Novelty Assessment Report

Paper: FutureFill: Fast Generation from Convolutional Sequence Models
PDF URL: https://openreview.net/pdf?id=t5GUEuIsxR
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We address the challenge of efficient auto-regressive generation in sequence prediction models by introducing FutureFill—a general-purpose fast generation method for any sequence prediction algorithm based on convolutional operators. FutureFill reduces generation time from quadratic to quasilinear in the context length. Moreover, when generating from a prompt, it requires a prefill cache whose size grows only with the number of tokens to be generated—often much smaller than the caches required by standard convolutional or attention-based models. We validate our theoretical claims with language modeling experiments and demonstrate substantial efficiency gains when generating from a deep convolutional sequence prediction model.

Core Task Landscape

This paper addresses: **Efficient Auto-Regressive Generation from Convolutional Sequence Models**

A total of **26 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

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

- **Core Convolutional Sequence-to-Sequence Architectures**
- **Efficient Generation and Decoding Methods**
- **Sequential Decoding in Convolutional Coding Theory**
- **Polarization-Adjusted Convolutional (PAC) Codes**
- **Domain-Specific Applications of Convolutional Sequence Models**

Complete Taxonomy Tree

- Efficient Auto-Regressive Generation from Convolutional Sequence Models Survey Taxonomy
- Core Convolutional Sequence-to-Sequence Architectures
  - Fully Convolutional Encoder-Decoder Models (2 papers)
  - [5] Convolutional sequence to sequence learning (Jonas Gehring, 2017) View paper
  - [8] Sequence-to-sequence speech recognition with time-depth separable convolutions (Awni Hannun, 2019) View paper
  - Hybrid Convolutional-Recurrent Architectures (2 papers)
  - [2] Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research (Susmita Das, 2023) View paper
  - [4] Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network (Qihang Yao, 2020) View paper
  - Specialized Convolutional Sequence Models (2 papers)
  - [1] Short-term wind speed interval prediction using improved quality-driven loss based gated multi-scale convolutional sequence model (Adnan Saeed, 2024) View paper
  - [7] Locally hierarchical auto-regressive modeling for image generation (T You, 2022) View paper
- Efficient Generation and Decoding Methods
  - Parallel and Blockwise Decoding Strategies ★ (2 papers)
  - [0] FutureFill: Fast Generation from Convolutional Sequence Models (Anon et al., 2026) View paper
  - [6] Blockwise Parallel Decoding for Deep Autoregressive Models (Stern, 2018) View paper
  - Adaptive Inference Strategies (3 papers)
  - [9] Sequential Analysis with Specified Confidence Level and Adaptive Convolutional Neural Networks in Image Recognition (Andrey Savchenko, 2020) View paper
  - [19] Transactions of the Association for Computational Linguistics, Volume 6 (L Lee, 2018) View paper
  - [23] Progressive Transmission of High-Dimensional Data Features for Inference at the Network Edge (Qiao Lan, 2022) View paper
  - Low-Complexity and Low-Latency Architectures (2 papers)
  - [18] Efficient low-latency speech enhancement with mobile audio streaming networks (Romaniuk, 2020) View paper
  - [20] A low-cost serial decoder architecture for low-density parity-check convolutional codes (S. Bates, 2008) View paper
- Sequential Decoding in Convolutional Coding Theory
  - Sequential Decoding Algorithms and Metrics (3 papers)
  - [10] Sequential Decoding of Convolutional Codes for Synchronization Errors (Anisha Banerjee, 2022) View paper
  - [11] Sequential Decoding of Multiple Sequences for Synchronization Errors (Anisha Banerjee, 2024) View paper
  - [17] Convolutional codes III. Sequential decoding (G. David Forney, 1974) View paper
  - Convolutional Code Construction for Sequential Decoding (2 papers)
  - [14] Construction of convolutional codes for sequential decoding (Costello, 1969) View paper
  - [21] Nonsystematic convolutional codes for sequential decoding in space applications (James L. Massey, 2003) View paper

- Polarization-Adjusted Convolutional (PAC) Codes
  - List Decoding for PAC Codes (2 papers)
  - [15] Fast List Decoding of PAC Codes With Sequence Repetition Nodes (Zi Wei Zhao, 2023) View paper
  - [22] Fast List Decoders for Polarization-Adjusted Convolutional (PAC) Codes (Zhu Hongfei, 2023) View paper
  - Sequential Decoding and Metric Design for PAC Codes (2 papers)
  - [16] On Sequential Decoding Metric Function of Polarization-Adjusted Convolutional (PAC) Codes (Mohsen Moradi, 2021) View paper
  - [24] Polarization-Adjusted Convolutional (PAC) Codes: Sequential Decoding vs List Decoding (Mohammad Rowshan, 2021) View paper
- Domain-Specific Applications of Convolutional Sequence Models (5 papers)
  - [3] Chinese grammatical error correction based on convolutional sequence to sequence model (Si Li, 2019) View paper
  - [12] Deep learning approaches for automated seizure detection from scalp electroencephalograms (Meysam Golmohammadi, 2020) View paper
  - [13] Spatio-temporal Fourier dynamic graph convolution network for traffic forecasting (Longfei Hu, 2025) View paper
  - [25] Natural Language Correction-Thesis Proposal (NÃ¡plava, n.d.) View paper
  - [26] DEEP LEARNING FOR SCENE TEXT DETECTION IN NATURAL IMAGES: A COMPREHENSIVE REVIEW (Murali, n.d.) View paper

Narrative

Core task: efficient auto-regressive generation from convolutional sequence models. The field encompasses several distinct branches that address different facets of this challenge. Core Convolutional Sequence-to-Sequence Architectures explore foundational designs such as fully convolutional networks for sequence learning (Convolutional Sequence Learning[5]) and specialized temporal structures (Time-Depth Separable[8], Hierarchical Autoregressive[7]). Efficient Generation and Decoding Methods focus on accelerating inference through parallel and blockwise strategies (Blockwise Parallel Decoding[6]), which reduce the sequential bottleneck inherent in auto-regressive generation. Meanwhile, Sequential Decoding in Convolutional Coding Theory and Polarization-Adjusted Convolutional (PAC) Codes branches draw from information theory, examining decoding algorithms for error-correcting codes (Sequential Decoding[17], Fast List PAC[15], PAC List Decoders[22]) that share structural similarities with sequence generation. Domain-Specific Applications demonstrate how convolutional sequence models are adapted to tasks ranging from speech enhancement (Low-Latency Speech Enhancement[18]) to medical signal processing (Arrhythmia Detection[4], Seizure Detection[12]) and natural language correction (Chinese Grammar Correction[3]).

A particularly active line of work centers on reducing the computational cost of generating long sequences token-by-token. Blockwise Parallel Decoding[6] exemplifies efforts to predict multiple positions simultaneously, trading off some model flexibility for substantial speed gains. FutureFill[0] sits squarely within this Parallel and Blockwise Decoding Strategies cluster, proposing mechanisms to fill future tokens in blocks rather than strictly left-to-right. Compared to earlier convolutional sequence-to-sequence frameworks like Convolutional Sequence Learning[5], which established the viability of purely convolutional architectures, FutureFill[0] emphasizes inference-time efficiency and parallelism. Its approach contrasts with hierarchical or multi-scale methods (Hierarchical Autoregressive[7]) that decompose generation into coarse-to-fine stages, instead focusing on direct blockwise prediction. This positioning reflects broader tensions in the field between maintaining generation quality, preserving model simplicity, and achieving low-latency deployment—a balance that remains an open question as convolutional models compete with transformer-based alternatives.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Blockwise Parallel Decoding for Deep Autoregressive Models

Authors: Stern, Mitchell, Mitchell Stern, Shazeer, Noam, et al. (9 authors total) | **Year/Venue**: 2018 • Neural Information Processing Systems | **URL**: View paper

Abstract

Deep autoregressive sequence-to-sequence models have demonstrated impressive performance across a wide variety of tasks in recent years. While common architecture classes such as recurrent, convolutional, and self-attention networks make different trade-offs between the amount of computation needed per layer and the length of the critical path at training time, generation still remains an inherently sequential process. To overcome this limitation, we propose a novel blockwise parallel decoding s...

Relationship Analysis

Both papers belong to the Parallel and Blockwise Decoding Strategies category, addressing sequential generation bottlenecks in auto-regressive models. While FutureFill focuses on efficient online convolution for convolutional sequence models through divide-and-conquer FFT-based algorithms achieving $O(L \log^2 L)$ generation time, Blockwise Parallel Decoding takes a different approach by predicting multiple tokens in parallel and validating them with a scoring model, applicable to various architectures including self-attention models. The key distinction is that FutureFill provides exact generation with algorithmic improvements specific to convolutions, whereas Blockwise Parallel Decoding uses speculative parallel prediction with validation across broader architecture classes.

Contributions Analysis

Overall novelty summary. The paper introduces FutureFill, a method to accelerate auto-regressive generation from convolutional sequence models by reducing complexity from quadratic to quasilinear in context length. It resides in the 'Parallel and Blockwise Decoding Strategies' leaf of the taxonomy, which contains only two papers total. This leaf sits within the broader 'Efficient Generation and Decoding Methods' branch, indicating a relatively sparse research direction focused specifically on overcoming sequential generation bottlenecks through parallel prediction schemes.

The taxonomy reveals neighboring work in 'Adaptive Inference Strategies' (three papers on early termination and confidence-based stopping) and 'Low-Complexity and Low-Latency Architectures' (two papers on parameter-efficient designs). FutureFill diverges from adaptive methods by targeting fixed-complexity blockwise generation rather than dynamic stopping criteria. The broader 'Core Convolutional Sequence-to-Sequence Architectures' branch (six papers across three leaves) establishes foundational designs, while FutureFill addresses inference-time optimization rather than base architecture innovation. The taxonomy's scope explicitly excludes domain-specific applications and coding-theoretic sequential decoding, clarifying that this work targets general-purpose neural sequence generation.

Among thirty candidates examined, the analysis identifies one refutable candidate for the core FutureFill method (ten candidates examined), while the two algorithmic variants—Epoched-FutureFill and Continuous-FutureFill—show no clear refutations among ten candidates each. The single sibling paper in the same taxonomy leaf represents the most directly comparable prior work on blockwise parallel decoding. The limited search scope means these statistics reflect top-ranked semantic matches rather than exhaustive coverage,

suggesting the core contribution has at least one overlapping predecessor within the examined set, while the specific algorithmic trade-offs appear less explored.

Given the sparse taxonomy leaf and limited literature search, FutureFill appears to address a recognized but under-explored problem space. The presence of one refutable candidate for the main contribution indicates some prior work on blockwise generation exists, though the algorithmic variants show fewer direct precedents among examined papers. The analysis covers top-thirty semantic matches and does not claim exhaustive field coverage, leaving open whether additional related work exists beyond this scope.

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

Contribution 1: FutureFill method for fast generation from convolutional sequence models

Description: The authors propose FutureFill, a general method that reduces auto-regressive generation time in convolutional sequence models from quadratic to quasilinear complexity relative to context length. The method applies to any convolution-based sequence prediction algorithm.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards Effective and Efficient Non-autoregressive decoders for Conformer and LLM-based ASR using Block-based Attention Mask

URL: View paper

Brief Assessment

Non-Autoregressive Conformer[56] focuses on ASR decoder architectures using block-based attention masks for speech recognition, not on general methods for reducing autoregressive complexity in convolutional sequence models for generation tasks.

---

2. XSpecMesh: Quality-Preserving Auto-Regressive Mesh Generation Acceleration via Multi-Head Speculative Decoding

URL: View paper

Brief Assessment

XSpecMesh[49] focuses on accelerating auto-regressive mesh generation through multi-head speculative decoding, not on reducing complexity in convolutional sequence models. The technical approaches are fundamentally different.

---

3. Fastwave: Accelerating autoregressive convolutional neural networks on fpga

URL: View paper

Brief Assessment

FastWave[53] focuses on FPGA hardware acceleration for WaveNet inference through fixed-point implementation and parallel architectures, not on algorithmic methods to reduce autoregressive complexity from quadratic to quasilinear.

---

4. Convolutional state space models for long-range spatiotemporal modeling

URL: View paper

Brief Assessment

Convolutional State Space[48] focuses on spatiotemporal modeling with parallel scans for training efficiency, not on reducing autoregressive generation complexity from quadratic to quasilinear for general convolutional sequence models.

---

5. Fasttalker: A neural text-to-speech architecture with shallow and group autoregression

URL: View paper

Brief Assessment

FastTalker[52] focuses on text-to-speech synthesis using group autoregression for acoustic feature generation, not on general convolutional sequence model generation methods.

---

6. Inference Acceleration of Autoregressive Normalizing Flows by Selective Jacobi Decoding

URL: View paper

Brief Assessment

Selective Jacobi Decoding[54] focuses on accelerating autoregressive normalizing flows through parallel iterative optimization, not convolutional sequence models. The technical approaches are fundamentally different: FutureFill addresses convolution operations in sequence prediction, while this candidate addresses inference in normalizing flow architectures.

---

7. Convolutional Sequence Generation for Skeleton-Based Action Synthesis

URL: View paper

Brief Assessment

Skeleton Action Synthesis[50] focuses on generating skeleton-based action sequences using Gaussian processes and graph convolutions for spatial-temporal modeling, not on reducing autoregressive generation complexity in convolutional sequence models.

---

8. Seq-u-net: A one-dimensional causal u-net for efficient sequence modelling

URL: View paper

Brief Assessment

Seq-U-Net[55] focuses on architectural efficiency through multi-scale U-Net processing for sequence modeling, not on reducing autoregressive generation complexity through algorithmic methods like FutureFill.

---

9. Fast Generation for Convolutional Autoregressive Models

URL: View paper

Prior Art Analysis

Fast Convolutional Autoregressive[47] demonstrates that prior work exists on accelerating generation in convolutional autoregressive models through caching mechanisms to avoid redundant computations. While the original paper proposes FutureFill as a novel method reducing complexity from quadratic to quasilinear, the candidate paper shows that the fundamental approach of caching hidden states to speed up generation in convolutional models was already established. Both papers address the same core problem: the inefficiency of naive generation in convolutional sequence models where computations are unnecessarily repeated.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental problem: slow generation in convolutional autoregressive models due to redundant computations in naive implementations. Fast Convolutional Autoregressive[47] explicitly states they 'describe a method to

speed up generation in convolutional autoregressive models' before the original paper's submission, demonstrating prior work on this exact problem. - **Original**: we address the challenge of efficient auto-regressive generation in sequence prediction models by introducing futurefill-a method for fast generation that applies to any sequence prediction algorithm based on convolutional operators. our approach reduces the generation time from quadratic to quasili... - **Candidate**: convolutional autoregressive models have recently demonstrated state-of-the-art performance on a number of generation tasks. while fast, parallel training methods have been crucial for their success, generation is typically implemented in a naive fashion where redundant computations are unnecessaril...

Evidence 2 - **Rationale**: Fast Convolutional Autoregressive[47] demonstrates that the core technique of caching to avoid redundant computation in convolutional models was already established, achieving significant speedups (21x-183x) on production models. This shows that methods for fast generation from convolutional models existed prior to FutureFill's proposal. - **Original**: convolutional models offer a more general framework than ssms because they can represent any linear dynamical system (lds) without requiring parameters that scale with the dimensionality of the hidden states (agarwal et al., 2023). this flexibility has led to recent developments that can handle long... - **Candidate**: the key idea is to cache hidden states to avoid redundant computation. we apply our fast generation method to the wavenet and pixelcnn++ models and achieve up to $21\times$ and $183\times$ speedups respectively.

---

10. Lightspeech: Lightweight and fast text to speech with neural architecture search

URL: View paper

Brief Assessment

LightSpeech[51] focuses on text-to-speech synthesis using neural architecture search for model compression, not on reducing autoregressive generation complexity in convolutional sequence models. The candidate addresses a different problem domain (speech synthesis) with different technical approaches (NAS, model efficiency).

Contribution 2: Epoched-FutureFill algorithm with runtime-memory trade-off

Description: The authors develop Epoched-FutureFill, an algorithmic variant that offers a flexible trade-off between computational complexity and memory usage, achieving O(L^(3/2)√log L) runtime with O(√L log L) memory when generating L tokens from scratch.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Headinfer: Memory-efficient llm inference by head-wise offloading

URL: View paper

Brief Assessment

HeadInfer[39] focuses on memory-efficient LLM inference through head-wise KV cache offloading to CPU RAM, not on algorithmic variants for sequence generation with runtime-memory trade-offs in convolutional models.

---

2. MOM: Memory-Efficient Offloaded Mini-Sequence Inference for Long Context Language Models

URL: View paper

Brief Assessment

MOM[43] focuses on memory-efficient inference for long-context language models through layer partitioning and KV cache offloading, not on algorithmic variants for convolutional sequence generation with runtime-memory trade-offs.

---

3. El-attention: Memory efficient lossless attention for generation

URL: View paper

Brief Assessment

EL-Attention[45] focuses on memory-efficient attention mechanisms for transformer models by reducing cache requirements for key-value pairs in multi-head attention, not on convolutional sequence models or runtime-memory trade-offs for generation algorithms like Epoched-FutureFill.

---

4. Informer: Beyond efficient transformer for long sequence time-series forecasting

URL: View paper

Brief Assessment

Informer[41] focuses on efficient transformer architectures for time-series forecasting using ProbSparse self-attention, not on memory-efficient algorithms with runtime-memory trade-offs for sequence generation from convolutional models.

---

5. RAP: Runtime-Adaptive Pruning for LLM Inference

URL: View paper

Brief Assessment

RAP[44] focuses on runtime-adaptive pruning for LLM inference by dynamically adjusting compression strategies for model parameters and KV-cache, not on memory-efficient algorithms for convolutional sequence generation with runtime-memory trade-offs.

---

6. Flashattention: Fast and memory-efficient exact attention with io-awareness

URL: View paper

Brief Assessment

FlashAttention[38] focuses on IO-aware exact attention for transformers using tiling and recomputation to reduce memory accesses between GPU HBM and SRAM. It does not address memory-efficient algorithms with runtime-memory trade-offs for sequence generation from convolutional models, which is the focus of the original paper's Epoched-FutureFill contribution.

---

7. Time-and memory-efficient genome assembly with Raven

URL: View paper

Brief Assessment

Raven[37] focuses on genome assembly from long sequencing reads, not on sequence generation algorithms with runtime-memory trade-offs for token generation in language models or convolutional sequence prediction.

---

8. LightCache: Memory-Efficient, Training-Free Acceleration for Video Generation

URL: View paper

Brief Assessment

LightCache[46] focuses on memory-efficient video generation through cache management strategies (asynchronous swapping, feature chunking, latent slicing) for diffusion models, not algorithmic variants for sequence generation with runtime-memory trade-offs in convolutional models.

---

9. Time-memory-and parameter-efficient visual adaptation

URL: View paper

Brief Assessment

Time-Memory Visual Adaptation[40] focuses on efficient visual adaptation methods for foundation models using frozen backbones and parallel networks. It does not address memory-efficient algorithms for sequence generation with runtime-memory trade-offs as proposed in the original paper's Epoched-FutureFill.

---

10. Gear: An efficient kv cache compression recipe for near-lossless generative inference of llm

URL: View paper

Brief Assessment

GEAR[42] focuses on KV cache compression for LLM inference using quantization, low-rank approximation, and sparse matrices. It does not address algorithmic variants for sequence generation with runtime-memory trade-offs in convolutional models.

Contribution 3: Continuous-FutureFill algorithm for quasilinear generation

Description: The authors introduce Continuous-FutureFill, which achieves quasilinear O(L log^2 L) total generation time with O(L) memory for generating L tokens from scratch, and O(L log L + K log^2 K) time with O(K) cache when generating K tokens from a prompt of length L.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Orchid: Flexible and data-dependent convolution for sequence modeling

URL: View paper

Brief Assessment

Orchid[27] focuses on data-dependent convolution for sequence modeling with O(L log L) complexity via FFT, not on token generation algorithms for convolutional sequence predictors. The candidate addresses a different problem domain.

---

2. Quantification of uncertainty associated with evidence layers in mineral prospectivity mapping using direct sampling and convolutional neural network

URL: View paper

Brief Assessment

Uncertainty Mineral Prospectivity[31] focuses on uncertainty quantification in mineral prospectivity mapping using direct sampling and CNNs for geological applications, not on token generation algorithms or convolutional sequence predictors for language modeling.

---

3. Neural machine translation in linear time

URL: View paper

Brief Assessment

Linear Time Translation[32] focuses on linear-time sequence-to-sequence translation using dilated convolutions for machine translation tasks, not on quasilinear token generation algorithms with specific O(L log^2 L) complexity or cache optimization strategies for auto-regressive generation.

---

4. Evaluating the Impact of Applied Sampling Algorithm on CNN and LSTM Models for Credit Card Fraud Analysis

URL: View paper

Brief Assessment

Sampling Fraud Detection[36] focuses on credit card fraud detection using CNN and LSTM with sampling algorithms for class imbalance. It does not address token generation algorithms or convolutional sequence models for language tasks.

---

5. Combining deep learning with token selection for patient phenotyping from electronic health records

URL: View paper

Brief Assessment

Token Selection Phenotyping[35] focuses on patient phenotyping from electronic health records using deep learning and token selection mechanisms, not on quasilinear time algorithms for convolutional sequence model generation.

---

6. LCformer: Linear Convolutional Decomposed Transformer for Long-Term Series Forecasting

URL: View paper

Brief Assessment

LCformer[33] focuses on time series forecasting with linear attention and local convolution for approximately linear complexity, not on quasilinear token generation algorithms for convolutional sequence predictors as described in the original paper's Continuous-FutureFill contribution.

---

7. Ddctrack: Dynamic token sampling for efficient uav transformer tracking

URL: View paper

Brief Assessment

DDCTrack[28] focuses on dynamic token sampling for UAV visual tracking using convolutional architectures, not on quasilinear time algorithms for token generation in sequence prediction models.

---

8. Optimal Linear MAP Decoding of Convolutional Codes

URL: View paper

Brief Assessment

Optimal Linear MAP[30] focuses on MAP decoding of convolutional codes using shift registers for error correction, not on token generation algorithms for sequence prediction models. The technical domains are fundamentally different.

---

9. Conv-basis: A new paradigm for efficient attention inference and gradient computation in transformers

URL: View paper

Brief Assessment

Conv-Basis[29] focuses on convolution-based attention mechanisms for transformers, not on quasilinear token generation algorithms for convolutional sequence predictors. The candidate's context does not contain sufficient detail about generation time complexity or algorithms comparable to Continuous-FutureFill.

---

10. Correlation embedding learning with dynamic semantic enhanced sampling for knowledge graph completion

URL: View paper

Brief Assessment

Correlation Embedding Learning[34] focuses on knowledge graph completion using correlation embeddings and sampling strategies, not on token generation algorithms for convolutional sequence models. The technical domains are entirely different.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] FutureFill: Fast Generation from Convolutional Sequence Models View paper
- [1] Short-term wind speed interval prediction using improved quality-driven loss based gated multi-scale convolutional sequence model View paper
- [2] Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research View paper
- [3] Chinese grammatical error correction based on convolutional sequence to sequence model View paper
- [4] Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network View paper
- [5] Convolutional sequence to sequence learning View paper
- [6] Blockwise Parallel Decoding for Deep Autoregressive Models View paper
- [7] Locally hierarchical auto-regressive modeling for image generation View paper
- [8] Sequence-to-sequence speech recognition with time-depth separable convolutions View paper
- [9] Sequential Analysis with Specified Confidence Level and Adaptive Convolutional Neural Networks in Image Recognition View paper
- [10] Sequential Decoding of Convolutional Codes for Synchronization Errors View paper
- [11] Sequential Decoding of Multiple Sequences for Synchronization Errors View paper
- [12] Deep learning approaches for automated seizure detection from scalp electroencephalograms View paper
- [13] Spatio-temporal Fourier dynamic graph convolution network for traffic forecasting View paper
- [14] Construction of convolutional codes for sequential decoding View paper
- [15] Fast List Decoding of PAC Codes With Sequence Repetition Nodes View paper
- [16] On Sequential Decoding Metric Function of Polarization-Adjusted Convolutional (PAC) Codes View paper
- [17] Convolutional codes III. Sequential decoding View paper
- [18] Efficient low-latency speech enhancement with mobile audio streaming networks View paper
- [19] Transactions of the Association for Computational Linguistics, Volume 6 View paper
- [20] A low-cost serial decoder architecture for low-density parity-check convolutional codes View paper
- [21] Nonsystematic convolutional codes for sequential decoding in space applications View paper
- [22] Fast List Decoders for Polarization-Adjusted Convolutional (PAC) Codes View paper
- [23] Progressive Transmission of High-Dimensional Data Features for Inference at the Network Edge View paper
- [24] Polarization-Adjusted Convolutional (PAC) Codes: Sequential Decoding vs List Decoding View paper
- [25] Natural Language Correction-Thesis Proposal View paper
- [26] DEEP LEARNING FOR SCENE TEXT DETECTION IN NATURAL IMAGES: A COMPREHENSIVE REVIEW View paper
- [27] Orchid: Flexible and data-dependent convolution for sequence modeling View paper
- [28] Ddctrack: Dynamic token sampling for efficient uav transformer tracking View paper
- [29] Conv-basis: A new paradigm for efficient attention inference and gradient computation in transformers View paper
- [30] Optimal Linear MAP Decoding of Convolutional Codes View paper
- [31] Quantification of uncertainty associated with evidence layers in mineral prospectivity mapping using direct sampling and convolutional neural network View paper
- [32] Neural machine translation in linear time View paper
- [33] LCformer: Linear Convolutional Decomposed Transformer for Long-Term Series Forecasting View paper
- [34] Correlation embedding learning with dynamic semantic enhanced sampling for knowledge graph completion View paper
- [35] Combining deep learning with token selection for patient phenotyping from electronic health records View paper
- [36] Evaluating the Impact of Applied Sampling Algorithm on CNN and LSTM Models for Credit Card Fraud Analysis View paper
- [37] Time-and memory-efficient genome assembly with Raven View paper
- [38] Flashattention: Fast and memory-efficient exact attention with io-awareness View paper
- [39] Headinfer: Memory-efficient llm inference by head-wise offloading View paper
- [40] Time-memory-and parameter-efficient visual adaptation View paper
- [41] Informer: Beyond efficient transformer for long sequence time-series forecasting View paper
- [42] Gear: An efficient kv cache compression recipe for near-lossless generative inference of llm View paper
- [43] MOM: Memory-Efficient Offloaded Mini-Sequence Inference for Long Context Language Models View paper
- [44] RAP: Runtime-Adaptive Pruning for LLM Inference View paper
- [45] El-attention: Memory efficient lossless attention for generation View paper
- [46] LightCache: Memory-Efficient, Training-Free Acceleration for Video Generation View paper
- [47] Fast Generation for Convolutional Autoregressive Models View paper
- [48] Convolutional state space models for long-range spatiotemporal modeling View paper
- [49] XSpecMesh: Quality-Preserving Auto-Regressive Mesh Generation Acceleration via Multi-Head Speculative Decoding View paper
- [50] Convolutional Sequence Generation for Skeleton-Based Action Synthesis View paper
- [51] Lightspeech: Lightweight and fast text to speech with neural architecture search View paper
- [52] Fasttalker: A neural text-to-speech architecture with shallow and group autoregression View paper
- [53] Fastwave: Accelerating autoregressive convolutional neural networks on fpga View paper

- [54] Inference Acceleration of Autoregressive Normalizing Flows by Selective Jacobi Decoding View paper
- [55] Seq-u-net: A one-dimensional causal u-net for efficient sequence modelling View paper
- [56] Towards Effective and Efficient Non-autoregressive decoders for Conformer and LLM-based ASR using Block-based Attention Mask View paper