Model Comparison and Performance Optimization in News Text Classification

Zelin Zhang，Jian Li，Yinglun Suo，Yanhua Zhang

April 2024

CONTENTS

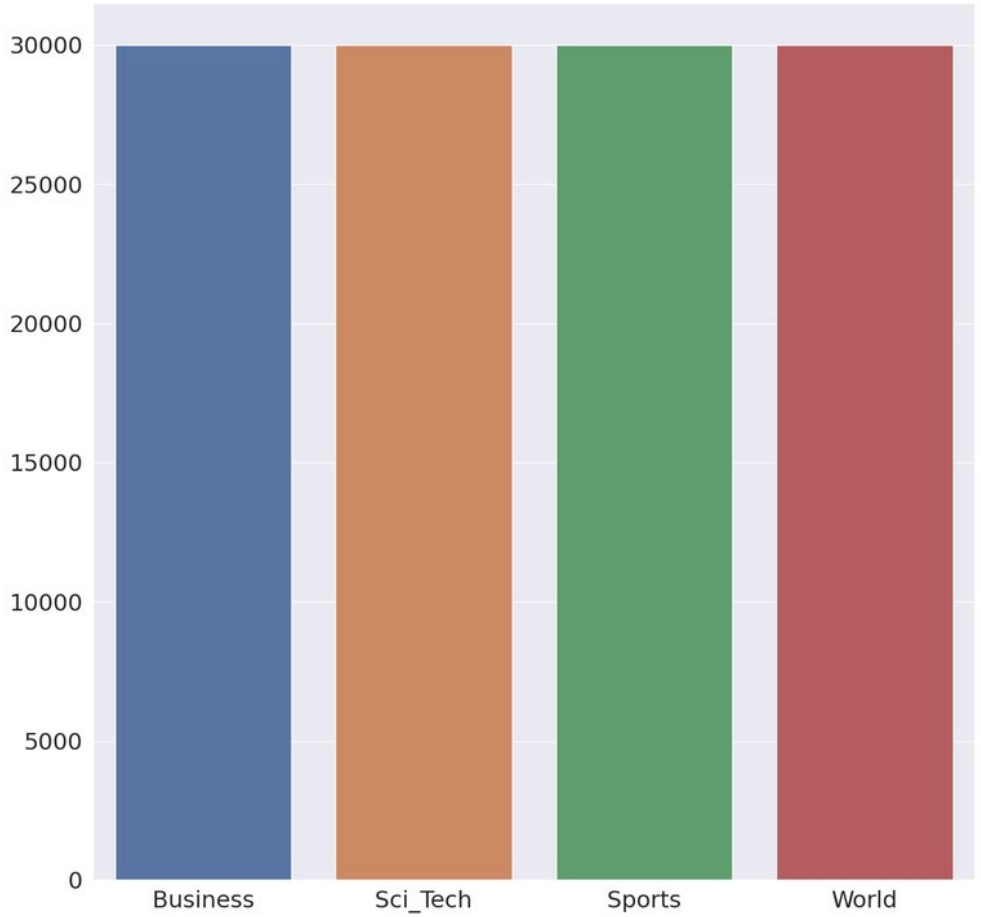
Introduction & Background

KEY WORDS

Deep Learning Natural Language Processing Text Classification Transfer Learning

Introduction

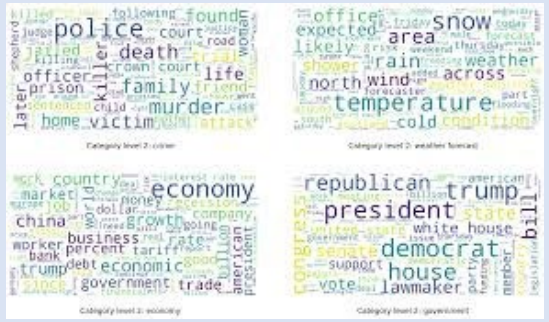
The rapid evolution of Natural Language Processing (NLP) technologies offers a unique chance to revisit the AGnews dataset, a benchmark for news text classification that has not fully exploited the latest advanced models. These pre-trained models, developed from extensive training on diverse textual data, have revolutionized the understanding and processing of human language. This thesis employs these cutting-edge models through transfer training to significantly enhance classification accuracy on the AGnews dataset. Our approach includes using both large and small model variants to investigate the balance between performance and computational efficiency. The goal is to not only surpass previously reported accuracy levels but also to demonstrate the practical applicability of these advanced models in real-world scenarios, thereby setting new benchmarks for future NLP tasks.



Research Target

Input
Chest X-Ray Image

Models
Transformers



Output
Label-

Performance and efficiency comparison between different models

We have conducted transfer training on many advanced models in recent years and successfully enabled them to achieve better results on AGnews compared to old models from many years ago. We analyze key indicators such as accuracy, training time, convergence speed, etc. for each model simultaneously. Finally, select the most suitable model.

Find the most suitable model for multiple environment

For some application scenarios where computing resources are relatively scarce, we have specifically set up a small scale group with fewer parameters outside of the large scale group. Their training time is shorter and they can adapt to more scenarios.

Experimental Design

Roadmap of CheXtNeXt

Dataset description and processing

AdamW loss function

Transfer learning

Save the best model

Dataset description

Find the best token length

In order to maximize our model performance and reduce training time, we need to choose a token length to ensure that this length can basically cover the items in the dataset. Reduce the potential loss of information without causing excessive computational resource waste.

	Average	Median	90%	99.5%	99.9%
percentage	38	37	48	80	112

tokenizer

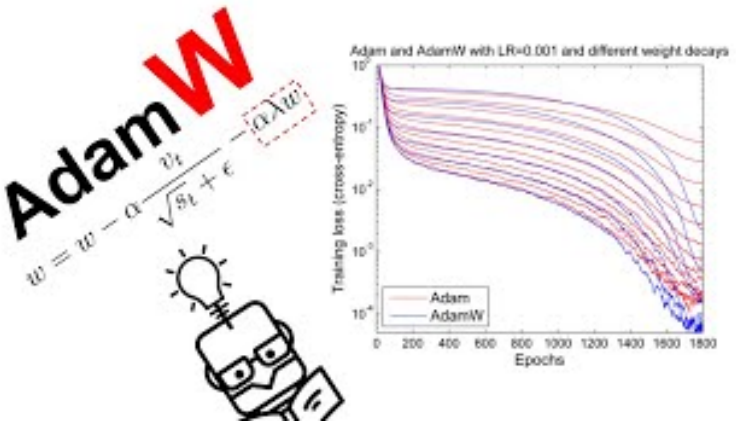
The size of the token

There are some differences in the tokens corresponding to each model, and we use the matching tokens provided by the pre trained model. Due to our previous finding that the length of 80 can cover 99.5% of the items in the AGnews dataset, we have set 80 as the size of each token. This can not only reduce information loss during the data preprocessing stage, but also prevent excessive performance overhead and delay in training time.

AdamW Optimizer

Why we choose AdamW

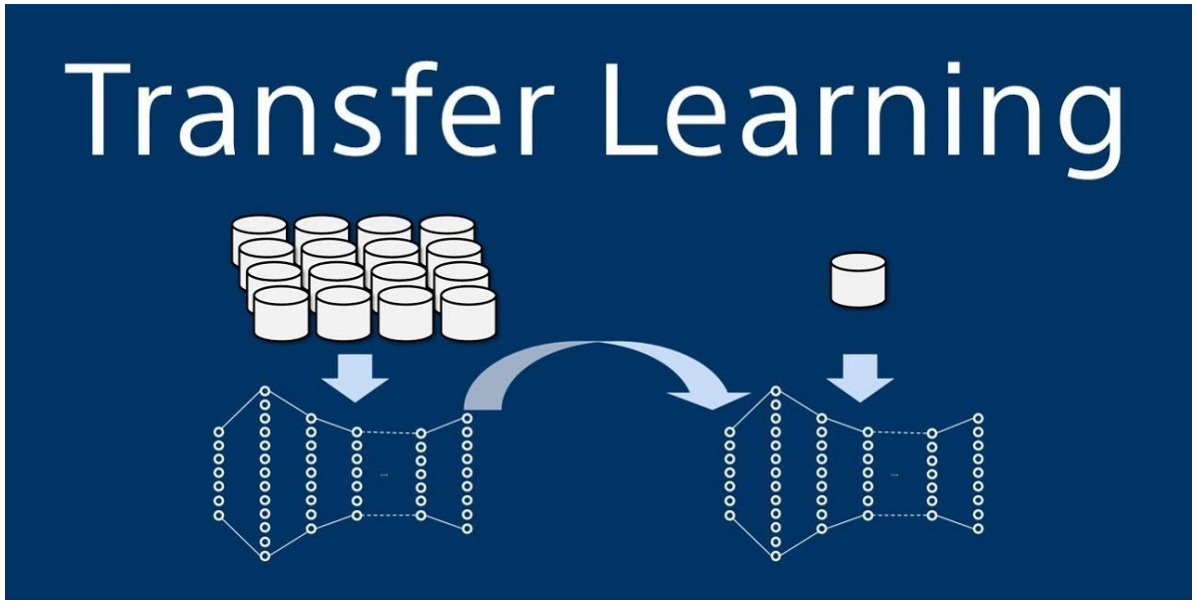
The AdamW optimizer, by decoupling weight decay from optimization steps, improves regularization and combats overfitting, making it especially suitable for handling the high-dimensional sparse data in news text classification. It adaptively adjusts the learning rate for each parameter, simplifies hyperparameter adjustment, and is easy to use. Compared to the traditional Adam optimizer, AdamW shows better training stability and performance, is more compatible with batch normalization techniques, and helps the model generalize better to unseen data. These features make it a powerful tool for dealing with rapidly evolving news content.



Transfer Learning

advantage

1)Efficiency: Transfer learning dramatically reduces the need for large amounts of labeled data, which is often a bottleneck in NLP tasks.
2)Improved Performance: Leveraging models pre-trained on extensive datasets provides a solid foundation, often leading to enhanced model performance on specific tasks.
3)Speed: It significantly accelerates the training process, making it feasible to deploy more sophisticated models within practical timeframes.

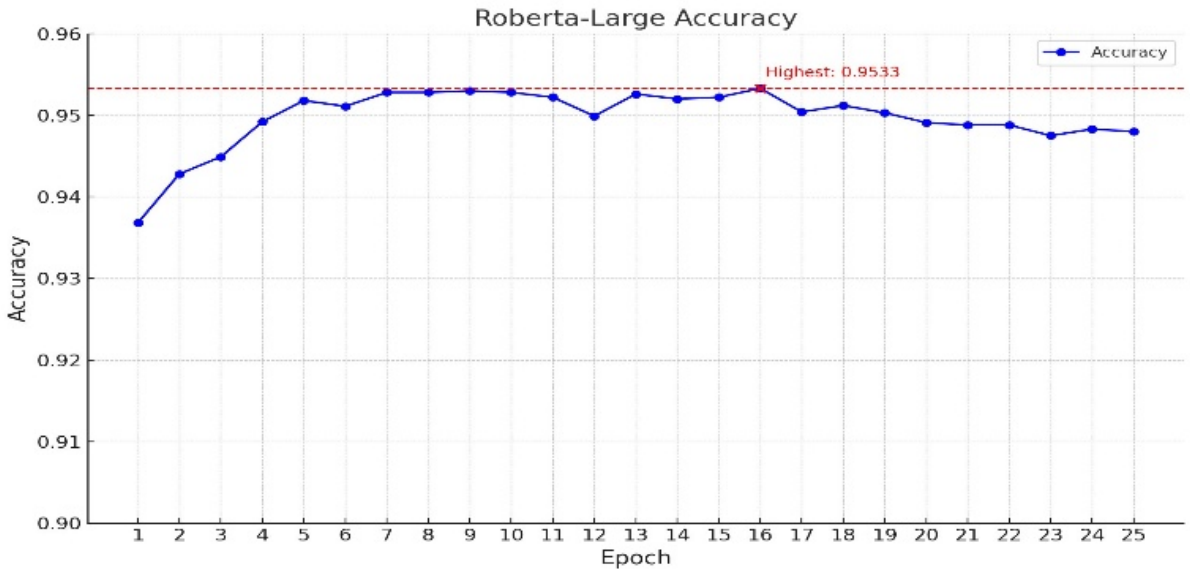


Stage4

Training and Saving the best model

Prevent overfitting

We only save the model as a. pt file when the training results improve. After experimental verification, the vast majority of models will reach the peak of accuracy in about 15 epochs, and then enter the downstream range.



Result & Evaluation

Experimental Environment

GPU

NVIDIA RTX4090

Version

PyTorch 2.1.0
Python 3.10
Cuda 12.1
Ubuntu 22.04

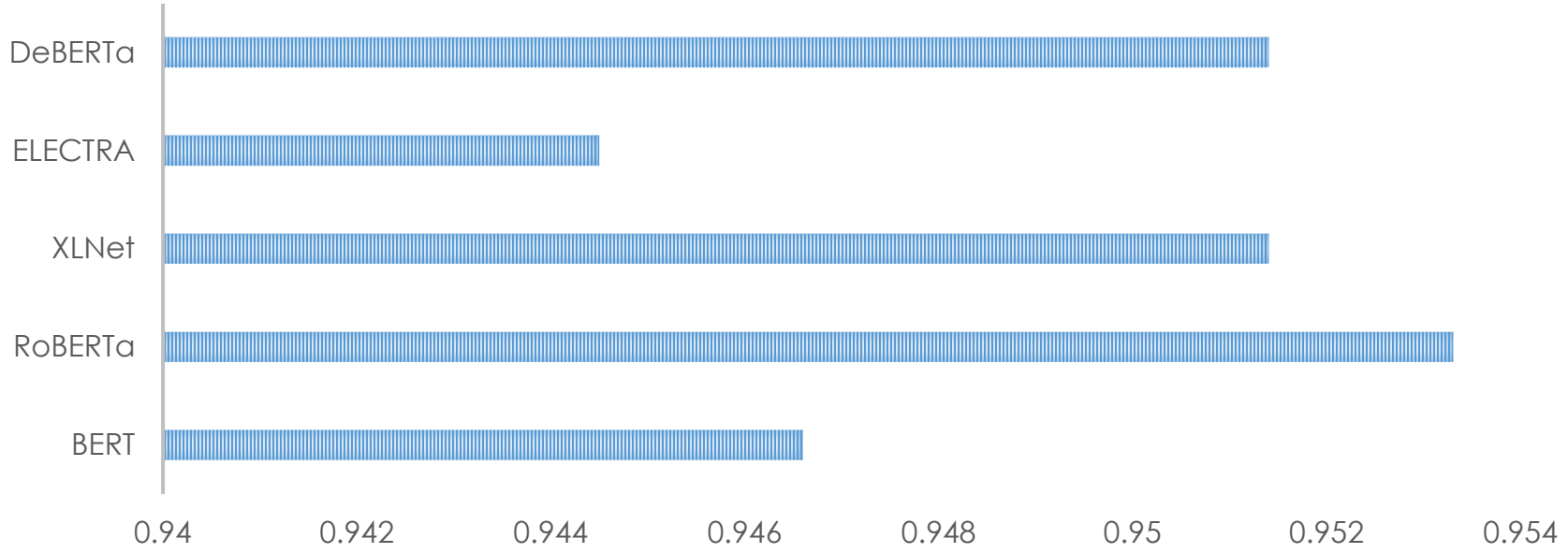
Float accuracy

We use mixed precision floating-point to accelerate the training process. Mixing FP16 and FP32

Experimental Result (float-point precision)

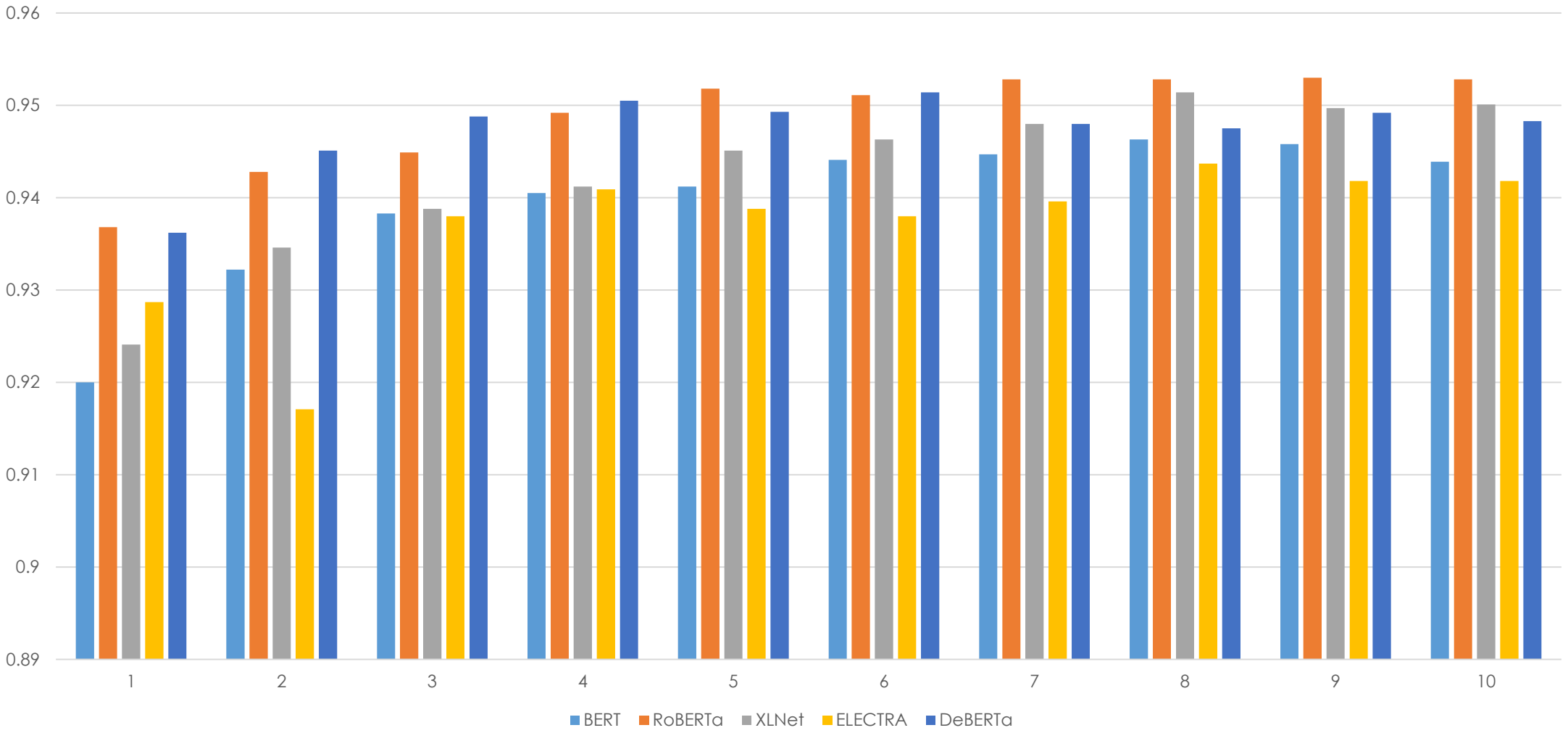
float precision	FP32	FP16
Max accuracy	0.9434	0.9433
Time of each epoch	1min 41sec	2min 57sec

Experimental Result (large-scale group)

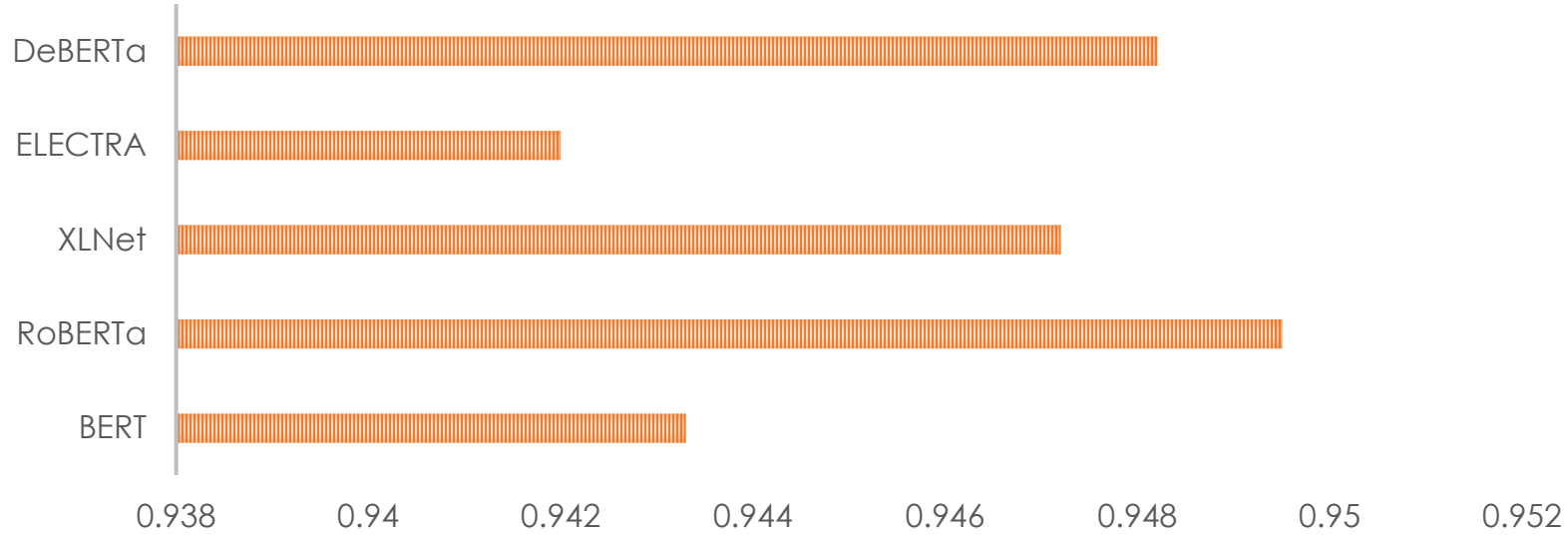


Model	BERT	RoBERTa	XLNet	ELECTRA	DeBERTa
Time	4min 26sec	4min 30sec	6min 16sec	4min 33sec	6min 49sec

Experimental Result (large-scale group)

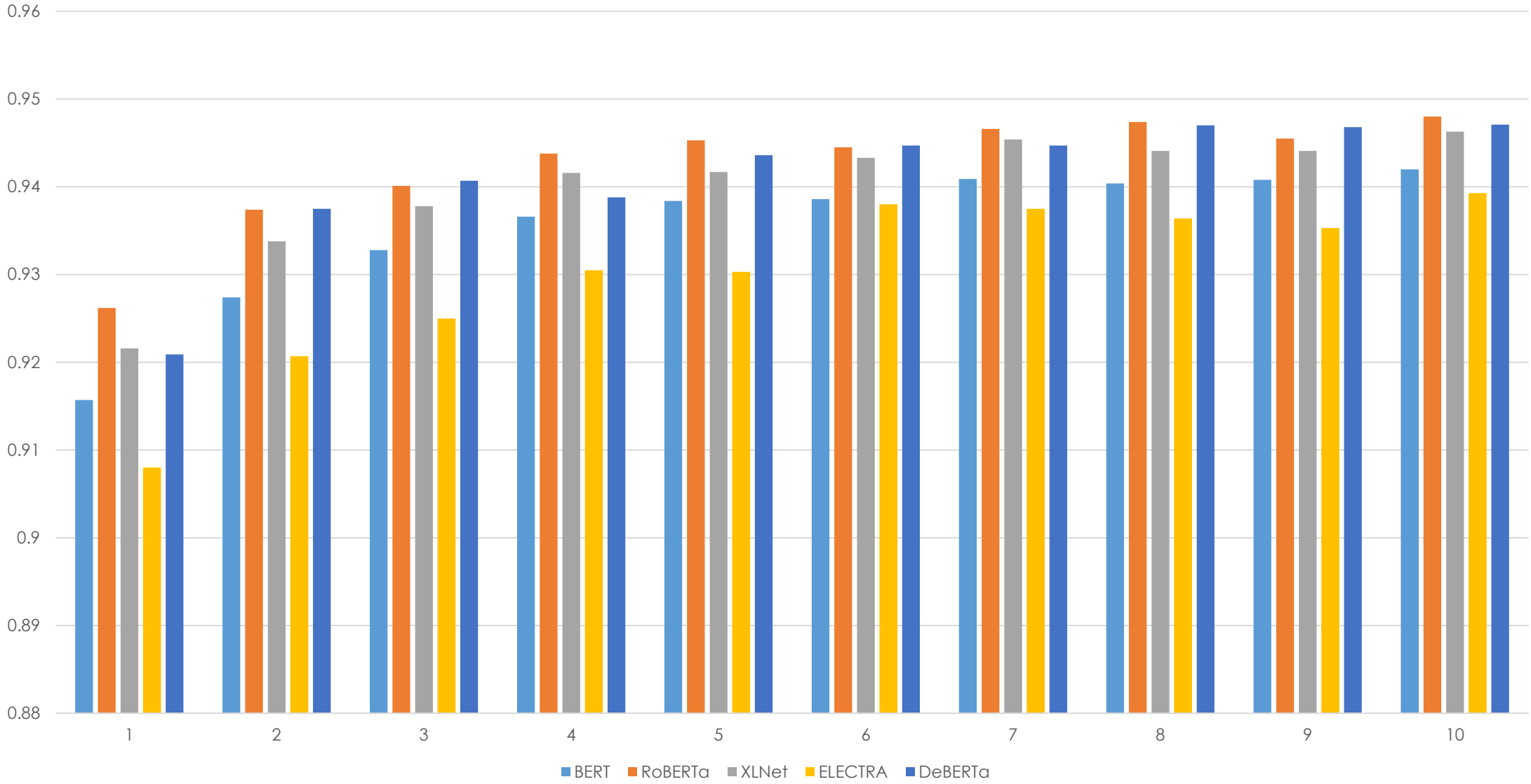


Experimental Result (small-scale group)



Model	BERT	RoBERTa	XLNet	ELECTRA	DeBERTa
Time	1min 41sec	1min 34sec	2min 25sec	1min 36sec	2min 33sec

Experimental Result (small-scale group)



Conclusion & Future Work

Conclusion

We can easily draw conclusions by comparing all the experimental results. RoBERTa is the most suitable dataset among several modern advanced models. Whether in a 300M parameter level large model for performance comparison or a 100M parameter level base model. RoBERTa Large has become the model with the highest performance, fastest convergence, and the second shortest training time per epoch. In the basic model group, we found that RoBERTa Base achieved the best performance in all three important indicators: accuracy, convergence speed, and training time per epoch. This fully demonstrates the strong ability of RoBERTa in the task of news text classification, making it the most competitive model structure in this sub field.

Future Work

Future Work

Model fine-tuning

Due to the time limit of the experiment, we have not further adjusted the hyperparameter. We will continue to improve in the future to make the performance of the model higher.

Expanding the training dataset

We did not use a dataset other than AGnews. In future we will test our model in more similar dataset to verify our conclusion.

Use more powerful computing resources

Our project is constrained by insufficient GPU computing power to implement the first two proposals, and we will use multi-GPU training to solve the problem in the future.

Acknowledgement

All members of our group in this project would like to express our gratitude to Dr. Ahmed Ibrahim, the instructor of the ECE9039 machine, for the knowledge he has taught and guidance on this project.

Western
UNIVERSITY · CANADA

Thank you!