Улучшение FineTuning LLM с помощью Multi Token Prediction

Мостовых Егор

1 апреля 2025

План презентации

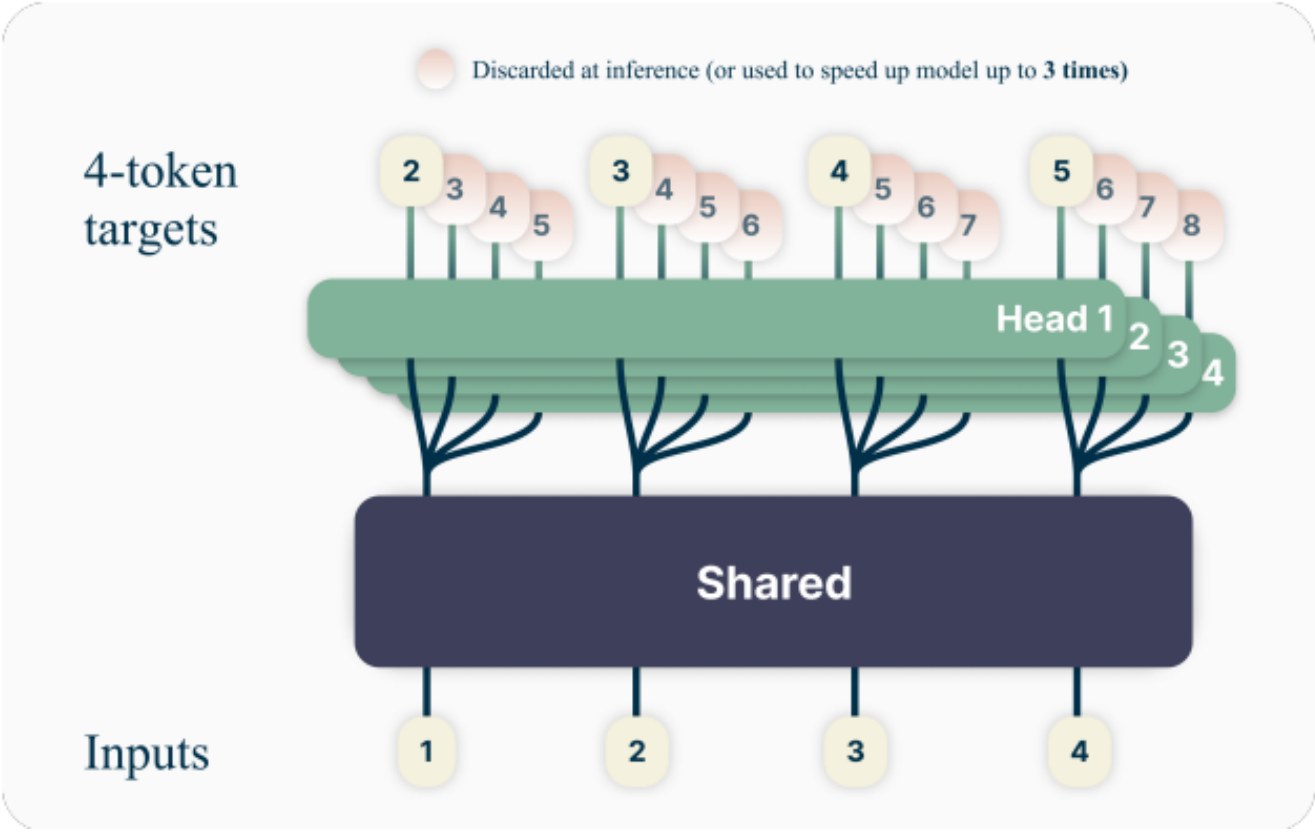
- Про то, как эта темы вытекает из исследований в этой области
- Суть идеи диплома
- Что сделано
- Как планирую развить тему дальше

Обзор литературы:

Спекулятивный декодинг (2022)

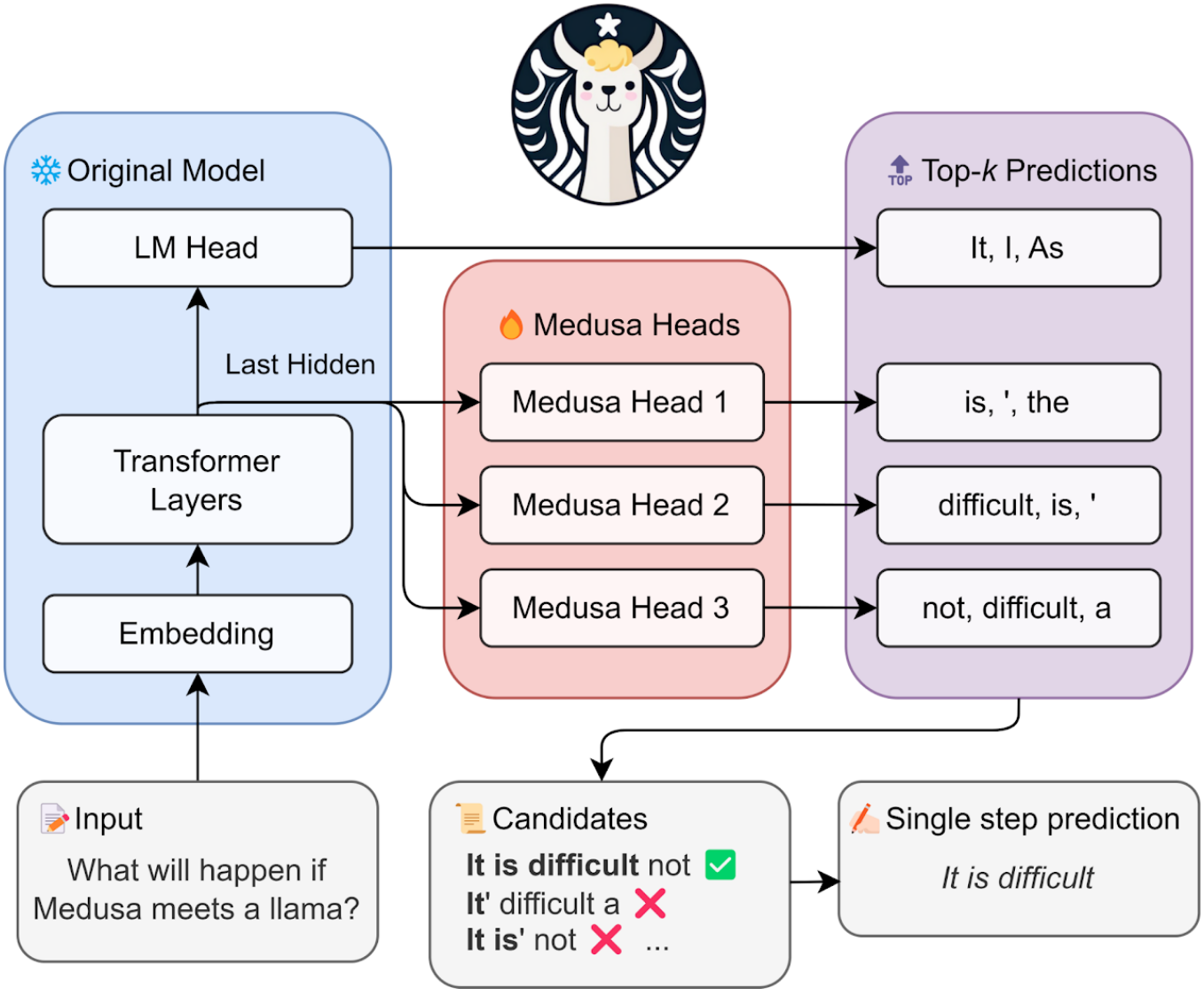
Обзор литературы:

Multi Token Prediction (2024)



Обзор литературы:

MEDUSA(2024) & EAGLE(2024)



Идея диплома:

Посмотреть как повлияет на качество/скорость FineTuning-а использование MTP.

Идея диплома:

Как именно это будет работать и что уже сделано

- Поступает предобученная модель
- Добавляем к ней несколько спекулятивных голов, обученных для архитектуры модели заранее нами
- Гипотеза: если мы будем дообучать такую модель на новую задачу - это получится быстрее и качественнее

Что планирую делать:

- Дообучить разное кол-во голов
- Попробовать разные методы FineTuning-а
- Провести несколько экспериментов с разными LLM на нескольких датасетах.

Last slide.