Megalodon: Efficient LLM Pretraining & Inference with Unlimited Context Length

Xuezhe Ma (Max)

collaborators at

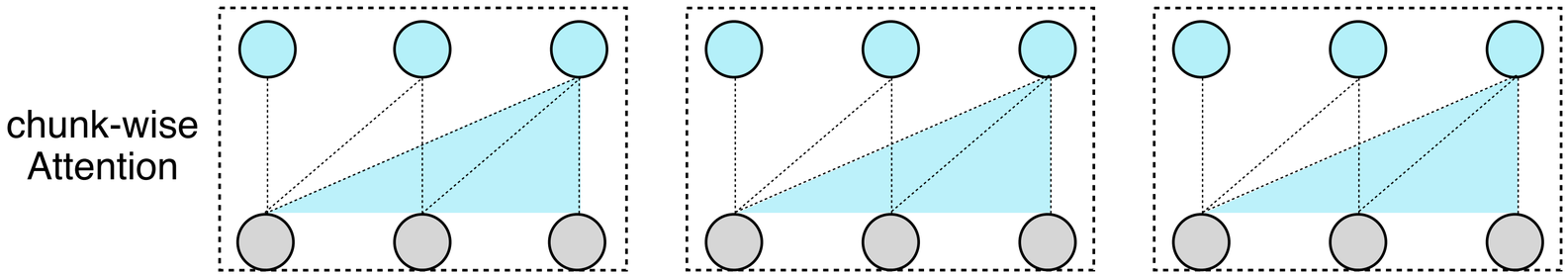
1. USC-ISI 2. Meta AI 3. CMU 4. UCSD 5. SJTU

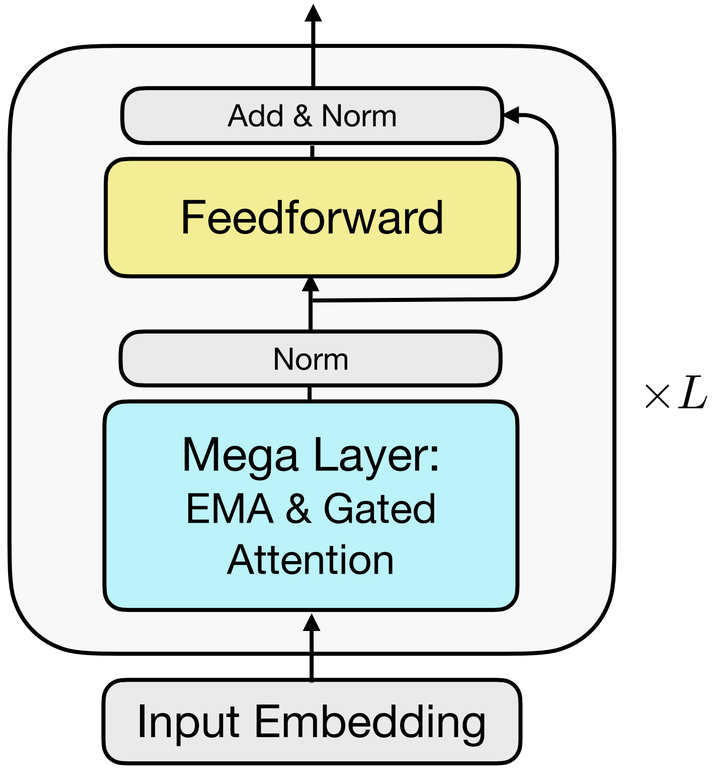
Carnegie Mellon University
Language Technologies Institute

上海交通大學
Shanghai Jiao Tong University

Mega: ICLR 2023

- **Effective and efficient drop-in replacement of attention for long sequence modeling**
  - Outstanding results on various types of data
    - text, images and audios
  - Exponential Moving Average (EMA)
  - Mega-chunk: linear complexity of time and space





Limitations of Mega

- **Mode Capacity vs. Full Attention**
  - Limited capacity of EMA sub-layer
    - Mega-chunk fails behind Mega w. full attention

Limitations of Mega

- **Mode Capacity vs. Full Attention**
  - Limited capacity of EMA sub-layer
    - Mega-chunk fails behind Mega w. full attention
- **Architectural Divergence on Difference Data Modalities**
  - Not only for Mega, but for almost all architectures
  - Different normalization layers on different positions
    - Batch Norm vs. Layer Norm vs. RMS Norm vs Scale Norm vs. …
    - Pre-Norm vs. Post-Norm vs. QK-Norm vs. …
  - Different attention functions
    - Softmax vs. ReLU2 vs. Laplace vs …
  - …

Limitations of Mega

- **Mode Capacity vs. Full Attention**
  - Limited capacity of EMA sub-layer
    - Mega-chunk fails behind Mega w. full attention
- **Architectural Divergence on Difference Data Modalities**
  - Not only for Mega, but for almost all architectures
  - Different normalization layers and positions
    - Batch Norm vs. Layer Norm vs. RMS Norm vs Scale Norm vs. …
    - Pre-Norm vs. Post-Norm vs. QK-Norm vs. …
  - Different attention functions
    - Softmax vs. ReLU2 vs. Laplace vs. …
  - …
- **No Evidences or Results for Mega's Scalability**
  - Large-scale pretraining with Mega?

Megalodon: An Improved Version of Mega

- **Complex Exponential Moving Average**
  - Extending EMA to the Complex Field
- **Timestep Normalization**
  - Auto-regressive group normalization across sequential dimension
- **Numerical Stability**
  - Normalized Attention
  - Pre-Norm w. Top-hop Residual

Megalodon: Results on Small-Scale Benchmarks

	LRA	SC	ImageNet	Enwiki8	WT103
S4	85.86	97.50	—	—	20.95
XFM	59.24	xx	81.8	1.08	18.66
Mega-chunk	85.66	96.92	—	1.02	18.07
Mega	88.21	—	82.35	—	—
Megalodon-chunk	**87.62**	**98.14**	**—**	**1.00**	**17.23**
Megalodon	**88.63**	**—**	**83.12**	**—**	**—**

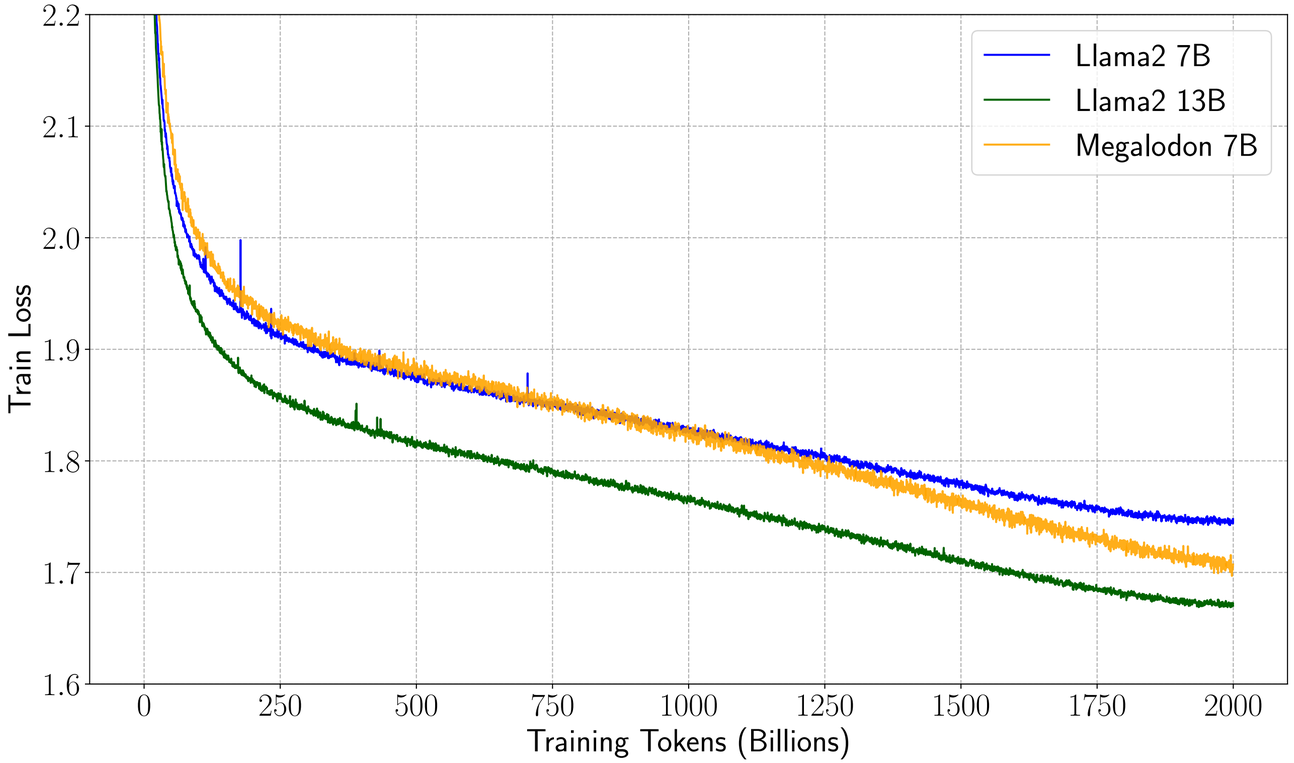
Importantly! All the experiments were conducted using exactly the same architecture!

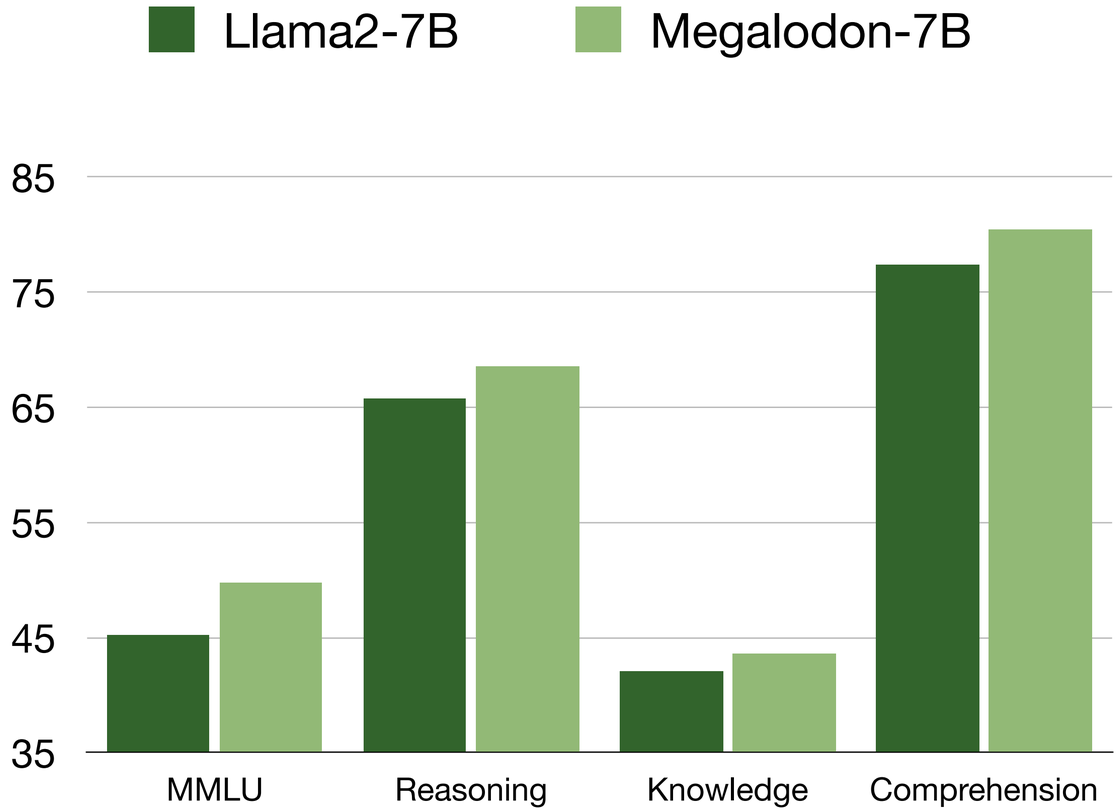
Megalodon on LLM Pretraining

- **Pretraining Data**
  - 2 trillion tokens
  - Exactly the same with Llama2 for head-to-head comparison
- **Architecture Hyperparameters**
  - Closely following Llama2
    - 7B parameters
    - 32 blocks, model dimension 4096
    - Rotary positional embedding (RoPE)
  - Differences
    - 4 attention heads (32 in Llama2)
    - 32K context length w. 4K attention chunk size (4K full attention in Llama2)

Efficiency of Megalodon-7B

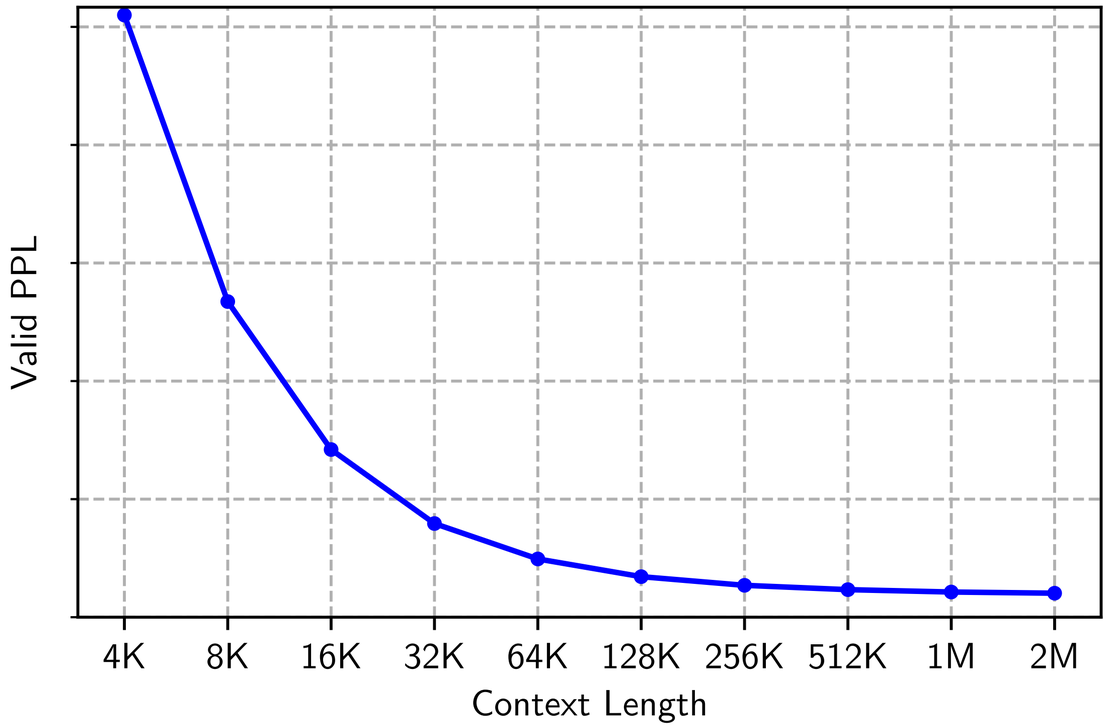
- **Data Efficiency**





Long-Context Modeling

- **Long-Context Evaluation of Megalodon-7B**
  - Perplexity w. various context lengths
  - Long-context QA tasks in Scrolls



Model	NaQA	Qasper	QMSum
Xgen	17.4	20.5	6.8
MPT	18.8	24.7	8.8
Yarn	20.9	26.2	11.4
LLAMA2	18.8	19.8	10.1
LLAMA2-L*	23.5	**28.3**	**14.5**
MEGALODON	**23.9**	28.0	13.1

Thanks!
Q&A

Code: https://github.com/XuezheMax/megalodon