OpenVINO™ 2024.0 の概要: パフォーマンスの強化とサポートの拡張で開発者を支援

この記事は、Medium に公開されている「Introducing OpenVINO 2024.0: Empowering Developers with Enhanced Performance and Expanded 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™ 2024.0 へようこそ。このリリースでは、動的量子化、GPU 最適化の向上、混合エキスパート・アーキテクチャーのサポートによる大規模言語モデル (LLM) のパフォーマンスの強化など、急速に進化する AI 環境で開発者を支援することを目的とした多くの機能が追加されました。OpenVINO™ 2024.0 (英語) により、開発者は AI アクセラレーションを効果的に活用できます。コミュニティーの継続的な貢献に感謝の意を表しつつご利用ください。

大規模言語モデルの推論の向上

LLM は、廃れることなく、そのモデルとユースケースは進化を続けています。我々は、モデルを高速化し、モデルの推論を手頃な価格で提供するという使命に継続して取り組んでいます。

パフォーマンスと精度の強化

このリリースでは、LLM のデフォルトのパフォーマンスの向上に取り組み、ランタイムとツールにいくつかの重要な変更を加えました。

まず、CPU プラットフォーム向けに動的量子化と KV キャッシュ圧縮のメカニズムを導入しました。KV キャッシュ圧縮機能により、大規模なシーケンスの生成をリソース効率良く、高いパフォーマンスで実行できます。動

的量子化は、一般に、モデルのほかの部分 (投影やフィードフォワード・ネットワークの埋め込み) の計算とメモリーの消費量を改善します。このアップデートにより、インテル® Xeon® Platinum 8490H プロセッサーでの生成レイテンシーが 4.5 倍改善され、mistral-7b-v0 では 28.2 トークン/秒を達成しました。[1]

GPU プラットフォームでは、カーネルおよびスタックを最適化することにより、生成特性も改善しました。また、ビーム検索を使用した生成を支援する、効率良いキャッシュ処理も実装しました。例えば、インテル® Core™ Ultra 7 プロセッサー 165H iGPU の llama-2-7b-chat INT4 精度モデルで最大 8.5 トークン/秒を達成しました。[1]

パフォーマンスは常に議論の対象となるトピックですが、精度も重要です。このリリースでは、NNCF 内の重み圧縮のアルゴリズムの精度を向上しました。データセットからの統計を使用して重みを圧縮する機能を導入し、精度を向上するために AWQ アルゴリズムを実装しました。

さらに、Hugging Face Optimum Intel との統合により、次のように Transformers API を利用してモデルを直接圧縮できるようになりました。

```
import nncf
from transformers import AutoTokenizer
from optimum.intel.openvino import OVModelForCausalLM

MODEL_ID = "databricks/dolly-v2-3b"
tokenizer = AutoTokenizer.from_pretrained(MODEL_ID)

model = OVModelForCausalLM.from_pretrained(
    MODEL_ID,
    export=True,
    load_in_4bit=True,
    quantization_config={"dataset": "ptb", "mode": nncf.CompressWeightsMode.INT4_ASYM, "ratio": 0.8},
)
```

コードのソース: https://github.com/huggingface/optimum-intel/pull/538 (英語)

注: **load_in_4bit** オプションを True に設定して from_pretrained メソッドの呼び出し内で **quantization_config** を直接渡す機能を使用することにより、すべての圧縮作業が行われます。さらに、Llama2、StableLM、ChatGLM、QWEN などを含む、人気の高いモデルの量子化構成も追加しました。これらのモデルでは、4 ビット圧縮を取得するための構成を渡す必要はありません。

アルゴリズムの品質の詳細は、OpenVINO™ ドキュメント (英語) または GitHub* の NNCF ドキュメント (英語) を参照してください。

混合エキスパート・アーキテクチャーのサポート

混合エキスパート (MoE) は、LLM の精度とパフォーマンスの向上をもたらす、次の主要なアーキテクチャーの進化を示します。Mixtral モデルから始まり、既存のモデルから MoE ベースのモデルを作成できるように、多くのモデルやフレームワークで急速に進化しました。

2024.0 リリースでは、これらのアーキテクチャーを有効にしてパフォーマンスを向上することに取り組みました。これらのモデルの効率良い変換に取り組んだだけでなく、ランタイム内でのエキスパートの動的な選択を適切に処理するために内部の動作を一部変更しました。

我々は、Hugging Face Optimum Intel への変更をアップストリームしており、これらのモデルの変換について深く理解しています。

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

インテルの NPU へのアクセス

インテル® Core™ Ultra プロセッサーのリリースにより、多くの開発者が NPU アクセラレーターを利用できるようになりました。NPU はソフトウェアとハードウェアのどちらから見ても進化中の製品であり、さまざまな機能を実現できます。すでに NPU 上で動作する OpenVINO™ ノートブック (英語) のデモを視聴した方もいるでしょう。

このリリースでは、人気の高いディストリビューション・チャネルである PyPI を利用して OpenVINO™ をインストールするときに NPU サポートを利用できるようにしています。いくつか注意すべき点があります。

- NPU を使用するには、システムにドライバーをインストールする必要があります。NPU を使用する場合は、必ずこの説明 (英語) に従ってください。
- 現在、NPU は自動デバイス選択 (英語) ロジックに含まれていないため、NPU でモデルを実行する場合は、次のようにデバイス名 (NPU など) を明示的に指定していることを確認してください。

```
compiled_model = core.compile_model(model=model, device_name="NPU")
```

ARM CPU のサポートの改善

スレッド化は、ARM プラットフォームで効率良く実装できていなかった機能の 1 つであり、パフォーマンス向上の妨げとなっていました。我々は oneTBB チーム (デフォルトのスレッド化エンジン・プロバイダー) と協力して ARM のサポートを変更し、パフォーマンスを大幅に向上することに成功しました。同時に、特定の演算の精度に関する作業を行った後、ARM CPU でのデフォルトの推論精度を fp16 にしました。

全体として、これは ARM CPU のパフォーマンスが向上することを意味しますが、マルチコア・プラットフォームで高いスループットを得ることができる OpenVINO™ ストリーム (英語) 機能の可用性も意味します。

古い機能の削除

2024.0 は次のメジャーリリースであり、ツールキットから古いコンポーネントを削除する適切な時機でもあります。

2 年前、我々はディープラーニング分野の進化に対応するため、API を大幅に変更しましたが、OpenVINO™ を使用する既存の開発者や製品への影響を最小限に抑えるために、API 1.0 をサポートしていました。その後、多くのことが変更され、現在では古い API を完全に削除しています。また、非推奨としてマークした次のようなツールも削除しています。

- トレーニング後の最適化ツール (POT)
- 精度チェック・フレームワーク
- デプロイメント・マネージャー

これらのツールは openvino-dev パッケージの一部であり、しばらくの間、このパッケージを使用する必要はありませんでした。オフラインモデル変換ツールのモデル・オプティマイザーを引き続き使用するユーザー向けには保持する予定です。

新しい API に移行できなかった場合でも、LTS リリースのいずれか (2023.3 など) を引き続き使用できる可能性が高くなります。

新しいノートブックと変更されたノートブック

AI 分野における最も重要なアップデートと、OpenVINO™ を活用してそれらのシナリオを進める方法を引き続き紹介します。我々は、次のようなことに取り組んできました。

- MobileVLM を使用したモバイル言語アシスタント (英語)
- DepthAnything を使用した深度の推定 (英語)
- マルチモーダル大規模言語モデル (MLLM) Kosmos-2 (英語)
- SigLIP を使用したゼロショット画像分類 (英語)
- PhotoMaker を使用したパーソナライズされた画像の生成 (英語)
- OpenVoice を使用した音声トーンのクローンの作成 (英語)
- Surya を使用した行レベルのテキスト検出 (英語)
- InstantID を使用したゼロショットのアイデンティティー保存の生成 (英語)
- LLM チャットボット (英語) と LLM RAG パイプライン (英語) は、新しいモデルの統合 (minicpm-2b-dpo、gemma-7b-it、qwen1.5–7b-chat、baichuan2–7b-chat) により更新されました。

開発者と貢献者の皆さんに感謝します!

我々は、物体検出から面接準備ツールまで、OpenVINO™ の歴史の中で多くのエキサイティングなプロジェクトを見てきました。そこで、OpenVINO™ を使用した素晴らしいプロジェクトのリスト (英語) をまとめることにしました。OpenVINO™ は現在も急速に成長し続けています。プロジェクトでプルリクエストを作成し、プロジェクトに「Mentioned in Awesome」バッジを使用して、その良さを共有してください。

我々の開発者ベースは拡大しており、コミュニティーが行っているすべての変更と改善に感謝しています。「OpenVINO™ の改善を支援するのに忙しい」と仰る方も目にしました。驚くとともに深く感謝しています。

貢献者によって行われた作業の一例は、openSUSE* プラットフォームでの OpenVINO™ サポート (英語) です。

この数週間我々は、Good First Issues の更新やプルリクエストのレビューを迅速に行えていません。我々はこの問題を認識しており、修正を行っています。続報をお待ちください。

また、準備を進めている Google Summer of Code (英語) プロジェクトに関して、皆さんから非常に興味深い提案をいただいています。皆さんのアイデアをご提案いただく時間はまだあります。

このリリースの貢献者のリストは、GitHub* (英語) の Acknowledgements セクションに記載されています。

[1] 結果は異なることがあります。システム構成はこちらを、ワークロードの説明はこちらを参照してください。法務情報はこちら (英語) を参照してください。

ベンチマークのシステム構成とワークロードの説明

システム構成

CPU 推論エンジン	インテル® Xeon® Platinum 8490H	MTL-H
マザーボード	Intel Corporation / Archer City	Intel Corporation CRB (Reef Ridge + Astral peak)
CPU	インテル® Xeon® Platinum 8490H プロセッサー @ 1.90GHz	インテル® Core™ Ultra 7 プロセッサー 165H
ハイパー スレッディング	有効	有効
ターボブースト	有効	有効
メモリー	16x16GB DDR5 4800MHz	2x16GB DDR5 5600MHz
オペレーティング ・システム	Ubuntu* 22.04.2 LTS	Windows* 11
カーネル バージョン	6.2.0-36-generic	10.0.22631 Build 22631
BIOS ベンダー	Intel Corporation	Intel Corporation
BIOS バージョン	EGSDREL1.SYS.9409.P31.2302280828	MTLPEMI1.R00.3323.D53.2310240712
BIOS リリース	2/28/2023	10/24/2023
バッチサイズ	1	1
精度	INT4	INT4
テスト日	2/27/2024	2/27/2024

表 1. ベンチマークのシステム構成

ワークロードの説明

ワークロードのパラメーターはモデルのパフォーマンス結果に影響を与えます。モデルはバッチサイズ 1 を使用して実行されます。生成 AI モデルのパラメーターを次に示します。

- 入力トークン: 1024
- 出力トークン: 128
- ビーム数: 1
- 生成 AI モデルのトークンは英語

使用したベンチマーク・アプリケーションの GitHub* リポジトリー:
https://github.com/openvinotoolkit/openvino.genai/tree/master/llm_bench/python (英語)

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