improving the short-dependency in the encoding process.

The results will also be applied to other natural language sentiment analysis tasks such as sentiment analysis and social network analysis.