The factsheet of Deep Video Compression based on the Context-Adaptive Entropy Model for P-frame Compression

Woonsung Park Sehwan Ki Munchurl Kim
Korea Advanced Institute of Science and Technology
335 Gwahak-ro, Yusong-Gu, Daejeon, 305-701, Korea
{pys5309, shki, mkimee}@kaist.ac.kr

Abstract

In this paper, we propose an end-to-end deep video compression method based on the context-adaptive entropy model for P-frame. First, we compress joint information which consist of motion information and residuals between the current frame and the predicted frame. Next, we propose a context-adaptive entropy model with the Gaussian modeling using the context information of the previous frame. Finally, we add an enhancement layer for the quality enhancement of the reconstructed frame. Our proposed deep video compression network are jointly optimized with both distortion and rate loss. Also, our method shows better compression performance than the baseline model of ours in terms of multi-scale structural similarity (MS-SSIM).

1. Proposed Method

Our team name is 'KAIST-VIC' for P-frame compression challenge. Our proposed deep video compression architecture for P-frame compression is illustrated in the Figure 1. As depicted in Fig. 1, the current frame x_0 and the previous frame x_{-1} with YUV format are converted to RGB formats for estimating the optical flow $F_{0\to-1}$. The joint information with the optical flow $F_{0\to-1}$ and the residual r_0 are mapped to the latent space y_0 through the encoder network. After the quantization step, we can obtain the quantized latent represenatiton $\hat{y}_0$. Then the reconstructed optical flow $\hat{F}_{0\to-1}$ and residual $\hat{r}_0$ are estimated by the decoder network with the entropy model of $\hat{y}_0$. The entropy model of $\hat{y}_0$ is based on a Gaussian model with mean μ and standard deviation σ, which are estimated through a hyperprior encoder-decoder network and a Context-Net with hyperprior $\hat{z}_0$ and the previous frame x_{-1}. Then the reconstructed frame $\hat{x}_0$ is obtained by adding the reconstructed residual $\hat{r}_0$ and prediction frame $\hat{p}_0$. Finally, the enhancement layer outputs enhanced frame $\tilde{x}_0$ from the reconstructed frame $\hat{x}_0$. The details of the proposed network are described in Section 1.1-1.4.

1.1. Compressing joint information

We proposed the structure that compresses joint information with motion information and residual, unlike the framework in conventional video compression. Conventional video compression is designed to compress optical flow and residual separately, but the proposed structure compresses joint information with motion information and residual to reduce coding efficiency under the assumption that the redundancy between motion information and residual exists. Note that we utilized PWC-Net [2] for estimating the optical flow between the current frame and previous frame.

1.2. The context-adaptive entropy model

In order to improve the coding efficiency, we proposed the context-adaptive entropy model of the quantized latent represenation $\hat{y}_0$. For the proposed entropy model, the hyperprior $\hat{z}_0$ and hyper encoder-decoder network are utilized. We follow the same design in [1] for the hyper encode-decoder network. Also, since there is redundancy between the current frame and the context information of the previos frame, we added Context-Net to estimate the mean and standard deviation of the Gaussian model for the context-adaptive entropy model. Our proposed Context-Net extracts the context information c of $\hat{y}_0$ and the context information of the previous frame x_{-1} and concatenates to obtain the mean and standard deviation of the Gaussian-based context adaptive entropy model. The details of Context-Net are illustrated in Figure 2.

1.3. The enhancement layer

To further improve the quality of the reconstructed frame, an Enhancement Net was added to enable the role of a deblocking filter or SAO in the traditional video compression. We used 5 layers of ResDenseNet [3] for our Enhancement Net, which 3 RDN blocks were used for each layer.

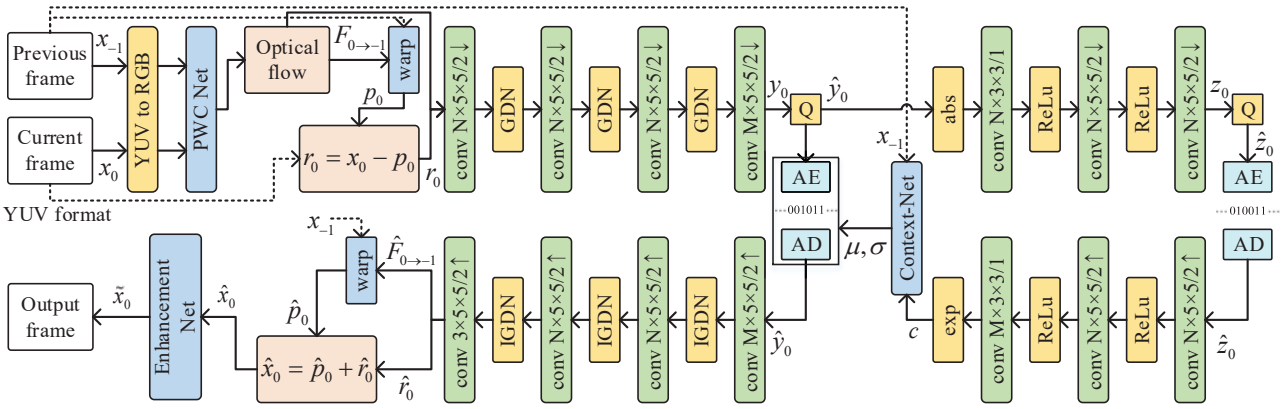

Figure 1. Overall architecture of the proposed deep video compression for P-frame compression

1.4. Training Procedure

The proposed structure was trained in an end-to-end manner, and we used rate-distortion loss as the loss function of the total network, which is frequently used in existing deep image compression [1]. For the distortion loss, 1-(MS-SSIM) was used for training in our experiment. Also, we used three different λ models ($\lambda = 10, 20, 40$) for matching the target bitrate. In our experiment, we use $N = 128$ and $M = 256$ for both training and validation phase.

2. Experiments

The best performance among the submitted decoders of our team 'KAIST-VIC' is 0.98717493 in terms of MS-SSIM (which is the leaderboard score). We think that there are some decoder bugs about entropy coding issues for our submitted decoder. In our experiment environment, the optimum performance for the same network with the submitted decoder is 0.993034158 in terms of MS-SSIM with the same bitrate. The performance of our baseline network which excludes three proposed contributions from the total network is 0.982968669 in terms of MS-SSIM with the similar bitrate.

3. Conclusion

In this paper, we proposed the end-to-end deep video compression method based on the context-adaptive entropy model for P-frame. Our architecture can reduce the coding efficiency of P-frame compression by compressing joint information with motion information and residual, using the context information of the previous frame for the context-adaptive entropy model, and adding the enhancement layer. Since we used a light decoder network with less than 12GB of memory, we think that better compression performance can be expected with more parameters of our decoder network.

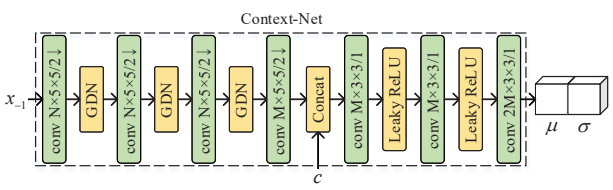

Figure 2. Our proposed Context-Net of the deep video compression for P-frame compression

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

[2] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[3] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.