In-model anti-expertによる大規模言語モデルのハルシネーション抑制

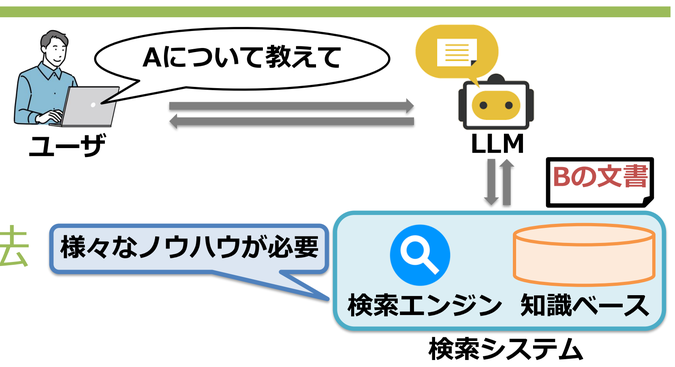
○門谷 宙, 西田 光甫, 西田 京介, 齋藤 邦子 (NTT)

概要

目標：外部知識を用いずに大規模言語モデル (LLM) のハルシネーションを抑制
既存手法：推論時に嘘つきLLM (anti-expert) を参照, 外部知識を用いず高性能だが適用コストが高い
提案手法：モデル内部にanti-expertの役割を担うモジュールを追加, モデル単体でのハルシネーション抑制を実現
実験結果：提案手法は低コストにハルシネーションを抑制し, LLMの真実性を向上させることを確認

背景

- RAG [Lewis+, 2020] など外部から知識を補うことでハルシネーションを抑制する手法の研究が盛んな一方で, 外部知識を用いない手法の研究は進んでいない
- RAGは検索に失敗すると機能しないため, RAGと併用可能な外部知識を用いない手法の研究も重要, 本研究で取り組む



既存手法：Anti-expert [Zhang+, 23]

手法概要

- LLMの真実性を直接向上させる学習は上手くいかない
- 元のLLM (base) を嘘が含まれるデータを用いてfine-tuningして嘘つきLLM (anti-expert) を構築
- 推論時にbaseとanti-expertのトークン予測確率を比較
- anti-expertの予測確率が閾値以上に高いトークンの生成確率を下げる

利点と課題

- ☺ 外部知識を用いないハルシネーション抑制手法の最高性能
- ☹ 適用コストが高い (パラメータサイズ：**2倍**, 生成時間：**1.6倍**)

提案手法：In-model anti-expert モデル単体でハルシネーションを抑制することで, 低コスト化を実現

手法概要

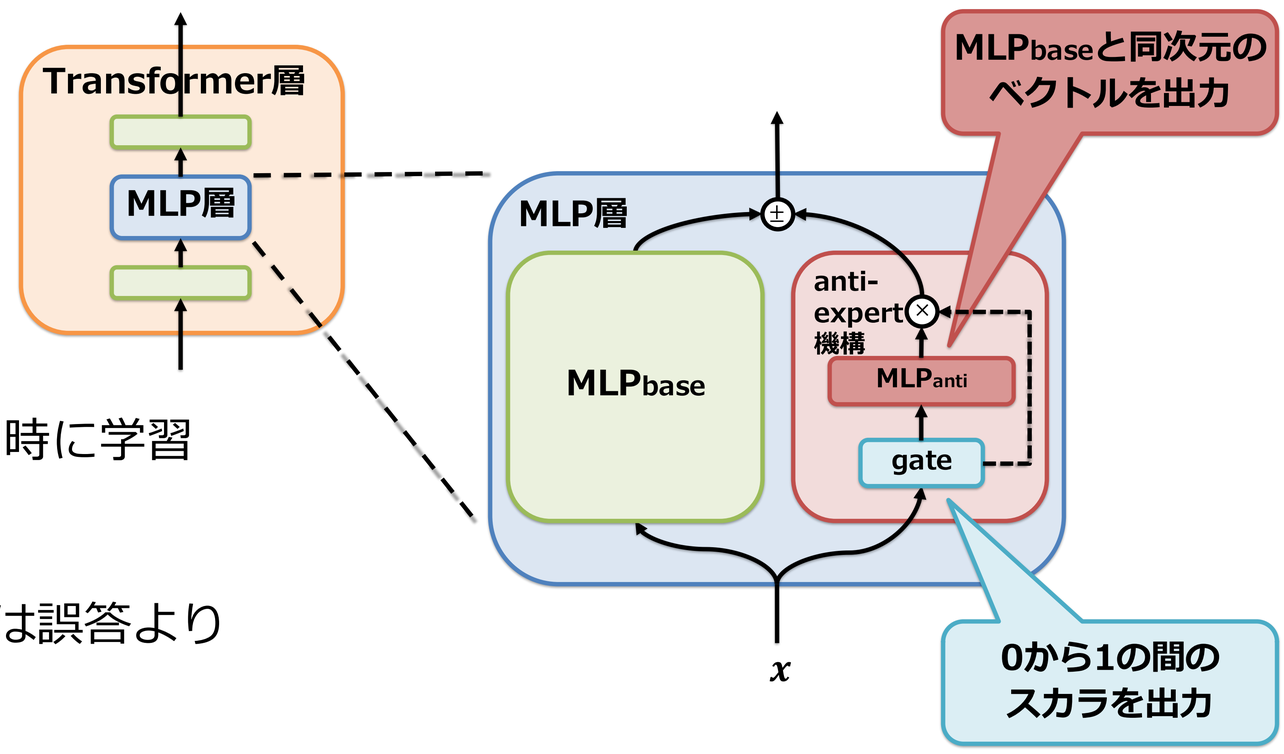
- baseのMLP層と並列にanti-expert機構を追加, anti-expert機構はゲートと小さなMLP層で構成
- baseのMLP層の出力にanti-expert機構の出力をプラスすれば誤答 (anti-expertモード), マイナスすれば正答 (expertモード) を生成するように学習, 推論時はexpertモードで生成

anti-expertモード：

$$\mathrm{MLP}(x) = \mathrm{MLP}_{\mathrm{base}}(x) + \mathrm{gate}(x) \cdot \mathrm{MLP}_{\mathrm{anti}}(x)$$

expertモード：

$$\mathrm{MLP}(x) = \mathrm{MLP}_{\mathrm{base}}(x) - \mathrm{gate}(x) \cdot \mathrm{MLP}_{\mathrm{anti}}(x)$$



学習方法

- baseは固定, anti-expert機構のパラメータのみを調整
- anti-expertモードとexpertモードをマルチタスクで同時に学習
- 入力文に正答と誤答が付与されたデータセットを使用
- 損失関数：normalized multiple choice
  anti-expertモードでは正答より誤答, expertモードでは誤答より正答の生成確率が高くなるように学習

実験 提案手法の効果をQAドメインで検証

設定

- 訓練データ：HaluEval, テストデータ：TruthfulQA
- ベースモデル：Llama2 (7B)
- 評価指標：MC1/2/3 (スコアが高いほど真実性が高い)

結果

- ベースモデルを正答で直接fine-tuningすると真実性が低下
- 提案手法はanti-expertの適用コストを改善
  (パラメータサイズ：2倍→**1.2倍**, 生成時間：1.6倍→**1.3倍**)
- 提案手法はハルシネーションを抑制し, anti-expert以外の既存手法を上回る性能を達成

	MC1	MC2	MC3
base	36.96	54.62	27.95
fine-tuning	27.78	45.21	22.31
DoLa [Chuang+, 23]	32.97	60.84	29.50
ITI [Li+, 23]	37.01	54.66	27.82
Anti-expert [Zhang+, 23]	46.32	69.08	41.25
In-model anti-expert (ours)	38.56	57.08	28.85

今後の展望

- anti-expert機構の小型化・一部の層のみへの追加による, さらなる低コスト化
- ゲート・損失関数の改善による, 抑制性能向上
- ゲートの出力値を用いたハルシネーション傾向分析