Contents

Longformer: The Long-Document Transformer

Paper Link: https://arxiv.org/abs/2004.05150

Authors: Iz Beltagy, Matthew E. Peters, Arman Cohan (Allen Institute for Artificial Intelligence)

Publication: ArXiv 2020

Code: https://github.com/allenai/longformer

Overview

Longformer addresses the fundamental limitation of Transformer models in processing long sequences by introducing a sparse attention mechanism that scales linearly with sequence length. Unlike standard self-attention's quadratic complexity, Longformer can efficiently process documents with thousands of tokens, making it practical for long-document understanding tasks.

Key Problem Addressed

Standard Transformer models face critical limitations with long sequences:

1. **Quadratic Complexity**: Self-attention scales as $O(n^2)$ with sequence length
2. **Memory Constraints**: Cannot process documents longer than 512-4096 tokens
3. **Computational Inefficiency**: Prohibitive costs for long document processing
4. **Limited Context**: Important information may be lost due to truncation

Core Innovation: Sliding Window Attention

Longformer introduces a **sliding window attention** mechanism that combines local and global attention patterns:

1. Sliding Window Attention

- Each token attends to `w/2` tokens on each side (total window size `w`)
- Typically w = 512 for most tasks
- Provides local context while maintaining linear complexity
- Enables information flow through multiple attention layers

2. Global Attention

- Specific tokens attend to all positions in the sequence
- All tokens can attend to these global tokens
- Task-specific: different tokens can be designated as global
- Examples: [CLS] token for classification, question tokens for QA

3. Dilated Sliding Window

- Some layers use dilated windows (gaps between attended positions)
- Increases receptive field without additional computation
- Similar to dilated convolutions

Attention Complexity Analysis

Standard Attention

- **Time Complexity**: $O(n^2)$
- **Space Complexity**: $O(n^2)$
- **Scalability**: Prohibitive for long sequences

Longformer Attention

- **Time Complexity**: $O(n \times w)$ where w is window size
- **Space Complexity**: $O(n \times w)$
- **Scalability**: Linear with sequence length

Memory Efficiency

- For n = 4096, w = 512: 8x memory reduction
- Enables processing of much longer documents
- Maintains computational efficiency

Architecture Details

Attention Patterns

1. **Local Windowed**: Each token attends to w neighboring tokens
2. **Global**: Selected tokens attend to entire sequence
3. **Dilated**: Some layers use gaps in attention windows

Implementation

- Drop-in replacement for standard self-attention
- Compatible with existing Transformer architectures

- Supports both encoder-only and encoder-decoder variants

Position Encoding

- Uses learned positional embeddings
- Supports sequences up to maximum trained length
- Can be extended for longer sequences

Experimental Results

Character-Level Language Modeling

- **text8**: Achieved state-of-the-art results
- **enwik8**: New best performance on character prediction
- Demonstrates effective long-range modeling

Reading Comprehension

- **WikiHop**: New state-of-the-art on multi-hop reasoning
- **TriviaQA**: Significant improvements over RoBERTa
- **HotpotQA**: Better performance on complex reasoning tasks

Document Classification

- **IMDb**: Improved sentiment analysis on full reviews
- **Hyperpartisan**: Better political bias detection
- **ArXiv**: Effective scientific paper classification

Long Document QA

- **Natural Questions**: Improved performance on long passages
- **MS MARCO**: Better document retrieval and ranking
- Consistent improvements over truncated approaches

Longformer-Encoder-Decoder (LED)

Architecture

- Encoder: Longformer with sliding window attention
- Decoder: Standard attention (shorter sequences)
- Designed for long-document generation tasks

Applications

- **Document Summarization**: arXiv paper summarization
- **Long-form Generation**: Extended text generation
- **Translation**: Long document translation

Performance

- **arXiv Summarization**: Strong results on scientific papers

- **CNN/DailyMail**: Competitive performance on news summarization
- Enables processing of full-length documents

Comparison with Other Approaches

vs. Big Bird

- **Attention Pattern**: Fixed sliding window vs. random + window + global
- **Theoretical Properties**: Both maintain linear complexity
- **Implementation**: Longformer is simpler to implement
- **Performance**: Comparable on most tasks

vs. Linformer

- **Approach**: Sliding window vs. low-rank projection
- **Theoretical Guarantees**: Different approximation strategies
- **Practical Performance**: Longformer often more effective

vs. Reformer

- **Memory**: Sliding window vs. locality-sensitive hashing
- **Complexity**: Both achieve linear scaling
- **Ease of Use**: Longformer more straightforward

Technical Implementation

Attention Computation

```
# Sliding window attention
attention_scores = compute_attention(query, key, value, window_size=w)

# Global attention (for special tokens)
if token_is_global:
    attention_scores = compute_full_attention(query, key, value)
```

Memory Optimization

- Sparse attention matrix representation
- Efficient implementation using sliding window operations
- Gradient computation optimized for sparse patterns

Practical Applications

1. Long Document Understanding

- Legal document analysis
- Scientific paper comprehension
- Book and novel analysis

2. Information Retrieval

- Document ranking
- Passage retrieval
- Question answering over long texts

3. Content Generation

- Document summarization
- Long-form content creation
- Multi-document synthesis

Limitations and Considerations

1. Window Size Selection

- Fixed window size may not be optimal for all tasks
- Requires task-specific tuning
- Trade-off between context and efficiency

2. Global Token Selection

- Choosing which tokens should be global is task-dependent
- Affects model performance significantly
- Requires domain expertise

3. Position Encoding

- Limited by maximum training length
- Extrapolation to longer sequences may be challenging
- Position bias in very long documents

Impact on the Field

1. Enabling Long-Document NLP

- Made long-document processing practical
- Influenced subsequent efficiency research
- Demonstrated viability of sparse attention

2. Practical Adoption

- Widely used in document understanding systems
- Integrated into popular NLP libraries
- Influenced commercial applications

3. Research Direction

- Inspired further work on efficient attention
- Contributed to understanding of attention patterns
- Influenced design of subsequent models

Key Insights

1. Local Context Sufficiency

- Most language understanding requires primarily local context
- Global attention only needed for specific tokens
- Sliding window captures most important patterns

2. Task-Specific Design

- Different tasks benefit from different attention patterns
- Global attention should be task-motivated
- One size doesn't fit all approaches

3. Simplicity Advantage

- Simple sliding window is often as effective as complex patterns
- Easier to implement and optimize
- More predictable performance characteristics

Why Read This Paper

Longformer is essential reading because:

1. **Practical Solution**: Addresses real limitations of Transformers
2. **Elegant Design**: Simple yet effective attention mechanism
3. **Strong Results**: Consistently outperforms baselines
4. **Broad Applicability**: Works across many long-document tasks
5. **Implementation Clarity**: Clear path to practical deployment

Key Takeaways

1. **Linear Scaling**: Sliding window attention enables linear complexity
2. **Local + Global**: Combining local and global attention is effective
3. **Task-Specific Design**: Global attention should be task-motivated
4. **Practical Impact**: Enables processing of real-world long documents
5. **Simplicity Works**: Simple patterns often outperform complex ones

Longformer represents a practical breakthrough in making Transformers applicable to long-document tasks, demonstrating that careful attention design can overcome fundamental scalability limitations while maintaining strong performance across diverse applications.