doi:10.19306/j.cnki.2095-8110.2023.02.007

基于 STDP 奖励调节的类脑面向目标导航

戴嘉伟,熊　智,晁丽君,杨　闯

(南京航空航天大学自动化学院导航研究中心,南京 211106)

摘　要:动物具有优秀的空间自主定位导航能力,能够实现在无先验环境信息下的导航定位和导航决策过程。针对智能体在连续空间中面向目标导航问题,研究了一种基于生物学放电时间依赖可塑性学习规则的智能体面向目标导航算法。首先分析了动物面向目标导航决策过程中的生理学机理,在此基础上,构建了基于脉冲神经网络的位置细胞和动作细胞模型。动作细胞间权值采用横向竞争模型更新,通过环境奖励信号的更新,采用放电时间依赖可塑性学习规则对位置细胞前馈动作细胞模型的突触权重进行权值调节,利用动作细胞群的脉冲放电现象表征智能体运动方向和速度。最后,对所提算法进行了仿真实验验证。仿真结果表明,所提出的类脑面向目标导航算法能够在单障碍环境中实现 30 ms 左右的规划速度,相比传统强化学习 Q 学习方法平均路径规划长度缩短了 15.9%。

关键词:类脑目标导航;放电时间依赖可塑性;智能体;脉冲神经网络;位置细胞;动作细胞

中图分类号:V249.32;TP18　　**文献标志码**:A　　**文章编号**:2095-8110(2023)02-0047-10

Brain-inspired target-driven navigation based on STDP reward modulation

DAI Jiawei, XIONG Zhi, CHAO Lijun, YANG Chuang

(Navigation Research Center, College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Animals have an excellent ability to perform autonomous localization and navigation, which can realize the navigation and decision-making process without prior environmental information. Aiming at the problem of target-oriented navigation of agents in continuous space, a target-driven navigation algorithm based on the rule of biological spiking-time-dependent plasticity (STDP) is studied. Firstly, the physiological mechanism in the decision-making process of target-driven navigation in animals is analyzed. On this basis, a place cell and action cell model based on spiking neural network is constructed. The weights between action cells are updated by the horizontal competition model, and the synaptic weights between the place cell and action cell model is updated by the rule of STDP. The movement direction and speed of the agent is represented by the pulse discharge phenomenon of the action cell group. Finally, the proposed algorithm is verified by simulation experiments. Simulation results show that the proposed brain-inspired target-driven navigation algorithm can achieve a planning speed of about 30 ms in a single obstacle environment. Compared with the traditional reinforcement learning method of Q learning, the average path plan-

收稿日期: 2022-08-22;**修订日期**:2022-11-07

基金项目:国家自然科学基金(61873125);国防基础科研计划项目(JCKY2020605C009);校创新基金项目(xcxjh20210334)

作者简介:戴嘉伟(1997-),男,硕士研究生,主要从事智能导航技术方面的研究。

通信作者:熊智(1976-),男,教授,主要从事惯性/组合/智能导航技术方面的研究。

ning length is reduced by 15.9%.

Key words: Brain-inspired navigation; Spiking-time-dependent plasticity; Agents; Spiking neural network; Place cell; Action cell

0 引言

面向目标导航是智能体自主执行任务(如自主侦察与救援)的前提和基础。随着人工智能的迅猛发展,在自然或人为灾害后的恶劣环境中,智能体能够代替人类最大限度地降低目标搜寻等任务难度,并提高任务效率。由于智能体工作逐渐面向非结构化和未知环境,如何快速准确地搜索出一条由初始状态到目标状态的安全路径成为当前智能体规划的技术难点,即面向目标导航问题。

1970 年左右,研究人员就已经开始了对智能体面向目标导航算法的研究[1]。传统方法中,文献[2]采用快速扩展随机树(rapidly-exploring random tree,RRT)算法,对规划航路点进行无人机飞行动力学约束和局部航路动态规划。RRT 算法[3-4]无需对规划空间进行预先处理且概率完备,但在节点拓展时盲目性和随机性大,缺乏较强的目的导向性。智能启发式方法[5-6]是受自然规律启迪而模仿出的算法,具备一定的自我学习、自我更新和记忆能力。文献[7]采用自适应学习粒子群算法,提出了一种基于协同进化的粒子群算法以解决机器人路径规划问题,更好地调整全局和局部搜索能力,解决了粒子群优化的停滞问题,但启发式算法[8]在未知环境下往往会陷入局部最小问题。除了传统和启发式方法之外,基于强化学习的规划算法如时间差分模型等也被广泛应用于各种自治系统的路径规划[9],但在连续状态空间中智能体会陷入维数灾难,收敛缓慢。此外,近年来基于深度强化学习的智能体导航方法解决了复杂目标任务难以建模的问题。文献[10]提出了一种基于优化深度 Q 网络(deep Q-network, DQN)算法的全局路径规划模型,解决了传统方法中路径冗余问题,但现实环境和模拟环境的差异性导致智能体可移植性差、计算量大,且训练过程复杂困难。这类基于传统冯诺依曼计算结构的规划方法在面对复杂目标导航问题时具备离散状态下的有效处理能力,但是其庞大的计算量导致计算效率低下及训练困难的问题,同时缺少生理学结构的研究,不具备生理学可解释性,因而需要探索发展基于新型计算模型的,能适应非结构化、未知环境的面向目标导航方式。

为解决现有面向目标导航方法存在的问题,本文提出了一种基于脉冲神经网络的智能体类脑面向目标导航方法。根据生物大脑海马体(hippocampus,HC)和腹侧被盖区(ventral tegmental area, VTA)到前额叶皮层(prefrontal cortex,PFC)中动作细胞(action cell, AC)调节现象,采用基于脉冲响应模型的脉冲时间依赖可塑性(spike-timing-dependent plasticity,STDP)学习规则,构建了前额叶皮层环状动作细胞的脉冲神经网络模型,利用动作细胞群脉冲放电现象表征智能体的运动方向和速度。本文所提模型能够同时记忆陌生环境中的障碍物和目标位置,通过动作细胞决策实现智能体的面向目标类脑导航功能,同时具备对于多种陌生环境下的面向目标导航能力,具有一定的模型泛化能力。

1 动物面向目标导航机理

生理学上的大脑导航关键区域结构示意图如图 1 所示。1971 年,J. O'Keefe 等发现在海马体中位置细胞(place cell, PC)存在着空间特定位置选择性放电现象[11]。动物在到达环境区域时,位置细胞会记忆特定环境信息信标点,迅速生成并且形成稳定的位置野[12-13],同时位置细胞群的放电活动随着动物到达特定信标点时显著提高,进而实现了对动物当前位置的编码[14]能力。动物导航以大脑海马区中的大量位置细胞集群放电为基础,逐渐形成稳定编码空间环境认知地图[15]的位置野。但是单一海马位置细胞对环境信息的表征能力并不能实现

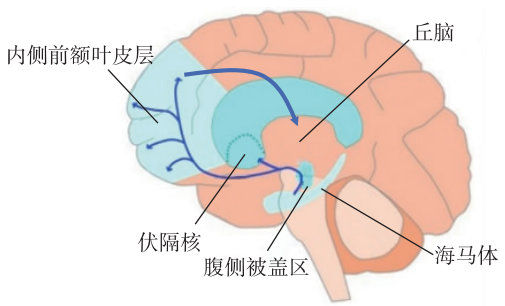


图 1 面向目标导航关键脑区

Fig. 1 Key brain regions for target-driven navigation

动物导航过程中的行为决策，需要通过和前额叶皮层构建特定的动态突触连接结构，形成大脑导航命令和控制中枢神经网络[16]。

生物在环境探索过程中进行目标导航的流程如下：1）由视觉皮层或感觉皮层等接收处理环境状态信息更新，向大脑腹侧被盖区传递环境奖励信号；2）腹侧被盖区中的多巴胺能神经元接收环境奖励信号生成奖励调节信息，海马体位置细胞生成空间认知信息实现位置信息编码，两者进一步通过伏隔核（nucleus accumben，NA）神经元形成前馈通路影响前额叶皮层动作细胞；3）前额叶皮层动作细胞激活依赖 PC-AC 前馈通路信号和 AC 横向通路信号，动作细胞群集群放电影响丘脑（thalamus，Tha）确定生物运动方向；4）通过运动皮层确定运动输出，更新智能体位置信息，最终完成面向目标导航的过程。具体流程如图 2 所示。

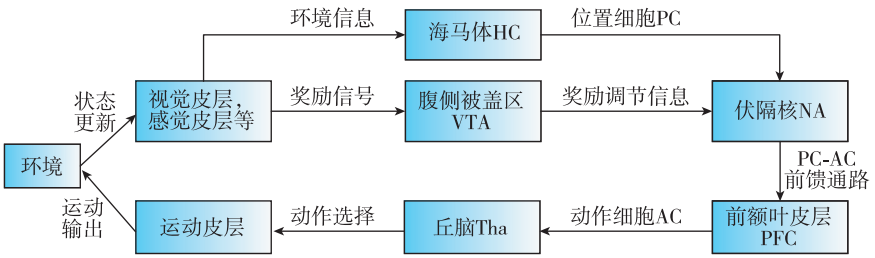


图 2　生物导航行为模型

Fig. 2　Model of biological navigation behavior

2　类脑目标导航模型

根据生物目标导航行为模型，设计了如图 3 所示的基于 STDP 奖励调节的类脑面向目标导航算法流程，主要内容为：1）构建了海马体位置细胞和前额叶皮层动作细胞的脉冲神经网络模型，分别表征智能体位置空间和动作空间信息；2）位置细胞采用前馈连接模型影响动作细胞激活，动作细胞群采用横向竞争模型输出动作细胞膜电位；3）根据动作细胞放电率，设计了智能体动作选择函数，同时基于脉冲神经网络权值更新方法，智能体接收到环境奖励调节信息后，采用 STDP 学习规则更新位置细胞到动作细胞的前馈突触权值。

2.1　位置细胞建模

当生物处在空间特定的范围内时，海马体内某些锥体细胞会出现最大频率放电现象[17-18]，而在其他位置则很少甚至没有放电现象，则该细胞被称作位置细胞，其放电现象所对应的环境生物活动范围则被称为该细胞的位置野。只要环境处于长期稳定状态，位置细胞的位置野在环境中的大小、形状、分布以及最大放电频率都可以维持较长时间的平稳状态，这一特性说明了位置细胞的位置表征能力具有良好的稳定性[19]。

在面向目标的导航任务需求下，根据大脑行为决策生理学依据，位置细胞采用位置野信息密集编

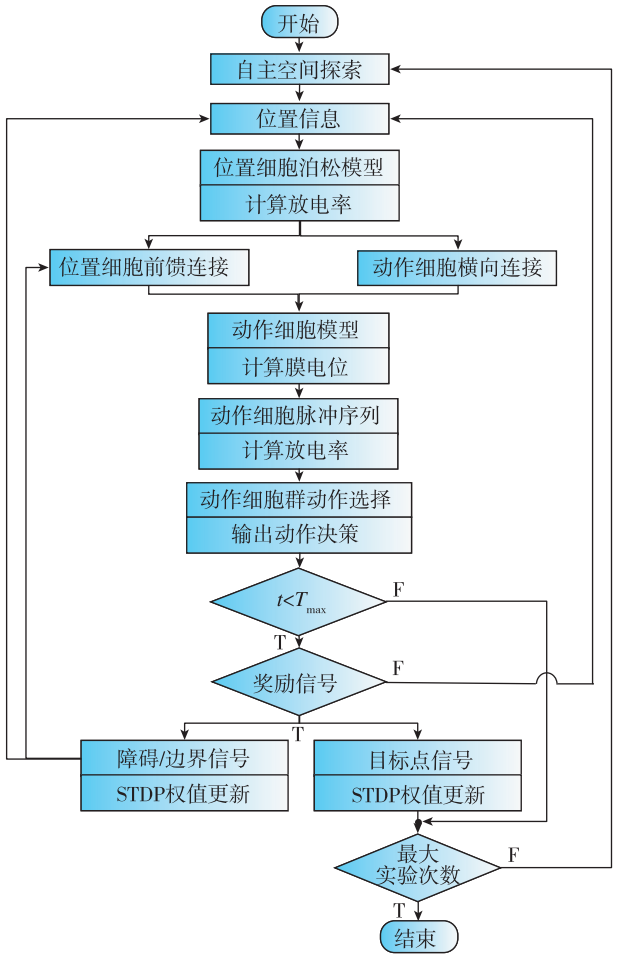


图 3　基于 STDP 奖励调节的目标导航算法流程

Fig. 3　Workflow of brain-inspired target-driven navigation algorithm based on STDP reward modulation

码智能体所处的整体空间环境。假设智能体在 t 时刻的位置由笛卡尔坐标系中的 $Pos(t)=(x(t),y(t))$ 来表示，智能体当前位置可由位置细胞群放电现象联合编码。假设在智能体所处空间环境中均匀分布着 $N_{pc}=121$ 个位置细胞，位置细胞的位置野半径为 $\sigma=0.4$ m，位置细胞的放电率 r_i 可建模为一个非齐次泊松过程

$$r_i(Pos(t))=\lambda\exp\left[-\frac{(x(t)-x_i)^2+(y(t)-y_i)^2}{\sigma^2}\right] \tag{1}$$

位置细胞放电率 r_i 由智能体当前位置 (x,y) 到位置细胞中心 (x_i,y_i) 的函数关系来表征，当智能体恰好位于位置野中心 (x_i,y_i) 时，位置细胞放电率最大，通过这种集群放电编码方式，位置细胞即可表征整个空间环境。为了保证在保持导航精度的同时缩短计算时间，令 $\lambda=400$ Hz，位置细胞的放电率会处在较高水平，并且放电率随着相对距离的增大而逐渐减小。

由于位置细胞建模为泊松神经元，则瞬时放电率为 r_i 的位置细胞在 $t_1\sim t_2$ 时间段内产生 n 个脉冲序列的概率为

$$P(n,t_1,t_2)=e^{-r_i\Delta t}\frac{r_i\Delta t^n}{n!} \tag{2}$$

由式(2)可知，在无穷小时间间隔内产生一个脉冲的概率为 $P(n=1)\approx r_i\Delta t$，当 P 大于一个在 0～1 之间均匀分布的随机数时，位置细胞会产生一个脉冲信号，并记录脉冲到达时间 $\bar{t}_i$。

2.2 动作细胞建模

如图 4 所示，位置细胞作为类脑面向目标导航系统的输入，通过加权系数 w^{ff} 投射到所有动作细胞。这些前馈加权系数初始化为 w_{in}，并且在最大权值 w_{max} 和最小权值 w_{min} 之间有界，这样使得兴奋性刺激和抑制性刺激均能通过位置细胞对动作细胞产生影响，同时动作细胞之间通过横向权重 w^{lc} 互相连接。根据神经科学理论，神经元在放电之后的短暂时间内存在不应期，即对输入信号不响应。为了在脉冲序列中模拟这个过程，在神经元放电之后的不应期内，将瞬时放电频率置为 0。在不应期结束之后，瞬时放电频率在限定时间内逐渐回到原始值。t 时刻动作细胞 j 的膜电位为

$$u_j(t)=\sum_i\sum_{t_i\in F_i}w_{ji}^{ff}\cdot\varepsilon(t-\bar{t}_i)+\sum_{k,k\neq j}\sum_{t_k\in F_k}w_{jk}^{lc}\cdot\varepsilon(t-\bar{t}_k)+p_{ref}H(t-\bar{t}_j)\exp\left(\frac{t-\bar{t}_j}{\tau_m}\right) \tag{3}$$

$$\varepsilon(t)=\frac{\tau_0}{\tau_m-\tau_s}(e^{\frac{-t}{\tau_m}}-e^{\frac{-t}{\tau_s}})H(t) \tag{4}$$

式中，w_{ji}^{ff} 为位置细胞 j 到动作细胞 i 的连接权值；w_{jk}^{lc} 为动作细胞 j 和 k 之间的横向连通权重；$\varepsilon(t)$ 为突触刺激反应核函数；$\bar{t}_i$ 和 $\bar{t}_k$ 表示位置细胞 i 和动作细胞 k 的脉冲到达时间；F_i 和 F_k 为包含 $\bar{t}_i$ 和 $\bar{t}_k$ 的集合；$p_{ref}=-5$ mV 为动作细胞不应期参数；$H(t)$ 为海维赛德阶跃函数；时间常数 $\tau_m=10$ ms，$\tau_s=5$ ms，$\tau_0=10$。

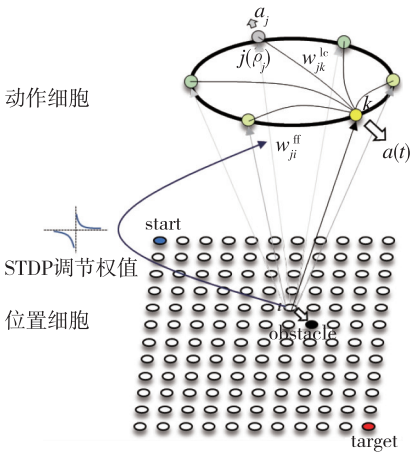


图 4 位置细胞-动作细胞模型

Fig. 4 Model of place cells to action cells

动作细胞脉冲响应处于随机状态，动作细胞放电率遵循依赖于动作细胞 j 膜电位的非齐次泊松过程

$$r_j(u_j(t))=r_0\exp\left(\frac{u_j(t)-\eta}{\tau}\right) \tag{5}$$

式中，$r_0=100$ Hz 为动作细胞最大放电率；$\tau=0.05$ 为脉冲响应过程中的脉冲时间间隔，确保了脉冲响应过程的随机性；$\eta=10$ mV 为动作细胞放电率调节阈值。动作细胞产生脉冲响应信号后，记录脉冲响应到达时间 $\bar{t}_j$。

定义动作细胞 j 和 k 之间的横向连通权重为

$$w_{jk}^{lc}=\frac{w_+f(j,k)}{Z}+\frac{w_-}{N_{ac}} \tag{6}$$

$$f(j,k)=(1-\delta_{jk})e^{\varphi\cos(\theta_j-\theta_k)} \tag{7}$$

式中，$\theta_j=2j\pi/N_{ac}$，$N_{ac}=40$ 为动作细胞个数；Z 为归一化因子；$w_-=-300$；$w_+=100$；$f(j,k)$ 为动作细胞 (j,k) 间横向连接函数。横向连接函数随动作选择方向相似度单调递增，δ 为狄拉克函数，$\varphi=20$ 为放电率调节因子。因此，当存在神经元同时

处于相似的放电频率时，动作细胞神经元会处于兴奋性刺激连接状态，否则处于相互抑制性状态，这保证了任意时间只会存在部分具有相似放电活动的动作细胞处于活跃状态，使得整体运动轨迹平滑且连续。

2.3　基于 STDP 奖励调节的面向目标类脑导航

2.3.1　面向目标类脑导航模型

在实验环境中，智能体位置信息由位置细胞编码，而智能体运动方向和速度决策由动作细胞决定。当遇到障碍物，环境边界或目标点获得奖励信号时，智能体通过 STDP 奖励调节规则调节位置细胞和动作细胞之间的前馈连接突触权重。动作细胞之间通过横向连接互相影响，当动作细胞神经元存在相似放电现象时，动作细胞神经元会处于兴奋性状态，否则处于抑制性状态。因此，智能体运动决策依赖于动作细胞，而动作细胞的激活依赖于位置细胞的前馈连接和动作细胞间的横向连接。

动作空间由脉冲神经网络建模的动作细胞表示。不同的动作细胞分别表示不同的运动方向，通过横向连接确保细胞间互相竞争，实现胜者为王的局面。来自位置细胞前馈连接和来自动作细胞横向的竞争连接共同作用，经式(3)输出动作细胞膜电位，联合决定动作细胞脉冲响应，最后由脉冲响应动作细胞放电率决定每个时刻前进的速度和方向。智能体的运动由动作细胞决定，设速度参数 $a_0=0.1$ m，采用动作细胞神经元 a_j 表示笛卡尔平面上不同的前进策略

$$a_j=a_0(\sin(\theta_j),\ \cos(\theta_j)) \tag{8}$$

智能体的动作选择过程根据动作细胞神经元放电率，由滤波脉冲序列 Y_j 和核函数 γ 决定

$$Y_j=\sum\nolimits_{\bar{t}_j\in F_j}\delta(t-\bar{t}_j) \tag{9}$$

$$\gamma(t)=\frac{e^{-t/\tau_\gamma}-e^{-t/v_\gamma}}{(\tau_\gamma-v_\gamma)}H(t) \tag{10}$$

式中，时间常数 $\tau_\gamma=50$ ms，$v_\gamma=20$ ms；Y_j 是动作细胞 j 在 $\bar{t}_j$ 时刻产生的全部突触后脉冲序列。

在连续运动情况下，需要动作细胞在每个时刻 t 都即时输出动作选择。每个动作细胞 j 表示了方向 a_j，t 时刻前额叶皮层动作选择过程中的动作细胞放电率为 $\rho_j(t)$，决定了最优的前进方向 $a(t)$，$a(t)$ 为所有动作神经元决策方向的加权均值，如式(12)所示

$$\rho_j(t)=(Y_j\circ\gamma)(t) \tag{11}$$

$$a(t)=\sum\frac{\rho_j(t)a_j}{N_{\mathrm{ac}}} \tag{12}$$

式中，N_{ac} 为动作细胞数量；$\circ$ 表示映射的乘积，即 $(Y_j\circ\gamma)(t)=Y_j(\gamma(t))$。在动作细胞数目足够多的情况下，该动作决策机制使得智能体具备了任意方向的连续移动能力，同时提高了导航定位和动作选择的准确性。当动作 $a(t)$ 确定之后，智能体的位置信息根据式(13)进行更新

$$\Delta x(t)=\begin{cases}a(t) & x(t+1)\text{ 在界限内}\\ d\cdot\boldsymbol{u}(x(t)) & \text{其他}\end{cases} \tag{13}$$

智能体根据 t 时刻动作选择 $a(t)$ 移动，当到达训练边界时，通过指向边界内部的单位向量 $\boldsymbol{u}(x(t))$ 与抗拒距离 $d=0.01$ m 转至训练区域内部。为避免较大的边界效应，边界上的位置细胞和指向边界外的动作细胞间的前馈连接权重设置为 0。

2.3.2　基于 STDP 奖励调节的突触权值更新方法

兴奋性和抑制性突触的权值变化效率受到多种可塑性机制的影响，其中 STDP 建立在神经元脉冲模式的相关性基础上，是赫布可塑性的一种形式。STDP 的确切形式会因为不同类型的突触形式而不同。在其最常见的形式中，突触时间依赖的可塑性表明，突触前脉冲发生后不久突触后脉冲就发生(前-后模式，pre-post)会导致突触权值的增加，即突触的长期增强(long-term potentiation，LTP)，突触权重的增加随着两次脉冲时间的不同呈指数衰减；反之，当突触前脉冲发生在突触后脉冲之后(后-前模式，post-pre)会导致神经元间突触经历一个长期抑制(long-term depression，LTD)。现在人们普遍认为，记忆和学习与 STDP 密切相关[20-21]。在数学上，突触强度的变化可以表示为

$$stdp=\Delta\omega=\sum_n\sum_m K(t^m_{\mathrm{post}}-t^n_{\mathrm{pre}}) \tag{14}$$

式中，t^m_{post}、t^n_{pre} 分别为突触后和突触前脉冲时间，m 和 n 分别为突触后和突触前脉冲神经元的计数；K 为 STDP 函数。相关的前-后模式脉冲对形成了突触的资格迹，呈现出指数衰减的过程，形成了多巴胺增强 LTP(或 LTD)的时间窗。

本文的学习模型考虑了突触前和突触后神经元之间的多个脉冲相互作用。在非对称形式学习规则中，STDP 函数由式(15)中函数定义

$$K=\begin{cases}A_+\exp(-\Delta t/\tau^+) & \text{if } \Delta t>0\\ A_-\exp(\Delta t/\tau^-) & \text{if } \Delta t<0\end{cases} \tag{15}$$

式中，$\Delta t = t_{\text{post}}^{m} - t_{\text{pre}}^{n}$ 为突触后和突触前脉冲时间。

如果 $\Delta t > 0$，即权值变化为正，则认为发生了突触的长期增强；另一方面，如果 $\Delta t < 0$，即发生了突触的长期抑制，那么突触权重减小。A_+ 和 A_- 分别是定义 LTP 和 LTD 窗口大小的标度常数，τ^+ 和 τ^- 定义了 2 个窗口的衰减率，其中 $A_+ = 0.1$，$A_- = -0.15$，$\tau^+ = \tau^- = 20$ ms。

STDP 规则中，突触的强度和突触后脉冲的概率之间存在线性关系：权重越大，下一个神经元就越有可能发生放电现象。因此，一旦突触增强，其后续增强的机会就会增加。然而，在生物学中，突触权重不能任意增大。因此，本文将兴奋性突触的大小限制在 0～3 mV 之间，抑制性突触的大小限制在 0～1 mV 之间。对于奖励调节 STDP 模型，在突触 ji 上从神经元 i 到神经元 j 的权值变化 w_{ji} 可以写成

$$\Delta w_{ji}(t) = e_{ji}(t) d(t) \tag{16}$$

式中，e_{ji} 表示 t 时刻从神经元 i 到神经元 j 的资格迹；$d(t)$ 为奖赏函数。资格迹函数由以下函数给出

$$\dot{e}_{ji} = -e_{ji}/\tau_{\text{c}} + stdp(t_{\text{post}} - t_{\text{pre}}) \tag{17}$$

式中，$stdp(t_{\text{post}} - t_{\text{pre}})$ 为根据 STDP 学习规则变化的突触权值；$\tau_{\text{c}} = 10$ ms 是表征资格迹衰减率的时间常数。

在获得奖励后，多巴胺（dopamine，DA）奖励函数 $d(t)$ 会随着时间的推移而增加，然后呈指数衰减到基础水平

$$\dot{d} = -d/\tau_{\text{d}} + DA(t) \tag{18}$$

$$DA(t) = \begin{cases} 0.01\ \mu\text{M/s} + 0.5\ \mu\text{M} & \text{reward} \\ 0.01\ \mu\text{M/s} & \text{else} \end{cases} \tag{19}$$

式中，$DA(t)$ 为多巴胺浓度（M）；$\tau_{\text{d}} = 0.2$ s 为 DA 时间常数，确保突触权重不会发生剧烈的跳变。图 5 所示为本文资格迹追踪影响下的突触强度变化示意图，pre-post 脉冲对产生了兴奋性刺激下的资格迹响应，并且在此期间使得受到多巴胺激励的突触强度增强。

3 仿真实验及结果分析

为验证本文所提智能体类脑面向目标导航算法的有效性，设计图 6 所示单障碍 4 m×4 m 的正方形实验环境，进行目标导航实验验证，实时记录并保存实验中动作细胞放电率和突触权重等相关参数。智能体的起点固定为环境边界左下角（0，0），圆形目标点半径为 0.25 m，正方形障碍物边长为 0.5 m。在智能体对环境的逐步探索过程中，获取环境反馈奖励信号，至智能体到达未知目标点或者最大探索时间结束时，采用第 2 章中 STDP 权重更新方法优化突触权值。多次训练后，智能体能够以较优路径到达未知目标点。

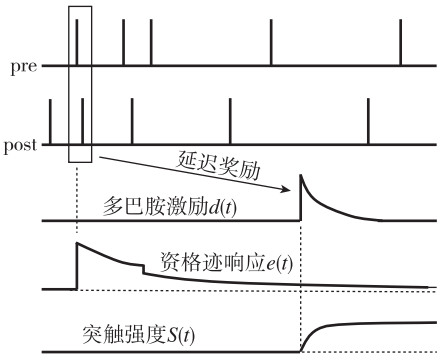
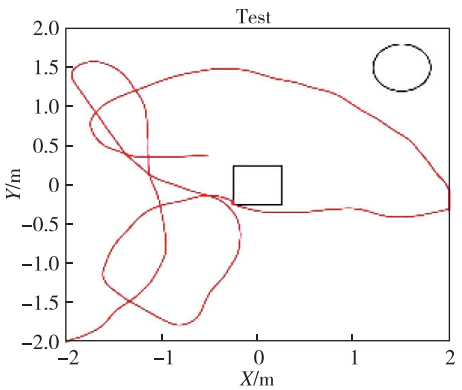
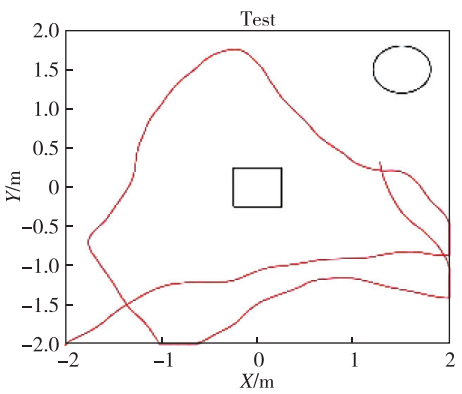


图 5 突触权值受资格迹影响示意图

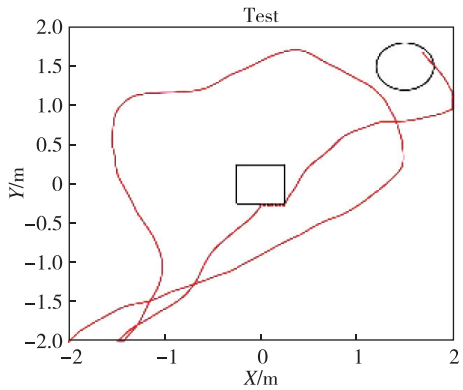
Fig. 5 Model of synaptic weights affected by the eligibility trace



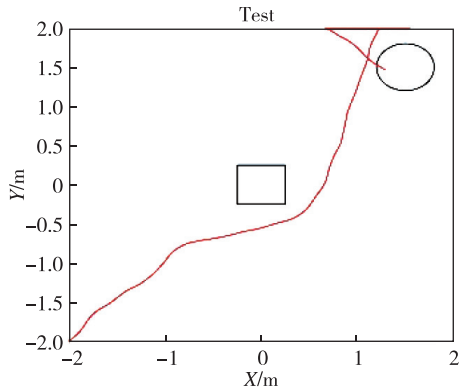
（a）1 次实验



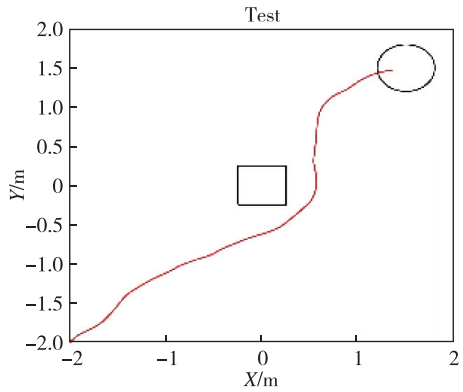
（b）4 次实验



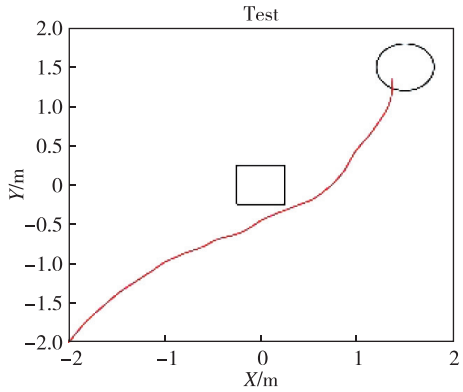
(c) 8 次实验



(d) 12 次实验



(e) 16 次实验



(f) 20 次实验

图 6　避障条件下的目标导航轨迹

Fig. 6　Target navigation trajectory in obstacle conditions

当训练开始时，智能体初始化前馈突触和横向突触权重，并采用随机策略对环境进行探索，同时学习从起点到未知目标点的导航方式。实验中将单次探索最大时间 T_{max} 设置为 50 s，智能体可以在 T_{max} 的最大持续时间内自由探索环境，如果在单次探索最大时间内发现未知目标并获得奖励，则一次探索提前终止，同时进入神经不应期，300 ms 后重新开始新一轮探索。为了在脉冲序列中模拟这个过程，在神经元放电之后的不应期内，通过抑制所有位置细胞的活性，将瞬时放电频率置为 0。

在 4 m×4 m 的正方形测试环境中，智能体通过多次训练学习，能够在陌生环境中迅速找到目标位置，并且实现从起始位置到目标位置的局部导航任务。图 6 所示为智能体在测试环境中不同训练次数下的目标导航轨迹，图 6(a)～(f)分别为 1 次、4 次、8 次、12 次、16 次及 20 次实验的智能体路径图，图 7 所示为对应的动作细胞导航策略图，图 8 所示为对应的位置细胞前馈突触平均权重图。在前 12 轮实验中，智能体由于初步进入陌生环境，尚未遍历整个环境，对于环境探索的随机策略导致了运动轨迹的随机性，同时运动策略和突触平均权重较为混乱，难以实现准确稳定的目标导航。在约 12 次训练后，智能体已经完成了隐藏目标点的探索过程，在面对环境中心的障碍物时，智能体运动策略已经显示出避让趋势，且障碍附近和远离最优路径的突触权值逐渐降低，此后实验中智能体具有了面向目标导航的能力。在第 20 次实验时，智能体已经实现了在障碍环境中的无碰撞面向目标导航任务。

根据实验结果可以看出，经过约 12 次训练之后，在没有外在路标参考情况下，智能体已经初步具备向目标点移动的目标导航能力，且靠近目标位置的前馈突触权值持续得到强化，表明智能体位置细胞-动作细胞模型已经记忆了障碍物和目标点位置，智能体在路径规划中动作细胞选择模型动作规划能力不断提高。经过 20 次左右的训练，智能体已经学会从起点以无碰撞路径实现面向目标的稳定避障导航。

为进一步验证本文提出的基于 STDP 学习规则的目标导航方法的有效性和收敛性能，在相同的单障碍实验环境中，采用目标导航算法中经典强化学习方法 Q-learning 算法对智能体进行路径寻优实验。对传统 Q-learning 模型和 STDP 模型分别进行10次80轮实验，再求取平均规划路径长度和

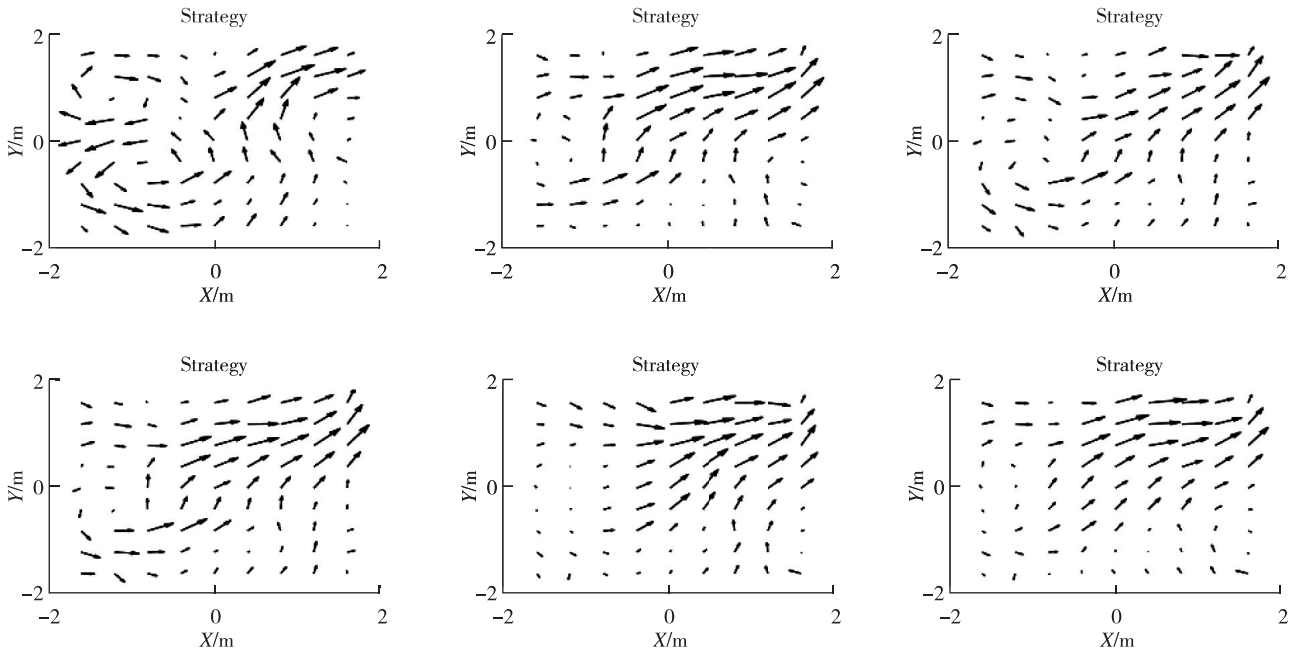


图 7 运动策略

Fig. 7 Behavioral strategy

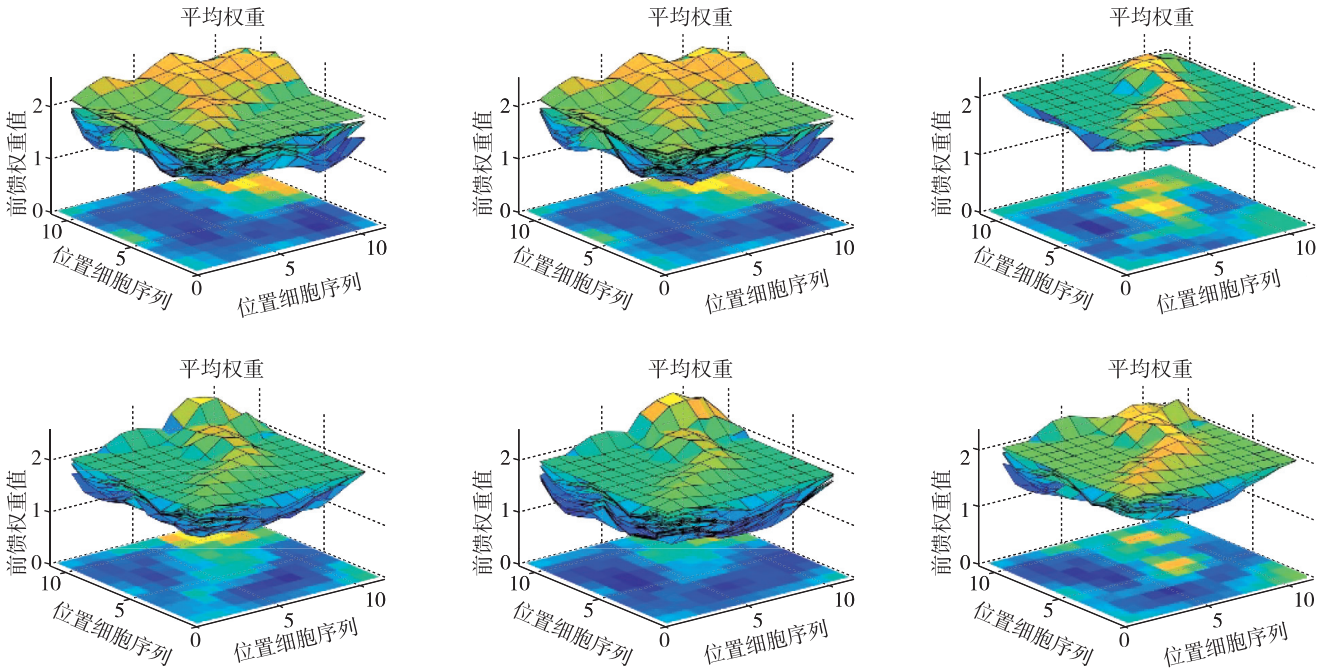


图 8 位置细胞前馈平均权重

Fig. 8 Average weight of place-cell feed-forward synapses

平均规划用时，其中平均规划路径长度 40 轮实验后均收敛，故截取前 40 轮实验结果。仿真实验结果如图 9 和图 10 所示。在更新地图动作细胞过程中，由于需要重复遍历整体陌生环境，采用 STDP 模型的智能体在初始路径规划长度上明显大于传统Q-learning 方法。而且，在后续得到目标点奖励后，通过 STDP 学习规则和资格迹延迟奖励，能够有效加速规划路径长度收敛，平均规划路径长度缩短了 15.9%，并且在算法规划时间上，STDP 模型对比传统 Q-learning 方法具有明显的优势。

为了研究 STDP 模型在复杂环境中的导航能力和环境适应性，通过迷宫仿真环境进行该问题的探索验证。模拟仿真环境如图 11(a) 迷宫环境所示，智能体从环境下方起点开始，且能够在迷宫中

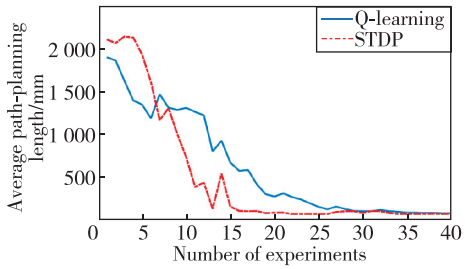


图 9　平均规划路径长度对比

Fig. 9　Comparison of average path-planning length

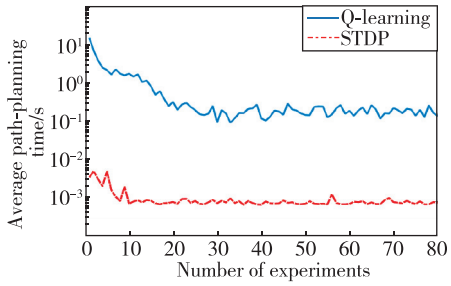
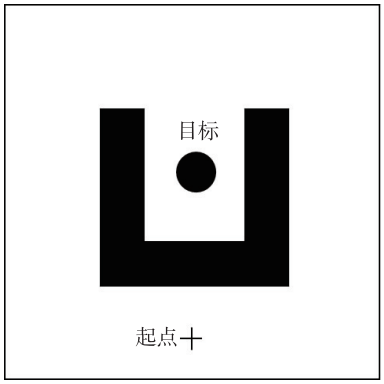


图 10　平均规划用时对比

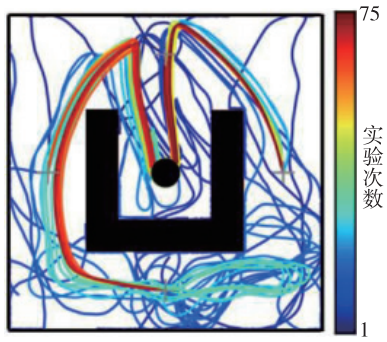
Fig. 10　Comparison of average path-planning time

自由探索。本实验在目标附近设置了黑色 U 形障碍，在智能体对环境的逐步探索过程中，获取环境反馈奖励信号，至智能体到达未知目标点或者最大探索时间结束时，采用第 2 章中 STDP 权重更新方法优化突触权值，多次训练后，智能体能够以较优路径到达未知目标点。图 11(b) 迷宫规划轨迹使用不同颜色表示了智能体从实验次数 1～75 的运行轨迹，仿真初始阶段(蓝线部分)学习如何避开墙壁和障碍物，当到达一次目标之后，后面的轨迹则会重复学习奖励高的轨迹，后续阶段(红色部分)表示智能体已学到的轨迹可以适应面向目标的迷宫环境导航。

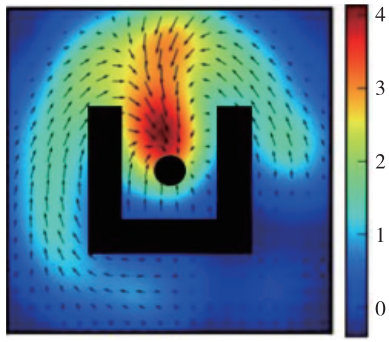
通过智能体中位置细胞到动作细胞的前馈连接权重大小，可以深入了解在导航过程中学习到的权重分布，导航运动策略如图 11(c) 所示。图中以不同颜色对智能体的权重强度进行区分，蓝色表示



(a) 迷宫环境



(b) 迷宫规划轨迹



(c) 迷宫运动策略

图 11　迷宫环境类脑目标导航实验

Fig. 11　Brain-inspired target navigation experiment in maze

强度最低，红色表示强度最高。在迷宫环境下的实验可以看出，智能体经过对环境的任意探索，40 次实验之后已经学习到面向目标导航的趋势，并学习到了适应 U 形迷宫的导航策略；在变更验证环境后，本文提出的 STDP 模型也能够适应多种障碍环境下的面向目标导航任务，初步具备多环境下的泛化导航能力。

4　结论

本文针对无先验知识空间中面向目标导航问题，主要工作如下：

1)根据动物导航过程生理学依据，构建了基于脉冲神经网络的海马体位置细胞和前额叶皮层动作细胞的特征表示模型，提出了一种基于 STDP 学习规则的面向目标类脑导航方法。

2)仿真实验表明，该模型能够有效地学习连续空间中面向目标位置的导航策略，实现障碍环境中稳定的学习和导航活动。本文所提出的类脑导航模型在单障碍环境中算法收敛性能优于传统 Q-learning 方法，平均路径规划长度缩短了 15.9%，平均路径规划用时为 30 ms，具有明显优势。迷宫环境中，本文模型在 40 次实验后也能适应面向目标导

航任务,对进一步发展未知环境下智能体面向目标导航方法具有较好的参考意义。

参考文献

[1] Kamil F, Hong T S, Khaksar W, et al. New robot navigation algorithm for arbitrary unknown dynamic environments based on future prediction and priority behavior [J]. Expert Systems with Applications, 2017, 86: 274-291.

[2] 马蓉. 基于改进 RRT 算法的无人机航路规划与跟踪方法研究[J]. 导航定位与授时, 2020, 7(1): 12-17.
Ma Rong. Research on a path planning and trajectory tracking method for UAVs based on improved RRT algorithm[J]. Navigation Positioning and Timing, 2020, 7(1): 12-17(in Chinese).

[3] Li J, Huang Y, Huang W, et al. Path planning for USV with FG-DA-RRT algorithm[C]// Proceedings of Chinese Automation Congress (CAC), 2019: 3211-3215.

[4] Hu B, Cao Z, Zhou M. An efficient RRT-based framework for planning short and smooth wheeled robot motion under kinodynamic constraints[J]. IEEE Transactions on Industrial Electronics, 2021, 68(4): 3292-3302.

[5] Lim H S, Fan S, Chin C K, et al. Particle swarm optimization algorithms with selective differential evolution for AUV path planning[C]// Proceedings of IAES International Journal of Robotics and Automation (IJRA), 2020.

[6] Li G, Chou W. Path planning for mobile robot using self-adaptive learning particle swarm optimization[J]. Science China (Information Sciences), 2018, 61(5): 267-284.

[7] Tang B, Xiang K, Pang M, et al. Multi-robot path planning using an improved self-adaptive particle swarm optimization[J]. International Journal of Advanced Robotic Systems, 2020, 17(5): 1-19.

[8] Zhou W, Xing Z, Bai W, et al. Route planning algorithm for autonomous underwater vehicles based on the hybrid of particle swarm optimization algorithm and radial basis function[J]. Transactions of the Institute of Measurement and Control, 2019, 41(4): 942-953.

[9] Kim B, Pineau J. Socially adaptive path planning in human environments using inverse reinforcement learning[J]. International Journal of Social Robotics, 2016, 8(1): 51-66.

[10] Guo S, Zhang X, Du Y, et al. Path planning of coastal ships based on optimized DQN reward function[J]. Journal of Marine Science and Engineering, 2021, 9(2): 210.

[11] O'Keefe J, Dostrovsky J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat[J]. Brain Research, 1971, 34(1): 171-175.

[12] Packard M G, Mcgaugh J L. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning[J]. Neurobiology of Learning and Memory, 1996, 65(1): 65-72.

[13] Franz M O, Mallot H A. Biomimetic robot navigation [J]. Robotics and Autonomous Systems, 2000, 30(1-2): 133-153.

[14] Wilson M A, Mcnaughton B L. Dynamics of the hippocampal ensemble code for space[J]. Science, 1993, 261(5124): 1055-1058.

[15] Wagatsuma H, Yamaguchi Y. Neural dynamics of the cognitive map in the hippocampus[J]. Cognitive Neurodynamics, 2007, 1(2): 119-141.

[16] Hyman J M, Zilli E A, Paley A M, et al. Working memory performance correlates with prefrontal-hippocampal theta interactions but not with prefrontal neuron firing rates[J]. Frontiers in Integrative Neuroscience, 2010, 4(2): 1-13.

[17] Zheng C, Hwaun E, Loza C A, et al. Hippocampal place cell sequences differ during correct and error trials in a spatial memory task[J]. Nature Communications, 2021, 12(1): 23765.

[18] Dong C, Madar A D, Sheffield M E J. Distinct place cell dynamics in CA1 and CA3 encode experience in new environments[J]. Nature Communications, 2021, 12(1): 2977.

[19] Cazin N, Scleidorovich P, Weitzenfeld A, et al. Real-time sensory-motor integration of hippocampal place cell replay and prefrontal sequence learning in simulated and physical rat robots for novel path optimization[J]. Biological Cybernetics, 2020, 114(2): 249-268.

[20] Izhikevich E M. Solving the distal reward problem through linkage of STDP and dopamine signaling[J]. Cerebral Cortex, 2007, 17(10): 2443-2452.

[21] Song S, Abbott L F. Cortical development and remapping through spike timing-dependent plasticity [J]. Neuron, 2001, 32(2): 339-350.

(编辑:孟彬)