E-016

ノイズに着目している Attention Head の同定とスケーリングによる大規模言語モデル出力の正確性と多様性のトレードオフ制御
Controlling the Trade-off between Accuracy and Diversity in LLM Outputs by Identifying and Scaling Noise-Focusing Attention Heads

下村 晃生‡　花沢 明俊‡
Teruki Shimomura　Akitoshi Hanazawa

概要

大規模言語モデル(LLM)は，汎用的なテキスト生成能力から，様々な分野への活用が広がってきている．しかし，不必要な情報が含まれる場合，推論能力が大幅に低下することが知られている．[1]本論文では，そのような現象を起こす原因として，LLM の中に無数にある Head の中に，文章中のノイズに対して過度に Attention を向けている Head があるからであるという仮説を立て，検証を行った．その結果，そのような Head は LLM の中間層に多く局在し，この Head を調整することによりモデルの出力を操作ができることを実証した．

1. 研究背景

大規模言語モデル(LLM)は，小学生レベルの算数の文章問題データセットである GSM8K[2]をはじめとした各種タスクにおいてその数理推論能力を評価されてきたが，GSM-Symbolic[1]では，回答に無関係な人物名や数量，背景説明といったノイズを付与するだけで，モデルの精度が著しく低下する現象が報告されている．このように，LLM は不必要な情報が含まれると出力品質が劣化するプロンプトロバスト性の低さが問題となっているが，モデル内部でノイズに反応している要素を特定し，制御する手法は未だ確立されていない．

一方，Transformer の Attention Head 単位での役割分析は，Olsson et al.[3]や Wu et al.[4]らによって進められており，それぞれ特定のタスクや構文的特徴を捉えるヘッドの存在が示されている．本研究では，「文脈中のノイズに対して Attention を向けている Head も存在するのではないか」という仮説をたて，その有無を検証し，特定した Head が持つ役割を検証するために，Head の調整による出力特性への影響を検証することを目的とした．

2. 提案手法

2.1 Attention Head の同定

本研究で行った実験には meta 社の Llama-3.1-8B-Instruct モデルを使用した．

2.1.1 GSM-NoOp データセット

まず，GSM-NoOp データセットを独自に再現した．このデータセットは，Mirzadeh らによって提案[1]された小学生レベルの算数の文章問題に対して，回答に無関係な人物名や数量，背景説明を加えたものである．入力は小学生レベルの算数文章問題xであり，文章中には解答に無関係なノイズNが混入している．モデルf_θは $\hat{y} = f_\theta(x, N(x))$を生成するが，理想的には$N(x)$の有無で$\hat{y}$は不変となるべきである．また，評価指標は正答率である．

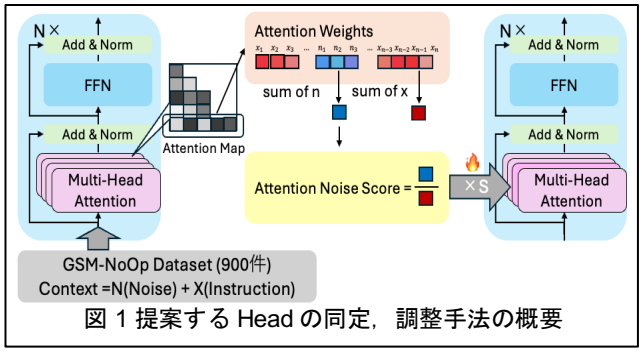

図 1 提案する Head の同定，調整手法の概要

2.1.2 GSM-NoOp データセットの再現

OpenAI 社の GPT-4o に Few-Shot を提示して 1300 件作成し，混入させたノイズが問題とは無関係であり，回答に影響を与えていないことは人手によって判断を行った．900 件を Head 同定用，400 問を性能評価用に分割した．混入させたノイズ部の平均トークンは 29.4，割合は 43.6%であった．

2.2 Noise-Focusing Attention Head の同定

図 1 にパイプラインを示した．

2.2.1 Noise Focusing Attention Head の同定

入力するトークン集合をノイズ部分Nと非ノイズ部分 C に分割し，レイヤl，ヘッドhの Noise-Focusing Attention Score (NF-Score)を式(1)で定義する．

$$NF - Score_{l,h} = \frac{\sum_{i \in N} \alpha_{l,h}(i)}{\sum_{j \in C} \alpha_{l,h}(j)} \quad (1)$$

ここで$\alpha_{l,h}(i)$はトークンiへ向けられた注意重みである．同定用の GSM-NoOp データセット 900 問全体で平均をとり，上位k %を NF-Head と呼ぶ．

2.3 Head スケーリングによる重み調整

特定した Head について，式(2)のように NF-Head が持つ役割を検証するために再学習を行わずに$value$行列にスカラーS (Scaling Factor)を掛けるだけで NF-Head の調整を行う．

$$H_{\mathrm{NF}} = softmax\left(\frac{W_q^i \times W_k^{i\,T}}{\sqrt{\mathrm{d_k}}}\right)\left(W_v^i \cdot S\right) \quad (2)$$

- $S = 1$: 元モデル
- $0 < S < 1$: NF-Head 抑制
- $S = 0$: NF-Head 無効化
- $S > 1$: NF-Head 増幅

2.4 評価指標

2.4.1 推論能力の評価

推論能力は，同定に使用していない GSM-NoOp データセット 400 件に対する正答率で評価を行った．

2.4.2 出力多様性の評価

Persona Hub[5]からランダムに抽出した 500 件のペルソナプロンプトを用いて小説作成用のプロンプトを作成し，小説を生成した．小説 500 件の生成文ベクトル $\{e_1, e_2, \ldots, e_N\}$ と，生成文集合$\{S_1, S_2, \ldots, S_N\}$に対して，Sentence-BERT を用いて式(3)に示した平均ペアワイズコサイン類似度と式(4)に示した Self-BLUE スコアを元に評価を行った．

$$AvgPairwiseCos = \frac{2}{N(N-1)} \sum \frac{e_i \cdot e_j}{||e_i|| \, ||e_j||} \quad (3)$$

$$Self-BLUE = \frac{1}{N} \sum_{i=1}^{N} BLUE(hyp = S_i, refs = \{S_j\}, j \neq i) \quad (4)$$

この時，平均ペアワイズコサイン類似度は意味的な多様性を，self-BLEU はトークンレベルの多様性を表し，いずれの指標も値が低いほど生成文の多様性が高いことを示す．

3. 結果

3.1 NF-Head の分布

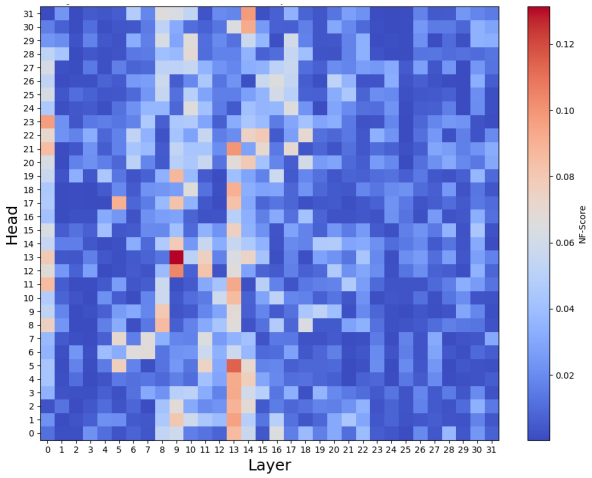

図 2 NF-Score の分布

図 2 にモデル内の NF-Score のヒートマップを示す．NF-Score が高い Head はレイヤ 9-12 付近に集中し，モデルの中間層に多く局在しており，多くの Head で極めて低い値を示している．

3.2 スケーリングによる出力の変化

表 1 スケーリングによる出力特性の変化

Scaling Factor	0.1	0.5	1(default)	1.5	2
GSM-NoOp Accuracy (%)	28.0	24.25	12	9.5	8.5
Self-BLUE	0.068	0.082	0.061	0.067	0.071
Avg Pairwise Cosine Distance	0.345	0.337	0.313	0.311	0.306

正確性 ⟷ 多様性

表 1 に，NF-Score が高い順に選んだ 10 個の Head に対してスケーリング操作を行った際の各評価指標の変化を示した．スケーリング操作前と比べると，Head を抑制することで，GSM-NoOp の正答率は上昇しているが，出力多様性は減少しており，逆に Head を増幅することで，GSM-NoOp の正答率は減少し，出力多様性の指標は上昇していることがわかる．

4. まとめ・今後の展望

不必要な情報に注意を向ける「Noise-Focusing Attention Head」を新たな概念として提案し，NF-Score によって定量的に同定した．Llama-3.1-8B-Instruct において，中間層に高スコアの NF-Head が集中して存在することが明らかになり，これらのヘッドがモデルの文脈理解のいて不必要な情報を取り込むボトルネックになっている可能性が示唆された．さらに，同定した NF-Head の value 行列にスケーリング係数 S を乗じる簡便な操作を行うことで，再学習を一切行わずに GSM-NoOp の正答率が 12%から 28%へ大幅に回復することを実証した．また，係数 S の値を変化させることで正答率(正確性)と出力の多様性との間にトレードオフが存在することを確認し，NF-Head が両者に寄与していることを明らかにした．

一方で，本実験では Llama 系モデルのみを対象としたため，他のアーキテクチャを持つ LLM でも同様の NF-Head の分布やスケーリング効果が再現されるかは未検証である．加えて，本研究ではヘッドの重みを単純にスケーリングする手法のみを検討したに留まり，ヘッド単位の再学習を使用した手法の有効性については今後の課題である．また，LLM の適用領域によって求められる性能特性は多岐に渡るため，本手法を合成データ生成や対話システムなど他タスクへと拡張し，その有効性を評価する必要がある．

参考文献

[1] I. Mirzadeh et al., “GSM-Symbolic: Understanding the Limitation of Mathematical Reasoning in Large Language Models”, arXiv preprint arXiv:2410.05229

[2] K. Cobbe et al., “Training Verifier to Solve Math Word Problems”, arXiv preprint arXiv:2110.14168

[3] C. Olsson et al., “In-Context Learning and Induction Head”, arXiv preprint arXiv:2209.11895

[4] W. Wu et al., “Retrieval Head Mechanistically Explains Long-Context Factuality”, arXiv preprint arXiv:2404.14474

[5] T Ge et al., “Scaling Synthetic Data Creation with 1,000,000,000 Personas”, arXiv preprint arXiv:2406.20094