

LLaMA: Open and Efficient Foundation Language Models

ユーゴ・トゥヴロン*、ティボー・ラブリル*、ゴーティエ・イザカール*、ザビエ・マルティネ
・マリー＝アンヌ・ラショー、ティモシー・ラクロワ、バティスト・ロジエール、ナマン・ゴヤ
ルエリック・ハンプロ、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・エドウアル・グラーヴ*、ギヨーム・ランブル*

メタAI

Abstract

7B から 65B 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作業が行われました (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あります。(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することです。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。この文脈では、目標レベルのパフォーマンスが与えられた場合、推奨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。また、一定のレベルに達するために大規模なモデルをトレーニングした方がコストが安くなる場合もありますが、

パフォーマンスに影響を与えるため、より長くトレーニングされた小規模なものほど、最終的には推論のコストが低くなります。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 トークンの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、通常使用されるトークン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 と比較して優れたパフォーマンスを備えた 7B ~ 65B パラメータの範囲にあります。たとえば、LLaMA-13B は、10 倍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、当社の 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競合することもできます。

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、私たちの作業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 (例: 「書籍 - 2TB」または「ソーシャル メディアでの会話」) に依存しています。いくつか存在します

例外は、特に OPT (Zhang et al., 2022)、GPT-NeoX (Black et al., 2022)、BLOOM (Scao et al., 2022) および GLM (Zeng et al., 2022) ですが、PaLM-62B や Chinchilla と競合するものはありません。

このペーパーの残りの部分では、変圧器アーキテクチャ (Vaswani et al., 2017) に加えた変更の概要と、トレーニング方法について説明します。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>