認知ファインチューニング：眼球運動による大規模言語モデルのファインチューニング

染谷大河　大関洋平
東京大学
{taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp

概要

近年、眼球運動データや脳波データなどの人間の生理指標データを用いて、人間の言語処理における特徴を反映し、より精度よく下流タスクを解くことができる言語モデルを構築しようとする試みがある。しかし、既存手法は特定の下流タスクでの性能向上に特化している、もしくはモデルのアーキテクチャ自体に変更が必要な手法となっており、既存の大規模言語モデルに適用して幅広い下流タスクにおける性能を向上しうる汎用的な手法となっていない。そこで本研究では、既存の大規模言語モデルに適用可能な新たな手法として、「認知ファインチューニング」を提案し、その方法論的妥当性を検証する。

1　はじめに

Transformer[1] をベースとした大規模言語モデル (LLM) が自然言語処理のあらゆるタスクで使用されるようになり、高い性能を発揮している (e.g., GPT-3 [2])。特に、近年では人間からの明示的なフィードバックを用いて、強化学習の枠組みで言語モデルの出力をより人間の価値基準に沿うように矯正するような手法群が研究されている [3]。

一方で、眼球運動データや脳波データなどの人間の生理指標データには、人間の言語処理に関する豊富な情報がエンコードされていることが知られており、例えば脳波データから発話内容や読んでいる単語の品詞を推定できることなどが報告されている [4, 5]。また一方で、人間の生理指標データを用いて言語モデルを訓練することで、人間の言語処理における特徴を言語モデルが首尾よく捉えられるように矯正し、品詞タグ付けや固有表現抽出など特定の下流タスクでの性能向上を試みる研究群がある [6, 7, 8]。

しかし、これらの研究で提案されている手法は、特定の下流タスクでの性能向上に特化している、もしくはモデルのアーキテクチャ自体に変更が必要な手法となっており、既存の大規模事前学習モデルに適用して幅広い下流タスクにおける性能を向上しうる汎用的な手法となっていない。また、推論時にも生理指標データが必要な手法も多く、活用範囲が必ずしも広くない。特に、アーキテクチャの変更が必要になることは、大規模データで学習した大規模モデルにて、幅広い下流タスクに対応しようとする現行の LLM 時代のパラダイムと不適合である。

そこで本研究では、特定のタスクに特化せず、かつ推論時に人間の生理指標データが不要である手法として、「認知ファインチューニング」を提案し、その方法論的妥当性を検証する。本手法は、言語モデルの埋め込みを入力とするフィードフォワードネットワークを通して、言語モデルの埋め込みから人間の読み時間を予測する追加学習を行うことにより、元の言語モデルがより人間の読みの特徴を捉えられるように矯正するものであり、既存の事前学習済み大規模言語モデルを人間の生理指標データを用いて追加学習可能な手法である。実験の結果、言語モデリングタスクと読み時間の予測タスクを組み合わせて学習することにより、単純に言語モデリングで学習した場合よりも、より人間の読みの特徴を捉えた言語モデルを構築できる可能性が示された。

2　先行研究

眼球運動データや脳波データなどの人間の生理指標データを用いて、言語モデルの性能向上を目指した研究群は数多く提案されている [6, 7, 8]。提案手法の多くは、品詞タグ付け [6] や固有表現抽出 [8] など特定の下流タスクでの性能向上を目指したものであり、一般に推論時にも生理指標データが必要な手法になっている。語タイプごとに読み時間を集約

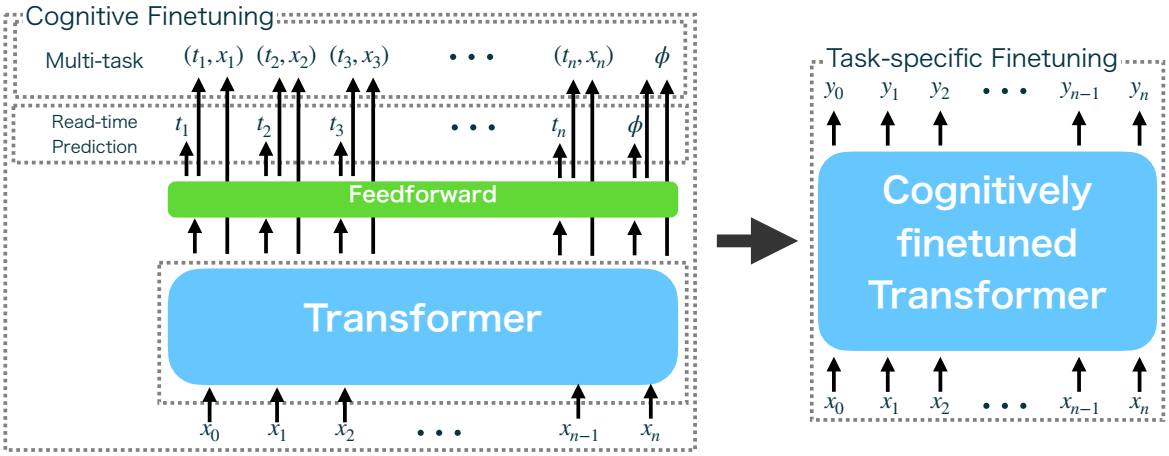

図 1　認知ファインチューニングの概念図。本研究では、1) 次トークン読み時間予測のみを行う手法 (Read-time Prediction) と、2) 読み時間予測と次トークン予測を同時に行う手法 (Multi-task) の二種類の手法を検証する。

するなどの工夫により、推論時に生理指標データが不要になる手法も提案されているが [9, 10]、基本的に LSTM ベースのモデルが用いられており、タスクの種類に依らない汎用的な手法となっていない。Transformer ベースのモデルを用いて、特定の下流タスクでの性能向上にとどまらない汎用的な手法を目指した手法も提案されているが [11]、当該手法でも元の言語モデルのアーキテクチャの変更が必要であり、既存の事前学習済み大規模言語モデルには適用不可である。

3　実験

3.1　認知ファインチューニング

図 1 に、本研究で検証する認知ファインチューニングの方法を示す。認知ファインチューニングは、Trasformer をベースとした事前学習済み言語モデルと、それにより得られた各トークンに対する言語モデルの埋め込みを入力とし読み時間を予測するフィードフォワードネットワークを用意し、その両方を人間の読み時間アノテーションつきコーパスで学習することで行われる。この方法により追加学習された言語モデルは、さらにタスク特化のファインチューニングを行うことにより任意の下流タスクで使用することが可能である。本研究では、1) 入力されたトークンの次のトークンに対する読み時間予測のみを行う手法と、2) 読み時間予測と次トークン予測を同時に行う手法の二種類の手法を用いて学習したモデルにおける、読み時間データモデリングの性能の向上を確認することで、認知ファインチューニングの方法論的妥当性を検証する。

3.2　データ

データ概要　本研究では、浅原ら [12] により構築された BCCWJ-EyeTrack を用いた。BCCWJ-EyeTrack は、『現代日本語書き言葉均衡コーパス』[13] の新聞記事サンプル 20 記事に対して, 日本語母語話者 24 人分の読み時間が付与されているコーパスである。18 記事（1053 文節）を学習用、2 記事（208 文節）を検証用、2 記事（217 文節）をテスト用データとして使用した。

読み時間データの前処理　BCCWJ-EyeTrack では 5 種類の読み時間データが計測されているが、本研究では TOTAL (Total Time) と FPT (First Pass Time) を予測対象とし、各読み時間 RT に対して $\log_2(RT + 1 \times 10^{-5})$ を予測すべき正解の読み時間とする。また、BCCWJ-EyeTrack では読み時間は文節を単位として付与されており、一般に文節の単位は言語モデルが入力として受け取るトークンの単位と異なる。従って、本研究では以下の方法で文節に付与された読み時間を、言語モデルの入力となるトークンごとの読み時間に分配した: まず、各文節に付与された読み時間をその文節の系列長で割った読み時間を各文字に対する読み時間として定義する。次に、その各文字に付与された読み時間を言語モデルの入力となるトークンごとに集約して、そのトークンに対する正解読み時間とした。

3.3　実験設定

言語モデル　本研究では、rinna 株式会社により公開されている埋め込み次元 1024 次元、16 ヘッド、

24 層の GPT-2 言語モデル[1)] を用いた。

実験条件　本研究では、二種類の認知ファインチューニングの手法と、ベースライン手法 4 つを合わせた以下の 6 つの学習方法を検証する:

Presurp
: 事前学習済み言語モデルを追加学習せずそのまま用いる。

TextOnly
: 事前学習済みモデルを BCCWJ-EyeTrack に含まれるテキストを対象に言語モデリングタスクでファインチューニングする。

EyeTrack
: 事前学習済みモデルを入力されたトークンの次のトークンに付与された読み時間を予測するタスクでファインチューニングする（図 1 における Read-time Prediction に対応）。

MultiTask
: 事前学習済みモデルを入力されたトークンの次のトークンに付与された読み時間を予測するタスクと、言語モデリングタスクのマルチタスクでファインチューニングする（図 1 における Multi-task に対応）。

EyeTrack (random)
: **EyeTrack** 条件をランダムな読み時間を用いて行う。

MultiTask (random)
: **MultiTask** 条件をランダムな読み時間を用いて行う。

ただし、各読み時間 (TOTAL/FPT) の学習データでの平均と分散と一致した平均と分散をもつ正規分布からサンプリングした値を、ランダムな読み時間として用いた。また、言語モデリングタスクでは次単語予測の損失 $\mathscr{L}_{\mathrm{NWP}}$、読み時間予測タスクでは正解時間との平均二乗誤差 $\mathscr{L}_{\mathrm{RTP}}$ を最小化するように学習した。また、マルチタスク条件では、$\mathscr{L}_{\mathrm{NWP}}$ と $\mathscr{L}_{\mathrm{RTP}}$ の重みつき和を最小化した:

$$\mathscr{L} = \alpha\mathscr{L}_{\mathrm{NWP}} + (1-\alpha)\mathscr{L}_{\mathrm{RTP}} \tag{1}$$

ただし、α は本実験では 0.5 とした。また、全ての条件において、5 つの異なるランダムシードを用いて学習した。その他全条件に共通のハイパーパラメータについては、付録 A に示した。

変数名	型	概要
prev_length	int	直前文節の文字数
length	int	文字数
prev_freq	int	直前文節内単語頻度の幾何平均
freq	int	文節内単語頻度の幾何平均
is_first	bool	行内最左要素
is_last	bool	行内最右要素
is_second_last	bool	行内右から 2 番目の要素
screenN	int	画面提示順
lineN	int	行提示順
segmentN	int	文節提示順
article	factor	記事情報
subj	factor	実験協力者 ID

表 1　本研究で用いた説明変数。各変数は既存研究 [12] で採用されているものである。

3.4　読み時間データのモデリング

本研究では、各条件で学習された言語モデルがどれほど人間の読みの特徴を捉えることができているかどうかを、読み時間データのモデリング性能を通して検証する。具体的には、あらかじめ既存研究 [12] で採用されているベースラインとなる特徴量（表 1）で読み時間をモデリングし、そこに各条件で学習された言語モデルから得られたサプライザルを特徴量として追加した際に、モデリング性能がどれほど向上するかを確認することで、その条件で学習された言語モデルがどれほど人間の読みの特徴を捉えられているかを評価する。先行研究にならい、各言語モデルによるサプライザルを追加した時の平均対数尤度の増加分 ΔLogLik を報告する。この値が大きいほど、より人間の読みの特徴を捉えられていることを意味する。ベースライン回帰モデルには、以下の線形混合モデルを用いる：

$$\begin{aligned}\log(\mathtt{RT}) \sim\ & \mathtt{prev\_length} + \mathtt{length} + \mathtt{prev\_freq} + \mathtt{freq} \\ & + \mathtt{is\_first} + \mathtt{is\_last} + \mathtt{is\_second\_last} \\ & + \mathtt{screenN} + \mathtt{lineN} + \mathtt{segmentN} \\ & + (1|\mathtt{article}) + (1|\mathtt{subj}).\end{aligned} \tag{2}$$

ただし、$(1|x)$ は x をランダム切片として組み込むことを指す。

4　結果

表 2 は、各条件における、ベースライン回帰モデルからの平均対数尤度の増加分 (ΔLogLik) を示す。TextOnly, EyeTrack, MultiTask 条件はいずれも Presurp 条件よりも ΔLogLik の値が高く、言語モデリングタ

1)　https://huggingface.co/rinna/japanese-gpt2-medium

	ΔLogLik (↑)	
	TOTAL	FPT
Presurp	13.99	13.62
TextOnly	14.26 ± 0.01	13.91 ± 0.01
EyeTrack	14.08 ± 0.20	13.70 ± 0.17
MultiTask	**14.32 ± 0.14**	**13.94 ± 0.12**
EyeTrack (random)	-180.06 ± 0.01	-123.29 ± 0.04
MultiTask (random)	-180.03 ± 0.01	-123.44 ± 0.02

表 2 各条件における、ベースライン回帰モデルからの平均対数尤度の増加分 (ΔLogLik)。

スクによる追加学習はモデリング性能の向上に寄与することが示された。TextOnly 条件が EyeTrack 条件よりも ΔLogLik の値が大きいことから、読み時間データによる学習単体による効果は、言語モデリングタスクによる学習単体の効果よりも小さいことが示唆された。一方で、EyeTrack 条件よりも MultiTask 条件における ΔLogLik の値が高いことから、言語モデリングタスクに加えて、読み時間データによる学習を行うことで、より人間の読み時間の特徴を捉えた言語モデルが構築できる可能性が示された。また、ランダムな視線データを用いた条件では、いずれも ΔLogLik は負の値となり、人間の読みの特徴をより反映できなくなることが示された。これは、実際の人間の読み時間データを用いることが精度向上に寄与することを意味する。

5 おわりに

本研究では、人間の読みの特徴を捉えたより「人間らしい」言語モデルを得るための新たなファインチューニングの方法として「認知ファインチューニング」を提案し、その方法論的妥当性を検証した。実験の結果、言語モデリングタスクと読み時間の予測タスクを組み合わせて学習することにより、単純に言語モデリングで学習した場合よりも、より人間の読みの特徴を捉えた言語モデルを構築できる可能性が示された。一方で、本研究では読み時間データのモデリングを通して、構築されたモデルが人間の読みの特徴をよりよく捉え、下流タスクをより精度よく解きう得るかを間接的に検証した。実際に、本研究で構築されたモデルを実際の下流タスクに応用して精度が向上するのかの検証や、脳波など他の人間の生理指標データを利用することは今後の課題としたい。また、読み時間データを用いて人間の読みの特徴を言語モデルに学習させるより効果的な手法の探索についても、今後の課題としたい。

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] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism, 2023.

[4] David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. **New England Journal of Medicine**, Vol. 385, No. 3, pp. 217–227, 2021. PMID: 34260835.

[5] Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. Decoding part-of-speech from human EEG signals. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2201–2210, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. Weakly supervised part-of-speech tagging using eye-tracking data. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 579–584, Berlin, Germany, August 2016. Association for Computational Linguistics.

[7] Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. Leveraging cognitive features for sentiment analysis. In Stefan Riezler and Yoav Goldberg, editors, **Proceedings of the 20th SIGNLL Conference on Computational Natural Lan-**

guage Learning, pp. 156–166, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Nora Hollenstein and Ce Zhang. Entity recognition at first sight: Improving NER with eye movement information. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1–10, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In Anna Korhonen and Ivan Titov, editors, **Proceedings of the 22nd Conference on Computational Natural Language Learning**, pp. 302–312, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[10] Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. Advancing nlp with cognitive language processing signals, 2019.

[11] Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 6327–6341. Curran Associates, Inc., 2020.

[12] 浅原正幸, 小野創, 宮本 エジソン正. Bccwj-eyetrack. 言語研究, Vol. 156, pp. 67–96, 2019.

[13] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. **Language Resources and Evaluation**, Vol. 48, No. 2, pp. 345–371, June 2014.

A　ハイパーパラメータ

全条件で共通のハイパーパラメータを表 3 に示す。

最適化手法	Adam
学習率	1e-6
エポック数	25
バッチサイズ	8
学習率スケジューラ	LinearLR

表 3　全条件で共通のハイパーパラメータ。