Cognitive Computation
https://doi.org/10.1007/s12559-023-10194-9

What AI, Neuroscience, and Cognitive Science Can Learn from Each Other: An Embedded Perspective

Tsvi Achler[1]

Received: 18 July 2022 / Revised: 9 August 2023 / Accepted: 12 August 2023

Abstract

Scientists studying in the fields of AI and neuroscience can learn much from each other, but unfortunately, since about the 1950s, it has been mostly one-sided: neuroscientists have learned from AI, but less so the other way. I argue this is holding back both brain understanding and progress in AI. Current AI ("neural network"/deep learning algorithms) and the brain are very different from each other. The brain does not seem to use trial-and-error–type learning algorithms such as backpropagation to modify weights and more importantly does not require the cumbersome rehearsal needed for trial-and-error implementation. The brain can learn information in a modular and true "one-shot" fashion as the information is encountered while the AI cannot. Instead of backpropagation and rehearsal, there is evidence that the brain regulates its inputs during recognition using regulatory feedback: form the outputs back to inputs—the same inputs that activate the outputs. This is observed through evidence from the fields of neuroscience and cognitive psychology but is not present in current algorithms. Thus, the brain provides an abundance of evidence about its underlying algorithms and while computer science tools and analysis are essential, algorithms guided by computer science should not be standardized into neuroscience theories.

Keywords Catastrophic forgetting · Independent and identically distributed · Regulatory feedback · Salience · Biased competition · Rehearsing

Background/Introduction

There are three predominant research fields that tie into brain science: *computer science* which for this discussion includes mathematics and applied AI, *neuroscience* which includes neuroanatomy and neurophysiology, and *cognitive psychology* which includes studying reaction times and error analysis of stimuli responses.

Using neuroscience as the only focus of brain science, it is difficult to reveal how the brain works. This is because the brain is very dense and currently tools are not available to record in detail millions of neurons at once. Although methods such as fMRI and scalp electrodes (e.g., EEG) record the contributions of large numbers of neurons, they are smeared together obfuscating specific circuits. More focal recordings such as single cell electrophysiology may record single neurons but not how the (potentially multiple of thousands) neurons that connect to the recorded neuron are contributing to the neuron's responses.

Likewise, cognitive science gives important clues on what affects processing but by itself also does not reveal the underlying neural mechanisms.

Thus, given the brain's density and lack of access, computational models need to be created, compared, and evaluated in order to understand which computational structures are practical from applied perspectives.

The field of computer science provides realistic environments, scenarios, and tools necessary to evaluate and benchmark complex models. Large scale AI and computer science models can provide insight and drive model selection; thus, such computational models are essential to assess understanding, research strategies, and guide experiments. To advance forward, computer scientists must incorporate algorithm design clues from brain scientists. Likewise, brain scientists, such as neuroscientists and cognitive psychologists (who could do better working with each other) must evaluate algorithm practicality using tools of computer scientists. Certain hypothetical configurations may seem like a good idea, work great as a small model of a limited

✉ Tsvi Achler
achler@optimizingmind.com

[1] Optimizing Mind Inc, Palo Alto, CA 94306, USA

Published online: 01 November 2023

experiment, but not scale in complexity. Thus, multiple perspectives in addition to appropriate integration are essential to study the brain.

However, the current trends in academia and funding seem to be very computer science centric and that seems to cause problems.

Problems from the Computer Science Perspective: Rehearsing

Old movies, cartoons, and science fiction depict robots interacting naturally in our environment; however, this has not panned out in reality. Although very capable robots have been created that dance, perform acrobatics, and gallop, they still do not interact naturally. Today's learning and recognition algorithms (which include unfortunately named "neural networks") may be powerful but they are deceivingly inflexible.

The inflexibility limits automated machine and robot application in real life environments safely, and even sometimes misleads experts: when are truly self-driving cars due? Training is arduous and requires big data and issues arise in real natural environments.

Currently, everything that may be in the real environment must be predicted during learning. Predicting the real environment is tricky. Moreover, if something is missing, new data must be captured and weights must be rerun and retrained from scratch. Thus, training for even semi-realistic natural environment interactions are beyond reach due to huge downtimes.

These difficulties ultimately exist because current algorithms require rehearsal that is unnatural and unintuitive.

In order to successfully implement trial-and-error backpropagation, the data needs to be rehearsed in fixed frequency and random order. This rehearsal is described in mathematical terms as establishing *i*ndependent and *i*dententically *d*istributed presentation order or *iid* rehearsal for short.

Implementations of *iid* rehearsal require mechanisms to:

1. Store all seen patterns for rehearsal.
2. Keep track of rehearsed presentations.
3. Retrieve stored patterns in random order and fixed frequencies.
4. Generate error signals predicted in backpropagation.
5. The brain would need to rehearse at extremely fast rates when it encounters new information, faster than today's computers in order to update the huge amount of data the brain stores.

The rehearsal paradigm is an attempt to solve an underlying problem with these models that is called catastrophic interference or forgetting [1–4]. Catastrophic forgetting describes how information fades away during learning. Rehearsal is a stopgap measure to manage this fading by periodically "reminding" the system with randomly selected patterns (because if the patterns are not random, something else may be forgotten). Beyond using rehearsal, this has not been solved satisfactorily [5].

Yet even if it is possible for the brain to rehearse patterns during learning using the *iid* criterion, this paradigm makes it difficult to quickly incorporate new information into the network as the organism encounters it. Any time a new piece of information needs to be learned, the new information must be trained with old information otherwise old information will be lost due to catastrophic forgetting.

This rehearsal is akin to a casino dealer shuffling training examples of everything we previously learned for anything new we want to learn. Imagine spending a summer in Hawaii, and by not rehearsing other environments (e.g., winter scenes and desert scenes) forgetting how to recognize other scenes.

These rehearsal difficulties occur even if the underlying models use what may appear as "natural" supervision through reinforcement learning and rapid reinforcement learning using inductive biases [6–8].

Some researchers continue to argue that backpropagation and rehearsal are possible in the brain (e.g., [1, 9]). Others attribute any rehearsal found in the brain that consolidates sequences of information such as paths in a maze or occurring uniquely in the hippocampus [10–15] as evidence for backpropagation and infrastructure for the unplausible amount rehearsal and *iid* shuffling. Despite the proposal of these models over 70 years ago, decisive evidence of backpropagation, the rehearsing needed in *iid* form, underlying shuffling and storing mechanisms, remain to be found.

There have been efforts in the computer science field to address this updatability problem primarily by doing transfer learning. For example, if all the layers of a deep network are trained on broad images found in the visual environment, it can be assumed that those will remain approximately the same. Thus, the bottom layers are fixed and only the top layer is learned, reducing the amount of learning. While reducing the problem, it does not solve it because if the top layer grows large (e.g., there are more than a handful say 10, 100, 1000 or 10,000 output classes) then the rehearsal issues remain.

The core issue is the ability of the learning algorithm and network to learn true "one-shot" learning and be modular: to have the ability to add new information to one labeled class without affecting the information of other classes.

Concerted efforts trying to address "one-shot" learning by: transfer learning, adding extra information and using alternate algorithms on the top layer [16] have not been very successful [17].

Thus, at their core backpropagation models and *iid* rehearsal are not optimal brain models because they do not display modularity.

Difficulties Tying the Brain Sciences Together

Subsequently, just about every neuroscientist, cognitive psychologist, and most computer scientists will agree that very little is currently understood computationally how the brain recognizes information from its environment and that neurons of the brain are more connected and self-regulated than current computational models. Thus, it is very important for all these fields to learn from each other.

However, insights from AI, neuroscience, and also cognitive psychology are poorly integrated together as each field has its own priorities, funding, terms, and metrics of evaluation. Many terms are defined differently in different fields making cross disciplinary study difficult.

Confusion Around Recurrence

For example, the term recurrent connections have different meanings in the computational neuroscience, computer science, and cognitive psychology communities.

In computer science, recurrent means an output used back as an input within a paradigm of delayed inputs. It is a method of representing time or sequences and also called LSTM networks. Unfortunately, recurrent connections in such neural networks are often confused with feedback back to the same inputs. Feedback back to the same inputs is actually never used in neural networks, because it forms an infinite loop and is not possible to rewind in order to generate an error signal through backpropagation.

Thus, recurrent feedback is not the same (and cannot be the same) as the output-modifying-its-own-input type of feedback, because such feedback blocks backpropagation, a key component of training these algorithms.

In cognitive psychology, recurrent connections mean lateral connections to neighbor neurons, such as the connections of lateral inhibition in neuroscience. In cognitive psychology, the term re-entrant connections are used to describe (regulatory) feedback back to the same inputs (Fig. 1).

So, to truly appreciate and integrate ideas, members of brain-focused fields need to be careful with and familiarize themselves with the terms and key ideas developed in other fields. This familiarity not only covers language but also concepts.

Confusion Around Salience and the Cognitive Phenomena of Salience

Another example is the multiple-defined term salience which is often overlooked but covers an essential guiding principle. Salience is a dynamic process which determines how well certain inputs are processed or perceived. Salience changes in the brain depending on what other inputs or features are concurrently present or what the person is instructed to focus on. Salience can change based on the patterns the person knows and the interaction between patterns that are recognized, so it is an integral part of processing. In cognitive psychology, there is a rich literature on salience, which is (again) different from salience in the computer science community.

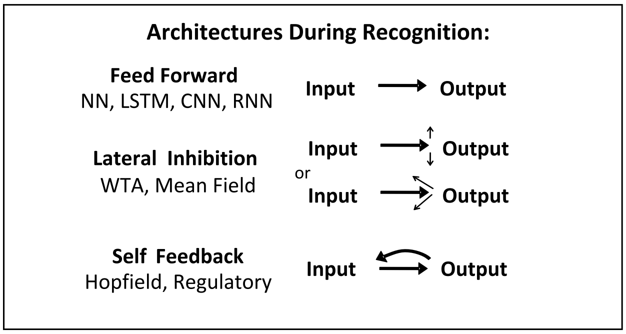

Fig. 1 Types of connections, architectures and terms used during recognition. In Feedforward architectures, information goes from inputs to outputs. In the case of sequential processing (LSTM, RNN), time is designated by a delay but can be unwound to a feedforward structure so backpropagation can be applied. In Lateral Inhibition, neurons inhibit their neighbors either directly or indirectly via neighbors' inputs, or diffuse inhibition. In Self-Feedback neurons affect *their own* inputs. This can be broken down to either positive feedback (e.g., Hopfield networks) or negative feedback (regulatory feedback)

Salience is an integral part of processing, suggesting it is not a simple learned spatial filter as hypothesized in current computer science–based neural network literatures. Subsequently, in cognitive science, salience is as much more fundamental property of recognition than the version of salience popular within the computer science literature (e.g., [18]).

Salience in cognitive science is associated with a signal-to-noise ratio during processing [19] which can be measured by the speed of processing or speed accuracy tradeoff given different inputs. These effects of salience can be measured both in spatial processing and by reaction times and errors in humans given fast stimuli.

It occurs automatically from a bottom up (through input interactions), which is a source of "pop-out" and difficulty with similarity [20, 21]. This can occur much before there is a chance to select a focus on spatial region. For example, it occurs when the display is too fast for spatial attention in fast-masking experiments but shows the similar interactions as spatial attention [22]. Salience seems to be generated "on-the-fly" as an inseparable part of recognition mechanisms. For example, a certain pattern may be salient in one context but not another.

Salience is also integrated into non-spatial modalities which are observed in non-visual modalities with poor spatial resolution such as olfaction, smell [23], and touch.

Thus, an account of salience must be integral to any computation account of recognition models. Subsequently, current AI-based models of the brain, even if they have the computer science term "salience," lack the cognitive version and concept of salience. Computer science salience is not the integrated mechanism seen in the brain, which is spatial and does not address salience based on dynamic state of mind of the perceiver (focus, goal, memory, etc.). A dynamic version of salience is not possible within the computer science version because deep and neural networks cannot change easily unless they trained with rehearsal, which would take too long in a real life–like environment.

There are models in computer science such as Bayesian networks that are more changeable and subsequently more brain-like but they are not a connectionist. They do not describe individual connections between neurons. The connections are more abstract and statistical; the connections are described as statistical likelihoods. Moreover, such models are not as scalable to large networks. Thus, while Bayesian networks are important conceptually, they are not a viable model of the brain either. Neural networks are more realistic than Bayesian networks because they are a connectionist, but neither of their learning paradigms are realistic. Thus, it is important to look for more models that are a connectionist but incorporate more feedback.

Two more guiding principles often overlooked in computer science but present in neuroscience and display ubiquitous evidence of feedback are excitation-inhibition balance and homeostatic plasticity.

Excitation-Inhibition Balance

If neurons of a brain are at a resting state and are presented with a new or unexpected stimulus, the neurons show a fast peak response to the stimulus followed by a slower change to a subsequent new steady state response. This is called excitation-inhibition balance (e.g., [24–27]).

Brain activation increases quickly then decreases slowly towards baseline, seen in variety of organisms, sensory modalities, isolated dissections, and neurons grown in dish. Excitation is balanced by inhibition through some sort of feedback during recognition. Excitation-inhibition balance is usually modeled with direct excitatory connections and diffuses poorly defined inhibition back. The feedback responsible is poorly understood and begs for a good scalable model that also reveals the feedback's mathematical and computational role.

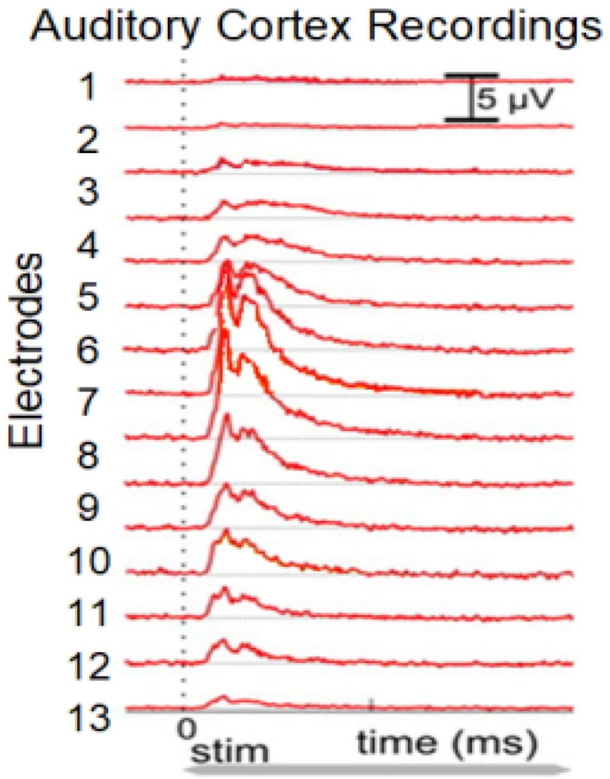

Fig. 2 Example of neuron recordings from auditory cortex and surrounding regions in response to sound stimuli (modified from [28])

Brain recordings commonly show a network-wide bursting with novel patterns where many neurons initially activate and the system slowly quiets down which is not present in "neural networks" models of recognition (Fig. 2).

Homeostatic Plasticity

Homeostatic plasticity refers to a form of neural plasticity that maintains the stability and balance of neuronal activity within a neural network. It is a fundamental mechanism that allows the nervous system to regulate its own activity and maintain optimal functioning. The key principle behind homeostatic plasticity is that it aims to maintain the overall activity level of a neuron or a network within a certain target range, often referred to as the set point. If neuronal activity becomes too low or too high, homeostatic mechanisms kick in to restore the activity back to the desired range. Classic neuroscience studies of the likes of Marder and Turrigiano reveal homeostatic plasticity the ability of connected pre-synaptic and post-synaptic (input and output) neurons to regulate their firing rates using multiple mechanisms such that activation and response tend towards a fixed setpoint. Homeostatic plasticity suggests every neuron connection is regulated, and the multiple mechanisms involved suggests this is a key configuration of the brain.

Thus, neurons do not exist in isolation instead are in a delicate excitation-inhibition balance that is found throughout the brain [24–27]. Neurons are regulated through presynaptic processes governing homeostatic plasticity [29–31]. This means inputs (neurons) are adjusted and closely regulated based on their connected outputs (neurons). In studies on homeostatic plasticity, this is revealed by making, for example, output neurons fire stronger. What is found is that over time input neurons fire less. They are regulated in a way where the pre-synaptic or input neuron must know about the post-synaptic or output neuron in order to balance the signaling.

Moreover a concert of internal neuron machinery may interchangeably contribute to this regulation; thus, it must be very important for the neural system.

Feedback directly back to inputs is counterintuitive because the neurons ultimately shut off inputs that activate them, but this can be done gradually over forward-backwards iterations setting up a tug of war situation between neurons that can perform an important role of recognition.

So maybe the core computational unit is not a feedforward "neural network" neuron but a combination of postsynaptic neurons coupled to pre-synaptic neurons in a regulatory way.

Computational Architectures Forming the Basis of Current Modeling

We designate architectures by connection types between neurons that process information (see Fig. 1) during recognition inference (as opposed to during learning). The simplest model is where neurons directly activate neurons downstream (from inputs to outputs) by their connection strength. This is designated as feedforward architectures and is the basis of not only deep networks and machine learning models but also computer science "recurrent" and reinforcement learning models. Computer science "recurrent" connections are considered feedforward because they designate inputs after a time-step. Those time-steps can be rewound in time determining a feedforward structure. These types of feedforward structures can then be trained by backpropagation.

Another type of connection, that is popular in brain architecture models, is lateral inhibition of neighbor neurons which is feedforward combined with side connections where neurons can inhibit other neurons in the same layer. Sometimes the inhibition is to other neuron's inputs but the key is that they are not back to the same inputs that activate them. So this kind of connection (whether directly to a neighbor or a neighbor's input) inhibits a neuron's neighbors.

Lateral inhibition or inhibition of neighbors makes sense in some cases such as low-level vision close to where information comes into the brain. This type of inhibition is one account for center-surround visual sensitivity patterns seen in retina and early experiments in the V1 region of the brain based on anesthetized animals. Not only may these patterns not hold in awake animals, but when considering higher level representations of the brain, where neurons represent more complex patterns, this can lead to a large number of connections and variables that do not make sense.

For example, if there is a neuron representing a square and a neuron representing a diamond, in this model, they would have to be connected inhibitively to each other. But then, a neuron representing a hexagonal pattern would have to connect to those other two, and any other pattern would also need to connect to the others, and so would the next and next and next (Fig. 3). Ultimately, there could be tens of thousands of representations, where a neuron would need to connect to every other representation. To the extreme, this would mean the brain would only be able to represent one idea at a time.

Moreover, what does a lateral connection and its weight really mean in the high level? How much should neurons that represent scissors inhibit the neurons that represent the bell? And does that number really mean anything useful for computation?

Neuroscience modelers that promote the lateral inhibition narrative use what is called mean-field models (e.g., [32, 33]). Mean field models are a class of mathematical models used in neuroscience to describe the collective behavior of large populations of neurons. They simplify the complex interactions between individual neurons and focus

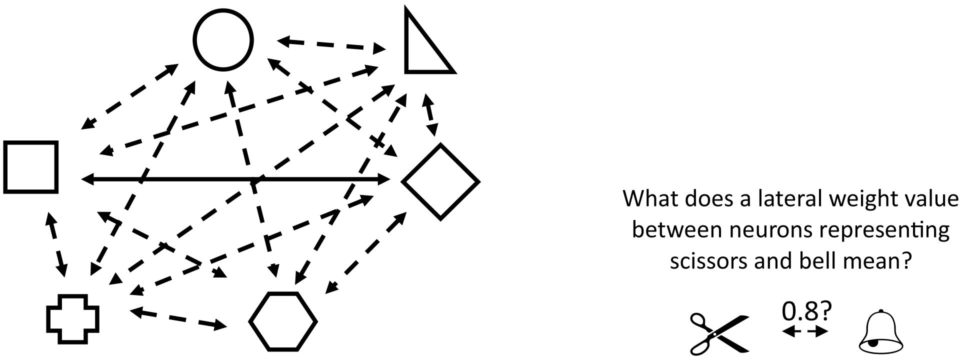
What does a lateral weight value between neurons representing scissors and bell mean?
0.8?

Fig. 3 Left: high level representations would ultimately need to connect to each other with lateral inhibition. Lateral inhibition is a reasonable model for center surround patterns seen in the retina but they make less sense for complex representations (right)

on characterizing the average behavior of the population. These models are based on the assumption that the behavior of individual neurons can be approximated by their average properties and that the interactions between neurons can be effectively described by their mean influences on each other. Mean-field models morph the lateral inhibition problem into another one by assuming that all of the connections are statistical. This causes a problem where these types of models are no longer a connectionist, like Bayesian models, and are just an estimate of how the network might function. Mean-field models not only lose computational functionality, they also do not really solve the problem of determining the significance or relevance of individual lateral connections.

An understanding of these connection architectures is essential to understanding the limits of theories and theoretical frameworks.

Theoretical Frameworks and the Roles of Underlying Architectures

There are several frameworks and theories that describe in words what may seem like feedback to the inputs during recognition but are implemented by the more-limiting connection types.

For example, according to predictive coding theory, the brain generates top-down predictions about the causes of sensory input based on prior knowledge and internal models [34]. These predictions are then compared with the actual sensory input received from the environment. Any discrepancies between the predictions and the actual input are referred to as prediction errors. The brain's goal is to minimize these prediction errors by updating its internal models and generating new predictions. This theory is primarily implemented by feedforward models and backpropagation. Feedforward recognition does not generate error, the backpropagation algorithms do, but they also require *iid* rehearsal to be successful.

Thus, predictive coding is a bit ambiguous because it can potentially describe a process that occurs during recognition (using feedback connections to inputs) or during learning (feedforward connections combined with backprop during learning). Most authors only consider predictive coding for learning, defaulting into standard feedforward models.

A similar theoretical framework is the free energy principle proposed by Friston [35]. It provides a unifying perspective on brain function, perception, and action, and it aims to explain how organisms maintain their internal states in a changing environment. At its core, the free energy principle is based on the idea that the brain's primary goal is to minimize surprise or uncertainty about the world; it is like predictive coding but applied broader. It suggests that the brain constantly generates and updates internal models of the world, referred to as generative models or "predictions." These models represent the brain's beliefs about the causes of sensory input and the consequences of its own actions. It argues that the brain minimizes a quantity called "free energy" which is a measure of the difference between the predicted sensory input generated by the internal models and the actual sensory input received from the environment. Minimizing free energy is equivalent to minimizing surprise and maintaining a coherent understanding of the world. Free energy is extended to active inference where perception and action are closely intertwined. Perception involves actively gathering sensory information that reduces uncertainty or surprise, while action serves to actively shape the sensory input by interacting with the environment and gathering evidence to confirm or update the internal models. The organism's behavior is driven by the desire to minimize surprise and maintain a coherent understanding of the world. By looking at behavior, this framework focuses more on what happens during recognition. The minimization of surprise is analogous to the minimization of error of predictive coding, and this theory is primarily implemented by Bayesian methods as opposed to feedforward models.

Yet another related framework is proposed using Hierarchical Temporal Memory (HTM) [36] in the book "On Intelligence," where Jeff Hawkins argues that the brain operates as a prediction machine, constantly generating and updating models of the world based on sensory input. This theory is primarily implemented by lateral inhibition.

The frameworks expand on how error is corrected and are implemented either with lateral inhibition, Bayesian models or feedforward models combined with lateral inhibition or backpropagation since feedforward models do not generate error in themselves. Thus, they suffer from scalability, connectivity, or *iid* rehearsal limitations and none implement feedback back to the inputs.

These ideas have been expanded upon with additional generative models through Bayesian methods (which are not scalable) or adversarial models such as using AutoAssociative encoder/decoders (some which are supervised and some unsupervised) which can find a solution then generate an approximation of the original input [37–39]. This regeneration allows feedback to inputs, in spirit, in the form of recreating the input, but it does not feedback to pre-synaptic inputs of neurons. Moreover, the fundamental mechanisms used remain backpropagation which suffers from *iid* rehearsal in both directions.

Biased Competition

A cognitive neuroscience framework of biased competition originally proposed by Desimone and Duncan [40] is even more difficult to implement in a scalable manner with feedforward methods or lateral inhibition. Biased competition describes how different neural representations or processes within the brain compete for limited neural resources and influence each other's activation or suppression. It refers to the idea that the processing and representation of sensory information are influenced by the relative strength or salience of competing inputs or stimuli. In the context of perception, biased competition suggests that when multiple sensory stimuli or inputs are present, they compete for processing resources, and the brain selectively amplifies or enhances the representation of the most relevant or salient stimulus while suppressing the processing of less relevant stimuli. The competition can be biased based on factors such as stimulus intensity, novelty, attentional focus, or task relevance. This biased competition occurs across different levels of neural processing, from sensory areas to higher order cortical regions. For example, in visual processing, neurons throughout visual areas compete for processing resources, and the neurons representing the attended or behaviorally relevant stimulus tend to be selectively enhanced, while the representations of irrelevant stimuli are suppressed. Biased competition is thought to play a crucial role in attentional processes and cognitive control but also recognition. It allows the brain to allocate its limited processing resources selectively and efficiently and is closely integrated with (cognitive) salience, and difficulty with similarity.

Potential Academic Pitfalls from Lack of Scalability and Too Many Parameters

Briefly stepping back from specific frameworks and architectures and looking at general research and modeling, too many parameters in data acquisition and modeling are a scourge. Although the experimental community provides the most important data, general models that the experimental communities prefer (both experimental neuroscience and cognitive psychology) are large overparameterized networks with a large number of parameters that can be used to match data of a narrow experiment perfectly. Thus, large-parameterized models are very popular to fit data. The problem with overparameterization is there may be many parameters that can solve the same problem and choosing between them is arbitrary. This is related to the concept of over-fitting; sometimes described in the computer science literature (but too often ignored in other fields). While neuroscience and cognitive communities can model any of their phenomena with enough parameters, models with huge parameter spaces that are changed for each type of experiment, are less desirable.

Even worse, data processing, analysis and results using too many parameters is vulnerable to human fallibility in the face of academic pressures. Science is based on repeatability and there is a scientific measure of repeatability that is uniformly adopted for experiments. This definition is formally described as a p value 0.05 (two standard deviations) and guided by the mathematics of chance. However, despite these requirements, published academic papers face a repeatability crisis where papers are 60–90% non-repeatable [41, 42] depending on the field. This does not seem to have changed even after the original expository studies [43, 44]. All fields are vulnerable including computer science [45]. One revealing example in cognitive neuroscience is the overparameterized methods used to get signals out of fMRI studies. This has resulted in dubious techniques culminating in the dead fish finding where in a standard experiment, it is possible to substitute a test subject with a dead fish and still get data [46, 47]. Note this exposition was not published officially in the field's literature but is essential to gain insight into underlying problems and insights in the field.

Overparameterization combined with human issues make multidisciplinary approaches even more difficult because it takes being immersed in the field to really understand what studies have been verified correctly.

Thus, from many perspectives, minimizing degrees of freedom (the number of parameters) is essential in research and models. Models that show the most amount of phenomena (e.g., multiple phenomena across multiple disciplines) with the least amount of free parameters are better. However, I have found too often that "we already have a working model" for a limited experiment is a justification for not looking any further regardless of number of parameters.

Returning to architectures of neural modeling, there is one architecture during recognition that is relatively overlooked that can fit many experiments, minimize parameters, and fit multiple frameworks.

Self-Feedback to Inputs

Although found throughout the brain, sensory recognition regions, the thalamus and cortex; the least likely considered connection in the theoretical community is pre-synaptic feedback, also known as top-down feedback (although this is another term that has different meanings in different fields). In this configuration, outputs feed back to their own inputs or as I call it here self-feedback.

So, for example, information will come from the eyes and neurons will make a connection in the thalamus, the thalamus projects to the cortex of the brain, and then the cortex feeds back to the thalamus, which then modulates those same neurons that feed up to the cortex.

Most theorists prefer feedforward and lateral inhibition connections. Part of the reason is that feedback connections are very difficult: they are very difficult to stabilize, analyze mathematically, and understand. The other reason is because of the popularity of feedforward models in computer science "deep neural networks" where tools are readily available.

But that does not change the fact that the brain has feedback connections and that they are important.

The most popular model that uses pre-synaptic feedback to the inputs is Hopfield networks [48]. Confusingly, this literature calls the connections recurrent like in psychology, but they are different than those in computer science that can use backpropagation. The feedback in Hopfield networks is self-reinforcing or borrowing nomenclature from engineering: positive feedback. An activated output activates its own inputs and accelerates its own activation, leading to runaway activation. This model is used as a memory system where degraded inputs can be given and the network completes the rest of the pattern.

Currently, very few popular models utilize negative or regulatory feedback back to inputs. Again, borrowing nomenclature from engineering, this is negative feedback. This type of feedback may be initially counterintuitive as neurons will inhibit their own inputs. However, examples of physical systems using this configuration are actually common and include a simple thermostat-heater configuration where the thermostat will send a signal to stop heating when a threshold is met.

My proposal at its core is an architecture change: that recognition networks are primarily based on pre-synaptic self-inhibition (not feedforward or lateral inhibition). Neuron activity is highly regulated by pre-synaptic feedback activity as evidenced by feedback loops in the brain and evidence for homeostatic plasticity and excitation-inhibition balance [49–53]. Computationally, regulatory feedback models using pre-synaptic inhibition do more work during recognition than feedforward models, but this allows simpler and faster learning without rehearsal or *iid* requirements, avoiding catastrophic forgetting. Moreover, without additional parameters, the architecture inherently produces phenomena such as network-wide bursting when a pattern is initially presented, and cognitive phenomena of salience and difficulty with similarity [49–51]. This is a connectionist method with no statistical assumptions of Mean Field or Bayesian methods.

From a framework perspective, this architecture fits into the frameworks of predictive coding, free energy, minimizing free energy, surprise, and unused inputs. The caveat is that it is achieving these processes during recognition (not during learning). The gradient during recognition occurring via feedback is not finding weights, it is finding activations using the input to output and output back to its pre-synaptic inputs. Through the bidirectional self-excitation inhibition balance, it determines the inputs are best used (with the best match and least duplicity) by the outputs. The bidirectional weights themselves can be learned with simple (original) classic Hebbian Learning [54] as described before that term was co-opted to fit feedforward learning, using "what fires together wires together." This simplifies learning and restrictive rehearsal.

Additionally, with only a parameter for bias (selectively sustained increased activation) of output neurons, regulatory feedback displays biased competition. These multidisciplinary results with one architecture change suggest that the brain is using negative feedback as a primary mechanism for recognition.

But the direction is from computer science to brain science, not from brain science to computer science. Introducing any other model than a feedforward model is an uphill battle, so science is stuck using computer science–based feedforward models for the foreseeable future.

Declarations

Human and Animal Rights and Informed Consent This article does not contain any studies with human participants or animals performed by any of the authors. The author is the founder of Optimizing Mind, Palo Alto, CA.

References

1. McClelland JL, McNaughton BL, O'Reilly RC. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev. 1995;102:419–57.
2. McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. The psychology of learning and motivation. 1989;24:109–65.
3. French RM. Catastrophic forgetting in connectionist networks. In: Encyclopedia of cognitive science, vol. 1. Nature Publishing Group, London; 2003. p 431–5.
4. Coop R, Arel I. Mitigation of catastrophic interference in neural networks using a fixed expansion layer. Midwest Symposium on Circuits and Systems. 2012;726–729.
5. Moe-Helgesen O-M, Stranden H. Catastrophic forgetting in neural networks. 2005. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.622.7385&rep=rep1&type=pdf.
6. Merel J, Botvinick M, Wayne G. Hierarchical motor control in mammals and machines. Nat Commun. 2019;10:5489. https://doi.org/10.1038/s41467-019-13239-6.
7. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Hassabis D. Reinforcement learning, fast and slow. Trends Cogn Sci. 2019;23:408–22. https://doi.org/10.1016/j.tics.2019.02.006.
8. Barretto A, Hou S, Borsa D, Silver D, Precup D. Fast reinforcement learning with generalized policy updates. PNAS. 2020;117:30079–87.
9. Scellier B, Bengio Y. Equilibrium propagation: bridging the gap between energy-based models and backpropagation. Front Comput Neurosci. 2016.
10. Derdikman D, Whitlock JR, Tsao A, Fyhn M, Hafting T, Moser MB. Fragmentation of grid cell maps in a multicompartment environment. Nat Neurosci. 2009;12(10):1325–32.
11. Klinzing JG, Niethard N, Born J. Mechanisms of systems memory consolidation during sleep. Nat Neurosci. 2019;22(10):1598–610.
12. Maingret N, Girardeau G, Todorova R, Goutierre M, Zugaro M. Hippocampo-cortical coupling mediates memory consolidation during sleep. Nat Neurosci. 2016;19(7):959–64.

13. Buzsáki G. Two-stage model of memory trace formation: a role for "noisy" brain states. Neuroscience. 1989;31(3):551–70.
14. Carr MF, Jadhav SP, Frank LM. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nat Neurosci. 2011;14(2):147.
15. Epsztein J. Mental replays enable flexible navigation. Nature. 2022;605:35–6.
16. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350(6266):1332–8.
17. Lake BM, Salakhutdinov R, Tenenbaum JB. The Omniglot challenge: A 3-year progress report. Behav Sci. 2019;335(29):97–104.
18. Koch I, Itti NL, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. V20, No. 11. 1998. p 1254–9.
19. Rosenholtz R. Search asymmetries? What search asymmetries? Percept Psychophys. 2001;63(3):476–89.
20. Duncan J, Humphreys GW. Visual-search and stimulus similarity. Psychol Rev. 1989;96(3):433–58.
21. Wolfe JM. Asymmetries in visual search: an introduction. Percept Psychophys. 2001;63(3):p381–9.
22. Francis G, Cho Y. Effects of temporal integration on the shape of visual backward masking functions. J Exp Psychol Hum Percept Perform. 2008;34:1116–28.
23. Rinberg D, Koulakov A. Gelperin A speed accuracy tradeoff in olfaction. Neuron. 2006;51(3):351–8.
24. Xue M, Atallah BV, Scanziani M. Equalizing excitation-inhibition ratios across visual cortical neurons. Nature. 2014;511:596–600.
25. Denève S, Machens C. Efficient codes and balanced networks. Nat Neurosci. 2016;19:375–82. https://doi.org/10.1038/nn.4243.
26. Shanglin Z, Yuguo Y. Synaptic E-I balance underlies efficient neural coding. Front Neurosci. 2018. www.frontiersin.org/articles/10.3389/fnins.2018.00046/full
27. Sadeh S, Clopath C. Excitatory-inhibitory balance modulates the formation and dynamics of neuronal assemblies in cortical networks. Nature Rev. 2021;22:21–37.
28. Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex J Neurophys. 2005;94:1904–11.
29. Marder E, Goaillard JM. Variability, compensation and homeostasis in neuron and network function. Nat Rev Neurosci. 2006;7:563–74. https://doi.org/10.1038/nrn1949.
30. Zenke F, Gerstner W, Ganguli S. The temporal paradox of Hebbian learning and homeostatic plasticity. Curr Opin Neurobiol. 2017.
31. Turrigiano G. Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function. Cold Spring Harb Perspect Biol. 2012.
32. Wilson HR, Cowan JD. Excitatory and inhibitory interactions in localized populations of model neurons. Biophys J. 1972;12(1):1–24.
33. Brunel N. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. J Comput Neurosci. 2000;8(3):183–208.
34. Rao & Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2(1):79–87.
35. Friston K. The free-energy principle: a unified brain theory. Nat Rev Neurosci. 2010;11:127–38. https://doi.org/10.1038/nrn2787.
36. Hawkins J, Blakeslee S. On intelligence. Times Books. ISBN 0–8050–7456–2. 2004.
37. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behavioral and Brain Sciences. 2013;36:181–253.
38. Butz MV. Event-predictive cognition: a root for conceptual human thought. Top Cogn Sci. 2021;13:10–24. https://doi.org/10.1111/tops.12522.
39. Butz MV. Towards strong AI. Künstlische Intelligenz. 2021;35:91–101. https://doi.org/10.1007/s13218-021-00705-x.
40. Desimone R, Duncan J. Neural mechanism of selective visual attention. Annu Rev Neurosci. 1995;18:193–222.
41. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124. PMC 1182327. PMID 16060722.
42. https://en.wikipedia.org/wiki/Replication_crisis. Accessed 13 Sept 2023.
43. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018;2(9):637–44. https://doi.org/10.1038/s41562-018-0399-z. PMID 31346273. S2CID 52098703.
44. Open Science Collaboration. Psychology. Estimating the reproducibility of psychological science. Science. 2015;349(6251).
45. Dacrema, P Cremonesi, D Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, vol. 13. 2019. p 101–9.
46. Neuroskeptic. fMRI gets slap in the face with a dead fish. Discover Magazine. 2009. www.discovermagazine.com/mind/fmri-gets-slap-in-the-face-with-a-dead-fish.
47. Madrigal A. Scanning dead salmon in fMRI machine highlights risk of red herrings. Wired. 2009. www.wired.com/2009/09/fmrisalmon/.
48. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities PNAS, v79.8. 1982. p 2554–8.
49. Achler T. Are assumptions of brain recognition correct? 2020. youtube.com/watch?v=F-GBIZoZ1mI&list=PL4nMP8F3B7bg3cNWWwLG8BX-wER2PeB-3&index=1.
50. Achler T. Neural phenomena focus. 2016. https://youtu.be/9gTJorBeLi8.
51. Achler T. Symbolic neural networks for cognitive capacities. Biol Inspired Cog Arch. 2014;9:71–81.
52. Spratling MW. A predictive coding model of gaze shifts and the underlying neurophysiology. Vis Cogn 2017;25(7-8):770–801.
53. Achler T. Input shunt networks. Neurocomputing. 2001;44–46c: 249–255.
54. Hebb DO. The organization of behavior. New York: Wiley & Sons; 1949.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.