Tencent 腾讯 interdigital

JVET-Z0092

EE1-related: additional results on Test 1.1 and Test 1.2 with 8-bit quantization

Liqiang Wang, Xiaozhong Xu, Shan Liu (Tencent)

Franck Galpin (InterDigital)

Outline

- Introduction
- Proposed method
- Results
- Conclusions

Introduction

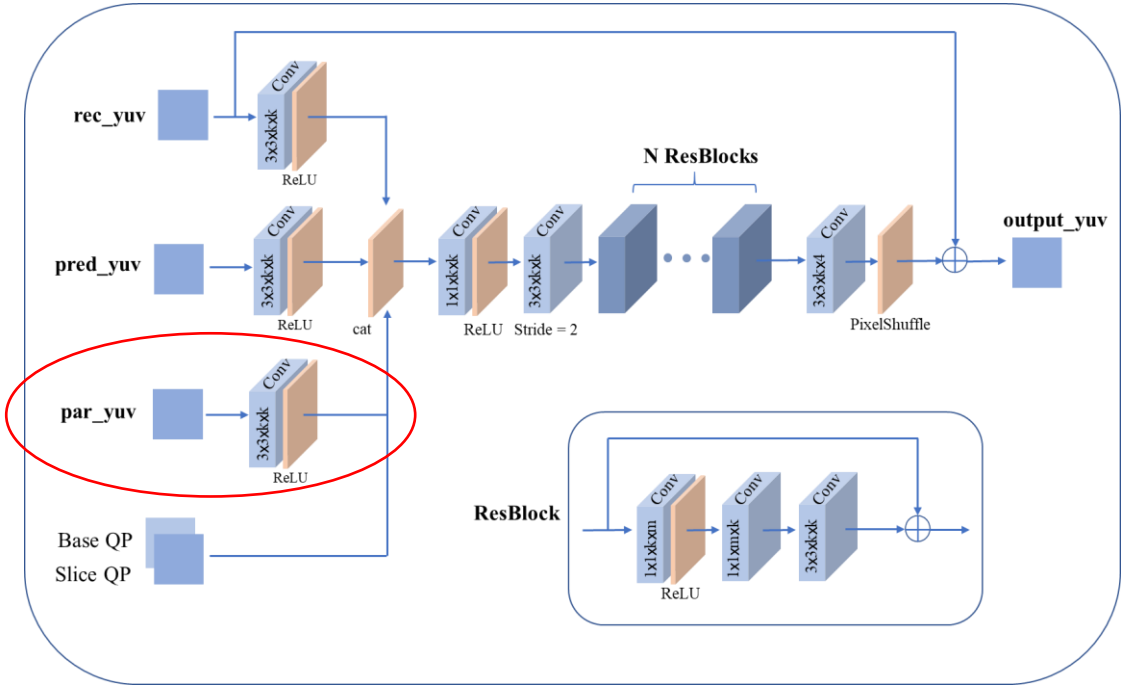
Model quantization is further studied based on EE1-1.1 and EE1-1.2 using 8-bit quantization.

- Only two or even single model is used in the proposed filter design.
- For both I and B slices, the optimal filtered result is decided between the two outputs from the proposed filter and SAO.

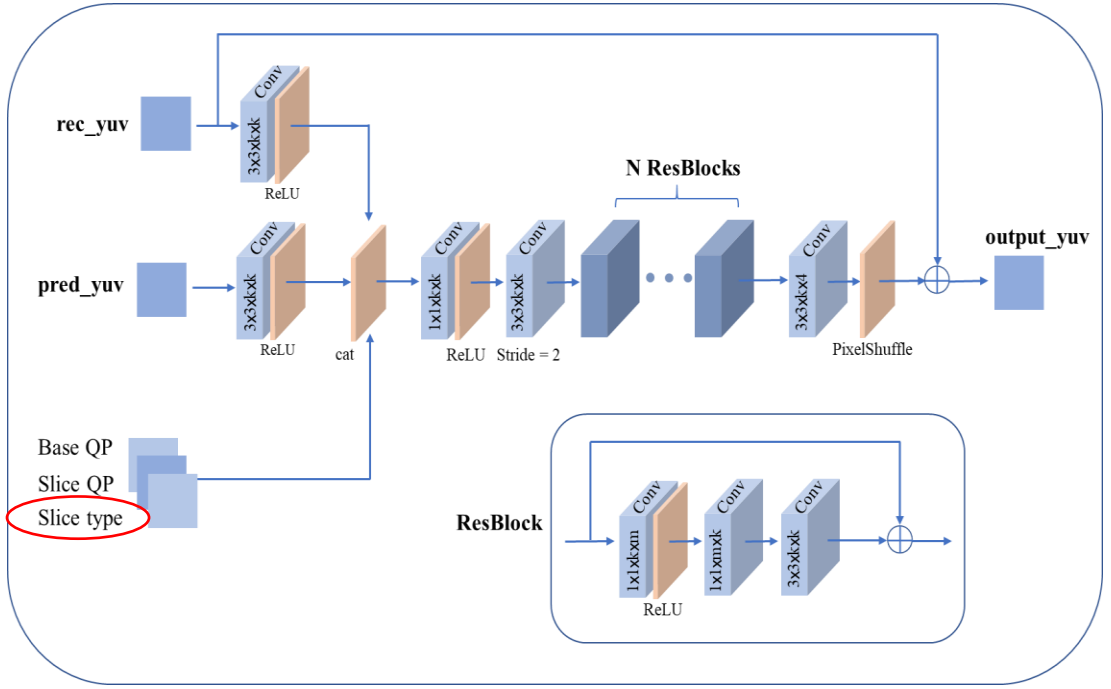
- More side information is utilized, such as the prediction image, the partition image, the slice QP, the based QP and the slice type.
- The main difference of EE1-1.1 and EE1-1.2 with the previous proposals is shown as follows:
  - The convolutional block attention module is removed for the simplicity. It seems to have a tiny influence on the performance.
  - A simpler Nearest method is used to replace Lanczos method in the resampling process for chroma components.
  - Multi-scaling refinement method, like the one in JVET-Y0098, is used to further improve the performance.
  - SADL deployment is studied.

Proposed Method

The network architecture of the proposed method.



The network architecture of the filter 1 with 2 models



The network architecture of the filter 2 with 1 model

The backbone of this network is cascaded by serval Resblocks. As for the structure of Resblock:

- It consists of two 1x1 and one 3x3 convolutional layers.
- The number of channels firstly goes up before the activation layer, and then goes down after the activation layer.

Proposed Method

Network information in the training and inference stages.

Network Information in Training Stage		
Mandatory	GPU Type	Tesla A100 40GB
	Framework:	Torch v1.9.0
	Number of GPUs per Task	1
	Epoch:	~1000
	Batch size:	64
	Loss function:	L1
	Training time:	~150h
	Training data information:	DIV2K, TVD, BVI-DVC
	Training configurations for generating compressed training data (if different to VTM CTC):	
Optional		
	Number of iterations	1200
	Patch size	144x144
	Learning rate:	1e-4
	Optimizer:	ADAM
	Preprocessing:	flipped, rotated
	Mini-batch selection process:	random cropped
	Other information:	

Network Information in Inference Stage			
	Test items	Filter 1	Filter 2
Mandatory	HW environment:		
	GPU Type	CPU only	
	Framework:	SADL	
	Number of GPUs per Task	0	
	Number of Parameters (Each Model)	1.54M	1.9M
	Total Number of Parameters (All Models)	3.08M	1.9M
	Parameter Precision (Bits)	8	
	Memory Parameter (MB)	1.54MB/model, 2 models in all	1.9MB/model, 1 model in all
	Multiplay Accumulate (MAC)/pixel, worst-case	401K	485K
Optional			
	Total Conv. Layers	78	101
	Total FC Layers	0	
	Total Memory (MB)		
	Batch size:	1	
	Patch size	144x144	
	Changes to network configuration or weights required to generate rate points		
	Peak Memory Usage (Total)		
	Peak Memory Usage (per Model)		
	Border handling		
	Other information:		

Proposed Method

Implementation

- For I slices and B slices, Deblock and SAO are both enabled. The optimal filtered result is decided between the two outputs from the proposed filter and the SAO.
- A scaling operation is carried out to refine the result of NN filter. The scaling factors are signaled in the slice header for each component. The specific process is the same as the scaling process used in JVET-W0130. Additionally, there are 3 fixed weights to blend the results of the proposed filter and the conventional filter. These fixed weights are 1, 0.75 and 0.5, which are similar as those in JVET-Y0098.
- The proposed filter can be turned on/off at the CTU level and slice level.

Result of filter 1 with 2 models

The result based on EE1-1.1.1 (Libtorch, flt32)

	All Intra Main10				
	BD-rate Over EE1-1.1.1				
	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	1.37%	4.33%	1.35%	232%	634%
Class A2	1.42%	2.83%	1.20%	184%	622%
Class B	1.04%	1.77%	1.25%	176%	615%
Class C	0.90%	1.02%	0.84%	147%	584%
Class E	1.71%	2.81%	3.60%	191%	623%
Overall	1.24%	2.38%	1.56%	181%	613%
Class D	0.93%	2.50%	1.36%	145%	606%
Class F	0.52%	1.06%	0.57%	140%	588%

	Random access Main10				
	BD-rate Over EE1-1.1.1				
	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	0.48%	1.22%	1.89%	179%	585%
Class A2	0.72%	2.00%	0.83%	174%	575%
Class B	0.61%	1.28%	2.09%	180%	594%
Class C	0.48%	0.60%	1.28%	162%	581%
Class E					
Overall	0.57%	1.23%	1.58%	174%	585%
Class D	0.52%	1.58%	1.39%	168%	593%
Class F	0.23%	0.39%	0.47%	218%	594%

The result based on the EE1 anchor

	All Intra Main10					
	BD-rate Over VTM-11.0_nnvc-1.0					
	YUV-PSNR	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	-7.44%	-5.27%	-11.36%	-16.52%	335%	126037%
Class A2	-7.53%	-5.22%	-15.40%	-13.53%	227%	103488%
Class B	-8.15%	-5.61%	-15.50%	-16.01%	211%	102701%
Class C	-9.62%	-6.97%	-16.47%	-18.68%	164%	77642%
Class E	-10.21%	-8.54%	-15.30%	-15.17%	235%	121950%
Overall	-8.60%	-6.28%	-14.98%	-16.14%	222%	102893%
Class D	-9.19%	-6.93%	-13.92%	-18.05%	160%	84747%
Class F	-5.54%	-3.94%	-9.86%	-10.80%	154%	70550%

	Random access Main10					
	BD-rate Over VTM-11.0_nnvc-1.0					
	YUV-PSNR	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	-10.62%	-8.84%	-14.66%	-17.24%	218%	175435%
Class A2	-10.60%	-8.76%	-17.81%	-14.44%	209%	169634%
Class B	-10.66%	-7.85%	-19.72%	-18.44%	220%	184679%
Class C	-11.32%	-7.88%	-22.10%	-21.24%	185%	168014%
Class E						
Overall	-10.82%	-8.24%	-18.96%	-18.14%	208%	175236%
Class D	-11.77%	-8.79%	-19.94%	-21.44%	192%	170133%
Class F	-5.37%	-3.42%	-11.67%	-10.74%	295%	73414%

Result of filter 2 with 1 model

The result based on EE1-1.2.1 (Libtorch, flt32)

	All Intra Main10				
	BD-rate Over EE1-1.2.1				
	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	0.21%	0.51%	0.91%	241%	631%
Class A2	0.28%	0.19%	-0.16%	192%	607%
Class B	0.14%	0.73%	0.66%	183%	625%
Class C	0.03%	0.70%	0.57%	154%	623%
Class E	0.26%	0.97%	1.65%	200%	643%
Overall	0.17%	0.64%	0.71%	188%	625%
Class D	0.02%	1.08%	1.14%	151%	622%
Class F	0.06%	0.59%	0.13%	145%	639%

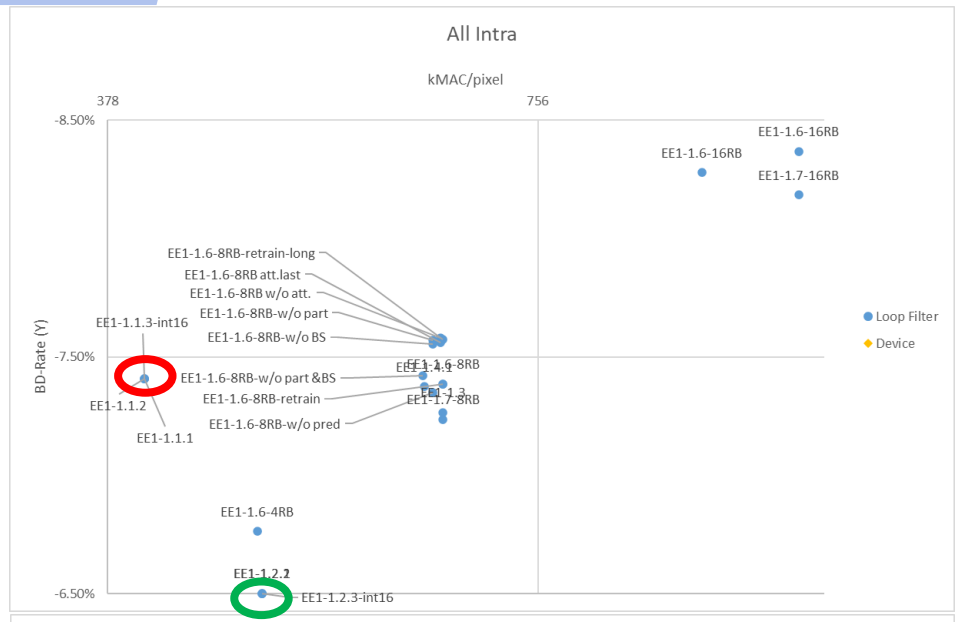
	Random access Main10				
	BD-rate Over EE1-1.2.1				
	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	0.17%	0.36%	1.00%	192%	609%
Class A2	0.09%	0.79%	0.29%	185%	591%
Class B	0.03%	0.82%	0.81%	195%	628%
Class C	0.00%	0.72%	0.53%	174%	621%
Class E					
Overall	0.06%	0.70%	0.67%	186%	615%
Class D	-0.22%	1.28%	1.06%	181%	627%
Class F	0.02%	0.42%	0.20%	238%	640%

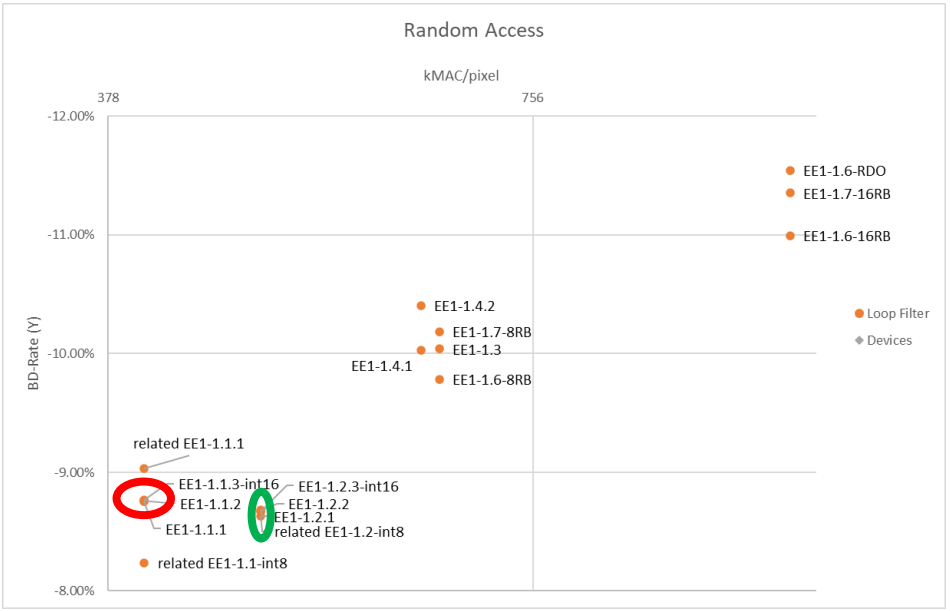
The result based on the EE1 anchor

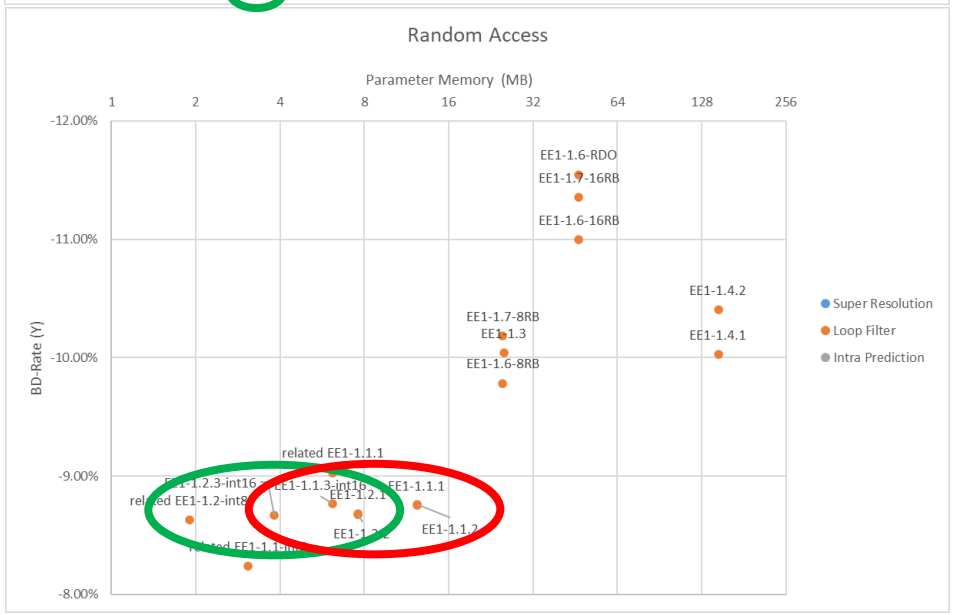
	All Intra Main10					
	BD-rate Over VTM-11.0_nnvc-1.0					
	YUV-PSNR	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	-8.03%	-5.98%	-13.06%	-15.34%	356%	142558%
Class A2	-7.50%	-5.54%	-14.60%	-12.18%	240%	114852%
Class B	-8.01%	-5.73%	-14.21%	-15.54%	221%	113155%
Class C	-8.91%	-6.41%	-15.35%	-17.49%	172%	85730%
Class E	-10.09%	-8.45%	-14.33%	-15.73%	248%	133525%
Overall	-8.48%	-6.34%	-14.36%	-15.41%	234%	113937%
Class D	-8.71%	-6.45%	-14.05%	-16.90%	167%	93771%
Class F	-5.76%	-4.00%	-10.70%	-11.33%	159%	96910%

	Random access Main10					
	BD-rate Over VTM-11.0_nnvc-1.0					
	YUV-PSNR	Y-PSNR	U-PSNR	V-PSNR	EncT	DecT CPU
Class A1	-11.10%	-9.19%	-15.72%	-17.95%	238%	211906%
Class A2	-10.90%	-9.23%	-17.23%	-14.60%	226%	200049%
Class B	-10.88%	-8.27%	-18.36%	-19.06%	241%	220104%
Class C	-11.28%	-8.20%	-20.18%	-20.88%	201%	198411%
Class E						
Overall	-11.04%	-8.63%	-18.09%	-18.43%	226%	208457%
Class D	-11.75%	-9.23%	-18.08%	-20.54%	209%	203885%
Class F	-5.71%	-3.74%	-11.73%	-11.47%	330%	96659%

The trade-off







Filter 1 Filter 2

As for the trade off between MAC and BD-rate

- Under AI, comparable performance can be achieved by filter 1 with saving more than 130K MACs.
- Under RA, similar trade-off is achieved by filter1 and filter 2.

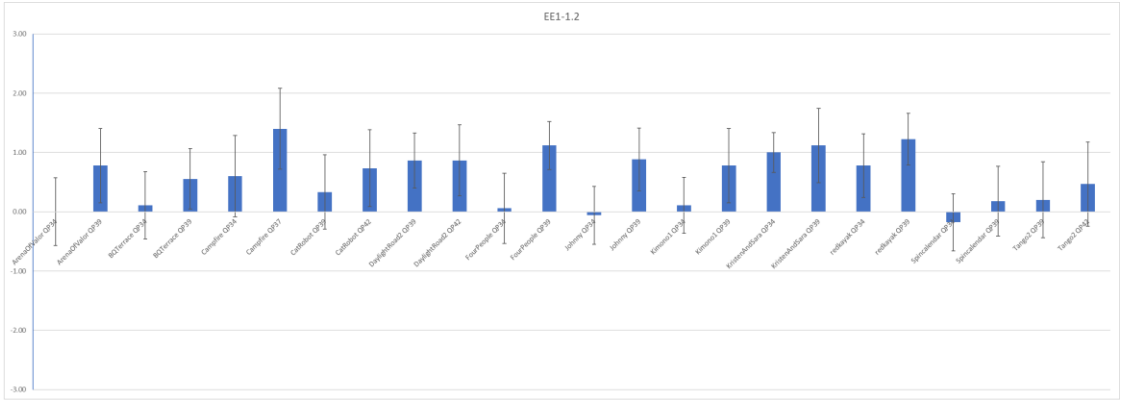
As for the trade off between MAC and BD-rate

- The memory size of parameters are serval times less than other filters.

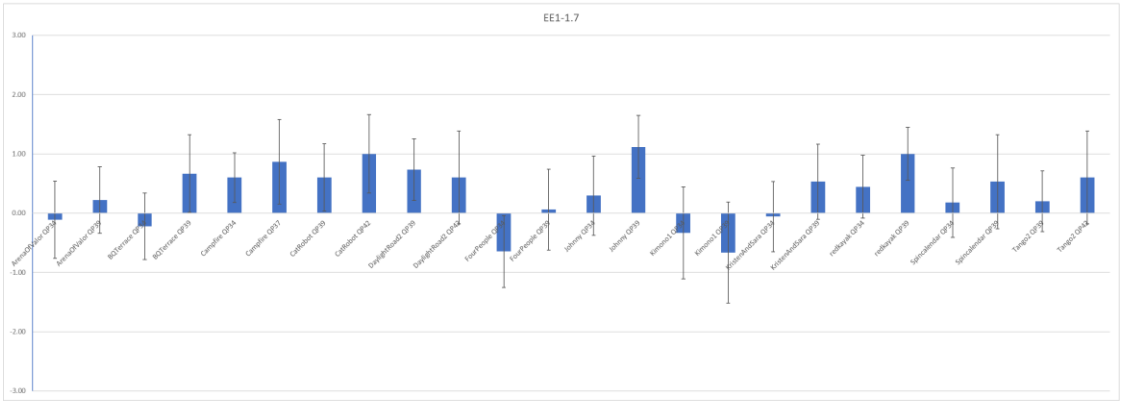
All the data are obtained from the EE1 summary in JVET-Z0023.

The result of subjective test for the filter 2 with 1 model

EE1-1.2 with 1 model



EE1-1.7 with 4 models



For the EE1-2 experiment, a significant visual benefit is reported for 14 out of 24 test cases. For the remaining cases, MOS=0 is included in the confidence interval such that comparable quality is indicated.

In the EE1-1.7 experiment, 9 test cases out of indicate a significant visual benefit while one case showed a significant loss. The remaining 14 cases indicate comparable quality.

It seems that EE-1.2 shows better subjective performance.

All the data are obtained from the REV Result in JVET-Z0053.

Conclusions

- With a constrained memory size and lower complexity, additional results for the two neural network based in-loop filters proposed in EE1 tests are presented in this contribution.
- Good trade-off between performance and complexity can be achieved by applying the proposed method.
- It seems that, compared with the implementation using the float or 16-bit quantization in EE1 tests, the trade-off can be further optimized by the 8-bit quantization, especially for the filter with the single model (the filter 2). That means not all the filters can maintain the performance after the model quantization.
- The filter with the fewer models and the lower MAC is easier to crosscheck training and do some study, because it costs fewer GPU resource or time to train the model.
- Recommend to adopt the proposed filter 1 or filter 2 as the base software to further explore the NN-based tools.

Thanks