Dlib-Transformer-extensions

Трансформеры и языковые модели для Dlib, от импорта модели до её поставки

Библиотека, на которую опирается всё, что мы публикуем. Этот временный форк Dlib, библиотеки машинного обучения на C++ с открытым кодом, служит местом, где каждый механизм проектируется, показывается на запускаемом примере, измеряется и стабилизируется, прежде чем его предлагают самой Dlib.

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

24
примеров программ, по одному на механизм

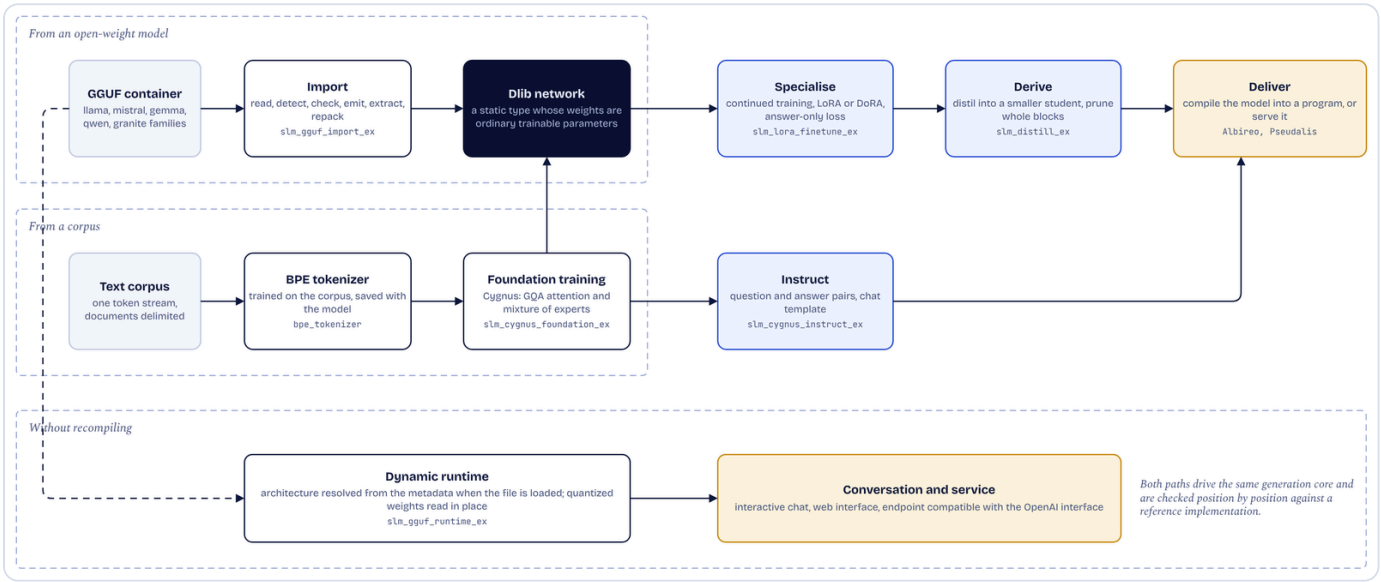
14
вкладов, уже принятых в Dlib

12
архитектур моделей, читаемых из файлов GGUF

0
зависимостей от Python в цепочке

Одна библиотека на всю жизнь модели

Модель с открытыми весами входит через свой файл GGUF и становится обычной сетью Dlib; наша собственная модель начинается с корпуса. Дальше обе идут одним путём: специализация, производные модели, поставка. Второй движок запускает любую поддерживаемую модель без перекомпиляции.



Возможности библиотеки

Зрение и мультимодальность
Vision Transformer как основа для изображений или как зрительная башня модели, читающей текст и изображения, с обучением от начала до конца.

Защита персональных данных
Правила и небольшие обученные распознаватели находят персональные данные на пяти языках; шифрование с сохранением формата обратимо заменяет их.

NPU и встроенный GPU
Каждая тензорная операция выполняется на устройстве, которое по замерам оказалось быстрее: хост, Intel NPU или встроенный GPU через OpenVINO.

Квантованные веса на карте
Ядра CUDA умножают активации на веса, хранящиеся в форматах GGUF, не разворачивая их.

Поиск
Модели эмбеддингов индексируют папку документов и отвечают на вопрос подходящим фрагментом.

Расширенная память GPU
V²RAM перемещает блоки между картой, памятью хоста и дисковым хранилищем, так что модель больше карты обучается на ней.

Ядра под конкретный процессор
Пути кода для AVX2, AVX-VNNI и AVX-512 VNNI выбираются во время выполнения по возможностям процессора.

Предсказание как сжатие
Обученная сеть и арифметический кодер образуют компрессор без потерь, который любой процессор воспроизводит бит в бит.

Подробнее:
cydral-technology.eu/ru/research/library/

Исходный код
github.com/Cydral/Dlib-Transformer-extensions

Контакты
cydral-technology.eu/ru/contact/

Бесплатно и с открытым кодом под лицензией Boost Software License 1.0. © 2026 Cydral Technology, Франция.