Implementation of a Fast AI Framework Using Vulkan
- Case Study on ailia SDK -

2021.04.23 ax Inc.
<table>
<thead>
<tr>
<th><strong>Name</strong></th>
<th>ax Inc.</th>
</tr>
</thead>
</table>
| **Location** | Tokyo, Japan  
19-22 Uguisudanicho Shibuya-ku  
Tosokaikan 4F |
| **Name/Title of representative** | TERADA Takehiko, President and CEO |
| **Business activities** | Consulting related to AI (development of AI applications, etc.),  
Development and sale of framework related to AI,  
Sale of middleware (AXIP) |
| **Capital** | 300 million yen (including capital reserve) |
| **Accounting period** | March 31 |
| **Date of establishment** | May 22, 2019 |
ailia SDK

ailia SDK is an AI framework leveraging the GPU to achieve high-performance AI inference.

It supports ONNX (opset 10 & 11) and enables high-performance inference using Vulkan.

It offers over 120 pre-trained models, ready to be integrated into our client’s application.

https://ailia.jp/
Why use ailia SDK?

**High-Performance Inference**
- Optimized for real-time AI inference
- Tuned for every platform to perform fast inference

**Cross-Platform**
- Compatible with servers / desktops / mobile devices / single-board computers
- On demand support for FPGAs, AI chips or embedded devices

**Multi-language**
- Available in C++ / C# (Unity) / Python / JNI
- Incoming support for Swift / Kotlin

**Simple Development Flow**
- Unified API for all platforms and using a single AI model format (ONNX)
- Emphasis on long-term support and backward compatibility.
Example of reduction of development cost by ailia SDK

Conventional
- search optimal model
- clone model repository
- setup development environment
- setup edge environment
- implement pre and post processing
- result verification
- install dependent libraries
- NG
- NG
- NG

ailia SDK
- ailia SDK install
- ailia MODELS download model
- ailia MODELS execute sample
- result verification
Case Study

CLIP STUDIO PAINT for Galaxy leverages ailia SDK

High-performance inference using GPU on Android


Launch of the “ailia AI showcase” demo

Adoption by UNIT.COM Inc. in their AI development/Deep Learning PC and workstations


AXELL and Blaize Collaborate to Integrate Blaize AI Processing Technology into AXELL ailia AI Framework

Use Case 1

ailia SDK that realizes high-speed deep learning inference is adopted in the CLIP STUDIO PAINT.

This time, the ailia SDK has been adopted as the AI execution library for CLIP STUDIO PAINT's “Smart Smoothing” function. “Smart smoothing” is a function that reduces the roughness (jaggy) caused by changing the resolution or enlarging the image and converts it into a beautiful image. When low-resolution illustrations or manga drawn for the Web are changed to a print size to make a book, the image becomes rough, or lines are blurred when the line being drawn is deformed (enlarged) It can be used for such cases.
Use Case 2

Launch of the “ailia AI showcase” demo

DEEP∞ is a series of PCs dedicated to AI development that comes with an "NGC" setting service that allows you to use the latest development environment locally. You can feel free to experience AI functions with DEEP∞, which is used by many users for AI development and deep learning.

In addition, UNITCOM Inc. has also started handling the "ailia SDK", and with the cooperation of both companies, we will provide total solutions related to machine learning such as framework provision and introduction support to meet the development needs of various AI applications. It is also possible to provide it.
Use Case 3

AXELL and Blaize Collaborate to Integrate Blaize AI Processing Technology into AXELL ailia AI Framework

• At CES 2020 (South Hall 2, Booth #25332), Blaize will showcase AXELL ailia™ AI solutions running on the Blaize Graph Streaming Processor™ (GSP) architecture

• Combining low latency with low energy consumption, Blaize GSP processes AI workloads at the system edge more efficiently than existing processing solutions

• Commercially available AXELL AI solutions using Blaize GSP technology are planned for the second half of 2020 from AXELL CORPORATION and ax Inc, a subsidiary company of AXELL CORPORATION.
ailia SDK Architecture

ailia MODELS

- ONNX (opset=10, 11) (supporting over 100 layers)

ailia SDK

- API (C++, Python, C#, JNI)
- Runtime Graph Optimization
- Accelerator (Convolution, Pooling, Resize, Add, etc.)

Vulkan, Metal, AVX / NEON
Benefits of Vulkan

Support for all major GPUs

GeForce, Radeon, IrisGraphics, Mali, Adreno,...

Support for all major OSs

Windows, Android, Linux (Mac, iOS support via Metal)

Easy installation for GPU inference

Being widely used for gaming, it only requires standard drivers to run

Little additional disk space usage

ailia_vulkan.dll is a mere 2.8MB compared to 444MB for cudnn_cnn_infer64_8.dll
ailia AI showcase

High-performance AI inference demo using Windows (RTX2080) + Vulkan

Fast inference through Vulkan can be achieved for all kinds of AI models

Demo developed in Unity and runs on Windows, Mac, iOS, Android, Linux

Object Detection: YOLOv3-tiny
Face Recognition: YOLOv3 Face
Semantic Segmentation: HRNetV2
Hair Segmentation: HairSegmentation
Crowd Counting: CrowdCounting
Object Recognition: ResNet50
Pose Estimation: LightWeightHumanPose
Image Super-Resolution: SRResNet
Image Denoising: Noise2Noise
Person Re-Identification: VGGFace2
AILIA MODELS

Over 120 models compatible with the ailia SDK are publicly available on github.

You can easily try out the latest models such as YOLOv4, MIDAS and PaddleOCR.

https://github.com/axinc-ai/ailia-models
ailia MODELS has samples for Python, C++ and Unity.

You can use accelerated inference using Vulkan in any environment.
AI Inference Acceleration with Vulkan

Fast AI inference achieved using Runtime Graph Optimization and Layer Fusion

Implementation of layers such as Convolution or Pooling using Vulkan’s Compute Shader

Implementation of the Winograd Algorithm with shaders to accelerate the heavy Convolution layer
The Winograd Algorithm

The arithmetic complexity of the convolution layer can be reduced by applying transforms to the inputs and the output.

Although the transforms are a little demanding, the matrix multiplication which takes up most of the computation time can be reduced.

https://arxiv.org/abs/1509.09308

For a 3x3 convolution yielding 4 outputs, the 36 multiply-accumulate operations can be reduced to 16.

Diagram:
- Filter weights
- Transform
- Image
- Transform
- Matrix multiplication
- InvTransform
Matrix Multiplication Optimization

A GEMM block size is $M=N=K=4$

In some architectures (e.g. Adreno), $M=8$

Computation of the matrix product for each block in every Invocation (equivalent to CUDA’s Thread)

No synchronization of workgroup

16-bytes memory alignment and reading as vec4 in a single instruction

Memory access optimization by keeping the vector format as much as possible in later processing
Command Buffer Management

The overhead of vkQueueSubmit being large, sub threading is used to reduce the load

1. Start a new thread to process the command buffer
2. When a request is made through ailia’s API, the command buffer is added to the sub thread’s queue
3. If the sub thread is idle, the command buffer queue is sent all at once to vkQueueSubmit
Summary

High-performance AI inference can be achieved using Vulkan regardless of which device or OS and with only standard drivers.

Easy integration into any application through ailia SDK.

Fast inference using Vulkan has already been implemented for more than 120 AI models.
ax Inc. provides total solutions related to AI, from consulting to model, SDK, AI applications / systems development and support. Should you have any inquiry, feel free to get in touch with us.

ax Inc. home page  https://axinc.jp/ (Inquiry)

ax Inc. BLOG  https://medium.com/axinc

ailia SDK https://ailia.jp/ (Free trial version available)

ailia MODELS https://github.com/axinc-ai/ailia-models

ailia AI showcase (Video) https://www.youtube.com/watch?v=lRnWX1rDRQU