Contextformer: A Transformer with Spatio-Channel Attention for Context Modeling in Learned Image Compression

A. Burakhan Koyuncu[1,2], Han Gao[4], Atanas Boev[2], Georgii Gaikov[3], Elena Alshina[2], and Eckehard Steinbach[1]
[1]Technical University of Munich [2]Huawei Munich Research Center [3]Huawei Moscow Research Center [4]Tencent America

ECCV TEL AVIV 2022

Motivation

In learned image compression frameworks, the context modeling is the one of the key components for a high-performance compression.

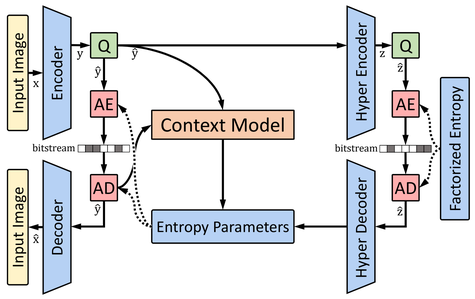
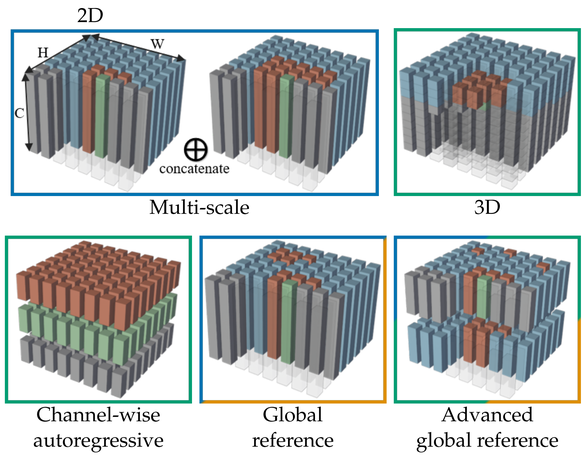


Figure 1: Commonly adopted compression framework [8]

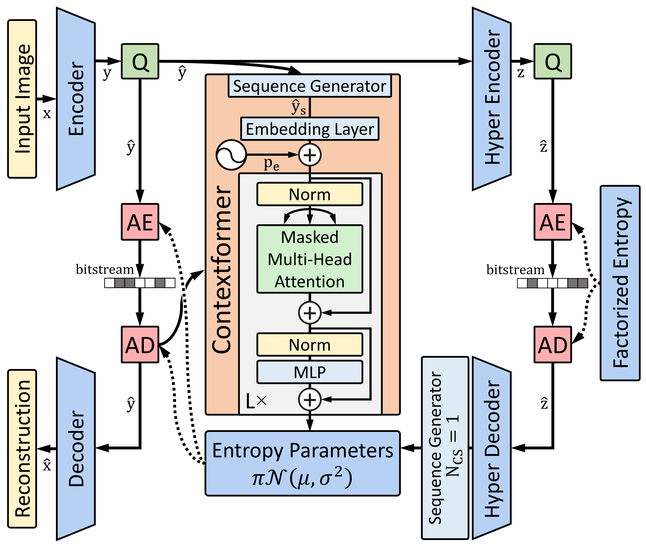


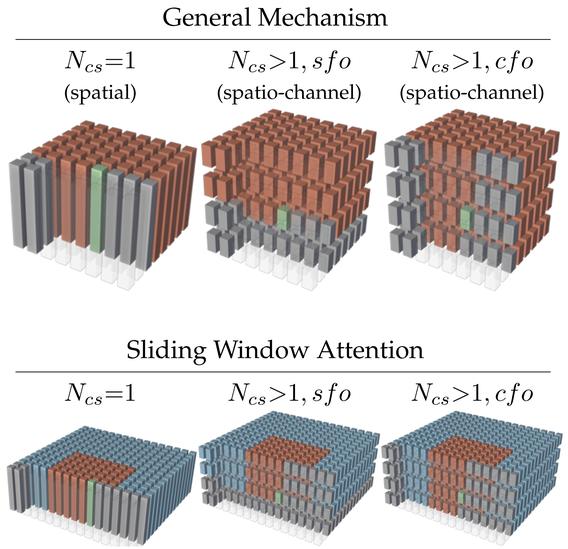
Recent proposals in context modeling improved the performance by:

- Increasing support for spatial dependencies
- Exploiting of cross-channel dependencies
- Increasing context adaptivity

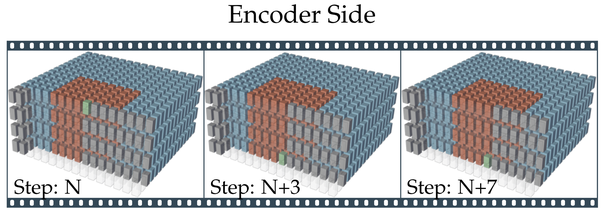
However, efficient exploitation of the latent space relations is still under-explored.

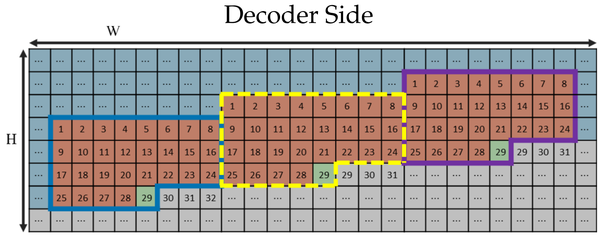
Our Method





Runtime Optimization





Encoder Side:

- Step N+3 contains the calculation for Step N
  ↪ Skip intermediate Channel Segment (SCS)
- Calculate Step N+3 and Step N+7 in parallel
  ↪ Batched Dynamic Sequence (BDS)

Decoder Side:

- Wavefront Coding

Method	Enc. Time [s] Kodak	Enc. Time [s] 4K	Dec. Time [s] Kodak	Dec. Time [s] 4K
w/o Optimization	56	1240	62	1440
BDS (ours)	32	600	–	–
BDS&SCS (ours)	8	120	–	–
Wavefront (ours)	40	760	44	820
3D context [3]	4	28	316	7486
2D context [4]	2	54	6	140
VTM 16.2 [1]	420	950	0.8	2.5

Experimental Results

Contextformer is implemented on top of Cui et al. [5].

Performance:

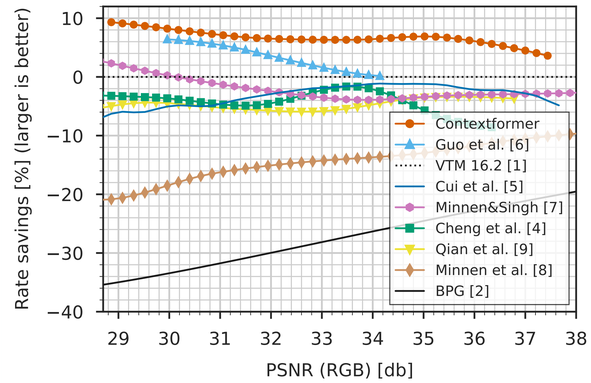


Figure 2: The rate savings relative to VTM 16.2 as a function of PSNR on the Kodak dataset

Method	BD-Rate [%] (lower is better) Kodak	CLIC-P/-M	Tecnick
VTM 16.2 [1]	0.0	0.0/ 0.0	0.0
BPG [2]	30.3	40.5 / 30.9	30.5
Minnen et al. [8] ●	14.7	11.6 / 14.1	13.7
Cheng et al. [4] ●	4.2	5.9 / 9.1	4.8
Cui et al. [5] ◆	3.2	–	–
Minnen&Singh [7] ✚	1.9	–	-2.1
Qian et al. [9] ■	4.7	–	–
Guo et al. [6] ★	-3.7	–	–
Contextformer ✖	**-6.9**	**-9.8** / **-5.9**	**-10.5**
Contextformer ▲	-1.8	–	–
Contextformer ■	-0.7	–	–

CNN-based
● 2D ◆ Multi-scale ✚ Channel-wise ★ Advanced global ref.
Transformer-based
■ spatial ▲ spatio-channel, *sfo* ✖ spatio-channel, *cfo*

Model Size:

Method	Auto-encoder	Hyper-prior	Context& Entropy
Contextformer ✖	17.5M	4.0M	15.9M
Cui et al. [5] ◆	17.5M	4.0M	17.2M
Qian et al. [9] ■	7.6M	25.1M	13.1M
Minnen&Singh [7] ✚	8.4M	11.0M	101.9M

Contact Information

Burak93/Contextformer burakhan.koyuncu@tum.de burakhan-koyuncu

References

[1] Versatile Video Coding. Standard, Rec. ITU-T H.266 and ISO/IEC 23090-3 (Aug 2020)
[2] Bellard, F.: Bpg image format (2015), accessed: 2022-06-01. URL https://bellard.org/bpg
[3] Chen, T., et al.: End-to-end learnt image compression via non-local attention optimization and improved context modeling. IEEE Trans. Image Process. (2021)
[4] Cheng, Z., et al.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog. (2020)
[5] Cui, Z., et al.: Asymmetric gained deep image compression with continuous rate adaptation. In: Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog. (2021)
[6] Guo, Z., et al.: Causal contextual prediction for learned image compression. IEEE Transactions on Circuits and Systems for Video Technology (2021)
[7] Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: IEEE Int. Conf. Image Process. (2020)
[8] Minnen, D., et al.: Joint autoregressive and hierarchical priors for learned image compression. In: NeurIPS (2018)
[9] Qian, Y., et al.: Entroformer: A transformer-based entropy model for learned image compression. In: Int. Conf. Learn. Represent. (2021)