|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligencewww.mpai.community |

|  |  |
| --- | --- |
| Nxxx | 2021/12/22 |
| Source | Video group |
| Title | MPAI-EVC Evidence Project report and plan |
| Target | MPAI Members |

The goal of the group is to enhance EVC (Essential Video Coding) using AI-tools to reach at least 25% improvement over the baseline profile. The group is currently working on three coding tool: Intra prediction, Super Resolution, and in-loop filtering. For each tool, in the following we describe the proposed approach and the steps of database building, learning phase and inference.

**Intra prediction tool**

A new EVC predictor, leveraging a CNN-based autoencoder is generated.

In the training phase the autoencoder is trained on AROD and BVI dataset by minimizing the Means Square Error (MSE) between its output and the original picture block.

In the inference phase the autoencoder is fed with blocks neighbouring the block being predicted, the so called context. The encoder sends the 64-by-64 decoded context of each 32-by-32 Coding Unit (CU) and 16-by-16 CU to the autoencoder and returns the new 32-by-32 or 16-by-16 predictor, depending on the case, to the EVC encoder. The EVC codeword used to signal the EVC mode zero (DC) is replaced to signal the new AI-based prediction mode to the decoder. The CNN-autoencoder output replaces the intra predictor unconditionally. The generated bitstream is fully decodable under the assumption that the autoencoder network is also available at the decoder side.

Lot of work was carried out on bug fixing. We discovered that we were improving the most probable mode between the 5 EVC modes and not the DC mode. Fixing this bud led to the improvement in Table 1.

Table 1 shows the BD-rate gains of the DC-enhanced EVC encoder over the 32x32 intra predictor only, with an improved gain from -0.61% to 0.95% on the considered QP range.


Table 1: BD-rate gains over 32x32 intra predictor after bug fixing

We trained over 128x128 context with no benefits as in Table 2.



Table 2: BD-rate gains over 128x128 context

Moreover, we tested the 128x128 context on 4K sequences of the Xiph Netflix 4K dataset (Table 3). Also in this case the benefits are negligible.



Table 3: BD-rate gains over 4K sequences

We worked on the fine tuning of the hyperparameters of the network, training the autoencoder with learning rate (LR) of 10^-4. Table 4 contains the results on HD and 4K sequences. In this case there is an improvement in BD-rate for both the class sequences.



Table 4: BD-rate gains with LR=10^-4 on HD (left) and 4K (right)



Figure 1: PSNR on Basketball sequence

In Figure 1 the PSNR curve of the Basketball sequence with LR=10^-4.

Finally, we worked on the loss function, using ABS (absolute difference) instead of MSE in the training phase. We improved the performance (on 900 epochs) as in Table 5.



Table 5: BD-rate gains using ABS on HD (left) and 4K (right)

Future plans include:

* extending the proposed approach also to 8x8 and 4x4 CUs
* experimenting with other network architectures than convolutional
* removing tanh that may improve convergence speed
* reducing number of parameters

**Super resolution tool**

We built a dataset to train the super resolution network with 3 resolutions (4k, HD, and SD), 4 values of picture quality, two coding tool sets (deblocking enabled, deblocking disabled) for a total of 170 GB dataset.

The super resolution step was added as a post processing tool. The picture before encoding with EVC baseline profile was downscaled and then the super resolution network was applied to the decoded picture to get the native resolution.

We have experimented with the performances of the trained network on 8 sequences for the super resolution SD to HD, and on 3 sequences for the super resolution HD 2 4K.

The group has worked on the computation of the BD-rate SD to HD and the result is -4.701% improvement.

The HD to 4K testing phase has been finalized on all QPs (15,30,37 and 45), and we are in the phase of calculating the BD-rate.

We are double checking our computations because it seems that the training distributed on more GPUs has altered the calculation of the metrics. Due to memory constraint during training on 4K pictures we had to distribute the training on two GPUs and this process somehow has affected the calculation of the PSNR as you can see in the Table 6.



Table 6: PSNR is no longer a monotonic function with respect to the quality factor

The next steps are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Date** | **Topic** | **Who** |
| Intra prediction | 1 meeting cycle | More experiments to improve the BD-rate | Attilio, Alessandra, Roberto |
| 1 meeting cycle | Experiments on 8x8 and 4x4 block size | Attilio |
| 2 meeting cycles | Removing tanh that may improve convergence speed | Attilio |
|  | 2 meeting cycles | Reducing number of parameters | Attilio |
|  | 2 meeting cycles | Find a proxy for the encoding rate | Attilio, Alessandra, Roberto |
| Super Resolution | 2 meeting cycles | More experiments to improve the BD-rate | Francesco, Antonio, Mattia and Alessandro |
|  | 1 meeting cycles | Solve the QP 37 training HD24K issue | Francesco, Antonio, Mattia and Alessandro |
|  | 1 meeting cycles | Visual evaluation of the compressed test sequences | All |
| Next candidate AI-tool | 2 meeting cycle | Evaluation of possible candidate (pros/cons in terms of open source, results..) | All |

**Future plan**

* motion compensation: improve the motion compensation using NN architecture
* inter prediction: use NN architectures to refine the quality of inter-predicted blocks; introduce new inter prediction mode which tries to predict a frame directly without the use of side information; leverage on Optical Flow algorithm for the motion estimation.
* quantization: uniform scalar quantization used in classical video codec standard does not conform to the characteristics of human visual system. It is possible to use a quantization strategy based on neural networks.
* arithmetic encoder: improve the CABAC performance by leveraging NN to directly predict the probability distribution of intra modes instead of the handcraft context models