Dlib-Transformer-extensions

Transformer und Sprachmodelle für Dlib, vom Import eines Modells bis zu seiner Auslieferung

Die Bibliothek, auf der alles beruht, was wir veröffentlichen. Dieser temporäre Fork von Dlib, der quelloffenen C++-Bibliothek für maschinelles Lernen, ist der Ort, an dem jeder Mechanismus entworfen, durch ein lauffähiges Beispiel gezeigt, gemessen und stabilisiert wird, bevor er Dlib selbst vorgeschlagen wird.

Lizenz **Boost Software License 1.0** C++14 **CUDA, OpenVINO**

24
Beispielprogramme, eines je Mechanismus

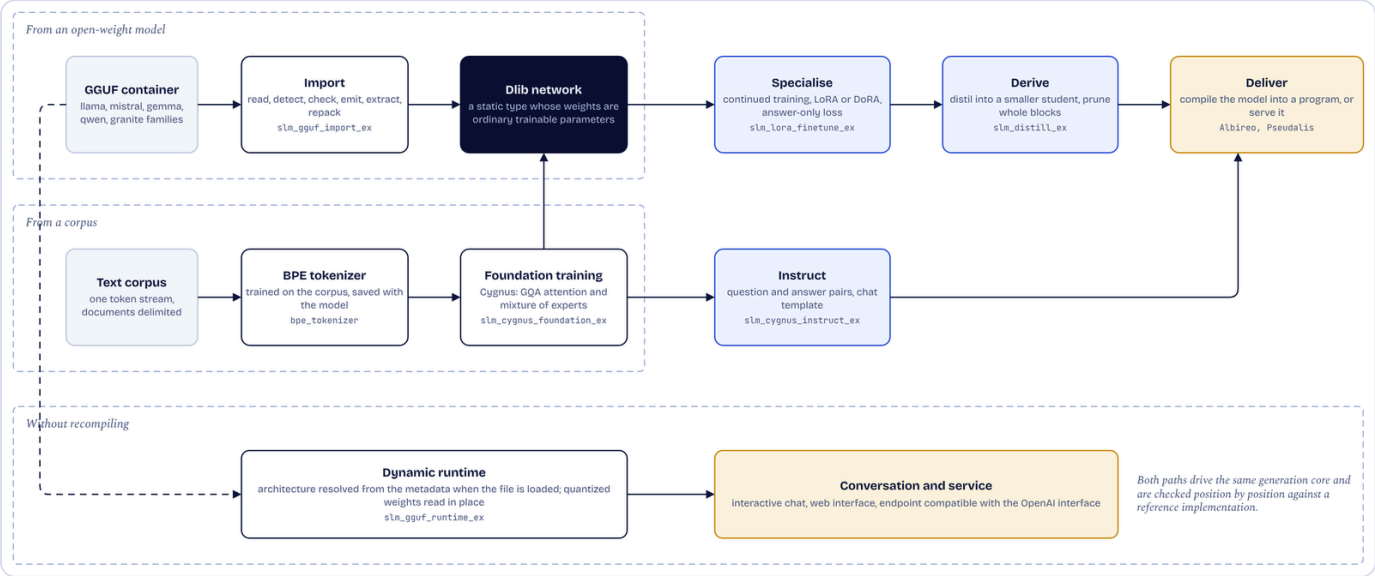
14
bereits in Dlib aufgenommene Beiträge

12
aus GGUF-Dateien gelesene Modellarchitekturen

0
Python-Abhängigkeiten in der Kette

Eine Bibliothek für das ganze Leben eines Modells

Ein Modell mit offenen Gewichten kommt über seine GGUF-Datei herein und wird zu einem gewöhnlichen Dlib-Netz; ein eigenes Modell beginnt mit einem Korpus. Beide gehen danach denselben Weg: spezialisieren, ableiten, ausliefern. Eine zweite Engine führt jedes unterstützte Modell ohne Neukompilierung aus.



Umfang der Bibliothek

Sehen und Multimodalität
Vision Transformer als Bild-Backbone oder als Vision-Turm eines Modells, das Text und Bilder liest, durchgehend trainierbar.

Schutz personenbezogener Daten
Regeln und kleine gelernte Erkenner finden personenbezogene Daten in fünf Sprachen; formaterhaltende Verschlüsselung ersetzt sie umkehrbar.

NPU und integrierte GPU
Jede Tensoroperation läuft auf dem Gerät, das als schnellstes gemessen wurde: Host, Intel-NPU oder integrierte GPU über OpenVINO.

Quantisierte Gewichte auf der Karte
CUDA-Kerne multiplizieren Aktivierungen mit Gewichten in ihren GGUF-Formaten, ohne sie zu entpacken.

Dokumentensuche
Embedding-Modelle indexieren einen Ordner mit Dokumenten und beantworten eine Frage mit der passenden Textstelle.

Erweiterter GPU-Speicher
V²RAM bewegt Blöcke zwischen Karte, Host-Speicher und einem Plattenspeicher, sodass ein Modell, das größer ist als die Karte, darauf trainiert.

Prozessorspezifische Kerne
Codepfade für AVX2, AVX-VNNI und AVX-512 VNNI werden zur Laufzeit nach dem gewählt, was der Prozessor bietet.

Vorhersage als Kompression
Ein trainiertes Netz und ein arithmetischer Kodierer bilden einen verlustfreien Kompressor, den jeder Prozessor Bit für Bit reproduziert.

Mehr dazu
cydral-technology.eu/de/research/library/

Quellcode
github.com/Cydral/Dlib-Transformer-extensions

Kontakt
cydral-technology.eu/de/contact/

Frei und quelloffen unter der Boost Software License 1.0. © 2026 Cydral Technology, Frankreich.