An Adaptive In-Loop Filter based on Neural Network for Video Coding

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

Abstract

Traditional video coding standards, such as HEVC and VVC, have achieved significant compression performance. To further improve the coding efficiency, this paper proposes an adaptive in-loop filter based on neural network (NNLF) for VVC. Specially, the neural network of NNLF mainly consists of residual blocks and 2-dimensional up-sampling convolution, which is implemented in VVC Test Model (VTM) between De-blocking and SAO. With CTU and frame level enabled flags, NNLF result is adaptively applied in CTU and frame reconstruction based on RDO and temporal id. The proposed scheme has achieved good performance in MS-SSIM in video compression track of challenge on learned image compression (CLIC)[1]. Compared with VTM-11.0, the proposed scheme not only has a smaller data size, but also 0.219dB higher in PSNR and 0.00126 higher in MS-SSIM, which demonstrates the superiority of our approach.

1. Introduction

The past few decades has witnessed the great progress in video compression, and many video coding standards have been released. Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC) are widely applied to video compression and transmission, which greatly promote the development of the video compression techniques. 7 years after HEVC publication, Versatile Video Coding (VVC) was finalized in 2020, providing about 50% bit-rate reduction over its predecessor (HEVC).

In recent years, deep learning technology has not only achieved great success in the field of artificial intelligence, but also brought new development opportunities to the field of video coding. More and more researchers have begun to focus on combining deep-learning technology with traditional video coding technology to improve video compression performance.

In order to further enhance the quality of compressed frames in VVC, a novel video compression method based on neural network is proposed in this paper. To reduce the compression artifacts and obtain compressed frames of better quality, we design an adaptive in-loop filter based on neural network to improve the quality of compressed videos. The architecture of NNLF contains the residual block (RB) and the 2-dimensional up-sampling convolution. In addition, the NNLF is integrated into VTM-11.0 to serve as a in-loop-processing module for better compression quality. Experimental results demonstrate that the proposed video compression approach can achieve good performance in the validation sets of CLIC.

The remainder of this paper is organized as follows: hierarchical B GOP structure and perceptual QP adaptation (QPA) encoding method in VTM will be reviewed in section II, and our video compression method based on neural network will be concretely described in section III. Experimental results will be presented and analyzed in Section IV and the conclusion will be given in the Section V.

2. Key Techniques in VTM

In video compression, GOP structure plays an important role, which determines the temporal reference. Also, a perceptual QP adaption (QPA) algorithm along with a correspondingly weighted PSNR (WPSNR) distortion measure is applied for improving subjective visual quality. In this section. we will concisely review the above key techniques in VTM.

2.1. Hierarchical B GOP Structure in VTM

For the random access configuration in VTM, a hierarchical B GOP structure is used for encoding [2]. A GOP size of 32 pictures is currently recommended in the JVET common test conditions, as shown in Figure 1.

The non-intra pictures in a GOP are encoded as B-pictures by default. The random access configuration defines a hierarchy among different B pictures whereby each hierarchical level is associated with a temporal identifier. Pictures with lower temporal id are shown towards the top in Figure 1 since they are used more often as reference for inter coding . The arrows depict reference frame of each picture with each arrow pointing to the reference picture. Meanwhile, reference pictures of some of

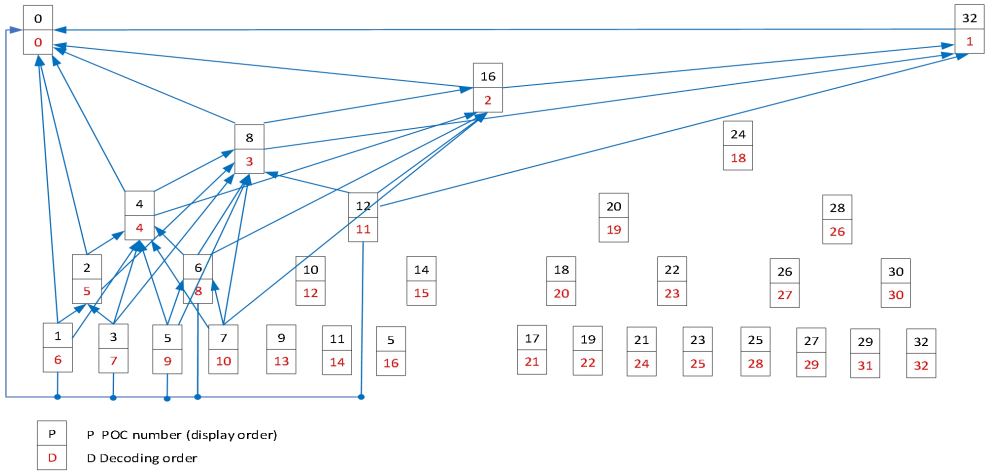

Figure 1. Hierarchical prediction structure in random access configuration

the pictures (but not all pictures) in the GOP are depicted for simplicity. The picture at temporal id 0 (commonly referred to as "Generalized B picture" or GBP) is used as the lowest temporal layer that can refer to intra or inter pictures for inter prediction. Low temporal layers (id 0 - 4) consist of referenced B pictures, while the highest temporal layer (id 5) contains non-referenced B picture only. For GBP pictures, each reference picture list is configured with four reference pictures and the two reference picture lists are identical. For all other B pictures, there are two reference pictures in each reference picture list.

The QP of each inter-coded picture is derived by adding an offset to the base QP. The higher a picture's temporal ID is, the larger its QP offset.

2.2. Perceptual QP Adaptation Method

In VTM encoder, a perceptual QP adaptation (QPA) algorithm along with a correspondingly weighted PSNR (WPSNR) distortion measure is applied for improving subjective visual quality [3]. In the algorithm, a subjectively motivated block-wise weighted distortion metric D^{W} derived from a local visual activity value as shown in the formula (1).

$$D_{\mathrm{pic}}^{\mathrm{W}} = \sum_{k} D_{k}^{\mathrm{W}} = \sum_{k} w_k \cdot \sum_{(x,y)\in B_k} (s[x,y]-s'[x,y])^2 \quad (1)$$

where $w_k = (\frac{a_{\mathrm{pic}}}{a_k})^{\beta}$ and B denotes the blocks (here CTUs) of the picture, indexed via k, w_k is the perceptual weight (also called visual sensitivity measure), a_k is the block's visual activity, a_{pic} is the picture's mean visual activity, s and s', are the original and reconstructed pel values, respectively, of the picture's luma component.

To determine the local visual activity a_k of a block B_k at index k, the luma input samples s are subjected to a 9-tap square-shaped filter kernel having the coefficients [–1, –2, –1, –2, 12, –2, –1, –2, –1].The "global" mean visual activity value a_{pic} is determined empirically for the purpose of controlling average bit-rate bias, as shown in the formula (2):

$$a_{\mathrm{pic}} = 2^{BD} \cdot \begin{cases} 16 & \text{for UHD sequences} > 2048\times1280 \text{ pels} \\ 32 & \text{for UHD sequences} > 1024\times640 \text{ pels} \\ 64 & \text{otherwise} \end{cases} \quad (2)$$

Where BD denotes the coding bit-depth.

Using D_{pic}^{wSSE}, the picture's width W, height H and component bit-depth BD, a weighted peak signal-to-noise ratio can be obtained in the formula (3):

$$WPSNR = 10 \cdot \log_{10}\left(\frac{W \cdot H \cdot 255^2 \cdot 2^{2BD-16}}{D_{\mathrm{pic}}^{wSSE}}\right) \quad (3)$$

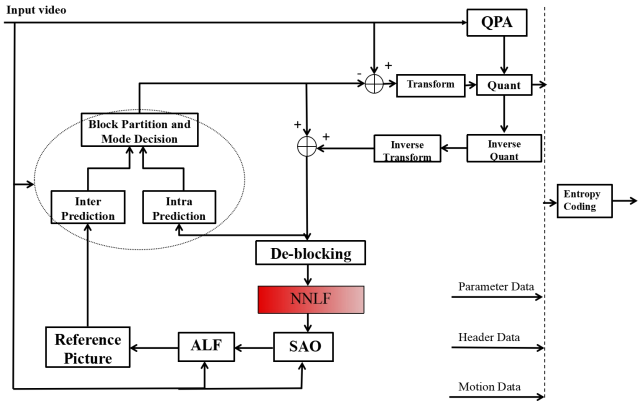

Figure 2. Proposed video encoder with NNLF

Notice that, when $w_k = 1$ for all blocks, this WPSNR definition reduces to the conventional PSNR metric used by the JVET. This is particularly true when $\beta = 0$. In VTM, the use of $\beta = 0.5$ was suggested. During encoding, the perceptual QPA makes use of w_k to adapt in each CTU at k. The perceptually optimized QP and Lagrange parameter based on the default pre-assigned fixed QP_{slice} and Lagrange parameter λ_{slice} is shown in the formula (4):

$$QP_k = QP_{slice} - [3\log_2 w_k], \lambda_k = \frac{\lambda_{slice}}{w_k} \qquad (4)$$

Using the adapted QP_k and λ_k in each CTU block ensures that, in terms of visual quality and WPSNR, the coding distortion is optimally distributed within the given picture and within the pictures of the video signal.

3. Adaptive In-Loop Filter based on Neural N etwork (NNLF)

Since the block based hybrid coding structure is adopted in VVC, major operations such as intra / inter prediction, transform and quantization are performed block by block. Consequently, the coding parameters vary by the blocks, which leads to blocking effects. In addition, high frequency components of the video will be lost during quantization process, which results in ringing and blurring effects. Aiming at eliminating these compression artifacts, an adaptive in-loop filter based on neural network is designed to improve the quality of compressed videos.

The proposed NNLF is located between de-blocking and SAO stages as shown in Figure 2. The reconstructed frames

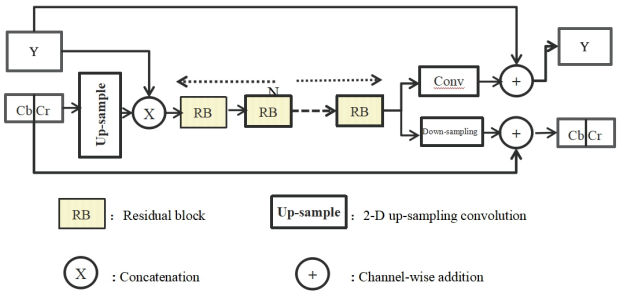

Figure 3. (a) NNLF architecture

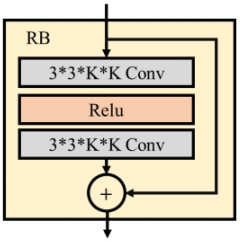

Figure 3. (b) Residual block

are firstly filtered by de-blocking before processed by NNLF, and SAO is employed after NNLF. During the process of NNLF, whether to apply proposed filter is based on the rate-distortion optimization (RDO) in CTU and frame level.

3.1. Network Architecture

The architecture of proposed in-loop filter is shown in Fig. 3 (a). The network mainly consists of residual blocks and the 2-dimensional up-sampling convolution[4]. The chroma samples are up-sampled from 64x64 to 128x128 and then concatenated with luma sample to form a 3x128x128 input features. In Fig. 3 (b), the RB contains two 3x3 Convolution filter with K input/output features and a rectified linear unit (ReLU) between them. In proposed network, N and K are set equal to 20 and 64, respectively.

3.2. Training Process

The proposed network is trained with the UGC data-set released by CLIC and DIV2K data-set[5]. The data set images are encoded and decoded by VTM-11.0 after converted into YUV420 format. In order to make NNLF obtain stronger generalization ability, QPs for training are set to 22, 27, 32, and 37. The reconstructed images are then split into 128x128 luma and 64x64 chroma blocks.

3.3. Adaptive Application of NNLF based on RDO

The proposed in-loop filter based on neural network is

implemented in VTM11.0 with CTU and frame level enabled flags. If the frame level flag is turn-off, all CTU in current frame are not applied with proposed filter. If the frame level flag is turn-on, the CTU level flag is signaled to indicate whether the proposed filter is applied.

The calculation method of the frame-level switch is determined by the rate-distortion optimization as described in the formula (5), where D_i refers to the change in mean square error distortion (MSE) of i-th CTU before and after neural network filtering compared with the source pixels, and n represents the number of CTU. If the rate-distortion cost is negative, turn on the frame-level filtering switch, otherwise turn off the frame-level filtering switch. For the CTU level switch, only the change in the mean square error distortion before and after the neural network filtering compared with the source pixels is considered. If the distortion after the neural network filtering is smaller, the switch is turned on, and vice versa. Both frame-level switches and CTU-level flags are coded into the stream.

$$RD\cos t = \sum_{i=1}^{n} \max(0, D_i) + \lambda * n \qquad (5)$$

3.4. Algorithm Acceleration

In the proposed neural network in-loop filter, the change of the neural network filter switch under the random access configuration is explored. On the one hand, as the temporal layer increases, the temporal information in the reference frame is more fully utilized. The number of neural network filtering blocks selected by the CTU switch gradually decreases, and the quality improvement brought by neural network filtering gradually decreases. On the other hand, as quantization step size increases with the increase of the temporal layer, the number of coded bits occupies more proportion in the rate-distortion optimization and syntax elements of CTU level flags bring more coded bits. After weighting the two factors of complexity and performance, we decide to turn off the frame-level NNLF where the temporal layer ID is 4 and 5 to speed up the algorithm. Because NNLF on those frames can only improve very limited visual quality, and most of those frames are not used as reference frames. Since the frames with frame-level NNLF enabled only occupies 1/4 of a GOP, encoding and decoding complexity of neural network loop filtering is greatly reduced.

4. Experimental Results

4.1. Implementation

The NNLF is implemented on top of the VTM-11.0 reference software. In order to compare the performance with the VTM-11.0, we adopt the same coding parameters under the default configuration of RA, and perceptual QP adaption is enabled. Then after converting the PNG images in the validation set to YUV videos, we encoded them with the above two different methods. Finally, coding performance is compared in both PSNR and MS-SSIM.

4.2. Compression Performance

As shown in Table.1, it can be evidently found that not only our proposed method has the smaller data size but also has the higher PSNR and MS-SSIM than VTM-11.0, which strongly proves the superiority of our method.

Table.1 The compression performance of in the validation sets of CLIC

Method	Data Size(bytes)	PSNR(dB)	MS-SSIM
VTM-11.0	24817697	35.482	0.98621
Proposed	24729616	35.701	0.98747

5. Conclusion

In this paper, we propose an adaptive in-loop filter based on neural network to improve the quality of compressed videos. The network architecture mainly consists of residual blocks and the 2-dimensional up-sampling convolution, which is implemented in VTM-11.0 with CTU and frame level enabled flags.Through data training and accelerated optimization of algorithms, this method has better performance in terms of MS-SSIM than the traditional VTM-11.0 in the validation sets of 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]. 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.

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

[4]. Tsung-Chuan Ma, Wei Chen, Xiaoyu Xiu, Yi-Wen Chen, Hong-Jheng Jhu, Che-Wei Kuo, Xianglin Wang, "*AHG11: In-loop filtering based on neural network*", JVET-T0094, 20th Meeting, by teleconference, 7 – 16 Oct. 2020.

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