YNU-HPCC at WASSA-2023 Shared Task 1: Large-scale Language Model with LoRA Fine-Tuning for Empathy Detection and Emotion Classification

Yukun Wang, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
wangyukun@mail.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper describes the system for the YNU-HPCC team in WASSA-2023 Shared Task 1: Empathy Detection and Emotion Classification. This task needs to predict the empathy, emotion, and personality of the empathic reactions. This system is mainly based on the Decoding-enhanced BERT with disentangled attention (DeBERTa) model with parameter-efficient fine-tuning (PEFT) and the Robustly Optimized BERT Pretraining Approach (RoBERTa). Low-Rank Adaptation (LoRA) fine-tuning in PEFT is used to reduce the training parameters of large language models. Moreover, back translation is introduced to augment the training dataset. This system achieved relatively good results on the competition's official leaderboard. The code of this system is available here.

1 Introduction

The purpose of WASSA-2023 Shared Task 1 (Barriere et al., 2023) is to use empathic reaction data to predict hidden sentiment and personality. This task consisted of five tracks:

- **Track 1:** Empathy and Emotion Prediction in Conversations (CONV), which consists in predicting empathy, emotion polarity, and emotional intensity in a conversation;
- **Track 2:** Empathy Prediction (EMP), which consists in predicting empathy, and personal distress in an essay;
- **Track 3:** Emotion Classification (EMO), which consists in predicting the emotion in an essay;
- **Track 4:** Personality Prediction (PER), which consists in predicting the conscientiousness, openness, extraversion, agreeableness, and stability of the essay writer;
- **Track 5:** Interpersonal Reactivity Index Prediction (IRI), which consists in predicting perspective-taking, personal distress, fantasy, and empathetic concern of the essay writer;

Although the prediction goals are different, all five tracks can be considered as either a sentiment classification (Peng et al., 2020) or regression task (Kong et al., 2022). One of the biggest challenges in this task lies in how to learn representation for the given text. The early exploration was based on text similarity (Jijkoun and Rijke, 2005) or text alignment (de Marneffe et al., 2008). With the development of neural networks, convolutional neural networks (CNN) (Kim, 2014) and recurrent neural networks (RNN) (Zaremba et al., 2014) and their variants are adopted to learn text representations. Both CNN and RNN are shallow models, which only incorporate previous knowledge in the first layer of the model. The models are also based on word embeddings that are useful in only capturing the semantic meaning of words without understanding higher-level concepts like anaphora, long-term dependencies, and many more.

Beyond word embeddings, recent studies proposed embedding from language models (ELMo), which can learn word embeddings by incorporating both word-level characteristics as well as contextual semantics (Zhang et al., 2021). This also led to the emergence of pre-trained models (PLM) using Transformers as basic units. The PLMs, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and DeBERTa (He et al., 2020), are first fed a large amount of unannotated data, allowing the model to learn the usage of various words and how the language is written in general. Then, they can be finetuned to be transferred to a Natural Language Processing (NLP) task where it is fed another smaller task-specific dataset. As the scale of PLMs increases, the model performance in downstream tasks becomes better and better. Nevertheless, the fine-

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 526–530
July 14, 2023 ©2023 Association for Computational Linguistics

tuning procedure brings about increased requirements for model training costs. For example, the large sequence-to-sequence model GPT-3 has 175B parameters (Brown et al., 2020). To reduce training costs, recent studies suggest using parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019) to enable the efficient adaption of PLMs to downstream applications without fine-tuning all the parameters of the PLMs.

To this end, this paper proposes to use DeBERTa fine-tuned with Low-Rank Adaptation (LoRA) (Hu et al., 2021) in PEFT and RoBERTa for all tracks in this competition. Both the DeBERTa and RoBERTa were initialized from a well-trained checkpoint, e.g., `deberta-v2-xxlarge` with 1.5B parameters and `roberta-base` with 125M parameters. For finetuning, LoRA only fine-tuned a small number of (extra) model parameters, thereby greatly decreasing the computational and storage costs. For classification tasks, a softmax head with the cross-entropy loss was applied, while a linear decoder head with the mean squared error was adopted for regression tasks.

The experimental results on the development dataset show that the XXL version of DeBERTa with LoRA and back translation achieves the best performance in tracks 1, 3, and 5. Although the number of trainable parameters decreases, the model achieves performance comparable to that of full fine-tuning. Additionally, RoBERTa with back translation achieved the best performance in tracks 2 and 4. The difference in the performance of the two models on different tracks may be due to the impact of the size of the training dataset.

The rest of this paper is organized as follows. Section 2 describes the system model and method. Section 3 discusses the specific experimental results. Conclusions are finally drawn in Section 4.

2 System description

The architecture of the proposed model is shown in Figure 1. The given text of conversations or essays is input into the tokenizer and then segmented into the corresponding token ID. Subsequently, DeBERTa or RoBERTa's encoder is used to extract the features of the text in a vector format. Meanwhile, LoRA is used to reduce fine-tuning parameters without degrading performance too much. Finally, the encoded hidden representation is used for both sentiment classification and regression.

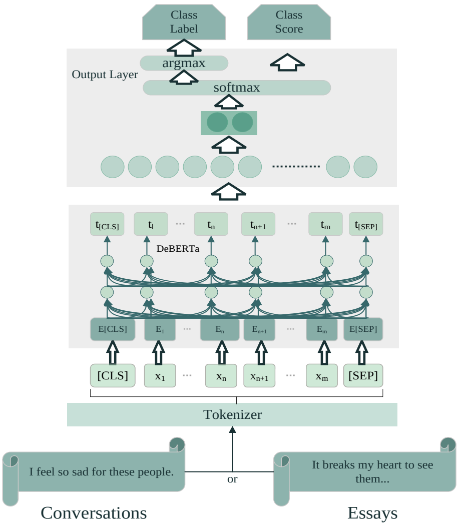

Figure 1: The structure for the system.

2.1 Tokenizer

SentencePiece and WordPiece were used for DeBERTa and RoBERTa to divide the text into subwords, respectively. The final output X of the tokenizer is denoted as,

$$X = [CLS]x_1x_2 \dots x_m[SEP] \qquad (1)$$

where m is the length of the given text, the [CLS] special tag is used to indicate the beginning of a text sequence, and the [SEP] special tag is used to indicate the separation of a text sequence.

2.2 RoBERTa

The RoBERTa used in this system is a model improved on BERT. BERT's pre-trained tasks include Masked Language Model (MLM) and Next Sentence Prediction (NSP). RoBERTa removed the NSP task, increased the batch size of the model, and used more training data. The performance improvement of RoBERTa has been demonstrated through experimental comparison. The RoBERTa used in this task was initialized from `roberta-base`, with the main structure of 12 layers, 768 hidden size, and 125M total parameters.

2.3 DeBERTa

DeBERTa used in this system improves the text representation capabilities of BERT and RoBERTa models using disentangled attention and enhanced mask decoder methods. Each word is represented using two vectors that encode its content and position, respectively. The attention weights among

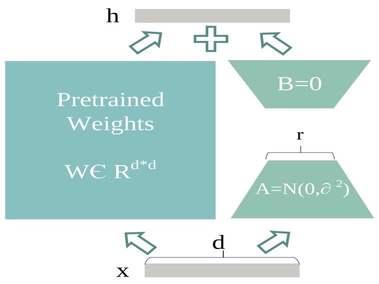

Figure 2: The conceptual diagram of parameter-efficient LoRA fine-tuning.

words are computed using disentangled matrices on their contents and relative positions. Then, an enhanced mask decoder is used to replace the output softmax layer to predict the masked tokens for model pretraining. It outperforms BERT and RoBERTa on many natural language understanding (NLU) tasks. The checkpoint of DeBERTa used in this system is `deberta-v2-xxl`, with the main structure of 48 layers, 1536 hidden size, and 1.5B total parameters.

2.4 LoRA

Transferring the models to the downstream tasks usually depends on the size of the training dataset and pre-trained model. However, the hardware cost of using large models is very significant. Meanwhile, large models are over-parameterized and have a smaller intrinsic dimension (Houlsby et al., 2019). Therefore, this system used LoRA to freeze most parameters and fine-tune the model through low-rank matrices. The LoRA decomposition is defined as,

$$W_0 + \Delta W x = W_0 + BAx \tag{2}$$

where W_0 represents the original parameter matrix. It is very huge and difficult to train. In this system, training updates to W_0 can be represented by ΔW. Therefore, W can be frozen to reduce a large number of training parameters. A and B represent the low-rank factorization matrix of W_0. A is initialized with random Gaussian and B is initialized with zero. Therefore, ΔW is initialized with zero.

LoRA reduces parameters by training a low-rank iterative decomposition matrix of the original parameter matrix. The original parameters of XXL DeBERTa used in this system are 1.5B, while the trainable parameters after LoRA processing are around 4 million. So, this method makes using a large language model on consumer-grade GPUs a reality.

2.5 Output Layer

The output layer is implemented in two distinct ways to accomplish classification and regression tasks.

Regression. Regression was performed for tracks 1, 2, 4, and 5. The training goal is to minimize the mean squared error (MSE) loss, denoted as,

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - P_i)^2 \tag{3}$$

where P_i is the predicted value, y_i represents the ground-truth, and n represents the number of training samples in a batch.

Classification. The classification was performed for track 3. A softmax function is used to predict probability distribution over the candidate labels. The training objective is to minimize the cross-entropy between the predicted labels and the ground truth, denoted as,

$$L_2 = -\frac{1}{N}\sum_{i}\sum_{c=1}^{C} y_{ic}\log P_{ic} \tag{4}$$

where C represents the number of categories classified, y_{ic} is the ground-truth label, and P_{ic} represents the prediction probability of the c-th class.

3 Experimental Results

This section evaluates the performance of the proposed system for both sentiment classification and regression tasks.

3.1 Datasets

This task is based on an Empathic Conversations dataset. The dataset marks conversations and essays after people read news stories about individuals, groups, or others who have been harmed (Omitaomu et al., 2022). This dataset for training contains two levels of sentiment classification: (1) Conversations between two users after reading the same news stories. The labels mainly include Emotional Polarity, Emotion, and Empathy. (2) Essays from each user. The labels mainly include Empathy, Emotion, Personality, and Interpersonal Reactivity Index. Each sentimental transition in user conversations or essays is interpreted as labels. The size of the training dataset for the conversation level is around 8700, while the size of the training

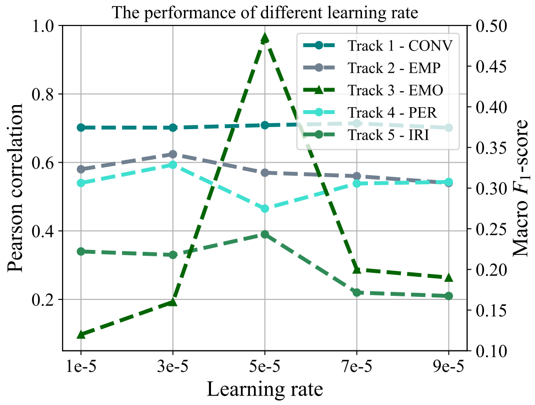

Figure 3: The performance of different learning rates on development dataset.

dataset for the essay level is around 770. Macro Correlation metric is used in tracks 1, 2, 4, and 5, Macro F1-score is used in track 3.

3.2 Implementation Details

The conversation-level dataset provided conversation text, and the essay-level dataset provided essay text and person-level demographic information (age, gender, ethnicity, income, and education level). In track 1, this system used conversation text as training data and used essay text as training data in tracks 2, 3, 4 and 5. All training datasets are first translated into Chinese and then translated back into English. This method of back translation can double the training datasets. Additionally, this system has chosen BERT as a baseline model.

The learning rate was fine-tuned on the development dataset. The results were shown in Fig. 3.

3.3 Comparative Results and Discussion

Tables 1 and 2 show the comparative results of BERT, RoBERTa, and DeBERTa with LoRA on different classification and regression tasks on the development dataset. It can be found that the average performance of the optimized RoBERTa and DeBERTa is better than BERT. DeBERTa's disentangled attention mechanism helps to improve the model's text representation ability because it not only calculates the attention weight of content and relative position for all word pairs but also considers the absolute positions of words. The results show that DeBERTa + LoRA performs better in tracks 1, 3, and 5, while RoBERTa performs better in tracks 2 and 4. This may be due to the relatively larger scale of training data for track 1, and the fact that track 3 is a complex 31-classification task. Therefore, DeBERTa+LoRA improves the performance of sentiment classification and regression tasks. We submitted the best results of each track on the leaderboard. The final results of the test dataset are shown in Table 3.

Track	BERT	RoBERTa	DeBERTa+LoRA
Track1-CONV	0.714	0.721	**0.767**
Track2-EMP	0.502	**0.624**	0.544
Track4-PER	0.342	**0.593**	0.508
Track5-IRI	0.278	0.353	**0.39**

Table 1: Comparative results using Pearson Correlation in the development dataset.

Track	BERT	RoBERTa	DeBERTa+LoRA
Track3-EMO	0.271	0.169	**0.486**

Table 2: Comparative results using Macro F_1 score in the development dataset.

Track	Score
Track1-CONV	0.730 (Pearson Correlation)
Track2-EMP	0.288 (Pearson Correlation)
Track3-EMO	0.514 (Macro F_1)
Track4-PER	0.252 (Pearson Correlation)
Track5-IRI	0.154 (Pearson Correlation)

Table 3: Final score in the test dataset.

4 Conclusion

This paper proposed a system submitted in shared task 1 of WASSA-2023, which uses RoBERTa and XXL version of DeBERTa as the pre-trained models and fine-tuning the DeBERTa model using LoRA. The experimental results indicate that this system has achieved good performance. In addition, this system has a lot of space for improvement compared to the top-ranked systems. Future works will attempt to try other text augmentation and generation methods to achieve better results.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task:

Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 1039–1047. The Association for Computer Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Valentin Jijkoun and Maarten Rijke. 2005. Recognizing textual entailment using lexical similarity. *Journal of Colloid and Interface Science - J COLLOID INTERFACE SCI*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical BERT with an adaptive fine-tuning strategy for document classification. *Knowl. Based Syst.*, 238:107872.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and Jo ao Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Bo Peng, Jin Wang, and Xuejie Zhang. 2020. Adversarial learning of sentiment word representations for sentiment analysis. *Information Sciences*, 541:426–441.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

You Zhang, Jin Wang, and Xuejie Zhang. 2021. Personalized sentiment classification of customer reviews via an interactive attributes attention model. *Knowl. Based Syst.*, 226:107135.