

独自のカスタム・チャットボットの作成

大規模な言語モデルをインテル・プロセッサで迅速かつ簡単にトレーニングする

Haihao Shen 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
Xinyu Ye、Kaokao Lv、および Xuhui Ren 同 AI ソフトウェア・エンジニア
Huma Abidi 同 AI ソフトウェア製品ジェネラル・マネージャー兼シニア・ディレクター

[ChatGPT*](#) (英語)、[GPT-4*](#) (英語)、[Bard*](#) (英語) などの対話エージェントの途方もないパフォーマンスにより、大規模言語モデル (LLM) が大きな注目を集めています。しかし、LLM のトレーニングや微調整には膨大なコストと時間が必要です。これは、モデルサイズとデータセットが大きいからです。

この記事では、簡単に入手できるハードウェアでカスタム・チャットボットをトレーニングして微調整する方法を説明します。第 4 世代インテル® Xeon® スケーラブル・プロセッサを使用して、体系的な手法を使用してチャットボットを作成し、ドメイン固有のデータセットと最適化された微調整コードベースを生成します。

アプローチ

スタンフォード大学が公開した [Alpaca](#) (英語) は、Meta の [LLaMA](#) (英語) モデルから微調整された、命令に従う言語モデルです。このプロジェクトからヒントを得て、我々はカスタムのドメイン固有チャットボットを作成する、強化された手法を開発しました。使用できる言語モデルは (パフォーマンスの高い言語モデルも含めて) いくつかありますが、オープンモデルである Alpaca を選択しました。

チャットボットのワークフローは、ガイド付きシード生成、ガイドなし (フリー) シード生成、サンプル生成、および微調整の 4 つのステップで構成されます (図 1)。

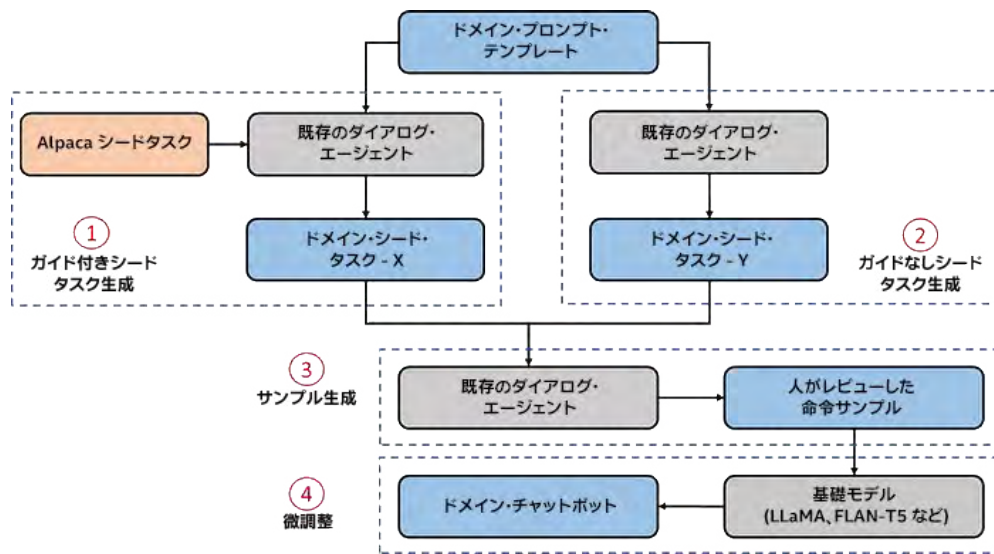


図 1. チャットボットの微調整の概要

これらのステップを説明する前に、シードタスクの生成に役立つプロンプト・テンプレートを紹介します。一般的なタスク向けの Alpaca のサンプルプロンプトを図 2 に示します。

```

You are asked to come up with a set of 20 diverse task instructions. These task instructions will be
given to a GPT model and we will evaluate the GPT model for completing the instructions.

Here are the requirements:
1. Try not to repeat the verb for each instruction to maximize diversity.
2. The language used for the instruction also should be diverse. For example, you should combine
questions with imperative instructions.
3. The type of instructions should be diverse. The list should include diverse types of tasks like
open-ended generation, classification, editing, etc.
...

List of 20 tasks:
    
```

図 2. シードタスク生成のプロンプト・テンプレート

「生成されるタスクの指示は <domain_name> の問題に関連している必要がある」という新しい要件を追加してテンプレートを変更しました。この要件は、指定されたドメインに関連するシードタスクの生成に役立ちます。多様なシードタスクを生成するため、ガイド付きシードタスク生成とガイドなし（フリー）シードタスク生成の両方を使用します。

ガイド付きシードタスク生成は、[Alpaca の既存のシードタスク](#)（英語）を活用します。シードタスクごとに、ドメイン・プロンプト・テンプレートのコンテンツを結合して、既存のダイアログ・エージェントにフィードします。対応する数のタスク（例えば、**図 2** ではプロンプト・テンプレートで定義した 20 のタスク）が生成されることが予想されます。このようなテキスト生成は、因果関係言語モデルの典型的なユースケースの 1 つです。

ガイドなしシードタスク生成は、追加のシードタスクを指定しないで、ドメイン・プロンプト・テンプレートをダイアログ・エージェントに直接フィードします。ここでは、ガイドなしシードタスク生成を「フリー」と呼びます。このアプローチを使用して、新しいドメイン・シード・タスクを生成します（**図 3**）。

```
{
  "instruction": "Summarize the key features of Intel's latest processor.",
  "input": "",
  "output": "Intel's latest processor features improved performance, higher energy efficiency, advanced AI capabilities, enhanced security, and support for cutting-edge memory and storage technologies."
},
```

図 3. ドメイン・シード・タスク

これらのシードタスクを使用して、既存のダイアログ・エージェントを再度利用し、指示サンプルを生成します。ドメイン・プロンプト・テンプレートを使用するため、出力は「命令」、「入力」、「出力」形式の要件に従います。このプロセスを繰り返して、微調整に使用する 2,000 個の命令サンプルを生成します（**図 4**）。

```
{
  "instruction": "Explain how Intel Xeon Scalable processor can improve machine learning tasks",
  "input": "",
  "output": "Intel Xeon Scalable processors feature advanced AI acceleration architectures, advanced caching solutions, and improved memory bandwidth, allowing them to handle heavy machine learning workloads quickly and efficiently."
},
```

図 4. ドメイン命令サンプル

ドメイン・シード・タスクと指示サンプルには類似性があります。これらはそれぞれ ChatGPT* プロンプトとその結果の出力であり、一方が他方に影響を与えると考えられます。

カスタム・チャットボットをトレーニングする

数十億のパラメーターを使用して LLM 全体を微調整するのではなく、[低ランク適応 \(LoRA\)](#) (英語) アプローチを使用して LLM を効率良く微調整します。LoRA は、事前トレーニング済みモデルの重みを固定し、トレーニング可能なランク分解行列をトランスフォーマー・アーキテクチャーの各層に挿入して、ダウンストリーム・タスクのトレーニング可能なパラメーターの数を大幅に削減します。

パラメーター効率に優れた微調整に加えて、ハードウェア・アクセラレーションとソフトウェア・アクセラレーションを利用して微調整プロセスを高速化します。ハードウェア・アクセラレーションの例として、AI パフォーマンスを高速化するために特別に設計された、第 4 世代 Intel® Xeon® スケーラブル・プロセッサで利用可能な Intel® アドバンスド・マトリクス・エクステンション (Intel® AMX) の [AMX-BF16](#) (英語) 命令があります。[PyTorch*](#) (英語)、[PyTorch* 向け Intel® エクステンション](#) (英語) および [Hugging Face* トランスフォーマー](#) (英語) に含まれるソフトウェアの最適化も、これらのフレームワークやライブラリーの最適化されていないバージョンと比較してパフォーマンスの高速化に役立ちます。

微調整プロセスをさらに改善するため、命令サンプルの連結を有効にすることもできます。基本的な考え方は、異なる長さの複数のトレーニング・サンプルを用意する代わりに、いくつかのトークン化された文を連結してトレーニング・サンプルとして長い連結された文にする、というものです。この手法は、基礎となるハードウェアの効率を最大化するのに役立ちます。

上記の最適化はすべて、単一の計算ノードで行われます。微調整プロセス中に分散データ並列処理を使用してマルチノード微調整を実行し、さらに多くの計算能力を活用することもできます。

では、ドメイン・チャットボットのトレーニングを開始しましょう。

- [トランスフォーマー向け Intel® エクステンション](#) (英語) のサンプルコードを確認します。
- [要件ファイル](#) (英語) に定義されている必要なソフトウェア・パッケージをインストールします。
- 事前トレーニング済みモデルをダウンロードします。FLAN-T5 は [Hugging Face*](#) (英語) からダウンロードできます。あるいは、[Meta*](#) (英語) から直接 LLaMA へのアクセスをリクエストするか、[Hugging Face*](#) (英語) の代替バージョンを使用します。
- **図 5** のスクリプトとコマンドを使用してトレーニングを開始します。詳細は、[ユーザーガイド](#) (英語) を参照してください。

```
CCL_WORKER_COUNT=1 python -m torch.distributed.launch --master_addr=<MASTER_ADDRESS> --nproc_per_node=
<NUM_PROCESSES_PER_NODE> --nnodes=<NUM_NODES> --node_rank=<NODE_RANK> \
  finetune_clm.py \
    --model_name_or_path "model_name" \
    --train_file "/path/to/training_data" \
    --dataset_concatenation \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 8 \
    --gradient_accumulation_steps 1 \
    --do_train \
    --learning_rate 2e-5 \
    --num_train_epochs 3 \
    --logging_steps 100 \
    --save_total_limit 2 \
    --overwrite_output_dir \
    --log_level info \
    --save_strategy epoch \
    --output_dir "model_output_dir" \
    --bf16 --no_cuda --xpu_backend ccl
```

図 5. マルチノード微調整スクリプト

結果

Alpaca からの 52,000 個の一般命令サンプルと生成した 2,000 個のドメイン命令サンプルを使用して、2 つの人気の高い LLM (LLaMA-7B および FLAN-T5 XL) で微調整をテストしました。前述のソフトウェア最適化手法を使用すると、第 4 世代 Intel® Xeon® スケーラブルプロセッサで最大 25 倍スピードアップしました (図 6)。

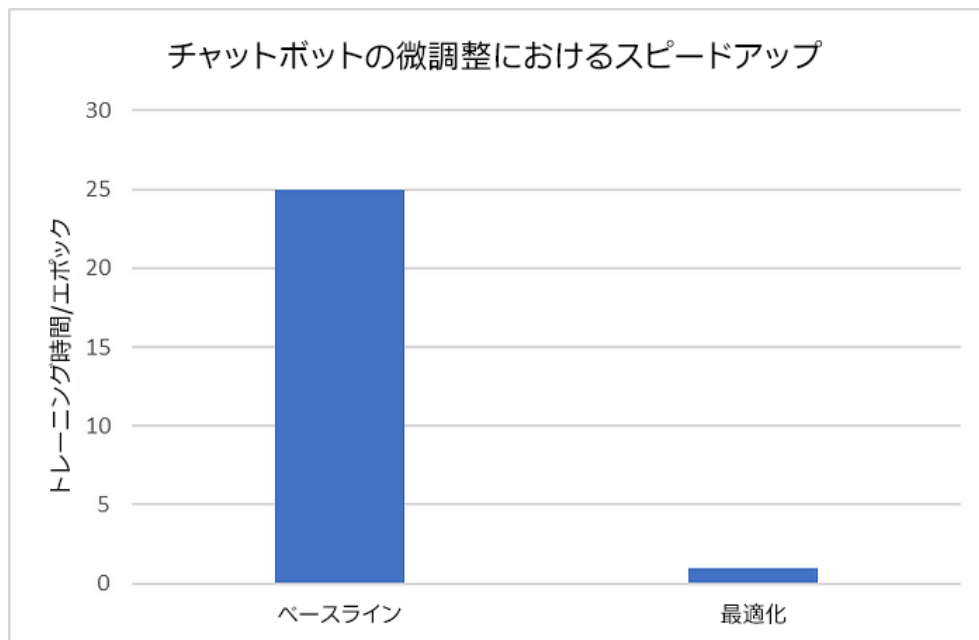


図 6. チャットボットの微調整におけるスピードアップ

ゼロショット設定と少数ショット設定でモデルを評価する新しい[大規模マルチタスク言語理解 \(MMLU\)](#) (英語) ベンチマークを使用して、事前トレーニング中にモデルが取得した知識を測定しました。**表 1** から、カスタム・チャットボットが平均してほとんどのドメインサブタスクで優れた結果を示していることが分かります。MMLU の結果は GPT-3* (175B) とほぼ同等ですが、パラメーター・サイズでは 25 倍効率的です。インテルのハードウェア・アクセラレーションとソフトウェアの最適化を活用しながら、ほかのオープン言語モデルを使用してこのビルドを試してみてください。

モデル	パラメーター	人文科学	STEM	社会科学	その他	平均
Alpaca	7B	38.3	33.2	44.5	47.0	40.6
カスタム	7B	39.2	32.9	44.7	47.3	40.9

表 1. MMLU ベンチマーク

MMLU ベンチマークに加えて、次のような会話評価も行いました。

入力プロンプト : Where to find the product information about Intel microchip?
(インテルのマイクロチップに関する製品情報はどこで見つかりますか?)

Alpaca の返答 : You can find product information about Intel microchips on their website, Intel.com. There you will find information about their processors, graphics cards, motherboards, and other products. You can also find product specifications, tutorials, and other resources. (インテルのマイクロチップに関する製品情報は、インテルのウェブサイト、Intel.com で見つけることができます。このサイトでは、インテルのプロセッサ、グラフィックス・カード、マザーボード、その他の製品に関する情報が見つかります。製品仕様、チュートリアル、その他のリソースも見つけることができます。)

カスタム・チャットボットの返答 : You can find the product information about Intel microchips on their website, <https://www.intel.com/content/www/us/en/support/processors/top-questions.html>. (インテルのマイクロチップに関する製品情報は、インテルのウェブサイト、<https://www.intel.com/content/www/us/en/support/processors/top-questions.html> で見つけることができます。)

まとめ

[トランスフォーマー向けインテル® エクステンション](#) (英語) で、カスタム・チャットボットの[ソースコード](#) (英語) と [ドメイン固有のデータセット](#) (英語) をリリースしました。インテルのプロセッサ上で独自のチャットボットを作成し、AI ワークフローの一部としてほかの[インテル® AI ツール](#) (英語) や最適化を検討してみてください。

次のステップは、量子化と圧縮技術を有効にして、インテルのプラットフォーム上でチャットボットの推論を高速化することです。最新の最適化に関する情報を受け取りたい場合は、[トランスフォーマー向けインテル® エクステンション](#) (英語) に星を追加してください。ご質問がある場合はお気軽にお問い合わせください。