GenAI HW7
Understanding what AI is thinking

TA: 方泓傑、李哲言、白宗民

ntu-gen-ai-2024-spring-ta@googlegroups.com

Deadline: 2024/5/16 23:59:59 (UTC+8)

Outline

- Introduction
- Task 1: Token Importance Analysis
- Task 2: LLM Explanation
- Submission & Deadline
- Contact

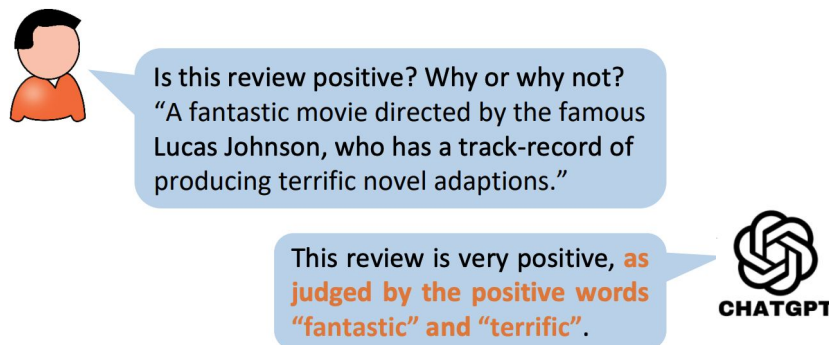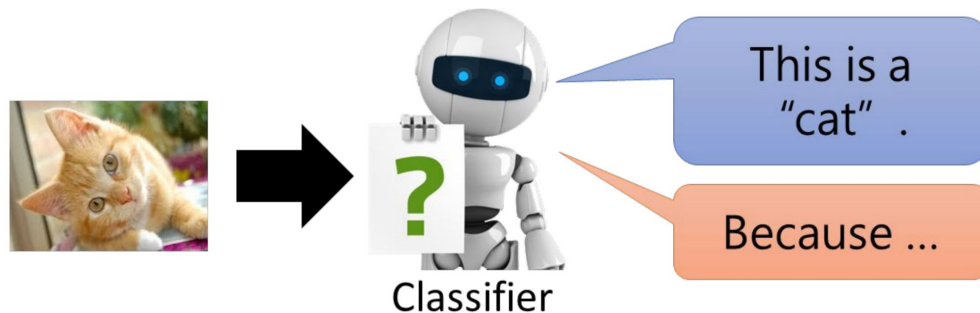# Link

[Colab](#)

[COOL Quiz](#)

[Questions](#)

Introduction

Why should we know what generative AI is thinking?

- Correct answer ≠ Intelligent

- Explanation is essential in high-stakes applications, e.g., medicine and law.

- We can improve our model based on our explanation.

Model Explanation

Image source: 2310.11207.pdf (arxiv.org), Interpreting Neural Network (ntu.edu.tw)
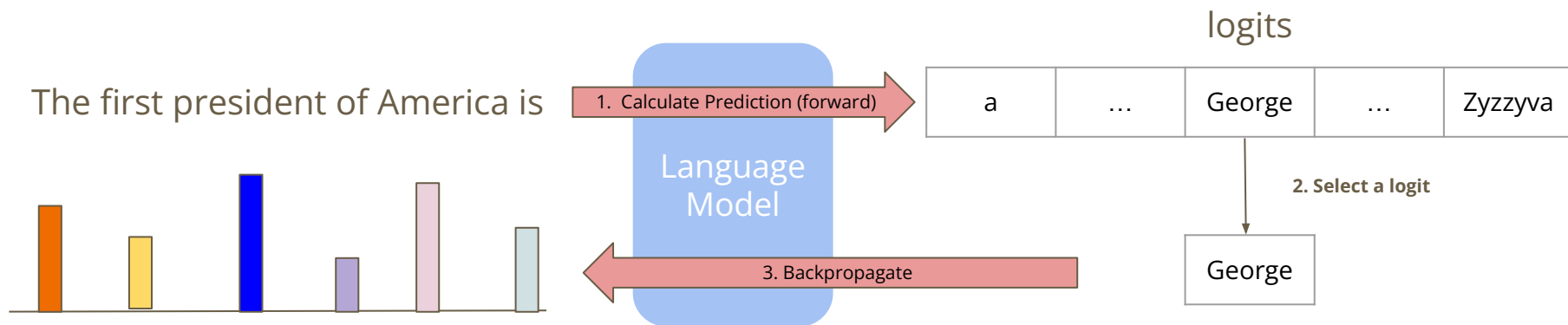
Task 1: Token Importance Analysis

Task Description

- In this task, we aim to understand what tokens play important roles in generating the response.

- We utilize **feature attribution methods** to analyze the importance.
  - Gradient-based approach
 - Attention-mechanism

- Run the sample code and finish question 1 to 7.

Gradient-based Approach (saliency)

- Compute the gradient of the target logit with respect to the input tokens.

The first president of America is

Language Model

1. Calculate Prediction (forward)

logits

| a | … | George | … | Zyzzyva |

2. Select a logit

George

3. Backpropagate

Attention-mechanism

- Commonly used in transformer-based models.

- Shows which tokens the model attends to when generating the output.
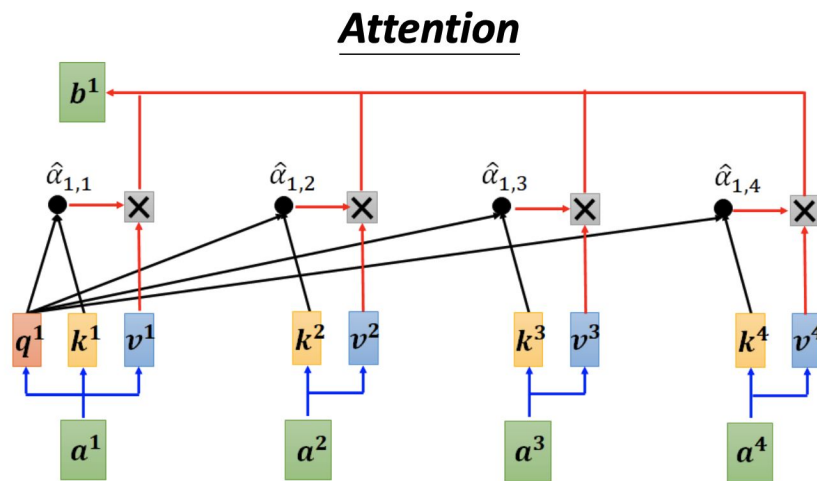
Attention

Token Visualization

- In this task, we use https://github.com/inseq-team/inseq/ to visualize the importance of token when generating the response.

- It supports many feature attribution methods, including gradient and attention, which we will use in this homework.

Inseq

Saliency map of machine translation task

Source Saliency Heatmap
x: Generated tokens, y: Attributed tokens

Output

	__美国	第一	任	总统	是	乔治	·	瓦	辛	顿	</s>
__The	0.052	0.089	0.204	0.193	0.056	0.035	0.113	0.01	0.02	0.066	0.109
__first	0.108	0.274	0.178	0.103	0.064	0.033	0.064	0.008	0.016	0.032	0.093
__president	0.191	0.314	0.202	0.285	0.093	0.057	0.072	0.014	0.023	0.036	0.129
__of	0.105	0.082	0.103	0.089	0.057	0.026	0.036	0.02	0.019	0.023	0.057
__America	0.308	0.073	0.061	0.089	0.09	0.066	0.047	0.022	0.027	0.031	0.107
__is	0.079	0.057	0.04	0.059	0.341	0.071	0.051	0.023	0.028	0.039	0.084
__George	0.052	0.035	0.044	0.048	0.142	0.364	0.139	0.079	0.079	0.061	0.114
__Wash	0.046	0.03	0.041	0.035	0.08	0.206	0.282	0.544	0.473	0.135	0.102
in	0.017	0.012	0.033	0.02	0.025	0.065	0.08	0.189	0.139	0.1	0.045
ton	0.015	0.017	0.044	0.032	0.02	0.049	0.066	0.072	0.142	0.401	0.075
.	0.026	0.018	0.049	0.046	0.031	0.028	0.049	0.018	0.035	0.076	0.084
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
probability	0.696	0.495	0.925	0.889	0.683	0.802	0.603	0.485	0.397	0.897	0.583

Input

The first president …

Translation Model

美國第一任總統 …

Saliency map of sentence completion task

Target Saliency Heatmap
x: Generated tokens, y: Attributed tokens

Generated tokens

Input

	__George	__Washington	.	__Unterscheidung	__between
__The	0.229	0.176	0.255	0.194	0.06
__first	0.107	0.093	0.079	0.059	0.031
__president	0.174	0.15	0.114	0.09	0.026
__of	0.141	0.142	0.116	0.085	0.083
__America	0.219	0.181	0.104	0.072	0.024
__is	0.129	0.12	0.145	0.1	0.069
__George		0.139	0.103	0.097	0.034
__Washington			0.084	0.103	0.031
.				0.2	0.353
__Unterscheidung					0.29
__between					
probability	0.908	0.992	0.66	0.023	0.535

Importance score

The	First	President	of	America	is	George

Autoregressive Model

First	President	of	America	is	George	Washington

Example of saliency map

Q: When generating the word "Washington", what's the importance score of "America"?

A: 0.181

Target Saliency Heatmap
x: Generated tokens, y: Attributed tokens

	__George	__Washington	.	__Unterscheidung	__between
__The	0.229	0.176	0.255	0.194	0.06
__first	0.107	0.093	0.079	0.059	0.031
__president	0.174	0.15	0.114	0.09	0.026
__of	0.141	0.142	0.116	0.085	0.083
__America	0.219	0.181	0.104	0.072	0.024
__is	0.129	0.12	0.145	0.1	0.069
__George		0.139	0.103	0.097	0.034
__Washington			0.084	0.103	0.031
.				0.2	0.353
__Unterscheidung					0.29
__between					
probability	0.908	0.992	0.66	0.023	0.535

Task 2: LLM Explanation

LLM Explanation

- LLMs have the ability to explain in **natural language.**

- It is much more straightforward to understand than prior methods.

Is this review positive? Why or why not? "A fantastic movie directed by the famous Lucas Johnson, who has a track-record of producing terrific novel adaptions."

This review is very positive, **as judged by the positive words "fantastic" and "terrific"**.

CHATGPT

Task Description

- In this task, we aim to assess the effectiveness of LLM explanation.

- We will explore two LLM explanation approaches.

  - Providing the explanation for the model's answer.

  - Simulating the feature attribution method in task 1 to see the importance of tokens.

- Run the given prompts on ChatGPT and finish Question 8 to 10.

  (No need of Colab)

Explain the model's answer

- Directly ask the LLM to explain its answer.

Prompt:

You are a creative and intelligent movie review analyst, whose purpose is to aid in sentiment analysis of movie reviews. Determine whether the review below is positive or negative, and explain your answers.

Review: This film is a compelling drama that captivates audiences with its intricate storytelling and powerful performances.

Simulate feature attribution methods with LLM explanation

- Ask the LLM to explain the importance of the input tokens in contributing to the answer, similar to what we do in task 1.

2310.11207.pdf (arxiv.org)

Simulate feature attribution methods with LLM explanation

Prompt:

You are a movie review analyst tasked with sentiment analysis. For each review, provide a list of tuples representing the importance of each word and punctuation, with values ranging from -1 (negative) to 1 (positive). Then, classify the review as positive (1) or negative (-1). The review is within <review> tags.

Example output:

[(<word or punctuation>, <float importance>), …]

<int classification>

<review> This film is a compelling drama that captivates audiences with its intricate storytelling and powerful performances. <review>

Note: ChatGPT's responses may vary due to randomness. If the format isn't as desired, please try again.

Submission & Deadline

Submission

- Finish questions on NTU COOL Quiz

- Unlimited times of submissions for the quiz, but only the **latest submission** will be considered when grading

- No late submission is allowed

Important dates

- Deadline for Submission (NTU Cool)

  2024/05/16 23:59:59 (UTC+8)

- Grading Release Date

  2024/05/31 23:59:59 (UTC+8)

Contact

If You Have Any Questions

- NTU Cool HW7 作業討論區
  - 如果同學的問題不涉及作業答案或隱私, 請**一律使用**NTU Cool 討論區
  - 助教們會優先回答NTU Cool討論區上的問題
- Email: ntu-gen-ai-2024-spring-ta@googlegroups.com
  - Title should start with [GenAI 2024 Spring HW7]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Time:
    - 5/3, 16:30 ~ 17:20
    - 5/10 13:20~14:10, 16:30 ~ 17:20
  - Location: 綜合大講堂