Contents

Big Bird: Transformers for Longer Sequences **1**
- Overview . . . 1
- Key Problem Addressed . . . 1
- Core Innovation: Sparse Attention Mechanism . . . 1
- Theoretical Contributions . . . 2
- Architecture Details . . . 2
- Experimental Results . . . 3
- Why Big Bird Works . . . 3
- Technical Insights . . . 4
- Practical Applications . . . 4
- Limitations and Considerations . . . 4
- Impact on the Field . . . 5
- Follow-up Work . . . 5
- Key Takeaways . . . 5
- Why Read This Paper . . . 6

Big Bird: Transformers for Longer Sequences

Paper Link: https://arxiv.org/abs/2007.14062

Authors: Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed (Google Research)

Publication: NeurIPS 2020

Code: https://github.com/google-research/bigbird

Overview

Big Bird addresses one of the fundamental limitations of Transformer models: the quadratic computational and memory complexity of the attention mechanism with respect to sequence length. This paper introduces a sparse attention mechanism that reduces complexity from $O(n^2)$ to $O(n)$ while maintaining the model's expressive power and theoretical properties.

Key Problem Addressed

Standard Transformer attention has several critical limitations:

1. **Quadratic Complexity**: Full attention requires $O(n^2)$ memory and computation
2. **Sequence Length Limitations**: BERT is limited to 512 tokens, limiting applications
3. **Scalability Issues**: Cannot process long documents, genomic sequences, or extended conversations
4. **Hardware Constraints**: Quadratic growth makes longer sequences prohibitively expensive

Core Innovation: Sparse Attention Mechanism

Big Bird introduces a sparse attention pattern that combines three types of attention:

1. Random Attention

- Each token attends to `r` randomly selected tokens
- Provides global connectivity across the sequence
- Enables information flow between distant positions

2. Window Attention

- Each token attends to `w` neighboring tokens (local window)
- Captures local dependencies and patterns
- Similar to convolutional operations

3. Global Attention

- Special global tokens attend to all positions in the sequence
- All tokens can attend to these global tokens
- Includes tokens like [CLS] and [SEP]

Combined Pattern

```
Total attention per token = r (random) + w (window) + g (global)
```

Where typically: - r = 3 random tokens - w = 3 window tokens (each side) - g = 2 global tokens

Theoretical Contributions

1. Universal Approximation

Big Bird proves that their sparse attention mechanism is a **universal approximator of sequence functions**, meaning it can approximate any sequence-to-sequence function to arbitrary accuracy.

2. Turing Completeness

The paper demonstrates that Big Bird attention is **Turing complete**, meaning it can compute any computable function given sufficient resources.

3. Global Token Analysis

Theoretical analysis reveals that having O(1) global tokens provides: - Information aggregation across the entire sequence - Broadcast mechanism for global information - Maintenance of full model expressiveness

Architecture Details

Attention Complexity Analysis

- **Standard Attention**: $O(n^2)$ time and space
- **Big Bird Attention**: O(n) time and space
- **Memory Reduction**: Linear scaling enables 8x longer sequences

Implementation Details

- Based on BERT architecture with sparse attention
- Maintains all other Transformer components (FFN, layer norm, etc.)
- Compatible with existing pre-training and fine-tuning procedures

Experimental Results

Long Document Tasks

Question Answering

- **Natural Questions**: Significant improvement on long-form QA
- **TriviaQA**: Better performance when full context is available
- **HotpotQA**: Improved multi-hop reasoning with longer context

Summarization

- **arXiv papers**: Achieved SOTA on scientific paper summarization
- **PubMed**: Strong performance on biomedical literature
- **CNN/DailyMail**: Competitive results on news summarization

Sequence Length Experiments

- **BERT**: Limited to 512 tokens
- **Big Bird**: Successfully handles 4,096 tokens (8x improvement)
- **Hardware**: Same GPU memory requirements as BERT-512

Genomics Applications

- **DNA Sequences**: Successfully applied to genomic data analysis
- **Protein Folding**: Demonstrated potential for biological sequences
- **Gene Expression**: Showed promise for regulatory sequence analysis

Why Big Bird Works

1. Sparse Connectivity

- Random attention provides global connectivity
- Window attention captures local patterns
- Global tokens aggregate and broadcast information

2. Information Flow

- Multi-hop information propagation through sparse connections
- Efficient routing of information across long sequences
- Maintains expressiveness of full attention

3. Computational Efficiency

- Linear complexity enables longer sequences
- Hardware-friendly sparse operations

- Reduced memory footprint

Technical Insights

1. Random Attention Benefits

- Provides probabilistic global connectivity
- Enables long-range dependency modeling
- Maintains theoretical properties of full attention

2. Global Token Importance

- Act as "hub" nodes in the attention graph
- Aggregate information from entire sequence
- Enable broadcast of global context

3. Window Attention Necessity

- Captures local dependencies crucial for language
- Provides structural inductive bias
- Ensures smooth information flow

Practical Applications

1. Long Document Processing

- Legal document analysis
- Scientific paper comprehension
- Book-length text understanding

2. Genomics and Biology

- DNA sequence analysis
- Protein structure prediction
- Gene regulatory network modeling

3. Conversational AI

- Multi-turn dialogue systems
- Long conversation history
- Context-aware responses

Limitations and Considerations

1. Attention Pattern Design

- Specific random/window/global pattern may not be optimal for all tasks
- Requires careful tuning of hyperparameters
- Pattern selection affects performance

2. Training Complexity

- Sparse attention requires specialized implementation
- More complex than standard attention
- Requires careful optimization

3. Task-Specific Performance

- May not improve all tasks equally
- Some tasks may not benefit from longer sequences
- Requires task-specific evaluation

Impact on the Field

1. Enabling Longer Sequences

- Opened up new application domains
- Made document-level processing feasible
- Influenced subsequent efficiency work

2. Theoretical Understanding

- Provided formal analysis of sparse attention
- Established theoretical foundations
- Influenced theoretical work on attention mechanisms

3. Practical Deployment

- Demonstrated feasibility of efficient attention
- Inspired other sparse attention mechanisms
- Contributed to efficiency research

Follow-up Work

1. Other Sparse Attention Mechanisms

- Longformer (similar sliding window approach)
- Linformer (low-rank attention approximation)
- Performer (kernel-based attention)

2. Efficiency Research

- Sparked interest in efficient transformer variants
- Influenced Flash Attention and other optimizations
- Contributed to scaling law research

Key Takeaways

1. **Sparse Attention Works**: Carefully designed sparse patterns maintain expressiveness
2. **Linear Scaling**: Enables processing of much longer sequences
3. **Theoretical Guarantees**: Sparse attention can preserve theoretical properties

4. **Global Tokens Matter**: O(1) global tokens provide crucial connectivity
5. **Application Diversity**: Benefits extend beyond NLP to genomics and other domains

Why Read This Paper

Big Bird is essential reading because:

1. **Fundamental Limitation**: Addresses core scalability issues of Transformers
2. **Theoretical Rigor**: Provides formal analysis of sparse attention
3. **Practical Impact**: Enables new applications with longer sequences
4. **Methodological Innovation**: Introduces principled approach to attention sparsity
5. **Broad Applications**: Demonstrates versatility across domains

The paper represents a significant advancement in making Transformers more scalable and applicable to real-world problems requiring long-range understanding, influencing subsequent research in efficient attention mechanisms and long-sequence modeling.