Dlib-Transformer-extensions

Transformers et modèles de langage pour Dlib, de l'import d'un modèle à sa livraison

La bibliothèque sur laquelle repose tout ce que nous publions. Ce fork temporaire de Dlib, la bibliothèque C++ open source d'apprentissage automatique, est le lieu où chaque mécanisme est conçu, démontré par un exemple exécutable, mesuré et stabilisé avant d'être proposé à Dlib.

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

24
programmes d'exemple, un par mécanisme

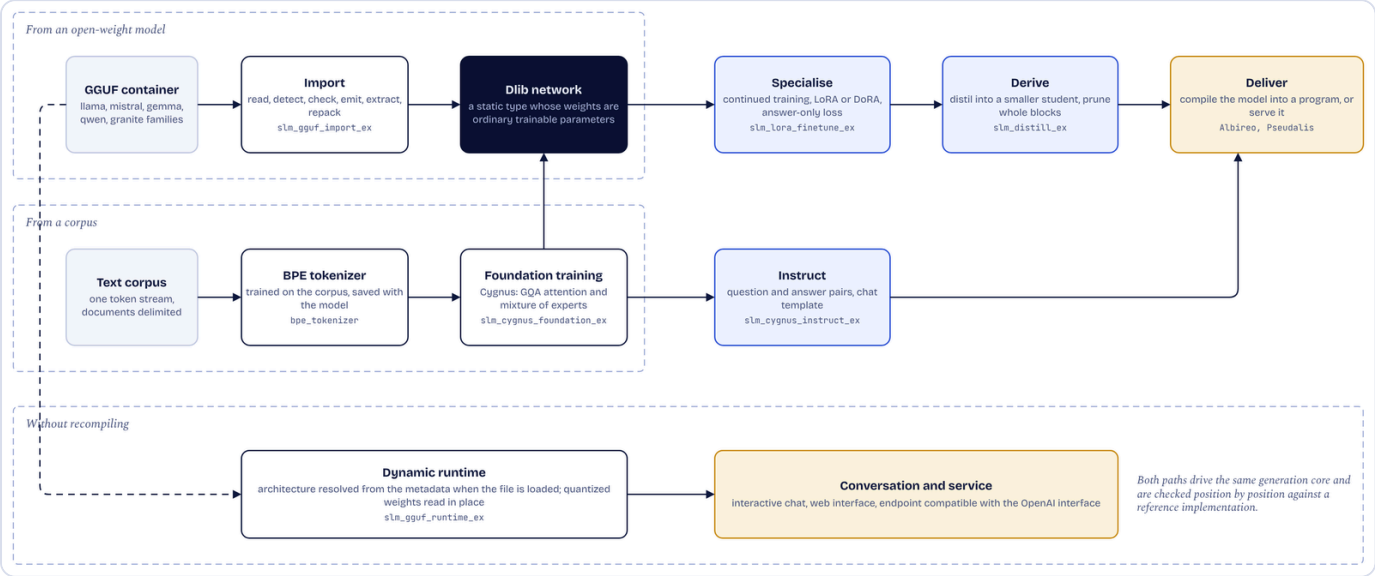
14
contributions déjà intégrées à Dlib

12
architectures de modèles lues depuis des fichiers GGUF

0
dépendance Python dans la chaîne

Une seule bibliothèque pour toute la vie d'un modèle

Un modèle à poids ouverts entre par son fichier GGUF et devient un réseau Dlib ordinaire ; un modèle maison part d'un corpus. Les deux suivent ensuite le même chemin : spécialiser, dériver, livrer. Un second moteur exécute tout modèle pris en charge sans recompilation.



Périmètre de la bibliothèque

Vision et multimodalité
Des Vision Transformers comme épine dorsale d'images ou comme tour de vision d'un modèle qui lit texte et images, entraînable de bout en bout.

Protection des données personnelles
Des règles et de petits modèles de reconnaissance trouvent les données personnelles en cinq langues ; un chiffrement préservant le format les remplace de façon réversible.

NPU et GPU intégré
Chaque opération sur les tenseurs s'exécute sur le périphérique mesuré comme le plus rapide : l'hôte, le NPU d'Intel ou le GPU intégré, par OpenVINO.

Poids quantifiés sur la carte
Des noyaux CUDA multiplient les activations par des poids gardés dans leurs formats GGUF, sans les développer.

Recherche documentaire
Des modèles d'embeddings indexent un dossier de documents et répondent à une question par le passage qui lui correspond.

Mémoire GPU étendue
V²RAM fait circuler les blocs entre la carte, la mémoire de l'hôte et un stockage sur disque, si bien qu'un modèle plus grand que la carte s'y entraîne.

Noyaux propres au processeur
Les chemins de code pour AVX2, AVX-VNNI et AVX-512 VNNI sont choisis à l'exécution d'après ce qu'offre le processeur.

La prédiction comme compression
Un réseau entraîné et un codeur arithmétique forment un compresseur sans perte que tout processeur reproduit bit pour bit.

En savoir plus
cydral-technology.eu/fr/research/library/

Code source
github.com/Cydral/Dlib-Transformer-extensions

Contact
cydral-technology.eu/fr/contact/

Libre et gratuit sous la Boost Software License 1.0. © 2026 Cydral Technology, France.