Basic concept of “islet of inter-LINKed spines” in the semblance hypothesis matches with that of the “attention heads” in Transformers

Kunjumon I. Vadakkan, SIMATS University, Chennai, India, dated 21st January 2024 (not peer-reviewed)

The exchange of ideas between neuroscience and artificial intelligence has been contributing to the progress of both fields. Since the unique feature of the nervous system is the generation of first-person properties, the neuroscience field heavily depends on artificial systems for replication of its major findings and hypotheses. The mutual interaction between fields began with the use of neurons that are connected by synapses (**Fig.1**) in neural networks (NNs). Later, a prototype of a general-purpose system was created by training a NN having a few neurons to learn sequential patterns to predict simple sequences of symbols (Jordan, 1986). Here, a set of neurons termed “state units” was added to the network with the aim of allowing the system to take the next action given what is currently observed. By adding connections from outputs to the state units and then to the middle of the network and finally to themselves led to the creation of a recurrent NN (RNN). When a network learns to perform a sequence, it learns to follow a trajectory through a state space that are called “attractors”. This helped neuroscience to view recurrent connections behind the formation of attractor networks to generate pattern completion and was a leading hypothesis for storage and retrieval of memories (Nakazawa et al., 2002; Rennó-Costa et al., 2014). Single neurons have been found to correlate with several functions in the nervous system and were named based on the function that they are associated with (O'Keefe and Dostrovsky, 1971; Liu et al., 2012). In parallel to these findings, when large RNN was used, it led to the identification of neurons deeper in the network that learn complex concepts called “sentiment neurons” (Radford et al., 2017) that directly correspond to the sentiment of the text. Large language models (LLMs) are the new versions of NNs that have revolutionized the field of AI with the property of generalization. Hence, this work examines how the concepts of the inner workings of Transformers in the LLMs match with the findings in neuroscience.

In neuroscience, arguments that motivated us to arrive at a testable mechanism that registers a signature change during associative learning between two stimuli are given in **Table 1**.

1	Associative learning between two sensory stimuli is expected to take place at the locations of their convergence. Hence, an operational mechanism is expected to take place at these locations.
2	Postsynaptic potential attenuates as the depolarization propagates along the dendritic branch towards the neuronal soma. Hence, any mechanism for information storage is expected to take place at its origin, which is dendritic spine head (spine or postsynaptic terminal or input terminal).
3	Arrival of any 140 input signals from nearly 4000 to 11000 input terminals (dendritic spines) of a neuron can lead to an output from that neuron in the form of an action potential (neuronal firing). If postsynaptic potential arriving from nearly any 140 of its dendritic spines can fire that neuron, then nearly $[1\text{x}10^4! \div (140! \text{ x } (1\text{x}10^4! - 140!))] \approx 2.79\text{x}10^{318}$ sets of combinations of input signals can fire that neuron. This shows that specificity of inputs is lost beyond the point of convergence of two associated stimuli where both information needs to be stored and outputs in response to a prompt to be made. Hence, a mechanism at the origin of postsynaptic potential is expected.
4	Conditioned learning experiments have informed us that during learning certain changes must take place that will allow later arrival of conditioned stimuli (CS) to generate features of both CS (conditioned stimulus) and US (**Fig.1**).
5	The mean inter-spine distance is more than the mean spine diameter among pyramidal neurons, which are the major neuronal types in the cortex (Konur et al., 2003).

Table 1. *Arguments that were made towards finding a solution for the nervous system.*

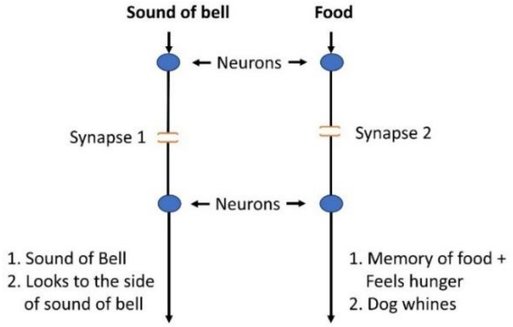


Figure 1. *In the conditioned learning paradigm, two different types of stimuli that can provide both first-person inner sensations and motor outputs can be used (total 4 features). Here, the commonly used association between the sound of a bell and food is shown. After learning, the arrival of the sound alone is expected to generate all the four features. What type of neuronal connection can provide this?*

To satisfy all the conditions given in **Table 1**, and to provide a structure-function mechanism to satisfy the requirements during conditioned learning (**Fig.1**), it is necessary that the dendritic spines (input terminals or postsynaptic terminals) that belong to different neurons along which the newly associated stimuli propagate to interact with each other at locations where they remain physically abutted (at the level of the spine heads) (**Fig.2**). Hence, as a general rule, inter-neuronal inter-spine interaction leading to inter-postsynaptic functional LINK (IPL) is expected to occur during associative learning. Arrival of one of the learned items reactivates the IPL and propagation of potential to the inter-LINKed second spine, which can generate both units of first-person inner sensations (due to a special feature at the synapses) and motor output reminiscent of the arrival of the second stimulus. This is the basis of the semblance hypothesis (Vadakkan, 2007; 2013; www.semblancehypothesis.org). Exceptions can also occur. For e.g. spines on two different dendritic branches of a neuron can interact and can generate first-person inner sensations, but without being able to generate motor actions. Following this, constraints from a large number of disparate findings from different levels of functioning of the system were examined to test whether inter-neuronal inter-spine interactions can explain all of them in an interconnected manner (Vadakkan, 2013; 2019). It was also possible to identify that the mechanism has features that qualify it to be an evolved mechanism (Vadakkan, 2019).

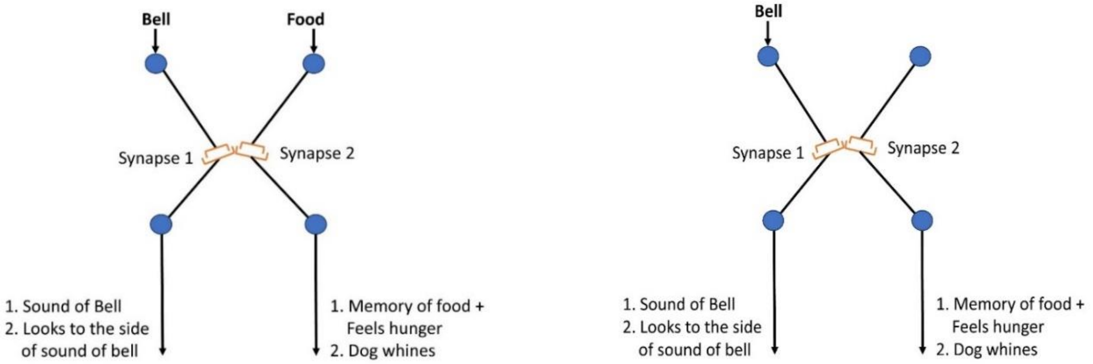

Bell
Food
Synapse 1
Synapse 2
1. Sound of Bell
2. Looks to the side of sound of bell
1. Memory of food + Feels hunger
2. Dog whines
Bell
Synapse 1
Synapse 2
1. Sound of Bell
2. Looks to the side of sound of bell
1. Memory of food + Feels hunger
2. Dog whines

Figure 2. Findings from a conditioned learning paradigm and the minimum necessary neurobiological conditions. Left*: Associative learning between sound of a bell (Conditioned stimulus, CS) and sight of food (Unconditioned stimulus, US) is expected to generate specific changes at the locations of their convergence. Based on the semblance hypothesis, an interaction between the postsynaptic terminals (dendritic spines) of neurons of two pathways at the locations of their convergence is expected to take place.* ***Right****: Later, when the sound of the bell alone (CS) arrives, the above change is expected to generate both first-person inner sensations of the sound of the bell and memory of food (and feeling of hunger) along with motor actions such as turning towards the bell and whining.*

Continued learning events can lead to formation of additional inter-LINKs between the already inter-LINKed spines with new ones that lead to formation of multiple inter-LINKed spines that are called islets of inter-LINKed spines (Vadakkan, 2011). A motor neuron being held at a sub-threshold activation state below the threshold will fire an action potential only when the remaining potential arrives at it (**Fig.3**). It can be anticipated that the remaining potential is generated by the arrival of a set of stimuli, that in turn leads to the firing of the motor neuron. Since the inter-LINKed spines form a hub, then all the neurons of the spines within that islet of inter-LINKed spines will receive a certain amount of potential that may allow them to cross the threshold for firing action potential. The propagation of potential within an islet of inter-LINKed spines and towards certain neurons of these spines depends on several factors, such as 1) how close spines are that receive inputs, and 2) the presence of inhibitory inputs in between them. This is expected to generate and control the production of motor outputs such as speech or behavioral motor actions. For example, production of a specific word in a sentence depends on several factors, such as 1) What is the question, 2) What is the expected answer? 3) What is the context? 4) What is the word used prior to it? , and 5) What is the word following it? Hence, motor neurons are expected to fire only when they receive all these inputs. If the motor neuron responsible for generating the word is being held at subthreshold potential and if it receives all the remaining potential from stimuli that prompt the production of that specific word, then only that motor neuron fires to produce that word.

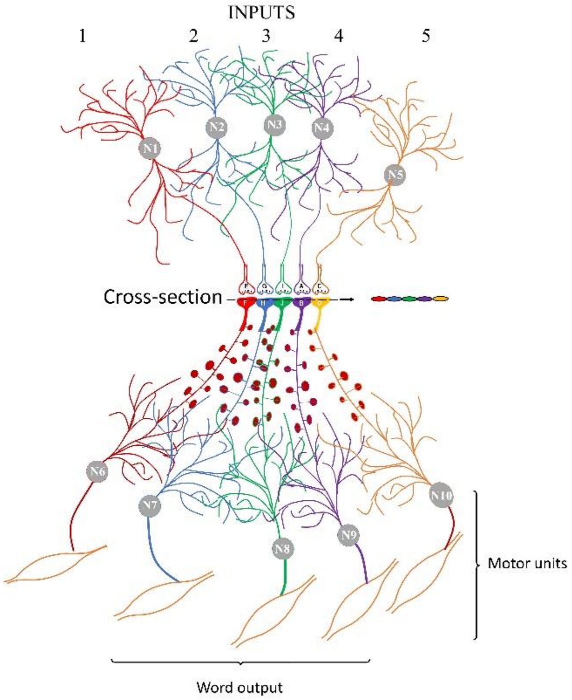


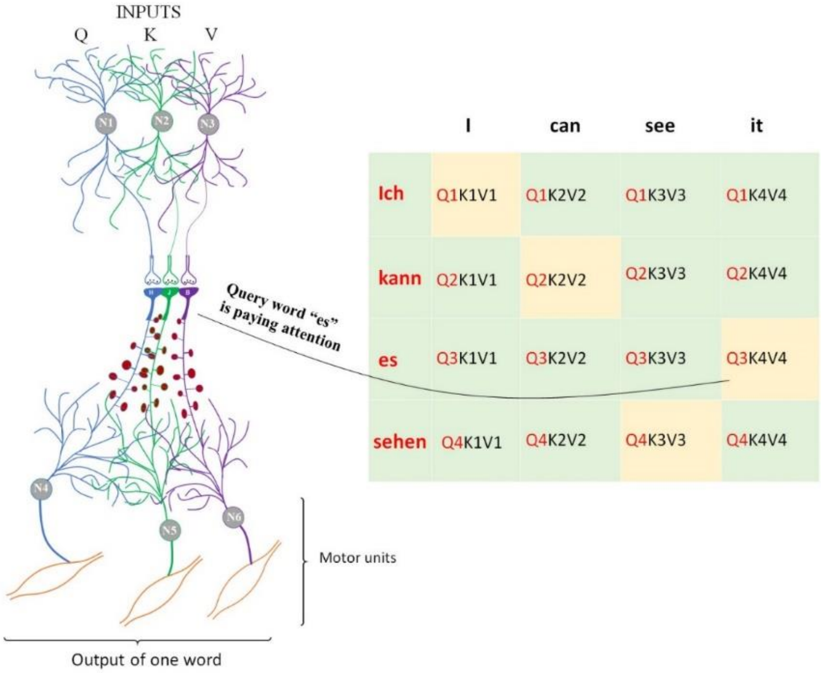
Figure 3. *Pictorial representation of the arrival of five inputs to motor neurons that generate a specific word. It is expected to go through an islet of inter-LINKed spines whose neurons fire to generate the word. Arrival of three input signals will reach the islet of inter-LINKed spines F-H-J-B-D. The first synapse E-F is an inhibitory synapse, meaning that any input coming through it will reduce the net potential moving out of this islet of inter-LINKed spines. When inputs arrive at spines H, J, B and D, then they get summated to provide maximal output that will pass on to their neurons. If their neurons are being held at a sub-threshold activation level short of certain potential and if activation of three inputs arriving through neurons N2, N3 and N4 can provide this input, then the neurons N7, N8 and N9 can fire to generate a specific word. Only a fraction of spines within an islet of inter-LINKed spines can be activated by a prompt and the outputs from their spines' neurons depend on the weights of their inter-LINKing, and inhibitory and promoting factors.*

Meaningful speech generation in Transformers

LLM uses Transformer for generating speech in response to a prompt using sentences by selecting and arranging words that provide meaningful outputs for a given context (for demonstration, see videos 1, 2). The developers of Transformer (Vaswani et al., 2017) took a sequence of measures to logically execute steps to generate sequences of sentences that have meaning. This includes methods to find a word depending on what needs to be said, the context, the word used before and the word after in a grammatically correct manner. The creators of the Transformer focused on the relationship between different words (self-attention) in building sentences with meaning (Decoder-Encoder attention). For this, signature relations between a specific set of inputs that determine which word to be chosen to make a sentence are to be marked during training and can be used during output generation. This was formalized in linear algebraic terms (Videos 1 and 2) to develop the Transformer.

An example of Transformer working can be done by showing how the English sentence "I can do it" is converted to German. Decoder-Encoder attention is estimated to select words in a specific order based on their attention scores. It starts with “word embedding” in numbers and then “positional encoding”. In this process, a 4 by 4 matrix is made. Using different vector components of both inputs and outputs, vector multiplication (first dot multiplication between Q and K followed by cross multiplication of the product with V) is carried out to find the matrix entry values, which are the attention scores. from which words with maximum attention scores are selected (**Fig.4**). Here, Q is the vector we compute attention for. This points to a German word to which a word in English needs to be translated. K is the vector we compute attention against, which means the English word that we need to translate to German. V is the learned vector corresponding to the best matching results in response to our query. The word with the highest attention score is used at each step to generate a sentence.

A common essential property of both the nervous system and the Transformer is a hub (islets of inter-LINKed spines and attention heads respectively) at the locations of convergence of input signals where the relationship between the input signals is registered and can be used in the future when a new prompt (cue stimulus) arrives. Even though an operational mechanism to generate first-person inner sensations is present in the inter-LINKed spines (Vadakkan, 2013), this function is not necessary to provide output of words for making sentences in a meaningful manner.



	I	can	see	it
Ich	Q1K1V1	Q1K2V2	Q1K3V3	Q1K4V4
kann	Q2K1V1	Q2K2V2	Q2K3V3	Q2K4V4
es	Q3K1V1	Q3K2V2	Q3K3V3	Q3K4V4
sehen	Q4K1V1	Q4K2V2	Q4K3V3	Q4K4V4

Figure 4. Comparison between interconnected spines in an islet of inter-LINKed spines and the relations between specific input signals that determine the correct word in the Decoder-Encoder attention.
Left: *Inter-LINKed spines within an islet of inter-LINKed spines. Spines of neurons N1, N2, and N3 interact to form an islet of inter-LINKed spines H, J and B.* ***Right:*** *Transduction of relations between inputs (via inter-LINKed spines) as entries of an attention matrix. Here, a matrix of vectors Q, K and V when an English sentence “I can see it” is translated to German is shown. Here, the word “see” has four different circumstances (vector components) that we compute attention for, four different circumstances (vector components) that we compute attention against and four different circumstances (vector components) that correspond to the best matching results in response to our query. Each entry in the matrix of vector products denotes a relation between at least three corresponding spines in different islets of inter-LINKed spines. Here, an islet of inter-LINKed spines (on the left) that corresponds to an entry in the 4x4 matrix (on the right) that provides an attention score to select the best matching word is shown. Note that entries in pink boxes have maximum values and are used to select words for the German translation.*

Discussion

Similarities between the concepts of islets of inter-LINKed spines and attention heads is that they act as hubs where the input signals can both register their relationships during training and use them when a new prompt (cue stimulus) arrives. Word embedding in Transformers is equivalent to conversion of a sensory stimulus to depolarization. The property of attention in Transformers is achieved by the formation of islets of inter-LINKed spines. The brain operates only when the frequency of oscillating extracellular potential remains within a narrow range. Since a corresponding intercellular oscillation between the connected neuronal processes is expected to occur and since the IPL mechanism can provide vector components of the oscillations, it is reasonable to expect that these oscillations contribute to an integration mechanism equivalent to positional encoding, and residual connections in the Transformers.

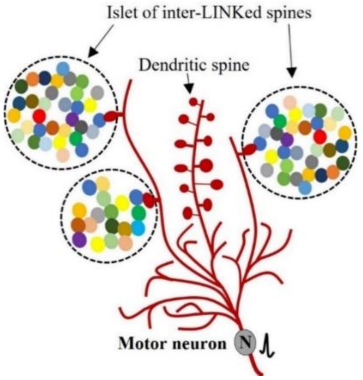


Figure 5. Each dendritic spine of a neuron can inter-LINK with abutted spines that belong to other neurons during a sequence of learning events. *A motor neuron N is shown, whose three spines are shown to be part of three different islets of inter-LINKed spines. Note that the small circles within each islet of inter-LINKed spines (shown in different colors) are spines that belong to other neurons that are not shown in the figure. One dendrite of the neuron N at the middle of the figure is shown to have dendritic spines on either side of it. Spines that are part of islets of inter-LINKed spines increase the possible number of circumstances under which the motor neuron can fire. i.e. when an output neuron is kept a certain subthreshold activation state and if a sufficient number of inputs arrive at an islet and reaches the axon hillock of that neuron to cross the threshold, that neuron can fire to generate a motor action such as speech. Potential reaching an islet of inter-LINKed spines can propagate through all the inter-LINKed spines depending on several factors. Even though the number of spines on a pyramidal neuron is limited to between 4000 and 11,000, inter-LINKing with spines of other neurons increases the input repertoire by the total number of inter-LINKed spines that belong to other neurons. More importantly, factors such as spatial relationships between spines within an islet, inhibitory inputs and dopaminergic inputs that can cause spine expansion can generate microdomains within each islet of inter-LINKed spines and control outputs for generating motor actions.*

Even though it is possible to identify abutted spines that belong to different neurons in an electron microscopic image, there were no attempts to visualize islets of inter-LINKed spines. Lack of a logical explanation for their presence warranting their visualization, their locations in different planes of a three-dimensional space and lack of methods to visualized them based on special electrical properties of inter-LINKed spines are factors that prevent us from identifying them. Alternatively, it may become possible to visualize them utilizing an anticipated property of oxidation state-dependent interactions (Vadakkan, 2023) between the inter-LINKed spines. The concept of islets of inter-LINKed spines forces us to view each spine of a neuron along with the remaining inter-LINKed spines that belong to other neurons that form islets of inter-LINKed spines of different sizes (**Fig.5**). It is possible that synchronized inputs to an islet of inter-LINKed spine in response to a prompt is responsible for the finding of a "dendritic spike". It is known that dendritic spikes can function as efficient detectors of specific input patterns, ensuring a neuronal output (action potential) (Gasparini et al., 2004). This matches with the expectation of a mechanism that can generate a specific motor output in response to a specific set of inputs arriving at an islet of inter-LINKed spines. Since some dendritic spikes occur in the absence of somatic action potentials (Golding and Spruston, 1998), it is possible to speculate that the large potential generated during a dendrite spike may get propagated to the neurons of other spines within the islet of inter-LINKed spines through a less resistant route.

The ability of any new prompt to reach the corresponding islets of inter-LINKed spines or attention heads and generate a potential or attention score to create an output is the basis of generalization. Similar to the fact that the presence of one variable in two equations of a system of equations allows finding a solution for the system, if an input signal arrives to more than one islet of inter-LINKed spines it will allow the system to find interconnections between stimuli. This will permit obtaining outputs in response to new prompts that the system was never exposed in the past. This property can be extended to explain the hypothesis generation expected by a brain mechanism (Abbott, 2008). Present work must undergo further verification.

Conflict of Interest: U.S. patent: number 9477924 pertains to a model of the inter-postsynaptic functional LINK.

References

1. Abbott LF (2008) Theoretical neuroscience rising. *Neuron* 60, 489–495.
2. de Almeida L, Idiart M, Lisman JE (2007) Memory retrieval time and memory capacity of the CA3 network: role of gamma frequency oscillations. *Learn. Mem.* 14(11):795–806.
3. Gasparini S, Migliore M, Magee JC (2004) On the initiation and propagation of dendritic spikes in CA1 pyramidal neurons. *J. Neurosci.* 24(49):11046–11056.
4. Golding NL, Spruston N (1998) Dendritic sodium spikes are variable triggers of axonal action potentials in hippocampal CA1 pyramidal neurons. *Neuron* 21(5):1189–200.
5. Jordan ML (1986) Serial order: A parallel distributed processing approach. ICS Report 8604. Institute for Cognitive Science, University of California, San Diego, La Jolla, California. 92093.
6. Konur S, Rabinowitz D, Fenstermaker VL, Yuste R (2003) Systematic regulation of spine sizes and densities in pyramidal neurons. *J. Neurobiol.* 56(2):95–112.
7. Liu X, Ramirez S, Pang PT, Puryear CB, Govindarajan A, Deisseroth K, Tonegawa S (2012) Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* 484(7394):381–385.
8. Nakazawa K, Quirk MC, Chitwood RA, Watanabe M, Yeckel MF, Sun LD, Kato A, Carr CA, Johnston D, Wilson MA, Tonegawa S (2002) Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* 297(5579):211–218.
9. O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34(1):171–175.
10. Radford A, Jozefowwics R, Sutskever I (2017) Learning to generate reviews and discovering sentiment. *arXiv*:1704.01444v2.
11. Rennó-Costa C, Lisman JE, Verschure PF (2014) A signature of attractor dynamics in the CA3 region of the hippocampus. *PLoS Comput. Biol.* 10(5):e1003641.
12. Vadakkan KI (2007) Semblance of activity at the shared post-synapses and extracellular matrices - A structure function hypothesis of memory. ISBN:978-0-5954-7002-0
13. Vadakkan KI (2011) Processing semblances induced through inter-postsynaptic functional LINKs, presumed biological parallels of K-lines proposed for building artificial intelligence. *Front. Neuroeng.* 4:8.
14. Vadakkan KI (2013) A supplementary circuit rule-set for the neuronal wiring. Front. Hum. Neurosci. 1;7:170.
15. Vadakkan KI (2019) From cells to sensations: A window to the physics of mind. *Phys. Life Rev.* 31:44–78.
16. Vadakkan KI (2023) Golgi staining of neurons: Oxidation-state dependent spread of chemical reaction matches with a testable property of the connectome. https://doi.org/10.31219/osf.io/zka2m
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. 31st Conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA.

Videos (not peer-reviewed)

1. https://www.youtube.com/watch?v=UPtG_38Oq8o&t=1902s 2. https://www.youtube.com/watch?v=zxQyTK8quyY