似た単語の知識ニューロンは似た形成過程を経る

有山知希 [1,2] Benjamin Heinzerling[2,1] 穀田一真 [1,2] 乾健太郎 [3,1,2]
[1] 東北大学 [2] 理化学研究所 [3]MBZUAI
{tomoki.ariyama.s3, kokuta.kazuma.r3}@dc.tohoku.ac.jp
benjamin.heinzerling@riken.jp, kentaro.inui@tohoku.ac.jp

概要

近年，言語モデルの高い性能を実現している仕組みを理解すべく様々な研究が行われているが，これらの研究の中には事前学習によって得られたモデルパラメータを“ニューロン”と見立て，入出力やタスク性能との関連を調べているものがある．本研究では，ニューロンとモデルの学習過程との関係に焦点を当て，ある単語に反応するニューロンが学習中にどのように形成されるのかを，その単語が持つ意味の側面と合わせて分析した．その結果，ある単語のニューロンの形成過程は，その単語と意味的に同類関係にある単語のニューロンの形成過程と類似する傾向が観察された．また，同類関係にある単語間では形成後のニューロンも共通する傾向が見られた．

1 はじめに

近年の言語モデルの性能向上は目覚ましく，その性能の高さを達成しているダイナミクスを解釈するために様々な研究が行われている．これらの研究の中には，モデルパラメータに着目して入出力 [1] やタスク [2] との関係を調査しているものがある．さらに Chen ら [3] や Voita ら [4] は，学習されたパラメータを“ニューロン”と捉え，言語モデルが持つ知識との関連やニューロンの役割を分析している．

こうした背景を踏まえ，本研究ではニューロンと言語モデルの学習過程との関係に焦点を当てる．具体的には，ニューロンがモデルの事前学習中にどのように形作られているのかを，そのニューロンが反応する単語と関連付けて調べることにした．

調査の結果，ある単語に反応するニューロンの形成過程は，その単語と同類関係にある単語（例えば，“Cat”に対する“Dog”のような動物グループなど）に反応するニューロンの形成過程と類似する傾向が見られることを確認した（図 1）．さらに，学習によって形成されたニューロンについても，同類関係にある単語間では共通するニューロンが多い傾向が観察された．なお，実験コードは全て公開する[1)]．

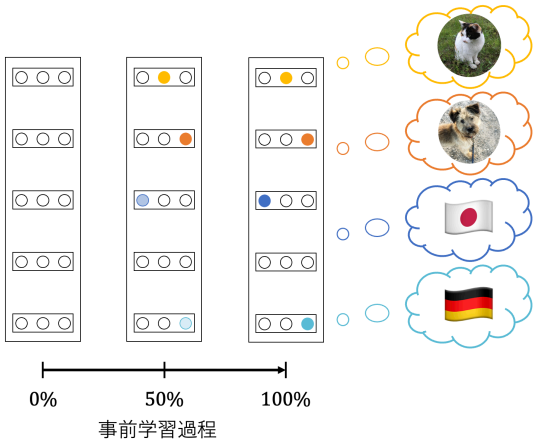

図 1 同類関係にある単語のニューロンは，類似した形成過程を経る傾向が観察された．

2 手法

2.1 ニューロン

本論文において，“ニューロン”とは Transformer[5] 内の Feed-Forward 層（以下，「FF 層」と呼ぶ）における，活性化関数を通過した後のベクトルの各要素のことを指す．これは，FF 層がキーバリューメモリと類似した働きをするという Geva ら [6] の主張から，バリューを引き出すキーの部分をニューロンの活性化に見立てる考え方による [7]．ここで，FF 層は入力を $\boldsymbol{x}$，第一・第二線形層の重みをそれぞれ $\boldsymbol{W}_1, \boldsymbol{W}_2$，バイアス項を $\boldsymbol{b}_1, \boldsymbol{b}_2$ で表し，活性化関数として GELU[8] を用いると，次の式 (1) で表される：

$$\mathrm{FF}(\boldsymbol{x}) = (\mathrm{GELU}(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1))\boldsymbol{W}_2 + \boldsymbol{b}_2 \qquad (1)$$

式 (1) 中のベクトルである“$\mathrm{GELU}(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)$”の各要素がニューロンに対応し，そのスカラ値がニューロンの活性値に対応する．

1) www.github.com/tomokiariyama/knowledge-neuron-sim

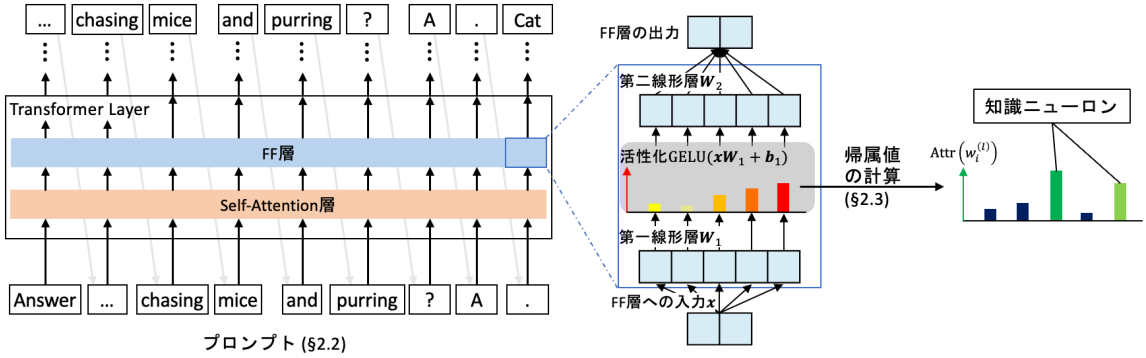

図 2　知識帰属法のイメージ．知識帰属法では，モデルがプロンプトの次トークンを予測する際の各ニューロンの活性値から帰属値を算出し，知識ニューロンを見つける．

2.2　単語の出力に寄与するニューロン

ニューロンの中には，ある単語の出力に特に寄与するものの存在が Dai ら [7] により報告されている．このようなニューロンを “寄与ニューロン” と呼ぶ．

寄与ニューロンを探すためのタスク　Dai ら [7] はマスク言語モデルを対象に実験を行っていたが，本研究では生成型言語モデルを扱うため，寄与ニューロンを探すためのタスクとして，ある単語が正解となる質問文を含むプロンプトを入力し，その続きをモデルに生成させるタスクを用いる．具体的なプロンプトの形式は，質問文の前後に指示文を追加したものとした．下記に一例を示す：

- “Answer the following question in one word: Q. What animal is famous for chasing mice and purring? A. ”

2.3　知識帰属法

本節では Dai ら [7] によって提案された，学習済み言語モデルから寄与ニューロンを探す手法である知識帰属法を，本研究に即した形で説明する [9](図 2 参照)．以下では，ある単語 T を扱うことを考え，その寄与ニューロンを探す方法について述べる．

まず，言語モデル内の l 番目の FF 層の中で i 番目に存在するニューロンを $w_i^{(l)}$ で，その活性値を $\hat{w}_i^{(l)}$ で表す．その上で，モデル内の各ニューロン $w_i^{(l)}$ のうち，“モデルがプロンプト s の続きとして正しい答え T を生成する確率 $\mathrm{P}_s(\hat{w}_i^{(l)})$” に大きな影響を与えているものを探す．影響の大きさは帰属値 $\mathrm{Attr}(w_i^{(l)})$ によって測るため，その計算方法を説明する．

帰属値 $\mathrm{Attr}(w_i^{(l)})$ の計算に必要な上述の確率 $\mathrm{P}_s(\hat{w}_i^{(l)})$ は，次の式 (2) で与えられる：

$$\mathrm{P}_s(\hat{w}_i^{(l)}) = p(T|s, w_i^{(l)} = \hat{w}_i^{(l)}) \tag{2}$$

この確率について，Sundararajan ら [10] の “Integrated Gradients” という帰属法を用い，あるニューロン $w_i^{(l)}$ について，学習済み言語モデルにおける活性値 $\bar{w}_i^{(l)}$ を，倍率 $\alpha = 0$ から 1 までで変化させたときの単語 T の生成確率の勾配を足し合わせた値として，帰属値 $\mathrm{Attr}(w_i^{(l)})$ が計算される（式 (3)）：

$$\mathrm{Attr}(w_i^{(l)}) = \bar{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial \mathrm{P}_s(\alpha \bar{w}_i^{(l)})}{\partial w_i^{(l)}} d\alpha \tag{3}$$

この帰属値が大きいほど単語 T に強く反応するニューロン，つまり寄与ニューロンと判断する．

本研究では，各単語について質問文を 4 文ずつ用意した上で，この知識帰属法によりモデル内の全ニューロンの帰属値を計算し，帰属値の平均が大きいニューロンをその単語の寄与ニューロンと判定する．なお，帰属値平均が大きい方からいくつまでのニューロンを寄与ニューロンとするかはハイパーパラメータであるが，本論文ではこのハイパーパラメータを 1,000 として調査した結果を報告する．また，説明の簡単のため，以下では特に学習済みモデルから発見される寄与ニューロンのことを，学習によって単語の知識をエンコードしているとの考えから “知識ニューロン” と呼んで区別する．

3　実験設定

3.1　データ

本研究では，実験対象の単語が答えとなるような質問文を ChatGPT [11] に生成させることでデータセットを作成した．なお，2.3 節で述べたように，質問文は 1 単語につき 4 文ずつ作成している[2)]．

単語のグループ – 同類関係　本研究で用いる単語は，知識ニューロンの形成過程との関連を調べ

2)　生成された質問文の妥当性は筆頭著者が目視で確認した．

表 1　使用した単語とそのグループ

グループ名	単語
動物	Cat, Dog, Fox, Ant, Bat, Rat, Hen, Wolf
色	Blue, Red, Green, Black, White, Brown, Gray
国	Canada, Germany, Brazil, France, England, India, China, Mexico, Russia, Iran, Japan, Israel, Australia, Spain
言語	English, Spanish, French, Chinese, Japanese, German, Russian, Italian, Portuguese

るため，意味的な関係に基づき，表 1 に示す 4 グループ計 38 個を用いることとした．ここで，同じグループに含まれる単語の組のことを “同類”，異なるグループに含まれる単語の組を “非同類” と呼ぶ．下記はその具体例である：

- 同類の例：(Cat, Dog), (Canada, India)
- 非同類の例：(Cat, Blue), (English, India)

3.2　モデル

実験で使用するモデルには，学習途中のチェックポイントが利用可能な Pythia [12] の 410M サイズ（ニューロン総数 = 98,304 個）を用いた．

チェックポイント　Pythia ではパラメータ更新回数である “step 数” に基づいてチェックポイントが公開されており，本研究ではそれらの中から下記に示す計 20 個を用いた：

- step0, 512, 1000, 3000, 5000
- step10,000 から step140,000 まで，10,000step 刻み
- step143,000

3.3　知識ニューロンの形成過程の定義

2.3 節で述べたように，ある単語についての知識ニューロンは学習済みモデルから 1,000 個発見されるが，これらのニューロンについてのチェックポイントを通じた帰属値変化のことを “知識ニューロンの形成過程” として定義する．すなわち，発見されたある 1 つの知識ニューロン n_1 に着目したとき，そのニューロンの形成過程 $\boldsymbol{p}_{n_1}$ は，最初のチェックポイントにおける帰属値 $\mathrm{Attr}_{n_1}^{\mathrm{step0}}$ から最後のチェックポイントにおける帰属値 $\mathrm{Attr}_{n_1}^{\mathrm{step143,000}}$ までを順に並べたベクトル（式 (4)）で表される：

$$\boldsymbol{p}_{n_1} = (\mathrm{Attr}_{n_1}^{\mathrm{step0}},\ \ldots,\ \mathrm{Attr}_{n_1}^{\mathrm{step143,000}}) \tag{4}$$

よって，ある単語 T の全知識ニューロンの形成過程 $\boldsymbol{P}_T$ は，式 (4) を用いて次の式 (5) で表される：

$$\boldsymbol{P}_T = (\boldsymbol{p}_{n_1}, \cdots, \boldsymbol{p}_{n_{1,000}})^\top \tag{5}$$

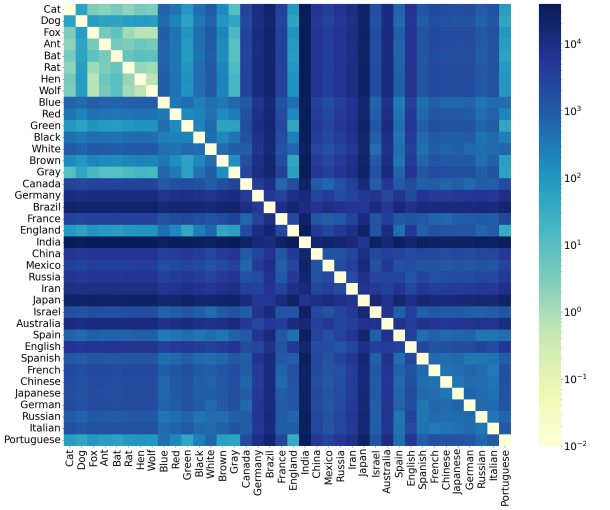

図 3　単語間の形成過程の距離（対数スケール）．対数プロットの都合で本来 0 の対角成分を 10^{-2} に置換している．

4　実験

4.1　知識ニューロンの形成過程と単語の同類関係

3.3 節の定義に則ると，表 1 内の全単語について知識ニューロンの形成過程 $\boldsymbol{P}_T$ が求められる．これらを用いて単語間での形成過程の距離を測定することを通じ，知識ニューロンの形成過程と単語の同類関係との間に関連性が見られるかを調査する．

ここで測定に用いる距離として，Rubner ら [13] によって提案された “Earth Mover’s Distance (EMD)” を用いる．この EMD は，2 つの分布が与えられた時，一方の分布を動かしてもう一方の分布に一致させるのに必要最小な距離を測定するものである．すなわち，2 つの分布を $\boldsymbol{X}, \boldsymbol{Y}$，分布の各要素を x_i, y_j，x_i から y_j への移動量を f_{ij} で表すと，EMD は次の式 (6) の分子が最小となる場合として定義される[3]：

$$\mathrm{EMD}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_i \sum_j f_{ij}}{\sum_j y_j} \tag{6}$$

本研究ではこの EMD を拡張した，分布の各要素がベクトルである場合の距離を測る 2 次元ワッサースタイン距離 EMD2 [14] を用い，単語 T_1 と T_2 の全知識ニューロン形成過程の距離 D_{T_1,T_2} を次の式 (7) で測定する：

$$D_{T_1,T_2} = \mathrm{EMD2}(\boldsymbol{P}_{T_1}, \boldsymbol{P}_{T_2}) \tag{7}$$

これはすなわち，各知識ニューロンの形成過程 $\boldsymbol{p}_n$ を 1 つの要素と捉えた上で，全知識ニューロン形成過程間の類似度を求めていることに相当する[4]．

3）　移動コストを定数とする等の詳細な制約は省略している．
4）　距離が小さいほど，類似度が高いことに相当する．

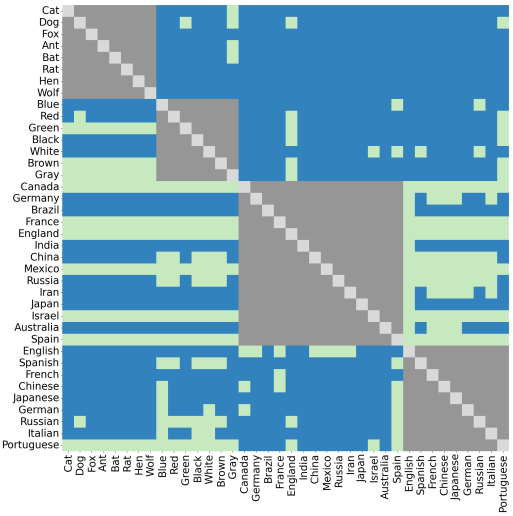

図 4　同類・非同類間の形成過程の距離の比較．同類間の距離は一律濃灰色で，非同類間の距離は，同類間の平均距離より大きい場合に青，小さい場合に薄緑で表している．

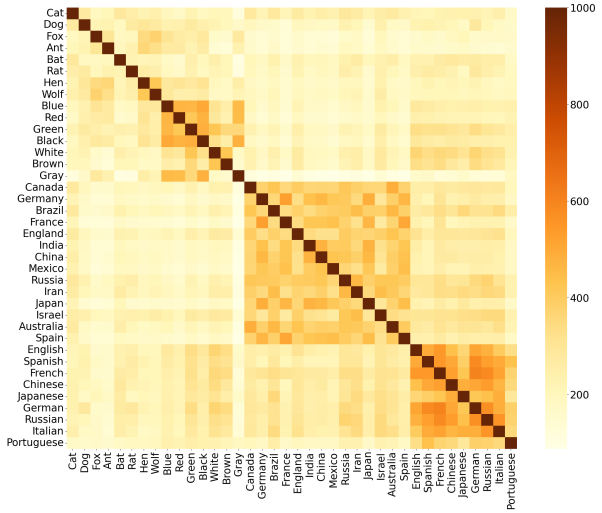

図 5　発見される知識ニューロンの積集合の要素数．

表 2　図 4 から算出した，単語グループごとの “同類間の平均距離 < 非同類間の距離” となっている割合．

グループ名	同類間の平均距離の方が小さい割合 [%]
動物	94.0
色	78.6
国	42.3
言語	76.3

4.2　形成された知識ニューロン

本節では追加の分析として，知識ニューロンが単語間でどの程度被っているかを調査する．具体的には，単語 T_1 と T_2 について，その知識ニューロン集合 S_{T_1} と S_{T_2} の積集合の要素数を調査する．

5　実験結果

5.1　知識ニューロンの形成過程と単語の同類関係

知識ニューロン形成過程の距離 D をプロットした結果を図 3 に示す．ここから更に解釈を深めるため，3.1 節で述べた “単語の同類・非同類” の観点から図 3 をプロットし直したものを図 4 に示す．

この図 4 の作成方法を説明する．最上行の単語 “Cat” に着目すると，この行で濃灰色で表された列の単語は “Cat” と同類の単語である．図 3 でこの濃灰色の部分に対応する距離の平均を計算することで，“Cat” と同類単語間の平均距離 $\overline{D}_{\mathrm{Cat}}$ を求めることができる．最上行における残りの各列，すなわち “Cat” と非同類の単語については，対応する図 3 の距離がこの平均距離 $\overline{D}_{\mathrm{Cat}}$ よりも大きい場合に青，小さい場合に薄緑でプロットをしている．この比較を図 3 の各行について行うことで図 4 が作成される．

図 4 から，単語グループごとに “同類間の平均距離 < 非同類間の距離” となっている割合を集計した結果を表 2 に示す．表 2 より，国グループを除くグループで，同類間の知識ニューロン形成過程が非同類間のそれよりも類似している傾向が見られる．

5.2　形成された知識ニューロン

4.2 節の調査を行った結果を図 5 に示す．この図 5 より，動物グループを除くグループでは，同類間でモデル内の同じニューロンが知識ニューロンになる傾向が認められる．

以上，5 章の結果から，ニューロンは事前学習によって同類関係を捉えていることが示唆される．

6　おわりに

本研究では，言語モデルの学習過程に着目した上で，モデル内に存在するニューロンの観点から分析を行った．分析の結果，事前学習の末にある単語の出力に寄与するニューロンは，その単語と同類関係にある単語のニューロンと学習中の形成過程が似ている傾向が観察された．また，学習によって形成される知識ニューロンは，同類関係にある単語同士でより共通している傾向も見られた．

本論文では，知識ニューロンとその形成過程について得られた結果を示したが，言語モデルの学習機構の観点などからこれらの結果により良い解釈を与えることは，重要な研究課題として残っている．

謝辞

本研究は，JST，CREST，JPMJCR20D2，および JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の支援を受けたものである．

参考文献

[1] Mengxia Yu, De Wang, Qi Shan, and Alvin Wan. The Super Weight in Large Language Models. **arXiv preprint arXiv:2411.07191**, 2024.

[2] Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding Skill Neurons in Pre-trained Transformer-based Language Models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[3] Lihu Chen, Adam Dejl, and Francesca Toni. Analyzing Key Neurons in Large Language Models. **arXiv preprint arXiv:2406.10868**, 2024.

[4] Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in Large Language Models: Dead, N-gram, Positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 1288–1301, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)**, 2017.

[6] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5484–5495.

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge Neurons in Pretrained Transformers. **arXiv preprint arXiv:2104.08696**, 2021.

[8] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). **arXiv preprint arXiv:1606.08415**, 2016.

[9] 有山知希, Benjamin Heinzerling, 乾健太郎. 言語モデルの学習における知識ニューロンの形成過程について. 言語処理学会 第 29 回年次大会 発表論文集, pp. 1169–1174, 2023.

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In **Proceedings of the 34th International Conference on Machine Learning**, Vol. 70, pp. 3319–3328, 2017.

[11] OpenAI. OpenAI: Introducing ChatGPT, 2022. https://openai.com/index/chatgpt/.

[12] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, KyleO'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff ほか. Pythia: A suite for analyzing large language models across training and scaling. In **International Conference on Machine Learning**, pp. 2397–2430. PMLR, 2023.

[13] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In **Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)**, pp. 59–66, 1998.

[14] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. **Journal of Machine Learning Research**, Vol. 22, No. 78, pp. 1–8, 2021.