Speculative decoding

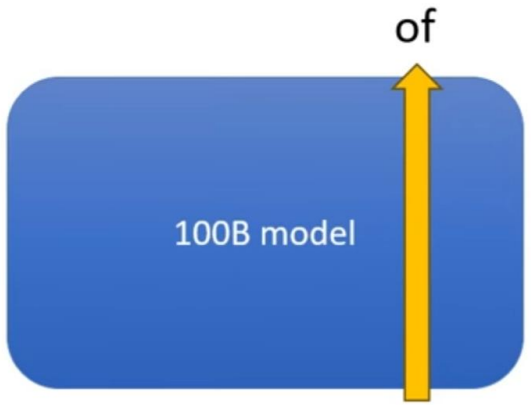
Кошелев Михаил

План

- Обзор проблемы
- Теоретическая основа спекулятивного декодирования
- Анализ и экспериментальные результаты
- Обзор аппроксимационных моделей
- Применимость к различным задачам и моделям
- Литература

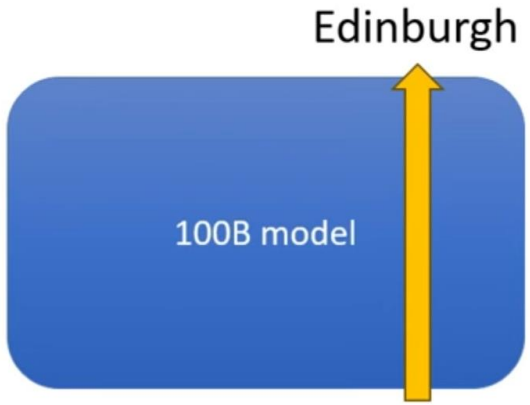
Проблема

GPT-3



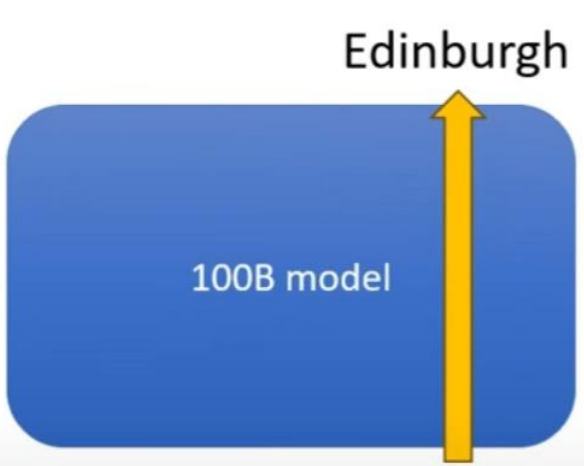
Geoffrey Hinton did his PhD at the University…

Very Easy

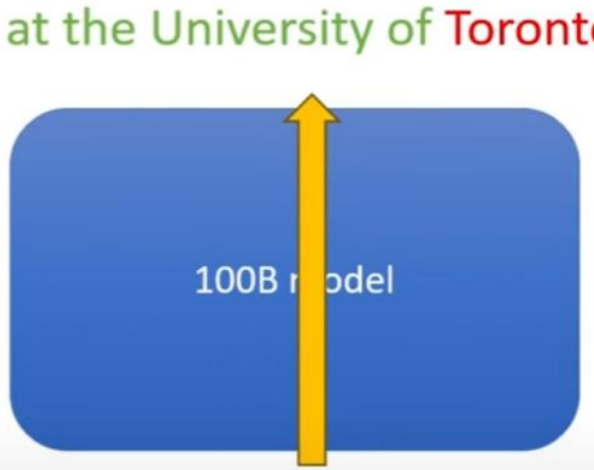


Geoffrey Hinton did his PhD at the University of…

Difficult



Geoffrey Hinton did his PhD
at the University of...



Geoffrey Hinton did his PhD
at the University of Toronto

Алгоритм

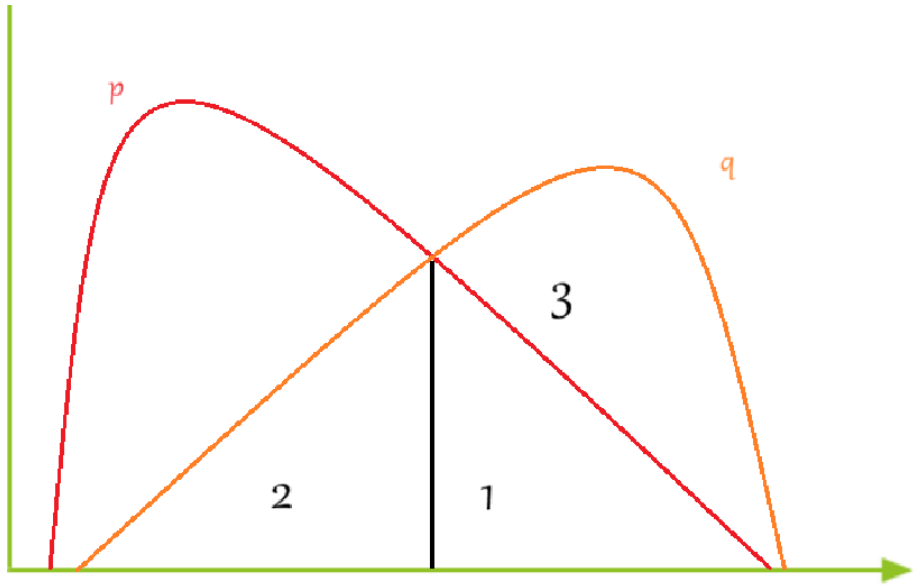
- M_p = черновая модель
- M_q = основная модель
- pf = префикс, γ = кол − во токенов

- $p_1(x) = M_p(pf) \rightarrow x_1$
- $p_2(x) = M_p(pf, x_1) \rightarrow x_2$
- …
- $p_\gamma(x) = M_p(pf, x_1, \ldots, x_{\gamma-1}) \rightarrow x_\gamma$

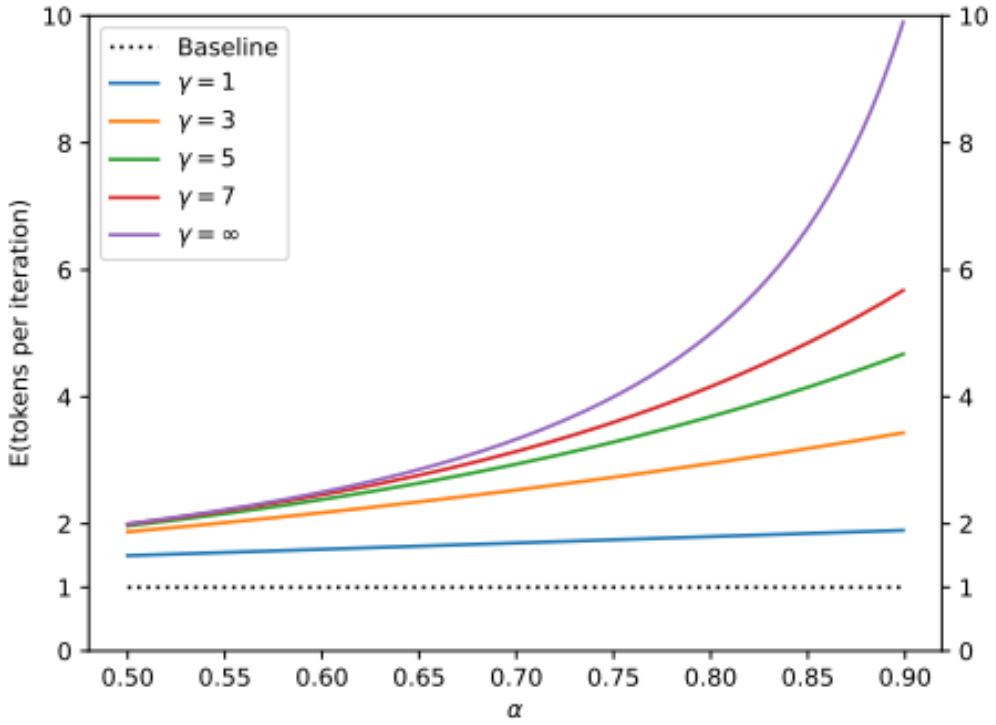
- $p_1(x) = M_p(pf)$ x_1
- $p_2(x) = M_p(pf, x_1)$ x_2
- …
- $p_\gamma(x) = M_p(pf, x_1, \dots, x_{\gamma-1})$ x_γ

- $q_1(x),\ q_2(x) \dots q_\gamma(x), q_{\gamma+1}(x) = M_q(pf, x_1, \dots, x_\gamma)$

- Цель: взять токен из q(x)
- Случай 1: q(x) ≥ p(x), принимаем токен
- Случай 2: q(x) < p(x), принимаем токен с вероятностью $\frac{q(x)}{p(x)}$
- Оставшийся случай: q(x) < p(x) и мы отвергли токен, тогда берем из распределения $(q(x) - p(x))_+$



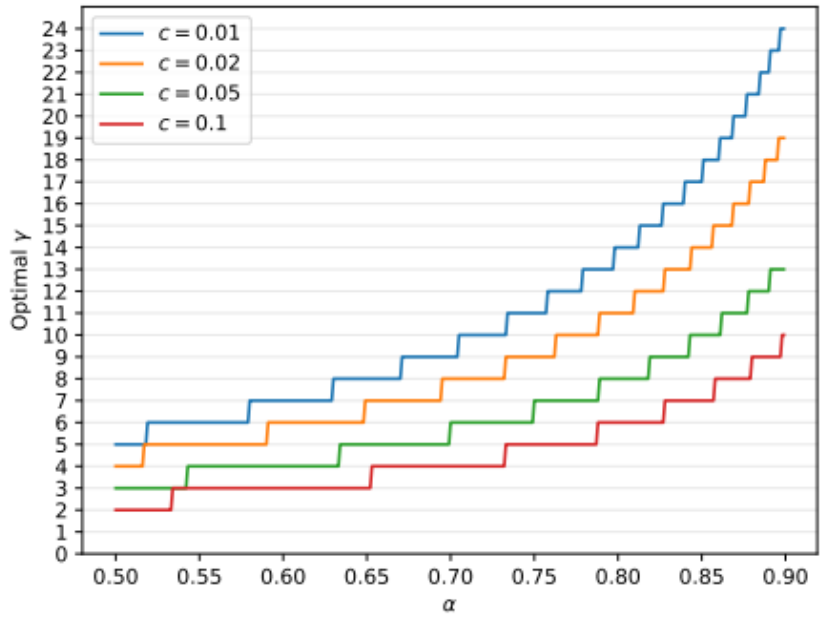
E(# generated tokens) = $\frac{1-\alpha^{\gamma+1}}{1-\alpha}$

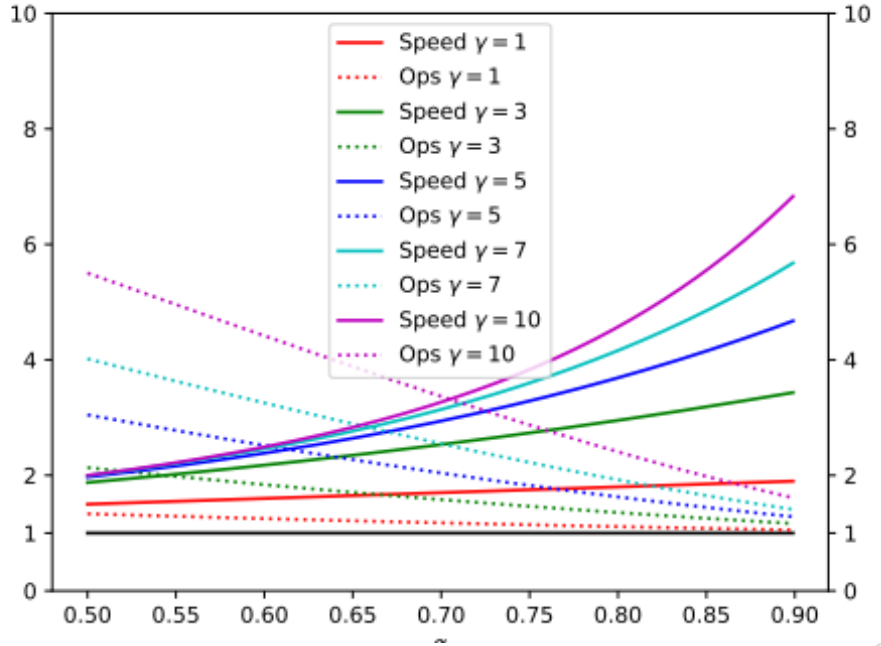


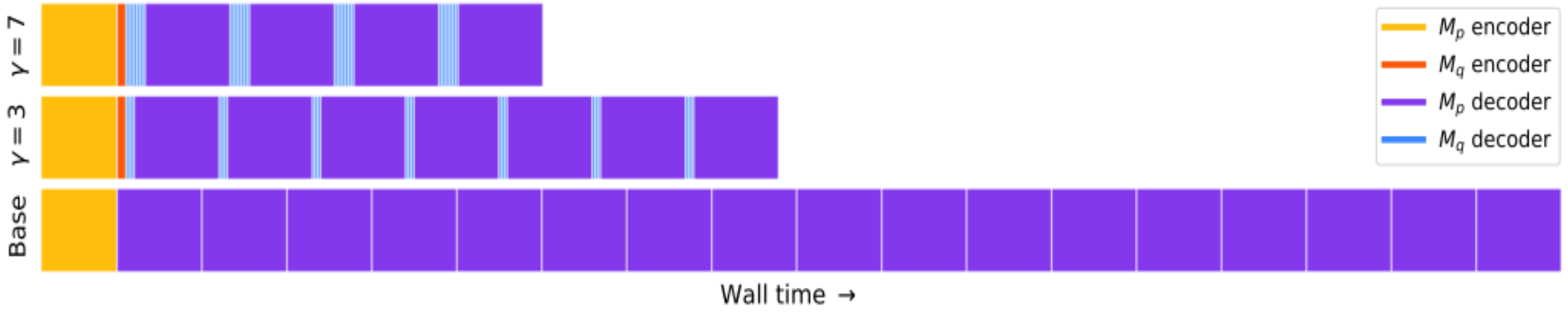
$\hat{c}$ - отношение арифметических операций
на токен аппроксимационной модели M_q к
арифметическим операциям целевой модели M_p.

Теорема 3.11. Ожидаемый коэффициент увеличения общего числа операций алгоритма равен $\frac{(1-\alpha)(\gamma\hat{c}+\gamma+1)}{1-\alpha^{\gamma+1}}$

Уравнение улучшения времени: $\frac{1-\alpha^{\gamma+1}}{(1-\alpha)(\gamma c+1)}$






γ = 7
γ = 3
Base
Wall time →
M_p encoder
M_q encoder
M_p decoder
M_q decoder

Обзор аппроксимационных моделей

- Биграммная модель $\longrightarrow$ T5-XXL
- M_q - могут быть простыми эвристиками
- Можно использовать нерекурсивные модели

TASK	M_q	TEMP	γ	α	SPEED
ENDE	T5-SMALL ★	0	7	0.75	**3.4X**
ENDE	T5-BASE	0	7	0.8	2.8X
ENDE	T5-LARGE	0	7	0.82	1.7X
ENDE	T5-SMALL ★	1	7	0.62	**2.6X**
ENDE	T5-BASE	1	5	0.68	2.4X
ENDE	T5-LARGE	1	3	0.71	1.4X
CNNDM	T5-SMALL ★	0	5	0.65	**3.1X**
CNNDM	T5-BASE	0	5	0.73	3.0X
CNNDM	T5-LARGE	0	3	0.74	2.2X
CNNDM	T5-SMALL ★	1	5	0.53	**2.3X**
CNNDM	T5-BASE	1	3	0.55	2.2X
CNNDM	T5-LARGE	1	3	0.56	1.7X

Литература

- Deepmind Paper: https://arxiv.org/abs/2302.01318
- Google Paper: https://arxiv.org/abs/2211.17192