Dlib-Transformer-extensions

Transformer e modelli linguistici per Dlib, dall'importazione di un modello alla sua consegna

La libreria su cui poggia tutto ciò che pubblichiamo. Questo fork temporaneo di Dlib, la libreria C++ open source di apprendimento automatico, è il luogo in cui ogni meccanismo è progettato, dimostrato da un esempio eseguibile, misurato e stabilizzato prima di essere proposto a Dlib.

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

24
programmi di esempio, uno per meccanismo

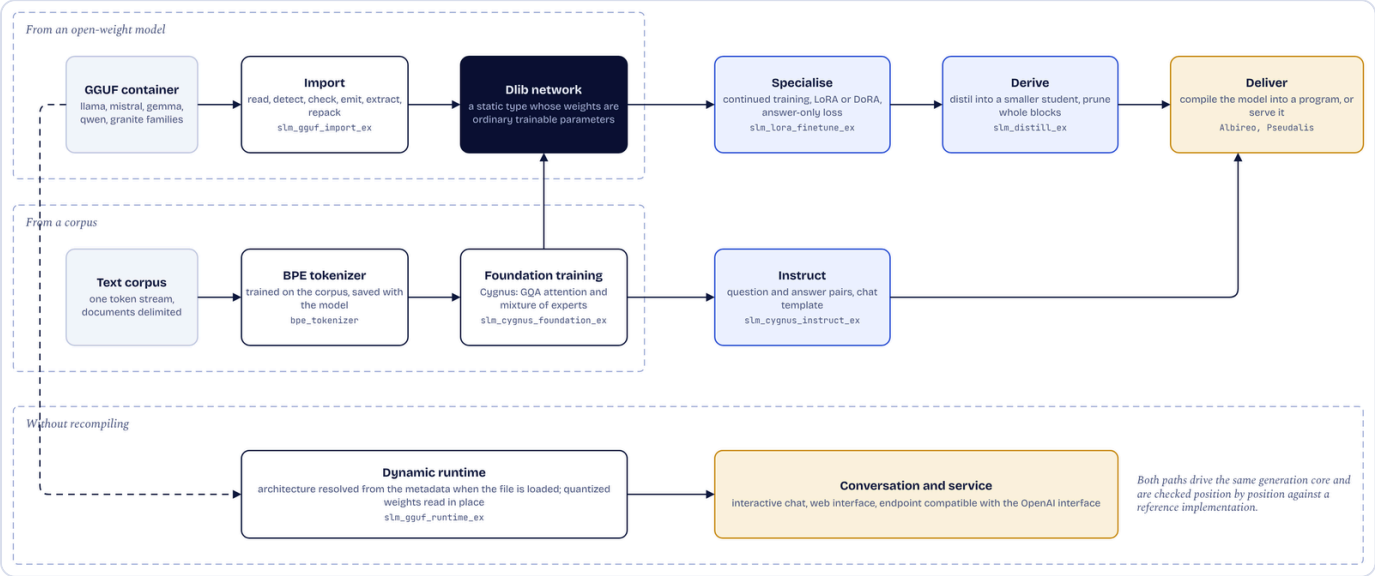
14
contributi già integrati in Dlib

12
architetture di modelli lette da file GGUF

0
dipendenze Python nella catena

Una sola libreria per tutta la vita di un modello

Un modello a pesi aperti entra dal suo file GGUF e diventa una normale rete Dlib; un modello nostro parte da un corpus. Entrambi seguono poi lo stesso percorso: specializzare, derivare, consegnare. Un secondo motore esegue qualsiasi modello supportato senza ricompilare.



Ambito della libreria

Visione e multimodalità
Vision Transformer come dorsale per immagini o come torre visiva di un modello che legge testo e immagini, addestrabile da un capo all'altro.

Protezione dei dati personali
Regole e piccoli riconoscitori appresi trovano i dati personali in cinque lingue; una cifratura che conserva il formato li sostituisce in modo reversibile.

NPU e GPU integrata
Ogni operazione sui tensori gira sul dispositivo misurato come il più veloce: l'host, la NPU di Intel o la GPU integrata, tramite OpenVINO.

Pesi quantizzati sulla scheda
Nuclei CUDA moltiplicano le attivazioni per pesi tenuti nei loro formati GGUF, senza espanderli.

Ricerca documentale
Modelli di embedding indicizzano una cartella di documenti e rispondono a una domanda con il passaggio che le corrisponde.

Memoria GPU estesa
V²RAM sposta i blocchi tra la scheda, la memoria dell'host e un archivio su disco, così un modello più grande della scheda vi si addestra.

Nuclei specifici del processore
I percorsi di codice per AVX2, AVX-VNNI e AVX-512 VNNI sono scelti in esecuzione secondo ciò che offre il processore.

La predizione come compressione
Una rete addestrata e un codificatore aritmetico formano un compressore senza perdita che ogni processore riproduce bit per bit.

Per saperne di più
cydral-technology.eu/it/research/library/

Codice sorgente
github.com/Cydral/Dlib-Transformer-extensions

Contatti
cydral-technology.eu/it/contact/

Libero e gratuito sotto la Boost Software License 1.0. © 2026 Cydral Technology, Francia.