Novelty Assessment Report

Paper: Continuum Transformers Perform In-Context Learning by Operator Gradient Descent

PDF URL: https://openreview.net/pdf?id=X63V2CWjj3

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Transformers robustly exhibit the ability to perform in-context learning, whereby their predictive accuracy on a task can increase not by parameter updates but merely with the placement of training samples in their context windows. Recent works have shown that transformers achieve this by implementing gradient descent in their forward passes. Such results, however, are restricted to standard transformer architectures, which handle finite-dimensional inputs. In the space of PDE surrogate modeling, a generalization of transformers to handle infinite-dimensional function inputs, known as "continuum transformers," has been proposed and similarly observed to exhibit in-context learning. Despite impressive empirical performance, such in-context learning has yet to be theoretically characterized. We herein demonstrate that continuum transformers perform in-context operator learning by performing gradient descent in an operator RKHS. We demonstrate this using novel proof strategies that leverage a generalized representer theorem for Hilbert spaces and gradient flows over the space of functionals of a Hilbert space. We additionally show the operator learned in context is the Bayes Optimal Predictor in the infinite depth limit of the transformer. We then provide empirical validations of this optimality result and demonstrate that the parameters under which such gradient descent is performed are recovered through the continuum transformer training.

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: **In-Context Operator Learning via Gradient Descent in Operator RKHS**

A total of **19 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Transformer-Based In-Context Operator Learning**
- **Operator Learning via Stochastic Gradient Descent**
- **Neural Network Operator Learning**
- **Gradient Descent Algorithms in RKHS**
- **Reinforcement Learning and Control in RKHS**
- **Nonparametric Differential Equation Learning**

Complete Taxonomy Tree

- In-Context Operator Learning via Gradient Descent in Operator RKHS Survey Taxonomy
- Transformer-Based In-Context Operator Learning
  - Continuum and Operator-Space Transformers ★ (2 papers)
  - [0] Continuum Transformers Perform In-Context Learning by Operator Gradient Descent (Anon et al., 2026) View paper
  - [17] In-Context Fine-Tuning for Neural Operators (Y Patel, n.d.) View paper
  - Finite-Dimensional In-Context Learning (3 papers)
  - [3] In-context learning with transformers: Softmax attention adapts to function lipschitzness (Liam Collins, 2024) View paper
  - [13] Transformers Can Perform Distributionally-robust Optimisation through In-context Learning (T Kim, 2024) View paper
  - [18] Functional Gradients and Generalizations for Transformer In-Context Learning (AT Wang, n.d.) View paper
  - Kernel-Based NLP Applications (1 papers)
  - [19] Support Vector Generation: Kernelizing Large Language Models for Efficient Zeroâ□□Shot NLP (Ohsawa, n.d.) View paper
- Operator Learning via Stochastic Gradient Descent
  - General Hilbert Space Operator Learning (1 papers)
  - [1] Learning operators with stochastic gradient descent in general hilbert spaces (Shi Lei, 2024) View paper
  - Operator-Valued Kernel Regression (1 papers)
  - [7] Learning Operators by Regularized Stochastic Gradient Descent with Operator-valued Kernels (Yang, 2025) View paper
- Neural Network Operator Learning
  - Neural Tangent Kernel Regime Analysis (1 papers)
  - [10] Optimal convergence rates for neural operators (Nguyen, 2024) View paper
  - Sequential and Transfer Learning Perspectives (2 papers)
  - [5] Gradient descent in neural networks as sequential learning in rkbs (Shilton, 2023) View paper
  - [6] Deep transfer operator learning for partial differential equations under conditional shift (S. Goswami, 2022) View paper
  - Meta-Learning Functional Transduction (1 papers)
  - [9] Learning functional transduction (Chalvidal, 2023) View paper
- Gradient Descent Algorithms in RKHS
  - Distributed and Spectral Regularization Methods (2 papers)
  - [8] Distributed kernel-based gradient descent algorithms (Shaobo Lin, 2018) View paper

    - [11] Optimal rates for vector-valued spectral regularization learning algorithms (Arthur Gretton, 2024) View paper
    - Importance Labeling and Active Learning (1 papers)
    - [14] Gradient Descent in RKHS with Importance Labeling (Tomoya Murata, 2020) View paper
    - Projected Mirror Descent in RKHS (1 papers)
    - [12] Projected Pseudo-Mirror Descent in Reproducing Kernel Hilbert Space (Abhishek Chakraborty, 2021) View paper
- Reinforcement Learning and Control in RKHS
    - Multi-Task and Quantum Policy Gradient (2 papers)
    - [2] Multi-task reinforcement learning in reproducing kernel hilbert spaces via cross-learning (Cervino, 2021) View paper
    - [4] Quantum Policy Gradient in Reproducing Kernel Hilbert Space (Bossens, 2024) View paper
    - Value Iteration in RKHS (1 papers)
    - [15] Value iteration for streaming data on a continuous space with gradient method in an RKHS. (Jiamin Liu, 2023) View paper
- Nonparametric Differential Equation Learning (1 papers)
    - [16] Learning nonparametric differential equations with operator-valued kernels and gradient matching (Heinonen, 2014) View paper

Narrative

Core task: in-context operator learning via gradient descent in operator RKHS. This field investigates how learning systems can adapt to new operator-valued tasks by leveraging gradient-based optimization in reproducing kernel Hilbert spaces (RKHS) and related functional frameworks. The taxonomy reveals several complementary perspectives: Transformer-Based In-Context Operator Learning explores how attention mechanisms can implicitly perform operator inference from context; Operator Learning via Stochastic Gradient Descent and Gradient Descent Algorithms in RKHS focus on optimization dynamics and convergence guarantees in infinite-dimensional spaces; Neural Network Operator Learning examines parameterized approximations of operators, often through deep architectures; Reinforcement Learning and Control in RKHS extends these ideas to sequential decision-making; and Nonparametric Differential Equation Learning targets discovery of governing equations from data. Representative works such as Stochastic Gradient Hilbert Spaces[1] and Sequential Learning RKHS[5] illustrate optimization foundations, while Functional Transduction[9] and Neural Operator Convergence[10] address approximation and generalization.

A particularly active line of research centers on bridging classical kernel methods with modern transformer architectures, asking whether in-context learning can be understood as implicit gradient descent in function space. Another contrasting direction emphasizes rigorous operator-theoretic guarantees, as seen in Regularized Operator-valued Kernels[7] and Vector-valued Spectral Regularization[11], which prioritize stability and convergence over architectural flexibility. The original paper, Continuum Transformers Operator Descent[0], sits within the Transformer-Based In-Context Operator Learning branch, specifically in the Continuum and Operator-Space Transformers cluster alongside In-Context Fine-Tuning Operators[17]. Compared to nearby works like Distributionally-robust In-context[13], which emphasizes robustness under distribution shift, or Softmax Adapts Lipschitzness[3], which studies adaptive smoothness, Continuum Transformers Operator Descent[0] appears to focus on the continuum limit of attention mechanisms and their connection to gradient flows in operator RKHS, offering a theoretical lens on how transformers implicitly navigate infinite-dimensional operator spaces during in-context adaptation.

Related Works in Same Category

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

1. In-Context Fine-Tuning for Neural Operators

Authors: Y Patel, A Mishra, A Tewari | **URL**: View paper

Abstract

â¦ in-context learning, when performed by transformer operators, is achieved by gradient descent in an operator RKHSâ¦ -context operator learning do so by performing gradient descent in a â¦

⚠ Similarity Notice

This paper appears to be a variant or earlier version of the original paper. Both papers have nearly identical titles (differing only by 'Continuum' vs 'Transformer Operators'), share the same core theoretical contribution (proving continuum/transformer operators perform in-context learning via operator gradient descent in RKHS), use the same mathematical framework (generalized continuum attention, operator RKHS, representer theorem), and present the same main theorem (Theorem 3.1 in both). The original paper appears to be a more complete version with additional theoretical results (Bayes optimality, pre-training convergence) and empirical validation that are absent or incomplete in this candidate paper.

Contributions Analysis

Overall novelty summary. The paper demonstrates that continuum transformers perform in-context operator learning by implementing gradient descent in an operator RKHS, extending prior finite-dimensional results to infinite-dimensional function spaces. It resides in the 'Continuum and Operator-Space Transformers' leaf, which contains only two papers total (including this work and one sibling). This represents a sparse, emerging research direction within the broader taxonomy of 19 papers across the field, suggesting the work addresses a relatively underexplored niche at the intersection of transformer architectures and operator theory.

The taxonomy reveals that the paper's immediate parent branch, 'Transformer-Based In-Context Operator Learning', contains three distinct leaves: the paper's own leaf (operator-space transformers), a sibling leaf on finite-dimensional in-context learning with three papers, and a third leaf on kernel-based NLP applications. Neighboring branches include 'Operator Learning via Stochastic Gradient Descent' (focusing on pure optimization without transformers) and 'Neural Network Operator Learning' (emphasizing NTK analysis and transfer learning). The paper bridges transformer architectures with classical operator-theoretic RKHS methods, diverging from purely optimization-focused or purely neural-network-centric approaches.

Among 30 candidates examined, each of the three contributions shows at least one refutable candidate from 10 examined per contribution. Contribution A (operator gradient descent in RKHS) found 1 refutable among 10 candidates; Contribution B (Bayes optimality recovery) similarly found 1 refutable among 10; Contribution C (parameter recovery via pre-training) also found 1 refutable among 10. The statistics suggest that while each contribution encounters some overlapping prior work within the limited search scope, the majority of examined candidates (9 out of 10 per contribution) do not clearly refute the claims, indicating partial novelty relative to the top-30 semantic matches.

Based on the limited search scope of 30 candidates, the work appears to occupy a sparsely populated research direction with modest but non-negligible prior overlap. The taxonomy structure confirms this is an emerging area, though the contribution-level statistics indicate that key claims have at least some precedent among closely related work. A more exhaustive literature search beyond top-30 semantic matches would be needed to fully assess novelty, particularly given the specialized intersection of continuum limits, operator theory, and transformer in-context learning.

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

Contribution 1: Continuum transformers perform in-context operator learning via operator gradient descent in RKHS

Description: The authors demonstrate that continuum transformers achieve in-context learning by implementing gradient descent steps in an operator-valued reproducing kernel Hilbert space. This characterization required novel proof strategies including a generalized representer theorem for Hilbert spaces and gradient flow analysis over functionals on Hilbert spaces.

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. Learning functional transduction

URL: View paper

Prior Art Analysis

Functional Transduction[9] demonstrates that transductive regression systems can be meta-learned to perform in-context operator learning using reproducing kernel Banach spaces (RKBS). The paper explicitly shows that their transducer model performs operator regression through kernel-based methods in function spaces, which directly challenges the novelty claim that the original paper was first to characterize in-context operator learning via gradient descent in RKHS. Both papers address the same fundamental problem of in-context learning for operators using reproducing kernel theory, with Functional Transduction[9] providing a complete framework including theoretical foundations (Theorem 2 on RKBS representer maps) and empirical validation on operator learning tasks.

Evidence

Evidence 1 - **Rationale**: Functional Transduction[9] provides a representer theorem for RKBS that characterizes solutions to operator learning problems, directly addressing the same theoretical framework claimed as novel in the original paper. - **Original**: we herein demonstrate that continuum transformers perform incontext operator learning by performing gradient descent in an operator rkhs. we demonstrate this using novel proof strategies that leverage a generalized representer theorem for hilbert spaces and gradient flows over the space of functiona... - **Candidate**: theorem 2 (rkbs representer map). let b be a u-valued rkbs from v to u, if for any dataset do p d, lp., doqis lower semi-continuous, coercive and bounded below, then there exists a function t : d þñ b such that t pdoq is a minimizer of equation (1). if l is of the form lp., doq= ˜l ◦tδvi uidn with ˜...

Evidence 2 - **Rationale**: Both papers develop mathematical frameworks using reproducing kernel theory to characterize in-context operator learning in transformers, with Functional Transduction[9] explicitly connecting attention mechanisms to reproducing kernels. - **Original**: we, therefore, herein extend this line of theoretical inquiry to characterize the continuum transformers that have been leveraged for in-context operator learning calvello et al. (2024). our contributions, therefore, are twofold: the insights afforded by such theoretical analysis and the development... - **Candidate**: in order to build such a model, we leverage the theory of reproducing kernel banach spaces (rkbs) (micchelli and pontil, 2004; zhang, 2013; lin et al., 2022) and interpret the transformer's (vaswani et al., 2017) attention mechanism as a parametric vector-valued reproducing kernel. while kernel regr...

Evidence 3 - **Rationale**: Functional Transduction[9] demonstrates that their model performs in-context operator learning through a single feedforward pass using kernel methods, achieving the same goal of in-context operator regression claimed as novel by the original paper. - **Original**: proving continuum transformers perform in-context operator learning by performing gradient descent in a reproducing kernel hilbert space of operators and that the resulting in-context predictor recovers the bayes optimal predictor under well-specified parameter choices of the transformer. - **Candidate**: our model is meta-trained to take as input any dataset do of pairs pvi, opviqqidì of some target function o together with a query element v1and produces directly an estimate of the image opv1q. after meta-training, our network is able to perform regression of unseen operators o1from varying dataset ...

2. Gradient-Based Non-Linear Inverse Learning

URL: View paper

Brief Assessment

Non-Linear Inverse Learning[34] focuses on statistical inverse learning using gradient descent for nonlinear inverse problems under random design, not on in-context learning mechanisms in transformer architectures for operator learning.

3. Transformers May Learn to Classify In-Context by Context-Adaptive Kernel Gradient Descent

URL: View paper

Brief Assessment

Context-Adaptive Kernel Descent[36] focuses on classification tasks with softmax attention implementing kernel gradient descent in RBF kernel feature spaces. The original paper addresses operator learning for PDEs with continuum transformers performing gradient descent in operator-valued RKHS, which is a fundamentally different mathematical framework and application domain.

4. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks

URL: View paper

Brief Assessment

Supervised Convolutional Kernels[37] focuses on supervised learning of convolutional kernel networks for image tasks, not on in-context learning or operator gradient descent dynamics in transformers for PDE surrogate modeling.

5. Learning surrogate potential mean field games via Gaussian processes: a data-driven approach to ill-posed inverse problems

URL: View paper

Brief Assessment

Surrogate Gaussian Processes[31] focuses on inverse problems in mean field games using Gaussian processes for parameter recovery, not on in-context learning mechanisms in transformers or operator gradient descent in RKHS.

6. Softmax Linear: Transformers may learn to classify in-context by kernel gradient descent

URL: View paper

Brief Assessment

Kernel Gradient Descent[35] focuses on classification tasks with softmax attention implementing kernel gradient descent in RBF kernel feature space, not operator learning in infinite-dimensional function spaces for PDEs.

7. Deep Learning: A (Currently) Black-Box Model

URL: View paper

Brief Assessment

Black-Box Model[32] discusses LLMs' in-context learning capabilities and reproducing kernel Hilbert spaces in a general context, but does not address continuum transformers, operator learning, or the specific characterization of gradient descent in operator-valued RKHS for infinite-dimensional function inputs.

8. In-context learning with transformers: Softmax attention adapts to function lipschitzness

URL: View paper

Brief Assessment

Softmax Adapts Lipschitzness[3] focuses on how softmax attention learns to adapt its attention window to function lipschitzness in standard (finite-dimensional) regression tasks. The original paper addresses continuum transformers for infinite-dimensional operator learning in PDE settings, which is a fundamentally different domain and architecture.

9. Transformer In-Context Learning for Categorical Data

URL: View paper

Brief Assessment

Categorical In-Context Learning[30] focuses on categorical outcomes with softmax modeling for classification tasks, while the original work addresses operator learning for PDEs in continuous function spaces. The technical frameworks are fundamentally different.

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

URL: View paper

Brief Assessment

Universality Next-Token Prediction[33] focuses on next-token prediction for autoregressive sequences xt+1 = f(xt) using causal kernel descent methods in vector-valued RKHS. The original paper addresses operator learning for PDEs using continuum transformers with operator-valued RKHS and operator gradient descent, which is a fundamentally different setting.

Contribution 2: In-context predictor recovers Bayes Optimal Predictor under well-specified parameters

Description: The authors prove that with appropriate parameter choices, the operator learned by continuum transformers in context converges to the Bayes Optimal Predictor as the number of transformer layers approaches infinity. This result leverages Gaussian measures over Hilbert spaces and connections to Hilbert space kriging.

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. Transformers as statisticians: Provable in-context learning with in-context algorithm selection

URL: View paper

Brief Assessment

Transformers as Statisticians[21] focuses on finite-dimensional statistical learning tasks (linear regression, classification) with standard transformers, not infinite-dimensional operator learning with continuum transformers as in the original paper.

2. One-Layer Transformers are Provably Optimal for In-context Reasoning and Distributional Association Learning in Next-Token Prediction Tasks

URL: View paper

Brief Assessment

One-Layer Optimal Reasoning[22] focuses on one-layer transformers for in-context recall tasks with finite-sample analysis, while the original work studies continuum transformers for operator learning with infinite-depth convergence to Bayes optimal predictors in function spaces.

3. Transformers learn variable-order Markov chains in-context

URL: View paper

Brief Assessment

Variable-order Markov In-context[24] focuses on variable-order Markov chains and context tree models for language modeling, not on continuum transformers for PDE operator learning. The technical settings and problem domains are fundamentally different.

4. Bayesian physics informed neural networks for reliable transformer prognostics

URL: View paper

Brief Assessment

Bayesian Physics Transformers[25] focuses on Bayesian physics-informed neural networks for transformer prognostics using PDEs for thermal modeling, not on continuum transformers performing in-context learning via operator gradient descent or convergence to Bayes optimal predictors in infinite depth limits.

5. Towards scalable Bayesian transformers: investigating stochastic subset selection for NLP

URL: View paper

Brief Assessment

Stochastic Subset Selection[27] focuses on Bayesian inference for transformer models in NLP tasks through parameter subset selection, not on in-context learning convergence to Bayes optimal predictors in infinite depth limits.

6. Bayesformer: Transformer with uncertainty estimation

URL: View paper

Brief Assessment

Bayesformer Uncertainty[29] focuses on uncertainty estimation in transformers using dropout-based Bayesian inference, not on in-context learning convergence to Bayes optimal predictors in operator learning settings.

7. A Bayesian adversarial probsparse Transformer model for long-term remaining useful life prediction

URL: View paper

Brief Assessment

Bayesian Probsparse Transformer[28] focuses on long-term remaining useful life (RUL) prediction for equipment maintenance, not on in-context learning or operator learning with transformers. The technical domains are entirely different.

8. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization

URL: View paper

Prior Art Analysis

Bayesian Model Averaging[26] demonstrates that transformers performing in-context learning implement Bayesian Model Averaging (BMA), which is the Bayes optimal predictor. The paper proves that perfectly pretrained LLMs perform ICL via BMA (Proposition 4.1), showing the prediction equals the Bayes optimal estimator. This directly refutes the novelty claim that the original paper was first to prove convergence to Bayes Optimal Predictor, as Bayesian Model Averaging[26] already established this connection between ICL and Bayes optimality through BMA.

Evidence

Evidence 1 - **Rationale**: This pair shows that Bayesian Model Averaging[26] already proved that ICL implements BMA, which is the Bayes optimal predictor. The candidate explicitly states that the prediction via BMA is the Bayes optimal estimator, establishing this result before the original paper. - **Original**: we further show the operator learned in context is the bayes optimal predictor in the infinite depth limit of the transformer. - **Candidate**: proposition 4.1 (llms perform bma). under the model in (4.1), it holds that p(rt+1 = · | promptt) = ∫ p(rt+1 = · | ˜ct+1, st, z)p(z | st)dz. we note that the left-hand side of (4.2) is the prediction of the pretrained llm given a prompt promptt. meanwhile, the right-hand side is exactly the predic...

9. Transformers can do bayesian inference

URL: View paper

Brief Assessment

Transformers Bayesian Inference[20] focuses on approximating posterior predictive distributions for supervised learning tasks using prior-data fitted networks, not on continuum transformers performing operator gradient descent in infinite-dimensional function spaces.

10. LLMs are Bayesian, in Expectation, not in Realization

URL: View paper

Brief Assessment

Bayesian Expectation not Realization[23] focuses on reconciling martingale violations with Bayesian-like behavior through information-theoretic analysis of positional encodings. It does not address continuum transformers, operator learning, or the specific convergence results for infinite-depth transformers in operator RKHSs that the original paper establishes.

Contribution 3: Parameters enabling gradient descent are recovered through pre-training

Description: The authors establish that the specific parameter configurations under which continuum transformers perform operator gradient descent are stationary points of the training objective. This required developing a novel gradient flow analysis over the space of functionals on a Hilbert space.

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. Recovering the pre-fine-tuning weights of generative models

URL: View paper

Brief Assessment

Recovering Pre-fine-tuning Weights[46] focuses on recovering pre-fine-tuning model weights from LoRA fine-tuned models, not on demonstrating that gradient descent parameters are stationary points of training objectives in continuum transformers.

2. Transformers learn to implement preconditioned gradient descent for in-context learning

URL: View paper

Prior Art Analysis

Preconditioned Gradient Descent[39] demonstrates that pre-training can recover parameters that implement gradient descent in transformers. The paper proves that specific parameter configurations are stationary points of the training objective and provides empirical validation that these parameters are recovered through training. This directly challenges the novelty claim of the original paper, as both works establish that gradient descent parameters emerge from the training process, though using different mathematical frameworks (Preconditioned Gradient Descent[39] uses Frobenius norm derivatives while the original uses gradient flows over functionals).

Evidence

Evidence 1 - **Rationale**: Both papers provide empirical validation that the theoretically predicted parameters are recovered through training. The original paper validates across operator RKHSs, while Preconditioned Gradient Descent[39] validates for linear regression, but both demonstrate that training recovers the gradient descent parameters. - **Original**: empirically validating that continuum transformers perform in-context operator gradient descent upon inference with the exhibited parameters and that such parameters are recovered with transformer training across a diverse selection of operator rkhss. - **Candidate**: we empirically validate the critical points analyzed in theorem 3 and theorem 4. for a transformer with three layers, our experimental results confirm the structural of critical points. furthermore, we observed the objective value associated with these critical points is close to 0, suggesting that ...

Evidence 2 - **Rationale**: Preconditioned Gradient Descent[39] characterizes global minimizers that implement gradient descent for transformers, establishing that training recovers these parameters. This demonstrates prior work on proving that pre-training recovers gradient descent parameters, though in a finite-dimensional rather than operator setting. - **Original**: we, therefore, herein extend this line of theoretical inquiry to characterize the continuum transformers that have been leveraged for in-context operator learning calvello et al. (2024). our contributions, therefore, are twofold: the insights afforded by such theoretical analysis and the development... - **Candidate**: theorem 1 (single-layer; non-isotropic data). assume that vector x(i) is sampled from n(0, σ), i.e., a gaussian with covariance σ = uλu⊤ where λ = diag(λ1, . . . , λd). moreover, assume that w⋆ is sampled from n(0, id). then, the following choice of parameters p0 = 0dxd 0 0 1 , q 0 = - udiag 1 n+1 n...

3. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models

URL: View paper

Brief Assessment

Delta Tuning[41] focuses on parameter-efficient fine-tuning methods for pre-trained language models, not on gradient descent recovery in continuum transformers or operator learning contexts.

4. OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks

URL: View paper

Brief Assessment

COVID Classification Networks[47] focuses on optimizing pre-trained Vision Transformers for COVID-19 image classification using stochastic configuration networks, not on analyzing gradient descent parameter recovery in continuum transformers for operator learning.

5. What learning algorithm is in-context learning? investigations with linear models

URL: View paper

Brief Assessment

Linear Models In-context[42] focuses on proving that transformers CAN implement gradient descent algorithms (constructive proofs), not on demonstrating that pre-training RECOVERS these parameters as stationary points through gradient flow analysis.

6. Deep compression of pre-trained transformer models

URL: View paper

Brief Assessment

Deep Compression Transformers[44] focuses on model compression techniques (quantization and pruning) for pre-trained transformers, not on theoretical characterization of in-context learning or gradient descent mechanisms during pre-training.

7. Understanding and minimising outlier features in transformer training

URL: View paper

Brief Assessment

Minimising Outlier Features[43] focuses on reducing outlier features in transformers to improve quantisation, not on characterizing in-context learning through gradient descent or analyzing pre-training recovery of gradient descent parameters in operator spaces.

8. Unraveling the gradient descent dynamics of transformers

URL: View paper

Brief Assessment

Gradient Descent Dynamics[38] analyzes gradient descent dynamics in standard transformers for finite-dimensional inputs, not continuum transformers for infinite-dimensional function spaces. The candidate focuses on convergence conditions and optimization landscapes for shallow transformers, while the original work addresses pre-training recovery of specific parameter configurations in continuum transformers that enable operator gradient descent.

9. Should I try multiple optimizers when fine-tuning pre-trained Transformers for NLP tasks? Should I tune their hyperparameters?

URL: View paper

Brief Assessment

Multiple Optimizers Fine-tuning[45] focuses on comparing different optimizers (SGD, Adam, etc.) when fine-tuning pre-trained transformers for NLP classification tasks, not on analyzing how pre-training recovers gradient descent parameters in transformer architectures.

10. Transformers learn to implement multi-step gradient descent with chain of thought

URL: View paper

Brief Assessment

Multi-step Chain Thought[40] focuses on training dynamics for chain-of-thought prompting in transformers, demonstrating convergence to multi-step gradient descent parameters. The original paper addresses continuum transformers for operator learning in PDE settings with novel functional analysis techniques, which is a fundamentally different domain and mathematical framework.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 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. In-Context Fine-Tuning for Neural Operators

Detected in: Core Task (sibling)

⚠ **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] Continuum Transformers Perform In-Context Learning by Operator Gradient Descent View paper
- [1] Learning operators with stochastic gradient descent in general hilbert spaces View paper
- [2] Multi-task reinforcement learning in reproducing kernel hilbert spaces via cross-learning View paper
- [3] In-context learning with transformers: Softmax attention adapts to function lipschitzness View paper
- [4] Quantum Policy Gradient in Reproducing Kernel Hilbert Space View paper
- [5] Gradient descent in neural networks as sequential learning in rkbs View paper
- [6] Deep transfer operator learning for partial differential equations under conditional shift View paper
- [7] Learning Operators by Regularized Stochastic Gradient Descent with Operator-valued Kernels View paper
- [8] Distributed kernel-based gradient descent algorithms View paper
- [9] Learning functional transduction View paper
- [10] Optimal convergence rates for neural operators View paper
- [11] Optimal rates for vector-valued spectral regularization learning algorithms View paper
- [12] Projected Pseudo-Mirror Descent in Reproducing Kernel Hilbert Space View paper
- [13] Transformers Can Perform Distributionally-robust Optimisation through In-context Learning View paper
- [14] Gradient Descent in RKHS with Importance Labeling View paper
- [15] Value iteration for streaming data on a continuous space with gradient method in an RKHS. View paper
- [16] Learning nonparametric differential equations with operator-valued kernels and gradient matching View paper

- [17] In-Context Fine-Tuning for Neural Operators View paper
- [18] Functional Gradients and Generalizations for Transformer In-Context Learning View paper
- [19] Support Vector Generation: Kernelizing Large Language Models for Efficient Zeroâ□□Shot NLP View paper
- [20] Transformers can do bayesian inference View paper
- [21] Transformers as statisticians: Provable in-context learning with in-context algorithm selection View paper
- [22] One-Layer Transformers are Provably Optimal for In-context Reasoning and Distributional Association Learning in Next-Token Prediction Tasks View paper
- [23] LLMs are Bayesian, in Expectation, not in Realization View paper
- [24] Transformers learn variable-order Markov chains in-context View paper
- [25] Bayesian physics informed neural networks for reliable transformer prognostics View paper
- [26] What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization View paper
- [27] Towards scalable Bayesian transformers: investigating stochastic subset selection for NLP View paper
- [28] A Bayesian adversarial probsparse Transformer model for long-term remaining useful life prediction View paper
- [29] Bayesformer: Transformer with uncertainty estimation View paper
- [30] Transformer In-Context Learning for Categorical Data View paper
- [31] Learning surrogate potential mean field games via Gaussian processes: a data-driven approach to ill-posed inverse problems View paper
- [32] Deep Learning: A (Currently) Black-Box Model View paper
- [33] Towards understanding the universality of transformers for next-token prediction View paper
- [34] Gradient-Based Non-Linear Inverse Learning View paper
- [35] Softmax Linear: Transformers may learn to classify in-context by kernel gradient descent View paper
- [36] Transformers May Learn to Classify In-Context by Context-Adaptive Kernel Gradient Descent View paper
- [37] End-to-End Kernel Learning with Supervised Convolutional Kernel Networks View paper
- [38] Unraveling the gradient descent dynamics of transformers View paper
- [39] Transformers learn to implement preconditioned gradient descent for in-context learning View paper
- [40] Transformers learn to implement multi-step gradient descent with chain of thought View paper
- [41] Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models View paper
- [42] What learning algorithm is in-context learning? investigations with linear models View paper
- [43] Understanding and minimising outlier features in transformer training View paper
- [44] Deep compression of pre-trained transformer models View paper
- [45] Should I try multiple optimizers when fine-tuning pre-trained Transformers for NLP tasks? Should I tune their hyperparameters? View paper
- [46] Recovering the pre-fine-tuning weights of generative models View paper
- [47] OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks View paper