N-O Cool-chic: reconcile fast encoding with lightweight decoding for neural image compression

Théophile Blard, Théo Ladune, Pierrick Philippe
Orange Innovation, France
firstname.lastname@orange.com

Xiaoran Jiang, Olivier Déforges
IETR, France
firstname.lastname@insa-rennes.fr

Abstract

Overfitted image codecs achieve strong compression performance and low decoder complexity by learning a lightweight decoder for each image. Such codecs include Cool-chic, which presents image coding performance on par with VVC while requiring around 2000 multiplications per decoded pixel. However, the encoding time associated with overfitted codecs may be prohibitively long for real-time applications, posing a challenge to their practical implementation in such scenarios. To address this issue, this paper proposes to decrease the encoding complexity of Cool-chic by bypassing the overfitting procedure and complementing the decoder with an encoder network. The proposed non-overfitted (N-O) Cool-chic, significantly reduces encoding complexity by a factor of 1000 compared to Cool-chic, while maintaining competitive performance.

Index terms

Neural image compression, low-complexity, overfitting

1 Introduction

Autoencoder-based codecs (ELIC [1], MLIC++ [2]) offer state-of-the-art compression results, outperforming conventional codecs (H.265/HEVC [3], H.266/VVC [4]). During training, autoencoder parameters are optimized following the rate-distortion cost computed on a large dataset of images. Once the training stage is completed, parameters are frozen and the autoencoder relies on *generalization* to compress unseen images. Generalization requires networks with many parameters, making the decoding particularly complex. Indeed, autoencoder-based codecs have millions of parameters and require up to a million multiplications to decode a single pixel. This decoding complexity might hinder their adoption especially when decoding happens on low-power devices such as smartphones.

Several studies have aimed to address the complexity constraint associated with autoencoder-based codecs. For instance, Johnston et al. [5] achieved a 50% reduction in the complexity of Ballé [6] through weight pruning applied to the decoder, without significant performance degradation. EVC [7] leveraged network distillation to reduce the complexity by a factor of 10 while maintaining performance comparable to VVC. More recently, Yang et al. [8] proposed the use of asymmetric architectures with shallow decoders, involving only 20 000 multiplications per decoded pixel, while still maintaining performance close to HEVC. However, it is important to note that these approaches still exhibit significantly higher complexity compared to conventional methods.

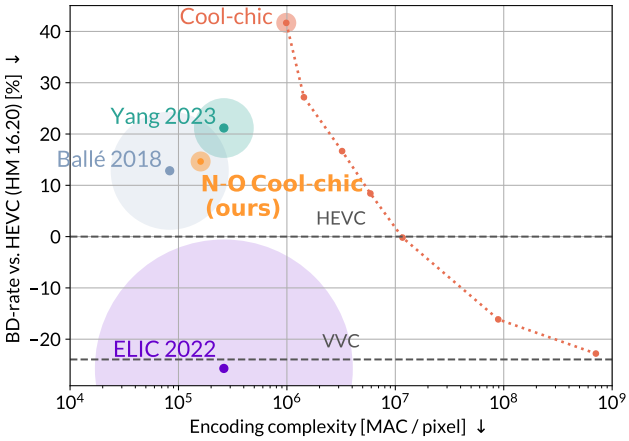

Figure 1: *Rate-distorsion performance as a function of the encoding complexity on CLIC 2020 validation set [9]. Negative results: less rate is required to get the same quality than HEVC. Cool-chic encoding complexity is varied by adjusting the training time. The circle radius denotes the decoding complexity (see Table 1).*

Overfitted codecs (Cool-chic [10, 11, 12], C3 [13]) have emerged as an alternative paradigm to autoencoders. To compress an image, overfitted codecs learn (overfit) a lightweight neural decoder and latent representation tailored for this image. The decoder parameters are then conveyed alongside the latents so that the receiver can reconstruct the image. Since overfitted codecs do not rely on generalization, their decoder are significantly lighter than autoencoders. As such, they offer compelling image coding performance on par with VVC with a decoder complexity of 2300 multiplications per pixel [12]. As for conventional codecs, this is obtained through an expensive encoding process, during which the decoder and latents are optimized according to the image rate-distortion cost.

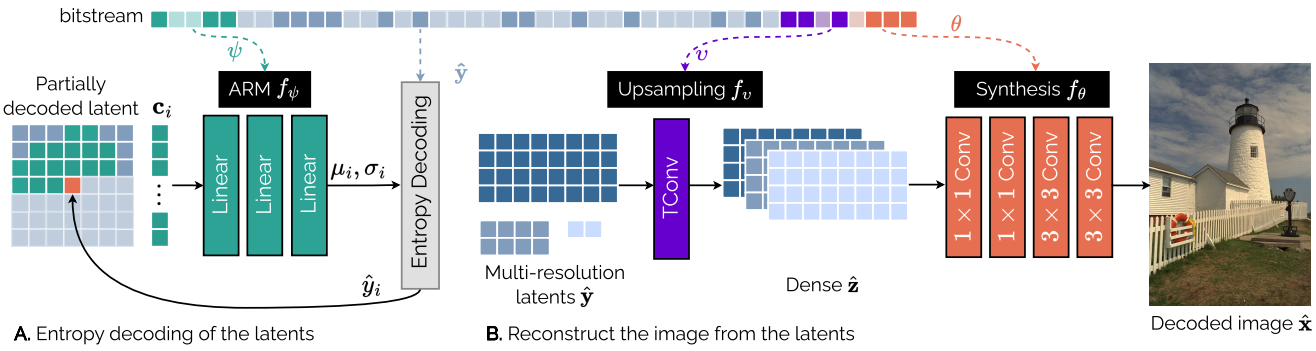

Figure 2: *Cool-chic decoding. ARM: Auto-Regressive Model.*

This paper aims to bridge the gap between autoencoder-based codecs and overfitted codecs by combining the existing Cool-chic lightweight decoder with an encoder which generates the latent representation in a single forward pass. The proposed network forms an autoencoder that offers the benefits of both low encoding and decoding complexity.

2 Background: Cool-chic

This section presents the Cool-chic encoding and decoding process.

Decoding. Figure 2 presents the decoding process of Cool-chic. It is composed of three main elements. *i)* L latent grids $\hat{\mathbf{y}}_l$ with different resolutions: $\hat{\mathbf{y}} = \{\hat{\mathbf{y}}_l \in \mathbb{Z}^{H/2^l \times W/2^l},\ l = 0, \dots, L-1\}$. *ii)* The latent grids are transmitted using an entropy coding algorithm, driven by an auto-regressive probability model $p_{\boldsymbol{\psi}}$ (ARM). This ARM models the distribution of one latent value conditioned on neighbouring values and is implemented as a MLP $f_{\boldsymbol{\psi}}$. *iii)* The upsampling $f_{\boldsymbol{\upsilon}}$ and synthesis $f_{\boldsymbol{\theta}}$ networks upsample the latents to a dense representation and synthesize the decoded image $\hat{\mathbf{x}}$.

Encoding. Cool-chic encodes an image $\mathbf{x}$ by simultaneously learning the latent and the different neural networks according to the image rate-distortion cost:

$$\begin{aligned} \hat{\mathbf{y}}^*, \boldsymbol{\psi}^*, \boldsymbol{\upsilon}^*, \boldsymbol{\theta}^* &= \underset{\hat{\mathbf{y}}, \boldsymbol{\psi}, \boldsymbol{\upsilon}, \boldsymbol{\theta}}{\operatorname{argmin}}\, \mathrm{D}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \mathrm{R}(\hat{\mathbf{y}}) \qquad (1) \\ &= \underset{\hat{\mathbf{y}}, \boldsymbol{\psi}, \boldsymbol{\upsilon}, \boldsymbol{\theta}}{\operatorname{argmin}}\, ||\mathbf{x} - f_{\boldsymbol{\theta}}(f_{\boldsymbol{\upsilon}}(\hat{\mathbf{y}}))||^2 - \lambda \log_2 p_{\boldsymbol{\psi}}(\hat{\mathbf{y}}). \end{aligned}$$

The Lagrange multiplier $\lambda \in \mathbb{R}$ balances the rate R and the distortion D, here the mean-squared error. The discrete latents $\hat{\mathbf{y}} = \mathrm{Q}(\mathbf{y})$ are optimized through the continuous version $\mathbf{y}$, using the method proposed in C3 [13] to obtain a differentiable proxy for the quantization Q. After the encoding, neural network parameters are quantized and entropy coded with an Exp-Golomb code since they usually represent less than 5 % of the total rate. The latents grids $\hat{\mathbf{y}}$ are entropy coded with a range coder driven by the probability model $p_{\boldsymbol{\psi}}$.

3 Non-overfitted Cool-chic

Cool-chic encoding complexity can be mitigated at the expense of the compression performance by reducing the training (i.e. encoding) time. While convenient, this is not enough for use-cases where *real-time* encoding is required. This section proposes a solution for situations where the encoding complexity constraint is paramount: a non-overfitted (N-O) Cool-chic.

Analysis transform. In order to reduce the encoding time, the overfitting process is bypassed by training a non-overfitted (N-O) Cool-chic, sharing identical parameters for all images. Taking inspiration from autoencoders, encoding an image with N-O Cool-chic relies on an additional analysis transform $f_{\boldsymbol{\alpha}}$ generating a Cool-chic-compatible latent representation $\hat{\mathbf{y}}$ from the input image $\mathbf{x}$:

$$\hat{\mathbf{y}} = \mathrm{Q}(f_{\boldsymbol{\alpha}}(\mathbf{x})). \qquad (2)$$

As a result, the iterative optimization of $\hat{\mathbf{y}}$ with gradient descent is replaced with a single forward pass in the analysis network.

Architecture overview. The proposed analysis transform is depicted in Fig 3. To produce a set of $L = 7$ hierarchical latent grids, a series of L residual blocks progressively downsample the input image $\mathbf{x}$ and extract relevant features for the different resolutions. After each downsampling step, a 1×1 convolution merges the $C = 64$ features into the l-th latent $\hat{\mathbf{y}}_l$. The proposed analysis hence produces a few latent grids with hierarchical resolutions, unlike autoencoder analysis which computes hundreds of small-resolution features.

Residual blocks. To increase the receptive field while maintaining low complexity, ConvNeXt blocks [14] are adopted. Downsampling with residual block is achieved by complementing the identity branch with a stride-2 2×2 average pooling and a 1×1 convolution as in [15]. The first residual block does not downsample so the pooling layer is removed. Following ELIC [1], additional residual blocks are stacked after each downsampling for more expressivity.

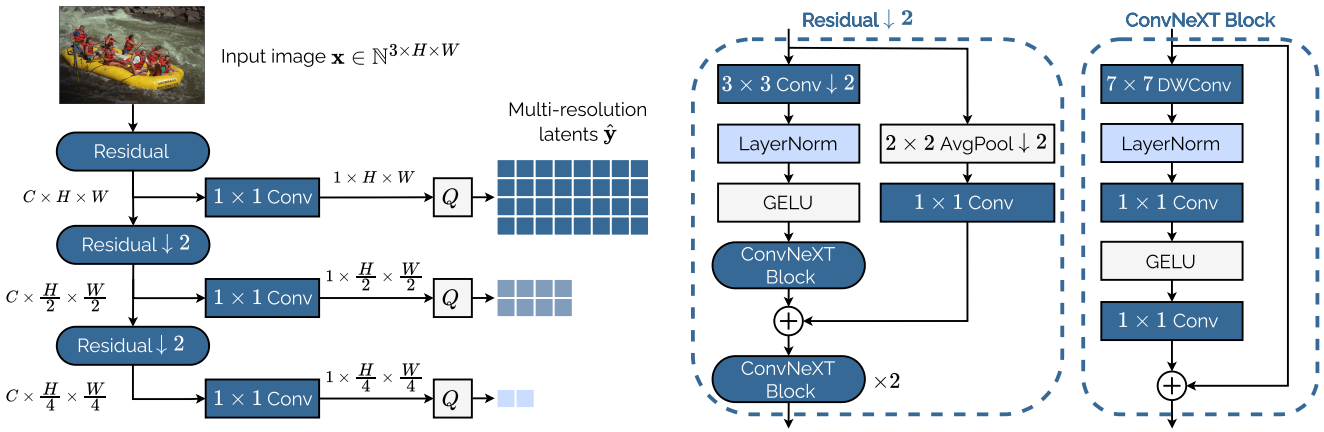

Figure 3: *Proposed analysis transform. DWConv stands for depth-wise convolution and ↓ denotes the stride.*

Decoder. The decoder parameters $(\boldsymbol{\psi}, \boldsymbol{\upsilon}, \boldsymbol{\theta})$ are learned alongside the analysis transform $f_{\boldsymbol{\alpha}}$. The architecture from [12] with a complexity of 2300 MAC (multiplication-accumulation) per decoded pixel is selected. It is worth noting that the network parameters no longer need to be transmitted alongside the latent representation since they are shared for all images.

4 Experiments

Training. The proposed N-O Cool-chic aims to obtain optimal parameters, generalizable to all possible images. To this end, it is trained with 256×256 patches of randomly cropped images from the CLIC 2019 training set [9]. Since the global rate-distorsion cost must be optimized, equation (1) becomes:

$$\begin{aligned}\boldsymbol{\alpha}^*, \boldsymbol{\psi}^*, \boldsymbol{\upsilon}^*, \boldsymbol{\theta}^* &= \underset{\boldsymbol{\alpha},\boldsymbol{\psi},\boldsymbol{\upsilon},\boldsymbol{\theta}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{x}}[\mathrm{D}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \mathrm{R}(\hat{\mathbf{y}})] \qquad (3)\\ &= \underset{\boldsymbol{\alpha},\boldsymbol{\psi},\boldsymbol{\upsilon},\boldsymbol{\theta}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{x}}[||\mathbf{x} - f_{\boldsymbol{\theta}}(f_{\boldsymbol{\upsilon}}(\hat{\mathbf{y}}))||^2 - \lambda \log_2 p_{\boldsymbol{\psi}}(\hat{\mathbf{y}})].\end{aligned}$$

Independent training with different rate constraints is performed to cover a wide range of rates, namely $\lambda = \{0.02, 0.004, 0.001, 0.0004, 0.0001\}$. Adam algorithm [16] is used, the learning rate starts from 10^{-3} and is dynamically decayed by a patience mechanism. When it reaches 10^{-6}, the training is terminated.

Encoding complexity. Figure 1 shows the compression performance of N-O Cool-chic against the encoding complexity. N-O Cool-chic encodes images in a single forward pass through the analysis. This reduces the complexity by a factor of 1000 compared to Cool-chic, requiring only 160 kMAC per pixel which is comparable to other autoencoder-based encoders. While reducing the training time can help decrease the encoding complexity of Cool-chic, it still requires 20 times more MACs compared to N-O Cool-chic to reach the same level of performance. Beside the reduction in MAC, encoding images with N-O Cool-chic is conceptually simpler since there is no more optimization through gradient descent. However, this reduces the performance significantly, falling behind state-of-the art codecs such as ELIC [1].

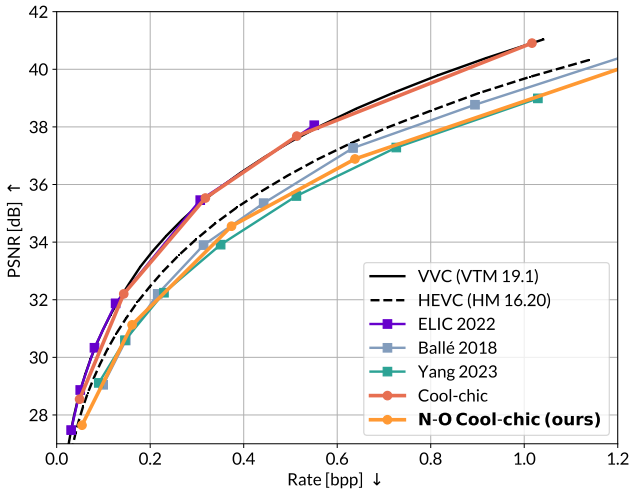

Figure 4: *Rate-distortion performance on the CLIC 2020 validation set [9].*

Rate-distorsion results. Figure 4 presents the rate-distortion curves. Compared to Cool-chic, N-O Cool-chic requires a 45% higher rate to achieve the same quality. Compared to Ballé [6], this gap is reduced to only 1%, with similar encoding complexity but with a decoder that is 30 times less complex (Table 1). On the other hand, N-O Cool-chic surpasses models that are specifically optimized for low complexity decoding, such as the 2-layer synthesis model proposed in [8]. It achieves the same quality with a 6% lower rate, while also having a decoder that is 8 times less complex.

Method	Complexity (kMAC / pixel)		BD-Rate vs HEVC (%) ↓
	Encoder	Decoder	
ELIC 2022 [1]	262.2	382.0	-25.7
Ballé 2018 [6]	80.0	83.0	12.9
Yang 2023 [8]	262.2	20.5	21.2
N-O Cool-chic (ours)	160.0	**2.3**	14.7

Table 1: *Encoder and decoder complexity v.s. average BD-rate relative to HEVC on CLIC 2020 validation set [9].*

5 Conclusion

This paper strives to make Cool-chic encoding faster. To this end N-O Cool-chic is proposed, where the encoding becomes a simple forward pass (less than 1 second), reducing the encoding complexity up to a factor of 1000. It is shown that although the decoding complexity is only 2300 multiplications per pixel, N-O Cool-chic achieves comparable performance to recent neural image codecs. However, it falls short of the performance achieved by the overfitted Cool-chic, underscoring the crucial role of overfitting in attaining optimal performance and adaptation.

References

[1] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[2] Wei Jiang and Ronggang Wang. MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*.

[3] Gary J. Sullivan et al. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012.

[4] B. Bross et al. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[5] Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé. Computationally efficient neural image compression. *arXiv preprint arXiv:1912.08771*, 2019.

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

[7] Wang Guo-Hua, Jiahao Li, Bin Li, and Yan Lu. EVC: Towards real-time neural image compression with mask decay. In *The Eleventh International Conference on Learning Representations*, 2023.

[8] Yibo Yang and Stephan Mandt. Computationally-efficient neural image compression with shallow decoders. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 530–540. IEEE.

[9] CLIC20. Challenge on learned image coding 2020. *http://clic.compression.cc/2021/tasks/index.html*, 2020.

[10] Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay. COOL-CHIC: coordinate-based low complexity hierarchical image codec. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*.

[11] Thomas Leguay, Théo Ladune, Pierrick Philippe, Gordon Clare, Félix Henry, and Olivier Déforges. Low-complexity overfitted neural image codec. In *25th IEEE International Workshop on Multimedia Signal Processing, MMSP 2023*.

[12] Théophile Blard, Théo Ladune, Pierrick Philippe, Gordon Clare, Xiaoran Jiang, and Olivier Déforges. Overfitted image coding at reduced complexity. *arXiv preprint arXiv:2403.11651*, 2024.

[13] Hyunjik Kim, Matthias Bauer, Lucas Theis, Jonathan Richard Schwarz, and Emilien Dupont. C3: high-performance and low-complexity neural compression from a single image or video. *CoRR*, 2023.

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[15] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019.

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