Dlib-Transformer-extensions

Dlib のための transformer と言語モデル。モデルのインポートから提供まで

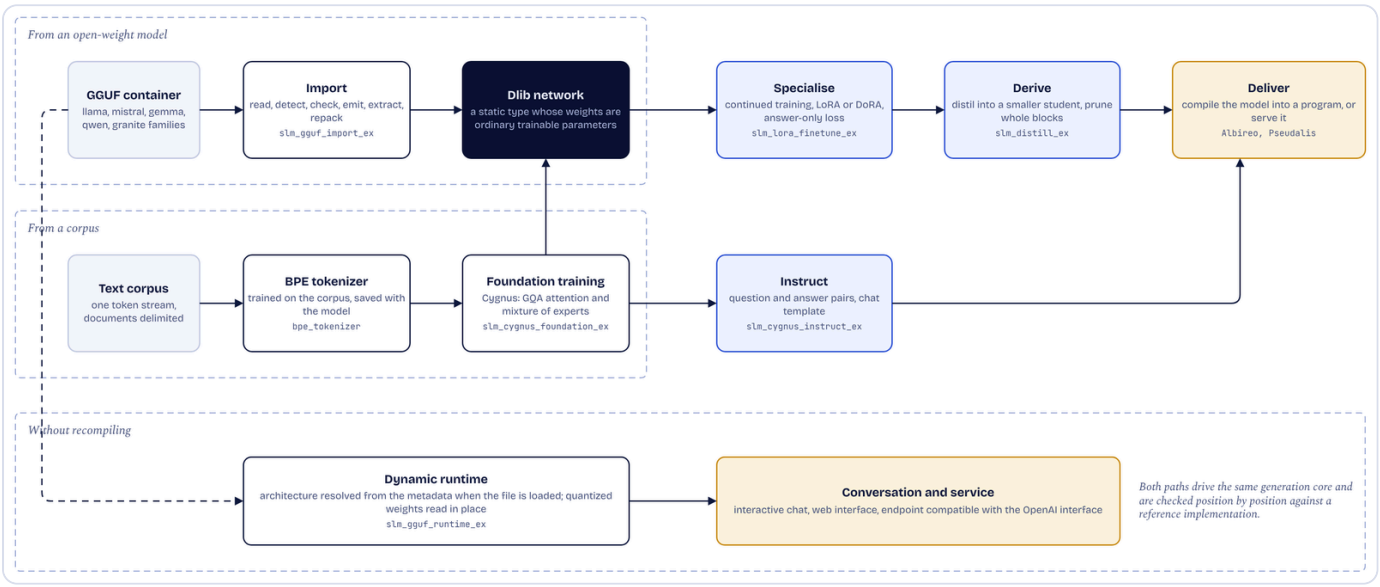
私たちが公開するすべてが、このライブラリの上に成り立っています。オープンソースの C++ 機械学習ライブラリ Dlib の一時的なフォークであり、各機構はここで設計され、実行できる例題で示され、測定され、安定してから Dlib 本体に提案されます。

ライセンス **Boost Software License 1.0**　　C++14 **CUDA, OpenVINO**

24	14	12	0
例題プログラム、機構ごとに 1 つ	Dlib にマージ済みの貢献	GGUF ファイルから読めるモデルアーキテクチャ	Python への依存

モデルのライフサイクル全体を 1 つのライブラリで

オープンウェイトモデルは GGUF ファイルから入り、通常の Dlib ネットワークになります。私たち自身のモデルはコーパスから始まります。その後はどちらも同じ道をたどります。特化、派生、提供です。もう 1 つのエンジンは、対応するどのモデルも再コンパイルなしで動かします。



ライブラリの対象範囲

ビジョンとマルチモーダル
Vision Transformer を画像のバックボーンとして、またはテキストと画像を読むモデルのビジョンタワーとして使い、エンドツーエンドで学習できます。

個人データの保護
ルールと小さな学習済み認識器が 5 言語で個人データを見つけ、フォーマット保持暗号が可逆的に置き換えます。

NPU と内蔵 GPU
各テンソル演算は、実測で最も速いデバイスで実行されます。ホスト、Intel NPU、または OpenVINO 経由の内蔵 GPU です。

カード上の量子化された重み
CUDA カーネルが、GGUF 形式のまま保持した重みを展開せずに活性値と掛け合わせます。

検索
埋め込みモデルが文書フォルダーに索引を付け、質問に最も合う一節で答えます。

GPU メモリの拡張
V²RAM がカード、ホストメモリ、ディスクストアの間でブロックを流し、カードより大きなモデルをそのカードで学習できるようにします。

プロセッサー別のカーネル
AVX2、AVX-VNNI、AVX-512 VNNI のコードパスを、プロセッサーの機能に応じて実行時に選択します。

予測という圧縮
学習済みネットワークと算術符号化器が、どのプロセッサーでもビット単位で再現できる可逆圧縮器になります。

詳しくは
cydral-technology.eu/ja/research/library/

ソースコード
github.com/Cydral/Dlib-Transformer-extensions

お問い合わせ
cydral-technology.eu/ja/contact/

Boost Software License 1.0 による無料のオープンソース。© 2026 Cydral Technology、フランス。