インテル® CPU における高速化された AI の新時代の到来

Numenta は CPU 上での AI 推論を高速化する神経科学ベースのソリューションでインテルと協力

Charmaine Lai Numenta マーケティング・マネージャー

問題ステートメント

2024 年を迎えて、さまざまな業界の企業が大規模言語モデル（LLM）の可能性を目の当たりにしています。しかし、主にコスト、複雑さ、高い消費電力、データ保護への懸念により、企業は LLM を自社の業務に統合することに苦労しています。GPU はこれまで LLM にとって主要なハードウェアでしたが、IT の複雑さとコストの増加により、成長する運用を維持することは容易ではありません。さらに、世界的な需要の増加と供給不足のため、企業は現在、AI プロジェクト向けの GPU を確保するのに 1 年以上待つこともあります。

新しいアプローチ

そこで、Numenta とインテルは、GPU への依存から企業を解放する、LLM 導入の新しいアプローチを作り出しました。Numenta Platform for Intelligent Computing（NuPIC*）は、神経科学に基づく概念をインテル® アドバンスト・マトリクス・エクステンション（インテル® AMX）命令セットにマップします。NuPIC* を使用すると、インテル® CPU 上に LLM を大規模にデプロイして、コストとパフォーマンスの劇的な向上を実現できます（第 4 世代インテル® Xeon® スケーラブル・プロセッサーのパフォーマンス・インデックス（英語）のベンチマーク P6 および P11 を参照）。

テクノロジー

インテル® AMX

インテル ® AMX は、専用のハードウェア・サポートを提供することにより行列演算を高速化します（**図 1**）。このテクノロジーは、ディープラーニング推論など、行列計算に大きく依存するアプリケーションでは特に有益です。インテル® AMX は、第 4 世代インテル® Xeon® プロセッサーで最初に導入され、新しい第 5 世代インテル® Xeon® プロセッサーでさらに高速化されました。

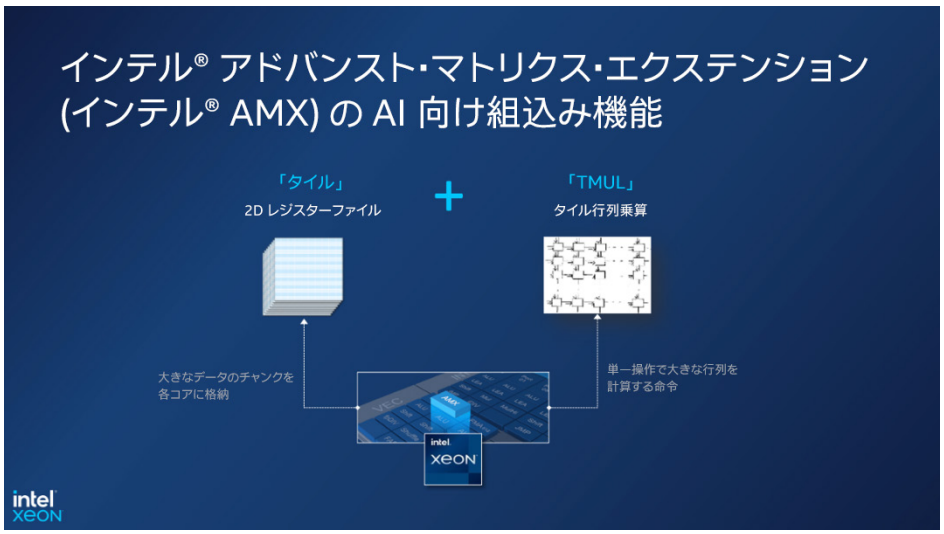

図 1. インテル® AMX 命令セットには、タイルと TMUL という 2 つの主要なコンポーネントが含まれています。タイルは、大きな行列の並列処理を可能にする 2D ユニットです。TMUL は、わずか 16 クロックサイクルで 16x16x32 の行列乗算を実行できます。

NuPIC*

NuPIC* は数十年にわたる神経科学研究に基づいて構築されており、強力な言語ベースのアプリケーションを迅速かつ容易に構築できます（**図 2**）。NuPIC* の中心となるのは、インテル® CPU 上で LLM を効率的に実行できるようにする、インテル® AMX を活用して高度に最適化された推論サーバーです。さまざまな自然言語処理のユースケースに合わせてカスタマイズできる、実用的な事前トレーニング済みモデルのライブラリーを使用して、NuPIC* 推論サーバーでモデルを直接実行できます。NuPIC* トレーニング・モジュールを使用してユーザーデータでモデルを調整し、そうしたカスタムモデルを NuPIC* モデル・ライブラリーにデプロイして推論サーバーで実行することもできます。

標準の推論プロトコルをベースに構築された NuPIC* は、標準の MLOps パイプラインにシームレスに統合できます。このソリューションは Docker* コンテナとしてデプロイされ、スケーラブル、セキュア、高可用性、ハイパフォーマンスの環境としてインフラストラクチャー内に保持できます。カスタムモデルは完全に制御できます。データは完全にプライベートのままで、システムの外にエクスポートする必要はありません。

インテル® Xeon® CPU 上の NuPIC* が AI 推論に最適な理由

Numenta とインテルは、この物語の新たな章を開き、非常にコスト効率に優れた方法で LLM を CPU 上に大規模にデプロイできるようにしました。その理由をいくつか紹介しましょう。

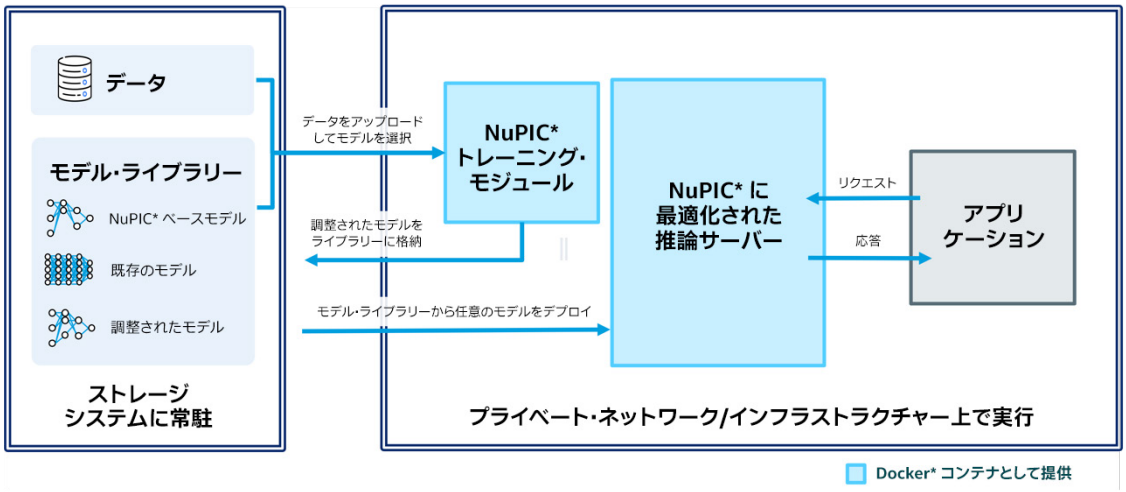

図 2. NuPIC* アーキテクチャー

パフォーマンス：NVIDIA* A100 Tensor Core GPU より 17 倍高速

Transformer 構造への最小限の変更により、NuPIC* は、インテル® AMX 対応の CPU で、前世代の CPU と比較して推論スループットを 2 桁以上向上し、GPU と比較しても大幅な速度向上を実現します(**表 1**)。BERT-Large で、第 5 世代インテル® Xeon® CPU 上の NuPIC* は、NVIDIA* A100 GPU よりも最大でおよそ 17 倍優れています。GPU で最高の並列パフォーマンスを実現するには、より大きなバッチサイズが必要です。しかし、バッチ処理は推論の実装が複雑になり、リアルタイム・アプリケーションでレイテンシーが生じます。対照的に、NuPIC* ではバッチ処理が必要ないため、アプリケーションは柔軟で、スケーラブルで、管理が容易になります。参考として、バッチサイズ 8 の NVIDIA* A100 のパフォーマンスをリストします。バッチサイズ 1 の NuPIC* は、このバッチサイズ 8 の NVIDIA* GPU 実装と比較した場合でも 2 倍以上優れています。

シーケンスの長さ	バッチサイズ	BERT-Large での 1 秒あたりの推論数		スピードアップ
		NuPIC* 第 5 世代インテル® Xeon®	Nvidia* A100	
64	1	2891	166**	17.4x
128	1	901	128	7.0x
128	8	-	495	2.2x

表 1. 第 5 世代インテル® Xeon® CPU 上の NuPIC* と NVIDIA* A100 の推論の結果。インテル® Xeon® CPU の結果は、シングルノード、2 ソケットのインテル® Xeon® Platinum 8592+ プロセッサー、メモリー 500GB のシステムで、Ubuntu* 22.04 カーネル 5.15.0-87、Numenta Platform for Intelligent Computing V1.0、NuPIC* に最適化された BERT-Large、シーケンス長 64/128、BF16、バッチサイズ 1 を使用して、2023 年 11 月 27 日に Numenta により生成されました。NVIDIA* A100 の結果は、こちら（英語）。 は推定。**

スケーラビリティー：複数のクライアントの管理

ほとんどの LLM アプリケーションは多くのクライアントをサポートする必要があり、それぞれ推論結果が生成されます。単一の第 5 世代インテル® Xeon® プロセッサー・ベースのサーバー上の NuPIC* で、アプリケーションは高いスループットを維持しながら、多くのクライアントのリクエストを処理できます（**図 3**）。クライアントが入力をバッチ処理する必要はなく、各クライアントを完全に非同期にできます。

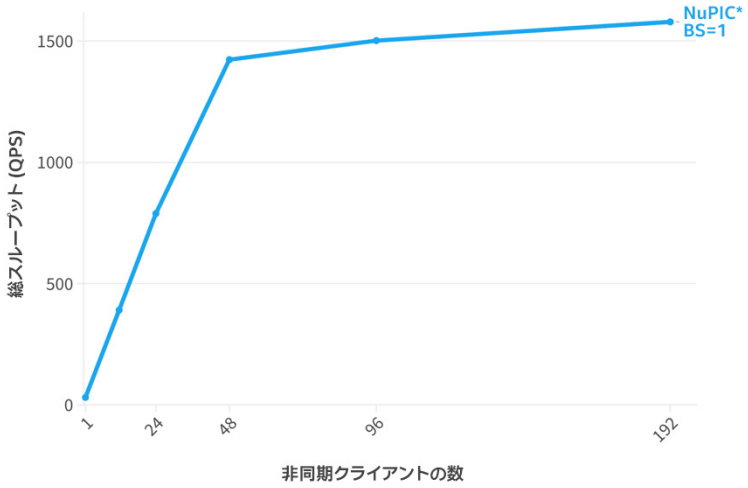

図 3. 非同期クライアントの数が増加した場合の、128 の独立した NuPIC* に最適化された BERT-Large モデルの総スループット。結果は、シングルノード、2 ソケットのインテル® Xeon® Platinum 9592+ プロセッサー、メモリー 500GB のシステムで、Ubuntu* 22.04 カーネル 5.15.0-87、Numenta Platform for Intelligent Computing v1.0、NuPIC* に最適化された BERT-Large、シーケンス長 64、BF16、バッチサイズ 1 を使用して、2023 年 11 月 28 日に Numenta により生成されました。

柔軟性：同一サーバー上で複数のモデルを実行

多くのビジネス・アプリケーションは、それぞれ異なるタスクを解決する複数の LLM を利用しています。NuPIC* を使用すると、非同期クライアントの管理に加えて、同一サーバー上で別々のモデルを同時に非同期に実行できますが、GPU でこの処理を効率的に管理することは困難です。**図 3** に示すように、単一サーバー上で 128 の独立したモデルを実行しました。異なるモデルでは重みのセットや計算要件も異なることがありますが、インテル® CPU 上の NuPIC* では容易に処理できます。さらに、データと計算のニーズが増加したときに、追加の GPU を統合するよりもシステムに CPU を追加する方が、より簡単で制限も少なくなります。これは、クリックするだけで計算能力を追加できるクラウドベースの環境に特に当てはまります。IT の観点から見ると、CPU は GPU よりもデプロイと管理が容易です。

コストと電力効率

CPU は一般に GPU よりも手頃な価格で入手できるため、多くの企業に推奨される選択肢と言えます。しかし、AI 推論を実行する際の有効性は、処理速度の面では制限されていました。NuPIC* はインテル® CPU 上での AI 推論を高速化し、推論にかかるコストと消費電力の両方で利点をもたらします。NuPIC* とインテル® AMX により、AI アプリケーションのコストと消費電力が大幅に削減され、持続可能性とスケーラビリティーが向上します。

実装例

NuPIC* の機能を活用する方法を示すため、容易にデプロイできるものから調整が必要なものまで、さまざまなユースケースのサンプルコードがあります。一例として、NuPIC* を使用して金融コンテキストで感情分析を実行する方法を示します。NuPIC* に最適化された BERT モデルは、ニュース記事や市場レポートなどのテキストデータを処理し、それらを異なる感情（肯定、否定、中立）に分類します。金融アナリストは市場全体の感情を迅速に把握し、投資戦略や市場の動向の予測に関して、十分な情報に基づいて判断できます。

NuPIC* をインストールして起動する

Numenta は、サンプルコードおよび調整とデプロイに必要なすべてのコンポーネントを含むスクリプトを提供しています。NuPIC* をインストールした後、次のコマンドを実行してトレーニング・モジュールと推論サーバーを起動します。

Unix +

```
./nupic_training.sh --start
./nupic_inference.sh -start
```

Docker* イメージがダウンロードされ、Docker* コンテナが起動します。デフォルトでは、トレーニング・モジュールと推論サーバーの Docker* コンテナは、API サーバーのポート 8321 と 8000 で実行されます。サーバーと Docker* コンテナは、`--stop` コマンドでいつでも停止できます。

データセットを準備する

この例では、金融感情データセットをモデルの調整に使用します。まず、データをトレーニング・セットとテストセットに分割します。

```python
dataset_file = f"{script_dir}/../../datasets/financial_sentiment.csv"

…

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

train = pd.concat([y_train, X_train], axis=1)
test = pd.concat([y_test, X_test], axis=1)

train.to_csv("financial_sentiment_train_dataset.csv", index=False)
test.to_csv("financial_sentiment_test_dataset.csv", index=False)
```

NuPIC* トレーニング・モジュールを使用してモデルをデータnい調整する

データセットを分割した後、NuPIC* のトレーニング・モジュールを使用してモデルを調整できます。このプロセスでは、パフォーマンスを最適化するため、学習率、バッチサイズなどのパラメーターを変更します。ここでは、NuPIC* に最適化された BERT モデルを金融感情データセットについて調整します。

```unix
python -m nupic.client.nupic_train  --train_path  ../../datasets/financial_sentiment_train_dataset.csv
--test_path ../../datasets/financial_sentiment_test_dataset.csv --url http\://localhost:8321
```

NuPIC* のトレーニング・モジュールは、金融感情データセットのテスト部分を使用して精度スコアを出力します。必要に応じてハイパーパラメーターを調整し、満足な結果が得られるまで再トレーニングします。調整が完了すると、新しいモデルが生成されます。上記のスクリプトは、トレーニング・モジュールからモデルを自動的に取得し、`.tar.gz` ファイルとしてローカル・ディレクトリーに保存します。後で、このファイルを NuPIC* の推論サーバーにデプロイできます。

調整したモデルを NuPIC* 推論サーバーにインポートする

調整したモデルをデプロイするには、まずそのモデルを推論サーバーで利用できるようにする必要があります。具体的には、最終フェーズの出力（`model_xxx.tar.gz`）を展開して、推論サーバーのコンテナにマップされる `models` ディレクトリーに移動します。次のコマンドを使用します。`xxx` を調整したモデルに置き換えてください。

```unix
cp model_xxx.tar.gz ../nupic/inference/models
cd ../../nupic/inference/models
tar -xzvf model_xxx.tar.gz
```

デプロイ : インテル® Xeon® CPU 上でのリアルタイム推論

次に、モデルをデプロイします。NuPIC* の Python* クライアントを利用して、簡単にデプロイできます。このクライアントは、エンドユーザー・アプリケーションとモデル間の通信を管理します。シンプルな API を使用して、埋め込みを自動的に生成し、推論サーバーにリアルタイムでリクエストを送信できます。NuPIC* クライアントをセットアップし、次のように推論を実行します。

```python
# Model name, you can use a model from NuPIC model libary, or your fine-tuned model
MODEL = "numenta-sbert-2-v1-wtokenizer"

# The URL to your inference server instance
URL = "localhost:8000"

# The supported protocol includes http and grpc
PROTOCOL = "http"

# Optional connection configuration, such as SSL protocol certificates.
CONNECTION_CONFIG = {}

#Connect to NuPIC client
numenta_client = ClientFactory.get_client(MODEL, URL, PROTOCOL, CONNECTION_CONFIG)

#Run inference
result = numenta_client.infer(["Semiconductor industry is doing great!"])
```

テキストデータを含むリクエストを API エンドポイントに送信し、感情予測を取得します。リアルタイム解析の場合、市場レポートやニュース記事を API に継続的にフィードして、感情予測を受け取ることができます。

まとめ

NuPIC* とインテル® AMX を利用して、インテル® CPU 上で LLM の素晴らしいスケーリング能力と組み合わせることにより、エンタープライズ IT に必要な汎用性とセキュリティーを得ることができます。このソフトウェアとハードウェア・テクノロジーの相乗的な組み合わせにより、CPU 上で LLM を実行できるようになるだけでなく、戦略的な利点が得られます。皆さんの環境で NuPIC* からどのようなメリットが得られるか確認したい場合は、こちらからデモをリクエスト（英語）してください。