生成AIは大きくなるのか、小さくなるのか

現在の生成AIを支えているのは「Transformerモデル」と呼ばれるアーキテクチャだ。このアーキテクチャの優秀性は大規模言語モデル（LLM）の登場で証明された。一方で、LLMの課題・限界も指摘されている。この課題・指摘を解消するために小規模言語モデル（SLM）と呼ばれるモデル構築が加速している。SLMは生成AIの景色を変えてしまうかもしれない。

「Transformerモデル」という革新

2017年、AI領域で革新が起きた。8名のAI研究者による“Attention is All You Need”[1]という論文が提示したTransformerと呼ばれるAIアーキテクチャによって一気に大規模言語モデル（LLM）の可能性が示された。このTransformerモデルは、それまでのディープラーニングモデルの性能を一気に引き上げ、生成AI領域での日進月歩を引き起こす契機となった。

実際、OpenAIやGoogle、Microsoft、Metaといった主要生成AIプロバイダーが提供している生成AIであるLLMのほぼすべてに、このTransformerモデルが採用されている。

Transformerモデルは、それまでのディープラーニングの学習過程に革新をもたらした。それまでのディープラーニングでは、テキストや画像といった入力データを分析する際に、単純な前後関係（相関関係）を計算していた。しかし、Transformerモデルは、データの構造を分析する際に分析対象をより前後に広げる（距離の拡張）ことに加え、分析対象であるデータの相互の関係をより精密に数値化することに成功した（位置の精緻化）。

この革新により、Transformerモデルは学習するデータを増やせば増やすほど（そして計算量を増やせば増やすほど）回答が正確になるという「スケーリング則」と呼ばれる性質を備えている。そして、このTransformerモデルを採用した生成AIは、幾何級数的に学習データを飲み込み、そして大量のGPUと電力を使いながら、驚くべき性能を発揮しだしたのである。その後の生成AIの進化のスピードはご存知のとおりだ。

LLMの限界？

一方でLLMの機能進化が今後も順調に続くのかという点に関して懐疑的な意見も増えてきた。LLMの成長を阻害する可能性を持つ要因は複数指摘されている[2]。ここではそのうちの2つを挙げる。

一つはLLMが引き起こす「ハルシネーション（捏造）」の問題だ。現在のTransformerモデルを元に作られたLLMは、確率的に「テキスト」を分析し、確率的に「テキスト」を回答せざるをえない。そのため、一定の確率で学習段階における誤った解釈が生じ、また回答段階でも確率的に誤回答を生み出してしまう。このハルシネーションを回避する手法が様々に検討されているが、どれも「ハルシネーションの頻度を下げる」というレベルにとどまっている。また、LLMはある時点の情報を元にモデルが構築されるため、モデル構築以降の最新のデータには対応できない。さらに、学習データに含まれていない情報はそもそもまともな回答が期待できない。Transformerモデルに立脚するLLMの機能進化にはどこかに限界があるのではないかとの危惧は自然だろう。

もう一つの課題は、モデルを生成する際に消費される資源が維持不可能な水準になりそうなことだ。ここでいう資源には「電力」「学習データ」「計算能力」が挙げられる。これらの物理的な資源の制約が、これ以上のLLMの発展の足かせになるのではないかと言われている。

そして、AIの究極的な目標とされている「汎用人工知能（AGI：Artificial General Intelligence）」は現行のLLMの延長線上にはないのではないかという懐疑論も根強い[3]。

NOTE

1) “Attention Is All You Need”
https://arxiv.org/abs/1706.03762
Transformerモデルの嚆矢となった論文。同論文の著者は8名なことから、日本の一部では「神エイト」と称されていたりする（AKB48の「神セブン」が元ネタなのだろう）。
2) 「AIの機能進化予測の難しさ」金融ITフォーカス2024年4月号。
https://www.nri.com/jp/knowledge/publication/fis/kinyu_itf/lst/2024/04/06
3) ChatGPTを世に送り出したOpenAIのサム・アルトマンCEOも現在のTransformerモデルの延長線上にAGIはないのではないかと懐疑的な意見を述べている。
“OpenAI CEO Sam Altman Discusses GPT-5, Sora, and the Road to AGI”
https://www.maginative.com/article/openai-ceo-sam-altman-discusses-gpt-5-sora-and-the-road-to-agi/
4) 例えば、Microsoftが発表したSLM「Phi-3」のコンセプトについては以下を参照。
「小さくても強力：小規模言語モデル Phi-3の大きな可能性」
https://news.microsoft.com/ja-jp/2024/04/24/240424-the-phi-3-small-language-models-with-big-potential/

新たな挑戦：RAG、SLM

当然ながらTransformerモデルに立脚するLLMの抱える問題に対して様々な解決策も模索されている。

ハルシネーションという課題に対しては、「検索拡張生成（RAG：Retrieval Augmented Generation）」という仕組みを導入することでハルシネーションの弊害をある程度は抑制できることがわかっている。このRAGとは、誤解を恐れずに言えば「AIに最初に答を教えてしまう」という手法だ。ただ、答そのものを教えては意味がないので、回答にたどり着くために必要な情報を人間側が質問時に提示するというやり方だ。このRAGはLLMの仕組み自体に手をいれることなく回答精度を高めることができることがわかっている。

もう一つのTransformerモデルにつきまとう資源消費の問題にも新たなアプローチが出始めている。その最たるものが「小規模言語モデル（SLM：Small Language Model）」と呼ばれるものだ。要は「大規模」ではなく「小規模」な学習データ・計算資源でも精度の高い回答を得られるようなモデルを構築できないかというチャレンジだ。

2024年4月16日、Googleは自社として初のSLMモデルとなる「RecurrentGemma」を発表した。またその8日後の4月26日には今度はマイクロソフトが「Phi-3」と名付けたSLMを新たに発表した。両モデルとも先行するGPT-4やLlama，GeminiといったLLMよりもはるかに少ない学習データで、LLMと遜色ない良好な性能を実現したとしている[4]。またSLMは消費するエネルギーも必要とする計算資源も少なくて済む。

「手のひらの上のAI」時代がくる？

現在のLLMを利用するにはWeb経由、もしくはAPI経由での利用が前提となっている。元となるLLMモデルは大規模なデータセンターで稼働しており、実は我々はあくまでその大規模なLLMモデルの機能をネット経由で「間借り」させてもらっているに過ぎない。

しかしスマートフォンや普通のPCに搭載可能で、大規模なモデルに接続する必要がないSLM技術がもし本物であれば、生成AIの利用スタイルは一気に変わる可能性が高い。スマートフォンやPC上のSLMに、自身のパーソナルデータを読み込ませて（RAG技術の活用）、自分用に最適化した「机の上・手のひらの上のパーソナルAI」が実現する可能性が出てきているのだ。

現在のLLMはすでに様々な業務・タスクをある程度の正確性をもって処理可能だ。文章理解、音声認識、文章作成、画像認識、タスクの自動処理などが代表例だが、このようなLLMが提供できる機能とほぼ同様の機能をSLMが持てるとすれば世界の景色は大きく変わる。

筆者はドラえもんに出てくる「ほんやくコンニャク」という「他言語同時通訳ツール」をずっと欲しいと思っていた。実はそういう未来はもうすぐそこまできているのかもしれない。

Writer's Profile

柏木 亮二 Ryoji Kashiwagi
金融デジタルビジネスリサーチ部
エキスパート研究員
専門はIT事業戦略分析
focus@nri.co.jp