A mathematical theory of misunderstanding: Energy efficiency considerations in communication

Abstract

New theoretical work in an old field shows that energy can be saved by foregoing precision and error checking in many important cases. For example, it is possible to derive theoretical results to justify what Google has done in practice with their TensorFlow architecture. This has implications for the operation of neural network architectures, but also for mobile devices and other energy-constrained systems. It is also applicable to biological systems, such as molecular genomics and neuronal signaling.

THE TOOLS of information theory have been around for 70 years, and have changed our world—for better and worse. They have been used to make electronic communication systems work, and are one of the spectacular engineering advances of the 20th century. Thanks to them, email, Youtube, and Facebook work—also for better and worse.

Those same tools of information theory can be used in a different way, to look at miscommunication and the energy use involved in making communication and computation work. In many cases, communication can be construed as taking place at several levels: electronic signals are interpreted as bits, then stuck together to make bytes, words, sentences, and so on. Phonological and visual processing work in a similar fashion, with signals climbing tiers of functionally similar syntactic operations feeding the processors of the next higher tier. New work by Tom Sgouros at the Brown University Computer Science Department,[1] looks at the choices a designer must make when deciding where to spend energy on error correction. While it is not universally true, for a large number of situations, there are sound theoretical reasons one can save energy by delaying or even omitting error correction in such a multi-level system.

A little more: Error correction in multiple levels

A MESSAGE is received by a multi-level system, and as it is passed up the levels, each level assumes a certain conditionality about the information it receives. To put bits together into bytes, one assumes that they come in sets of eight, for example. To put letters into words, one assumes they are likely to come from a certain dictionary. In each case, the receiver is adding assumptions about the data stream it receives. In information theoretical terms, these receivers are assuming the information received is conditional on factors invisible to the senders, such as letter frequency according to word position, or correlations between collections of bytes. A sender of bytes need have no idea they are to be assembled into words, or images, or anything else, but the receiver does, and this colors how it interprets what it detects.

Conditionality reduces the effective information content in a message, therefore as one travels up the levels, the number of bits that need to be corrected is reduced at each step along the way. Therefore, assuming some base level of correctness that allows the communication to work at all, when given a choice between correcting at a lower level or a higher one, it will often save energy to wait to correct at the higher level. Where the circumstances demand some correction be done at the lower level, an incomplete correction will likely be adequate.

[1] See `https://arxiv.org/abs/1810.08017`

Furthermore, one can construct a multidimensional space to represent an arbitrarily complex message. As one climbs the levels of analysis, the space becomes larger and more complex at the same time the number of bits of information shrinks. As a result, the space becomes very sparse, with lots of conceptual distance between appropriate interpretations. As an example, consider that the distinction between an 'n' and an 'm' sound is fairly subtle. Distinguishing between "nearly" and "merely" can be challenging because there are usually few contextual clues to help. In the conceptual space of adverbs, these two nodes would be quite near one another, so precision counts. That is, without being sure of the consonant, it is challenging to calibrate the degree of insult intended by calling a colleague's work "[m]erely adequate." By contrast, it is generally much easier to distinguish between "Nat" and "Matt" because in the conceptual space of people you know—a richer space than adverbs—these two nodes might be quite distant from each other. Perhaps you don't even know someone named Nat, in which case, you'll think of Matt whether or not your ear successfully distinguishes the 'n' from the 'm'. In other words, as one climbs the conceptual levels, a first guess at interpreting a message becomes more likely to be correct, and the need for error correction is reduced, even in the presence of noise. In other words, not only are there fewer bits of information to correct at the higher levels, but correction may not be necessary at all.

Energy costs of computing

ERROR CORRECTION has two components: finding errors and fixing them. One can assign an energy cost to each, depending on general parameters relating to the kind of checking and correcting. For example, it usually costs less energy to correct a message in place than to request a retransmission. One can thus write an expression for the energy spent checking errors at two levels. If one assumes that errors are corrected to some known rate, then it is straightforward to find a minimum for this energy expenditure and equally clear that in most cases, it makes sense to skimp on error correction at level one—if not skip it entirely—in favor of level two.

This is more or less the opposite of how computers are organized today. Robust error correction typically happens at the circuit level, and the higher levels of communication are implemented accordingly. High-level internet communication protocols, like HTTP, FTP, and SMTP (email) typically assume that the data they control is sent without bit-level errors, though maybe sometimes the connection will drop a block. The consequences are a computing environment where the high-level players assume perfect communication below them. As a result, energy use is large and performance is brittle in the face of noise beyond the hardware design specifications.

In a practical sense, these results have been anticipated for some time. Researchers in "approximate computing" have found several opportunities for energy savings by relaxing various error checking standards. Much has been accomplished, but mostly in an empirical and somewhat *ad hoc* fashion with researchers hypothesizing and trying different strategies according to their areas of expertise. In a similarly practical sense, the value of precision in a deep learning neural network is a target of suspicion. Google engineered their TensorFlow architecture to use low-precision arithmetic and it works as well as before, at far lower energy cost. Others have reported similar results for neural nets for portable and automotive sensors. Such energy savings are a readily predictable consequence of this analysis.

The limitations of current computer architectures become especially apparent in tasks such as perception and image analysis, where flexibility and good performance in the face of noise are important design goals, but remain elusive. Resilience in the face of errors is an important goal, for saving energy, but also for better performance in navigating a complex world. Algorithms that can reliably extract sense from noise are also able to extract sense from confounding data and malicious attacks.

Quantum computing

DESPITE the existence of actual machines making computations today, quantum computing remains almost as much chimera as practical reality. Small capacity machines exist, but how easily they can be scaled up remains an open question. Error correction is an important preoccupation of the field, for obvious reasons. Quantum events are described with probabilities; there is always a chance of incorrect results. Unfortunately, error correction without corruption of the quantum state that embodies a particular computation is a substantial theoretical and practical challenge. Many of the available solutions seem to solve the problem almost at the expense of the original promise of this kind of computing, locking down states that ought to remain fluid and multiplying qubits to create artificial stability. These measures drastically increase the size and reduce the speed and efficiency of the resulting machines.

An alternative view is that it might be possible to build a computing stack from the ground up that attempts to produce reliable results from very unreliable quantum components, to create applications that use multiple layers of analysis, where one can delay or skip error correction. Results exist to suggest that it is a potentially viable approach to establish a second level of analysis to overcome errors at the first level in a quantum system.Such a stack may not be a general-purpose computer as we have come to know them, but there are a host of computational problems that might be addressed by such a machine. Simulations of natural processes, machine learning, image analysis, and Bayesian networks are all important problem spaces addressed with multiple layers and thus potentially susceptible to this approach.

Saving information

THE STATEMENT that one should delay error correction when possible is of only limited value without also being able to identify when it is possible. Accompanying work in the theory of information related to multiple sources makes it easier to reliably identify signals whose errors can be neglected in the context of some large analysis. Again, in a system of multiple levels, assumptions about conditionality differ from one level to another. Analyses about which inputs to a calculation “matter” are not exactly subjective, but they can differ according to the assumptions relative to the level. At one level, an input that appears to be well correlated with a distant computational output can seem quite important to that result and thus worthy of close attention to its errors. But if the view from the level of that result reveals the input to be only one among dozens of nearly identical signals, then it becomes instead a candidate for this form of benign neglect. One can think of a pixel of an image where a tan color is well correlated with the identification within the image of a camel. This does not imply that pixel is an important contributor to the identification, but a naïve application of information theory might suggest exactly that. As a result, such tools are not generally used for the analysis of those calculations.

Borrowing ideas from quantum information theory, it is possible to formulate an understanding of classical multi-variable information theory that facilitates just such analyses, making it clear where accuracy is worth extra effort. It also reveals that the manifestations of information theory in the two fields are not so completely distant from one another. In some aspect, quantum information can be understood as a special case of a more general formulation of classical information theory, though this remains to be worked out in more detail.

Agenda

THE WORK described here has been submitted to journals, but there is much work still to do, developing the details and ramifications. The way to advance theoretical ideas at this stage is to find concrete practical problems to which they can be applied. Through experience, the theory will be refined and advanced. This is already underway through an analysis of some molecular genomic pathways and with a neuronal signaling lab. We are looking for opportunities in industry where energy or time savings might be found by redistributing the responsibility of error correction, re-engineering applications to be more resilient, or simply reducing the information content of important messages.

As noted above, there seem to be important possibilities worth exploration in quantum computation. Beyond that, the likeliest fields to make advances using these observations are in mobile communication and robotics, where power consumption (i.e. battery life) is a vital consideration, and in the operation of large neural network applications. There are also important potential applications in information security, since awareness of the appropriate range of some source of input data is an essential part of being able to handle incorrect or malicious transmissions. There are likely to be to be many other applications of this analysis. After all, saving energy, saving time, and engineering resilience are important goals for virtually all of the tech industry.