Exploring smaller batch sizes for a high-performing BabyLM model architecture

Sharid Loáiciga, Eleni Fysikoudi and Asad Sayeed

Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg

CLASP centre for linguistic theory and studies in probability

Vetenskapsrådet
Swedish Research Council

Marianne och Marcus Wallenbergs Stiftelse

tl;dr

We study **ELC-BERT**—the 2023 BabyLM winner—under **constrained compute**, focusing on **batch size**. Removing the original system's long training time and large batches largely eliminates its advantage on fine-grained grammaticality **BLiMP**, but some **smaller effective batches** remain competitive on **GLUE** and **MSGS**.

Research Questions

1. How does **batch size** affect ELC-BERT under compute constraints?
2. Which **hyperparameter settings** remain competitive *without* the long-training advantage?

Experimental Setup

- `strict-small` track with 2023 evaluation tools.
- A100 GPUs for pre-training; RTX 3090s for fine-tuning.
- Pre-training hyperparameters follow Charpentier & Samuel (2023); default BabyLM evaluation hyperparameters for fine-tuning.
- We vary **batch size**, **training steps**, and **gradient accumulation**.

Results

Pre-training							Fine-tuning	
Batch size	**Training steps**	**Grad. accu.**	**Epochs**	**Time**	**BLiMP**	**BLiMP suppl.**	**GLUE**	**MSGS**
Original 8096	31250	1	>2000	–	80.00	67.00	73.7	29.4
32	15625	1	4	21m	51.03	47.08	55.93	46.94
32	31250	1	7	44m	50.18	46.89	57.89	43.67
32	15625	12	41	2h39m	50.53	50.70	63.20	43.71
256	15625	1	27	1h7m	44.85	50.59	63.23	43.62
256	31250	1	53	19h57m	50.37	47.07	65.46	39.62
256	125000	1	218	8h31m	44.85	50.59	65.46	39.62
256	250000	1	437	17h4m	44.17	49.49	65.46	39.62
256	15625	12	333	5d10h5m	47.72	49.41	63.66	39.31
512	15625	1	55	1h49m	50.04	46.94	62.38	43.17
512	31250	1	109	3h37m	52.22	45.65	63.80	43.15
253	31250	32	1479	5d22h29m	46.95	49.88	63.72	39.31
506	31250	16	1736	3d18h42m	49.03	49.36	63.64	39.31

Averaged accuracies of ELC-BERT (Charpentier & Samuel, 2023) re-runs with varying batch sizes.

Findings, conclusions and future work

- **Performance trade-offs:** **BLiMP** accuracy drops under constrained runs, while **GLUE** and **MSGS** remain relatively stable.
- **Efficiency:** Smaller effective batches (e.g., 32 × accum 12; effective 384) offer competitive results with far shorter training times.
- ELC-BERT's headline gains appear **compute-sensitive**; under limited constraints, hyperparameters matter.
- **Future work:** exploring learning-rate adjustments under small batches and longer-but-feasible training schedules.

Acknowledgements

Supported by the Swedish Research Council (VR 2014-39) for CLASP and the Marianne and Marcus Wallenberg Foundation grant 2019.0214 (GRIPES). Compute via NAISS (grant 2022-06725). Thanks to ELC-BERT developers and BabyLM organizers for discussions.

{sharid.loaiciga,asad.sayeed}@gu.se,gusfysel@student.gu.se