Neural Network-based In-Loop Filter for CLIC 2022

Yonghua Wang, Jingchi Zhang, Zhengang Li, Xing Zeng, Zhen Zhang, Diankai Zhang, Yunlin Long, Ning Wang
State Key Laboratory of Mobile Network and Mobile Multimedia Technology
ZTE Corporation
{wang.yonghua1, zhang.jingchi ,li.zhengang1, zeng.xing1, zhang.zhen23, zhang.diankai, long.yunlin, wangning}@zte.com.cn

Abstract

A hybrid video codec comprised of an optimized VVC codec and a convolutional neural network-based loop filter (CNNLF), was submitted in the video compression track in Challenge on Learned Image Compression (CLIC) 2022[1].

This paper presents the traditional methods and deep learning scheme in video coding optimization, which were adopted in the hybrid codec based on VTM-15.0. Traditional methods include QP adaptive adjustment of I frame and rate-distortion optimization based on SSIM. Meanwhile, the deep learning scheme proposes an adaptive CNNLF, which is turned on / off based on the rate-distortion optimization at CTU and frame level. The network architecture mainly consists of the attention residual module and the convolution feature maps module, which help extract image features and improve image quality. To balance performance and complexity, the proposed scheme sets different training parameters for 0.1 Mbps and 1 Mbps, respectively. The experimental results show that compared with VTM-15.0, the proposed traditional methods and adding CNNLF improve the PSNR by 0.4dB and 0.8dB at 0.1Mbps, respectively; 0.2dB and 0.5dB at 1Mbps, respectively, which proves the superiority of our method.

1. Introduction

The past few decades have witnessed great progress in video compression, and many video coding standards have been released by ISO and ITU-T. The latest is Versatile Video Coding (VVC) [2], finalized in 2020. Before, High Efficiency Video Coding (HEVC) [3] and Advanced Video Coding (AVC) [4] are widely applied to video compression and transmission, which greatly promote the development of video compression techniques and related industry development. The experimental results show that compared with the HEVC reference software HM-16.24, the VVC reference software VTM-15.0 saves bitrate about 37% in random access (RA) configuration. Therefore, the method proposed in this paper selects VTM-15.0 as the baseline.

Nowadays, more and more researchers focus on combining traditional video coding technology with deep-learning technology to improve video compression performance [5]. Likewise, this competition encourages deep learning solutions. Therefore, to further enhance the quality and reduce the compression artifacts of compressed frames in VVC, an adaptive in-loop filter based on a convolutional neural network (CNN) is proposed in this paper.

The remainder of this paper is organized as follows: traditional encoding method optimization will be introduced in Section 2, including QP adaptive adjustment for I-frame and Rate-distortion optimization based on improved SSIM. The video compression method based on the neural network will be concretely described in section 3. Experimental results will be presented and analyzed in Section 4 and the conclusion will be given in Section 5.

2. Optimization of traditional encoding

In terms of encoding configuration, the goal of CLIC is to maximize the human rating scores, so a perceptual QP adaptation algorithm targeting subjective effect maximization should be applied [6]. For the best encoding performance, a Group Of Pictures (GOP) size of 32 pictures [7] is recommended in the JVET common test conditions in RA configuration. Besides, only one Intra frame at the beginning can save bit allocation at the same subjective effect.

2.1. QP adaptive adjustment of I-frame

Considering the reference influence relationship of the encoding quality of the I-frame to the subsequent encoded frames, the QP of the first frame is very important. We know that the rate control (RC) can determine the I-frame QP for encoding, or set the initial QP for the perceptual quantization encoding of the I-frame without RC, but no one has tried to combine the two. Here, rate control is used to set an initial I-frame QP that takes into account the characteristics of the specific sequence content, followed by perceptual quantization coding.

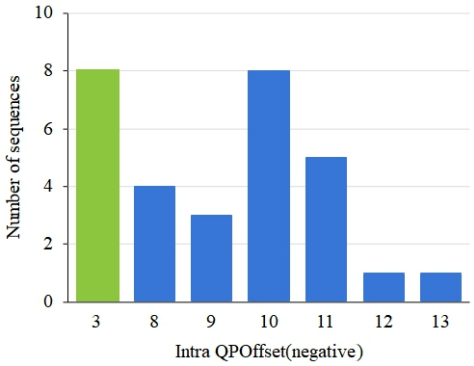

Figure 1. Selected QP Offset values

It is crucial to select the appropriate I-frame QP by considering the complexity and motion characteristics of the content in different sequences. For I-frame content complexity: using the RC method to determine QP, which considered the Hadamard transform cost of I-frame. And sequence motion characteristics are measured by the rate of change of sequence content to distinguish static or slowly changing sequences from dynamic sequences.

After testing, the final selection range of QP offset values at 0.1 Mbps is shown in Figure 1, the abscissa represents QP offset, the ordinate represents the number of sequences in the validation set. Considering the reference influence relationship, this paper only uses the above method to set the QP offset value for static or slowly changing sequences, as shown in the blue column in Figure 1. And the original method of the encoder is used for dynamic sequences, as shown in the green column in Figure 1, the QP offset value is -3. This method can make the subjective viewing effect of overall coded frames better.

2.2. Rate-distortion optimization based on SSIM

Rate-distortion optimization (RDO) [8] plays an important role in video coding. In the current RDO criteria of VTM-15.0, MSE is used as the measurement standard, as shown in formula (1), where N is the number of pixels in the coding unit.

$$J_{SSE} = SSE + \lambda_{SSE} \times R = N \cdot MSE + \lambda_{SSE} \times R \tag{1}$$

This competition focuses more on people's visual ratings. Since SSIM is more in line with the human visual perception system, it is widely used for subjective quality assessment of encoded videos. Therefore, this paper proposes to use SSIM as the optimization goal in the RDO process. In fact, there is an approximate conversion relationship between the SSIM metric and the MSE metric, the formula is as follows in formula (2), where σ_x and σ_y represent the variance of the original image x and the reconstructed image y, respectively, σ_{xy} represents the covariance of the two, c2 represents a very small constant.

$$SSIM \approx \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \approx \frac{2\sigma_{xy} + c_2}{2\sigma_{xy} + MSE + c_2} \tag{2}$$

On the basis of obtaining the relationship between the distortion (D) and the bit rate (R), the rate-distortion optimization coding based on SSIM is carried out. Formula (1) is rewritten as follows:

$$\begin{aligned} J &= N \cdot RD\_SSIM + \lambda \times R \approx N\left(1 + \frac{MSE}{2\sigma_{xy} + c_2}\right) + \lambda \times R \\ &= N + \frac{1}{2\sigma_{xy} + c_2}\left(SSE + \left(2\sigma_{xy} + c_2\right) \cdot \lambda \times R\right) \end{aligned} \tag{3}$$

Finally, the Lagrange multiplier in the rate-distortion function optimized based on SSIM can be obtained, and the formula is as follows:

$$\lambda_{SSIM} = \lambda_{SSE} \cdot \frac{2\sigma_x^2 + c_2}{exp\left(\frac{1}{N}\sum_{1}^{N} log\left(2\sigma_{xy} + c_2\right)\right)} \tag{4}$$

Where λ_{SSE} represents the original Lagrange multiplier. After optimization by this method, the overall visual effect is improved.

3. Adaptive In-Loop Filter based on CNN

VVC follows the block-based hybrid coding structure, the main operations such as intra prediction, inter prediction, and quantization are performed block by block. Therefore, the coding parameters vary from block to block, resulting in blocking effects. In addition, high-frequency components of the video will be lost during the quantization process, which leads to ringing and blurring effects.

Aiming at eliminating these compression artifacts, an adaptive in-loop filter based on neural network is adopted. The CNNLF is integrated into VTM-15.0 to serve as an in-loop processing module for better compression quality. After CNNLF processing, on the one hand, the image quality of this frame can be improved, on the other hand, it can provide a better reference for other frames. This paper comprehensively considers the coding time complexity and coding performance, so different training methods are implemented at 0.1 Mbps and 1 Mbps respectively. The architecture of network and training process will be introduced separately.

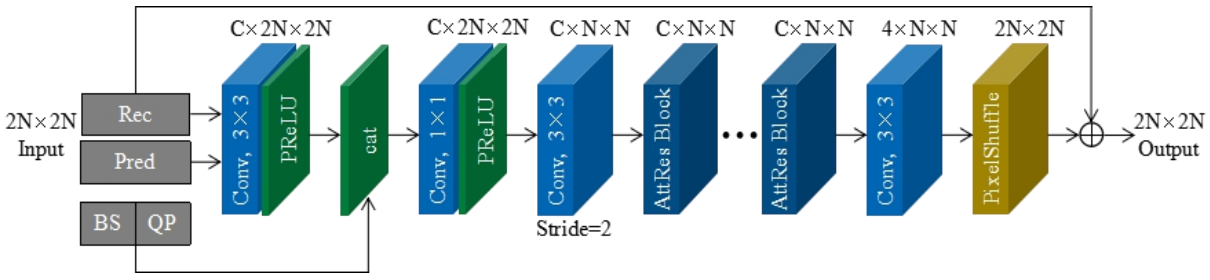

Figure 2(a). CNNLF architecture

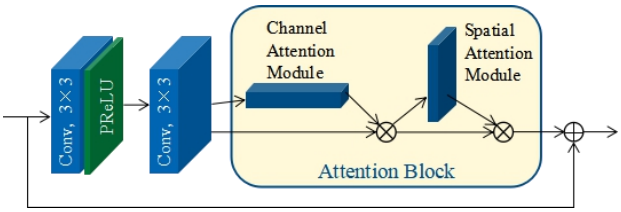

Figure 2(b). Attention Residual module

3.1. Architecture of Network

The proposed CNN filter structure diagram is shown in Figure 2(a). The input of the network contains reconstructed image (Rec), predicted image (Pred), boundary strengths (BS), and quantization parameter (QP). C denotes the channel number of feature maps, is set as 96, N represents the size of patch size in one dimension, the number of attention residual blocks is set as 8.

The network architecture is mainly composed of an attention residual module [5][9] and a convolution feature maps module, as shown in Figure 2(b). The calculation process in the attention module can be written as:

$$F\_out = F\_in \times f\,(\text{Rec}, \text{Pred}, \text{BS}, \text{QP}) + F\_in \qquad (5)$$

where F_in and F_out represent the input and the output of the attention module, respectively. f contains 2 convolutional layers, where an activation function is applied after the first convolutional layer. The attention module has two sub-modules: channel and spatial attention, focusing on 'what' and 'where' is meaningful given an input image respectively. Besides, two sub-modules is placed sequentially and the channel-first order performs slightly better [10]. The goal of f is to generate a spatial attention map from external information and then recalibrate the feature map F_in. After CNNLF processing, image quality can be improved.

The CNN architecture is conditioned on QP, leading to a unified model to handle different quality levels.

3.2. Training Process

For the Network, PyTorch is used as the training platform. In order to make CNNLF obtain stronger generalization ability, after converting the data-set images to YUV420 format, use VTM-15.0 configured with RA for encoding and decoding. For the training data, we randomly extracted and cropped consecutive frames from the CLIC 2022 data-set and DIV2K data-set [11] about 200k samples. We trained for 90 epochs using the Adam optimizer with a starting learning rate of 1e−4, a batch size of 64, and reconstructed images are split into 128x128 luma and 64x64 chroma blocks. We decay the learning rate by 0.1 whenever the loss reaches a plateau. The network is first optimized for MSE to improve convergence and stability, then switch to the SSIM metric at 80% of the total training steps, and also increase the patch size to reduce border artifacts. The CNN filters of I slices and B slices are trained respectively. Training is performed on a single Tesla-V100-32GB GPU.

There are some differences: for 0.1Mbps, the input is one single-channel image, luma and chroma CNN models are trained separately to adapt to different QP points, QP points include 32, 37, 43, 48; for 1Mbps, in order to reduce the coding time complexity, the input is the 3 channel image, luma and chroma share the same model, which can reduce the complexity of training and inference, QP points include 22, 27, 32 and 37.

4. Experimental Results

4.1. Implementation

The proposed traditional methods and CNNLF are implemented on top of the VTM-15.0 reference software. Deblocking filtering and SAO are disabled while ALF is placed after the proposed CNNLF. During the process of CNNLF, whether to apply the proposed filter is based on the rate-distortion optimization in CTU and frame level. In addition to CNNLF with CTU and frame-level flags, the flag for on / off CNNLF is also in the encoder conditional parameter index.

In order to compare the performance with the VTM-15.0,

we adopt the same coding parameters under the default configuration of RA, and perceptual QP adaption is enabled. Then after converting the mp4 videos in the validation set to YUV videos, we encoded them at 0.1 Mbps and 1 Mbps, respectively. Finally, the objective coding performance index PSNR and human visual subjective effects are compared.

4.2. Compression Performance

The target bitrate is approximately 0.1 Mbps and 1 Mbps for the 30 sequences of 10-second. Consequently, the limit for the Submission Size is set to 35.7628 Mbytes at 1 Mbps and 3.5763 Mbytes at 0.1 Mbps, respectively.

Experimental results demonstrate that the proposed video compression approach can achieve good performance in the validation sets of CLIC 2022 under the condition of limiting the bitstream size. On objective indicators, for 0.1 Mbps, as shown in Table1, the proposed traditional optimization method, named Trad_Opt, improve PSNR by 0.4 dB, and adding CNNLF improves PSNR by 0.8 dB. And for 1 Mbps, as shown in Table2, Trad_Opt improves PSNR by 0.2 dB and adding CNNLF improves PSNR by 0.5 dB. In subjective evaluation, the compression artifacts were greatly reduced and the visual effects were significantly improved.

In conclusion, the proposed method not only has a suitable amount of data but also has a better objective and subjective performance than VTM-15.0, which strongly proves the superiority of our method.

Table1. The compression performance of in the validation sets of CLIC at 0.1 Mbps

Method	Data Size(Mbytes)	PSNR(dB)
VTM-15.0	3.54	29.515
Trad_Opt	3.56	29.887
Trad_Opt+CNNLF	3.57	30.333

Table2. The compression performance of in the validation sets of CLIC at 1 Mbps

Method	Data Size(Mbytes)	PSNR(dB)
VTM-15.0	35.5	36.266
Trad_Opt	35.7	36.380
Trad_Opt+CNNLF	35.7	36.776

5. Conclusion

In this paper, on traditional encoders, an adaptive offset method that adjusts the QP of I-frame and a rate-distortion optimization method based on SSIM are implemented. In the depth learning scheme, CNNLF is proposed, the network architecture mainly consists of the attention residual module and the convolution feature maps module, which is implemented in VTM-15.0 with CTU and frame-level enabled flags.

The proposed method has better performance in terms of PSNR and subjective evaluation than the traditional VTM-15.0 in the validation sets of the challenge on learned image compression (CLIC), which demonstrates the superiority of our approach.

References

[1] Workshop and challenge on learned image compression (CLIC). *http://www.compression.cc/challenge/*.

[2] B. Bross, J. Chen, J. R. Ohm, G. J. Sullivan, and Y. K. Wang. Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). *Proceedings of the IEEE*, 1–31, 2021

[3] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12): 1649–1668, 2012.

[4] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra. Overview of the H.264/AVC video coding standard, *IEEE Trans. Circuits Syst. Video Technol*., 13(7): 560-576, Jul. 2003.

[5] L. Wang, W. Jiang, X. Xu, S. Liu. AHG11: neural network based in-loop filter. *JVET-W0113, 23rd Meeting*, by teleconference, 7–16, July 2021.

[6] C. Helmrich, H. Schwarz, D. Marpe, T. Wiegand. Improved perceptually optimized QP adaptation and associated distortion measure, *JVET-K0206*, Ljubljana, SI, July 2018.

[7] J. Chen, Y. Ye, S. H. Kim. Algorithm description for Versatile Video Coding and Test Model 11 (VTM 11), *JVET-T2002, 20th JVET meeting by teleconference*, Oct. 2020.

[8] G. J. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6): 74–90, Nov 1998.

[9] He K , Zhang X , Ren S , et al. *Deep Residual Learning for Image Recognition*[J]. 770-778, 2015.

[10] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module, *Proceedings of the European conference on computer vision (ECCV)*. 3-19, 2018.

[11] R. Timofte, E. Agustsson, S. Gu, J. Wu, A. Ignatov, L. V. Gool, https://data.vision.ee.ethz.ch/cvl/DIV2K/