International Journal of Electrical Engineering and Technology (IJEET)
Volume 11, Issue 5, July 2020, pp. 137-143, Article ID: IJEET_11_05_016
Available online at http://www.iaeme.com/ijeet/issues.asp?JType=IJEET&VType=11&IType=5
ISSN Print: 0976-6545 and ISSN Online: 0976-6553
DOI: 10.34218/IJEET.11.5.2020.016

PERFORMANCE ANALYSIS OF ADVANCEMENTS IN VIDEO COMPRESSION WITH DEEP LEARNING

Sangeeta and Preeti Gulia
Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India

ABSTRACT

Video content over the internet is increasing day by day with the increasing trends of live video streaming services. People use to capture, share and save their various moments of life using videos. The main challenge before video compression emerged to deal with high quality video content. This led to the emersion of highly efficient and powerful video compression techniques. Videos are disseminated over the internet using efficient and powerful video compression techniques. Existing video compression techniques are designed and optimized manually. Recent researches have shown that deep learning based video compression techniques are giving comparable and better results in comparison to the existing traditional techniques. These results showed the ways to the researchers to work in the direction of applying deep learning concepts in video compression for their practical applicability. This paper gives an insight into the various recent deep learning based video compression techniques and their comparative analysis based on various parameters pertaining to their architectures, compression results, training set, data set, VQMs etc. The comparative and performance analysis presents a future endeavor for scope of further enhancements and optimizations.

Key words: Deep Learning, Video Compression, CNN (Convolutional Neural Network), MS-SSIM (Multi-Scale Structural Similarity Index), PSNR.

Cite this Article: Sangeeta and Preeti Gulia, Performance Analysis of Advancements in Video Compression with Deep Learning. *International Journal of Electrical Engineering and Technology*, 11(5), 2020, pp. 137-143.
http://www.iaeme.com/IJEET/issues.asp?JType=IJEET&VType=11&IType=5

1. INTRODUCTION

Video content on internet contributes more than 75 percent to all of IP traffic (both business and consumer) and as per cisco study, it will grow fourfold and will take more than 82 percent by 2022 [1]. This led to the need of efficient and powerful video compression techniques which will consume less internet traffic and save the storage space. These techniques will prove as a milestone in various domains like 3D gaming and real time video streaming. Some

of the currently used video compression techniques are manually designed. Manual designing require lot of expertise and also results in slow development. In the last few years, researchers have shown the capabilities vest in deep learning based video compression techniques. Various international funding agencies are also supporting the research in this new field for its applicability in real time applications. In this paper, various techniques like interpolation, importance maps and priming etc. pertaining to video compression domain have been discussed.

2. LITERATURE SURVEY

The need of highly efficient and automated video compression techniques motivated the researchers to explore the new avenues. Several new models were proposed based on different deep learning techniques and they were experimented on different data sets.

According to Chao-Yuan Wu *et al*., the concept of image interpolation can be used in deep learning based video compression [2]. End to end trainable architecture was used. Firstly, the the key frames are encoded using deep image compression, and then reconstruct the remaining frames using vanilla U-net interpolation architecture. In this architecture, offline motion estimation techniques like block motion estimation or optical flow are also used as the used interpolation architecture cannot properly disambiguate the trajectories of moving patterns. The remaining spatial redundancy is reduced and compression is done by using the same architecture and adaptive arithmetic coding technique as used by Oord *et al*. in [3]. Moreover, the image interpolation is used in hierarchical manner for the further reduction in bit rate. This model was trained on only one data set, but applied to three different datasets. The results have shown that this proposed model performs better in comparison to HEVC (High Efficiency Video Coding) and MPEG-4 Part 2. Moreover, the model also provides good compression quality when compared to H.261 and H.264. The compression quality was measured in terms of PSNR when applied on high high-resolution UVG dataset. These results provide the new paths for the further research in this field as the model achieves state-of-the-art performance in comparison to the existing traditional codecs without sophisticated heuristics or excessive engineering.

Nick Johnston *et al*. incorporated three additional features in the previous model to improve lossy image compression as proposed in [4]. This model was based on recurrent CNN architecture. Firstly, the network was trained with pixel-wise loss measured in terms of SSIM. Moreover, the recurrent architecture was slightly modified to make the spatial diffusion better and the hidden states of the network can effectively capture and propagate the image information. Thirdly, spatially adaptive bit allocation algorithm is used so that the image can be encoded efficiently with minimum number of bits. The encoding is achieved progressively with the recurrent iterations. Each layer propagates only the subset of its code; hence different bit rates are achieved. The proposed model in this paper uses the same recurrent architecture as in [5]. In this architecture, the missing codes are overlooked and the decoder runs only fewer iterations for low bit rate encodings, results into less accurate but valid reconstruction. This model uses the concept of priming for initial iterations. The concept of priming is used for the first iteration only to improve its recurrent depth, but these additional steps are used by the encoder to avoid the additional bandwidth and to learn for the actual bit stream of the image. With the multiple processing of the image, the successive encoders will ignore the bits that were generated but only track the changes in the hidden state of the recurrent encoders. On the decoder side, the first convincing set of bits transmitted is taken and decoded image is created multiple times and final image reconstruction is achieved. The amalgam of three improvements hidden-state priming, perceptually weighted training

loss and spatially adaptive bit rates led to significant improvement in the reconstruction. The model performs better than many traditional image codecs.

Generally, it is observed that the information content within the image is not uniformly distributed but it is spatially variant. Mu Li *et al.* used the same concept and created the importance map of the image [6]. To highlight the different parts of the image, a content weighted importance map is generated. This importance map is used for discrete entropy estimation and thereafter compression rate and the binarizer are adapted accordingly. The importance map, binarizer, encoder and decoder can be optimized jointly in an end-to-end manner. The compression level of different parts of the image is different. In [7, 8], the code length after quantization is spatially invariant, and entropy coding is then used for further compression. The even areas of the image have less content hence those areas can be easily compressed, but the areas having salient objects or rich textures, have large and complex content, more bits will be required for their representation, hence led to the more complexity in compression. The experiments show that this model provides good image quality in comparison with traditional image codecs when measured in terms of SSIM.

The field of end-to-end image compression, using autoencoder like techniques, has emerged since last two years, providing further directions to the researchers in the field of video compression. The major challenge in the learning based video compression is the motion prediction. Zhibo Chen *et al*. introduced the concept of Pixel Motion CNN (PMCNN) for modeling the spatio-temporal coherence between the images to effectively perform predictive coding inside the learning network [9]. The concept is based on three concepts-predictive coding, iterative analysis/synthesis and binarization. The predictive coding is based on the model of PMCNN. The iterative analysis/synthesis is based on the model as proposed by Toderici et al. in [10], which comprises of several LSTM-based auto-encoders adjacently connected. The differences between reconstruction and target are analyzed and synthesized iteratively to give a variable-rate compression. The quantitative results have shown the efficiency of this model over the existing codecs and provide new directions to further researches in the field of video compression by incorporating various aspects.

The researchers also attempted to use autoencoders in the image compression to analyze its performance and applicability and later can be further extended for videos, if results in significant outcomes. Thierry Dumas *et al.* in their paper [11] tries to explore whether the learning in image compression is dependent on quantization. All image coding standards transform the image into a form which can be then scalar quantized, and then compression algorithm is applied. This paper focuses to learn transform and quantization. The quantization step size doesn't keep same during training and testing but varied and it was found that this learned approach gives comparable results with the existing codecs. This attempt encouraged the use of autoencoders for videos also.

Guo Lu *et al.* in [12] presented a new end to end trained framework for video compression based on the conventional methods of predictive coding using deep learning. The model relies on the optical flow map for motion estimation and reconstruction of frames. All the modules in the network are jointly trained and optimized using the same loss function. In the network of this proposed model, all the modules are deep learning based and developed in one to one correspondence with the conventional architectures so that better performance can be achieved by replacing existing modules with their better architectures like better optical flow estimation. This framework is modeled using Tensorflow. The ablation study and testing results reveal that the proposed architecture performs better than H.264 and gives comparable results with H.265 when compared in PSNR and MS-SSIM. This framework provides new opportunities for further research in this field by incorporating new techniques of different modules.

Oren Rippel *et al.* in their paper [13] presented an enhanced deep learning based architecture for low latency mode. This model also use the arbitrary information learned by the network along with the reference frames. Moreover, the optical flow and the residuals are compressed jointly to enhance the performance. Machine Learning based spatial bit allocation is used for the efficient bit rate and bit rate controller algorithm is used to control the bit rate. The experimental results show the efficiency and potential of the proposed model for future endeavors in this field for its applicability in real time applications. This model is slow to be deployed in practical applications but advancements in its computational domain may give better results.

Other than designing new video compression techniques, there have been developments on improving existing codecs which are used by mass consumers. Researchers are working in the field of applying learning based optimization techniques to enhance the performance of traditional codecs [14, 15]. Every image has some sensitive area that should be less deformed. A deformation-insensitive error measure is proposed to represent the different parts of the image in terms of deformation insensitivity. This technique represents the preprocessing of image before compression. It can be easily embedded with the existing compression schemes [16]. This paper mainly focuses on the visual quality of the image. The main concept is that prior to optimal compression, the input image is slightly deformed. After such deformation, the image can be more compressed and these deformations are so small to be noticed. Moreover, they facilitate the codec to keep the details that are otherwise completely lost. The experimental results show that this technique works powerfully to improve the visual quality of the image.

Motivated by the recent advances in the field of deep learning based approaches for video compression, Jun Han *et al.* proposed and experimented the use of Variational Autoencoders for unsupervised learning in video compression. This method is not based on block motion estimation but rely on assigning probabilities to the segments. In this approach, entropy coding is done along with transform learning to reduce the video frames in low dimensions. It works well if trained with same type of videos and performance is limited in diversified categories of videos [17]. These experiments show various recent advancements in the field of video compression using machine learning based techniques.

3. COMPARATIVE ANALYSIS

Some of the recent major breakthrough researches in this field of video compression are summarized in the table given below. The table presents the different techniques proposed, developed and experimented with their relative datasets used for both training and testing. It shows the trade-off between the performances of different architecture with their complexities. Some of the outperforming architectures suffer from complexity concerns and in some modules, the performance limited to small video clips and homogeneous video content only. This comparative study and analysis of different architectures reveals the relative scope of improvement in different modules.

Table 1 Comparative of some recent researches in video compression

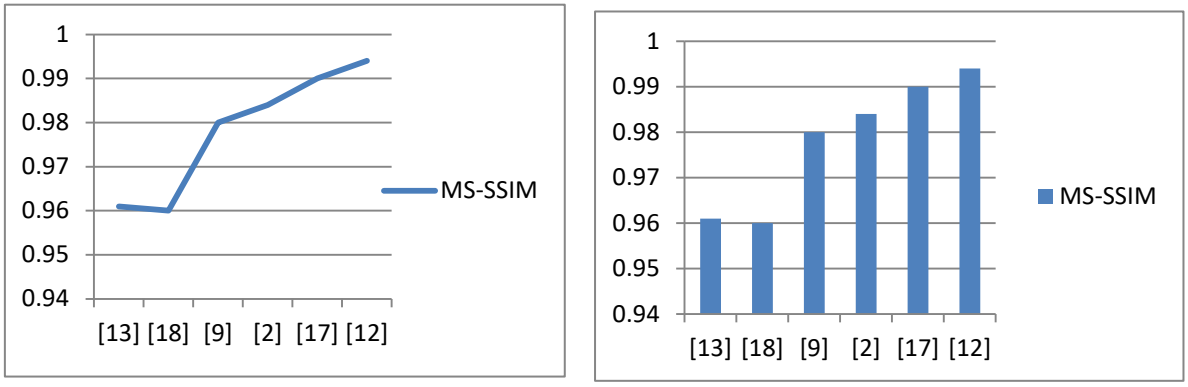
Authors' Name	Paper Name	Techniques Used	Dataset Used
Guo Lu et al. (2019).	DVC: An End-to-end Deep Video Compression Framework	Learning based optical flow estimation Entropy Coding	Vimeo-90k dataset. UVG dataset
Sungsoo Kim et al. (2018)	Adversarial Video Compression Guided by Soft Edge Detection	Video compression framework using conditional Generative Adversarial Networks (GANs).	KTH dataset YouTube Pose datasets

		Soft Edge Detection using additional encoder.	
Zhibo Chen et al. (2018)	Learning for Video Compression	Autoencoder Pixel Motion CNN Predictive coding Iterative analysis/synthesis Binarization	Flickr-530,000 color images UCF-101 - 10,000 samples
Chao-Yuan Wu et al. (2018)	Video Compression through Image Interpolation	Image Interpolation Vanilla U-net interpolation architecture End to end trained	UCF-101 HMDB-51 Charades dataset: Training videos-7,985 and test videos-1,863
Jun Han et al. (2019)	Deep Probabilistic Video Compression	Variational autoencoders (VAEs) Entropy coding using temporally-conditioned probabilistic model.	Sprites: Samples from video game BAIR, an action-free robot pushing dataset Kinetics600: collection of samples from YouTube videos comprising various human physical actions.
Oren Rippel et al. (2018)	Learned Video Compression	Video compression with framework for Machine Learning-based spatial rate control	Action scenes downloaded from YouTube. In HD: Xiph 1080p video dataset In SD, VGA resolution dataset from the Consumer Digital Video Library (CDVL).

Authors' Name	Performance MS-SSIM	Comparison with Conventional Codecs	Remarks
Guo Lu et al (2019).	MS-SSIM - 0.961 at 0.0529bpp	Outperforms H.264	Provide a platform for plugging of some other concepts into this framework. Uses HEVC Standard Test Sequences
Sungsoo Kim et al. (2018)	For 64x64 size MS-SSIM :0.94 For 256x256 size MS-SSIM: 0.96	Higher quality scores below 10 Kbps comparable to H.264. Above 10Kbps, H264 outperforms.	At bitrates below 7.5 Kbps, produced viable reconstructions while H.264 failed (no output).
Zhibo Chen et al. (2018)	MS-SSIM 0.98 at 155kbps	Superior performance compared with MPEG-2 and achieve comparable results with H.264 codec	Computational complexity is about 141 times that of H.264. Increment in BD Rate in comparison to H.264 codec: 8.175% Decrease in BD-PSNR in comparison to H.264 codec: 0.41dB
Chao-Yuan Wu et al. (2018)	MS-SSIM = 0.984 at 0.080 BPP	Outperforms codecs such as H.261, MPEG-4 Part 2, and performs on par with H.264.	Improvement on HMDB-51: 5.8% Improvement on UCF-101: 2.7% Network computation complexity and accuracy: 4.6x more efficient than 3D CNN.

Jun Han et al. (2019)	MS-SSIM=0.99	Outperforms H.264. Comparable visual quality on generic video content to VP9.	Extreme compression performance for small scale and specialized content videos. Undesirable for general purpose codec.
Oren Rippel et al. (2018)	For SD dataset, MS-SSIM values: 0.990-0.998. For HD dataset, MS-SSIM values: 0.980-0.994.	Upto 37.5% smaller file size on SD videos relative to HEVC/H.265, AVC/H.264 and VP9 codecs.	The current speed of encoder and decoder is too slow to be used for practical applications.

The relative performance of some of the major techniques in MS-SSIM is presented in the form of graph. With the more progress and advancements in deep learning techniques, the performance is gradually improving.



Graph 1, 2 Graphical and Histogram representation of max MS-SSIM achieved in above mentioned works

The above graphical comparison and analysis reveals that MS-SSIM of the evolving architectures approximate towards one, resulting in achieving better visual perception of the frames. These positive results of researches motivated further exploration in the application of deep learning concepts for its practical applicability. A number of advanced approaches have also been proposed to consume videos using compressed videos directly. The approaches are also shown to not only be faster than existing action recognition approaches but also gives state-of-the-art results. Consuming compressed video already removes superfluous information. Deep learning based compression may emerge as a powerful and significant domain in future.

4. CONCLUSION

This study provides an insight into the recent machine learning based techniques proposed and developed for video compression. The experimental results reveal that learning based optimization and compression techniques are giving satisfactory results and henceforth providing further directions of research in the field of video compression. Their comparative analysis throws light on the relative performance of different architectures along with the issues need to be further addressed and explored related to complexity and limitations in practical applicability. Further researches in deep learning based techniques may supersede the manual architectures in future and would also results in efficient use of decompressed video for video analytics. This study provides a premise for future research on optimizing the existing architectures for their use in real time applications.

REFERENCES

[1] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper - Cisco https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html.

[2] Chao-Yuan Wu, Nayan Singhal, Philipp Krahenbhul. Video Compression through Image Interpolation, arXiv :1804.06919v1 [cs.CV] 18 Apr 2018.

[3] Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K. Pixel recurrent neural networks. In: ICML (2016).

[4] Nick Johnston Google Inc., Damien Vincent, David Minnen, Michele Covell, Saurabh Singh *et al.* Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks. arXiv: 1703.10114v1 [cs.CV] 29 Mar 2017.

[5] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. CVPR, abs/1608.05148, 2017.

[6] Mu Li, Wangmeng Zuo, Shuhang Gu, et al.; Learning Convolutional Networks for Content-weighted Image Compression. arXiv :1703.10553v2 [cs.CV] 19 Sep 2017.

[7] J. Ball´e, V. Laparra, and E. P. Simoncelli. End to end optimized image compression. arXiv preprint arXiv:1611.01704, 2016.

[8] Tamar Rott Shaham, Tomer Michaeli, Deformation Aware Image Compression. arXiv:1804.04593v1 [cs.CV] 12 Apr 2018.

[9] Zhibo Chen, Senior Member, IEEE, Tianyu He, Xin Jin, Feng Wu, Fellow, IEEE. Learning for Video Compression. arXiv:1804.09869v2 [cs.MM] 9 Jan 2019.

[10] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. Computer Vision and Pattern Recognition (CVPR), July 2017.

[11] Thierry Dumas, Aline, Roumy and Christine Guillemot. Autoencoder based image compression. Can the learning be quantization dependent? arXiv: 1802.09371v1 [eess.IV] 23 Feb 2018.

[12] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned Video Compression. arXiv:1811.06981v1 [eess.IV] 16 Nov 2018.

[13] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai and Zhiyong Gao. DVC: An End-to-end Deep Video Compression Framework. arXiv: 1812.00101v3 [eess.IV] 7 Apr 2019.

[14] L. Theis, W. Shi, A. Cunningham, and F. Husz´ar. Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395, 2017.

[15] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar. Variable rate image compression with recurrent neural networks. arXiv preprint arXiv:1511.06085, 2015.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[17] Jun Han, Salvator Lombardo, Christopher Schroers, Stephan Mandt. Deep Probabilistic Video Compression. arXiv:1810.02845v1 [cs.CV] 5 Oct 2018.

[18] Sungsoo Kim, Jin Soo Park, Christos G. Bampis, Jaeseong Lee, Mia K. Markey, Alexandros G. Dimakis, Alan C. Bovik. Adversarial Video Compression Guided by Soft Edge Detection. arXiv:1811.10673v1 [eess.IV] 26 Nov 2018.