Dlib-Transformer-extensions

Transformers y modelos de lenguaje para Dlib, de la importación de un modelo a su entrega

La biblioteca en la que se apoya todo lo que publicamos. Este fork temporal de Dlib, la biblioteca C++ de código abierto de aprendizaje automático, es el lugar donde cada mecanismo se diseña, se demuestra con un ejemplo ejecutable, se mide y se estabiliza antes de proponerlo a Dlib.

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

24
programas de ejemplo, uno por mecanismo

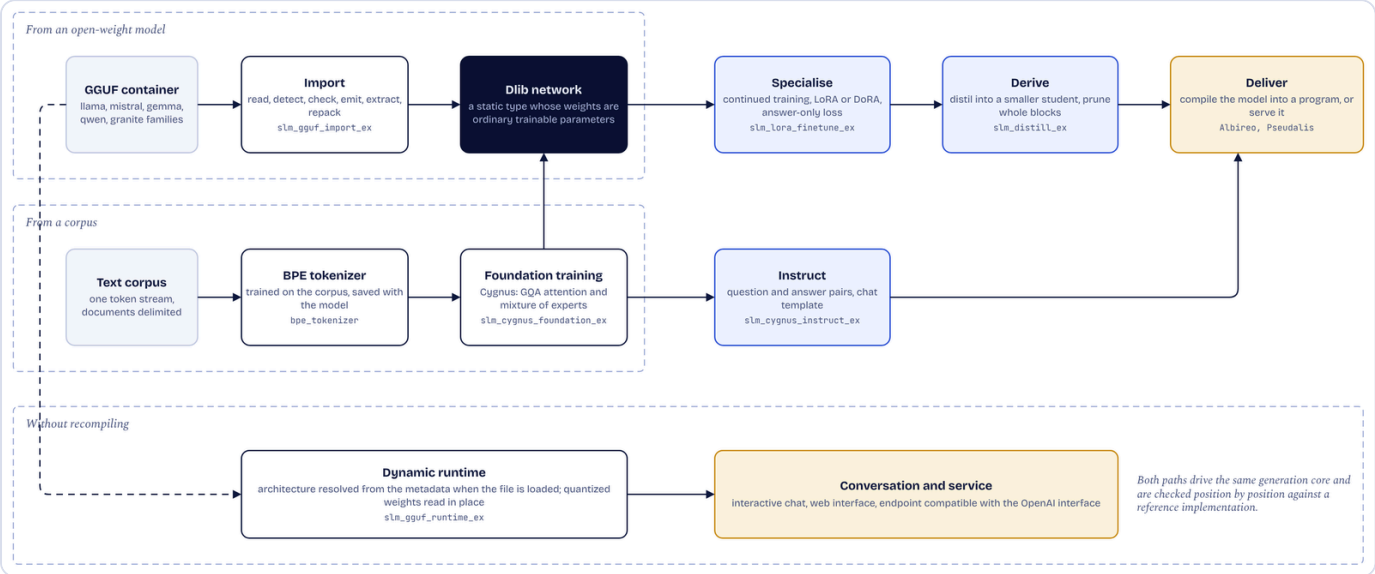
14
contribuciones ya integradas en Dlib

12
arquitecturas de modelos leídas de archivos GGUF

0
dependencias de Python en la cadena

Una sola biblioteca para toda la vida de un modelo

Un modelo de pesos abiertos entra por su archivo GGUF y se convierte en una red Dlib ordinaria; un modelo propio parte de un corpus. Ambos siguen después el mismo camino: especializar, derivar, entregar. Un segundo motor ejecuta cualquier modelo compatible sin recompilar.



Alcance de la biblioteca

Visión y multimodalidad
Vision Transformers como columna vertebral de imágenes o como torre de visión de un modelo que lee texto e imágenes, entrenable de extremo a extremo.

Protección de datos personales
Reglas y pequeños reconocedores aprendidos encuentran los datos personales en cinco idiomas; un cifrado que conserva el formato los sustituye de forma reversible.

NPU y GPU integrada
Cada operación con tensores se ejecuta en el dispositivo medido como más rápido: el anfitrión, la NPU de Intel o la GPU integrada, mediante OpenVINO.

Pesos cuantizados en la tarjeta
Núcleos CUDA multiplican las activaciones por pesos mantenidos en sus formatos GGUF, sin expandirlos.

Búsqueda documental
Modelos de embeddings indexan una carpeta de documentos y responden a una pregunta con el pasaje que le corresponde.

Memoria de GPU ampliada
V²RAM mueve los bloques entre la tarjeta, la memoria del anfitrión y un almacén en disco, de modo que un modelo mayor que la tarjeta se entrena en ella.

Núcleos propios del procesador
Los caminos de código para AVX2, AVX-VNNI y AVX-512 VNNI se eligen en ejecución según lo que ofrece el procesador.

La predicción como compresión
Una red entrenada y un codificador aritmético forman un compresor sin pérdida que cualquier procesador reproduce bit a bit.

Más información
cydral-technology.eu/es/research/library/

Código fuente
github.com/Cydral/Dlib-Transformer-extensions

Contacto
cydral-technology.eu/es/contact/

Libre y gratuito bajo la Boost Software License 1.0. © 2026 Cydral Technology, Francia.