Neural networks grown and self-organized by noise

Guruprasad Raghavan
Department of Bioengineering
Caltech
Pasadena, CA 91125
graghava@caltech.edu

Matt Thomson
BBE
Caltech
Pasadena, CA 91125
mthomson@caltech.edu

Abstract

Living neural networks emerge through a process of growth and self-organization that begins with a single cell and results in a brain, an organized and functional computational device. Artificial neural networks, however, rely on human-designed, hand-programmed architectures for their remarkable performance. Can we develop artificial computational devices that can grow and self-organize without human intervention? In this paper, we propose a biologically inspired developmental algorithm that can 'grow' a functional, layered neural network from a single initial cell. The algorithm organizes inter-layer connections to construct a convolutional pooling layer, a key constituent of convolutional neural networks (CNN's). Our approach is inspired by the mechanisms employed by the early visual system to wire the retina to the lateral geniculate nucleus (LGN), days before animals open their eyes. The key ingredients for robust self-organization are an emergent spontaneous spatiotemporal activity wave in the first layer and a local learning rule in the second layer that 'learns' the underlying activity pattern in the first layer. The algorithm is adaptable to a wide-range of input-layer geometries, robust to malfunctioning units in the first layer, and so can be used to successfully grow and self-organize pooling architectures of different pool-sizes and shapes. The algorithm provides a primitive procedure for constructing layered neural networks through growth and self-organization. Broadly, our work shows that biologically inspired developmental algorithms can be applied to autonomously grow functional 'brains' in-silico.

1 Introduction

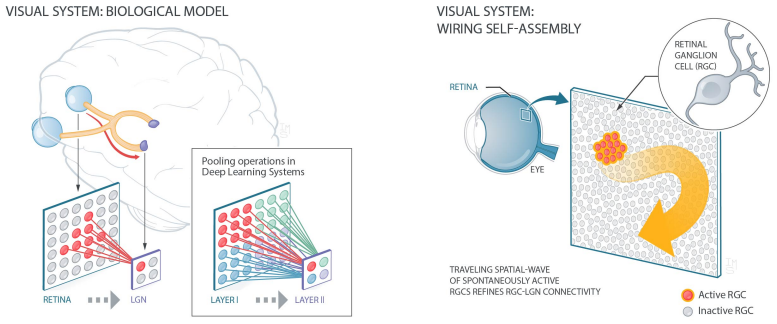
Living neural networks in the brain perform an array of computational and information processing tasks including sensory input processing [1, 2], storing and retrieving memory [3, 4], decision making [5, 6], and more globally, generate the general phenomena of "intelligence". In addition to their information processing feats, brains are unique because they are computational devices that actually self-organize their intelligence. In fact brains ultimately grow from single cells during development. Engineering has yet to construct artificial computational systems that can self-organize their intelligence. In this paper, inspired by neural development, we ask how artificial computational devices might build themselves without human intervention.

Deep neural networks are one of the most powerful paradigms in Artificial Intelligence. Deep neural networks have demonstrated human-like performance in tasks ranging from image and speech recognition to game-playing [7, 8, 9]. Although the layered architecture plays an important role in the success [10] of deep neural networks, the widely accepted state of art is to use a hand-programmed network architecture [11] or to tune multiple architectural parameters, both requiring significant engineering investment. Convolutional neural networks, a specific class of DNNs, employ a hand

Preprint. Under review.

programmed architecture that mimics the pooling topology of neural networks in the human visual system.

Here, we develop strategies for *growing a neural network* autonomously from a single computational "cell" followed by *self-organization* of its architecture by implementing a wiring algorithm inspired by the development of the mammalian visual system. The visual circuity, specifically the wiring of the retina to the lateral geniculate nucleus (LGN) is stereotypic across organisms, as the architecture always enforces pooling (retinal ganglion cells (RGC's) pool their inputs to LGN cells) and retinotopy. The pooling architecture (figure-1a) is robustly established early in development through the emergence of spontaneous activity waves (figure-1b) that tile the light insensitive retina [12]. As the synaptic connectivity between the different layers in the visual system get tuned in an activity-dependent manner, the emergent activity waves serve as a signal to alter inter-layer connectivity much before the onset of vision.



(a) (b)

Figure 1: **Wiring of the visual circuitry** (a) Spatial pooling observed in wiring from the retina to LGN and in CNN's. (b) Synchronous Spontaneous bursts (retinal waves) in the light-insensitive retina serve as a signal for wiring retina to the brain.

The main contribution of this paper is that we propose a developmental algorithm inspired by visual system development to *grow and self-organize a pooling architecture*, a key feature of the convolutional neural network (CNN). Once a pooling architecture emerges, any non-linear function can be implemented by units in the second layer to morph it into functioning as a convolution or a max/average pooling. We show that our algorithm is adaptable to a wide-range of input-layer geometries, robust to malfunctioning units in the first layer and can grow pooling architectures of different shapes and sizes, making it capable of countering the key challenges accompanying growth. We also demonstrate that 'grown' networks are functionally similar to that of hand-programmed pooling networks on conventional image classification tasks. As CNN's represent a model class of deep networks, we believe the developmental strategy can be broadly implemented for the self-organization of intelligent systems.

2 Related Work

Self-organization of neural networks dates back many years, with the first demonstration being Fukushima's neocognitron [13, 14], a hierarchical multi-layered neural network capable of visual pattern recognition through learning. Although weights connecting different layers were modified in an unsupervised fashion, the network architecture was hard-coded, inspired by Hubel and Wiesel's [15] description of simple and complex cells in the visual cortex. This development inspired modern-day convolutional neural networks (CNN) [16]. Although CNN's performed well on image-based tasks, they had a fixed, hand-designed architecture whose weights were altered by back-propagation. This changed with the advent of neural architecture search [17], as neural architectures became malleable to tuning by neuro-evolution strategies [18, 19, 20], reinforcement learning [21] and multi-objective searches [22, 23]. These strategies have been successful in training networks that perform significantly much better on CIFAR-10, CIFAR-100 and Image-Net datasets. As the objective function being maximized is the predictive performance on these datasets the networks evolved may not generalize well to multiple datasets. On the contrary, biological neural networks in the brain grow

architecture that can generalize very well to innumerable datasets. Neuroscientists have been very interested in how the architecture in the visual cortex emerges during brain development. Meister et al [12] suggested that spontaneous and spatially organized synchronized bursts prevalent in the developing retina guide the self-organization of cortical receptive fields. In this light, mathematical models of the retina and its emergent retinal waves were built [24], and analytical solutions were obtained regarding the self-organization of wiring between the retina and the LGN [25, 26, 27, 28, 29]. These models have been essential for understanding how self-organization functions in the brain, but haven't been generalized to growing complex architectures that can compute. One of the most successful attempts at growing a 3D model of neural tissue from simple precursor units was demonstrated by Zubler et. al [30] that defined a set of minimal rules that could result in the growth of morphologically diverse neurons. Although their networks were grown from single units, they weren't functional as they weren't equipped to perform any task. To bridge this gap, in this paper we attempt to grow and self-organize functional neural networks from a single precursor unit.

3 Bio-inspired developmental algorithm

In our procedure, the pooling architecture emerges through two processes, growth of a layered neural network followed by self-organization of its inter-layer connections to form defined 'pools' or receptive fields. As the protocol for growing a network is relatively straightforward, our emphasis in the next few sections is on the self-organization process, following which we will combine the growth of a layered neural network with its self-organization in the penultimate section of this paper.

We, first, abstract the natural development strategy as a mathematical model around a set of input sensor nodes in the first layer (similar to retinal ganglion cells) and processing units in the second layer (similar to cells in the LGN).

Self-organization comprises of two major elements: (1) A **spatiotemporal wave generator** in the first layer driven by noisy interactions between input-sensor nodes and (2) A **local learning rule** implemented by units in the second layer to learn the "underlying" pattern of activity generated in the first layer. These two elements are inspired by mechanisms deployed by the early visual system, which spontaneously triggers retinal waves that tile the light-insensitive retina, that further serve as signals to wire the retina to higher visual areas in the brain [31, 32].

3.1 Spontaneous spatiotemporal wave generator

The first layer can serve as a noise-driven spatiotemporal wave generator when (1) its constituent sensor-nodes are modeled via an appropriate dynamical system and (2) when these nodes are connected in a suitable topology. In this paper, we model each sensor node using the classic Izikhevich neuron model [33] (dynamical system model), while the input layer topology is that of local-excitation and global-inhibition, a motif that is ubiquitous across various biological systems [34, 35]. A minimal dynamical systems model coupled with the local-excitation and global-inhibition motif has been analytically examined in the supplemental materials to demonstrate that these key ingredients are *sufficient* to serve as a spatiotemporal wave generator.

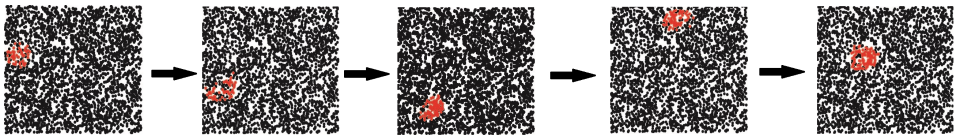


Figure 2: **Emergent spatiotemporal waves tile the first layer.** The red-nodes indicate active-nodes (firing), black nodes refer to silent nodes and the arrows denote the direction of time.

The **Izhikevich model** captures the activity of every sensor node ($v_i(t)$) through time, noisy behavior of individual nodes (through $\eta_i(t)$) and accounts for interactions between nodes within the same layer defined by a synaptic adjacency matrix ($S_{i,j}$). These equations are elaborated in Box-1. The **input layer topology** (local excitation, global inhibition) is defined by the synaptic adjacency matrix ($S_{i,j}$). Every node in the first layer makes excitatory connections with nodes within a defined local excitation radius. $S_{i,j}$ = 5, when distance between nodes i and j are within the defined excitation radius of 2 units; $d_{ij} \leq 2$. Each node has decaying inhibitory connections with other nodes present

above a defined global inhibition radius ($S_{i,j}$ = -2 exp(-d_{ij}/10), when distance between nodes i and j are above a defined inhibition radius of 4 units; $d_{ij} \geq 4$) (see supporting information).

On implementing a model of the resulting dynamical system, we observe the emergence of spontaneous spatiotemporal waves that tile the first layer for specific parameter regimes (see figure 2 and videos in supplemental materials).

Dynamical model for input-sensor nodes in the lower layer (layer-I):

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = 0.04v_i^2 + 5v_i + 140 - u_i + \sum_{j=1}^{N} S_{i,j}\mathcal{H}(v_j - 30) + \eta_i(t) \tag{1}$$

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = a_i(b_i v_i - u_i) \tag{2}$$

with the auxiliary after-spike reset:

$$v_i(t) > 30, \text{then} : \begin{cases} v_i(t + \Delta t) = c_i \\ u_i(t + \Delta t) = u_i(t) + d_i \end{cases}$$

where: (1) v_i is the activity of sensor node i; (2) u_i captures the recovery of sensor node i; (3) $S_{i,j}$ is the connection weight between sensor-nodes i and j; (4) N is the number of sensor-nodes in layer-I; (5) Parameters a_i and b_i are set to 0.02 and 0.2 respectively, while c_i and d_i are sampled from the distributions $\mathcal{U}(-65, -50)$ and $\mathcal{U}(2, 8)$ respectively. Once set for every node, they remain constant during the process of self-organization. The initial values for $v_i(0)$ and $u_i(0)$ are set to -65 and -13 respectively for all nodes. These values are taken from Izhikevich's neuron model [33]; (6) $\eta_i(t)$ models the noisy behavior of every node i in the system, where $< \eta_i(t)\eta_j(t') > = \sigma^2\ \delta_{\mathrm{i,j}}\delta(\mathrm{t} - \mathrm{t}')$. Here, $\delta_{\mathrm{i,j}}$, $\delta(\mathrm{t} - \mathrm{t}')$ are Kronecker-delta and Dirac-delta functions respectively, and σ^2 = 9; (7) $\mathcal{H}$ is the unit step function:

$$\mathcal{H}(v_i - 30) = \begin{cases} 1, & v_i \geq 30 \\ 0, & v_i < 30. \end{cases}$$

3.2 Local learning rule

Having constructed a spontaneous spatiotemporal wave generator in layer-I, we implement a local learning rule in layer-II that can learn the activity wave pattern in the first layer and modify its inter-layer connections to generate a pooling architecture. Many neuron inspired learning rules can learn a sparse code from a set of input examples [36]. Here, we model processing units as rectified linear units (ReLU) and implement a modified Hebbian rule for tuning the inter-layer weights to achieve the same. Individual ReLU units compete with one another in a winner take all fashion.

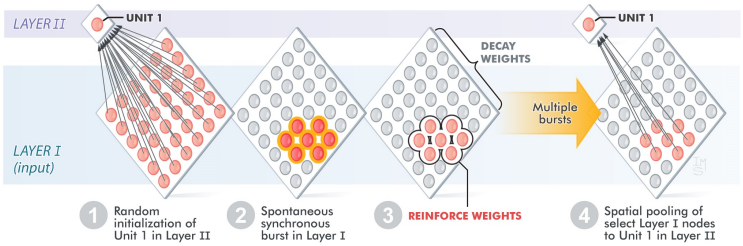


Figure 3: Learning rule

Initially, every processing unit in the second layer is connected to all input-sensor nodes in the first layer. As the emergent activity wave tiles the first layer, at most a single processing unit in the second layer is activated due to the winner-take-all competition. The weights connecting the activated unit in the second layer to the input-sensor nodes in the first layer are updated by the modified Hebbian rule (Box-2). Weights connecting active input-sensor nodes and activated processing units are reinforced

while weights connecting inactive input-sensor nodes and activated processing units decay (cells that fire together, wire together). Inter-layer weights are updated continuously throughout the self-organization process, ultimately resulting in the pooling architecture (See figure-3 and supplemental materials).

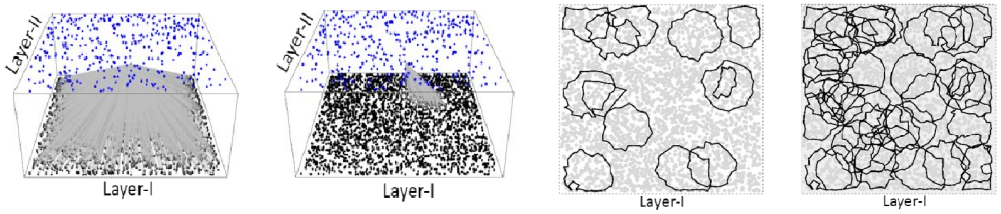
Modifying inter-layer weights

$$w_{i,j}(t+1) = \begin{cases} w_{i,j}(t) + \eta_{learn}\mathcal{H}(v_i(t) - 30)y_j(t+1) & y_j(t+1) > 0 \\ w_{i,j}(t) & \text{otherwise} \end{cases}$$

where: (1) $w_{i,j}(t)$ is the weight of connection between sensor-node i and processing unit j at time 't' (inter-layer connection); (2) η_{learn} is the learning rate; (3) $\mathcal{H}(v_i(t) - 30)$ is the activity of sensor node i at time 't'; and (4) $y_j(t)$ is the activation of processing unit j at time 't'.

Once all the weights $w_{i,j}(t+1)$ have been evaluated for a processing unit j, they are mean-normalized to prevent a weight blow-up. This ensures that the mean strength of weights for processing unit j remains constant during the self-organization process.

Having coupled the spontaneous spatiotemporal wave generator and the local learning rule, we observe that an initially fully connected two-layer network (figure-4a) becomes a pooling architecture, wherein input-sensor nodes that are in close proximity to each other in the first layer have a very high probability of connecting to the same processing unit in the second layer (figure-4b & 4c). More than 95% of the sensor-nodes in layer-I connect to processing units in layer-II (higher layer) through well-defined pools, ensuring that spatial patches of nodes connected to units in layer-II tile the input layer (figure-4d). Tiling the input layer ensures that most sensor nodes have an established means of sending information to higher layers after the self-organization of the pooling layer.



(a) (b) (c) (d)

Figure 4: **Self-organization of Pooling layers.** (a) The initial configuration, wherein all nodes in the lower layer are connected to every unit in the higher layer. (b) After the self-organization process, a pooling architecture emerges, wherein every unit in layer-II is connected to a spatial patch of nodes in layer-I. (a,b) Here, connections from nodes in layer-I to a single unit in layer-II (higher layer) are shown. (c) Each contour represents a spatial patch of nodes in layer-I connected to a single unit in layer-II. (d) More than 95% of the nodes in layer-I are connected to units in the layer-II through well-defined pools, as the spatial patches tile layer-I completely.

4 Features of the developmental algorithm

In this section, we show that spatiotemporal waves can emerge and travel over layers with arbitrary geometries and even in the presence of defective sensor-nodes. Since activity waves can form independent of macroscopic features of the input layer, our algorithm can construct pools over sensor layers with curved or irregular geometries and also in the presence of defects or holes. As the local structure of sensor-node connectivity (local excitation and global inhibition) in the input layer is conserved over a broad range of macroscale geometries (Figure-5a), we observe a traveling activity wave in input layers with arbitrary geometries, which when coupled to a learning rule in layer-II forms a pooling architecture (refer to supplemental information for an analytical treatment). Furthermore, we demonstrate that the size and shape of the emergent spatiotemporal wave can be tuned by altering the topology of sensor-nodes in the layer. Coupling the emergent wave in layer-I with a learning rule in layer-II leads to localized receptive fields that tile the input layer.

Together, the wave and the learning rule endow the developmental algorithm with useful properties: (i) **Flexibility**: Spatial patches of sensor-nodes connected to units in layer-II can be established over arbitrary input-layer geometries. In Figure-5a, we show that an emergent spatiotemporal wave on a ring-shaped input layer coupled with the local learning rule (section-3.2) in layer-II, results in a pooling architecture. Flexibility to form pooling layers on arbitrary input-layer geometries is useful for processing data acquired from unconventional sensors, like charge-coupled devices that mimic the retina [37]. (ii) **Robustness**: Spatial patches of sensor-nodes connected to units in layer-II can be established in the presence of defective sensor nodes in layer-I. As shown in figure-5b, we initially self-organize a pooling architecture for a fully functioning set of sensor-nodes in the input-layer. To test robustness, we ablate a few sensor-nodes in the input-layer (captioned 'DN'). Following this perturbation, we observe that the pooling architecture re-emerges, wherein spatial-pools of sensor-nodes, barring the damaged ones, re-form and connect to units in layer-II. (iii) **Reconfigurable**: The size and shape of spatial pools generated can be modulated by tuning the structure of the emergent traveling wave (figure-5c & 5d). In figure-5e, we show that the size of spatial-pools can be altered in a controlled manner by modifying the topology of layer-I nodes. Wave-x in the legend corresponds to an emergent wave generated in layer-I when every node in layer-I makes excitatory connections to other nodes in its 2 unit radius and inhibitory connections to every node above x unit radius. This topological change alters the properties of the emergent wave, subsequently changing the resultant spatial-pool size. The histograms corresponding to these legends capture the distribution of spatial-pool sizes over all pools generated by a given wave-x. The histogram also highlights that the size of emergent spatial-pools are tightly regulated for every wave-configuration.

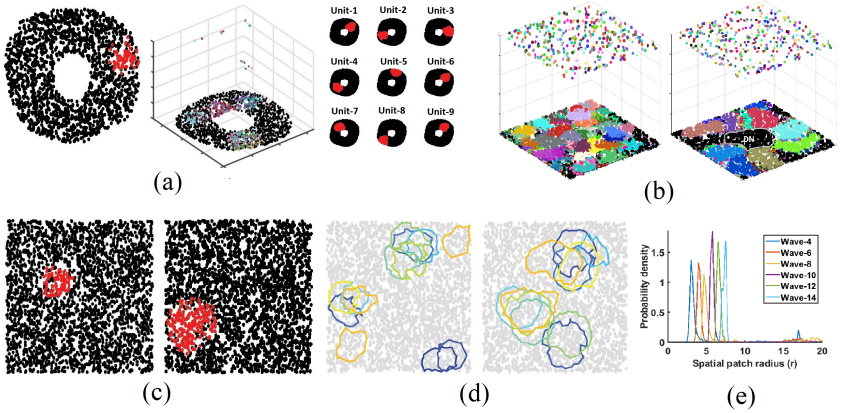
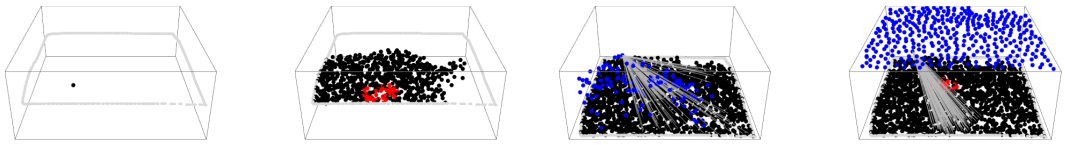


Figure 5: **Features of the developmental algorithm.** (a) **Self-organization of pooling layers for arbitrary input-layer geometry.** (a) The left most image is a snapshot of the traveling wave as it traverses layer-I; Layer-I has sensor-nodes arranged in an annulus geometry; red nodes refer to firing nodes. On coupling the spatiotemporal wave in layer-I to a learning rule in layer-II, a pooling architecture emerges. The central image refers to the 3d visualization of the pooling architecture, while each subplot in the right-most image depicts the spatial patch of nodes in layer-I connected to a single processing unit in layer-II. (b) **Self-organization of pooling layers are robust to input layer defects** (b) The figure on the left depicts a self-organized pooling layer when all input nodes are functioning. Once these inter-layer connections are established, a small subset of nodes are damaged to assess if the pooling architecture can robustly re-form. The set of nodes within the grey boundary, titled 'DN', are defective nodes. The figure on the right corresponds to pooling layers that have adapted to the defects in the input layer, hence not receiving any input from the defective nodes.(c,d,e) **Pooling layers are reconfigurable.** (c) By altering layer-I topology (excitation/inhibition radii), we can tune the size of the emergent spatial wave. The size of the wave is 6 A.U (left) and 10 A.U (right). (d) Altering the size of the emergent spatial wave tunes the emergent pooling architecture. The size of the pools obtained are 4 A.U (left), obtained from a wave-size of 6 A.U and a pool-size of 7 A.U (right), obtained from a wave-size of 10 A.U. (e) A large set of spatial-pools are generated for every size-configuration of the emergent wave. The distribution of spatial-pool sizes over all pools generated by a specific wave-size are captured by a kernel-smoothed histogram. Wave-4 in the legend corresponds to a histogram of pool-sizes generated by an emergent wave of size 4 A.U (blue line). We observe that spatial patches that emerge for every configuration of the wave have a tightly regulated size.

5 Growing a neural network

As the developmental algorithm is flexible to varying scaffold geometries and tolerant to malfunctioning nodes, it can be implemented for growing a system, enabling us to push AI in the direction towards being more 'life-like' by reducing human involvement in the design of complex functioning architectures. The growth paradigm implemented in this section has been inspired by mechanisms that regulate neocortical development [38, 39].

The process of growing a layered neural network involves two major sub-processes. One, every 'node' can divide horizontally to produce daughter nodes that populates the same layer; Two, every node can divide vertically to produce daughter processing units that migrate upwards to populate higher layers. Division is stochastic and is controlled by a set of random variables. Having defined the 3D scaffold, we seed a single unit (figure-6a). As horizontal and vertical division ensues to form the layered neural network, inter-layer connections are modified based on the emergent activity wave in layer-I and a learning rule (section-3.2) in layer-II, to form a pooling architecture. A detailed description of the growth rule-set coupled with a flow chart governing the growth of the network is appended to the supplemental materials.



(a) (b) (c) (d)

Figure 6: **Growing a layered neural network** (a) A single computational "cell" (black node) is seeded in a scaffold defined by the grey boundary. (b) Once this "cell" divides, daughter cells make local-excitatory and global-inhibitory connections. As the division process continues, noisy interactions between nodes results in emergent spatiotemporal waves (red-nodes). (c) Some nodes within layer-I divide to produce daughter cells that migrate upwards to form processing units (blue nodes). The connections between the two layers are captured by the lines that connect a single unit in a higher layer to nodes in the first layer (Only connections from a single unit are shown).(d) After a long duration, the system reaches a steady state, where two layers have been created with an emergent pooling architecture.

Having intertwined the growth of the system and self-organization of inter-layer connections, we make the following interesting observations: (1) spatiotemporal waves emerge in the first layer much before the entire layer is populated (figure-6b), (2) self-organization of inter-layer connections commences before the layered network is fully constructed (figure-6c) and (3) Over time, the system reaches a steady state as the number of 'cells' in the layered network remains constant and most processing units in the second layer connect to a pool of nodes in the first layer, resulting in the pooling architecture (figure-6d). Videos of networks growing on arbitrary scaffolds are added to the supplemental materials.

6 Growing functional neural networks

In the previous section, we demonstrated that we can successfully grow multi-layered pooling networks from a single unit. In this section, we show that these networks are functional.

We demonstrate functionality of networks grown and self-organized from a single unit (figure-7c) by evaluating their train and test accuracy on a classification task. Here, we train networks to classify images of handwritten digits obtained from the MNIST dataset (figure-7e). To interpret the results, we compare it with the train/test accuracy of hand-crafted pooling networks and random networks. Hand-crafted pooling networks have a user-defined pool size for all units in layer-II (figure-7b), while random networks have units in layer-II that connect to a random set of nodes in layer-I without any spatial bias (figure-7d), effectively not forming a pooling layer.

To test functionality of these networks, we couple the two-layered network with a linear classifier that is trained to classify hand-written digits from MNIST on the basis of the representation provided by these three architectures (hand-crafted, self-organized and random networks). We observe that

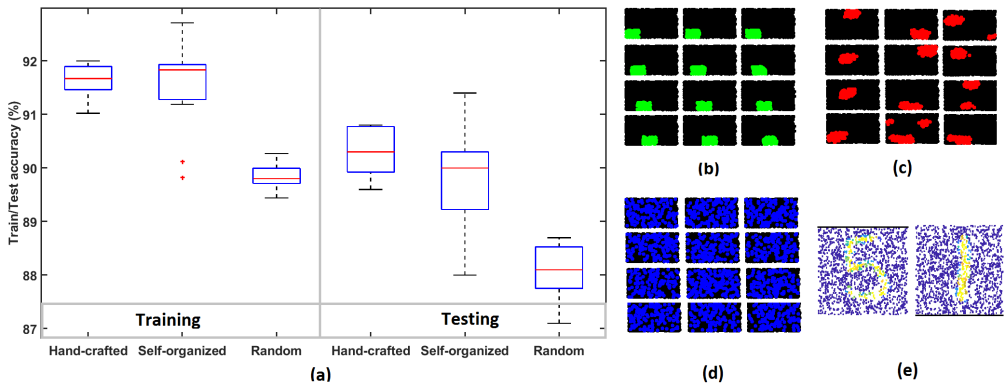


Figure 7: **Networks grown from a single unit are functional.** Three kinds of networks are trained and tested on images obtained from the MNIST database. We use 10000 training samples and 1000 testing samples. The 3 kinds of networks are: (i) Hand-crafted, (ii) Self-organized networks and (iii) random networks. This procedure is run over n=11 networks to ensure that the developmental algorithm always produces functional networks. (a) The box-plot captures the training and testing accuracy of these 3 networks. We notice that the testing accuracy of self-organized networks is comparable to that of to that of hand-crafted networks (p-value = 0.1591>0.05) and are much better than random networks (p-value = 5.6 x 10^{-5}). (b,c,d) Each unit in the second layer is connected to a set of nodes in the lower layer. The set it is connected to are defined by the green, red or blue nodes in the subplots shown. (b) Hand-crafted (c) Self-organized and (d) Random-basis.(e) Two MNIST images as seen in the first layer.

self-organized networks classify with a 90% test accuracy, are statistically similar to hand-crafted pooling networks (90.5%, p-value = 0.1591) and are statistically better than random networks (88%, p-value = 5.6 x 10^{-5}) (figure-7a). This performance is consistent over multiple self-organized networks. These results show that self-organized neural networks are functional and can be adapted to perform conventional machine-learning tasks, with the big add-on of being autonomously grown from a single unit.

7 Discussion

In this paper, we address a pertinent question of how artificial computational machines could be built autonomously with limited human intervention. Currently, architectures of most artificial systems are obtained through heuristics and hours of painstaking parameter tweaking. Inspired by the development of the brain, we have implemented a developmental algorithm that enables the robust growth and self-organization of functional layered neural networks.

Implementation of this framework brought many crucial questions concerning neural development to our attention. Neural development is classically defined by discrete steps, one proceeding the other. However this isn't the case, as development is a continuous flow of events with multiple intertwined processes [40]. Our work on growing artificial systems got us interested in how critical times of different developmental processes are controlled, and whether they were controlled by an internal clock.

The work also reinforces the significance of brain-inspired mechanisms for initializing functional architecture to achieve generalization for multiple tasks. A peculiar instance in the animal kingdom would be the presence of precocial species, animals whose young ones are functional immediately after they are born. One mechanism that enables functionality immediately after birth is spontaneous activity that assists in maturing neural circuits much before the animal receives any sensory input. Although we have shown how a layered architecture (mini-cortex) can emerge through spontaneous activity in this paper, our future work will focus on growing multiple components of the brain, namely a hippocampus and a cerebellum, followed by wiring these regions in a manner useful for an organism's functioning. This paradigm of growing mini-brains in-silico will allow us to (i) explore how different components in a biological brain interact with one another and guide our design of neuroscience experiments and (ii) equip us with systems that can autonomously grow, function and interact with the environment in a more 'life-like' manner.

Acknowledgments

We would like to thank Markus Meister, Carlos Lois, Erik Winfree, Naama Barkai for their invaluable contribution for shaping the early stages of the work. We also thank Alex Farhang, Jerry Wang, Tony Zhang, Matt Rosenberg, David Brown, Ben Hosheit, Varun Wadia, Gautam Goel, Adrianne Zhong and Nasim Rahaman for their constructive feedback and key edits that have helped shape this paper.

References

[1] Lindsey L Glickfeld and Shawn R Olsen. "Higher-order areas of the mouse visual cortex". In: *Annual review of vision science* 3 (2017), pp. 251–273.

[2] Jonathan W Peirce. "Understanding mid-level representations in visual processing". In: *Journal of Vision* 15.7 (2015), pp. 5–5.

[3] Hui Min Tan, Thomas Joseph Wills, and Francesca Cacucci. "The development of spatial and memory circuits in the rat". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 8.3 (2017), e1424.

[4] Christine A Denny, Evan Lebois, and Steve Ramirez. "From Engrams to Pathologies of the Brain". In: *Frontiers in neural circuits* 11 (2017), p. 23.

[5] Timothy D Hanks and Christopher Summerfield. "Perceptual decision making in rodents, monkeys, and humans". In: *Neuron* 93.1 (2017), pp. 15–31.

[6] Camillo Padoa-Schioppa and Katherine E Conen. "Orbitofrontal cortex: A neural circuit for economic decisions". In: *Neuron* 96.4 (2017), pp. 736–754.

[7] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. 2015.

[8] William Song and Jim Cai. "End-to-end deep neural network for automatic speech recognition". In: *Standford CS224D Reports* (2015).

[9] David Silver et al. "Mastering the game of go without human knowledge". In: *Nature* 550.7676 (2017), p. 354.

[10] Andrew M Saxe et al. "On Random Weights and Unsupervised Feature Learning." In: *ICML*. Vol. 2. 3. 2011, p. 6.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[12] Markus Meister et al. "Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina". In: *Science* 252.5008 (1991), pp. 939–943.

[13] Kunihiko Fukushima. "Neocognitron: A hierarchical neural network capable of visual pattern recognition". In: *Neural networks* 1.2 (1988), pp. 119–130.

[14] Kunihiko Fukushima and Nobuaki Wake. "Handwritten alphanumeric character recognition by the neocognitron". In: *IEEE transactions on Neural Networks* 2.3 (1991), pp. 355–365.

[15] David H Hubel and TN Wiesel. "Shape and arrangement of columns in cat's striate cortex". In: *The Journal of physiology* 165.3 (1963), pp. 559–568.

[16] Yann LeCun et al. "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems*. 1990, pp. 396–404.

[17] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. "Neural architecture search: A survey". In: *arXiv preprint arXiv:1808.05377* (2018).

[18] Kenneth O Stanley and Risto Miikkulainen. "Evolving neural networks through augmenting topologies". In: *Evolutionary computation* 10.2 (2002), pp. 99–127.

[19] Esteban Real et al. "Large-scale evolution of image classifiers". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 2902–2911.

[20] Esteban Real et al. "Regularized evolution for image classifier architecture search". In: *arXiv preprint arXiv:1802.01548* (2018).

[21] Barret Zoph and Quoc V Le. "Neural architecture search with reinforcement learning". In: *arXiv preprint arXiv:1611.01578* (2016).

[22] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. "Efficient Multi-objective Neural Architecture Search via Lamarckian Evolution". In: *arXiv preprint arXiv:1804.09081* (2018).

[23] Yanqi Zhou and Gregory Diamos. “Neural architect: A multi-objective neural architecture search with performance prediction”. In: *Proc. Conf. SysML*. 2018.

[24] Keith B Godfrey and Nicholas V Swindale. “Retinal wave behavior through activity-dependent refractory periods”. In: *PLoS computational biology* 3.11 (2007), e245.

[25] AF Haussler. “Development of retinotopic projections: an analytical treatment”. In: *Journal of Theoretical Neurobiology* 2 (1983), pp. 47–73.

[26] David J Willshaw and Christoph Von Der Malsburg. “How patterned neural connections can be set up by self-organization”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 194.1117 (1976), pp. 431–445.

[27] Stephen J Eglen and Julijana Gjorgjieva. “Self-organization in the developing nervous system: Theoretical models”. In: *HFSP journal* 3.3 (2009), pp. 176–185.

[28] NV Swindale. “The development of topography in the visual cortex: a review of models”. In: *Network: Computation in neural systems* 7.2 (1996), pp. 161–247.

[29] NV Swindale. “A model for the formation of ocular dominance stripes”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 208.1171 (1980), pp. 243–264.

[30] Frederic Zubler and Rodney Douglas. “A framework for modeling the growth and development of neurons and networks”. In: *Frontiers in computational neuroscience* 3 (2009), p. 25.

[31] Rachel OL Wong. “Retinal waves and visual system development”. In: *Annual review of neuroscience* 22.1 (1999), pp. 29–47.

[32] Noelia Antón-Bolaños et al. “Prenatal activity from thalamic neurons governs the emergence of functional cortical maps in mice”. In: *Science* (2019), eaav7617.

[33] Eugene M Izhikevich. “Simple model of spiking neurons”. In: *IEEE Transactions on neural networks* 14.6 (2003), pp. 1569–1572.

[34] Brett Kutscher, Peter Devreotes, and Pablo A Iglesias. “Local excitation, global inhibition mechanism for gradient sensing: an interactive applet”. In: *Sci. STKE* 2004.219 (2004), pl3–pl3.

[35] Yuan Xiong et al. “Cells navigate with a local-excitation, global-inhibition-biased excitable network”. In: *Proceedings of the National Academy of Sciences* 107.40 (2010), pp. 17079–17086.

[36] Bruno A Olshausen and David J Field. “Emergence of simple-cell receptive field properties by learning a sparse code for natural images”. In: *Nature* 381.6583 (1996), p. 607.

[37] Giulio Sandini et al. “Retina-like CCD sensor for active vision”. In: *Robots and Biological Systems: Towards a New Bionics?* Springer, 1993, pp. 553–570.

[38] P Rakic. “Radial unit hypothesis of neocortical expansion”. In: *Novartis Foundation Symposium*. Wiley Online Library. 2000, pp. 30–52.

[39] Irina Bystron, Colin Blakemore, and Pasko Rakic. “Development of the human cerebral cortex: Boulder Committee revisited”. In: *Nature Reviews Neuroscience* 9.2 (2008), p. 110.

[40] Paola Arlotta and Pierre Vanderhaeghen. “Editorial overview: Developmental neuroscience 2017.” In: *Current opinion in neurobiology* 42 (2017), A1.

Supplementary Material

S1 Mathematical model

S1.1 Dynamical model for input sensor nodes

Input sensor nodes are modeled using the Izhikevich neuron model. This is used primarily because it has the least number of parameters for accurately modeling neuron-like activity and the parameter regimes that produce different neuronal firing states have been well characterized earlier [1].

Dynamical model for input-sensor nodes in the lower layer (layer-I):

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = 0.04v_i^2 + 5v_i + 140 - u_i + \sum_{j=1}^{N} S_{i,j}\mathcal{H}(v_j - 30) + \eta_i(t) \tag{S1}$$

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = a_i(b_i v_i - u_i) \tag{S2}$$

with the auxiliary after-spike reset:

$$v_i(t) > 30, \text{then} : \begin{cases} v_i(t + \Delta t) = c_i \\ u_i(t + \Delta t) = u_i(t) + d_i \end{cases}$$

where: (1) v_i is the activity of sensor node i; (2) u_i captures the recovery of sensor node i; (3) $S_{i,j}$ is the connection weight between sensor-nodes i and j; (4) N is the number of sensor-nodes in layer-I; (5) Parameters a_i and b_i are set to 0.02 and 0.2 respectively, while c_i and d_i are sampled from the distributions $\mathcal{U}(-65, -50)$ and $\mathcal{U}(2, 8)$ respectively. Once set for every node, they remain constant during the process of self-organization. The initial values for $v_i(0)$ and $u_i(0)$ are set to -65 and -13 respectively for all nodes. These values are taken from Izhikevich's neuron model [1]; (6) $\eta_i(t)$ models the noisy behavior of every node i in the system, where $< \eta_i(t)\eta_j(t') > = \sigma^2\ \delta_{\mathrm{i,j}}\delta(\mathrm{t} - \mathrm{t}')$. Here, $\delta_{\mathrm{i,j}}$, $\delta(\mathrm{t} - \mathrm{t}')$ are Kronecker-delta and Dirac-delta functions respectively, and σ^2 = 9; (7) $\mathcal{H}$ is the unit step function:

$$\mathcal{H}(v_i - 30) = \begin{cases} 1, & v_i \geq 30 \\ 0, & v_i < 30. \end{cases}$$

S1.2 Topology of input-sensor nodes

The nodes in the lower layer (layer-I) are arranged in a local-excitation, global inhibition topology, with a ring of nodes that have neither excitation or inhibition (zero weights) between the excitation and inhibition regions. We have observed that this ring of no connections between the excitation and inhibition regions gives us a good handle over the emergent wave size. This is detailed in Box-S1.2 and depicted in figure-S1a.

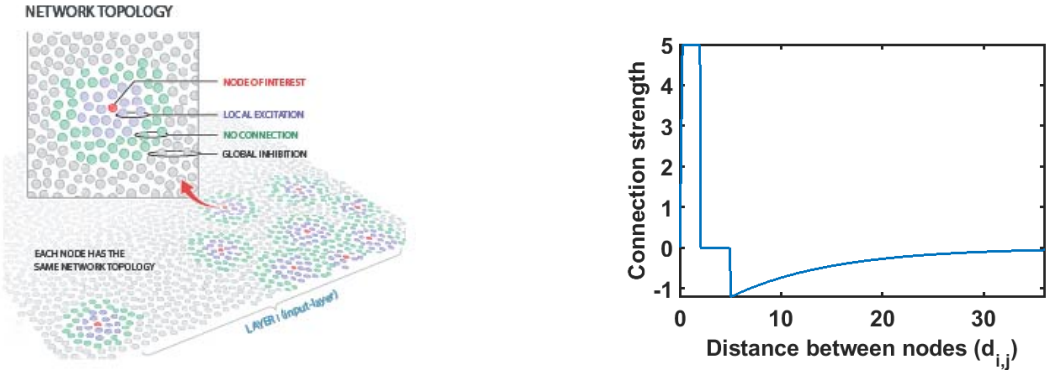
Topology of input-sensor nodes in layer-I:

This topology is pictorially depicted in figure-S1a and mathematically defined below:

$$S_{i,j} = \begin{cases} l, & d_{i,i} \leq r_e \\ m \exp(\frac{-d_{i,j}}{10}), & d_{i,j} \geq r_i \\ 0 & r_e < d_{i,j} < r_i \end{cases}$$

where:

- $S_{i,j}$ is the connection weight between sensor-nodes i and j
- $d_{i,j}$ is the Euclidean distance between sensor-nodes i and j in layer-I
- r_e is the local excitation radius (r_e = 2)
- r_i is the global inhibition radius (all nodes present outside this radius are inhibited) (r_i = 4)
- l is the magnitude of excitation (l = 5)
- m is the magnitude of inhibition (m = -2)



(a) (b)

Figure S1: Topology of sensor-node connections: Every node is connected to other nodes in the layer within a radius r_e via a positive weight, not connected to nodes positioned at a distance between r_e and r_i and connected to nodes at a distance larger than r_i with a decaying negative weight.

S1.3 Modeling Processing units and winner-take-all strategy

Processing units are modeled as Rectified linear units (ReLU) associated with an arbitrary threshold. Although the threshold is randomly initialized, it is updated during the process of self-organization. Its update depends entirely on the activity trace of the processing unit it is associated with. We also require that at every time point, at most a single processing unit in layer-II be activated by the emergent patterned activity in layer-I. To enforce this, we let the processing units, modeled as ReLU units compete with each other in a winner-take-all (WTA) manner. This ensures that at every time point, at most a single unit in layer-II responds to the patterned activity in the input layer.

Each processing unit in layer-II is modeled by the equation given below:

$$y_j(t) = \mathcal{W}[\max(0, \sum_{i=1}^{N} w_{i,j}(t)\mathcal{H}(v_i(t) - 30))] \tag{S3}$$

Here, the $\max(0, x)$ is the implementation of a rectified linear unit (ReLU); $\mathcal{H}(v_i(t) - 30)$ is the threshold activity of sensor node i (in layer-I) at time 't'; $y_j(t)$ is the activation of processing unit j (in layer-II) at time 't'; $w_{i,j}^t$ is the connection weight between sensor-node i and processing unit j at time 't'; N is the number of sensor-nodes in layer-I and $\mathcal{W}$ refers to the winner-take-all mechanism that ensures a single winning processing unit.

The winner-take-all function implemented in layer-II is mathematically elaborated below:

$$\mathcal{W}[y_j(t)] = \begin{cases} \max(0, y_j(t) - c_j(t)), & \text{if } y_j(t) > y_k(t) \ \ \forall k \in [1, ...j-1, j+1, ..., M] \\ 0 & \text{otherwise} \end{cases}$$

Here, $y_j(t)$ is the activation of processing unit j (in layer-II) at time 't'; $c_j(t)$ is the threshold for processing unit j at time 't' and M is the number of processing units in layer-II. Every processing unit is modeled as a ReLU with an associated threshold (c_j). Although this threshold is arbitrarily initialized, they are updated during the process of self-organization. The update depends on the number of times the connections between processing units and nodes in layer-I are updated, and it's described below.

To implement this, we keep track of the number of times connections between a specific processing unit and sensor nodes in layer-I are updated over the course of 1000 time-points. $z_j(t)$ captures the number of times connections between processing unit-j and sensor-nodes in layer-I are updated.

Keeping track of the synaptic changes per processing unit:

$$z_j(t+1) = \begin{cases} z_j(t) + 1 & \text{if } (y_j(t) > 0) \\ 0 & \text{if } (t \bmod 1000) = 0 \\ z_j(t) & \text{otherwise} \end{cases}$$

The threshold for a processing unit is updated based on the number of connections that were altered in the past 1000 time points between that processing unit and sensor-nodes in layer-I.

Updating the threshold for every processing unit:

$$c_j(t+1) = \begin{cases} \max(y_j(t), y_j(t-1), ..., y_j(0))/5, & \text{if } (t \bmod 1000) = 0 \text{ AND } z_j(t) < 200 \\ c_j(t) & \text{otherwise} \end{cases}$$

Here, $w_{i,j}(t)$ is the weight of connection between sensor-node i and processing unit j at time 't'; η_{learn} is the learning rate; y_j^t is the activation of processing unit j at time 't'; $z_j(t)$ is the number of synaptic modifications made to unit j until time 't'; (t mod 1000) is the remainder when t is divided by 1000 and $c_j(t)$ is the activation threshold for processing unit j at time 't'.

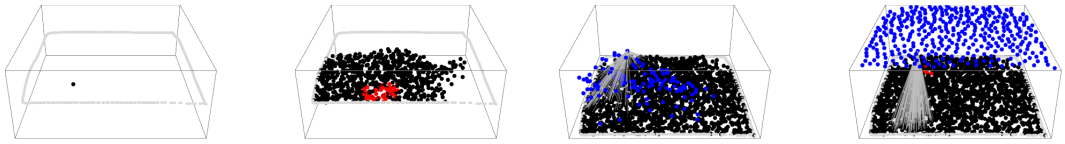
The emergent wave in layer-I coupled with the learning rule implemented by processing units in layer-II are sufficient to self-organize pooling architectures.

S2 Growing a neural network

We demonstrate that by defining a minimal set of 'rules' for a single computational 'cell', we can grow a layered network, followed by the self-organization of its inter-layer connections to form pooling layers.

In order to grow a layered network, we define a 3D scaffold as well as seed the first layer in the scaffold with a computational 'cell' (figure-6a). The major attributes of nodes in the first layer are:

- $v_i(t)$: activity of node i modeled by the Izhikevich equation [1]
- $clockH_i$: records the age of the 'cell', allowing horizontal division (division within the same layer) until it reaches a certain age
- $HFlim_i$: the maximum divisions permitted for node i
- VCD_i : a binary variable that records whether node i has vertically divided or not. Vertical division is the process when a 'cell' divides and its daughter 'cells' migrate upwards to form processing units that populate higher layers.



(a) (b) (c) (d)

Figure S2: **Growing a layered neural network** (a) A single computational "cell" (black node) is seeded in a scaffold defined by the grey boundary. (b) Once this "cell" divides, daughter cells make local-excitatory and global-inhibitory connections. As the division process continues, noisy interactions between nodes results in emergent spatiotemporal waves (red-nodes). (c) Some nodes within layer-I divide to produce daughter cells that migrate upwards to form processing units (blue nodes). The connections between the two layers are captured by the lines that connect a single unit in a higher layer to nodes in the first layer (Only connections from a single unit are shown).(d) After a long duration, the system reaches a steady state, where two layers have been created with an emergent pooling architecture.

S2.1 User-defined Growth Parameters

Parameter	Value	Description
HCD_AGE	25	The maximum time a cell can pursue horizontal division
HF_MAX	40	The maximum number of divisions a single cell can pursue
R_HDIV	1	Critical radius I
R_VDIV	1	Critical radius II
THRESH_HDIV	3	The maximum number of cells permitted within a radius (R_HDIV)

S2.2 Growth Process

Step: 1:

A single computational 'cell' endowed with the following attributes is seeded on a 3D scaffold. The attributes and values that a seeded computational 'cell' is endowed with is mentioned in the table below. The first column indicates attributes, second column denotes the initial values that they take and the third column is a description of the attribute.

Cell attribute	Initialization	Description
v_i	-65	Initialize activity of node i
clockH_i	0	Initializing clock to 0, for every newly divided daughter cell
HFlim_i	HF_MAX	Initializing the max divisions to HF_MAX for the seeded cell.
VCD_i	0	Before vertical division, $\text{VCD}_i = 0$; After vertical division, $\text{VCD}_i = 1$;

S2.2.1 Step: t $\rightarrow$ t+1

A random cell i is sampled from the input layer.

If the cell hasn't crossed the critical age threshold ($\text{clockH}_i <$ HCD_AGE) and the number of cells within a radius (R_HDIV) is below the density threshold (numCells_i(R_HDIV) $<$ THRESH_HDIV), the cell divides horizontally to form daughter cells that populate the same layer. The clockH is reset to zero for the daughter cells, however the HFlim attribute of the daughter cells is one less than their parent to keep track of the number of divisions.

If it hasn't reached the critical age threshold, but has a local density above the defined density threshold, it remains quiescent and a new 'cell' is sampled.

A cell i can divide vertically only if the cell has reached the critical age threshold (clockH_i = HCD_AGE) and cells in its local vicinity (with radius :- R_VDIV) haven't divided vertically. As

mentioned in an earlier section, a binary variable VCD_i keeps track of whether a cell has divided vertically or not.

When a cell divides vertically, one daughter cell occupies the parent's position on layer-I, while the other daughter cell migrates upwards. The daughter cell that migrates upwards initially makes a single connection with its twin on layer-I, which gets modified with time, resulting in a pool of nodes in layer-I making connections with a single unit in the higher layer (pooling architecture).

S2.2.2 Termination condition

The local rules that control horizontal division and vertical division are active throughout and prevent the system from blowing up, with respect to the number of nodes in each layer. It has been observed that the system reaches a steady state, as the number of 'cells' in both layers remain constant.

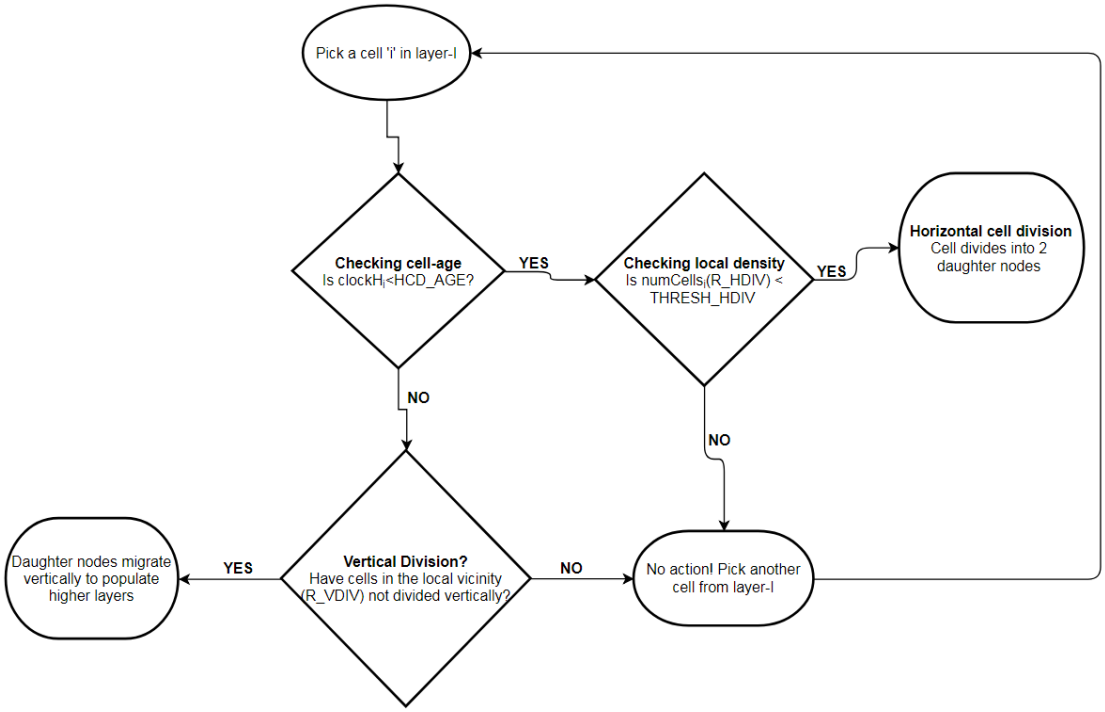


Figure S3: Growth flowchart

S2.3 Growing neural networks on arbitrary scaffolds (Results)

Videos of multi-layered networks growing on arbitrary scaffolds can be viewed by visiting this link: [`https://drive.google.com/open?id=1YtFEvWHTU9HWl760V8lEr9Heapx0sUdh`]

S3 Minimal model for observing emergent spatiotemporal waves

In this section, we provide an analytical solution for the emergence of a spatiotemporal wave through noisy interactions between constituent nodes in the same layer.

As we stated in the main-text, the key ingredients for having a layer of nodes function as a spatiotemporal wave generator are:

- Each sensor-node should be modeled as a dynamical systems model
- Sensor-nodes should be connected in a suitable topology (here, local excitation ($r_e < 2$ and global inhibition ($r_i > 4$).

On modeling all nodes in the system using a simple set of ODE's, we highlight the conditions required for observing a stationary bump in a network of spiking sensor-nodes and to observe instability of the stationary bump resulting in a traveling wave.

S3.1 Arranging sensor-nodes in a line

We choose a configuration where N sensor-nodes are randomly arranged in a line (as shown in figure-S4).

Figure S4: Sensor nodes arranged in a line

The activity of N sensor nodes, arranged in a line as in figure-S4, are modeled using a minimal ODE model as described below:

$$\tau_d \frac{\mathrm{d}x(u_i,t)}{\mathrm{d}t} = -x(u_i,t) + \sum_{u_j \in \mathcal{U}} S(u_i,u_j)\mathcal{F}(x(u_j,t)) \quad \forall i \in 1,...,N \tag{S4}$$

Here, u_i represents the position of nodes on a line; $x(u_i,t)$ defines the activity of sensor node positioned at u_i at time t; S_{u_i,u_j} is the strength of connection between nodes positioned at u_i and u_j; τ_d controls the rate of decay of activity; $\mathcal{U}$ is the set of all sensor nodes in the system (u_1,u_2,...,u_N) for N sensor nodes; and $\mathcal{F}$ is the non-linear function required to convert activity of nodes to spiking activity. Here, $\mathcal{F}$ is the heaviside function with a step transition at 0.

Each sensor-node has the same topology of connections, ie fixed strength of positive connections between nodes within a radius r_e, no connections from a radius r_e to r_i, and decaying inhibition above a radius r_i. This is depicted in figure-S5

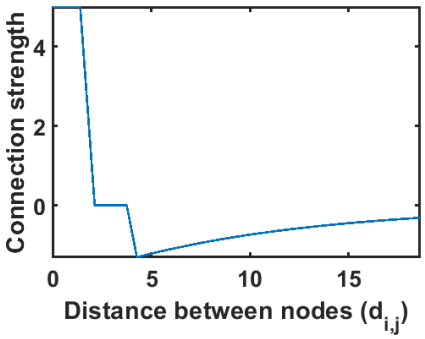


Figure S5: strength of connections between sensor-nodes

S3.1.1 Fixed point analysis

We determine the stable activity states of nodes placed in a line by a fixed point analysis, similar to what Amari developed in [2] for the case when there are infinite nodes.

$$x(u_i) = \sum_{u_j \in \mathcal{U}} S(u_i,u_j)\mathcal{F}(x(u_j)) \quad \forall i \in 1,...,N \tag{S5}$$

On solving this system of non-linear equations simultaneously, we get a fixed point ie a vector $x^* \in \mathcal{R}^N$, corresponding to the activity of N sensor nodes positioned at (u_1,u_2,...,u_N). To assess their spiking from the activity of sensor-nodes, we have

$$s_i = \mathcal{F}(x(u_i)) \quad \forall i \in 1,...,N \tag{S6}$$

As the weight matrix (S_{u_i,u_j}) used incorporates the local excitation ($r_e < 2$) and global inhibition ($r_i > 4$) (figure-S5), we get solutions with a single bump of activity (figure-S6a), two bumps of activity (figure-S6c) or a state when all nodes are active.

(a) (b) (c)

Figure S6: **Fixed points:** Multiple fixed points are obtained by solving N non-linear equations simultaneously. Some of the solutions obtained are: (a) a single bump at the center, (b) a single bump at one of the edges and (c) two bumps of activity.

S3.1.2 Stability of fixed points

To assess the stability of these fixed points, we evaluate the eigenvalues of the Jacobian for this system of differential equations. As there are N differential equations, the Jacobian ($\mathbb{J}$) is an NxN matrix.

$$\begin{aligned}
\frac{\mathrm{d}x(u_i,t)}{\mathrm{d}t} &= \frac{-x(u_i,t)}{\tau_d} + \sum_{u_j \in \mathcal{U}} \frac{S(u_i,u_j)\mathcal{F}(x(u_j))}{\tau_d} \\
\frac{\mathrm{d}x(u_i,t)}{\mathrm{d}t} &= f_i(u_1,u_2,...,u_N) \\
f_i(u_1,u_2,...,u_N) &= \frac{-x(u_i)}{\tau_d} + \sum_{u_j \in \mathcal{U}} \frac{S(u_i,u_j)\mathcal{F}(x(u_j))}{\tau_d} \\
\mathbb{J}(i,j) &= \frac{\partial f_i(u_1,u_2,...,u_N)}{\partial x(u_j)}
\end{aligned} \tag{S7}$$

On evaluating the Jacobian ($\mathbb{J}$) at the fixed points obtained (x*), we get:

$$\begin{aligned}
\mathbb{J}(i,i) &= \frac{\partial f_i}{\partial x(u_i)} \\
\mathbb{J}(i,i) &= \frac{-1}{\tau_d} \\
\mathbb{J}(i,j) &= S(u_i,u_j)\mathcal{F}^{'}(x(u_j))\frac{\partial x(u_j)}{x(u_j)} \\
\mathbb{J}(i,j) &= S(u_i,u_j)\delta(x(u_j)) \\
\mathbb{J}(i,j) &= 0 \;\; \forall x(u_j) \neq 0
\end{aligned} \tag{S8}$$

Here, $\mathcal{F}$ is the Heaviside function and its derivative is the dirac-delta(δ); where, $\delta(x)$ = 0, for $x \neq 0$ and $\delta(x) = \infty$ for $x = 0$.

For a fixed point, where $x^*(u_k) \neq 0, \forall k \in 1,...,N$, the Jacobian is a diagonal matrix with $\frac{-1}{\tau_d}$ in its diagonals. This implies that the eigenvalues of the Jacobian are $\frac{-1}{\tau_d}$ (τ_d >0), which assures that the fixed point $x* \in \mathcal{R}^N$ is a stable fixed point.

S3.1.3 Destabilizing the fixed point

With the addition of high amplitude of gaussian noise to the ODE's described earlier, we can effectively destabilize the fixed point, resulting in a traveling wave. The equations with the addition of a noise term are:

$$\tau_d \frac{\mathrm{d}x(u_i,t)}{\mathrm{d}t} = -x(u_i,t) + \sum_{u_j \in \mathcal{U}} S(u_i,u_j)\mathcal{F}(x(u_j,t)) + \eta_i(t) \quad \forall i \in 1,...,N \tag{S9}$$

Here, $\eta_i(t)$ models the noisy behavior of every node i in the system, where $< \eta_i(t)\eta_j(t') >$ = σ^2 $\delta_{\mathrm{i,j}}\delta(\mathrm{t}-\mathrm{t}')$. Here, $\delta_{\mathrm{i,j}}$, $\delta(\mathrm{t}-\mathrm{t}')$ are Kronecker-delta and Dirac-delta functions respectively, and σ^2 captures the magnitude of noise added to the system.

The network of sensor nodes is robust to a small amplitude of noise ($\sigma^2 \in$ (0,4)), while a larger amplitude of noise (σ^2>5) can destabilize the bump, forcing the system to transition to another bump

in its local vicinity. Continuous addition of high amplitudes of noise forces the bump to move around in the form of traveling waves. The behavior is consistent with the linear stability analysis because noise can push the dynamical system beyond the envelop of stability for a given fixed point solution.

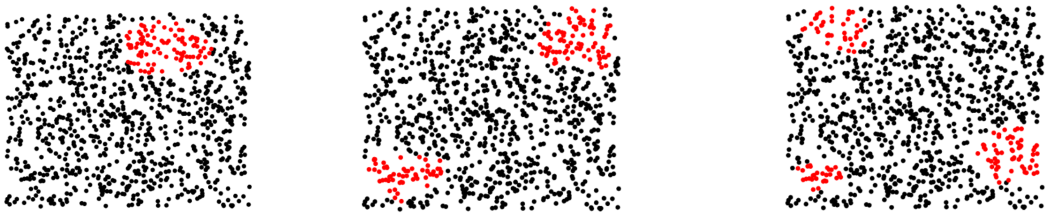
S3.2 Arranging sensor nodes in a 2D square

In this section, we arrange N sensor nodes arbitrarily on a 2-dimensional square as shown in figure-S7, with the same local structure (local excitation and global inhibition).

The activity of these sensor nodes are modeled using the minimal ODE model described earlier (in equation-4).

Figure S7: Sensor nodes placed arbitrarily on a square plane

We obtain the fixed points ($x^* \in \mathcal{R}^N$), by solving N simultaneous non-linear equations using BBsolve [3]. We notice that the fixed point solutions have a variable number of activity bumps in the 2D plane as shown in figure-8a,8b & 8c.



(a) (b) (c)

Figure S8: **Stable Fixed points:** Multiple fixed points are obtained by solving N non-linear equations simultaneously. Some of the solutions obtained are: (a) a single bump, (b) two bumps and (c) three bumps of activity.

S3.3 Arranging sensor nodes on a 2D sheet of arbitrary geometry

In this section, we arrange sensor nodes on a 2D sheet in any arbitrary geometry as shown in figure 9. Although the macroscopic geometry of the sheet changes, the local structure of sensor nodes in conserved (ie local excitation and global inhibition).

The fixed points are evaluated by simultaneously solving the non-linear system of equations. We notice that the bumps are stable fixed points even when sensor nodes are placed on a 2-dim sheet of arbitrary geometry.

(a) (b) (c) (d)

Figure S9: **Stable Fixed points:** Multiple fixed points are obtained by solving N non-linear equations simultaneously. Some of the solutions obtained are: (a,b) a single bump for a circular geometry (c,d) two bumps of activity for arbitrary geometry

S4 Growing functional neural networks

We estimate functionality of networks grown and self-organized from a single unit by evaluating their train and test accuracy on a classification task. Here, we train networks to classify images of handwritten digits obtained from the MNIST dataset. To interpret the results, we compare it with the train/test accuracy of hand-crafted pooling networks and random networks. Hand-crafted pooling networks have a user-defined pool size for all units in layer-II, while random networks have units in layer-II that connect to a random set of nodes in layer-I without any spatial bias, effectively not forming a pooling layer.

To test functionality of these networks, we couple the two-layered network with a linear classifier that is trained to classify hand-written digits from MNIST on the basis of the representation provided by these three architectures (hand-crafted, self-organized and random networks).

The first two layers in the network serve as feature extractors, while the last layer behaves like a perceptron. The optimal classifier is learnt by minimizing the least square error between the output of the network and a desired target. However, there isn't any back-propagation through the entire network. In essence, the architecture grown through the developmental algorithm remains fixed, performing the task of latent feature representation, while the classifier learns how to match these latent features with a set of task-based labels.

S4.1 Setting up the pooling architecture

The first two layers of the network correspond to the pooling architecture grown by the developmental algorithm. The input is fed to the first layer, while the units in the second layer, that are connected to spatial pools in layer-I, extract features from these inputs.

Let $x \in \mathcal{R}^N$ be the input data (for N sensor nodes) and the weights connecting the first and second layer be $W_1 \in \mathcal{R}^{M \times N}$ (for M processing units). The features extracted in layer-II are: $y = \mathcal{F}(W_1 x)$. Here, $\mathcal{F}$ is any non-linear function applied to the transformation in order to map all the values in layer-II within the range [-1,1].

S4.2 Appending a fully connected layer

The pooling architecture sends its feature map through a fully connected layer with L nodes, with the weights connecting the set of processing units and the fully connected layer being randomly initialized as $W_2 \in \mathcal{R}^{L \times M}$. The features extracted by the fully connected layer are: $y_{FC} = \mathcal{F}(Wy)$. $\mathcal{F}$ is the same as the one used in section-4.1.

S4.3 Classification accuracy

The final set of weights connecting the fully connected layer to the 10 element vector (as there are 10 digit classes in the MNIST dataset) is denoted by $W_3 \in \mathcal{R}^{10 \times L}$. The output generated by the network is $y_O = W_3 y_{FC}$. Let us denote the target output as y_T.

As we want to minimize the least square error between the target output (y_T) and output of the network (y_O), conventionally, we can perform a gradient descent. However, as it is a linear classifier, we have a closed form solution for the weight matrix (W_3).

$$
\begin{aligned}
y_O &= W_3 y_{FC} \\
y_T &= W_3 y_{FC} \qquad \text{for zero error, } y_0 = y_T \\
y_T y_{FC}^T &= W_3 y_{FC} y_{FC}^T \\
W_3 &= y_T y_{FC}^T (y_{FC} y_{FC}^T)
\end{aligned}
$$

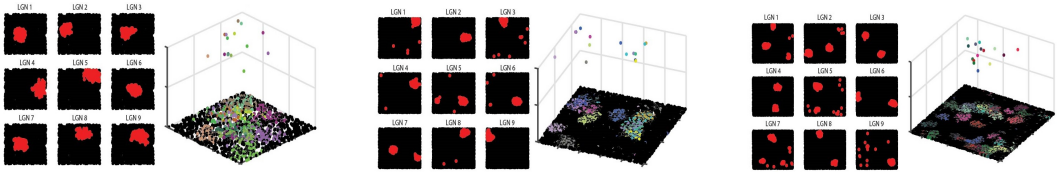
Setting the weights between the fully connected layer and the output layer ($W_3 = y_T y_{FC}^T (y_{FC} y_{FC}^T)$, we evaluate the train and test accuracy for 3 kinds of networks. (Hand-crafted pooling, self-organized and random networks). These networks differ primarily in how their first two layers are connected. The hand-programmed pooling networks are those that have a fixed size of spatial pool that connects to units in layer-II, while the random networks have no spatial pooling.

The results are described in the main-paper and we observe that self-organized networks classify with a 90% test accuracy are statistically similar to hand-crafted pooling networks (90.5%, p-value = 0.1591) and statistically better than random networks (88%, p-value = 5.6 x 10^{-5}) (figure-7a). This performance is consistent over multiple self-organized networks. The train/test accuracy of self-organization networks highlights that growing networks through a brain-inspired developmental algorithm is potentially useful to building functional networks.

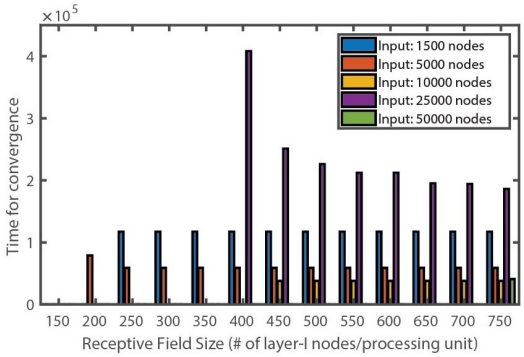
S5 Scalability: Determining the speed of self-organization of the pooling architecture as the size of the input-layer increases

Here, we demonstrate that the pooling layers can be self-organized for very large input layers. Large layers are defined based on the number of sensor nodes in the layer. We observe that enforcing a spatial bias on the initial set of connections from units in layer-II to the nodes in the input layer, enables us to speed up the process of self-organization.

Our simulations show that the self-organization of pooling layers can be scaled up to large layers (with upto 50000 nodes) without being very expensive, as an increase in number of sensor-nodes results in multiple simultaneous waves tiling the input layer, effectively forming a pooling architecture in parallel.



(a) Input layer: 1500 nodes (b) Input layer: 5000 nodes (c) Input layer: 10000 nodes



(d) Time complexity for self-organization of pooling layers

Figure S10: **Developmental algorithm scales efficiently to very large input layers:** (a) Layer-I has 1500 nodes and layer-II has 400 nodes. The emergent wave in layer-I results in a single traveling wave that tiles layer-I. (b) Layer-I has 5000 nodes and layer-II has 400 nodes. The emergent wave in layer-I results in a single traveling wave that tiles layer-I. (c) Layer-I has 10000 nodes and layer-II has 400 nodes. The emergent wave in layer-I results in a multiple traveling wave that tile layer-I simultaneously. This results in a single processing unit receiving pools from different regions. (d) The histogram captures the time taken for a pooling layer to form for variable number of input sensor nodes (1500, 5000, 10000, 25000 and 50000 nodes). With an increase in the number of sensor-nodes, the speed of self-organization increases as multiple waves tile the input layer simultaneously.

References

[1] Eugene M Izhikevich. "Simple model of spiking neurons". In: *IEEE Transactions on neural networks* 14.6 (2003), pp. 1569–1572.

[2] Shun-ichi Amari. “Dynamics of pattern formation in lateral-inhibition type neural fields”. In: *Biological cybernetics* 27.2 (1977), pp. 77–87.

[3] Ravi Varadhan, Paul Gilbert, et al. “BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function”. In: *Journal of statistical software* 32.4 (2009), pp. 1–26.