

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
 Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
 Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
 Edouard Grave*, Guillaume Lample*

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community¹.

1 Introduction

Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a few examples (Brown et al., 2020). These few-shot properties first appeared when scaling models to a sufficient size (Kaplan et al., 2020), resulting in a line of work that focuses on further scaling these models (Chowdhery et al., 2022; Rae et al., 2021). These efforts are based on the assumption that more parameters will lead to better performance. However, recent work from Hoffmann et al. (2022) shows that, for a given compute budget, the best performances are not achieved by the largest models, but by smaller models trained on more data.

The objective of the scaling laws from Hoffmann et al. (2022) is to determine how to best scale the dataset and model sizes for a particular *training* compute budget. However, this objective disregards the *inference* budget, which becomes critical when serving a language model at scale. In this context, given a target level of performance, the preferred model is not the fastest to train but the fastest at inference, and although it may be cheaper to train a large model to reach a certain level of

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance, LLaMA-13B outperforms GPT-3 on most benchmarks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU. At the higher-end of the scale, our 65B-parameter model is also competitive with the best large language models such as Chinchilla or PaLM-540B.

Unlike Chinchilla, PaLM, or GPT-3, we only use publicly available data, making our work compatible with open-sourcing, while most existing models rely on data which is either not publicly available or undocumented (e.g. “Books – 2TB” or “Social media conversations”). There exist some exceptions, notably OPT (Zhang et al., 2022), GPT-NeoX (Black et al., 2022), BLOOM (Scao et al., 2022) and GLM (Zeng et al., 2022), but none that are competitive with PaLM-62B or Chinchilla.

In the rest of this paper, we present an overview of the modifications we made to the transformer architecture (Vaswani et al., 2017), as well as our training method. We then report the performance of our models and compare with others LLMs on a set of standard benchmarks. Finally, we expose some of the biases and toxicity encoded in our models, using some of the most recent benchmarks from the responsible AI community.

* Equal contribution. Correspondence: {htouvron, thibautlav, gizacard, egrave, glample}@meta.com

¹<https://github.com/facebookresearch/llama>

LLaMA: Open and Efficient Foundation Language Models

ユーゴ・トウヴロン , ティボー・ラブリル , ゴーティエ・イザカール , ザビエル・マルティネ・マリー=アンヌ・ラショー, テイモシー・ラクロワ、パティスト・ロジエール、ナマン・ゴヤル エリック・ハンブロ、ファイサル・アズハル、オーレリアン・ロドリゲス、アルマン・ジュラン・エドゥアール・グレイヴ , ギヨーム・ランプル

Meta AI

Abstract

7B から 65B までのパラメータにわたる基礎言語モデルのコレクションである LLaMA を紹介します。私たちは何兆ものトークンでモデルをトレーニングし、独自のアクセスできないデータセットに頼ることなく、公開されているデータセットのみを使用して最先端のモデルをトレーニングできることを示しました。特に、LLaMA-13B はほとんどのベンチマークで GPT-3 (175B) を上回り、LLaMA-65B は最高のモデルである Chinchilla-70B および PaLM-540B と競合します。私たちはすべてのモデルを研究コミュニティに公開しています¹。

1 Introduction

膨大なテキストのコーパスで訓練された大規模言語モデル (LLM) は、テキストの指示またはいくつかの例から新しいタスクを実行する能力を示しています (Brown et al., 2020)。これらの少数ショットの特性は、モデルを十分なサイズにスケーリングするときに初めて現れ (Kaplan et al., 2020)、その結果、これらのモデルをさらにスケーリングすることに焦点を当てた一連の研究が行われました (Chowdhery et al., 2022; Rae et al., 2021)。これらの取り組みは、パラメータが多いほどパフォーマンスが向上するという前提に基づいています。しかし、ホフマンらの最近の研究では、(2022) は、特定のコンピューティング バジェットにおいて、最高のパフォーマンスは最大のモデルによって達成されるのではなく、より多くのデータでトレーニングされた小規模なモデルによって達成されることを示しています。ホフマンらのスケーリング則の目的は次のとおりです。(2022) は、特定のトレーニング コンピューティング 予算に合わせてデータセットとモデルのサイズを最適にスケーリングする方法を決定します。ただし、この目標では推論バジェットを無視しています。推論バジェットは、言語モデルを大規模に提供する場合に重要になります。これに関連して、目標レベルのパフォーマンスが与えられた場合、推論されるモデルはトレーニングが最も速いモデルではなく、推論が最も速いモデルです。特定のパフォーマンス レベルに達するまでに大規模なモデルをトレーニングする方がコストが安くなる可

能性がありますが、小規模なモデルをより長くトレーニングした方が、最終的には推論のコストが安くなります。たとえば、Hoffmann et al. (2022) は 200B トークンで 10B モデルをトレーニングすることを推奨していますが、7B モデルのパフォーマンスは 1T トークン後でも向上し続けることがわかりました。

この研究の焦点は、通常使用されるものよりも多くのトークンでトレーニングすることにより、さまざまな推論予算で可能な限り最高のパフォーマンスを達成する一連の言語モデルをトレーニングすることです。結果として得られる LLaMA と呼ばれるモデルは、既存の最高の LLM と比較して競争力のあるパフォーマンスを備えた 7B ~ 65B パラメーターの範囲にあります。たとえば、LLaMA-13B は、10 分の 1 小さいにもかかわらず、ほとんどのベンチマークで GPT-3 を上回ります。このモデルは単一の GPU 上で実行できるため、LLM へのアクセスと研究の民主化に役立つと考えています。スケールのハイエンドでは、私たちの 65B パラメータ モデルは、Chinchilla や PaLM-540B などの最高の大規模言語モデルとも競合します。

Chinchilla、PaLM、または GPT-3 とは異なり、私たちは公開されているデータのみを使用するため、私たちの研究はオープンソースと互換性がありますが、既存のモデルのほとんどは、公開されていない、または文書化されていないデータに依存しています（例：「書籍 - 2TB」または「ソーシャルメディアでの会話」）。いくつかの例外、特に OPT (Zhangら, 2022)、GPT-NeoX (Blackら, 2022)、BL0OM (Scaoら, 2022) および GLM (Zengら, 2022) が存在するが、PaLM-62B やチンチラと競合するものはない。

このページの残りの部分では、変圧器アーキテクチャ (Vaswani et al., 2017) に加えた変更の概要と、トレーニング方法を示します。次に、モデルのパフォーマンスをレポートし、一連の標準ベンチマークで他の LLM と比較します。最後に、責任ある AI コミュニティからの最新のベンチマークのいくつかを使用して、モデルにエンコードされたバイアスと毒性の一部を明らかにします。

* Equal contribution. Correspondence: {htouvron, thibautlav, gizacard, egrave, glample}@meta.com

¹ <https://github.com/facebookresearch/llama>