Penn & BGU BabyBERTa+ for Strict-Small BabyLM Challenge

Yahan Yang
University of Pennsylvania
yangy96@seas.upenn.edu

Elior Sulem
Ben-Gurion University
eliorsu@bgu.ac.il

Insup Lee
University of Pennsylvania
lee@cis.upenn.edu

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

The BabyLM Challenge aims at pre-training a language model on a small-scale dataset of inputs intended for children. In this work, we adapted the architecture and masking policy of BabyBERTa (Huebner et al., 2021) to solve the strict-small track of the BabyLM challenge. Our model, Penn & BGU BabyBERTa+, was pre-trained and evaluated on the three benchmarks of the BabyLM Challenge. Experimental results indicate that our model achieves higher or comparable performance in predicting 17 grammatical phenomena, compared to the RoBERTa baseline. [1]

1 Introduction

With the emergence of deep-learning techniques (Liu et al., 2019; Vaswani et al., 2017), large language models pre-trained on massive datasets containing billions or trillions of words have achieved remarkable performance across various downstream tasks. However, the BabyLM challenge (Warstadt et al., 2023) highlights the importance of investigating the impact of small-scale pretraining and cognitive modeling. BabyBERTa (Huebner et al., 2021), a variant of the RoBERTa architecture in a smaller size, demonstrated superior performance and data efficiency in learning grammar phenomena with child-directed inputs compared to RoBERTa-base (Liu et al., 2019). Inspired by this work, we propose a model named Penn & BGU BabyBERTa+[2] (encoder-only), which shares the architecture and pretraining policies, for the BabyLM challenge with BabyBERTa. In this work, we consider the strict-small challenge which contains approximately 10M words for small-scale pretraining. We provide the details of our BabyBERTa+ in Section 2 and show the result of the BabyLM challenge in Section 3.

2 Methodology

In this section, we provide the descriptions of our BabyBERTa+ model including the architectures, tokenizers, training objectives and so on. As shown in Table 1, our model is much smaller compared to RoBERTa-base in terms of depth and width but uses a different masking policy[3]. The pre-training hyperparameters are the same as in the RoBERTa baseline as used in Warstadt et al. (2023) if not specified in Table 1 and the architecture choices are based on (Huebner et al., 2021). The model is pre-trained on the dataset ($\sim$ 10M words) provided in the strict-small track of the challenge. In other words, BabyBERTa+ differs from BabyBERTa (Huebner et al., 2021) by its vocabulary size and training corpus.

	RoBERTa-base	BabyBERTa+
layers	12	8
attention heads	12	8
hidden size	768	256
intermediate size	3072	1024
vocabulary size	50265	30000
epochs	20	100

Table 1: Comparison of RoBERTa and BabyBERTa+ in terms of their architectures.

2.1 Tokenizer

Following previous work (Liu et al., 2019; Huebner et al., 2021), our model utilizes Byte-Pair Encoding to create a vocabulary containing both words and subwords. We create a tokenizer with a vocabulary size of 30,000 and train it on the strict-small dataset.

2.2 Unmasking Removal Policy

To train the masked language model, the standard RoBERTa masking strategy replaces 80% of the corrupted tokens with the "<mask>" token, while 10% of the tokens are replaced with random tokens, and the remaining 10% are left unchanged. The

[1] Our Dynabench submission ID is 1372. The link to access the model is https://huggingface.co/yangy96/BabyLM_strict_small_Penn-BGU-BabyBERTa/tree/main.

[2] In our paper, we use Penn & BGU BabyBERTa+ and BabyBERTa+ interchangeably.

[3] Our implementation of the model is based on the Huggingface transformer library (Wolf et al., 2020).

Acc.	Ana Agr.	Agr. Str	Binding	C/R	D-N Agr.	Ellipsis	Filler-Gap	Irregular	Isl. Eff
R	81.5	67.1	67.3	67.9	90.8	76.4	63.5	87.4	39.9
B+	83.6	68.2	66.9	65.9	92.4	82.5	65.8	92.1	39.7
	NPI	Quan.	S-V Agr.	Hypernym	QA (easy)	QA (tricky)	Subj.-Aux.	Turn Taking	
R	55.9	70.5	65.4	49.4	31.3	32.1	71.7	53.2	
B+	68.8	75.9	68.1	50.2	71.9	40.6	87.6	67.5	

Table 2: Zeroshot performance of RoBERTa (R) and BabyBERTa+ (B+) on the BLiMP benchmark. The performances were reported in terms of accuracy.

Acc.	CoLA	SST-2	MRPC	QQP	MNLI	MNLI-mm	QNLI	RTE	BoolQ	MultiRC	WSC
R	70.8	87	79.2	73.7	73.2	74	77	61.6	66.3	61.4	61.4
B+	69.48	86.42	82	81.08	70.39	71.18	69.29	51.52	61.69	60.02	61.45

Table 3: Comparison of RoBERTa (R) and BabyBERTa+ (B+) on the SuperGLUE benchmark. The performances were reported in terms of accuracy, except for MRPC and QQP, where the F1 score was used instead.

Acc.	CR_C	LC_C	MV_C	RP_C	SC_C	CR_LC	CR_RTP	MV_LC	MV_RTP	SC_LC	SC_RP
R	84.1	100	99.4	93.5	96.4	67.7	68.6	66.7	68.6	84.2	65.7
B+	85.6	100.0	98.7	96.2	87.3	66.3	66.7	66.6	66.9	67.4	64.2

Table 4: Comparison of RoBERTa (R) and BabyBERTa+ (B+) on the MSGS benchmark. The performances were reported in terms of accuracy.

unmasking removal policy proposed in Huebner et al. (2021) takes a different approach by removing the prediction for unchanged tokens. In this case, 90% of the corrupted tokens are masked with the "<mask>", and the remaining are replaced with random tokens. We utilize the same masking policy when pre-training BabyBERTa+.

3 BabyBERTa+ on downstream tasks of BabyLM challenge

In this section, we evaluate our pre-trained models on the tasks in BabyLM challenges of the strict-small tracks. There are three different evaluation benchmarks: BLiMP test suites (Warstadt et al., 2020a), SuperGLUE (Wang et al., 2019) and Mixed Signals Generalization Set (MSGS) (Warstadt et al., 2020b). The BLiMP test suite evaluates the ability of language models to handle grammar. MSGS is a syntactic dataset to test the inductive bias for downstream tasks. SuperGLUE is a standard benchmark to evaluate the capabilities of the pre-trained language models on natural language understanding downstream tasks. We presented the detailed performance of predicting grammatical phenomena in Table 2 and downstream tasks in Table 3 and 4. We use the default hyperparameters as defined in (Warstadt et al., 2023) to fine-tune our system on BabyLM challenges. Our model gets 6% improvement on BLiMP test suite compared to the baseline RoBERTa (69.86% vs 63.02%). The average score on all SuperGLUE tasks in Table 3 is 69.50% while the performance of the baseline model is 71.42%. The average score on MSGS is 78.72% while the RoBERTa-base's score is 81.35%.[4]

[4]The RoBERTa's results are provided in the BabyLM challenge.

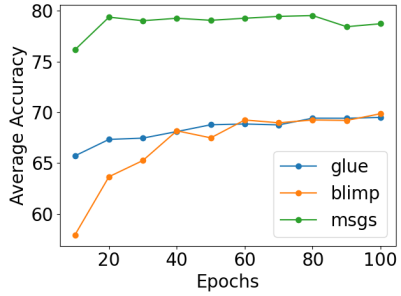

Figure 1: Average accuracy of BabyBERTa+ on three tasks versus the number of pre-training epochs.

We additionally plot the average accuracy on grammaticality tests and downstream tasks versus the number of pre-training epochs in Figure 1. We observe that when continually pre-training with more epochs, both grammatical phenomena prediction and SuperGLUE downstream task performance improve.

4 Conclusion

In this study, we propose a model named BabyBERTa+ by adapting BabyBERTa (Huebner et al., 2021) for the BabyLM challenge (Warstadt et al., 2023) on strict-small tasks and demonstrate the effectiveness of pre-training a smaller model in learning grammatical phenomena compared to RoBERTa (Liu et al., 2019) and other baselines. However, while our model exhibits promising results in learning grammatical features, its performance on downstream tasks remains lower than larger models like RoBERTa. In the future, we aim to explore the impact of the various pre-training factors when pre-training the small model on a limited size of child-directed data corpora and enhance the small model's performance on downstream tasks.

Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080. It was also supported by Contracts FA8750-19-2-0201 and FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA) as well as by grants from the Israeli Ministry of Innovation, Science & Technology (#000519) and the BGU/Philadelphia Academic Bridge (The Sutnick/Zipkin Endowment Fund). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Army Research Office, the Department of Defense or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This research was also supported by a gift from AWS AI for research in Trustworthy AI.

References

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. 2020b. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.