公開

研究終了報告書

「（研究課題名）」
研究期間：2019 年 10 月～2023 年 3 月
研 究 者：　孫　鶴鳴

1. 研究のねらい

Video consumes more than 80% of the internet traffic, therefore, video compression technology is very important to reduce the burden of video storage and transmission. The traditional video compression standards have been developed for more than 30 years, and the recent one VVC was standardized in 2020. In the past five years, with the rapid developments of neural network, learned video compression has shown a promising ability on the coding gain. Latest learned codec has outperformed VVC.

Learned codec can be classified into two categories. One is the hybrid scheme which enhances the traditional codec by using neural networks for each component such as intra prediction and in-loop filtering. The other is the end-to-end scheme which is fully based on neural network. Both schemes have shown a great success in the coding gain. Because of the high coding ability, learned codec has drawn attention from both academy and industry. Google launched the workshop and challenge on learned image compression (CLIC) at top-tier conference CVPR from 2018. JPEG has launched a standardization JPEG-AI which will be finalized next year. MPEG is now also trying to enhance the coding gain of VVC by using neural network-based filter.

Though the coding ability is very high, the complexity is too high to realize the real-time processing. Even running on a very powerful GPU, most learned codec consumes much larger time than the traditional codec. In addition to the high complexity, there are many other practical issues remaining unsolved. For example, cross-platform cannot be ensured for learned codec with floating-point arithmetic; multiple models need to be prepared to realize a variable bitrate model, and so on.

The research target is to develop a real-time low-power learned codec system, which can achieve a better coding compression ratio than the existing video compression standard HEVC. Our research content includes both algorithm and architecture development. For the algorithm, we plan to do the network pruning and quantization to reduce the neural network complexity. In addition, we will solve the cross-platform coding problem by quantizing all the arithmetic operation in the network. For the architecture, we plan to develop a specific hardware accelerator to accelerate the processing of learned codec. After designing the architecture, we will do the algorithm-architecture co-design to reach a better trade-off between arithmetical accuracy (i.e. coding gain) and architectural performance (i.e. speed and power). Finally, we plan to build a real coding system which includes both neural computing and entropy coding.

2. 研究成果

（1）概要

Research results can be categorized to three parts: algorithm, architecture and system.

For the algorithm, the developments can be divided to two categories: hybrid scheme and end-to-end scheme. For the hybrid scheme, we developed learned intra prediction [J3,C4] and learned filtering [J2]. For the end-to-end scheme, we developed the world-first learned image compression which outperforms VVC intra [C5]. We proposed a fixed-point arithmetic system to solve the cross-platform coding problem [C3]. In addition, we applied the network quantization to learned codec [J1]. For the architecture, we developed an FPGA accelerator with fine-grained pipeline. [C2]. Finally, we built a real-time codec system in [C1].

（2）詳細

Learned intra prediction [J3, C4]

Intra prediction is an essential component in the image coding. We proposed an intra prediction framework completely based on neural network modes (NM). Each NM can be regarded as a regression from the neighboring reference blocks to the current coding block. (1) For variable block size, we utilize different network structures. For small blocks 4×4 and 8×8, fully connected networks are used, while for large blocks 16×16 and 32×32, convolutional neural networks are exploited. (2) For each prediction mode, we develop a specific pre-trained network to boost the regression accuracy. When integrating into HEVC test model, we can save 3.55%, 3.03% and 3.27% BD-rate for Y, U, V components compared with the anchor.

Learned filter [J2]

Convolutional neural network (CNN)-based filters have achieved great success in video coding. However, in most previous works, individual models were needed for each quantization parameter (QP) band, which is impractical due to limited storage resources. To address this issue, (1) we propose a frequency and spatial QP-adaptive mechanism (FSQAM), which can be directly applied to the vanilla convolution to help any CNN filter handle different quantization noise. (2) We propose QA-filter based on FSQAM. With only one QA-filter used for all the filtering, average 5.25% and 3.84% BD-rate reductions are achieved for luminance under AI and RA configurations.

World-first learned image compression outperforming VVC intra [C5]

To enhance the coding gain of learned image compression, we proposed to use discretized Gaussian Mixture Likelihoods to parameterize the distributions of latent codes, which can achieve a more accurate and flexible entropy model. Besides, we take advantage of recent attention modules and incorporate them into network architecture to enhance the performance. Experimental results demonstrate our proposed method achieves a state-of-the-art performance compared to existing learned compression methods on both Kodak and high-resolution datasets. To our knowledge, our approach is the first work to achieve comparable performance with latest compression standard Versatile Video Coding (VVC) regarding PSNR. More importantly, our approach generates more visually pleasant results when optimized by MS-SSIM.

Learned image compression with fixed-point arithmetic [C3]

Most LIC frameworks are based on floating-point arithmetic which has two potential problems. First is that using traditional 32-bit floating-point will consume huge memory and computational cost. Second is that the decoding might fail because of the floating-point error coming from different encoding/decoding platforms. To solve the above two problems. 1) We linearly quantize the weight in the main path to 8-bit fixed-point arithmetic, and propose a fine tuning scheme to reduce the coding loss caused by the quantization. Analysis transform and synthesis transform are fine tuned layer by layer. 2) We exploit look-up-table (LUT) for the cumulative distribution function (CDF) to avoid the floating-point error. When the latent node follows non-zero mean Gaussian distribution, to share the CDF LUT for different mean values, we restrict the range of latent node to be within a certain range around mean. As a result, 8-bit weight quantization can achieve negligible coding gain loss compared with 32-bit floating-point anchor. In addition, proposed CDF LUT can ensure the correct coding at various CPU and GPU hardware platforms.

Network quantization for LIC [J1]

Learned image compression (LIC) has reached a comparable coding gain with traditional hand-crafted methods such as VVC intra. However, the large network complexity prohibits the usage of LIC on resource-limited embedded systems. Network quantization is an efficient way to reduce the network burden. This paper presents a quantized LIC (QLIC) by channel splitting. First, we explore that the influence of quantization error to the reconstruction error is different for various channels. Second, we split the channels whose quantization has larger influence to the reconstruction error. After the splitting, the dynamic range of channels is reduced so that the quantization error can be reduced. Finally, we prune several channels to keep the number of overall channels as origin. By using the proposal, in the case of 8-bit quantization for weight and activation of both main and hyper path, we can reduce the BD-rate by 0.61%-4.74% compared with the previous QLIC. Besides, we can reach better coding gain compared with the state-of-the-art network quantization method when quantizing MS-SSIM models. Moreover, our proposal can be combined with other network quantization methods to further improve the coding gain. The moderate coding loss caused by the quantization validates the feasibility of the hardware implementation for QLIC in the future.

FPGA accelerator with fine-grained pipeline [C2]

FPGA is appropriate for fix-point neural networks computing due to high power efficiency and configurability. However, its design must be intensively refined to achieve high performance using limited hardware resources. We present an FPGA-based neural networks accelerator and its optimization framework, which can achieve optimal efficiency for various CNN models and FPGA resources.

In detail, we have several proposals. First, we propose a cascading DSP to handle two 8-bit multiplications in one DSP. Moreover, the kernel-width addition is also handled inside the DSP. For a specific input channel and output channel, the MAC operation for the kernel width is computed in parallel. Following the weight stationary mechanism, each input activation is multiplied with all the kernels in the width-wise. Two weights from two output channels share one DSP to multiply with the same input activation. In addition, we propose the cascading DSP to calculate the kernel-width-wise partial sum inside DSP.

Second, we propose the deconvolution with zero skipping. Given that the stride is two and kernel is 5x5, for the input 3x3, it is first dilated to double size by inserting zero elements. The dilated input is then convolved with the kernel.

World-fastest learned image codec on FPGA [C1]

In the proposed system, the convolutional neural network is computed on FPGA, while the range codec is performed on CPU.

For the encoding, analysis transform hyper analysis and hyper synthesis are allocated on FPGA, while range encoder is performed on host CPU. One hardwired IP of the proposed architecture is instantiated with three output taps. The first tap is to extract the results y of analysis transform. The second tap is to obtain the results z of hyper analysis. By using the CDF which is prepared offline, the host PC can generate the bitstream of z by a range encoder. The third tap is to generate mean and scale to construct the CDF of y. Based on the CDF, we can generate the bitstream of y. For the decoding, synthesis transform and hyper synthesis are allocated on FPGA, while range decoders are performed on CPU. Two hardwired IPs are instantiated to be in charge of hyper synthesis and synthesis transform respectively. There are four steps for the whole decoding process. 1) Host CPU generates z by a range decoder and sends the result to FPGA. 2) FPGA generates mean and scale of y and send back to host CPU. 3) Host CPU generates y and send to FPGA. 4) FPGA generates decoded image x'.

Table I. Codec system comparison for 720P

	Platform	Technology	Speed (fps)	Latency (ms)	Power per frame (J/f)
[TCSVT'22]	ZCU104	16nm	3.90	1126	Enc: 1.68 Dec: 1.72
Proposal	VCU128	16nm	30.28	562	Enc: 1.44 Dec: 1.51

Table II. Coding gain comparison for Kodak

	bpp	PSNR (dB)
[TCSVT'22]	0.2565	29.8941
Proposal	0.2542	29.94

We first built a real-time 720P@30fps codec system on two FPGA boards VCU128. The performance is given in Table I. Both encoder and decoder can reach 30fps, and the end-to-end latency from the input camera in the encoder side to the output display in the decoder side is about 562ms. About the power consumption, we used two power meters to measure the real power of encoder and decoder respectively. Compared with a very innovative work [TCSVT'22], our proposed codec system can reach 7.76x faster speed with about half latency. Regarding the power efficiency, ours is just slightly better than [TCSVT'22] since we did not adopt the lower power schemes such as data gating.

In addition, we also build a demo system on two different FPGA boards. Encoder is performed on VCU128, while decoder is performed on KU115. The correct decoding illustrates that the proposed system is capable of cross-platform coding.

In Table II, we provide the coding result of our quantized model on Kodak dataset. Compared with [TCSVT'22], we can reach slightly better coding gain with smaller bpp and higher PSNR.

[TCSVT'22] C. Jia, X. Hang, S. Wang, Y. Wu, S. Ma, W. Gao, "FPX-NIC: An FPGA-accelerated 4K Ultra-high-definition Neural Video Coding System," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 9, pp. 6385-6399, 2022.

3. 今後の展開

About the extension of the research, we aim to focus on video coding for machine. It is expected that more than 50% connections will be the connections among the machines. Therefore, reducing the bandwidth of machine connection becomes more and more important. We have developed an automated tool which can generate the RTL codes for any convolutional neural networks. Therefore, we aim to develop a real-time VCM system. I plan to finish this work in 2023, and I believe this system can contribute to the community of MPEG-VCM.

The extension to VCM can be used in a wide field in IoT society to improve the security. One user case is video surveillance which is an inevitable component in IoT city. Originally, 30fps videos are coded with the traditional standard. In the backend, human will observe the videos and perform tasks such as abnormal detection. In our proposal, with a higher compression ratio, 240fps videos can be transmitted to the backend. By reducing the frame interval from 33ms to 4.17ms, the security can be significantly enhanced. Meanwhile, due to the compatible bitstream format, 30fps videos can also be generated for human vision in the backend.

4. 自己評価

It is my first time to get such a big project. The self-evaluation is conducted from two aspects. First aspect is about the research content. I am personally not so satisfied with the research result, while I think it is still qualified. I created the world-fastest codec system on FPGA, which was presented at ASSCC and got VCIP best demo award.

The second aspect is about the contribution to the society and community. The current research results have not shown direct contribution to the society. However, as explained in the last section, the extension to machine task is very potential to contribute to MPEG-VCM community. Furthermore, considering that many machine tasks such as object detection and image classification are related with security issue. Therefore, I believe that this research is also very important for the next-generation IoT society.

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:10件

[J1] **Heming Sun**, Lu Yu, Jiro Katto, “Q-LIC: Quantizing Learned Image Compression with Channel Splitting”, IEEE Transactions on Circuits and Systems for Video Technology, early access. **(Impact factor: 5.859)**
This paper presents a quantized learned image compression by channel splitting. First, we explore that the influence of quantization error to the reconstruction error is different for various channels. Second, we split the channels whose quantization has larger influence to the reconstruction error. Finally, we prune several channels to keep the number of overall channels as origin. By using the proposal, in the case of 8-bit quantization for weight and activation of both main and hyper pat, we can reduce the BD-rate by 0.61-4.74% compared with the previous QLIC. Besides, we can reach better coding gain compared with the state-of-the-art network quantization method when quantizing MS-SSMI models. Moreover, our proposal can be combined with other network quantization methods to further improve the coding gain. The moderate coding loss caused by the quantization validates the feasibility of the hardware implementation for QLIC in the future.
[J2] Chao Liu, **Heming Sun (corresponding author)**, Jiro Katto, Xiaoyang Zeng, Yibo Fan, “QA-Filter: A QP-Adaptive Convolutional Neural Network Filter for Video Coding”, IEEE Transactions on Image Processing, Vol. 31, pp. 3032-3045, Apr. 2022. **(Impact factor: 11.041)**
Convolutional neural network (CNN)-based filters have achieved great success in video coding. However, in most previous works, individual models were needed for each quantization parameter (QP) band, which is impractical due to limited storage resources. To explore this, our work consists of two parts. First, we propose a frequency and spatial QP-adaptive mechanism (FSQAM), which can be directly applied to the (vanilla) convolution to help any CNN filter handle different quantization noise. From the frequency domain, a FQAM that introduces the quantization step (Qstep) into the convolution is proposed. When the quantization noise increases, the ability of the CNN filter to suppress noise improves. Moreover, SQAM is further designed to compensate for the FQAM from the spatial domain. Second, based on FSQAM, a QP-adaptive CNN filter called QA-Filter that can be used under a wide range of QP is proposed. By factorizing the mixed features to high-frequency and low-frequency parts with the

pair of pooling and upsampling operations, the QA-Filter and FQAM can promote each other to obtain better performance. Compared to the H.266/VVC baseline, average 5.25% and 3.84% BD-rate reductions for luma are achieved by QA-Filter with default all-intra (AI) and random-access (RA) configurations, respectively. Additionally, an up to 9.16% BD-rate reduction is achieved on the luma of sequence BasketballDrill. Besides, FSQAM achieves measurably better BD-rate performance compared with the previous QP map method.

[J3] **Heming Sun**, Zhengxue Cheng, Masaru Takeuchi, and Jiro Katto, “Enhanced Intra Prediction for Video Coding by Using Multiple Neural Networks”, IEEE Transactions on Multimedia, Vol. 22, No. 11, pp. 2764-2779, Nov. 2020. **(Impact factor: 8.182)**

This paper enhances the intra prediction by using multiple neural network models (NM). Each NM serves as an end-to-end mapping from the neighboring reference blocks to the current coding block. For the provided NMs, we present two schemes (appending and substitution) to integrate the NMs with the traditional modes (TM) defined in high efficiency video coding (HEVC). For the appending scheme, each NM is corresponding to a certain range of TMs. The categorization of TMs is based on the expected prediction errors. After determining the relevant TMs for each NM, we present a probability-aware mode signaling scheme. The NMs with higher probabilities to be the best mode are signaled with fewer bits. For the substitution scheme, we propose to replace the highest and lowest probable TMs. New most probable mode (MPM) generation method is also employed when substituting the lowest probable TMs. Experimental results demonstrate that using multiple NMs will improve the coding efficiency apparently compared with the single NM. Specifically, proposed appending scheme with seven NMs can save 2.6%, 3.8%, and 3.1% BD-rate for Y, U, and V components compared with using single NM in the state-of-the-art works.

（2）特許出願

該当なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

主要な学会発表

[C1] **Heming Sun**, Qingyang Yi, Fangzheng Lin, Lu Yu, Jiro Katto, Masahiro Fujita, “Real-time learned image codec on FPGA”, IEEE International Conference on Visual Communications and Image Processing (VCIP), Dec. 2022. **(Best demo award)**

[C2] **Heming Sun**, Qingyang Yi, Fangzheng Lin, Lu Yu, Jiro Katto, Masahiro Fujita, “F-LIC: FPGA-based Learned Image Compression with a Fine-grained Pipeline”, IEEE Asian Solid-State Circuits Conference (ASSCC), Nov. 2022.

[C3] **Heming Sun**, Lu Yu, and Jiro Katto, “Learned Image Compression with Fixed-point Arithmetic”, Picture Coding Symposium (PCS), pp. 1-5, June 2021. **(Top-10 best paper)**

[C4] **Heming Sun**, Lu Yu, and Jiro Katto, “Fully Neural Network Mode Based Intra Prediction of Variable Block Size”, IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 21-24, Dec. 2020. **(Best paper award)**

[C5] Zhengxue Cheng, **Heming Sun**, Masaru Takeuchi, Jiro Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7939-7948, June 2020. **(Acceptance rate: 22%, citation: 295)**