２０２５年１２月２５日発行　第５７１号

2025
12

ふじさん

fujijoho group monthly magazine

富士情報

～ ２０２５年指針 ～

開　点

［今月のひとこと］

TPU GPU

・国字

・冬至

水かけ菜の収穫

写真提供：都留市　産業課

今月のひとこと

TPU GPU

社 長 渡 辺 直 企

11 月 19 日に Google が新しい AI、Gemini3 をリリースしました。この AI はほとんどすべてのテストにおいて Chat GPT5.1、Grok 4.1 などの競合モデルを上回るスコアを記録しています。これを受けて ChatGPT を提供する OpenAI のサム・アルトマンは同日「Congrats to Google on Gemini 3! Looks like a great model.」とお祝いをツイートしていました。12 月 2 日にウォール・ストリート・ジャーナルはサム・アルトマンが社内向けに「コード・レッド（緊急事態）」宣言したと報じています。その後 12 月 11 日に OpenAI は性能を知的労働で専門家を超えるレベルまで大幅強化した GPT5.2 を発表しています。

OpenAI をはじめ Meta の Llama や xAI の Grok などほとんどの AI は NVIDIA の GPU（Graphics Processing Unit）を使用しています。その名の通り元々は 3D グラフィックスの描画を高速に処理するために作られ、CPU に比べ膨大な並列処理を得意としています。2009 年スタンフォード大学のアンドリュー・ン教授は、機械学習において CPU の代わりに GPU を使用することで 1 億パラメータのニューラルネットワークの学習に必要な時間を数週間から約一日に短縮できると発表しました。このとき使用しているのは白黒 32x32 ピクセルの画像でした。

機械学習では用途別に画像認識向けの CNN(Convolutional Neural Network)、音声認識、株価予測、機械翻訳など時系列データ向けの RNN(Recurrent neural network)、SNS の人間関係、交通網などグラフ(網)学習向けの GNN(Graph neural networks)など多くのニューラルネットワークを開発していました。2017 年 Google は Transformer アーキテクチャを発表しました。Transformer は RNN 同様言語など時系列データを扱いますが、並列処理を強化し、データ全体を俯瞰することで、RNN が苦手だった長文理解や長期記録が可能となり、汎用性を実現しました。また、言語のみならず画像なども処理が可能となりました。Transformer は LLM（Large Language Model）の基礎となっています。

OpenAI は 2018 年に Transformer を予測機能(生成)に特化した生成 AI、GPT-1（Generative Pre-trained Transformer）を発表しました。OpenAI をはじめ多くの生成 AI ベンダーは NVIDIA の GPU で生成 AI を開発するために CUDA(Compute Unified Device Architecture)という開発環境を使用しています。一方、Google は 2013 年から独自で開発している TPU（Tensor Processing Unit）を使用しています。Tensor は行列を多次元に一般化した概念で、TPU はニューラルネットワークを行列とみなした計算に特化したプロセッサです。GPU は画像処理向けですので 4x4 などの小規模な行列演算を得意としています。GPU で生成 AI の大規模な行列演算する場合は、小さい行列演算に分解して計算する必要があり、計算のたびにメモリを使用する必要があります。一方、TPU は 256x256 という大規模な行列演算をメモリの使用を最小限に押さえて処理可能なので高効率になります。生成 AI の処理に関しては TPU は GPU よりも 1.2～2 倍程度、消費電力に優れています。GPU は TPU に比べ汎用的で、処理できる機能も多い一方、生成 AI 向けの半導体という観点では効率性に課題があります。GPT-4 や Gemini2.5 のパラメータは 1 兆を超え、博士号レベルの知識を得るに至っています。GPT-5 や Gemini3 のエージェント機能の拡充で自律的に仕事を任せられるようになってきました。今後、一層便利になるので、積極的に活用していけるようにしていきたいと考えています。