Title	文脈内学習のメカニズム的基盤
Author(s)	趙, 羽風
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	ETD
URL	https://hdl.handle.net/10119/20578
Rights	
Description	Supervisor: 井之上 直也, 先端科学技術研究科　, 博士

氏名	ZHAO, Yufeng		
学位の種類	博士（情報科学）		
学位記番号	博情第 565 号		
学位授与年月日	令和 8 年 3 月 25 日		
論文題目	The Mechanistic Basis of In-context Learning		
論文審査委員	井之上　直也	北陸先端科学技術大学院大学	准教授
	NGUYEN, Le Minh	同	教授
	白井　清昭	同	教授
	KERTKEIDKACHORN, Natthawut	同	准教授
	HEINZERLING, Benjamin	東北大学言語 AI 研究センター	特任准教授

論文の内容の要旨

In-context Learning (ICL) has recently emerged as a powerful paradigm enabling Language Models (LMs) to perform few-shot learning without parameter updates. However, despite the success in application, the underlying mechanism of ICL, i.e., why and how LMs perform ICL operation, remains poorly understood. Therefore, this dissertation, which consists of 4 major works, aims to present a systematic investigation into the inner dynamics of LMs under the ICL scenario, and also utilize the gained insights to develop practical applications to improve ICL performance.

(Work 1) Motivated by the gap between the current interpretation of ICL and the observed ICL inference phenomena, first, we establish a comprehensive inference circuit based on the previous works of induction circuits, revealing how the induction circuits work in ICL on real data samples and real LMs. Through careful measurements and causal analysis, we identify three essential operations: (1) Input Text Encode: LMs encode every input text (in the demonstrations and queries) into a linear representation in the hidden states with sufficient information to solve ICL tasks. (2) Semantics Merge: LMs merge the encoded representations of demonstrations with their corresponding label tokens to produce joint representations of labels and demonstrations. (3) Feature Retrieval and Copy: LMs search the joint representations of demonstrations similar to the query representation on a task subspace, and copy the searched representations into the query. Then, language model heads capture these copied label representations to a certain extent and decode them into predicted labels. This decomposition unifies previously fragmented observations of ICL and is empirically validated via ablation, which shows substantial degradation when any step is disabled. Additionally, we confirm and list some bypass mechanisms that solve ICL tasks in parallel with the proposed circuit.

(Work 2) However, operation 3 above requires the ground-truth label in the demonstrations as the source for copying; otherwise, according to our proposed theoretical framework, LMs should fail to make correct predictions in the unseen-label scenario, where the ground-truth labels are absent from the demonstrations. Surprisingly, we find that LMs can still achieve significantly higher accuracy even under this unseen-label scenario, which suggests the existence of another significant mechanism in the aforementioned bypasses. Therefore, we hypothesize another mechanism, called Task-oriented Information Removal, which runs parallel with induction circuits. Specifically, we demonstrate that in the zero-shot scenario, LMs encode queries into non-selective representations in hidden states containing information for all possible tasks, leading to arbitrary outputs without focusing on the intended task, resulting in low accuracy. Meanwhile, we find that selectively removing specific information from hidden states by a low-rank filter effectively steers LMs toward the intended task. Building on these findings, by measuring the hidden states on carefully designed metrics, we observe that

few-shot ICL effectively simulates such task-oriented information removal processes, selectively removing the redundant information from entangled non-selective representations, and improving the output based on the demonstrations, which constitutes a key mechanism underlying ICL. Moreover, we identify essential attention heads inducing the removal operation, termed Denoising Heads, which enables the ablation experiments blocking the information removal operation from the inference, where the ICL accuracy significantly degrades, especially when the correct label is absent from the few-shot demonstrations, confirming both the critical role of the information removal mechanism and denoising heads. Notice that such an information removal mechanism does not require the ground-truth labels in the demonstrations, thus explaining the ICL performance in the unseen-label scenario.

(Work 3) Then, we leverage the mechanistic insights gained from the above studies to develop practical applications to enhance ICL performance. Based on the conclusion of "LMs encode every input text into a linear representation in the hidden states with sufficient information to solve ICL tasks", we develop a new output calibration method called Hidden Calibration, which renounces token probabilities and uses the nearest centroid classifier on the LM's last hidden states. In detail, we assign the label of the nearest centroid previously estimated from a calibration set to the test sample as the predicted label. Our experiments on 6 models and 10 classification datasets indicate that Hidden Calibration consistently outperforms current token-based baselines by about 20%∼50%, achieving a strong state-of-the-art in ICL. Our further analysis demonstrates that Hidden Calibration finds better classification criteria with less inter-class overlap, and LMs provide linearly separable intra-class clusters with the help of demonstrations, which supports Hidden Calibration and gives new insights into the principle of ICL.

(Work 4) Moreover, based on the conclusion of "induction head copy label information to the output", we develop an ICL-oriented fine-tuning method called Attention Behavior Fine-Tuning, building training objectives on the attention scores instead of the final outputs, to force the attention scores of induction heads to focus on the correct label tokens presented in the context and mitigate attention scores from the wrong label tokens. Our experiments on 9 modern LMs and 8 datasets empirically find that ABFT outperforms in performance, robustness, unbiasedness, and efficiency, with only around 0.01% data cost compared to the previous methods. Moreover, our subsequent analysis finds that the end-to-end training objective contains the ABFT objective, suggesting the implicit bias of ICL-style data to the emergence of induction heads. Also, ABFT demonstrates the possibility of controlling specific module sequences within LMs to improve their behavior, opening up the future application of mechanistic interpretability.

In summary, this dissertation improves the mechanistic interpretation of ICL by fitting induction circuits onto the previous observations, then proposing a new mechanism to supply the unseen label scenario. Moreover, based on such a mechanistic understanding, we propose some applicable methods to improve the performance of ICL, and also open up a new research direction, Mechanistic Controllability by the proposed prototypes.

Keywords: Mechanistic Interpretability, In-context Learning, Language Models, Transformer, Low-resource Model Controllability, Representation Learning

論文審査の結果の要旨

本論文は、大規模言語モデル（LLM）がパラメータ更新を伴わずに新たなタスクを習得する能力である「文脈内学習（In-context Learning; ICL）」の背後にあるメカニズムを解明するとともに、その知見に基づいたICLの性能向上手法を提案したものである。

ICLは、少数の例示のみで多様なタスクに適応できる強力な学習パラダイムとして注目されているが、その高い性能の一方で、モデル内部で具体的にどのような演算が行われているかという推論機序については未解明な点が多く残されていた。これに対し本論文では、ICLを単なる現象として捉えるのではなく、LLM内部における一連の推論回路としてメカニズムを体系化した。特に、正解ラベルが欠如した条件下での挙動を補完する「情報除去メカニズム」の本質を突き止めた点は、学術的に極めて高い価値を有する。また、因果解析やアブレーション実験といった厳密な手法によって推論機序の妥当性を検証しており、提案された理論の信頼性は極めて高い。さらに、得られた知見を応用し、従来手法よりもデータ量、及び計算量の面で効率的にICLの性能向上を実現しており、計算リソースの節約とAIの社会実装の加速という観点から、実用的にも多大な意義を持つ。

本論文は全6章で構成され、その内容は以下の通りである。第1章および第2章では、ICLの現状と機序解明の重要性を整理した。第3章では、実データを用いた介入実験と因果解析を通じて、ICLが「例示の入力テキストとクエリのエンコード」「ラベル情報の統合」「特徴量の検索とコピー」という3つの演算ステップからなる推論回路によって実現されていることを明らかにした。第4章では、例示の中に正解ラベルが含まれない未知ラベル設定においてもICLが機能する機序として、不要な情報を選択的に除去しタスクに集中させる「タスク指向の情報除去」メカニズムを特定し、これを司る特定の注意機構ヘッド（Denoising Head）を同定した。第5章と第6章では、これらの機序に関する知見に基づき、ICLの性能を向上させる二種類の手法を提案している。第5章では、中間層の表現を用いてICLのラベル予測確率を較正する「Hidden Calibration」手法を提案し、従来のトークン確率に基づく手法を大幅に上回る精度を実現した。第6章では、ICLで重要な役割を担う注意機構ヘッドの振る舞いを最適化する「Attention Behavior Fine-tuning」を提案し、少量の計算資源で効率的にICLの性能向上が可能であることを示した。

以上、本論文は、ICLの内部機序の解明とその応用手法について、多角的な理論解析と実験的検証を行い、その結果を報告したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。