Transformer-Based Learned Image Compression with Cross-Scale Attention

Background

Transformers have revolutionised visual representation learning by modelling long-range dependencies through self-attention [Dosovitskiy et al., 2021]. In image compression, attention mechanisms can capture complex spatial correlations that conventional CNNs overlook. Recent hybrid methods such as ViT-VAE [Li et al., 2023] and SwinT-Compression [Chen et al., 2024] demonstrate state-of-the-art performance at low bit-rates.

Problem Description

This project aims to design a **transformer-based image compression architecture** that integrates **cross-scale self-attention** to replace or complement convolutional transforms. The student will investigate rate–distortion–perception trade-offs and compare against baseline CNN/entropy-bottleneck codecs.

Milestones and Extensions

- Literature study: learned compression and ViT architectures.
- Implement transformer encoder–decoder within the CompressAI framework.
- Explore hybrid CNN + ViT blocks for local/global correlation modeling.
- Evaluate RD performance (PSNR/LPIPS/BD-Rate) on Kodak, CLIC, and Light-Field patches.
- **Extension:** perceptual fine-tuning using adversarial or perceptual losses.

Tools, Qualifications, and Outcomes

- **Skills:** Python, PyTorch, understanding of self-attention and information theory.
- **Tools:** CompressAI, Hugging Face Transformers, GPU cluster.

- **Outcomes:** A ViT-based compression prototype and reproducible benchmarks.

References

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D. and Zhai, X., 2021. Thomas unterthiner mostafa dehghani matthias minderer georg heigold sylvain gelly jakob uszkoreit and neil houlsby. An image isworth 16× 16 words: transformers for image recognition atscale. In *International Conference on Learning Representations* (Vol. 1, No. 2, p. 3).
- https://www.kaggle.com/code/olgakozhushko/vit-vae-vision-transformer-variational-autoencoder
- Wang, S., An, P., Yang, C., Huang, K. and Huang, X., 2024. STSIC: Swin-transformer-based scalable image coding for human and machine. *Journal of Visual Communication and Image Representation*, *98*, p.104016.