NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Transformers Acceleration

For AutoNLP application in Document classification

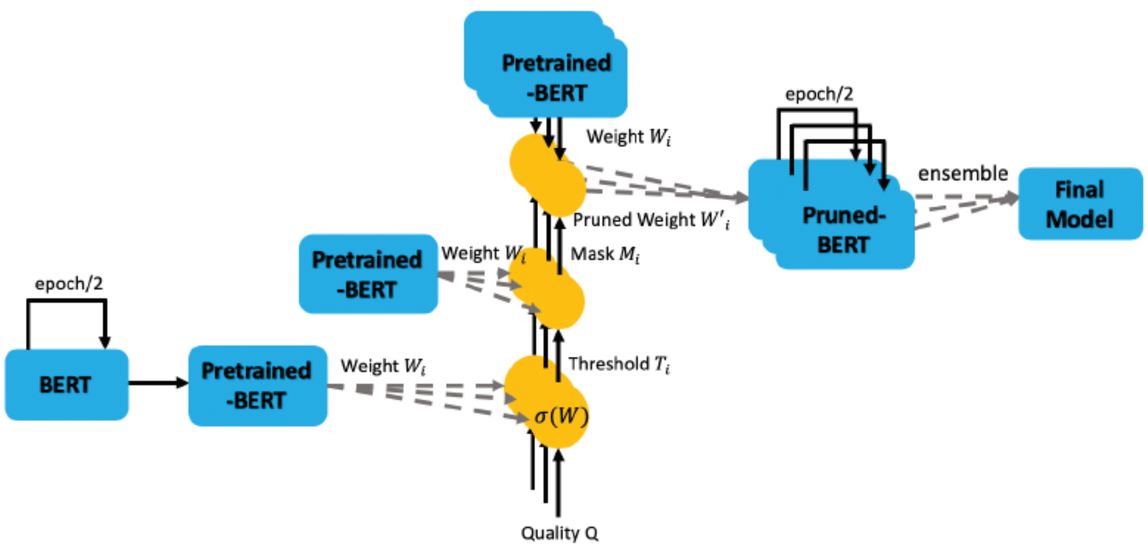
Student: Cao Hannan **Supervisor:** Assoc Prof Sinno Pan

1. Objective

This project aims to develop an acceleration method for Transformers in order to increase the model performance, reduce the model's training time and its weight. The acceleration method is based on weight pruning and its applied to the state-of-the-art model, BERT. Ensemble method has also been applied to the final models
An AutoNLP platform is built making use of this method, which can accept user's specification towards training process and return the prediction result and the model weight back to the user using our acceleration method.

2. Approach

The new method break the training process into two parts, first the connectivity is trained, then weak connections is removed. Finally the network is retrained and ensembled.

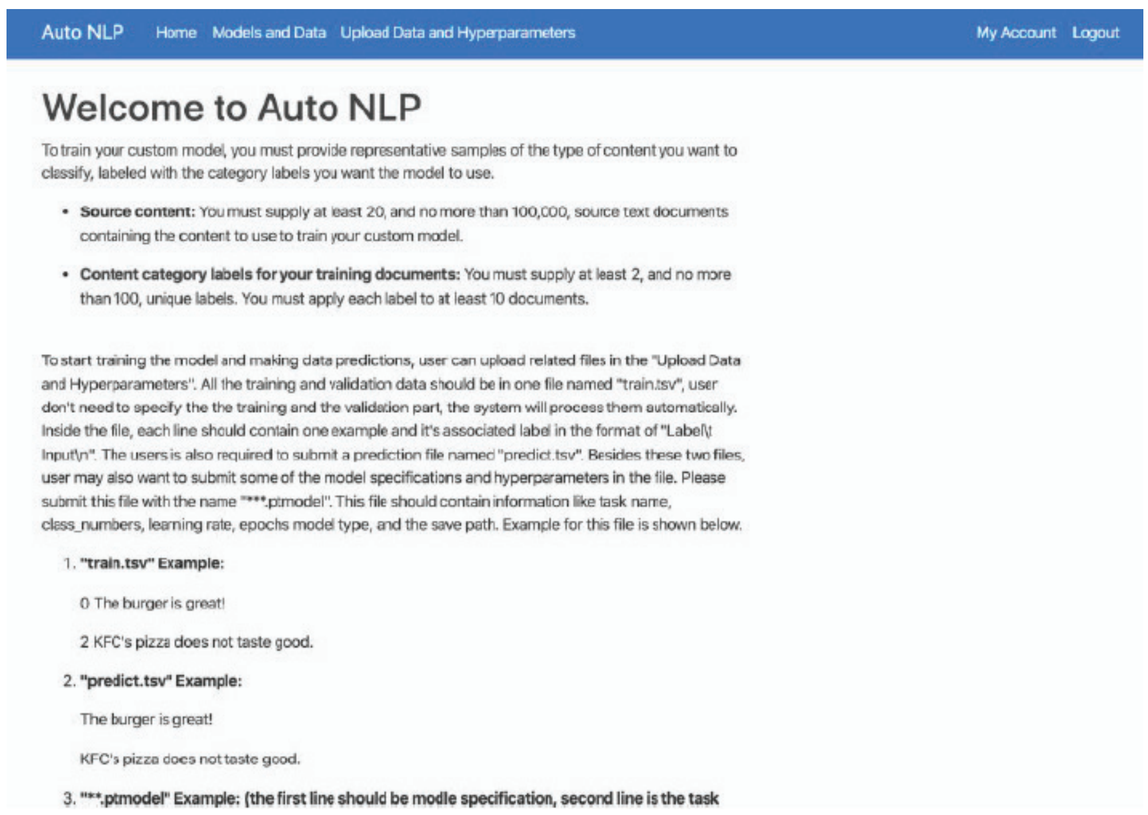


3. Result

We tested the performance on 20 NewsGroup and Reuters datasets

	Hyperparameter			20 NewsGroup				
	Model	Epoch	Learning Rate	Val Acc	Training Time(min)	Weight Reduced	Accuracy Increase	Time Reduction
1	$Ori_{\text{bert-base-uncased}}$	14	2e-5	83.03%	304	–	–	–
2	bert-base-uncased	14	2e-5	83.9%	178.09	**19.64%**	**0.87%**	**41.4%**
3	$Ori_{\text{bert-large-uncased}}$	8	3e-5	83.18%	551.85	–	–	–
4	bert-large-uncased	8	3e-5	83.9%	439.65	**20.14%**	**0.72%**	**20.3%**
5	$Ori_{\text{bert-large-uncased}}$	14	2e-5	83.43%	798.85	–	–	–
6	bert-large-uncased	14	2e-5	84.77%	606.18	**20.14%**	**1.34%**	**24.21%**

	Hyperparameter			Reuters				
	Model	Epoch	Learning Rate	Dev Acc	Training Time(min)	Weight Reduced	Accuracy Increase	Time Reduction
1	$Ori_{\text{bert-base-uncased}}$	90	1.75e-5	90.24%	341.09	–	–	–
2	bert-base-uncased	90	1.75e-5	92.15%	317.26	**19.63%**	**1.91%**	**7.3%**
3	$Ori_{\text{bert-large-uncased}}$	60	2.25e-5	92.39%	500.69	–	–	–
4	bert-large-uncased	60	2.25e-5	92.66%	444.73	**20.13%**	**0.27%**	**11.1%**



4. Conclusion

The newly proposed acceleration method could reduce the training time by about 17% and total weight by 20%. Moreover, this method could increase the overall performance in the field of Document Classification for BERT.