OpenVINO™ 2024.1 の概要: LLM のパフォーマンスの強化とサポートの拡大で生成 AI ワークロードを活用

この記事は、Medium に公開されている「Introducing OpenVINO™ 2024.1: Enable Your Generative AI Workloads with Enhanced LLM Performance and Broadened Support」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。

OpenVINO™ ツールキットの最新リリースへようこそ。OpenVINO™ ツールキットは、新しいリリースごとに可能性の裾野を広げるだけでなく、進化し続ける AI 推論とデプロイの分野においても、確固たる地位を示しています。この記事では、OpenVINO™ 2024.1 の最新機能を紹介します。世界中の開発者に強化されたパフォーマンス、柔軟性、使いやすさを提供するという我々の取り組みは、このリリースでも継続しています。

OpenVINO™ 2024.1 のリリースは、コミュニティーからの強力なフィードバックと、開発者に大規模言語モデル(LLM)のパフォーマンスの向上、モデルサポートの拡張、クラウドやローカルのさまざまなプラットフォームでの合理化されたデプロイを提供するという明確なビジョンにより形成されています。このリリースの最も重要なアップデートについて見ていきましょう。

大規模言語モデルの推論の向上

LLM の開発は、依然として驚異的な速度で進んでいます。LLM の強力な機能はパフォーマンスの向上によりさらに強化されています。その一方で、OpenVINO™ の最適化と推論の高速化は、これらの複雑なモデルの実行を改善します。高速で効率的な処理、計算オーバーヘッドの削減、ハードウェアの潜在能力の最大化が可能になり、LLM のスループットの向上と低いレイテンシーがもたらされます。

圧縮された埋め込みによる追加の最適化により、LLM のコンパイル時間とメモリー・フットプリントが削減されます。LLM の最初のトークンのパフォーマンスが、インテル® アドバンスト・マトリクス・エクステンション (インテル® AMX) を搭載した第 4 世代および第 5 世代のインテル® Xeon® スケーラブル・プロセッサーとインテル® Arc™ GPU で向上しました。

oneDNN により、LLM の圧縮が改善され、パフォーマンスが向上しました。現在、量子化または圧縮後の INT4 および INT8 精度の LLM がインテル® Arc™ GPU でサポートされています。統合 GPU を搭載したインテル® Core™ Ultra プロセッサーでは、一部の小さな生成 AI モデルでメモリーが大幅に削減されます。

また、トレーニング後の量子化後に INT8 PyTorch* モデルに微調整を適用して、モデルの精度を向上し、トレーニング後の量子化からトレーニングを考慮した量子化へ簡単に移行できるようになりました。詳細は、追加されたサンプル (英語) を参照してください。

より多くの生成 AI カバレッジとフレームワークの統合

OpenVINO™ を使用した生成 AI の分野をさらに深く掘り下げました。この新しいリリースで、OpenVINO™ は生成 AI の範囲を広げ、幅広いニューラル・アーキテクチャーとアプリケーションをカバーするようになりました。

新しくリリースされた最先端の Llama 3 モデル (英語) と Phi-3 (英語) は、OpenVINO™ でサポートされ最適化されています。Mixture of Expert (MoE) LLM アーキテクチャーを備えた Mixtral、および URLNet モデルは、インテル® Xeon® プロセッサー上でパフォーマンスが向上するように最適化されています。テキストから画像を生成するモデルの Stable Diffusion 1.5 と、LLM の ChatGLM3–6b および Qwen-7B モデルは、統合 GPU を搭載したインテル® Core™ Ultra プロセッサー上で推論速度が向上するように最適化されています。

優れたパフォーマンス・メトリックを備えた、すぐに使えるチャット/Instruct (指示) モデルの生成 AI LLM、Falcon-7B-Instruct がサポートされ、OpenVINO™ で利用できるようになりました。

このほか、YOLOv9、YOLOv8 指向性バウンディング・ボックス検出 (OOB)、Keras での Stable Diffusion、MoblieCLIP、RMBG-v1.4 バックグラウンド除去、Magika、TripoSR、AnimateAnyone、LLaVA-Next、OpenVINO™ と LangChain を使用した RAG システムなどのモデルもサポートされました。OpenVINO™ ノートブック・リポジトリー (英語) では、Jupyter* Notebook のサンプルも提供しています。

新しいプラットフォームの変更と既存のプラットフォームの強化

インテル® Core™ Ultra プロセッサー向けのプレビュー NPU プラグイン (英語) が、PyPI のメイン OpenVINO™ パッケージに加えて、OpenVINO™ オープンソース GitHub* リポジトリーでも利用できるようになりました。

JavaScript* API (英語) が npm リポジトリーからさらに簡単にアクセスできるようになりました。JavaScript* 開発者は OpenVINO™ API にシームレスにアクセスできます。JavaScript* アプリケーションと OpenVINO™ の統合を始める開発者を支援するため、ドキュメントが拡張されました。

ARM プロセッサー上の FP16 推論が、畳み込みニューラル・ネットワーク (CNN) でデフォルトで有効になりました。ARM デバイス上の広範なモデルのパフォーマンスが大幅に向上しました。FP16 精度の推論が、ARM デバイス上のすべてのタイプのモデルのデフォルトになりました。CPU アーキテクチャーに依存しないビルドが実装され、異なる ARM デバイスで統一されたバイナリーを配布できるようになりました。

新しいノートブックと変更されたノートブック

OpenVINO™ ノートブック(英語)は、依然として AI 分野の最も重要な進歩における OpenVINO™ の活用を示す貴重なリソースです。最近、OpenVINO™ ノートブック・リポジトリーに、デフォルトのブランチの 'main' から 'latest' への変更、「notebooks」フォルダー内のノートブックの命名構造の改善など、いくつかの変更を加えました。

コンテンツを確認するには、ローカルの README.md(英語)ファイルおよび GitHub* ページの OpenVINO™ ノートブック(英語)を使用してください。

次のノートブックが更新または新しく追加されました。

- Grounded Segment Anything(英語)
- MobileCLIP を使用したビジュアルコンテンツ検索(英語)
- YOLOv9 の最適化(英語)
- YOLOv8 指向性バウンディング・ボックス検出の最適化(英語)
- Magika: AI を活用した高速で効率的なファイルタイプの識別(英語)
- Keras での Stable Diffusion(英語)
- RMBG バックグラウンド除去(英語)
- AnimateAnyone: ポーズガイド付き画像からビデオを生成(英語)
- LLaVA-Next ビジュアル言語アシスタント(英語)
- TripoSR: 単一画像から 3D 画像を再構築(英語)
- OpenVINO™ と LangChain を使用して RAG システムを作成(英語)
- こんにちは、NPU!(英語)

貢献者の皆さんに感謝します!

OpenVINO™ で達成された最新のマイルストーンを祝うにあたり、貢献者の皆さんの努力こそ称賛されるべきです。皆さんの貴重な貢献は、OpenVINO™ ツールキットを強化しただけでなく、コミュニティーを発展させて、革新とコラボレーションの環境を育成しました。コードの提出やコミュニティーでの活発なアイデアの交換など、貢献していただいたすべての方々に心より感謝します。

この先も、皆さんと一緒にこの旅を続けることをこれまで以上に楽しみにしています。経験豊富な開発者も新しい開発者も、引き続き貢献していただけることを願っています。皆さんの独自の視点と革新的なアイデアにより、OpenVINO™ は、開発者が AI のビジョンを現実のものにする優れたツールとして進化を続けるでしょう。

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