Разработка методов усиления следования контексту больших языковых моделей для достоверной генерации

Кратков Егор Вадимович
kratkov.ev@phystech.edu

Центр когнитивного моделирования МФТИ

2024 – Научно-исследовательский семинар

Введение

Мотивация

- Ответы на вопросы по документам в специфичной предметной области
- У модели помимо контекста присутствуют внутренние знания. Знания модели могут быть устарелые и отсутствовать в специфичных предметных областях
- Модели часто галлюцинируют и их ресурсозатратно дообучать
- Часто модели недостаточно полно раскрывают контекст

Цель работы

Усиление следования и отражения полноты контекста и уменьшение контекстных галлюцинаций в задачах ответа на вопросы на основе документов без дообучения модели

Постановка задачи

Дано

- LLM $\mathcal{M}$ с параметрами θ
- Запрос: $x = [x_1, x_2, ..., x_k]$, где x_i - i-ый токен. $x = prompt(qn)$ - текстовая инструкция (англ. prompt) от вопроса qn
- Контекст: c, содержащий документы, мы его считаем правильным и подобранным заранее
- На вход подается $q = f(c, x)$, обычно $q = [c, x] = [c, prompt(qn)]$
- Токен LLM: генерируется $y_t \sim p_\theta(y_t|[q, y_{<t}])$, где $p_\theta(y_t|[q, y_{<t}]) = softmax(logit_\theta(y_t|[q, y_{<t}]))$
- $\mathcal{M}_\theta(q) \to Y_{raw}$ - сгенерированный моделью ответ, выбирая на каждом шаге токен из распределения p_θ
- Y - может отличаться от Y_{raw}, т. к. модель может использваться несколько раз, иначе $Y = Y_{raw}$

Результат

$\mathcal{F}_\theta(q, c, qn) \to Y$, где $\mathcal{F}$ - алгоритм

Метрики

- RAGAs [1] (оценивание генерации, дополненной поиском, англ. RAG assesment)
  - $faithfulness = \frac{|V|}{|S|}$, где S - утв. V - утв., опирающихся на контекст,
  - $context\ relevance \frac{|R|}{|S|}$, где R - релевантные(уместные) утв. относительно контекста
  - $answer\ relevance = \frac{1}{n}\sum_{1}^{n} sim(qn, q_i)$, где q_i - вопросы, составленные по ответу
- RePASs [2] (Показатель стабильности ответа на нормативный акт) $RePAS(c, a) = \frac{E - C + OC + 1}{3}$, где E - следствие (англ. entailment), C - противоречие (англ. contradiction), OC - покрытие обязательств (англ. obligation coverage) с исходным контекстом

[1] Shahul Es, RAGAS: Automated Evaluation of Retrieval Augmented Generation

[2] Tuba Gokhan, RIRAG: Regulatory Information Retrieval and Answer Generation

Classifier-Free Guidance

Изначально CFG[3] был использован для диффузионных моделей

- Определили новое распределение $\hat{p}_\theta(y|c,x) \propto p_\theta(y|x) \cdot p_\theta(c|x,y)^\gamma$
- $log\ \hat{p}_\theta(y_t|c,x,y_{<t}) =$
  $log\ p_\theta(y_t|x,y_{<t}) + \gamma(log\ p_\theta(y_t|c,x,y_{<t}) - log\ p_\theta(y_t|x,y_{<t}))$,
  где $p_\theta(y_t|x,y_{<t})$ - безусловная генерация $p_\theta(y_t|c,x,y_{<t})$ - условная генерация
- Данную формулу можно интерпретировать как шаг γ в латентном пространстве от безусловной к условной генерации

[3] Guillaume Sanchez, Stay on topic with Classifier-Free Guidance

Context-Aware Decoding

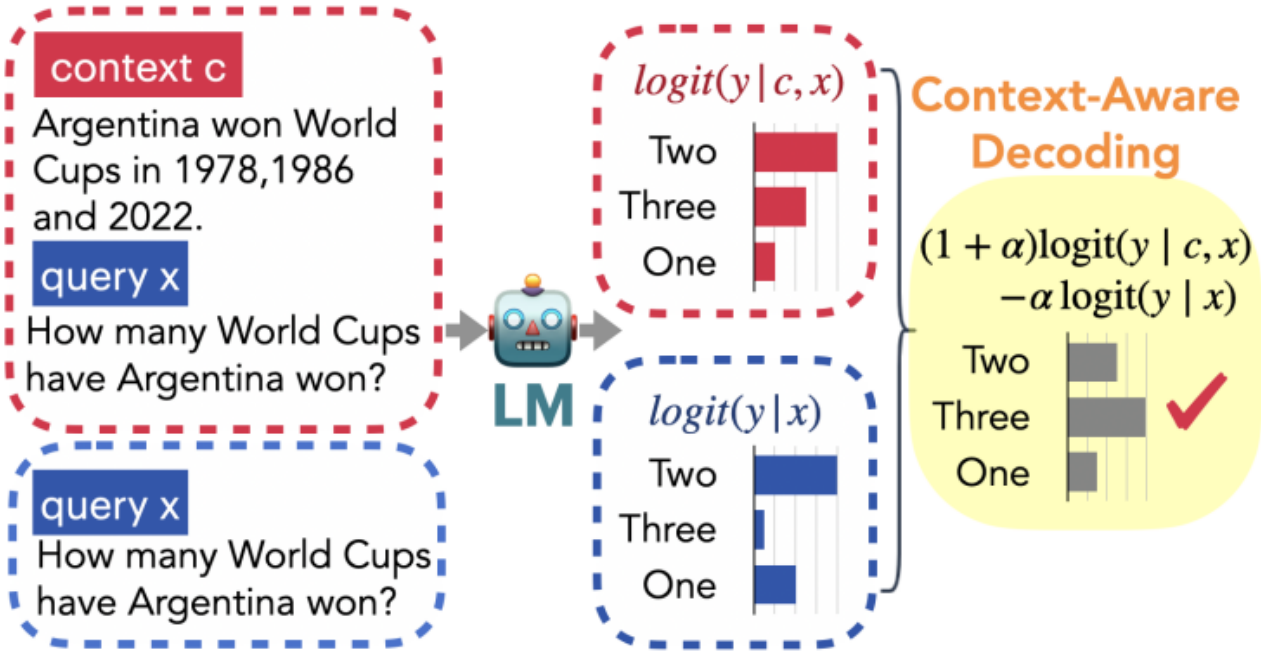

Рис. 1: Идея работы CAD [4]

$$\hat{p}_\theta(y_t|c, x, y_{<t}) \propto softmax((1+\alpha)logit_\theta(y_t|c, x, y_{<t}) - \alpha\ logit_\theta(y_t|c, x, y_{<t}))$$

[4] Shi, Weijia, Trusting Your Evidence: Hallucinate Less with Context-aware Decoding

DoLa: Decoding by Contrasting Layers

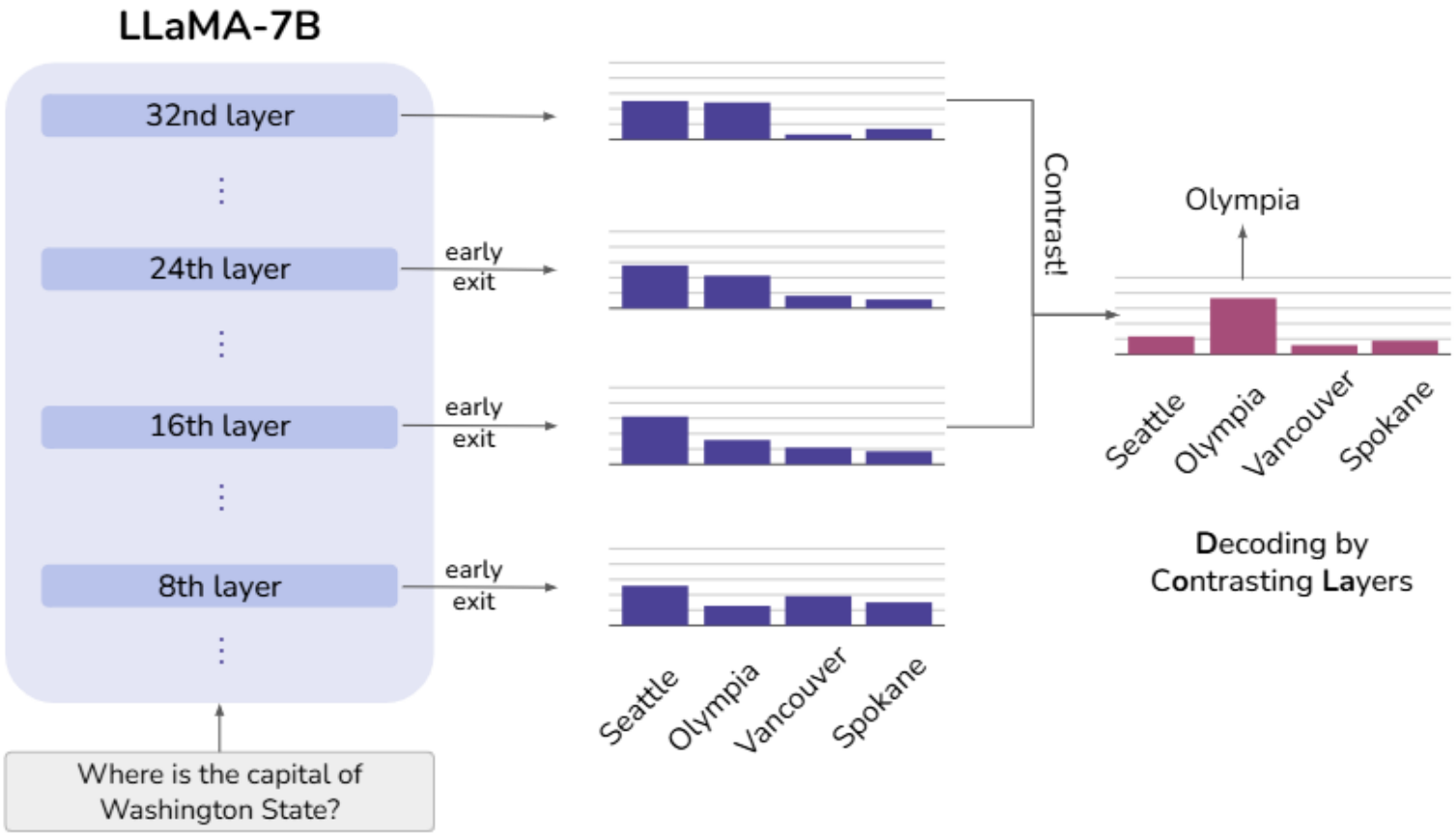

Рис. 2: Принцип работа Dola [5]

[5] Yung-Sung Chuang, DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models

Lookback Lens

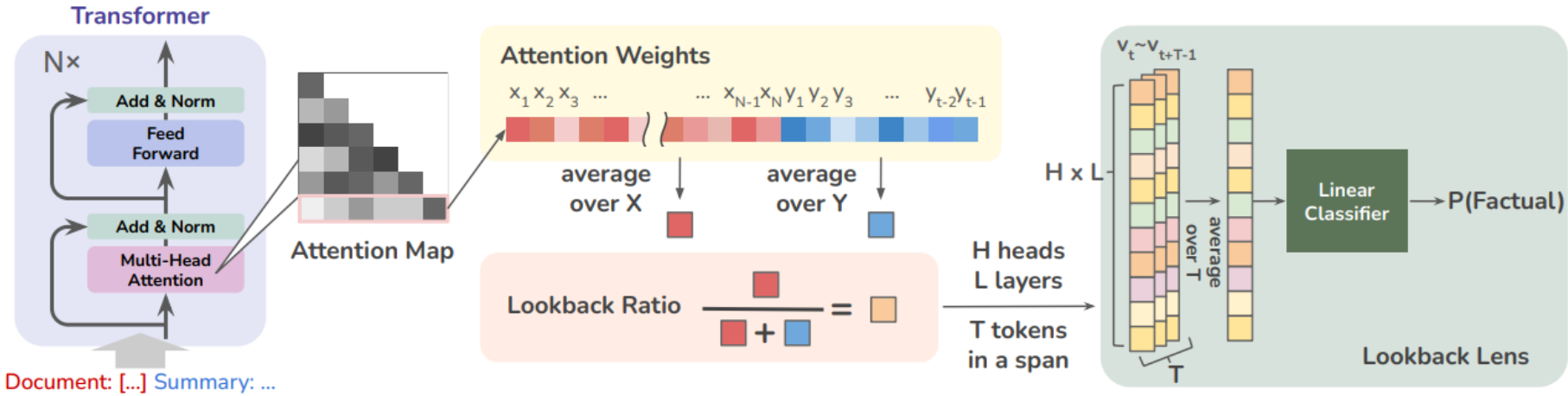

Рис. 3: Обнаружение галлюцинаций в Lookback Lens [6]

$$v_t = [LR_t^{1,1}, LR_t^{1,2}, ..., LR_t^{L,H}];\ \hat{v} = [v_t, v_{t+1}, ..., v_{t+T-1}]$$

$$C^* = \underset{C^j \in \{C_1,...,C_k\}}{argmax} \mathcal{F}(\hat{v}^j)$$

[6] Chuang, Yung-Sung, Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

Self-Refine

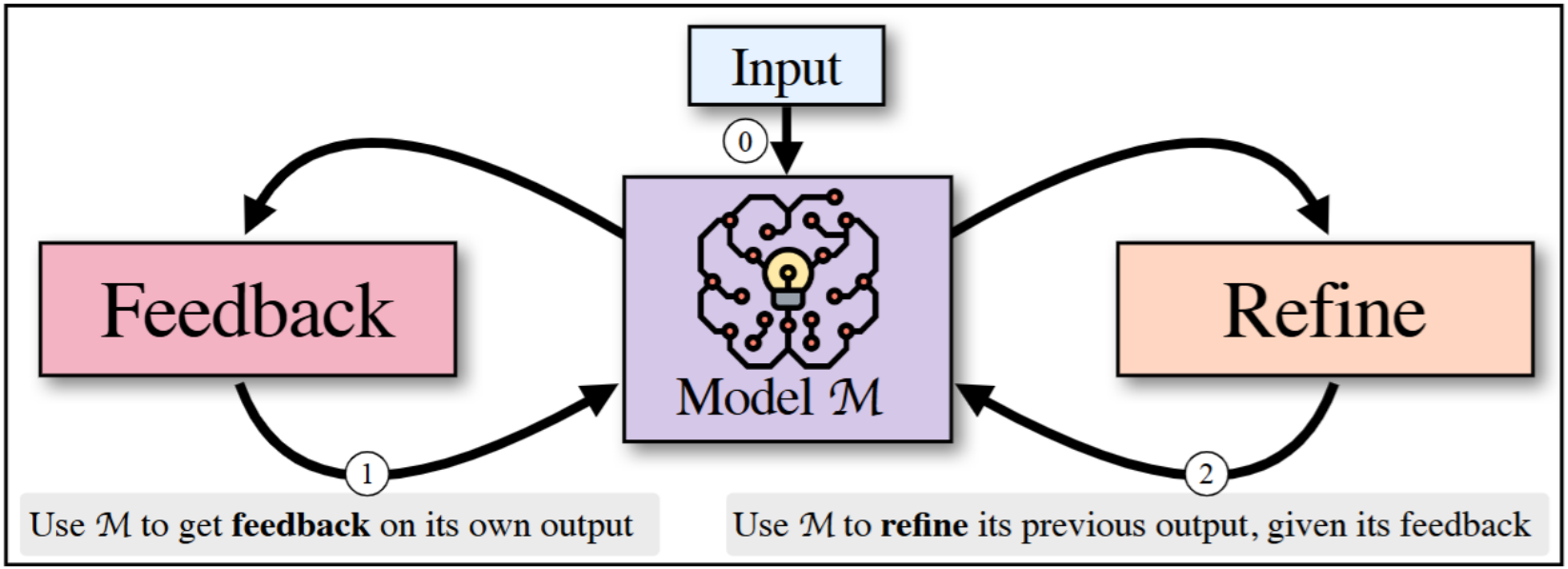

Рис. 4: Идея метода Self-Refine[7]

$$y_0 = \mathcal{M}(p_{gen}, x)$$
$$fb_k = \mathcal{M}(p_{fb}, x, y_0)$$
$$y_{k+1} = \mathcal{M}(p_{refine}, x, y_0, fb_0, ..., y_k, fb_k),$$

[7] Madaan, Aman and Tandon, Self-Refine: Iterative Refinement with Self-Feedback

Chain of NLI

Идея метода[8] состоит в том, чтобы поделить сгенерированный текст на части и рассматривать следование контексту отдельно

- Определение(детекция) галлюцинаций
  - сгенерированный текст Y_{raw} разбивается на предложения и добавляется во множество гипотез
  - Определение на уровне предложений: с помощью LLM
  - Определение на уровне сущностей для не галлюцинаций аналогично
  - В итоге формируется множество с аргументацией по каждому предложению
- Устранение галлюцинаций происходит с помощью LLM, которая доверяет детектору

[8] Deren Lei Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations

Классификация



Рис. 5: Классификация методов

Сравнение в рамках RiRag [9]

Таблица 1: Следование контексту на датасете ObliQA

Модель	Тип	RePASs
Llama-3.1-8B	–	48.64
	BS	<u>70.09</u>
	CFG	59.22
	CAD	64.32
Qwen2-72B	–	61.1

[9] Tuba Gokhan, RIRAG: Regulatory Information Retrieval and Answer Generation, RegNLP 2025 workshop

Выводы

- Существуют методы, усиляющие следование контексту без дообучения модели
- Методы стратегии декодирования:
  - работают в среднем быстрее
  - могут сильнее изменить модель
  - но могут привести к генерации грамматических ошибок и эмпирически ненатурального текста
- Методы улучшения ответа:
  - не требуют доступа к внутренней модели и могут применяться с использованием LLM на сторонних сервисах
  - ответ получается более естественным

Спасибо за внимание!

kratkov.ev@phystech.edu