生成 AI の基礎: OpenVINO™ を使用した LLM のデプロイ

この記事は、Medium に公開されている「Generative AI Fundamentals: Deploying LLMs with OpenVINO™」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

生成 AI の台頭とともに、大規模言語モデル (LLM) は、仕事から研究、電子商取引、エンターテインメントまで、あらゆるものに革命をもたらしています。少し前までは、サイズ、コスト、およびトレーニング、最適化、デプロイの複雑さにより、LLM と生成 AI で成功を収めることは困難でした。

ところが、この 1 年の大幅な進歩により状況は大きく変わり、これらのモデルを簡単に高速化してデプロイできるようになりました。

この記事では、生成 AI の基礎について説明し、OpenVINO™ ツールキットを使用してさまざまなプラットフォームでパフォーマンスと柔軟性を向上させる方法を紹介します。DevCon に登録 (英語) して、LLM をローカルで実行する詳細なコードを含むチュートリアル、AI エバンジェリストによる実例のデモを含む、ライブおよびオンデマンドのウェビナーを視聴できます。

OpenVINO™ で LLM の可能性を解き放つ

OpenVINO™ を使用することの素晴らしい点は、AI 開発者が「1 つのコードでどこでもデプロイできる」ことです。以前にトレーニングされたモデルを中間表現 (IR) 形式に変換して実際のタスクに合わせて最適化し、コードを変更することなく、通常の CPU から FPGA まで、あらゆるアーキテクチャーで推論を実行できます。

OpenVINO™ チームは、最新のテクノロジーのトレンドとイノベーションに対応するため、AI ツールキットを継続的に改良および更新しています。一例として、OpenVINO™ 2024.2 リリース (英語) では、処理の効率化、計算オーバーヘッドの削減、スループットの向上、レイテンシーの改善など、LLM のパフォーマンスが向上しています。

OpenVINO™ は、ニューラル・ネットワーク圧縮フレームワーク (NNCF) (英語) も活用して、重み圧縮、キー値キャッシュ、ステートフル変換により、IR モデルと推論プロセスを最適化します。多くの実際のシナリオでは複数の異なるモデルを接続する必要があるため、開発者は OpenVINO™ モデルサーバー (OVMS) (英語) を利用して、さまざまなアーキテクチャーに生成 AI モデルを簡単にデプロイできます。

単純な API 呼び出しを行うだけで OpenVINO™ で AI モデルを処理できる Hugging Face Optimum Intel ツール (英語) を利用すると、わずか 5 行のコードで、これらの手法を使用したモデルを実装してデプロイできます。

この点を実証するため、OpenVINO™ DevCon のビデオ (英語) では、チャットボットとイメージ・ジェネレーターで、大規模なモデルに匹敵する品質でユーザーのリクエストに応じて複雑なテキストを要約およびイメージを作成する方法、ローカル PC でわずか数秒で実行する方法を紹介しています。

「GenAI Fundamentals with OpenVINO (OpenVINO™ を使用した生成 AI の基礎)」ウェビナーでは、OpenVINO™ で 3 つの異なる AI モデル (埋め込み、ランキング、LLM) のパイプラインを最適化する方法を確認できます。

生成 AI LLM を実行する場所

巨大な汎用生成 AI モデルに代わる優れた代替手段の開発を目標にしたときに、自問すべき最初の質問は、「生成 AI サービスをどこで実行すべきか、またはどこで実行できるか」です。

クラウド・プラットフォームは、非常に大規模な生成 AI ワークロードを一元管理するための最も一般的な選択肢の 1 つですが、必ずしも最適なオプションであるとは限りません。例えば、機密データをクラウドに送信することは、多くのミッション・クリティカルなアプリケーションで常に許容されるとは限りません。また、異なるクラウド間でサブスクリプションを移行できないことが問題になるユーザーもいるでしょう。

エッジサーバーは、計算能力の高さよりもレイテンシーの低さを優先する小規模モデルや、分散処理ベースのサービスに適しています。

しかし、クラウドとエッジサーバーが唯一のソリューションであるという考えや、クラウドで実行する巨大な LLM のみが優れた結果をもたらすという考えは、時代遅れになっています。DevCon のビデオで詳しく説明しているように、OpenVINO™ などのツールを使用すると、低電力、高スループット、帯域幅、レイテンシーなど、あらゆる要件に応じて AI アプリケーションを最適化し、PC を含むさまざまな場所で実行できます。

これは、小型で強力な生成 AI サービスの需要の高まりが続く中の、素晴らしいニュースと言えます。

その理由は、オフラインでも常に利用でき、機密データをインターネットに公開しないサービスを実現することにより、生成 AI アプリケーションが個人の創造性と生産性を大幅に向上させるという約束を達成できるからです。さらに重要なのは、これらのサービスは、飛行機で移動中にメモを要約する、遠隔地で休暇の次の行動を計画するなど、各ユーザーが実際に必要とする少数の非常に具体的なタスクに合わせて、可能な限りカスタマイズする必要があることです。

ユビキタスな可用性、最小限のリソース消費、データ保護、カスタマイズ (最適化されたローカル推論が必要) に対するこれらのニーズから、将来、非常に特化した生成 AI モデルが必要とされる十分な余地があるという、明確な結論が導き出されます。これらのモデルは、トレーニングから最適化まで、特殊なタスク (特定の分野のテキスト要約など) 向けに特別に構築されていて、エッジサーバーはもちろん、ローカルの AI PC でも実行できます。

LLM をローカルで実行する方法、AI PC の新時代、モデル圧縮、その他の AI および生成 AI のトピックの詳細については、OpenVINO™ ワークショップ・シリーズ (英語) のライブおよびオンデマンドのウェビナーを視聴してください。

関連情報 (英語)

OpenVINO™ ドキュメント
Jupyter* Notebook
インストールとセットアップ
製品ページ

OpenVINO™ ツールキットとは

AI を加速する無償のツールである OpenVINO™ ツールキットは、インテルが無償で提供しているインテル製の CPU や GPU、VPU、FPGA などのパフォーマンスを最大限に活用して、コンピューター・ビジョン、画像関係をはじめ、自然言語処理や音声処理など、幅広いディープラーニング・モデルで推論を最適化し高速化する推論エンジン/ツールスイートです。

OpenVINO™ ツールキット・ページでは、ツールの概要、利用方法、導入事例、トレーニング、ツール・ダウンロードまでさまざまな情報を提供しています。ぜひ特設サイトにアクセスしてみてください。

https://www.intel.co.jp/content/www/jp/ja/internet-of-things/openvino-toolkit.html