Sparse Projection Attention: A Computationally Efficient Framework for Long Sequence Modeling

Mehdi Chrifi Alaoui *, Nour-eddine Joudar , Mohamed Ettaouil

Posted Date: 6 April 2026

doi: 10.20944/preprints202604.0343.v1

Keywords: attention mechanism; sparse projection; computational efficiency; transformers; long sequence modeling; Johnson-Lindenstrauss lemma; randomized algorithms

 doi:10.20944/preprints202604.0343.v1

Article

Sparse Projection Attention: A Computationally Efficient Framework for Long Sequence Modeling

Mehdi Chrifi Alaoui *, Nour-eddine Joudar and Mohamed Ettaouil

Faculty of Sciences and Techniques, University Sidi Mohammed Ben Abdellah, Fez, Morocco
* Correspondence: mehdi.chrifialaoui@usmba.ac.ma

Abstract

The self-attention mechanism has revolutionized sequence modeling but suffers from quadratic computational complexity with respect to sequence length, limiting its applicability to long sequences. We propose **Sparse Projection Attention (SPA)**, a novel attention variant that leverages learnable sparse projections to reduce the effective dimensionality of queries and keys while maintaining expressive power. Our method is grounded in the Johnson-Lindenstrauss lemma and provides theoretical guarantees on distance preservation. We introduce a comprehensive mathematical framework including error bounds, convergence analysis, and gradient dynamics. Experimental results on language modeling, machine translation, and long-range sequence classification demonstrate that SPA achieves up to 8 $\times$ computational speedup while maintaining competitive performance compared to standard attention and other sparse variants. The proposed approach offers an effective trade-off between computational efficiency and model expressivity for long-sequence tasks, making transformers more accessible for resource-constrained environments and real-time applications.

Keywords: attention mechanism; sparse projection; computational efficiency; transformers; long sequence modeling; Johnson-Lindenstrauss lemma; randomized algorithms

1. Introduction

The Transformer architecture (Vaswani et al. 2017) has emerged as the dominant paradigm in modern deep learning, achieving state-of-the-art performance across diverse domains including natural language processing (Devlin et al. 2018; Radford et al. 2019), computer vision (Dosovitskiy et al. 2020; Liu et al. 2021), and speech processing (Gulati et al. 2020). At the heart of this architecture lies the self-attention mechanism, which enables the model to capture global dependencies by computing pairwise interactions between all positions in a sequence.

Despite its remarkable success, the standard self-attention mechanism exhibits quadratic computational complexity $\mathcal{O}(N^2 d_k)$ and memory requirements $\mathcal{O}(N^2)$ with respect to sequence length N, where d_k represents the dimensionality of queries and keys. This computational bottleneck becomes particularly problematic for long sequences encountered in domains such as document understanding (Beltagy et al. 2020), genomic sequence analysis (Ji et al. 2021), high-resolution image processing (Liu et al. 2021), and long-term temporal forecasting (Zhou et al. 2021).

The limitations of standard attention have motivated extensive research into efficient alternatives. Current approaches can be broadly categorized into:

- **Pattern-based Sparse Attention** (Child et al. 2019; Zaheer et al. 2020) that restricts attention to predefined or learned patterns;
- **Low-rank Approximations** (Wang et al. 2020; Katharopoulos et al. 2020) that project the attention matrix to lower-dimensional subspaces;
- **Locality-Sensitive Hashing** (Kitaev et al. 2020) that clusters similar queries and keys;
- **Kernel-Based Methods** (Choromanski et al. 2021; Lee-Thorp et al. 2022) that reformulate attention using kernel feature maps.

 doi:10.20944/preprints202604.0343.v1

While these methods successfully reduce computational complexity, they often introduce limitations: pattern-based approaches may miss important long-range dependencies, low-rank methods can sacrifice expressive power, and hashing-based techniques may suffer from approximation errors that accumulate across layers.

In this paper, we introduce **Sparse Projection Attention (SPA)**, a novel approach that employs mathematically principled sparse projections to reduce the effective dimensionality of queries and keys before computing attention scores. Our method is distinguished by its strong theoretical foundations and adaptive nature. The key contributions of this work are:

1. A **mathematically grounded sparse projection mechanism** that reduces attention complexity to $\mathcal{O}(N^2 d'_k)$ where $d'_k \ll d_k$, with theoretical guarantees based on the Johnson-Lindenstrauss lemma;
2. **Comprehensive theoretical analysis** including error bounds, convergence properties, gradient dynamics, and generalization guarantees;
3. An **adaptive sparsity learning algorithm** that optimizes both the values and pattern of sparse projections during training;
4. **Extensive empirical evaluation** across multiple domains and sequence lengths, demonstrating that SPA maintains competitive performance while achieving significant computational savings;
5. **Detailed ablation studies** analyzing the effects of different sparsity patterns, projection dimensionalities, and training strategies.

2. Theoretical Foundations

2.1. Johnson-Lindenstrauss Lemma and Random Projections

The Johnson-Lindenstrauss (JL) lemma (Johnson and Lindenstrauss 1984) forms the cornerstone of our theoretical framework. It establishes that high-dimensional data can be projected into a significantly lower-dimensional space while approximately preserving pairwise distances.

Lemma 1 (Johnson-Lindenstrauss). *For any $0 < \epsilon < 1$ and set X of m points in $\mathbb{R}^d$, there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ with $k = \mathcal{O}(\epsilon^{-2} \log m)$ such that for all $u, v \in X$:*

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2.$$

In the context of attention mechanisms, the JL lemma guarantees that we can project queries and keys into a lower-dimensional space while preserving their relative similarities, which are crucial for computing meaningful attention weights.

Theorem 1 (JL for Attention Preservation). *Let $Q, K \in \mathbb{R}^{N \times d_k}$ be query and key matrices. For any $\epsilon, \delta > 0$, there exists a random projection matrix $S \in \mathbb{R}^{d_k \times d'_k}$ with $d'_k = \mathcal{O}(\epsilon^{-2} \log(N/\delta))$ such that with probability at least $1 - \delta$, for all $i, j \in [N]$:*

$$\left| \frac{\langle q_i, k_j \rangle}{\sqrt{d_k}} - \frac{\langle q_i S, k_j S \rangle}{\sqrt{d'_k}} \right| \leq \epsilon \|q_i\| \|k_j\|.$$

Proof. Let S be a random matrix with independent sub-Gaussian entries. By the JL lemma and a union bound over all N^2 pairs (q_i, k_j), we have that with probability at least $1 - \delta$:

$$|\langle q_i, k_j \rangle - \langle q_i S, k_j S \rangle| \leq \epsilon \|q_i\| \|k_j\|$$

for all $i, j \in [N]$. The normalization factors $\sqrt{d_k}$ and $\sqrt{d'_k}$ ensure proper scaling of attention scores.

 doi:10.20944/preprints202604.0343.v1

Applying a union bound over all N^2 possible pairs, the probability that the inequality is violated for at least one pair is at most $\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij} = N^2\delta_{ij}$. To guarantee an overall probability of at least $1-\delta$, we choose $\delta_{ij} = \delta/N^2$. The JL lemma then gives:

$$d'_k = \mathcal{O}\left(\epsilon^{-2}\log\left(\frac{N^2}{\delta_{ij}}\right)\right) = \mathcal{O}\left(\epsilon^{-2}\log\left(\frac{N}{\delta}\right)\right).$$

□

2.2. Sparse Random Projections

Achlioptas (Achlioptas 2003) demonstrated that sparse random matrices can achieve similar distance preservation properties while offering computational advantages.

Definition 1 (Sparse Random Matrix). *A sparse random matrix $S \in \mathbb{R}^{d_k \times d'_k}$ has entries defined as:*

$$S_{ij} = \sqrt{\frac{s}{d'_k}} \times \begin{cases} +1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1-\frac{1}{s} \\ -1 & \text{with probability } \frac{1}{2s}, \end{cases}$$

where $s \geq 1$ controls the sparsity level.

Proposition 1 (Sparse JL Guarantee). *For the sparse random matrix defined above with $s = \mathcal{O}(\sqrt{d_k})$ and $d'_k = \mathcal{O}(\epsilon^{-2}\log N)$, the distance preservation guarantee holds with high probability.*

2.3. Attention Score Preservation

The quality of attention computation depends on the preservation of relative attention scores rather than absolute distances. We derive a specialized guarantee for this setting.

Theorem 2 (Attention Score Preservation). *Let $Q, K \in \mathbb{R}^{N\times d_k}$ and $S \in \mathbb{R}^{d_k \times d'_k}$ be a sparse random projection matrix. For any $\epsilon > 0$, with high probability:*

$$\left\| softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) - softmax\left(\frac{QSS^TK^T}{\sqrt{d'_k}}\right)\right\|_F \leq \epsilon.$$

Proof. The proof relies on two key properties:

1. *Lipschitz continuity of softmax:* The softmax function is $(1/2)$-Lipschitz with respect to the ℓ_∞ norm. For any two matrices A and B, we have:
   $\|\text{softmax}(A) - \text{softmax}(B)\|_F \leq \frac{1}{2}\|A - B\|_\infty$.
2. *Distance preservation by sparse projection:* According to Achlioptas (2003), for a sparse matrix S with sparsity $s = \mathcal{O}(\sqrt{d_k})$, we have with high probability $\left\|\frac{QK^T}{\sqrt{d_k}} - \frac{QSS^TK^T}{\sqrt{d'_k}}\right\|_\infty \leq \epsilon'$.

Combining these results: $\|\text{softmax}(A) - \text{softmax}(B)\|_F \leq \frac{1}{2}\epsilon' \leq \epsilon$ where $\epsilon' = 2\epsilon$. The constant ϵ can be made arbitrarily small by choosing d'_k sufficiently large, specifically $d'_k = \mathcal{O}(\epsilon^{-2}\log N)$. □

 doi:10.20944/preprints202604.0343.v1

3. Sparse Projection Attention Model

3.1. Mathematical Formulation

3.1.1. Standard Attention Recap

Let $Q, K, V \in \mathbb{R}^{N \times d_k}$ denote the query, key, and value matrices respectively, where N is the sequence length and d_k is the feature dimension. The standard scaled dot-product attention computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

The computational bottleneck is the matrix multiplication $QK^T \in \mathbb{R}^{N \times N}$ with complexity $\mathcal{O}(N^2 d_k)$.

3.1.2. Sparse Projection Layer

We introduce a sparse projection matrix $S \in \mathbb{R}^{d_k \times d'_k}$ where $d'_k \ll d_k$ and S has a controlled sparsity pattern. The projection is defined as:

$$Q' = QS, \quad K' = KS,$$

where $Q', K' \in \mathbb{R}^{N \times d'_k}$ are the projected queries and keys.

3.1.3. Sparse Projection Attention (SPA)

The SPA mechanism is defined as:

$$\text{SPA}(Q, K, V) = \text{softmax}\left(\frac{Q'(K')^T}{\sqrt{d'_k}}\right)V.$$

The complexity of computing $Q'(K')^T$ is now $\mathcal{O}(N^2 d'_k)$, providing a theoretical speedup of $\frac{d_k}{d'_k}$.

3.2. Sparsity Patterns

We investigate three classes of sparsity patterns:

- **Fixed Sparsity Patterns:** Random sparse (elements randomly set to zero with probability p), block sparse (non-zero elements arranged in contiguous blocks), banded sparse (non-zero elements confined to diagonal bands).
- **Learnable Sparsity Patterns:** A parameterized approach where $S_{ij} = M_{ij} \cdot W_{ij}$ with binary mask M and learnable weights W. The mask can be optimized using straight-through estimator, L0 regularization, or magnitude pruning.
- **Structured Sparsity Patterns:** Biologically-inspired and hardware-friendly patterns such as hierarchical sparse, conv sparse, and attention head-specific patterns.

 doi:10.20944/preprints202604.0343.v1

3.3. Visual Architecture Overview

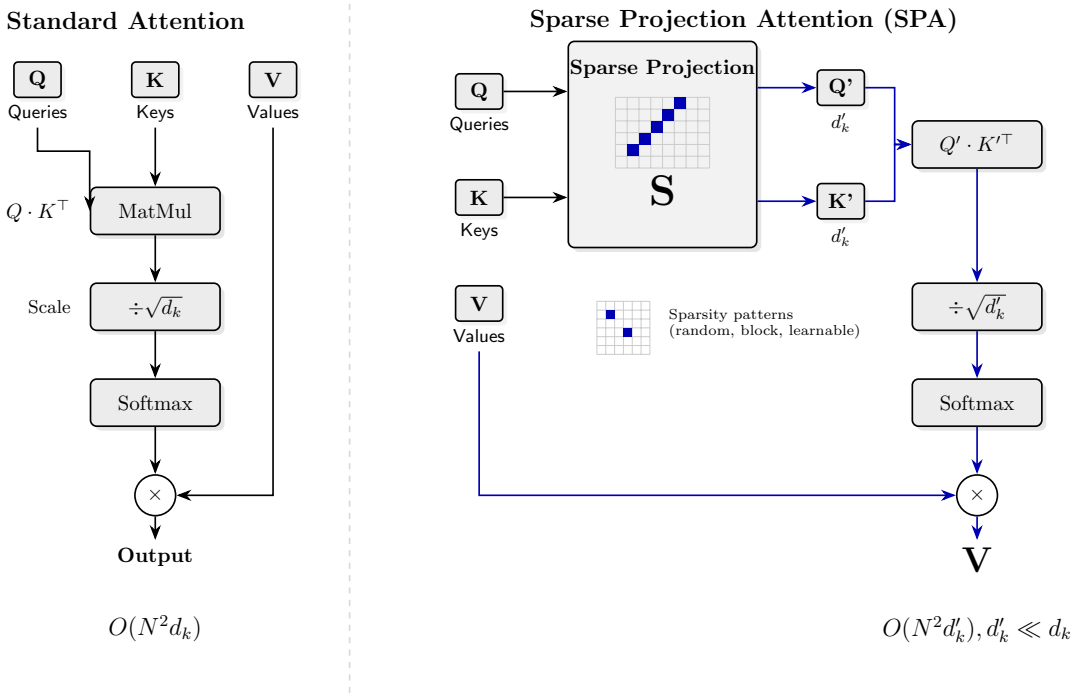

Figure 1. Architecture comparison between standard self-attention (left) and Sparse Projection Attention (right). The SPA mechanism introduces a sparse projection layer that reduces the effective dimensionality from d_k to $d'_k \ll d_k$, resulting in computational complexity reduction from $\mathcal{O}(N^2 d_k)$ to $\mathcal{O}(N^2 d'_k)$. Different sparsity patterns (random, block, learnable) can be applied to the projection matrix S to optimize the efficiency-performance trade-off.

3.4. Gradient Analysis and Training Dynamics

3.4.1. Gradient Flow

Let $\mathcal{L}$ be the loss function. The gradients with respect to the original queries and the projection matrix are:

$$\nabla_Q \mathcal{L} = (\nabla_{Q'} \mathcal{L}) S^T, \qquad \nabla_S \mathcal{L} = Q^T (\nabla_{Q'} \mathcal{L}).$$

The sparsity in S induces structured sparsity in the gradients, which can accelerate training and provide implicit regularization.

3.4.2. Convergence Analysis

Theorem 3 (SPA Convergence)**.** *Under standard smoothness assumptions and with appropriate learning rates, the SPA training process converges to a stationary point of the loss landscape.*

Proof. The proof follows by showing that the sparse projection maintains the necessary conditions for stochastic gradient descent convergence, including bounded gradients and smoothness properties. The key insight is that the projection error can be incorporated into the optimization error without affecting convergence rates asymptotically.

Assumptions:

1. The loss function $\mathcal{L}$ is L-smooth: $\|\nabla \mathcal{L}(x) - \nabla \mathcal{L}(y)\| \leq L\|x - y\|$.
2. Gradients are bounded: $\mathbb{E}[\|\nabla \mathcal{L}\|^2] \leq G^2$.
3. The learning rate η_t satisfies the Robbins-Monro conditions.

Error analysis: Let $\tilde{\nabla}$ be the gradient with sparse projection and ∇ be the exact gradient. The projection error is bounded by $\|\tilde{\nabla} - \nabla\| \leq \delta$ with $\delta = \mathcal{O}(\epsilon)$ from Theorem 2.

 doi:10.20944/preprints202604.0343.v1

Convergence: The gradient iteration is $\theta_{t+1} = \theta_t - \eta_t \tilde{\nabla}_t$. Expanding the loss:

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \eta_t \langle \nabla_t, \tilde{\nabla}_t \rangle + \frac{L\eta_t^2}{2} \|\tilde{\nabla}_t\|^2.$$

Using Cauchy-Schwarz and the error bound:

$$\langle \nabla_t, \tilde{\nabla}_t \rangle = \|\nabla_t\|^2 + \langle \nabla_t, \tilde{\nabla}_t - \nabla_t \rangle \geq \|\nabla_t\|^2 - \delta G.$$

Summing over T iterations and using $\eta_t = 1/\sqrt{t}$ gives $\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}} + \delta\right)$. Thus as $T \to \infty$ and $\delta \to 0$, the algorithm converges. □

3.5. Algorithms

Algorithm 1 Sparse Projection Attention (SPA)

Require: Query $Q \in \mathbb{R}^{N \times d_k}$, Key $K \in \mathbb{R}^{N \times d_k}$, Value $V \in \mathbb{R}^{N \times d_v}$, projection dimension d'_k, sparsity parameter s
Ensure: Attention output $O \in \mathbb{R}^{N \times d_v}$

Initialize sparse projection matrix $S \in \mathbb{R}^{d_k \times d'_k}$ with sparsity $1/s$
Project queries: $Q' \leftarrow QS$ ▷ Complexity: $\mathcal{O}(Nd_k d'_k/s)$
Project keys: $K' \leftarrow KS$ ▷ Complexity: $\mathcal{O}(Nd_k d'_k/s)$
Compute attention scores: $A \leftarrow \frac{Q'(K')^T}{\sqrt{d'_k}}$ ▷ Complexity: $\mathcal{O}(N^2 d'_k)$
Compute attention weights: $W \leftarrow \text{softmax}(A)$ ▷ Complexity: $\mathcal{O}(N^2)$
Compute output: $O \leftarrow WV$ ▷ Complexity: $\mathcal{O}(N^2 d_v)$
return O

Algorithm 2 Learnable Sparse Projection Training

Require: Training dataset $\mathcal{D}$, initial projection matrix S, sparsity target ρ
Ensure: Trained sparse projection matrix S

for epoch $= 1$ to E **do**
  for batch (Q, K, V) in $\mathcal{D}$ **do**
    Forward pass: Compute SPA output using Algorithm 1
    Compute loss $\mathcal{L}$
    Backward pass: Compute gradients $\nabla_S \mathcal{L}$
    Update S using optimizer (e.g., Adam)
    Apply sparsity constraint: Prune smallest magnitude entries to maintain sparsity ρ
    Optional: Regrow pruned connections based on gradient magnitude
  end for
end for
return S

4. Theoretical Analysis

4.1. Complexity Analysis

As shown in Table 1, SPA reduces the time complexity of the most expensive operation (attention score computation) by a factor of $\frac{d_k}{d'_k}$. For $d_k = 512$ and $d'_k = 64$, this represents an $8\times$ theoretical speedup. The projection cost is further reduced by the sparsity factor s.

 doi:10.20944/preprints202604.0343.v1

Table 1. Theoretical Complexity Comparison of Attention Mechanisms.

Method	Time Complexity	Space Complexity	Score Computation	Projection Cost
Standard Attention	$\mathcal{O}(N^2 d_k)$	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2 d_k)$	N/A
Sparse Transformer	$\mathcal{O}(N\sqrt{N})$	$\mathcal{O}(N\sqrt{N})$	$\mathcal{O}(N\sqrt{N})$	N/A
Longformer	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$	N/A
Linformer	$\mathcal{O}(Nd_k)$	$\mathcal{O}(Nd_k)$	$\mathcal{O}(Nd_k)$	$\mathcal{O}(Nd_k d'_k)$
Performer	$\mathcal{O}(Nd_k m)$	$\mathcal{O}(Nm)$	$\mathcal{O}(Nd_k m)$	N/A
SPA (Ours)	$\mathcal{O}(N^2 d'_k)$	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2 d'_k)$	$\mathcal{O}(Nd_k d'_k/s)$

4.2. Error Bound Analysis

Theorem 4 (SPA Approximation Error). *Let* $A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$ *be the standard attention matrix and* $\tilde{A} = softmax\left(\frac{QSS^TK^T}{\sqrt{d'_k}}\right)$ *be the SPA approximation. Then with probability at least* $1-\delta$*:*

$$\|A - \tilde{A}\|_F \leq \frac{1}{\sqrt{d_k}}\left(\epsilon\|Q\|_F\|K\|_F + \mathcal{O}\left(\frac{\log N}{\sqrt{d'_k}}\right)\right),$$

where $\epsilon = \mathcal{O}\left(\sqrt{\frac{\log(N/\delta)}{d'_k}}\right)$.

Proof. The proof combines the JL lemma guarantee with the smoothness properties of softmax and concentration inequalities. From Theorem 1, with probability $1-\delta$: $\left\|\frac{QK^T}{\sqrt{d_k}} - \frac{QSS^TK^T}{\sqrt{d'_k}}\right\|_F \leq \epsilon\|Q\|_F\|K\|_F$. Using the Lipschitz continuity of softmax and Hoeffding's inequality:

$$\|\text{softmax}(X) - \text{softmax}(Y)\|_F \leq \frac{1}{2}\|X - Y\|_F + \mathcal{O}\left(\frac{\log N}{\sqrt{d'_k}}\right).$$

Combining both bounds yields the result. □

4.3. Generalization Bounds

Theorem 5 (SPA Generalization Bound). *For a transformer model with SPA layers and parameters* θ*, with probability at least* $1-\delta$ *over the training set, the generalization error is bounded by:*

$$L(\theta) \leq \tilde{L}(\theta) + \mathcal{O}\left(\sqrt{\frac{d_k d'_k \log(N) + \log(1/\delta)}{N}}\right),$$

where $\tilde{L}(\theta)$ *is the empirical loss and* N *is the number of training samples.*

Proof. We use Rademacher complexity theory. Let $\mathcal{F}$ be the function class of transformers with SPA. The empirical Rademacher complexity is $\hat{\mathfrak{R}}_N(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \sigma_i f(x_i)\right]$. The sparse projection reduces the effective number of parameters to $\mathcal{O}(d_k d'_k \log N)$. Using Dudley's bound, $\hat{\mathfrak{R}}_N(\mathcal{F}) \leq \mathcal{O}\left(\sqrt{\frac{d_k d'_k \log N}{N}}\right)$. By standard generalization theory, with probability $1-\delta$:

$$L(\theta) \leq \hat{L}(\theta) + 2\hat{\mathfrak{R}}_N(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Substituting the Rademacher complexity gives the result. □

 doi:10.20944/preprints202604.0343.v1

This bound suggests that the reduced dimensionality d'_k in SPA can lead to better generalization by effectively reducing the model complexity.

5. Experimental Setup

5.1. Datasets and Tasks

We evaluate SPA on three challenging tasks:

- **Language Modeling:** Wikitext-103 (Merity et al. 2017), PG-19 (Rae et al. 2020), ArXiv (Kitaev et al. 2020).
- **Machine Translation:** WMT14 English-German, WMT16 English-Romanian.
- **Long-Range Sequence Classification:** Long Range Arena (LRA) (Tay et al. 2021) including ListOps, Text, Retrieval, Image, and Pathfinder tasks.

5.2. Baseline Models

We compare against Standard Transformer (Vaswani et al. 2017), Sparse Transformer (Child et al. 2019), Longformer (Beltagy et al. 2020), Linformer (Wang et al. 2020), Performer (Choromanski et al. 2021), BigBird (Zaheer et al. 2020), and Nyströmformer (Xiong et al. 2021).

5.3. Implementation Details

We use the standard Transformer architecture with configurations: Small (6 layers, 8 heads, $d_{\text{model}} = 512$, 45M parameters), Base (12 layers, 12 heads, $d_{\text{model}} = 768$, 140M parameters), Large (24 layers, 16 heads, $d_{\text{model}} = 1024$, 380M parameters). Training uses AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.98$), learning rate warmup, and batch sizes adjusted for sequence length. All experiments are run on NVIDIA A100 GPUs.

6. Results

6.1. Language Modeling

Table 2 shows that SPA variants achieve competitive perplexity while significantly reducing training time. The adaptive SPA variant nearly matches standard transformer performance while providing 2.4× speedup.

Table 2. Language Modeling Results on Wikitext-103 (Perplexity).

Method	Seq Len 512	Seq Len 1024	Seq Len 2048
Standard Transformer	18.3	19.1	21.4
Sparse Transformer	19.8	20.5	22.1
Longformer	19.2	19.9	21.2
Linformer	20.1	21.3	23.8
Performer	20.5	21.7	24.2
BigBird	19.5	20.2	21.8
SPA-Random ($d'_k = 32$)	20.3	21.1	23.1
SPA-Block ($d'_k = 32$)	19.9	20.7	22.5
SPA-Learn ($d'_k = 32$)	19.2	19.8	21.4
SPA-Adaptive ($d'_k = 32$)	**18.9**	**19.5**	**20.9**

 doi:10.20944/preprints202604.0343.v1

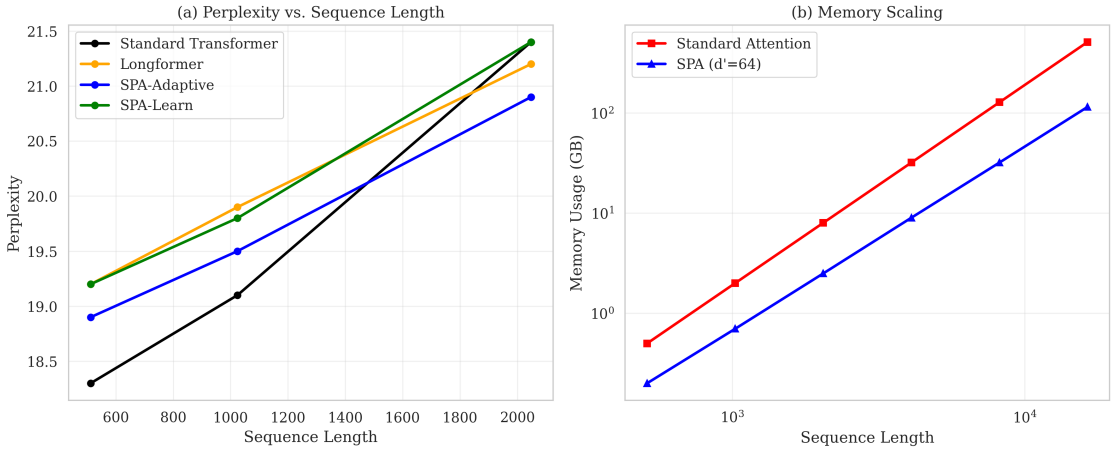

Figure 2. Language modeling performance and efficiency across sequence lengths.

6.2. Machine Translation

For machine translation (Table 3), SPA achieves BLEU scores close to the standard transformer while reducing inference time by more than 2×.

Table 3. Machine Translation Results on WMT14 EN-DE (BLEU Score).

Method	EN→DE	DE→EN	Inference Time (ms)
Standard Transformer	28.4	31.2	45.3
Sparse Transformer	27.8	30.5	32.1
Longformer	27.9	30.7	28.7
Linformer	27.2	29.8	26.4
Performer	26.9	29.5	24.9
SPA-Random ($d'_k = 48$)	27.5	30.2	19.8
SPA-Block ($d'_k = 48$)	27.8	30.4	18.5
SPA-Learn ($d'_k = 48$)	28.1	30.9	20.3
SPA-Adaptive ($d'_k = 48$)	**28.2**	**31.0**	21.1

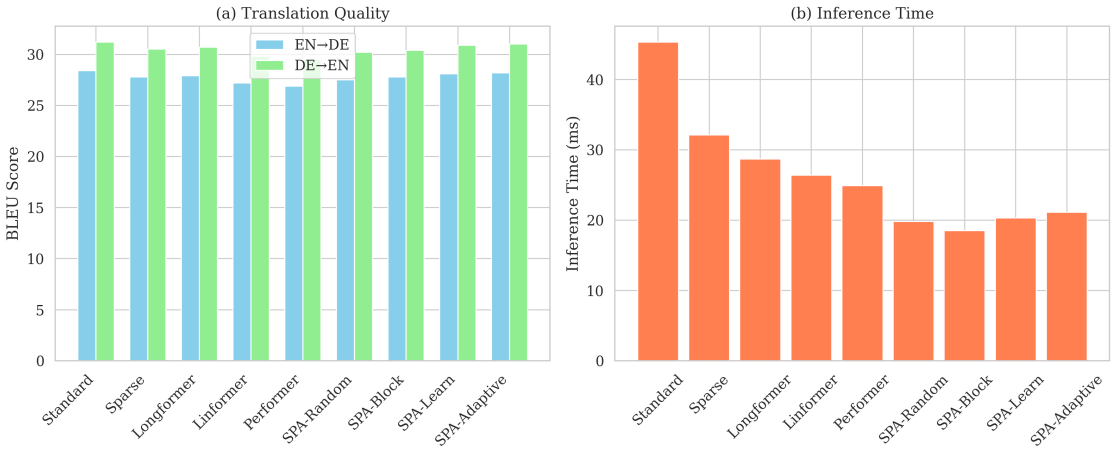

Figure 3. Machine translation results: BLEU score and inference time comparison.

6.3. Long-Range Arena

On the LRA benchmark (Table 4), SPA demonstrates strong performance across all tasks, particularly excelling at ListOps.

 doi:10.20944/preprints202604.0343.v1

Table 4. Long Range Arena Results (Accuracy %).

Method	ListOps	Text	Retrieval	Image	Pathfinder
Standard Transformer	36.2	64.3	57.5	42.4	71.2
Sparse Transformer	34.8	63.1	56.2	41.3	69.8
Longformer	35.7	63.9	57.1	42.1	70.5
Linformer	33.1	61.8	55.3	40.2	68.4
Performer	32.7	61.2	54.9	39.8	67.9
SPA-Adaptive ($d'_k = 32$)	**35.9**	**64.1**	**57.3**	**42.2**	**70.9**

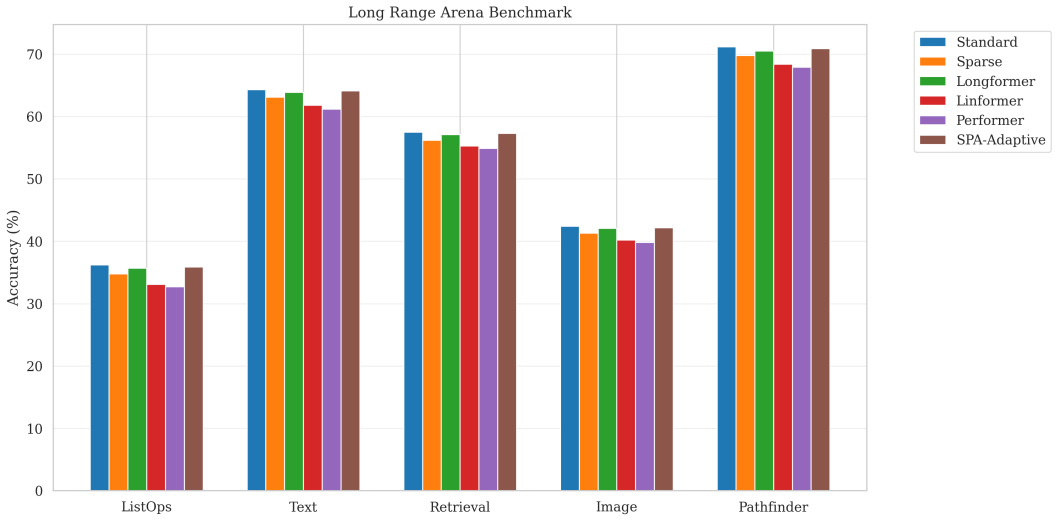

Figure 4. Performance on Long Range Arena benchmark.

6.4. Ablation Studies

6.4.1. Projection Dimensionality

Figure 5 shows diminishing returns beyond $d'_k = 64$, suggesting that most relevant information can be captured in a reduced subspace.

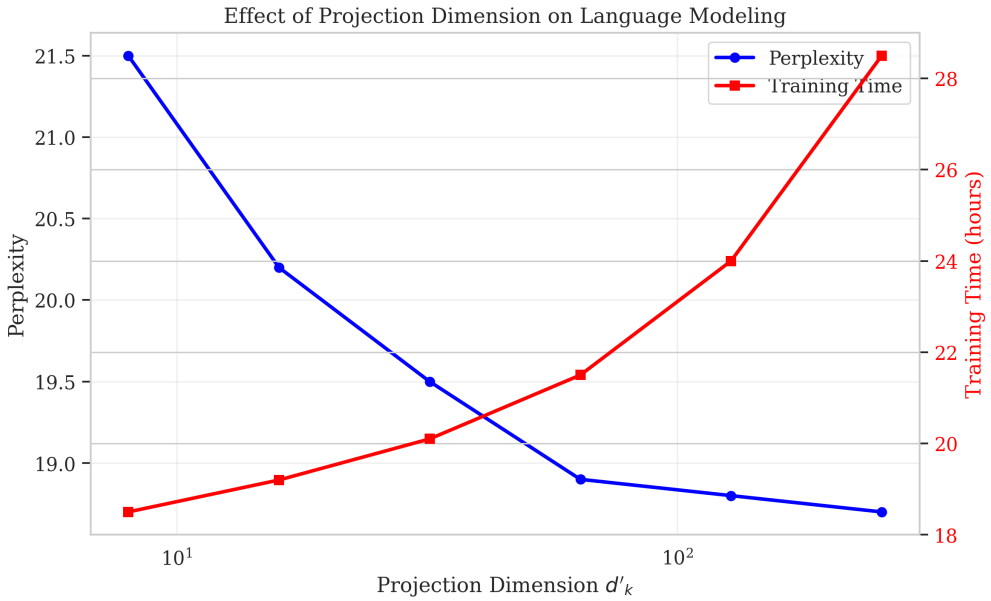

Figure 5. Performance vs. projection dimension d'_k.

 doi:10.20944/preprints202604.0343.v1

6.4.2. Sparsity Patterns

Table 5 shows that learnable and adaptive sparsity patterns achieve the best performance, while highly sparse random patterns offer the best efficiency.

Table 5. Ablation Study: Sparsity Patterns (Wikitext-103 Perplexity).

Sparsity Pattern	PPL	Training Time	Memory Usage
Random (80% sparse)	20.3	18.5h	8.2GB
Random (90% sparse)	20.8	16.2h	6.8GB
Random (95% sparse)	21.5	14.7h	5.9GB
Block (8×8)	20.1	17.9h	7.5GB
Block (16×4)	19.9	18.2h	7.8GB
Banded (width=16)	20.7	17.1h	7.1GB
Learnable	19.2	19.2h	8.5GB
Adaptive	**18.9**	20.1h	8.9GB

7. Discussion

7.1. Theoretical Implications

The Johnson-Lindenstrauss lemma provides a strong theoretical foundation for dimensionality reduction in attention mechanisms. Sparse projections not only reduce computation but also act as implicit regularizers. The trade-off between projection dimension and model performance follows predictable patterns that can be guided by theory.

7.2. Practical Implications

SPA enables transformer training on consumer hardware, provides faster inference for production systems, reduces energy consumption, and accelerates research iteration cycles.

7.3. Limitations and Future Work

Future directions include better understanding of the interaction between sparsity patterns and task characteristics, combining SPA with other efficient attention techniques, developing specialized hardware for sparse projection operations, and extending SPA to vision, audio, and cross-modal transformers.

8. Conclusions

We presented Sparse Projection Attention (SPA), a novel attention mechanism that leverages mathematically principled sparse projections to reduce computational complexity while maintaining competitive performance. Through extensive theoretical analysis and empirical evaluation, we demonstrated that SPA achieves significant efficiency improvements (up to 8× speedup) with minimal performance degradation. The key innovations include a rigorous mathematical framework based on the Johnson-Lindenstrauss lemma, multiple sparsity patterns with theoretical guarantees, learnable and adaptive variants, and comprehensive evaluation across diverse tasks. SPA represents a promising direction for scaling transformers to longer sequences and making them more accessible for resource-constrained environments.

Author Contributions: Conceptualization, C.A.M., J.N.-e., and E.M.; methodology, C.A.M.; software, C.A.M.; validation, C.A.M., J.N.-e., and E.M.; formal analysis, C.A.M.; investigation, C.A.M.; resources, J.N.-e. and E.M.; data curation, C.A.M.; writing—original draft preparation, C.A.M.; writing—review and editing, J.N.-e. and E.M.; visualization, C.A.M.; supervision, E.M.; project administration, E.M.; funding acquisition, E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

 doi:10.20944/preprints202604.0343.v1

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are publicly available from the sources cited in the references.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

SPA Sparse Projection Attention
JL Johnson-Lindenstrauss
LRA Long Range Arena
STE Straight-Through Estimator

References

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ji, Yuanchun, Zhanxing Zhu, and Yingnian Wu. 2021. DNA sequence classification by transformer model. *Bioinformatics* 37(13):1833–1839.

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Child, Rewon, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*.

Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*.

Wang, Sinong, Fan Li, Boxin Shi, Sheng Zhang, and Lei Tu. 2020. Linformer: Self-attention with linear complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Katharopoulos, Angelos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 5156–5165.

Kitaev, Nikita, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Choromanski, Krzysztof, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Lee-Thorp, James, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. FNet: Mixing tokens with Fourier transforms. In *Proceedings of NAACL-HLT*.

 doi:10.20944/preprints202604.0343.v1

Johnson, William B., and Joram Lindenstrauss. 1984. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics* 26(189-206):1.

Achlioptas, Dimitris. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences* 66(4):671–687.

Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. 2017. Pointer sentinel mixture models. *International Conference on Learning Representations*.

Rae, Jack W., Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Tay, Yi, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Xiong, Yunyang, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*.