TÉCNICO LISBOA

Application of Artificial Neural Networks for modelling cognitive dimensions

Pedro Henrique Carvalho de Paula Ferreira da Costa

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Robert Leech

Prof. Rita Homem de Gouveia Constanzo Nunes

Examination Committee

Chairperson:	Prof. José Paulo Sequeira Farinha
Supervisor:	Prof. Rita de Gouveia Constanzo Nunes
Member of the Committee:	Dr. Ana Catarina Fidalgo Barata

November 2018

Acknowledgments

None of this work would have been possible without the guidance and tutoring from my supervisors in England. I am truly grateful to Rob, Romy and Sebastian for the hours invested in this project and the vast knowledge that they untiringly passed onto me. I'd also like to express my deepest gratitude and esteem to Rob for trusting me when I approached him to join his lab. To my home supervisor, Professora Rita Nunes, I would like to thank for the openness, guidance and counseling provided.

Going abroad is always challenging, but I was lucky enough to join C3NL where I found a group of outstanding and friendly people, from the Post-Docs to my friends from the MSc and MRes in Neurosciences, who always made me feel at home. To Rodrigo and Carolina, I must say that there are no words that can express my gratitude for your generosity and for letting me disrupt your life. I will always be in debt to you.

Gratitude is all I have for all who accompanied me through this journey. Thank you, Margarida, for invariably being an ear and a voice which guided me through good and bad times. Thank you to the very best FCTenses for showing me that an important part of University is companionship. Thank you, every one of my friends I bring from before University, whether it be from hockey or from school.

Years from now I hope to look back and see that all the people who were paramount through this stage of my life, will still be an important part of it, whether they are on the other side of the world or they are under the same roof as me.

Ultimately but most importantly, I cannot be thankful enough to my parents for dedicating such a large part of your life to make sure I have everything I need. Obrigado.

Resumo

A relação cérebro-cognição mantém-se desconhecida apesar de décadas de investigação em neuroimagiologia funcional. Uma das limitações prende-se com o facto dos processos cognitivos que tentamos associar com a actividade cerebral basearem-se em conceitos psicológicos apresentados previamente à introdução de técnicas de imagem médica. Neste projeto seguiu-se uma abordagem diferente, tirando vantagem dos desenvolvimentos com redes neurais artificiais (RNA) para aprender mecanismos partilhados. Pretendeu-se avaliar o mecanismo de execução de múltiplas provas cognitivas sem pré-definir os domínios cognitivos envolvidos. Assim sendo, um modelo de redes neurais recurrentes que resolve seis testes cognitivos foi desenvolvido com uma precisão de 93% abrangendo processos de reacção, inibição e memória de trabalho. Tendo em conta o modelo obtido, testou-se se o mecanismo providencia uma boa explicação dos padrões de actividade nas regiões cerebrais associadas previamente aos processos cognitivos. Embora as comparações entre os mecanismos biológicos e artificiais expressem poucas semelhanças, os mecanismos utilizados pelo modelo representam um sistema eficaz para interpretar cada prova. É aparente, por análise da interpretação do *input* do modelo, que os conceitos de prova e momento de reacção são importantes para obter a solução correcta. Através do estudo dos nós para cada prova, detectou-se que um dos nós (unidade 28) apresentou um comportamento semelhante ao processo de controlo de inibição em sistemas biológicos. Este trabalho pretendeu apresentar uma nova perspectiva na análise da relação cérebro-cognição, sugerindo um potencial nível de paralelismo entre o modelo neural desenvolvido e os processos biológicos. Esta perspetiva pode contribuir para uma clara interpretação dos processos cognitivos.

Palavras-chave

Percursos Neuronais; Ontologias; Aprendizagem Automática; Processos Cognitivos

Abstract

The relationship between the brain and cognition remains unclear, despite several decades of functional neuroimaging research. One limitation is that the cognitive processes we attempt to match to brain activity are taken from psychological constructs derived in a somewhat ad hoc manner. This project took a different approach, taking advantage of developments with artificial neural networks (ANNs) to learn shared mechanisms. The purpose was to evaluate the execution mechanisms between multiple cognitive tasks without relying on predefined cognitive domains. Therefore, a Recurrent Neural Network was developed to perform six cognitive tasks with an accuracy of 93%, that tapped on the processes of reaction, inhibition and working memory. With regard to the obtained model, it was tested if the mechanism provides a good explanation for the activation patterns in the brain regions previously associated to the cognitive processes. Although comparisons between the model's activations and real brain data bared little similarities, the model's mechanisms expressed an effective system of interpreting each task. It was clear, by analysis of the model's interpretation of the input dataset, that the concept of task and moment of reaction were important factors for the correct solution. From the study of node variation along trials, one stood out (unit 28) by displaying a behaviour similar to inhibition control in biological systems. This work intended to provide novel insights into both brain and cognition, suggesting a potential parallelism between the artificial model and the biological processes. This perspective can contribute to a clearer interpretation of the cognitive processes.

Keywords

Neural Pathways; Ontologies; Machine Learning; Cognitive Processes

Table of Contents

List of Figures

List of Tables

I Introduction

For millennia scholars have wondered what makes the remarkable homo sapiens sapiens be able to apply logic, speak, think and think about thinking, traits that seem absent or greatly reduced in other animals. These questions are at the core of humanity's insatiable curiosity and their complexity has resulted in different answers throughout the ages. In Ancient Egypt, the heart was thought to be the seat of intelligence [1] and in the 16th century, recognized philosophers such as Descartes defended the existence of an ethereal organ that holds consciousness and self-awareness: the mind. Today it is well established that consciousness is not an immaterial construct but a process of the brain. The dissemination of the scientific methods and the evolution of technology has allowed different fields such as psychology and neuroscience to study some of the mysteries of the brain but with limited success on the overall mechanisms that result in intelligence. Cognitive neuroscience, a research field at the intersection of the two, concerns the study of the neural processes and mechanisms that give rise to thought and understanding.

During everyday life, humans perform complex behaviours, such as navigating through a city, memorising names or planning the dinner for the family. This behaviour generally depends on multiple overlapping faculties and usually allows for several solution strategies. As an example, the simple task of crossing the street requires motor dexterity from the legs, requires attention mechanisms to observe when the pedestrian light turns green and ultimately requires memory skills on interpretation that the green light means it is safe to move.

To overcome this challenge in the study of human cognition, over the last 150 years the scientific discipline psychology has developed cognitive tasks aimed at isolating a given cognitive ability (e.g., sustained attention, working memory, response inhibition). With these efforts a broad cognitive ontology was developed and is still used in cognitive neuroscience to date. It is a formal representation of the types of entities in the cognitive domain and the relationship between each other. However, this cognitive ontology has been arrived at in a relatively ad hoc way and does not take into consideration the functional organization of the human brain. This hinders the process of understanding the neural mechanisms and prevents a full mapping between functional and anatomical structures [2], [3].

With the recent surge in applications of artificial neural networks to solve cognitive tasks, there is an increasing amount of literature that highlights similarities in biological and artificial structures. Machine Learning might prove to be the conductive thread between current ontologies (i.e. based on cognitive paradigms) and a more structural (i.e. focus on networks activations) approach to cognitive ontologies by effectively mapping one to the other [4]. Contrarily to brain systems, artificial neural networks allow full access to its inner calculations. Even though the interpretation of how each component contributes to the general outcome remains challenging, the understanding of its mechanisms might provide an alternative way to formulate a cognitive ontology and shed light on the relationship between cognitive paradigms and the structures that originate function.

Based on recent developments from deep learning (e.g., Google's DeepMind [2], [5]), this work has attempted to use artificial neural network approaches to learn the complex relationships in cognitive tasks. Such artificial neural networks have previously been shown to be able to simultaneously capture

multiple cognitive tasks in a single model [6]. This allowed modelling the compositionality and clustering of task representations [2], [7], which allowed us to understand from a bottom-up, purely mechanism-driven way how different tasks group together. These prior studies indicate that the proposed work was plausible and forms the starting point for the translation into more complex cognitive tasks thought to be related to frontoparietal networks.

This project set out to build and train a neural network to solve six classical cognitive tasks in order to better understand the mechanisms intrinsically created by the algorithm for addressing each cognitive process. This project was designed as a first stepping-stone towards a larger goal of creating a mechanistically based cognitive ontology, using neural networks as the medium for an effective mapping between cognitive paradigms and its representations without relying on predefined cognitive processes. This approach has the potential for providing a radically different understanding of the relationship between brain networks and cognition.

I.1 Cognitive Neuroscience

Cognitive neuroscience is an academic field concerned with explaining the neurological mechanisms involved in cognition (i.e. thinking and understanding). It is intrinsically multidisciplinary, emerging from the interception of the fields of neuroscience and psychology.

This field emerged in the 1950s as the advances in neuroscience and the development of the computational sciences changed the focus in the study of cognition from its behaviouristic approach (i.e. concerned only with the observable stimulus-response behaviours) towards the issue of localizability of mental processes. The field aimed at mapping each cognitive process to a specific region of the brain. This was done by identifying task-dependent signal changes associated to brain activity in order to assess if a given brain area is specialized in a particular function [8]. However this type of univariate analysis did not consider how possible communications between neural populations would contribute to cognition [9].

Advances in the field caused a shift from this univariate analysis towards connectivity studies and multivariate approaches. This breakthrough, allied to novel brain imaging techniques, unveiled that cognitive functions do not map to specific brain regions, but that brain regions participate in many functions which can also be carried out by many regions [10]. This resulted in a shift of focus in cognitive studies from searching for specialized brain regions towards identifying networks of regions in which functional interactions involved diverse cognitive domains.

I.1.1 The Role of Cognitive Tasks

The study of cognition has long been faced with the challenge of understanding the cognitive faculties required to solve a given task. This poses a complex problem because generally tasks depend on multiple abilities and admit multiple solution strategies. In order to avoid these ambiguities, cognitive psychology developed rigorously controlled laboratory-based experiments that addressed one specific

cognitive ability. These tests, commonly referred as cognitive tasks, allowed for systematic and consistent research throughout the field.

Because these tasks were developed before the advent of neuroimaging techniques, they were mostly based on intuition and behavioural response. Therefore, when using state-of-the-art techniques, the same tasks failed to map consistently onto brain structures and concepts that were deemed as elemental paradigms of cognition were proven to recruit different networks for different tasks. As an example, Kane et al. proved that two standard Working Memory tasks presented low correlation in behavioural results, which suggests that they do not reflect one single construct [11]. Also, Adam et al. showed that intelligence, thought to be one single construct, maps onto different networks in the brain both structurally and behaviourally [12].

It becomes apparent that classical cognitive ontologies currently used in the field, are not based or validated on their neural states and need to be evaluated and reformed if necessary to keep pace with the evolution of the field [13].

I.1.2 Limitations of current cognitive ontology

The simplest way to map the mental processes onto the brain would be to assume a one-to-one mapping in which each behaviour would fit an independent brain area. Even though this is the case in some sensory skills (e.g. the visual cortex is linked to visual perception), there are brain regions that participate in many functions, named pluripotent (e.g. the dorsal medial prefrontal cortex is recruited for a number of cognitive operations and for emotional processing) and there are different regions that are active for a single function, named degenerate (e.g. both parietal and frontal regions are activated in attentional processes) [10]. A single task fails to conceptualize the function of a network which is only derived by studying the interception of different cognitive processes. Conventional brain scanning techniques are slow and expensive processes and the recording of activity can take several hours in order to obtain data consistency for a single task. This results in a single experiment tackling a single task or a handful of tasks, which cannot be used to understanding by itself the functional role of a network.

The restrictions mentioned above currently limit the obtainment of a clear neurobiologically-derived ontology that would draw consensus in the field. Nonetheless, new methods of brain-data meta-analysis, that take advantage of the growing literature in neuroimaging, show promising results that might advance the field's understanding of cognition.

I.2 Meta-analysis of neuroimaging literature

Meta-analysis in neuroimaging consists of synthesizing findings across different studies, different laboratories and different task variants. This analysis overcomes most of fMRI studies setbacks, by conclusively defining which brain regions or networks are consistently associated with particular cognitive faculties, an achievement impossible for single studies as they are yet to be validated by replications of their findings outlined above. Moreover, meta-analysis is also able to identify specificity

of brain regions or networks for individual functions. By taking advantage of these techniques Yarkoni et al. created Neurosynth, an open-sourced meta-analysis of functional brain imaging [14].

I.2.1 Neurosynth

Neurosynth [1] is an online platform for large-scale analysis and synthesis of functional Magnetic Resonance Imaging (fMRI) that allows the identification of regional brain activation patterns based on specific terms. The Neurosynth project takes advantage of the rapid growth of published neuroimaging studies using non-invasive techniques, predominantly fMRI. The Neurosynth database covers 11,406 neuroimaging studies to represent 413,429 brain activations. This large data trove creates difficulties for synthesizing findings. Therefore, Neurosynth uses automatic data science techniques to perform meta-analysis on its database to relate brain activity to verbal descriptions of cognitive, behavioural processes and other concepts. These are detailed below:

First, Neurosynth uses text-mining techniques to identify terms of interest in relevant studies based on its frequency in a given article. Second, the activation coordinates, extracted from group level brain images reported in the paper, are collected. This allows matching terms with the pattern of brain activations in an automated manner, which is an advantage over manual meta-analyses. Third, a meta-analysis is run combining hundreds of psychological terms, to produce a quantitative relationship between the brain regions and cognition. Finally, Naïve Bayes classification was applied to match activation maps with their psychological concepts [15]. The database currently has meta-analyses of 3,107 terms. Each term is associated with two 3-dimensional brain images, one for brain activation assessed by forward inference and another by reverse inference. While the forward inference displays brain regions constantly active across studies, reverse inference displays active brain regions that are only specific to that given term [14].

Despite meta-analysis being the most powerful approach to date to tackle research questions about broad cognitive domains, it is not without limitations as it fails to extract finer-grained cognitive states [15] and is prone to biases in the literature regarding the report of results (e.g. file-drawer effect [16], selection of contrasts and inconsistent labelling of brain areas and cognitive states [13], [17]). Meta-analysis also fails to provide mechanistic insights to the process of cognition.

These issues might be surpassed by incorporating different types of brain data acquisition and by considering the integration of different techniques. This project took special attention to how artificial neural networks might provide a mechanistic explanation for different cognitive processes.

I.3 Cognitive Ontologies

An ontology is a formal and controlled collection of concepts that facilitate description of knowledge based on its meaning and how it relates to others. This allows not only interoperability between agents

[1] www.neurosynth.org

across different domains, which is a requirement in science research, but also for the information to be machine-readable [18]. An effective ontology should integrate new knowledge effortlessly without disrupting concepts and their unique identifiers. This is especially challenging in the biological sciences mainly due to the complexity of concepts' relationships [19].

The term "cognitive ontology" was introduced into cognitive neuroscience by Price and Friston following on the success of ontologies in genetics [20]. Gene Ontology (GO) combined the effort of different ontology developers to build a terminology of around 16,000 terms and became the standard in genetic research [21]. In cognitive neuroscience there is not a standard preferred ontology, a matter of concern across the field [13], [22], [23]. Different domains in the study of the brain have adopted different categories that fail to map onto each other, such as the Neuroscience Information Framework for neuroscientific ontology [24] and the Neural Electromagnetic Ontologies which address an EEG and MEG ontology [25]. Moreover, there is no unanimity on what counts as cognitive domain nor at what level it is best analysed [22]. As an example of the necessity for a standard ontology, a behavioural task can be called by its name (e.g. Stroop task) or by function (processing speed, selective attention); however, the stimuli presented, the subject's response and the instructions can vary extensively for the same task and with it, the active brain regions vary too. This stems from a larger issue previously addressed that the tasks used in experiments were designed prior to the existence of brain acquisition techniques and therefore, consist of multiple cognitive processes that map onto multiple brain networks.

Recent advances in neuroimaging and dissemination of fMRI are leading to a surge in brain data that promises to unveil how mental processes shizzle maps onto the brain [13]. Taking this trend into account, the most prominent ontologies in cognitive neuroscience have taken a data-driven approach to their concepts' structure by running meta-analytical studies [3], [26].

There are two main distinct approaches to cognitive ontologies, a top-down approach, where a cognitive set is distinguished based on a behaviour dissociation [27], [28], and a structural bottom-up approach, where the anatomical set is identified on the regions of activations on a given task [13], [29].

The first methodology involves the standardization of cognitive paradigms, with the labels set out by the researchers. Turner and Laird presented in 2012 the Cognitive Paradigm Ontology (CogPO), a domain ontology for application in the functional neuroimaging community [28]. This cognitive ontology allowed for the integration of two brain imaging databases, the Functional Imaging Biomedical Informatics Research Network (FIBIRN) and the BrainMap database and was built in compliance with the Open Biomedical Ontology (OBO) Foundry, a set of principles designed to foster interoperability of ontologies within the OBO framework. Its design focuses on categorization of practical aspects of any cognitive tasks: the stimulus, the requested instructions and the subject's response. Currently CogPO is composed by 84 classes of paradigms.

In 2011, Poldrack et al. published the Cognitive Atlas, a cognitive ontology project aimed at characterizing the current state on cognitive sciences with manual input from researchers across the globe [30]. It relies on taxonomy on both concepts and tasks to define the properties of cognitive processes. It states as its guiding principle the design of a distinction between mental tasks and mental processes, as these terms do not present specificity on each other. The Cognitive Atlas presents itself

as a solution for a need of adopting formal knowledge that provides a systematic approach to cognitive theories and how it relates to empirical data.

In contrast, the structural approach defends that cognitive ontologies should not define a taxonomy focused on cognitive paradigms but on predicting them [29]. It critiques the psychology field's randomized and tightly controlled experiments as a means of providing theories on psychological mechanisms but failing to relate them to the underlying structure. To bypass issues of context-dependent multi-functional regions of the brain, this approach builds on data-driven methods, of large scale imaging meta-analysis to create an ontology that effectively predicts real-world behaviour [3]. It sets a bottom-up approach relating active regions in the brain to its predicted human reaction.

As an example of a structural approach, Yarkoni et al. [15] took advantage of the large-scale database of neuroimaging in Neurosynth and through meta-analysis of the activation coordinates generated a consistent mapping of brain structure associated with the terms present in these studies abstracts. By doing so, a term database was generated that reliably associated cognitive taxonomy to the networks in the brain associated to them. Despite this project's success, it fails to provide mechanistic accounts of how networks of brain region generate the cognitive processes.

Another example is Yeo et al.'s work [31], where 1000 subjects fMRI data on resting-state functional connectivity was analysed and through a clustering approach, networks with functionally coupled regions were mapped across the cerebral cortex. The study was able to identify regions confined to sensory and motor cortices and networks of association regions. The resulting parcellations were separated into a coarse solution, dividing the cortex into 7 functionally connected networks, and a finer solution which identified 17 networks.

In addition to these two prominent approaches, there a third approach aiming to create a cognitive taxonomy relevant to both brain systems and psychological functions based on computational mechanisms [22]. Weiskopf presented the concept for a network-based model grounded on a taxonomy linked by several levels of theoretical significant properties through network categories [4]. This mechanistic approach, as Weiskopf regards, is not without complications as it is unclear how some psychological models can be integrated on neural models. He defends that a new set of analytical tools might prove useful to solve the cases of concept entanglement between fields. With the development of artificial neural networks, it is valid to ponder whether this new field might hold the key to combine the different approaches to cognitive ontologies using a mechanistic approach. Table I-1 presents a summary of current cognitive ontologies, its baseline approach and the authors that support them.

Table I-1 - Overview of the different approaches to cognitive ontologies and supporting authors.

Approach	Authors	Taxonomies
Top-down	Turner et al.[28], Bilder et al.[27], Price et al[32].	Cognitive Paradigm Ontology (CogPO)[2]
	Poldrack et al.[30]	Cognitive Atlas[3]
Bottom-up	Eisenberg et al.[3], Yeo et al. [31], Yarkoni et al.[14], Anderson et al.[33].	Neurosynth terms Yeo's Cortical Parcellations[4]
Mechanistic	Weiskopf[4]	-

I.4 Artificial Neural Networks

The nervous system is composed of neurons, unique cells that are electrically excitable and responsible for transmitting information in the biological system through electrical or chemical signals that transverse the neural circuits. Neurons are interconnected to each other through synapses that facilitate signal transmission between cells. If an input signal overcomes a certain threshold, an imbalance of ions will result in an electrical potential that is transmitted through the axon onto its next connection. The mechanism can be simplified into a logical gate determining if the neuron fired or not. Synapses that result in neurons firing will strengthen as those that do not will weaken. This process is believed to be the foundation for learning and memory in biological systems [34].

In 1943, McCulloch and Pitts created the predecessor of the artificial neural network, a computational model based on a linear classifier that would output a binary result dependent on the unit's surpassing a given threshold [35]. Modern Artificial Neural Networks (ANN) still follow the same paradigm but using a more sophisticated non-linear activation functions, putatively inspired by neurobiological features, such as the interconnection between neurons, the learning-dependant variance of synaptic weight or the firing upon reaching a given threshold. Together, these networks constitute a computational model that, through mathematical operations, allows information to be channelled towards the general goal of classifying a given input. More concretely, ANN consist of nested

[2] http://www.cogpo.org
[3] https://www.cognitiveatlas.org
[4] https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation_Yeo2011

composite functions with trainable parameters that vary to minimize a certain cost function, which relates the model's output to its expected result [36]. This is possible through training procedures such as back-propagation [37], allied to gradient descent mechanisms [38] which consist of using the error variance for a given output to vary the function's parameters towards a lower error rate.

Classical (feed-forward) deep neural networks are not able to solve temporal dependencies, as the model assumes independence between input iterations. This proves to be a major shortcoming as most cognitive tasks require the processing and retention of sequential information. Recurrent Neural Networks (RNNs) are a class of ANN that exhibit dynamic temporal behaviour for time sequences by concatenating the previous output to the input of a given unit, effectively creating a loop of information. Different varieties of RNNs have been developed in order to allow for longer temporal dependencies and avoid computational difficulties such as the exponential changes in the gradient across time (i.e. the vanishing or exploding gradient problem). One well-known solution called Long Short-Term Memory (LSTM) unit was developed by Hochreiter and Shmidhuber [39] and consists of a gating mechanism applied to the RNN in order to facilitate the passing of information through time. Each unit saves a cell state transmitted across iterations that is updatable and accessible when convenient for the model to deliver a correct output. This cell state functions as the flexible working memory of the network, avoiding long dependency problems.

I.5 Networks simulating behaviour

Recent advances in machine learning have led neural networks to successfully execute highly complex tasks, such as playing Atari [40] or winning the human Go championship [41], which involve a number of cognitive skills. Due to their level of complexity, these models remain hard to analyse. Simpler tasks designed in the psychology field to isolate elements of cognitions allow for a clearer comparison between the human and the machine performer [2].

Despite substantial advances in neuroscience, the mechanisms employed by the brain to perform these tasks are not yet understood. While the importance of pre-frontal cortex and working memory on context-dependent cognitive task performance is established, the computation underlying it at a more elemental level is still largely unknown [42]. In contrast, neural networks allow complete access to its computation and the neural circuit, allowing the model's behaviour to be understood and manipulated [43].

There are, nonetheless, clear differences between the brain and the mathematical model that make the latter biological implausible, such as its learning mechanism (i.e. through backpropagation of the error signal), the absence of neuromodulators or excitatory/inhibitory balance [44], and yet most neuroscientific models that intend on explaining neural computation in a top-down framework consider the simplistic notion that the brain optimizes a single cost function for a single computational architecture [45], [46], analogous to the artificial network models. Moreover, different neural networks optimized for distinct tasks, such as object recognition [47], decision-making, working memory [48], timing [49] and motor tasks [50], display different features of neural activity or behaviour found in the biological model.

Research on more biologically plausible models have been mounting and the most promising results are subsequently analysed.

Song et al. trained an RNN model using reinforcement learning to perform on classical cognitive tasks [43]. The network was composed by two distinct subnetworks, a decision network and a value network. The results exhibit features of neural activity of behaving animals with particular consideration for the similarity of the electrophysiological response on the orbitofrontal cortex (OFC), a structure implicated in the representation of reward-related signals. As with the animal counterpart, a longer stimulus results in improved performance. The tasks ranged from perceptual decision-making, working memory and multisensorial tasks.

Yang et al. built an RNN which learned, using supervised learning, to perform 20 different tasks that tackled psychological concepts as inhibitory control, decision-making, reaction-time, context-dependent analysis and working memory tasks [2]. The trained recurrent units displayed compositionality of task representation (i.e. the combination of rules of two different tasks resulted in the model performing a third unique task), a characteristic paramount for cognitive flexibility. In this sense, this study was able to correctly perform tasks by combining other tasks' instructions.

Cueva et al. applied a recurrent neural network to perform spatial localization in a limited two-dimensional area with square, triangular and hexagonal shapes [51]. Some RNN units presented a grid-like spatial response, while others presented a border spatial response, which is consistent with the neural representation on cells on the Entorhinal Cortex of rodents, appropriately named grid cells. The mechanism behind the biological network and the functional utility of the grid representation are still a mystery in neuroscience, but these results imply that they provide an efficient solution for navigation tasks and that recurrent neural networks might prove useful to understand certain neural mechanisms.

Carnevale et al. studied how both monkeys and neural networks solved a detection task with variable stimulation times [48]. The stimulus, when present, was provided in a two second time window which resulted in both electrophysiological analysis of the prefrontal cortex and analysis of the model's activations displaying temporal expectation, by drawing closer to the reaction threshold when the stimulus is expected. It is unclear how neural mechanisms take advantage of temporal structures at intermediate stages of sensorimotor events but by modelling an ANN on the same task, Carnevale et al. were able to unveil the dynamic mechanism of this neural structure.

Hong applied SpikeProp to implement three classical machine-learning problems, MNIST digits recognition (i.e. a famous database of hand-written numbers for visual identification), spatial coordinate transformation, and motor sequence generation [52]. SpikeProp is a Spike Neural Network (SNN) with a soft threshold model applied to allow for backpropagation. SNNs are regularly used by neuroscientists to simulate biological models as the model's spiking activity displays a similar output to the neuron than the regular artificial neural unit does. The trained model expressed features from the biological model, such as selective activity and excitatory-inhibitory balance.

In 2018 Google DeepMind, an artificial intelligence company, open-sourced Psychlab, a simulated psychology laboratory inside a first-person three-dimensional environment [5]. Psychlab allows for a common platform between man and machine on solving classical laboratory psychological

experiments in a controlled and consistent protocol. It allows for both the mechanical and the biological model to receive the same inputs while executing the tasks that have been validated by years of research in the psychology field. By tackling known classical cognitive tasks, the platform allows for a clearer analysis of the cognitive and perceptual faculties at play.

Research in ANNs applied to neural simulations is gaining traction in recent years and the amount of information on the subject is growing by the day. Here the analysis was restricted to the bigger breakthroughs and to the results most relevant to this project. There is a clear trend in applying RNNs for solving classic cognitive tasks as their time-dependency properties are desirable for simulating how biological systems solve the same assignments. Table I-2 presents a brief overview of the networks applied and the most significant results that were reported in each study described in this section.

Table I-2 - Overview of state-of-the-art literature regarding the application of neural networks to solve cognitive tasks.

Author	Architecture	Tasks	Results and Conclusions
Song et al. [43]	RNN trained with Reinforcement Learning.	Perceptual decision-making, context-dependent integration, multisensory integration, and parametric working memory tasks.	Similar response between the artificial network model and the electrophysiological response of the OFC.
Yang et al. [2]	RNN trained with supervised learning.	20 tasks trained simultaneously using concepts of inhibition control, working memory, decision-making and reaction-time.	Model's units displayed compositionality of task representation, a characteristic paramount for cognitive flexibility.
Cueva et al. [51]	RNN trained with Reinforcement Learning.	Spatial navigation tasks.	Some units presented a grid-like spatial response, similar to the representation found in the Entorhinal Cortex of rodents.
Carnevale et al. [48]	RNN trained with Supervised Learning.	Detection tasks with variable stimulation times.	Both the model and the primate display temporal expectation mechanisms for faster reactions.

Author	Architecture	Tasks	Results and Conclusions
Hong [52]	Spiking Neural Networks trained with supervised learning.	MNIST digits recognition, spatial coordinate transformation and motor sequence generation.	Without being constrained the model presented selective activity, excitatory-inhibitory balance and weak pair-wise correlation similarly to neurons.
Leibo et al. – PsychLab [5]	Deep Reinforcement Learning agents installed on a virtual environment.	Psychlab has several classical psychological experiments implemented and its API allows for the installation of others.	As a proof of concept a study was conducted that concluded that the agent would learn more quickly the larger the target stimuli, similarly to biological systems.

I.6 Overview

Recent neuroscience developments allied to the proliferation of neuroimaging literature have directed the field to a meta-analytical approach of cognition, in which neuroimaging databases (e.g. Neurosynth) take advantage of their vast resources to allow for a combinatorial analysis of multiple studies in order to define activation consistency across studies. This analysis is not straight-forward due to the heterogeneity between fields that study cognition (such as cognitive neuroscience and psychology), which results in a number of non-convertible taxonomies and hinders the meta-analytical process.

Recently there have been efforts in integrating the different fields in the science of cognition by mapping cognitive paradigms on to their anatomical constructs, a task that has been proven challenging mainly due to variability in both tasks' nomenclature and brain network activations. With the goal of standardizing both approaches, a number of cognitive ontologies were created. These differed in their baseline approach from a top-down concept (i.e. fixing on the functional mechanisms) to a bottom-up concept (i.e. fixing on the structural dependencies). There is, however, a possible third approach that might combine the other two. By using a mechanistical analysis it intends on mapping functional regions to their structural concepts. Recent developments in ANN might prove helpful to achieve this combination by shedding light on the mechanistic origin of the cognitive paradigms.

Moreover, despite the braodly biologically implausible learning mechanism, ANNs have shown promising parallels with their neurobiological counterparts. Research in the application of machine learning to the understanding of the brain networks is gaining pace and these breakthroughs might assist in providing a clearer understanding of the brain's mechanisms and a deeper association between behaviour and structure.

I.7 Objectives

This projects' primary aim was to create a working model that effectively solves a set of six cognitive tasks adapted from psychology research. To do so, a dataset that automatically generated the desired tasks and registered each trial variability was coded. The tasks address the following cognitive paradigms: working memory, reaction-time and inhibitory control.

A secondary aim concerns trying to interpret how the network models the inputs to obtain the solutions and how task representations are displayed across unit's multidimensionality. Through analysis of the model's activations for each given task and its parameters, this project intended to elucidate how each cognitive faculty is interpreted by the algorithm.

The project's final goal was to study similarities in task representation between the artificial model and the biological system when solving the same tests. To do so, open-sourced meta-analysis of cognitive paradigms were considered for the biological counterpart.

I.8 Thesis Outline

This dissertation is composed by the following structure:

- In Chapter I, important concepts for the understanding of the dissertation is described and previous work relevant to this project is examined and detailed. In Section I.1 the field of cognitive neuroscience is introduced. In Section I.2 the meta-analytical approach to neuroimaging is described with the working example of Neurosynth. Section I.3 focuses on cognitive ontology history and the different approaches taken. Section I.4 gives a brief description of ANNs and its history and finally, Section I.5 describes the most prominent literature regarding the use of neural networks to model biological systems with special emphasis on models built to solve cognitive tasks;
- Chapter II describes, in depth, the methods applied in the execution of the project and the considerations regarded when defining each decision. It provides a thorough description of the dataset created, the reasoning behind the chosen tasks, the neural network model's structure and the terms associated from the Neurosynth database. Finally, it describes a pipeline for analysis of the model's mechanisms and its relationship with the biological model;
- In Chapter III, the experimental evaluation of the proposed methodology is presented. The model is appraised, and its activations and parameters are analysed. The results obtained using the pipeline described in Chapter II are described;
- Chapter IV discusses the results obtained and how they relate to the available literature. It debates on some aspects of the model and the function of relevant units. This chapter also discusses how the comparison between the model and the Neurosynth terms should be regarded as well as the project's limitations;
- Chapter V considers the project's main conclusions. It delineates work that is ongoing or planned that will build upon this project to make a more mechanistic cognitive ontology a reality.

II Methods

This Chapter describes the methodology behind the project for the creation of the model and its subsequent analysis. Section II.1 details the structure of the dataset, how it was created and the flexibility it allows for adding new tasks. Section II.2 presents a description of the tasks that were coded into the dataset for the model to learn. Section II.3 presents a brief explanation on the mathematical concepts of neural networks on which the model depends. Section II.4 describes the model developed for this project and the hyperparameters considered for the effective learning of the tasks. Section II.5 reports the methodology of the model analysis. Section II.6 explains the decision process behind the choice of the Neurosynth terms used to describe the learned tasks and Section II.7 describes the methodology applied to compare these terms to our model's results.

The source code related to the development of the project is made available on Github[5]. This consists of source code to generate the datasets, for the implementation of the proposed neural network architecture and for the posterior analysis and results obtained.

II.1 The Dataset

For the purpose of training and testing the model, a dataset was generated with 2000 trials covering 6 tasks that the model was intended to learn: The Go Task, the Anti Go Task, the Memory Go Task, the Memory Anti Go Task, the Reaction Go Task, and the Reaction Anti Go Task. The aforementioned were based on tasks described in the work of Yang et al. [2].

Each trial was composed of two separate matrices - the input matrix, which was fed to the model, and the output matrix, which was employed to evaluate the model's prediction and define the error function. Independently of the task of the respective trial, all input matrices have the dimension 71 by 600 and all output matrices have the dimension 33 by 600. The matrices' lines represent nodes and the columns represent time-steps. The reasoning behind each trial lasting for 600 time-steps is the considered limit for LSTMs to process long-term patterns (between 500 and 1000) wherein RNNs are only able to process dependencies of up to 100 time-steps [53]. If each time-step is considered to last 10ms, then each trial has a reasonable duration of 6 seconds.

The input signal (71 x 600) can be decomposed into 4 different aspects – the rule signal, the fixation unit and the two modalities.

The rule signal corresponds to the first 6 nodes and is implemented as a one-hot vector (in which one value is expected to be 1 and the rest to be 0) to represent which one of the six tasks is being executed in a given trial. The rule signal is constant across every time-step of a trial.

The 7th node represents the fixation unit – a binary value that cues when and whether the model should react to the stimulus or not. Across time-steps, if the value of the unit is 1, the model should not

[5] https://github.com/PedroFerreiradaCosta

react and should maintain fixation. When the value drops to 0, this signals that the model should react to the stimulus according to the task (i.e., consistent with the rule representation).

The latter 64 nodes represent two separate modalities of 32 nodes each. Each modality can display an arbitrary signal, by activating one of its 32 nodes. In the original paper, from where this project's tasks are modelled, they proxy rodent behavioural tasks, with the inputs representing a direction in a circle that a rodent should take in order to collect its reward [2]. In a similar way, the signal can represent the direction to which the behavioural response should be executed (Figure II-1). It is in modality one or modality two that the stimulus, to which the model should react, is represented. There is only one stimulus per trial and its duration depends on the task in hand. It is important to note that the need for both modalities instead of just one arises from the next stages of the project in which new tasks will require the model to discriminate between stimuli in each modality. In order to maintain the dataset structure, it was defined to already account for these new tasks.

To simulate noise in underlying neural representations, the input matrix is combined with additive Gaussian noise (i.e. $N(0; 0.1)$). In total there are $N_{in} = 1 + 6 + 32 \times 2 = 71$ input units.

The output signal (33 x 600) is a one-hot vector representation of the desired action of the model. It is divisible into two components – the fixation unit and a 32-unit reaction output. As with the input's fixation unit, the output's fixation signal sets a unitary value for when the model stays put (fixates) and drops to 0 to react in a given direction. In a computational sense, it allows the model to have an output signal even when it is not required to respond. The reaction output is structured similarly to the input's modalities and it corresponds to the input representation where the model should define the response to the stimulus when the fixation is dropped. In total there are $N_{out} = 1 + 32 = 33$ output units.

Figure II-1 depicts an example of a Go Task trial that serves as both input and output for the ANN.

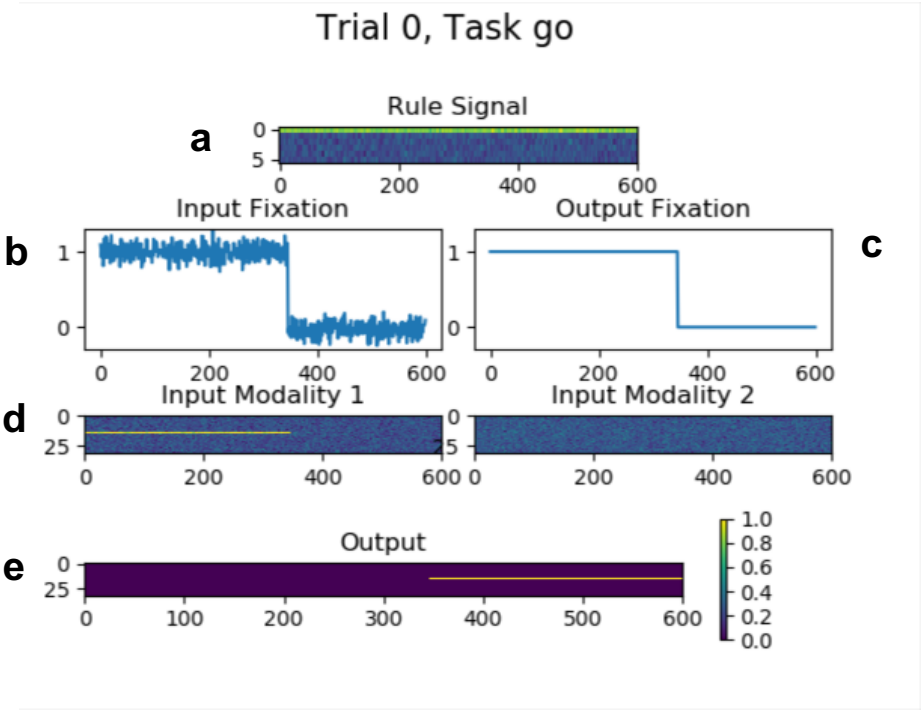


Figure II-1- Representation of the dataset provided to the model for a trial of a Go Task. a: Rule signal indicates which of the 6 trials is being considered, in this case unit 0 (i.e. Go task) is on throughout the trial. b: Input of the fixation unit. It falls to zero when the model is expected to react. c: Expected output of the fixation unit. The model is expected to mimic the fixation unit, in order to understand when it should react. d: The input modalities consist, each one, of a 32-vector input per time step that simulates the direction of the stimulus. For this project, only one modality has a signal for each trial. e: The expected output signal represents the direction the model should take for a given trial, depending on the task. In this case, the model is expected to follow the stimulus in modality 1, when the fixation reaches zero. Examples for the other 5 cognitive tasks are set in VII.1.

II.2 The Tasks

In choosing the learning tasks for the model, three aspects were taken into consideration: each task's commonality across neurophysiological studies, its simplicity for posterior analysis and the flexibility to allow the same input dimensions for different tasks. Based on the work by Yang et al. [2], 6 tasks were chosen and implemented that included variants of saccade, inhibitory control, parametric working memory, memory-guided response and reaction time.

All tasks have the same input and output dimensions, as described in Section II.1 so what differentiates between tasks is the rule signal and how the stimuli are expressed across the modalities of the input and the expected reaction to the stimulus in the output. For each task only one dedicated unit could be active in the rule signal, as was previously explained. In order to assure reproducibility and to facilitate posterior analysis, the random values in each trial are registered and saved (these include the duration of the stimulus, the time of reaction or the task being covered by the trial). It is

subsequently detailed how each task simulation was modelled and the dataset for each task is shown in the Annex A.

The **Go Task**, which could be described with the instructions: “react in the direction of the stimulus when the fixation stops”, is a fundamental common across many behaviour tasks [54]. In the present dataset, the stimulus is randomly assigned an input representation in one of the two modalities. The stimulus duration is sampled from a uniform distribution between 100 time-steps and 500 time-steps. When the stimulus disappears, the fixation stops, and the model is supposed to react in the direction of the stimulus (one of the 32 units).

The **Anti Task** can be described as: “react in the opposite direction of the stimulus when the fixation stops”. This involves an inhibitory reaction from the participant, not directing their gaze in the direction of the stimulus [55]. The task is the inverse of the Go Task. The stimulus duration is also sampled from a uniform distribution between 100 and 500 time-steps, but the expected reaction should be taken at 180° from the original stimulus. For instance, the reaction to a stimulus in the 16th unit should be taken towards the 32nd unit.

The **Delay Go Task** can be described with the instructions: “memorize the direction of the stimulus and react to it when the fixation drops” [56].This type of cognitive task would only be solvable with a model with some form of memory of past events, such as the LTSM recurrent network presented here (see Section II.3). The stimulus duration is uniformly sampled between 100 and 300 time-steps and the delay period, when the model should maintain fixation, is obtained from a discrete distribution from the values of (50, 100, 150, 200, 250) time-steps. After this period fixation stops and the network should react in the direction of the memorized stimulus.

The **Delay Anti Task** is the inverse of the previous task, being described as: “memorize the direction of the stimulus and react in its opposite direction”. Similarly, the stimulus duration is uniformly sampled between 100 and 300 time-steps and the delay period is also sampled from the discrete interval (50, 100, 150, 200). Unlike the previous task, though, when the fixation stops, the model should react in the opposite direction to the memorized stimulus.

The **Reaction Time Go Task** is similar to the Go Task but the reaction to the stimulus should be taken as fast as possible, while the fixation never drops. The stimulus is present for a duration chosen from a uniform distribution between 100 and 500, but while the stimulus drops, the fixation is kept throughout the whole trial and the reaction should be taken as fast as possible when the stimulus is detected.

The **Reaction Time Anti Task** can be described as: “react in the opposite direction of the stimulus, as fast as possible”; as with the React-Go task, the fixation signal never drops to 0. The stimulus duration is defined by a normal distribution between 100 and 500 time-steps.

II.3 Neural Networks and LSTMs

As mentioned in Chapter 1, ANNs are computing systems inspired by the biological neural networks that are able to perform tasks without explicit direction on how to execute them. They are a crucial aspect of this project as the analysis of the mechanistic properties of task-solving models might help explain biological networks.

The simplest neural network is composed of a single unit that applies a linear transformation to the input data given a few adjustable parameters and employs a given activation function to output a single value. This computation is described mathematically in Equation II.1, where $x \in \mathbb{R}^n$ is the input data, y is the network's output, $w \in \mathbb{R}^n$ is the vector of weights, b is the bias term and $\varphi(.)$ is a non-linear activation function (e.g. a logistic sigmoid or a hyperbolic tangent).

$$y = \varphi\left(\sum w_i \times x_i + \mathrm{b}\right) = \varphi(w^T \times x_i + \mathrm{b}) \qquad (\mathrm{II}.1)$$

This mathematical concept is the building block for larger and deeper networks and is called both neural unit or node. In this project both terms are used interchangeably. The output of a given unit or sets of units (i.e. layers), can serve as input for a deeper layer, resulting in a more complex and abstract representation of the input values. This is the concept behind Deep Neural Networks (DNN) which have displayed remarkable success in different supervised machine learning problems [57]. The model is trained by adapting the units' weights towards the desired solution using back-propagation, a two-step algorithm. The first step consists of a forward-pass, in which the network's output is obtained by flowing the input data along the units and the error is calculated based on its variance from the expected results. The second step consists of a backward pass, where the partial derivative of the cost function with respect to a given weight or bias is calculated to determine how much did the weight influence the error and in which direction it should change its value for a better performance of the network. The parameter's partial derivatives with respect to the cost function are calculated using the chain rule of differentiation.

The weights and biases updates are done by a gradient-based optimization algorithm. A commonly used optimization algorithm due to its efficacy in training DNNs is the Adaptive Moment Estimation (Adam) algorithm. The Adam algorithm updates the parameter by considering an exponentially decaying average of the previous gradients, while applying adaptive rates for each parameter. This allows the algorithm to apply larger updates on parameters that are used with less frequency [58].

Because cognitive tasks tend to be time-dependent (i.e. past information is required to solve the task in hand), this project applied RNNs with LSTM units to solve the generated dataset. LSTM units apply a different mathematical algorithm than simple artificial neural units, by applying a gating mechanism to a cell state, which simulates memory.

Each unit is composed of three gate layers (i.e. the forget, the insert and the output gate) with different functions of processing the cell state data. The forget gate (f_t) consists of a logistic sigmoid function (i.e. $\sigma(\cdot)$) applied to the linear transformation from the recurrent unit inputs with its parameters (Equation II.2). The result is a scalar ($f_t \in [0,1]$) that determines how much information of the previous cell state will be preserved. The insert gate (i_t) is, similarly, a scalar computed by a sigmoid function applied to the linear transformation of the input with the gate's parameters (Equation II.3). It determines how much of the new candidate values ($\widetilde{C_t}$) should be considered to pass onto the cell state. The candidate values are obtained by applying a hyperbolic tangent to the linear transformation between the input of the recurrent unit and its parameters (Equation II.4). The new cell state is then a consideration between the previous cell state and the candidate values, after multiplied by their respective gates (Equation II.5). The output gate is a scalar obtained by the logistic sigmoid function applied to the linear transformation between the unit's input and the gate's parameters (Equation II.6) and will determine the proportion of the new cell-state (C) that will serve as the unit's output (h_t) after a hyperbolic tangent transformation (Equation II.7). In the following equations, W stands for each gate's weight matrix and b for the bias term.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (II.2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (II.3)$$

$$\widetilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (II.4)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (II.5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (II.6)$$

$$h_t = o_t * tanh(C_t) \quad (II.7)$$

Besides allowing for larger time-dependencies, this model permits the inspection of the cell state which will prove useful to understand what information the model decides to keep in each iteration and what it deems dispensable. For these two reasons, an LSTM network was considered for the baseline of our model when solving cognitive tasks.

II.4 The Model

The following image is a pictorial representation of the model that was developed for this project with the goal of solving 6 cognitive tasks with temporal dependencies.

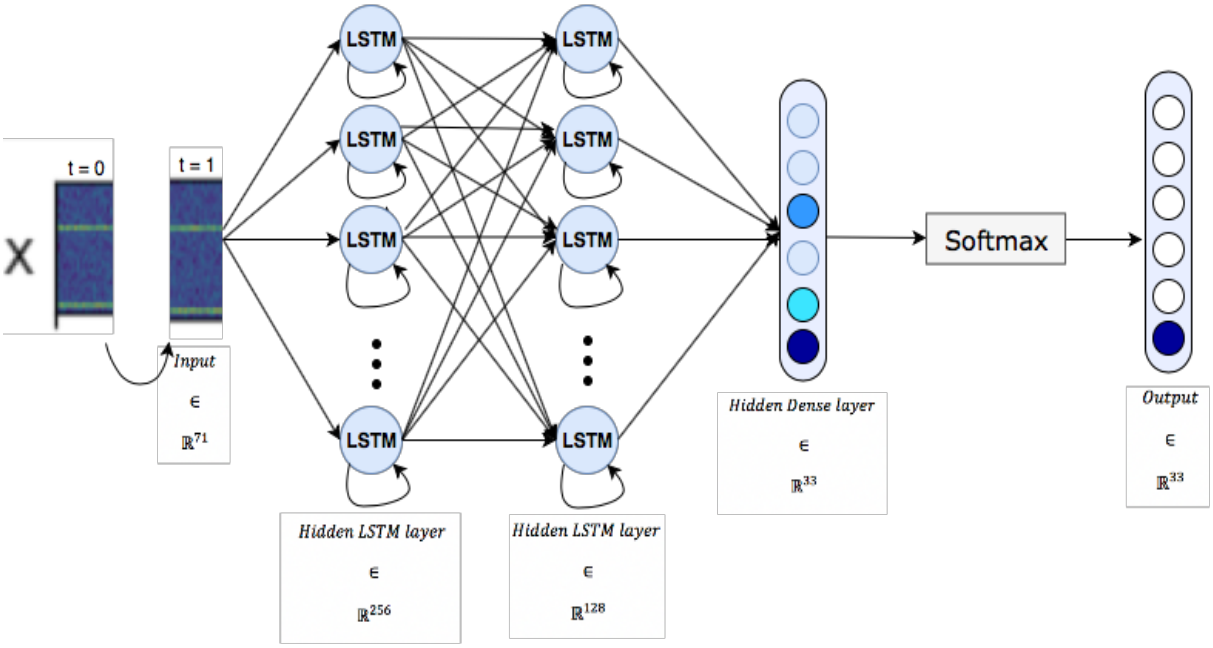


Figure II-2 - The proposed neural network architecture. The model is composed by two densely connected LSTM layers and a third dense layer that, after a softmax activation outputs a vector of size 33.

After building the dataset, the deep RNN structure was defined and implemented using Keras[6], a high-level neural networks API with TensorFlow[7] as its computational backend. The basic recurrent unit applied was an LSTM, as it is able to maintain information for longer periods than regular recurrent units and store a cell state, simulating the network's memory storage. Overall, the software architecture involved the use of 4 specific Python packages (i.e. keras, numpy[8], sklearn[9], and matplotlib[10]). The entire model was trained end-to-end from the dataset created for the project.

Most of the model's hyperparameters were chosen using empirical heuristic methodology. The model was trained with an increasing number of layers and units per layer until it was able to learn the tasks with an acceptable accuracy. The final network was composed of 3 layers (see Figure II-2):

- A first layer with 256 LSTM units;
- A second layer with 128 LSTM units;
- A third dense layer of 33 units with a softmax as the activation function.

[6] https://keras.io
[7] https://www.tensorflow.org
[8] http://www.numpy.org
[9] http://scikit-learn.org/stable/
[10] https://matplotlib.org

The *softmax* function, also known as the normalized exponential function [59], allows the representation of a categorical distribution, that is, a probability distribution over the 33 possible outcomes. The desired outcome per time-step is a one-hot vector of 33 values, whereby the model's response is the unit with the highest probability, independently of how high the other units' probabilities are. Mathematically, the *softmax* function can be written as in Equation II.8, where x stands for the layer's input and w stands for j^{th} unit's trainable weight.

$$P(y = j \mid x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \qquad (II.8)$$

The model was fully interconnected between layers 1, 2 and 3 with 537,249 weights, all of which were trained using a categorical cross-entropy loss function (Equation II.9), where q is the approximate distribution and p is the real distribution. The optimization function corresponds to the minimization of the sum of the loss functions across every time-step j and trial i of a given batch. Each batch was composed of 64 randomly chosen trials – Equation II.10.

$$H(p, q) = -\sum_{x} p(x) \log q(x) \qquad (II.9)$$

$$min \sum_{i} \sum_{j} H\left(p_{i,j}, q_{i,j}\right) \qquad (II.10)$$

In order to effectively train the model, the Adam optimizer method was used, a type of stochastic gradient descent, with the parameters following those proposed by the seminal paper of Kingma and Ba [58]. The learning rate used was 0.001 with a decay rate for the 1st and 2nd moments of 0.9 and 0.999 respectively.

Assuming a reasonable data normalization, the weights are expected to be zero or close to it. It could seem that it might be reasonable then to set all weights to zeros for its initialization. This, however, would result in all weights having the same error and the weights would be equal to each other across training – there would be no break of symmetry. Therefore, weights were initialized as follows:

- The kernel weights were initialized with the Glorot Uniform Initializer [60], an effective method that draws samples from a uniform distribution;
- An orthogonal initialization was utilized for the recurrent weights to avoid vanishing or exploding gradients [61];
- The bias weights were all initialized to zeros, as the break in symmetry is guaranteed by the previous initializations.

Table II-1 - Parameters considered for the correct learning of the six cognitive tasks by the model.

Parameter		Value
Batch Size		64 trials
Number of epochs trained		250
Loss Function		Categorical Cross Entropy
Gradient Optimizer		Adam
	Learning rate	0.001
	Decay rate - β_1	0.9
	Decay rate - β_2	0.999
Weight Initializer	Kernel Weights	Glorot Uniform Initializer
	Recurrent Weights	Orthogonal Initialization
	Bias Weights	Initialized to zero

II.5 Model Analysis

After successfully implementing the model to correctly learn the six cognitive tasks, it was this project's intention to understand the mechanisms used by the model to correctly solve the tasks, based solely on the input it is given. Towards this goal, a number of analyses on the model's activations, parameters and unit variance were run on Python.

Understanding how the activation for each layer varied according to the different variables present in the dataset (i.e. task, moment of reaction, delay period, stimulus duration) is a multidimensional problem. Therefore, t-distributed stochastic neighbor embedding (t-SNE), an algorithm that allows representation of high-dimensionality data in two dimensions, was applied to the model while solving the testing set. T-SNE is widely used in machine learning and consists of a first stage of constructing a probability distribution over pairs of high-dimensional object based on similarity (in our case, pairs of trials), and a second stage where the algorithm defines a similar probability distribution over the objects in a low-dimensional map. This algorithm was used to study both output and cell state representation on each layer.

To study the importance of each unit for the correct execution of different functions required in each task, a variance study was conducted. Variance can be written as in Equation II.11, where μ is the set's average, n is the number of trials and x_i relates to a single trial.

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (II.11)$$

For each unit, the value of interest was the moment after the model reacted to the stimulus. For this reason, for the variance study, the moment of 10 time-steps after initial reaction was considered. Different moments were also analysed and are reported in Appendix B. In order to understand units concerned with functions present in every task and functions present only for a specific task, a second variance analysis was run on each task's average of variance.

Finally, the most relevant units were turned off in order to study the model's ability to maintain efficacy on the cognitive tasks without these mechanisms.

II.6 Neurosynth Terms

In order to investigate if the partition of task space based on how the ANN solves multiple tasks is consistent with how tasks are represented in the brain, the meta-analyses of existing task fMRI data was considered. To do so we used the online platform Neurosynth and considered the unthresholded inference z-maps associated to the relevant terms in the MNI (i.e a standard brain widely considered in Neuroscience to facilitate reproducibility of results) space. This resulted in neural similarity matrices between terms that could be compared to the model's similarity matrices. The criteria for choosing the terms to search for in the Neurosynth database were their relationship to the behaviour being mimicked by the tasks, the number of studies the given term had supporting its brain map representation and the similitude of the most relevant studies' tasks to our own. Because the tasks modelled are very simplistic, it is difficult to find terms that are complete analogues for human performances. Also, for some tasks there is more than one suitable term and, in these cases, the coactivation maps are combined together by adding the two activation values. In this regard, the brain representations of the terms "finger tapping", "motor task", "reaction time", "reaction times", "switch", "switching", "wm task", "working memory", "nogo" and "incongruent" were obtained from the platform. The following table explains the reasoning behind the choice of terms for each specific task.

Table II-2 - Tasks and respective Neurosyth terms used for analysis and the reasoning behind each term choice.

Tasks	Neurosynth terms	Description
Go Task	finger tapping, motor task	Finger tapping in response to a stimulus or simple motor tasks are some of the easiest cognitive paradigms for healthy participants to perform. Most commonly, participants just need to tap their fingers in response to a stimulus.

Tasks	Neurosynth terms	Description
Delay Go Task	working memory, wm task	Working memory tasks require participants to temporarily hold information for cognitive processing at a later stage in the experiment.
Reaction Time Task	reaction time, reaction times	In contrast to simple finger tapping paradigms, cognitive tasks measuring reaction times may include instructions for participants to respond as fast as possible to a given stimulus.
Anti Task	no go	While the identification of corresponding Neurosynth terms with the 'go' tasks could be accomplished in a relatively straightforward and direct manner, the 'anti' tasks have no ideal correspondence to standard cognitive paradigms. Therefore, an attempt was made to identify Neurosynth terms that display different levels of complexity for the anti task, the Delay Anti task and Reaction Time Anti task. In this respect, for the Reaction Time Anti task, the emphasis was put on identifying cognitive paradigms that highlight processing speed of the 'anti' response. For the Delay Anti task, the emphasis was put on more complex processing in terms of suppressing the response to stimuli that are incongruent with the target stimulus. Therefore, for the 'simplest' version of the Anti task category, the Anti task, a cognitive paradigm was chosen that does not emphasise reaction times (Reaction Time Anti task), or the processing of complex stimuli arrangements (Delay Anti task), and the No go task paradigm was selected that only involves participants to withhold a button press in response to a stimulus.
Delay Anti Task	incongruent	Cognitive paradigms involving incongruent stimuli measure the ability of participants to suppress inappropriate responses. Commonly, a target stimulus is flanked by non-target stimuli that in the case of incongruent flankers calls for the opposite response to the target. Typical cognitive tasks are the Eriksen flanker task or the Simon task.

Tasks	Neurosynth terms	Description
Reaction Time Anti Task	switch, switching	The switch task involves participants reversing their response mapping for trials. As an example, if a participant had to respond with the right hand to a red 'x', then for switch trials, the participants needs to respond with the left hand to a red 'x'. Switch tasks usually involve the instruction for participants to respond as fast and accurately as possible and switch costs are measured as the difference in reaction time between 'non-switch' and 'switch' trials.

II.7 Correlation Studies

It was this project's goal to understand if task representations would be similar between artificial models and biological models. For this purpose, a number of parametric and non-parametric correlations were run, and the Fisher z-transformation was applied to the results.

For the artificial model a similarity matrix was built on the correlations of variance across tasks. For the biological counterpart, the relevant Neurosynth terms were used to simulate the brain's activation structure for each task. Then, three different methods were used to analyse the similarity matrix between tasks' correlations: considering the whole brain, considering a brain region at a time and considering different brain networks separately. Finally, both similarity matrices were studied and compared.

II.7.1 Whole Brain Correlation

The first analysis was run by considering the similarity matrices of the correlations between whole brain maps for the different terms. For each term-associated task, the 3-dimensional cube is flattened into a vector and correlated with every other. This method considers the variations between tasks as a whole, without focusing on a specific region or process, as the activations report different functions required to solve a given task. The similarity matrix is then flattened and correlated to the similarity matrix obtained from the task correlations in the artificial model.

II.7.2 Searchlight Method

The searchlight method considers a smaller dimension from the MNI space and iterates through the whole brain maps, correlating one at a time for each subset. The user defines the length of the cube and the stride it takes after each iteration.

For each iteration, a cube with the given length is considered as a subset of the brain maps of the Neurosynth terms. This subset is then correlated, generating a similarity matrix for a smaller region of the MNI space. The similarity matrix is correlated to the artificial model's similarity matrix and the

obtained value is added to the respective subset of a new MNI space. A new iteration is then run by repeatedly moving the searchlight cube until the whole of the brain has been covered. As a last step, the correlations in the MNI space are weighted according to the number of correlations that were added to each voxel.

The final result is a brain map where the correlations between the artificial model and the Neurosynth terms are considered for each separate region. This method will look for similar task representations at a regional level and assumes the artificial model might simulate specific functions instead of the whole process taken by the brain when solving cognitive tasks.

II.7.3 Network Correlation

The last method takes a more functional approach to the correlation analysis. Instead of individually comparing brain regions or comparing the brain as a whole, this method considers as a correlation space the set of brain networks identified by Yeo et al. [31]. His work consisted of exploring the cerebral cortex using resting-state functional connectivity MRI and, by employing a clustering approach, identifying several local networks of functionally coupled regions. A coarser parcellation resulted in the identification of 7 networks (Figure II-3 a) and a finer solution identified 17 networks (Figure II-3 b). Both sets of networks were considered in this analysis.

Within each network, the brain map activations for the Neurosynth terms are correlated, generating a similarity matrix that is then correlated with the tasks' similarity matrix from the artificial model. These networks are identified based on function and, consequently, the comparison between the artificial and the biological structures might reveal how they relate to the model mechanism for solving similar cognitive processes. Afterwards the results were tested for statistical significance by employing the "max statistic" approach to control the family-wise error rate. The p-values were adjusted by considering the null hypothesis that every possible order of the task correlations was equally likely. This was tested by performing a large number of the same correlation test for sets with their data order randomized.

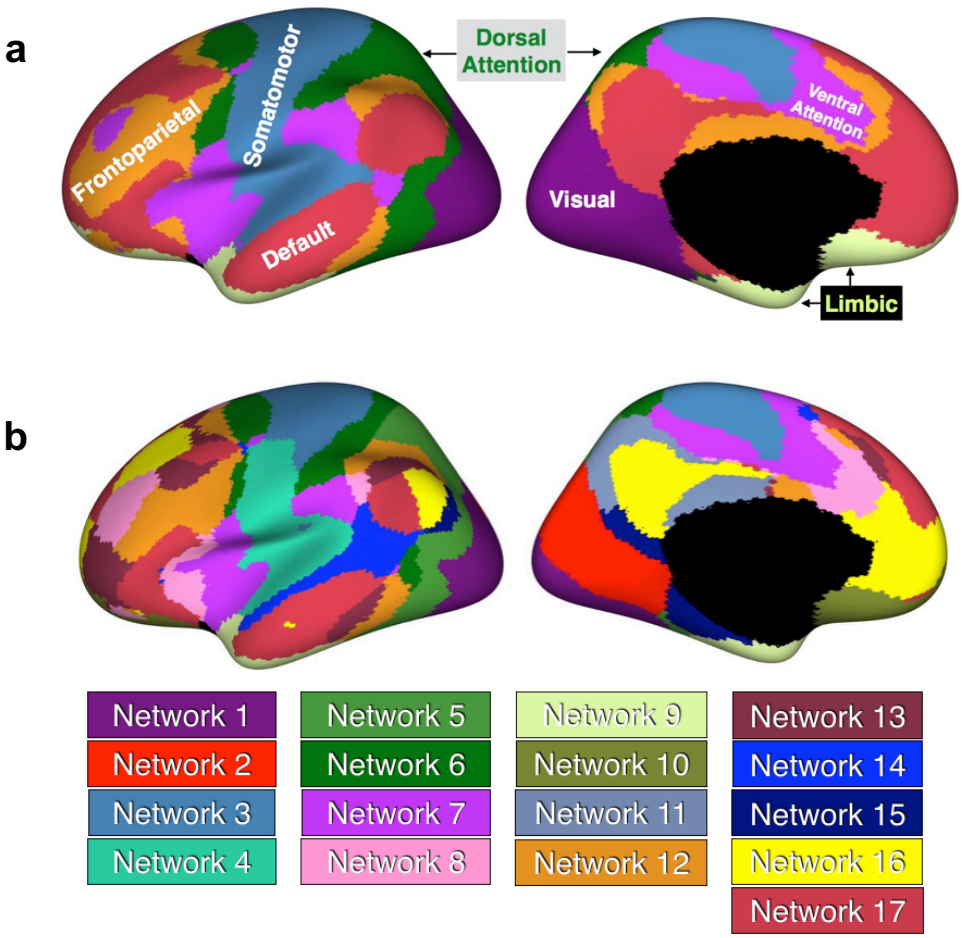


Figure II-3 - Yeo Networks with the respective colour mapping. a: Medial and lateral views of the 7-area atlas on the inflated FreeSurfer average brain. b: Medial and lateral views of the 17-area atlas. Figure adapted from [31].

III Results

This chapter describes the experimental evaluation of the proposed method. Section III.1-presents the process of building a functioning model capable of performing the 6 tasks and assesses its general performance. Section III.2 analyses the model's patterns of activations (i.e., a way of understanding how the model represents different tasks). Section III.3 studies how each unit of the model varies its activations depending on which task it is executing. Finally, Section III.4 describes relationships between the model and the human brain when considering similarity matrices between tasks.

III.1 Evaluation of the trained model

The first experimental objective was to successfully train a model to perform 6 different cognitive tasks. In order to do so, different architectures with an increasing number of units were tested and the model accuracies at performing each task are reported in Table III-1. The dataset on which the models were trained contained 2000 trials, divided into a training set of 1600 trials (i.e. 80% of the full dataset), a validation set of 200 trials assessed at the end of each epoch, and a testing set containing 200 trials, to define how the model generalizes for values not seen before.

Table III-1 - Different models and their accuracy after training.

			Accuracy (%)		
Model	**LSTM Structure (hidden units)**		**Training Set**	**Validation Set**	**Testing Set**
1	1 Layer	64	65.67	59.1	61.9
2	1 Layer	128	64.14	60.42	56.34
3	1 Layer	256	42.55	51.77	36.61
4	2 Layers	64 128	50.75	48.79	48.43
5	**2 Layers**	**128** **256**	**98.84**	**93.54**	**93.28**
6	2 Layers	256 512	91.05	83.44	80.94

In theory, an increasing number of units allows for higher information processing capacity which allows the model more flexibility to perform each task [62]. Although this is not verified between Model 1 and 4, the model with 2 layers of 256 and 128 hidden units (i.e. Model 5) has the highest accuracy of the trained models. However, a model with a higher number of hidden units is prone to overfitting the training data, which is the case in Model 6. Model 5 results in a good trade-off between a slight overfit and a good performance as the model is able to attain a 93% accuracy in the testing set. A 2-layered structure allows the model to attain a higher level of abstraction, allowing for more complex representations of rules and variabilities from each task.

To demonstrate that the LSTM approach is preferable to alternative (simpler models) we trained alternative neural networks models substituting the LSTM units for nodes with simpler activation units (e.g., just recurrent connections rather than memory gates) while maintaining other training conditions. The results obtained are shown in Table III-2. The simpler models were not able to learn the complex dependencies required to correctly generalize for all the trials, resulting in an accuracy below 50% for the testing set. It is important to note that only the Model 5 architecture was tested for these different networks and that a different structure might have resulted in higher accuracy. Despite this, for the mentioned architecture, the results strengthen the confidence in the presented model and its efficacy on solving the cognitive tasks.

Table III-2 - Models with different networks and respective accuracy during training.

		Accuracy (%)		
			Validation	
Model	**Units**	**Training Set**	**Set**	**Testing Set**
Model 5	LSTMs	98.84	93.54	93.28
RNN	Simple RNNs	40.22	38.61	37.8
DNN	Activation Gates	50.47	47.87	47.18

To establish how well the model has captured the different tasks, it was run on 6 task-specific datasets of 200 trials each of unseen data. The accuracies are reported on Table III-3. Comparatively, the model underperforms on the Reaction Time tasks. It is also notable how, for the other 4 tasks, the accuracies on the 200 trials are alike, with special emphasis on the Go and Anti tasks (i.e. 85.9% and 85.8% respectively).

Table III-3 - Model 5's accuracy for the different tasks learned.

	Accuracy (%)					
	Go Task	**Delay Go Task**	**Reaction Time Go Task**	**Anti Task**	**Delay Anti Task**	**Reaction Time Anti Task**
Model 5	85.9	86.3	61.9	85.8	87.1	71

III.2 Analysis of the model's activations

To characterise how each layer contributed to the model performance, the different layers' activations were analysed with the non-linear dimensionality reduction technique, t-SNE [63]. t-SNE facilitates the interpretation of multidimensional arrays, as is the case of the layers' activations, by placing each multidimensional object into a two-dimensional space based on its similarity to the other objects. In this analysis, each point represents one of the trials from the testing set (200 in total).

The trials are colour labelled by the different variables that the model was hypothesized to have to learn for correct performance (i.e. the task, time the model should react, delay period between stimulus and reaction and duration of the stimulus). Through this method it is possible to analyse both the final outcome and the activations of the inner layers, including the cell states (see Figure III-1.

It is evident from the Figure III-1 a, that there is a significant similarity between trials of a given task across the cell states of the first layer of the LSTM. The 6 clusters are grouped by the task being performed, which demonstrates how the model represents how different tasks relate to different expected reactions. It is also evident from Figure III-1 b that trials are organised by when the model has to react, with later responses occupying the distal region of each cluster and the earlier ones being represented in the proximal regions. This is consistent with the cell state of the first layer which allows the information to be retained with minimal losses between time steps, allowing for its later retrieval.

The t-SNE applied to the output activation patterns of the first LSTM layer (Figure III-1 c and Figure III-1 d) highlights a similar result, which is expected as the layer is directly influenced by the values of the cell-state. Despite this, there is no clear distinction between task representations, with some tasks appearing more overlapping (i.e., Go and Delay Go tasks and the Anti and the Delay Anti tasks). Furthermore, there is an apparent symmetry between both Go/Anti tasks and Delay Go and Delay Anti tasks, an interesting result as the tasks behave symmetrically (i.e. the Go task should react towards the stimulus and the Anti task should react against it). Both Reaction Time Go and Reaction Time Anti tasks are representationally distinct from each former, an outcome present throughout Section III and that will be addressed in the Section IV. Figure III-1 d delineates a clearer representation of a gradient suggesting when the model intends to react. A number of other variables were colour coded to

the presented t-SNE plots without expressing any meaningful representation and are displayed in the Appendix C.

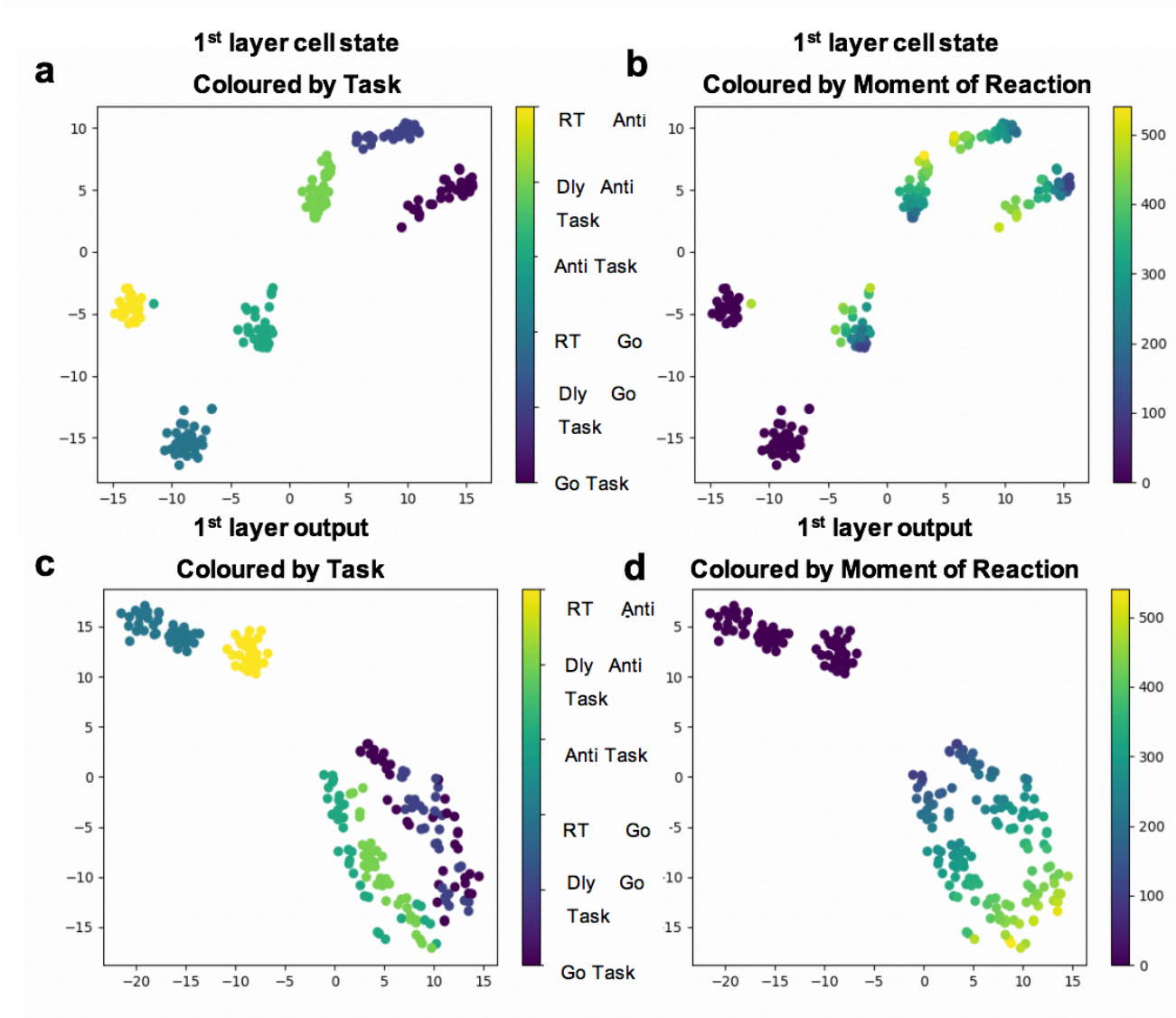


Figure III-1 - Four t-SNEs plots, each representing 200 trials as scattered points relating to Model 5's first layer. a: cell state activations across time with the colour label of the respective task. b: cell state activations across time with the colour label of the moment of reaction (Tgo). c: LSTM cell output across time with the colour label of the respective task. d: LSTM cell output across time with the colour label of the moment of reaction (Tgo).

The same analysis was executed for the second layer of the LSTM and the most significant results are displayed in Figure III-1 . The output of the model, which corresponds to the effective prediction of the desired outcome, demonstrate a much greater variation between trials, and there are no well-defined clusters from the t-SNE analysis that relate to known sources of variation across the trials. This result is expected due to the fact that the higher representational levels of LTSM networks encode increasingly complex combinations of ascending input and so are expected to correspond to

more unique and abstract representations. This is apparent even for the same task, with trials varying their properties, such as the moment of reaction and the duration of the stimulus, each resulting in a different 2nd level representation. This trial uniqueness forces the model to learn a more general concept to apply in the testing set instead of copying the tasks presented to it in the training set.

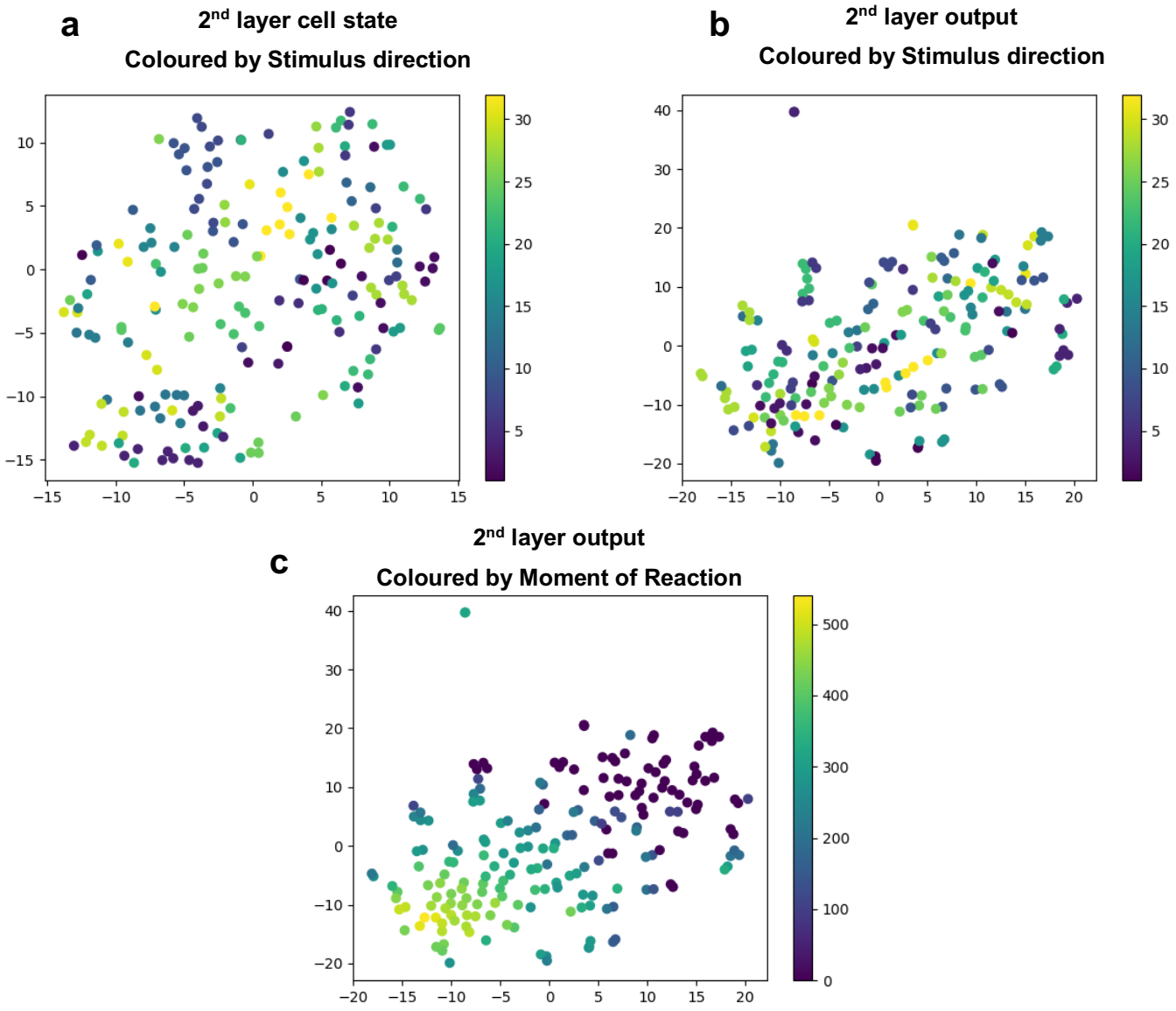


Figure III-2 - Three t-SNEs plots, each representing 200 trials as scattered points relating to Model 5's second layer. a: cell state activations across time with the colour label of the stimulus direction. b: LSTM cell output across time with the colour label of the stimulus direction. c: LSTM cell output across time with the colour label of the moment of reaction (Tgo).

Similar to the previous layers, there is a gradient in the second layer related to the moment the model should react, something that appears important for the correct execution of the task. It is also notable that the direction of the stimulus, which directly determines which unit the model should activate,

does not have a big impact on the general 2-D representation of the model's output. Despite this, it is still identifiable in the cell-state of the second LSTM layer as an agglomeration between some task directions, allowing for this information to be partially maintained across time steps even when the stimulus is no longer present.

III.3 Variability of individual units

To help us better understand the importance of each unit for the correct execution of a given task, the variance of each model's units was studied across trials for a given task. Each trial is composed of 600 time-steps with varying time of reaction and stimulus period. Therefore, the moment chosen for the variance analysis was locked to 10 time-steps after the model was expected to react (Tgo), although other moments were considered for the analysis and are reported in Appendix B; this allowed for the activations to settle into the correct response signal. High intra-task variance is expected to reveal the units responsible for functions such as the moment of reaction or the direction of the reaction (i.e. variables that change between trials of the same task). The results obtained are depicted in Figure III-3.

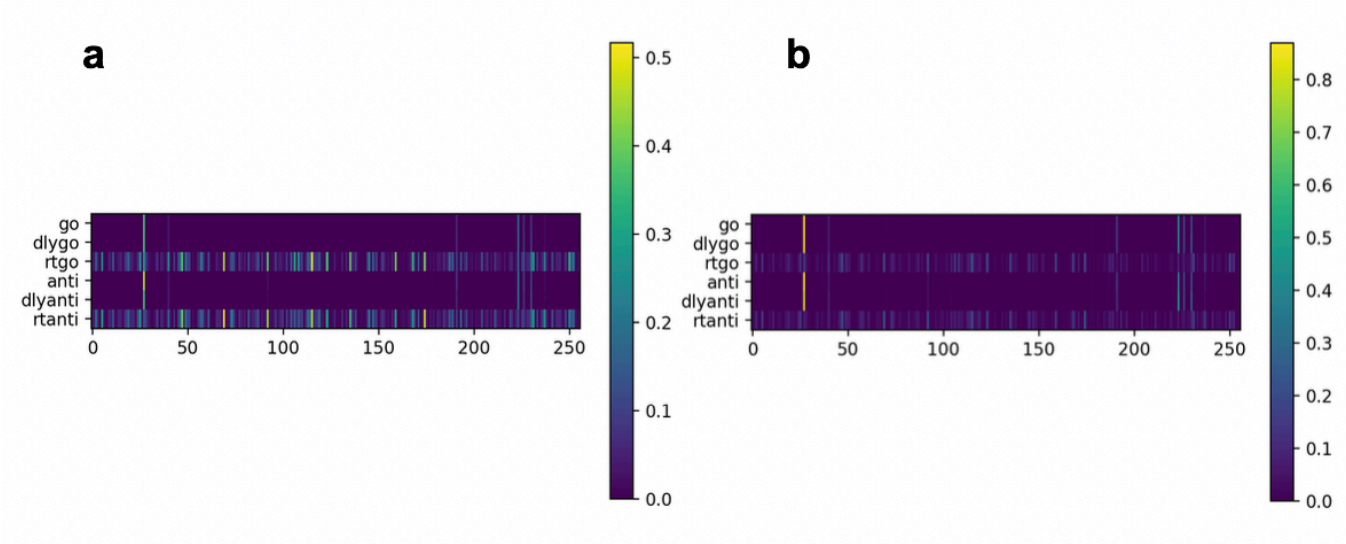


Figure III-3 - Variance of each first-layer unit across each task (a) and same results normalized across each task (b). Colour bars represents the variance intensity. Y-axis refers to the model's tasks.

The Reaction-Time and the Anti Reaction-Time tasks display a higher variance between trials along the units of its first LSTM layer. This finding agrees with the t-SNE representation that displayed the two tasks further from the rest. Due to the high variance present in both Reaction Time tasks, the lower variance for the units in other tasks are not visible in Figure III-3 a. For this reason, the variance was normalized across each task to gain a relative understanding of which units varied more for one specific task (Figure III-3 b). It is apparent that some units (e.g. units 28, 224, 227 and 231), display a function that is common across every task.

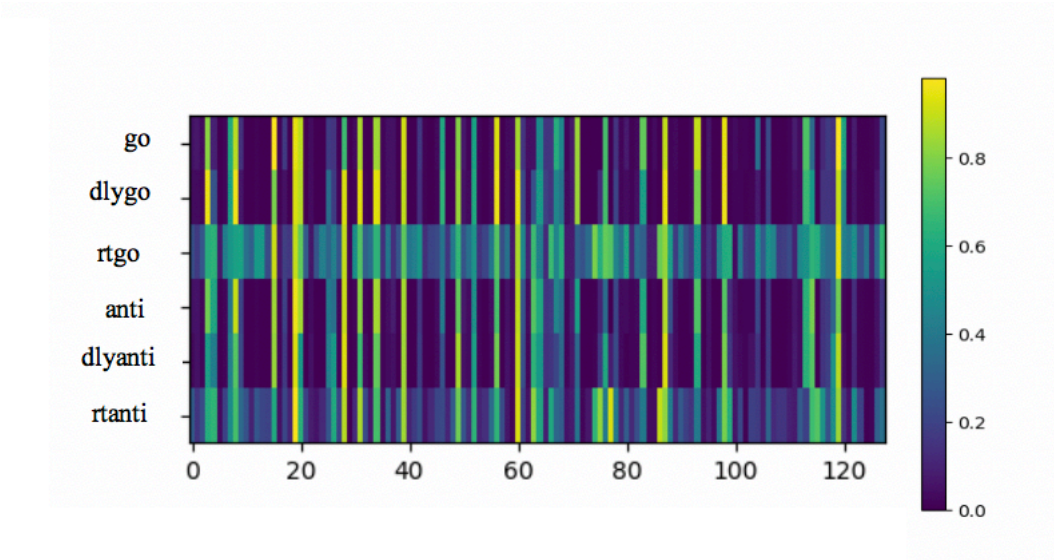


Figure III-4 - Variance of each second-layer unit across each task. Colour bar represents the variance intensity. Y-axis refers to the model's tasks (in descending order: Go task, Delay Go task, Reaction Time Go task, Anti task, Delay Anti task, Reaction Time Anti task.

In the second layer (Figure III-4) the variability within individual nodes across different tasks is more prominent, since activity is passed through a non-linear function between the first and second layers. There is a clear divide between units with high variance, which can be understood as units responsible for functions independent of the task (e.g. define the direction of the reaction; define the moment of reaction), and units with low to no variance, which might be responsible for functions specific to a given task (e.g., react towards the stimulus; react against the stimulus; react immediately). Subsequently, the variance of the node's activations across every trial is compared, independently of its task, to the variance between averages of the task specific trials. This allows us to quantify how each node relates to the desired outcome, as a node with a greater variance across trials has a more significant role on the final output than one with little variation. Also, by observing how the variance changes between different tasks, it is possible to discover whether a specific node is responsible for a function common to every task (the variance across task-averaged trials would be low), or alternately, for a function that is directly responsible for different aspects of a specific tasks (higher variance between task-averaged trials). This analysis is depicted in Figure III-5.

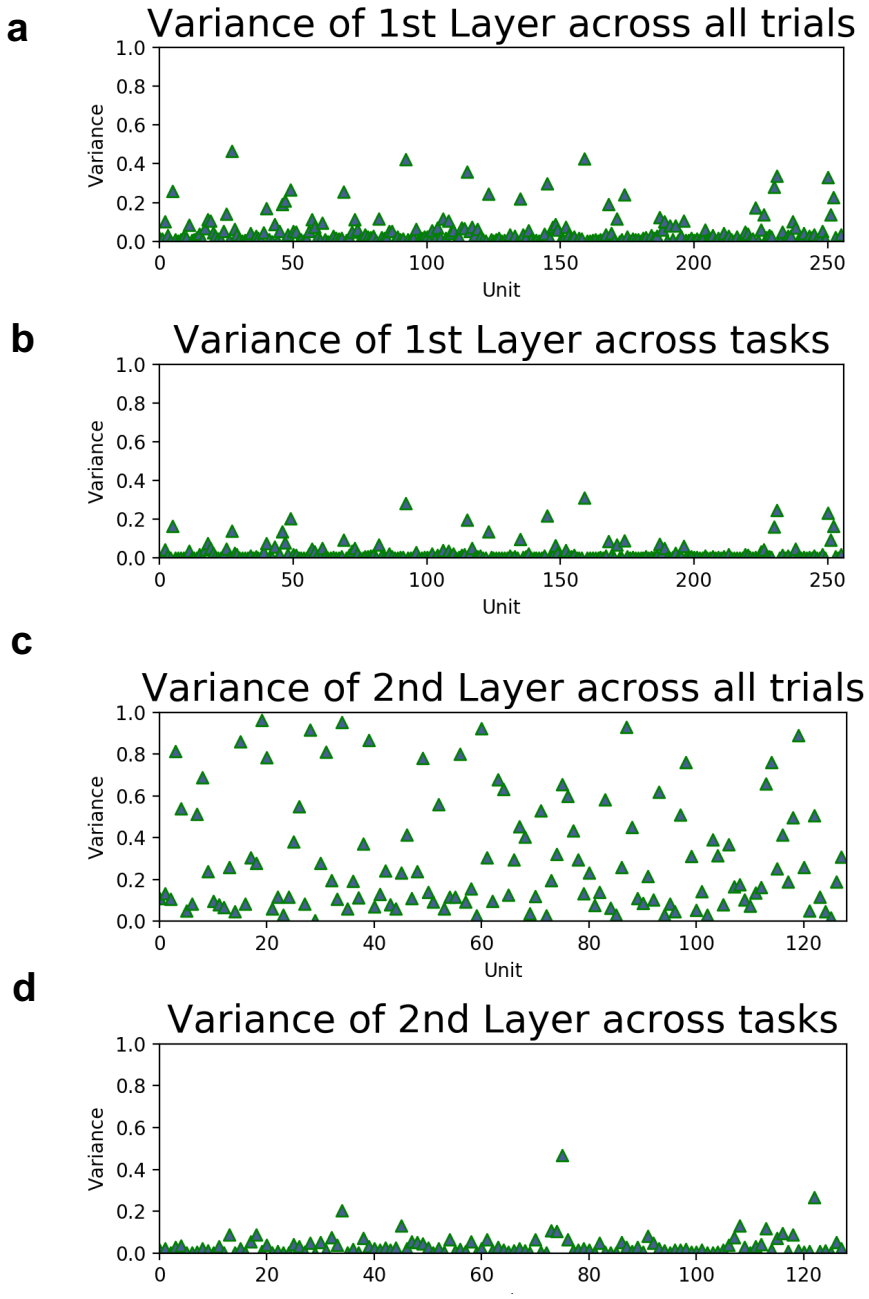


Figure III-5 - Four plots, analysing the variance (σ^2) for every unit of the model. a: variance of the first-layer units accounting for every trial. b: variance of the first-layer units accounting for the average for the trials in each task. c: variance of the second-layer units accounting for every trial. d: variance of the second-layer units accounting for the average for the trials in each task.

Although most of the nodes in Figure III-5 a display little variance, some display a clear variation of the value outputted from the first layer of LSTMs to the second layer. The most noteworthy ones are, in descending order, units 28, 160 and 93 with variance of σ^2 = 0.466, σ^2 = 0.427 and σ^2 = 0.424 respectively. As expected, in Figure III-5 b the general variance of the nodes decreases, as the comparison between 6 examples (one for each

task averaged across trials) are now considered instead of the 200 examples considered for Figure III-5 a. There is also an abrupt decrease in variability of unit 28 from σ^2 = 0.466 (the highest in the first analysis) to σ^2 = 0.142, significantly lower than the variance for unit 93 and 160 – σ^2 = 0.282 and σ^2 = 0.312 respectively. It is reasonable then, to consider that unit 28's function is related to something common to every task (e.g., the direction of the stimulus).

The second layer displays generally higher variance when compared to the first one, due to each unit accounting for the 256 flexible values as input instead of just 33. Therefore, 37 out of the 128 nodes present a variance higher than 0.40. Curiously, this high variance decreases significantly when the trials of a given task are averaged together (Figure III-5), resulting in an absolute variance across tasks even lower than the variance across tasks for the first layer (i.e. σ^2 = 4.39 and σ^2 = 5.80 respectively). This result leads to the assumption that the variance across the second layer is more relevant for solving issues common to every task, than to characteristics that vary between task (e.g. following the direction of the stimulus or the opposite one).

The Reaction Time task elicits a larger variation across trials for a given unit. For this reason, the same analysis was run without considering trials from this task. The results obtained are depicted in the Figure III-6. The variability drops to close to 0 in almost every node of the first layer as represented in Figure III-6 a, as the Reaction Time tasks account for most of the variance. Exception being the unit 28 with a variance of σ^2 = 0.396. As was the case when accounting for every task, there is a heavy drop in variance for unit 96, as expressed in Figure III-6 b, but because the variance is close to 0 for every unit, it is still the unit with highest variance at σ^2 = 0.009. For the second layer, in the Figure III-6 c, it is again noticeable a high variance for most nodes across trials, with a heavy decrease when considering variation between averaged-by-task trials in the Figure III-6 d. It is worth mentioning, that when studying the variance in-between tasks, only variance between 4 values is considered. The nodes with the highest variance in the second layer are unit 20, with a variance of σ^2 = 0.97 that decreases to σ^2 = 0.008 when considering task-averaged trials, and unit 29 with a variance of σ^2 = 0.96 which subsequently decreases to σ^2 = 0.05 which is the highest variance across tasks.

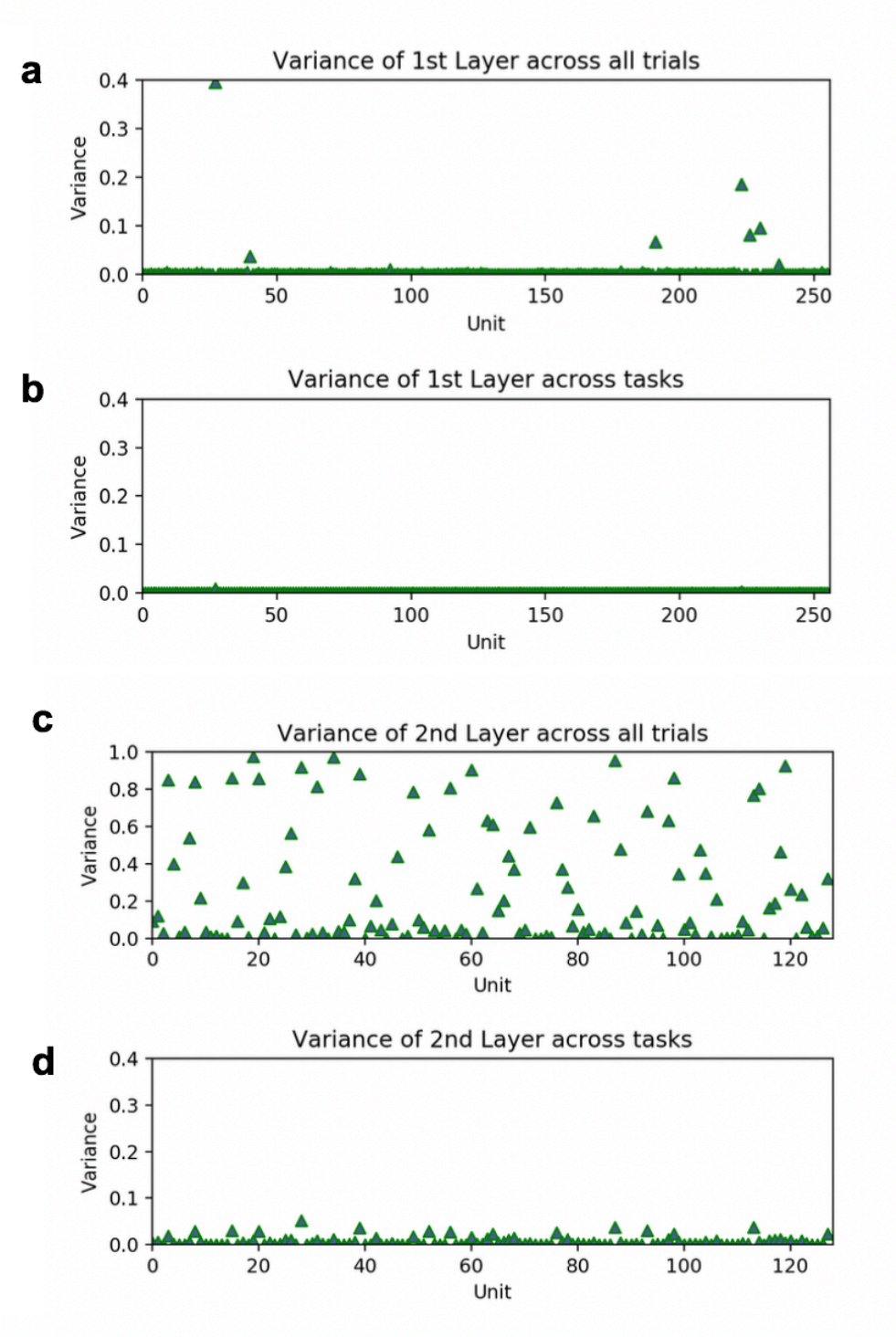


Figure III-6 - Four plots, analysing the variance for every unit of the model without considering the Reaction Time tasks. a: variance of the first-layer units accounting for every trial. b: variance of the first-layer units accounting for the average for the trials in each task. c: variance of the second-layer units accounting for every trial. d: variance of the second-layer units accounting for the average for the trials in each task.

III.3.1 Removal of the 28th unit from the first layer

Unit 28 from the first layer of the model displayed interesting patterns of activations. The variance displayed across trial activations exhibited a significant magnitude between tasks, with special emphasis on the Go, Delay Go, Anti and Delay Anti tasks. To further understand its effects on the overall output of the model, the unit was eliminated by reducing its weights to zero. The model was subsequently analysed on the different tasks and the accuracy is displayed in Table III-4.

Table III-4 - Accuracy for the original model and for the model with unit 28 from the first layer removed on the different tasks learned.

	Accuracy (%)					
	Go Task	**Delay Go Task**	**Reaction Time Go Task**	**Anti Task**	**Delay Anti Task**	**Reaction Time Anti Task**
Original Model	85.9	86.3	61.9	85.8	87.1	71
Model without unit 28	46.1	36.9	60	44.1	41.4	71.2

It is evident a clear decrease in the overall accuracy of the model when unit 28 is dropped. This drop is supported by a decrease of accuracy in all tasks but the Reaction Time Go and Reaction Time Anti tasks. These two have nearly reported the same percentage of accuracy with and without the 28th unit from the first layer. The largest drop in accuracy is verified in the Delay Go task (i.e. from 86.3% to 36.9%). The model is composed of 384 units so the impact of a single unit in the overall accuracy is remarkable.

With the goal of further analysing the effects of the elimination of unit 28, the model outputs were compared to the original model with special focus on the tasks where accuracy dropped heavily. The most significant results are displayed in Figure III-7.

The analysis made clear that without unit 28 the model was not able to follow fixation before the expected reaction. The altered model reacts to the stimulus the moment it is shown, resulting in a lower accuracy in the Go, Anti and Delay tasks and, logically, not affecting the Reaction Time tasks as they do not depend on the fixation input. The examples of Figure III-7 are verified across the dataset.

In summary, for the first LSTM layer, unit 28 displays importance across every trial that requires to follow fixation, which is not the case for units 93 and 160 which appear to be only relevant for the correct solution of the Reaction Time tasks. These units have relatively transparent functional roles,

whereas other nodes are also functionally involved despite not varying their activations a lot between trials. This would be the case for units responsible for functions common to every trial (e.g. related to the fixation unit; regard stimulus direction). The second LSTM layer presents more significance for the desired output as expected, although most node variance seems to be accounted for intra-task purposes as it decreases to close to 0 when only variance between tasks is accounted for.

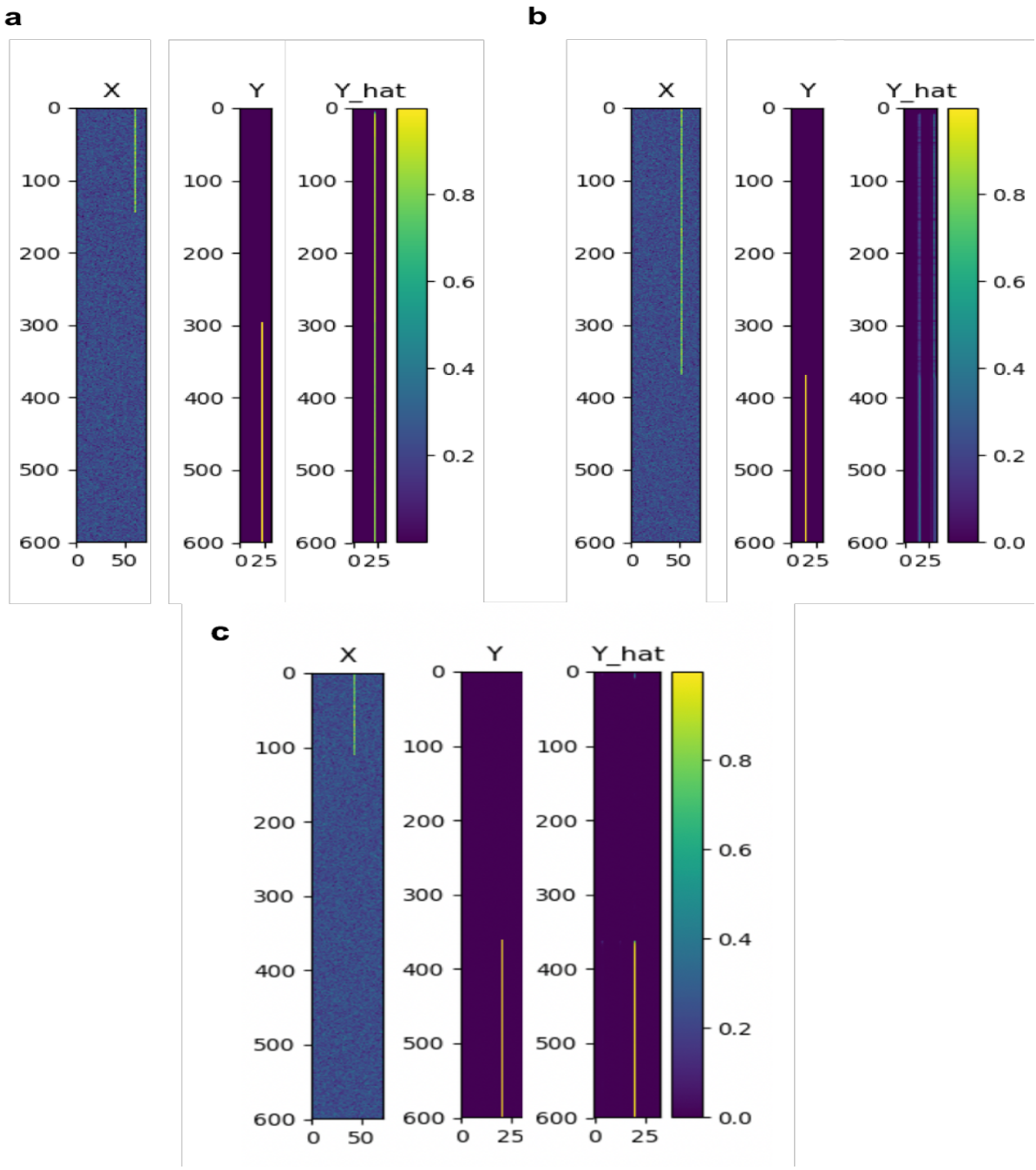


Figure III-7 – Representeation of the data set of the model's response for three trials. X represents the input signal past onto the network, Y is the expected output signal from the dataset used to calculate the loss function and Y_hat is the model's true output.

III.4 Exploring similarities with brain activations

We considered whether and where the neural network's approach for modelling tasks would be similar to how the human brain processes similar types of tasks. To do this, it was first analysed how the model's activations correlated for different tasks. This similarity matrix between tasks could then be compared to an equivalent similarity matrix created from neural data taken from a large meta-analytic database of brain activation patterns (the Neurosynth database).

We focused on the model's activation for each trial measured at 10 time-steps after when the model was expected to react to the specific task. Moreover, only the 128 units of the second layer were considered; this layer was chosen because, as the previous results demonstrated, most of the model's variability occurred in its activations. Model activation patterns were averaged for every trial of a given task and the 6 task-averaged sets were then correlated between themselves. Pearson and non-parametric Spearman-rank correlation analyses were applied to assess the similarity between tasks for neural and model data, and to compare neural with model similarity matrices. Pearson and Spearman analyses in general produced similar results. To approximately correct for non-normality in the distribution of correlation coefficients, the Fisher z-Transformation was applied to all correlation results before they were used in higher level statistical analyses. Figure III-8 presents the intercorrelation between tasks for the model activation patterns.

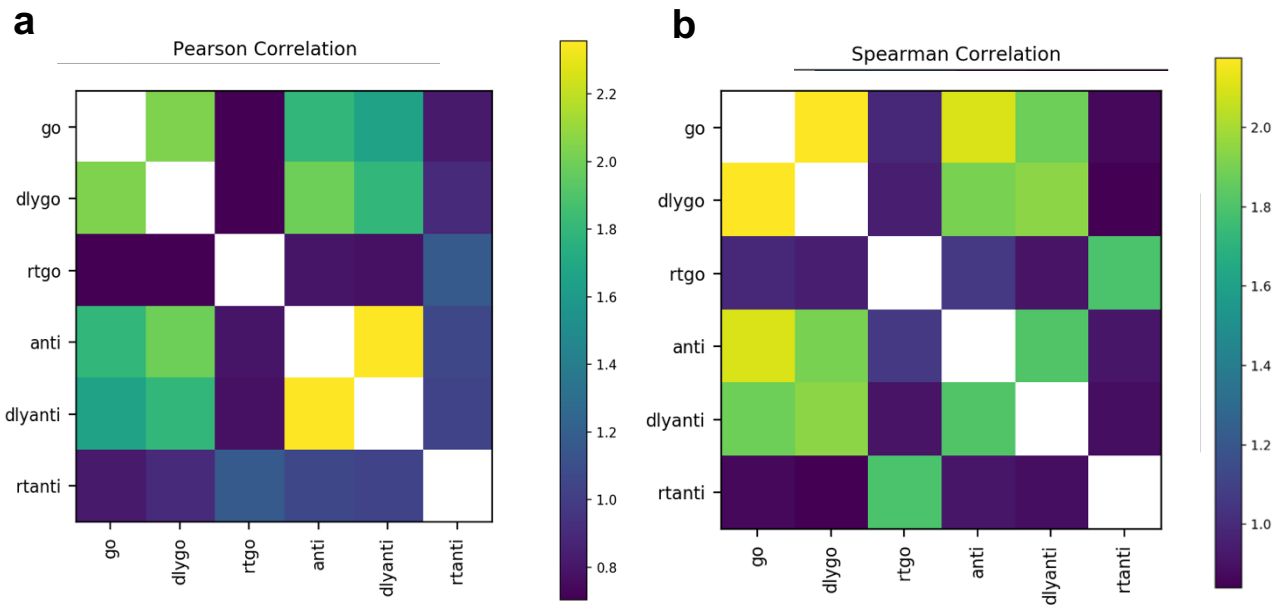


Figure III-8 - Similarity matrix for the Model tasks using Pearson Correlation (a) and Spearman Correlation (b). Colours code correlation intensity.

The most salient result is that Reaction Time tasks were very different from every other. This is apparent as the very low correlation values, even between the pair of reaction time tasks that were expected to be similar. As expected, every other task presents high correlations, especially with its counterparts.

In order to compare the model activation patterns with neural activation patterns from meta-analyses, we needed to select cognitive terms from the Neurosynth database (as detailed in the methods section II.6). There are clear differences between both the model and the neural similarity matrices - Figure III-9. While previously the Reaction Time tasks presented the lower degree of correlations, for the Neurosynth terms it is the pair Go and Anti tasks that report a lower correlation with the other tasks.

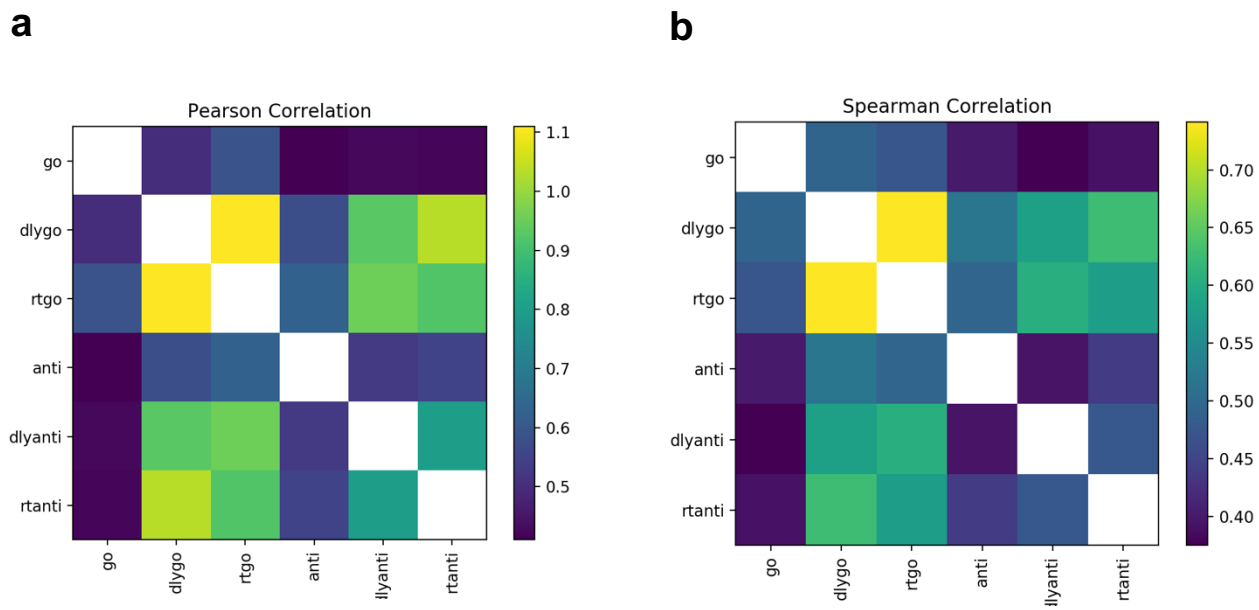


Figure III-9 - Similarity matrix for the Neurosynth meta-analytical maps divided into tasks using fisher-transformed Pearson Correlation (a) and Spearman Correlation (b). Colours code correlation intensity.

When considering the brain as a whole, activations for each defined task do not appear to correlate well to their model's task counterpart. Despite this, there may be smaller regions or brain networks that better correspond to the model activation patterns. Therefore, a searchlight method, that scanned iteratively through the whole brain map, was applied in order to find regions where both similarity matrices presented high correlation. Figure III-9 presents the results for the searchlight method analysing cubes with 20mm of length and its correlation with the model's similarity matrix and with a stride of 10mm iteratively running through the whole brain.

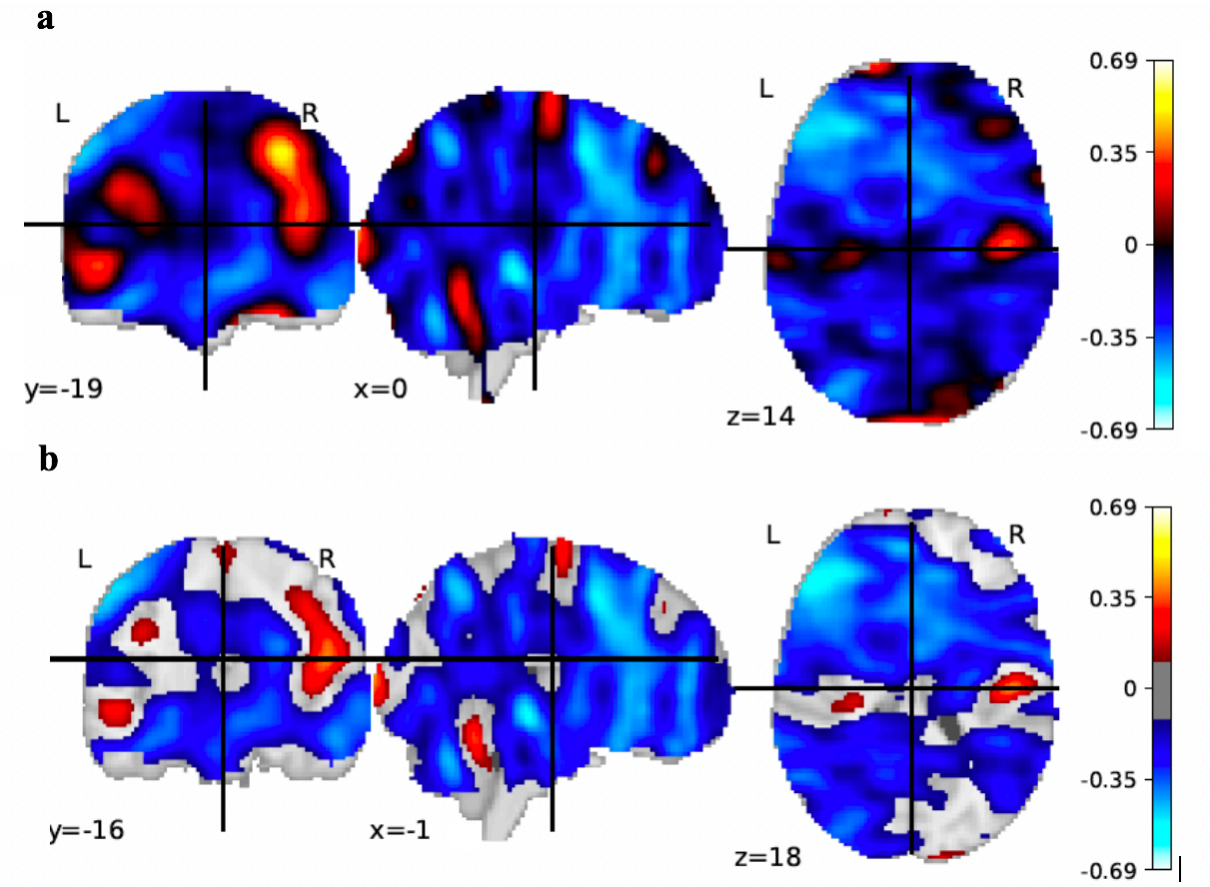


Figure III-10 - Regional correlations using the Searchlight method. a: coronal, sagittal and axial plane for the unthresholded correlation. b: coronal, sagittal and axial plane for the fisher-transformed correlations with a threshold automatically set by the software to facilitate the analysis.

The upper images present the correlation study with no threshold applied and the lower images display the same study ignoring correlations between 0.2 and -0.2 in order to focus on the highest correlations. A region in the cerebellum displays high positive correlation in this analysis. Positive correlations extend into the posterior insula region achieving a correlation of r = 0.4 after the Fisher Transform. Negative correlations are observed in frontal regions including dorsal anterior cingulate and most strongly in medial parietal cortex. These values have yet to be corrected for multiple comparisons.

In addition, large scale brain networks were also considered. To this end, an existing resting state of frequently occurring brain networks [64] were utilised to mask the brain activation regions required for each of the correlation analysis. These are subdivided into two sets that cover differently the brain map, a set of 7 networks and a set of 17 networks. Both sets were utilised. For each of the 7 resting state networks, a similarity matrix was obtained for the Neurosynth maps and was subsequently correlated with the model's similarity matrix. Subsequently, to assess statistical significance (and correct for multiple comparisons) permutation was run using the "max statistic" approach [65] to obtain the p-values for the correlation between similarity matrices. The results are reported in Figure III-11 and Table III-5.

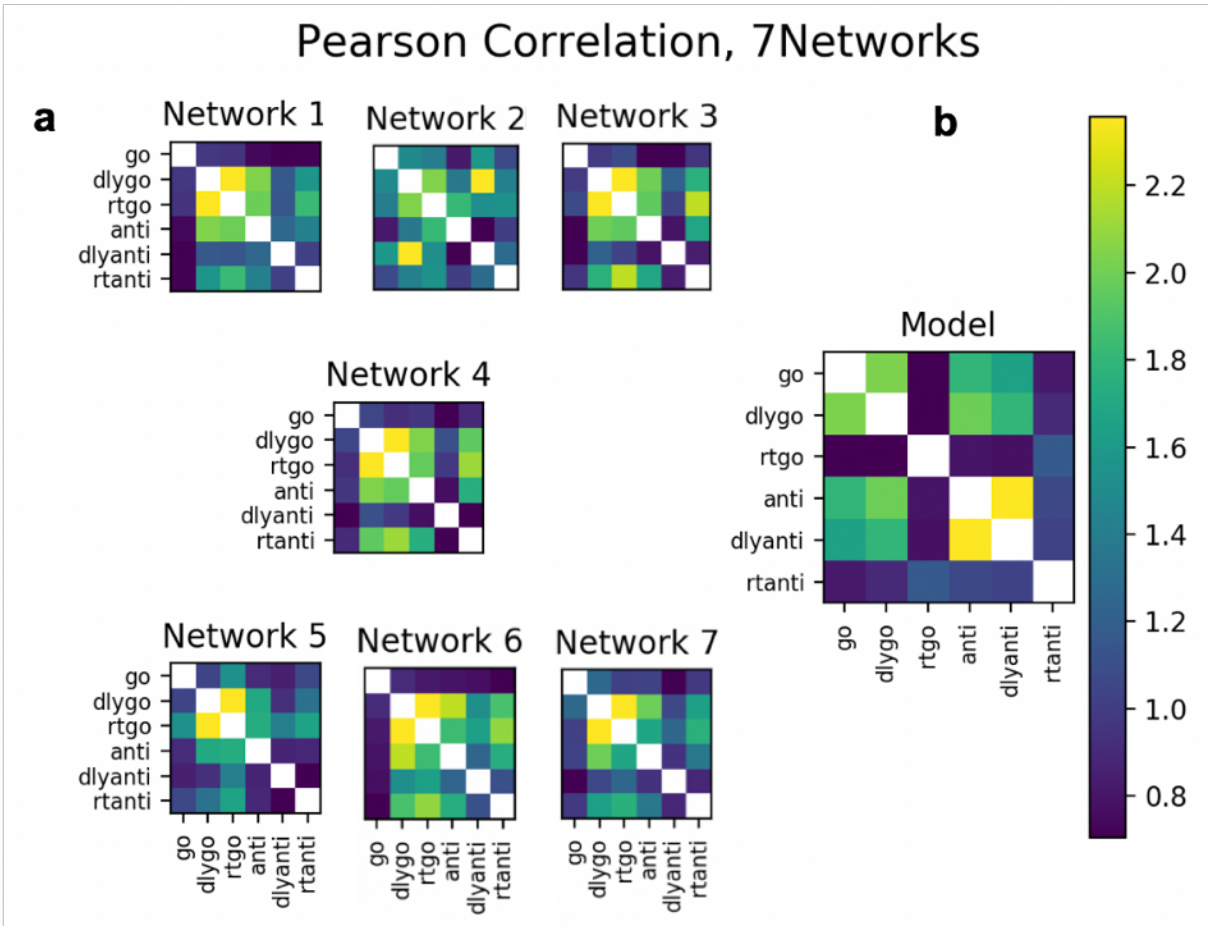


Figure III-11 - Similarity matrices for the Neurosynth meta-analytical maps on the 7 different Yeo networks using Pearson Correlation (a) and for the Model's tasks (b). Colours code correlation intensity.

Table III-5 - Correlation values between each of the 7 network and the model's similarity matrices and p-values for the given correlations.

Network	Correlation (r)	p-value
1	-0.431	0.34
2	-0.381	0.44
3	-0.504	0.22
4	-0.432	0.34
5	-0.139	0.97
6	-0.462	0.28
7	-0.446	0.31

As was the case with the whole brain analysis, there is no significant correlation between the model and the network similarity matrices. By analysing each network similarity matrix, we observe that it resembles the whole-brain similarity matrix in most cases. The low correlation of the Reaction Time tasks with other tasks in the model stands out as markedly different to the neural results. None of the similarities between model and neural data were significant. Figure III-12 and Table III-6 report the same analysis on the 17 Networks provided by Yeo.

By analysing the similarity matrices, the resemblance between the Model and the network correlations was assessed; as with the previous analysis, no significant results were found. Primarily, the Go task presents low correlation on the networks and high correlation on the Model. The opposite is apparent on the Reaction Time tasks, with high values on the networks and low values on the Model.

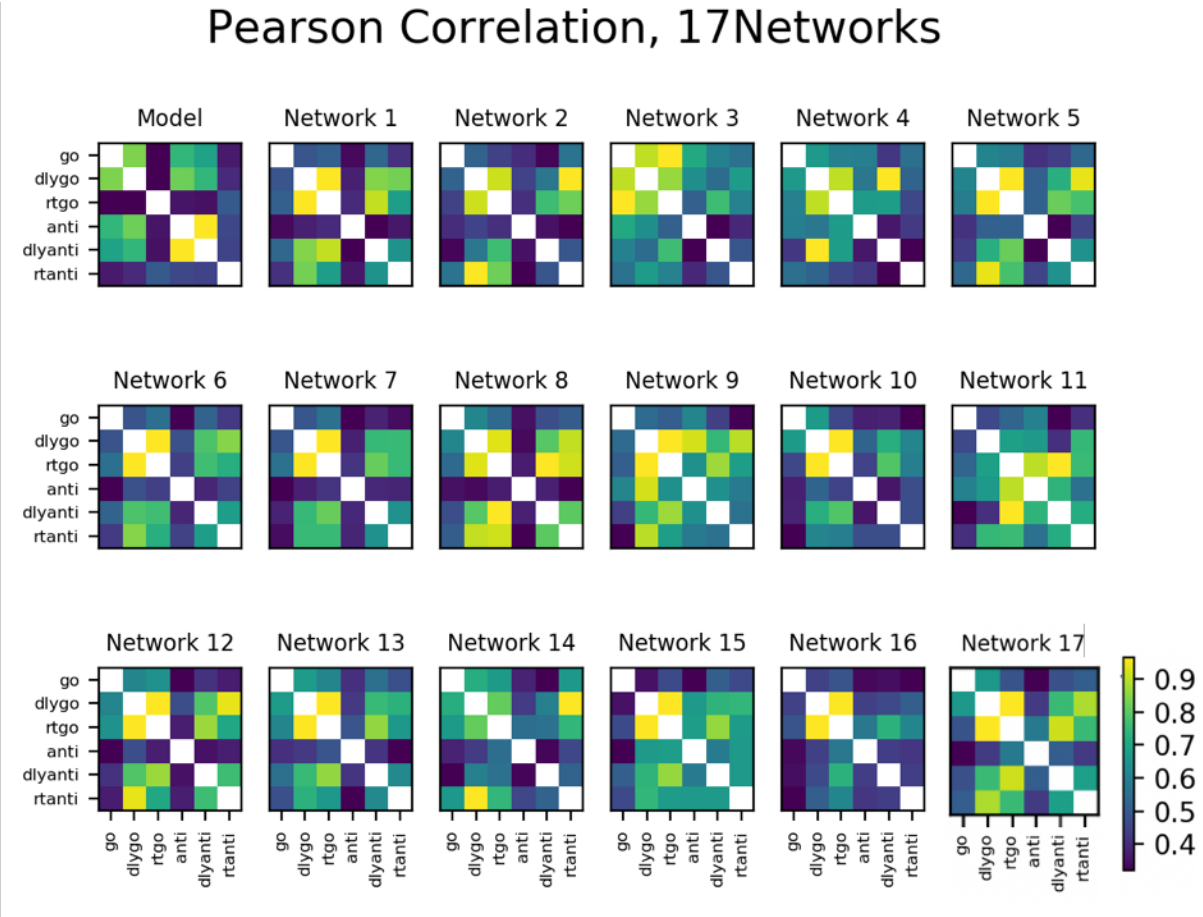


Figure III-12 - Similarity matrices for the Model's tasks and for the Neurosynth meta-analytical maps on the 17 different Yeo networks using Pearson Correlation. Colours code correlation intensity.

Table III-6 - Correlation values between each of the 17 network and the model's similarity matrices and p-values for the given correlations.

Network	Correlation (r)	p-value
1	-0.456	0.52
2	-0.491	0.96
3	-0.264	0.83
4	-0.123	0.57
5	-0.600	0.68
6	-0.479	0.24
7	-0.470	0.53
8	-0.446	0.39
9	-0.041	0.56
10	-0.238	0.44
11	-0.327	0.93
12	-0.433	0.99
13	-0.392	0.24
14	-0.596	0.47
15	-0.451	0.49
16	-0.513	0.54
17	-0.438	0.99

IV Discussion

In this Chapter the results obtained are interpreted and discussed in order to explain how they report towards the goal of understanding the model mechanisms and their relationship with the brain. Section IV.1 decodes the model structure and how it conveys information. In section IV.2 the difference between Reaction Time tasks and every other task are discussed and interpreted. Section IV.3 describes unit 28, its relationship with the model's output and how other units can be interpreted. Finally, section IV.4 describes the study of correlations between the artificial and the biological systems and assesses the limitations of the analyses.

IV.1 Interpreting the model

By comparing different model structures, this project was able to find a model architecture that successfully generalized for new examples without overfitting the training data (i.e. low bias and variance). The model accounted for a high number of trainable parameters (i.e. 537,249), mainly due to the number of variables between trials of the same task, such as stimulus direction and moment of reaction. This also allowed the model to obtain low variance without using regularization methods. Despite this, the high number of weights and units made the model analysis difficult.

The study of trial representations on both layers revealed the expected outcome – the first layer's activations displayed a clear task clustering and a gradient variation on the moment of reaction while the second layer's activations across trials were more dissimilar and did not appear to cluster significantly. The variance of unit activations across trials is also more dominant on the second layer than on the first. Both results are justified by the second layer activations depending directly from the output of the first layer with each unit connected to every unit from the previous layer.

Due to the previous outcomes and the first layer being connected directly to the model's input, it is possible to assert that the first layer allows for a clearer interpretability of each unit's function. The first layer appears to display an objective representation of the different modelled variables, with its units representing clearer roles in the overall solution (as is the case of unit 28). The second layer's interpretation is more challenging as the variables do not match the layer's representation and the different processes result from the influence from different units, which also contribute to more than one function. Therefore, the first step to fully understand the underlying computational mechanisms of how the model captures each type of task is to examine the first layer.

These mechanisms display a resemblance to the brain's functionality representation where the somatosensory periphery of the brain allows for a one-to-one mapping between regions and basic functions, but central areas are more difficult to interpret as functions and regions display many-to-many mappings and a single structure might not relate to an objective function [10].

IV.2 The curious case of the Reaction Time tasks

Out of the six cognitive tasks the model learned, there was a clear distinction between how the model processed the RT tasks when compared to other tasks. Analyses on the trial representation using the t-SNE algorithm for the first layer revealed that trials clustered according to the tasks and that the activations for trials of RT tasks greatly differed from the other tasks both on the cell state and on the layer's output. Besides this differentiation, the variance analysis concluded that RT task's activations for the first layer varied greatly across the 256 units relatively to the other tasks. This allows the assumption that the model has greater ease on solving the four other tasks, as the trials in a given task do not change their activations by a large margin to successfully account for the intra-task variability (i.e. moment it should react, stimulus direction). This assumption is supported by the model's accuracy which is 20% lower for the RT tasks than for the other tasks.

There are ultimately two differences between the RT tasks and the other four. These are the necessity to respond from the first iteration without any interval to interpret the stimulus and the requirement to ignore the fixation input which stays on for the whole trial and, contrarily to other tasks, does not convey the moment to react. Both differences should account for the change in representation. This work established the role of unit 28 for the trailing of the fixation input. Therefore, it can be assumed that this unit and other units that contribute to the execution of the fixation input mostly need to be ignored, resulting in an essentially different activation map.

Despite the higher variability present in the second layer of the model for every task, both RT tasks present high variance for a larger number of units but lower variance in some specific units (e.g. unit 20). These results are senseful as the first layer, which depicts higher variance across units, is densely connected to the second layer resulting in variance transmission. It is possible that the units with lower variance are a result of the model learning to deactivate units that are responsible for functions not involved in the RT tasks.

From the analyses of the similarity matrices on task representations for both the biological and artificial model, it is possible to asses that there is a large disparity regarding the RT tasks' correlations in both models. In fact, the RT tasks' brain maps have high correlations with the brain maps of the Delay tasks, meaning that a lot of the same functions are employed. It is, then, possible to assert that the mechanisms undertaken by the artificial model are intrinsically different to the ones used by the brain in regard to the RT tasks.

In conclusion, as the RT tasks are currently modelled, they present little relation to other tasks, which does not resonate with the brain's interpretation of RT tasks. They also appear to be unreasonably complicated for the model's execution, resulting in the need to recruit a large number of units and a lower accuracy rate. A different approach to the modelling of the task should be considered in the future, which might result in a simpler model that as the other four tasks might require a lower number of units to represent successfully the appropriate outcome.

IV.3 The functionality of unit 28

Out of all the units analysed, unit 28 from the first layer stood out due to its high variance across non-RT tasks. In fact, when studying the variance across all trials and doing the same study without regarding RT tasks, unit 28 was one of the few whose variance remained virtually unchanged. It was possible to assert that its function related directly to a process not required by the RT tasks. Furthermore, the decrease in variance when studying the values averaged across each of the tasks, allows the assumption that unit 28's function regards a process similar to every task, as the averages of variances reported similar results.

In order to better assess unit 28's functionality and understand how it might influence the general outcome of the model, its weights were set to 0 (effectively "lesioning" it). Its importance for the model's success was immediately asserted, as the overall accuracy dropped to 49%. In comparison, the removal of another random unit resulted in a decrease of only 2% when compared to the original model. By assessing how the model fared in each task separately, in became clear that the performance on the RT tasks remained virtually unchanged. In contrast, every other task's performance was heavily diminished by the unit's absence with the Delay Go task decreasing its accuracy by 49.4%. The output of the model revealed the direct impact of the unit's removal – the model failed consistently to follow fixation and regarded the stimulus from the moment it was presented. As the Go and Delay tasks' success depends on the model only reacting to the stimulus when the fixation input is set to zero, the inability to maintain this capacity resulted in the drop in accuracy on the overall model.

Using the traditional cognitive ontology, unit 28 seems to control the model's inhibition system or sustained attention, allowing the model to wait for a go signal to define when it should react and responding preemptively when this system is off. Response inhibition is a well-documented cognitive process that allows an individual to control their behavioural responses and impulses. In the brain, the process is associated with the prefrontal cortex, caudate nucleus and subthalamic nucleus [48], and damages to these regions result in a lack of control similar to the behaviour observed by the unitless model. Despite the complexity of the inhibitory system in animals not being comparable to the complexity of a 2 layered artificial neural network, the understanding of this process in the artificial system will express a mechanistic solution to inhibition control and might provide hypothesis for the biological system.

Unit 28 has a clear role in the model's outcome but there are certainly other units that contribute to the following of the fixation input. Furthermore, there are other units, such as unit 93 and 160, that exhibit high variance across trials and report a function apparently associated with a process common for every trial of RT tasks. This is not the general case and most do not have transparent functional mappings, with a large number of units in the first layer reporting low to no variance. This does not mean that they have no function, as the model applies backpropagation while learning to maximise every unit's contribution to the final outcome. What is certainly the case in most of the tasks' functions is that their processes are distributed across a number of different units, each contributing to the correct execution.

This process of cooperation is common in trained ANNs and is also displayed in brain structure as the processes depend not on a single region but on networks of brain structures.

IV.4 Biological comparisons and limitations

In order to study resemblances between the model and the brain in solving cognitive tasks, this project compared correlations of task representations for both cases. In addition, three analysis were considered to study if different brain regions or networks would displayed different similarity matrices and how they would relate to the computational model. Ultimately, this analysis found significant dissimilarities between the two systems, without the obtaining clear evidence of similarities. It is clear that the main source of differentiation between both models is the representation of the RT tasks. Although the Go and Anti tasks also display a lack of correlation with other tasks in the brain, a result not verified by the artificial model. The study considering different brain networks reported similar results as the task correlations did not differ much between networks.

There are a number of explanations for a lack of representations for the computational and neural description. The most intuitive one would be to disregard the artificial model as a bad proxy for the brain. Although this is a possibility, there are some other considerations that should be taken into account.

The dataset that serves as input and output for the model to learn the cognitive tasks might not represent correctly the given cognitive processes, despite being influenced by the work of Yang et al. [2]. They tried to model the interaction between an animal and the environment when solving these tasks in a simplified manner so that the information could be provided to the artificial model. If so, a different task setup might result in a more significant correlation between models.

The lack of similarity can also be attributed to the terms chosen from the Neurosynth meta-analytical data to represent each task. Each term was chosen with prior knowledge of the brain structures associated but isolating cognitive processes is challenging and it is impossible to assert if undesirable processes are not represented in the brain maps of interest. Furthermore, some Neurosynth terms overlap with each other (e.g. ‘wm task’ and ‘working memory’ or ‘switch’ and ‘switching’), a problem that was addressed by accounting for more than one term for some cognitive tasks. Finally, meta-analytical studies can reflect some biases that skew true results such as publication bias (i.e. the success of publishing is influenced by the significance of the results obtained).

There might also be a lack of representation similarity due to the low number of cognitive tasks being solved by the artificial model. The animal brain is able to solve a number of cognitive tasks and so, the brain is optimized to efficiently solve each of them. By using a larger number of cognitive tasks, the model will be forced to optimize its parameters for different processes, resulting in compositionality of tasks and, possibly, in higher similarity to equivalent brain mappings.

Furthermore, although the model is supposed to represent a simplified version of the brain, there are significant structural dissimilarities that can be addressed. The layers in this model are densely connected (i.e. each unit connects to every unit of the following layer), which is not verified in the brain.

The supervised learning mechanism and the backpropagation algorithm are not biologically plausible and the artificial model does not account for neural mechanisms such as neuromodulation or spiking. It is also important to note that the model does not intend to mimic directly neurons or networks, but only the mechanisms used by the brain despite of the level of analysis. Finally, although backpropagation does not appear biologically logical, the mechanisms the brain uses to learn are still not fully understood and it is impossible to tell if the backpropagation algorithm does not present the same results as the brain mechanism.

V Closure

V.1 Conclusion

This project's main contribution is demonstrating an ANN architecture capable of effectively solving six cognitive tasks with an accuracy rate of 93% developed for this work. There is no literature available on using LSTM units to simulate multiple cognitive tasks as this project does, allowing it to deal with longer dependencies, retaining information for a larger number of iterations, useful features when solving cognitive tasks as most require the engagement of working memory for its successful execution in biological systems. On that note, LSTM units characteristically store a memory value that is recurrently passed along iterations called cell state. This state is seen as a proxy of memory in biological systems, so its analysis allows for interesting conclusions.

A second contribution concerns the analysis between the artificial model activations and the brain region activations. As far as it was asserted, there is no literature studying the parallelism between both task representations as of yet. Therefore, as the pipeline structured for this project allows a comparison between two different mediums that relate to the same cognitive paradigms, it arguably serves as an example for further studies regarding similar associations.

In a more general manner, this work operates as a first step towards mapping the artificial task representation to its biological counterpart and consequently, mapping the cognitive paradigms to the brain networks activations and their implied mechanisms. It is through understanding the mechanisms behind cognitive processes that an effective cognitive ontology can be built without disregarding any level of brain studies. If, with further work, the dissimilarities in representations of different tasks between the biological and artificial systems are still present, the artificial model will, nonetheless, convey a successful mechanistic solution for the cognitive processes, a concrete step towards demystifying intelligence.

V.2 Future Work

This work opens way for further analysis and new projects with the goal of better understanding the artificial model and possibly, the brain. There are several model optimizations that can be considered when the model analysis is taken into account.

The RT tasks account for a disproportionate amount of the model's variance and so it should be considered if its removal or substitution by different tasks in the training set would require a smaller number of units to be successfully trained. Furthermore, future work could investigate how units would behave mechanistically, by applying an analysis similar to the one reported for unit 28. Another approach could be to delineate a simplified version of the RT tasks which would prove easier for the model to learn.

As unit 28 was analysed attentively, there are a number of other units that can have a significant and transparent role for the model's success. The same method of analysis can be undertaken for

different units and, ultimately, even the units whose functions are more abstract or result in a small contribution towards a given process, should be analysed and understood in order to fully map out the model's mechanistic solution for each of the cognitive tasks. A more targeted weight analysis with focus on the first seven input units (i.e. rule inputs and fixation) may allow further understanding on how each unit prioritizes its processes.

Another variation of this project that is interesting to consider is the application of dropout [66] to the units during training (i.e. randomly choosing a fixed percentage of units to become inactive for each iteration of the training set). Neural network units tend to collaborate towards the desired outcome, as backpropagation impact every unit's weight simultaneously. This results in units which do not depict a specific process for general outcome, but that contribute minimally to every process. By using dropout, units will be forced to rely less on each other and will develop specific functions useful to solving the tasks. This specialization would draw parallels with the brain as it is understood that when solving a specific task, not every region of the brain is engaged.

In order to avoid difficulties in drawing datasets that would correctly simulate correctly cognitive tasks and difficulties in choosing Neurosynth terms that would correctly map out the desired brain structures, a more reliable solution would be to collect fMRI data of subjects solving the exact same dataset that serves as input for the model. Psychlab [5] software facilitates this process as it creates an open-source virtual environment that can also be fed into a neural network, successfully replicating the same environment provided for both the user and the machine. Additionally, the model can be adapted towards a more biologically plausible learning method with reinforcement learning, where the model only knows if it has successfully executed the trial when it is finished. Currently, with supervised learning, the model's weights are updated every time-step depending on following the correct solution or not.

Finally, a larger number of cognitive tasks should be considered and trained. The source code that generates the dataset was built to allow flexibility for the user to implement different tasks allowing for desirable reaction manipulation and two different modalities that are paramount for multisensory or decision-making tasks. The larger number of tasks learned will force the model to efficiently allocate units function, resulting in a mechanism that competently solves cognitive processes.

VI References

[1] Ziskind, Bernard and Halioua, Bruno, "La conception du coeur dans l'Égypte ancienne," *Med Sci*, vol. 20, no. 3, pp. 367–373, 2004.

[2] G. R. Yang, H. F. Song, W. T. Newsome, and X.-J. Wang, "Clustering and compositionality of task representations in a neural network trained to perform many cognitive tasks," *bioRxiv*, p. 183632, 2017.

[3] I. Eisenberg *et al.*, "Uncovering mental structure through data-driven ontology discovery," *PsyArXiv*, p. n. pag., 2018.

[4] D. A. Weiskopf, "Integrative Modeling and the Role of Neural Constraints," *Philos. Sci.*, vol. 83, no. 5, pp. 674–685, 2016.

[5] J. Z. Leibo *et al.*, "Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents," *arXiv*, pp. 1–28, 2018.

[6] R. Caruana, "Multitask Learning," *Learn. to Learn*, vol. 75, pp. 41–75, 1997.

[7] T. Miconi, "Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks," *Elife*, vol. 6, pp. 1–22, 2017.

[8] K. J. Friston, "Modalities, modes, and models in functional neuroimaging," *Science*. p. 326, 399–403., 2009.

[9] B. Mišić and O. Sporns, "From regions to connections and networks: New bridges between brain and behavior," *Curr. Opin. Neurobiol.*, vol. 40, pp. 1–7, 2016.

[10] L. Pessoa, "Understanding Brain Networks," *Phys. Life Rev.*, vol. 11, no. 3, pp. 400–435, 2015.

[11] B. M. J. Kane, A. R. A. Conway, T. K. Miura, G. J. H. Colflesh, and G. J. H. Working, "Working Memory, Attention Control, and the N-Back Task: A Question of Construct Validity," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 33, pp. 615–622, 2007.

[12] A. Hampshire, R. R. Highfield, B. L. Parkin, and A. M. Owen, "Fractionating Human Intelligence," *Neuron*, vol. 76, no. 6, pp. 1225–1237, 2012.

[13] R. A. Poldrack and T. Yarkoni, "From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure," 2016.

[14] T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager, "Large-scale automated synthesis of human functional neuroimaging data," *Nat. Methods*, vol. 8, no. 8, pp. 665–670, 2011.

[15] T. Yarkoni, R. Poldrack, T. Nichols, D. Van Essen, and T. Wager, "NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data," in *Frontiers in*

Neuroinformatics Conference Abstract: 4th INCF Congress of Neuroinformatics, 2011.

[16] J. D. Scargle, "Publication Bias (The 'File-Drawer Problem') in Scientific Inference," *J. Sci. Explor.*, vol. 14, no. 1, pp. 91–106, 1999.

[17] R. A. Poldrack, "Can cognitive processes be inferred from neuroimaging data?," *Trends in Cognitive Sciences*. p. no pag., 2006.

[18] J. Busse *et al.*, "Actually, what does 'ontology' mean?: A term coined by philosophy in the light of different scientific disciplines," *J. Comput. Inf. Technol.*, vol. 23, no. 1, pp. 29–41, 2015.

[19] J. B. L. Bard and S. Y. Rhee, "Ontologies in biology: design, applications and future challenges," *Nat. Rev. Genet.*, vol. 5, no. 3, pp. 213–222, 2004.

[20] C. J. Price and K. J. Friston, "Functional ontologies for cognition: The systematic definition of structure and function," *Cogn. Neuropsychol.*, vol. 22, no. 3–4, pp. 262–275, May 2005.

[21] M. Bada *et al.*, "A short study on the success of the Gene Ontology," *Web Semant.*, vol. 1, no. 2, pp. 235–240, 2004.

[22] A. Janssen, C. Klein, and M. Slors, "What is a cognitive ontology, anyway?," *Philos. Explor.*, vol. 20, no. 2, pp. 123–128, May 2017.

[23] J. B. McCaffrey and E. Machery, "The reification objection to bottom-up cognitive ontology revision," *Behav. Brain Sci.*, vol. 39, p. e125, Jun. 2016.

[24] D. Gardner *et al.*, "The neuroscience information framework: A data and knowledge environment for neuroscience," *Neuroinformatics*, vol. 6, pp. 149–160, 2008.

[25] G. Frishkoff, R. Frank, and P. Lependu, "Ontology-based Analysis of Event-Related Potentials," *ICBO*, p. n. pag., 2011.

[26] Y. Schwartz, B. Thirion, and G. Varoquaux, "Mapping cognitive ontologies to and from the brain," *arXiv*, pp. 1–9, 2013.

[27] R. M. Bilder *et al.*, "Cognitive ontologies for neuropsychiatric phenomics research," *Cogn. Neuropsychiatry*, vol. 14, no. 4–5, pp. 419–450, Jul. 2009.

[28] J. A. Turner and A. R. Laird, "The cognitive paradigm ontology: design and application.," *Neuroinformatics*, vol. 10, no. 1, pp. 57–66, Jan. 2012.

[29] T. Yarkoni and J. Westfall, "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning," *Perspect. Psychol. Sci.*, vol. 12, no. 6, pp. 1100–1122, Nov. 2017.

[30] R. A. Poldrack *et al.*, "The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience," *Front. Neuroinform.*, vol. 5, p. 17, Sep. 2011.

[31] B. T. Thomas Yeo *et al.*, "The organization of the human cerebral cortex estimated by intrinsic

functional connectivity," *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, Sep. 2011.

[32] C. J. Price and K. J. Friston, "Functional ontologies for cognition: The systematic definition of structure and function," *Cogn. Neuropsychol.*, vol. 22, no. 3–4, pp. 262–275, May 2005.

[33] J. R. Anderson, J. M. Fincham, Y. Qin, and A. Stocco, "A central circuit of the mind," *Trends Cogn. Sci.*, vol. 12, no. 4, pp. 136–143, Apr. 2008.

[34] A. Citri and R. C. Malenka, "Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms," *Neuropsychopharmacology*, vol. 33, p. 18, Aug. 2007.

[35] W. S. McCulloch and W. Pitts, "A Logical Calculus of the Idea Immanent in Nervous Activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.

[36] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," *J. Artif. Intell. Res.*, vol. 57, pp. 1–75, 2015.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, p. 533, Oct. 1986.

[38] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv*, pp. 1–14, 2016.

[39] S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, p. 529, Feb. 2015.

[41] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[42] O. Barak, D. Sussillo, R. Romo, M. Tsodyks, and L. F. Abbott, "From fixed points to chaos: Three models of delayed discrimination," *Prog. Neurobiol.*, vol. 103, pp. 214–222, 2013.

[43] H. F. Song, G. R. Yang, and X. J. Wang, "Reward-based training of recurrent neural networks for cognitive and value-based tasks," *Elife*, vol. 6, pp. 1–24, 2017.

[44] A. Marblestone, G. Wayne, and K. Kording, "Towards an integration of deep learning and neuroscience," *Front. Comput. Neurosci.*, vol. 10, no. September, pp. 1–41, 2016.

[45] S. Pinker, "How the mind works," *Ann. N.Y. Acad. Sci.*, p. n.pag., 1999.

[46] K. Friston, "The free-energy principle: a unified brain theory?," *Nat. Rev. Neurosci.*, vol. 11, p. 127, Jan. 2010.

[47] H. Hong, D. L. K. Yamins, N. J. Majaj, and J. J. DiCarlo, "Explicit information for category-orthogonal object properties increases along the ventral stream," *Nat. Neurosci.*, vol. 19, p. 613, Feb. 2016.

[48] F. Carnevale, V. deLafuente, R. Romo, O. Barak, and N. Parga, "Dynamic Control of Response Criterion

in Premotor Cortex during Perceptual Detection under Temporal Uncertainty," *Neuron*, vol. 86, no. 4, pp. 1067–1077, 2015.

[49] K. Rajan, C. D. D. Harvey, and D. W. W. Tank, "Recurrent Network Models of Sequence Generation and Memory," *Neuron*, vol. 90, no. 1, pp. 128–142, 2016.

[50] D. Sussillo, S. D. Stavisky, J. C. Kao, S. I. Ryu, and K. V Shenoy, "Making brain–machine interfaces robust to future neural variability," *Nat. Commun.*, vol. 7, p. 13749, Dec. 2016.

[51] C. J. Cueva and X.-X. Wei, "Emergence of grid-like representations by training recurrent neural networks to perform spatial localization," in *ICLR 2018 Conference*, 2018, pp. 1–19.

[52] C. Hong, "Training Spiking Neural Networks for Cognitive Tasks: A Versatile Framework Compatible to Various Temporal Codes," *arXiv*, p. n.pag., 2017.

[53] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN," *arXiv*, no. 1, p. n.pag., 2018.

[54] J. M. Findlay, "The Visual Stimulus for Saccadic Eye Movements in Human Observers," *Perception*, vol. 9, no. 1, pp. 7–21, Feb. 1980.

[55] D. P. Munoz and S. Everling, "Look away: the anti-saccade task and the voluntary control of eye movement," *Nat. Rev. Neurosci.*, vol. 5, p. 218, Mar. 2004.

[56] S. Funahashi and C. J. Bruce, "Mnemonic coding of visual space in the monkey ' s dorsolateral prefrontal cortex," *J. Neurophysiol.*, vol. 61, no. 2, pp. 331–349, 1989.

[57] P. Sharma and A. Singh, "Era of deep neural networks: A review," in *8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2017, pp. 1–5.

[58] D. P. Kingma and J. L. Ba, "Adam : a method for stochastic optimization," *ICLR*, pp. 1–15, 2015.

[59] C. M. Bishop, *Pattern Recognition and Machine Learning*. 2006.

[60] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Pmlr*, vol. 9, pp. 249–256, 2010.

[61] M. Henaff, A. Szlam, and Y. LeCun, "Recurrent Orthogonal Networks and Long-Memory Tasks," *arXiv*, p. n.pag., 2016.

[62] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, 2000.

[63] L. J. P. Van Der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

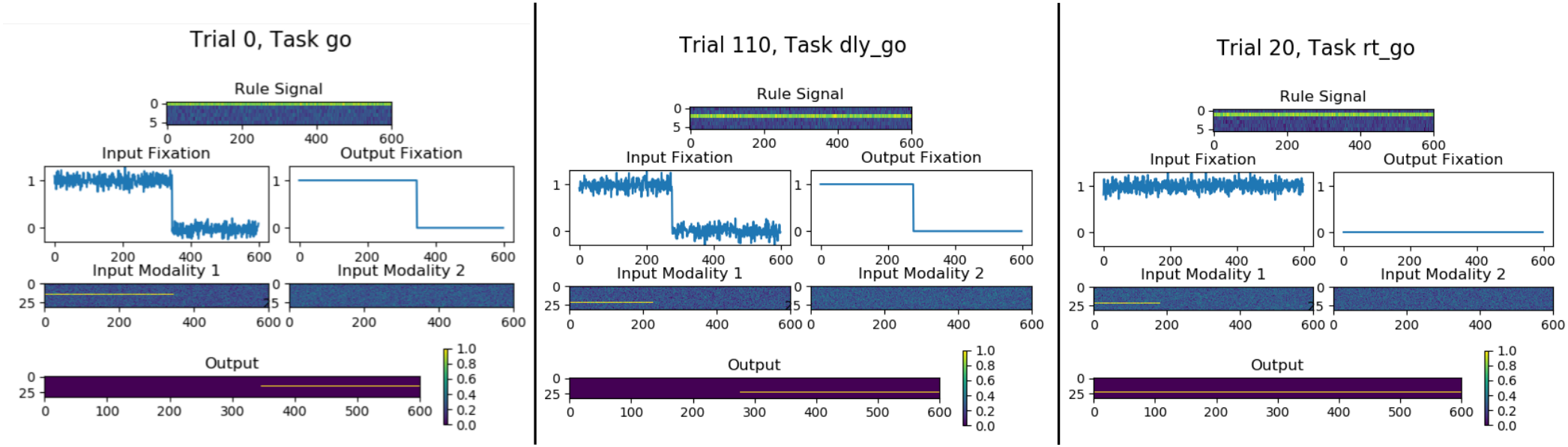
[64] B. T. T. Yeo *et al.*, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity.," *J. Neurophysiol.*, 2011.

[65] T. Nichols and A. Holmes, “Nonparametric Permutation Tests for Functional Neuroimaging,” in *Human Brain Function: Second Edition*, 2003, pp. 887–910.

[66] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

VII Appendix

VII.1 Appendix A - Representation of the 6 Cognitive Tasks



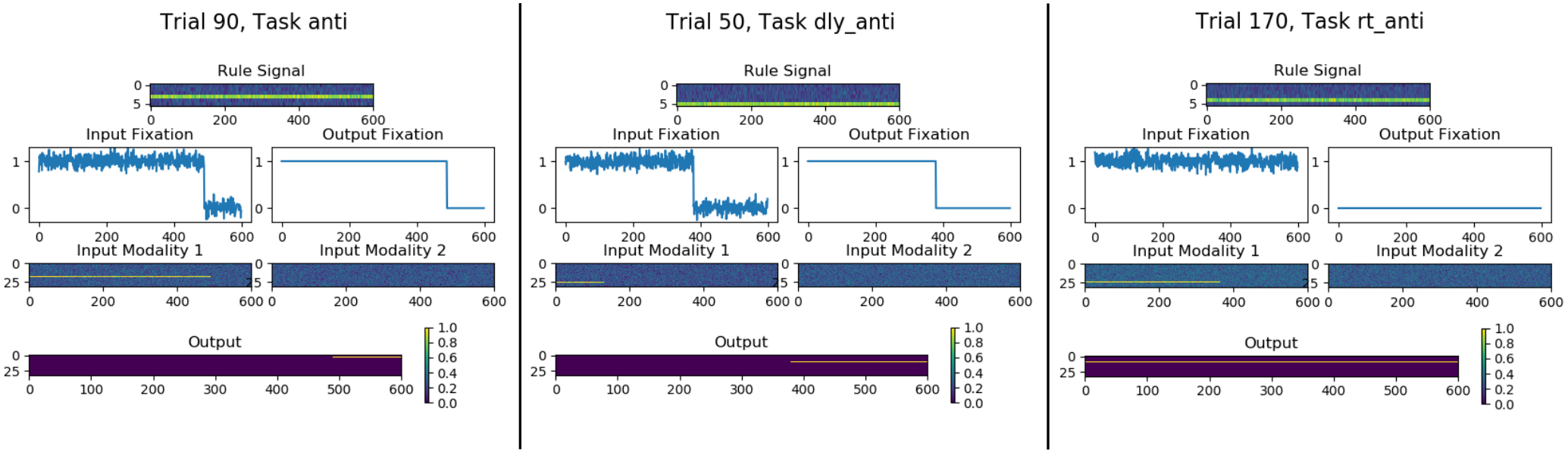


Figure VII-1 - Six examples of trials that represent the 6 cognitive tasks employed by the model.

VII.2 Appendix B - Study of Variation and Correlation for the Moment of 50 timesteps after Moment of Reaction

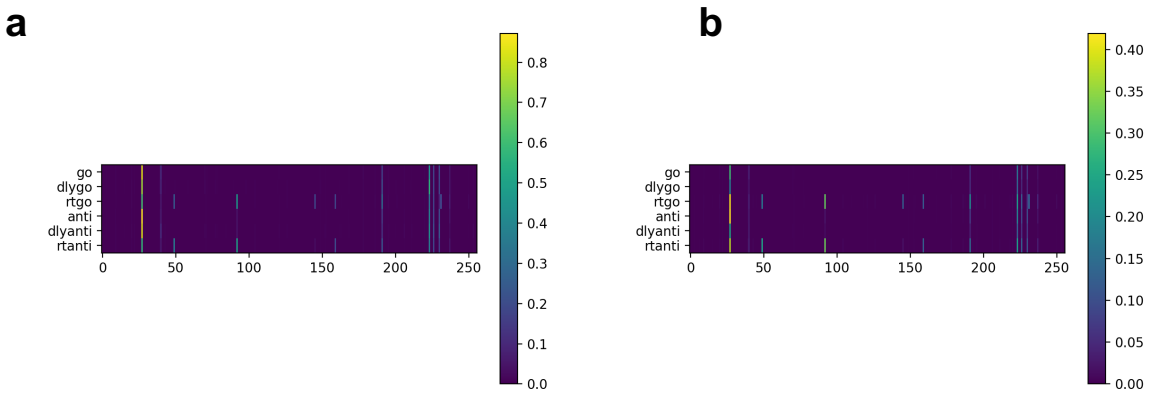


Figure VII-3 - Variance of each first-layer unit across each task (a) and same results normalized across each task (b). Colour bars represents the variance intensity. Y-axis refers to the model's tasks.

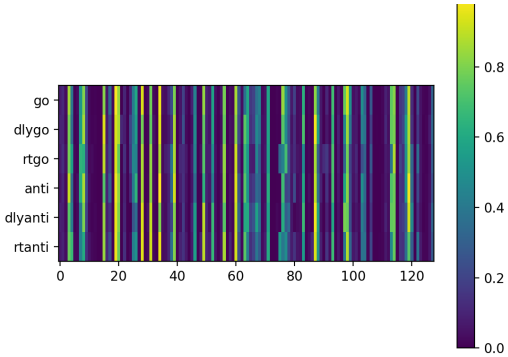


Figure VII-2 - Variance of each second-layer unit across each task. Colour bar represents the variance intensity. Y-axis refers to the model's tasks (in descending order: Go task, Delay Go task, Reaction Time Go task, Anti task, Delay Anti task, Reaction Time Anti task.

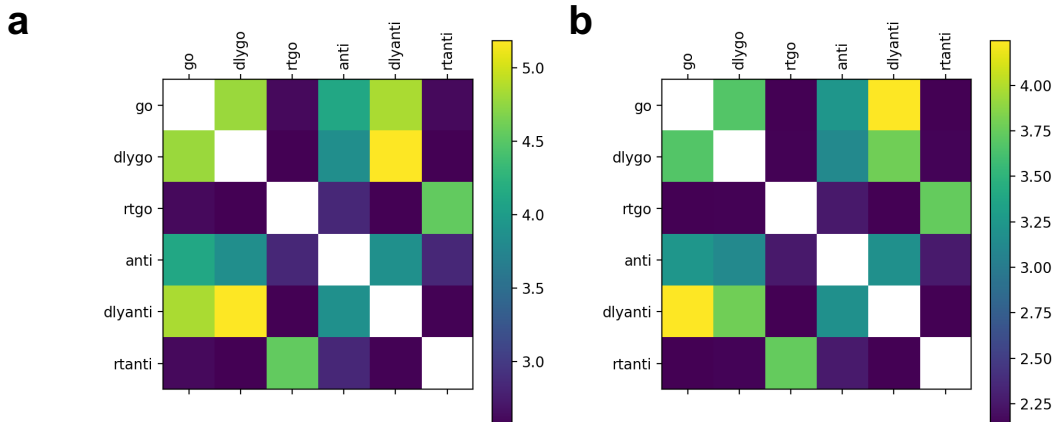
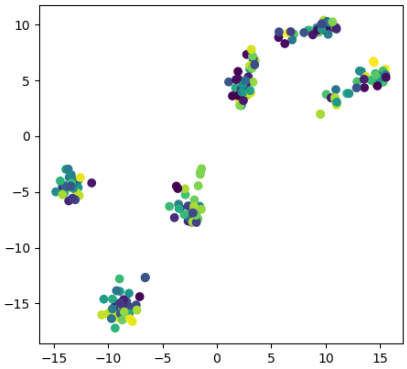


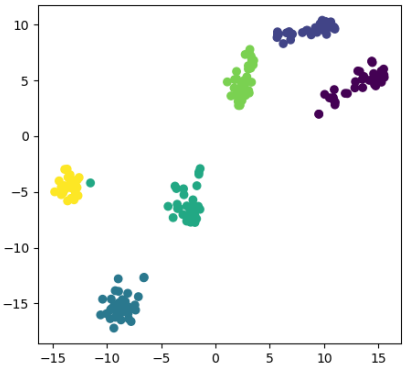
Figure VII-4 - Simlarity matrix for the Model tasks using Pearson Correlation (a) and Spearman Correlation (b). Colours code correlation intensity.

VII.3 Appendix C - Model's t-SNE plots for different variables

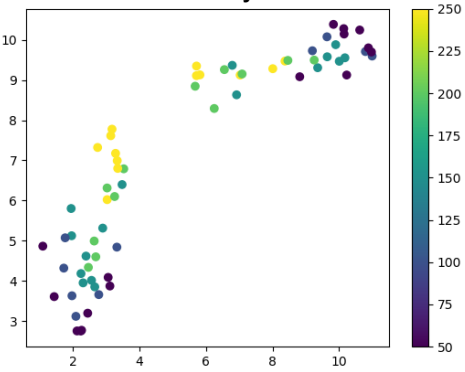
1st layer cell state
Coloured by Stimulus direction



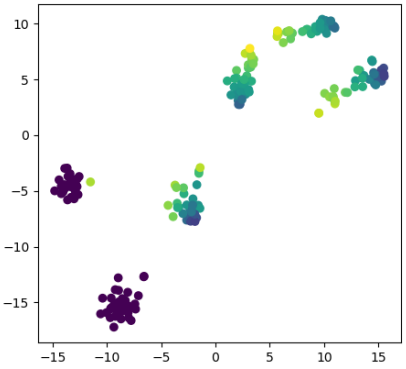
1st layer cell state
Coloured by Task



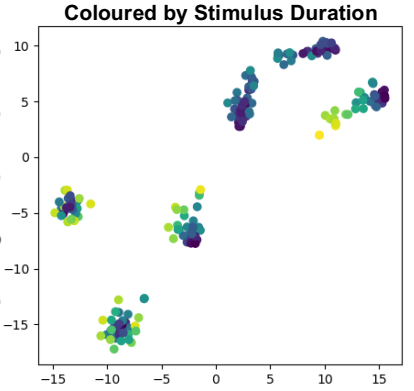
1st layer cell state
Coloured Delay Period



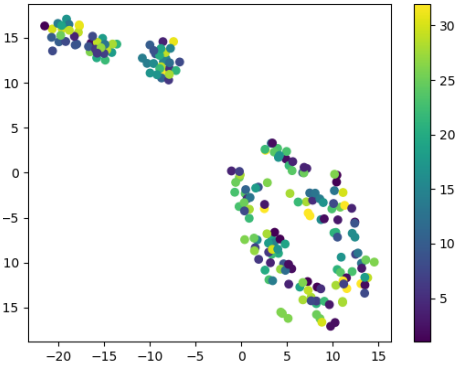
1st layer cell state
Coloured by Moment of Reaction



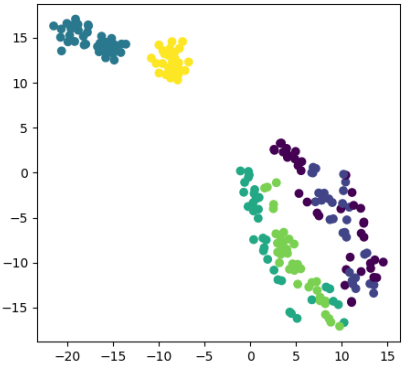
1st layer cell state
Coloured by Stimulus Duration



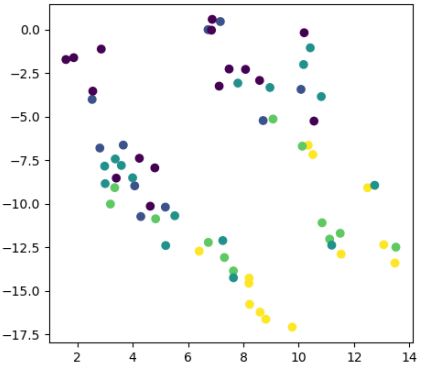
1st layer output
Coloured by Stimulus Direction



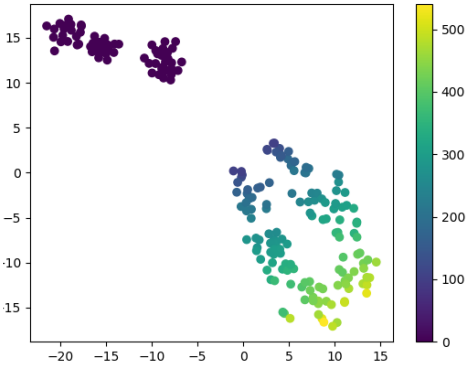
1st layer output
Coloured by Task



1st layer output
Coloured by Delay Period

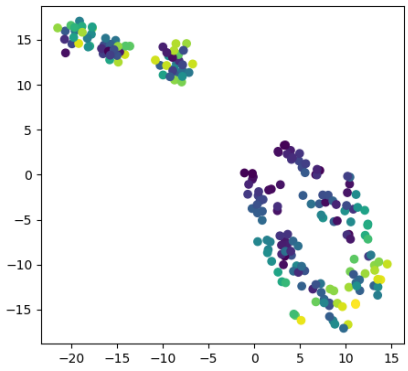


1st layer output
Coloured by Moment of Reaction



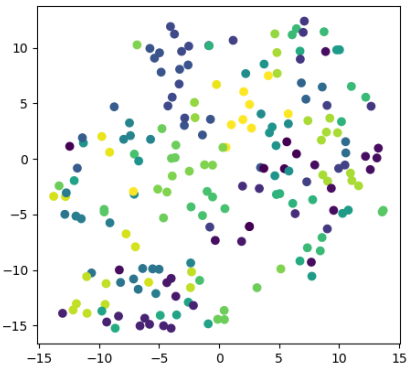
1st layer output

Coloured by Stimulus Duration



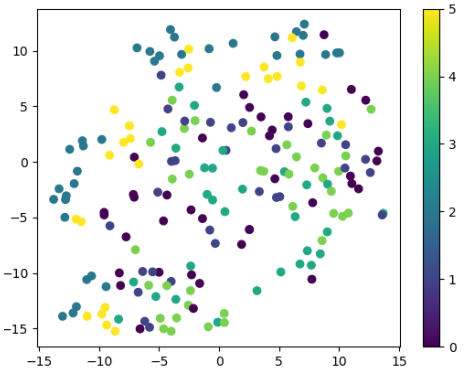
2nd layer cell state

Coloured by Stimulus Direction



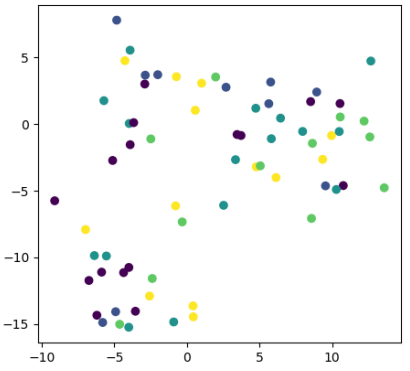
2nd layer cell state

Coloured by Task



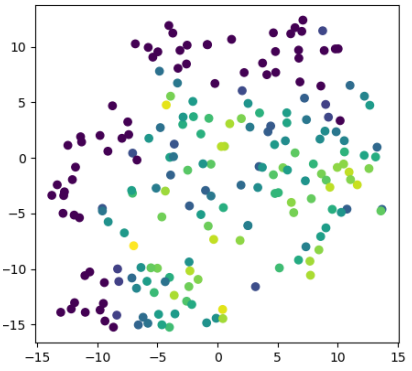
2nd layer cell state

Coloured by Delay Period



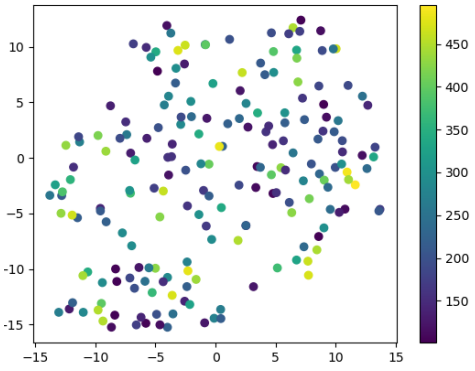
2nd layer cell state

Coloured by Moment of Reaction



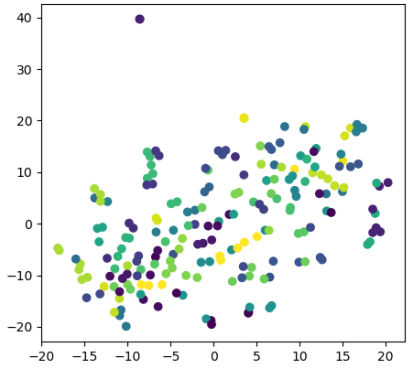
2nd layer cell state

Coloured by Stimulus Duration



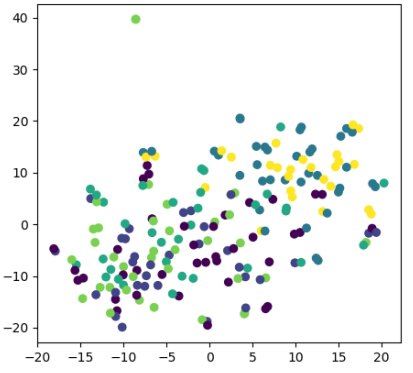
2nd layer output

Coloured by Stimulus Direction



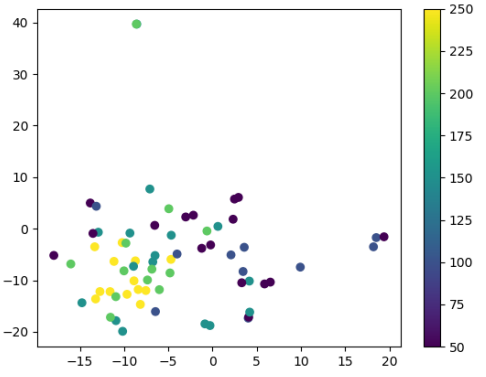
2nd layer output

Coloured by Task



2nd layer output

Coloured by Delay Period



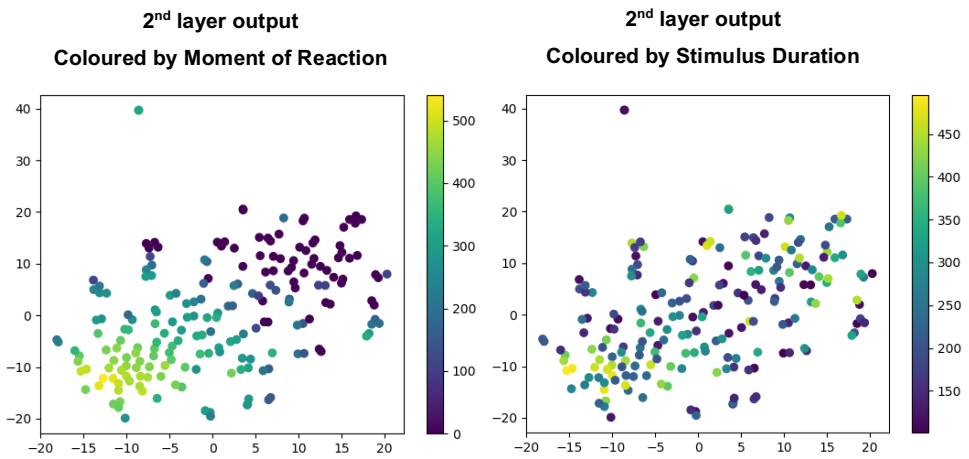


Figure VII-5 - t-SNEs plots for every variable present in each task and for every activation from the model (both cell state and hidden output). Each plot represents 200 trials as scattered points relating to Model 5's first and second layers.