

解題

証券アナリストジャーナル編集委員会
第三小委員会委員 岡田 克彦

1. 画像識別・自然言語処理の大幅な技術進化と オルタナティブデータ

最近のAI研究領域における諸技術の進化速度はすさまじい。最新技術に関する論文を読んでいるうちに次の技術が開発され、すぐに陳腐化してしまうほどである。このような指数関数的新技術の発展は、社会における潜在的な影響度の大きさから、産業革命にも匹敵すると評されることがある。この始まりはどこにあったのだろうか。

AI研究の歴史は1950年代にさかのぼるが、最もエポックメイキングだったのは2012年である。この年開催された画像識別（computer vision）の国際会議、ILSVRC（ImageNet Large Scale Visual Recognition Challenge）は、画像識別研究の歴史において長く記憶されるだろう。2012年以前の画像識別モデルは、対象となる画像の特徴量を研究者が緻密に設計し（feature engineering）（注1）、独自のモデルを組むことで識別正答率を向上させてきた。わずかな改善を研究者間で競い合い、徐々に精度が向上してきたと

いう研究領域であった。しかし、2012年のILSVRCでは驚くべきことが起こった。ジェフリー・ヒント博士とその学生たちが開発したアレックスネット（AlexNet）が、従来のものとは全く違うアプローチを採用し、2位以下に圧倒的な差をつけて優勝したからである。アレックスネットとその他2位以下のモデルとの違いは、人間の研究者が画像の特徴量を考えてモデルに組み込むのではなく、ビッグデータ（注2）から自動的に機械が特徴量を学ぶ仕組み（深層学習、Deep Learning）を導入した点だ。深層学習の考え方は以前からあったが、圧倒的なパフォーマンスの違いを見せつけたのはこれが初めてであった。したがって、2012年は画像識別の世界で技術的ブレイクスルーが起こった年だといえよう。この後、同技術は大きく注目され、深層学習の仕組みと、畳み込みフィルター（Convolutional Filter）を組み合わせたニューラルネットが登場。識別精度は短期間にさらに向上した。現在、その精度は実用レベルまで向上し、私たちが日々用いるスマー

（注1） 特徴量職人と呼ばれる研究者が、画像識別させるための特徴量を考えてモデルに組み込む作業を feature engineering と呼ぶ。例えば、猫と犬の画像識別タスクにおいては、猫と犬を識別する特徴は何かについて考え、モデルに組み込むのである。多くの研究者が独自の工夫を凝らしコンテストに参加するが、精度の大きな向上は長年みられなかった。

（注2） <https://www.image-net.org/> に登録することで、大量のラベル付き画像を入手することができる。

トフォンの顔認証機能など、様々な領域において応用されている。この画期的技術革新の資産運用や金融への応用可能性については、約10年後の Jiang *et al.* [2020]、Okada *et al.* [2023] で示唆されるまで待たねばならず、技術の金融実務への応用は少し先になるだろう。

画像識別よりも少し遅れて、2017年に自然言語処理 (NLP) の領域にブレイクスルーが起こる。詳細に説明する紙幅はないが、注意機構 (Attention Mechanism) (Vaswani *et al.* [2017]) というメカニズムの導入である。注意機構とそれを利用したアーキテクチャであるトランスフォーマーの出現以降、NLPの研究は加速度的に進化した。特に、巨大コーパスを、巨額の費用をかけ、1兆を超えるパラメータを持つ巨大モデルで事前学習させたモデル (PTM) が研究者向けに公開されたことで、言語構造そのものを理解したモデルは誰でも使えるようになった (注3)。大規模言語モデル (LLM) とは、事前学習の仕組みを使った様々な言語モデルの総称である。一般の研究者は、こうした巨大企業の費用でトレーニングされたPTMを利用することで、小さなデータしかなくとも、領域特化した言語モデルを開発できるようになったのだ。

研究者が特徴量を考えていた時代の画像識別の研究や、言葉の一つ一つに着眼する方法 (Bag of Words : BoW) で行われてきた自然言語処理 (NLP) の研究水準は、2012年以降、指数関数的に質が高まっていった。その進化のスピードはあまりにも速く、応用研究が全く追いついていない状況だ。

この10年間で起こった画像識別と自然言語処理のブレイクスルーは、間違いなく社会を変える潜在力を持つ。とりわけ、オルタナティブデータ

は画像 (衛星画像、チャート画像等々) やテキスト (ESG、統合報告書、ニュース、SNS) の形式をとることが多く、昨今のAIの技術革新との相性は良い。これからのオルタナティブデータの活用は、こうした技術革新と切り離して考えることはできない。

2. 現状におけるオルタナティブデータの活用

これまで、オルタナティブデータの価値を定量的に示すことを目的にした研究はいくつか行われている。とりわけ直感的に有用性が高そうなオルタナティブデータは、POSデータであろう。いち早く店舗売り上げが分かれば、それを集計することで誰よりも早く企業の売り上げを予想することができるように思えるからだ。ただ、企業全体の売り上げの一部でしかないPOSデータの扱い方と結果の解釈については注意が必要だ。複数の注意点が吉野・山本論文「POSデータ分析と運用実務への応用可能性について」で示唆されている。彼らはさらに、POSデータそのものよりも、JANコードの変動に着眼し、JANコードの増減が激しいほど企業の株価パフォーマンスが劣ることを発見した。この事例から得られる示唆は、単純に販売されているデータをそのまま使ったとしても、効果は一時的で持続性は担保されないということだ。 α を継続的に獲得するためには、オルタナティブデータを直接使うのではなく、運用者が他の市場参加者には見えていない重要な情報を可視化するツールとして使う、という姿勢が必要であろう。販売されているオルタナティブデータは直感的に分かりやすく、効果があるものであればあるほど、競争相手も購入すると考えられる。その場合、 α の源泉は徐々に枯渇する。効果が高いもの

(注3) すべてではないが、Google、Facebook、OpenAIの3社は、それぞれBERT、fastText、GPTをオープンソースとしている。

であればあるほど、枯渇が早いというジレンマに陥ってしまう点に注意が必要だ。McLean and Pontiff [2016] がアノマリーについて、論文発表後に効果が減衰する性質を実証しているが、オルタナティブデータについても同様の原理が働く。本特集の座談会において諏訪部氏が、オルタナティブデータを使って独自性が高い運用をすることが重要であると示唆しているように、 α を獲得するためには独自の解釈・方法に基づいた活用が必要になろう。

森田・臼井論文「特許データを活用した企業の研究開発活動の分析」は、観察できないR&D活動の変数を、特許データと入手可能なデータから推測している。そして、R&D活動がゼロと報告されている企業でも、現実にはR&D活動を行っている企業を特定し、データを修正した上で回帰することで、R&Dと企業価値の価値関連性を明らかにしている。この論文で見えてくるオルタナティブデータの有用性は、企業価値には関連していそうだが、観察できない要因や、観察できるがデータが公開されていないものを分析の材料にできるということである。これも、見えないものを可視化してくれるオルタナティブデータの効用である。したがって、数値化はされていなくとも経営者の研究開発に対するビジョンなどが記述されているテキスト情報があれば、それらをLLMで解釈し、本論文のロバスト性を高めることができるだろう。

野崎・田代・岩田論文「オルタナティブデータのエンゲージメントへの活用」は、残余利益モデルを用いて企業の資本コストを推定し、その推定値を財務・非財務情報を用いて要因分解している。筆者は、企業のインプライド資本コストをマケ

ット要因、財務要因、非財務要因、企業固有要因に分解し、資本コストの約76.5%しか説明できていないとしている。この結果から、投資家評価に影響を与えており、かつ定量化できていない非財務要因はまだ多く存在すると主張している。彼らはその一例として統合報告書のマテリアリティ欄の充実度と資本コストとの関連性を検証した。そして、「記載が充実している企業ほど資本コストが下がる」という回帰結果を得ている。現在のLLMであれば、統合報告書をそのまま解析し、充実度の部分をより詳しく評価できるだろう。

3. 今後のオルタナティブデータの活用の方向性

座談会「資産運用におけるオルタナティブデータ活用の現在と今後」においては、オルタナティブデータの中でも、とりわけテキストデータの活用可能性について、活発に議論されている。テキストデータは、POSデータや駐車場付き店舗の衛星画像から売り上げを予想するような、直接的に業績予測に資するものではないが、その応用可能性は幅広く、独自の運用戦略を考えるに役立つものだ。一方で、テキストデータに関しては、出所のよく分からないデータやSNSデータが含まれるため、相応の注意を払って扱うべきでもある。

最も安心なテキストの利用法は、信頼できるニュースソースだけを使用する、廣瀬氏の所属企業における活用法だ。機械が世界のニュースを網羅的に解析し、人間（運用担当者）にアラートを投げる仕組みを導入しているという。これは業務効率を上げ、運用担当者が本来やるべき仕事に集中できるというメリットをもたらすことだろう。しかし、もう一步踏み込み、より積極的に α を獲得するために、オルタナティブデータはどのように活用すべきだろうか。解題者は、画像識別と自然言語処理で起こった進化の成果をいち早く利

用し、試行錯誤を含めた実証研究を蓄積していくべきだと考える。米国のアカデミアは応用研究に貪欲である。Lopez-Lira and Tang [2023] は、大規模言語モデル (ChatGPT) が2022年11月に公開された半年後には、ファイナンス応用におけるその能力と既存手法との比較検証結果を報告している。彼らによると、LLMの解析能力は驚くほど正確だという。従来データベンダーが有料で展開しているセンチメント指数よりもChatGPTで判断したセンチメントの方がリターンの予測力が高いという。Hansen and Kazinnik[2023]は、FRBの金融政策決定会合後に公開される非常に曖昧な表現で満ちているFedspeakをChatGPTに説明させたところ、FEDの専門家とほぼ同じコメントを返したという。また、Itoh and Okada[2023]は全上場銘柄の四季報コメントを16年分ChatGPTに解釈させたところ、リターンの予測可能性を持つことを報告している。

こうした優れたツールが入手可能となった現在、画像識別でも自然言語処理でも積極的に活用されていくだろう。そうなれば、オルタナティブデータの価値は大きく変化していく。従来はテキストの良しあしを単語レベルでしか判断できなかったものが、文脈の中で正確に意味を抽出できるようになったのだ。この応用範囲は広く、オルタナティブデータを持っていないことは、競争上、大きな問題になってくるかもしれない。

今回の座談会では、オルタナティブデータの有用性をどのように宣伝していくかという問題についても語られていた。いま時代が変わり、精緻なテキスト解析ができるツールが誰でも使える状態になった。言語構造を理解したLLMであれば、オルタナティブデータ、中でもナラティブなテキスト情報の解析は正確かつ簡単である。これを使った研究活動の中から α の源泉となるものも生ま

れてくるだろう。ただ、市場は高度に競争的だ。これまで、新しいツールを活用して得られた超過リターンや、研究活動によって既知となったアノマリー群についてもそうであったように、新技術による α の源泉も長く続くとは限らない。生き残る運用者は、常に最新のツールを活用しながら創造的な人間であることは忘れてはならない。

(参考文献)

- Hansen, A. L. and S. Kazinnik [2023] “Can ChatGPT Decipher Fedspeak?,” Working Paper. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4399406>
- Itoh S. and K. Okada [2023] “The Power of Large Language Models: A ChatGPT-driven Textual Analysis of Fundamental Data,” Working Paper. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4535647>
- Jiang, J., B. T. Kelly and D. Xiu [2020] “(Re-) Imag(in)ing Price Trends,” Chicago Booth Research Paper No.21-01. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3756587>
- Lopez-Lira, A. and Y. Tang [2023] “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models,” Working Paper. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4412788>
- McLean, R. D. and J. Pontiff [2016] “Does Academic Research Destroy Stock Return Predictability?” *Journal of Finance* 71(1), pp.5-31.
- Okada, K., Y. Hamuro and M. Nakasuji [2023] “Decoding the Unique Price Behavior in the Japanese Stock Market with Convolutional Neural Networks,” Working Paper. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4478013>
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin [2017] “Attention Is All You Need,” arXiv:1706.03762v.5.