2024 年 3 月 25 日

株式会社ヘッドウォータース

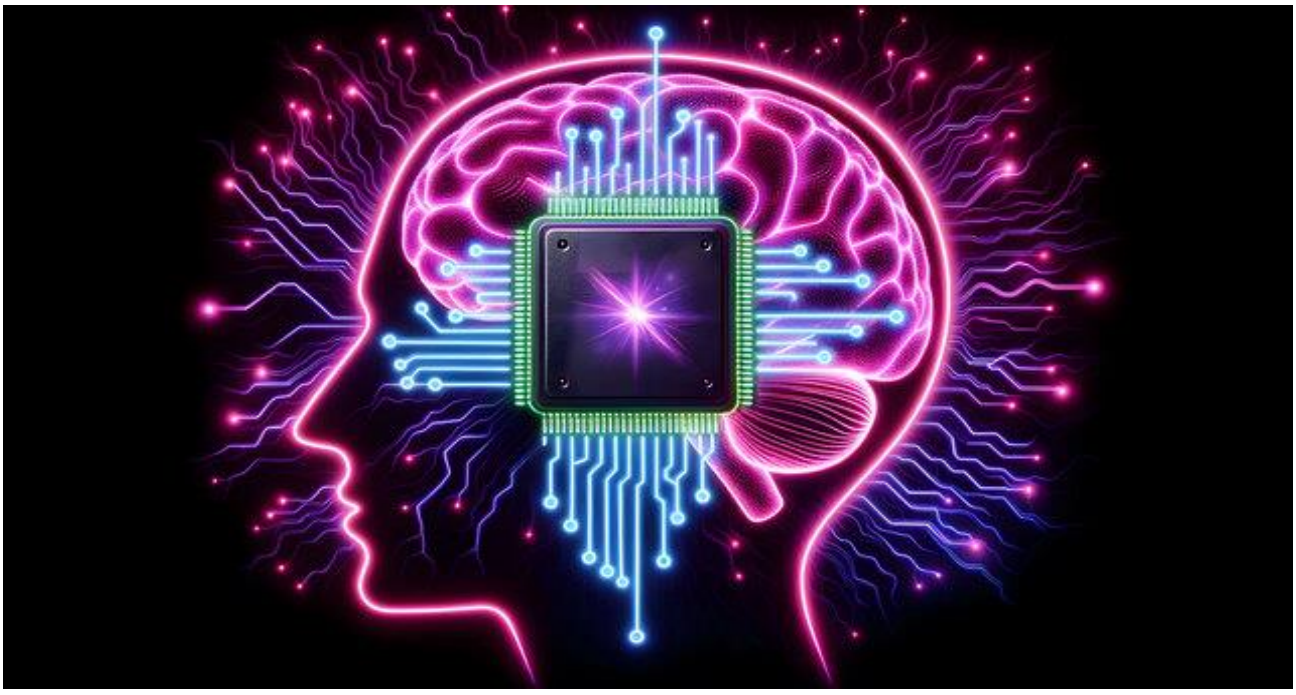
(コード番号：4011 東証グロース)

生成 AI×エッジ AI に向けて

「NVIDIA® Jetson Orin™ Nano」上で稼働する

小規模言語モデル SLM と画像言語モデル VLM の検証を開始

AI ソリューション事業を手掛ける株式会社ヘッドウォータース（本社：東京都新宿区、代表取締役：篠田庸介、以下「ヘッドウォータース」）は、日本マイクロソフトが提供する SLM（Small Language Models:小規模言語モデル）「Phi-2」と、Meta 社が提供する「LLaMA」をベースとしたオープンソース SLM「TinyLlama」、ならびに VLM(Vision-Language Model:画像言語モデル)「LLaVA」を NVIDIA 提供の小型エッジデバイス「NVIDIA® Jetson Orin™ Nano」上で稼働させた動作検証を開始しました。



■検証開始の経緯

ヘッドウォータースでは、「Azure OpenAI Service」による企業向け GPT サービスラインナップの拡充を行っており、企業向けに生成 AI、ならびに LLM（大規模言語モデル）と当社の技術力を活かした RAG（Retrieval Augmented Generation）システム、伴走支援型ラボなど多くのソリューションを企業に提供してまいりました。

また、NVIDIA とのコラボレーションによって、「NVIDIA Jetson」シリーズを活用したエッジ AI ソリューションの開発に取り組み、スマート化を推進する企業に対しても同様にソリューションを提供して参りました。

そのような状況の中、当社が強みとしているエッジ AI 領域で、スマートストア、スマートファクトリー、スマートシティ、スマートモビリティを提供する顧客企業から「生成 AI を使って、さらにスマート〇〇化を進められないか？」というご相談をいただく機会が増えております。

このような声に応えるため、ヘッドウォータースでは、生成 AI×エッジ AI 領域の取り組み強化を目的に、日本マイクロソフトの SLM「Phi-2」と、Meta 社の「LLaMA」をベースとしたオープンソース SLM「TinyLlama」、ならびに VLM「LLaVA」を NVIDIA の「NVIDIA® Jetson Orin™ Nano」上で検証することによって、エッジ AI と生成 AI を組み合わせたビジネス活用方法の整理と活用拡大に向けたアーキテクチャーの確認を行ってまいります。

■検証内容について

SLM の主な利点は、「LLM（大規模言語モデル）の軽量化」にあります。通常 LLM を運用するには莫大なコストが必要となりますが、これは AI が膨大な量のデータを処理するために高価なコンピューティングリソースが必要になるためです。一方、SLM は小規模なデータ処理となるため、消費電力が少ないエッジデバイスのような小型コンピュータ上で言語モデルを稼働させることができ、さらにクラウドを経由せずローカル（オフライン）環境で言語モデルを扱える事から、セキュアでコスト効率も良いという特徴があります。

これらの特徴を最大限に発揮させるため、オープンソースの VLM「LLaVA」を「NVIDIA® Jetson Orin™ Nano」で稼働できるように 4 ビット精度で量子化してメモリ使用量を削減させます。これによってパフォーマンスを向上させた「nanoVLM」「Live LLaVA」を活用でき、画像や映像、テキストを読み込むマルチモーダルな生成 AI の稼働が可能となります。

検証を通して、

- 自動車×生成 AI による「音声対話できる自動車」
- スマートファクトリーにおけるセキュリティを考慮したオンプレミス型生成 AI
- 音声によるロボティクスの機械制御
- スマートシティの都市 OS データを活用した生成 AI による案内
- スマートストアにおいて自動接客を行う生成 AI

といったソリューションの展開を目論み、このような事例で生成 AI×エッジ AI 領域の言語モデルテクノロジーを活用してまいります。

■SLM（小規模言語モデル）とは

SLM（小規模言語モデル）は、LLM（大規模言語モデル）よりもサイズが小さく軽量化された言語モデルです。高速なトレーニングと推論が可能で、リソース効率も高まり、コストパフォーマンスに優れています。また、リソースに制約のあるデバイスやエッジコンピューティングに適しており、セキュアで機密性が高いと言った様々な特徴があります。より小型となる言語モデルの可能性が生成 AI カテゴリーで注目されており、小規模言語モデルの採用が増加しております。

■Phi-2 とは

日本マイクロソフトが提供する小規模言語モデルで、優れた推論能力と言語理解能力を実証する 27 億パラメータの言語モデルで、130 億未満のパラメータを持つ基本言語モデルの中で最先端のパフォーマンスを示します。複雑なベンチマークでは、Phi-2 は最大 25 倍のモデルと同等、またはそれを上回るパフォーマンスを発揮します。

コンパクトなサイズの Phi-2 は、機構の解釈可能性、安全性の向上など、言語モデルの研究開発を促進するために、Azure AI Studio モデルカタログ（MIT ライセンス）で利用できます。

■LLaVA とは

マイクロソフト、ウィスコンシン大学、コロンビア大学の研究者が公開したオープンソースのマルチモーダルな LLM です。meta 社が提供する「LLaMA」をベースにした大規模な言語モデルと画像分析機能を持つ視覚モデルであり、ScienceQA ベンチマークで最先端の精度を達成しています。

■NVIDIA® Jetson Orin™ Nano とは

NVIDIA® Jetson Orin™ Nano は、NVIDIA Ampere アーキテクチャーGPU を採用し、電力効率に優れた小型のフォームファクターで従来のエントリーレベル向けエッジ AI の常識を覆す性能を発揮します。

最大毎秒 40 兆回の演算性能を持ち、前世代の NVIDIA® Jetson Nano™ と比較して最大 80 倍のパフォーマンス向上を実現。これまで以上に複雑な AI モデルを活用できるようになります。

■今後について

ヘッドウォータースでは、SLM や MiniVLM といった小型化された言語モデルは生成 AI をあらゆるプラットフォームに適用させるために必要なテクノロジーであると考えております。

今後は、生成 AI×エッジ音声/画像解析、生成 AI×オンプレミス、TinyGPT-V 利用やモバイル VLM 推進、RAG システムに SLM を活用する「ハイブリッド RAG」、Databricks のデータ連携、NVIDIA 社の生成 AI アプリマイクロサービス「NIM」に関するソリューション展開を図ってまいります。

また、当社の掲げるアライアンス戦略では、顧客企業ともビジネスパートナーとなり共に生成 AI 経済圏を拡大する取り組みを行ってまいります。

なお、本件による当社の当期業績に与える影響は軽微であります。今後開示すべき事項が発生した場合には速やかにお知らせいたします。

■参考情報

Azure OpenAI Service Advanced パートナー認定について
https://www.headwaters.co.jp/news/azure_openai_service_advanced_partner.html

「Azure OpenAI Service ×音声」による企業向け GPT サービスラインナップについて
https://www.headwaters.co.jp/news/azure_openai_service_gpt_voice.html

NVIDIA Inception のパートナー企業に認定
https://www.headwaters.co.jp/news/nvidia_inception.html

NVIDIA の「Metropolis Partner Program」に参画
https://www.headwaters.co.jp/news/nvidiametropolis_partner_program.html

データブリックスの SI コンサルティングパートナーに認定
https://www.headwaters.co.jp/news/databricks_partner.html

■商標について

<会社情報>

会社名：株式会社ヘッドウォータース
所在地：〒163-1304　東京都新宿区西新宿 6-5-1 新宿アイランドタワー 4 階
代表者：代表取締役　篠田 庸介
設立　：2005 年 11 月
URL　：https://www.headwaters.co.jp

<本件のお問い合わせ>

株式会社ヘッドウォータース
メール：info@ml.headwaters.co.jp