

What Is the Latent Space of an Image Synthesis System?

Originally published at <https://blog.metaphysic.ai/what-is-the-latent-space-of-an-image-synthesis-system/>

November 17, 2022



If you have any interest in image synthesis architectures such as Generative Adversarial Networks ([GANs](#)) or latent diffusion frameworks like [Stable Diffusion](#), you'll have noticed references to the 'latent space' of these systems. The latent space, which we'll look into more deeply here, is the 'subconscious' and overarching understanding of relationships between learned data points that a machine learning system has been able to derive from the information that it gets fed.



CLICK TO VIEW ANIMATED GIF

The latent space of number representations/embeddings in the MNIST dataset, primarily used to develop and improve optical character recognition (OCR), visualized in TensorFlow. The relationship between the images of characters and their accompanying text is explicit and obvious, but the data has been ordered by the AI so that the most 'similar' data points are grouped together (low dimensionality, i.e. the '2s' are relatively near each other), whilst maintaining the higher groupings (high dimensionality, i.e., the higher numbers are broadly clustered together). Source: <http://projector.tensorflow.org/>

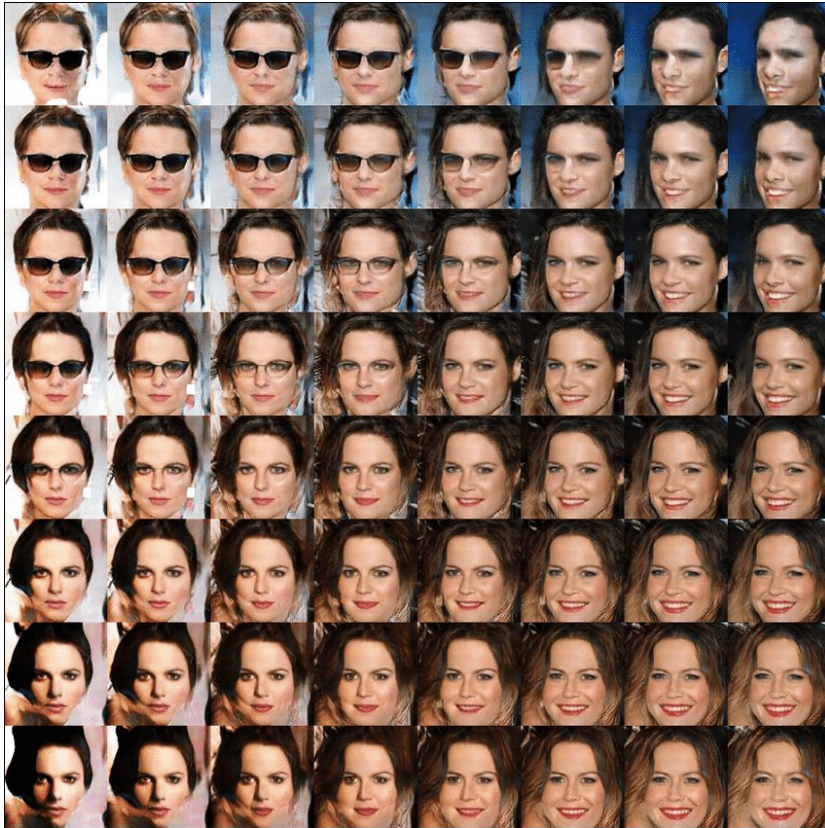
Arguably, we also put our own picture of the world together in this way, according to imperfect information that enters our developing consciousness either in an ordered way (didactic education) or an unordered way (random chance and happenstance). With our very survival at stake, we're likewise challenged to categorize and order incoming information into a cohesive and functional network of relationships which will eventually inform our rational processes and cognitive abilities. Though we can all probably individuate pivotal events and data that defined our development, most of us have little better than an intuitive or vague understanding of our own model of the world. Like trained AI systems, we demonstrate these nascent connections most [in our choices and interpretations](#) of events and data. The operations of the supporting mechanisms, many of which were formed before our age of reason, are not always clear to us.

So it is with the latent space, which is not a pre-formed 'mail-room' waiting for parcels to put into cubby-holes, but which will be *architecturally defined* by the data that fills it – and which tends to resist active investigation.

The Utility of the Latent Space

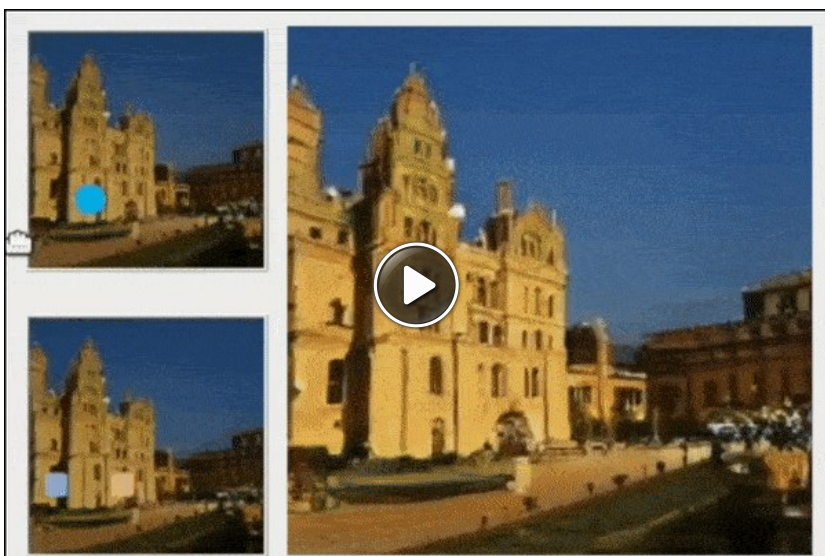
Being able to operate at will in the latent space is a powerful method of gaining increased and more granular control over a trained system, because you can potentially tell a process or agent where to start, and what it is that you want to achieve.

For example, in the image below, we observe the process of cycling through embeddings of faces trained into the latent space of a Wasserstein GAN with Gradient Penalty ([WGAN-GP](#)) model:



Identifying one embedding and 'sending it' somewhere else in its latent space - an example of why the latent space is worth exploring in image synthesis. Source: https://www.researchgate.net/figure/Bilinear-interpolation-on-latent-space-for-random-noise-vectors-Dataset-used-is-CelebA_fig17_324057819

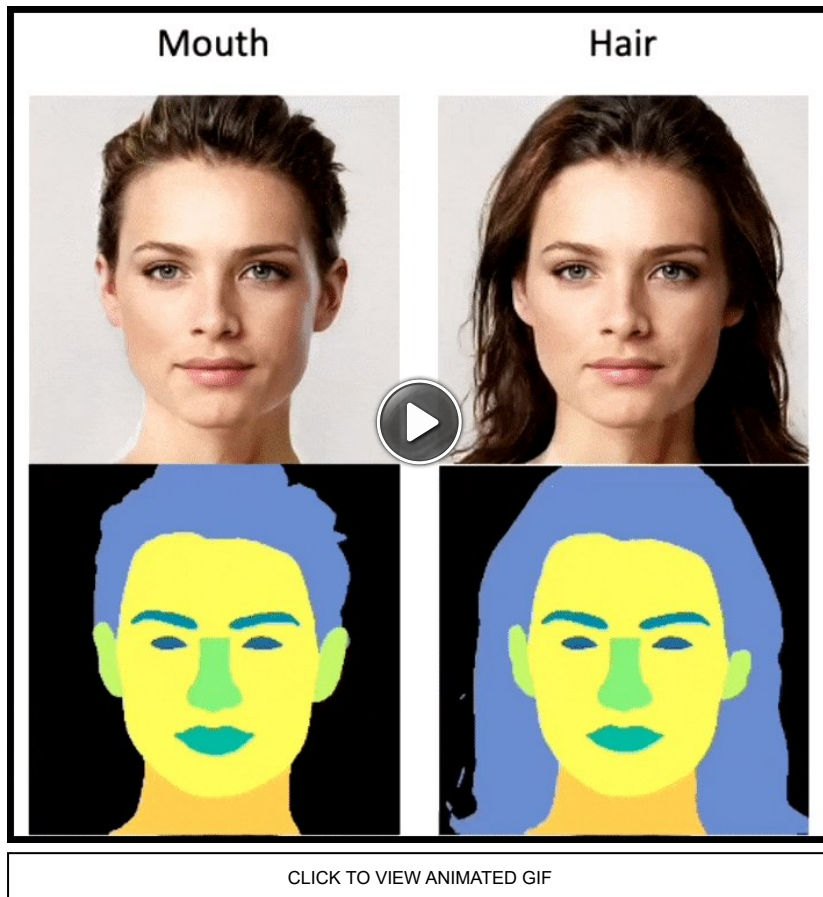
In the animation below, we see researchers from the Chinese University of Hong Kong and the Australian National University cycling through church designs in a trained GAN using a simple 'hand' cursor – a feat made possible by [using 'heat maps'](#) to illuminate the otherwise-hidden routes between embeddings in the latent space, and to instrumentalize them:



CLICK TO VIEW ANIMATED GIF

The EqGAN method, revealed in 2022 by researchers at the Chinese University of Hong Kong and the Australian National University, uses the GradCAM system like a kind of barium meal for GANs, revealing the location of specific information in the latent space, so that the data can be addressed and transformed, by 'dragging' it over to other locations. Source: <https://genforce.github.io/eqgan-sa/>

A little closer to our own area of interest, greater control of the latent space means potential mastery of the transformational powers of a trained image synthesis system, such as changing *just one aspect* of a facial representation, by knowing exactly where, for instance, the 'hair' or 'mouth expression' codes are located, and operating exclusively on them:



Semantic StyleGAN, a paper from December 2021, addresses the problem of entanglement by training the system compositionally, with separate attention to areas which may need to be addressed individually, such as hair or lips. Source: <https://semanticstylegan.github.io/>

The Latent Maze

However, the latent space is a very strange place. You can't explicitly design one, except to the extent that you refine and curate the data that it feeds on. Rather, it's assembled by the training algorithm over a period of some hours (or even some weeks or months), and represents a vast, multi-dimensional array of values that the AI was able to extract from the data that you gave it.



By inputting new search terms on a trained database, the TensorFlow embedding projector easily drills down to sub-topics and classes in the dataset, revealing where they are positioned among the quadrants. Source: <http://projector.tensorflow.org/>

The latent space is 'multi-dimensional' for two reasons: firstly, because many of the embeddings (i.e. the extracted information) that occupy it belong in more than one place, and can't be accommodated by something as simple as a Venn diagram or a magic quadrant. Secondly, and more importantly, because it can cohesively represent a very broad grouping (such as 'people'), a narrower group (such as 'women'), and also represent a single embedding that belongs to all these categories (such as 'Taylor Swift', a valid entry in all three levels of dimensionality).

The diversity of categories in a complex latent space, such as a general latent diffusion image synthesis system like Stable Diffusion, can place embeddings in multiple 'drill-down categories' of this kind, such as *people>women>Taylor Swift*, *workers>entertainers>Taylor Swift*, or *composers>modern>Taylor Swift*.

In the animated image above, we see several searches, for the terms 'man', 'woman' and 'person', drilling down into the latent space of the [Word2Vec](#) 10K Natural Language Processing dataset, and revealing where these values lie in what otherwise appears to be a vast and messy self-assembled cloud of embeddings.

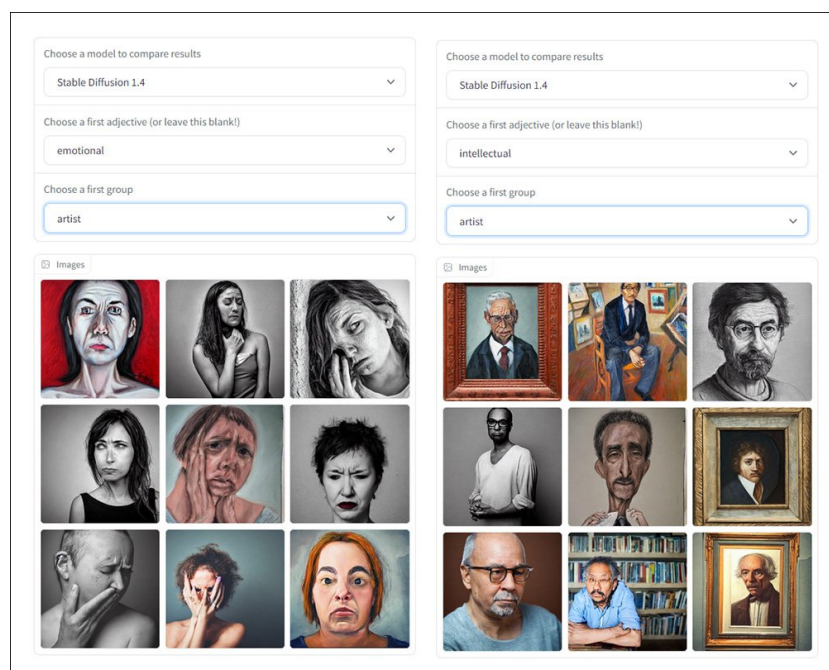
If we just stay out of the latent space of a trained system and run queries on it, we can get the useful results that were the objective of training the system in the first place, such as generating [novel photorealistic faces](#), or obtaining [conversational reasoning](#) from a system such as OpenAI's GPT-3.

However, compared to the extent to which we can control information in pre-AI systems such as Photoshop, or in CGI workflows, or in a program such as Microsoft Word, we have remarkably little ability to intervene in a latent space once it has formed, or even to understand how the interrelationships between the data points operate.

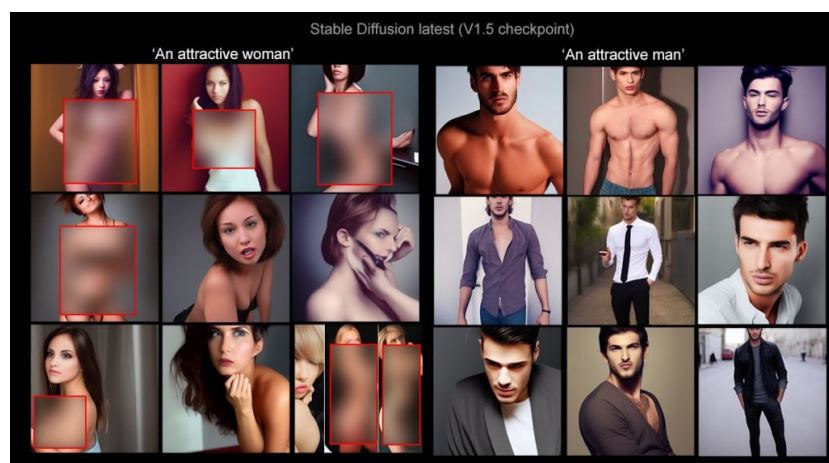
This makes for powerful but [opaque](#) systems, an uncomfortably Druidic workflow, and for systems that cannot easily be investigated for bias. The latter issue has made explainable AI (XAI) a driving concern for state and responsible private sector use of machine learning systems trained on under-curated, web-scraped data that's likely to embed undesirable biases into the output of trained systems.

A Crude Flashlight in the Latent Space

Most methods for understanding the disposition of a latent space are *diagnostic* in nature, merely testing the system through inference, such as provoking a natural language processing (NLP) system to [reveal intrinsic racial prejudice](#), [reinforce stereotypes](#), or sexualize women (see images below).



The 'emotional artist' and the 'intellectual artist' are strictly gender-defined, according to a recent build of the standard Stable Diffusion (1.4) checkpoint. Source: <https://huggingface.co/spaces/sasha/StableDiffusionBiasExplorer>



Two simple (non-cherry-picked) prompts from a vanilla install of the latest version of Stable Diffusion - the official V1.5 checkpoint released in October. We can see that 70% of the 'attractive men' are fully-dressed, while the women are all either nude, semi-nude, or in revealing and sexualized attire. These emphases emerge from the fevered tags that followed the images into the training process, many of which are user-contributed. Even where tags were originally 'guessed' by AI, these algorithms too are following human trends in small and biased communities, and delivering that bias and emphasis straight into the heart of the AI system.

Actually tracing the sequence of events that define an AI's decisions at inference time is more challenging, due to the 'holographic' nature of the latent space, and the complexity of the interrelationships between the embeddings that it contains.

One popular and (by now) quite mature solution, frequently featured in new research aimed at demystifying the latent space, and also used in the 'morphing church' project featured earlier, is *Gradient-weighted Class Activation Mapping* ([Grad-CAM](#)), an academic collaboration between the Chinese University of Hong Kong, the Australian National University, and the University of California at Los Angeles.

Grad-CAM: Gradient-weighted Class Activation Mapping

Grad-CAM highlights regions of the image the captioning model looks at while making predictions.

Try Grad-CAM: Sample Images

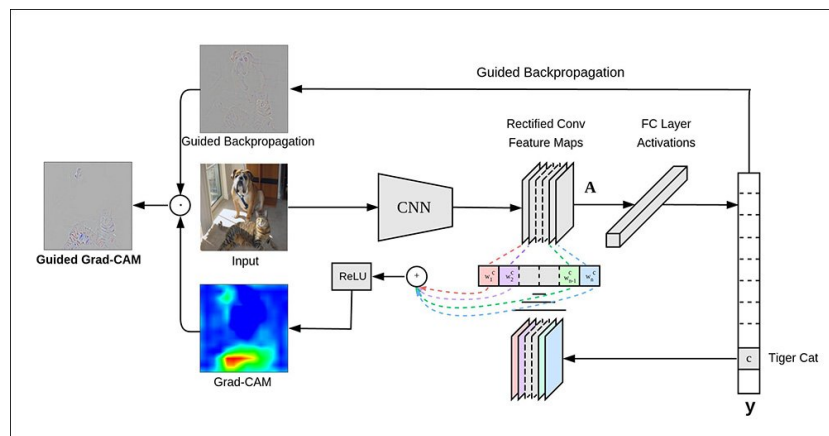
Click on one of these images to send it to our servers (Or upload your own image below)



Show Demo Images

CLICK TO PLAY VIDEO ON SITE

Grad-CAM uses guided [backpropagation](#) to generate 'heat-maps' that can make explicitly visible the way that the trained neurons formed associations in response to a request from the user, thus providing some kind of 'justifying rationale' that could help researchers not only identify the ways in which undesirable bias is formed, but also prove an aide for image synthesis systems that wish to individuate and disentangle certain visual elements, without adversely affecting other elements of a generated image or video.



The conceptual architecture for Grad-CAM. Source: <https://github.com/ramprs/grad-cam/>

However, Grad-CAM was released five years ago now, and the majority of new research into XAI and conceptual mapping of the latent space continues to be passive and interpretive, and the latent space itself a challenging mystery.

Since one objective of machine learning research is to understand the world as it is, 'anti-bias' latent space exploration systems are operating under a kind of ideological conflict: if they permit generative AI systems to become entirely bowdlerized (or, in the terminology of the worst of Reddit and Twitter, 'woke'), then whatever appeal or unfettered creative impetus popularized the system will presumably migrate to less locked-down frameworks and outlets.

On the other hand, if they do nothing but observe and analyze the output of freely available latent spaces, such as Stable Diffusion, the resulting controversies seem likely only to proliferate. This characterizes the challenge as a cultural and ethical one, far beyond the purview of the 'enabling' technologies, which would seem to be a barometer of culture rather than a 'villainous' transformative technological phenomenon.

Conclusion: The Price of Automation

If you ask Mary Poppins to go in and magically clean up your messy bedroom, you're saving a lot of personal labor; but you're not going to know where anything is later, because she follows her own system. It's probably a better system than yours, but that doesn't help you find your clean socks.

So it is with the latent space, which, in formation, has traversed and categorized millions, or even billions of data points at a post-human scale and speed, deriving the requisite compositional logic and relationships from cheaply, often badly-labeled data which would be unconscionably expensive to curate manually (if far more accurately and fairly).

Whether for addressing bias or gaining greater control over image synthesis, new methodologies and more transparent tools than Grad-CAM will be needed to gain a deeper understanding of the way that the latent space functions. The challenge is compounded by the fact that the architecture of a latent space emerges directly from the data, instead of according to some predefined and universal logic, the imposition of which could hinder a machine learning system's most fruitful and inspired insights – as well as some of its most objectionable standpoints.