Abstract—Ever since the advent of the neuron doctrine more than a century ago, information processing in the brain is widely believed to mainly follow the forward pre to post-synaptic neurons direction. Challenging this prevalent view, in this paper, we emit the backpropagation-based recollection hypothesis as follows: weak and fast fading Action Potentials following the (highest weight) post to pre-synaptic backward pathways, mediate explicit cue-based memory recall. This includes also the tasks of imagination, future episodic thinking, language understanding and associating names to various stimuli. These signals originate in highly invariant neurons, which uniquely respond to some specific stimuli (e.g. image of a cat). They then travel backwards to reactivate the same populations of neurons that uniquely respond to this specific stimuli during perception, thus recreating “offline” an experience that is similar. After stating our hypothesis in details, we challenge its assumptions through a thorough literature review. We find abundant evidence that supports most of its assumptions, including the existence of backpropagating signals that have interesting properties. We then leverage simulations based on existing spiking neural network models with STDP learning to show the computational feasibility of using such a mechanism to map the image of an object to its name with the same high accuracy as a state of the art machine learning classifier. Although not yet a theory, we believe this hypothesis presents a paradigm shift that is worth investigating: it opens the way, among others, to new models with STDP learning to show the computational feasibility of using such a mechanism to map the image of an object to its name with the same high accuracy as a state of the art machine learning classifier. Although not yet a theory, we believe this hypothesis presents a paradigm shift that is worth investigating: it opens the way, among others, to new interpretations of language acquisition and understanding, the interplay between memories encoding and retrieval, as well as reconciling the apparently opposed views between sparse coding and distributed representations.

I. INTRODUCTION

Human brains process sensory visual input and learn to extract invariant representations from such input in an unsupervised manner. However, mapping signifiers 1, i.e. mental representations of the image-sounds or "words", to the signified mental representations that such words refer to, needs an interaction with an external agent that supervises the learning. Such “teacher” simultaneously generates the sound-image related to the signifier in the presence of the actual stimuli that relates to the signified: this can happen for example by generating the sound "cat" or the writing of the word "cat" in the presence of an actual image of a cat. The teacher shall repeat this procedure until both are associated. We say that the agent has learned to map the signifier to the signified.

In this context, it is tempting to think at first sight that the repeated co-occurrence of both stimuli reinforces their connection, “in a hebbian manner”, thus allowing the mapping of signifiers to their signified representations. However, it is commonly accepted that neurons process sensory information mainly in a forward manner, i.e. from pre-synaptic to post-synaptic neurons. Yet, in the case of the signifier and the signified “problem”, the two co-occurring signals would eventually reach a connection point, where both signals have followed only forward paths. Now, assuming that memories are stored and encoded in the same area 1 that uniquely responded to the memorized stimuli during the first encounter, a challenging question arises: What neural mechanism allows to associate one to the other, such that the activation of the signifier can trigger back the activation of the signified and vice versa.

We argue that this problem is a particular case of a more general one that occurs whenever the recollection of previously encountered and stored stimuli is needed. This is the case of explicit memory where sensory stimuli, such as a visual scene or a smell, act as a cue to trigger the retrieval of past related events. We further argue that this is also the case in imagination where different parts of previously encoded stimuli are recalled and “merged” together to generate an “imagined” new experience that was not exactly met before.

In this paper, we hypothesize that weak and fast-fading backpropagating action potentials (APs) (from post to presynaptic neurons) whose strength is proportional to presynaptic “weights” are the medium by which previously encoded information is recollected. Such signals start in what we call source pointer neurons and travel all the way back, selectively reactivating on their path the neurons which uniquely responded to the stimuli during the first encounter, thus creating a similar experience to that of the first time. A source pointer neuron, as we will develop later is a neuron that specializes, thanks to past memorized co-occurrences, in selectively and invariantly 2 responding only to the retrieval cue (e.g. signifier for language) or its associated memories to be retrieved (e.g. signified). We refer to this as the backpropagation-based recollection hypothesis throughout the paper.

Prior work has extensively studied the interplay between visual perception and retrieval as happens in visual imagery (See for example Pearson’s recent review 4). For example, in addition to the high overlap between areas involved in retrieval

1 See the discussions in Sec. III-C1 for more elaboration on this assumption
2 As we will show later in Sec. III-D the existence of such selective neurons has been widely observed 2, 3
and perception, it has been observed—thanks to Dynamical Causal Modeling (DCM) of activation patterns, that there exists indeed a reverse top-down signaling pathway, from higher-level to lower-level cortical areas, that is responsible for recollection of visual images [5], [6]. However, following the traditional conception of forward propagation, these observed top-down activation patterns were attributed (wrongly, we argue) to backward recurrent feedback connections. Providing a biologically plausible computational model at the neuronal level that explains how backward recurrent connections (that use the pre to post synaptic path) can re-activate a previously encoded stimuli was beyond the scope of their work and remains to the best of our knowledge unsolved. For the sake of completeness, it is worth mentioning that recent years have actually seen the rise of “forward-based” computational generative models that come from machine learning and that can generate realistic images, the most notorious being Variational Autoencoders (VAEs) [7] and Generative Adversarial Nets (GANs) [8]. However, having been designed for a different purpose, it is not clear how they can be put together to implement retrieval tasks, even in machine learning. Second and most importantly, their complexity and the supervised mechanisms they employ make them less likely to be biologically plausible [9], [10], [11]. We argue instead in this paper in favour of a simpler unsupervised mechanism where no local error information and no output target for supervision are needed: the same forward paths used for perception are simply used backward for retrieval.

After stating our hypothesis, in the general case (Sec. II-A) and in the particular case of language understanding and naming (Sec. II-B2), we discuss its verifiability and run it against a thorough review of related literature (Sec. III). We found no evidence that allows to rule it out, but a significant body of experimental evidence that corroborates the plausibility of many of its assumptions. We find, for example, that fading away action potentials backpropagating to the dendrites have been widely measured and shown to exhibit interesting properties: they are stronger when the postsynaptic neuron is firing and most of all, they can be controlled by neuromodulation so as to increase their strength or disinhibit them (see Sec. III-A). We further review the literature about neural correlates of cue-based explicit memory retrieval, as well as the existence of sparse pointer neurons, and find further evidence supporting the framework of the hypothesis.

We then focus in the remainder of the paper on a particular case of our problem which is language understanding and particularly naming, which we computationally model (Sec. IV) and then simulate (Sec. V). In this context, we define naming as the act of retrieving the representation of the sound-image (signifier) that refers best to a presented visual stimuli (signified). We define understanding, on the other hand, as the task of retrieving the signified representation that corresponds to a presented auditory or visual stimuli of a signifier. We leverage recent success [12], [13] in training Artificial Spiking Neural Networks (SNNs) with Spike Timing Dependent Plasticity (STDP) learning to simulate a neural network that implements our hypothesis. We verify the computational efficiency of backpropagation-based recollection by comparing its accuracy in correctly naming a visual object, to that of a state of the art Machine learning algorithm. To further challenge the computational ability of our hypothesis, we test it on an extreme learning task, which is naming objects after seeing a single instance of each class. We find that backpropagation-based recollection leads on average and maximum to a higher accuracy compared to a Support Vector Machine (SVM) classifier. We are of course aware that the SNN models we use in this paper are not the brain, let alone from the particular implementation we leverage. We believe, however, that the simulations hint towards the computational efficiency of the mechanism. That, especially tied with our literature review, calls for a serious further exploration of this path. Given the breadth of the potential implications (as we discuss in Sec. VI-A), the backpropagation-based recollection hypothesis deserves at least to be rigorously and explicitly proven wrong through further studies.

II. THE BACKPROPAGATION-BASED RECOLLECTION HYPOTHESIS

A. General case

We posit our hypothesis and its assumptions as follows. When a stimuli is presented to a (“trained”) brain, the latter responds by a unique activation pattern, resulting from the activity of an ensemble of neurons that signal the various salient features of the stimuli. Memories of past encountered stimuli are stored in a distributed fashion by the same neural ensembles that detected such stimuli when encountered and encoded for the first time. Recollection of memories is therefore a process by which the appropriate population of neurons is re-activated again so as to “re-live”, offline, an experience that is similar to the first encounter. We hypothesize that weak and fading-away backpropagating signals from post-synaptic to pre-synaptic neurons, whose strength is proportional to the post-synaptic neuron’s firing rate and pre-synaptic weights, is the mechanism by which the brain performs generative tasks. By generative tasks, we mean the regeneration of a previously lived and memorized stimuli (e.g. recollection in explicit declarative memory), the generation of a plausible future stimuli (e.g. future episodic memory) or the regeneration and combination of previous separately-lived and memorized stimuli, a process we refer to as imagination (e.g. imagining a cat that laughs by combining a memorized mental image of a cat with that of the act of laughing).

The recollection process starts by the presentation of a retrieval cue, which activates few sparse neurons that uniquely identify both the retrieval cue and the to-be retrieved memory. We assume that these neurons learned to respond only to the presence of either of both stimuli thanks to a low-level learning rule such as STDP [15]; due to the (repeated or modulated)
co-occurrence of both stimuli in the past: the cue and the to-be retrieved. We further hypothesize that the retrograde signal is initiated in such source “pointer” neurons (e.g. “Jennifer Anniston” cells [3] if the goal is to recall prior stimuli related to, say, Jennifer Anniston). The signal then travels backwards following the paths with higher weights, activating on its way all the various neurons that compose the mental images to be recalled. We hypothesize finally that such backpropagation can be controlled remotely via neuromodulation so as to invoke it, increase its strength or to inhibit it. This neuromodulation acts thus as a “switch” to control whether to retrieve or not.

B. Case of language acquisition, understanding and naming

We further argue that a particular case of the above-mentioned tasks, which we later computationally simulate, is a form of explicit semantic memory related to language understanding and naming. We first start with some terminology.

1) Terminology: We adhere to the conceptualization and terminology introduced by swiss linguist Ferdinand de Saussure [1] and build on the distinction he introduced between the signifier and the signified. Since several interpretations of Saussure’s work could perhaps be made, we clarify ours. We refer to the signifier as the mental representation of the sound-image of the word. By sound-image, we mean either the phonetic sound resulting from the word, or the image of the letters that form the word. We refer to the signified as the mental representation of the actual object that the sound-image and its mental concept refer to. Both the signifier and the signified are concepts, one represents the word, the other represents the mental image(s) that this word often refers to. In this context, we refer to understanding as the act of mapping the signifier to its signified. Naming is the act of retrieving the signifier that corresponds to a given mental representation or to a presented visual stimuli.

2) Illustration in the case of language: When it comes to language, our hypothesis implies that backpropagating APs mediate understanding and naming. Fig. 1 illustrates our modeling through the example of three concurrent sensory inputs that are presented to a learner and that need to be “permanently associated”. This exemplifies a “teacher” that shows the “learner” an image of a cat, simultaneously to how the word cat is written, as well as the sound of the word. We assume as illustrated that there exists an area where the visual and sound activation pathways intersect (here, at the last “backpropagation root” layer). Such intersection areas are the root of backpropagating APs. During the “understanding” task, the retrieval cue is a word-related stimuli (e.g. sound “cat” or image of the word “cat”) and the retrieved memory is the signified representation (illustrated by the Understanding backward pathway in the figure). In the task of naming, the retrieval cue is the signified object (here a cat) and the retrieved memory is the name or signifier of the object (illustrated by the Naming backward pathway).

In more details, the sensory input of a sound “cat” as well as an image of the word are processed through consecutive feedforward neural layers. Similarly to what happens in primates’ visual cortex [17], neurons in the earlier layers have learned (thanks to a simple unsupervised rule such as STDP) to respond to simple features and the deeper we go, the more selective the neurons become and the more they respond to more complex ones. We assume that at a later processing stage, there exists fewer “sparse” neurons that selectively respond only to the presence of the sound “cat”, we refer to such neurons as the “sound signifier” pointer neurons of the word cat. Similarly, we assume that there exists neurons that selectively respond to the “image” of the word and we refer to such family of neurons as the “Image signifier” pointer neurons. Finally, neurons that selectively respond to both are referred to as the “sound-image signifier” or simply signifier pointer neurons. Following a similar pattern, the visual image of the cat itself is processed through various neural stages until certain neurons (referred to as signified in the figure) respond selectively only to the presence of the image of a cat.

This is how, as illustrated in the figure, the above described visual and sound pathways reach a common connection point, in a subsequent feedforward neural layer. We hypothesize that the repeated, or the neuromodulated [4] co-occurrence of signifier and signified stimuli (i.e. saying “cat” in the presence of an actual cat) reinforces their connection, at this junction point layer, in “a hebbian manner” thanks to a simple rule such as STDP learning. Next, when presented with either signifier or signified related stimuli, the source pointer neuron(s) [5] in the backpropagation root layer will fire, resulting in backpropagating action potential(s) that are proportional to presynaptic weights. The latter will cause the backward activation of the appropriate neurons, thus recalling the signified if the presented stimuli relates to the signifier, and vice versa.

It goes thus without saying that we adhere to the view that

Fig. 1: Illustration in the case of naming and understanding

---

4The connection can be reinforced either by mere repetition, e.g. repeating many times the word cat in the presence of an image cat, or by neuromodulation, in which case a single co-occurrence can be enough.

5One in theory should be enough but there must be many in reality, e.g. at least for redundancy to prevent failures.
there exists few neurons that respond selectively to complex stimuli such as (i) sound signifier stimuli, (ii) image signifier stimuli, (iii) signified or (iv) uniquely to any of the three previous ones. During backpropagation-based retrieval, these neurons act as pointers to selectively reactivate an appropriate population of pre-synaptic neurons that uniquely characterizes the memory trace to be retrieved; thus creating a similar experience to that of the first encounter(s) when the memory was encoded. For example, what is retrieved could be an experience of the sound of the word "cat" with the particular voice or conditions in which it was encoded (in line with what is called the encoding specificity principle [18]).

This is how, as we will discuss later, our hypothesis reconciles (i) sparse/localist and (ii) distributed representation theories in the brain, promising to end a long debate between cognitive psychologists and neuroscientists [19], [20]. In our framework, there is no need to choose between them as both are needed, but for different purposes: Sparse coding, illustrated here by the presence of highly selective neurons, is needed for backpropagation-based retrieval, while the encoding of the entire memory trace is still done via a distributed set of neurons. The latter can be selectively reactivated "on-demand", from source pointer neurons backwards. It is thus the simultaneous activation of an entire population of neurons that forms the entire memory trace, and single highly selective neurons are only pointers, helpful for retrieval.

Finally, and anecdotally, the fact that backpropagating signals are "fading-away" in nature could explain the ephemeral nature of the experience of recalled memories or visual mental imagery's lack of vividness: the latter are not as persistent as the experience of live sensory stimulation.

C. What this hypothesis is not about

Finally, we clarify that this hypothesis covers only the recollection processes. This means, in the particular case of explicit recall, the reactivation, as close as possible, of the same neurons that were activated during previous encounters. As a consequence, in the case of language, what is covered by our hypothesis is how to (i) learn the name association and how to (ii) recall the name, not yet actually how to produce it (e.g. emitting the sound or writing the letters). Further investigations are needed to reassess production tasks in light of our new hypothesis. Nonetheless, it occurs to us that learning to produce the right sounds, i.e. speaking, happens through a trial and error, reward-based basis process in which the goal is to "mimic". The study of such mechanism is beyond the scope of this paper. Our hypothesis covers how the reactivation of source pointer neurons leads to recollection. How such pointers further participate to invoke motor areas to produce sounds or write letters is left for future work.

6For simplicity, we focus in this toy example, on a single encounter that, we assume, was "encoded right away". In reality, stored memory traces might evolve with repeated exposure, such that recalling what is meant by the word cat, leads to the recall of a memory of the signified, that is statistical in nature (e.g. one of the many cats met before, or an average abstract image of a cat)

Next, when we talk about understanding, we mean the modality-specific features of semantic memory [21]: i.e. recalling the details of how a face or an emotion look or feel like exactly, as opposed to other aspects of semantic memory like finding abstract relationships between words. We leave the latter aspect also for future work. Worth mentioning though, Patterson et al. reviewed semantic knowledge organization in the human brain [21] and reported that all theories agreed on the fact that modality-specific recall is implemented by a distributed brain network, a fact that is coherent with a backpropagation-based recollection hypothesis.

Then, our hypothesis assumes that there are centers that remotely control via neuromodulation whether or not to call the recollection (by either increasing the backpropagation or inhibiting it). Our hypothesis does not cover what mechanisms control those control centers, and under which conditions recollection is favoured or shut down. What our hypothesis predicts is that the task of these "control centers" can be extremely easy to implement: the diffuse and "untargeted" remote generation of an excitatory neuromodulator favours further recall (i.e. recall of whatever cues are active or firing at the moment). The same applies for inhibition.

Finally, our hypothesis, being only focused on the recollection process from sparse pointer neurons backwards, does not directly explain the mechanisms involved in sparse neurons formation, novelty or familiarity detection, and the interplay between short (e.g. few days back) and long (e.g. few years) term memories. Nonetheless, it still offers a ground to reason about these issues. For example, the fact that the same sparse neurons keep being used to signal the presence of the same familiar stimuli throughout the years (e.g. face of own child) could explain why humans are unable to remember much younger versions of these faces (in the absence of photos): memories are updated in situ and familiar faces will always lead to the activation of the very same “familiar” invariant sparse neurons, not to the activation of “novel” ones.

III. Review of evidence in the literature

In this section, we position our hypothesis in the literature and show evidence that backs it up.

A. Existence of backpropagating action potentials

Despite the forward-processing prevalent view, a plethora of studies have measured, in vitro and in vivo in anesthetized and awake mammals, action potentials that backpropagate to apical and distal dendrites, and this for various classes of neurons. We cite in the following some of them. For a more complete list, the reader can refer to the reviews of Stuart et al. or Waters et al. which summarized the findings about the measurements and hypothesized few roles of backpropagated action potentials. For a more general review of retrograde signals, i.e. not only activity-dependent but also during synaptogenesis etc., the reader can refer to Tao et al.

Williams and Stuart performed simultaneous somatic and dendritic recordings from Thalamocortical (TC) neurons
and measured that action potentials, both due to sensory information or cortical excitatory postsynaptic neurons, backpropagate into the dendrites. In another work, the same authors [29] measured the same phenomenon in neocortical pyramidal neurons. Interestingly for our hypothesis, the authors have found that action potentials due to physiological patterns of firing, backpropagate three to four times more effectively compared to action potential pertaining to mean firing rates. This observation is confirmed by several studies (reviewed by Waters et al. [27]), which found that backpropagation was modulated by synaptic input. For example, properly timed excitatory input leads to the amplification of backpropagation, whereas inhibitory input might block it. More interestingly for our hypothesis, many neuromodulators have an influence on backpropagation, often leading to its enhancement, but in more complex ways. For example and interestingly, given the supposed role of the hippocampus in retrieval (see later), in hippocampal CA1 pyramidal neurons, it has been observed that muscarinic agonists enhance the backpropagation in a progressive manner having a stronger and stronger effect on subsequent action potentials [31], [32]. This suggests that neuromodulation therein can act as a “switch” to enable Action Potential backpropagation in a selective manner, a feature that is necessary for our hypothesis. As described in Sec. II, not every presentation of a visual stimuli would systematically lead to the activation of “naming”. And probably not every presentation of the signifier stimuli should lead to the evocation of its signified representation. Similarly, not every exposure to a familiar stimuli (known) automatically leads to explicitly remembering its context.

Nonetheless, to the best of our knowledge, the role of such activity-dependent backpropagation of Action Potentials, as reviewed for example by Waters et al. [27], has been so far hypothesized to be local, acting as a feedback loop from postsynaptic to pre-synaptic neurons, to regulate spiking activity or support synaptic plasticity. In this work, we emit the hypothesis that such retrograde signals play a more explicit and direct role in higher-level cognitive tasks, such as naming, understanding, and other generative processes like imagination and explicit memory retrieval. We next oppose our hypothesis to the state of knowledge in experimental cognitive sciences about explicit memory (Sec. III-B). We later dive deeper in details and oppose our assumptions to what is known about the neural correlates of explicit memory (Sec. III-C).

B. Cue-based retrieval: a cognitive sciences perspective

Our hypothesis applies to any task in which the reconstruction of previous encoded stimuli is needed. The literature on this topic found its origins in early experimental cognitive science research (e.g. [33], [34]) before recent advances, driven by neuroimaging and optogenetic stimulation, allowed to shed more and more light on some actual neural correlates [35]. We start by surveying the first and linking it to our hypothesis.

Since the seminal work of Endel Tulving, long-term human memory is widely classified into explicit (declarative) memory and implicit (procedural) memory. While procedural memory relates to long-term acquired skills such as driving or playing an instrument, declarative memory, relates to the explicit recollection of memories about facts, words, images and events etc. In this paper, we focus on the latter and provide a hypothesis about how the retrieval of these memories happens at the neuronal level. It is worth mentioning that Tulving also played a role in further dividing explicit memories into episodic and semantic ones [34]. Our hypothesis is orthogonal to the difference between them.

One crucial principle in this area was formulated by Tulving and Thomson under the name of encoding specificity [18]. The principle stresses the importance of retrieval cues and the entire context that is perceived during encoding for later retrieval: the surrounding context that was present during the first perception and encoding moment can act as an efficient retrieval cue in the future. Although it might sound straightforward today, such early work [33], [18] played a role in differentiating between availability of memories and their accessibility, thanks to cues, let them be internal or external stimuli-based: the inability to recall does not necessarily mean that the memory is not available but could be also due to the lack (inactivation) of appropriate cues.

Our hypothesis offers a neurobiological ground to interpret and simulate encoding specificity: for us, any surrounding context during encoding can act as a retrieval cue as long as it activates the source pointer neurons that uniquely identify the retrieval cue and all the remaining surrounding context to be retrieved. Backpropagating action potentials can then “travel backwards” to reactivate networks of neurons which uniquely responded to the stimuli during the encoding moment, thus creating, again, a similar experience.

In fact, the process by which an internal or external (sensory) cue activates the stored memory trace is well known and has been called ephory [36], [37]. Ephory is this interaction between trace and cue, is described as the first stage of memory retrieval before conversion actually happens and the recollection experience is “lived”. In our hypothesis, the cue activates mainly the source pointer neurons, which in turn activate back the appropriate presynaptic populations, resulting in the recall of all the related traces. As such, following our hypothesis, it can be easily seen how two components may affect memory retrieval as already predicted by Tulving: the lack of the appropriate cue or a decay of the synaptic weights of the concerned neural networks. Our hypothesis announces others that could be verified. For example, a decline in the intensity or extent of the backpropagation (e.g. impairment in the neuromodulation that is supposed to facilitate it) can hamper retrieval. Conversely, an excess of such backpropagation might lead to higher levels of intrusive thoughts.

Today, the encoding specificity view has passed the test

The selectivity of source pointer neurons is a consequence of encountered prior stimuli, and this defines what particular contextual cues are efficient at recalling what particular encoded trace.

A term that Tulving revived together with the forgotten work of german scientist Richard Semon, who first coined it and stressed the importance of retrieval cues.
of time (see [35] for a recent review) and even its “opponents” [38], [39], [40] do not question the necessity of some degree of match between encoding and retrieval conditions, but rather stress the importance of additional factors that influence the performance of later retrieval; the most important being the “discriminative” power of the retrieval cue or its distinctiveness. Accordingly, recalling performance is not only related to the amount of match between encoding and retrieval conditions, as thought first, but also to cue overload and hence to what extent the cue is “discriminative”. This latter view can be easily observed under our framework: let’s have in mind the illustration in Fig. 7 and the image of a cat as a cue. If this image appears, during learning, simultaneously with all sorts of names, and not only the signer “cat”, then the backpropagating action potentials would potentially simultaneously activate many “Signifier” neurons (and not only that of the cat), making it hard to distinguish and hence correctly name. The necessity of being discriminative can be seen in our simulations of naming later in Sec. 7 after backpropagating action potentials back to the “categories” layer, the “signifier neuron” that gets the highest “votes” is elected to signal the name of the object. If all neurons receive “equal votes” because of cue overload, it would be impossible to retrieve the right name.

C. Explicit memory: neural correlates

We now dive more into the neurobiological foundations of encoding specificity, ecphory and the processes involved in explicit memory. This field has seen tremendous advances driven by both neuroimaging and optogenetic stimulation.

We start from a recent thorough review [35] in which a large body of research strongly supported Semon’s and Tulving’s cognitive theories: namely, that (i) the success of accessibility depends on the interaction between cues and memory traces and that (ii) there are strong ties between encoding and retrieval, including at the level of activated neural ensembles. In particular, it was shown, using artificial optogenetic stimulation [17] that it is possible to either disrupt or mimic ecphoric processes by activating or inhibiting the same specific neural ensembles that were active during encoding.

In one experimental study [41] that we further analyze below, blocking the neural ensembles that were used during encoding to recognize the cue, resulted in impairment in retrieval. In the experiments, mice were conditionized to produce a fear response whenever placed in a particular context, a context which stands here for the cue. At the same time, CA1 neural ensembles that were particularly active during encoding were optogenetically tagged. Whenever placed in the same context again, mice successfully freeze as a sign that they recognize the environment. However, placing them while inhibiting the previously tagged neural ensembles considerably reduces freezing levels. This means that if the neural ensembles that recognize the cue do not activate, the memory is not retrieved. Other studies (e.g. [42]) similarly used optogenetics to demonstrate the “opposite”: artificially reactivating the neural ensembles that recognize the cue, thus inducing the retrieval of the memory trace (i.e. causing freezing), even outside the context in which the conditioning happened (i.e. in the absence of a natural cue). Now, in the previous two families of experiments, (i) mapping the cue and the trace was learned naturally, (ii) neural ensembles were tagged depending on their activity during encoding, and (iii) artificial inhibition or excitation was used to disrupt or elicit retrieval. A last recent family of experiments [43] showed that it is even possible to associate a cue and a trace artificially and to later elicit retrieval in natural conditions. In particular, they repeatedly used photo-stimulation to artificially activate a neural ensemble that usually recognizes a special smell, simultaneously to photo-stimulating another memory trace that elicits avoidance. After this co-occurrence based conditioning happened, exposure of mice to the real smell caused an avoidance reaction: mice “remembered” to avoid although they’ve never experienced the smell in real conditions.

Finally, note how, in optogenetic stimulation above, light simultaneously and selectively activates a set of neurons that were prepared in advance to become light sensitive. This allowed scientists to dissect the interactions between memory traces, retrieval cues and encoding. However, it is not clear today what mechanism in the brain can cause a similar selective reactivation, equivalent to optogenetic stimulation. That’s exactly the purpose of our hypothesis: hence, our answer to the question “how does a neural ensemble activates in a selective way” lies in backpropagating action potentials following the paths with highest presynaptic weights. Our hypothesis also offers a framework to simulate (as we do later) and understand these issues at the level of single neurons.

As a first summary, the studies above [41], [42], [43] and many others reported in the review [35] (which we encourage the reader to check) confirm the importance of cues and encoding specificity. But beyond that, they suggest that retrieval reactivates what was active during encoding, in a process referred to, sometimes, as neural reinstatement. This principle is at the heart of our hypothesis as backpropagated APs should follow exactly the reverse path that uniquely led to the activation of source neurons during encoding. It turns out that many arguments support this reinstatement principle. We review them in what follows.

1) Retrieval as top-down re-activation: First, historically, the oldest (yet weak) supporting fact is that simply reinstating the encoding context during the recollection time enhances retrieval performance and quality, as reported by some reviews [44], [45]. Second, more recently and more strongly, a significant body of research intentionally studied the overlap between encoding and retrieval and provided large evidence in favour of the principle using various ensemble tagging [42], [41], [46], [47], [48], [49], [50], [51], [52], [53]. EEG [54] and neuroimaging [55], [56], [57], [58], [59], [60], [61], [62].

9A technique that allows to later activate or inhibit precisely only selected populations of neurons that were initially selectively tagged, depending on their activity. 10Note that, here, we use interchangeably encoding, learning and conditioning to accommodate different terminologies used in different papers.
techniques. Furthermore, it was shown that the reactivation overlap between encoding and retrieval influences also the perceived quality of the retrieval. For example, in the case of visual imagery (i.e. attempting to mentally visualize an image), it has been shown that activation overlap in visual cortex increased visual imagery vividness, or the subjective intensity of the remembered image[65], [62], [61]. We refer the reader to the references above for more information and cite in what follows only few examples of each technique.

In terms of neuroimaging, Dijkstra et al. [5] used Dynamical Causal Modeling (DCM) [60] to infer coupling between cortical regions involved in the tasks of visual perception as opposed to visual imagery. They measured that visual imagery vividness correlated more with top-down connectivity patterns (from high level cortical areas to lower level areas) as opposed to perception itself. Many other studies suggest such top down mechanism during visual imagery [67], [68], [69], [6], [4]. The reader can refer to Pearson’s recent review [4] of the cognitive neuroscience of visual mental imagery for more details about the top down reverse hierarchy of information and the fact that the process seems to be a weak form of the bottom-up perception. However, in general, due to the widespread view that neural computation is mainly forward from pre to postsynaptic neurons, this top down activation cascade has been always interpreted, in the literature, as the result of feedforward feedback connections from higher-level cortical layers to lower ones.

Then, and perhaps more convincing than neuroimaging, neural ensemble tagging techniques also confirm the principle. In addition to work we described above [41], [42], [43], another recent example is the work of Guskjolen et al. [53] who performed contextual fear conditioning experiments on young mice while tagging the neural ensembles which were active during encoding. As happens with infantile amnesia in humans, the infant mice later exhibited forgetfulness. However, photo stimulation of tagged neurons, only in the hippocampal formation (Dentate Gyrus in particular), induced memory recovery and reactivation of broader areas which were tagged during conditioning including hippocampal CA1 and C3, and cortical neurons. This finding is inline again with the idea that traces are distributed in neural ensembles that span many cortical brain regions [70], each responsible of one aspect of information (sensory, motor, visual, emotional etc).

2) Existence of “backpropagation root layers”: The last paper leads us to the last point we review in this section: the role of the Medial Temporal Lobe and its relationship to our backpropagation root layers, where source pointer neurons are located. Indeed, our hypothesis assumes the existence of an area where source pointer neurons lie and where the backpropagation starts. If our hypothesis is correct, this area should form the “glue” between cues and retrieved traces, and should observe a reversal of the flow of information. Interestingly, the Medial temporal lobe and the hippocampus in particular has been shown to (i) play this role and (ii) exhibit a similar reversal behaviour.

For (i), many theories [71], [72], [73], [74], [75] support that the hippocampus performs exactly the task of reinstating patterns of activity in the cortex that were alive during encoding. This can be already seen from the study of Tanaka et al. [51] which we reported above. What they did by actually monitoring cortical activity while inactivating hippocampal CA1 cells on rodents, shows how the hippocampus is likely responsible for reinstating the patterns that were active at encoding. By permanently tagging neurons which were active during encoding (in a fear conditioning experiment), they were able to silence them with laser stimulation, up to several days later. When silencing only the tagged CA1 cells (and not the entire engram or trace ensemble), memory retrieval was impaired; and the rest of the neural ensemble in the cortex and amygdala, which used to reactivate during retrieval, was not reactivated again. Many other studies [76], [77], [78] also showed that retrieval success depended on whether or not the hippocampus was concurrently solicited or not, and this during both encoding and retrieval. Horner et al. [77] showed further evidence that the hippocampus binds together all elements composing a trace that is stored in distributed regions in the cortex, playing as a hub to perform what is also called pattern completion task. Finally, interestingly for our hypothesis, Staresina et al. [78] observed a reversible signal flow from the cue region to the “to be recalled” target region, through the hippocampus. This puts the HC and MTL in the position of good candidates to be root backpropagation areas as per in our hypothesis: they seem to implement the link between cues’ networks and traces’ networks and they seem to be the place where flow reversal happens.

Citing verbatim a thorough review and perspective from Moscovitch[79]: “Retrieval occurs when an external or internally generated cue triggers the hippocampal index, which in turn activates the entire neocortical ensemble associated with it. In this way, we recover not only the content of an event but the consciousness that accompanied our experience of it”. Moscovitch refers later to the hippocampal memory indexing theory of Tevler and DiScenna [73], [72] as follows: “Memory traces in the HC/MTL are encoded in sparse, distributed representations that act as an index or pointers to the neocortical ensembles that mediate the attended information”. This claim, which is inline with our hypothesis, leads to the next assumption which we further oppose to the literature: the existence of sparse source pointer neurons.

D. Existence of highly selective source Pointer Neurons

We assume that, at a certain deep level of processing, certain neurons will become highly selective and invariant: they serve, in our hypothesis, as pointers to reconstruct the encoded stimuli they represent. For example, in the particular case of language, some neurons will respond only to the signified or only to the signifier (one of the cues), and some would respond to both the signifier and the signified (i.e. what is in common between the cue and the “to be recalled”). In this section, we scan the literature about concept representation in the brain to assess the plausibility of the existence of such neurons. It turns out that similar neurons have been documented and that their
role is not yet well understood, given the still ongoing debates between two opposed views on the matter: the “distributed representations” view [80], [81] and the “sparse coding” view [82], [83].

Indeed, the distributed representations view defends that concepts in the brain are represented by the unique activation patterns of entire and large populations of neurons. It is thus the pattern uniqueness across a large population that defines complex concepts, not particular single neurons. The sparse coding view defends instead that there exists few neurons that represent selectively particular items or concepts. The extreme version of sparseness would be that there is a unique cell that responds to each single unique concept, a version that is pejoratively known as the grandmother cell hypothesis [19], [20]. Our hypothesis promises to reconcile both views as follows: information is stored in distributed networks, and sparse neurons, which also exist, play the role of hubs to connect such distributed networks, easing retrieval by being sources of backpropagated APs.

Now looking at the literature, the first accounts of sparseness are quite old already. In practice, since the seminal work of Hubel and wiesel [84], it became mainstream that neurons tend overall to respond to more and more complex features, the deeper we go in the processing layers of sensory input [17], [85]. Indeed, evidence suggests the existence of a hierarchy starting from the primary visual cortex V1, where basic features are encoded, until the inferior temporal (IT) cortex, where neurons selectively respond to complex shapes like hands and faces [86], [87], [88].

Other known examples of sparse coding, for spatial representation, are place and grid cells [89]. Place cells for instance are single neurons which signal specific places in the environment: as the individual navigates in its environment, only the neurons that signal the current place field fire. Interestingly, such highly selective neurons have been found within the hippocampal formation, the area which seems to be a good candidate for being a backpropagation root layer as discussed in Sec. III-C2 above.

In general, the literature is rife with studies that have measured such selective neurons, in ways that fit our hypothesis, and interestingly in these same MTL areas. For example, Fried et al. [90] have measured neurons that selectively discriminated humans (faces) from inanimate objects, and this, interestingly, during both encoding and retrieval. Others distinguished specific facial expressions. A little later, Kreiman et al. [91] have measured neurons that highly responded only to specific categories such as animals, houses and celebrities.

In a continuous line of work, Quiroga and colleagues [2], [92], [93], [94] have set to understand how the visual features we mentioned above are passed to upper layers of the hierarchy so as to understand how they are later used by higher cognitive processes: a question to which we hypothesize as an answer in this paper. It is in one of these works, which became popular, that Quiroga et al. [2] reported the existence of highly selective neurons that responded to the presence of specific stimuli related to places or individuals such as Bill Clinton and Jennifer Anniston. One of the found selective neurons even exhibited highly selective responses to any stimuli that is related to Halle Berry, let it be the face, or even the written words. The latter neuron exhibits striking similar properties to our Source pointer neurons as described in the language understanding and naming tasks. Then, given that this work reminds the widely unaccepted grandmother cell hypothesis, further clarification have followed up.

Waydo et al. [95], with Quiroga as a co-author, later used a probabilistic approach to explore a bit more rigorously Quiroga et al.’s original findings [2]. Indeed, the latter obviously did not test for all MTL neurons and all possible categories of objects. Hence, (i) a found selective invariant neuron could respond to other untested categories, and (ii) there might exist many neurons, and not only one as found by the authors, that would selectively respond to the same stimulus. Authors thus developed a probabilistic model to estimate the odds, and confirmed the sparseness hypothesis (yet, arguing also against single grandmother cells [82] as done first by Quiroga and co-authors). Nonetheless, the model leads to, as the authors conclude, only a bound on the true sparseness: the neural coding could be in reality even much sparser than they estimated. In another follow up work, Quiroga et al. [92] insisted, already in the title, on the fact that it is sparse but not grandmother cells and argued against the unlikely possibility that a single unique neuron responds to each stimulus. In our case, although we simulate the hypothesis using single neurons in Sec. [IV] our hypothesis is inline with existence of many of such sparse invariant neurons. On the contrary, we believe that multiple neurons shall be needed, at least to guarantee resiliency if some neurons fail. The authors however conclude with a set of difficult open questions. Our hypothesis suggests already few answers to the following ones: “How are MTL cells involved in learning associations? How are MTL cells involved in free recall or the spontaneous emergence of recollection in the human mind?” As discussed, we hypothesize it is: backpropagated APs, triggered through neuromodulation by remote “control centers” that simply decide whether or not to facilitate the recall, the naming etc. The same principle should apply to free recall, where said centers activate sequentially related concepts (see our discussion on mind wandering in Sec. [VI-A]).

Last but not least, in a subsequent work [94], Quiroga et al. have measured that the MTL selective neurons reflect the subjects’ decisions about the stimuli rather than visual features themselves. They’ve put this in evidence by performing experiments in which they present subjects with vague stimuli that is a mixture of different celebrities (e.g. a picture that is the mixture of presidents Bush and Clinton). As expected from previous studies, exposing users to one of the celebrities first, leads them to later see the morphed image as pertaining to the other celebrity. This is probably due to the fact of “tiring” the neurons that encode this celebrity. Later by recording Clinton’s and Bush’s neurons, they concluded that such MTL neurons fire in accordance to the decision, not the features. In an interesting follow up comment, Reddy et al. [95] remarked that
 damages to the MTL area cause subjects to have memory impairments, yet having perfect perceptual awareness and consciousness. This is inline with our hypothesized role of source pointer neurons that allow to map related stimuli with backpropagated APs.

As a summary, sparse neurons that respond selectively to complex concepts exist, in areas where we suspect them to do, with properties that are inline with our backpropagation-based recollection hypothesis.

E. Summary of arguments in favour

We now summarize, as simplified in Table I, the arguments in favour of our hypothesis. First, as seen in Sec. III-A, activity-dependent backpropagating action potentials are biologically plausible. Moreover, it has been observed that these APs are stronger when neurons are firing, a necessary condition for our hypothesis. Second, neuromodulation can enhance such backpropagation in a selective and progressive manner, thus acting as a switch to inhibit it or to strengthen it. This “feature” is necessary as humans can control whether or not to favour the retrieval of some explicit memory after the exposure to a cue. Indeed, not each time we see a screen of laptop, we recall its name, or other related memories. Interestingly enough, this modulatory phenomenon on backpropagated APs has been observed in hippocampal cells, an area that is known to be crucial in memory retrieval, and especially known in some theories as the place that stores the indexes that allow to retrieve memories that are stored in other cortical areas. We found also evidence that such sparse indexes which we called pointer neurons have been also observed and well documented. Even more interestingly, we’ve seen that a reversal of information flows has been observed in the hippocampus which is believed to act as a glue between the cues and the engrams to be retrieved. This brings us to another assumption of our hypothesis which is that retrieval is the reactivation of the same areas and neurons that were used during encoding, a task sometimes called neural reinstatement, or pattern completion. We find a large body of optogenetic-based and neuroimaging-based evidence that confirms such assumption. There is indeed a high overlap between the areas involved in these two tasks, and optogenetic-based experiments cited above are indeed based on tagging the specific neural ensembles that were active during encoding. Additionally, and perhaps as an added bonus, we will computationally show later that this hypothesis is an effective computational method to associate names to visual input, with the same high accuracy of a supervised machine learning algorithm.

Additionally and may be anecdotally, the fact that these signals are weak and fading away might explain why imagination and memory recollection elicit subjective experiences that are themselves transient and fading away in nature: the recollection of an image of a cat is much less vivid and persistent than the subjective experience that is due to the sensory input of a real cat.

Finally, not reviewed in details above, if our hypothesis proves true for cue-based recollection, it becomes more than reasonable to embrace the view that it also mediates other generative tasks such as mind wandering, intentional creative thinking, dreaming, as well as future episodic thinking or imagining the future. Existing neural correlates studies of such generative tasks can be leveraged to further verify our hypothesis.

To summarize, we interpret all the arguments in favour as an encouraging call for future work and further investigations. In particular, it should be verifiable experimentally if the extent of the action potential backpropagation is proportional to pre-synaptic weights, and whether and to what extent backpropagation can be far reaching (e.g. eventually one or more pre-synaptic hop(s) away). The latter is a crucial point that can drastically increase or decrease the plausibility of our hypothesis, and for which we did not find evidence neither in favour nor against.

IV. NAME ASSOCIATION: MODELING WITH SPIKING NEURAL NETWORKS

We now focus on the task of retrieving object names using their image as a cue. We set out to simulate our hypothesis and assess whether it is a computationally efficient strategy for this task. To this end, we leverage existing artificial Spiking Neural Networks (SNNs) trained with Spike Timing Dependent Plasticity (STDP) learning and simulate a “teacher” that simultaneously shows to the SNN, during learning, the images and their corresponding names. Then during test, backpropagated action potentials are used to retrieve the right name. We compare the accuracy of a naming mechanism employing our hypothesis to that of a machine learning classifier. This section focuses on our modeling. We first describe recent existing SNN models we build on (Sec. IV-A). We then critically review their limits and plausibility (Sec. IV-B). Finally, we detail how we use them to build our model (Sec. IV-C).

A. Image classification with existing Spiking Neural Networks

SNNs are a class of biologically inspired computational models in which spiking neurons communicate information through individual spikes that propagate from one neuron to the next. Such spikes simulate APs, happening when the membrane potential of the neuron crosses a certain threshold. In reality, both the rates at which spikes are generated and the temporal patterns of spikes are believed to carry information about the input stimuli. The artificial SNNs we

| Assumption                              | Evidence   |
|-----------------------------------------|------------|
| Backpropagating Action Potentials       | ✓ Sec III-A|
| Backpropagation stronger when neurons fire | ✓ Sec III-A |
| Backpropagation can be selectively modulated | ✓ Sec III-A |
| High overlap between retrieval and encoding | ✓ Sec III-C |
| Information flow reversal (at pointer neurons) | ✓ Sec III-C |
| Backpropagation effects can be far reaching | No, but verifiable |
| Backpropagation proportional to presynaptic weights | No, but verifiable |
| Existence of source pointer neurons | ✓ Sec III-D |
leverage in this paper simulate a simpler version of this process, yet still offering higher biological plausibility \[101\] compared to other artificial models. Indeed, for training, the SNN we leverage uses the more plausible STDP learning rule \[102\], \[16\], \[103\], \[104\], \[105\]. Under this rule, synaptic weights are updated according to the relative spike times of pre and post synaptic neurons: if the pre-synaptic spike occurs slightly before the post-synaptic spike, then a persistent strengthening of synapses called long-term potentiation (LTP) occurs \[99\]. In the other case, the result is a long-term depression (LTD) which leads to a persistent depotentiation of synapses.

Two recent models in particular provided background for our simulations \[12\], \[13\]. We build in particular on that of Kheradpisheh et al. \[12\] which achieved impressive accuracy on simple datasets. We reuse “as-is” its feature extraction layers. The latter are illustrated in the “feature learning (STDP)” upper part of Fig.\[2\]. As can be seen, it consists of consecutive layers of neural processing. The first is a temporal coding layer, that is meant to somewhat simulate retinal ganglion cells firing moments. It is followed by a cascade of convolutional and pooling layers to extract visual features. In more details, the first layer is responsible of encoding the input signal into discrete spike trains in the temporal domain. For this, it uses Difference of Gaussian (DoG) filters\[11\]. This layer detects positive and negative contrasts in the input image and encodes them in spike latencies, according to their strengths. Next, each neuron in the convolutional layer receives input spikes from the neurons located in a certain window and emits a spike when its potential reaches a specific threshold. Pooling layers perform a nonlinear max pooling operation in which they only propagate the first spike emitted. In this model, STDP learning only occurs in convolutional layers and it is done layer by layer. For each image presented to the neural network, there is a “competition” between the neurons of a convolutional layer and those which fire earlier trigger STDP and learn the input pattern. Finally, the last layer is a global pooling layer which performs a global max pooling. The role of these feature extraction layers is just to learn visual features: they are trained without name labels by propagating many images through the layers and adjusting the weights with STDP.

Next, unlike what happens within our hypothesis, in Kheradpisheh’s model \[12\], the trained output of this final layer is used to train a linear Support Vector Machine (SVM) classifier. The SVM classifier is of course not biologically plausible but the goal of Kheradpisheh et al. was only to assess the ability of SNNs and STDP to extract salient visual features that are good enough to discriminate images. And they actually found that they were good enough in terms of classification accuracy: their implementation reached 99\%, and 98.4\% of accuracy in the face/motorbike and MNIST datasets, respectively. We reproduced their results using Perez’s available implementation \[14\]. After some search of best parameters\[15\], we reached around 96\% of accuracy on the face/motorbike dataset with an SVM classifier.

Finally, worth mentioning, Mozafari et al. \[13\] proposed a 4 layers SNN with STDP, whose classification last layer is trained this time using reinforcement learning, instead of the SVM classifier. Their final layer is a decision making layer that performs a global pooling operation. Each neuron in it is assigned to a category and the neuron which fires first indicates the network decision. This work, unlike the previous one does not thus rely on an external, biologically not plausible, classifier. Indeed, weight change in the last layer is modulated by a reward/punishment signal which depends on the correctness/incorrectness of the network’s decision. However, the paper lacks plausibility in that it does not answer the challenging question of how and who generates the reward and punishment signals and more crucially: how does it “know” which neurons to punish and which neuron to “reward”. Later, we present instead an end-to-end model, at the neuronal level, from learning associations to naming. Indeed, we postulate that the simple repeated co-occurrence of signifier and signified is enough to tie them together as in passive learning, and this, in a bidirectional way (both naming and understanding). Hence in our framework, first, no unknown reward signal or mechanism is needed. Second, the naming of the object does not implicate a feed-forward mechanism but rather the backpropagation of action potentials.

**B. Plausibility of the above SNN models**

Next, before using Kheradpisheh’s model \[12\] as a basis, we briefly discuss its (lack of) plausibility, as this allows us to later better gauge the plausibility of our hypothesis. In a nutshell, we are aware that the SNN model above lacks plausibility in many aspects, despite being unsupervised, and despite using a simple rule like STDP learning. For example, it uses only a spike-time neural coding. It also uses convolutional neural networks (CNNs) with weight sharing, which is biologically not plausible. However, all these “problems” do not impact our hypothesis since we are mainly interested in the last layers of the neural networks (where the retrograde signaling or backpropagation of the action potentials will actually initiate). Besides, and very interestingly, recent work \[106\] has shown that training using “properly translated data” such as the case of “correlated” images in video relieves the need of using CNN weight sharing, and results in an approximate form of it.

Our position here is as follows. The above simple neural network model trained only with unsupervised STDP learning achieves good performance on what was 20 years ago a difficult problem. This means that it extracts visual features of fairly good quality. The latter are of course far from being perfect; the accuracy does not reach 100\% even on the simplest motor/face dataset. Nonetheless, we set out to verify if our hypothesis with STDP learning can successfully use these same features to find the right name association. For fairness, we compare our performance results to those of the SVM classifier.
The second step is a semi-supervised learning one, whereby a teacher shows a learner an image and the right name that refers to it. We call this co-occurrence learning. Humans for instance can learn from a single example to map a new object to its new name. Sometimes, when no external reinforcement happens, the exposition needs to be repeated multiple times until it is remembered. We model this step by simply adding a new “categories layer” and emulating the right spike each time an image is propagated through the SNN, i.e. generate a spike for “cat” category neuron while the cat image is propagated through the SNN.

The last step is simply the recollection of the name. It is in this step that a retrograde signaling from all neurons in the backpropagation root layer is sent backwards to the categories layer. The neuron(s) which receive the highest “vote” signal the network decision. We now describe the three steps of Figure 2 in more details.

1) Feature Learning: For feature learning, we used almost as-is the SNN model described above [12]. The reader can refer to the original paper for more details. This phase starts with the input image being encoded into discrete spike events in the temporal domain. Each “spatial position” or pixel in the input image will have its own spiking time. Such time-encoding process is performed by using Difference of Gaussian (DoG) filters, where spike times are computed according to the output of the DoG filter. More precisely, let r be the value at a certain position index after having applied the DoG filter. Then, the firing time t is defined to be $t = \frac{1}{r}$. This corresponds to encoding higher contrast areas of the image to lower spike times (i.e., latency is inversely proportional to the contrast).

As a result, each single image is transformed into several waves of spikes that propagate, one by one, through the layers: spikes that signal higher contrast areas being the first to enter the network. Next, Convolutional layers are arranged in a feedforward manner. Between two consecutive convolutional layers, a pooling layer performs a max operation to compress visual data and provide translation invariance. The task of a neuron in a pooling layer simply consists in propagating the first spike received from a receptive window of the previous convolutional layer. Neurons in all the convolutional layers are non-leaky integrate and fire neurons. They integrate input spikes and emit a spike as soon as they reach their threshold. The latter is a hyperparameter to set. Immediately after a spike occurs, weights are updated accordingly, using a simplified version of the STDP learning rule. Let $i, j$ be the indices of the post and pre-synaptic neurons, respectively, and let $t_i, t_j$ be their corresponding spike times. The synaptic weight $w_{ij}$ is updated by adding a modification factor $\Delta_{ij}$ computed as follows according to a simplified version of STDP [12], [107].

$$\Delta_{ij} = \begin{cases} \alpha^+ \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j \leq t_i \\ -\alpha^- \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j > t_i \end{cases}$$

$\alpha^+, \alpha^- \in \mathbb{R}_{\geq 0}$ are two parameters that specify the learning rate or by how much the weights are changed. The latter factor impacts a lot the learning. Indeed, small values would lead to a slow learning process, they simulate a neural network that is confident in its prior “beliefs and decisions” (weights). High values allow to learn very fast information about the current stimuli, but they can have as a consequence to “forget” what they learned with previous stimuli. Note that this simplified version of the STDP rule does not take into account the absolute time difference between post and pre-synaptic spikes. What matters instead is only the order, or the sign of the difference. In practice, this is not a problem for our model.

This feature learning process goes on, by propagating training images one by one. Each time the different spiking waves of a new image have been fully processed, and weights updated and stored, the potential of each neuron is reset to 0, preparing for the next image. Initially, the synaptic weights are chosen at random from a normal distribution with some mean and standard deviation. The STDP rule ensures that they always remain in the range $[0, 1]$. Within each image, learning is done layer by layer: learning at layer $\ell$ begins when learning at layer $\ell - 1$ has terminated.

The intent of this feature learning phase is to learn the synaptic weights of each neuron in all the convolutional layers. As observed previously with this SNN model [12], neurons in the first layer converged to the simple four oriented edges. Neurons in the successive layers learned more complex ones by integrating spikes from previous layers. We remind that this phase is totally unsupervised as the network only learns frequent features associated with images and it requires no knowledge about the input image categories. The next learning step includes such categories, we qualify it as semi-supervised.

2) Co-occurrence learning: Once the SNN has learned the right weights and hence visual features, the second co-occurrence learning step can begin. For this step, we assume as per our hypothesis, that there is a layer which encodes the object categories or names (equivalent of sparse neurons that respond to image or sound signifiers or both). The latter layer is connected, as shown in the figure, to the last layer of the image processing network. Then, an image (e.g. a cat as illustrated in the figure) is propagated through the image processing neural network, while at the same time, the...
1. Feature learning (STDP)

![Diagram of feature learning (STDP)]

2. Co-occurrence learning (STDP)

![Diagram of co-occurrence learning (STDP)]

3. Naming (Backpropagated APs)

![Diagram of naming (Backpropagated APs)]

Fig. 2: Learning under our hypothesis, simulated with the proposed SDNN with its three main parts.

We simulate a single neuron for simplicity, but same reasoning applies multiple ones

In more details, as in the previous phase, train images are considered one by one. Using the weights learned in the first phase, each image passes through the network until it reaches the last layer where a max pooling operation is performed. During co-occurrence learning, a neuron of the last pooling layer would thus receive two spikes: one propagated by the neuron associated to the class of the image, and one input from the last convolutional layer. This entire process simulates the teacher that simultaneously shows the image and its name.

Note that, implementation-wise [15], this is equivalent to having a matrix with the same shape as the last pooling layer that is associated to each image category (one weight per category neuron per neuron in the backpropagation layer). As in the first phase, weights are initially random and are updated only using the STDP learning rule, as defined in [1]. As a result of this STDP learning, according to the order of post and pre-synaptic spike times and to the index of the spiking neuron, what will happen is the following: the weight matrix of the correct image category is always strengthened (LTP), while the weight matrices of the other categories will be weakened (LTD).

Contrarily to the previous one, this phase is supervised in that it requires knowledge of the image category in order to be able to link it with the corresponding image features. Indeed, the aim here is to learn associations between the previously learned features and image categories by using only the simple STDP rule. At the end of the second phase, training has completed and the SNN can proceed to the naming task using the backpropagation principle.

One extreme version of this second phase is what is called one-shot learning: the network is given only a single example of each category. We will vary the number of such training examples in Sec. [V] effectively trying one-shot and few-shot learning scenarios, hence why we consider this task to be “semi-supervised”.

3) Naming: Once the learning is done with the previous two phases, we are ready now for the naming task, following the principles of backpropagated action potentials.

In this task, the image to name is first propagated through the fully trained neural network until spikes start to happen in the last pooling layer, which is our backpropagation root layer. We consider that all neurons in this layer are source pointer neurons. This means, as per our hypothesis, that we allow them (whenever they fire) to send backpropagated action potentials, modulated by the presynaptic weights learned in the previous second step, to the previous layer that encodes the labels or names. Neurons in the “categories/signifiers” layer will integrate such received signals and the category which has the highest vote is the retained name for the image. Namely, let $C_i$ for $i = 1, ..., k$ be the class associated to the neuron with the highest class score. Then, if $i$ is chosen to be the class the image belongs to. It is this score that can be used as the accumulated potential that brings the right neuron closer to its firing threshold, leading when it fires to the class decision $[14]$

We show later in the next section that such a simple mechanism allows to label the images as accurately as the SVM. More interestingly, by using high learning rates during the first co-occurrence (e.g. neuromodulating to increase synaptic strength), it is possible to learn to name objects, with

In practise, implementation-wise, the score is simply the sum of weights. Indeed, the SNN model considers a spike to be a binary decision that happens when the accumulated potential reaches a threshold (i.e. we do not consider the rate). We tried a version that uses the exact value of the internal potential instead of the unit value 1 but results were similar.
maximum accuracy, by showing the neural network only a single instance of the image class; an extreme learning task in which the SVM classifier seems to have more difficulty.

V. SIMULATION RESULTS

We now evaluate the accuracy of the SNN model above when using our hypothesis to learn and name (the three steps described in Fig. 2). We compare its accuracy to that of Kheradpisheh et al. [12] (the same feature learning as step 1 above, followed by an SVM classifier on the feature vectors).

A. Experimental Setup

Experiments have been performed on a server with 5 Intel 2.10 GHz CPU, 32 GB of memory and a GPU Nvidia Tesla P100 SXM2 with 16 GB of dedicated memory. We evaluate the accuracy reached by our model using the Caltech motor face dataset [108] considering two classes: Faces and Motorbikes.

B. Overall accuracy with backpropagated APs

We first focus on the overall accuracy and the case where we show the SNN many examples of each class during learning. In particular, for each class, we select 398 images among which 200 are reserved for training (feature learning and co-occurrence learning likewise) and 198 are left for testing. After a search of parameters [15], we set $\alpha^+$ and $\alpha^-$ to 0.007 and 0.003, respectively. The firing thresholds for the first, second, and third convolutional layers are set to 6, 21, and 10. Similar to prior work [12], max pooling is not performed for this Caltech dataset because images have low resolutions. Our last layer uses thus a pooling window of size 1x1. Finally, the synaptic weights of the class matrices are initialized at random from a normal distribution with mean 0.5 and standard deviation 0.05.

In this setting, using the backpropagation-based recollection, we reach an accuracy of 96.7 and 95.7% on the train and test datasets, respectively. This performance is on par with that of the SVM classifier (96%), thus confirming the computational efficiency of backpropagated action potentials.

In more details, Fig. 3 shows the class scores for both train and test datasets. Recall that these scores correspond to the “votes” received by backpropagation from the last layer. Each point in the figure corresponds to an image with motorbike points placed on the left and faces on the right. Values in the ordinate represent the difference between the scores associated to the Motorbike and Face classes. Therefore, images with positive values are associated to the Motorbike class, while images with negative values are associated to the Face one. As it can be observed, backpropagated action potentials allow to separate the two classes in a net way for most of the images. However, for some of the images, this distinction is not clear. We believe as we discussed earlier that the problem comes from the SNN feature extractor which, although performing, does not yet learn good representations. Nonetheless, only few images are classified incorrectly with a large relative error. Such a good performance shows overall the computational plausibility of the backpropagation-based recollection.

C. Accuracy in few-shot learning

We now focus on the case where the “teacher” shows the SNN only few examples.

Varying the Number of Images in co-occurrence learning

We first vary the number of labeled examples given in the co-occurrence based learning phase. In the remainder, whenever we talk about training, we refer to the supervised “teacher-based” co-occurrence learning where a label is given. Fig. 4 shows the accuracy on the train and test datasets as a function of the number of labeled train images. We vary the number of labeled train images per experiment from 5 to 200. As the number of train images increases, the accuracies on both the train and test images increase as well. The train and test scores pass from 90% and 85% after 25 images per category up to the 96.75% and 95.7%, from above, after having used the full train dataset. This shows that the SNN learns but slowly as we feed images and labels. One way however to speed up this process is to increase the learning rates.

Varying the learning rate.

Increasing the learning rate can simulate a neuromodulatory action that strengthens a connection suddenly, without the need for repeated exposure. We hence vary $\alpha^+$ and $\alpha^-$ used in the co-occurrence phase and observe a considerable impact on the accuracy. Fig. 5 illustrates the impact of the modulated learning rate on the accu-
accuracy level reached as a function of the number of train images used. As a baseline, we use $\alpha^+ = 0.007, \alpha^- = 0.003$ and we multiply both values by some factor $\lambda \in \{\frac{1}{10}, 1, 2, 3, 5, 10\}$.

When $\lambda < 1$, the learning is obviously slower. Indeed, the accuracy varies from 50% with 0 training images (i.e. random), to 85% using the entire train dataset and grows in a linear way. It will probably reach higher values with more training time. When $\lambda = 1$, we reach the maximum possible test score using all the train images.

An interesting behaviour can be observed when $\lambda > 1$. The learning is at first faster, as it can reach high accuracy after having seen only a small sample of train images. However, it then starts decreasing as the number of train images increases. We recall that the used STDP rule keeps the weights within the range $[0, 1]$. Thus, starting with high values for $\alpha^+$ and $\alpha^-$ allows to faster associate discriminant features with the images categories. However, as the number of train images increases, weights might become less helpful if they tend to reach the maximum value of 1 and to have thus smaller intermediate values. The scores would become closer and less distinguishable. Another factor is that, as we will see later, the SNN is better at recognizing and learning from certain particular images compared to others (see description of Fig. 6). Hence, being exposed to a good image with a higher learning rate leads to reaching a good accuracy. But being exposed afterwards to a “bad” image will lead to unlearning the good weights, thus decreasing the performance.

When $\lambda = 10$ and with only 20 train images per category, it is possible to reach an accuracy of 95.5%. With $\lambda = 2, 3,$ and 5, the number of necessary train images per category to reach the same accuracy are 45, 75, and 100, respectively. These results hint to the following direction: the “best” approach in terms of few-shot learning would be to first start with a high learning rate or $\lambda$ but then to stop changing the weights by either suddenly decreasing $\lambda$, or by simply freezing the learning and making the network always stick to the old beliefs.

D. One-shot learning: Machine learning vs. Backpropagation

In this final section, we set the bar high and propose to train and test the SNN in a one shot learning task, meaning that we show the neural network only one single image from each class, together with their correct names. We then test the accuracy of the network on the entire 198 images of the test set. We compare SVM and our model on this task.

As explored earlier, reaching good performance in this task goes through even higher learning rates than tried previously. We experiment with various $\lambda$s and various (motorbike,face) image couples. We find for instance that $\lambda = 65$ yields good results among many other values, so we pick it. But this is where we interestingly discover that the performance highly depends on which couple of (motorbike,face) were used for the one-shot learning task. We found that, using the backpropagation-based recollection, certain single couples of motorbike and face yielded an accuracy of 96.2% on all the remaining unseen 198 test images. Note that this is higher than what we achieved earlier when training with all images and not only one or few shots (we obtained 95.7% on test set). At the same time, the maximum we could achieve with the SVM classifier on a single example was 84.5% of accuracy.

To assess this more systematically, we test both SVM and backpropagation in the one-shot exercise on more than 1500 different image couples of motorbike and face photos. We plot in Fig. 6 the resulting empirical cumulative distribution functions. The figure shows that both approaches yield different distributions, with most of SVM results being less dispersed around a little higher than 80% accuracy. The observed curious disparity between pairs of images suggests that the SNN models are still not good enough in feature extraction. The learned representations are probably not as invariant as they should be, hence the variability and differences between images. Further investigating the differences between these successful and less successful image couples might help enhancing current SNN models.

For these reasons, our simulations should be viewed only as an additional argument that shows the computational effectiveness of the backpropagation-based recollection mechanism, compared to an SVM classifier on the same extracted visual features. Our results above show indeed that with the right images used for training, backpropagation outperforms by far the SVM in the one shot task (96.2% accuracy on 198 images against only 84.5%).

VI. DISCUSSION

For more than a century [109], information processing in the brain has been mainly believed to follow the forward pre to post-synaptic neurons direction. In this work, we emitted the hypothesis that the backpropagation of action potentials in the reverse direction mediates all “offline” generative tasks where
the simultaneous activation of specific targeted populations of neurons is needed. This is, we claimed, the case of the retrieval of past memories or mental images, retrieval of signification of words, retrieval of names and even the mixture of distinct past memories into imagination. We reviewed in sec. III-A evidence that calls for taking the hypothesis seriously (see summary in Sec. III-E). As an added bonus, we computationally showed in Sec. VII that our hypothesis can be more efficient than a machine learning algorithm in retrieving the category name of an object. If confirmed true, this would have tremendous implications, considerably improving our understanding of neural encoding and how high cognitive functions are implemented from a low-level neural perspective.

A. Possible implications

The first big implication of this hypothesis is the promise to bring answers to the neural encoding problem, closing the old debate in cognitive sciences between localist and distributed representation theories. If our hypothesis is true, the answer to the representation problem becomes simpler: (i) representations of concepts are distributed over a population of neurons but at the same time, (ii) there exists highly selective neurons that respond to unique concepts, but that serve a specific retrieval purpose. The latter act as hubs between various related concepts, playing the role of source pointer neurons to retrieve the entire concept’s features (encoded by an entire population of neurons). For example, there should exist relatively few neurons that selectively respond to the image of a cat. But such neurons serve as pointers or hubs to easily connect such image to other related memories. Similar sparse neurons act for instance as pointers to retrieve the visual features of a cat through backpropagated APs that travel backwards to reactivate selectively the neurons that represent

the right lines, shapes[^17] and colours that define a cat. Hence, to be able to recall through mental visual imagery the image of a cat, the brain does not need to activate only the sparse neurons that respond to “cats”, but, as the optogenetic studies above also found, the activation of an entire larger population is needed. And if neurons that represent, say, vertical lines, are not activated during this process, the recalled mental image will lack them.

Now beyond vision, the cat’s toy example could apply to any mental state and any couple of (lived stimuli, later memory of that stimuli), let them be smells, affects or even impressions of movements. In accordance with the principle of grounded cognition [110], for which we believe our hypothesis framework applies, discrete concepts are grounded in the sensorimotor experiences that were encoded with them, such that the activation of a signifier of a concept leads to the activation of the experiences that are grounded with it. Hence, the cue that is the words “moving” or “tickling” is correlated with areas that encode actual moving or tickling.

Next, another related hard problem in cognitive sciences that can benefit from our hypothesis is the binding problem or how does the brain binds higher-level concepts to more elementary ones and particularly, how does it associate the right features (e.g. colors) to the right discrete objects or concepts, for example in an image composed of many objects. The fact that the brain needs some time to correctly perform the binding [111] suggests that this operation is not forward-based but generatively and iteratively happens later in a second stage. If our hypothesis is correct, this should happen through slow and repetitive runs of top down action potential backpropagation, starting from the right source pointer neurons that define uniquely the discrete object, all the way backwards, activating all the neurons that describe its attributes. In fact, two famous competing (and high level) theories on this problem are still the feature-integration theory [112] and temporal synchronization theory [113], [114]. Our hypothesis offers perspectives to reconcile them as well. Indeed, both admit the involvement of different runs of bottom up and top down hierarchies (e.g. attention in feature integration theory) to implement the binding problem. However, the exact physical mechanisms that would implement them are still unknown today. Our hypothesis constitutes a good candidate for such mechanisms, with backpropagated APs implementing such top down hierarchies to bind objects to their features. Moreover, under this new realm, one does not have anymore to chose exclusively between binding-by-synchrony and feature-integration theory: attention with backpropagated APs could synchronously activate selectively all the features related to a given discrete object. An implicit consequence of this is that attention itself is likely implemented through top-down backpropagated APs.

In general, backpropagation could constitute an ubiquitous and easy-to-implement simple mechanism that underlies a
diverse set of cognitive tasks such as offline thinking or mind wandering, imagination, episodic memory retrieval and future episodic thinking. In our framework, imagination would be “simply” the resulting activation patterns of a mixture of usually unrelated concepts: for example, an imagined “laughing cat” results from simultaneously back-activating the “laughing” and “cat” concept neurons. The same applies for mind wandering, where backpropagated APs should induce activation patterns that follow “the most likely neural pathways”, thus generating what seems to be coherent thoughts.

Our hypothesis assumes the existence of (backpropagation root) areas where the flow reverses during cue-based retrieval (from the cue recognition networks, back to the networks of the memory trace). Our review of the literature confirms that a similar reversal happens in the MTL and the hippocampal formation. However, all this predicts also the existence of generators, or specialized centers, that release neuromodulators remotely to control the generation process by either inhibiting or facilitating backpropagation at such root layers. This is where more advanced modeling and computer simulations could help tremendously in future work. For example, it would be helpful to understand the interplay between backpropagation and forward propagation, since the first can result as well in feedforward propagation, that in turn might cause backpropagation etc. Such “lateral” activation patterns could be useful to find associated concepts, as opposed to “digging into the details” of a single concept.

Moreover, if proven true, our hypothesis would open new ways to better understand some recollection-related pathologies. In this case, understanding what factors impact the inhibition or excitation of backpropagated APs could help understanding probably-related disorders such as obsessive thinking or intrusive thoughts. Similar (rarer) dysfunctions happen in mental visual imagery as well, such as the absence (aphantasia) or excess (hyperphantasia) of visual imagery experiences [4]. In all these cases, the neuromodulation mechanisms that control backpropagation should be first suspected.

Last and not least, another implication is that the retrieval process is stochastic in nature: the retrieved memory trace looks like the original first perception but, depending on past experiences (and hence the weights of the neural connections due to past experiences), the reactivation might not be exactly the same. This can be seen most in the case of language where the same word (e.g. “a screen”) was seen multiple times during encoding, in the presence of multiple similar stimuli (e.g. many different types of screens), and where the same concept is further grounded in different neural network connections from one subject to the other. Such considerations offer new perspectives to apprehend language as a system of common cues, useful to make others live intended target experiences through the invocation of the right cues.

B. Verifiability and further investigations

But before drawing a bright future to our hypothesis, at least two directions must be seriously taken to further investigate it and confirm its plausibility or infirm it.

1) Empirical methods: This is perhaps the most important direction. We verified in the literature the existence of retrograde signals that satisfy many of the assumptions of our hypothesis. Further targeted empirical studies must verify the remaining ones so as to support or invalidate our hypothesis. In particular, first, it is crucial to understand if the backpropagating signal is stronger on paths with “higher synaptic weights”. Second, it is necessary to measure how far-reaching the backpropagating signal can be, beyond solely the previous connection. In our literature review, we found no evidence for nor against the latter points.

2) Computational effectiveness: Another line of work could verify in parallel the computational effectiveness of our hypothesis in implementing its other target goals. Here, we verified (as a first step) the ability of retrograde APs to perform the object recognition or the naming task. We opted for this comparison because of the presence of a baseline to compare to (an existing image classifier) and a metric (accuracy) to quantify the computational power of the mechanism. Not simulated in this work, retrograde APs can be used symmetrically for the task of “understanding”: i.e. an activation of the signifier (word) neuron that leads automatically to the activation of the signified concept and its, say, visual features. This needs however, to rely on Artificial Neural Networks that have good and plausible feature extraction capabilities. The SNNs we use are promising and close but they do not yet fully satisfy the last property (see Sec. IV-B).

Finally, in the event that such artificial backward reconstruction works successfully with some SNN, our hypothesis could become then computationally verifiable for the imagination task. One interesting experiment could be to train artificial neural networks to recognize two separate concepts from images, exactly as we do for ”Motor” and ”Face” in Sec. V. Then, instead of backward constructing only one concept at a time, it would be interesting to simultaneously activate two concepts and see the effect on the backward reconstructed images. As exemplified above, one could activate a concept like “laughing” and a concept like ”cat” and visualize the results of the “competition” between backpropagating signals on the backward-constructed images. Note that similar work can be achieved with today’s state of the art deep neural networks such as GPT-3’s DALL-E [11] which uses a transformer decoder architecture. However, the latter employs supervised mechanisms that lack biological plausibility [9], [10], [11], which is our focus in this work.

ACKNOWLEDGEMENT

The author would like to acknowledge the contributions of Dr. Andrea Tomassilli for his help in the setup of the experiments and analysis of the results during his internship in fall 2018. We are grateful to Dr. Alessandro Finamore for interesting feedback on an earlier draft of the paper.

REFERENCES

[1] F. De Saussure, Course in general linguistics. New York: McGraw-Hill., 1959.
robust activity marking system for exploring active neuronal ensembles,” *Elife*, vol. 5, p. e33904, 2016.

[49] A. F. Lacagnina, E. T. Brockway, C. R. Crovetti, F. Shue, M. J. McCurry, K. P. Satter, S. C. Lim, S. L. Santos, C. A. Denny, and M. R. Drew, “Distinct hippocampal engrams control extinction and relapse of fear memory,” *Nature neuroscience*, vol. 22, no. 5, pp. 753–761, 2019.

[50] O. Khalaf, S. Resch, L. Dixsaut, V. Gorden, L. Glauser, and J. Gräff, “Reactivation of recall-induced neurons contributes to remote fear memory attenuation,” *Science*, vol. 360, no. 6394, pp. 1239–1242, 2018.

[51] S. Ramirez, X. Liu, P.-A. Lin, J. Suh, M. Pignatelli, R. L. Redondo, T. J. Ryan, and S. Tonegawa, “Creating a false memory in the hippocampus,” *Science*, vol. 341, no. 6144, pp. 387–391, 2013.

[52] K. K. Taylor, K. Z. Tanaka, L. G. Reijmers, and B. J. Witgen, “Reactivation of neural ensembles during the retrieval of recent and remote memory,” *Current Biology*, vol. 23, no. 2, pp. 99–106, 2013.

[53] A. Guskolen, J. W. Kenney, J. de la Parra, B.-r. A. Yeung, S. A. Josselyn, and P. W. Frankland, “Recovery of “lost” infant memories in mice,” *Current Biology*, vol. 28, no. 14, pp. 2283–2290, 2018.

[54] G. T. Walhauser, V. Braun, and S. Hanslmayr, “Episodic memory retrieval functionality relies on very rapid reactivation of sensory information,” *Journal of Neuroscience*, vol. 36, no. 1, pp. 251–260, 2016.

[55] A. Iafarpour, L. Fuentemilla, A. J. Horner, W. Penny, and E. Duzel, “Replay of very early encoding representations during recollection,” *Journal of Neuroscience*, vol. 34, no. 1, pp. 242–248, 2014.

[56] J. D. Johnson, S. G. McDuff, M. D. Rugg, and K. A. Norman, “Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis,” *Nature*, vol. 63, no. 5, pp. 697–708, 2009.

[57] J. R. Manning, S. M. Polyn, G. H. Balthuch, B. Litt, and M. J. Kahana, “Oscillatory patterns in temporal lobe reveal context reinstatement during memory search,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 31, pp. 12893–12897, 2011.

[58] M. Ritchey, E. A. Wing, K. S. LaBar, and R. Cabeza, “Neural similarity between encoding and retrieval is related to memory via hippocampal interactions,” *Cerebral cortex*, vol. 23, no. 12, pp. 2818–2828, 2013.

[59] B. P. Staresina, R. N. Henson, N. Kriegeskorte, and A. Alink, “Episodic reinstatement in the medial temporal lobe,” *Journal of Neuroscience*, vol. 32, no. 50, pp. 18150–18156, 2012.

[60] R. B. Yaffe, M. S. Kerr, S. Damera, S. V. Sarma, S. K. Inati, and K. A. Zaghloul, “Reinstatement of distributed cortical oscillations occurs with precise spatiotemporal dynamics during successful memory retrieval,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, pp. 18727–18732, 2014.

[61] J. Fuldor, F. Milton, D. Salas, A. Smith, A. Simler, C. Winlove, and A. Zeman, “The neural correlates of visual imagery vividness—an fmri study and literature review,” *Cortex*, vol. 105, pp. 26–40, 2018.

[62] N. Dijkstra, S. E. Bosch, and M. A. van Gerven, “Vividness of visual imagery depends on the neural overlap with perception in visual areas,” *Journal of Neuroscience*, vol. 37, no. 5, pp. 1367–1373, 2017.

[63] N. Dijkstra, S. E. Bosch, and M. A. van Gerven, “Shared neural mechanisms of visual perception and imagery,” *Trends in cognitive sciences*, 2019.

[64] N. Dijkstra, P. Mostert, F. P. de Lange, S. Bosch, and M. A. van Gerven, “Differential temporal dynamics during visual imagery and perception,” *Elife*, vol. 7, p. e33904, 2018.

[65] M. St-Laurent, H. Abdi, and B. R. Buchsbaum, “Distributed patterns of reaction prediction vividness of recollection,” *Journal of Cognitive Neuroscience*, vol. 27, no. 10, pp. 2000–2018, 2015.

[66] K. J. Friston, L. Harrison, and W. Penny, “Dynamical causal modelling,” *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[67] S. Hochstein and M. Aihussar, “View from the top: Hierarchies and reverse hierarchies in the visual system,” *Neuron*, vol. 36, no. 5, pp. 791–801, 2002.

[68] T. Serre, A. Oliva, and T. Poggio, “A feedforward architecture accounts for rapid categorization,” *Proceedings of the national academy of sciences*, vol. 104, no. 15, pp. 6424–6429, 2007.

[69] J. Linde-Domingo, M. S. Treder, C. Kerrén, and M. Wibmer, “Evidence that neural information flow is reversed between object perception and object reconstruction from memory,” *Nature communications*, vol. 10, pp. 1–13, 2019.

[70] J. Fulford, F. Milton, D. Salas, A. Smith, A. Simler, C. Winlove, and A. Zeman, “The neural correlates of visual imagery vividness—an fmri study and literature review,” *Cortex*, vol. 105, pp. 26–40, 2018.

[71] R. C. Decharms and A. Zador, “Neural representation and the cortical code,” *Annual review of neuroscience*, vol. 23, no. 1, pp. 613–647, 2000.

[72] H. B. Barlow, “Single units and sensation: a neuron doctrine for perceptual psychology,” *Perception*, vol. 1, no. 4, pp. 371–394, 1972.

[73] B. A. Olshausen and D. J. Field, “Sparse coding of sensory inputs,” *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[74] D. H. Hubele and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of Physiology*, vol. 160, pp. 1–126, 1962.

[75] M. Vishkin, L. G. Ungerleider, and K. A. MacKo, “Object vision and spatial vision: two cortical pathways,” *Trends in neurosciences*, vol. 6, pp. 414–417, 1983.

[76] C. G. Gross, D. B. Bender, and C. D. Rocha-Miranda, “Visual receptive fields of neurons in inferotemporal cortex of the monkey,” *Science*, vol. 166, no. 3910, pp. 1303–1306, 1969.

[77] J. Tanaka, “Inferotemporal cortex and object vision,” *Annual review of neuroscience*, vol. 19, no. 1, pp. 109–139, 1996.

[78] N. K. Logothetis and D. L. Sheinberg, “Visual object recognition,” *Annual review of neuroscience*, vol. 19, no. 1, pp. 577–621, 1996.

[79] E. I. Moser, E. Kropff, and M.-B. Moser, “Place cells, grid cells, and the brain’s spatial representation system,” *Annu. Rev. Neurosci.*, vol. 31, pp. 69–89, 2008.

[80] I. Fried, K. A. MacDonald, and C. L. Wilson, “Single neuron activity in human hippocampus and amygdala during recognition of faces and objects,” *Neuron*, vol. 18, no. 5, pp. 753–765, 1997.

[81] G. Kreiman, C. Koch, and I. Fried, “Category-specific visual responses of single neurons in the human medial temporal lobe,” *Nature neuroscience*, vol. 3, no. 9, pp. 946–953, 2000.

[82] R. Q. Quiroga, G. Kreiman, C. Koch, and I. Fried, “Sparse but not ‘grandmother-cell’ coding in the medial temporal lobe,” *Trends in cognitive sciences*, vol. 12, no. 3, pp. 87–91, 2008.

[83] S. Waydo, A. Krasok, R. Q. Quiroga, I. Fried, and C. Koch, “Sparse representation in the human medial temporal lobe,” *Journal of Neuroscience*, vol. 26, no. 40, pp. 10232–10234, 2006.
R. Q. Quiroga, A. Kraskov, F. Mormann, I. Fried, and C. Koch, “Single-cell responses to face adaptation in the human medial temporal lobe,” *Neuron*, vol. 84, no. 2, pp. 363–369, 2014.

L. Reddy and S. J. Thorpe, “Concept cells through associative learning of high-level representations,” *Neuron*, vol. 84, no. 2, pp. 248–251, 2014.

K. Christoff, Z. C. Irving, K. C. Fox, R. N. Spreng, and J. R. Andrews-Hanna, “Mind-wandering as spontaneous thought: a dynamic framework,” *Nature Reviews Neuroscience*, vol. 17, no. 11, pp. 718–731, 2016.

A. Kucyi, “Just a thought: How mind-wandering is represented in dynamic brain connectivity,” *Neuroimage*, vol. 180, pp. 505–514, 2018.

D. R. Addis, A. T. Wong, and D. L. Schacter, “Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration,” *Neuropsychologia*, vol. 45, no. 7, pp. 1363–1377, 2007.

A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. S. Maida, “Deep learning in spiking neural networks,” *arXiv preprint arXiv:1804.08150*, 2018.

W. Gerstner and W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

S. Ghosh-Dastidar and H. Adeli, “Spiking neural networks,” *International journal of neural systems*, vol. 19, no. 04, pp. 295–308, 2009.

M. Taylor, “The problem of stimulus structure in the behavioural theory of perception,” *South African Journal of Psychology*, vol. 3, pp. 23–45, 1973.

N. Caspore and Y. Dan, “Spike timing–dependent plasticity: a hebbian learning rule,” *Annu. Rev. Neurosci.*, vol. 31, pp. 25–46, 2008.

S. Huang, C. Rozas, M. Trevino, J. Contreras, S. Yang, L. Song, T. Yoshioka, H.-K. Lee, and A. Kirkwood, “Associative hebbian synaptic plasticity in primate visual cortex,” *Journal of Neuroscience*, vol. 34, no. 22, pp. 7575–7579, 2014.

D. B. McMahon and D. A. Leopold, “Stimulus timing-dependent plasticity in high-level vision,” *Current biology*, vol. 22, no. 4, pp. 332–337, 2012.

J. Ott, E. Linstead, N. LaHaye, and P. Baldi, “Learning in the machine: To share or not to share?,” *Neural Networks*, 2020.

T. Masquelier and S. J. Thorpe, “Unsupervised learning of visual features through spike timing dependent plasticity,” *PLoS Comput Biol*, vol. 3, no. 2, p. e31, 2007.

L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.

G. Berlucchi, “Some aspects of the history of the law of dynamic polarization of the neuron. from william james to sherrington, from cajal and van gehuchten to golgi,” *Journal of the History of the Neurosciences*, vol. 8, no. 2, pp. 191–201, 1999.

L. W. Barsalou, “Grounded cognition,” *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, 2008.

C. Von der Malsburg, “The what and why of binding: the modeler’s perspective,” *Neuron*, vol. 24, no. 1, pp. 95–104, 1999.

A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

P. M. Milner, “A model for visual shape recognition,” *Psychological review*, vol. 81, no. 6, p. 521, 1974.

M. N. Shadlen and J. A. Movshon, “Synchrony unbound: a critical evaluation of the temporal binding hypothesis,” *Neuron*, vol. 24, no. 1, pp. 67–77, 1999.

A. Ramesh, M. Pavlov, G. Goh, and S. Gray, “Dall-e: Creating images from text.” https://openai.com/blog/dall-e/. Accessed: 2021-01-06.