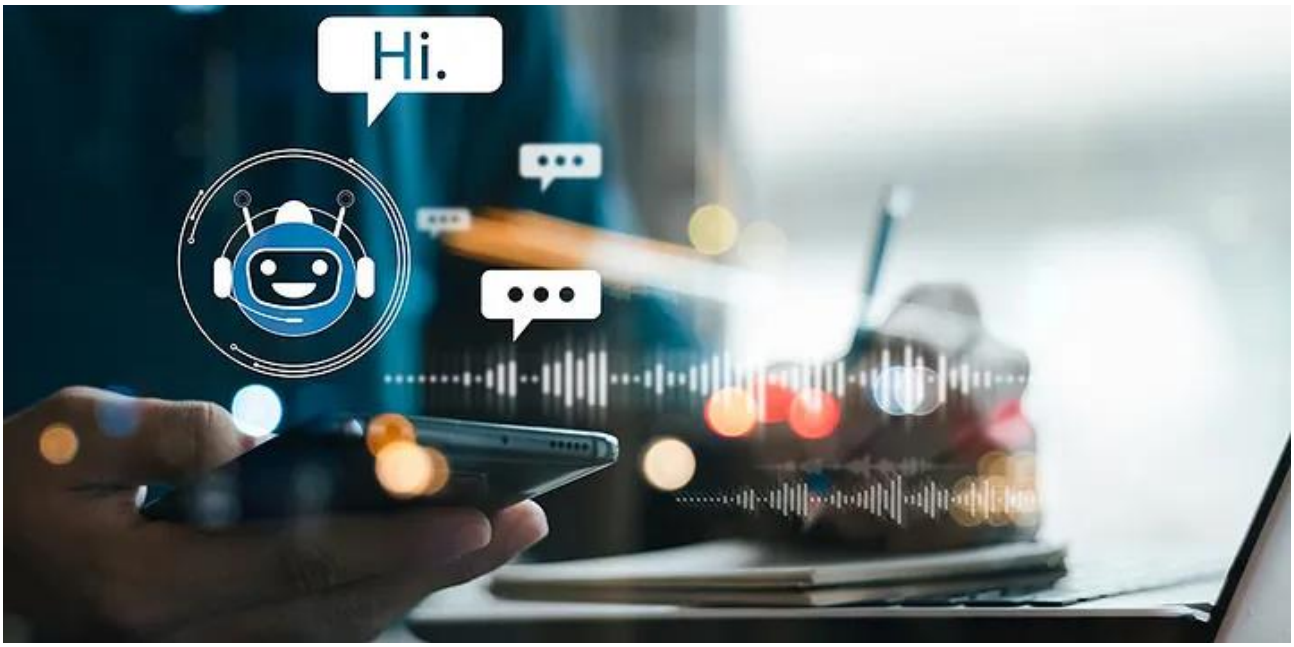
正しい手法を使用して優れた生成 AI アプリケーションを作成する

この記事は、Medium に公開されている「Build Better GenAI Applications with the Right Techniques」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。

著者: Ria Cheruvu インテル コーポレーション AI ソフトウェア・アーキテクト & エバンジェリスト



生成 AI は、現在の AI テクノロジーの中で最も急速に進歩している分野です。生成 AI を初めて使用する場合や従来のマシンラーニングから移行する場合、生成 AI を使用することが困難に思えるでしょう。

AI ソフトウェア・アーキテクトとしての私の役割は、派手に宣伝することも含め、現在の研究トレンドを中心として全体像を考えることです。

多くの業界が大規模言語モデル (LLM) と生成 AI ツールの使用を検討しているため、モデルだけでなく、これらの生成 AI エクスペリエンスを高める手法を、機能、制限、可能性とともに考慮することが重要です。

これらのアプリケーションの背後にある重要な要素は、**専門化**、**コンテキスト化**、**マルチモダリティー**などの手法をトレーニング・パイプラインに導入するための適切なデータの必要性です。それぞれについて見てみましょう。

独自の特殊な生成 AI のトレーニング

言語ベースの AI プロジェクトを開始するユーザーは通常、GPT-4、Llama 2、Mistral 7B、ChatGPT* を含む、いくつかの事前トレーニング済みの高度な LLM から選択します。それぞれ長所と短所がありますが、(少なくとも基本的な形式に) 共通する特徴の 1 つは、言語機能を提供するが焦点や具体性に乏しい、広範で一般的なデータセットに基づいて意図的にトレーニングされていることです。この種のモデルは基礎モデルと呼ばれます。基礎モデルは複数のタスクを実行できる大規模な AI モデル (英語) であり、幅広い下流アプリケーションで有益です。

ただし、独自のモデルをトレーニングすると、次のような大きな利点を得ることができます。

1. **データのプライバシー**: 機密情報や専有情報を第三者に公開しないようにできます。
2. **パフォーマンスの向上**: 特定のタスクに最適化して、低コストで良い結果を得ることができます。
3. **コンテンツの制御**: 特定の値または標準に合わせてモデルをトレーニングできます。
4. **バイアスの制限**: 優れた公平性と中立性が実現するようにトレーニング・データセットを整理できます。

独自のモデルをゼロからトレーニングする場合の短所は、多大な労力と、かなりの専門知識が必要になることです。そのため、生成 AI モデル最適化の将来の課題として微調整が注目されています。

人間と機械のフィードバックを利用した生成 AI の微調整

微調整は、事前トレーニング済みモデルを開始点として使用し、新しい特定のトレーニング・データセット向けに調整を行います。この一連の手法により、開発時間が大幅に短縮され、コストも削減されます。トレーニング・プロセスの大部分を短縮できるため、同じモデルを最初からトレーニングするよりも LLM を微調整するほうがコストが桁違いに安くなる可能性があります。

多くの LLM があるように、微調整にも多くのアプローチがあります。これらのアプローチにはすべて、事前トレーニング済みモデルを新しいデータセットに公開することと、次のことが含まれます。

1. **再利用**: 関連するタスクにモデルを適応させます。
2. **完全なモデルの微調整**: すべてのパラメーターを調整して、新しく大幅に異なるタスクを実行します。
3. **命令の微調整**: 特定のガイドラインに従うようにモデルをトレーニングして、動作を制限します。
4. **教師あり微調整**: ラベル付きデータセットを使用して、目的の結果が明確に定義されるようにタスクを最適化します。
5. **人間のフィードバックによる強化学習 (RLHF)**: 人間による評価を使用して、複雑なタスクに対する微妙なフィードバックを提供します。
6. **パラメーター効率の良い微調整 (LoRA など)**: モデルの一部のみ調整します。大規模モデルの調整の課題を克服するのに役立ちます。

RLHF は、人間のような論法と意思決定を行う能力により、最も注目を集めています。例えば、OpenAI のチームは、パラメーターが 100 分の 1 以上少ないにもかかわらず (英語)、GPT-3 モデルからの出力よりも RLHF モデルからの出力の方がユーザーに好まれていると述べています。RLHF の短所は、大量の人的リソースと計算リソースが必要なことです。そこで、効率を重視して設計された、新しく特殊な手法を用いた LoRA を考えてみました。

LoRA は、モデルのトランスフォーマー・アテンションとフィードフォワード・ブロックに注目しています。モデルの重みを調整するほかの微調整手法とは異なり、LoRA はこれらの値を固定し、代わりに追加のトレーニング可能なレイヤーを挿入します。これらの追加のレイヤーのトレーニングに必要な計算は非常に少なくなりますが、その結果は完全なモデルの微調整に匹敵します。我々のチームは最近、Stable Diffusion + ControlNet と OpenVINO™ 最適化を組み合わせて異なるスタイルのイメージを生成するパイプラインを利用して LoRA の可能性を実証 (英語) しました。

Stable Diffusion v1.5 + ランタイム LoRA Safetensors の重み + ControlNet

微調整における最適化と意思決定

最適化はコストだけでなく生成 AI の柔軟性も決定するため、重要な考慮事項です。モデルのパラメーターの精度 (INT8、FP16、FP32 など) を最適化することにより、モデルの速度、メモリー・フットプリント、スケーラビリティーを大幅に向上することができます。

LoRA では必要なトレーニングが大幅に限定されますが、どのパラメーターを固定するかという課題も生じます。新しい API と、Hugging Face のライブラリーのような抽象化は、開発者に「既製の」最適化のパスを提供します。インテルは同社と協力し、OpenVINO™ を使用して Hugging Face のモデルを最適化 (英語) することにより AI の普及を進めてきました。OpenVINO™ を使用することにより、開発者は最適化済みライブラリーを活用して、インテル® Arc™ グラフィックスを搭載したシステムでモデルをトレーニング (英語) したり、インテル® Xeon® プロセッサー上のクラウドでモデルをトレーニングすることができます。

マルチモーダル・アプローチ

生成 AI に対して行われるもう 1 つの大きな変更は、複数のデータソースへの移行です。この様子は、ChatGPT* などの LLM のユーザーに馴染み深いマルチモーダル機能に見ることができます。ここでは、テキストベースの機能が、イメージやサウンドなどのほかのデータ型を取り込む機能により補完されます。

現在の注目はデータ表現に移っており、異なる形式を単一のデータセットに統合することを目標としています。これにより、モデルを多様なデータ形式を同時に処理できるようになり、アシスタントとして機能する、洗練された有能な AI システムが実現します。

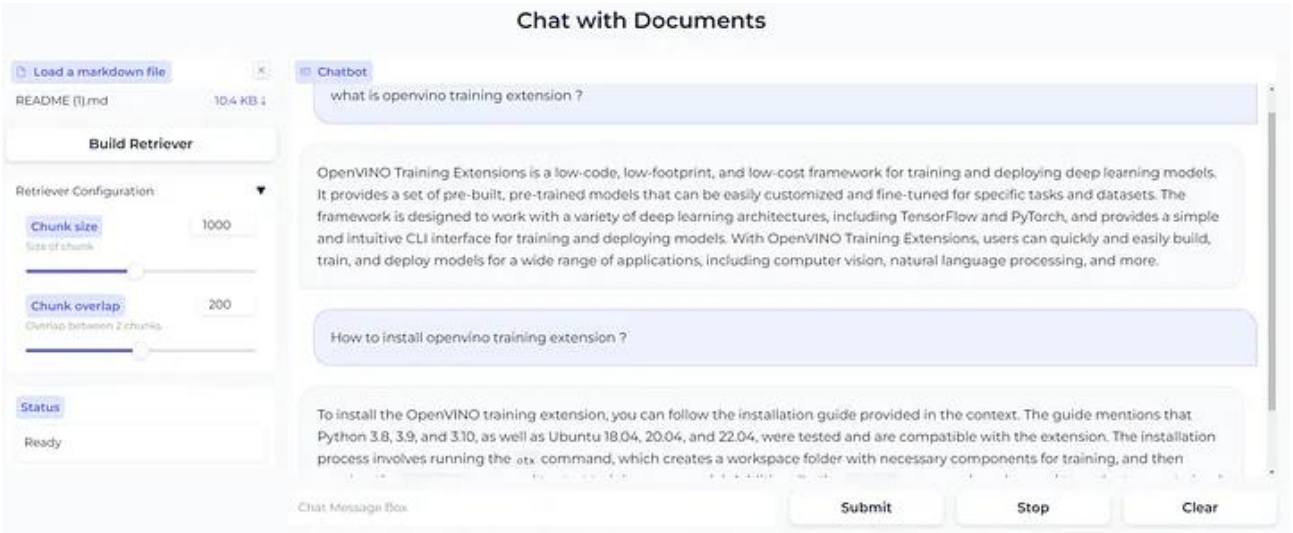
マルチモーダル・モデルの課題の 1 つは、新しいデータ構造の導入がパフォーマンスと精度に影響を与える可能性があることですが、OpenVINO™ を使用すると、開発者はビジュアルデータやその他の複雑なデータの推論とベンチマークを簡単に高速化できます。

例えば、我々は最近、ビジュアルとイメージの入力を受け付けるマルチモーダル・システムの LLaVa と OpenVINO™ を使用した仮想アシスタントの作成を検討 (英語) しました。OpenVINO™ NNCF を使用してモデルの重みを (4 ビットと 8 ビットに) 圧縮した後、インタラクティブな仮想アシスタントで推論を実行し、イメージに関する質問を行いました。

コンテキスト内学習

モデルで複数のデータソースを利用するもう 1 つの方法は、コンテキスト内学習を使用することです。この手法は、LLM とデータベースまたはほかのデータ・リポジトリーを組み合わせます。このアプローチではモデルそのものは変更されません。代わりに、リポジトリーからのデータをユーザーのクエリーに追加し、応答に適したコンテキストを LLM に提供します。これは微調整を補完するものであり、LoRA などの手法と組み合わせることができます。

検索拡張生成を使用してアップロードしたドキュメントからテキストを検索し、インタラクティブなインターフェイスで質問に効率良く回答する LLM の例として、OpenVINO™ ノートブック・リポジトリーの llm-chatbot (英語) を次に示します。



Mistral 7B や Zephyr 7B などの大規模言語モデルと検索拡張生成を使用

生成 AI の未来を加速

生成 AI 革命は、モデルのトレーニングと調整手法の急速な進化と、異なる AI 分野の融合を推進しています。業界がこれらの進歩をどのように利用して新しいレベルのインテリジェンスを可能にするのか、見ることができるのを楽しみにしています。

生成 AI の旅を始めるには、新しい生成 AI アプリケーションを含む、OpenVINO™ ノートブック (英語) をチェックすることを推奨します。可能性は無限です。この記事が皆さんのアイデアを実現するのに役立つことを願っています。

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