UNIFIED MATRIX CODING FOR NN ORIGINATED MIP IN H.266/VVC

Junyan Huo, Yu Sun, Haixin Wang, Shuai Wan, Fuzheng Yang, Ming Li
Email: *jyhuo@mail.xidian.edu.cn* & *ysun_xd@stu.xidian.edu.cn*

Paper ID: 5290

Backgrounds

1. In order to further improve the coding performance, VVC (released in 2020) introduces a large number of new coding tools compared with HEVC, including traditional coding algorithms and NN-based coding algorithms.
2. Matrix-based intra prediction (MIP) is the first tool to apply neural network to video coding in H.266/VVC, originally originating from the complex nonlinear network. After many simplifications, MIP with matrix as the core competes with the traditional mode and is jointly used with the latter in the intra prediction of VVC.
3. When incorporating MIP into H.266/VVC, to handle the diversity of video content, 30 matrices are trained and stored to derive predicted samples.
4. Since matrices from training are usually floating-point values, which should be avoided in H.266/VVC, two parameters, shift and offset, are introduced for each matrix to convert floating-point values to integers.
5. MIP consumes more memory and computing resources because of the need to store matrix, offset, and shift values.

It is of great importance for MIP to design an advanced algorithm which can not only reduce computing and memory overhead but also ensure coding performance as much as possible.

Contributions

1. Based on the analysis of the weight distribution of MIP, the input vector of MIP is improved to make the weight range more concentrated.
2. Based on the improved input vector, a unified integer conversion scheme is proposed, which removes the offset and shift lookup table, and reduces the memory and computing overhead of MIP.
3. Due to its unification, memory reduction, and no coding loss, the proposed scheme has been adopted into H.266/VVC.

Results

Table 1. BD-rate results of the proposed algorithm over VTM7.0 for Each Class in AI configuration

Class	Y	Cb	Cr	enctime	dectime
A1	-0.02%	-0.19%	0.15%	100%	103%
A2	0.00%	0.01%	-0.03%	99%	99%
B	0.01%	0.00%	-0.05%	99%	97%
C	0.00%	-0.19%	0.02%	97%	99%
D	-0.02%	0.09%	0.14%	101%	99%
E	-0.04%	0.16%	0.05%	100%	100%
F	-0.02%	-0.03%	0.05%	99%	99%
Avg	**-0.01%**	**-0.19%**	**0.04%**	**99%**	**99%**

1. The proposed algorithm improves input vector calculation, deriving weight matrices with a small range.
2. The proposed algorithm removes the dependency between the shift and offset with the matrix, saving 300 bits of memory which includes 90 bits of shift and 210 bits of offset, achieving hardware-friendliness.
3. Compared to VTM7.0, the proposed algorithm brings -0.01%, 0.01% and 0.04% BD-rate on average for Y, Cb and Cr components, respectively, with 99% encoding time and 99% decoding time.

Proposed Method

Improved input vector calculation: In order to adapt to the prediction of different blocks, MIP designs different input vector processing methods. By analyzing the weight distribution of MIP, we **unify the calculation method** of input vector, as shown in the following formula, reducing the range of weight values.

$$D_j=\begin{cases} Y'_0-0.5Y_{\max} & j=0 \\ Y'_j-Y'_0 & j=1,\ldots,size(Y'_N)-1 \end{cases} \quad \textbf{and} \quad D_j=Y'_{j+1}-Y'_0 \quad j=0,1,\ldots,size(Y'_N)-2$$

$$D_j=\begin{cases} 0.5Y_{\max}-Y'_0 & j=0 \\ Y'_j-Y'_0 & j=1,\ldots,size(Y'_N)-1 \end{cases} \quad \textbf{and} \quad D_j=Y'_{j+1}-Y'_0 \quad j=0,1,\ldots,size(Y'_N)-2$$

Small _ blcksize *Big _ blocksize*

Unified MIP parameters design: Based on the proposed unified input algorithm, we obtain the space that can improve the accuracy of MIP weight representation. Here, we set **shift to 6** and offset to 32. The MIP prediction vector, P, in the proposed algorithm is derived as:

$$P_i=\left(\left(\sum_j\left(W_{i,j}-offset\right)\cdot D_j\right)>>shift\right)+Y'_0 \quad \Rightarrow \quad P_i=\left(\left(\sum_j\left(W_{i,j}-32\right)\cdot D_j\right)>>6\right)+Y'_0$$

MIP Architecture