ofinno

Architecture of Deep Neural Network-based Video Compression

Young-Yoon Lee

Ofinno • ofinno.com • October 2021

Architecture of Deep Neural Network-based Video Compression

1. Introduction

In our previous article [1], we introduced the status of Deep Neural Network-based Video Coding (DNNVC) approaches in the Moving Picture Expert Group (MPEG), one of the most important standardization groups for video compression technologies. In principle, video compression systems seek to minimize the end-to-end reconstruction distortion under a given bit rate budget, called a rate-distortion (R-D) optimization problem. To this end, a lot of efforts in video compression had been focused on the development of video coding tools, such as prediction, transform, entropy coding, and visual quality enhancement. These tools are devised to exploit spatial, temporal, and statistical redundancies in video signals. FIG.1 illustrates a conventional video compression system.

In recent years, the utilization of the increasing computational resources and the tremendous volume of data has led to the unprecedent success of Deep Neural Network (DNN) technologies in the field of artificial intelligence (AI); DNN has brought new opportunities in the development of video compression. MPEG started an exploration experiment (EE) on DNNVC at the 133rd MPEG meeting in January 2021. From the architectural viewpoint, DNN plays a role in two different ways in video compression: Hybrid block-based coding with DNN (or Hybrid coding), and End to End learning based coding (or E2E coding). In the following, we review the architectures of DNNVC including two approaches: Hybrid and E2E coding approaches.

2. Hybrid block-based coding with DNN approaches

In hybrid coding approaches, some of the conventional video coding tools are replaced by DNNs while the architecture of the conventional video codec is preserved. To get an additional gain to the conventional video codec, individual coding tools are optimized independently. The improvement of components in FIG. 1 has been widely and effectively exploited using DNNs owing to a remarkable progress in DNN-based image enhancement, such as image filtering [2] and image super resolution [3]. For example, DNN-based Loop Filter (DNNLF) can reduce the distortion by enhancing the visual quality of the reconstructed video of the conventional video codec. The breakthroughs of DNN-based super resolution (DNNSR) can make it possible to recover the finer details from a spatially down sampled image. DNNSR can be appended to the conventional video codec for a down sampled input frame in FIG. 2. In recent EE activities of MPEG [4], DNNLF and DNNSR have demonstrated up to 12.3% and 9.8% bit rate savings compared with the Versatile Video Coding (VVC or H.266) video codec, respectively. .

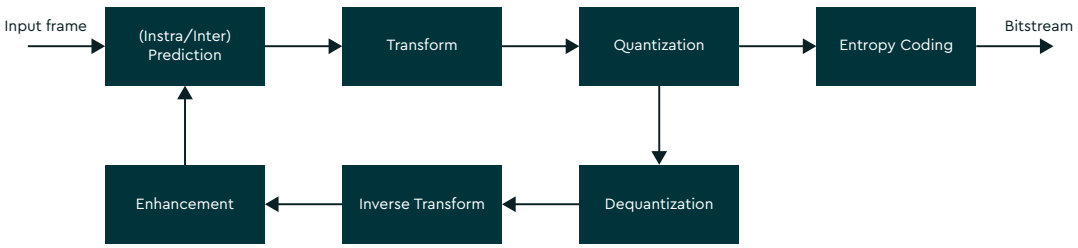

FIG. 1: A conventional video compression system

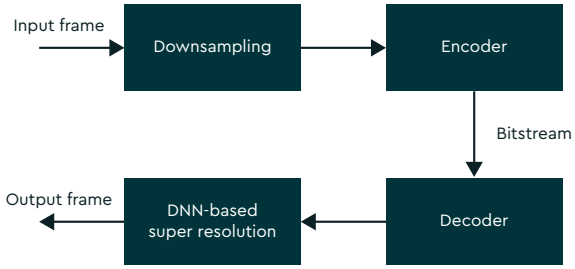

FIG. 2 : A video coding architecture with DNN-based super resolution

3. End to End learning based coding approaches

In E2E coding approaches, input images and videos are compactly represented by end-to-end supervised learning scheme. Most E2E coding approaches still follow the conventional coding paradigm in which the modules, such as transformation, quantization, entropy estimation, and enhancement can be jointly optimized.

In contrast to the linear transform in the conventional video codec, nonlinear transforms have been investigated based on the autoencoder with an entropy bottleneck [5]. Analysis transform was trained so that it could map the image into a low entropy latent representation from which synthesis transform could reconstruct the image in the left side of FIG. 3. In another aspect, this method has paved the way for many subsequent approaches which predict the latent space distribution conditioned by a context model, called a hyperprior model in the right side of FIG. 3 [6, 7]. For analysis transform and synthesis transform, the recurrent neural network (RNN) [8] as well as the convolutional neural network (CNN) has been studied to perform the R-D optimization adaptively.

To improve the visual quality of the reconstructed image, there has been much research taking advantage of the generative capability of generative adversarial network (GAN) and finding optimal loss functions. FIG. 4 illustrates the autoencoder architecture in which the generative network of a GAN can be interpreted as synthesis transform in autoencoder.

Quantization is required to assign values into discrete symbols for entropy coding in data compression. In the aspect of E2E learning scheme, quantization-related operation should be differentiable to optimize the neural network parameters by backpropagation.

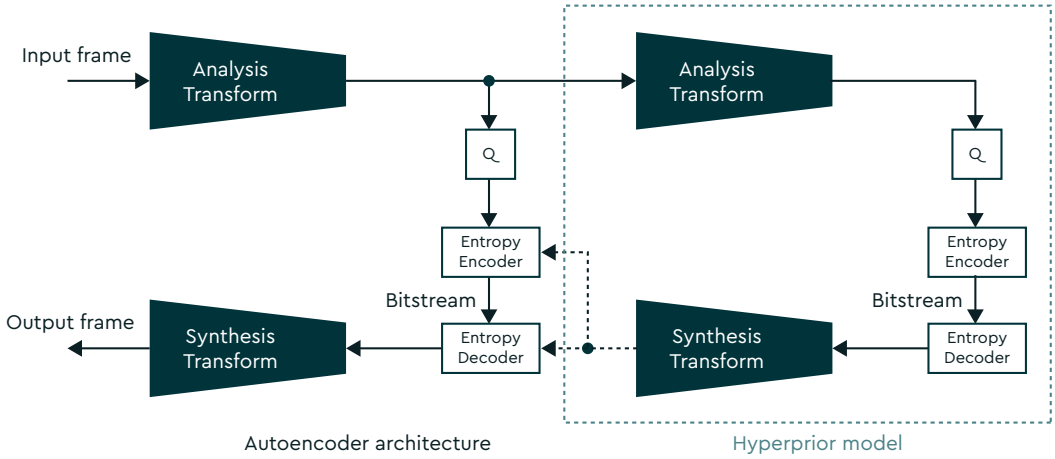

FIG. 3: Autoencoder architecture with a hyperprior model

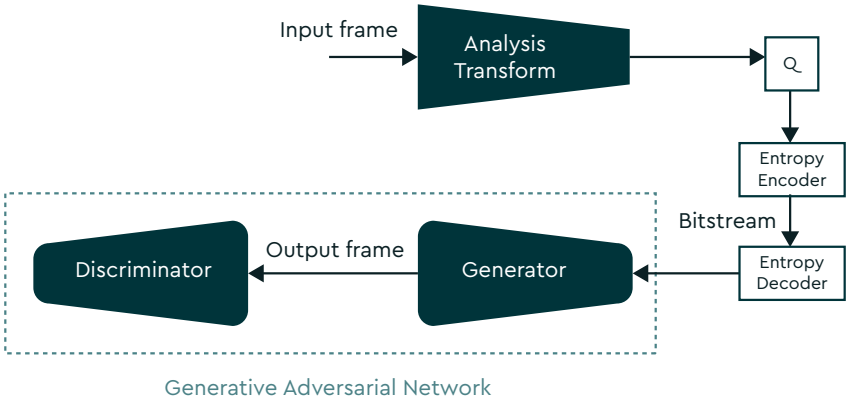

FIG. 4: Autoencoder architecture with GAN

To this end, several approaches, such as adding uniform noise [5], stochastic round operation [8], and soft quantization [9], were exploited.

In comparison to E2E image compression, E2E video compression has additionally focused on delving into temporal redundancy, which can be delicately reduced by learning-based temporal interpolation. FIG. 5 illustrates a simplified architecture of inter frame coding with inter prediction network and residual coding network. The inter prediction network and the residual coding network can adopt autoencoder architecture to compress motion information and residual information, respectively [10, 11, 12]. Reference frames can be reconstructed using E2E image compression or generated autoregressively. Using input frames and reference frames, inter prediction network estimates the optical flow and finds a compact motion representation from which inter prediction frame can be interpolated by synthesis transform of autoencoder. Residual, the difference between a prediction by inter prediction network and the input frame, can be compressed via residual coding network in the same manner as E2E image compression.

4. Conclusion

In this article, we reviewed the advances in DNNVC architectures. Hybrid coding approaches are briefly summarized and then E2E coding approaches are discussed in detail. Though E2E coding is still in its infant stage, this approach can bring fast-growing compression efficiency because all the parameters in an autoencoder architecture can be learned in an automatic and unsupervised manner, thereby it is expected to become more generalized and more efficient.

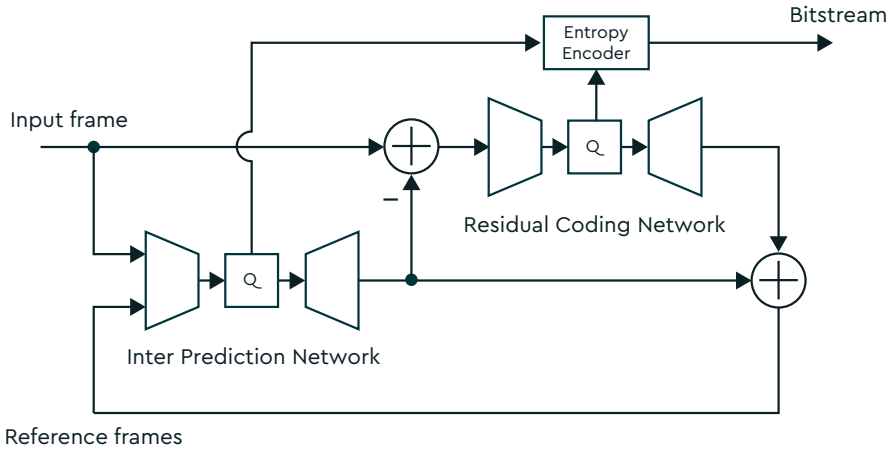

FIG. 5: A simplified architecture of inter frame coding with inter prediction network and residual coding network

5. Acronym List

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DNNLF	Deep Neural Network-based Loop Filter
DNNSR	Deep Neural Network-based Super Resolution
DNN	Deep Neural Network
DNNVC	DNN-based Video Coding
E2E	End to End
EE	Exploration Experiment
GAN	Generative Adversarial Network
MPEG	Moving Picture Experts Group
R-D	Rate-Distortion
RNN	Recurrent Neural Network
VVC	Versatile Video Coding

References

1. "Deep Neural Network based Video Compression for Next Generation MPEG Video Codec Standardization," Tae Meon Bae, http://ofinno.com/article, Oct. 2020
2. "A machine learning approach for non-blind image deconvolution", Christian J. Shuler et al., CVPR 2013
3. "Photo-realistic single image super-resolution using a generative adversarial network," Christian Ledig et al., CVPR 2017
4. "BoG Report: Neural Networks Video Coding Analysis and Planning," m57597, 135th MPEG meeting, online, July 2021
5. "End-to-end optimized image compression," Johannes Ballé et al., ICLR 2017
6. "Variational image compression with a scale hyperprior," Johannes Ballé et al., ICLR 2018
7. "Joint autoregressive and hierarchical priors for learned image compression," David Minnen et al., NIPS 2018
8. "Variable rate image compression with recurrent neural networks," George Toderici et al., ICLR 2016
9. "Conditional Probability Models for Deep Image Compression," Fabien Mentzer et al., CVPR 2018
10. "Deepcoder: A deep neural network based video compression," Ruihao Gong et al., VCIP 2017
11. "DVC: An End-to-End Deep Video Compression Framework," Guo Lu et al., CVPR 2019
12. "Learning for Video Compression with Hierarchical Quality and Recurrent Enhancement," Ren Yang et al., CVPR 2020

About the Author:

Young-Yoon Lee is currently a principal scientist in Ofinno, Reston, VA. Before joining Ofinno in Sep. 2020, he was a senior engineer and a principal engineer in Samsung Electronics Co., Ltd., South Korea from 2008 and 2017, respectively. From 2014 to 2015, he was a visiting scholar with MediaX, Stanford University, Stanford, CA. He has published more than 20 journals and conference papers. His current research interests include image and video processing, computer vision, machine learning applications, and optimization.

About Ofinno:

Ofinno, LLC, is a research and development lab based in Northern Virginia, that specializes in inventing and patenting future technologies including 5G, IEEE, and video compression. Ofinno's researchers create technologies that address some of the most important issues faced by wireless device users and the carriers that serve them. Ofinno's research involves technologies such as 5G Radio and Core networks, IoT, V2X, and ultra-reliable low latency communications. Our innovators create the technologies and oversee the entire process from design to the time the technology is sold. For more information about Ofinno, please visit www.ofinno.com.