


ELT YAML Description

TREX-TS

Exported on 11/29/2024

Table of Contents

No headings included in this document

 Release **7.11.22.21** based information

Converter Section (important: BOLD)

key	type	explanation	values for Root	Comments
activation_bitwidth	[8, 16]	Bitwidth of intermediate feature map(A)	8	<p>asymm에서만 사용되며 fp16에서는 무시됨</p> <p>액티베이션 값의 비트 너비 설정. 모델 연산 효율성과 정확성 사이의 트레이드오프 관리</p>
debug	bool	dump layerwise sqnr between new snc and hw quantized snc	false	<p>레이어별 SQNR(Signal-to-Quantization-Noise Ratio) 덤프. 레퍼런스 snc와 하드웨어 양자화 snc 사이의 차이 표시</p> <p><목적></p> <ul style="list-style-type: none"> • 양자화 손실이 큰 레이어 식별 • 전체 양자화 품질 평가 • 최적화 방향 설정(특정 레이어 정밀도 향상, 추가 양자화 기법 적용)
input_dtype	float32, float16, uint8, uint16, none	You can set model input datatype as float32, uint8(only Asym)(if none take from model)	float32	<p>float32/float16 정밀도 높고 원본 특성 유지</p> <p>uint8은 비대칭 양자화(양자화 범위가 0을 중심으로 대칭이 아님)만 지원</p>
output_dtype	float32, float16, uint8, uint16, none	You can set model output datatype as float32, uint8(only Asym)(if none take from model)	float32	<p>float32/float16 정밀도 높고 원본 특성 유지</p> <p>uint8은 비대칭 양자화(양자화 범위가 0을 중심으로 대칭이 아님)만 지원</p>

key	type	explanation	values for Root	Comments
profile_batchsize	unsigned int	Batch size for profile (value 100 is recommended)	2	<p>asymm에서만 사용되며 fp16에서는 무시됨</p> <p>이 배치 크기를 기준으로 모델의 처리 속도를 측정. 예로 100이면 한번에 100개의 샘플을 처리할 때의 성능 지표를 계산</p>
snc_converter	bool	True, convert old snc to new snc, set it to false when input is new snc	true	
test_vector_gen	bool	Enable layerwise test vector generation after quantization.	true	
tv_input	path	Input data file path for test vector generation (default path is {workspace}/DATA/database.h5)	DATA/ database. h5	
use_randomdb	bool	Use randomdb for profiling data set	true	<p>FP16은 true로 고정</p> <p>false는 asymm에서 사용</p> <ul style="list-style-type: none"> • true 랜덤 데이터를 기반으로 프로파일링 수행 • false 실제 데이터셋으로 프로파일링 수행
userdb	path	Profiling data set path (default path is {workspace}/DATA/database.txt)	DATA/ database. txt	
device	Gen-2, Gen-2a, Gen-3, Gen-3DSP, Gen-4, Gen-4DSP, Gen-4Multi(only for fp16), Gen-5, Gen-5a, Gen-5b, Gen-6, Gen-7	Soc type Gen-6: Root	Gen-6	

key	type	explanation	values for Root	Comments
do_quantize	bool	Enable quantization	true	
mode	elt, eht_cnnx, eht_snc	conversion mode	elt	현재 버전에서는 elt
onnx_simplify	bool	enable onnx_simplify process	true	<p>true로 고정</p> <ul style="list-style-type: none"> • true: onnx 모델 내 불필요/중복 연산자를 제거하여 모델 구조를 단순화/최적화한다.
optimize	bool	Use graph optimization	true	<p>그래프 최적화를 활성화한다. 모델 계산 그래프를 분석하여 불필요한 연산을 제거하거나 여러 연산을 하나로 병합하는 작업을 통해 효율성을 높인다.</p>

key	type	explanation	values for Root	Comments
quantize_type	symm, asymm, fp16, qat	Select quantization type, quantized model (include caffeQAT) is "qat"	fp16	<p>현재 버전 기준으로는 fp16 only인데 예외적으로 tflite 모델 중에 qat 학습된 것은 qat 적용 필요</p> <p>양자화 방식 옵션으로 다음 유형이 있음</p> <ul style="list-style-type: none"> 대칭 양자화: 양자화 범위가 원점을 중심으로 대칭 (INT8?) ==> 현재 버전에는 이 symm 옵션 없음 비대칭 양자화: 양자화 범위가 원점 중심이 아니고 데이터 최소/최대에 맞춰짐 (UINT8? asymm) FP16 양자화: FP32 대신 FP16 양자화, 메모리 사용량을 절반으로 줄이면서 표현력 유지 양자화 인지 학습 (QAT, Quantization-Aware Training): QAT 되어 있는 모델을 변환.

Compiler Section (important: BOLD)

key	type	explanation	values for Root	comments
assign_dsp	str	Assign specific layer to dsp device	null	
assign_ve	str	Assign specific layer to ve device	null	
best_fit_generalized	bool	Control whether generalized best fit allocation is to be used.	false	

key	type	explanation	values for Root	comments
cluster_execution	str	Only for Solomon		remove "" when applying
convert_to_rshape	str	Only for Solomon		remove "" when applying
debug_str	str	debug str for compiler	null	remove "" when applying
device	Gen-3, Gen-3b, Gen-4, Gen-5, Gen-5a, Gen-5b, Gen-6, Gen-7	System on Chip type	Gen-6	
enable_active_stratum	bool	this is optimization parameter for stratum tiling	false	
enable_hw_cost_based_stratum	bool	this is optimization parameter for stratum tiling based on HW cost	false	
enable_ofm_reuse	bool	Enable the reuse of OFM region for IMFM.	false	
enable_stm	bool	Generate compile log including L1 tiling information	false	
gradual_core_start		Only for Solomon		remove "" when applying
multi_vc	bool	Introduce Multi-VC for OFM, IFM, and weight transfer	true	
multicore	bool	Enable NPU multicore	false	

key	type	explanation	values for Root	comments
schema_version	v2, v3	schema version (need to check the ENN framework version in the target board)	v2	
strict_dfs	bool	this is optimization parameter for tiling scheduling	false	
sync_npu_deps	bool		false	
use_fine_grained_deps	bool	this is optimization parameter for tiling scheduling	false	