One-time learning in a biologically-inspired Salience-affected Artificial Neural Network (SANN)

Leendert A Remmelzwaal

Department of Electrical Engineering, University of Cape Town, Cape Town 7700, South Africa

E-mail: leenremm@gmail.com

George F R Ellis

Department of Mathematics and Applied Mathematics, University of Cape Town 7700, Cape Town, South Africa

E-mail: george.ellis@uct.ac.za

Jonathan Tapson

MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

E-mail: J.Tapson@westernsydney.edu.au

* Corresponding author: leenremm@gmail.com

Key Words: artificial neural networks, salience effects, one time learning

Abstract

Standard artificial neural networks (ANNs), loosely based on the structure of columns in the cortex, model key cognitive aspects of brain function such as learning and classification, but do not model the affective (emotional) aspects of brain function. However these are a key feature of the brain (the associated 'ascending systems' have been hard-wired into the brain by evolutionary processes). These emotions are associated with memories when neuromodulators such as dopamine and noradrenaline affect entire patterns of synaptically activated neurons. Here we present a bio-inspired ANN architecture which we call a Salience-Affected Neural Network (SANN), which, at the same time as local network processing of task-specific information, includes non-local salience (significance) effects responding to an input salience signal. During pattern recognition, inputs similar to the salience-affected inputs in the training data will produce reverse salience signals corresponding to those experienced when the memories were laid down. In addition, training with salience affects the weights of connections between nodes, and improves the overall accuracy of a classification of images similar to the salience-tagged input after just a single iteration of training. Note that we are not aiming to present an accurate model of the biological salience system; rather we present an artificial neural network inspired by those biological systems in the human brain, that has unique strengths.

Glossary

AI = artificial intelligence

AND = affective neural darwinism [28]
ANN = artificial neural network [28]
BBD = brain-based device [22]
MNF = non-negative matrix factorization [49]
MSE = mean squared error
NN = neural network
RNN = recurrent neural network
SANN = salience-affected neural network
SNN = spiking neural network

1. Introduction

Some aspects of human cognition, such as the ability to learn and classify input signals, can be modelled with artificial neural networks (ANN). However, standard ANNs do not incorporate the non-specific, diffuse dispersion of neuromodulators in the cortex. The presence of neuromodulators in the cortex, originating from the arousal system, allow memories to be tagged with emotion and salience. In this paper we present a *Salience Affected Neural Network* (SANN), an artificial neural network (ANN) with an additional trainable dimension of *salience* or significance compared to standard models. This is a novel ANN architecture model inspired by the effect that neuromodulators have on memories - a crucial feature of the way the human brain functions [28]. It enables powerful single-exposure learning in ANNs, affecting memories proportional to the strength of a global salience signal, and retrieval of that salience signal when the same image is encountered later (a *Reverse Salience Signal*). It is important to note that we are not presenting an accurate model of the biological salience system, rather we present an artificial neural network inspired by the salience systems in the human brain; this allows it to have unique attributes.

Our paper provides a proof of principle that this proposal can be made to work, and could be employed in far more complex reinforcement learning dynamics in Deep Neural Networks. A key feature of the paper is the distinction made between Generic and Specific Semantic Memory (§2.2), and our use of the SANN to determine both.

1.1. Affective systems, neurotransmitters, and neuromodulators

Human cognition does not just consist of cortical processing, but also includes sub-cortical processing such as the primary and secondary emotions of the affective system [55] [17] [72] [35] [15] [16]. Primary emotions, for example, are critical to guiding the cortex in the context of ongoing interaction with the physical, ecological, and social environment [61] [27]. Emotions are associated with memories when neuromodulators such as dopamine and noradrenaline affect patterns of synaptically activated neurons. Neuromodulators are dispersed in the cortex by diffuse projections originating from the arousal system (see Figure 1). The release of neuromodulators in the cortex are the source of primary emotions [61], and they have the effect of altering neural network weights as described in Edelman's concept of "Neural Darwinism" [20] [21]. Thus emotions affect the cortex via Affective Neural Darwinism (AND) [28].

Like most discussion of brain function [13] [36], most artificial neural network structures focus on information processing tasks such as pattern recognition, classification, information storage, and sequence prediction. Inter alia this involves them storing information ("memories") and retrieving them ("remembering"). However, standard artificial neural networks omit representing a crucial aspect of cognitive function: namely "affect" [61] [16] [29], i.e. the effect of primary and secondary emotions. In this paper we focus on the impact of neuromodulators in the cortex, and it is this diffuse effect of post-synaptic gain that we model in a SANN.

Each memory has associated with it an emotional tag that is recalled when the memory is recalled, which triggers feelings associated with the relevant context or event. According to Edelman [20] [24] [21], emotional tags are a dimension of *salience* that is present in addition to standard neural network training (e.g. pattern recognition) that takes place in the cortex.

In this paper we clearly distinguish between '*neurotransmitters*' and '*neuromodulators*'. Neurotransmitters are chemicals responsible for a targeted, rapid signaling between one synapse and another post-synaptic neuron. Neurotransmitters are only released at the synaptic cleft, and are only received by specific receptors at the post-synaptic terminal. Neurotransmitters are said to have an impact on the 'channel' [69] [55] [42] [56].

By comparison, neuromodulators are messenger molecules that are diffusely dispersed across the cortex via the 'ascending systems' (Fig.1). Neuromodulators can impact many neurons simultaneously, which may result in post-synaptic gain. Some neuromodulators have a gradual impact over time, while others (e.g. those responsible for registering fear) can be responsible for single-exposure learning. The impact of neuromodulators depend on the chemical, the quantity and the location in the cortex that it is dispersed. In contrast to neurotransmitters, neuromodulators are described as having an impact on the 'state' of a neuron - post-synaptic gain - rather than an impact on the 'channel' [56] [33] [55] [6].

Neuromodulators not only impact the synapses in the cortex, but they also impact the rest of the body (e.g. adrenaline). As a result, the diffuse projections from the arousal system can be considered as both ascending and descending (see Figure 1). In addition to neuromodulators, other chemicals produced in the human body can impact cognition (such as hormones from the pituitary, and peptides from the gut) [42] [9]. However our focus in this paper is on the 'ascending systems' and their function. It is worth remembering here that the hippocampus is a key target of the arousal system. [63]

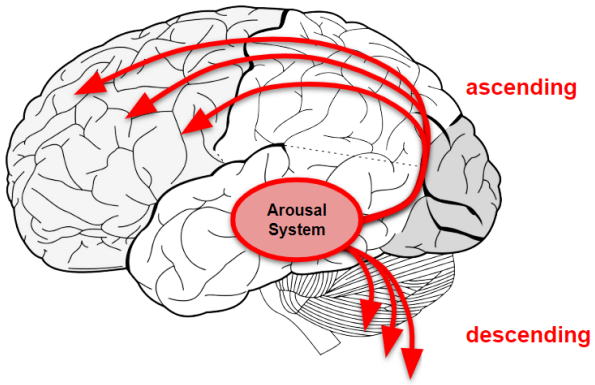

Fig 1. *The release of neuromodulators by the arousal system in the cortex are the source of primary emotions. The arousal system sends neuromodulators (e.g. dopamine and noradrenaline) into the cortex by means of diffuse projections. These neuromodulators simultaneously affect patterns of activated neurons at the time the neuromodulators are delivered, acting on the memory of the object currently being observed.*

1.2. Salience Affected Neural Network

In this paper we present a Salience Affected Artificial Neural Network that represents the non-local signaling in the cortex via neuromodulators associated with rewards and

reinforcement signals broadcast from the arousal system, and demonstrates that this makes single-exposure learning possible. The SANN model we present is not trying to be an accurate model of the biological salience system, rather it is an architecture for an artificial neural network inspired by the salience systems in the human brain. This work is based on the relation between these important structural and functional aspects of the human brain, and thereby opens up a new class of ANN that may be useful in computational applications.

In particular a SANN demonstrates that these kinds of networks not only enable single-exposure learning by laying down memory patterns associated with strong salience signals, but enables recall of those salience signals when the relevant stimulus is encountered at later times. The SANN models both the way memories associated with emotionally-laden events [17] may be embodied in neural networks, and the way those emotional associations can be recalled when the events are remembered and so alter immediate responses appropriately.

The architecture we propose in this paper is a key breakthrough in ANN structuring, because our model adds salience as an additional dimension to the network, and allows for salience training by single exposure. By doing so, we demonstrate how a neural network can be modified to include a salience signal without needing to re-train it on the original data set. While salience can be applied post-training by single exposure, the SANN also allows for salience to be applied during the initial training. Single exposure training has practical application for large neural networks which have already been trained on extensive data sets, and would not wish to repeat the training for specific new additions.

We present a one-dimensional affective system (just positive/negative affect). In reality the innate primary emotional systems are multi-dimensional [61] [79] because this enables innate responses to a variety of circumstances. Extensions of the model proposed here should be able to handle this higher dimensional emotional structure.

1.3. Significance: Robotics

The emotional systems in humans have been developed by evolutionary processes over millions of years because they produce mental processes that significantly enhance survival rates [27]. It might therefore be expected that using simulated emotional systems in robots would also improve their abilities to learn and survive in a variety of different contexts, where the 'affective' experiences they have guide them as to how to behave, even though no qualia are associated with them. As in the human case, they would be provided with a set of primary pre-primed systems giving guidance to the main ANNs as to what kind of behaviour is likely to enhance survival in the context of particular experiences. This would then shape both immediate behaviour and alter neural connection strengths so that the pattern of active connections as a whole at that time would be enhanced, perhaps even being learned at one go. Thus what is demonstrated here as a matter of principle could be incorporated as aspects of far more complex neural networks or robotic applications. Thus while this paper uses a simple SANN to demonstrate the effect, it is a proof in principle of the possibility of existence of applications where inclusion of the architecture we lay out here in deep learning reinforcement neural networks [30] could enhance their capabilities considerably. This is a subject for future research.

There is now a substantial literature on 'emotional robots' that simulate emotions in order to have a more effective robot-human interaction [3] [7]. However they do not simulate the affect-neocortical relations of the kind we discuss here, where emotions have an important effect on cognition and memory [17] [61] [27]. Such 'affective' systems as proposed here could potentially further enhance robot-human interactions by inclusion in such robots, allowing them to relate to humans more meaningfully on an

emotional level (for example, retrieved memories would have an associated emotional tone that would affect behaviour).

1.4. The Big Picture

The big picture here is how genuine complexity emerges from the underlying physics, the two key cases (apart from life in general) being the brain on the one hand and digital computers on the other.

The way the functioning of the mind/brain occurs at the micro level is discussed in [26], where the physical nature of bio-molecules such as voltage gated ion channels in axons and dendrites is identified as the key enabler allowing branching logic to emerge from the underlying unitary physics. However this only leads to our ability to think when billions of neurons are linked by synapses to form immensely complex hierarchically structured neural nets such as occur in neocortical columns [42]. Given the neural and synaptic structure, that architecture is the key to brain function.

The way digital computers are able to perform the logical operations on the basis of the underlying physics is discussed in [25], where it is the detailed physical structure of transistors in integrated circuits that is the key enabler allowing digital logic to emerge from the underlying unitary physics. These again have to be built into a dense network of interconnected transistors forming a hierarchically structured whole [77] in order to perform the magic of turning the logic of an algorithm into its representation in terms of digitally coded electron flows in transistors. Given the functioning of transistors, it is that detailed architecture that is the key to computer function.

Now one can model biological neural networks through suitable computer programs that code for Artificial Neural Networks (ANNs) (see Figure 6), and this gives remarkable properties such as pattern recognition and classification. In essence they are a simplified representation of the structure of cortical columns in the brain. These models can be spiking or non-spiking; the majority in broad use are non-spiking, although increasingly Spiking Neural Networks (SNNs) are being used in artificial intelligence applications [64].

The key point is that neither ANNs or SNNs in use at present represent the architecture embodied in the "ascending systems" in the brain, with diffuse projections that we model here. The burden of this paper is that doing so has the potential to greatly increase the power of both ANNs and SNNs. We demonstrate this in the case of an ANN by implementing the architecture whose essence is shown in Figure 7. That is, we show that this highly simplified version of the arousal system in the brain - which is there for very good functional reasons, leading to such enhanced survival prospects that its existence has been hardwired into our genes [27] - can in the case of the simple ANN we implement be demonstrated to give greatly improved learning performance. It can be expected that the same will be true if this architecture is implemented in the much more complex deep learning networks in use in the AI community.

1.5. Structure of this paper

Using a cognitive architecture approach [52], we consider the effect of emotions on memory in the neocortex (Section 2) and associated cortical structures of diffuse projections in the human brain. We conduct an investigation of existing comparable frameworks (Section 3) and then present the expected characteristics of the SANN model and its design (Section 4), and its implementation (Section 5). Thereafter we present the results from simulations (Section 6) and discuss conclusions and future work (Section 7).

2. The effect of emotions on memory

This section considers the way emotional effects affect cortical function, and how this relates to semantic memory (generic and specific).

2.1. Edelman's value system

In this paper, we build on Edelman's framework of the arousal system (which he calls the "value system") [20].

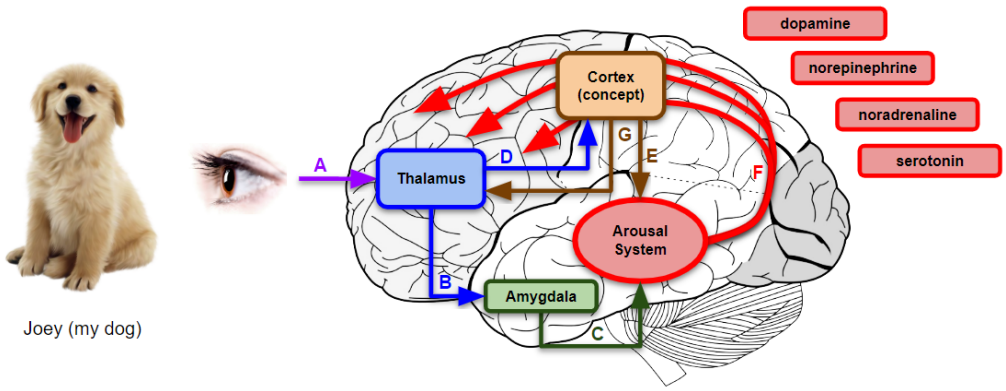

Fig 2. *There are four stages of emotional effect in the cortex. Firstly, external experiences and internal thoughts 5 give rise to an image in the cortex. External experiences are first received by the senses and are then sent via the thalamus (A) to both the amygdala (B), generating an unconscious response, and to the cortex (D), responsible for conscious response. Secondly, the experience can then cause a conscious (E), and/or a sub-conscious (C) emotional response. At the same time, top-down corticothalamic connections send information from the cortex back to the thalamus (G) [67]. Thirdly, the arousal system releases neuromodulators into the neocortex via diffuse ascending connections (F), having the effect of associating an emotional tag with a memory [22]. Fourthly and finally, if the same or similar events contexts/places/people are encountered later, the memory is recalled with the same emotional tag that was stored.*

According to Edelman [24] [21], there are four stages of emotional influence in the cortex (shown in Figure 2):

(i) External experiences (seeing your dog) and internal thoughts (remembering something good or bad) activates an image in the cortex. External experiences are first received by the senses and are then sent via the thalamus [73] (label "A" in Figure 2) to both the the amygdala (generating a unconscious response, label "B" in Figure 2), and to the cortex (responsible for conscious response, label "D" in Figure 2) [48]

(ii) The experience can then cause a conscious emotional response (label "E" in Figure 2), and/or a sub-conscious emotional response (label "C" in Figure 2). This emotional response takes the form of the release of neuromodulators (such as dopamine, noradrenaline and serotonin) from nuclei in the arousal system. This process of triggering is modulated by experience, and can be altered through learning (for example, learning to identify your mother's face in a crowd). At the same time, top-down corticothalamic connections send information from the cortex back to the thalamus (label "G" in Figure 2), which allow for contextual prediction to take place in the thalamus [67].

(iii) The associated spread of neuromodulators from these nuclei to the neocortex via diffuse ascending connections [44] (labelled as "F" in Figure 2) has three effects:
   (a) It induces emotional feelings (qualia), and can also possibly initiate immediate action in response in a sub-cortical (fast) way;
   (b) It alters neural connection strengths [20] so that memories of the event and associated contexts/places/people are stored together with the appropriate associated emotional tag (positive or negative) [22]; this alteration of cortical connection weights is the AND affect-shaped mechanism of neural plasticity [28]; In addition, once-off learning takes place if a strong enough emotional tag is attached to the event.
   (c) It results in a cognitive response, as the conscious brain responds to what has happened [17]. This can then lead to resulting considered action.

(iv) If the same or similar events contexts/places/people are encountered later, the memory is recalled with the same emotional tag that was stored.

In this paper we focus on the effect that neuromodulators have on memories (Item 3b), the storing of emotional tags associated with memories, and their recall (Item 4). We demonstrate how once-off learning takes place if a strong enough emotional tag is attached to the event. We do not model the immediate effects of the emotions (Item 3a), nor do we attempt to model how the emotional system is triggered (Item 2).

2.2. Types of memory

Long term memory is commonly classified as either *explicit* (declarative) or *implicit* (non-declarative) memory [5] (see Figure 3). Non-declarative memory is considered to act unconsciously, and covers activities like tying a shoelace, driving a car, or riding a bike. By comparison, declarative memory is available to the conscious mind, and can be broadly categorized as either *semantic* or *episodic* memory [75]. Semantic memory on the one hand describes facts, ideas, meaning, and concepts [65], and on the other specific members of classes identified in semantic memory, with an emotional tag attached to each. Episodic memory places specific items in a specific context, connecting specific semantic memories in a sequence, with a location, time, and with emotions [80].

To illustrate the difference between semantic and episodic memory, we provide an example. Semantic memory described unique classes of objects (e.g. a woman, your mother, a book, a bench). Concepts can be structured in a hierarchy of complexity, and (as with objects inheritance in computer science) semantic memories can inherit from parent classes. Semantic memories can be further classified as either Generic (classes of concepts) or Specific (instances of a concept); this corresponds to the distinction between a class and an instance of a class in object oriented programming [11]. For example, your mother's face could be a specific instance of the generic concept of a woman's face, which in turn could be a sub-class of the generic concept of a human face. An example of a computational model of semantic memory would be an Artificial Neural Network (ANN) that is designed to classify objects.

By comparison an episodic memory places a memory in a context (e.g. a location, with other objects, and in a particular instance in time). For example, a memory of "Mother reading while sitting on a bench in the park, at 2pm on 14th June." would be an episodic memory. This concept can be expressed in computer science as an instance of a class. An example of a computational model of episodic memory would be a Recurrent Artificial Neural Network (RNN), that is able to classify a sequence of objects.

In this paper we focus on modelling the value system associated with semantic memories, with the aim to extend this to episodic memory in the future. In the following section we take a closer look at the structure of diffuse projections, and the effects of salience in the cortex.

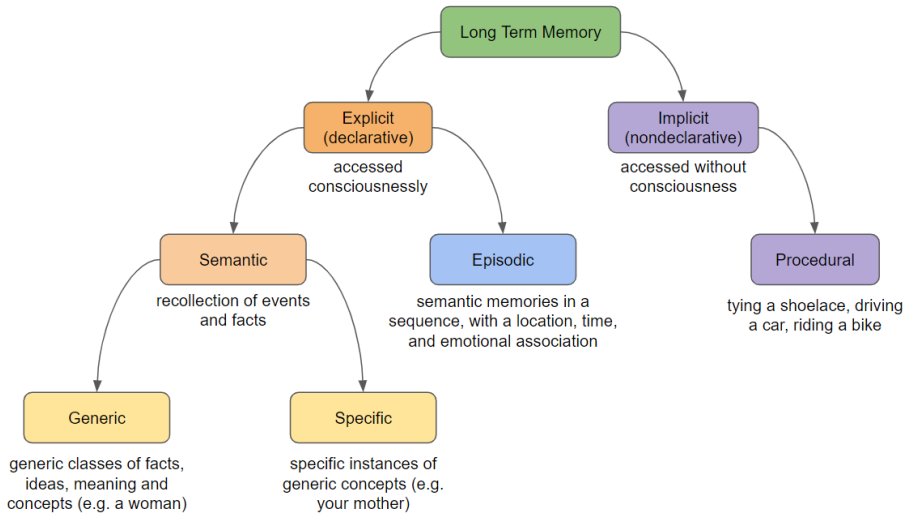

Fig 3. *Types of long term memory include Explicit and Implicit memories. Explicit memories can be further divided into Semantic and Episodic memories. Semantic memory be divided into Generic (classes) and Specific (instances of those classes). While emotions can be associated with both Explicit and Implicit memories, in this paper we focus on Semantic memories, and specifically how neuromodulators associated with emotions impact Generic and Specific memories.*

2.3. Cortical structures of diffuse projections

Within the motor, sensory and cognitive systems, the human brain has at least three classes of connections relating to the cortex [21]. These are all crucial to brain function, which is why they have been built into our brain through a variety of evolutionary processes [27].

The first class consists of localized inter-neuronal connections forming layered neural networks, which are modeled by standard artificial neural networks (ANNs). Standard ANNs [10] model the way signals flow locally in the columns in the cortex through local connections in those columns, allowing complex processes such as pattern recognition [10], image classification [14] [46] and naming [36] to occur.

The second class are recurrent connections between different cortical areas [21]. These are important in terms of cortical functioning, but are not the concern of the present paper.

A third class consists of diffuse projections from the arousal system [54] to the cortex, known as '*modulatory*' systems [44] or '*ascending*' systems [24]. From their nuclei of origin they send axons up and down the nervous system in a diffused spreading pattern (see Figure 1). The effect of the modulatory systems projecting profusely is that each associated neuromodulator (for example norepinephrine and dopamine) affects large populations of neurons, allowing non-local interactions to occur in the brain. The release of the neuromodulators affects both the neurons, and the strength of the connections between neurons (see Figure 4). Specifically, neuromodulators affect the probability that neurons in the neighborhood of value-system axons [20] [21] will fire after receiving excitatory input, thus they are an important mechanism effecting neural plasticity in the long term. These systems bias neuronal responses affecting both learning and memory by guiding neuronal group selection, and for this reason that they are sometimes termed value systems [23] [24]. Standard ANNs do not model this non-local flow of information represented by the ascending systems, which are a significant feature of the structure of the brain. The aim of this paper is to model the

effects of such connections in a new class of ANNs, which we refer to as *Salience-affected neural networks* (SANNs) because they allow representation of salience effects.

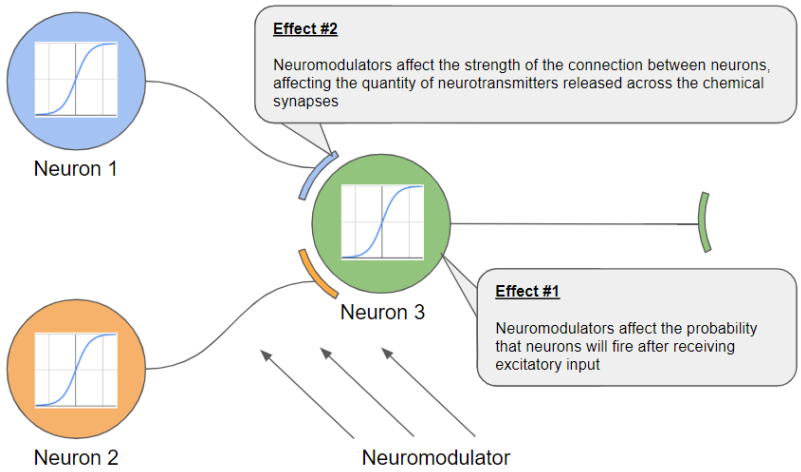

Fig 4. *The release of neuromodulators by the arousal system affects a whole region in the cortex simultaneously via the ascending systems, impacting those neurons that are active at the time the neuromodulators are released in proportion to their level of activation. A released neuromodulator affects both the neurons themselves (their behaviour) as well as the strength of the connections between neurons at synapses. Specifically, neuromodulators affect the probability that neurons will fire after receiving excitatory input, and they also affect the quantity of neurotransmitters released across the chemical synapses between neurons.*

The salience of an entity refers to its state or quality of standing out relative to neighboring entities. For example, a salient memory might be one that significantly stands out among others because of its emotional content. Beebe and Lachmann defined the three principles of salience that describe the interaction structures in the first year of life [8]. These are the principles of ongoing regulations, disruption and repair, and heightened affective moments. Ongoing regulation describes the characteristic pattern of repeated interactions, such as a child interacting with their mother or father. This is well represented by standard ANNs, which allow associational learning with multiple-trial (i.e. ongoing regulations). However heightened affective moments are dramatic moments standing out among other events [8], often leading to single-exposure learning, which is an established psychological phenomenon [70], and has been the subject of discussion in the psychology literature for over 50 years.

While the details of how single-trial training works in the cortex are in dispute, the fact that single-trial training sometimes occurs is well established [4] [51]. When it occurs, single-trial learning is often associated with the neuromodulator dopamine [81]. This is one of the neuromodulators released throughout the cortex by the ascending systems, and is associated both with the neural coding of basic rewards, and in reinforcement signals defining an agent's goals [72] [18] [57]; hence it is known to play an important role in brain functioning. Standard ANNs do not currently model the salience effects of neuromodulators such as dopamine, and are unable to implement single-exposure learning; but SANNs do both.

2.4. Characteristics of neuromodulators

A key characteristic of a neuromodulator is that it is present or absent in a *region* of the cortex, rather than being associated with specific synaptic connections. It can occur with various strengths in such a region, therefore we model the input salience signal of a

neuromodulator associated with positive salience in the SANN model as a positive signal with value between 0 and 1, and the input salience signal of a neuromodulator associated with negative salience as a negative signal with value between -1 and 0.‡

A second key characteristic of the effect of neuromodulators is that they affect the the entire pattern of activated nodes according to the level of activation of each synapse. Therefore, in the presence of a salience signal, both the connection weights and the sigmoidal activation function were adjusted depending on the current level of activation of each neuron/synapse (see Equation 1 and Equation 6 below); hence there is no change if the synapse is not active. The key feature of the SANN is that it is able to strengthen an entire pattern of activated nodes proportional to the strength of the salience signal, rather than just individual connections.

In the human brain, neuromodulators only influence those neurons that possess the relevant receptors for each modulator. Furthermore, neuromodulators receptors come in different sub-types. However, in modelling terms for the purposes of demonstrating the concept, we assume that neuromodulators are global in nature, affecting all nodes and weights in a network. The salience signal is therefore modelled as a single value, and associated with each node and weight in the network (see Figure 7).

3. Comparison with other Frameworks

In this section we compare SANNs with other neural network models.

3.1. Standard ANNs

Most neural network learning processes take many repetitions before a memory is stored. But in real life, some memories are stored after just one experience ('once-off' learning). In this paper we demonstrate that once-off learning takes place if a strong enough emotional tag is attached to the event.

3.2. Husbands' model of gas diffusion

The only other class of ANNs of which we are aware that represent non-local effects are models of the effects of Nitric Oxide (NO), which is a freely diffusing neuromodulator [39]. A model of gas diffusion is used in a class of ANNs in which nodes are capable of non-locally modulating the behaviour of other nodes. These models are called GasNets [39], as they simulate the presence of the gas NO in the environment surrounding the neurons. Like SANNs, this form of modulation allows a kind of plasticity in the network in which the intrinsic properties of nodes are changing as the network operates [39]. However to the best of our knowledge, GasNets have not been modified to train and test an ANN with specific salience, nor do they have the potential for single-trial learning in ANNs.

Our model differs from the research on GasNets by Husbands [39] as it models both the way memories associated with emotionally-laden events may be embodied in neural networks, and the way those emotional associations can be recalled when the events are remembered. Furthermore, this research introduces an additional input salience signal during training, and each node produces a nodal reverse salience signal during testing. This research also demonstrates that these kinds of networks enable single-exposure learning by associating strong salience signals with input combinations.

‡ In future work we will model a multidimensional salience system, with different neuromodulators associated with each dimension [61] [79]; however for this paper it will be adequate to think of a single neuromodulator that can have positive or negative valence.

3.3. Juvina's model of valuation and arousal

Juvina [41] presents an approach to adding primitive evaluative capabilities to a cognitive architecture in which two sub-symbolic quantities called valuation and arousal are learned for each declarative memory element, based on usage statistics and the reward it generates. Consequently each memory element can be characterized as positive or negative and having a certain degree of affective intensity. In turn, these characteristics affect the latency and probability of retrieval for that memory element. This is similar to what we do, but using a very different cognitive architecture. Ours is based closely on the way the diffuse systems interact with cognitive functioning in the human brain [20] [21], which is an affective process because these systems are those implicated in the primary emotional systems identified by Panksepp [61].

3.4. Edelman's Robots

Gerald Edelman created a range of brain-based devices (BBDs) including Darwin VII and Darwin X. [22] [45]. Edelman created 'Darwin VII' to capture a holistic picture of its environment, incorporating 3 different senses (vision, auditory, conductance or "taste"), a motor system, and a "value system" (attempting to model an ascending neuromodulatory value or reward system). Edelman then attempted to model long-term episodic memory (behaviour attributed to the hippocampus) in his model 'Darwin X'.

Edelman's robot 'Darwin VII' contained a computational nervous system of 20,000 neuronal units, and 450,000 synaptic connections. After some training 'Darwin VII' successfully learned to associate the taste value of the blocks with their visual patterns [22]. The sensory systems interpreted incoming signals (such as conductance of a block) as a 'value signal". In Edelman's model of the value system, the strength of connections between two neural units could change based on the value signal (i.e. a positive value signal would strengthen a connection).

Key difference relative to our paper: Edelman allowed the 'salience' signal in 'Darwin VII' to directly impact the strength of synaptic connections between neurons. By doing so, Edelman's model incorporates 'salience' as just another input signal (along with the visual, auditory and sensory signals) that requires extensive training. As a result, Edelman's BBD models required many iterations of training for the salience to be embedded in the model ,and associated with sensory input patterns. Due to the design of his model, Edelman's model was not able to demonstrate single exposure learning on an previously trained dataset.

By contrast, in this paper we demonstrate how the 'salience' signal from the arousal system (which Edelman called the 'value' system) can be modelled as an additional dimension to the neural network model. We demonstrate how the 'salience' signal can be either part of the training, or applied afterwards as a single exposure. By modelling the 'salience' signal as an additional dimension to the neural network model, we are able to demonstrate how the arousal system can affect neural activation patterns (e.g. memories) during single exposure, without affecting synaptic weights of the neural network.

4. Salience-affected neural network (SANN)

To model the effect that neuromodulators have on semantic memories we present a neural network that allows for feature extraction training on a standard data set, while in addition also allowing for the network to be affected globally by a *salience* signal related to a specific individual item.

Definition of terms We define the *salience* of an entity as its state or quality of standing out relative to neighboring entities, representing a positive or negative evaluation of that entity relative to individual welfare, and a salience signal (S) as an additional input signal to a SANN representing such a quality that affects each node during training. We also define the *reverse salience signal* (S') as the sum of the reverse salience signals produced by each node during testing. In practice this means that the SANN will allow for a salience signal to be associated with a specific image (i.e. a weighted combination of features), and for that same image (for example, a specific face) to product a reverse salience signal if that specific combination of features is experienced again in the future.

4.1. Expected performance characteristics

A Salience-affected neural network (SANN) that models diffuse projections in the cortex must be able to demonstrate the following characteristics:

(i) The salience signal should affect entire patterns of activated nodes, not just individual nodes, and not all the nodes in the network equally: rather the entire current pattern of activation of nodes is either strengthened or weakened.

(ii) The network should be able to be trained with the standard training data set, as well as an additional dimension of an incoming salience signal associated with each image.

(iii) The network should be able to produce a reverse salience signal (as an output) for a specific input and the associated pattern of activation.

(iv) Training with salience should have an overall positive impact on the classification accuracy of the network as a whole, and for images in the same class.

(v) Training with salience should have a positive impact on the classification confidence of the specific image tagged with salience during training.

It is important to note that in this paper we do not model how salience activation occurs; rather we model its outcomes.

4.2. Design: SANN

The Salience-affected Neural Network (SANN) presented in this paper is designed to classify handwritten digits (0-9). As part of the pre-processing, non-negative matrix factorization [49] (NMF) was applied to extract features from the dataset, reducing the size of the input layer.

In addition to a standard ANN, each node of the neural network is adapted to accept a global salience signal (S) during training, and to respond with a reverse salience signal (S') during classification (see Figure 5). Furthermore, when exposed to salience signal, the SANN also globally strengthens the connection weights in the network proportional to the activation levels of the parent nodes.

4.3. Dataset

We use the MNIST database of handwritten digits [47], consisting of 70,000 labelled, 28x28 grayscale images of the 10 handwritten digits (0-9). The training set consists of 60,000 images while the testing set consists of 10,000 images (see Figure 9). This dataset has been widely used in conjunction with supervised and unsupervised neural networks models [12] [43] [59], to provide a realistic dataset for simulations of biologically-inspired models, similar to the one we present in this paper.

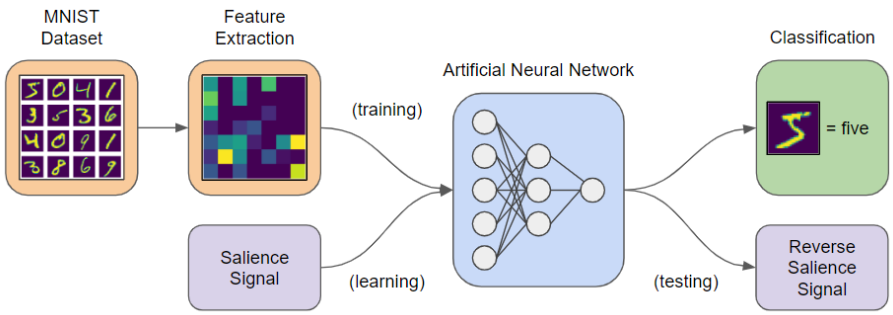

Fig 5. *We present a salience-affected neural network (SANN). The architecture of the model starts with a feature extraction algorithm, to extract features from the dataset. Feature extraction also has the impact of reducing the size of the input layer required for the neural network. Each image is then reduced to a weighted set of features, and this input is fed to the input layer of the neural network. The neural network is a standard feed-forward neural network, and is trained as a classifier. In addition to the input values, we add an additional dimension: a global salience signal. This salience signal (S) consists of a single value, and affects all nodes in the neural network. Also, the neural network is designed to produce a reverse salience signal (S') when an image is presented to the model, so that each image produces both an output (classification) as well as a reverse salience signal (S').*

4.4. Feature extraction (NMF)

As part of the pre-processing, non-negative matrix factorization [49] (NMF) was applied to extract features from the dataset. A NMF component layer size of 49 features was chosen. The MNIST dataset was therefore reduced from a set of images of 784 pixels each, to a set of images with with 49 features per images. The input layer of the neural network was reduced from 784 to 49 nodes (see Figure 6).

NMF pre-processing was chosen for a two reasons. Firstly, it is known that 2-layer feed forward networks can achieve 98.4 - 99.3 % classification accuracy [74]. We require a wider range of accuracy to demonstrate the impact of an SANN on classification accuracy, hence we degrade the maximum accuracy by reducing the dimensionality from 784 (28x28) input values, to 49 input values using NMF. Secondly, NMF is known to produce "human like" features from raw data, so it is a reasonable technique to use for the purpose of demonstrating the SANN model.

4.5. Neural network model

The model we use is a fully-connected artificial neural network (ANN), configured with 49 input nodes, a hidden layer of 100 nodes, and an output layer of 10 nodes (see Figure 6). Each node in the output layer represents a specific digit (0-9).

We chose an ANN because it is a well-characterized and common neural network from which many other neural networks are derived [2]. For this research, the neural network was adapted from an open-source neural network model [71]. The neural network model could be extended by addition of many more layers, or even as a deep neural network [60].

In this model we chose to use the sigmoidal activation function. We felt that the sigmoidal activation function was better suited than other activation functions (e.g. maxout functions [34], ReLu functions [58], PLUs [78]) to the SANN model for two reasons. Firstly, because sigmoidal activation functions are more biologically realistic: most biological systems saturate at some level of stimulation (where activation functions like ReLu do not). And secondly, sigmoidal functions allow for bipolar activations,

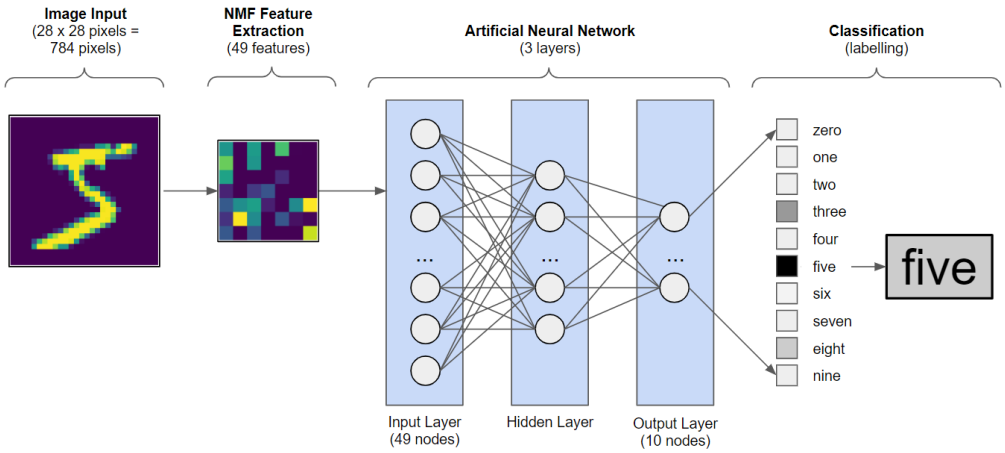

Fig 6. *Each layer on the SANN model has different dimensions. The input layer consists of 784 values per image (28 x 28 pixels). Each image is passed through a feature extraction algorithm, reducing its dimension to a weighted set of 49 features. The input layer of the neural network is 49 nodes wide, to accept the set of 49 features. The hidden layer of the neural network is 100 nodes, and the output layer of the neural network is 10 nodes, each node in the output layer representing a specific digit (0-9).*

whereas functions like ReLu are effectively monopolar and hence not useful for a reverse salience signal with both polarities of activation.

4.6. Salience signal

To incorporate salience, an additional signal was embedded in a standard ANN [2] (the salience signal), such that it directly affected every individual node during training, and collectively produced a reverse salience signal during testing (see Figure 7).

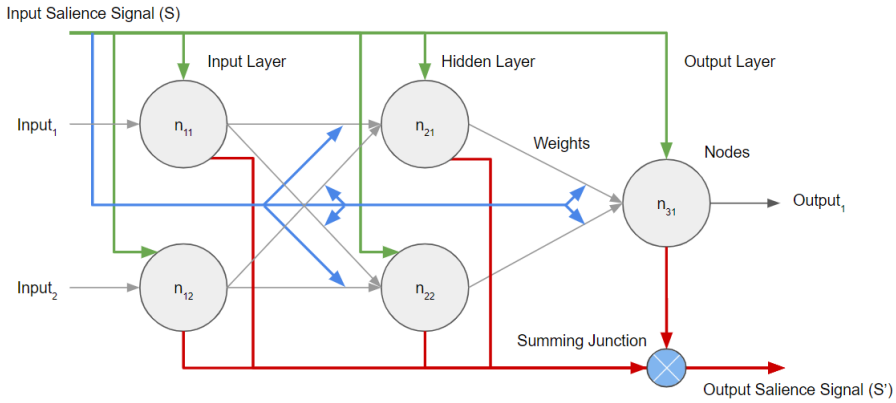

Fig 7. *The Salience-affected neural network (SANN) model is an artificial neural network (ANN) with an additional signal dimension added: the salience signal. The salience signal affects all nodes (green arrows) in the network directly during training. Each node in the network then produces a reverse salience signal during testing, which signals are summed to produce a single reverse salience signal for the network. During training, the salience signal also has the impact of altering the weights (blue arrows) of the connections between all nodes in the network simultaneously by an amount proportional to the activation level of the subsequent node and the salience signal strength.*

Similar ANN input combinations would be expected to produce similar reverse salience signals, because the reverse salience signal is defined as the summation of the nodal reverse salience signals observed at each node.

5. Implementing a salience signal in the SANN

The addition of a global salience signal in the SANN introduces an additional dimension to the standard neural network (ANN). The global salience signal (S) was designed to have two distinct effects.

First, to accept a global salience signal (S) which would affect each node in the network simultaneously, and produce a reverse salience signal for future inputs. To achieve this we enable the global salience signal to affect the sigmoidal activation function of each node in the ANN. In this section we discuss and evaluate 3 different ways in which we allow the salience signal to impact the activation function.

Secondly, the global salience signal (S) was also designed to to alter the strength (weights) of all connections between nodes, by an amount proportional to the activation level and the salience signal strength (Equation 6).

The variations of the activation function proposed in this research for use with the SANN closely mirror the variations described by Husbands [39]. This research focuses firstly on the threshold offset variable, referred to as b_i by Husbands [39], and secondly on the alteration of node weight, corresponding to Edelman's idea of Neural Darwinism [20] [28].

In this section we discuss the two effects the input salience signal has in the SANN, firstly the effect on the activation function at each node, and secondly, the effect on the weight of each connection in the network. We also discuss the reverse salience signal, which can occur because of the alteration to the activation function.

5.1. Effect 1: Activation Function

Each node was assigned a salience value. This salience value (S_i) was defined as a value between -1 and 1, calculated relative to the input salience signal (S_{global}), and the current activation level of the node (A_i). A constant (β) was introduced to allow for different learning rates, which might differ from neuron type to neuron type, or with the kind or neuromodulator modelled.

$$S_i(new) = S_i(old) + (1 - S_i(old)) \times A_i \times S_{global} \times \beta \tag{1}$$

In the SANN, the input salience signal (S_{global}) affects the network such that the network produces a reverse salience signal (S') for future input values. As previously mentioned, neuromodulators affect the probability that neurons will fire after receiving excitatory input. To maintain a link to biological behaviour of neurons, salience was implemented such that it affected the activation function. The activation function chosen in this paper is a sigmoid activation function (see Equation 2).

$$y(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Modifications We investigate three possible modifications to the activation function, namely (a) a horizontal offset, (b) a change in the gradient, and (c) a change in the amplitude (see Figure 8).

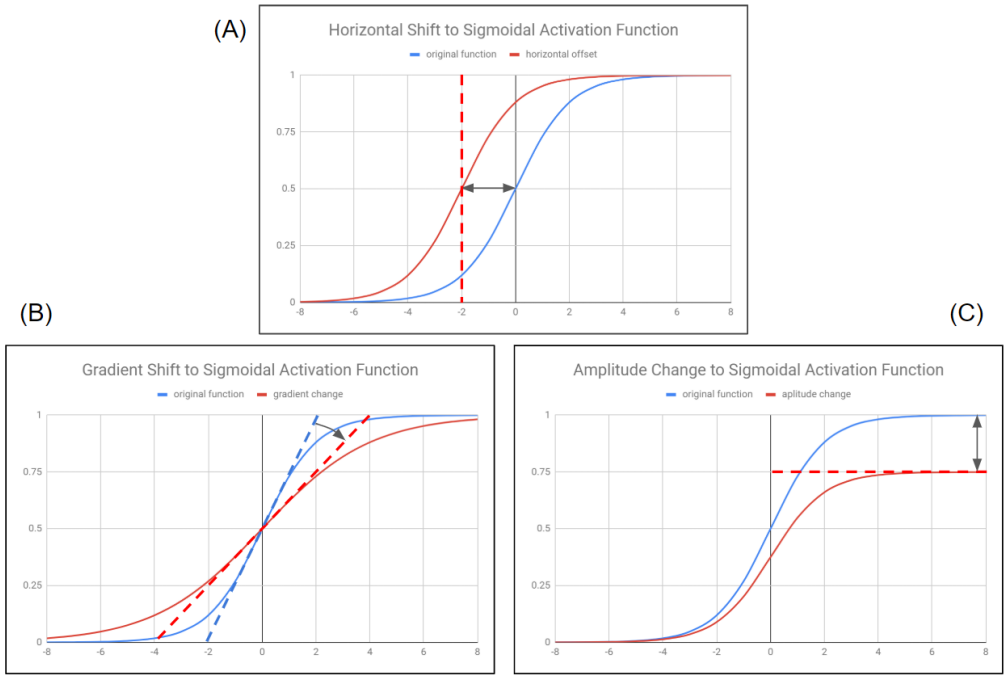

Fig 8. *Three changes to the activation function were explored: (A) Change in the horizontal offset of the activation function: A positive salience signal would shift the activation function to the left, resulting in a higher output from the activation the next time. (B) Change in the gradient of the activation function: A positive salience signal would reduce the gradient of the activation function, resulting in a higher output from the activation the next time. (C) Change in the amplitude of the activation function: A positive salience signal would increase the amplitude of the activation function, resulting in a higher output from the activation the next time.*

Effect 1a: Horizontal Offset The first of three effects was to introduce a horizontal shift of the activation function along the x-axis (see Figure 8A). A positive salience signal would shift the activation function to the left, resulting in a higher output from the activation the next time. The activation function incorporated the horizontal offset as shown in Equation 3.

$$y(x) = \frac{1}{1 + e^{-(x+S_i)}} \tag{3}$$

Effect 1b: Gradient Change The second of three effects was to introduce a change in the gradient of the activation function (see Figure 8B). A positive salience signal would reduce the gradient of the activation function, resulting in a higher output from the activation the next time. The activation function incorporated the gradient change as shown in Equation 4.

$$y(x) = \frac{1}{1 + e^{-(0.5^{-S_i} \times x)}} \tag{4}$$

Effect 1c: Amplitude Change The third of three effects was to introduce a change in the amplitude of the activation function (see Figure 8C). A positive salience signal would increase the amplitude of the activation function, resulting in a higher output from the activation the next time. The activation function incorporated the amplitude

change as shown in Equation 5.

$$y(x) = \frac{0.5^{-S_i}}{1 + e^{-x}} \tag{5}$$

5.2. Effect 2: Strengthening of connection weights

In addition to producing a reverse salience signal (S'), the global salience signal (S_{global}) also has the impact of altering the weights of the connection between all nodes (W_{ij}) in the network simultaneously, by an amount proportional to the current weight of the connection (W_{ij}), the strength of the input salience signal (S_{global}), and the activation level of the destination node (A_j).

$$W_{ij}(new) = W_{ij}(old) \times (1 + \alpha \times |S_{global} \times A_j|) \tag{6}$$

Here α is a constant which might differ from neuron type to neuron type, or with the kind or neuromodulator modelled. Note that because the updating equation for each synapse weight depends on the current state of activation A_i of that synapse, just as in the case of Neural Darwinism [20], it is the entire current pattern of activation that gets reinforced by the salience signal.

5.3. Classification accuracy

To measure the impact of salience on the neural network, we measured the classification accuracy (τ) of the class, and the classification confidence (ϕ) of the individual image.

Classification accuracy (class): The classification accuracy (τ) of a class was calculated as the number of correct classifications, divided by the total number of classifications performed. The inverse of the classification accuracy is often referred to as the error rate.

Classification confidence (individual image): The classification confidence (ϕ) of the individual image (i) described how confident the classification algorithm is to assigning an image to a class. The classification confidence (ϕ) was calculated as the highest output value the neural network ($max(o)$). The amplitude change was normalized when calculating the classification accuracy by a factor of $0.5^{(-S_{global})}$, because it was the only function to change the amplitude of the signal propagating through the network.

5.4. Reverse Salience Signal

The changes made to the activation function allow the value of salience associated with a specific image to be read off when that image is presented. This is the *reverse salience signal* ($S'_{network}$). If an image (or a member of its class) had high salience signal (S_{global}) associated with it during training, then we would expect a high reverse salience signal ($S'_{network}$) during testing. The reverse salience signal for each node (S'_i) was calculated as the product of the current activation level A_i, and the reverse salience signals of each node (S_i) (see Equation 7). The reverse salience signal for the network ($S'_{network}$) was calculated as the sum of the reverse salience signal of each node in the network (see Equation 8).

$$S'_i = |A_i| \times S_i \tag{7}$$

$$S'_{network} = \sum S'_i \tag{8}$$

It was observed that, due to the nature of the definition of reverse salience signal value, once the SANN was trained with salience, the reverse salience signal values would have a global offset. We refer to this offset as the *residual reverse salience value*. The reverse salience signal was defined as the average reverse salience signal over a set of test images. The reverse salience signal was then adjusted to exclude the residual reverse salience signal.

6. Results

In this section we describe the results obtained from training the SANN with salience. We are concerned with the effect of a salience signal applied to a particular single representative image within a class of images both on (a) recognition of that class of images (*Semantic memory: Generic*, see Figure 3), and (b) recognition of that particular image (*Semantic memory: Specific*, see Figure 3), so we clearly separate these out in what follows.

6.1. Neural network training

In this section, we review the impact of the changes to the activation function indicated in Fig. 8 (Eqns. (3)-(5)), as well as the change in weights (Eqn. (7)). We clearly separate effects on class recognition and on individual image recognition in our results below.

Classification training: The neural network was designed with 49 inputs, a single hidden layer of 100 nodes and an output layer of 10 nodes (see Figure 7). The neural network was trained on 1,000 images, over a period of 25 epochs, with a learning rate of 0.5. The trained network was then tested on 1000 test images, and correctly classified 776 of the 1000 test image (categorization accuracy of 77.6% for the MNIST dataset). Our aim in this paper is not to optimize the performance of the neural network by changing the size of the neural network layers, the size of the training set, the number of epochs, or the learning rate. Rather we establish the baseline performance of a sufficiently accurate neural network classifier, and document the impact of training the network with a global salience signal (S).

Single exposure salience training: After the SANN had been trained to classify the images in the MNIST dataset, the network was exposed to a salience signal as discussed above in previous sections. To achieve this, a single image in the training set was tagged with salience, and then the network underwent a single-exposure training event (see Figure 9). We repeated this process for every individual image in the training set.

6.2. Effect 1: Activation Function

In this section we will discuss the results observed for (a) class recognition, and for (b) individual image recognition. Specific measures we will discuss include reverse salience signal value, and classification accuracy. We will also discuss the results of the three possible modifications to the activation function, namely (a) the horizontal offset, (b) the gradient, and (c) the amplitude.

Reverse salience signal value (class): The results show that after a single exposure training iteration, the SANN associates a strong reverse salience value (S') to all inputs that are similar to the salience-affected value. The reverse salience signal (S') was notably grouped by similarity and by class (see Figure 10). The more similar an input image was to the image trained with salience, the higher the reverse salience

Fig 9. *The SANN model was trained using the MNIST database of handwritten digits [47], consisting of 70,000 labelled, 28x28 grayscale images of the 10 handwritten digits (0-9). The training set consists of 60,000 images while the testing set consists of 10,000 images. A single image in the training set chosen for salience tagging in one particular run is shown here.*

signal. Similarity was calculated using MSE. This result supports the principle that salience applied to the member of a class, will have an impact on similar inputs, having the highest impact on the other members of the same class.

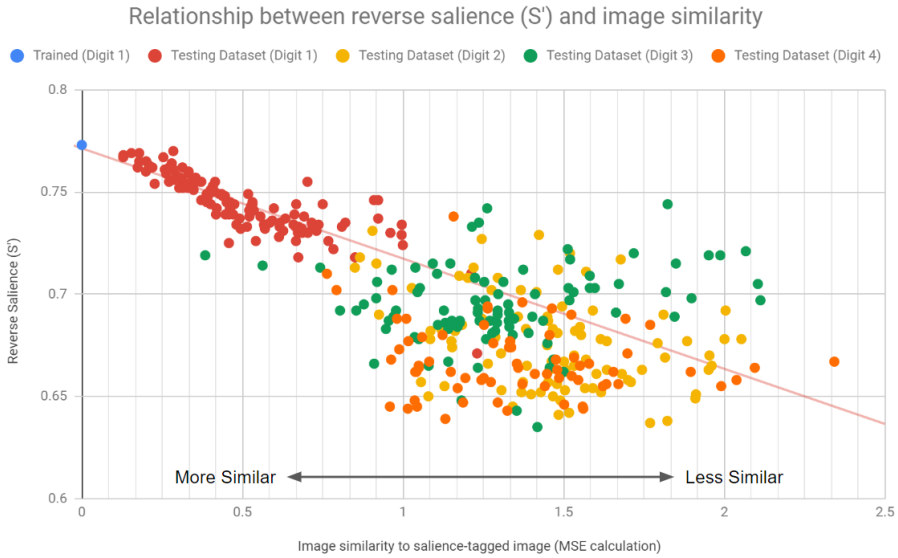

Fig 10. *After salience training, the neural network produced a reverse salience signal (S') for every input image. Input images most similar to the salience-tagged image (based on an MSE calculation) presented with the highest reverse salience signal (S'). The reverse salience signal was also notably grouped by classification (i.e. digit). In this diagram image #7 in the training set was tagged with salience. Image #7 was correctly classified as belonging to class: Digit #1.*

Reverse salience signal value (individual image): The results shows that the image trained with salience produced the highest reverse salience signal (S') (see Figure

10). This result supports the principle that if an input is accompanied with salience, then that same input will produce a strong reverse salience signal if it is seen again.

Classification accuracy (class): At the class (generic) level, we observed that each of the three modifications to the activation function had varying effects on the network's classification accuracy (ϕ). The horizontal offset and amplitude change of the activation function both had a negative impact on the average classification accuracy, however, the gradient change had a positive impact on the classification accuracy (see Figure 11). It was observed that the higher the salience magnitude (for gradient change), the higher the overall improvement in network classification. The highest average improvement in network classification accuracy was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement of classification accuracy (ϕ) of 0.61%, from the benchmark performance of 77.60% to 78.21%.

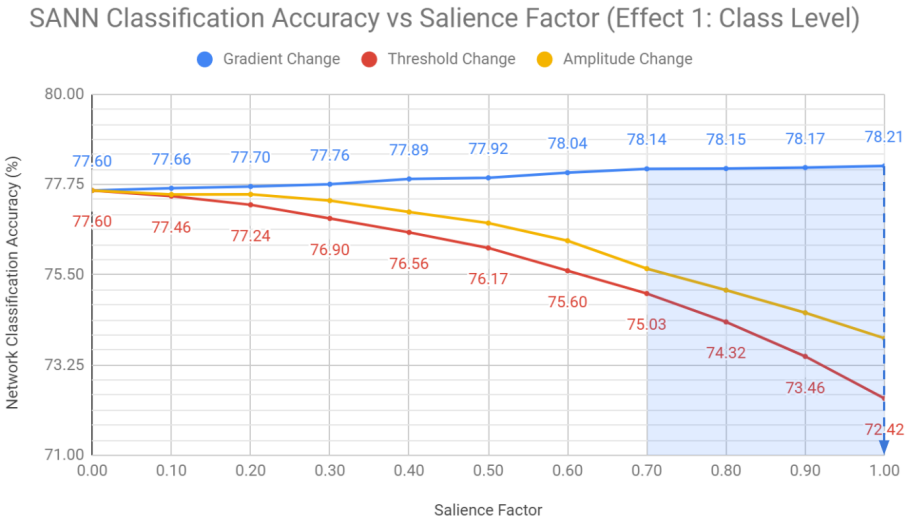

Fig 11. *Each of the three modifications to the activation function had varying effects on the network's classification accuracy (ϕ) at an individual image level. For each modifications to the activation function, the input salience factor (S) was varied from 0 - 1, and the network performance was recorded. It was observed that the horizontal offset and amplitude change of the activation function both had a negative impact on the average classification accuracy, however, the gradient change had a positive impact on the classification accuracy. The highest average improvement in network classification accuracy was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement of 0.61%, from the benchmark performance of 77.60% to 78.21%.*

Classification accuracy (individual image): At the individual image (instance) level, we observed that each of the three modifications to the activation function had varying effects on the network's classification confidence (τ) for the salience-tagged image. The horizontal offset and amplitude change of the activation function both had a negative impact on the average classification confidence. However, the gradient change had a positive impact on the classification accuracy (see Figure 12). It was observed that the higher the salience magnitude (for gradient change), the higher the overall improvement in the classification confidence for the salience-tagged image. The highest average improvement in network classification confidence for the salience-tagged image was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement in classification confidence (τ) of 4.89%, from the benchmark performance of 90.41% to 95.20%.

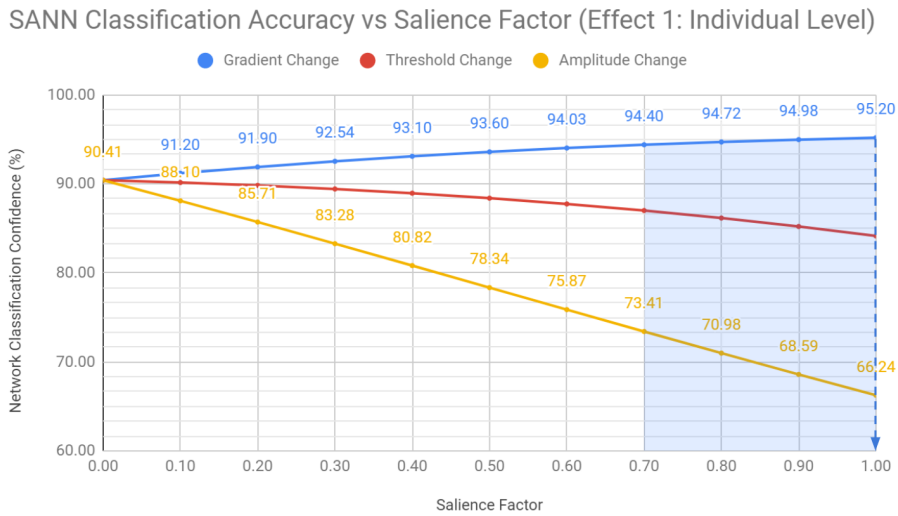

Fig 12. *Each of the three modifications to the activation function had varying effects on the network's classification confidence (τ) for the salience-tagged image. The horizontal offset and amplitude change of the activation function both had a negative impact on the average classification confidence. However, the gradient change had a positive impact on the classification accuracy. The highest average improvement in network classification confidence for the salience-tagged image was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement in classification confidence (τ) of 4.89%, from the benchmark performance of 90.41% to 95.20%.*

In this section we have observed the effect of a global salience signal on the activation function of each node and shown how a salience signal can improve the average classification accuracy (ϕ) by 0.61%, and the average classification confidence (τ) by 4.89%. In the next section we discuss the effects of the global salience signal on the connection weights.

6.3. Effect 2: Strengthening of connection weights

In addition to impacting the activation function of each node, the global salience signal (S) also affects each of the connection weights. It is important to note that this process is very different from back-propagation because it affects all weights globally at the same time, only factoring in the global salience signal and the current activation level of each node.

Classification accuracy (class): During a single exposure of global salience (S), the salience signal affects the connection weights proportional to the current activation level of the nodes. The impact of this single iteration of training on the overall network was a clear improvement in the classification accuracy (ϕ) of images. The highest average improvement in network classification accuracy (ϕ) was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement of classification accuracy (ϕ) of 0.61%, from the benchmark performance of 77.60% to 78.21%.

Classification confidence (individual images): At the individual image (instance) level, we also observed an improvement in the classification confidence (τ) of images in response to a change in weights (see Figure 14). The highest average improvement in network classification confidence (τ) for the salience-tagged image was observed with a gradient change only and a salience factor of 1.0, resulting in an

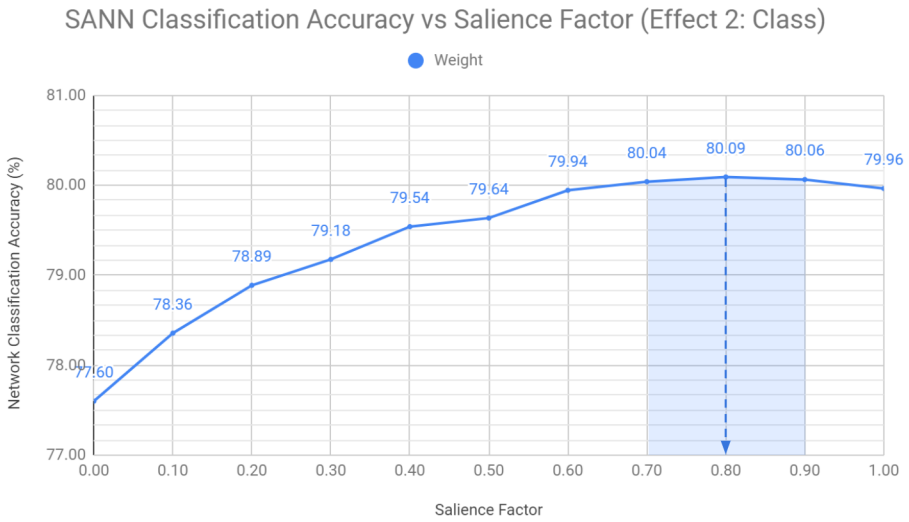

Fig 13. *In addition to affecting the activation function of each node in the network, the salience signal (S) also strengthens the weights of connections between nodes, proportional to the activation level of the post-synaptic node. The input salience factor (S) was varied from 0 - 1, and the network performance was recorded. It was observed that for a salience magnitude of 0.8 resulted in the greatest overall improvement in network classification: 2.49%. After salience training, the SANN network was able to accurately categorize 80.06% for the MNIST dataset, compared to the benchmark of 77.6% accuracy before the salience training.*

average improvement in classification confidence (τ) of 5.56%, from the benchmark performance of 90.41% to 95.97%. At a salience factor of 0.8, the average improvement in classification confidence (τ) is still significant: 5.48%, from the benchmark performance of 90.41% to 95.89%.

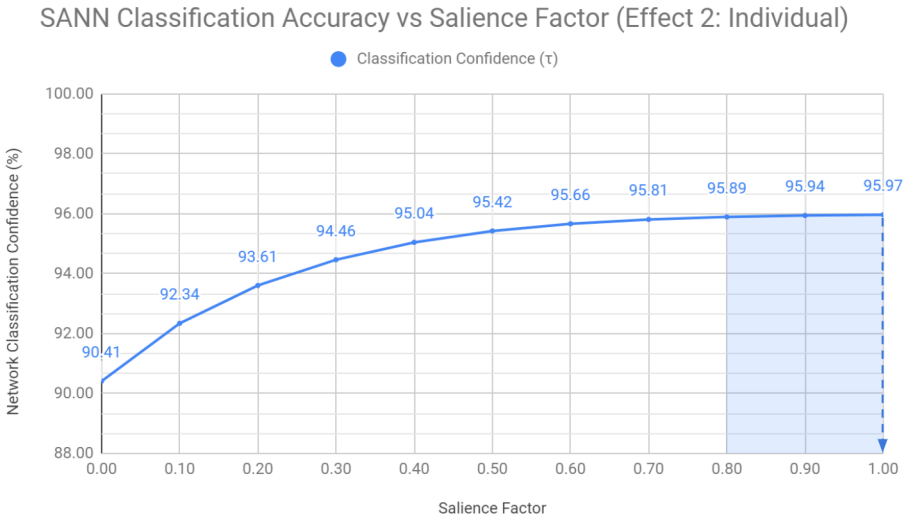

Fig 14. *At the individual image (instance) level, we also observed an improvement in the classification confidence (τ) of images in response to a change in weights. The highest average improvement in network classification confidence (τ) for the salience-tagged image was observed with a gradient change only and a salience factor of 1.0, resulting in an average improvement in classification confidence (τ) of 5.56%, from the benchmark performance of 90.41% to 95.97%.*

In this section we have observed the effect of a global salience signal on the activation function on the connection weights in the network. With a salience factor of 0.8, this effect improves the average classification accuracy (ϕ) by 2.49%, and the average classification confidence (τ) by 5.48%.

6.4. Combining Effect 1 and Effect 2

Having reviewed the results of Effect 1 and Effect 2 separately, we then combined Effect 1 (combined gradient and amplitude change) and Effect 2 (changing connection weights). By combining these effect we were able to test the impact of single-iteration training of a global salience signal on both the activation function of each node (Effect 1) and the connection weight of each connection in the network (Effect 2).

Benchmark performance: The benchmark performance of the neural network before salience training was a classification accuracy (ϕ) of 77.6%, and a classification confidence (τ) of 90.41%. The best performance for Effect 1 alone took place with a salience magnitude of 1.0: an improvement in network classification accuracy (ϕ): 0.61%, and an improvement in network classification confidence (τ): +4.89%. The best performance for Effect 2 alone took place with a salience magnitude of 0.8: an improvement in network classification accuracy (ϕ) of 2.49%, and an improvement in network classification confidence (τ) of 5.56%.

Classification accuracy (class): We combined Effect 1 and Effect 2, and varied the salience factor for each effect to create a performance heat map. We noticed an improved overall higher classification accuracy (ϕ) across the entire training set, compared to either effect alone (see Figure 15A). The best performance for Effect 1 and Effect 2 combined was an improvement in network classification accuracy (ϕ) 2.56% (an improvement from 77.6% to 80.16%). This result is 0.07% higher than Effect 2 alone, and 1.95% higher than Effect 1 alone.

In addition to calculating average performance, we also recorded the standard deviation of the results (see Figure 15C). The standard deviation of the dataset describes how similar the results were within the result set. The smaller the standard deviation, the closer to the average the results in the dataset were. While we would like to maximize the average performance improvement, we will seek to minimize the standard deviation of the performance improvement. The lowest standard deviation of classification accuracy (ϕ) observed for Effect 1 and Effect 2 combined was 0.00% (when both salience factors were set to 0). The higher the salience factors, the higher the standard deviation for the result set.

Classification confidence (individual images): After combining Effect 1 and Effect 2, we also noticed an improvement in average classification confidence (τ) across each individual image in the training set (see Figure 15B). The best performance for Effect 1 and Effect 2 combined was an improvement in classification confidence (τ) from 90.41% to 96.0% (an improvement of 5.59%). This was 0.03% higher than Effect 2 alone, and 0.7% higher than Effect 1 alone.

We also performed a standard deviation analysis on the classification confidence (τ) (see Figure 15D). The lowest standard deviation of classification accuracy (ϕ) observed for Effect 1 and Effect 2 combined was 19.54% (when salience factor of Effect 1 was 0.0 and the salience factor for Effect 2 was 1.0). However, there were several values of salience factors that produced this result.

Optimization (overall performance): To optimize the salience factors for the SANN model as a whole, an optimization value (γ) was calculated to maximize the average values of classification accuracy (ϕ) and classification confidence (τ), and to

minimize the standard deviations of classification accuracy (ϕ) and classification confidence (τ) (see Equation 9).

$$\gamma = \frac{\overline{\phi} \times \overline{\tau}}{stdev(\phi) \times stdev(\tau)} \tag{9}$$

The optimization value was calculated for varying salience factors, and a heatmap of this optimization value was created to visualize the optimal salience factors for the SANN model as a whole (see Figure 15E). The optimal salience factors to maximize the average values of classification accuracy (ϕ) and classification confidence (τ), and to minimize the standard deviations of classification accuracy (ϕ) and classification confidence (τ) were a salience factor of 0.7 for Effect 1, and 0.4 for Effect 2. At these salience factor values, the network demonstrates an improvement in average classification accuracy (ϕ) of 80.1%, and an improvement in average classification confidence (τ) of 96.0%.

Benchmark comparison: Compared to the benchmark performance of the neural network before salience training, a salience factor of 0.7 for Effect 1, and 0.4 for Effect 2 resulted in the following performance improvement: average classification accuracy (ϕ) improvement of 2.51%, and an improvement in average classification confidence (τ) of 5.60%.

6.5. Discussion

We made notable observations at various levels, namely at the level of (a) the entire network, (b) each classification category, and (c) individual images.

Level A: entire network: After salience training, the SANN model as a whole experienced an additional dimension: salience. A global salience signal (S) was added as a single input during training, and a reverse salience signal (S') was added as an additional output. The results shows that the image trained with salience produced the highest reverse salience signal (S') (see Figure 10) during testing, and that images similar to the salience-tagged image also produce a relatively high reverse salience signal (S') during testing.

Level B: classification category (generic memories): An individual image tagged with salience would experience, on average, an improvement in classification accuracy (ϕ) improvement of 2.51% (see Figure 15A).

Level C: individual images (specific memories): Crucially, the specific image tagged with salience was positively affected by the salience training. An individual image tagged with salience would experience, on average, an improvement in average classification confidence (τ) of 5.60% (see Figure 15B).

A classification accuracy (ϕ) improvement of 2.51% at the class level may not seem significant at first, but this result is in fact quite significant. The classification accuracy benchmark for the neural network before salience training was 77.6%, which was achieved after 50 epochs of training with a training set of 1000 images per epoch. To achieve a 2.51% improvement in accuracy, the same network would require an additional 38 epochs of back propagation training, with 1000 input images per epoch: a total total exposure of 38,000 images (see Figure 17). The results show that the SANN model is able to achieve this significant improvement in overall classification accuracy with just a single iteration of global salience training.

Having demonstrated that a classification accuracy (ϕ) improvement of 2.51% at the class level is significant, it is then clear that an average classification confidence (τ)

Average Classification Accuracy (class level) (φ)

A

Salience Factor (Effect 2 - weights) \ Salience Factor (Effect 1 - activation)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	77.60	78.36	78.89	79.18	79.54	79.64	79.94	80.04	80.09	80.06	79.96
0.1	77.66	78.36	78.90	79.25	79.52	79.69	79.93	80.08	80.06	80.06	79.95
0.2	77.70	78.42	78.95	79.36	79.52	79.78	79.94	80.08	80.08	80.04	79.93
0.3	77.76	78.46	78.98	79.46	79.52	79.82	79.97	80.11	80.04	80.00	79.86
0.4	77.89	78.50	79.08	79.53	79.57	79.84	79.99	80.11	80.00	79.98	79.82
0.5	77.92	78.57	79.10	79.59	79.63	79.88	79.99	80.08	79.99	79.90	79.74
0.6	78.04	78.64	79.18	79.62	79.68	79.91	80.04	80.08	80.02	79.86	79.66
0.7	78.14	78.71	79.22	79.68	79.76	79.93	80.08	80.10	80.00	79.84	79.57
0.8	78.15	78.75	79.30	79.67	79.83	79.97	80.13	80.07	79.96	79.77	79.54
0.9	78.17	78.80	79.36	79.71	79.92	80.01	80.14	80.07	79.88	79.71	79.47
1	78.21	78.82	79.46	79.74	79.97	80.08	80.16	80.06	79.84	79.64	79.43

Average Classification Confidence (individual image level) (τ)

B

Salience Factor (Effect 2 - weights) \ Salience Factor (Effect 1 - activation)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	90.41	92.34	93.61	94.46	95.04	95.42	95.66	95.81	95.89	95.94	95.97
0.1	91.20	92.93	94.05	94.79	95.28	95.58	95.77	95.87	95.93	95.96	95.98
0.2	91.90	93.45	94.43	95.07	95.47	95.71	95.84	95.92	95.96	95.98	95.99
0.3	92.54	93.91	94.76	95.30	95.62	95.80	95.90	95.95	95.98	95.99	95.99
0.4	93.10	94.30	95.04	95.48	95.73	95.86	95.93	95.97	95.99	95.99	96.00
0.5	93.60	94.64	95.26	95.62	95.81	95.91	95.96	95.98	95.99	96.00	96.00
0.6	94.03	94.93	95.44	95.72	95.87	95.94	95.97	95.99	96.00	96.00	96.00
0.7	94.40	95.16	95.58	95.80	95.91	95.96	95.98	95.99	96.00	96.00	96.00
0.8	94.72	95.36	95.69	95.86	95.94	95.98	95.99	96.00	96.00	96.00	96.00
0.9	94.98	95.51	95.78	95.90	95.96	95.98	95.99	96.00	96.00	96.00	96.00
1	95.20	95.63	95.84	95.93	95.97	95.99	96.00	96.00	96.00	96.00	96.00

Classification Accuracy **Standard Dev** (class level) (φ)

C

Salience Factor (Effect 2 - weights) \ Salience Factor (Effect 1 - activation)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0.00	1.06	1.39	1.70	1.98	3.32	1.94	1.98	2.53	2.13	3.20
0.1	0.57	1.23	1.52	1.77	2.23	3.57	1.78	1.85	2.54	2.48	3.16
0.2	0.94	1.10	1.42	1.79	2.38	3.03	2.00	1.80	2.44	2.98	3.28
0.3	1.10	1.60	1.59	1.57	2.57	2.92	2.34	1.70	2.56	3.57	3.61
0.4	1.42	1.75	1.62	1.66	2.93	2.53	2.02	1.66	2.73	3.47	3.67
0.5	1.83	1.57	1.99	2.14	2.62	2.73	2.04	2.33	2.76	3.26	3.94
0.6	1.77	2.10	2.30	2.19	2.53	2.52	2.04	2.53	3.11	3.49	4.05
0.7	2.32	2.10	2.58	2.06	3.01	2.24	1.87	2.56	3.07	3.48	4.13
0.8	2.52	1.86	2.92	2.24	2.77	2.09	1.97	2.66	3.10	3.85	4.35
0.9	2.60	1.91	3.40	2.06	2.83	2.30	2.14	2.54	3.52	4.12	4.47
1	2.53	2.34	3.03	2.04	2.66	1.90	2.48	2.68	3.77	4.04	5.10

Classification Confidence **Standard Dev** (individual image level) (τ)

D

Salience Factor (Effect 2 - weights) \ Salience Factor (Effect 1 - activation)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	21.32	20.55	20.01	19.71	19.59	19.55	19.55	19.57	19.58	19.59	19.59
0.1	20.93	20.26	19.83	19.62	19.56	19.55	19.56	19.57	19.58	19.59	19.59
0.2	20.59	20.02	19.70	19.57	19.55	19.56	19.57	19.58	19.59	19.59	19.59
0.3	20.30	19.83	19.61	19.55	19.55	19.56	19.58	19.59	19.59	19.59	19.60
0.4	20.06	19.70	19.57	19.54	19.56	19.57	19.58	19.59	19.59	19.60	19.60
0.5	19.87	19.62	19.55	19.55	19.57	19.58	19.59	19.59	19.59	19.60	19.60
0.6	19.74	19.57	19.54	19.56	19.57	19.58	19.59	19.59	19.60	19.60	19.60
0.7	19.64	19.54	19.54	19.56	19.58	19.59	19.59	19.60	19.60	19.60	19.60
0.8	19.58	19.54	19.55	19.57	19.58	19.59	19.59	19.60	19.60	19.60	19.60
0.9	19.55	19.54	19.56	19.58	19.59	19.59	19.60	19.60	19.60	19.60	19.60
1	19.54	19.55	19.57	19.58	19.59	19.59	19.60	19.60	19.60	19.60	19.60

Calculated Optimization Value (γ)

E

Salience Factor (Effect 2 - weights) \ Salience Factor (Effect 1 - activation)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0.00	3.20	4.32	4.43	4.75	2.99	5.91	6.03	4.83	5.66	3.62
0.1	0.43	2.85	4.05	4.50	4.20	2.86	6.41	6.58	4.75	4.86	3.64
0.2	0.48	3.49	4.56	4.76	3.93	3.52	5.73	6.76	4.98	4.01	3.48
0.3	0.65	2.56	4.20	5.78	3.67	3.71	4.95	7.25	4.66	3.30	3.06
0.4	0.94	2.47	4.42	5.67	3.29	4.34	5.81	7.43	4.31	3.36	2.97
0.5	0.82	2.98	3.68	4.54	3.80	4.09	5.75	5.23	4.24	3.46	2.66
0.6	1.20	2.41	3.34	4.51	4.03	4.50	5.86	4.79	3.81	3.17	2.50
0.7	1.12	2.57	3.07	4.95	3.51	5.10	6.51	4.78	3.83	3.16	2.33
0.8	1.05	3.02	2.84	4.54	3.95	5.54	6.29	4.55	3.72	2.76	2.18
0.9	1.07	3.07	2.53	5.02	4.02	5.13	5.83	4.76	3.18	2.51	2.05
1	1.18	2.56	3.01	5.14	4.37	6.40	5.06	4.48	2.91	2.47	1.76

Fig 15. *Having tested the independent impact of salience on the node's activation function (Effect 1) and impact on the connection weights (Effect 2), these two effects were then combined. Simulations were run varying the salience factor for both Effect 1 and Effect 2, to observe the effects of the average (A) and standard deviation (B) of the classification accuracy (ϕ) of the entire network, as well as the average (A) and standard deviation (B) of the classification confidence (τ) of each individual images. These results were combined as described in Equation 9 to calculate the optimal salience factors to ensure optimal performance of the SANN. The optimal salience factors to maximize the average values of classification accuracy (ϕ) and classification confidence (τ), and to minimize the standard deviations of classification accuracy (ϕ) and classification confidence (τ) were a salience factor of 0.4 for Effect 1, and 0.7 for Effect 2. The values with a darker border indicate the maximum values in each grid.*

improvement at the individual level of 5.60% is even more significant. This means that a strong salience signal attached to an individual image makes its recognition much more likely. This demonstrates the efficacy of the mechanism whereby a strong salience signal attached to an individual image strengthens the entire activation pattern associated with that specific image, which is the central message of this paper.

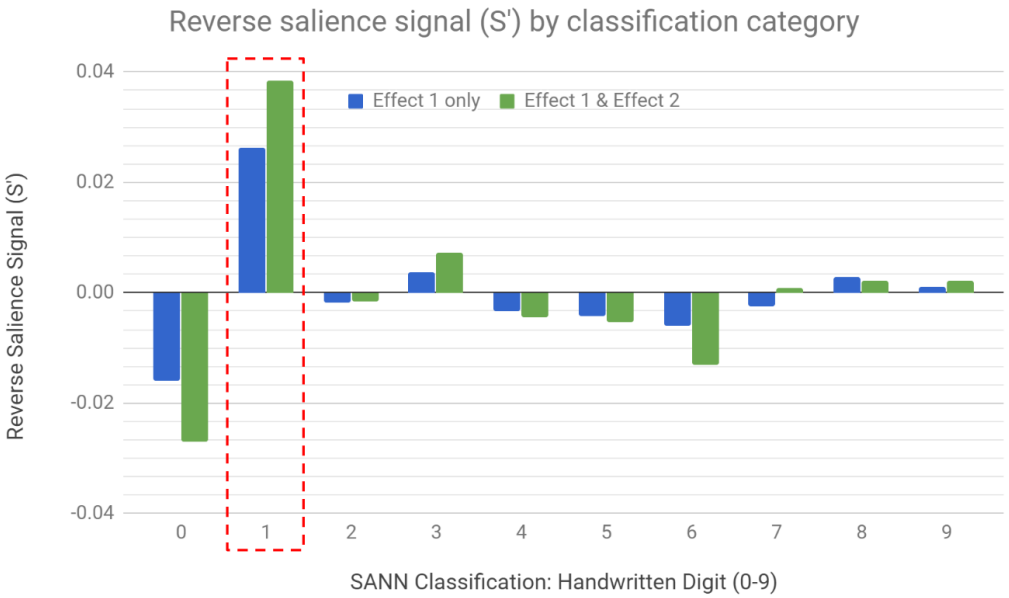

Fig 16. *After only a single iteration of salience training, the SANN produces a reverse salience signal (S'). This reverse salience signal is present after Effect 1 only, but is amplified with addition of Effect 2. The reverse salience signal (S') is the highest for the handwriting digits in the same classification group as the salience-tagged image. In this figure, we demonstrate the reverse salience signal for the scenario where Image #7 (classified as a handwriting digit #1) in the training set was tagged with salience during training.*

7. Conclusion

In this paper we have presented a ANN that incorporates some fundamental features and behaviors that we associate with salience and learning in the cortex, as indicated in Figure 7. We do not claim that the SANN accurately models the biological salience system, but rather that the SANN is an artificial neural network model with architecture inspired by the salience systems in the human brain, leading to important performance enhancement. Specifically, the SANN incorporates a global salience signal, modelled on the diffuse projections from the arousal system observed in the cortex.

A SANN accepts a global salience signal in addition to input and output values representing incoming sensory data. We have not modelled what leads to affect, or its effect on behaviour, but rather two distinct effects in the cortical network related to memory. Firstly, it affects each node by altering the activation function. Secondly, it affects the weights of every connection in the network. In both cases, the impact on the network is proportional to the activation level of the nodes in the network, which closely models the effect of the neuromodulators associated with the ascending systems on the cortex. This is a powerful mechanism underlying learning via neural plasticity, because it reinforces an entire neural activation pattern in one go.

After the neural network had undergone standard supervised training (on the MNIST dataset), the SANN model was trained with a single iteration of salience. Following this training the SANN model produced a reverse salience signal for similar images to the salience-affected input image during training. In addition, training with salience strengthened the weights of connection during the single iteration of salience training, which further increased the reverse salience signal and the classification accuracy for images similar to the salience-tagged image.

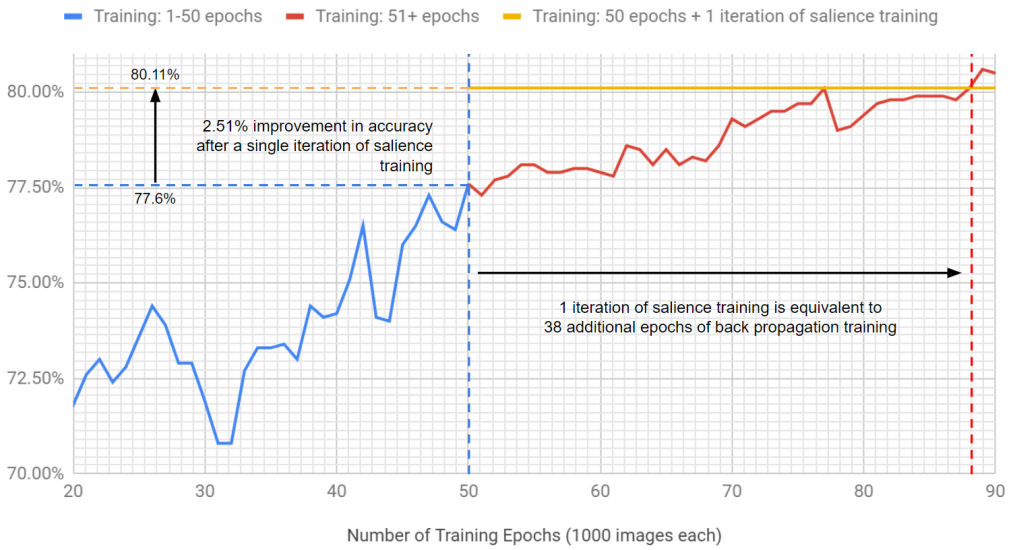

Fig 17. *Combining Effect 1 and Effect 2, the SANN model was able to demonstrate a 2.51% improvement in accuracy for the class after just a single iteration of training (a jump from 77.6% to 80.11% accuracy). To highlight the significance of this achievement, we measure how many additional training epochs using back-propagation the SANN would have required to achieve the same improvement. To achieve the same improvement with only back-propagation training, the same network would require an additional 38 epochs of training, with 1000 input images per epoch: a total additional exposure of 38,000 images.*

Notably, we demonstrated that the SANN can be trained with just a single training iteration. It was found that the reverse salience signal was proportional to the magnitude of the input salience signal. We associate this result with the expectation that single events of high salience can produce a learned response; to the best of our knowledge, this is the first demonstration of this phenomenon in an ANN.

7.1. Measured performance characteristics

We demonstrated the following characteristics:

(i) The salience signal affect *patterns* of activated nodes, not just individual nodes, and not all the nodes in the network equally: rather the entire current pattern of activation of nodes is either strengthened or weakened.

(ii) The network was able to be trained with the standard training data set, as well as an additional dimension of an incoming salience signal associated with each image.

(iii) The network produced a reverse salience signal (as an output) for a specific input and the associated pattern of activation.

(iv) Training with salience had a overall positive impact (2.6% improvement) on the classification accuracy of the network as a whole.

(v) Training with salience had a positive impact on the classification confidence of the specific image tagged with salience during training, raising the confidence from 99.85% to 100%.

Specific and generic images: Results The models and results published here link back into the distinction between generic and specific images (Figure 3). While salience affects the generic class (e.g. women), it has a heightened effect on the specific memory

(e.g. your mother) [75] [65]. This also links back to the differentiation between a class and an instance of a class in object oriented programming [11]. The SANN has shown that single iteration salience training can affect both the generic class of memories (e.g. the class: Digit 1), as well as the memory of a specific memory (e.g. Digit 1: Image 7 / 1000). The effect on the specific image that has been tagged with salience is significantly greater than the effect for the class to which it belongs. This is a key feature of the way affect (salience) affects cognitive functioning: we recognize specific contexts or individuals as safe and friendly or threatening and dangerous on the basis of previous experience, and act accordingly. Enhanced recognition and the reverse salience signal are both key to this dynamic.

7.2. Research Implications

The research explored here has many implications, especially at the computational, biological and psychological levels.

At a computation level, the modeling of ongoing regulations in neural networks will allow us to extract more information (in the form of salience) from a standard ANN, without significantly adding to the complexity of the ANN structure. Furthermore, exploring the effect of heightened affective moments on a neural network will enable the future training of neural networks with a single training iteration. A major point is that salience signals can strengthen an entire pattern of neural activation at one go. This is the key feature that makes this architecture so powerful.

At the biological level, one of the puzzling issues about the brain is why evolutionary selection lead to the very complex "wet" structure of chemical synapses, involving transformation of the signal from electrical to chemical form and then back again, as opposed to the much more direct gap junction (electrial) synapses, which are much faster [42] [9]. Our analysis, following on the work of Edelman [20] [24] [21], provides an answer: signal transduction via neuromodulators at synapses allows the non-local effects of the ascending systems to implement the very effective one-time learning process we have demonstrated in this paper, rather than the customary laborious and computationally intensive gradient descent algorithms like back-propagation involving thousands of training events in conventional ANNs. It also allows activation of the reverse salience signals, a key aspect of how these systems are related to Panksepp's primary emotional systems [61] [27], as elucidated in [28].

At a behavioural level, a SANN allows the modeling of the effect of affective states on memory (for example, adding an emotional tag to significant memories), because the ascending systems originate in the arousal system which is the seat of affective states. Thus a SANN potentially represents a key effect of emotions on cortical activity, which is known to be significant. It does not however model how affect alters perception and action, which is a different topic [17]; nor does it model what effects trigger salience signals.

7.3. Future Work

Future work to be done includes:

(i) Extending the SANN to include episodic memory, and not just semantic memory. This could be achieved by feeding the output back into the input to create a recurrent neural network. Salience signals would then apply to a sequence of events, rather than a single event.

(ii) Investigating alternative variations to the sigmoidal activation function at each node in response to the input salience signal, for example activation function gradient.

(iii) Extending the SANN to multi-dimensional salience signals [79].
(iv) Extending the hidden layer size, and the number of hidden layers.
(v) Extending the salience signal to a biologically-accurate spiking-neuron SNN.
(vi) Extension of deep learning reinforcement neural networks [30] to include the principles laid out here

This research will also eventually be incorporated in a model two of the authors of this paper (LR and GE) are proposing jointly with A. Mishra of cortico-thalamic interactions [67]. This models how perception works [31] via top-down predictive connections from the cortex to the thalamus [1]. Conjoining this with the results of this paper would open the way to modelling how affect alters perception.

7.4. Supporting Material

The source code for the SANN, as well as records of the tests conducted in this paper, are publicly available online [66]. For additional information, please contact the authors.

Acknowledgements

A preliminary version of this work was the subject of an MSc thesis of one of us (LR), supervised by JT and GE, which was put on the internet as a preprint a while ago [68]. The present version is so improved and updated that it is essentially a new paper, in particular because, apart from a greatly improved presentation of the logic of the project and its relation to brain structure and function, it has added the dynamics of a shift of weight in proportion to the salience signal and two further effects on the activation function, as well as extensive testing of how this works out in practice.

We are grateful to Mark Solms for very helpful comments.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We declare no conflicting interests.

[1] Henry J Alitto and W Martin Usrey. Corticothalamic feedback and sensory processing. *Current opinion in neurobiology*, 13(4):440–445, 2003.
[2] Michel A Arbib. The elements of brain theory and neural networks. *The handbook of brain theory and neural networks*, pages 2–23, 2003.
[3] Michael A Arbib and Jean-Marc Fellous. Emotions: from brain to robot. *Trends in cognitive sciences*, 8(12):554–561, 2004.
[4] Caren M Armstrong, Loren M DeVito, and Thomas A Cleland. One-trial associative odor learning in neonatal mice. *Chemical Senses*, 31(4):343–349, 2006.
[5] RC Atkinson and RM Shiffrin. Human memory: A proposed system and its control processes. in press on the psychology of learning and motivation: Advances in research and theory, 1968.
[6] Cornelia I Bargmann. Beyond the connectome: how neuromodulators shape neural circuits. *Bioessays*, 34(6):458–465, 2012.
[7] Aryel Beck, Lola Cañamero, and Kim A Bard. Towards an affect space for robots to display emotional body language. In *19th International symposium in robot and human interactive communication*, pages 464–469. IEEE, 2010.
[8] Beatrice Beebe and Frank M Lachmann. Representation and internalization in infancy: Three principles of salience. *Psychoanalytic psychology*, 11(2):127, 1994.
[9] Michael J Berridge. *Cell signalling biology*. Portland Press, 2007.
[10] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
[11] Grady Booch. *Object oriented analysis & design with application*. Pearson Education India, 2006.
[12] Kendra S Burbank. Mirrored stdp implements autoencoder learning in a network of spiking neurons. *PLoS computational biology*, 11(12):e1004566, 2015.
[13] Patricia S Churchland and Terrence J Sejnowski. Perspectives on cognitive neuroscience. *Science*, 242(4879):741–745, 1988.
[14] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[15] Antonio Damasio. *Self comes to mind: Constructing the conscious brain.* Vintage, 2012.
[16] Antonio Damasio and Gil B Carvalho. The nature of feelings: evolutionary and neurobiological origins. *Nature reviews neuroscience*, 14(2):143, 2013.
[17] Antonio R Damasio and Elwin Marg. Descartes' error: Emotion, reason, and the human brain. *Optometry and Vision Science*, 72(11):847–847, 1995.
[18] Jonathan Downar, David J Mikulis, and Karen D Davis. Neural correlates of the prolonged salience of painful stimulation. *Neuroimage*, 20(3):1540–1551, 2003.
[19] George H Dunteman. *Principal components analysis.* Number 69. Sage, 1989.
[20] Gerald M Edelman. *Neural Darwinism: The theory of neuronal group selection.* Basic books, 1987.
[21] Gerald M Edelman. *Wider than the sky: The phenomenal gift of consciousness.* Yale University Press, 2004.
[22] Gerald M Edelman. Learning in and from brain-based devices. *Science*, 318(5853):1103–1105, 2007.
[23] Gerald M Edelman, George N Reeke, W Einar Gall, Giulio Tononi, Douglas Williams, and Olaf Sporns. Synthetic neural modeling applied to a real-world artifact. *Proceedings of the National Academy of Sciences*, 89(15):7267–7271, 1992.
[24] Gerald M Edelman and Giulio Tononi. *Consciousness: How matter becomes imagination.* Penguin UK, 2013.
[25] George Ellis and Barbara Drossel How Downwards Causation Occurs in Digital Computers. philarchive.org (2019).
[26] George Ellis and Jonathan Koppel The Dynamical Emergence of Biology From Physics: Branching Causation via Biomolecules *Frontiers in physiology* **9**:1966
[27] George Ellis and Mark Solms. *Beyond Evolutionary Psychology.* Cambridge University Press, 2017.
[28] George FR Ellis and Judith A Toronchuk. Neural development affective and immune system influences. *Consciousness & Emotion: Agency, conscious choice, and selective perception*, 1:81, 2005.
[29] Elaine Fox. Perspectives from affective science on understanding the nature of emotion. *Brain and Neuroscience Advances*, 2:2398212818812628, 2018.
[30] Vincent Francois-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An Introduction to Deep Reinforcement Learning. *Foundations and Trends in Machine Learning*. 11 (3–4): 219–354, 2018.
[31] Chris Frith. *Making up the mind: How the brain creates our mental world.* John Wiley and Sons, 2013.
[32] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.
[33] Ann M Graybiel. Neurotransmitters and neuromodulators in the basal ganglia. *Trends in neurosciences*, 13(7):244–254, 1990.
[34] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.
[35] Stanley I Greenspan and Stuart Shanker. *The first idea: How symbols, language, and intelligence evolved from our primate ancestors to modern humans.* Da Capo Press, 2009.
[36] Jeff Hawkins and Sandra Blakeslee. *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines.* Macmillan, 2007.
[37] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.
[38] Patrick O Hoyer. Nmf pack. `https://github.com/aludnam/MATLAB/tree/master/nmfpack`, 2006. [Online; accessed 24-February-2019].
[39] Phil Husbands. Evolving robot behaviours with diffusing gas networks. In *European Workshop on Evolutionary Robotics*, pages 71–86. Springer, 1998.
[40] Ian Jolliffe. *Principal component analysis.* Springer, 2011.
[41] Ion Juvina, Othalia Larue, and Alexander Hough. Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research*, 48:4–24, 2018.
[42] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, and AJ Hudspeth. *Principles of neural science*, 4th Edition. McGraw-hill New York, 2000.
[43] David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network plasticity as bayesian inference. *PLoS computational biology*, 11(11):e1004485, 2015.
[44] Robert E Kingsley, Steven R Gable, et al. *Concise text of neuroscience.* Lippincott Williams and Wilkins, 1996.
[45] Jeffrey L Krichmar and Gerald M Edelman. Brain-based devices for the study of nervous systems and the development of intelligent machines. *Artificial Life*, 11(1-2):63–77, 2005.
[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[47] Y. LeCun, C. Cortes, C. Burges *The MNIST database of handwritten digits* http://yann.lecun.com/exdb/mnist/

[48] Joseph LeDoux. The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology*, 23(4-5):727–738, 2003.

[49] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[50] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[51] Evelyn Lee-Teng and S Murray Sherman. Memory consolidation of one-trial learning in chicks. *Proceedings of the National Academy of Sciences of the United States of America*, 56(3):926, 1966.

[52] Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 48:39–55, 2018.

[53] Wolfgang Maass. Computation with spiking neurons. *The handbook of brain theory and neural networks*, pages 1080–1083, 2003.

[54] Paul D MacLean. *The triune brain in evolution: Role in paleocerebral functions*. Springer Science & Business Media, 1990.

[55] M-Marsel Mesulam. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of neurology*, 28(5):597–613, 1990.

[56] M-Marsel Mesulam. *Principles of behavioral and cognitive neurology*. Oxford University Press, 2000.

[57] P Read Montague, Steven E Hyman, and Jonathan D Cohen. Computational roles for dopamine in behavioural control. *Nature*, 431(7010):760, 2004.

[58] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[59] Bernhard Nessler, Michael Pfeiffer, Lars Buesing, and Wolfgang Maass. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS computational biology*, 9(4):e1003037, 2013.

[60] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press USA, 2015. http://neuralnetworksanddeeplearning.com/index.html.

[61] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004.

[62] Jaak Panksepp and Lucy Biven. *The archaeology of mind: Neuroevolutionary origins of human emotions*. WW Norton & Company, 2012.

[63] James W Papez. A proposed mechanism of emotion. *Archives of Neurology & Psychiatry*, 38(4):725–743, 1937.

[64] JingPei *et al*. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572** 106, 2019.

[65] Daniel Reisberg. *The Oxford handbook of cognitive psychology*. Oxford University Press, 2013.

[66] Leendert Remmelzwaal. Salience affected neural network - python code. `https://bitbucket.org/leenremm/sann_python/`, 2019.

[67] Leendert Remmelzwaal, Amit Mishra, and George FR Ellis. Corticothalamic neural network model. *Submitted*, 2019.

[68] Leendert Remmelzwaal, Jonathan Tapson, and George FR Ellis. The integration of diffusely-distributed salience signals into a neural network. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.9331&rep=rep1&type=pdf`, 2010; https://arxiv.org/abs/1001.3246.

[69] Harald Reuter. Calcium channel modulation by neurotransmitters, enzymes and drugs. *Nature*, 301(5901):569, 1983.

[70] Irvin Rock and Walter Heimer. Further evidence of one-trial associative learning. *The American Journal of Psychology*, 72(1):1–16, 1959.

[71] Neil Schemenauer. A back-propagation neural network written in python. `http://arctrix.com/nas/python/bpnn.py`. [Online; accessed 24-February-2019].

[72] Wolfram Schultz. Getting formal with dopamine and reward. *Neuron*, 36(2):241–263, 2002.

[73] S Murray Sherman and RW Guillery. Exploring the thalamus and its role in cortical function. mit press. *Cambridge, MA*, 2009.

[74] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3, 2003.

[75] Julia Spaniol, David J Madden, and Andreas Voss. A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1):101, 2006.

[76] Olaf Sporns, Nikolaus Almássy, and Gerald M Edelman. Plasticity in value systems and its role in adaptive behavior. *Adaptive Behavior*, 8(2):129–148, 2000.

[77] Andrew S Tanenbaum (2006) *Structured Computer Organisation* (Prentice Hall, Englewood

Cliffs), 5th Edition.

[78] MohamadAli Torkamani, Phillip Wallis, Shiv Shankar, and Amirmohammad Rooshenas. Learning compact neural networks using ordinary differential equations as activation functions. *arXiv preprint arXiv:1905.07685*, 2019.

[79] Judith A Toronchuk and George FR Ellis Affective neuronal selection: the nature of the primordial emotion systems. *Frontiers in psychology*, 3:589, 2013.

[80] Endel Tulving. Precis of elements of episodic memory. *Behavioral and Brain Sciences*, 7(2):223–238, 1984.

[81] Pascale Waelti, Anthony Dickinson, and Wolfram Schultz. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412(6842):43, 2001.