2024年4月26日
株式会社オルツ

ニュースリリース

オルツ、数兆パラメータ規模の大規模言語モデル構築に着手

～ユースケースから逆算した設計で世界最高峰のスピードとコストパフォーマンスを追求～

P.A.I.®（パーソナル人工知能）をはじめ、AIクローン技術でつくり出すパーソナルAIの開発および実用化を行う株式会社オルツ（本社：東京都港区、代表取締役：米倉 千貴、以下、オルツ）は、数兆パラメータ規模の大規模言語モデル（LLM）の構築に着手したことを発表いたします。約10年間、自然言語処理を含む本領域の研究開発を行ってきており、また早い段階からLLMの研究・開発・運用を手掛けてきた当社として、実際のビジネスシーンや実生活におけるユースケースから逆算したLLM設計・構築が肝要であると考えています。本開発は、単にパラメータ数を膨大にするだけではなく、生成AIのエンドユースケースから逆算した目標であり、パラメータ数に加えて、スピードと計算効率、コストパフォーマンスという実運用時に重要となる指標でも世界最高峰レベルを目指してまいります。また、当社は、本開発を通して、すでに展開している生成AIプロダクト群をより費用対効果の高いサービスとしてエンドユーザーに提供すること、ならびに、グローバルに先駆けた日本発の生成AIのエンドユースケースの確立を目指します。

＜足許のLLM競争に対する課題・対策意識＞

オルツは、大規模言語モデル（LLM）開発において、より複雑な表現力と高度なカスタマイズ性の両面を追求しています。デベロッパー及びエンドユーザー両方の観点から、GPTなどの既存モデルを超える使いやすさを実現するため、数兆パラメータ規模を持ち、日本語に優位なモデルの構築を構想してまいりました。大規模言語モデル（LLM）では、スケーリング則の元となる「パラメータ数」「データ量」「計算量」に加えて、スピードとコストのバランスを取ることが重要です。当社は、これらを適切に管理しながら、ユーザーにとって実用的なサービスを提供することを目指してまいります。

①パラメータ数

ーより高度で複雑な表現力を目指すために多パラメータ（数兆規模）のモデル開発と、実用性を兼ね備えた軽量・高速モデルの開発の両面が鍵となります。

②データ量

ーノイズ混じりのデータではスケーリング則が効かなくなってきています。同じような種類のデータを増やしていても飽和状態がきます*。よってより質の高いデータをいかに投入するかが鍵となり、高品質な事前学習データ及びバリエーションに富んだインストラクションデータの投入による精度向上の実績とノウハウ、またデータ作成人員の教育で競争力を持つ必要があります。

*データ数を半分以上削ったとしても、削った後のクリーンなデータの方が効果的であることを示す論文参考：https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/A3-2.pdf

③計算量

ー計算資源の枯渇と高コストという大きな課題に直面しています。特にエンタープライズ向けGPUの不足とクラウドリソースの価格高騰が課題で、これらの問題を解決するために、我々は分散コンピューティング基盤を活用し、世界中のGPU資源を効率的に利用することを目指しています（EMETHプロジェクト*）。この取り組みは、計算資源の流動性を改善し、誰もが簡単に高度なAI技術を利用できる世界を実現することを目標としています。

*オルツの分散コンピューティング基盤EMETプロジェクト参考：Alt EMETH — Super High Speed Distributed Computating Platform

＜当社の目指すLLMの方向性～意外と見過ごされるスピードとコストの重要性～＞

LLMの開発では、モデルのパラメータ数を増やすことで精度や表現力が向上することは確かですが、それだけでは実用性に欠ける場合があります。例えば、1プロンプトに対して30分もの時間がかかるようなモデルでは、ユーザー体験が損なわれ、実際には誰も利用しないので、リアルタイムでの反応速度が求められています。また、1プロンプトあたり1万円もかかるようなモデルでは誰も利用しません。

サービスとしての反応速度を実用的なレベルに保つには、単にLLMの処理能力を高めるだけでは不十分であり、ハードウェアレベルでの最適化、例えばLLM特化チップの採用などが必要になります。さらに、APIとしての高速化アーキテクチャの構築、通信技術基盤、分散化技術を用いた可用性の向上など、ソフトウェア面での工夫も欠かせません。

また、ネットワークレイテンシーの問題を克服するためには、クラウドだけでなく、エッジ側（端末側）での処理*を活用した高速化も重要です。これにより、ユーザーの要求に即時に応答できるシステムを構築することが可能となります。

*業界動向”Apple、オンデバイス処理に特化したフランスのAI企業を買収”参考：https://gori.me/apple/apple-news/152746#google_vignette

つまり、LLMの競争においては、モデルの精度や表現力の向上だけでなく、スピードとコストのバランスを取ることが重要です。これらを適切に管理しながら、ユーザにとって実用的なサービスを提供することが求められています。

＜根底にあるLLM/GPUのエネルギー問題＞

加えて、LLMの運用とGPUリソースの使用は、大量の電力を消費します。このエネルギー問題は、環境への影響だけでなく、運用コストにも大きく関わってきます。特に、LLMのトレーニングや推論には膨大な計算量が必要で、これを支えるGPUリソースは多くのエネルギーが必要となります。

地理的な電気代の差異を考慮すると、電力コストが比較的低い地域でのデータセンターの運用が経済的です。このため、リソースの地理的分散化は、エネルギーコストを削減する有効な戦略となります。しかし、これにはデータ転送の遅延や、特定地域の法規制などの課題も伴います。

一方で、すべてをサーバー側で処理するのではなく、エッジ側（端末側）での処理を行うことによる消費エネルギーの分散も重要です。エッジコンピューティングにより、データ転送の量を減らし、応答時間を短縮することができるので、サーバー側の負担を軽減することで、全体的なエネルギー消費を削減することが可能となります。また、エッジデバイスの進化により、より高度な処理を端末側で行えるようになることも、このアプローチの可能性を広げることができます。

総じて、LLMとGPUリソースのエネルギー問題に対処するには、電力コストの低い地域へのリソース分散化と、エッジコンピューティングを活用した消費エネルギーの分散が有効な戦略となります。これにより、環境への影響を減らしつつ、運用コストの削減を図ることが可能となります。

上述の課題意識と現状を踏まえ、当社は、エンドユースケースの最適化を実現するために重要な「クオリティ」「スピード」「コスト」に対して、以下の取り組みを加速いたします。

- 学習データ（学習トークン）の大規模な構築
- インストラクションデータの構築と自動化
- プロンプトエンジニアリングの自動化
- 既存モデルの改良に資する生涯学習、メタ認知の研究加速（モデルのリアルタイム継続学習）
- 軽量モデルによる大規模モデル同様の出力再現（知的蒸留）に資する研究加速
- 量子化導入による推論効率化
- メタ認知プロセス導入による推論精度と品質の向上
- LLM 特化チップ（TPU, LPU, NPU）の研究開発
- 日本語に関するRAGデータベースの整備
- 分散コンピューティング基盤の更なる整備

当社オルツは、本取り組みを通じて、AI技術の可能性を広げ、ビジネス実装力を強化し、社会に新たな価値を提供してまいります。また、本取組みに資する協奏パートナーとの連携も積極的に推進してまいります。

（参考情報）

スケーリング則

- Chinchilla則
    - パラメータ数を活かすために、最低限必要な学習トークン数との比率の法則
        - パラメータ数の20倍のトークン数が必要（なお、最近ではchinchilla則を超えたトークン数で小規模モデルを構築するのがトレンド）
        - このため、1兆パラメータを目指す場合、20兆トークンの事前学習データが必要
            - swallowで当てはめると、1トークン1.5文字
            - 本一冊10万文字とすると、6.6万トークン
            - kindle日本語は47万冊で31億トークン

        - kindle全部で700万冊で4620億トークン
    - 高品質な書籍データを集めるとともに、独自のクローリングも必要
        - commoncrawlだけでは足りない
- その他参考数値
    - 特許：30万件、特許明細書の平均文字数7000~16000（年代による）
    - 新聞：25万記事（主要5社）、50万文字
    - 国会図書館の著作物以外のファイル数十億件
        - NIIに対してのみ提供
        - https://www.ndl.go.jp/jp/news/fy2023/240130_01.html

→よって、日本語は20兆トークン獲得のために相応な規模の構築が必要

生涯学習

- LLMの学習は膨大なコストを要するため、何度も１から作るのは困難である
- よって、新しいデータが出てきたら、既存のモデルを改良できると良いという発想
  →そこで、生涯学習の研究開発が必要
- この生涯学習にはメタ認知の研究が深く関わってくるため、ここに着目する場合は東北大 乾先生（当社Head of AI）の研究にかなり価値が出てくると考えている
- 生涯学習ができるようになるとリアルタイムにモデルが学習し続けられるようになるので、より人間に近いものになっていく

知識蒸留

- 学習が出来ても、例えば１兆パラメータをそのままサービス展開できない
- 知識蒸留して軽量にする必要がある（例としてはGPTのturboモデル）
    - 知識蒸留は大きなモデルを教師として、小さいパラメータのモデルを学習し、小さいモデルで大きいモデルの出力結果を再現できるようにする手法

→サービス展開から逆算すると、なるべく小さいパラメータサイズのモデルにしていく必要があるため、知識蒸留の研究開発が必要

▶LHTM-2／LHTM-OPT／GPT など大規模言語処理ソリューションに関するお問い合わせ先
https://alt.ai/aiprojects/gpt/

■株式会社オルツについて
2014年11月に設立されたオルツは、P.A.I.®（パーソナル人工知能）、AIクローンをつくり出すことによって「人の非生産的労働からの解放を目指す」ベンチャー企業です。生成AI、独自開発LLM及び音声認識技術をはじめとするAI要素技術を豊富に保有し、それらを活用した多くのAI Productsを開発・提供しています。2024年4月までの累計調達額は約100億円超に達しています。
https://alt.ai/

＜報道関係者からのお問い合わせ先＞
株式会社オルツ　広報 西澤
e-mail：press@alt.ai

＜アライアンスに関するお問い合わせ先＞
株式会社オルツでは、IT・金融・建設・物流・メディア・製造・小売・サービス業など、
ジャンルを問わずAIソリューションの提供および支援を行っております。
お気軽にお問い合わせください。

株式会社オルツ　AI Solutions事業部 小村
e-mail：gptsolutions@alt.ai