Dlib-Transformer-extensions

面向 *Dlib* 的 *transformer* 与语言模型，从模型导入到交付

我们发布的一切都建立在这个库之上。它是开源 C++ 机器学习库 Dlib 的一个临时分支：每种机制都在这里设计，用可运行的示例演示，经过测量并稳定下来，然后才提交给 Dlib 本身。

许可证 **Boost Software License 1.0** C++14 **CUDA, OpenVINO**

24
个示例程序，每种机制一个

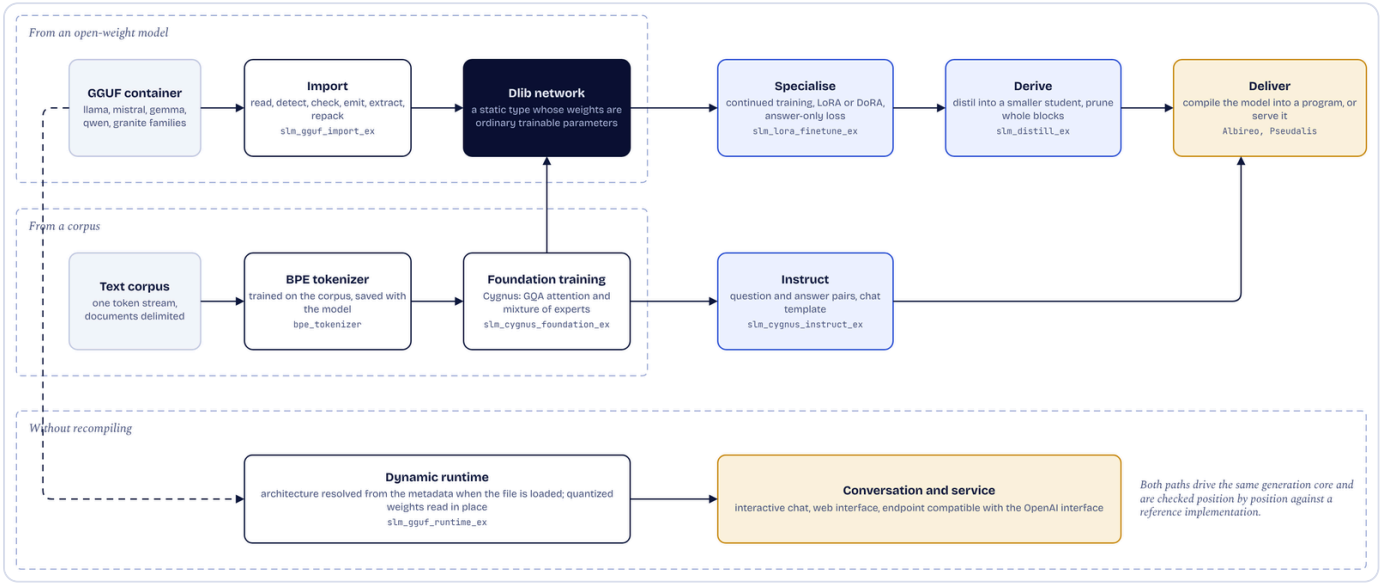
14
项贡献已合并进 Dlib

12
种可从 GGUF 文件读取的模型架构

0
全程零 Python 依赖

一个库，覆盖模型的整个生命周期

开放权重模型通过它的 GGUF 文件进入，成为一个普通的 Dlib 网络；我们自己的模型则从语料起步。此后两者走同一条路：专门化、派生、交付。另有第二个引擎，无需重新编译即可运行任何受支持的模型。



核心库的覆盖范围

视觉与多模态
Vision Transformer 可作为图像骨干网络，也可作为图文模型的视觉塔，并支持端到端训练。

个人数据保护
规则加上小型学习识别器，在五种语言中找出个人数据；保留格式加密以可逆的方式替换它们。

NPU 与集成 GPU
每个张量运算都在实测最快的设备上执行：主机、Intel NPU，或通过 OpenVINO 使用的集成 GPU。

显卡上的量化权重
CUDA 内核直接用保持 GGUF 格式的权重与激活值相乘，无需展开。

检索
嵌入模型为一个文档文件夹建立索引，并用最匹配的段落回答问题。

扩展 GPU 显存
V²RAM 在显卡、主机内存和磁盘存储之间流式调度数据块，让比显卡更大的模型也能在其上训练。

面向处理器的计算内核
AVX2、AVX-VNNI 和 AVX-512 VNNI 的代码路径在运行时根据处理器的能力选择。

预测即压缩
训练好的网络加上算术编码器构成无损压缩器，任何处理器都能逐比特复现。

了解更多
cydral-technology.eu/zh/research/library/

源代码
github.com/Cydral/Dlib-Transformer-extensions

联系
cydral-technology.eu/zh/contact/

免费且开源，采用 Boost Software License 1.0。© 2026 Cydral Technology，法国。