Transformer-Based Image Compression Without Positional Encoding

Bouzid Arezki, Fangchen Feng, Anissa Mokraoui
L2TI, Université Sorbonne Paris Nord
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
{bouzid.arezki, fangchen.feng, anissa.mokraoui}@univ-paris13.fr

Abstract

In this paper, we address the image compression problem and introduce the Swin Non-Positional Encoding (SwinNPE) transformer. SwinNPE improves the efficiency of the SwinT transformer while reducing the number of model parameters. We generalize the Swin cell and propose the Swin convolutional block, which can better handle the local correlation between image patches. Additionally, the Swin convolutional block can capture the local context between tokens without relying on positional encoding, reducing the model complexity. Preliminary results show that SwinNPE outperforms state-of-the-art CNN-based architectures in terms of the trade-off between bit-rate and distortion, achieving results comparable to SwinT with 16% less computational complexity on the Kodak dataset.

Key Words

Image Processing, Image Compression, Transformer, Transform Coding, Attention Mechanism.

1 Introduction

Transform coding is a widely used approach for image compression and forms the basis for many popular coding standards, such as JPEG. Codecs based on transform coding typically comprise three components for lossy compression : transform, quantization, and entropy coding. These components have all been improved using deep neural networks through end-to-end training, as demonstrated by various works [1, 2, 3, 4, 5, 6].

As one of the first works, the authors of [1] proposed a CNN-based two-level hierarchical variational autoencoder with hyper-prior as the entropy model. This architecture consists of two pairs of encoders/decoders, one for the generative model and another for the hyper-prior model.

Recently, transformers [7] have had great success in the computer vision area including neural image compression. The authors of [8] incorporated the attention mechanism into the image compression framework by introducing self-attention in the hyper-prior model. The more sophisticated Swin block [9] is also used in [10] in both the generative and the hyper-prior model to adopt shift window-based attention to restrict the attention in local windows. Indeed, with the attention mechanism that can better handle global context compared to convolutional neural networks, transformers have the ability to adapt the receptive field depending on the task conversely to CNNs where the kernel size is fixed. This better understanding of global information allows for capturing long-range dependencies in image compression applications.

Positional encoding is a vital component of transformers. The original ViT transformer [7] breaks down images into non-overlapping series of patches mapping each patch to a token. The standard transformer layers are then used to read the whole sequence of tokens at once. The positional encoding, therefore, plays a crucial role to maintain the sequence order, and different variations are proposed for better modeling the positional information of the sequence and maintaining the local context [11, 12, 13]. In the context of image compression, the benefits of positional encoding have been demonstrated in terms of Rate-Distortion (RD) performance in works such as [8, 10]. In particular, the authors of [8] have shown that a 2D diamond-shaped relative position encoding is useful and has particular advantages.

Despite its many advantages, using positional encoding in transformers can increase the dimensionality of embeddings, leading to higher computational costs during training and limiting the flexibility of the models. Recently, the authors of [14] demonstrated that the positional encoding can be abandoned in the attention module for image classification without any drop in performance. This was achieved by introducing convolution in the tokenization process of patches and in the self-attention block to maintain local spatial information. It is claimed that this combination of convolution and the attention mechanism benefits from both the advantages of convolutional neural networks and transformers.

In this paper, we present a new image compression framework called SwinNPE. It is based on our proposed *convolutional Swin block* which combines patch convolution and shift window-based attention in Swin without positional encoding. We believe that this framework can better capture spatial contextual information. Our preliminary experiments show that SwinNPE achieves comparable results to the SwinT architecture [10], without the need for positional encoding and with fewer parameters.

2 Proposed framework

The proposed SwinNPE uses the same architecture as in [10], which is shown in Figure 1. Specifically, the in-

put image x is first encoded by the generative encoder $y = g_a(x)$, and the hyper-latent $z = h_a(y)$ is obtained. The quantized version of the hyper-latent $\hat{z}$ is modeled and entropy-coded with a learned factorized prior to passe through $h_s(\hat{z})$ to obtain μ and σ which are the parameters of a factorized Gaussian distribution $P(y|\hat{z}) = \mathcal{N}(\mu, diag(\sigma))$ to model y. The quantized latent $\hat{y} = Q(y - \mu) + \mu$ is finally entropy-coded and sent to $\hat{x} = g_a(\hat{y})$ to reconstruct the image $\hat{x}$. We use the classical strategy of adding uniform noise to simulate the quantization operation which makes the operation differentiable. The channel-wise autoregressive block [2, 3] is designed to learn the auto-regressive prior which factorizes the distribution of the latent as a product of conditional distributions incorporating prediction from the causal context of the latents [4, 5, 6].

The generative and the hyper-prior encoder, g_a and h_a, are built with the patch merge block and the convolutional Swin block. The patch merge block contains the *Depth-to-Space* operation [10] for down-sampling, a normalization layer, and a linear layer to project the input to a certain depth C_i. In g_a, the depth C_i of the latent representation increases as the network gets deeper which allows for getting a more abstract representation of the image. The size of the latent representation decreases accordingly. In each stage, we down-sample the input feature by a factor of 2.

The proposed *convolutional Swin block* is a generalization of the Swin cell [9]. As shown in Figure 2, we use convolutions instead of position-wise linear projections to project the K, Q, and V matrices in the multi-head attention block. This makes the attention module more sensitive to spatial context. Instead of using hand-crafted positional encoding, we let the convolution layer capture the positional information. In this paper, we use depth-wise separable convolution [14] due to its parameter efficiency. More specifically, the depth-wise separable convolution first applies a 2D convolution in each feature channel independently. The outcome is then concatenated and passed through another convolution layer, such convolution reduces the number of parameters and computation while increasing representational efficiency where it deals not just with spatial dimension but with depth dimension already. It's important to note that the proposed block is not limited to convolution operations. Different forms of convolution [15, 16] are possible, making the proposed convolutional Swin block particularly flexible. Compared to the convolutional attention block in [14], we keep the shift window structure which allows cross-window connections.

The generative and the hyper-prior decoder, g_s and h_s, are built with the patch split block and the convolutional Swin block. In the patch split block, we reverse the merging sequence and use *Space-to-Depth* operation [10] for up-sampling.

3 Experiment and Analysis

3.1 Experiment configuration

This section presents an assessment of the SwinNPE architecture and a comparison of its image compression results against state-of-the-art approaches. The SwinNPE was trained on the CLIC2020 training set for 3.3 million steps. During training, each batch consisted of eight randomly cropped images with a size of 256×256 pixels.

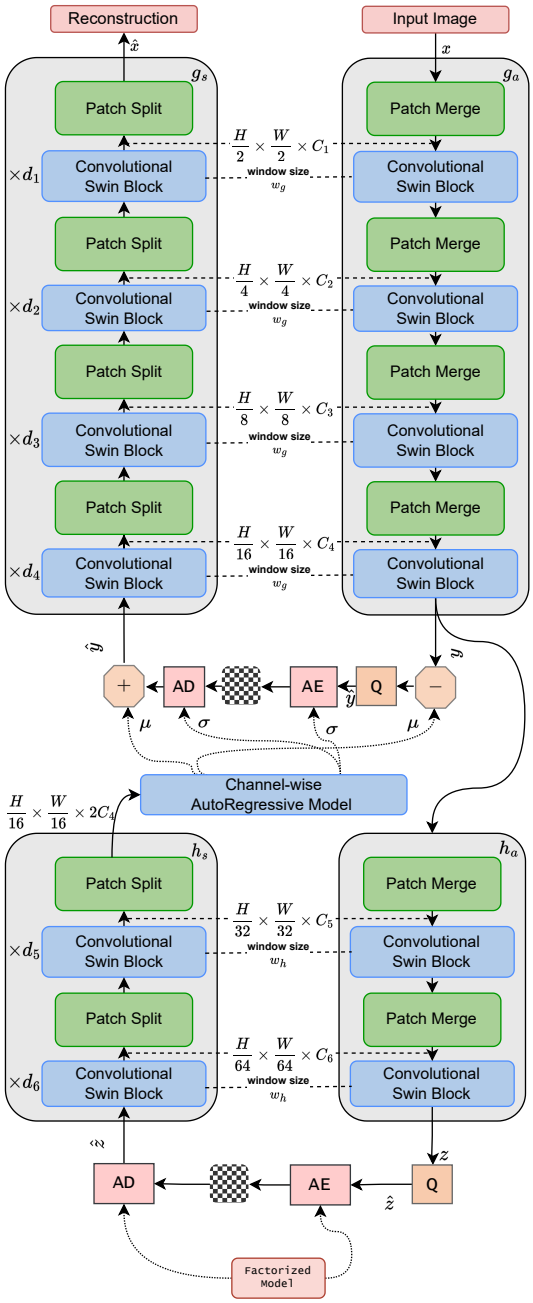

FIGURE 1 – *Network architecture of our proposed SwinNPE.*

The SwinNPE's performance was evaluated on the Kodak [17] dataset and we center-cropped all images to mul-

tiples of 256 to avoid padding. We choose the following loss function to optimize the trade-off between the bit-rate R and the quality of reconstruction D which corresponds to the Mean Squared Error (MSE) in RGB color space :

$$L = D + \beta R, \tag{1}$$

with $\beta \in \{0.003, 0.001, 0.0003, 0.0001\}$.
The learning rate starts at 10^{-4} and the hyper-parameters of the architecture shown in Figure 1 are as follows. $(d_1, d_2, d_3, d_4, d_5, d_6) = (2, 2, 6, 2, 5, 1)$, $(w_g, w_g) = (8, 8)$, $(w_h, w_h) = (4, 4)$, and $(C_1, C_2, C_3, C_4, C_5, C_6) = (128, 192, 256, 320, 192, 192)$.
For the autoregressive model, we use the model proposed in [6] with 10 slices. The kernel size in all convolutional Swin blocks for depth-wise separable convolution is set to 3.

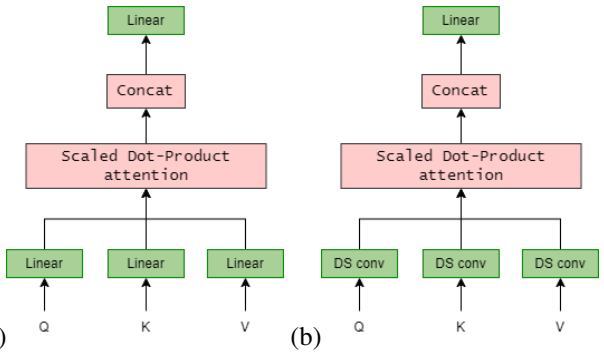

FIGURE 2 – *(a) The attention mechanism scheme for multi-head attention (b) The attention mechanism scheme for convolutional Swin. DS conv means depthwise separable convolution.*

3.2 Analysis

We compare our proposed SwinNPE with the results of two transformers-based architectures [10, 8] and some of the most used CNN-based image compression architectures and standard codecs on the Kodak dataset [17]. The rate-distortion curves of different methods are shown in Figure 3. We summarize the number of parameters of the tested transformer-based architectures in Table 1 where we also illustrate the Bijonteguard metric [18] using the SwinT-CHARM as the reference.
From Figure 3, we can clearly see that the SwinNPE outperforms all of the tested CNN-based architectures in terms of the bit-rate/distortion tradeoff. It is particularly interesting to notice that our proposed approach obtains almost the same results as Entroformer [8] (orange dashed line in Figure 3) with much less model parameters (see Table 1). Specifically, the saving bit-rate of SwinNPE is 5.46% less than SwinT-CHARM (optimal saving bit-rate) which is at the same level as Entroformer with 4.33% more bit-rate saving compare to SwinT-CHARM. We argue that it is due to the fact that the convolutional layer in the proposed convolutional Swin block can capture the local contextual information. With fewer parameters, the proposed SwinNPE has results comparable to SwinT-CHARM. We emphasize that our proposed architecture is particularly advantageous compared to SwinT-based architecture without positional encoding [1] validating the advantages of combining convolutions and transformers for image compression.

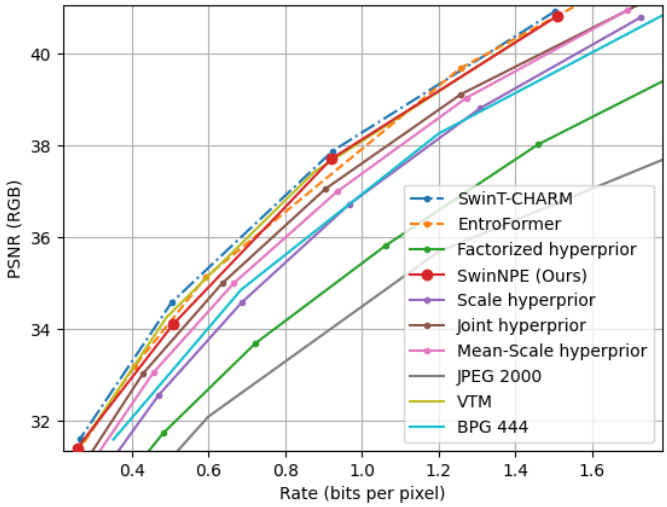

FIGURE 3 – *SwinNPE achieves nearly the same results as Entroformer [8] and SwinT-CHARM [10] that relying on Positional encoding and better RD performance than CNNs-based methods Factorized [19], Scale [1], Mean-Scale [4], Joint hyperprior [4] and standard codecs on the Kodak image set.*

4 Conclusion

In this paper, we propose SwinNPE, a transformer-based image compression model built with convolutional Swin blocks without positional encoding. SwinNPE achieves comparable results to state-of-the-art methods while using fewer model parameters and outperforming CNN-based architectures. The proposed convolutional Swin block allows for better exploitation of spatial context without the need for positional encoding, resulting in greater flexibility and fewer parameters.
For future work, it would be interesting to explore the use of different convolution operations and sizes in the proposed SwinNPE model. This could allow for more accurate modeling of complex spatial relationships and patterns, leading to improved performance in image compression. Additionally, incorporating the convolution operation into the patch merge/split module could benefit from the advantages of CNN. The proposed SwinNPE model with convolutional Swin blocks provides a promising direction for the development of efficient and effective transformer-based models for image compression.

1. The results are shown in the ablation studies in [10].

Network	#Param. (M)	Positional encoding	Bijonteguard Metric	
			$\Delta PSNR$	$\%\ \Delta rate$
SwinT-CHARM [10]	32	Positional Relative Encoding 2D	0	0%
Entroformer [8]	142.7	Positional Relative Encoding 2D + Diamond	-0.228	4.33%
SwinNPE (Ours)	**27**	-	-0.311	5.46%

TABLE 1 – *Performance comparison using Bijonteguard metric [18] where ΔPSNR measures the average PSNR difference and % Δ rate the average rate saving in percent between SwinT-CHARM [10] (selected as the reference network) and another given network .*

Références

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, et Nick Johnston. Variational image compression with a scale hyperprior. Dans *6th Inter. Conf. on Learning Representations (ICLR)*, 2018.

[2] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, et David Zhang. Learning convolutional networks for content-weighted image compression. Dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018.

[3] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, et Luc Van Gool. Conditional probability models for deep image compression. Dans *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 4394–4402, 2018.

[4] David Minnen, Johannes Ballé, et George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. Dans *Advances in Neural Information Processing Systems*, 2018.

[5] Jooyoung Lee, Seunghyun Cho, et Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. Dans *International Conference on Learning Representations*, 2019.

[6] David Minnen et Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. Dans *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, et Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. Dans *Inter. Conf. on Learning Representations*, 2021.

[8] Yichen Qian, Ming Lin, Xiuyu Sun, Tan Zhiyu, et Rong Jin. Entroformer : A transformer-based entropy model for learned image compression. Inter. Conf. on Learning Representations (ICLR), 02 2022.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, et B. Guo. Swin transformer : Hierarchical vision transformer using shifted windows. Dans *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[10] Yinhao Zhu, Yang Yang, et Taco Cohen. Transformer-based transform coding. Dans *Inter. Conf. on Learning Representations*, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, et Illia Polosukhin. Attention is all you need. Dans I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett, éditeurs, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Peter Shaw, Jakob Uszkoreit, et Ashish Vaswani. Self-attention with relative position representations. Dans *Proceedings of the 2018 Confe. of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics, Juin 2018.

[13] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, et Chunhua Shen. Conditional positional encodings for vision transformers. Dans *The Eleventh Inter. Conf. on Learning Representations*, 2023.

[14] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, et Lei Zhang. Cvt : Introducing convolutions to vision transformers. Dans *Proceedings of the IEEE/CVF Inter. Conf. on Computer Vision (ICCV)*, pages 22–31, October 2021.

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, et Yichen Wei. Deformable convolutional networks. Dans *Proceedings of the IEEE inter. conf. on computer vision*, pages 764–773, 2017.

[16] Lu Chi, Borui Jiang, et Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33 :4479–4488, 2020.

[17] Kodak. Kodak test images. `http://r0k.us/graphics/kodak/`, 1999.

[18] Gisle Bjøntegaard. Calculation of average psnr differences between rd-curves. 2001.

[19] Johannes Ballé, Valero Laparra, et Eero P Simoncelli. End-to-end optimized image compression. *5th Inter. Conf. on Learning Representations (ICLR)*, 2017.