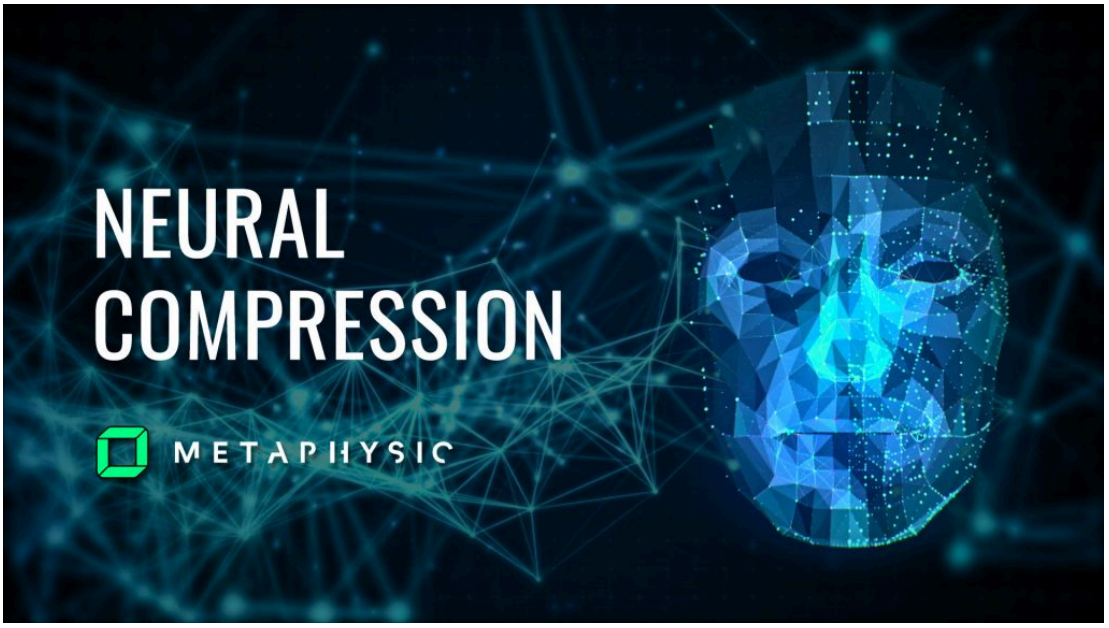
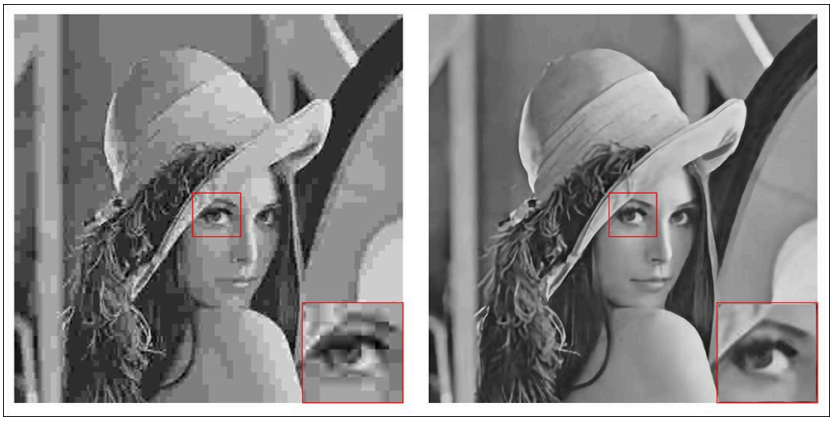
What is Neural Compression?

Originally published at https://blog.metaphysic.ai/what-is-neural-compression/

October 31, 2022

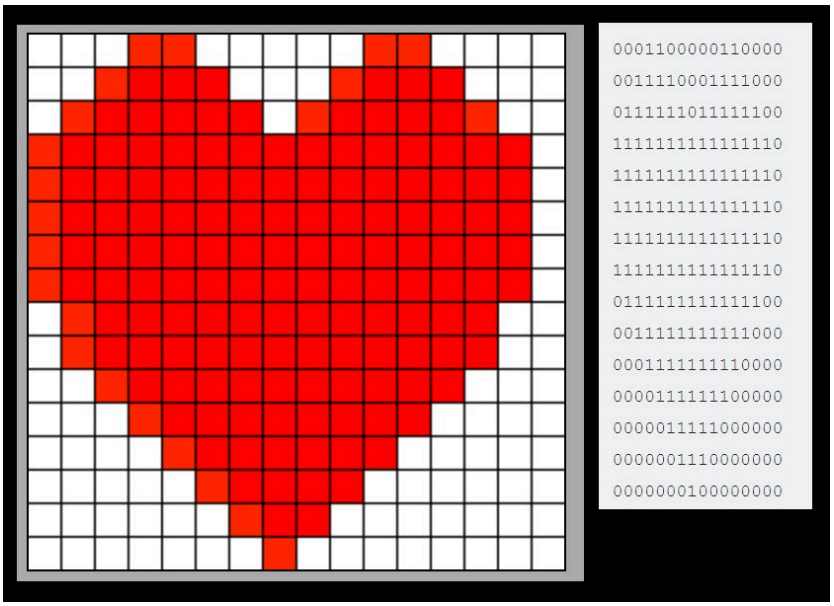


Neural Compression is the conversion, via machine learning, of various types of data into a representative numerical/text format, or *vector format*.



The classic computer vision test case, Lena, compressed (left) by a conventional bitmap-based codec, and by neural compression (right), which does not store bitmapped data, but rather extracts generalized 'features' from which the original image can be reconstructed. Source: https://arxiv.org/pdf/1708.00838v1.pdf

Normally, images are saved as pixel data, which has hard practical limits on compressibility – no matter how ingenious your bitmap-based compression algorithm or video codec is, it eventually has to resolve to an array of hard values, and those hard values can't be compressed any further by traditional methods.

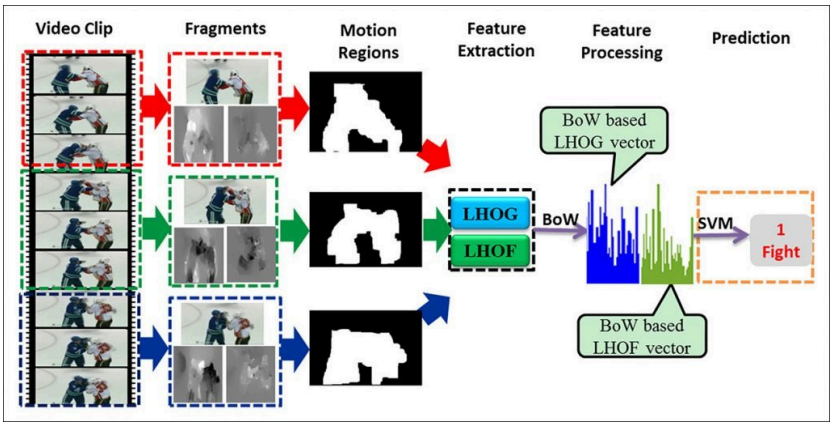


Left, non-reducible raw pixel values, as seen by the viewer (if they zoom right in) and, left, the same pixels in the form of their equivalent hex code. Source: https://www.khanacademy.org/computing/computers-and-internet/xcae6f4a7ff015e7d:digital-information/xcae6f4a7ff015e7d:data-compression/a/simple-image-compression

To boot, almost any kind of video or image compression entails throwing away some of the data permanently. This is known as lossy encoding, where, for instance, in the hex dump visualized above, the sequence '0000000' might be reduced, effectively, to the phrase 'three zeroes'. However this type of run-length encoding does not save enough space for commercial use, and does not capture extensive detail economically. Recorded as raw data, without any compression, an hour of 1920×1080 HD video would likely take up 0.61TB of disk space, depending on the complexity of the content. This means that a single two-hour movie would occupy more hard disk space than the average laptop computer currently contains – and would additionally have such a high bitrate that it would be difficult to play, and almost certainly impossible to stream.

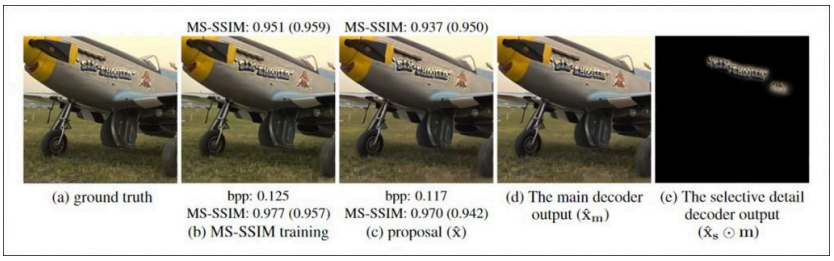
'Features' in Neural Compression

Therefore in recent years, interest has grown in storing image content by some other method than dumping pixels into files (with the ensuing loss of quality associated even with the best traditional image and video codecs).
Interest in neural compression has grown notably in the research community, not least because of the technique's potential to save complex information, including video information, in a potentially truly 'lossless' format that could be rendered back for viewing at the full capture resolution (or more, with upscaling), whilst occupying just a fraction of the hard disk space of its pre-AI equivalent.



In this example from a violence detection system, pixel data is converted into vector (mathematical) data which is far easier to manipulate and store. Source: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0203668&type=printable

The process of translating image data into vector data may seem unintuitive, but really it's analogous to describing a scene for vision-impaired viewers, or, arguably, describing music without the use of audio – except that it is far more successful in representing the original than these techniques.
Interestingly, some neural compression researchers are showing interest in the same kind of detail-focused lossy compression that typifies historical image compression codecs such as JPEG.
Under this approach, the neural compression process identifies and gives more weight and emphasis to *perceptual loss* – the fact that our eyes tend to notice certain details more than others. By increasing the fidelity of these details at the expense of surrounding image information, we are likely to 'feel' that the resulting image or video is more detailed than it actually is.



Research from a 2020 paper which uses neural network-based learned image compression to accentuate details that give the impression of a 'feature rich' image, whilst making economies elsewhere. Source: https://tinyurl.com/462fsby8

Though the science repositories are full of putative systems that may one day power the way we watch, create and manipulate video and image content, there are more central and conceptual uses for feature extraction in the field of image synthesis, in architectures such as autoencoders, Neural Radiance Fields (NeRF) and Generative Adversarial Networks (GANs).
In the case of these systems, the fact that neural compression happens to produce extraordinarily compressed representations of images is only an added side-benefit, though a welcome one. Let's now take a look at some neural compression implementations where versatility is favored over fidelity.

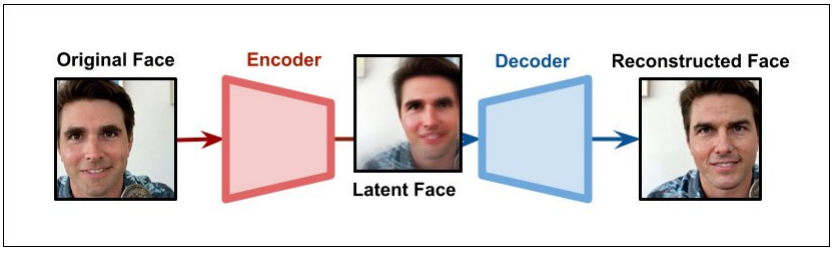
Neural Compression as a Creative Tool

Once a video is neurally compressed, it becomes far more motile, in terms of its potential for creative interpretation.
In the active research sector of Space-Time Video Super-Resolution (STVSR), it's possible to reduce even further the amount of necessary recorded information, since the learned features of the image frames and their temporal behavior can be manipulated not only into higher resolutions, but also converted into higher frame rates, by interpolating between existing frames – a technology that has been available for some years through frameworks such as DAIN (Depth-Aware Video Frame Interpolation).

https://www.youtube.com/watch?v=n0J5H-F_s0k

Above is an example of one STVSR project, from 2022, which is capable of arbitrary upscaling and ad hoc frame interpolation. In the above example, a low frame-rate clip has been massively interpolated into slow-motion. In the example below, a 'jerky' frame-rate has been transformed into a more acceptable appearance for a modern viewing audience, by the same method.

In regard to facial synthesis, it is easy to confuse 'features' (in the sense described earlier) with 'facial features', but they are not the same thing. A derived 'feature', in the computer vision sense, may represent anything, *including* a face.
The most famous neural compression systems in the world, currently, are the open source software distributions derived from the controversial deepfakes code released to Reddit in 2017 code – though neural compression is, in this use case, only a means to an end.



The 'encoding' part of the deepfakes creation pipeline extracts essential and characteristic facial identity features from hundreds or even thousands of source data images, until a trained (and difficult to visualize) representation of that person exists within the latent space of the autoencoder system (represented in the middle image above).
Though pixel-based morphing has been possible at least since the 1980s (when some of the earliest neural compression papers began to emerge), the extraordinary ductility of extracted features makes the deepfake process far more powerful and potentially transformative than 'frozen' pixels.
At the same time, thanks to neural compression, the entirety of the two identities in a trained deepfake model typically occupy less than a gigabyte of disk space.

The Efficacy of Neural Compression

Likewise, a typical checkpoint (model) for the hugely popular Stable Diffusion latent diffusion text-to-image framework weighs less than 4GB, despite containing extracted features from over 2.3 billion images in the LAION 5B-based dataset that powers the system.
Assuming that each contributing source image weighed no more than 100kb (which is an absurdly conservative estimate), storing that extent of pixel data in a single database capable of reproducing the source images in some way would result in a file weighing *420 terabytes*.

In fact, the neural compression evident in Stable Diffusion, though relatively typical of encoders that generate a latent space, has even been used experimentally as an image compression technique in its own right.

On the left, an image compressed via the JPEG compression algorithm; in the middle, the original ('ground truth') image, which itself has a number of compression artifacts already; and, right, the image reproduced via Stable Diffusion by one curious developer, showing superior resolution – though some of it may arguably be 'fictional', derived from similar and better-detailed instances of the same kind of subject matter (i.e., 'distant city skylines') that exist in the trained database. Source: https://pub.towardsai.net/stable-diffusion-based-image-compresssion-6f1f0a399202

Neural Compression 'Puppetry' for Video-Conferencing

Several works and initiatives of recent years have concentrated on the minimal transmission of actual information over a network, positing that the receiving apparatus will be capable of inferring the correct image information, based on augmentation of the scant transmitted data via slimmed-down local neural networks.

The idea received its most popular proponent in 2020, in the form of NVIDIA's demonstration of the potential 'virtual meetings of tomorrow' via its Maxine system. Maxine effectively uses a kind of deepfake puppetry in order to transmit only human body and facial motion information, and some minimal keyframes; subsequently, the receiver's local equipment 'tweens' and interprets the movement of their correspondent, with very little actual data-rich information passing between the two communicants:

This is equivalent of an encoded message that tells the user to look at page 273 of a book that they already own: the message is mere bytes, but the resulting experience is rich. In a sense, the principle is unchanged from traditional video codecs, which already require local support (i.e., in the web browser or local operating system) in order to play back codec-encoded video.

The Maxine system offers a 10x reduction in video data transmission over traditional VOIP platforms such as Zoom, claiming to require only a few kb per frame. Neural compression is central to an autoencoder-driven system of this type; not only to minimize data transmission, but also to ensure that each end user is not required to install terabytes of data in support of the local neural network that is doing the heavy lifting for the process.

However, noted above in regard to Stable Diffusion, there is always the risk with a generalized trained model that it will not or cannot reproduce exactly the image that was fed into it, but may go hunting around for 'similar content' in its trained database that will augment the user experience at the cost of fidelity.

Therefore, if neural compression is to become a new standard in the years ahead, it will need to address some notable public and legal concerns about the potential for AI to interfere in the veracity (rather than 'authenticity') of transmitted or recorded content.

The Future of Neural Compression

Video codecs that use neural compression are not exempt from some of the tiresome challenges and tribulations that continue to face the pixel-based compression research community, such as the need to trade off detail and fidelity against other factors, including compression time, and the minimum expectations of resources on the host system.

In the case of some of the more bleeding-edge initiatives in neural compression, getting the local resources requirements to a rational level represents a particular challenge, though the increased use of dedicated local neural network modules in modern consumer hardware promises to improve the situation.

At the moment, the challenge is being met primarily by creating neural compression codecs that are targeted at very specific use-cases, wherein the codec may be optimized not only for a particular view of a surveillance camera, but for particular hardware that it is running on.

If the evolution of neural compression follows trends in other kinds of revolutionary technology, we can expect an early multiplicity of dedicated codecs designed to operate more generally across a far wider range of domestic and professional computer hardware, before an 'acceptable open source standard emerges to sideline the early attempts at monopoly and market capture.

As with VHS vs. Betamax, and DALL-E 2 vs Stable Diffusion, the best product may turn out to be the most available, rather than the most capable.

But in any case, an effective and widespread neural compression codec will need to take in a far greater range of potential use cases than many of the most efficient and impressive efforts currently do. For the time being, neural compression is likely to remain a nascent codec technology, but an active creative tool.