Novelty Assessment Report

Paper: Deconstructing Positional Information: From Attention Logits to Training Biases

PDF URL: https://openreview.net/pdf?id=D0u0glT060

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Positional encodings, a mechanism for incorporating sequential information into the Transformer model, are central to contemporary research on neural architectures. Previous work has largely focused on understanding their function through the principle of distance attenuation, where proximity dictates influence. However, the interaction between positional and semantic information remains insufficiently explored, and the complexity of mainstream corpora hinders systematic, comparative studies of these methods. This paper addresses these challenges through a deconstruction of the attention-logit computation and a structured analysis of all mainstream positional encodings. A key focus is placed on Rotary Positional Embedding (RoPE), whose product-based structure uniquely facilitates a direct interaction between position and content. To probe this characteristic, we designed a novel synthetic task that explicitly demands a strong synthesis of positional and semantic information. As theoretically predicted, RoPE demonstrates a significant performance advantage over other encodings on this specialized task. Concurrently, this targeted evaluation uncovers an implicit training issue: a hidden bias manifesting as a distinct information aggregation phenomenon in the model's shallow layers, which we term the "single-head deposit pattern." Through subsequent ablation studies, we analyze this pattern and identify a method for its mitigation. These findings highlight the need for a deeper investigation into the training dynamics of positional encodings to bridge the gap between their theoretical design and practical implementation.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Positional Encoding Mechanisms in Transformer Attention**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Positional Encoding Design and Theoretical Foundations**
- **Generalization and Extrapolation Capabilities**
- **Application Domains**

Complete Taxonomy Tree

- Positional Encoding Mechanisms in Transformer Attention Survey Taxonomy
- Positional Encoding Design and Theoretical Foundations
  - Comparative Analysis and Taxonomies (2 papers)
  - [3] Position information in transformers: An overview (Dufter, 2022) View paper
  - [4] Positional encoding in transformer-based time series models: a survey (Metsis, 2025) View paper
  - Attention Mechanism Interactions ★ (3 papers)
  - [0] Deconstructing Positional Information: From Attention Logits to Training Biases (Anon et al., 2026) View paper
  - [18] Understanding the Expressive Power and Mechanisms of Transformer for Sequence Modeling (Weinan E, 2024) View paper
  - [30] A Free Probabilistic Framework for Analyzing the Transformer-based Language Models (Swagatam, 2025) View paper
  - Novel Encoding Schemes
  - Relative and Rotary Encodings (4 papers)
    - [5] Rotary position embedding for vision transformer (Byeongho Heo, 2024) View paper
    - [23] HoPE: Hyperbolic Rotary Positional Encoding for Stable Long-Range Dependency Modeling in Large Language Models (Dai Chang, 2025) View paper
    - [24] Positional Encoding via Token-Aware Phase Attention (Wang Yu, 2025) View paper
    - [38] Rethinking and improving relative position encoding for vision transformer (Wu Kan, 2021) View paper
  - Absolute and Learnable Encodings (3 papers)
    - [19] Rethinking positional encoding in language pre-training (Guolin Ke, 2020) View paper
    - [34] Learning to encode position for transformer with continuous dynamical model (Liu, 2020) View paper
    - [40] A simple and effective positional encoding for transformers (Tsai, 2021) View paper
  - Hybrid Encoding Strategies (2 papers)
    - [16] Position embedding fusion on transformer for dense video captioning (Sixuan Yang, 2020) View paper
    - [20] Two stones hit one bird: Bilevel positional encoding for better length extrapolation (He Zhenyu, 2024) View paper
  - Dynamic and Adaptive Encodings (2 papers)
    - [47] Dynamic positional attention fusion (DPAF): Adaptive encoding and weighted attention for ship motion attitude prediction (Huachuan Zhao, 2024) View paper
    - [50] HDTSLR: A Framework Based on Hierarchical Dynamic Positional Encoding for Sign Language Recognition (Jiangtao Zhang, 2024) View paper
  - Empirical Studies of Encoding Behavior (4 papers)

  - [13] Understanding How Positional Encodings Work in Transformer Model (T Miyazaki, 2024) View paper
  - [21] Positional encoding is not the same as context: A study on positional encoding for Sequential recommendation (Lopez-Avila, 2024) View paper
  - [39] An Empirical Study on the Impact of Positional Encoding in Transformer-Based Monaural Speech Enhancement (Qiquan Zhang, 2024) View paper
  - [43] What do position embeddings learn? an empirical study of pre-trained language model positional encoding (Yu-An Wang, 2020) View paper
- Generalization and Extrapolation Capabilities
  - Length Generalization (2 papers)
  - [1] The impact of positional encoding on length generalization in transformers (Kazemnejad, 2023) View paper
  - [6] Randomized positional encodings boost length generalization of transformers (DelÃ©tang, 2023) View paper
  - Arithmetic and Algorithmic Tasks (1 papers)
  - [15] Positional description matters for transformers arithmetic (Shen, 2023) View paper
  - In-Context Learning and Induction (1 papers)
  - [35] Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers (Si-yu Chen, 2024) View paper
- Application Domains
  - Computer Vision Applications
  - Image Recognition and Classification (4 papers)
    - [25] Novel Deepfake Image Detection with PV-ISM: Patch-Based Vision Transformer for Identifying Synthetic Media (Orkun Ã□Ä±nar, 2025) View paper
    - [26] StyTr2: Image Style Transfer with Transformers (Deng Yingying, 2021) View paper
    - [33] Rethinking position embedding methods in the Transformer architecture (Xin Zhou, 2023) View paper
    - [42] Reviving Shift Equivariance in Vision Transformers (Ding, 2023) View paper
  - Video and Temporal Vision (2 papers)
    - [8] Aiatrack: Attention in attention for transformer visual tracking (Shen-Yuan Gao, 2022) View paper
    - [10] An improved end-to-end multi-target tracking method based on transformer self-attention (Hong Yong, 2022) View paper
  - 3D Vision and Point Clouds (3 papers)
    - [12] Self-positioning point-based transformer for point cloud understanding (Jin-Young Park, 2023) View paper
    - [22] MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo (Cao Chenjie, 2024) View paper
    - [28] 3DPPE: 3D Point Positional Encoding for Transformer-based Multi-Camera 3D Object Detection (Changyong Shu, 2023) View paper
  - Specialized Vision Tasks (3 papers)
    - [27] Attention and positional encoding are (almost) all you need for shape matching (Alessandro Raganato, 2023) View paper
    - [37] VPDETR: End-to-End Vanishing Point DEtection TRansformers (Guo Ruohao, 2024) View paper
    - [41] MCT-Grasp: A Novel Grasp Detection Using Multimodal Embedding and Convolutional Modulation Transformer (Guowei Yang, 2024) View paper
  - Natural Language Processing
  - Language Pre-training and Representation (1 papers)
    - [29] Self-adapted positional encoding in the transformer encoder for named entity recognition (Kehan Huangliang, 2023) View paper
  - Sequence Labeling and Classification (2 papers)
    - [31] Improving part-of-speech tagging with relative positional encoding in transformer models and basic rules (Jerome Aondongu Achir, 2025) View paper
    - [36] CovTiNet: Covid text identification network using attention-based positional embedding feature fusion (Md. Rajib Hossain, 2023) View paper
  - Question Answering and Generation (1 papers)
    - [48] On solving textual ambiguities and semantic vagueness in MRC based question answering using generative pre-trained transformers (Muzamil Ahmed, 2023) View paper
  - Time Series and Temporal Data (4 papers)
  - [9] Enhancing multivariate time-series anomaly detection with positional encoding mechanisms in transformers (Abdul Amir AlÄ±oghlÄ±, 2024) View paper
  - [32] Heterogeneous multivariate time series imputation by transformer model with missing position encoding (Caizheng Liu, 2025) View paper
  - [44] An Adaptive Transformer Model for Long-Term Time-Series Forecasting of Temperature and Radiation in Photovoltaic Energy Generation (El-Saieed, 2025) View paper
  - [49] Lightweight Hierarchical Transformer Combining Patch-Random and Positional Encoding for Respiratory Sound Classification (Jianhong Wang, 2024) View paper
  - Graph and Structured Data
  - Graph Neural Networks and Transformers (3 papers)
    - [2] Graph representation learning via enhanced GNNs and transformers (Hongrui Mu, 2025) View paper
    - [7] Graphit: Encoding graph structure in transformers (Mialon, 2021) View paper
    - [45] What Improves the Generalization of Graph Transformers? A Theoretical Dive into the Self-attention and Positional Encoding (LI Hongkang, 2024) View paper
  - Tree-Structured Data (1 papers)
    - [11] Novel positional encodings to enable tree-based transformers (Vighnesh Leonardo Shiv, 2019) View paper
  - Process and Workflow Modeling (1 papers)
    - [14] Structural positional encoding for knowledge integration in transformer-based medical process monitoring and trace classification (Christopher Irwin, 2024) View paper
  - Scientific and Engineering Applications (2 papers)
  - [17] Positional knowledge is all you need: Position-induced transformer (PiT) for operator learning (Chen Jun-feng, 2024) View paper
  - [46] SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery (Cong, 2022) View paper

Narrative

Core task: positional encoding mechanisms in transformer attention. The field has organized itself around three major branches. The first, Positional Encoding Design and Theoretical Foundations, encompasses works that propose novel encoding schemes—ranging from sinusoidal and learned embeddings to rotary and relative approaches—and investigate their theoretical properties, including how they interact with attention mechanisms and what expressive power they confer. Representative efforts include Position Information Overview[3], Positional Encoding Survey[4], and Rotary Vision Transformer[5], which illustrate the diversity of design choices. The second branch, Generalization and Extrapolation Capabilities, focuses on whether models can handle sequences longer than those seen during training or transfer positional knowledge across domains; Length Generalization Positional[1] and Randomized Positional Encodings[6] exemplify this line of inquiry. The third branch, Application Domains, explores how positional encodings adapt to specialized settings such as graphs, vision, speech, and medical imaging, with works like Enhanced GNNs Transformers[2] and Anomaly Detection Positional[9] demonstrating domain-specific innovations.

Within the design and theoretical foundations branch, a particularly active area examines how positional information flows through and modulates attention computations. Deconstructing Positional Information[0] sits squarely in this cluster, analyzing the interplay between encoding schemes and attention weights to understand what makes certain designs effective. It shares thematic ground with Expressive Power Mechanisms[18], which investigates the representational capacity conferred by different positional strategies, and Free Probabilistic Framework[30], which offers a probabilistic lens on how position signals propagate. These works collectively address open questions about whether positional encodings should be baked into embeddings, injected into attention scores, or both, and how such choices affect downstream performance and interpretability across varied sequence lengths and task complexities.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Understanding the Expressive Power and Mechanisms of Transformer for Sequence Modeling

Authors: Weinan E, Mingze Wang | **Year/Venue**: 2024 • Neural Information Processing Systems | **URL**: View paper

Abstract

We conduct a systematic study of the approximation properties of Transformer for sequence modeling with long, sparse and complicated memory. We investigate the mechanisms through which different components of Transformer, such as the dot-product self-attention, positional encoding and feed-forward layer, affect its expressive power, and we study their combined effects through establishing explicit approximation rates. Our study reveals the roles of critical parameters in the Transformer, such as...

Relationship Analysis

Both papers belong to the 'Attention Mechanism Interactions' category, investigating how positional encodings interact with self-attention mechanisms. The original paper deconstructs attention logits to analyze additive versus multiplicative positional encodings (like RoPE) and their training biases, while the candidate paper takes a broader theoretical approach by studying the approximation properties and expressive power of Transformers for sequence modeling. The key difference is that the original paper focuses on empirical phenomena (single-head deposit patterns) and mechanistic interpretations of specific PE schemes, whereas the candidate paper establishes formal approximation rates and theoretical bounds on Transformer capabilities.

2. A Free Probabilistic Framework for Analyzing the Transformer-based Language Models

Authors: Das Swagatam | **Year/Venue**: 2025 • Statistics & Probability Letters | **URL**: View paper

Abstract

We present a formal operator-theoretic framework for analyzing Transformer-based language models using free probability theory. By modeling token embeddings and attention mechanisms as self-adjoint operators in a tracial (W^*)-probability space, we reinterpret attention as non-commutative convolution and describe representation propagation via free additive convolution. This leads to a spectral dynamic system interpretation of deep Transformers. We derive entropy-based generalization bounds ...

Relationship Analysis

Both papers belong to the 'Attention Mechanism Interactions' category, investigating how positional encodings interact with self-attention mechanisms. The original paper empirically deconstructs attention logits by categorizing positional encodings into additive and multiplicative forms, revealing training biases like the single-head deposit pattern in RoPE through synthetic tasks. The candidate paper takes a purely theoretical approach using free probability theory and operator-theoretic frameworks to model attention as non-commutative convolution, focusing on spectral dynamics and entropy-based generalization bounds rather than empirical phenomena or specific encoding comparisons.

Contributions Analysis

Overall novelty summary. The paper proposes a unified Toeplitz-based framework for analyzing positional encodings, with particular emphasis on RoPE's product-based structure and its interaction with semantic information. It resides in the 'Attention Mechanism Interactions' leaf under 'Positional Encoding Design and Theoretical Foundations,' alongside two sibling papers. This leaf represents a focused research direction within the broader taxonomy of 50 papers across 22 leaf nodes, indicating a moderately populated area dedicated to theoretical investigations of how positional encodings modulate attention computation rather than empirical performance studies or application-specific implementations.

The taxonomy reveals that neighboring leaves include 'Comparative Analysis and Taxonomies' (2 papers) and 'Empirical Studies of Encoding Behavior' (4 papers), while the broader 'Novel Encoding Schemes' branch contains multiple subtopics examining relative, absolute, hybrid, and dynamic encodings. The paper's focus on RoPE's unique product-based structure positions it at the intersection of theoretical analysis and encoding design, distinguishing it from purely comparative surveys or empirical behavior studies. Its synthetic task methodology bridges theoretical predictions with targeted evaluation, connecting to the 'Arithmetic and Algorithmic Tasks' leaf in the generalization branch.

Among the 9 candidates examined through limited semantic search, all three contributions show evidence of prior work overlap. The Toeplitz framework contribution examined 2 candidates with 1 refutable match; the single-head deposit pattern discovery examined 6 candidates with 1 refutable match; and the causal demonstration examined 1 candidate with 1 refutable match. These statistics suggest that within the limited search scope, each core contribution encounters at least one paper providing overlapping prior work, though the scale of examination (9 total candidates) means substantial relevant literature may remain unexamined.

Based on the top-9 semantic matches examined, the work appears to build incrementally on existing theoretical frameworks for positional encoding analysis, with each contribution finding at least one overlapping prior study. The taxonomy context shows this is an active research area with established theoretical foundations, though the limited search scope prevents definitive assessment of whether the specific combination of Toeplitz analysis, RoPE-focused investigation, and synthetic task design represents a novel synthesis or extends known approaches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Unified Toeplitz-based framework for analyzing positional encodings

Description: The authors introduce a framework that decomposes attention logit computation using Toeplitz matrix structures to systematically distinguish between additive positional encodings (e.g., T5, ALiBi) and multiplicative encodings (e.g., RoPE), revealing how each type couples content and position information differently.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Unpacking Positional Encoding in Transformers: A Spectral Analysis of Content-Position Coupling

URL: View paper

Prior Art Analysis

Spectral Analysis Positional[52] presents a nearly identical Toeplitz-based framework for analyzing positional encodings. Both papers decompose attention logits using Toeplitz matrices to distinguish additive (T5, ALiBi) from multiplicative (RoPE) encodings. The candidate paper uses the same mathematical formulation (Equations 8-10), the same theoretical assumptions (representation decomposition and Toeplitz structure), and applies the same spectral analysis techniques (Szegő's theorem) to analyze how different PE schemes couple content and position. The extensive overlap in framework structure, mathematical notation, and analytical approach demonstrates that this contribution was not novel to the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce a unified Toeplitz-based framework for analyzing positional encodings, with identical categorization into additive and multiplicative mechanisms. - **Original**: we propose a unified analytical framework using toeplitz matrices that categorizes positional encodings into additive and multiplicative mechanisms, clarifying their distinct effects on attention logits. - **Candidate**: we present a unified framework that analyzes pe through the spectral properties of toeplitz and related matrices derived from attention logits. we show that multiplicative content-position coupling-exemplified by rotary positional encoding (rope) via a hadamard product with a toeplitz matrix-induces...

Evidence 2 - **Rationale**: The mathematical formulation for additive mechanisms is identical, using the same notation and decomposition structure. - **Original**: for additive mechanisms (e.g., absolute or relative pe), the logit matrix is a sum of components: ladditive = gqc,kc + gqc,kp + gqp,kc + gqp,kp + b (1) here, b denotes the explicit relative-bias matrix used only by relative encodings such as t5 and alibi. - **Candidate**: under this notation, for pe methods that operate via additive modification of token representations or attention biases (excluding rope), the attention logit matrix can be expressed as: lnon-rope = gqc,kc + gqc,kp + gqp,kc + gqp,kp + b (8) here, b denotes an explicit bias matrix (e.g., from t5 or al...

Evidence 3 - **Rationale**: Both papers use identical mathematical formulation for RoPE's multiplicative structure with Hadamard product and Toeplitz matrix. - **Original**: in contrast, multiplicative mechanisms like rope induce a different structure. following its formulation using complex numbers in su et al. (2024), the logit matrix becomes: lrope = re {(gqc,kc + gqc,kp + gqp,kc + gqp,kp) ◦ ge} (2) where ◦ is the hadamard product and ge is a toeplitz matrix. - **Candidate**: in contrast, rope applies complex-valued rotations to token representations. by pairing adjacent dimensions into complex numbers, we construct complex-valued vectors qi, ki, and pi, and define a modulation vector ei such that e(j) i = eiθj . the logit matrix then takes the form: lrope_complex = (gq,...

Evidence 4 - **Rationale**: The foundational assumption for the framework is identically stated in both papers. - **Original**: assumption 3.1 (representation decomposition). any token representation xi can be conceptually decomposed into a position-independent content component ci and a position-dependent position component pi, such that xi = ci + pi. - **Candidate**: assumption 3.1 (representation decomposition). each token representation xi at position i can be decomposed into a position-independent content component ci and a position-dependent component pi, such that xi = ci + pi. (5)

Evidence 5 - **Rationale**: The second foundational assumption is also identically formulated in both papers. - **Original**: assumption 3.2 (toeplitz structure from positional interaction). since the positional contribution to attention depends only on the relative displacement j - i, the gram matrix formed by the position-dependent components naturally takes a toeplitz form. - **Candidate**: assumption 3.2 (toeplitz structure from positional interaction). the attention score matrix induced solely by the position-dependent components pi exhibits a toeplitz structure: attention(pi, pj) = [ai-j]. (7)

... and 1 more evidence pairs

2. Lightweight structure-aware attention for visual understanding

URL: View paper

Brief Assessment

Lightweight Structure-Aware Attention[51] focuses on visual understanding tasks using structure-aware attention mechanisms with Toeplitz matrices for relative position embeddings in vision transformers, not on analyzing or comparing additive versus multiplicative positional encodings in language models.

Contribution 2: Discovery and empirical analysis of single-head deposit pattern in RoPE

Description: Through carefully designed synthetic tasks requiring content-position coupling, the authors discover that RoPE concentrates nearly all positional processing into a single attention head in shallow layers, a phenomenon they term the single-head deposit pattern, which explains RoPE's performance paradoxes.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Unpacking Positional Encoding in Transformers: A Spectral Analysis of Content-Position Coupling

URL: View paper

Prior Art Analysis

Spectral Analysis Positional[52] reports the identical 'single-head deposit pattern' phenomenon in RoPE through the same experimental methodology. Both papers use synthetic tasks requiring content-position coupling, conduct head-wise ablation studies, and observe that RoPE concentrates positional processing into a single attention head in shallow layers. The candidate paper uses the same terminology ('deposit pattern'), the same visualization methods (violin plots), and reaches identical conclusions about this being unique to RoPE on position-sensitive tasks. The extensive methodological and observational overlap demonstrates prior discovery of this phenomenon.

Evidence

Evidence 1 - **Rationale**: Both papers claim to discover and identify the same 'single-head deposit pattern' in RoPE through empirical analysis. - **Original**: we empirically identify and analyze rope's performance paradox through targeted synthetic tasks, revealing a unique 'single-head deposit pattern' where positional logic becomes highly concentrated in shallow layers. - **Candidate**: our experiments reveal strong alignment with theory: rope consistently outperforms other methods on position-sensitive tasks and induces 'single-head deposit' patterns in early layers, indicating localized positional processing.

Evidence 2 - **Rationale**: Both papers observe and report the same phenomenon: ablating a single shallow-layer head in RoPE causes catastrophic accuracy loss, while other heads have negligible impact. - **Original**: in the rope model, nearly all positional logic is localized to a single head in the first layer. ablating this one head causes a catastrophic drop in accuracy (≈60%), while ablating other heads has a negligible effect. we term this phenomenon the 'single-head deposit pattern.' - **Candidate**: figure 4 confirms that under rope, ablating a single shallow-layer head causes a pronounced accuracy loss, indicating early-layer specialization. by contrast, no pronounced deposit pattern appears under no-rope methods like nope or alibi.

Evidence 3 - **Rationale**: Both papers establish that the deposit pattern is unique to RoPE on position-sensitive tasks and does not appear in other PE methods or on position-agnostic tasks. - **Original**: crucially, this pattern is unique to the combination of the rope architecture and a position-sensitive task. as shown in figure 4, the pattern does not emerge in the nope model on the same task, nor does it appear in the rope model trained on the position-agnostic task 2. - **Candidate**: by contrast, no pronounced deposit pattern appears under no-rope methods like nope or alibi. in the main text we illustrate nope's task 1 ablation and rope's task 2 ablation in figure 4; the full set of non-rope results (absolute, alibi, relative, random) is provided in appendix c.

Evidence 4 - **Rationale**: The experimental methodology for discovering the deposit pattern is identical: head-wise ablation by zeroing outputs and measuring accuracy drops. - **Original**: to find direct evidence for this phenomenon, we conduct a head-wise causal ablation study on the models trained on task 1. for each head in the network, we zero out its output and measure the resulting drop in test accuracy. a large drop indicates that the head is critical to the task. - **Candidate**: to complete our theoretical argument, we predict that pe mechanisms with spectrally stable position coupling-such as rope-will localize essential positional computations into a few specialized attention heads, especially in early layers, yielding a deposit pattern. we test this via a causal, headwis...

Evidence 5 - **Rationale**: Both papers use identical visualization methods (violin plots) and identify the same three characteristic patterns: deposit, average, and shock. - **Original**: each violin shows the distribution of accuracies after ablating each head in that layer; long vertical tails correspond to outlier heads whose removal causes a large accuracy drop. the three insets illustrate the characteristic shapes of these distributions: a 'deposit' shape with a single deep outl... - **Candidate**: deposit pattern average pattern shock pattern head layer accuracy figure 4: head-wise ablation violin plot for three situations.

2. Head-wise Adaptive Rotary Positional Encoding for Fine-Grained Image Generation

URL: View paper

Brief Assessment

Head-Wise Adaptive Rotary[54] focuses on improving RoPE for image generation through learnable transformations, not on analyzing attention head specialization patterns in shallow layers or discovering the single-head deposit phenomenon.

3. Vision Transformer-Based Deepfake Detection: A Self-Attention Approach for Classification of Real and Synthetic Facial Images

URL: View paper

Brief Assessment

Vision Transformer Deepfake[55] focuses on deepfake detection using vision transformers for facial image classification, not on analyzing attention head specialization or positional encoding mechanisms in transformers.

4. Transformer with Syntactic Position Encoding for Machine Translation

URL: View paper

Brief Assessment

Syntactic Position Encoding[57] focuses on incorporating dependency tree structures into transformer attention for machine translation, not on analyzing attention head specialization patterns or positional information processing in RoPE. The paper does not examine how RoPE concentrates positional processing into specific heads.

5. ComplexFormer: Disruptively Advancing Transformer Inference Ability via Head-Specific Complex Vector Attention

URL: View paper

Brief Assessment

ComplexFormer Head-Specific[56] focuses on a novel complex-valued attention mechanism with per-head adaptive transformations for improved performance. It does not discuss or analyze the single-head deposit pattern phenomenon in shallow layers of RoPE-based transformers.

6. Context-aware Biases for Length Extrapolation

URL: View paper

Brief Assessment

Context-Aware Biases[53] proposes a novel additive relative positional encoding method for length extrapolation. It does not analyze attention head specialization patterns or the single-head deposit phenomenon in RoPE's shallow layers.

Contribution 3: Causal demonstration that deposit pattern is intrinsic to RoPE architecture

Description: Through ablation studies and theoretical gradient analysis, the authors prove that the single-head deposit pattern arises inherently from RoPE's multiplicative structure rather than being a training artifact, providing a mechanistic explanation for why RoPE sometimes underperforms despite strong theoretical properties.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Unpacking Positional Encoding in Transformers: A Spectral Analysis of Content-Position Coupling

URL: View paper

Prior Art Analysis

Spectral Analysis Positional[52] provides the same causal analysis demonstrating that the deposit pattern is intrinsic to RoPE's multiplicative structure. Both papers conduct identical ablation studies: (1) injecting absolute positional encodings to show the pattern re-emerges in deeper layers, (2) applying RoPE to subsets of heads to show minimal heads suffice, and (3) testing MLA architecture to show distributed alternatives. The candidate paper also provides theoretical gradient analysis linking multiplicative structure to the deposit pattern. The extensive overlap in experimental design, theoretical analysis, and conclusions demonstrates prior establishment of this causal relationship.

Evidence

Evidence 1 - **Rationale**: Both papers provide causal analysis linking RoPE's multiplicative structure to the deposit pattern through spectral properties. - **Original**: we conduct a causal analysis to demonstrate that the deposit pattern is an intrinsic property of rope's multiplicative architecture, offering a mechanistic explanation for its observed performance paradox. - **Candidate**: spectral contraction-manifested as a reduced eigenvalue range or lower condition number-is well known to accelerate and stabilize gradient-based learning [30, 26]. the spectral properties associated with rope's coupling mechanism, through this contraction, facilitate faster and more stable learning ...

Evidence 2 - **Rationale**: Both papers conduct the identical ablation experiment of adding absolute PE to RoPE to test if the deposit pattern is intrinsic. - **Original**: to test this, we inject a redundant signal by augmenting a rope-equipped transformer with an additional absolute pe at the input layer. this directly provides an explicit pi component before the rope-enabled attention layers. we then observe if this alters the deposit pattern. - **Candidate**: first, we augment rope by adding an absolute positional embedding before the first layer, which provides an initial position cue that can be adjusted by deeper layers.

Evidence 3 - **Rationale**: Both papers observe and report identical results: absolute PE disrupts the deposit pattern in layer 1, but it re-emerges in deeper layers, proving the pattern is intrinsic to RoPE. - **Original**: the results, shown in figure 5, are revealing. the explicit absolute pe signal disrupts the deposit pattern in the first layer, distributing positional responsibility more evenly. however, the pattern re-emerges in deeper layers. this strongly suggests that while an initial explicit signal can be ut... - **Candidate**: we present the result of the early method in figure 6. due to the absolute position encoding, the deposit pattern no longer appears in the first layer. ablating each head does not affect the overall performance. however, starting from the second layer, the deposit pattern reappears, which also indir...

Evidence 4 - **Rationale**: Both papers conduct the identical 'partial RoPE' ablation to demonstrate that minimal RoPE-enabled heads suffice, proving the deposit pattern is intrinsic. - **Original**: we test this by systematically reducing the number of attention heads that utilize rope, from all heads down to just one, with the rest operating as nope heads. we evaluate performance on task 1. - **Candidate**: we therefore conduct a 'partial rope' experiment, applying rope to k heads and nope to the rest. as figure 5 shows, even with only a handful of rope-enabled heads, task 1 performance remains below-near-optimal

Evidence 5 - **Rationale**: Both papers test MLA architecture and observe identical results: it mitigates the deposit pattern by distributing positional responsibility, confirming the pattern is specific to RoPE's structure. - **Original**: we replaced our standard attention with the mla module and evaluated it on both tasks. the mla architecture successfully mitigates the deposit pattern. figure 7 shows that no single head is indispensable, and positional responsibility is diffused across the model. - **Candidate**: crucially, head-wise ablation under mla (figure 7) shows no single head is indispensable for task 1, demonstrating effective mitigation of the earlier deposit pattern.

Appendix: Text Similarity Detection

Textual similarity detection checked 9 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Unpacking Positional Encoding in Transformers: A Spectral Analysis of Content-Position Coupling

Detected in: Contribution: contribution_1, Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Deconstructing Positional Information: From Attention Logits to Training Biases View paper
- [1] The impact of positional encoding on length generalization in transformers View paper
- [2] Graph representation learning via enhanced GNNs and transformers View paper
- [3] Position information in transformers: An overview View paper
- [4] Positional encoding in transformer-based time series models: a survey View paper
- [5] Rotary position embedding for vision transformer View paper
- [6] Randomized positional encodings boost length generalization of transformers View paper
- [7] Graphit: Encoding graph structure in transformers View paper
- [8] Aiatrack: Attention in attention for transformer visual tracking View paper
- [9] Enhancing multivariate time-series anomaly detection with positional encoding mechanisms in transformers View paper
- [10] An improved end-to-end multi-target tracking method based on transformer self-attention View paper
- [11] Novel positional encodings to enable tree-based transformers View paper
- [12] Self-positioning point-based transformer for point cloud understanding View paper
- [13] Understanding How Positional Encodings Work in Transformer Model View paper
- [14] Structural positional encoding for knowledge integration in transformer-based medical process monitoring and trace classification View paper
- [15] Positional description matters for transformers arithmetic View paper
- [16] Position embedding fusion on transformer for dense video captioning View paper
- [17] Positional knowledge is all you need: Position-induced transformer (PiT) for operator learning View paper
- [18] Understanding the Expressive Power and Mechanisms of Transformer for Sequence Modeling View paper
- [19] Rethinking positional encoding in language pre-training View paper
- [20] Two stones hit one bird: Bilevel positional encoding for better length extrapolation View paper
- [21] Positional encoding is not the same as context: A study on positional encoding for Sequential recommendation View paper
- [22] MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo View paper
- [23] HoPE: Hyperbolic Rotary Positional Encoding for Stable Long-Range Dependency Modeling in Large Language Models View paper
- [24] Positional Encoding via Token-Aware Phase Attention View paper
- [25] Novel Deepfake Image Detection with PV-ISM: Patch-Based Vision Transformer for Identifying Synthetic Media View paper
- [26] StyTr2: Image Style Transfer with Transformers View paper
- [27] Attention and positional encoding are (almost) all you need for shape matching View paper
- [28] 3DPPE: 3D Point Positional Encoding for Transformer-based Multi-Camera 3D Object Detection View paper
- [29] Self-adapted positional encoding in the transformer encoder for named entity recognition View paper
- [30] A Free Probabilistic Framework for Analyzing the Transformer-based Language Models View paper
- [31] Improving part-of-speech tagging with relative positional encoding in transformer models and basic rules View paper
- [32] Heterogeneous multivariate time series imputation by transformer model with missing position encoding View paper
- [33] Rethinking position embedding methods in the Transformer architecture View paper
- [34] Learning to encode position for transformer with continuous dynamical model View paper
- [35] Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers View paper
- [36] CovTiNet: Covid text identification network using attention-based positional embedding feature fusion View paper
- [37] VPDETR: End-to-End Vanishing Point DEtection TRansformers View paper
- [38] Rethinking and improving relative position encoding for vision transformer View paper

- [39] An Empirical Study on the Impact of Positional Encoding in Transformer-Based Monaural Speech Enhancement View paper
- [40] A simple and effective positional encoding for transformers View paper
- [41] MCT-Grasp: A Novel Grasp Detection Using Multimodal Embedding and Convolutional Modulation Transformer View paper
- [42] Reviving Shift Equivariance in Vision Transformers View paper
- [43] What do position embeddings learn? an empirical study of pre-trained language model positional encoding View paper
- [44] An Adaptive Transformer Model for Long-Term Time-Series Forecasting of Temperature and Radiation in Photovoltaic Energy Generation View paper
- [45] What Improves the Generalization of Graph Transformers? A Theoretical Dive into the Self-attention and Positional Encoding View paper
- [46] SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery View paper
- [47] Dynamic positional attention fusion (DPAF): Adaptive encoding and weighted attention for ship motion attitude prediction View paper
- [48] On solving textual ambiguities and semantic vagueness in MRC based question answering using generative pre-trained transformers View paper
- [49] Lightweight Hierarchical Transformer Combining Patch-Random and Positional Encoding for Respiratory Sound Classification View paper
- [50] HDTSLR: A Framework Based on Hierarchical Dynamic Positional Encoding for Sign Language Recognition View paper
- [51] Lightweight structure-aware attention for visual understanding View paper
- [52] Unpacking Positional Encoding in Transformers: A Spectral Analysis of Content-Position Coupling View paper
- [53] Context-aware Biases for Length Extrapolation View paper
- [54] Head-wise Adaptive Rotary Positional Encoding for Fine-Grained Image Generation View paper
- [55] Vision Transformer-Based Deepfake Detection: A Self-Attention Approach for Classification of Real and Synthetic Facial Images View paper
- [56] ComplexFormer: Disruptively Advancing Transformer Inference Ability via Head-Specific Complex Vector Attention View paper
- [57] Transformer with Syntactic Position Encoding for Machine Translation View paper