Novelty Assessment Report

Paper: Universal Approximation with Softmax Attention
PDF URL: https://openreview.net/pdf?id=8cj7ydwaaK
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-08

Abstract

We prove that with linear transformations, both (i) two-layer self-attention and (ii) one-layer self-attention followed by a softmax function are universal approximators for continuous sequence-to-sequence functions on compact domains. Our main technique is a new interpolation-based method for analyzing attention's internal mechanism. This leads to our key insight: self-attention is able to approximate a generalized version of ReLU to arbitrary precision, and hence subsumes many known universal approximators. Building on these, we show that two-layer multi-head attention or even one-layer multi-head attention followed by a softmax function suffices as a sequence-to-sequence universal approximator. In contrast, prior works rely on feed-forward networks to establish universal approximation in Transformers. Furthermore, we extend our techniques to show that, (softmax-)attention-only layers are capable of approximating gradient descent in-context. We believe these techniques hold independent interest.

Core Task Landscape

This paper addresses: **Universal Approximation of Attention Mechanisms for Sequence-to-Sequence Functions**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Theoretical Foundations of Attention and Transformer Expressiveness**
- **Architectural Variants and Enhancements**
- **Training Methodologies and Optimization**
- **Analysis and Interpretability of Attention Mechanisms**
- **Application Domains**

Complete Taxonomy Tree

- Universal Approximation of Attention Mechanisms for Sequence-to-Sequence Functions Survey Taxonomy
- Theoretical Foundations of Attention and Transformer Expressiveness
  - Universal Approximation Theory for Attention-Based Architectures ★ (7 papers)
  - [0] Universal Approximation with Softmax Attention (Anon et al., 2026) View paper
  - [4] Are Transformers universal approximators of sequence-to-sequence functions? (Yun, 2022) View paper
  - [23] A unified framework on the universal approximation of transformer-type architectures (J Cheng, 2025) View paper
  - [31] Universal Approximation Theorem for a Single-Layer Transformer (Gumaan, 2025) View paper
  - [33] Theory, Analysis, and Best Practices for Sigmoid Self-Attention (Ramapuram, 2024) View paper
  - [41] Prompting a Pretrained Transformer Can Be a Universal Approximator (Petrov, 2024) View paper
  - [50] Universal Approximation and Optimization Theory for Multi-Head Self-Attention: Theoretical Foundations and Scaling Laws (Kodela, 2024) View paper
  - Computational Power and Formal Language Capabilities (4 papers)
  - [7] Theoretical limitations of self-attention in neural sequence models (Hahn Michael, 2020) View paper
  - [9] The power of hard attention transformers on data sequences: a formal language theoretic perspective (Pascal BergstrÃ¤Ã er, 2024) View paper
  - [40] On the computational power of transformers and its implications in sequence modeling (Bhattamishra, 2020) View paper
  - [49] Your transformer may not be as powerful as you expect (Luo, 2022) View paper
  - Computational Complexity and Efficiency Analysis (3 papers)
  - [24] O (n) connections are expressive enough: Universal approximability of sparse transformers (Yun, 2020) View paper
  - [25] Sumformer: Universal Approximation for Efficient Transformers (Alberti, 2023) View paper
  - [42] Fundamental Limits of Prompt Tuning Transformers: Universality, Capacity and Efficiency (Hu, 2024) View paper
  - Expressive Power Mechanisms and Component Analysis (1 papers)
  - [13] Understanding the expressive power and mechanisms of transformer for sequence modeling (Weinan E, 2024) View paper
- Architectural Variants and Enhancements
  - Efficient Attention Architectures (3 papers)
  - [20] NystrÃ¶mformer: A nystrÃ¶m-based algorithm for approximating self-attention (Yunyang Xiong, 2021) View paper
  - [27] Contextual priority attention enables linear time sequence modeling in transformers (Karim Ben Khaled, 2025) View paper
  - [44] Diffuser: Efficient Transformers with Multi-hop Attention Diffusion for Long Sequences (Feng, 2022) View paper
  - Multi-Scale and Hierarchical Attention Mechanisms (2 papers)
  - [21] MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning (Zhao, 2022) View paper
  - [34] A Hierarchical Neural Network for Sequence-to-Sequences Learning (Zuo, 2022) View paper
  - Positional and Sequential Modeling Enhancements (1 papers)

- [26] Improving self-attention networks with sequential relations (Zaixiang Zheng, 2020) View paper
- Cross-Attention and Multi-Sequence Architectures (2 papers)
- [28] Layer-wise cross-view decoding for sequence-to-sequence learning (Fenglin Liu, 2020) View paper
- [43] A cross-attention transformer encoder for paired sequence data (Ceder Dens, 2023) View paper
- Convolutional and Hybrid Sequence-to-Sequence Models (2 papers)
- [2] Convolutional Sequence to Sequence Learning (Jonas Gehring, 2022) View paper
- [19] Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction (Maha Elbayad, 2022) View paper

• Training Methodologies and Optimization
- Exposure Bias and Scheduled Sampling Techniques (2 papers)
- [10] Improving Attention-based Sequence-to-sequence Models (Dou, 2022) View paper
- [36] Attention forcing for sequence-to-sequence model training (Qingyun Dou, 2019) View paper
- Variational and Probabilistic Attention (1 papers)
- [3] Variational Attention for Sequence-to-Sequence Models (Bahuleyan, 2022) View paper
- Attention Guidance and Constraint Methods (1 papers)
- [14] Guiding attention in Sequence-to-sequence models for Dialogue Act prediction (Colombo, 2022) View paper

• Analysis and Interpretability of Attention Mechanisms
- Error Detection and Analysis in Attention (1 papers)
- [12] Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech (Cassia Valentini-Botinhao, 2021) View paper
- Confidence Estimation for Sequence-to-Sequence Models (1 papers)
- [11] Confidence Estimation for Attention-Based Sequence-to-Sequence Models for Speech Recognition (Qiujia Li, 2021) View paper
- Attention Mechanism Behavior and Interpretability (1 papers)
- [18] An Analysis of" Attention" in Sequence-to-Sequence Models. (Rohit Prabhavalkar, 2017) View paper
- Explainability in Sequence-to-Sequence Learning (1 papers)
- [17] Introducing explainability in sequence-to-sequence learning for short-term load forecasting (GÃ¼rses-Tran, 2022) View paper

• Application Domains
- Natural Language Processing Applications (4 papers)
- [16] Bangla text generation system by incorporating attention in sequence-to-sequence model (Nayan Banik, 2022) View paper
- [38] Abstractive Gujarati Text Summarization Using Sequence-To-Sequence Model and Attention Mechanism (Shilpa Serasiya, 2025) View paper
- [39] A Review on Transformer Models: Applications, Taxonomies, Open Issues and Challenges (Nidhi Passi, 2024) View paper
- [48] Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning (Chen, 2022) View paper
- Speech Processing Applications (3 papers)
- [6] Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition (Linhao Dong, 2018) View paper
- [15] Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System (Seyed Reza Shahamiri, 2023) View paper
- [22] FastS2S-VC: Streaming Non-Autoregressive Sequence-to-Sequence Voice Conversion (Kameoka Hirokazu, 2022) View paper
- Computer Vision Applications (2 papers)
- [1] Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers (Sixiao Zheng, 2021) View paper
- [5] Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers (Li, 2021) View paper
- Time Series Forecasting and Prediction (6 papers)
- [8] State-of-charge estimation for lithium-ion batteries based on attentional sequence-to-sequence architecture (Yong Tian, 2023) View paper
- [29] Forecasting photovoltaic power production using a deep learning sequence to sequence model with attention (May Daniel, 2020) View paper
- [32] A Multistep Sequence-to-Sequence Model With Attention LSTM Neural Networks for Industrial Soft Sensor Application (Lianwei Ma, 2023) View paper
- [35] An attention-based Bayesian sequence to sequence model for short-term solar power generation prediction within decomposition-ensemble strategy (Fei Xiao, 2023) View paper
- [37] A CNN-Sequence-to-Sequence network with attention for residential short-term load forecasting (Mosbah Aouad, 2022) View paper
- [46] Attention-based sequence to sequence model for machine remaining useful life prediction (Mohamed Ragab, 2021) View paper
- Time Series Imputation (1 papers)
- [45] Attention-Based Sequence-to-Sequence Model for Time Series Imputation (Yurui Li, 2022) View paper
- Audio and Music Processing (1 papers)
- [30] Sequence-to-Sequence Piano Transcription with Transformers (Hawthorne, 2022) View paper
- Sensor Data and Activity Recognition (1 papers)
- [47] Wearable Sensor Based Human Activity Recognition with Transformer (Iveta DirgovÃ¡ LuptÃ¡kovÃ¡, 2022) View paper

Narrative

Core task: universal approximation of attention mechanisms for sequence-to-sequence functions. The field structure reflects a balance between rigorous theoretical inquiry and practical deployment. The taxonomy organizes work into several main branches: Theoretical Foundations of Attention and Transformer Expressiveness examines the representational capacity and computational limits of attention-based architectures, often drawing on universal approximation results (e.g., Transformers Universal Approximators[4], Computational Power Transformers[40]); Architectural Variants and Enhancements explores modifications such as sparse patterns (Sparse Transformers[24]), alternative attention formulations (Sigmoid Self-Attention[33]), and efficiency improvements (Nystromformer[20]); Training Methodologies and Optimization addresses learning dynamics and convergence; Analysis and Interpretability of Attention Mechanisms investigates what attention weights reveal about model behavior (Analyzing Attention[18]); and Application Domains spans speech (Speech Transformer[6]), vision (Semantic Segmentation Transformers[1]), forecasting (Solar Power Prediction[35]), and text generation tasks.

A particularly active line of work centers on understanding the expressive power of transformers and their components, with studies probing whether single-layer or simplified architectures suffice for certain function classes (Single-Layer Transformer[31], Transformer Expressive Power[13]) and whether prompting alone can achieve universal approximation (Prompting Universal Approximator[41],

Prompt Tuning Limits[42]). Universal Approximation Softmax[0] sits squarely within this theoretical branch, contributing formal guarantees on the capacity of softmax-based attention to approximate sequence-to-sequence mappings. Its emphasis on foundational approximation properties aligns closely with Transformers Universal Approximators[4] and Multi-Head Self-Attention Theory[50], which similarly establish representational bounds, yet it contrasts with works like Hard Attention Transformers[9] or Variational Attention[3] that modify the attention mechanism itself. By anchoring its analysis in classical approximation theory, Universal Approximation Softmax[0] helps clarify which architectural features are essential for expressiveness and which are primarily optimization or efficiency concerns.

Related Works in Same Category

The following **6 sibling papers** share the same taxonomy leaf node with the original paper:

1. Are Transformers universal approximators of sequence-to-sequence functions?

Authors: Yun, Chulhee, Bhojanapalli, Srinadh, Chulhee Yun, et al. (14 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Despite the widespread adoption of Transformer models for NLP tasks, the expressive power of these models is not well-understood. In this paper, we establish that Transformer models are universal approximators of continuous permutation equivariant sequence-to-sequence functions with compact support, which is quite surprising given the amount of shared parameters in these models. Furthermore, using positional encodings, we circumvent the restriction of permutation equivariance, and show that Tran...

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. The original paper focuses on proving that softmax attention alone (without feed-forward networks) can achieve universal approximation through an interpolation-based method that shows attention approximates generalized ReLUs, requiring only one or two attention layers. The candidate paper proves universal approximation for Transformers including both self-attention and feed-forward layers, demonstrating that fixed-width self-attention computes contextual mappings while feed-forward layers contribute to the overall approximation property, and additionally addresses permutation equivariance constraints through positional encodings.

---

2. A unified framework on the universal approximation of transformer-type architectures

Authors: J Cheng, Q Li, T Lin, Z Shen | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ We investigate the universal approximation property (UAP) â¦ networks to models incorporating attention mechanisms. Our â¦ any continuous sequence-to-sequence function over compact â¦

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. The original paper focuses on proving that softmax attention with linear transformations (one or two layers) can approximate continuous functions through an interpolation-based technique that shows attention approximates generalized ReLUs, while the candidate paper develops a unified framework based on token distinguishability conditions that applies to diverse transformer architectures (kernel-based, sparse attention, etc.) and incorporates permutation equivariance under group actions. The key difference is that the original paper provides a constructive interpolation method for minimal attention-only architectures, whereas the candidate paper offers non-constructive sufficient conditions applicable to broader transformer variants with different attention mechanisms.

---

3. Universal Approximation Theorem for a Single-Layer Transformer

Authors: Esmail Gumaan | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Deep learning employs multi-layer neural networks trained via the backpropagation algorithm. This approach has achieved success across many domains and relies on adaptive gradient methods such as the Adam optimizer. Sequence modeling evolved from recurrent neural networks to attention-based models, culminating in the Transformer architecture. Transformers have achieved state-of-the-art performance in natural language processing (for example, BERT and GPT-3) and have been applied in computer visi...

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. The original paper focuses on proving that softmax attention with linear transformations (either two-layer self-attention or one-layer self-attention followed by softmax) can approximate continuous functions using a novel interpolation-based method that shows attention approximates generalized ReLUs. The candidate paper proves that a single-layer Transformer (one self-attention layer plus a feed-forward network) is a universal approximator, but critically relies on the feed-forward network component in addition to attention, whereas the original paper demonstrates universality using attention mechanisms alone without requiring feed-forward layers.

---

4. Theory, Analysis, and Best Practices for Sigmoid Self-Attention

Authors: Ramapuram, Jason, Danieli, Federico, Jason Ramapuram, et al. (33 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Attention is a key part of the transformer architecture. It is a sequence-to-sequence mapping that transforms each sequence element into a weighted sum of values. The weights are typically obtained as the softmax of dot products between keys and queries. Recent work has explored alternatives to softmax attention in transformers, such as ReLU and sigmoid activations. In this work, we revisit sigmoid attention and conduct an in-depth theoretical and empirical analysis. Theoretically, we prove that...

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. The original paper focuses on softmax attention with linear transformations, proving that two-layer self-attention or one-layer self-attention followed by softmax suffices for universal approximation using an interpolation-based technique. The candidate paper examines sigmoid attention as an alternative to softmax, proving universal approximation properties for transformers with sigmoid attention while also analyzing regularity properties and providing practical implementation improvements through hardware-aware kernels.

---

5. Prompting a Pretrained Transformer Can Be a Universal Approximator

Authors: Petrov, Aleksandar, Aleksandar Petrov, Torr, Philip H. S., et al. (9 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Despite the widespread adoption of prompting, prompt tuning and prefix-tuning of transformer models, our theoretical understanding of these fine-tuning methods remains limited. A key question is whether one can arbitrarily modify the behavior of pretrained model by prompting or prefix-tuning it. Formally, whether prompting and prefix-tuning a pretrained model can universally approximate sequence-to-sequence functions. This paper answers in the affirmative and demonstrates that much smaller pretr...

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. The original paper focuses on demonstrating that softmax attention with linear transformations alone (without feed-forward networks) can achieve universal approximation through an interpolation-based technique that approximates generalized ReLUs, while the candidate paper examines universal approximation through prefix-tuning of pretrained transformers, showing that a single attention head can approximate functions on hyperspheres and extending this to sequence-to-sequence settings. The key difference lies in their approach: the original paper analyzes the intrinsic approximation power of attention mechanisms themselves, whereas the candidate paper studies how prefixing (modifying input context without changing model parameters) enables universal approximation in pretrained models.

---

6. Universal Approximation and Optimization Theory for Multi-Head Self-Attention: Theoretical Foundations and Scaling Laws

Authors: KC Kodela | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ Our contributions include: 1) Universal Approximation Theory: We prove multihead self-attention â¦ efficiency advantages over feedforward networks for sequence-tosequence mappings. â¦

Relationship Analysis

Both papers belong to the Universal Approximation Theory for Attention-Based Architectures category, establishing formal proofs that attention mechanisms can universally approximate sequence-to-sequence functions. They share the core focus of proving universal approximation capabilities of multi-head self-attention without relying on feedforward networks. The key difference is that the original paper develops an interpolation-based method showing that two-layer or one-layer attention with softmax suffices for universal approximation with O(1/n) to O(1/(nH)) precision, while the candidate paper appears to additionally address optimization theory and scaling laws beyond the pure approximation guarantees.

Contributions Analysis

Overall novelty summary. The paper establishes that two-layer self-attention and one-layer self-attention followed by softmax can universally approximate continuous sequence-to-sequence functions on compact domains. It resides in the 'Universal Approximation Theory for Attention-Based Architectures' leaf, which contains seven papers total. This leaf sits within the broader 'Theoretical Foundations' branch, indicating a moderately populated research direction focused on formal expressiveness guarantees. The taxonomy shows this is a core theoretical area with active inquiry into what minimal architectural components suffice for universal approximation.

The taxonomy reveals neighboring leaves examining computational power and formal language capabilities, computational complexity analysis, and expressive power mechanisms. The paper's focus on approximation guarantees distinguishes it from complexity-oriented work and from studies of Turing-completeness or formal language recognition. Its sibling papers in the same leaf explore related universal approximation questions, suggesting a coherent research thread investigating which transformer components are theoretically necessary. The taxonomy structure indicates this theoretical branch is well-developed but not overcrowded, with clear boundaries separating approximation theory from architectural design and empirical applications.

Among twenty-two candidates examined across three contributions, none were found to clearly refute the paper's claims. The interpolation-based analysis method examined ten candidates with zero refutations; the generalized ReLU approximation result examined two candidates with zero refutations; and the two-layer sufficiency claim examined ten candidates with zero refutations. This limited search scope suggests that within the top semantic matches and citation neighborhood, no prior work directly establishes the same results. However, the modest candidate pool means the analysis cannot rule out relevant prior work outside this search radius.

Given the limited literature search covering twenty-two candidates, the paper appears to occupy a relatively novel position within its theoretical niche. The absence of refutable prior work among examined candidates, combined with the moderately populated taxonomy leaf, suggests the specific combination of techniques and results may be new. However, the search scope leaves open the possibility of related approximation results in adjacent theoretical areas not captured by semantic similarity or immediate citation links.

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

Contribution 1: Interpolation-based method for analyzing attention's internal mechanism

Description: The authors introduce a novel interpolation selection technique that partitions the target function's output range into uniform anchors, embeds them into attention's key-query-value transformations, and uses softmax to approximate argmax-style selection. This method demonstrates that attention can simulate piecewise linear behavior without relying on auxiliary feed-forward layers.

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. Exact Sequence Interpolation with Transformers

URL: View paper

Brief Assessment

Exact Sequence Interpolation[63] focuses on exact sequence-to-sequence interpolation for transformers using geometric constructions and clustering effects, not on analyzing attention mechanisms through interpolation selection techniques as described in the original paper.

---

2. From interpolation to extrapolation: Complete length generalization for arithmetic transformers

URL: View paper

Brief Assessment

Length Generalization Arithmetic[67] focuses on arithmetic tasks and length generalization using attention bias calibration, not on proving universal approximation properties of attention mechanisms through interpolation-based analysis of key-query-value transformations.

---

3. Video Frame Interpolation Transformer

URL: View paper

Brief Assessment

Video Frame Interpolation[60] focuses on video frame interpolation using transformers for computer vision tasks, not on analyzing attention mechanisms' internal workings or universal approximation theory. The interpolation here refers to generating intermediate video frames, not to the mathematical interpolation technique for attention analysis.

---

4. Assigning channel weights using an attention mechanism: an EEG interpolation algorithm

URL: View paper

Brief Assessment

EEG Interpolation Attention[61] focuses on EEG signal reconstruction using attention mechanisms for channel weight assignment in neuroscience applications, not on analyzing transformer attention's internal mechanisms or universal approximation theory.

---

5. Performance of imputation techniques: A comprehensive simulation study using the transformer model

URL: View paper

Brief Assessment

Imputation Techniques Performance[59] focuses on missing data imputation methods (LOCF, NOCB, mean imputation, linear interpolation, etc.) for time series analysis, not on analyzing attention mechanisms in transformers. The paper does not address attention's internal mechanism or key-query-value transformations.

---

6. Parallel spatio-temporal attention transformer for video frame interpolation

URL: View paper

Brief Assessment

Parallel Spatio-Temporal Attention[66] focuses on video frame interpolation using transformers for spatio-temporal feature aggregation in computer vision, not on analyzing attention mechanisms' internal workings through interpolation selection techniques for universal approximation theory.

---

7. Extending context window of large language models via positional interpolation

URL: View paper

Brief Assessment

Positional Interpolation[64] focuses on extending context windows in LLMs through position index interpolation for RoPE encodings, not on analyzing attention's internal mechanism through interpolation selection of key-query-value transformations as described in the original contribution.

---

8. A Spatial Downscaling Approach for Enhanced Accuracy in High Wind Speed Estimation Using Hybrid Attention Transformer

URL: View paper

Brief Assessment

Wind Speed Estimation[68] focuses on spatial downscaling for wind speed prediction using interpolation methods in a different context (meteorological data processing), not on analyzing attention mechanisms' internal workings or universal approximation theory.

---

9. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation

URL: View paper

Brief Assessment

Inter-Frame Attention[65] focuses on video frame interpolation using attention for motion and appearance extraction in computer vision, not on analyzing attention mechanisms in transformers for universal approximation theory.

---

10. Enhancing video frame interpolation with region of motion loss and self-attention mechanisms: A dual approach to address large, nonlinear motions

URL: View paper

Brief Assessment

Region Motion Loss[62] focuses on video frame interpolation using self-attention for motion analysis in computer vision, not on theoretical analysis of attention mechanisms in transformers or universal approximation theory.

Contribution 2: Single-head and multi-head attention approximate generalized ReLUs

Description: The authors prove that single-head attention approximates n generalized ReLU functions (truncated linear models) with O(1/n) precision, and that H-head attention improves this to O(1/(nH)) precision. This establishes that attention mechanisms can replicate the behavior of known universal approximators like ReLU networks.

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. Visual Analytics for Taxi Dispatching Based on Multi-Agent Reinforcement Learning

URL: View paper

Brief Assessment

Taxi Dispatching Analytics[52] focuses on multi-agent reinforcement learning for taxi dispatching systems, not on theoretical analysis of attention mechanisms approximating ReLU functions with precision bounds.

---

2. Advancing Deep Learning for Multiagent AI: Mechanisms, Organizations, and Dynamics

URL: View paper

Brief Assessment

Multiagent Deep Learning[51] focuses on multiagent AI systems and deep learning mechanisms for agent coordination. The candidate's full text context contains only fragments about approximation errors and network performance, with no substantive discussion of attention mechanisms approximating ReLU functions or establishing precision bounds like O(1/n) or O(1/(nH)).

Contribution 3: Two-layer attention suffices for sequence-to-sequence universal approximation

Description: The authors demonstrate that either two stacked attention layers or one attention layer followed by softmax can universally approximate any continuous sequence-to-sequence function on compact domains. This result shows that attention alone provides the core expressiveness without requiring feed-forward networks or deep attention stacks.

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. Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input

URL: View paper

Brief Assessment

Infinite Dimensional Input[58] focuses on approximation of shift-equivariant sequence-to-sequence functions with infinite dimensional inputs under smoothness assumptions (anisotropic/mixed smoothness), not universal approximation of arbitrary continuous functions on compact domains. The technical approaches and problem settings differ fundamentally.

---

2. Approximation rate of the transformer architecture for sequence modeling

URL: View paper

Brief Assessment

Approximation Rate Transformer[54] focuses on Jackson-type approximation rates with explicit complexity measures for temporal structures, not on minimal-layer universal approximation results. The candidate establishes rates governed by low-rank structure in pairwise coupling, while the original demonstrates that attention alone (without feed-forward networks) achieves universality with two layers.

---

3. Big bird: Transformers for longer sequences

URL: View paper

Brief Assessment

Big Bird[56] focuses on sparse attention mechanisms for handling longer sequences in transformers, not on universal approximation properties of attention layers. The paper does not address whether two stacked attention layers or one attention layer followed by softmax can universally approximate sequence-to-sequence functions.

---

4. Prompting a Pretrained Transformer Can Be a Universal Approximator

URL: View paper

Brief Assessment

Prompting Universal Approximator[41] focuses on prefix-tuning of pretrained transformers for universal approximation, where the model parameters are fixed and only prefixes are modified. The original paper studies universal approximation of attention mechanisms themselves without prefix-tuning constraints.

---

5. Are Transformers universal approximators of sequence-to-sequence functions?

URL: View paper

Brief Assessment

Transformers Universal Approximators[4] focuses on establishing universal approximation using self-attention combined with feed-forward layers, not on demonstrating that attention alone (without FFN) suffices for universal approximation as claimed in the original contribution.

---

6. Attention Mechanism, Max-Affine Partition, and Universal Approximation

URL: View paper

Brief Assessment

Attention Max-Affine Partition[53] focuses on single-layer, single-head attention achieving universal approximation through max-affine partitioning. The original paper demonstrates two-layer attention or one-layer plus softmax suffices, while [53] shows even single-layer single-head attention alone can achieve this, representing a different (and potentially stronger) architectural result rather than refuting the novelty of the two-layer claim.

---

7. STaRFormer: Semi-Supervised Task-Informed Representation Learning via Dynamic Attention-Based Regional Masking for Sequential Data

URL: View paper

Brief Assessment

STaRFormer[57] focuses on practical sequential modeling for intent prediction with spatiotemporal data, not on theoretical universal approximation properties of attention mechanisms.

---

8. Towards understanding the universality of transformers for next-token prediction

URL: View paper

Brief Assessment

Universality Next-Token Prediction[55] focuses on next-token prediction in autoregressive sequences ($x_{t+1} = f(x_t)$), not general sequence-to-sequence universal approximation. The candidate's framework and theoretical results address a different problem setting than the original paper's universal approximation claims for arbitrary continuous sequence-to-sequence functions.

---

9. Sumformer: Universal Approximation for Efficient Transformers

URL: View paper

Brief Assessment

Sumformer[25] focuses on efficient transformers (Linformer, Performer) and proves that one attention layer suffices for universal approximation, whereas the original paper demonstrates that two stacked attention layers or one attention layer followed by softmax can universally approximate sequence-to-sequence functions. The technical approaches differ: Sumformer[25] uses multisymmetric polynomials and piecewise constant approximations, while the original paper uses an interpolation-based method with attention as a selection mechanism.

---

10. A unified framework on the universal approximation of transformer-type architectures

URL: View paper

Brief Assessment

Unified Transformer Framework[23] focuses on establishing universal approximation through token distinguishability conditions and group-equivariant architectures, rather than analyzing the minimal layer requirements for attention-only mechanisms. The candidate's framework requires both attention and feedforward layers working together, whereas the original contribution demonstrates that attention layers alone (without feedforward networks) provide core expressiveness.

Appendix: Text Similarity Detection

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

References

- [0] Universal Approximation with Softmax Attention View paper
- [1] Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers View paper
- [2] Convolutional Sequence to Sequence Learning View paper
- [3] Variational Attention for Sequence-to-Sequence Models View paper
- [4] Are Transformers universal approximators of sequence-to-sequence functions? View paper
- [5] Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers View paper
- [6] Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition View paper
- [7] Theoretical limitations of self-attention in neural sequence models View paper
- [8] State-of-charge estimation for lithium-ion batteries based on attentional sequence-to-sequence architecture View paper
- [9] The power of hard attention transformers on data sequences: a formal language theoretic perspective View paper
- [10] Improving Attention-based Sequence-to-sequence Models View paper
- [11] Confidence Estimation for Attention-Based Sequence-to-Sequence Models for Speech Recognition View paper
- [12] Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech View paper
- [13] Understanding the expressive power and mechanisms of transformer for sequence modeling View paper
- [14] Guiding attention in Sequence-to-sequence models for Dialogue Act prediction View paper
- [15] Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System View paper
- [16] Bangla text generation system by incorporating attention in sequence-to-sequence model View paper
- [17] Introducing explainability in sequence-to-sequence learning for short-term load forecasting View paper
- [18] An Analysis of" Attention" in Sequence-to-Sequence Models. View paper
- [19] Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction View paper
- [20] NystrÃ¶mformer: A nystrÃ¶m-based algorithm for approximating self-attention View paper
- [21] MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning View paper
- [22] FastS2S-VC: Streaming Non-Autoregressive Sequence-to-Sequence Voice Conversion View paper
- [23] A unified framework on the universal approximation of transformer-type architectures View paper
- [24] O (n) connections are expressive enough: Universal approximability of sparse transformers View paper
- [25] Sumformer: Universal Approximation for Efficient Transformers View paper
- [26] Improving self-attention networks with sequential relations View paper
- [27] Contextual priority attention enables linear time sequence modeling in transformers View paper
- [28] Layer-wise cross-view decoding for sequence-to-sequence learning View paper
- [29] Forecasting photovoltaic power production using a deep learning sequence to sequence model with attention View paper
- [30] Sequence-to-Sequence Piano Transcription with Transformers View paper
- [31] Universal Approximation Theorem for a Single-Layer Transformer View paper
- [32] A Multistep Sequence-to-Sequence Model With Attention LSTM Neural Networks for Industrial Soft Sensor Application View paper
- [33] Theory, Analysis, and Best Practices for Sigmoid Self-Attention View paper
- [34] A Hierarchical Neural Network for Sequence-to-Sequences Learning View paper
- [35] An attention-based Bayesian sequence to sequence model for short-term solar power generation prediction within decomposition-ensemble strategy View paper
- [36] Attention forcing for sequence-to-sequence model training View paper
- [37] A CNN-Sequence-to-Sequence network with attention for residential short-term load forecasting View paper
- [38] Abstractive Gujarati Text Summarization Using Sequence-To-Sequence Model and Attention Mechanism View paper
- [39] A Review on Transformer Models: Applications, Taxonomies, Open Issues and Challenges View paper
- [40] On the computational power of transformers and its implications in sequence modeling View paper
- [41] Prompting a Pretrained Transformer Can Be a Universal Approximator View paper
- [42] Fundamental Limits of Prompt Tuning Transformers: Universality, Capacity and Efficiency View paper
- [43] A cross-attention transformer encoder for paired sequence data View paper
- [44] Diffuser: Efficient Transformers with Multi-hop Attention Diffusion for Long Sequences View paper
- [45] Attention-Based Sequence-to-Sequence Model for Time Series Imputation View paper
- [46] Attention-based sequence to sequence model for machine remaining useful life prediction View paper
- [47] Wearable Sensor Based Human Activity Recognition with Transformer View paper
- [48] Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning View paper
- [49] Your transformer may not be as powerful as you expect View paper
- [50] Universal Approximation and Optimization Theory for Multi-Head Self-Attention: Theoretical Foundations and Scaling Laws View paper
- [51] Advancing Deep Learning for Multiagent AI: Mechanisms, Organizations, and Dynamics View paper
- [52] Visual Analytics for Taxi Dispatching Based on Multi-Agent Reinforcement Learning View paper
- [53] Attention Mechanism, Max-Affine Partition, and Universal Approximation View paper
- [54] Approximation rate of the transformer architecture for sequence modeling View paper
- [55] Towards understanding the universality of transformers for next-token prediction View paper
- [56] Big bird: Transformers for longer sequences View paper
- [57] STaRFormer: Semi-Supervised Task-Informed Representation Learning via Dynamic Attention-Based Regional Masking for Sequential Data View paper
- [58] Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input View paper
- [59] Performance of imputation techniques: A comprehensive simulation study using the transformer model View paper
- [60] Video Frame Interpolation Transformer View paper
- [61] Assigning channel weights using an attention mechanism: an EEG interpolation algorithm View paper

- [62] Enhancing video frame interpolation with region of motion loss and self-attention mechanisms: A dual approach to address large, nonlinear motions View paper
- [63] Exact Sequence Interpolation with Transformers View paper
- [64] Extending context window of large language models via positional interpolation View paper
- [65] Extracting motion and appearance via inter-frame attention for efficient video frame interpolation View paper
- [66] Parallel spatio-temporal attention transformer for video frame interpolation View paper
- [67] From interpolation to extrapolation: Complete length generalization for arithmetic transformers View paper
- [68] A Spatial Downscaling Approach for Enhanced Accuracy in High Wind Speed Estimation Using Hybrid Attention Transformer View paper