Longformer: The Long-Document Transformer

Iz Beltagy* **Matthew E. Peters*** **Arman Cohan***
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,matthewp,armanc}@allenai.org

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

Published at NeurIPS 2020

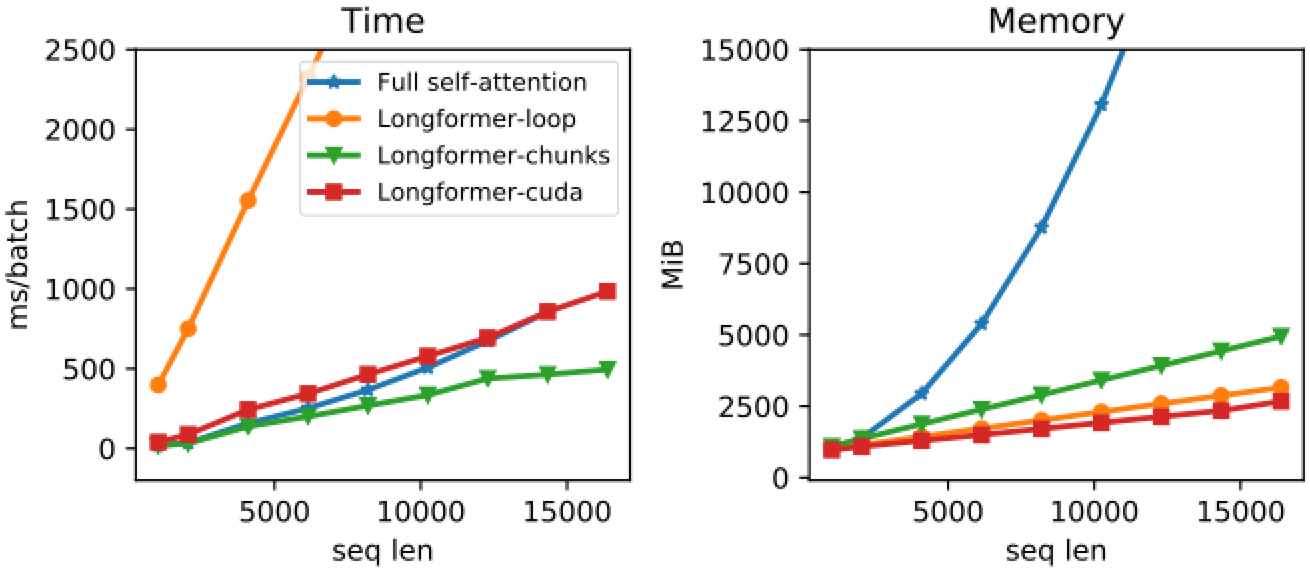
Rethinking Attention with Performers

Krzysztof Choromanski*[1], Valerii Likhosherstov*[2], David Dohan*[1], Xingyou Song*[1]
Andreea Gane*[1], Tamas Sarlos*[1], Peter Hawkins*[1], Jared Davis*[3], Afroz Mohiuddin[1]
Lukasz Kaiser[1], David Belanger[1], Lucy Colwell[1,2], Adrian Weller[2,4]
[1]Google [2]University of Cambridge [3]DeepMind [4]Alan Turing Institute

Published at ICLR 2021

Long Context: Challenges

- Computation and Memory $\propto n^2$

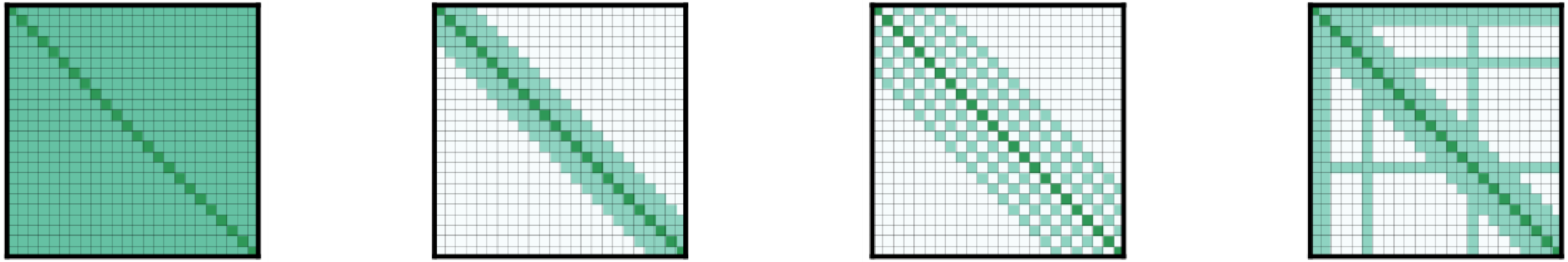


Long Context: Strategies

- Memory based: processes the document in chunks moving from left-to-right
  - LSTM, Transformer-XL
- Sequence based:
  - Sparse Attention: avoid computing the full quadratic attention matrix multiplication
    - Longformer, Big Bird
  - Approximate Attention:
    - Performer

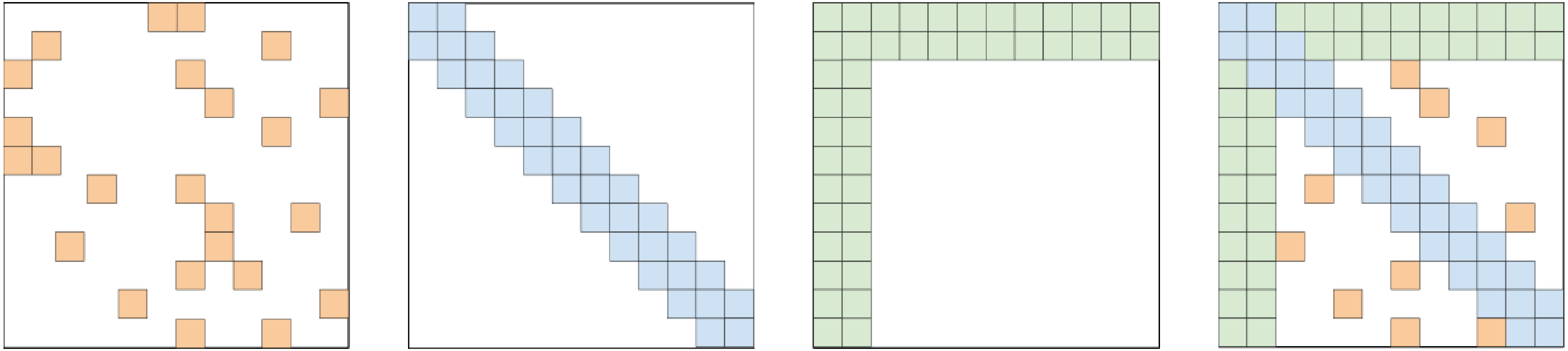
Longformer Sparse Attention

- Sliding Window: the importance of local context
- Dilated Sliding Window: increase the receptive field without increasing computation
- Global Attention: end task motivated, encodes inductive bias about the task.
- Linear Projections for Global Attention: additional projections provide flexibility to model the different types of attention



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Big Bird Sparse Attention

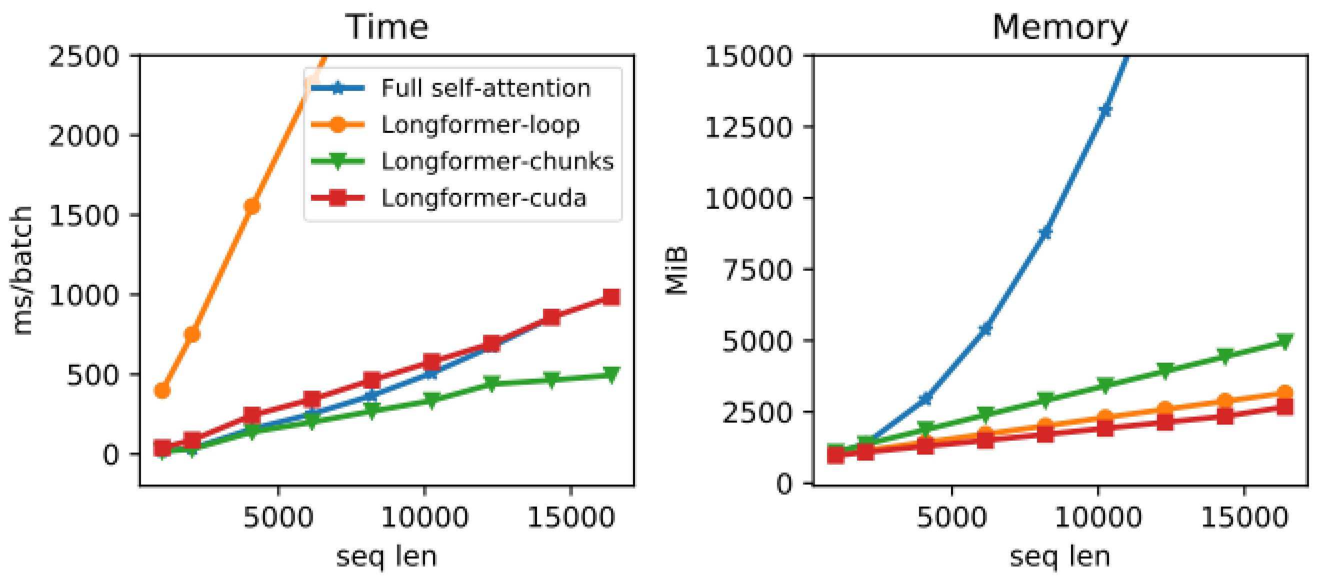


(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

- Prooved to have the same expressiveness, and prooved to be Turing Complete

Longformer Implementation

- Longformer-loop: computes each diagonal separately in a loop
- Longformer-chunks: chunks Q and K into overlapping blocks of size w and overlap of size $\frac{1}{2}w$, multiplies the blocks, then mask out the diagonals.
- Longformer-cuda: a custom CUDA kernel that we implement using TVM (Tensor Virtual Machine, a deep learning compiler)



Performer

Find feature map (kernel function) $\phi(x)$, such that:

$$A = \exp\left(QK^T\right) \approx \phi(Q)\phi(K)^T = Q'\left(K'\right)^T$$

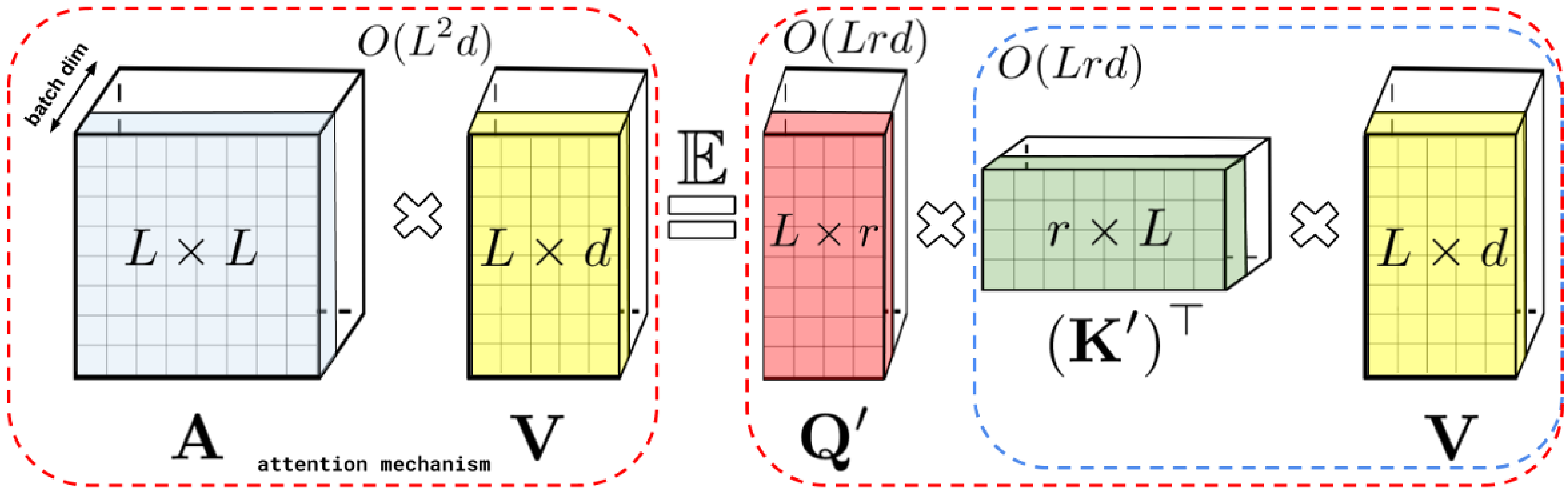


Figure 1: Approximation of the regular attention mechanism $\mathbf{AV}$ (before $\mathbf{D}^{-1}$-renormalization) via (random) feature maps. Dashed-blocks indicate order of computation with corresponding time complexities attached.

Performer Timing

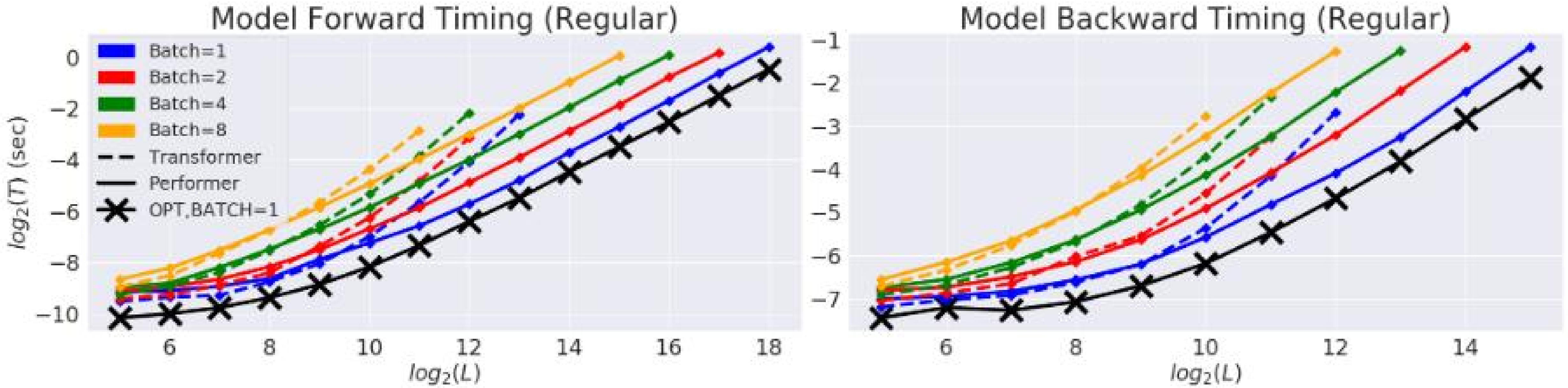


Figure 3: Comparison of Transformer and Performer in terms of forward and backward pass speed and maximum L allowed. "X" (OPT) denotes the maximum possible speedup achievable, when attention simply returns the $\mathbf{V}$-matrix. Plots shown up to when a model produces an out of memory error on a V100 GPU with 16GB. Vocabulary size used was 256. Best in color.

Longformer Experiments

- Character-level Autoregressive Language Modeling
- Longformer Finetuning:
  - Question answering
  - Coreference Resolution
  - Text classification
- Longformer-Encoder-Decoder (LED):
  - Summarization

Character-level Autoregressive Language Modeling

- window size: balance between efficiency and performance
  - small window sizes for the lower layers and increase window sizes as we move to higher layers
- dilated sliding windows:
  - For lower layers, do not use dilated sliding windows: maximize their capacity to learn local context
  - For the higher layers, use a small amount of increasing dilation only on 2 heads.
- Staged training procedure: on each phase, double the window size and the sequence length, and halve the learning rate.
  - The model needs a large number of gradient updates to learn the local context first, before learning to utilize longer context
- Dataset: text8, enwik8

Param	Value
Position Embeddings	Relative and Sinusoidal as in Dai et al. (2019)
Small model config	12 layers, 8 heads, 512 hidden size as in Dai et al. (2019)
Large model config	30 layers, 8 heads, 512 hidden size as in Child et al. (2019)
Optimizer	AdamW
Dropout	0.2 (small model), 0.4 (large model)
Gradient clipping	0.25
Weight Decay	0.01
Layernorm Location	pre-layernorm (Xiong et al., 2020)
Activation	GeLU
Number of phases	5
Phase 1 window sizes	32 (bottom layer) - 8,192 (top layer)
Phase 5 window sizes	512 (bottom layer) - (top layer)
Phase 1 sequence length	2,048
Phase 5 sequence length	23,040 (gpu memory limit)
Phase 1 LR	0.00025
Phase 5 LR	000015625
Batch size per phase	32, 32, 16, 16, 16
#Steps per phase (small)	430K, 50k, 50k, 35k, 5k
#Steps per phase (large)	350K, 25k, 10k, 5k, 5k
Warmup	10% of the phase steps with maximum 10K steps
LR scheduler	constant throughout each phase
Dilation (small model)	0 (layers 0-5), 1 (layers 6-7), 2 (layers 8-9), 3 (layers 10-11)
Dilation (large model)	0 (layers 0-14), 1 (layers 15-19), 2 (layers 20-24), 3 (layers 25-29)
Dilation heads	2 heads only

Evaluation

(follow Transformer-XL)

- metric: BPC (bit per character)
- split the dataset into overlapping sequences of size 32,256 with a step of size 512, and report the performance on the last 512 tokens on the sequence.

Ablation

Model	Dev BPC
Decreasing w (from 512 to 32)	1.24
Fixed w (= 230)	1.23
Increasing w (from 32 to 512)	**1.21**
No Dilation	1.21
Dilation on 2 heads	**1.20**

Longformer Experiments

- Character-level Autoregressive Language Modeling
- Longformer Finetuning:
  - Question answering
  - Coreference Resolution
  - Text classification
- Longformer-Encoder-Decoder (LED):
  - Summarization

Further Pretraining and Finetuning

- continue MLM pretraining from the RoBERTa
- Attention Pattern:
  - window size: 512 (same amount of computation as RoBERTa)
  - dilation hurt performance: not compatible with the pretrained RoBERTa weights
- Position Embeddings: leverage RoBERTa's pretrained weights
  - copy the 512 position embeddings from RoBERTa multiple times, support up to position 4096

Question answering: WikiHop

WikiHop: a question, answer candidates (2~79 candidates), supporting contexts (3~63 paragraphs)

```
[q] question [/q] [ent] candidate1 [/ent] ... [ent] candidateN [/ent]
</s> context1 </s> ... </s> contextM </s>
```

- global attention on the entire question and answer candidate sequence
- attach a linear layer to each `[ent]`

Question answering: TriviaQA

TriviaQA: 100K question, answer, document triplets. Documents are Wikipedia articles, and answers are named entities mentioned in the article.

```
[s] question [/s] document [/s]
```

- truncate the document at 4096 wordpiece to avoid it being very slow
- global attention on all question tokens
- add one layer that predicts the beginning and end of the answer span

Question answering: HotpotQA

HotpotQA: answering questions from a set of 10 paragraphs from 10 different Wikipedia articles where 2 paragraphs are relevant to the question and the rest are distractors.

```
[CLS] [q] question [/q] <t> title1 </t> sent1,1 [s] sent1,2 [s] ... <t> title2 </t> sent2,1 [s] sent2,2[s] ...
```

two-stage Longformer model:

1. identify relevant paragraphs
2. find the final answer span and evidence

Text classification

Datasets:

- IMDB: sentiment classification datasets consisting of movie reviews (only 13.6% of them are larger than 512 wordpieces)
- Hyperpartisan news detection: 645 long documents

Method:

- addition of global attention to `[CLS]`
- binary cross entropy loss on top of a first `[CLS]` token

Ablations on WikiHop

Model	Accuracy / Δ
Longformer (seqlen: 4,096)	73.8
RoBERTa-base (seqlen: 512)	72.4 / -1.4
Longformer (seqlen: 4,096, 15 epochs)	75.0 / +1.2
Longformer (seqlen: 512, attention: n^2)	71.7 / -2.1
Longformer (seqlen: 2,048)	73.1 / -0.7
Longformer (no MLM pretraining)	73.2 / -0.6
Longformer (no linear proj.)	72.2 / -1.6
Longformer (no linear proj. no global atten.)	65.5 / -8.3
Longformer (pretrain extra position embed. only)	73.5 / -0.3

- performance gains are not due to additional pretraining

Longformer Experiments

- Character-level Autoregressive Language Modeling
- Longformer Finetuning:
  - Question answering
  - Coreference Resolution
  - Text classification
- Longformer-Encoder-Decoder (LED):
  - Summarization

Longformer-Encoder-Decoder (LED)

- Encoder: Longformer local+global attention
  - window size 1024, global attention on the first `<s>` token
- Decoder: full self-attention and cross-attention
- initialize LED parameters from the BART
- Position Embeddings: leverage BART's pretrained weights
  - copy the 1K position embeddings from BART multiple times, support up to position 16K

Summarization

arXiv summarization dataset: summarization in the scientific domain, 90th percentile of document lengths is 14.5K tokens

- Training: teacher forcing on gold training summaries
- Inference: beam search