LangChain が OpenVINO™ を公式にサポート

この記事は、Medium に公開されている「LangChain Officially Supports OpenVINO™ Now!」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。

LangChain は、開発者が言語モデルを使用してエンドツーエンドのアプリケーションを構築できるように設計された強力なフレームワークで、大規模言語モデル(LLM)を利用したアプリケーションの作成プロセスを簡素化するツール、コンポーネント、インターフェイスのスイートを提供します。LangChain を使用すると、開発者は RAG やエージェント・パイプラインなどの高レベルのアプリケーションを簡単に構築できます。LangChain が OpenVINO™ を公式にサポートしたことで、LLM、テキスト埋め込み、リランカーを含む OpenVINO™ ベースのコンポーネントを LangChain で直接呼び出すことができるようになりました。この統合により、ローカル RAG およびエージェント・サービスのパフォーマンスが向上します。

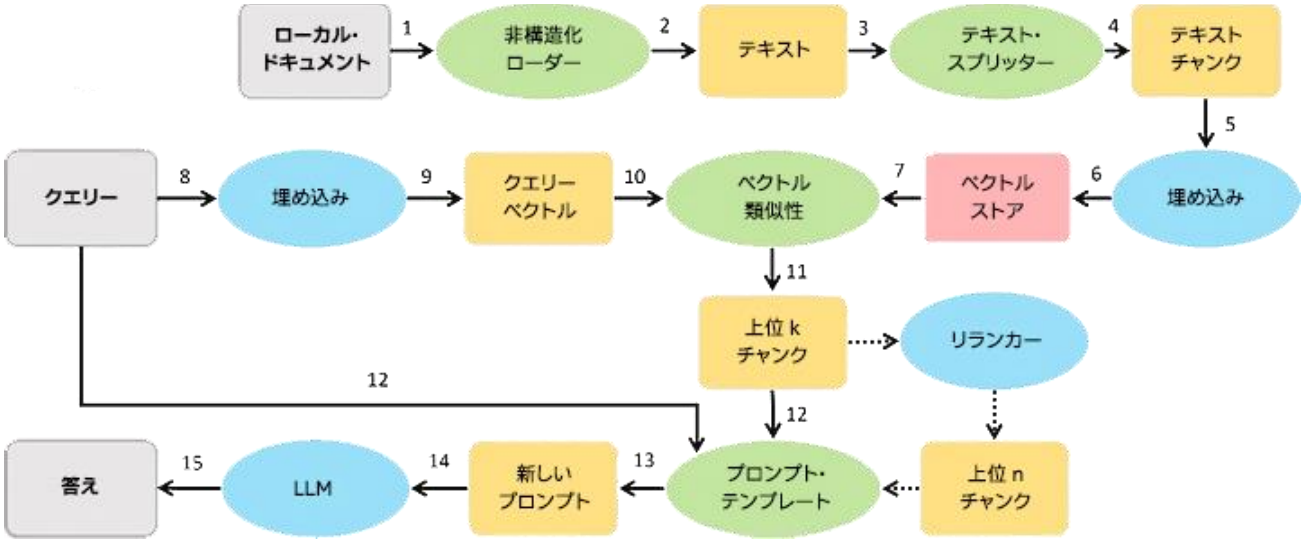

図: RAG システム・リファレンス・パイプライン

インストール

LangChain で OpenVINO™ を呼び出すには、通常の LangChain のインストール手順に加えて、Optimum Intel ライブラリーをインストールします。Optimum Intel には、モデル・コンバーター、ランタイム、NNCF などの OpenVINO™ のすべての依存ファイルがすでに含まれています。

```
pip install langchain
pip install --upgrade-strategy eager "optimum[openvino,nncf]"
```

LLM

大規模言語モデルは LangChain フレームワークのコア・モデル・コンポーネントで、RAG システムで最終的な答えを生成したり、エージェント・システムで計画を立ててツールを呼び出すことができます。HuggingFace パイプラインにバックエンドとして OpenVINO™ を追加し、そのコードを直接再利用することで、開発者は LangChain の HuggingFace パイプラインで OpenVINO™ を使用して LLM を初期化できます。model_id に

は、HuggingFace のモデル ID、ローカルの PyTorch* モデルパス、または OpenVINO™ モデルパスを指定します。

```
from langchain_community.llms.huggingface_pipeline import HuggingFacePipeline

ov_config = {"PERFORMANCE_HINT": "LATENCY", "NUM_STREAMS": "1", "CACHE_DIR":
""}

ov_llm = HuggingFacePipeline.from_model_id(
    model_id="gpt2",
    task="text-generation",
    backend=" openvino",
    model_kwargs={"device": "CPU", "ov_config": ov_ config},
    pipeline_kwargs={"max_new_tokens": 10},
)
```

OpenVINO™ LLM モデル・オブジェクトを作成した後、推論タスクを LangChain のほかの LLM コンポーネントとしてデプロイできます。

```
from langchain_core.prompts import PromptTemplate

template = """Question: {question}

Answer: Let's think step by step."""
prompt = PromptTemplate.from_template(template)

chain = prompt | ov_llm

question = "What is electroencephalography?"

print(chain.invoke({"question": question}))
```

LLM をインテルの GPU にデプロイする場合は、model_kwargs={“device”: “GPU”} と指定して、推論を GPU で実行します。Optimum Intel のコマンドライン・ツールを使用して INT4 の重みのモデルをローカルフォルダーに直接エクスポートすることもできます。

```
optimum-cli export openvino --model gpt2  --weight-format int4 ov_model_dir
```

OpenVINO™ LLM コンポーネントと使用方法の詳細は、次のウェブサイトを参照してください。

https://python.langchain.com/v0.1/docs/integrations/llms/openvino/ (英語)

テキスト埋め込み

テキスト埋め込みモデルは、テキストを特徴ベクトルに変換するために使用されます。テキストの類似性に基づいてリトリーバーを作成するためにも使用できます。このモデルは RAG システムで幅広く使用されていて、テキスト埋め込みタスクから上位 k の候補コンテキストを生成することが期待されます。テキスト埋め込みモデルは、Optimum Intel を利用して特徴抽出タスクによりエクスポートできます。

```
optimum-cli export openvino — model BAAI/bge-small-en — task feature-extraction
```

LangChain では、OpenVINOEmbeddings クラスと OpenVINOBgeEmbeddings クラスを利用して従来の BERT 埋め込みモデルと BGE ベースの埋め込みモデルをデプロイできます。次の手順は BGE 埋め込みモデルの例です。

```
model_name = "BAAI/bge-small-en"
model_kwargs = {"device": "CPU"}
encode_kwargs = {"normalize_embeddings": True}
ov_embeddings = OpenVINOBgeEmbeddings(
    model_name_or_path=model_name,
    model_kwargs=model_kwargs,
    encode_kwargs=encode_kwargs,
)

embedding = ov_embeddings.embed_query("hi this is harrison")
```

OpenVINO™ 埋め込みコンポーネントと使用方法の詳細は、次のウェブサイトを参照してください。

https://python.langchain.com/v0.1/docs/integrations/text_embedding/openvino/（英語）

リランカー

リランカーはテキスト分類モデルの一種で、各候補コンテキストとクエリー間の類似性のリストを取得し、並べ替えた後、RAG システムでコンテキストをさらにフィルタリングできます。リランカーモデルは、Optimum Intel のテキスト分類タスクを利用してエクスポートできます。

```
optimum-cli export openvino --model BAAI/bge-reranker-large --task text-
classification
```

モデルのデプロイのプロセスで、OpenVINO™ ベースのリランクタスクを OpenVINOReranker クラスで作成し、ContextualCompressionRetriever で呼び出して、リトリーバーの検索結果を圧縮できます。次の例は、リトリーバーの上位 k 件の検索結果を並べ替え、クエリーとの類似性に基づいて上位 4 件の結果を選択して、入力プロンプトの長さをさらに圧縮します。

```
model_name = "BAAI/bge-reranker-large"

ov_compressor = OpenVINOReranker(model_name_or_path=model_name, top_n=4)
compression_retriever = ContextualCompressionRetriever(
    base_compressor=ov_compressor, base_retriever=retriever
)
```

OpenVINO™ リランカー・コンポーネントと使用方法の詳細は、次のウェブサイトを参照してください。

https://python.langchain.com/v0.1/docs/integrations/document_transformers/openvino_rerank/（英語）

まとめ

OpenVINO™ ベースのモデルのタスクが LangChain フレームワークに統合されたことにより、開発者は LangChain を使用して主要なモデルのタスクの推論パフォーマンスをより簡単に向上できるようになります。

関連情報(英語)

- LangChain と OpenVINO™ ベースの RAG の例: https://github.com/openvinotoolkit/openvino_notebooks/tree/latest/notebooks/llm-rag-langchain
- LangChain と OpenVINO™ ベースのエージェントの例: https://github.com/openvinotoolkit/openvino_notebooks/tree/latest/notebooks/llm-agent-langchain

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