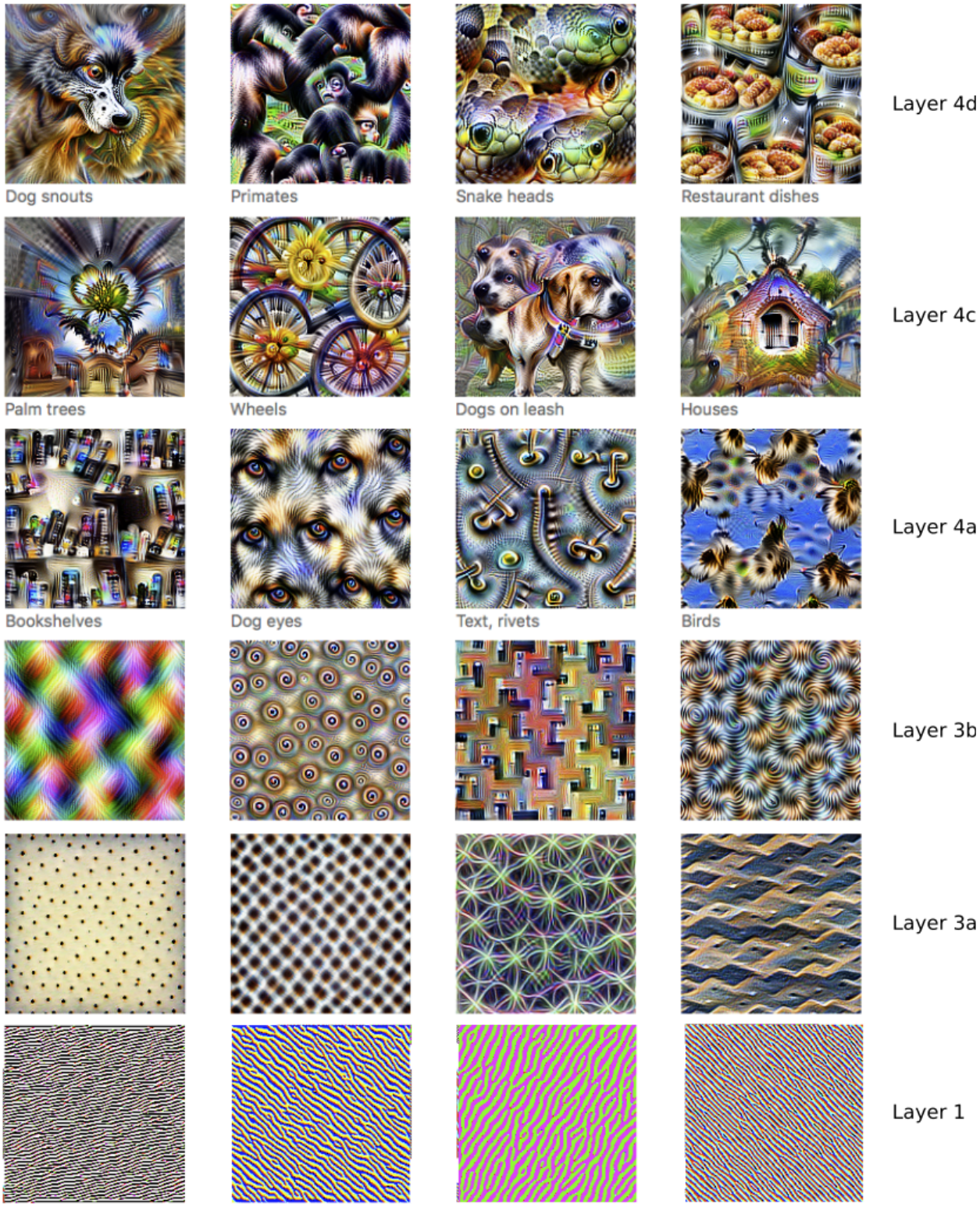
Magic Numbers

How can creativity, emotion, a view of the world, reside in a list of numbers? Those of us employed in the 21st Century showbusiness of making computers sing and dance call such lists, that stream through our algorithms at impossible speed, by various names: representations, embeddings, feature maps, latent variables, projections, codes. Geoffrey Hinton, as skilled at turning emotive phrases as he is at dreaming up new forms of neural network, has enlivened our mathematics with the term “thought vectors”. But they are still lists of numbers.

Yet this conjuring trick, thought-rabbit from number-hat, should not be so mystifying to those of you involved in the more venerable stagecraft of making words laugh and cry. We are, as Nabokov reminded us “absurdly accustomed to the miracle of a few written signs being able to contain immortal imagery, involutions of thought, new worlds with live people, speaking, weeping, laughing.” To a neural network such signs—whether Latin letters or Chinese logographs—are number lists too, with the specific property that all the numbers are zero apart from a single, uniquely placed one; a structure that identifies them, in network-ese, as *discrete*.

The dichotomy between the continuous and the discrete, between things that can be continuously modified into one another, and those that remain stubbornly distinct, enforcing difficult either-or decisions, continues to exercise the scientists and engineers of AI as it has philosophers for millenia—a point we will return to later. For now consider that when a neural network “reads” a text (when translating a web-page from one language to another, for example) what it sees is in essence no different from what a human reader sees: a series of arbitrary symbols with no inherent significance beyond their distinctness from one another. The “meaning” of these symbols, as we all know, resides in a prior pact between the writer and the reader, sender and receiver, to identify certain symbol sequences with certain objects, colours, states of mind etc. How could such a pact be made with a neural network that is born, so to speak, with a scrambled mind (with randomised parameters, to be a little less poetic) and no linguistic or sensory apparatus with which to build links from signifier to signified? This is exactly the quandary that Wittgenstein faced when he considered the origin of language: how can a word be defined without words? His solution—at least the cartoon of it that has reached us philosophobic scientists—is that words don’t need to be defined at all. They simply need to have a utility, a goal that can only be achieved by the transmission of certain information. The syntax and vocabulary required for that transmission then evolves spontaneously from a “language game” played by A and B (uniglyph neonates who would later grow up to be the Alice and Bob of communication theory). Where Wittgenstein’s house-builders learned the language of blocks and beams, of picking up and setting down, a neural network—whom we may soon consider our labouring class—might be tasked with classifying an image. The image is presented to the network, of course, as a list of numbers (in this case a two-dimensional list, an array). The array is then iteratively transformed by a succession of processing *layers* within the network (the “deep” in “deep learning”), each receiving a description of the image from the layer below, and passing on a rephrased description to the layer above. A hierarchical vocabulary emerges to fit the requirements the classification game, with visual primitives such as edges and textures at the bottom, and a composite pictography at the top, sufficiently expressive to distinguish cats from dogs, friends from strangers, pornography from adverts for shoes.



The visual vocabulary of a convolutional neural network. For each layer of the network, images are generated that maximally activate particular neurons. The response of these neurons to other images can then be interpreted as the presence or absence of visual "words": textures, bookshelves, dog snouts, birds — *Feature Visualization, Olah et. al. (2017)*

The striking similarity between these visual “words” and the activity patterns present in the human visual cortex has been repeatedly noted. Indeed a wish to emulate cortical behaviour was a prime motivator for Yann LeCun and others to develop *convolutional neural networks*, now the engine behind image search, face recognition, and everything else that suddenly gave us the spooky sense that our computers can see. Tellingly, a different terminology is used for picture words from the more inscrutable representations found in networks processing text or other symbols: these are “feature maps”, not thought vectors. The connotation is that of signal processing, of low-level, unconscious, feedforward reflexes, as opposed to high-level, self-conscious, recursive symbol manipulation. But the underlying object is still a list of numbers.

So *how* does this vocabulary emerge? It isn’t sufficient for the task to demand them, there must be a path from random patterns to useful representations—much as natural selection requires both the favouring of certain genetic traits and a mechanism for genetic change. This path exists because neural networks are “universal function approximators”: *any* input (image, sound, text, or other) can be transformed into any representation, given the right network parameters. These parameters, known as “weights” for their affinity to synaptic strengths in the human brain, are just another list of numbers. But they have the magical, metalogical property of defining the way one list of numbers—one image, one symbol, one thought—is turned into another. A change in weights is therefore a transmutation of transmutations. By making such changes, by taking steps along what we researchers—with our human weakness for thinking in spaces—sometimes dub the *surface* of the network parameters, we can discover which representations are suited to our problem. Wittgenstein’s game is recast as a *search* through the dizzyingly high-dimensional space of neural languages.

The search is abetted by what we call differentiable objective functions. That is, by tasks whose outcome is not simply success or failure, cat or dog, but a continuum from failure to success: how confident were you that that cat was a dog? This smoothness of merit allows us to follow *gradients*, paths of optimal improvement along the parameter surface, rather than the blind biological route of mutating and surviving, sucking and seeing. Thus the philosophical chestnut of continuous versus discrete is for us a highly practical matter: it could be the difference between running a simulator for hours, or weeks. A good part of the pleasure of deep learning research lies in watching networks flow from one mode of behaviour to another as they trace these gradients. The idea of a plastic deformation from spitting out random words to translating languages, from calling everything a kettle to distinguishing a hundred breeds of dog, from going round in circles to finding treasure in a maze, still enchants me after sixteen years in the field.

But perhaps the simplest of a neural network’s magic numbers is their size, the quantity of neurons and synapses bound together in their virtual web. The deep learning boom that started in the late 2000s and continues to gather momentum today was largely precipitated by an abrupt increase in computational power brought by repurposing the custom hardware of 3D computer games for neural number crunching. Neural nets suddenly got much bigger; in particular they got much deeper, with more processing layers stacked on top of each other. Had this growth spurt led only to quantitative improvements in accuracy and speed, it would have attracted little attention beyond nerdy engineering circles. But there was a qualitative shift in behaviour. Tasks like speech transcription, text translation, image recognition, went from hardly working to working uncannily well. Cognitive scientists were once again intrigued—after the connectionist doldrums of the late 90s and early 2000s—by the possibility of studying their subject *in silico*. Other scientists embraced a new set of instruments to probe their oceans of data. Neural networks were suddenly a tool of great commercial and academic value, and an object of study in their own right. And all because the numbers governing the network dimensions—*hyper-parameters* as they’re sometimes

known, a level of metaphysical remove from the worldly parameters in which learning is ingrained—were suddenly increased.

The idea of intelligence as a bulk property, a mass of thinking stuff that can be shrunk or expanded on demand may strike us as distasteful: a flattening of the miracle of consciousness—the coffee we drank this morning, the chime of its taste with the tastes and impressions, the lantern shadows, of other coffees and other mornings—down to a single neural currency, a fistful of cognitive dollars. But we should remember that this number is simply a container, an expression of the quantity of those other numbers in which the real magic is diffused, and whose vital property is irreducibility. If one network weight can be predicted from the others it does not need to be there. Estimates as to the population of synapses in the human brain vary wildly, but there is no doubt that it dwarfs the number of numbers needed to encode a lifetime's listening and reading. We should not expect, then, a concise description of the contents of our brains—even if we manage to elucidate the principles by which those contents accrued. If artificial neural networks have taught us anything about their biological role models, it's that very complex behaviours can emerge from very simple structures. Just three mathematical primitives suffice to cast all the deep learning spells, the synthetic voices, defeated Go champions, generated celebrities, that we read about in the news: addition, multiplication, and any of a disruptive clan of functions known as nonlinearities, whose *raison d'etre* is to squeeze some numbers together, push others further apart.

To return to the question that opened this essay, if we want to understand *how* the magic of intelligence can be encoded in numbers, we should give up asking *where*. To attempt to locate consciousness in a network of interlinked signals, whether simulated or real, is to fall into the homunculus fallacy, the search for a mind within a mind. Instead we should embrace the quality of quantity, the necessity for life of patterns too intricate to put into words, and patterns of patterns, and so on up. The reductive drift of Western thought has tended to leave the celebration of complexity to artists and poets: Hopkins' "pied beauty" or Whitman's "I contain multitudes". Which builds a somewhat wobbly bridge to the topic of drawing. Tracing the structures we find in our minds, committing them to a white page or a cave wall, remains as good a way as any of grappling with the inexhaustible richness of experience and thought. And it seems to me that it will continue to do so even if we are surpassed by artificial or augmented brains. After all, they will have better things to draw.

Alex Graves is an artificial intelligence researcher. His contributions to the field include connectionist temporal classification (widely used for commercial speech and handwriting recognition), neural Turing machines and the closely related differentiable neural computer. He is the author of the textbook *Supervised Sequence Labelling with Recurrent Neural Networks* (2014).