SMLab short talk on

NeoBERT

by
ADHILSHA ANSAD

06. 03. 2025

Stand on the shoulders of GIANTS

arXiv:2502.19587v1 [cs.CL] 26 Feb 2025

NeoBERT: A Next-Generation BERT

Lola Le Breton[1,2,3] **Quentin Fournier**[2] **Mariam El Mezouar**[4] **Sarath Chandar**[1,2,3,5]

[1]Chandar Research Lab [2]Mila – Quebec AI Institute [3]Polytechnique Montréal
[4]Royal Military College of Canada [5]Canada CIFAR AI Chair

PREREQUISITES & MOTIVATION

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Masked Language Model (MLM) objective and Next Sentence Prediction (NSP) objective
- Produces different <u>*semantic*</u> meanings for same word by context
- ***Motivation***: Recent innovations in architecture, pre-training, and fine-tuning

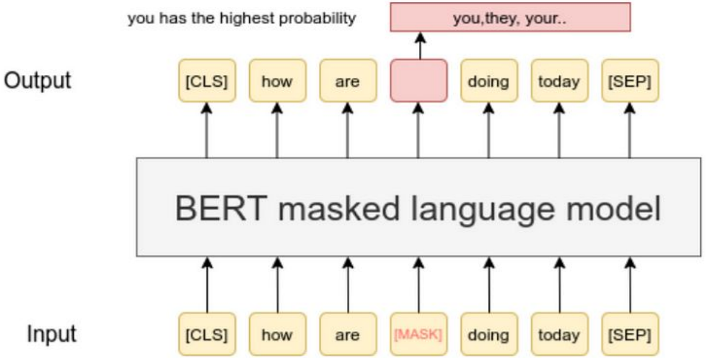

Fig. 1: MLM objective

Fig. 2: NSP objective

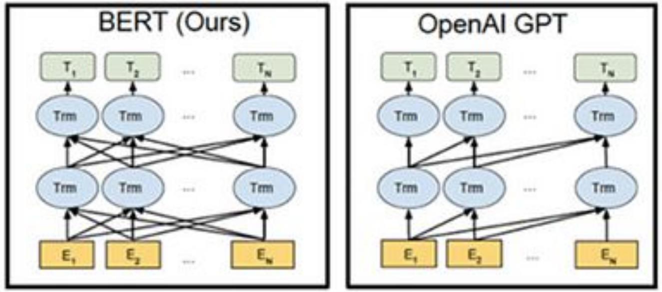

Fig. 3: Bidirectional BERT

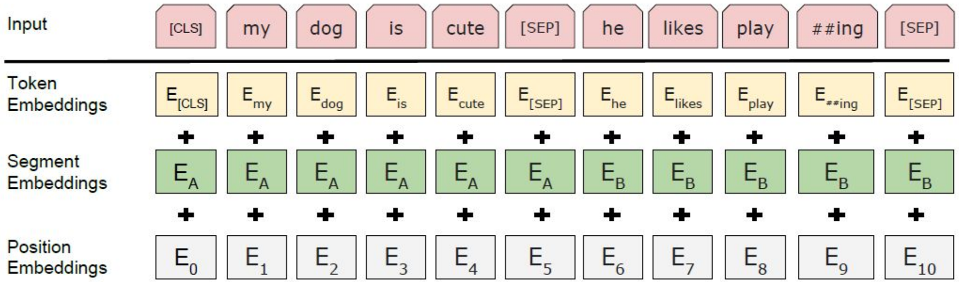

Fig. 4: Input embeddings

RELATED WORKS

Table 1: Comparison of Model Architectures, Training Data, and Pre-Training Configurations.

	BERT		RoBERTa		NomicBERT	ModernBERT		NeoBERT
	base	*large*	*base*	*large*	*base*	*base*	*large*	*medium*
Layers	12	24	12	24	12	22	28	28
Hidden Size	768	$1,024$	768	$1,024$	768	768	$1,024$	768
Attention Heads	12	16	12	16	12	12	16	12
Parameters	120M	350M	125M	355M	137M	149M	395M	250M
Activation Function	GeLU				SwiGLU	GeGLU		SwiGLU
Positional Encoding	Positional Embeddings				RoPE	RoPE		RoPE
Normalization	Post-LayerNorm				Post-LayerNorm	Pre-LayerNorm		Pre-RMSNorm
Data Sources	BooksCorpus Wikipedia		BooksCorpus OpenWebText Stories / CC-News		BooksCorpus Wikipedia	Undisclosed		RefinedWeb
Dataset Size	13GB		160GB		13GB	-		2.8TB
Dataset Year	2019		2019		2023	-		2023
Tokenizer Level	Character		Byte		Character	Character		Character
Vocabulary Size	30K		50K		30K	50K		30K
Sequence Length	512		512		$2,048$	$1,024 \rightarrow 8,192$		$1,024 \rightarrow 4,096$
Objective	MLM + NSP		MLM		MLM	MLM		MLM
Masking Rate	15%		15%		30%	30%		20%
Masking Scheme	80/10/10		80/10/10		-	-		100
Optimizer	Adam		Adam		AdamW	StableAdamW		AdamW
Scheduler	-		-		-	WSD		CosineDecay
Batch Size	131k tokens		131k		8M	448k to 5M		2M
Tokens Seen	131B		131B		-	$\sim$ 2T		2.1T
Training	DDP		DDP		DeepSpeed FlashAttention	Alternate Attention Unpadding FlashAttention		DeepSpeed FlashAttention

IMPORTANT DESIGN ASPECTS

Optimal Depth-to-Width ratio

- BERT-like models in a width-inefficiency regime
- increase depth to achieve optimal ratio

RoPE and YaRN Positional Encodings

- well-suited for tasks requiring extended context.

Pre-Layer Normalization

- improves stability, allows for larger learning rates, and accelerates model convergence
- substitute the classical LayerNorm with RMSNorm

SwiGLU Activation Function

- scale the number of hidden units to keep the number of parameters constant. (extra weight matrix in SwiGLU)

TRAINING

- two-stage pre-training (compute-efficient strategy):
  - 1M steps (2T tokens) with sequences of 1, 024 tokens
  - 50k steps (100B tokens), with sequences upto 4, 096 tokens
- Mitigate the distribution shift when filtering for longer sequences, by sampling from different lengths with certain probabilities.
- Pseudo-perplexity:
  - independently masked each token
  - Compute perplexity calculation with cross-entropy losses of these tokens
- Extensive Results on GLUE benchmark and MTEB benchmark

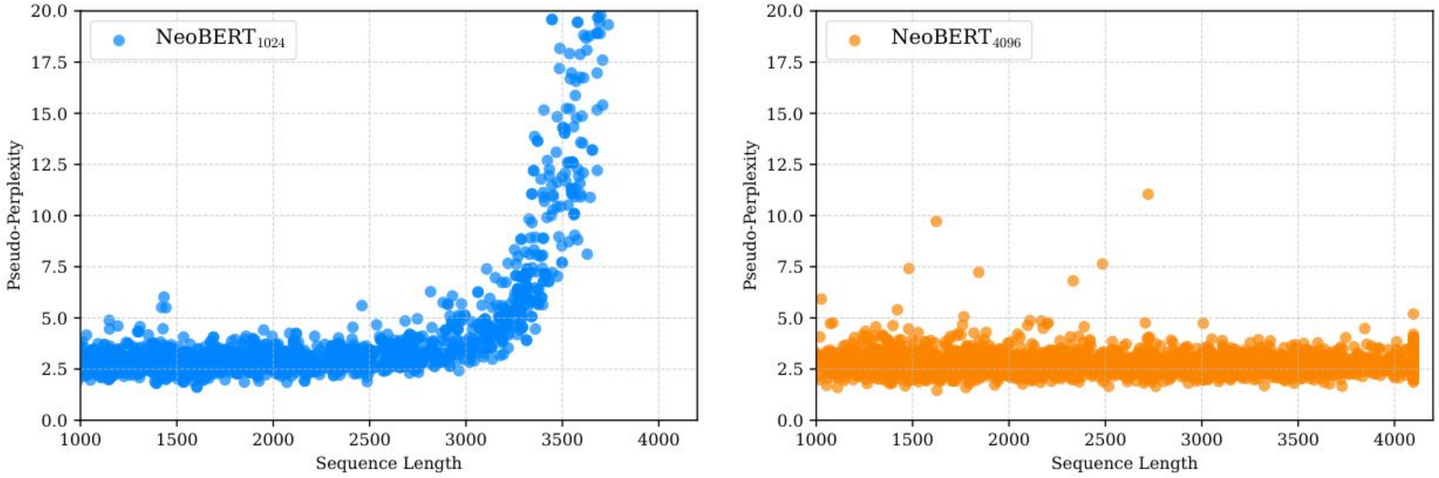

Fig. 5: Stage 1 (1024 tokens) on the left, and stage 2 (upto 4096 tokens) on the right

TRAINING DETAILS

- 8 H100 for 1,050,000 steps, for a total of 6,000 GPU hours
- Fine-tune strategy:
  - cosine similarity and τ = 0.07 as a temperature parameter in the contrastive learning loss

$$\mathcal{L} = -\log \frac{e^{s(q,d^+)/\tau}}{e^{s(q,d^+)/\tau} + \sum_{d^- \in N_q} e^{s(q,d^-)/\tau}}$$

- Sampled datasets with a multinomial distribution based on their sizes, with α = 0.5.

$$\pi = \frac{n_i^\alpha}{\sum_{j=1}^{m} n_j^\alpha}$$

- fine-tune every model for 2,000 steps and evaluate on MTEB in float16.

CONCLUSIONS & LIMITATIONS

GLUE benchmark

- NeoBERT outperforms BERT(L) and NomicBERT
- comparable with RoBERTa(L) - 100M params lesser and supporting 8x longer sequences

MTEB benchmark

- isolating the effects of pre-training and fine-tuning
- NeoBERT consistently outperforms all baselines

Biases

- inherits the biases and limitations of its pre-training data.
- retraining will likely be needed once newer, larger, and more diverse datasets become available

Open Source!

- all code, data, model checkpoints, and training scripts available.
- fully open-source model

REFERENCES

Papers

- Breton, L. L., Fournier, Q., Mezouar, M. E., & Chandar, S. (2025). **NeoBERT: A Next-Generation BERT**. *arXiv preprint* arXiv:2502.19587.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). **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, volume 1 (long and short papers) (pp. 4171-4186).

Figures

- [Fig. 1] - Sentence Transformers. *MLM.*
- [Fig. 2] - Qiu, X., Liao, S., Xie, J., Nie, J.. (2022). *Tapping the Potential of Coherence and Syntactic Features in Neural Models for Automatic Essay Scoring.* 10.48550/arXiv.2211.13373.
- [Fig. 3] - *AraBERT transformer model for Arabic comments and reviews analysis - Scientific Figure on ResearchGate.* Available from: https://www.researchgate.net/figure/Differences-in-pre-training-model-architectures_fig1_357796027
- [Fig. 4] - OpenGenus IQ | Adith Narein T. *Embeddings in BERT.*

Model & Codes

- https://github.com/chandar-lab/NeoBERT
- https://huggingface.co/chandar-lab/NeoBERT

THANK YOU…