Sample-Efficient Language Modeling with Linear Attention and Lightweight Enhancements

Patrick Haller, Jonas Golde, Alan Akbik
Humboldt Universität zu Berlin

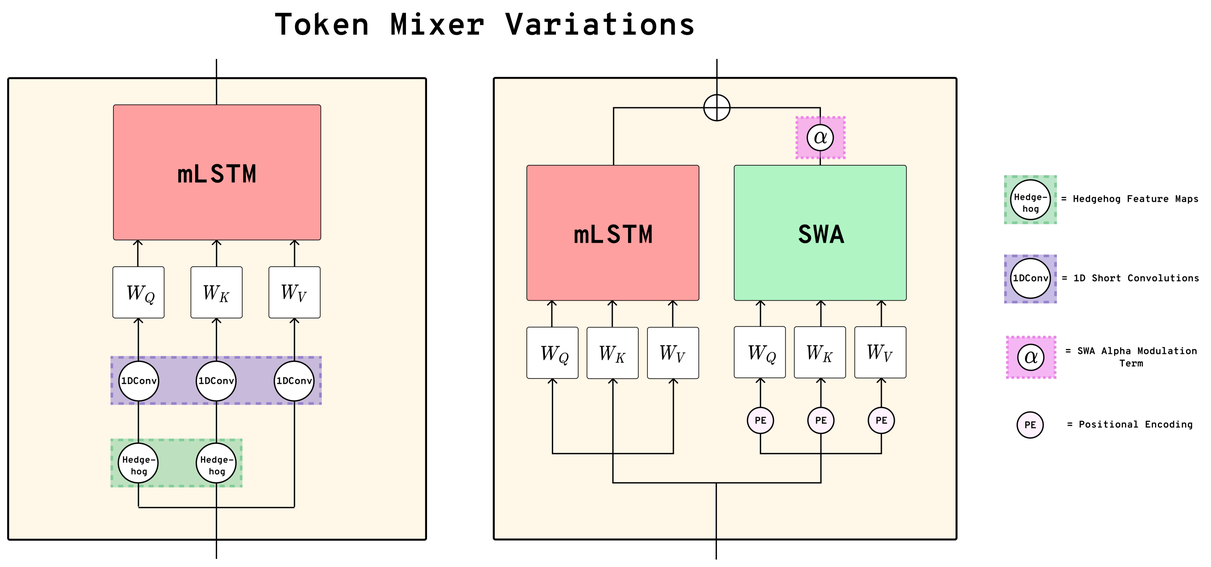
Improving Subquadratic Language Models under a Tiny Data Budget.

Why this matters: BabyLM

- Strict budgets: 10M or 100M words, **10 epochs**, small models.
- We study **sample-efficient** training without scaling: replace self-attention with **mLSTM** and add *lightweight* tweaks.
- Goal: Practical wins on zero-shot linguistic/educational tasks under tight compute.

Architecture	STRICT-SMALL	STRICT
Transformer	32.27	35.03
mLSTM	35.96	35.42

Architecture



$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = conv(\mathbf{Q}), conv(\mathbf{K}), conv\mathbf{V}) \quad \mathbf{Q}, \mathbf{K} = hedgehog(\mathbf{Q}), hedgehog(\mathbf{Q})$$

$$h_{total} = h_{LA} + tanh(\alpha) \cdot h_{SWA}$$

BLaLM: Baby Linear Attention Language Model

Modications and their use:

- ShortConv[1]: Local Pattern Extraction
- Sliding-Window Attention (SWA)[2]: Short-range token interaction
- DynMod: Learns when to rely on mLSTM vs. SWA
- HedgehogMaps[3]: Mimic Softmax-Attention

	STRICT-SMALL		STRICT	
MECHANISM	PPL.	AVG.	PPL.	AVG.
BLaLM	20.01	**37.27**	7.95	35.08
- *ShortConv*	12.37	36.41	6.48	34.57
- *SWA*	12.08	36.16	7.38	35.86
- *SWA with Memory*	10.08	34.96	6.67	37.21
- *SWA DynMod*	9.44	36.15	7.76	**38.82**
- *SWA DynMod Bounded*	8.58	34.41	6.84	36.21
- *Hedgehog*	6.18	33.58	6.68	36.65
- *Hedgehog + SWA*	7.27	36.25	6.63	34.20

Layerwise learned weighting of SWA and mLSTM improves performance.

Optimizer Choice

OPTIMIZER	PPL	AVG.
AdamW	11.21 ±0.11	35.75 ±1.74
Muon	7.95 ±0.15	36.24 ±1.16

Muon[4]: Scale invariant, norm-preserving updates
→ Optimizer consistently improves perplexity and zero-shot performance.

Takeaways

- **Linear-time token mixers (mLSTM)** are a *viable drop-in* for sample-efficient training.
- **Local attention (SWA) + Dynamic Modulation** improves downstream generalization, especially at 100M.
- **Muon** stabilizes and accelerates training in low-data regimes.

References & Acknowledgments: Alan Akbik and Patrick Haller are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant "Eidetic Representations of Natural Language" (project number 448414230). Alan Akbik is furthermore supported under Germany's Excellence Strategy "Science of Intelligence" (EXC 2002/1, project number 390523135). Jonas Golde is supported by the Bundesministerium für Bildung und Forschung (BMBF) as part of the project "FewTuRe" (project number 01IS24020).**References:** [] Short-Long Convolutions Help Hardware-Efficient Linear Attention to Focus on Long Sequences by Zicheng Liu et al., 2024, [] LoLCATs: On Low-Rank Linearizing of Large Language Models (Zhang et al., 2024/25), [] The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry (Zhang, Bhatia, Kumbong, Ré, 2024), LoLCATs: On Low-Rank Linearizing of Large Language Models (Zhang et al., 2024/25)