Image coding algorithm for DNA data storage combining JPEG and autoencoders

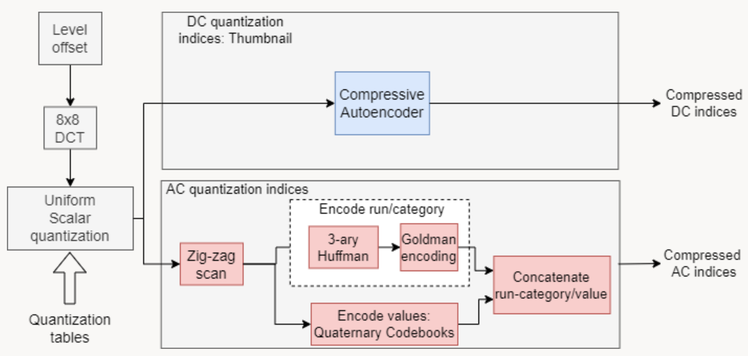
Xavier Pic, Marc Antonini

Image coding algorithm for DNA data storage combining JPEG and autoencoders

Xavier Pic [1] Marc Antonini [1]

[1]I3S laboratory, Côte d'Azur University and CNRS
UMR 7271, Sophia Antipolis, France

Proposed workflow



- Based on the work of Melpomeni et al.[1]
- Block DCT based encoding into a DNA sequence
- DNA: sequence of nucleotides A, T, C and G : AACTCAGCATGCAGGG....
- Formatted oligos synthesized into DNA molecules
- DNA molecules sequenced to retrieve the oligos
- Decoding the oligos to retreive the original image

JPEG thumbnails

- Critical information
- Thumbnail can be encoded seperately

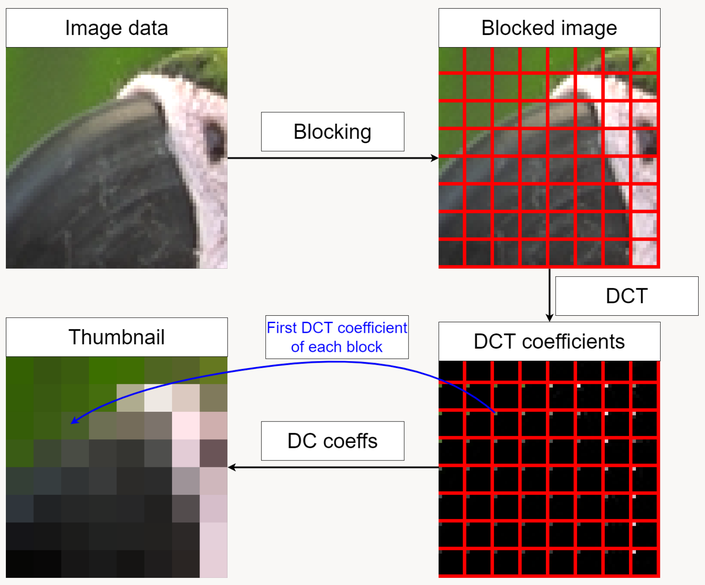


Figure 1. JPEG proedure and the thumbnail

Data Compression

- Thumbnail (DC coefficients): Compressive Autoencoder
- AC coefficients : Goldman-based [3] Variable Length coding

Image reconstruction

Image retrieved by decoding and recombining the Thumbnail (DC coefficients) and the AC coefficients and applying inverse DCT to it.

Figure 2. Original and Reconstructed image at 37.2 dB, compression rate: 33 bits/nt

References

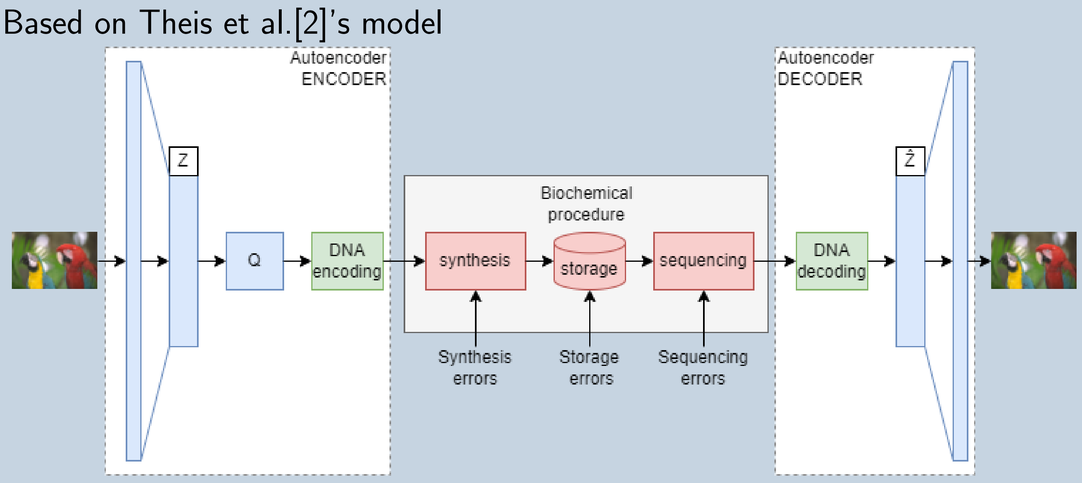
[1] M. Dimopoulou, M. Antonini, P. Barbry, and R. Appuswamy. "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA". In: *European Signal Processing Conference (EUSIPCO)* (2019).

[2] L. Theis, W. Shi, A. Cunningham, and F. Huszár. "Lossy image compression with compressive autoencoders". In: *International Conference on Learning Representations* (2017).

[3] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA". In: *Nature* 494.7435 (2013), p. 77.

Compressive autoencoders

- Based on Theis et al.[2]'s model



- Trained to adapt to the DNA storage channel errors:

Original	ATACTAGCT
Substitution	ATACAAGCT
Insertion	ATACGTAGCT
Deletion	ATAC_AGCT

- Reconstruct image from noised data

Model mathematical definition

$$\hat{I} = g(\overline{\beta_{DNA}}(\alpha_{DNA}(Q(f(I)) + \epsilon)) \quad (1)$$

- f : encoding part of the autoencoder
- g : decoding part of the autoencoder
- Q : quantization ($Q(z) = q \times \left\lfloor \frac{z}{q} + \frac{1}{2} \right\rfloor$)
- f_{DNA} : DNA encoding function
- g_{DNA} : DNA decoding function
- ϵ : Noise generated by the DNA storage channel

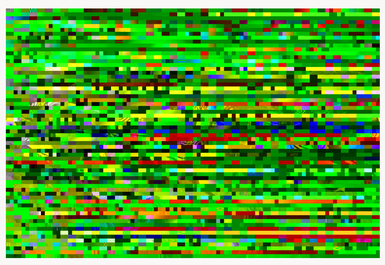
Training process

- Training Loss: $L = ||I - \hat{I}||^2 + \lambda H(Q(Z))$
  - $||.||^2$: Mean Squarred Error
  - Z : latent space ($Z = f(I)$)
  - H : entropy $H(Q(Z)) = -\sum_{i=1}^{n} Pr\{Q(z_i)\} log_4 Pr\{Q(z_i)\}$
  - λ : weight parameter
- Optimizer: Adam
- Training Dataset: Thumbnails of Flickr30k (∼30k images from social media)
- Validation Dataset: Thumbnails of Kodak Image dataset (24 images)
- DNA storage chanel noise model

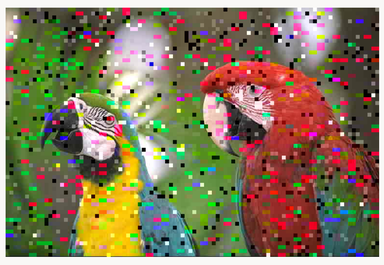
Results

- The compression in terms of performance is lower than the jpegdna one
- Achieving lower bit-rates is harder because of the size of the compressed representation of the tumbnail
- Better performance on noisy channels
- More resistant to noise on the critical data managed by the autoencoder
- More adapted to a noisy channel like synthetic DNA data storage
- 1.5% subsitution rate corresponds to the Nanopore sequencing method's substitution rate

JPEG DNA, no noise

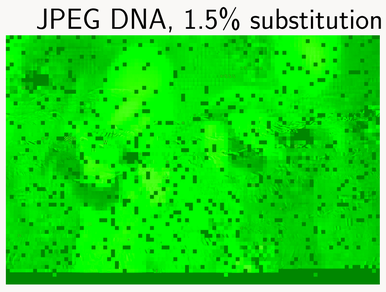


JPEG DNA, 1.5% substitution

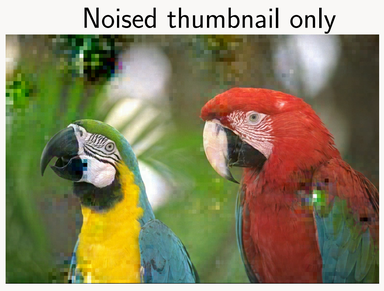


Noised thumbnail only

Our method, no noise



Our method, 1.5% substitution



Our method, noised thumbnail only