Исследование масштабирования размера оптимального батча в больших языковых моделях

Дипломная работа

Бессонов А.С.
МГУ им. М.В. Ломоносова,
кафедра математических методов
прогнозирования, 417 группа
beccohov.a@yandex.ru

Дьяконов А.Г.
д-р физ.-мат. наук
Центральный университет Тинькофф *
djakonov@mail.ru

Аннотация

В данной статье исследуется связь между оптимальным размером батча при обучении языковых моделей. Нами были проведены эксперименты на широком спектре моделей, варьирующихся от 180M до 4B параметров. Каждая модель обучалась с использованием различных значений размера батча. Количество токенов, применяемых для предварительного обучения модели, зависит только от ее размера и задается величиной $P \times 50$, где P - количество параметров модели. Результаты каждого измерения дали теоретическую оценку оптимального батча для каждой модели, подтверждающую универсальность использованной аппроксимации. Кроме того, была установлена экспериментальная зависимость между размером оптимального теоретического батча и размером модели, а также разработана математическая модель, позволяющая практически оценивать размер батча для больших моделей. Полученные значения приближенно соотносятся с текущими практическими значениями, которые широко принимаются в последнее время.

Keywords Масштабирование моделей · Языковые модели

1 Введение

Большие языковые модели в последнее десятилетие занимают доминирующую роль как в сфере исследований, так и практическом применении. При этом обучение таких моделей требует значительных вычислительных ресурсов, что усложняет эксперименты с подбором гиперпараметров при приобучении. В недавних работах было уделено внимание тому, как масштабируются гиперпараметры подобных моделей при увеличеснии количества параметров. В данной работе исследуется, как должен изменяться размер обучающего батча модели в зависимости от количества параметров в модели.

Недавно стало ясно, что подбор гиперпараметров имеет значительное влияние на качество финальной модели. Например, Hoffmann et al. [2022] указывает на то, что количество обучающих данных должно расти квазилинейно с увеличением размера модели. В OpenAI [2023] также отмечается, что одним из методов для предсказания оптимальных гиперпараметров является построение моделей, способных предсказывать эти гиперпараметры.

Языковые модели начали быстро развиваться с появлением архитектуры трансформера (Vaswani et al. [2017]). В последствии стало ясно, что главную роль в качестве таких моделей играет их размер (ссылки). Новую эру в развитии языковых моделей ознаменовало появление архитекруры GPT-2 (Radford et al. [2019]), которая в то время стала SOTA среди языковых моделей. Последующее использование RLHF ознаменовало новую эпоху персональных ассистенов (Ouyang et al. [2022]).

*Научный руководитель

В настоящее время размеры языковых моделей превышают триллионы параметров, что создает множество инженерных трудностей при их обучении. Подбор гиперпараметров для таких моделей практически невозможен. Вместо этого необходимо использовать эвристики и предсказывать оптимальные значения гиперпараметров на основе меньших моделей.

В данной работе исследуется и предлагается эвристика для масштабирования размера обучающего батча. В экспериментах используются оптимальные модели с большим количеством параметров, а размеры моделей варьируются от 180M до 4B параметров. Проведенные эксперименты подтверждают характер масштабирования размера батча и дают оценки оптимальных размеров батчей для очень больших языковых моделей (70B, 140B и более).

2 Постановка задачи

Для более эффективной утилизации GPU требуются большие размеры батчей. С другой стороны, размер батча является гиперпараметром, значительно влияющем на качество итоговой обученной модели. В связи с этим возникает необходимость эвристик подбора и масштабирования размеров батча в зависимости от размера модели. Будет показано, что для получения оценки оптимальности батча недостаточно обучения на 10-15 % от общего количества итераций. Соответственно, только проводя полные эксперименты, можно найти оптимальный размер батча. С другой стороны, с увеличением размеров модели такие эксперименты на практике не применимы. Это обосновывает потребность в наличии указанных эвристик.

3 Вычислительные эксперименты

Как упоминалось ранее, в данной задаче используется модель трансформерного декодера. В качестве обучающих данных использовался набор из пяти датасетов - книги, веб-страницы, википедия, учебники, датасет гитхаба. Все датасеты были разделены на обучающую и валидационную выборки. Поскольку размер датасетов очень большой (несколько Тб), в качестве валидационной выборки бралось порядка 50 миллионов токенов с конца каждого датасета. При этом для обучения каждой из модели использоватись данные из начала каждого датасета в равной пропорции (семплирование данных из каждого датасета равномерно). В качестве критерия качества использовалось значение перплексии на валидационной части данных.

Для каждой модели размера N параметров в качестве обучающей выборки бралось $30 \times N$ токенов, итого поряка $6 \times N$ токенов из кажого датасета. Таким образом, число токенов на каждый параметр сохранялось постоянным. Для каждой модели были взяты от трех до пяти различных значений размера батча. Далее, мерой качества модели полагалось среднее значение перплексии на пяти датасетах на валидации. Таким образом, получаем от трех до пяти пар (Q, B), где Q - качество модели, B - размер батча. По данному множеству точек строится приближение параболой. Для этой параболы вычисляется минимальное значение перплексии и размер батча, на котором она достигается. Полученное значение называется "оптимальным размером батча". Далее по полученному набору (OB, M), где OB - оптимальный размер батча, M - размер модели, строится зависимость оптимального размера батча от размера модели. Полученная зависимость позволяет оценить оптимальный размер батча для больших моделей (больше 4B параметров).

4 Экспериментальные данные

На Рис. 2 показаны примеры аппроксимаций для мододей. Всего было обучено 7 моделей с разными размерами батчей. Аппроксимация строилась по двум точкам, а третья попадала на построенную кривую, что подтверждает приемлемость выбора порядка апроксимации. Также это наименьший порядок аппрокимации, который согласуется с требованиями к аналитический зависимости - наличие минимума.

Таким образом, полученные аппроксимации помогают получить теоретическую оценку оптимального размера батча. Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula

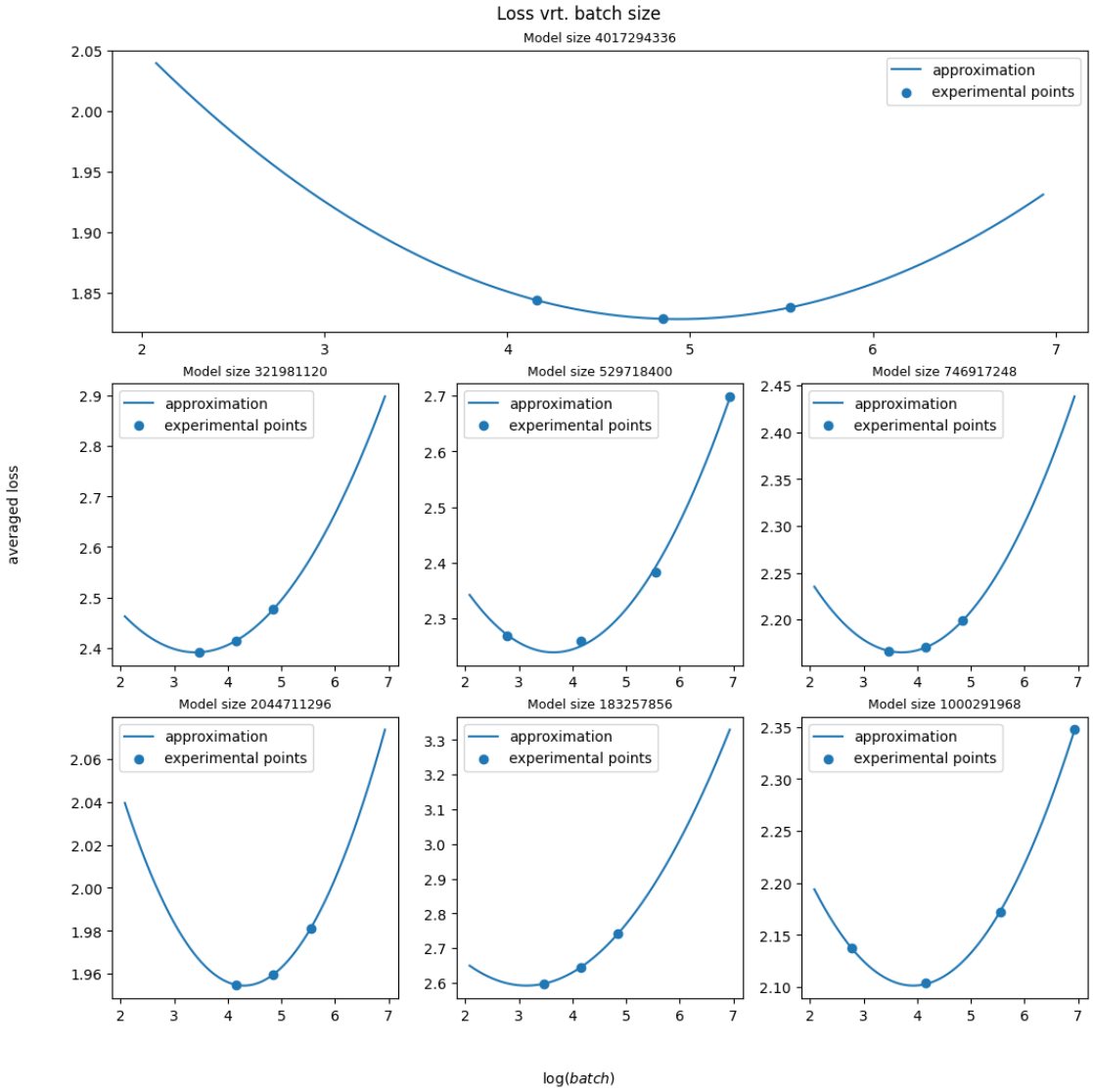

Рис. 1: Sample figure caption.

sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

4.0.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque

Рис. 2: Sample figure caption.

non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

5 Examples of citations, figures, tables, references

5.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [??] but other people thought something else [?]. Many people have speculated that if we knew exactly why ? thought this...

5.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 2. Here is how you add footnotes. [2] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

5.3 Tables

See awesome Table 1.

[2] Sample of the first footnote.

Таблица 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

5.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

Список литературы

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, 2022.

OpenAI. Gpt-4 technical report. ArXiv, abs/2303.08774, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.