Projet PRIM 2024-2025

Supervisor : Attilio Fiandrotti, Alaa Mazouz, Sumanta Chaudhuri

Contact : *sumanta.chaudhuri@telecom-paris.fr* *alaa.mazouz@telecom-paris.fr* *attilio.fiandrotti@telecom-paris.fr*

Field : : AI, Multimedia, Embedded Systems

Key words : : Neural Networks, CNN, Learned Image Compression, FPGA

No. of Students: 1-3

Learned Image Compression on FPGA

In Learnable Image Compression (LIC), the image is projected to a low-dimensional latent space by a convolutional encoder at the source side. Such representation is quantized and entropy-coded in the form of a binary bitstream. At the receiver, the bitstream is entropy-decoded, a convolutional decoder projects such representation back to the pixel domain, recovering an approximate representation of the image. Early seminal works accounted for a unique latent representation modelled with a fully factored distribution [Balle2016]. Since then, much of the research in the field has focused on improving the compression efficiency by refining the entropy model. This basic scheme was then improved by introducing an auxiliary latent space called hyperprior capturing spatial correlation within the image, furthering compression efficiency [Balle2018]. LIC has shown the ability to outperform standardised video codecs in compression efficiency, fostering the demand for embedded hardware implementations.

Achieving realtime coding on resource constrained platforms such as FPGAs demands ad-hoc design choices such as in the state of the art LIC implementations [Jia2022, Sun2024]. However, FPGA implementations have been lagging behind recent research in LIC due to the increasing complexity of implementing in hardware recent LIC models. For example, [Minnen2020] further improves the RD efficiency by introducing slice-based latent channel conditioning and latent residual prediction with an approach suitable for parallel execution. The RD efficiency is further boosted in [Zou2022] by introducing a Window Attention Module in the autoencoder architecture and experimenting with a transformer-based architecture in place of the traditional convolutional architecture.

Objectives and Methods:

The goal of this PRIM project is

- To study state of the art LIC models ([Minnen2020],[Zou2022]) w.r.t the suitability of their FPGA implementation.
- Applying frugal AI techniques such as Pruning/Quantization/Kowledge Distillation within acceptable loss of RD performance.
- Accelerating parallelizable parts of the algorithms using FPGA to meet realtime coding targets.

The methods/tools used in this project will be

- Pytorch for model exploration/quantization/pruning.
- Xilinx Vitis-AI for FPGA implementation.
- The final design and performance testing will be done on Xilinx ZCU102 and KV260 AI SoM (System on Module)

References :

[Balle2016] Ballé, Johannes, Valero Laparra, and Eero P. Simoncelli. "End-to-end optimized image compression." *arXiv preprint arXiv:1611.01704* (2016).

[Balle2018] Ballé, Johannes, et al. "Variational image compression with a scale hyperprior." *arXiv preprint arXiv:1802.01436* (2018).

[Minnen2020] Minnen, David, and Saurabh Singh. "Channel-wise autoregressive entropy models for learned image compression." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

[Zou2022] Zou, Renjie, Chunfeng Song, and Zhaoxiang Zhang. "The devil is in the details: Window-based attention for image compression." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[Jia2022] Jia, Chuanmin, et al. "Fpx-nic: An fpga-accelerated 4k ultra-high-definition neural video coding system." *IEEE Transactions on Circuits and Systems for Video Technology* 32.9 (2022): 6385-6399.

[Sun2024] Sun, Heming, Qingyang Yi, and Masahiro Fujita. "FPGA Codec System of Learned Image Compression with Algorithm-Architecture Co-Optimization." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (2024).