Neural Mechanisms of Working Memory

王宇哲 1800011828
College of Chemistry and Molecular Engineering, Peking University

Introduction

工作记忆（Working Memory, WM）是人类对信息进行短期存储和操作的能力，在视觉处理、语言理解和情景记忆等大量认知任务中发挥着关键作用（Miller et al., 1960; Cowan, 2001）。认知神经科学实验证明，工作记忆的容量极其有限，对于大多数健康的人类受试者而言在$3\sim6$项之间（Cowan, 2001）。大量研究对工作记忆的机制提出了不同假说，例如Cowan等提出大脑具有专门的缓冲区，需记忆的项目可以在该区域短期存放，在需要时移除，工作记忆容量对应该缓冲区的大小（Cowan, 2001; Oberauer, 2002）。但上述模型未明确记忆项目从缓冲区中存放和移除的具体神经机制，假说未得到实验证实。

2008年，Mongillo et al. 提出了工作记忆的突触理论（Synaptic Theory of Working Memory），认为工作记忆与突触的短期可塑性（Short-term Plasticity, STP）有着直接联系，工作记忆的形成不需要神经元的持续发放（persistent firing），神经元之间连接的短期促进介导了工作记忆的形成（Mongillo et al., 2008）。上述理论得到了实验证据和计算模拟结果的支持，成为学界普遍接受的关于工作记忆的神经机制的理论。后续研究在此基础上成功解释了工作记忆的有限容量及工作记忆中信息的维持和操纵机制等问题（Mi et al. 2017; Masse et al. 2019）。本文试图结合相关文献简要阐述基于突触短期可塑性的工作记忆的神经机制，并对神经网络通过工作记忆维持和操纵信息的机制加以说明。

Synaptic Theory of Working Memory

工作记忆的突触理论基于STP，后者包括神经元脉冲发放造成神经递质减少导致的短期抑制（Short-term Depression, STD）效应和神经元脉冲发放导致Ca^{2+}浓度升高、增大神经递质释放概率导致的短期促进（Short-term Facilitation）效应。STP的数学模型由下列方程给出（Markram et al., 1998）：

$$
\begin{aligned}
\frac{du}{dt} &= -\frac{u}{\tau_f} + U(1-u^-)\delta(t-t_{sp}) \\
\frac{dx}{dt} &= \frac{1-x}{\tau_d} - u^+x^-\delta(t-t_{sp}) \\
\frac{dI}{dt} &= -\frac{I}{\tau_s} + Au^+x^-\delta(t-t_{sp})
\end{aligned} \tag{1}
$$

其中模型变量u表征神经递质的释放概率，x表征剩余神经递质的比例，I为输出到突触后神经元的突触电流，t_{sp}为脉冲发放的时刻。模型参数U表征单个脉冲导致u的增量，τ_d为STD的时间常数，τ_f为STF的时间常数。在t_{sp}时，脉冲发放所产生的突触电流由

$$\Delta I(t_{sp}) = Au^{+}x^{-} \tag{2}$$

给出，其中A表征对神经递质释放的响应幅度。

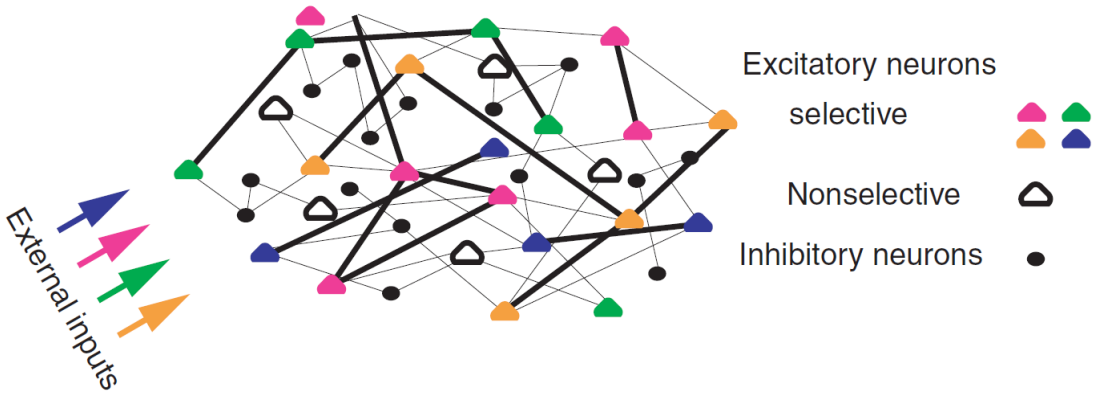

Figure 1

如Figure1所示，Mongillo et al. (2008) 提出，神经网络通过随机选取的部分兴奋性神经元，组成不同的、具有选择性的兴奋性神经簇（excitatory neural cluster）对不同的记忆项目进行编码，编码相同记忆项目的神经元之间的连接由于STF而加强，而不同的兴奋性神经簇与抑制性神经元连接，导致不同记忆项目之间的竞争。

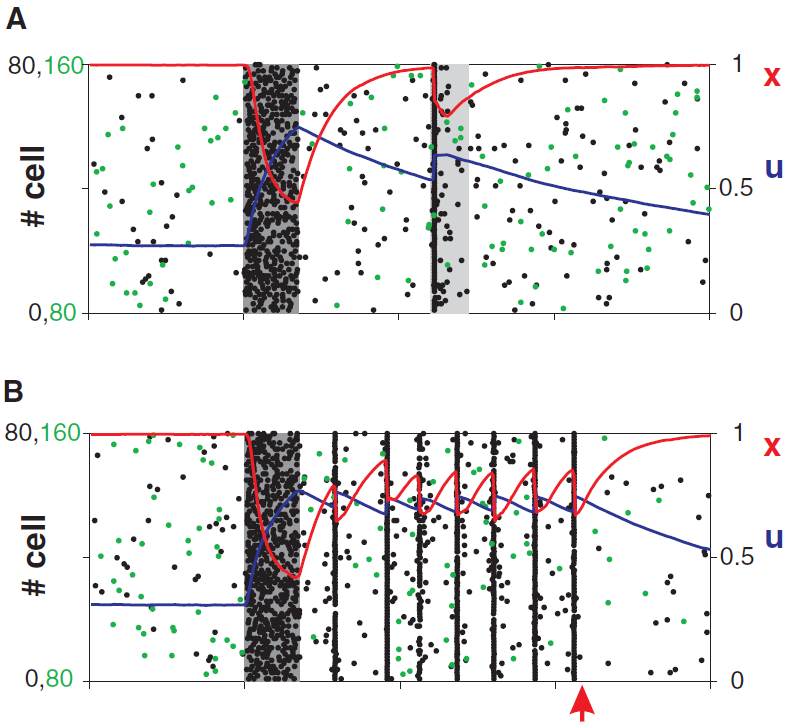

Figure 2

具体地，如Figure 2A所示，通过对相应的兴奋性神经元施加短时外部刺激（左侧灰色阴影）将一个记忆项目载入工作记忆中，该过程使得这部分神经元的活动增强，从而改变了突触连接的状态，使得u增大、x减小。控制$\tau_f \gg \tau_d$使得STF占主导，从而使这些神经元之间较强的连接持续存在。在对神经网络施加较弱的非特异性外部刺激（右侧灰色阴影）时，这些兴奋性神经元发生特异性的同步响应，产生群体发放（population spike, PS），从而使得记忆项目被重新激活。更一般地，如Figure 2B所示，适当增加背景输入的强度，神经网络可以进入双稳定状态（bistable activity regime），即使没有非特异性外部刺激，兴奋性神经簇的周期性群体发放也能在较长时间内稳定存在，直到停止神经网络激发（红色箭头）。

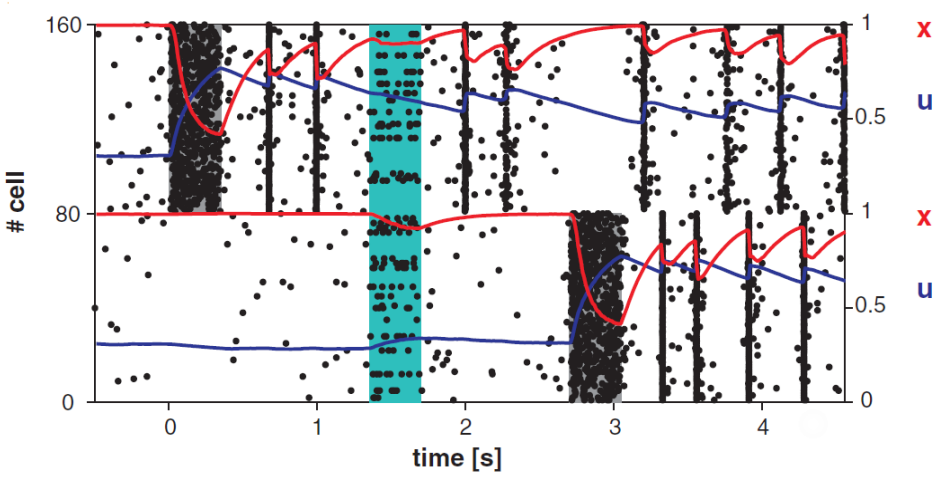

Figure 3

如Figure 3所示，当神经网络处理多个记忆项目（对于Figure 3, $t = 0$ 和 $t = 2.7$ s 两个记忆项目）时，不同的兴奋性神经簇通过相互交错的群体发放维持不同的记忆项目。

上述基于STP的工作记忆的突触理论很好地解释了工作记忆的神经机制，对于相关的认知神经科学实验结果具有较强的解释力。

Working Memory Capacity

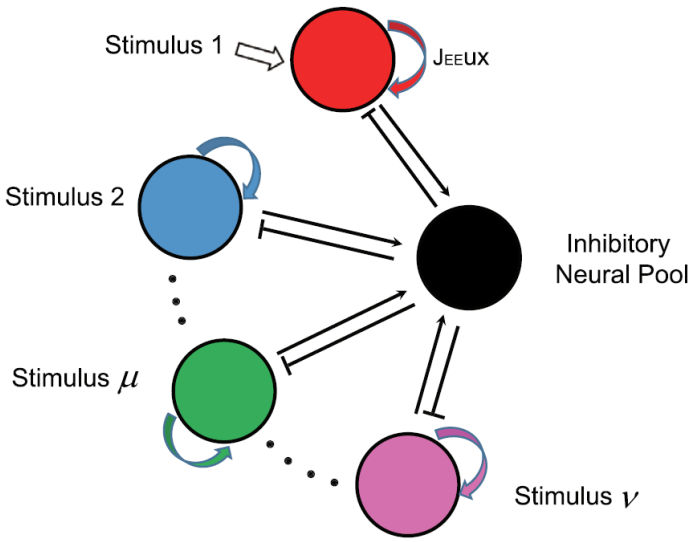

Figure 4

根据前述Mongillo et al. 提出的工作记忆的突触理论，可以进一步表示基于STP的工作记忆神经网络模型。如Figure 4所示，多个兴奋性神经簇与抑制性神经元相连接，对于每个兴奋性神经簇μ，考虑表征神经递质释放概率的变量u_μ、表征可用神经递质的比例的变量x_μ和突触电流h_μ，构建数学模型如下（Tsodyks et al., 1998）：

$$\begin{aligned}
\tau\frac{dh_\mu}{dt} &= -h_\mu + J_{EE}u_\mu x_\mu R_\mu - J_{EI}R_I + I(b) + I_e(t) \\
\frac{du_\mu}{dt} &= \frac{U-u_\mu}{\tau_f} + U(1-u_\mu)R_\mu \\
\frac{dx_\mu}{dt} &= \frac{1-x_\mu}{\tau_d} - u_\mu x_\mu R_\mu \\
\tau\frac{dh_I}{dt} &= -h_I + J_{IE}\Sigma_\nu R_\nu
\end{aligned} \tag{3}$$

其中J为两个神经元之间的绝对突触效能（absolute synaptic efficacy），表征突触连接的强度，角标E表示兴奋性神经元，I表示抑制性神经元，τ为时间常数；I_b为反映网络注意力状态的恒定背景激发，I_e为将需记忆的项目载入神经网络中的外部输入，神经元增益函数$R(h)=\alpha\ln\left(1+\exp\left(h/\alpha\right)\right)$。

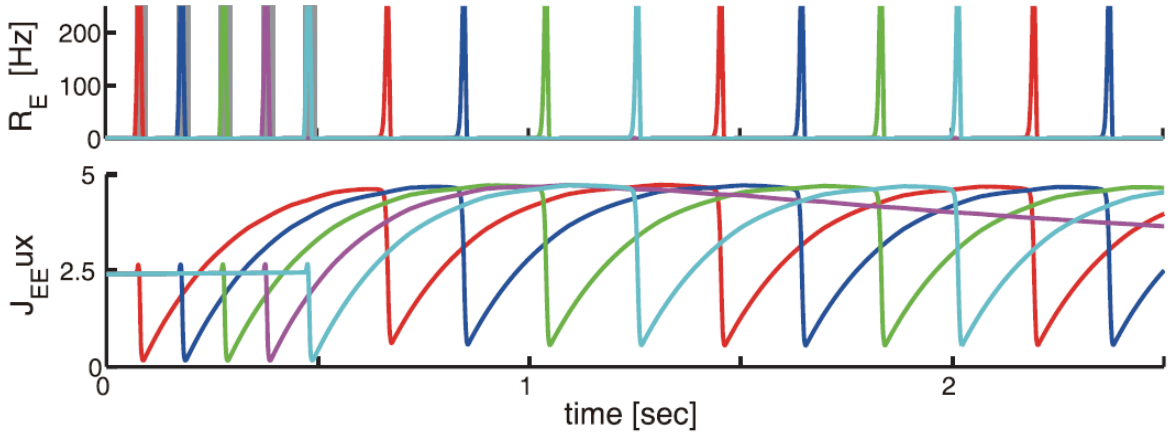

Figure 5

在上述模型和前文所述Mongillo et al. (2008) 开创性研究的基础上，Mi et al. (2017) 试图结合计算模拟，对工作记忆的有限容量作出解释。具体地，如Figure 5所示（上图为神经元的发放情况，下图为与不同记忆项目对应的、不同兴奋性神经簇的瞬时突触效能变化），使用与前额叶皮质锥体神经元连接的实验测量值相符合的数据进行人工神经网络模拟，通过向神经网络施加瞬时外部刺激将5个不同的记忆项目载入工作记忆中，可以看出只有4个记忆项目以兴奋性神经簇群体发放的方式得到成功维持，说明该神经网络工作记忆容量为4。进一步实验证明，若继续增加记忆项目的数量，第1个和最后1个记忆项目保留在工作记忆中的概率最高，计算证明可以合理假设当记忆项目的数量超出工作记忆容量，以前的项目之一以相等的概率从工作记忆中删除。

以(3)的数学模型为基础，Mi et al. (2017) 定量计算了工作记忆的容量N_C。记同一个兴奋性神经簇连续两次产生群体发放的最大时间间隔为T_{max}，不同兴奋性神经簇相邻两次群体发放的时间间隔为t_s，则

$$N_C \approx \frac{T_{max}}{t_s} \tag{4}$$

首先对T_{max}进行估计。由Figure 5可以看出，兴奋性神经簇激活并产生群体发放的最长时间大致与突触效能曲线由极小达到峰值所需的时间相等，而后者可以通过(3)近似定量解出，计算得

$$T_{max} \approx \tau_d \ln \frac{\tau_f / \tau_d}{1-U} \tag{5}$$

再对t_s进行估计。分析可知，t_s包含3个部分：前一项群体发放峰的宽度，该群体发放触发的抑制脉冲的延迟和宽度，以及下一项所对应的兴奋性神经簇从抑制中恢复、启动新的群体发放所需要的时间。计算表明，前两项正比于突触时间常数τ，第三项远大于其他两项，且可以通过近似求解(3)获得。具体地，对于该兴奋性神经簇的突触电流h，有

$$\begin{aligned} \tau \frac{dh}{dt} &= F(h) \\ F(h) &= -h + J_{max} R(h) + (I_b - I_{inh}) \end{aligned} \tag{6}$$

而h_{min}对应的时刻t与初始超极化电流h_0对应的时刻t_0之差近似即为t_s，计算得

$$t_s \approx \tau (\ln \frac{|h_0|}{I_b - I_{crit}}) + C \tag{7}$$

其中$I_{crit} \approx I_{inh} - \alpha \ln (J_{max} - 1)$为背景激发的临界值。根据(4)(5)(7)各式，有

$$N_C \approx \frac{\tau_d}{\tau} \frac{\ln \frac{\tau_f / \tau_d}{1-U}}{\ln \frac{|h_0|}{I_b - I_{crit}} + C} \tag{8}$$

根据以上结果可知，工作记忆容量N_C正比于两个时间常数之比$\frac{\tau_d}{\tau}$，τ_d和τ分别表征突触抑制和突触电流衰减；背景激发I_b应大于临界值I_{crit}，否则分母对数发散，使得$N_C \to 0$，无法维持工作记忆。该结论与本文第1部分的结果（参考Figure 2）是完全一致的。

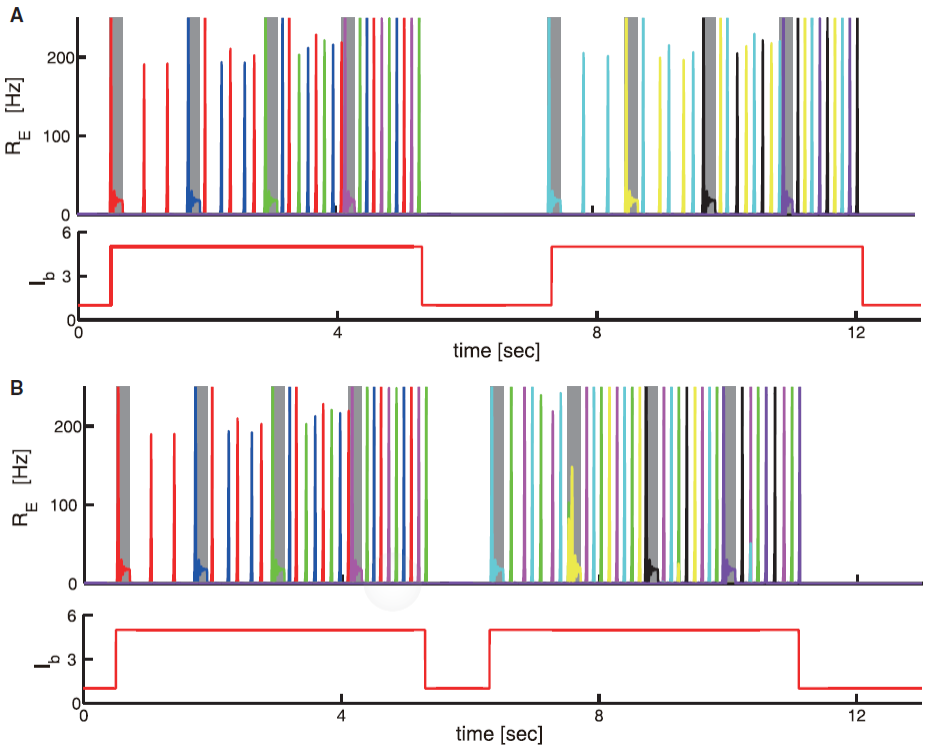

Figure 6

上述结果说明，工作记忆容量N_C直接取决于大脑皮层神经网络的基本参数，因此简单的后天练习不能实现N_C的显著提高，这与实际实验结果也具有一致性。实际上，如Figure 6所示，由于人脑的工作记忆容量N_C有限，当需要记忆的项目较多时，需要适当调制背景输入、在输入之间插入足够长的时间间隔（Figure 6上图），否则不同的记忆项目会发生混合（Figure 6下图）。

RNN with STP for Working Memory

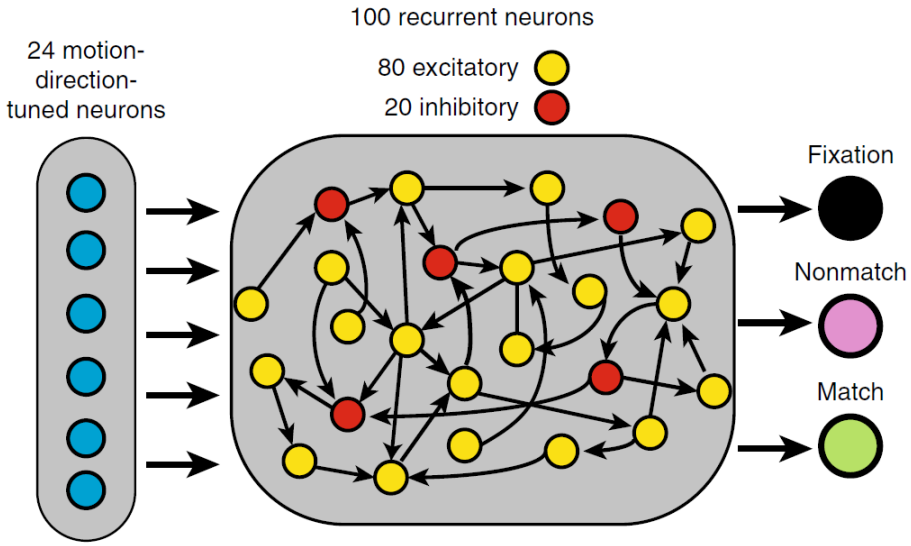

Figure 7

记忆对于人工循环神经网络（Recurrent Neural Network, RNN）有着重要意义，例如基于长短期记忆（Long Short-term Memory, LSTM）的RNN架构对于需要用到长时间间隔信息的时序性任务有着很好的表现，而RNN也适于对实际的神经科学机制进行研究。Masse et al. (2019) 通过将RNN与STP结合，对工作记忆中信息的维持和操纵机制进行研究，网络架构如Figure 7所示，输入层由24个兴奋性的方向调谐神经元组成，输入到由80个兴奋性神经元和20个抑制性神经元组成的循环神经网络中，这些神经元之间的连接权重通过STP进行动态调节。

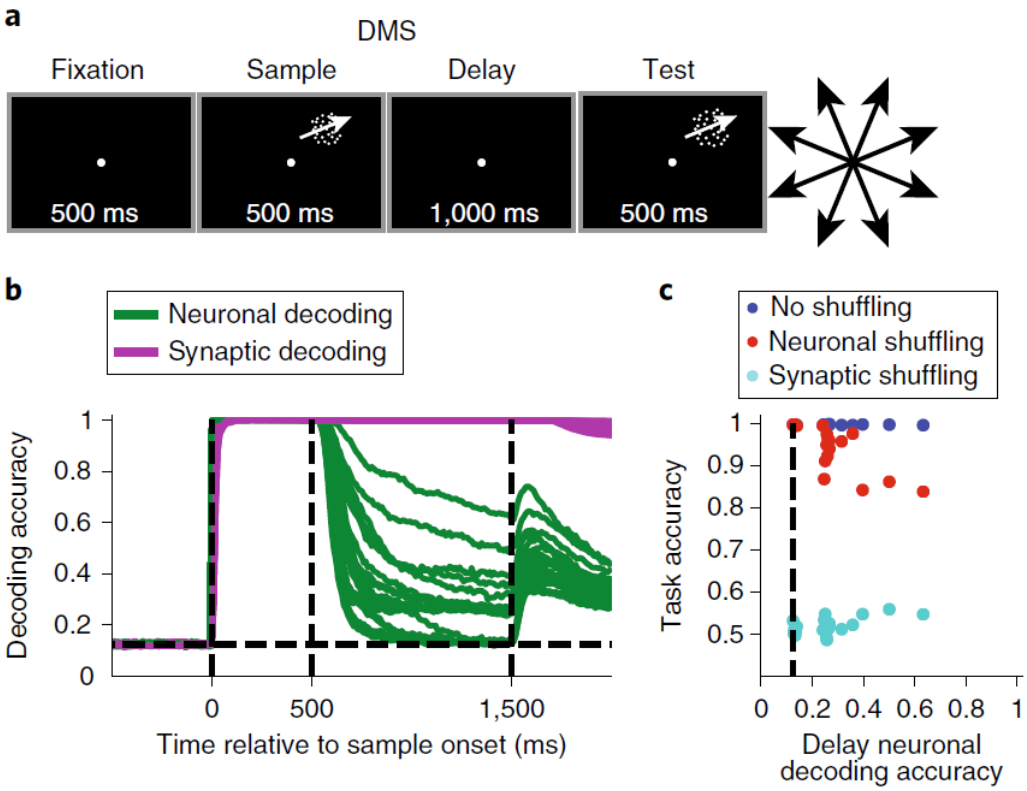

Figure 8

Masse et al. (2019) 训练上述网络完成延迟样本匹配（delayed match-to-sample, DMS）任务，对依次出现的样本和测试图像是否匹配进行判断（Figure 8a）。为研究工作记忆维持信息的机制，作者分别取100个神经元和100个通过STP进行调节的突触效能，对其所记录的样本方向信息进行解码，结果如Figure 8b所示，使用突触效能进行解码，解码准确率接近1.0，而通过神经元活动解码的准确率显著较低。上述结果证明，样本作为工作记忆被准确地通过突触效能进行编码，而在神经元活动中几乎没有编码。为进一步确证神经网络用以解决DMS任务的神经机制，作者分别打乱神经元活动（"Neuronal shuffling"）、打乱突触效能（"Synaptic shuffling"）、不作任何打乱（"No shuffling"），再判断网络输出是否正确。如Figure 8c所示，打乱神经元活动几乎不影响准确率，而打乱突触效能显著降低了准确率，从而有力证明了在延迟期间工作记忆通过突触效能维持信息的神经机制，而神经元活动保持在较低水平，从而实现了不需要持续神经活动的信息维持。

作者随后研究了该RNN模型通过STP操纵信息的机制，实验证明，对于需要对工作记忆中的信息进行操作的任务，持续的神经活动是必要的，该结论解释了不同的日常任务中，神经活动的强度显著不同的背后机制。限于篇幅，本文不再展开对于该部分内容的介绍。

Masse et al. (2019) 的工作将RNN与STP相结合，对人脑利用工作记忆完成任务的过程进行了有效而成功的模拟。值得说明的是，该实验中Figure 8b的数据测量在实际的神经科学实验中几乎是不可能的，而在人工RNN中则极易测量，可见人工神经网络建模能够为神经科学的研究提供极大便利，同时基于人工RNN研究发现的生物体内的神经机制也能够反过来促进研发更好的神经网络算法。作者在文中认为，如果添加神经元到自身的连接，结合STP的RNN能够潜在地解决具有长期时间依赖性的任务，接近甚至超越RNN+LSTM的效果。这对于后续研究具有极高的启发性。

Conclusion

本文结合相关文献，简要阐述了基于突触短期可塑性的工作记忆的神经机制，通过对工作记忆容量的定量计算确证了这一神经机制的合理性，并对神经网络通过工作记忆维持和操纵信息的机制加以说明。

本文认为，工作记忆背后的神经生物学机制能够更加深入地与人工神经网络算法相结合，启发后者在长时间间隔的时序任务上取得更优表现，有待后续进一步的研究。

References

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Markram, H., Wang, Y., & Tsodyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences*, *95*(9), 5323-5328.

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, *22*(7), 1159-1167.

Mi, Y., Katkov, M., & Tsodyks, M. (2017). Synaptic correlates of working memory capacity. *Neuron*, *93*(2), 323-330.

Miller, G. A., Eugene, G., & Pribram, K. H. (2017). *Plans and the Structure of Behaviour* (pp. 369-382). Routledge.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543-1546.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411.

Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural computation*, *10*(4), 821-835.