FERI INFORMATION

エグゼクティブのための生成 AI 講座
その１ LLM（大規模言語モデル）とは何か

副島 豊（SBI 金融経済研究所・研究主幹）

生成 AI、とくに LLM（大規模言語モデル）を金融ビジネスに活用し、１）既存業務の効率化や高度化を図ること、２）新しい金融サービスを創造すること。日々様々なニュースリリースが出ているように、多くの金融機関がこの課題に取り組んでいる。

競争力の強化や新しい収益源の模索に直結するものであり、他社がやって自社がやらない状態が続けば、やがては競争力や顧客基盤、ひいては収益力が失われていく。2000 年代初期のネットバンキングやネット金融サービスの幕開け期に近いのかもしれない。

生成 AI は、ハイプサイクルの盛り上がりと幻滅期を通過しつつ、スマートフォンやキャッシュレス決済のように着実に企業活動や社会生活に定着し、金融ビジネスにおいても当たり前のように使われる技術となっていこう。

金融機関では「生成 AI への取り組みを」というトップダウンの号令がかかり、各事業部門も何に使えるのか、どのようなビジネスモデルが可能となるのか、技術キャッチアップと活用のトライアルを続けている。こうした事業アイディアや実装の試み、その成果を取り上げた記事も多く出ている。

その一方で、生成 AI がどのような仕組みで動いているのか、どうやって顧客向けサービスを提供するシステムを構築すればいいのか、そうした技術面に踏み込んだ記事を経営層やビジネス現場に向けて提供したものは少ない。

ネットには膨大な一次情報（技術開発者/社が提供する使用法や解説）や利用者が作成した解説記事があるが、殆どは開発者向けのものである。書籍として ChatGPT の使い方指南書などが多数出版されているが、個人が生産性を向上させる目的のものが多く、ビジネス展開というサービス提供側の視点に立ったものは少ない。

そこで本稿では、エグゼクティブ向け・ビジネスパーソン向けに、LLM とは何か、どのような技術が使われ、どう実装されているのかを、技術面に踏み込みながら、かつできるだけ平易に解説することを試みた。紹介すべき情報量が多いため、シリーズ形式で掲載する。

LLM とは

LLM とは何か。端的にいうと、この言葉がきたらこんな言葉が続く可能性が高い、この話の流れからはこのような展開になるだろう、こんな話や指示が振られているからこの内容を返すのが適切そうだ、そうした適切な言葉の繋がりを確率計算で求め、次にくる言葉を連続的に生成することで文章を作り出すモデルが LLM である。本当にそのようなアプローチで上手く文章が作れるのか直感的には疑問であるが、とても上手くいったというのが LLM の衝撃的な誕生と発展のストーリーである。

こうしたアプローチが可能となるためには、膨大な文章を学習情報として収集し、文章のパターン性を精緻に学習する必要がある。ある話の流れがあって次にくる言葉を適切に選ぶといった場合、その「適切」とは何を意味するのか。

教師（正解）あり学習モデルでは、学習用の膨大な文章において言葉があるパターン性をもって並んでおり、その並び方、登場の仕方、組み合わせ方、前の受け方、後ろへの続き方などを上手く模倣できることが「適切」を意味する。次にくる確率が最も高い言葉を適切に選ぶようモデルのパラメータを設定していくことが自然言語モデルにとっての「学習」であり、人間の学習のイメージのように知識や体系を学ぶ・記憶する行為とは概念的に異なる。

画像情報からそれが何であるかを判別する学習では正解データが必要となる（犬の写真に犬というラベル、トイプードルの写真にトイプードルというラベル）。この膨大な情報を学習用に作成するのは大変であるが、文章サンプルでは次にくる言葉は既に分かっている。しかも、切る場所を変えれば 1 文で多くの学習用問題が作れる。

日本語 Wikipedia では 140 万近い項目について膨大な解説文がある。Wikipedia は文章や言葉の使い方を大規模に収集した「コーパス」の代表例として LLM の学習に利用されてきた。また、ネットにある膨大な文章を集めて整理した更に大規模なコーパスも複数ある。例えば、国立情報学研究所（NII）が 10 月 20 日に公開した LLM-jp-13B では、パラメータ数 130 億、学習データとしては約 3000 億トークンを利用している（トークンは後述、日本語では 1 トークンは概ね 1 文字に相当）。

まとめると、LLM とは、自然言語を生成するプログラムを、大規模な学習データを用いて、大規模なモデルとして作成したものである。

ChatGPT と LLM の関係

生成 AI では OpenAI 社の ChatGPT が有名である。OpenAI 社の Web サイトを訪れ、Web アプリケーションサービスとして利用する。同社は他の様々な LLM を開発・提供

し続けており、ChatGPT サービスの裏側で動いているのはこうした LLM であり、その最新型が GPT-4 である。SaaS として Web アプリケーションサービスを利用する以外に、OpenAI 社のサーバに API でアクセスし、特定の LLM を指定して利用することもできる。

LLM には、チャットが得意なもの、日本語が得意なもの、特定専門分野に特化したもの、文章を数字に置き換えるという前処理を行うのが得意なものなど様々な種類がある。また、例えば GPT-3 といってもモデル規模や性能進化に応じて種々なバリエーションが存在している。Google や Meta ほか様々な企業が LLM を開発しており、日本でも日本語の文章処理能力を高めた多くの LLM が開発され続けている。

モデルとパラメータをオープンソースとして公開した LLM も少なくない。こうした LLM は、パブリッククラウドや自前のサーバにシステムを構築することができる。軽量なものはノート PC でも稼働する。オープンソースとして公開されてはいないがパブリッククラウド上で LLM を活用したシステムを構築できるようサービス提供されているものがある。Azure OpenAI がその代表例である。

ところで、ChatGPT が注目され始めた当初、Google 検索エンジンのように調べものに使った結果、嘘を教えられた、あれはダメだという話がよく聞かれた。LLM は文章の展開が尤もらしくなるよう単語や文を並べているだけであり、その尤もらしさが事実である保証はない。

学習に用いたコーパスにその事実に関する記述が大量にあり、A といえば B が出てくるように学習結果がパラメータに反映されているのでなければ、文章として尤もらしいが内容は嘘という現象は当然生じる。元になる情報がなければ適切な回答は返せない。こうした汎用 LLM の限界を改善していく方法として様々なアプローチが提案され、実用化されている。この話はシリーズの後半で解説する。

LLM のベースはニューラルネットワーク

ここまでの説明で「モデル」という言葉を使ってきたが、LLM のベースにあるモデルの先祖はニューラルネットワーク（以下、NN）である。1980 年代前半に登場し、第二次 AI ブームを牽引した NN は、脳のシナプスの構造を単純化して表現したモデルであった。

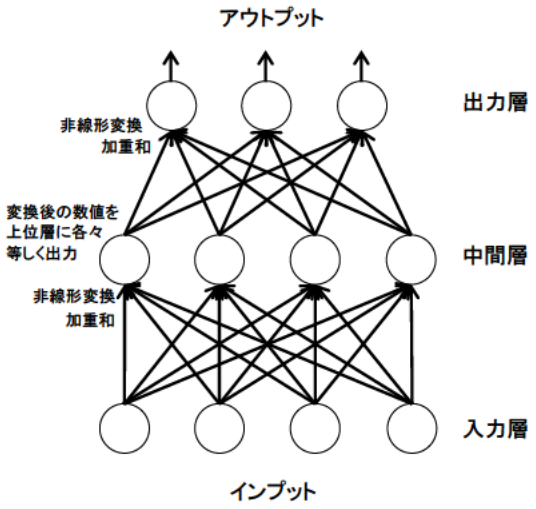
次図のようにインプットとアウトプットを中間層を使って繋ぎ、アウトプットが正解と合致するようパラメータを調整する（学習する）仕組みである。例えば、植物のアヤメの特徴を現すデータとして花びらやガク片の長さと幅のデータがあったとする。これがインプットで、アウトプットがアヤメの種類である。花びらやガク片のデータから種類を正しく言い当てられるよう NN のパラメータが推計される。

金融分野ではデフォルトの判別分析や格付けモデルがこれに近く、データによる判別機械といえる。ちなみに判別分析は機械学習の最も代表的な用途であり、NN 以外にも様々なモデルが考案されている。

計量経済学でも質的選択モデル/離散選択モデルとして昔から開発されてきたが、NN は

関数の非線形性の表現力が高い点が注目された。筆者は 1990 年代前半に、金融政策反応関数（当時は公定歩合の上げ/下げ/据え置き）に NN を適用し、政策変更を高い精度でトレースするモデルを推計したことがあり、論文中でその柔軟な非線形性を可視化している（リンク先の図 17・18）。

ニューラルネットワークの基本構造



出所：筆者作成

NN のインプットやアウトプットは、数値もしくは数値化された画像や音声が中心であった。例えば、アウトプットについて正常先を 1、破綻懸念先を 4 とインデックス化し、インプットを企業の PL/BS の様々な指標などから集めてきたような信用リスクモデルをイメージすると解りやすい。画像も数値化が容易である。1 万画素の 24 ビットカラー画像は、各画素を RGB（Red/Green/Blue）3 色の 0～255 段階色調で表現することによって数値化することができる。

インプットとアウトプットを繋ぐのは四則演算に過ぎない。指数と分数を使うことで非線形性が導入されているが、各過程の計算は非常にシンプルである。インプット変数を様々なウエイト（パラメータ）を付けて合算し、それを非線形関数で変換したものを中間層へのインプットとし、これを次の中間層への出力として同じことを繰り返し、最終的にアウトプット層から出力がなされる。単純な関数を大量に組み合わせることで複雑な関数を作りだす仕組みとなっている。このため、モデルは多くのパラメータを持つ。

NN は中間層の階層を増やしたり、インプット要素を増やしたりするとモデルの精度や表現力が高まることが期待されたが、計算技術やコンピュータの計算能力の制約からそうはならず、80 年代の第二次 AI ブームは徐々に収束していった。その後、長い AI 冬の時代が訪れる。

2000 年代後半にこの限界をブレイクする技術が現れた。ディープニューラルネットワーク（DNN）であり、これが第三次 AI ブームを切り開いた。中間層を数十も積み重ねてもパラメータの学習が上手く行われる手法の開発や計算能力の確保によって、NN のパフォーマンスが飛躍的な向上を辿り始めた。LLM には NN/DNN を発展させたモデルが利用されている。

次のレポートでは、自然言語モデルに NN を適用する際の鍵となった要素を解説し、LLM がどのように発展し現在のような高性能を持つに至ったかを順に紹介していく。

（その 2 に続く）