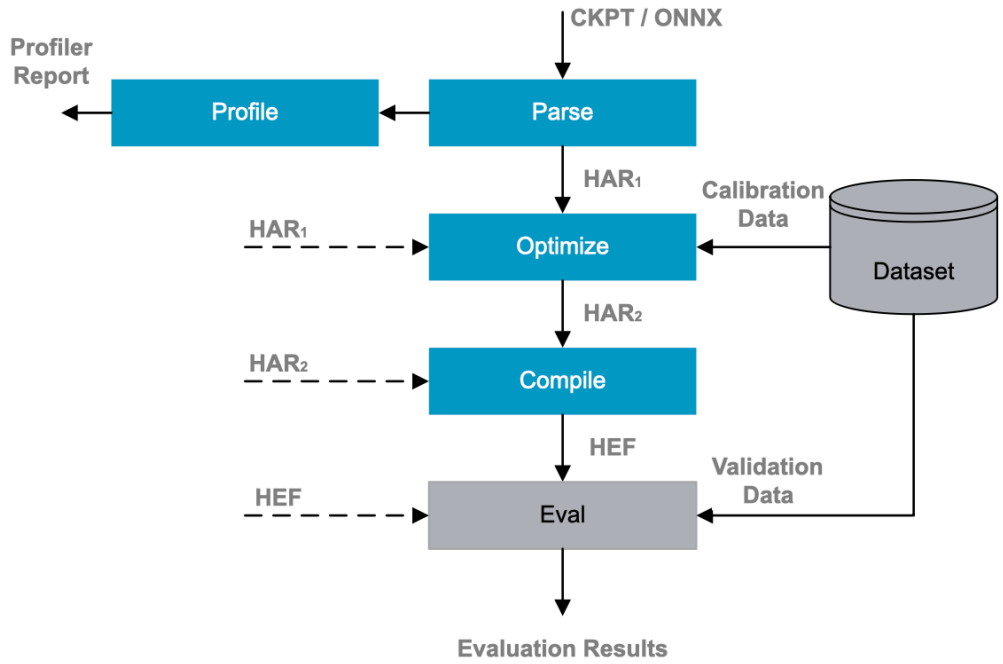
Smart Digital Junction

Running custom models on the Hailo-8 AI Accelerator

Full Conversion Process

Flow Diagram

The following scheme shows high-level view of the model-zoo evaluation process, and the different stages in between.



NOTE: Hailo Model Zoo provides the following functionality for Model Zoo models only. If you wish to use your custom model, use the Dataflow Compiler directly.

Full Conversion Process

Following this project:

1. **Model Parsing:** Converting from ONNX to HAR (these models also supported)
    1. Parses the ONNX model into Hailo's internal representation
    2. Generate the Hailo Archive (HAR) file.
2. **Model Optimization:** HAR to Quantized HAR
    1. Convert HAR from full precision into integer representation
    2. Generate a quantized Hailo Archive (HAR) file.
    3. This includes the model input normalization and the non-maximum suppression (NMS) on the model output.
    4. Requires calibration dataset (numpy-specific binary format) for non-HMZ models
3. **Model Compilation:** Quantized HAR to HEF
    1. Compile the quantized Hailo Archive (HAR) and generate the Hailo Executable Format
4. **Model Execution**
    1. Run inference on Hailo hardware (with a compatible HailoRT version)
    2. Models need to be compiled for the specific hardware (current standard is H8L, we have H8)

Current Issues

1. **Version incompatibility**
    1. Hailo Dataflow Compiler only runs on Linux x86 (officially only Ubuntu)
    2. Some dependencies only work on DFC v3.27.0 (latest is 3.29.0)
    3. Some dependencies only work on python v3.9 and below (most Linux distros come with 3.12)
2. **Rapid Development**
    1. Major changes being made to Hailo demo programs
        1. Outdated demo programs throw errors (<1 month of commit history)
    2. Most likely significant changes being made to underlying Hailo APIs
        1. APIs seem to be highly coupled
3. **Documentation issues**
    1. Decentralized documentation
        1. Some span multiple repos in GitHub, some only available on HDZ, some unavailable
    2. Existing documentation not in-depth enough
    3. Configuration specific issues not well documented
        1. Some ad-hoc fixes provided by the mods on HDZ
        2. Some reported (and solved) by users on HDZ

Release versions compatibility for Accelerators

AI SW Suite	Dataflow Compiler	HailoRT	Integration Tool	Model Zoo	TAPPAS
2024-10	v3.29.0	v4.19.0	v1.19.0	v2.13.0	v3.30.0
2024-07.1	v3.28.0	v4.18.0	v1.18.0	v2.12.0	v3.29.1
2024-07	v3.28.0	v4.18.0	v1.18.0	v2.12.0	v3.29.0
2024-04	v3.27.0	v4.17.0	v1.17.0	v2.11.0	v3.28.0
2024-01	v3.26.0	v4.16.0	v1.16.0	v2.10.0	v3.27.0
2023-10	v3.25.0	v4.15.0	v1.15.0	v2.9.0	v3.26.0
2023-07.1	v3.24.0	v4.14.0	v1.14.1	v2.8.0	v3.25.0
2023-07	v3.24.0	v4.14.0	v1.14.0	v2.8.0	v3.25.0
2023-04	v3.23.0	v4.13.0	v1.13.0	v2.7.0	v3.24.0
2023-01.1	v3.22.1	v4.12.1	v1.12.0	v2.6.1	v3.23.1
2023-01	v3.22.0	v4.12.0	v1.12.0	v2.6.0	v3.23.0
		v4.11.0		v2.5.0	v3.22.0
2022-10	v3.20.0	v4.10.0	v1.10.0	v2.4.0	v3.21.0
		v4.9.0			v3.20.0

Pre-trained Hailo Models

Task Type	Hailo-8	Hailo-8L	Hailo-15H	Hailo-15M	Hailo-10
Classification	Link	Link	Link	Link	Link
Object Detection	Link	Link	Link	Link	Link
Semantic Segmentation	Link	Link	Link	Link	Link
Pose Estimation	Link	Link	Link	Link	Link
Single Person Pose Estimation	Link	Link	Link	Link	Link
Face Detection	Link	Link	Link	Link	Link
Instance Segmentation	Link	Link	Link	Link	Link
Depth Estimation	Link	Link	Link	Link	Link
Facial Landmark Detection	Link	Link	Link	Link	Link
Person Re-ID	Link	Link	Link	Link	Link
Super Resolution	Link	Link	Link	Link	Link
Face Recognition	Link	Link	Link	Link	Link
Person Attribute	Link	Link	Link	Link	Link
Face Attribute	Link	Link	Link	Link	Link
Zero-shot Classification	Link	Link	Link	Link	Link
Stereo Depth Estimation	Link	NA	NA	NA	NA
Low Light Enhancement	Link	Link	Link	Link	Link
Image Denoising	Link	Link	Link	Link	Link
Hand Landmark detection	Link	Link	Link	Link	Link
Zero-shot Instance Segmentation	Link	Link	Link	Link	Link

Supported Models

COCO

Network Name	mAP	Quantized	FPS (Batch Size=1)	FPS (Batch Size=8)	Input Resolution (HxWxC)	Params (M)	OPS (G)	Pretı
centernet_resnet_v1_18_postprocess	26.37	1.41	366	366	512x512x3	14.22	31.21	downl
centernet_resnet_v1_50_postprocess	31.77	2.54	78	146	512x512x3	30.07	56.92	downl
damoyolo_tinynasL20_T	42.8	0.59	130	309	640x640x3	11.35	18.02	downl
damoyolo_tinynasL25_S	46.53	1.22	228	228	640x640x3	16.25	37.64	downl
damoyolo_tinynasL35_M	49.7	1.86	61	127	640x640x3	33.98	61.64	downl
detr_resnet_v1_18_bn	33.91	2.43	29	75	800x800x3	32.42	61.87	downl
detr_resnet_v1_50	35.38	0.4	10	20	800x800x3	41.1	120.4	downl
efficientdet_lite0	27.32	0.78	90	250	320x320x3	3.56	1.94	downl
efficientdet_lite1	32.27	0.45	62	164	384x384x3	4.73	4	downl
efficientdet_lite2	35.95	1.2	43	107	448x448x3	5.93	6.84	downl
nanodet_repvgg ⭐	29.3	0.67	820	820	416x416x3	6.74	11.28	downl
nanodet_repvgg_a12	33.73	2.24	400	400	640x640x3	5.13	28.23	downl
nanodet_repvgg_a1_640	33.28	0.34	280	280	640x640x3	10.79	42.8	downl
ssd_mobilenet_v1 🚀 ⭐	23.19	0.77	1015	1015	300x300x3	6.79	2.5	downl
ssd_mobilenet_v2	24.18	1.16	140	358	300x300x3	4.46	1.52	downl
tiny_yolov3	14.66	0.25	1044	1044	416x416x3	8.85	5.58	downl
tiny_yolov4	19.18	1.37	1299	1299	416x416x3	6.05	6.92	downl
yolov10b	52.0	0.85	29	67	640x640x3	20.15	92.09	downl
yolov10n	38.5	1.38	166	427	640x640x3	2.3	6.8	downl
yolov10s	45.86	0.7	96	210	640x640x3	7.2	21.7	downl
yolov10x	53.7	1.75	16	32	640x640x3	31.72	160.56	downl
yolov3	38.42	0.04	31	47	608x608x3	68.79	158.10	downl
yolov3_416	37.73	0.16	47	95	416x416x3	61.92	65.94	downl
yolov3_gluon	37.28	1.5	36	57	608x608x3	68.79	140.7	downl

Moving Forward

- Reaching out to Hailo directly for more information about their APIs
- ML expertise for unpredictable errors that require ML-knowledge to debug