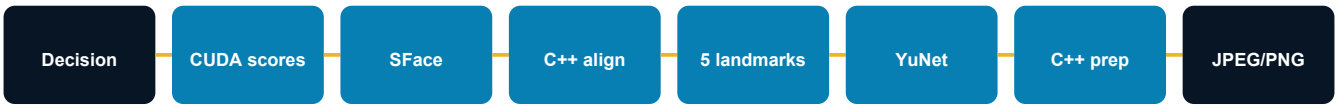
21.08.2026 - ינכדעו ירוקמ שומימ

CUDA-בו CPU ++C-ב סינפ יוהיז

CPU ויב תיתימאה הקולחה, 500-batch ו 50 batch, הדיחי הצרה: Hive לש ליעפה דוקל רצק ינכט ץירדמ הדדמנש הצאההו GPU-ל.

תחא השקב לש לעופב לולסמה



מאתומ kernel-score ו ONNX Runtime CUDA ץרד YuNet/SFace :GPU | הטלחהו רושיי YuNet, חונעפ ,הנכה ,חונעפ :CPU

החיכומ המגדהה המ

תחא הנומת לע תותשרה תא הליעפמ CPU תסרג .מיסוחיי מתואו תונומת ץתוא לע YuNet + SFace תמישמ התוא תועצבמ תואסרגה יתש תצירטמ תבשחמ וכמ רחאלו, מייירוקמ ONNX Runtime CUDA sessions לא תונומת B לש אלמ batch תחלוש CUDA תסרג .מעפ לכב ץרותמ מאתומ kernel תועצמאב B x N ץוימד sface_cuda.cu.

identity_cpu.exe מה מיליעפה בושיחה יכילהת .HTTP/Ursina רובע הרובעת תבכש ראשנ אוה .תשרה תא ץירמ וניא Python ו-identity_cuda.exe, ונבנש מ-++C מו-CUDA C++.

לעופב דבוע GPU וכיהו דבוע CPU וכיה

השוע דוקה המ	וקתה	בלש
.RGBA-ל JPEG/PNG חנעפמ image_io.cpp	CPU / C++	הנומת תניעט
.padding-ו NCHW BGR רוזנט רצוי prepare_detection_batch	CPU / C++	YuNet תנכה
.B dynamic batch רובע CUDAExecutionProvider-ב שמתשמ Ort::Session	CUDA תסרגב GPU	YuNet inference
.וויצ תודוקנ שמחו מינפ רחובו YuNet יטלפ 12 ארוק דוקה	CPU / C++	מייוהיז חונעפ
.bilinear המיגדו similarity transform בשחמ align_faces	CPU / C++	112 x 112 רושיי
.batch-ה לכל 128D לדוגב embedding רצוי ינשה Ort::Session	CUDA תסרגב GPU	SFace inference
.1 ררואל למרונמ 128D רוטקו לכ	CPU / C++	לומרנ
.מייולת יתלב וויצ יבושיח B x N רגשמ sface_cuda.cu	GPU / CUDA C++	Cosine scores
.accepted-ו margin, score, runner-up, identity מירחבנ	CPU / C++	JSON-ו ספ

embeddings לש תינומהה האוושההו YuNet/SFace מידבכה תשרה יפרג .GPU-ב ער בלש לכש ועוט וניא יחכונה דוקה :בושח CPU. לע ירוקמ C++ מיראשנ תיפוס הטלחהו רושיי, הנכה; מיצאומ

500-ו 50 ,1 רובע דוק ותוא

תועמשמ	לעופב הצרה	Batch
.GPU-ה תא תאלממ הניא ויידע תויליבקמה; ילמינימ היהשה ומזו תונוכנ קדוב	CUDA provider ررד דיחי לולסמ	1
.מינוויצ N x 50 בשחמ kernel-ה זאו, תחא השקבב מיטירפ 50 מילבקמ YuNet/SFace	B=50 דיחי dynamic batch	50
תויוליפכ; תשרב מירבוע תומוקמה 500 לכ, 10 לופכ 50 frames-ב מישמתשמ רשאכ תוגלודמ וניא.	B=500 דיחי dynamic batch	500

CUDA תויליבקמ שוריפ המ

תויצולובנוק קרפמ ONNX Runtime .clock ותואב תיזיפ תועצובמ תונומתה 500 לכש חיטבמ וניאו תודרפנ תוינכות 500 וניא Batch 500 :ידממ-וד grid רגושמ מאתומה score kernel-ב. תונימז SM תודיחי ויב threads יפלא קלחמ GPU-הו, CUDA kernels-ל הצירטמ תולועפו .וקת thread לכל דחא score מע, מיסוחיי לש מיקולב ceil(N/256) לופכ התליאש תורוש B

B קוידב ריזחמ worker-ה, מיביתנ B רצוי benchmark_native.ps1. תומישמ רשע תראשנ מימעפ רשע הנומת התוא לע הרזח מימאות מניא expected identity וא count מא תלשכנ הקידבהו, תואצות.

תויזכרמ תודוקפו סיליעפ סיצבק

קיודמ דיקפת	הדוקפ	ץבוק
.SFace-ו YuNet לש sessions-ל CUDAExecutionProvider רבחמ	`options.AppendExecutionProvider_CUDA`	sface_engine.cpp
.אלמ dynamic batch רובע ONNX יפרג ינש תא ץירמ	`yunet.Run(...) / sface.Run(...)`	sface_engine.cpp
.האירק לכל דחא CPU: frame-ב יתרדס baseline הנווכב רצוי	`for (...) embed_images({images[index]})`	sface_engine.cpp
לדוגב references-ו Bx128 לדוגב queries הלעמו ,קתה ירגאמ הצקמ Nx128.	`cudaMalloc / cudaMemcpy H2D`	sface_cuda.cu
.רגשמ kernel מאתומ: block=256, grid=(ceil(N/256), B).	`cosine_scores_kernel<<<grid, block>>>`	sface_cuda.cu
query/reference גוזל multiply-add תולועפ 128 עצבמ דיחי thread דחא.	`dot += query[d] * reference[d]`	sface_cuda.cu
.GPU-ה ירגאמ תשולש לכ תא ררחשמו BxN ינויצ ריזחמ	`cudaMemcpy D2H / cudaFree`	sface_cuda.cu
.Hive תושקב ,יב מינועט סוחיי לש embeddings-ו מילדומ ריאשמ	`--serve-tsv`	sface_identity_cli.cpp

הז בשחמב תתמואמ האוושה

לכל מדקומ מומיחו, מיסוחייו תונומת, מילדומ מתוא, מייוקמ Release יצבוק. Intel Core i9-14900K + NVIDIA GeForce RTX 4060 תרוצ batch.

תונוכנ	הצאה	C++/CUDA batch	יתרדס C++ CPU	סמוע
50/50	2.31x	2840.1 ms	6551.4 ms	50
500/500	6.97x	9365.7 ms	65304.2 ms	500

identity_cpu.exe-ל קר תסחייתמ הלבטה .ורסוה מדוקה PDF-המ Python/PyTorch לש מינשיה מירפסמה תוינש 9.37-כל 65.3-כמ ,מזה תא הניטקמ CUDA batch 500-ב .מייחכונה identity_cuda.exe-לו.

batch סע לדג ורתיה עודמ

רבסה	סרוג
.תמדוקה ירחא הליחתמ האבה היצרטיאה ;SFace-ו רושיי ,YuNet ךרד דרפנב רבוע frame לכ	CPU baseline
.תונומת רתוי ויב תקלחתמ runtime-ו kernels רוגיש תולעו ,GPU-ה תא רתוי בוט אלממ לודג רוזנט	CUDA batch
.ויקת גוז לכל דרפנ thread הצקמ CUDA ;סייולת יתלב תוגוזה B x N לכ	Score kernel
.הדיחי הנומתל רתוי ריהמ תויהל בייח וניא GPU ;חוורה תא סיתיחפמ ץטק batch-ו חראמה דצב רושיי ,ורכיז תורבעה	הלבגמ

רזחשל דציכ

תיינב Release: .\setup_windows.ps1 -CudnnBin PATH_TO_CUDNN_BIN	1
.9 cuDNN-ו CUDA runtime, CUDA DLL, ONNX Runtime identity_cuda.exe, identity_cpu.exe סימייקש אדוול שי	2
.\benchmark_native.ps1 -RuntimePaths PATH_TO_CUDNN_BIN :וחבמ ותוא תצרה	3
.images_per_second-ו recognition_ms, errors, correct, batch קודבלו results/native_face_cpu_cuda_results.csv חותפל שי	4
.Python-ב inference עצבמ וניאו ץרשמתמו ירוקמ worker ליעפמ AI_MIPS_NATIVE_FACE=1; server.py ריאשהל שי Hive רובע	5

טקיורפה תנגהל סוכיס

לע תסוסבמ הניא האצותה .המישמ התוא רובע batch-ב C++/CUDA לומ יתרדס C++ CPU :תרקובמ האוושה גיצמ טקיורפה תיזכרמה הביסה .סייוהיז 500 רובע 6.97x-ו 50 רובע 2.31x איה הצאהה RTX 4060-ב .תובושת לש cache לע וא תויוליפכ גוליד cosine. ינויצ B x N לש יולת יתלב יליבקמ בושיחו CUDA לע batch-ב YuNet/SFace תצרה איה