# A multi-stage recurrent neural network better describes decision-related activity in dorsal premotor cortex

**Michael Kleinman (michael.kleinman@ucla.edu)**
University of California, Los Angeles
Los Angeles, CA, 90095, USA

**Chandramouli Chandrasekaran (cchandr1@bu.edu)**
Boston University
Boston, MA, 02118, USA

**Jonathan Kao (kao@seas.ucla.edu)**
University of California, Los Angeles
Los Angeles, CA, 90095, USA

**Abstract:**

**We studied how a network of recurrently connected artificial units solve a visual perceptual decision-making task. The goal of this task is to discriminate the dominant color of a central static checkerboard and report the decision with an arm movement. This task has been used to study neural activity in the dorsal premotor (PMd) cortex. When a single recurrent neural network (RNN) was trained to perform the task, the activity of artificial units in the RNN differed from neural recordings in PMd, suggesting that inputs to PMd differed from inputs to the RNN. We expanded our architecture and examined how a multi-stage RNN performed the task. In the multi-stage RNN, the last stage exhibited similarities with PMd by representing direction information but not color information. We then investigated how the representation of color and direction information evolve across RNN stages. Together, our results are a demonstration of the importance of incorporating architectural constraints into RNN models. These constraints can improve the ability of RNNs to model neural activity in association areas.**

**Keywords: decision making; neural representations; recurrent neural network; dimensionality reduction**

## Introduction

As neuroscientists gain the capability to record from large populations of neurons and measure the connectivity between the population, there is a need for simplified descriptions that link these growing observations with the observed behavior (Gao & Ganguli, 2015). Recurrent neural networks (RNNs) trained through optimization to perform a behavioral task of interest are increasingly used as models for cognitive (Mante, Sussillo, Shenoy, & Newsome, 2013; Song, Yang, & Wang, 2016), timing (Goudar & Buonomano, 2018; Laje & Buonomano, 2013) and motor tasks (Hennequin, Vogels, & Gerstner, 2014; Stroud, Porter, Hennequin, & Vogels, 2018; Sussillo, Churchland, Kaufman, & Shenoy, 2015). In many cases, the artificial units of these trained RNNs have exhibited similarities with neural recordings from brain regions thought to be performing similar tasks. These networks can then be "reverse engineered" to understand how the RNN's solution can be explained in terms of the representation of the internal units and the connectivity profile – the same goal as experimental neuroscience.

To date, the dominant approach is to use single trained RNNs as candidate models for a local, recurrently connected brain region. In line with this approach, we studied the properties of a single RNN trained to perform a perceptual decision-making task (Chandrasekaran, Peixoto, Newsome, & Shenoy, 2017; Wang et al., 2019). In this task, the monkey discriminated the dominant color of a central static checkerboard and reported his decision by reaching to a target whose color matched the dominant color of the checkerboard (Fig. 1a). Since the target configuration was randomized across trials, this task separates the action decision from the color decision. We wished to develop a model capable of replicating the responses in dorsal premotor cortex, an important decision-related brain region, during this task. In particular, we wished to replicate the observation that PMd neurons covary with the action decision, but not with the decision about the dominant color of the checkerboard (Chandrasekaran et al., 2017; Wang et al., 2019).

After training the RNN to perform an analogous task, we found that the activity of the artificial units of our single RNN differed from recordings in PMd. In particular, units in the single RNN showed selectivity for both color and direction, suggesting that inputs to the dorsal premotor cortex differ from the inputs to the RNN. In the brain, decision-making arises from distributed interconnected circuits. Inputs into PMd are likely to



This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

emerge from previous regions such as dorsolateral prefrontal cortex that transform color and target identity information.

To more faithfully model this distributed network, we expanded our architecture into a multi-stage recurrent neural network (Michaels, Schaffelhofer, Agudelo-Toro, & Scherberger, 2017). We chose the connectivity of this network using published anatomical data regarding feedforward and feedback connectivity between PMd and frontal areas (Markov et al., 2014). With this modification, we found that the activity of artificial units of the last stage of processing more closely resembled activity in dorsal premotor cortex compared to the single RNN. Consistent with neural data recorded in PMd, the last stage of our multi-stage RNN model retained only direction related information and did not reflect signals related to the color of the checkerboard.

## Model description and training

Our RNN model was the standard rate model for neural networks (Sompolinsky, Crisanti, & Sommers, 1988), shown below using the rectified nonlinearity $f(x) = relu(x)$. We refer to x as the hidden state. The output y was defined as a linear readout of the rates *f(x)*.

$$\tau\dot{x} = -x + W_{rec}f(x) + W_{in}u + b \quad (1)$$

$$y = W_{out}f(x) \quad (2)$$

We generated synthetic data as illustrated in Fig. 1a. The inputs were a 4-dimensional vector, with the inputs denoting (1) the color of the right target, (2) the color of the left target, (3) the proportion of red squares, and (4) the proportion of green squares. In line with previous literature, we rescaled the proportion of red and green squares to compute the signed coherence, calculated as $100 \times (R - G)/(R + G)$, with R (G) denoting the number of red (green) squares. A red (green) target was denoted by a value of -1 (+1). Thus, the inputs describe the target configuration and checkerboard color. We added zero-mean Gaussian noise (SD = 0.1) to the checkerboard color inputs to model the noisy and time-varying perception of the static checkerboard. The output of the network was two decision variables accumulating evidence for a left and right reach. To train the network, we defined the desired decision variable output as 1 for the desired reach, and 0 for the undesired reach, after a 200ms delay from the presentation of the checkerboard. We optimized the parameters b, $W_{rec}$, $W_{in}$, and $W_{out}$ using backpropagation through time with Adam optimization (Pascanu, Mikolov, & Bengio, 2013) to minimize the mean square error between the desired output and the output produced by the network (Song et al., 2016).

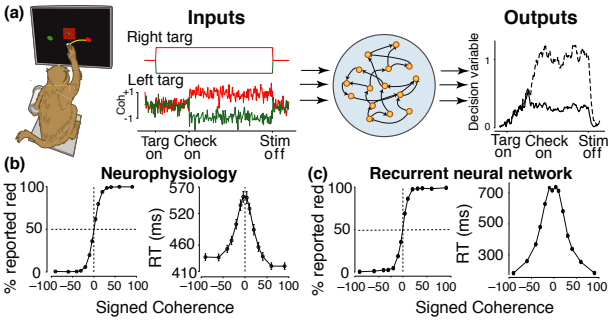


**Fig. 1.** Task design and behavior. (a) The task consists of a monkey reaching to the color of the target corresponding to the dominant color of the checkerboard. We modeled the task with four inputs, corresponding to the orientation of the targets and the color composition of the checkerboard, and two outputs, corresponding to decision variables for left and right reaches. The psychometric curve and corresponding reaction time curve of (b) the monkey, and (c) the network as a function of signed coherence.

We used 100 units in each stage of our multi-stage model, and 300 units for our single RNN. We implemented Dale's law with an 80/20 excitatory/inhibitory ratio (Song et al., 2016). The network time-constant $\tau$ was 50ms and when simulating the RNN, we used a first-order approximation of Equation 1 with step size of 10ms. When training the networks, we terminated training early to approximately reproduce the behavior of the animal. The RNN demonstrated similar psychometric and chronometric behavior as the animal, shown in Fig. 1b (compare with Fig. 1c).

## Results

### Single RNN

After optimizing the single RNN to perform the perceptual decision-making task, we examined the PSTHs of example units of the single RNN (Fig. 2a). In these representative examples, we observed that some units modulate their rates to encode the color choice (Fig. 2a, left), direction choice (Fig. 2a, right), or a mixture. Consistent with these single examples, we found that the choice probability (not shown) for these units could be strong for color, direction, or both. However, in contrast PMd activity only shows units that encode direction (e.g., Fig. 2b (Chandrasekaran et al., 2017)).

We next investigated how the population was representing task information. To this end, we embedded the rates of the neural population from independent trials near movement onset (averaged from a 400ms window) using the tSNE dimensionality reduction technique (Van der Maaten & Hinton, 2008). We found that there were four separable clusters, organized by both the direction (dot vs x) and color of the reach (Fig. 3a). Importantly, the tSNE embedding is

unsupervised, with the labelling being performed afterwards to assist with visualization. The four separable clusters indicate that on the population level, the single network represented both color and direction information to solve the task.

These results are in contrast to physiological recordings which suggest there is no representation of the dominant color in the checkerboard in PMd (Chandrasekaran et al., 2017; Wang et al., 2019) which was the region we were interested in modeling during this task. The absence of color representation in PMd suggests that PMd is receiving modified task related inputs to our single artificial RNN, possibly transformed upstream in areas such as the dorsolateral prefrontal cortex.

## Multi-stage RNN

To test this hypothesis and allow for the flexibility for the task inputs to be modified before they reached the last layer of processing, we imposed architectural constraints so that instead of a single RNN, there was a multi-stage RNN. Our multi-stage RNN consisted of three stages of processing. The connectivity between the stages of the network was loosely based on published connectivity matrices between PMd and area 9, and between area 9 and the dorsolateral prefrontal cortex (Markov et al., 2014). We trained the multi-stage RNN to reproduce the same behavior as the single RNN.

Fig. 2c shows PSTHS of three neurons, one from each stage. In contrast to the single RNN, in the last stage of processing of our multi-stage model, we observed similar single-unit representations (Fig. 2c, right panel) to PMd. In particular, units in the last stage (Fig. 2c, right panel) did not demonstrate color selectivity, as evidenced by the separation of colors in the PSTHs. However, importantly, these units did demonstrate direction selectivity and also exhibited slower rate changes for less discriminable

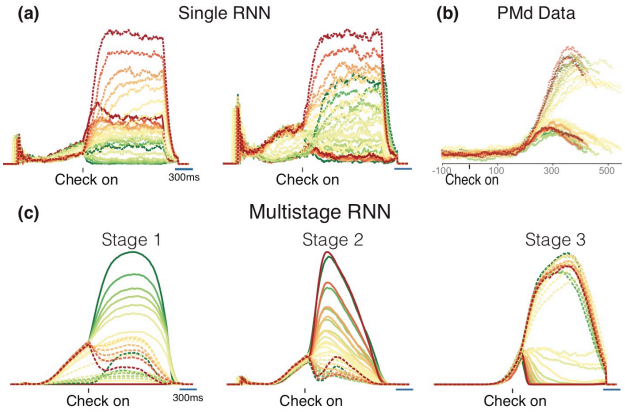


**Fig. 2**. PSTHs. Dotted lines denote left reaches and solid lines denote right reaches. Red (green) denotes reaches to the red (green) target, with darker shades corresponding to higher signed coherences. Example units from (a) single RNNs, (b) PMd data, and (c) multi-stage RNNs.

checkerboards, consistent with activity in PMd (compare Fig. 2c right and Fig. 2b).

Finally, when embedding the representation near movement onset in each of the three layers using tSNE, we observed only direction relevant information in layer 3. This can be seen by the mixing of both red and green crosses and dots in the right panel of Fig. 3b suggesting that it is not possible to read out information relating to color from the activity in layer 3. In both layer 1 and layer 2, we observed a mixture of color and direction related information. This visualization shows that the color information filters out before movement time in the third layer, indicating that the RNN no longer represents color but does maintain direction information. This provides a more faithful replication of activity in PMd.

## Discussion

We observed that, when building a multi-stage RNN, color information was filtered by the last layer in the

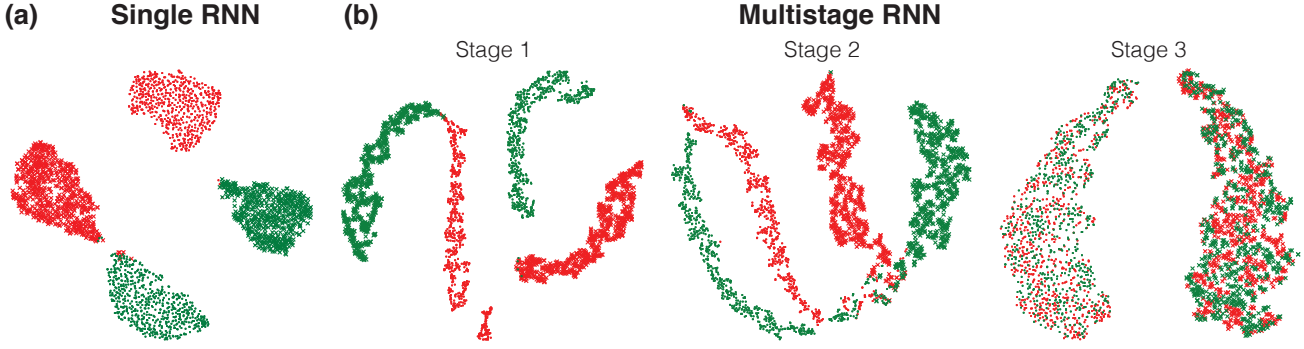


**Fig. 3**. tSNE embedding of RNN rates near movement time into a two-dimensional space. Trials to the right are denoted with an 'x' and trials to the left are denoted by a dot. (a) For the single RNN, the four clusters indicate both color and direction information are represented, unlike recordings in PMd. (b) For the multi-stage RNN, in layer 1, there are unique clusters which correspond to an encoding of both direction and color information. By layer 3, there is only an encoding of direction information, consistent with PMd data.

multi-stage RNN suggesting that the internal representation has become useful to solve our perceptual decision-making task (i.e., provide the correct direction output). Interestingly, this is similar to what is observed in later stages of cortical processing. In PMd, the decision direction information is present, however the color information is no longer represented.

Our multi-stage RNN makes several important physiological predictions. In particular, the results suggest that areas upstream of PMd should show mixed selectivity for both color and direction information. These representations are likely to be complex. We anticipate testing these predictions using recordings from dorsolateral prefrontal cortex and posterior parietal cortex.

Our results also suggest that multi-stage RNNs might be a better way to model cognitive tasks that involve complex transformations of multiple inputs into behavioral outputs. For instance, this approach has already been useful in describing the dynamics of the grasp network (Michaels et al., 2017).

Our results also appear to be related to recent theoretical results linking the performance of deep learning systems with the information bottleneck framework (Shwartz-Ziv & Tishby, 2017; Tishby, Pereira, & Bialek, 1999). These results suggest that the performance of deep learning systems depends on how subsequent stages of processing discard aspects of the input that are not necessary for the output, with each stage of processing becoming a more useful representation to produce the output (Achille & Soatto, 2018). Our results, showing the multi-stage RNN no longer represented color information, seem to be consistent with these theoretical results.

In order to provide neuroscientific insight and propose hypotheses as to how information may be filtered across stages in cortical processing, it is critical to understand how the learned weights allow for the propagation of information *necessary* for the output (e.g., direction) into the last stages of the network and filter out information *unnecessary* for the output (e.g., color). Further, to potentially provide insight to deep learning systems, it is important to study how these representations evolve during learning.

## Acknowledgments

CC and JCK conceived of the study. CC collected data in PMd and trained monkeys. JCK trained models. MK and JCK performed analyses. MK, CC and JCK discussed results and wrote the paper. MK was supported by the National Sciences and Engineering Research Council (NSERC). CC was supported by an NIH/NINDS R00 Grant 4R00NS092972-03.

## References

Achille, A., & Soatto, S. (2018). Emergence of Invariance and Disentanglement in Deep Representations. *Journal of Machine Learning Research*, *18*, 1–34.

Chandrasekaran, C., Peixoto, D., Newsome, W. T., & Shenoy, K. V. (2017). Laminar differences in decision-related neural activity in dorsal premotor cortex. *Nature Communications*, *8*(1).

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, *32*, 148–155.

Goudar, V., & Buonomano, D. V. (2018). Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks. *eLife*, *7*, 1–28.

Hennequin, G., Vogels, T. P., & Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, *82*(6), 1394–1406.

Laje, R., & Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, *16*(7), 925–33.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., … Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, *522*(1), 225–259.

Michaels, J., Schaffelhofer, S., Agudelo-Toro, A., & Scherberger, H. (2017). A modular neural network model of the primate grasping circuit (nanosymposium). *46th Annual Meeting of the Society for Neuroscience. Washington, November 14th, 2017*.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training Recurrent Neural Networks.

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information, 1–19.

Sompolinsky, H., Crisanti, A., & Sommers, H. (1988). Chaos in Random Neural Networks. *Physical Review Letters*, *61*(3), 259–262.

Song, H. F., Yang, G. R., & Wang, X. J. (2016). Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework. *PLoS Computational Biology*, *12*(2), 1–30.

Stroud, J. P., Porter, M. A., Hennequin, G., & Vogels, T. P. (2018). Motor primitives in space and time via targeted gain modulation in cortical networks. *Nature Neuroscience*, *21*(12), 1774–1783.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method arXiv : physics / 0004057v1 [ physics . data-an ] 24 Apr 2000, 1–16.

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning …*, *9*, 2579–2605.

Wang, M., Montanède, C., Chandrasekaran, C., Peixoto, D., Shenoy, K. V., & Kalaska, J. F. (2019). Macaque dorsal premotor cortex exhibits decision-related activity only when specific stimulus–response associations are known. *Nature Communications*, *10*(1).