神経科学に着想を得たシナプス刈り込みによる大規模言語モデルの原理解明

原田宥都　前田ありさ　森田早織　大関洋平
東京大学
{harada-yuto,alyssa-maeda,msaori6012, oseki}@g.ecc.u-tokyo.ac.jp

概要

言語モデルが大規模化し、高度な言語能力を獲得していくにつれ、モデルの内部処理は更にブラックボックス化している。これに対して近年の言語モデルの説明性は、内部情報を直接取得し解釈するというよりはむしろ、入力プロンプトに対するモデルの振る舞いをもって担保する傾向にある。そこで本研究では、大規模言語モデルの新たな説明手法としてプルーニングを利用した方法を提案する。意図的にモデルを損傷させ、その振る舞いの変化を定量的に評価することで、モデルの内部処理についての示唆を得ることを目的とする。実験の結果、モデルをプルーニングする際には、対象とするレイヤーによってベンチマークでの性能に影響が出るかどうかが異なる可能性が示唆された。

1　はじめに

近年の大規模言語モデルは高度な言語理解能力を実現しているが、その仕組みには人間の脳との相違点が多く存在する。例えば、ニューロン同士のネットワークの発達過程に着目すると、脳においてシナプスの形成ペースが変化していくことが知られている [1]。そのペースは脳の領域ごとに異なるが、出生直後からある一定の時期までは非常に密なシナプス形成が行われ、その後に不要なシナプスだけが大幅に減少する。この動的なネットワークの形成により、機能的な神経回路が完成していく。この過程は「シナプス刈り込み (Synapse elimination, Synaptic pruning)」と呼ばれ、生後発達期の神経系で普遍的に起こる重要な現象であり、効率的な神経回路を形成するための基本的な過程であるとされている [2]。また、一部の精神神経疾患の原因として発達期の異常なシナプス刈り込みが関連しているということも指摘されており [3]、脳機能におけるシナプス刈り込みは普遍的で重要な要素であるようである。

一方、機械学習の研究においても、プルーニング (刈り込み) という技術があり、こちらは量子化 (Quantize) や蒸留 (Distillation) などと並んでモデルを軽量化することを目的とした手法の代表的なものである。プルーニングは自然言語処理モデルにおいても有効であり、現在盛んに研究されている大規模言語モデルにおいても様々なアルゴリズムを通じて適用可能である [4]。機械学習におけるプルーニングは、脳機能におけるシナプス刈り込みとは異なる形ではあるが、自然言語処理技術の発展に貢献する重要な技術である。

本研究では、本来モデルを軽量化するための工学的手法であるプルーニングを、モデルの説明手法として利用する方法を提案する。これは、脳においてシナプス刈り込みの異常 (過剰や不足) が一部の精神疾患の原因となっている可能性があることに着想を得ており、言語モデルを意図的にプルーニングによって部分的に損傷させることで、モデルの振る舞いの変化を評価するものである。手法としては神経科学における脳の損傷研究 (Lesion Studies) [5] のアナロジーとして捉えることができ、損傷研究とは特定の脳部位や機能に損傷を受けた場合の対象の振る舞い (Behavior) の変化を観察することでその関係性を明らかにしようとする方法論である。

本稿では、脳機能と機械学習における 2 つの「刈り込み」について整理したのち、既存のモデルの説明手法を分類しつつ紹介し、その後、プルーニングを用いて部分的に損傷した状態の大規模言語モデルの評価実験を行う。実験の結果、プルーニングによって起こるモデルの性能低下は特に中間層以下で顕著であり、評価タスクによって低下の度合いが異なることがわかった。これらの結果は大規模言語モデルの内部でどのような処理が行われているかについて新たな示唆をもたらす。

2 関連研究

2.1 脳機能におけるシナプス刈り込み

ニューロン同士の情報伝達を担うシナプスは、出生直後からある程度の期間、過剰に形成されることが知られる。シナプスの数は人間の場合、大脳皮質視覚野では 6 から 8 ヶ月の間で最大になり、前頭前野等の領域などでは 3 年を経て最大になるなど、領域によってそのペースは異なるものの、出生時に対しておおよそ 10 倍程度の数になる [6]。最大数に達したシナプスはその後不要なものから除去され、大幅な減少に転じ、最終的に最大数の半数程度に落ち着く。この過程はシナプス刈り込みと呼ばれ、神経系の発生過程において広く見られる現象であることから、脳の神経回路網の形成に重要な役割を果たしていると考えられている。ただし、不必要なシナプスがどのような単位で選択され、削除が起きるのかは現在も研究がなされている。

発達段階でのシナプス刈り込みの異常は自閉スペクトラム症や統合失調症といった社会性障害をきたす代表的な疾患の病因の一つと見られており [3]、シナプス刈り込みが正常に機能しなかった場合、発達過程において精神神経疾患として問題をきたしてしまうことが指摘されている。このような脳における神経回路網の形成過程は、明示的には、学習の完了までパラメータ数が一定である深層学習モデルとは異なっている。ただし、機械学習におけるプルーニング技術は、不要な重みをネットワークから削除した上で再学習を行い、モデルのサイズを小さくするという点で、脳機能におけるシナプス刈り込みと対応する。

2.2 機械学習におけるプルーニング

効率的な神経回路網を形成するためのシナプス刈り込みというアイデアは、1980 年代後半の初期のニューラルネットワーク研究にも応用された。これらは現在の機械学習におけるプルーニング技術の源流となる研究であり、Optimal Brain Damage[7] を始めとして、様々な手法が検討されている [8, 9, 10, 11]。当時からモデルの軽量化を目的とした手法であり、あるいはその軽量化によるニューラルネットワークの解釈性の向上 [9] を目指した面もあった。その後深層学習モデルの台頭により、2000 年台に再び関心が集まった [12]。

プルーニングには、まず、不要な重みを選択するための刈り込みプロセスがあり、その次に再学習によって精度を回復するプロセスがある [13]。必要に応じてこの一連のプロセスを繰り返し、モデルのサイズを軽量化することができる。不要な重みの選択にはニューロン単位やレイヤー単位など様々な単位があり、また刈り込みの基準としても重みのノルムから大きさを見て判定する単純なアルゴリズムから、タスクへの寄与度を計算して不要な重みを選定するアルゴリズムなど、複数の手法が存在する。

機械学習におけるプルーニングはモデルの軽量化を目的とした工学的な手法であるが、本研究ではプルーニングによってモデルを意図的な単位で過剰に損傷し、その際のモデルの振る舞いの変化を評価タスクを通じて定量的に観察することを目的とする。

2.3 自然言語処理モデルの解釈性

深層学習モデルの台頭と共に、そのモデル内部のブラックボックス性が問題となった。モデルが学習する膨大な数のパラメータは人間にとって直感的に理解できる形式ではないため、様々な方法でモデルの説明性を担保するための手法が開発されてきた。それらの手法は、モデルの個々の入出力において局所的に説明するローカルなものと、モデル全体の内部構造を説明するためのグローバルなものに分けて整理することができる [14]。

例えば、ローカルな手法としては、BERT[15] のような Masked Language Model が主な対象だったときには、Attention に基づいた可視化手法 [16] や、Feature Attribution を利用した勾配ベースの判断根拠の可視化 [17]、Adversarial Example を利用した Example ベースの手法 [18] などが利用されてきた。グローバルな手法としては、モデルの埋め込みを利用したプロービングタスクなどが存在する。BERT においては、プロービングタスクを利用した分析によって、レイヤーごとに保持している情報が異なることが指摘されている [19]。

ただし、これらの従来手法は近年のよりスケールアップした大規模言語モデルに対してはあまり適切でないことが指摘されている [14]。近年のモデルでは高度な推論能力が示されており、局所的な例を利用した説明はあまり本質的な説明にはならず、また、計算負荷の高い説明手法は大規模なモデルに対して適用が難しい。その他にも、高度な文章生成の仕組みや、Hallucination を起こす仕組み、高度な推

ベンチマーク	オリジナルモデル	プルーニングモデル									
		0.5%					2.0%				
		4-8 層	10-14 層	14-18 層	20-24 層	24-28 層	4-8 層	10-14 層	14-18 層	20-24 層	24-28 層
JNLI	0.36	0.16	0.34	0.16	**0.36**	0.34	0.12	0.14	0.22	**0.40**	0.32
JSTS (Pearson)	0.32	0.34	0.34	**0.38**	0.33	0.26	0.08	0.06	0.01	**0.09**	0.05
JCommonsenseQA	0.66	0.30	0.48	0.66	**0.68**	**0.68**	0.32	0.32	0.48	0.64	**0.66**

表 1　評価実験の結果の概要。太字はそのプルーニング割合における最も高いパフォーマンス。

論能力を引き出すためのより良い方法など、従来とは解明したい内容も異なってきているという背景がある。そのため、このような大規模言語モデルに対しては、現在、主により入力プロンプトを重視した説明性の担保が試みられている。例えば、出力結果に大きな影響を与えるような入力トークンをスコアで評価する技術などがある [20]。

このように、モデルがより大規模になり、高度な言語能力を獲得するにつれて、その仕組みを理解するための手法は、モデル内部の情報に直接アクセスするというよりはむしろモデルの入力と出力（振る舞い）へ着目する向きが強くなっている。このような潮流を考慮し、本研究では、これまでは複雑な研究対象である脳に対して行われてきた損傷研究 (Lesion Studies) を参考にして、プルーニング技術を用いた言語モデルの説明手法を検討する。

3　実験

3.1　実験の目的

本稿では、プルーニングを用いた大規模言語モデルの検証のための最初の簡易的な実験を行う。BERT を対象とした先行研究 [19] ではレイヤーごとに保持している情報が異なる可能性が指摘されていることを踏まえ、更に大規模なモデルにおいても、レイヤーごとに処理している情報の種類が異なることを仮定する。異なるレイヤーをプルーニングで損傷した際のモデルの振る舞いの変化を、既存のベンチマークを利用して評価する。

3.2　実験設定

実験に用いるモデルとしては、Llama2[21] をベースに日本語による追加事前学習を行ったモデルである、ELYZA-japanese-Llama-2-7b-instruct[22] を使用する。また、プルーニングのためのツールとしては、LLM-pruner[4] を用いた。ニューロン単位の構造的プルーニングを行い、重み削除の基準としては単純な L1 ノルムを参照した。また、MLP レイヤと Attention レイヤの両方の重みを削除の対象としている。モデルは全体で 32 層あり、そのうち 5 層の重みに対してプルーニングを行う。浅い層 (1-3 層) と最も深い層 (32 層) へのプルーニングはモデルの性能を著しく低下させることが知られている [4] ため、削除対象のレイヤは 4-8 層、10-14 層、14-18 層、20-24 層、24-28 層の 5 通りとした。また、それぞれの設定において、モデル全体のパラメータの 0.5%をプルーニングした場合と、2.0%をプルーニングした場合とがある。

3.3　評価データ

ベンチマークとしては、llm-jp-eval[1)] を用いた。タスクとしては JNLI(自然言語推論)、JSTS(文ペアの意味的類似度の判定)、JCommonsenseQA(5 択での QA) の 3 つを採用した。プルーニング後のモデルは出力を途中で止めることができなくなるケースが散見され、実験時間が大幅に増えてしまうため、すべての実験設定で最大出力系列長を 30 に設定した。その他の設定については、全て llm-jp-eval のデフォルトの値を利用している。

JNLI　Japanese Natural Language Inference(JNLI) は、2 つの文のペア (前提・仮説) が与えられたときに、前提の文が仮説の文に対してどのような推論関係を持つかを回答するタスクである。回答となる推論関係には「含意 (entailment)」「矛盾 (contradiction)」「中立 (neutral)」の 3 つの値がある。

JSTS　Japanese Semantic Textual Similarity(JSTS) は、与えられた 2 つの文のペアに対して、意味的な類似度を推定し付与するタスクである。正解の類似度は、0 (意味が完全に異なる) 5 (意味が等価) の間の値として付与されており、正解の類似度とどの程度同様の推定を行えるかを測る。

JCommonsenseQA　JCommonsenseQA は、モデルの常識推論能力を評価するための 5 択 QA 問題である。自由回答でないため、回答は 1 から 5 までの数字で行う。

1）　https://github.com/llm-jp/llm-jp-eval

4 結果

評価実験の結果を表 1 にまとめた。JNLI は exact match でスコアが算出されており、JSTS については Pearson 相関係数でスコアが算出されている。まず、プルーニングの割合については、モデル全体のパラメータに対して 0.5%のプルーニングを行なった際は、ほとんどの設定でオリジナルモデルとの性能の違いが現れなかった。これに対して 2.0%をプルーニングした際には、全てのタスクで性能が低下する傾向があることがわかる。3 つのタスクにおいて、性能が大きく低下しているケースとあまり影響がないケースが見られるが、性能が低下している場合では、どのタスクにおいても、few-shot で与えた問題例の回答をそのまま繰り返したり、few-shot で与えた問題の例の中にある単語を無意味な回答として提示するような挙動がほとんどであった。少なくとも、与えられた例に従って個々の問題に取り組むことはできなくなるようである。ただし、2.0%のプルーニングにおいても、JNLI と JCommonsenseQA においては、20-24 層、24-28 層という比較的深い層のプルーニングではほとんど影響がなかった。このことは、同じ数のパラメータの削除を行ったとしても、どの層に対してプルーニングの操作を行なうかによってモデルへの影響が異なる可能性を示している。また、タスクによるプルーニングの影響の差を見ると、JSTS においてのみ、深い層のプルーニングをした際にも大幅な性能低下が起きている。出力を確認すると、この場合の性能低下も他のケースと似ており、事前に与えたプロンプトでの例の回答をただ繰り返してしまっている (ほとんどの問題に対して直前に見た例の回答である類似度の数値を回答してしまっている)。このような結果から、3 種類実施したいずれのタスクにおいても同様の理由による性能の低下が見られ、それぞれのベンチマークで本来調べたい能力 (常識的な知識や、文章の類似度を測る能力) が低下しているというわけではないようだった。

5 今後の課題

今回の実験では、32 層のモデルを 5 層ずつに区切り一部をプルーニングすることでモデルを損傷させたが、特にモデル全体のパラメータのうち 2.0%をプルーニングする実験の際に、レイヤーによって損傷の影響が異なる可能性が示された。その場合、20-24 層や 24-28 層などの深い層は、今回のタスクで正解するためにあまり影響していないといえ、このことはさらに多くのベンチマークによってモデルの性能を評価することで、より深く調査することができる。今回は小規模な 3 つのベンチマークのみを採用したが、今後はより多くのベンチマークでの実験を行っていきたい。

また、プルーニングによってモデルの性能が低下したケースでは、タスクの種類に関わらず、入力プロンプトの意図を適切に処理出来ず、few-shot として与えられた回答の例をただ繰り返してしまう、つまりほとんどの問題で同じ回答を繰り返してしまうことにより性能が大幅に低下していることがわかった。このことは、比較的浅い層が入力プロンプトの意図の理解に関連している可能性を示唆しているものの、どのタスクにおいても同様の振る舞いを行なっているため、タスクごとで異なる振る舞いを見ることはできなかった。今後は、プルーニング対象の重みを選定するアルゴリズムの変更しつつ、プルーニングの割合を更に調整することによって、タスクが本来求めている能力 (知識や推論の能力など) がモデルの損傷によりどれくらい失われたのかを調べられるようにしたい。

6 おわりに

本稿では、脳機能におけるシナプス刈り込みと機械学習におけるプルーニング技術について類似点と相違点を整理し、シナプス刈り込みにアイデアを得ているプルーニングを用いて、大規模言語モデルの意図的な損傷を行い、振る舞いの変化を定量的に評価した。プルーニングを利用した言語モデルの調査は近年の大規模な言語モデルに対する新たな説明的手法となりうるもので、実験の結果、レイヤーごとに性能低下の影響が異なること可能性が示された。ただし、複数のタスクにおいて同様の理由による性能低下を起こしているため、より深くモデルへの影響を調査するためには、手法の改善が必要である。

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Hua JY and Smith SJ. Neural activity and the dynamics of central nervous system development. **Nat Neurosci.**, 2004.

[2] 渡邉貴樹, 上阪直史, 狩野方伸. 生後発達期の小脳におけるシナプス刈り込みのメカニズム. **Journal of Japanese Biochemical Society**, Vol. 88, No. 5, pp. 621–629, 2016.

[3] Peter Penzes, Michael E Cahill, Kelly A Jones, Jon-Eric VanLeeuwen, and Kevin M Woolfrey. Dendritic spine pathology in neuropsychiatric disorders. **Nature neuroscience**, Vol. 14, No. 3, pp. 285–293, 2011.

[4] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. **arXiv preprint arXiv:2305.11627**, 2023.

[5] Avinash R Vaidya, Maia S Pujara, Michael Petrides, Elisabeth A Murray, and Lesley K Fellows. Lesion studies in contemporary neuroscience. **Trends in cognitive sciences**, Vol. 23, No. 8, pp. 653–671, 2019.

[6] Christopher A Walsh. Peter huttenlocher (1931–2013). **Nature**, Vol. 502, No. 7470, pp. 172–172, 2013.

[7] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. **Advances in neural information processing systems**, Vol. 2, , 1989.

[8] Steven A Janowsky. Pruning versus clipping in neural networks. **Physical Review A**, Vol. 39, No. 12, p. 6600, 1989.

[9] Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. **Advances in neural information processing systems**, Vol. 1, , 1988.

[10] MICHAEL C. MOZER and PAUL SMOLENSKY. Using relevance to reduce network size automatically. **Connection Science**, Vol. 1, No. 1, pp. 3–16, 1989.

[11] Ehud D Karnin. A simple procedure for pruning back-propagation trained neural networks. **IEEE transactions on neural networks**, Vol. 1, No. 2, pp. 239–242, 1990.

[12] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? **Proceedings of machine learning and systems**, Vol. 2, pp. 129–146, 2020.

[13] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. **Advances in neural information processing systems**, Vol. 28, , 2015.

[14] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. **ACM Transactions on Intelligent Systems and Technology**, 2023.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[16] Jesse Vig. Bertviz: A tool for visualizing multihead self-attention in the bert model. In **ICLR workshop: Debugging machine learning models**, Vol. 23, 2019.

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **International conference on machine learning**, pp. 3319–3328. PMLR, 2017.

[18] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? natural language attack on text classification and entailment. **arXiv preprint arXiv:1907.11932**, Vol. 2, p. 10, 2019.

[19] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. **arXiv preprint arXiv:1905.05950**, 2019.

[20] Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. **arXiv preprint arXiv:2307.13339**, 2023.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[22] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023.