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

|  |  |
| --- | --- |
| Nxxx | 2022/02/22 |
| 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 tools: 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.

**BVI dataset preparation**

MPAI-EVC has decided to use: **BVI-DVC Part 1 (University of Bristol)**

Data from<https://data.bris.ac.uk/data/dataset/3h0hduxrq4awq2ffvhabjzbzi1>

The sequences need a pre-processing to be used in the MPAI-EVC experiments. In the following the steps adopted by the group to prepare the sequences for the future training.

The Figure 1 illustrates the processing workflow.



Figure 1 processing workflow

YUV content is encoded at fixed QP (as per the Common Test Conditions) to provide references and the downsampled videos to be processed by the super-resolution network. Since the super-res network needs RGB input, the color space conversion process is critical, because conversion between YUV 4:2:0 to RGB 4:4:4 and re-conversion to YUV 4:2:0 causes a loss of information to be minimized.

This is specifically relevant when it comes to the BVI-DVC dataset pre-processing. The decompression from mp4 to YUV format must be carried out using the appropriate color matrix (BT.709 or BT.2020) to avoid loss of information.

Due to the lack of specific information on this subject, we have adopted a heuristic approach by trying to convert content using both matrices and measuring the PSNR between the original sequence and the result of the conversion. Sequences have been classified as BT.709 or BT.2020 depending on the calculated PSNR.

**Further datasets**

The BVI-DVC dataset contains 193 different sequences in four resolutions. To increase content diversity in network training, three other datasets with suitable licenses are being considered:

* Ultravideo dataset, containing 16 4K 10-bit raw sequences, available here <http://ultravideo.fi/>
* The opensource SVT datasets (7 new plus 5 old 4K sequences), available here <https://www.svt.se/opensource/content>
* The Tencent video dataset (85 4K sequences), available here <https://multimedia.tencent.com/resources/tvd>

Other datasets, such as Youtube UGC dataset are being investigated.

**Intra prediction tool**

We address the challenge of predicting a block given its context (Intra prediction) as an image inpainting problem, i.e. recovering pixels of an image that are unavailable due to, e.g. occlusions or information loss. Masked convolutional neural networks have been recently proposed for image inpainting exploiting the apriori information from the context to recover the missing image area.

The method we propose relies on masked convolutions to generate the block predictor starting from a decoded context of 64 × 64 pixel (Figure 2). For example, for each 32x32 coding unit a 64x64 context is sent to the autoencoder. Then the autoencoder return a 32x32 new predictor and we replaced EVC predictor mode 0 (i.e., the DC mode predictor) with a new predictor that is computed by a masked convolutional autoencoder. We replace the DC predictor unconditionally so we don’t need additional signaling to the decoder and the bitstream is decodable. The generated bitstream is fully decodable under the assumption that the autoencoder network is also available at the decoder side.



Figure 2: context con the left and the predictor on the right

The masked autoencoder (Figure 3) is trained in a supervised manner extracting random patches

from about 800 images representing various types of contents.

In the training phase the autoencoder is trained on the BVI dataset by minimising absolute error (ABS) between its output and the original picture block. The training lasts 1000 epochs.



Figure 3: the schema of the masked autoencoder

In the following tables the results of the experiments of the DC-enhanced EVC encoder, minimising ABS, over 32x32+16x16+8x8+4x4 block size on JVET test sequences.



Table 1 JVET Class A, 3840x2160, **QP 22-37** (left table), **QP 22-47** (right table)



Table 2 JVET Class B, 1920x1080, **QP 22-37** (left table), **QP 22-47** (right table)



Table 3 JVET Class C, 832x480, **QP 22-37** (left table), **QP 22-47** (right table)



Table 4 JVET Class D, 416x240, **QP 22-37** (left table), **QP 22-47** (right table)



Table 5 JVET Class E, 1280x720, **QP 22-37** (left table), **QP 22-47** (right table)



Table 6 JVET Class F, computer screen, **QP 22-37** (left table), **QP 22-47** (right table)



Table 7 Recap on JVET sequences

Table 7 shows that we obtain low performance on low resolution contents and low performance on some screen contents.



Table 8 JVET Class B, 1920x1080, Oracle, **QP 22-37** (left table), **QP 22-47** (right table)

We are working on adding a 6th intra mode and we have estimated the maximum gain. Table 8 shows the Oracle mode, i.e. a non-decodable bit stream, in order to get an upper bound for the gain.

Future plans include:

* adding a 6th EVC intra mode
* reducing number of parameters
* encode sequences for crop mode

**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 carried out an extensive training of the selected deep-learning approach for super resolution on 4 QPs (15,30,37 and 45). We have tested its performances on 8 test sequences for the case of SD to HD, and on 3 test sequences for the case of HD to 4K.

The group has worked on the computation of the BD-rate SD to HD and this has shown an improvement of -4.701% when compared with the ground truth EVC.

For the HD to 4K we are in the phase of calculating the BD-rate.

Due to the fact that the training step for the case HD 2 4K, was carried out on different machines we have been carrying out substantial work for verification purposes. We have divided the verification process in two steps. First, we have verified that the training carried out by the two partners using a training set composed of exacted patches provided by one partner, provides similar results. This has shown that the results were similar among the partners involved in this activity. Second, due to the fact that there are different implementations and parameters settings for the patch extractions, we are verifying which one gives better training results. This activity is still ongoing and results will be available soon.

The next steps are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Date** | **Topic** | **Who** |
| Intra prediction | 1 meeting cycle | More experiments to improve the BD-rate | Attilio, Alessandra, Roberto |
| 2 meeting cycles | Adding 6-th Intra mode  | 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 | Verify patch extraction among partners | 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