Dlib-Transformer-extensions

Dlib के लिए ट्रांसफ़ॉर्मर और भाषा मॉडल, मॉडल के इम्पोर्ट से उसकी डिलीवरी तक

वह लाइब्रेरी जिस पर हमारा हर प्रकाशन टिका है। ओपन-सोर्स C++ मशीन लर्निंग लाइब्रेरी Dlib का यह अस्थायी फ़ोर्क वह जगह है जहाँ हर व्यवस्था डिज़ाइन की जाती है, चलाए जा सकने वाले उदाहरण से दिखाई जाती है, मापी और स्थिर की जाती है, और उसके बाद ही स्वयं Dlib को प्रस्तावित की जाती है।

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

24
उदाहरण प्रोग्राम, हर व्यवस्था के लिए एक

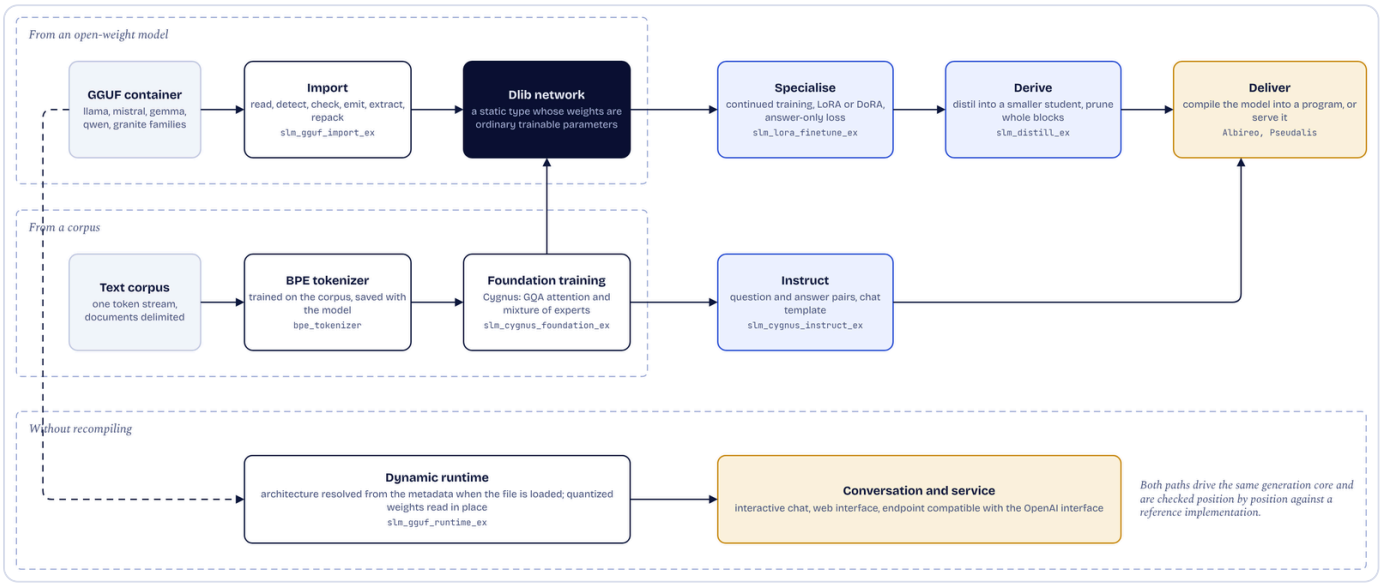
14
योगदान, जो Dlib में पहले ही मर्ज हो चुके हैं

12
मॉडल आर्किटेक्चर, GGUF फ़ाइलों से पढ़े गए

0
Python निर्भरता, पूरी शृंखला में

मॉडल के पूरे जीवन के लिए एक ही लाइब्रेरी

ओपन-वेट मॉडल अपनी GGUF फ़ाइल के ज़रिए प्रवेश करता है और एक सामान्य Dlib नेटवर्क बन जाता है; हमारा अपना मॉडल एक कॉर्पस से शुरू होता है। इसके बाद दोनों एक ही राह चलते हैं: विशेषज्ञ बनाना, व्युत्पन्न करना, पहुँचाना। एक दूसरा इंजन किसी भी समर्थित मॉडल को दोबारा कंपाइल किए बिना चलाता है।



लाइब्रेरी का दायरा

विज़न और मल्टीमोडैलिटी
Vision Transformer, इमेज बैकबोन के रूप में या टेक्स्ट और चित्र पढ़ने वाले मॉडल के विज़न टावर के रूप में, एंड-टू-एंड प्रशिक्षण के साथ।

व्यक्तिगत डेटा की सुरक्षा
नियम और छोटे प्रशिक्षित रिकग्नाइज़र पाँच भाषाओं में व्यक्तिगत डेटा खोजते हैं; फ़ॉर्मेट-संरक्षित एन्क्रिप्शन उसे ऐसे बदलता है कि वापस लाया जा सके।

NPU और इंटीग्रेटेड GPU
हर टेंसर ऑपरेशन उस डिवाइस पर चलता है जो माप में सबसे तेज़ निकला: होस्ट, Intel NPU या OpenVINO के ज़रिए इंटीग्रेटेड GPU।

कार्ड पर क्वांटाइज़्ड वेट
CUDA कर्नेल, GGUF फ़ॉर्मेट में रखे वेट को बिना फैलाए एक्टिवेशन से गुणा करते हैं।

खोज
एम्बेडिंग मॉडल दस्तावेज़ों के फ़ोल्डर को इंडेक्स करते हैं और प्रश्न का उत्तर उससे मेल खाते अंश से देते हैं।

विस्तारित GPU मेमोरी
V²RAM ब्लॉकों को कार्ड, होस्ट मेमोरी और डिस्क स्टोर के बीच प्रवाहित करता है, ताकि कार्ड से बड़ा मॉडल भी उसी पर प्रशिक्षित हो सके।

प्रोसेसर-विशिष्ट कर्नेल
AVX2, AVX-VNNI और AVX-512 VNNI के कोड पथ रन टाइम पर प्रोसेसर की क्षमता के अनुसार चुने जाते हैं।

पूर्वानुमान ही संपीड़न
एक प्रशिक्षित नेटवर्क और एक अरिथमेटिक कोडर मिलकर ऐसा लॉसलेस कंप्रेसर बनाते हैं जिसे कोई भी प्रोसेसर बिट-दर-बिट दोहराता है।

और जानकारी:
cydral-technology.eu/hi/research/library/

सोर्स कोड
github.com/Cydral/Dlib-Transformer-extensions

संपर्क
cydral-technology.eu/hi/contact/

Boost Software License 1.0 के अंतर्गत निःशुल्क और ओपन सोर्स। © 2026 Cydral Technology, फ़्रांस।