Compositionality

Stuart Geman
Division of Applied Mathematics

We know how to build machines that learn. In fact, we know how to build machines that learn to perform *optimally* in any motor or sensory task. This is part of the theory of statistical inference. The theory includes *constructive* theorems about learning from example---in effect blueprints for building devices that turn a sequence of examples into a sequence of better and better solutions. Some of these devices are called "artificial neural networks" because they are inspired by what their proponents guess to be the workings of the brain.

Consider a vision task, such as the classification of scenes as either including or not including a tree. A neural network (actually, a sequence of neural networks) can be devised that will "learn from examples" to make *optimal* decisions in this task. This is supervised learning: starting from a sequence of pre-classified ("tree," "no tree") and suitably representative scenes, the parameters of the network are incrementally adjusted in such a way that network performance converges to optimal performance. In particular, the network will eventually classify scenes at least as well as people. This is called nonparametric, or model-free, inference, and it applies to more than just detecting trees: optimal performance is learnable in any task, even in artificial tasks that could not be solved by humans.

Real brains are much less versatile, but much better at just about every real-world task that one can think of. What is most surprising is their speed---not their processing speed, which is not so remarkable by the standards of today's computing machines, but their learning speed. Artificial neural networks never get very good at hard tasks because they never get enough examples. In fact, for most real-world problems it would be infeasible to include enough "suitably representative" examples for the theory of nonparametric inference to be of practical value. It is unfortunate, then, that nonparametric inference is mostly an *asymptotic* theory, about what happens when "N" (the number of samples) goes to infinity.

In contrast, children often learn to recognize letters and other objects from just a few examples. *When ready,* a child can learn to recognize an upper case 'E' from a single example. Just one example, and the child is ready to find the letter in complex scenes, regardless of color, contrast, texture, style, size, or orientation. There is no computer vision system that can solve this single task as well as a typical four-year-old. Forget the hype: when it comes to vision, there is nothing like the real thing.

Consider that every presentation of an object, such as a tree in a natural scene, is essentially unique, from instance to instance, characterized by, among other things, a particular shading or texturing, a particular conformation of the constituents, and a particular pose. How then do biological vision systems construct object recognition algorithms from just a few examples? Could it be, as my brother Donald Geman has argued, that the interesting limit is not “N goes to infinity,” but rather “N goes to zero”? This would pretty much rule out the theory of nonparametric inference, which, as I have said, is mostly about asymptotics (large N).

On the other hand, it is easy to construct artificial classification tasks that are pretty much unlearnable by humans, at least in a reasonable amount of time. On these tasks computers may do better. These are classification tasks that lack structure: there are no familiar parts to build from. Apparently, the real world is more regular. And apparently, regularities are anticipated and exploited by real brains.

Insight into these regularities may be gained from the cognitive sciences, through a theory known as c*ompositionality.* Compositionality refers to our evident ability to construct hierarchical representations, whereby constituents are used and reused in an essentially infinite variety of relational compositions. An upper case E consists of four suitably arranged and suitably sized line segments. The line segments are themselves nameable and useable objects, and in fact could be used in any of a nearly infinite variety of other compositions. And the E is a reusable part as well---a component of any one of tens of thousands of words. More primitive building blocks for vision include intensity gradients and other discontinuities, which, if properly aligned, make up the boundary segments of lines and curves, which are, in turn, the constituents of line segments and other strokes.

It certainly seems as though cognition is organized around hierarchical representation. And it seems as well as though learning is incremental in the sense that compositions are constructed from familiar constituents. It is easy to see how children might learn upper-case letters if they already recognize the building blocks---perhaps line segments and other forms of strokes. The child is ready to learn an E when these constituents are already recognizable. The incremental task is small: learn the allowed relationships among the poses of the constituents.

Fine motor skills, on the other hand, are learned slowly. But the argument for hierarchy, building from pieces that are already well learned and reusable, still seems right. As our colleague at Brown in Neuroscience, John Donoghue, points out, learning to type is a matter of going from an isolated execution of individual letters, gradually to a single, smooth, integrated execution of entire words. It is hard to imagine that the neural program for typing ‘THE’ does not build upon---employ as some kind of subroutines---the existing programs for ‘T’, ‘H’, and ‘E’. Of course there are critical issues of feedback, timing, and control. Indeed, so much more seems to be involved than in a typical perceptual task. Yet the arguments for hierarchy and incremental learning are the same. How else could such complexity and finely tuned action emerge?

Mathematically, the difficulty of a learning task can be quantified through a “dimensionality” or “complexity” measure. This complexity relates to the number of samples or trials needed to achieve any given level of precision. Complexity sets the speed limit for learning. I have argued that hierarchy decomposes the learning task into manageable, low-complexity increments. How does this fit with the richness of actions and presentations? After all, the ensemble of executable actions is essentially infinite, as is the ensemble of presentations of even a simple object such an upper-case ‘E’.

The answer is that the complexities in a hierarchical representation are multiplicative, not additive. The richness of a category emerges, one step at a time with each embedded composition, as the number of the instantiations of constituents are multiplied by the number of meaningful arrangements of the constituents into a composite.

It is true, then, that hierarchical systems can represent rich categories. But not all categories can be represented hierarchically. More precisely, not all categories can be represented in a hierarchy with a fixed and bounded number of levels and a fixed and bounded complexity at each level. Roughly speaking, the issue is one of local versus global complexity. Is it true that (humanly) learnable categories can always be represented by hierarchies that are *locally* simple?

My collaborator Elie Bienenstock and I are proceeding under the assumption that what is learnable is what is representable as a hierarchy of more-or-less simple composition rules. This is speculation, but let us suppose that it is right. What can be made of it? Can we, for one thing, build better vision machines, and might we also be able to make statements, or testable predictions, about representations and computations in real nervous systems?

Our approach is to begin with a mathematical formulation. The mathematical study of compositionality started with Chomsky, and that is where we start---with formal grammars very much of the type that Chomsky proposed for natural language. But the grammars we construct are intended primarily for modeling and analyzing scenes. Composition rules specify sets of constituents (such as line segments) and their relationships (such as their relative positions) that can be allowably combined into a well-formed composition. Of course not all relationships are geometric. In general, composition rules will appeal to other attributes, such as color, style, shape, and texture. Recursive application of the composition rules to the most primitive constituents (local features, such as discontinuities in image intensity) defines the set of well-formed compositions. This is the repertoire of recognizable and nameable objects.

The linguist constructing a grammar for natural language faces a dilemma: the richer the grammar the more ambiguously it parses a typical sentence. Linguists have discovered that any grammar rich enough to accommodate natural language will be highly ambiguous for all but the most simple of sentences. Often, thousands of parses are grammatical for a sentence of only modest complexity. Something very much like this happens in vision. Given the raw evidence, which is to say the pixel intensities, there are an amazing and bewildering number of “correct” compositional hierarchies consistent with the data and with any reasonable set of composition rules.

This is where statistics and the theory of inference come in: all but a few parses are statistically unlikely. Most involve one or more unusual compositions. What is needed is a distribution, one for every composition rule, that assigns a likelihood to every allowable combination of constituents. And this brings us back to our central hypothesis: that learning such a distribution is entirely manageable. Although a composition rule may reference any attribute of the constituents (which puts us outside of the class of grammars that Chomsky called “context-free”), the claim is that the rules can be devised in such a way that any one rule makes reference to only one or a few attributes, and that the resulting inference problem is of low complexity.

There are many issues to attend to. Above all there is the question of scope: what are the categories that admit representation with a bounded local complexity and bounded hierarchical depth? Do these include the everyday categories of our perceptual and motor experience? In short, is our “working hypothesis” sound?

And there is the very practical issue of computation. How hard is it to compute the best (most probable) parse? Provably, very hard by standard computation theory, at least in the case of vision. But maybe we are asking for too much. We are currently studying something more modest---a multi-resolution approach in which the best coarse-scale parses emerge quickly. Detail comes later. A line is seen quickly and coarsely as a line, but the details of its width, intensity, straightness, position and so on are seen more slowly. Composition rules are constructed simultaneously and invariantly at each of multiple scales. In principle, the process eventually terminates with an optimal parse, regardless of scale; in practice, computing never ends. “The more you look the more you see.” Maybe it is a mistake to think about “the best parse” or “the optimal interpretation” of a scene. We should perhaps think instead about a process that is finite only in theory. The discernable detail and compositional complexity of even a single scene is nearly infinite.

What about biological vision? What are the principles of computation and representation in real brains? New methods allow neurophysiologists to monitor dozens of individual neurons simultaneously as an animal performs a perceptual or motor task. Composition theory makes certain predictions about the nature of representations of compositions in terms of the representations of their constituents. But testing these predictions is hard. For one thing, by the standards of conventional statistics, multi-neuronal recordings are extremely high dimensional. This introduces theoretical as well as computational problems. Furthermore, there is very little agreed upon theory to guide the search for meaningful patterns and associations.

But there are some hints---hints from a number of laboratories that point towards part-whole type representations. The interpretation is controversial, but experiments suggest that compositions might have a statistical signature. Specifically, they suggest that the neural representations of constituents are statistically dependent exactly when they are bound into a composition. Whatever the final conclusion about these recent experiments, the new recording technologies raise the hope that we might now begin to meaningfully connect theories about patterns and regular structure in the world with theories about representation and computation in the nervous system.