Samuel Belkadi
Frenciel Anggi

Are Bigger Models Always Better? A Study on Natural Language Inference

The University of Manchester

Motivation

- (Very) Large Language Models are witnessing an exponential surge in model size, implying expensive high-end hardware and expansive training datasets.
- Consequently, training processes become lengthy, entailing considerable ecological and environmental impacts (e.g. training GPT3 emits ~1.9 tons of CO_2)
- Could multiple smaller models be considered solid alternatives by working out together on a given task?

Low-Rank Adaptation (LoRA)

- We use the LoRA technique as an alternative because we cannot afford the training requirements of very large language models.
- The pretrained weights of RoBERTa-Large are frozen, and trainable low-rank matrices (or layers) are introduced within the model.
- Only the low-rank matrices are fine-tuned for the given task, which accounts for about 0.5% of the overall model size.
- This way, the pretrained knowledge can be used during fine-tuning while the resources and time requirements remain acceptable.

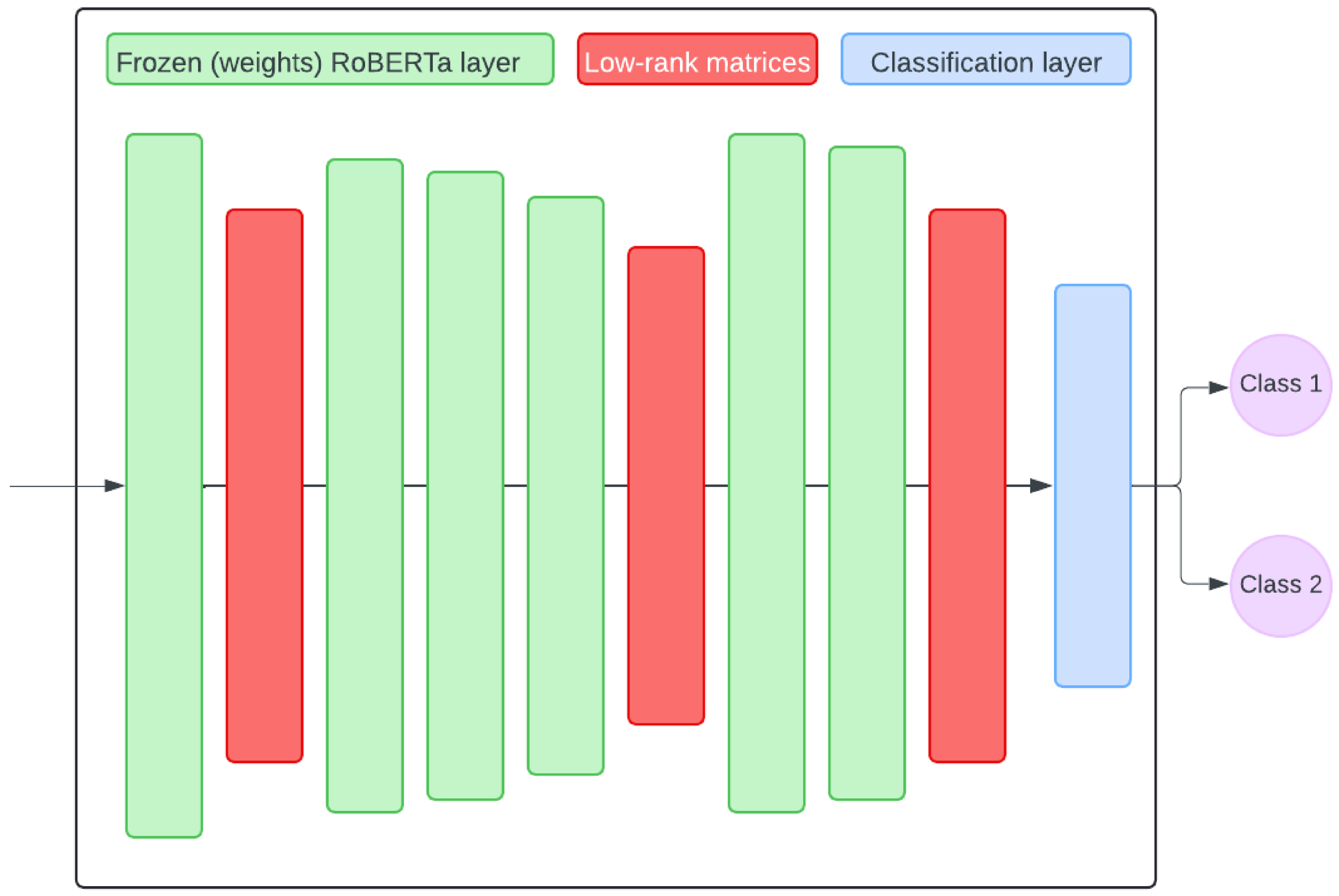

Figure 1. Illustration of a RoBERTA-Large model with additional LoRA layers.

Ensemble Model

- Ensemble models adopt a strategy where multiple small models, called base learners, undergo parallel training.
- Each base learner either trains on a random subset of the data to encourage diversity within the ensemble or targets weaknesses observed in other base learners.
- The ensemble leverages a normalized geometric mean to aggregate predictions from individual learners.
- By averaging out outlier predictions, the ensemble achieves a robust overall performance, leveraging individual learner's strengths.

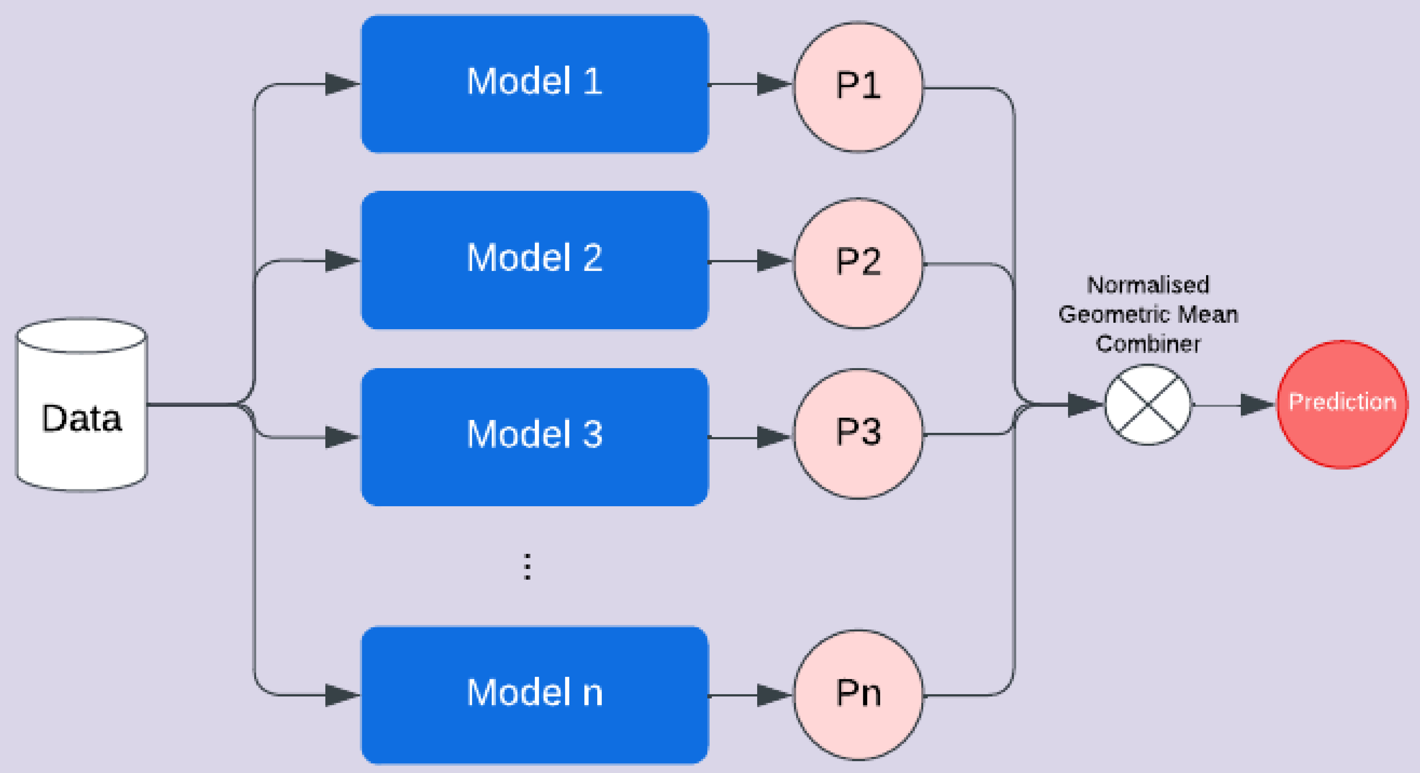

Figure 2. Ensemble architecture combining learners' predictions using the normalised geometric mean. Each base learner is a BiLSTM.

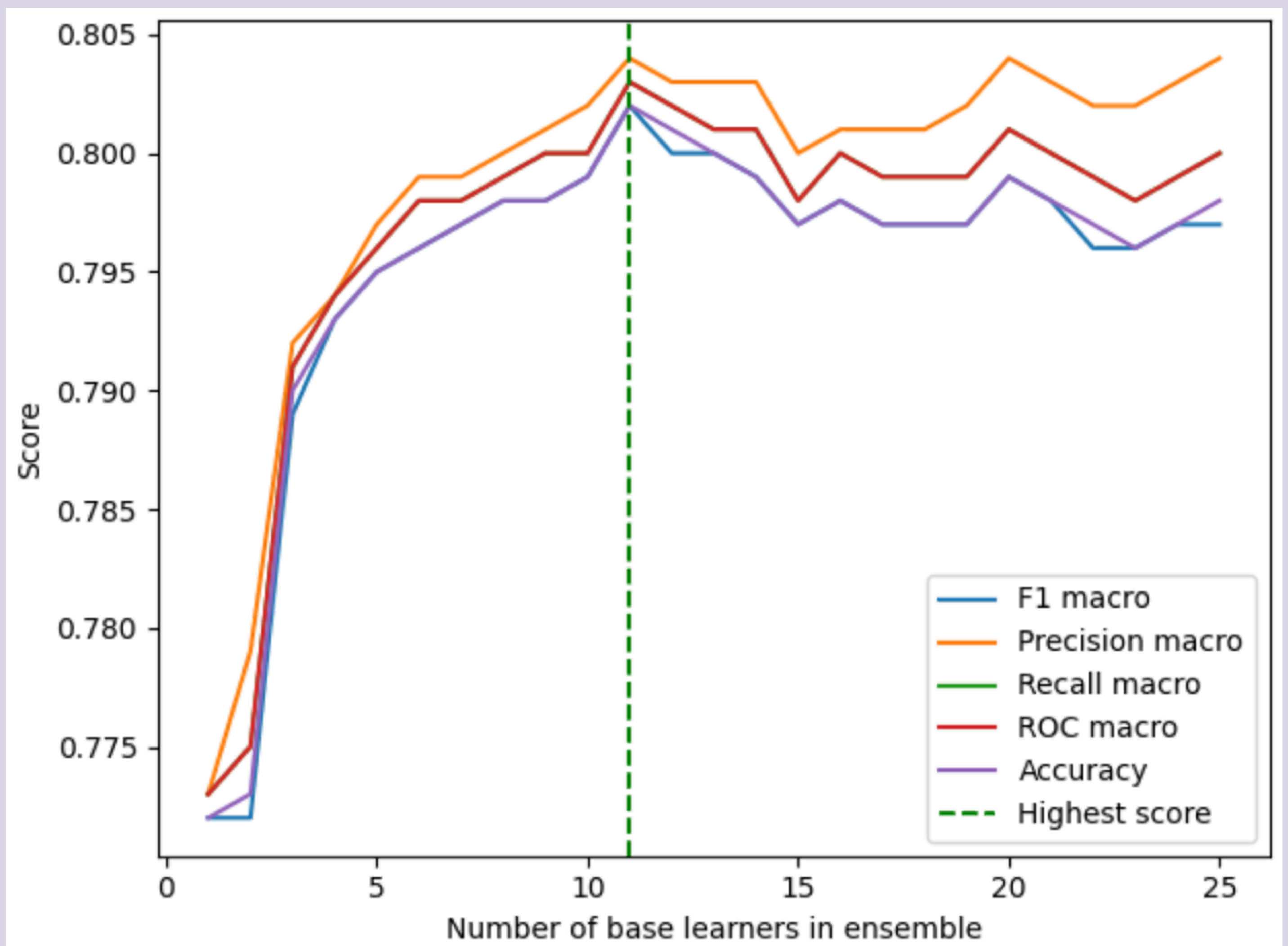

Figure 3. Evaluation results of the ensemble model as the number of base learners increases. Best result at n_learners=11 and epochs=8.

Results

	Category B		Category C	
Metrics	BL (BiLSTM)	Ensemble	BL (BERT)	LoRA
F1 (weighted)	0.5615	0.802	0.7864	**0.9158**
F1 (macro)	0.5617	0.802	0.7860	**0.9157**
Precision (macro)	0.5627	0.804	0.7874	**0.9160**
Recall (macro)	0.5626	0.803	0.7856	**0.9155**
ROC (macro)	-	0.803	-	**0.9155**
Accuracy	0.5616	0.802	0.7867	**0.9158**

Table 1. Evaluation results on the validation dataset. Proposed models show significant improvement in all performance metrics over baseline models.

Accuracies of:

- BiLSTM (baseline): 56.2%
- Ensemble of BiLSTM (11M params): 80.2%
- BERT-base (baseline) (110M params): 78.7%
- RoBERTa-large (with LoRA) (356M params): 91.6%

Our proposed models significantly enhanced all five classification metrics on the NLI task, compared to the baseline performance.

Conclusion

- Ensembles offer a cost-effective and ecological-friendly alternative solution.
- With a highly parallelizable structure and cluster-based architecture, the ensemble models outperformed individual BiLSTM models.
- Despite its modest size, the ensemble could outperform the BERT-Base LLM with 11M vs 110M parameters.
- However, against extremely large models (e.g. RoBERTa Large), ensembles may still lag behind in performance.
- Nonetheless, it remains an interesting solution for balancing training cost and performance, particularly in domains where performance is not imperative.

References

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Wood, D., Mu, T., Webb, A.M., Reeve, H.W., Lujan, M. and Brown, G., 2023. A unified theory of diversity in ensemble learning. Journal of Machine Learning Research, 24(359), pp.1-49.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.