**When a neuron tells you what it wants**

When we look at the world before us, we experience a unified scene of shapes and colors. But neurons in our brain do not – they are microscopic automata that can only emit electrical impulses when specific shape patterns appear in their little territory of visual space, their *receptive fields*. Billions of these neurons work together to give rise to visual recognition, and yet they are still outnumbered by the practical infinity of images impinging on the retina over time. That means that the brain must be efficient about allocating its neurons – it cannot assign a neuron to detect every pattern in the world. Neurons should be *selective* about which shape patterns to learn to recognize in scenes –shapes that better be diagnostic of entities essential to the animals’ survival.

The foundation of visual neuroscience rests on understanding the *selectivity* of individual neurons. So how do we discover what any given cell signals about the world? Most of us use a powerful experimental design developed by David Hubel and Torsten Wiesel in the 1950s, where an animal faces a screen, displaying pre-selected pictures, while cortically placed microelectrodes report the presence of neuronal action potentials (“spikes”). But experimental sessions are time-limited and therefore we must have a pretty good idea of what pictures to show the neurons in the first place: lines, angles, pictures of faces, animals, places…miss the right picture, miss the correct conclusion. The problem is that there is an astronomical number of pictures worth testing. Famously, Hubel and Wiesel themselves ran into this obstacle in early experiments, when they tried to use spots (painted on glass slides) to stimulate primary visual cortex cells – only to find that the cells responded better to the actual *slide edge*. How likely is it that most of us would miss that proverbial edge?

What if instead of trying to guess what a neuron encodes, we just let it *tell* us by guiding the development of images based on its own responses? This was the core idea behind our project. Our co-authors Will Xiao and Gabriel Kreiman had found that deep learning models called *generative adversarial networks* (*GANs*), which can synthesize images ranging from abstract to photorealistic, could be directed to create images that made model “neurons” respond more strongly than to any other picture. As part of the Margaret Livingstone laboratory, we set out to discover if this could also work in biological neurons. We used a GAN pre-trained by Jeff Clune’s group at the University of Nebraska, which functioned by taking in a list of 4096 numbers (an *input code*) and outputted an image. The GANwas remarkable in its ability to create abstract new images, not just “remember” photorealistic fragments from training. Our first experiments began as the GAN created random images of formless textures, which were then presented to a monkey keeping a steady gaze on the screen. The images were shown within an inferotemporal cortex neuron’s receptive field, which elicited responses that could be used to evaluate the images: the more spikes elicited by the image, the more likely its GAN input code would be passed on to the next generation (thus the name of the process: a *genetic* algorithm). Over tens of minutes, the neuron-directed GAN images began to acquire form: first a dark spot, then another aligned horizontally; then a convex line, surrounding them both from above – it looked like a *face*, which made sense to us because that particular cell also responded most strongly to pictures of actual faces. But the neuron responded to this abstract image more so than to any face pictures; it was as if the neuron was yelling “yes! This is what I’ve been trying to tell you all this time…!” We replicated this process across neurons and individual monkeys. The neuron-driven synthetic images are a fascinating mix of the familiar and the bizarre: many contain patterns we can easily recognize in monkeys –face-like patterns surrounded by the tan-reddish brown texture reminiscent of monkey fur- but many contain patterns that we cannot easily refer to any given object category. To interpret these synthetic patterns, we used two approaches: first, after a neuron led to a synthetic image, we showed it as many natural images as possible, then compared its synthetic image to its top “preferred” natural images. Second, we showed the synthetic image to a pre-trained neural network, along with another 100,000 natural images, and then we queried the neural network to indicate which natural images were closest to the synthetic image. Both approaches revealed that neurons in monkey’s brains frequently led to the synthesis of patterns present in pictures of monkeys, consistent with our understanding of monkeys as highly social animals. Thanks to machine learning, we are finally solving the 70-year-old problem of visual selectivity in the primate brain. Currently we are working to identify the full diversity of patterns represented in the macaque brain, extending from Hubel and Wiesel’s primary visual cortex to the end of the visual recognition pathway at inferotemporal cortex. Finding this full set of patterns – what some scientists have called *the visual alphabet* – will allow us to train deep network models that better resemble the primate brain.