

Transformers Acceleration For AutoNLP application in Document classification

Student: Cao Hannan

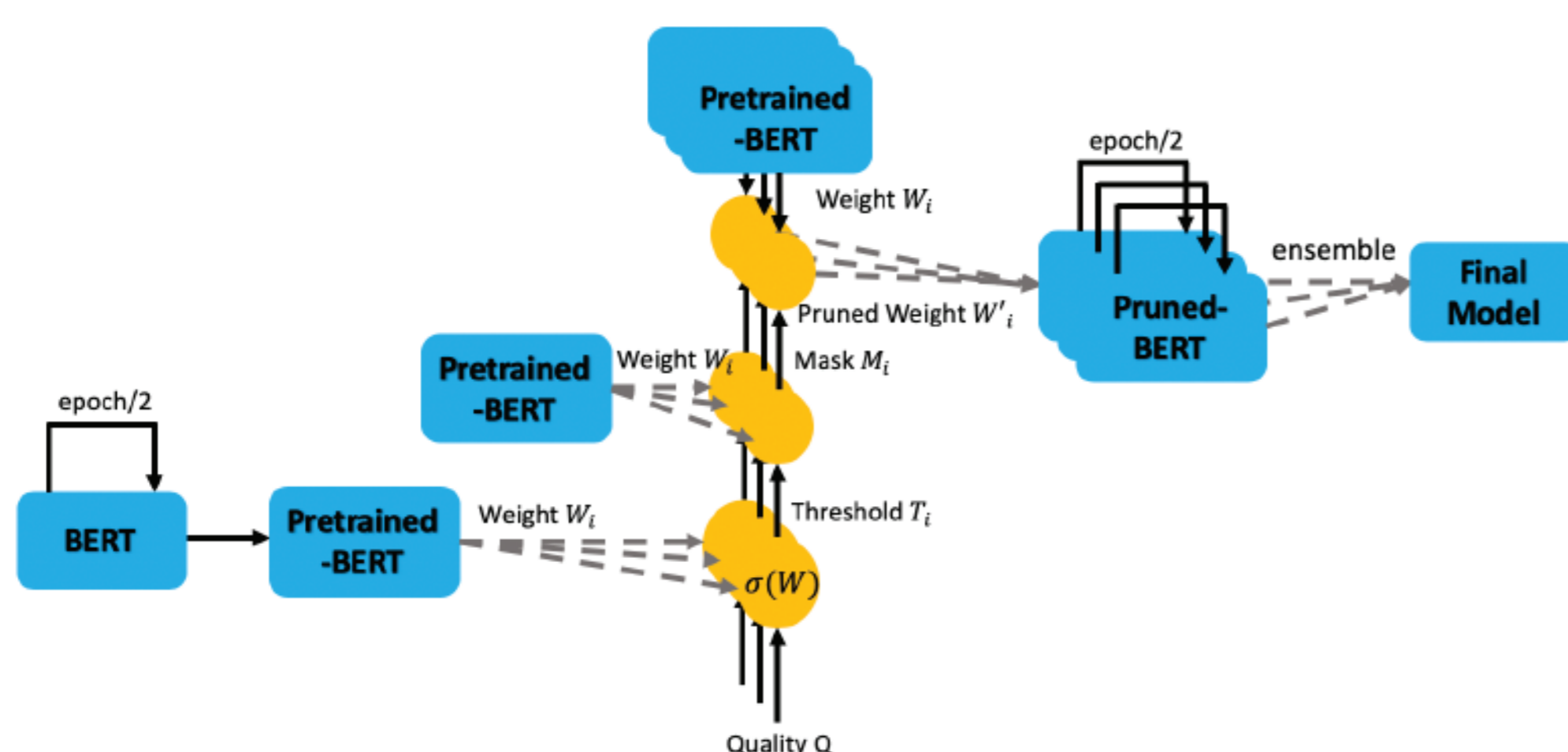
Supervisor: Assoc Prof Sinno Pan

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 _{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 _{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 _{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 _{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 _{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%

Auto NLP Home Models and Data Upload Data and Hyperparameters

My Account Logout

Welcome to Auto NLP

To train your custom model, you must provide representative samples of the type of content you want to classify, labeled with the category labels you want the model to use.

- **Source content:** You must supply at least 20, and no more than 100,000, source text documents containing the content to use to train your custom model.
- **Content category labels for your training documents:** You must supply at least 2, and no more than 100, unique labels. You must apply each label to at least 10 documents.

To start training the model and making data predictions, user can upload related files in the "Upload Data and Hyperparameters". All the training and validation data should be in one file named "train.tsv", user don't need to specify the training and the validation part, the system will process them automatically. Inside the file, each line should contain one example and its associated label in the format of "Label!\nInput\n". The users is also required to submit a prediction file named "predict.tsv". Besides these two files, user may also want to submit some of the model specifications and hyperparameters in the file. Please submit this file with the name "****.ptmodel". This file should contain information like task name, class_numbers, learning rate, epochs model type, and the save path. Example for this file is shown below.

1. "train.tsv" Example:


```
0 The burger is great!
2 KFC's pizza does not taste good.
```
2. "predict.tsv" Example:


```
The burger is great!
KFC's pizza does not taste good.
```
3. "****.ptmodel" Example: (the first line should be model specification, second line is the task)

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.