ENHANCING HEVC SPATIAL PREDICTION BY CONTEXT-BASED LEARNING

Li Wang[*]*, Attilio Fiandrotti*[†]*, Andrei Purica*[†]*, Giuseppe Valenzise*[*]*, Marco Cagnazzo*[†]

[*] CNRS L2S, CentraleSupelec, Université Paris-Sud
[†] LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

ABSTRACT

Deep generative models have been recently employed to compress images, image residuals or to predict image regions. Based on the observation that state-of-the-art spatial prediction is highly optimized from a rate-distortion point of view, in this work we study how learning-based approaches might be used to further *enhance* this prediction. To this end, we propose an encoder-decoder convolutional network able to reduce the energy of the residuals of HEVC intra prediction, by leveraging the available context of previously decoded neighboring blocks. The proposed context-based prediction enhancement (CBPE) scheme enables to reduce the mean square error of HEVC prediction by 25% on average, without any additional signalling cost in the bitstream.

Index Terms— Spatial prediction, HEVC Intra, deep generative models, video coding

1. INTRODUCTION

Modern image and video codecs strongly rely on spatial prediction as a fundamental tool to achieve high rate-distortion performance. Conventionally, spatial prediction leverages an ensemble of simple linear models to interpolate information from a context of already decoded pixels, with the goal to obtain a prediction residual which is simpler to code than the original signal. These prediction models have been improved and optimized for several decades, by continuously adding new modes, block partitions and prediction directions, e.g., 33 directional modes and up to 32×32 prediction units are employed in the HEVC video standard [1].

Despite the high number of available prediction modes, current spatial prediction approaches assume the underlying signal can be approximated by a simple linear combination of a few (reconstructed) pixels. Increasing further the number of prediction modes might guarantee a better signal approximation; however, this leads to continuously increase the computational cost, and is still ineffective when the signal to predict requires more complex representations than simple bilinear interpolation.

Recently, deep generative models such as auto-encoders have been employed to learn effective representations for very complex signals such as natural images [2]. Differently from the signal models used in video spatial prediction, deep generative models are much more complex and highly non-linear, and can potentially approximate any class of signal provided that enough training samples are available to learn the original data distribution. In the last couple of years, these models have been applied to image compression [3, 4, 5, 6, 7, 8, 9], producing image representations able to provide, in some cases, equivalent or better visual quality than traditional image codecs [10].

In this paper, we consider how spatial prediction in image/video coding can be improved by means of a deep generative network. Differently from previous work that concentrated on either directly optimizing end-to-end reconstructed pixels [11] or residuals [12], in this work we aim at improving the spatial prediction produced by state-of-the-art codecs such as HEVC, in order to reduce the bitrate of the prediction residual. To this end, we employ a convolutional encoder-decoder neural network to predict a block of an image based on a context of already decoded pixels *and* the spatial prediction of the same block as produced by the rate-distortion optimization of the video encoder. This is motivated by the practical observation that the spatial prediction of HEVC is indeed a good initial solution from a rate-distortion perspective. In particular, we found that the block partitioning of HEVC prediction enables to capture some general structure of the block to predict, while the deep network can later further refine this initial guess by filling in more complex details. Our encoder-decoder architecture is somewhat inspired by the context encoder proposed in [13] for the purpose of image inpainting, with the important difference that we aim at obtaining a prediction residual which is *easier to code*, rather than a reconstruction that is visually plausible. Therefore, in the following we will refer to the proposed scheme as *context-based prediction enhancement* (CBPE). Our experiments on a dataset of natural images show that the proposed CBPE reduces the mean square error of HEVC spatial prediction by 25% on average.

2. RELATED WORK

Generative models aim at learning the distribution of input data, which is intrinsically linked to finding the best source code in information theory [14]. Hence, it is not surprising

that the recent advances in deep generative models, such as auto-encoders and generative adversarial networks, have stimulated research towards applying these tools to image and video compression [11, 12, 6]. Auto-encoder architectures [2, 15], in particular, are especially effective to obtain compressed latent representations, by forcing the output to reproduce the input image through an information bottleneck whose dimensionality is much smaller than the original input space. Image compression methods based on auto-encoders have been shown to yield coding gains compared to legacy image codecs such as JPEG and JPEG 2000, and competitive results with more recent image compression algorithms such as BPG [7, 12, 10].

One of the first image compression methods based on deep learning architectures was presented by Toderici et al. [12], who proposed a set of full-resolution lossy image compression methods based on recurrent auto-encoders. At each stage of the network, a residual is computed; afterwards, the residue can be further compressed by successive iterations of the network, enabling progressive encoding and reconstruction of the input image. This architecture has been expanded later to include a spatial context predictor [5]: instead of predicting an image block directly, the authors used a modified version of the context encoder proposed by [13] to produce an initial residual. While we also employ similar tools, we found that a blank initialization of the predicted block requires a substantially larger amount of data to train, while providing HEVC prediction as input to the network enables to generalize well even on relatively small datasets.

The use of an encoder-decoder scheme to inpaint missing pixels in images was initially proposed by Pathak et al. [13]. Their context encoder is able to predict the content of an arbitrary image region conditioned on its surroundings, by minimizing a loss function composed of a fidelity (L2) term, and an adversarial term to pick the mode of the distribution which is more representative of the training data. While this is shown to produce sharper and more visually convincing results than a simple L2 loss, we found that the adversarial term tends to produce structures and details which are not present in the original frames. This is especially undesirable in predictive coding, due to the cumulative effect of the error in subsequent predicted units.

In a very recent work, Schiopu et al. [9] have presented an image prediction scheme fusing a CNN-based predictor with conventional lossless image codecs. Although a similar principle might be applied to spatial prediction in video coding, their architecture is designed to predict pixel lines and integrate a state-of-the-art lossless codec such as CALIC. In this work we focus on the lossy scenario only.

3. NETWORK ARCHITECTURE

This section describes our proposed convolutional network architecture for context-based predictor enhancement, then it details the relative training procedures.

3.1. Network Architecture

Our proposed CBPE network architecture is shown in Fig. 1 and is composed of one *encoder* connected to a *decoder* via a *bottleneck* [13]. For the sake of simplicity, in the following all operations are relative to 8-bit grayscale images. The encoder takes in input a 64×64 image $\mathcal{I}$ and projects it on a latent feature space. The encoder is composed of 4 cascaded convolutional layers with 64, 128, 256 and 512 filters respectively. Filters have size 4×4 and have a horizontal and vertical stride equal to 2 pixels, which avoids the need for pooling layers. Each convolutional layer has Rectified Linear Unit (ReLU) activation functions and we use Batch Normalization for accelerating the learning process. The output of the encoder consist of 512 feature maps of size 4×4, for a total of about 8k features. The encoder output is provided in input to a fully connected bottleneck layer composed of $B = 1000$ units which performs a reduction of the feature space size to keep the complexity low and reduce the risk of overfitting. The decoder finally takes in input the features produced by the bottleneck layer and recovers a 32×32 image. The decoder architecture mirrors that of the encoder and is composed of 4 deconvolutional layers with 512, 256, 128 and 64 filters respectively. Each deconvolutional layer can be seen as a learnable interpolating filter that doubles the resolution of the input feature maps, recovering the spatial information lost at the encoder. The output of the encoder is finally the 32×32 predicted block $\mathcal{Y}$.

3.2. Training Procedure

Our network is trained end-to-end to reconstruct an original (uncompressed) 32×32 block $\mathcal{O}$ starting from an input 64×64 image $\mathcal{I}$, as shown in Figure 1. The bottom-right quadrant of the input block is the HEVC prediction $\mathcal{P}$ obtained after rate-distortion optimization, while the remaining quadrants are the decoded (thus, noisy) causal context of the block to be predicted. The loss function is defined as the mean square error (MSE) between $\mathcal{Y}$ and $\mathcal{O}$, that is:

$$L(w, y, o) = \frac{1}{K} \sum_k (y_k(w) - o_k)^2, \tag{1}$$

where y_k and o_k are pixels of $\mathcal{Y}$ and $\mathcal{O}$, respectively, and $K = 32^2$. The network is trained to minimize the above loss function with Stochastic Gradient Descent using the ADAM optimizer [16] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate equal to $2 \cdot 10^{-4}$ and batch size of 64 images. All the network weights w are randomly initialized according to the Xavier algorithm [17] and the training procedure ends when the loss over a set of validation images ceases decreasing or a maximum number of iterations has elapsed.

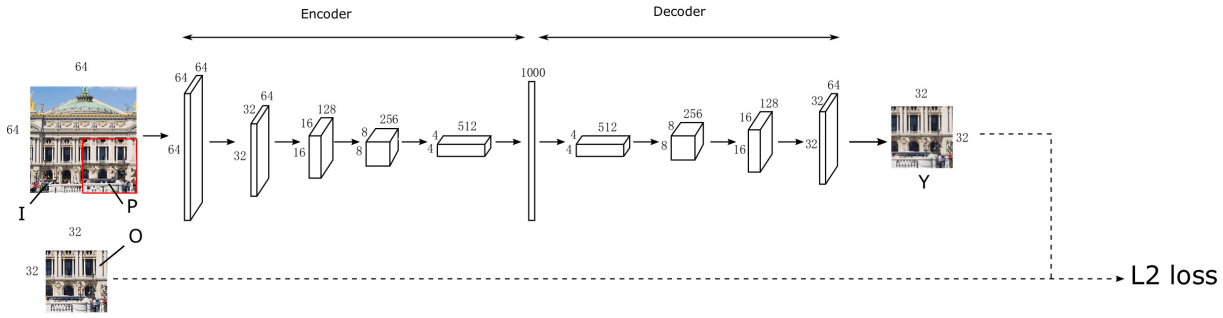

Fig. 1. The proposed encoder-decoder convolutional network architecture.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

In order to train and validate our CBPE model, we draw at random about 16k images from the dataset originally proposed in [18] for aesthetic quality evaluation.[1] The dataset contains natural images downloaded from Flickr, spanning a wide range of semantic classes and acquisition quality. The images come in a JPEG compressed format, with the original quality/resolution of the Flickr source, thus providing a large variety of train/test conditions. Each image is independently compressed and decoded using the H.265/HEVC HM reference software (version 16.0), with a fixed QP=15. All prediction unit sizes, from 4×4 to 32×32 are enabled in the rate-distortion optimization. Next, for each decoded image, we extract a number of non-overlapping 64×64 patches aligned with the HEVC CTU grid, along with the HEVC predictors $\mathcal{P}$. Since images have different spatial resolutions, for each picture we ensure that the overall number of pixels in the extracted patches does not exceed 10% of the total pixels of that image. Following this protocol, a total of 405k patches are extracted from a first set of randomly drawn images, of which 324k (80%) are used for training and 81k (20%) for validation. Finally, about 50k patches are extracted from a different set of randomly drawn images for testing.

4.2. Experimental Results

In order to test the performance of the proposed CBPE, we provide in input to our trained network the test patches and, for each test patch, we measure the MSE between the network output $\mathcal{Y}$ and the ground-truth, uncompressed reference $\mathcal{O}$, i.e., the energy of the prediction residual obtained by CBPE. For comparison, we also compute for each test patch the energy of the HEVC prediction residual, i.e., the MSE between HEVC predictor $\mathcal{P}$ and $\mathcal{O}$. Figure 2 compares the energy of residuals of the conventional HEVC predictor with that of residuals after CBPE. Among the 50954 patches, there are

[1]The dataset can be downloaded from `https://github.com/cgtuebingen/will-people-like-your-image`

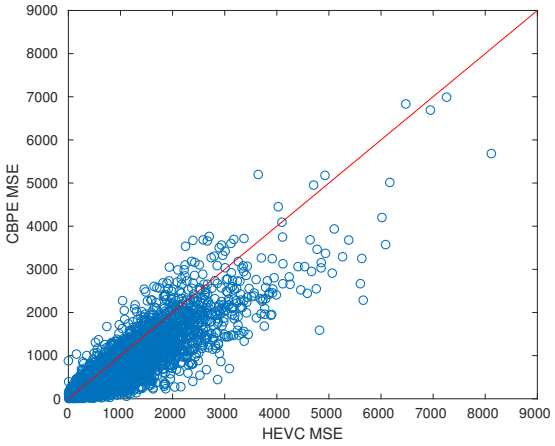

Fig. 2. Comparison of the mean squared error between the original and predicted blocks by HEVC and after the CBPE refinement.

33518 cases (approximately 66% of the cases) where CBPE enhances the HEVC predictor by reducing the energy of the prediction residuals. The MSE of CBPE reduces the average HEVC predictor MSE from 190.8 to 142.0, which corresponds to a 25% reduction of the prediction residual energy. Although quantifying precisely the end-to-end coding gain provided by CBPE would require integrating it into a whole coding chain, we notice that reducing the energy of prediction residuals is directly related to improving rate-distortion performance in predictive coding [19]. In addition, this reduction comes at *zero* additional bitrate cost: assuming the CBPE trained network is known at the decoder, no further signalling nor side information needs to be transmitted in order to reproduce CBPE at the decoder side.

In order to illustrate qualitatively the prediction improvement brought by CBPE, we report in Figure 3 a few examples of predicted blocks. From left to the right, we show the original content, the HEVC predictor, the predictor refined by CBPE, the HEVC prediction residual and the residual after CBPE, for three different patches, with the corre-

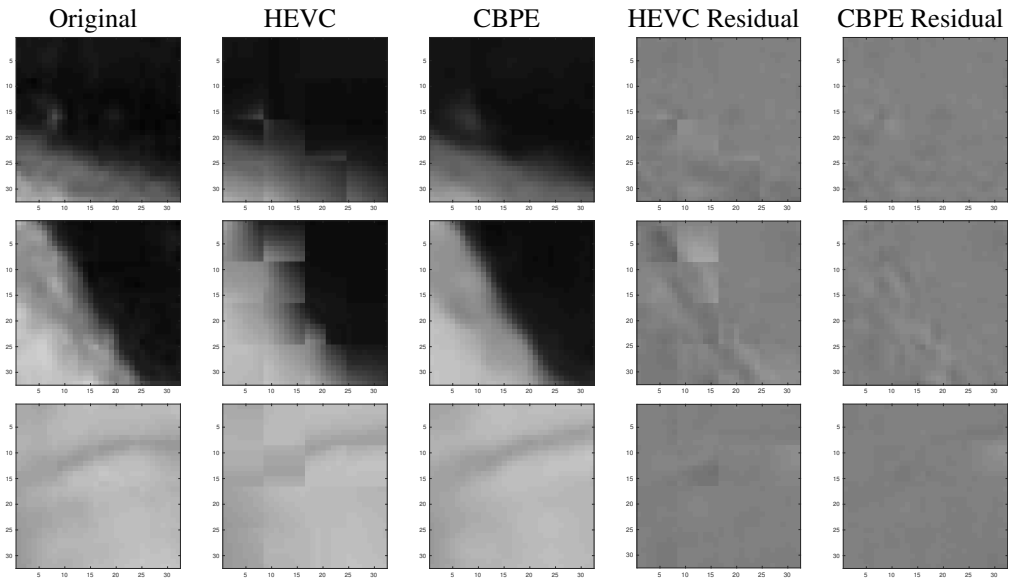

Fig. 3. Comparison of predicted patches by using HEVC prediction and proposed scheme prediction methods. The MSE of prediction blocks using HEVC and CBPE for each blocks are: top: 83.99 (HEVC), 34.99 (CBPE); middle: 321.71 (HEVC), 57.37 (CBPE); bottom: 36.63 (HEVC), 27.41 (CBPE).

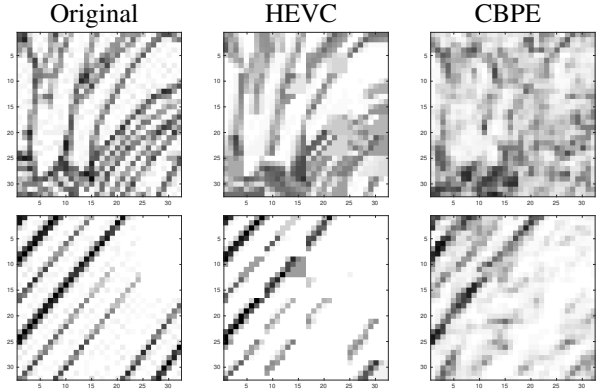

Fig. 4. Some failure cases where CBPE is not able to improve HEVC spatial prediction. The MSE of prediction blocks using HEVC and CBPE for each blocks are: top: 2.25e+03 (HEVC), 3.53e+03 (CBPE); bottom: 1.45e+03(HEVC), 2.54e+03 (CBPE).

sponding prediction MSE. We observe that, in these cases, the HEVC predictor can capture the overall structure of the block. However, due to the limited directional prediction modes and the block-based predictions, the HEVC prediction alone introduces some visible artifacts, and fails in capturing fine-grained structures of the content. Conversely, the CBPE can enhance this prediction, smoothing out the HEVC blocking and recovering somehow better the original image structure. Interestingly, the CBPE predictor has a more natural aspect, confirming previous findings on the ability of deep generative models to learn image "naturalness" [10].

It is also instructive to analyze cases where CBPE fails and degrades the quality the HEVC predictor. Figure 4 shows two of the worst cases from the scatter plot in Figure 2. We observe that CBPE tends to over-smooth patches with periodic or high-frequency structures and sharp edges, and in some cases to add some low-frequency noise which was not present in the original signal. This might be caused by the lack of sufficient training data. An interesting solution to explore would be to add a regularization term in the loss function (1) in order to preserve sharper structures and penalize noise, similar to what is done in total-variation denoising.

5. CONCLUSIONS

In this paper, we proposed a context based prediction enhancement (CBPE) model to reduce the energy of HEVC prediction residuals and thus improve coding performance. Differently from recent learning-based image coding schemes, our approach is applied on top of HEVC rate-distortion optimization, leading to an average reduction of 25% of the prediction residual energy without any additional signalling cost in the bitstream. To the authors' knowledge, this is the first work to employ deep generative models to the enhancement of spatial prediction in a video coding scenario. As that, there are many challenging questions to answer, e.g., how CBPE does perform when the context quantization is more severe – our initial experiments show that this requires increasing the number of training samples; and to which extent context-based prediction might be used to partially or completely replace the HEVC Intra prediction modes.

6. REFERENCES

[1] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan, "High efficiency video coding (HEVC)," *Integrated Circuit and Systems, Algorithms and Architectures. Springer*, vol. 39, pp. 40, 2014.

[2] Ian Goodfellow, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[3] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy image compression with compressive autoencoders," in *Int. Conf. on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[4] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.

[5] David Minnen, George Toderici, Michele Covell, Troy Chinen, Nick Johnston, Joel Shor, Sung Jin Hwang, Damien Vincent, and Saurabh Singh, "Spatially adaptive image compression using a tiled deep network," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 2796–2800.

[6] Oren Rippel and Lubomir Bourdev, "Real-time adaptive image compression," *arXiv preprint arXiv:1705.05823*, 2017.

[7] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *Int. Conf. on Learning Representations (ICLR)*, Vancouver, CA, May 2018.

[8] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," *arXiv preprint arXiv:1804.02958*, 2018.

[9] Ionut Schiopu, Yu Liu, and Adrian Munteanu, "CNN-based prediction for lossless coding of photographic images," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 16–20.

[10] Giuseppe Valenzise, Andrei Purica, Vedad Hulusic, and Marco Cagnazzo, "Quality assessment of deep-learning-based image compression," in *Multimedia Signal Processing*, Vancouver, Canada, Aug. 2018.

[11] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," in *Int. Conf. on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[12] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell, "Full resolution image compression with recurrent neural networks," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, July 2017, pp. 5435–5443.

[13] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[14] Antti Honkela and Harri Valpola, "Variational learning and bits-back coding: an information-theoretic view to bayesian learning," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 800–810, 2004.

[15] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *Int. Conf. on Learning Representations (ICLR)*, Banff, CA, Apr. 2014.

[16] Diederik P Kingma and Jimmy Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[18] Katharina Schwarz, Patrick Wieschollek, and Hendrik PA Lensch, "Will people like your image?," *arXiv preprint arXiv:1611.05203*, 2016.

[19] Nuggehally S Jayant and Peter Noll, "Digital coding of waveforms: principles and applications to speech and video," *Englewood Cliffs, NJ*, pp. 115–251, 1984.