Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Modelling of Depth Prediction Algorithm for Intra Prediction Complexity Reduction

Helen K. Joy[1*], Manjunath R. Kounte[2]

[1] School of Electronics and Communication Engineering, REVA University, 560064 Bengaluru, India
[2] Department of Electronics and Computer Engineering, School of ECE, REVA University, 560064 Bengaluru, India

[*] Correspondence: Helen K. Joy (R19PEC09@ece.reva.edu.in)

Received: 09-20-2022 **Revised:** 10-15-2022 **Accepted:** 11-12-2022

Citation: H. K. Joy and M. R. Kounte, "Modelling of depth prediction algorithm for intra prediction complexity reduction," *Acadlore Trans. Mach. Learn.*, vol. 1, no. 2, pp. 81-89, 2022. https://doi.org/10.56578/ataiml010202.

Abstract: Video compression gained its relevance with the boon of the internet, mobile phones, variable resolution acquisition device etc. The redundant information is explored in initial stages of compression that's is prediction. Inter prediction that is prediction within the frame generates high computational complexity when working with traditional signal processing procedures. The paper proposes the design of a deep convolutional neural network model to perform inter prediction by crossing out the flaws in the traditional method. It briefs the modeling of network, mathematics behind each stage and evaluation of the proposed model with sample dataset. The video frame's coding tree unit (CTU) of 64x64 is the input, the model converts and store it as a 16-element vector with the goodness of CNN network. It gives an overview of deep depth decision algorithm. The evaluation process shows that the model performs better for compression with less computational complexity.

Keywords: Intra prediction; CTU; Computational complexity; Deep learning

1. Introduction

Video compression in its actual form helps the data transfer in fast way maintaining the quality. When it comes to video, quality is an important parameter. The codecs should have a compression procedure that produces better quality output with less computational cos, considering recent video codec like H.264 [1], H.265(HEVC) [2], H.266 [3]. Versatile video coding (VCC) the issue facing is computational complexity, compression artifact, low coding efficiency etc., the total compression procedure when split into blocks the 70% of compression happens in the initial stage i.e., prediction stage. The main predictions in video compression are intra prediction and inter prediction. Intra prediction compares the similarities within the frame and inter prediction corelate the similarities between the frames. The more the similarities the chance of compression is more. As video compression advances, expect more accuracy with lower bitrates. The compression age is actively investigating VVC, advancement of existing HEVC, and much more to focus on developing. When the conveyance of data at a low bit rate became necessary, several compression methods emerged, but image quality was also a concern. As a result, video compression and its variations became prominent in the study field.

The traditional prediction method compares the similarity within/ between the frames and calculates the RDO [4] (rate distortion cost). Based on the value of rate distortion RD cost of the blocks, the split of the blocks is decided. The frame is split or not is decided by rate distortion algorithm, for that the frame is split to blocks of coding unit (CU). The coding units are arranged according to pixel length like 32*32, 16*16, 8*8, 4*4. After splitting the rate distortion optimization of best CU is calculated and compared. If the parent CU is having less value of RDO than the child CU [5] then the split is not done else the frame will be split. This will be done for the entire frame. The total procedure is complicated with many calculations. The computational complexity [5] is more in this case.

This motivated to frame a new concept to reduce computational complexities in inter prediction, out of all traditional signal processing scheme, the computational complexity reduction was a hard task to be done. So the focus turned to artificial neural network. The neural network in its traditional form couldn't be able to satisfy the

need as it wanted more layers to complete the procedure. The deep layer helped to satisfy this need. The deep learning and its ability to extract the features, analyze the content and classify helps it strong to be used for this purpose.

Because of its dense layer, the deep learning computational model was ideal for video coding. Since the 1990s, there has been research on neural network-based compression in both video and picture, but it has not been able to prove itself good by delivering a greater compression efficiency since its network was not very deep. As time passed, computational power increased, and it could now handle and train massive databases with deep or many layers. With this information, an overview of NN/DL based compression algorithms and their scope is a worthwhile topic to investigate. Some study schemes are focusing on the implementation of deep learning in traditional coding approaches, i.e., by incorporating deep learning techniques into the existing video coding procedures.

The modelling of deep learning algorithm chooses convolutional neural network as it was good in classifying the details. The network is deep and should extract and analyze the features correctly using the selected kernels, compressing the total content and narrowing down helps the total reduction of the computational complexity. Analyzing the features of deep network with CNN helps the modelling of a new network for reducing the computational complexity, time for encoding is a smart way.

The paper is divided into 3 parts chapter 1 gives an overview of motivation to use CNN network for inter prediction, chapter 2 analyses the proposed design and its features, chapter 3 analyses and evaluate the model with dataset, followed by the results and conclusion.

2. Motivation Behind the Deep CNN Network Design

Signal processing was meant to be the master in video compression field. Traditional signal processing techniques uses the neighboring pixel information to do the predictions [6] in video compression. The shift and rate distortion optimization cost comparison between the pixels helps in intra prediction. The intra prediction procedure uses the similarity index between the pixels within the frame and decides split of the CTU is required or not. Here in this traditional technique the calculations are more that is the computational complexity of the total procedure is high. This was noted as an issue in the HEVC video compressions intra prediction. This research gap was a motivation to design a network with less complexity by maintaining the quality. The overall idea is presented with Figure 1. The proposed model was able to reduce the computational complexity. It is proved by the difference in time of encoding.

The deep learning came into the picture with its goodness in dealing with the content-based analysis than pixel-based calculations. This feature of deep learning helps in resolving the issue faced in intra-prediction in HEVC. The feature extraction [7] in deep learning make the procedure easy and reduce computation. Considering these features, the deep convolutional neutral network [8] showed its power in the designing of this deep depth decision algorithm for intra prediction. The kernel size of each layer decides which features needs to be extracted from that. Here for the design low to high resolution kernels are used to extract the multiple features [9] in the frame. The kernels used for the design are 5x5,3x3,4x4,8x8,16x16 [10] etc. The low sized kernel helps in the extraction of local data while the high sized kernel derives the global information's from the given frame. The design needs a mix of local and global information so a mixture of various size id preferred in the design.

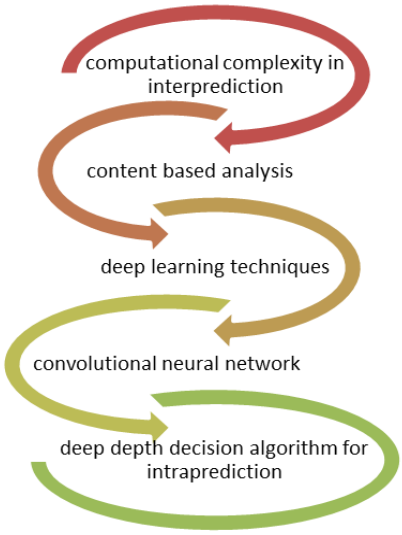

Figure 1. Flow of designing inter prediction with deep depth decision algorithm

The convolutional neural network [8] extract features with kernel and compress it by Maxpool or averaging the information and works in both forward and backward propagation [11]. The activation function is used to excite the neurons. The activation function helps to boost the Maxpooled values to fire the neurons. Here the activation function decided for the design of deep depth decision algorithm is ReLu (rectified linear unit) that maintain the value in the range 0 and infinity as shown in Figure 2. This function helps to the values to remain in positive coordinate thus by helping the firing of neurons. The last stage is flattening the information and crunching it to vector of lower size that holds the depth information of the total frame. This helps in the reduction of computational complexity as the network can predict the depth when input is given as video frame.

The path of development started from a con in the intra prediction that is the computational complexity because of the calculation of RDO cost and its comparison between parent CTU and CU. This leads to the concept of content-based analysis than pixel-based analysis with traditional signal processing. The content-based analysis for comparison makes the procedure simple and calculations less tedious. As the procedure needs an output of correlation between the contents within the frame. Deep learning techniques will be a good resolving this issue [7, 8, 12]. The deep layers help to extract the features and recover the perfect match for intra prediction. The CNN as it is good for comparison it showed its power to match for this. The depth of each CTU is calculated and kept as a 16-length vector by this algorithm thus by reducing the complexity.

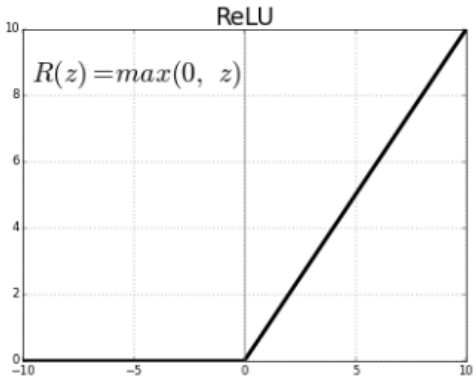

Figure 2. Activation function rectified linear unit

3. Modeling of Sample Deep CNN for Inter Prediction

Wherever TimesFor modeling the network Figure 3 shows the flow of development in the design. The input to the system is a video frame with specific features. Its send through multiple convolutional layers to extract the feature by using various kernels. The information is then Maxpooled and send to fully connected network for flattening the information and for softmax. The output derived is a 16 length vector of information with depth details of 64x64 CTU.

The network that extracts the feature should be flattened by using SoftMax. The data in the model are flattened and SoftMax to a 16-length vector. The whole input is crunched to a 16-length vector by reducing the complexity.

To model the network initially input is chosen with defined format that goes to filter if the input is i filter is represented by f, the convolution layer output z.

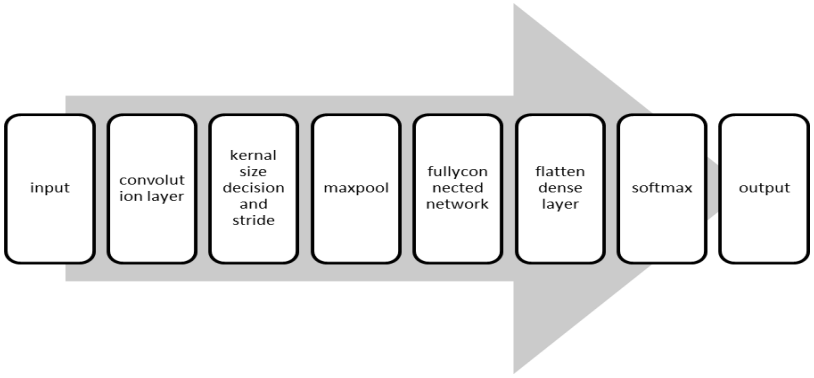

Figure 3. Designing steps for CNN based deep depth decision algorithm

$$z = \mathrm{i} * \mathrm{f} \tag{1}$$

where, * = convolution operation, the size of the output depends on input if input is of size (n x n) and if the size is (m x n) then the output is a size of (n – m + 1, n – m +1).

The pooling layer averages the output of convolution layer, the pooling layer output can be represented as.

$$p = \frac{1}{k}\sum_{a=1}^{k} z_a \tag{2}$$

The fully connected network flattened the information.

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & . & . & p_{1x} \\ p_{21} & p_{22} & p_{23} & . & . & p_{2x} \\ & & . & & & \\ & & . & & & \\ & & . & & & \\ p_{x1} & p_{x2} & p_{x3} & . & . & p_{xy} \end{bmatrix} \rightarrow \begin{bmatrix} p_{11} \\ p_{12} \\ . \\ . \\ . \\ p_{xy} \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ . \\ p_z \end{bmatrix} \tag{3}$$

$$y = \omega^T . p + b \tag{4}$$

where, *y*= output, *w*=weight (randomly initiated), *b*= bias.

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} & . & . & w_{1x} \\ w_{21} & w_{22} & w_{23} & . & . & w_{2x} \\ & & . & & & \\ & & . & & & \\ & & . & & & \\ w_{n1} & w_{n2} & w_{n3} & . & . & w_{nz} \end{bmatrix} . \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ . \\ p_z \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \tag{5}$$

n is the required flattening required.

Activation function used is ReLu:

$$R(z) = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases} \tag{6}$$

The error function can be calculated by:

$$E = |y - y\hat{}| \tag{7}$$

In Eq. (7) the *y* represents the actual output and *y*^ is predicted output. The difference between the actual and predicted input is calculated and with the modules operator analyses the network designed. If error value is high it represents that the variations are more and perdition efficiency is less, whereas if error value is less that shows the similarity index between input and perdition is more and that shows the high efficiency of the designed network.

4. Design of Deep CNN Inter Prediction Network

The model designed for inter-prediction focus on computational complexity reduction [13]. The model's input is video frame and the output is its representation as 16 length vectors. The convolution [14] procedure converts it to a minimum of 16 length vector by using various filter n 4 layers by extracting global and local information. Figure 4 represent the total design and layers in deep depth decision algorithm.

Initially after preprocessing the input to YUV the frame is cropped to 64x64, its send to a convolution layer

with filter size 5x5 to extract the global content and is maxpooled to 4x4 range resulting in a 16x16 image patch. Meanwhile a cropped 32x32 part is convoluted with same filter of 5x5 and pooled with 2x2 to get same 16x16 patch. Both are concatenated and send to various convolution layer of different filters. The activation function used is ReLu and the stride as the length of the filter to reduce computation. The output is sent to fully connected network that flatten the input. The flattened input with weight and bias is compressed to lower sizes. SoftMax is done in last stage to generate the 16-length vector.

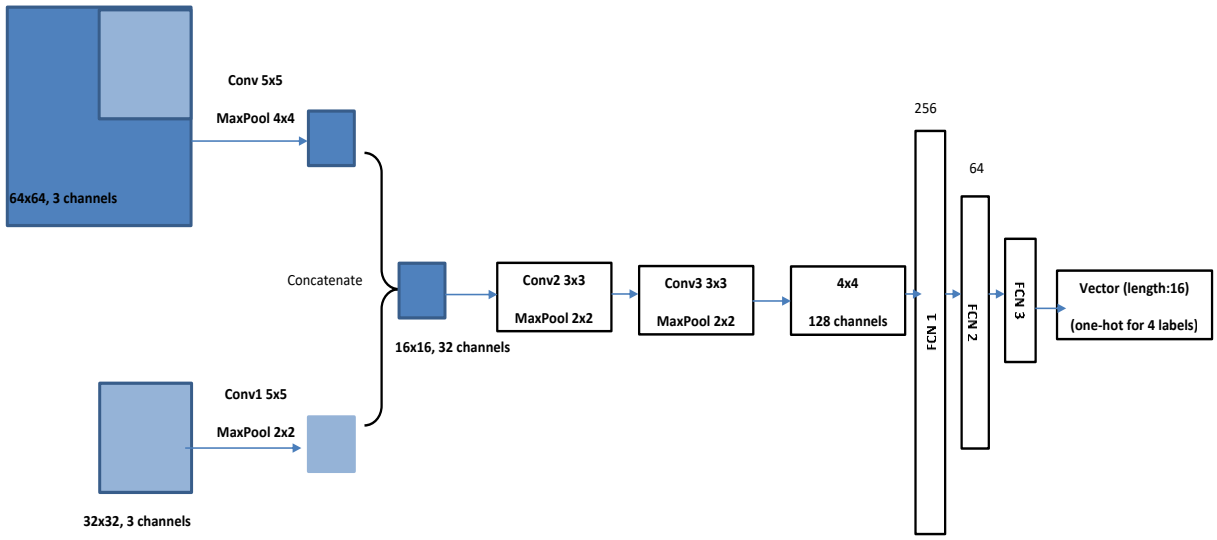

Figure 4. Deep depth decision algorithm with convolutional neural network a sample model [12]

The model designed for inter-prediction focus on computational complexity reduction. The model's input is video frame [15] and the output is its representation as 16 length vectors. The convolution procedure converts it to a minimum of 16 length vector by using various filter n 4 layers by extracting global and local information. Figure 4 represent the total design and layers in deep depth decision algorithm.

Initially after preprocessing the input to YUV [16] the frame is cropped to 64x64, its send to a convolution layer with filter size 5x5 to extract the global content and is maxpooled to 4x4 range resulting in a 16x16 image patch. Meanwhile a cropped 32x32 [17] part is convoluted with same filter of 5x5 and pooled with 2x2 to get same 16x16 patch. Both are concatenated and send to various convolution layer of different filters. The activation function used is ReLu and the stride as the length of the filter to reduce computation. The output is sent to fully connected network that flatten the input. The flattened input with weight and bias is compressed to lower sizes. SoftMax is done in last stage to generate the 16-length vector.

5. Evaluation of Deep Depth Decision Algorithm

The proposed model is evaluated on a sample data set of various video frames. The dataset is rich with various resolution video frames for the test, train and validation. It consists of a video frame and its labels. The frame is split to 64x64, 32x32 etc and its corresponding label s saved in another folder. For each image patch, a python list of 16 element length is saved. Using this input and output folder, the model designed is trained, tested and validated.

The data set used for this purpose has wide range of video frames split into image frames the size of the dataset is 110,000 images for testing and 40,000for validation of various resolution. The images are split into test train and validate to check the performance of the algorithm. Each image has a label associated with it which hold the frame number, CTU number, video number etc.

The image file is further split into 64x64,32x32 size patches for analysis. The label representing it is a python list. Image and label are interlinked. The resolution is of wide range from 4K to lower resolution for evaluation. The images and video frames in the dataset are of YUV with luminance and chrominance data. This data set support only YUV file videos for analysis. The test to validation split ratio is maintained properly and overlapping of members in the data set are avoided to get genuine output. The dataset can be extended for further analysis too. Adding the information to the dataset can enhance the performance of the algorithm.

The loss observed for the testing for cross entropy loss and can be represented as:

$$L = \frac{1}{R}\sum_{\gamma=1}^{R}[y^r log y^{\wedge r} + (1 - y^r)\, log(1 - y^{\wedge r})] \tag{8}$$

The model designed is tested for various input in the data set and the output observed can be noted as follows.

The training loss of the deep depth decision algorithm is calculated as 3.1049. The prediction accuracy of each label in the algorithm is 66.12%. The evaluation of the designed model can be done by using pipelining the proposed algorithm to the original HEVC system and evaluate the performance of both. The performance evaluation is given in Figure 5. Subgraph (a) of Figure 5 represents various video frame samples compressed with original HEVC and subgraph (b) of Figure 5 represents the results of compression by pipelining deep depth decision algorithm in it. The Figure 6 represent the comparison chart showing the 'time of encoding" and "Bitrate" of samples with original method and our proposed method. The comparison chart showing the 'YUV-PSNR" and "Y-PSNR" of samples with original method and our proposed method is plotted in Figure 7. Figure 8 shows the comparison chart showing the 'time of encoding of HEVC and CNN" and "Bitrate of HEVC and CNN" of various samples. The results clearly shows that the encoding time have drastically reduced in proposed method. This proves that proposed method helps in reducing the computational complexity of the intra predation in HEVC.

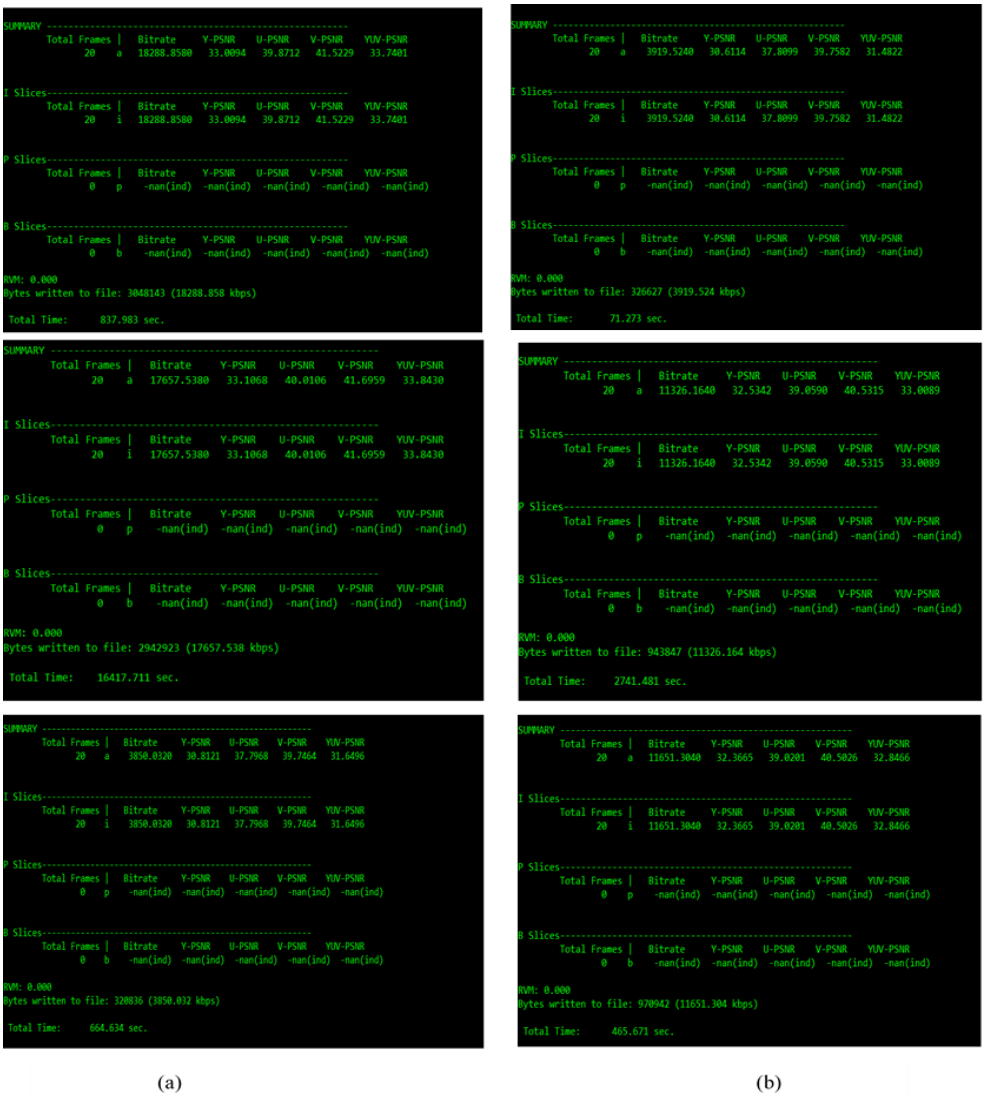

Figure 5. Output window showing the encoding time, Y-PSNR, U-PSNR, V-PSNR, YUV-PSNR of a sample video with 20 frame: (a) using HEVC; (b) by deep depth decision algorithm

The results clearly show the reduction in encoding time thus proves the complexity reduction in computation of intra perdition.

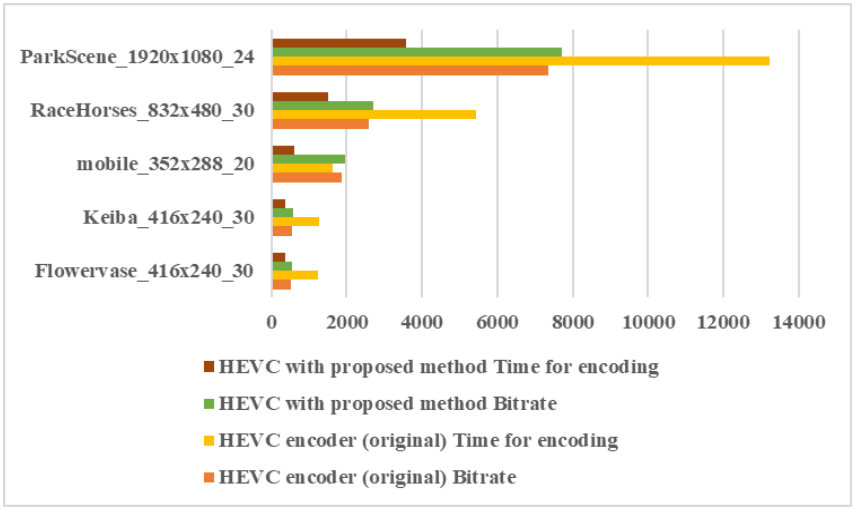

Figure 6. The comparison chart showing the 'time of encoding" and "Bitrate" of samples with original method and our proposed method

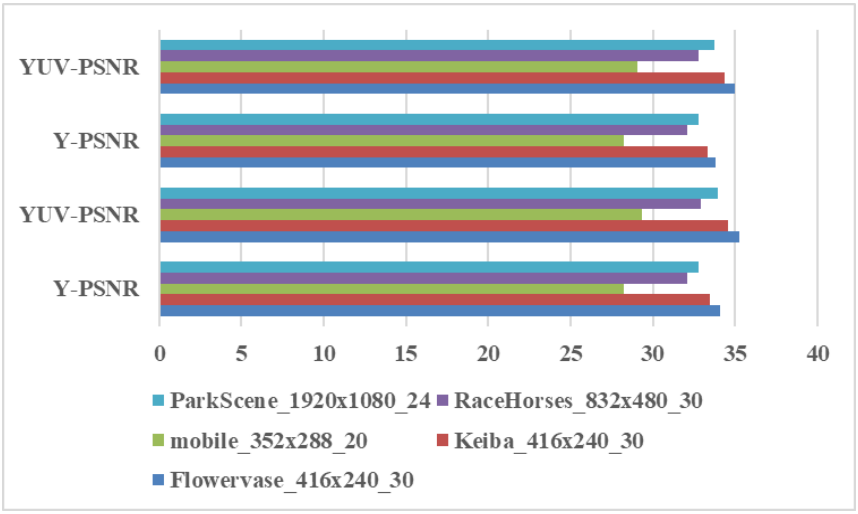

Figure 7. The comparison chart showing the 'YUV-PSNR" and "Y-PSNR" of samples with original method and our proposed method

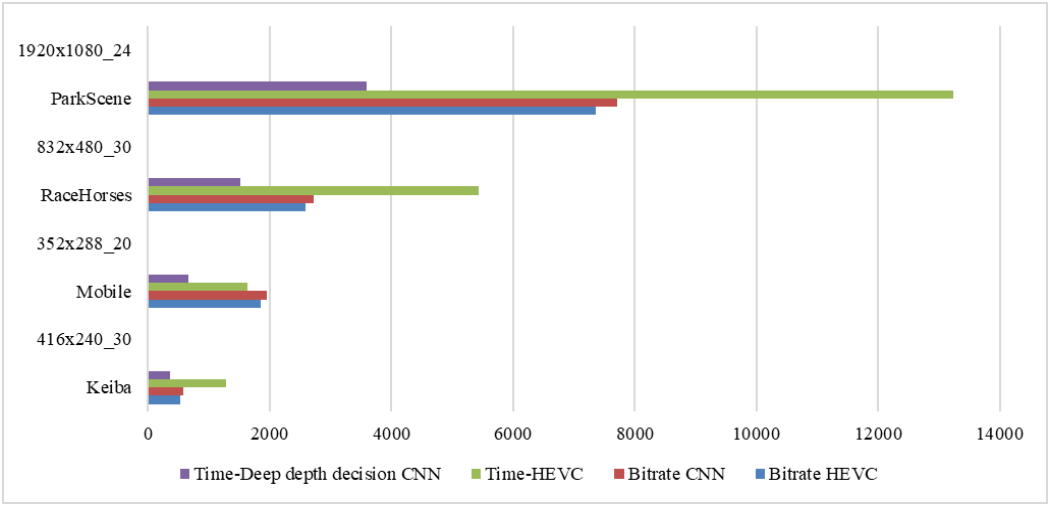

Figure 8. The comparison chart showing the "time of encoding of HEVC and CNN" and "Bitrate of HEVC and CNN" of samples

6. Conclusion

The need for compression in a smarted way leads to the plan of developing a model based on convolutional neural network to reduce the computational complexity in intra-prediction. The complexity of the calculation in HEVC prediction is closely connected to encoding time. A deep depth decision approach is proposed here to evaluate the depth of CTU with a deep CNN network and store it as a 16-element vector by deleting superfluous elements instead of closing it as a 16*16 matrix. This saves not only the bit but also the computation time. The proposed approach is included into the original HEVC encoder (through HM software) to test its performance in a real-world setting. The time for encoding, bit rate, Y-PSNR, and YVU PSNR with and without the deep depth decision technique in the HEVC encoder are all evaluated. The proposed model, its motivation, design and evaluation id incorporated in this paper. The model is designed with the support of deep convolutional neural network with the input as 64X64 coding tree unit and output as 16-length vector. The model chooses each element in the design carefully to extract the essence of the coding tree unit to compress and reconstruct with high efficiency. The kernels, depth, pooling, stride is selected rightly to extract the features. The fully connected network compresses the information. The model possesses a loss of 3.1049 on test model. The accuracy of the prediction while using the dataset on a sample test model is 66.12%. This model can be used to pipeline with the existing HEVC encoder to check its compatibility and efficiency with the existing model.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Acknowledgements

The authors are greatly indebted to the anonymous reviewers whose thought-provoking and encouraging comments have motivated them to modify significantly and update the paper. They also like to express their gratitude to REVA University for extending research facilities to carry out this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] H. Kalva, "Standards: The H.264 video coding standard," *IEEE Multimed.*, vol. 13, no. 4, pp. 86-90, 2006. https://doi.org/10.1109/MMUL.2006.93.

[2] H. K. Joy and M. R. Kounte, "A comprehensive review of traditional video processing," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 6, pp. 274-279, 2020. http://dx.doi.org/10.25046/aj050633.

[3] B. Bross, Y. K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J. R. Ohm, "Overview of the Versatile Video Coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736-3764, 2021. https://doi.org/10.1109/TCSVT.2021.3101953.

[4] A. Dhanalakshmi and G. Nagarajan, "Combined spatial temporal based In-loop filter for scalable extension of HEVC," *ICT Express*, vol. 6, no. 4, pp. 306-311, 2020. https://doi.org/10.1016/j.icte.2020.04.006.

[5] S. Bouaafia, R. Khemiri, F. E. Sayadi, and M. Atri, "Fast CU partition-based machine learning approach for reducing HEVC complexity," *J. Real-Time Image Process.*, vol. 17, no. 1, pp. 185-196, 2020. http://dx.doi.org/10.1007/s11554-019-00936-0.

[6] L. Zhao, S. Q. Wang, X. F. Zhang, S. S. Wang, S. W. Ma, and W. Gao, "Enhanced Ctu-level inter prediction with deep frame rate up-conversion for high efficiency video coding," In 2018 25th IEEE International Conference on Image Processing, (ICIP), Athens, Greece, October 7-10, 2018, IEEE, pp. 206-210. https://doi.org/10.1109/ICIP.2018.8451465.

[7] O. Alharbi, "A deep learning approach combining CNN and Bi-LSTM with SVM classifier for Arabic sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 165-172, 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120618.

[8] H. K. Joy, "Deep learning approach in intra -prediction of high efficiency video coding," In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics, (ICSTCEE), Bengaluru, India, December 8, 2020, IEEE, pp. 134-138. https://doi.org/10.1109/ICSTCEE49637.2020.9277189.

[9] D. Liu, Z. Chen, S. Liu, and F. Wu, "Deep learning-based technology in responses to the joint call for proposals on video compression with capability beyond HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1267-1280, 2020. https://doi.org/10.1109/TCSVT.2019.2945057.

[10] X. Li and N. Gong, “Run-time deep learning enhanced fast coding unit decision for high efficiency video coding,” *J. Circuits Syst. Comput.*, vol. 29, no. 3, pp. 1-19, 2020. https://doi.org/10.1142/S0218126620500462.
[11] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, “Learning convolutional networks for content-weighted image compression,” In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA, December 16, 2018, IEEE, pp. 3214-3223. https://doi.org/10.1109/CVPR.2018.00339.
[12] H. K. Joy, M. R. Kounte, and B. K. Sujatha, “Design and Implementation of deep depth decision algorithm for complexity reduction in High Efficiency Video Coding (HEVC),” *Int. J. Adv. Comput. Science Appl.*, vol. 12, no. 1, pp. 553-560, 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130168.
[13] M. U. K. Khan, M. Shafique, and J. Henkel, “An adaptive complexity reduction scheme with fast prediction unit decision for HEVC intra encoding,” In 2013 IEEE International Conference on Image Processing, (ICIP), Melbourne, VIC, Australia, September 15-18, 2013, IEEE, pp. 1578-1582. https://doi.org/10.1109/ICIP.2013.6738325.
[14] Z. Liu, X. Yu, S. Chen, and D. Wang, “CNN oriented fast HEVC intra CU mode decision,” In 2016 IEEE International Symposium on Circuits and Systems, (ISCAS), Montreal, QC, Canada, May 22-25, 2016, IEEE, pp. 2270-2273. https://doi.org/10.1109/ISCAS.2016.7539036.
[15] Z. Liu, X. Yu, Y. Gao, S. Chen, X. Ji, and D. Wang, “CU partition mode decision for HEVC hardwired Intra encoder using convolution neural network,” *IEEE T. Image Process.*, vol. 25, no. 11, pp. 5088-5103, 2016. https://doi.org/10.1109/tip.2016.2601264.
[16] Y. Zhang, S. Kwong, X. Wang, H. Yuan, Z. Pan, and L. Xu, “Machine learning-based coding unit depth decisions for flexible complexity allocation in high efficiency video coding,” *IEEE TIP*, vol. 24, no. 7, pp. 2225-2238, 2015. https://doi.org/10.1109/tip.2015.2417498.
[17] S. Bouaafia, R. Khemiri, F. E. Sayadi, and M. Atri, “Fast CU partition-based machine learning approach for reducing HEVC complexity,” *J. Real-Time Image Process.*, vol. 17, no. 1, pp. 185-196, 2020. https://doi.org/10.1007/s11554-019-00936-0.

Nomenclature

CTU	Coding Tree Unit
CNN	Convolutional Neural Network
VCC	Versatile Video Coding
HEVC	High Efficiency Video Coding
RDO	Rate distortion optimization cost
CU	Coding Unit
NN/DL	Neural Network/Deep Learning
ReLu	Rectified Linear Unit