基于人工智能的战场态势感知和作战辅助决策

陈晓婧，朱德政
（中国电子科技集团公司第二十八研究所，江苏南京 210007）

摘 要：现代战争对指挥人员的作战决策能力具有更高的要求，借助人工智能技术感知战场态势并给出最优策略有利于提高智能化作战水平。文章首先阐述了基于卷积神经网络的深度学习方法，该方法具备对复杂战场环境特征的感知和提取能力；然后基于DQN（Deep Q-Learning）算法，提出一个解决智能化决策问题的方法；最后分析了深度强化学习应用于作战辅助决策设计时的关键问题和解决思路。

关键词：人工智能；战场态势；卷积神经网络；深度强化学习

Battlefieldsituation sensing and operational auxiliary decision-making based on artificial intelligence

CHEN Xiao-jing，ZHU De-zheng
（The 28th Research Institute of China Electronics Technology Group Corporation，Nanjing Jiangsu 210007）

Abstract：Modern war has higher requirement on the operational decision-making ability of commanders. It is beneficial to improve the level of intelligent combat by using artificial intelligence technology to perceive the battlefield situation and give the optimal strategy. Firstly，the deep learning method based on convolutional neural network is described，which has the ability to perceive and extract the features of complex battlefield environment. Then，based on DQN（Deep Q-Learning）algorithm，a method to solve intelligent decision problem is proposed. Finally，the key problems and solutions in the application of deep reinforcement learning to combat aided decision design are analyzed.

Keywords：artificial intelligence；battlefield situation；convolutional neural network；deep reinforcement learning

0 引言

随着指挥体系和战争形态不断发展演化，战场逐渐进入“秒杀”时代，高速度、大机动和远射程装备发展趋势使得战争节奏显著加快，大规模体系作战、一体化联合作战，使得战争复杂性大幅上升，战场的非线性、跨域、网络化、无人化等特点，在时空范围、要素种类、行动节奏上都对决策、指挥和协同提出了极高要求[1]。在战场上，指挥员面对错综复杂、瞬息万变的战争数据和信息，人脑已经无法快速容纳和高效处理，严重影响了指挥人员对战场态势认知的准确性，进而影响指挥决策的及时性。为了解决上述问题，借助人工智能辅助决策是当今研究的热点，已取得了一定理论研究成果[2]。在作战指挥领域，通过运用人工智能，有效缩短观察—判断—决策—行动（OODA）环的时间，极大提升态势感知、情况研判、任务规划、方案生成、分析决策、行动管控等能力，提高作战指挥的效率和决策的科学性[3]。

如今，深度强化学习是一种将深度学习的感知能力和强化学习的决策能力相结合的人工智能技术[4]。深度学习和强化学习均为机器学习的分支，深度学习赋予机器人类的学习方式，通过训练大量的样本数据获得其内在规律，已经在图像识别、语音识别等领域超越了人类的感知能力；强化学习的基本思想是智能体在与环境的交互过程中，以获取最多的累计奖励值为目标，学习得到最优策略，在

作者简介：陈晓婧（1993— ），女（汉），甘肃兰州人，助理工程师，硕士，主要研究领域为指挥信息系统，邮箱 chen_ xj2021@163.com；
朱德政（1989— ），男（汉），江苏盐城人，工程师，硕士，主要研究领域为指挥信息系统，邮箱 zhudezheng1989@163.com。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

智能博弈和自动控制领域具有广泛的应用前景[5]。深度强化学习自提出以来，已经在许多需要感知高维度原始输入数据和决策控制的任务中取得了很好的效果。利用深度强化学习，在对物理域、社会域、知识域、认知域的解析及建模基础上，根据指挥员的意图进行快速优化，实现对战场态势精准感知和精确筹划。

基于以上军事需求和技术基础，本文针对深度强化学习技术在战场环境感知和作战辅助决策方面的应用方法进行研究。本文基于卷积神经网络的深度学习方式，感知海量的、复杂的战场环境信息，从中提取出有助于指挥决策的特征，以这些特征为输入进行强化学习，进而给出最优的决策。

1 基于深度学习的战场态势感知

深度学习是一种基于神经网络的机器学习方法，模仿人类大脑的“神经元网络”建立一个类似的复杂网络模型，该模型具有学习和分析复杂问题的能力。

1989 年 Yann Le Cun 等提出了卷积神经网络的概念[6]。2012 年 Hinton 等利用多层卷积神经网络解决了 ImageNet 数据库分类问题后，卷积神经网络受到了极大的关注[7]。卷积神经网络可以自适应学习多层网络获取的图像特征，为图像特征的甄别和提取工作带来了极大的便利。

1.1 卷积神经网络基本原理

神经网络通常是由输入层、输出层和多个隐层构成的，如图 1 所示。相邻的神经元之间通过权重连接，每个神经元采用非线性的激活函数计算输出值。

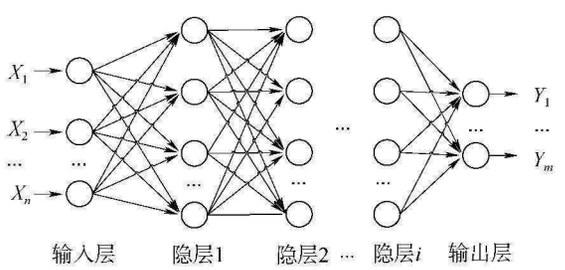

图 1 神经网络结构图

卷积神经网络的隐层包括卷积层、激活层、池化层和全连接层[8]。卷积层逐层对感知到的图像局部特征进行综合操作，最终获取全局特征信息；由于神经元的卷积操作是线性操作，激活层的目的就是引入非线性激活函数，以解决线性模型不能解决的特征分类问题；池化层利用某个区域中图像特征的平均值或最大值作为这个区域的特征值，避免数据量过大带来的过拟合问题，增强了网络的适应性；全连接层是卷积神经网络的最后一部分，是两层神经元之间全连接的神经网络，一般情况下卷积神经网络通过该层对图像特征进行分类和回归。

卷积层是卷积神经网络的核心功能模块[9]。卷积层两个主要结构特点是稀疏连接和权值共享，区别于全连接神经网络相邻两层的连接方式。

稀疏连接是指卷积层的神经元仅和相邻层的部分神经元相连。如图 2 所示，第 m 层的神经元只与第 $m-1$ 层的若干个相邻的神经元相连接，第 $m+1$ 层与第 m 层之间的连接也按照同样的规则，多个这样的层堆叠后，尽管神经元之间是局部连接的，仍然可以实现整体网络的相互关联。

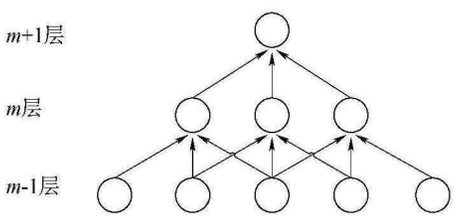

图 2 稀疏连接示意图

权值共享是指具有相同权值连接的神经元同属一种特征映射。如图 3 所示，m 层的三个神经元与各自相邻的三个神经元连接权重均为 W_1、W_2、W_3，权值共享后可以不考虑相同神经元在网络中的具体位置，避免了重复计算，减少了需要计算的参数个数。图中的情况如果只通过稀疏连接，需计算共 9 个权值参数，权值共享后只需要计算 3 个参数。从图像特征的角度分析，共享权值的神经元可以从二维图像中提取出相同的图像特征，多种初级特征通过后续网络的组合可以合成更为高级、抽象的特征，使得神经网络对于图像特征具有平移不变性，更加有利于增强网络对于特征的识别能力。

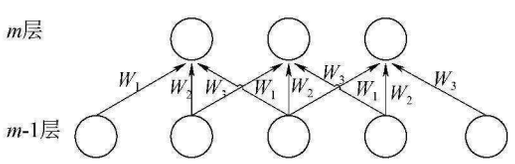

图 3 权值共享示意图

卷积神经网络在训练阶段，采用梯度下降法求得损失函数的最小值，通过网络的逐层反向传播来更新权值。该训练过程类似人脑的学习过程，通过反馈误差来不断调整学习策略，从而得到最优的神

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

经网络模型。

1.2 战场环境特征提取

对作战活动有影响的各种因素均属于战场环境的范畴，随着战争的发展演变，战场环境的内涵也不断拓宽。传统的战场环境主要包括地形地貌、敌我双方的兵力部署等信息，现代战场还需特别关注电磁环境特征。充分感知和理解战场态势是做出正确决策的必要条件，战场环境特征的主要分类结构如图 4 所示，其中每一类别包含若干要素。

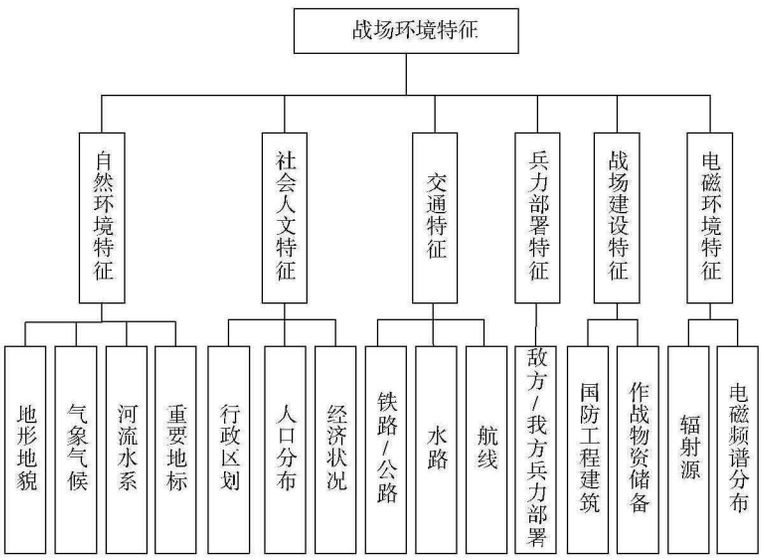

图 4 战场环境特征分类图

卷积神经网络的网络结构是影响网络模型性能的重要因素，网络性能随着深度的增加而提高，但过于复杂的网络计算复杂度较高，信息传播过程中会出现损失。常用的卷积神经网络结构有 LeNet、AlexNet、GoogleNet、VGGNet 等，大多数用于图像特征提取分类的卷积神经网络都设置为卷积层和池化层交替出现的形式，最后经过若干全连接层，输出特征分类结果。例如 LeNet 的网络结构共 7 层，各层的输入输出及卷积核参数如表 1 所示，32×32×1 意味着输入图像为 32×32 像素的单通道图像，最终得出 10 种图像特征，通常用于处理简单的图像。AlexNet 网络输入为 227×227 的 RGB 图像，最后输出 1000 种图像特征，通常用于处理较为复杂的图像。

表 1 LeNet 各层的输入输出

层序号	输入	核尺寸	核种类	输出
卷积层 1	32×32×1	5×5	6	28×28×6
池化层 1	28×28×6	2×2	6	14×14×6
卷积层 2	14×14×6	5×5	16	10×10×16
池化层 2	10×10×16	2×2	16	5×5×16
卷积层 3	5×5×16	5×5	120	120
全连接层 1	120	1×1	84	84
全连接层 2	84	1×1	10	10

针对战场环境特征提取，应首先对图像样本做出预先分类，根据图像包含的信息量及复杂程度选用适宜的卷积神经网络结构，然后对样本做预处理以满足网络的输入条件。

2 基于深度强化学习的作战辅助决策

传统的基于先验知识的作战辅助决策方法存在样本特征维度过高、决策时效性差和作战辅助决策模型普适性不强等诸多缺点，本文提出一种基于强化学习的思路解决智能辅助决策问题，采用卷积神经网络和 Q-learning 相结合的 DQN 算法，以卷积神经网络的输出代替 Q 值矩阵，有效解决了样本特征维度过高问题。DQN 算法是一种端到端的无模型算法，即通过深度学习感知战场环境特征，然后根据感知到的信息做出决策，不存在大量的中间运算过程，相比于传统的作战辅助决策方法，更具高效性和泛化能力。

2.1 强化学习基本设定

使用马尔科夫随机过程描述强化学习任务，如图 5 所示。指挥中心在做出作战决策时，给出动作策略 A_t 与战场环境进行交互，战场环境产生新的状态 S_{t+1} 并给出新的奖励 R_{t+1}，指挥中心根据新的数据修正自身的动作策略，如此循环迭代数次，指挥

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中心学习到所需的动作策略[10]。

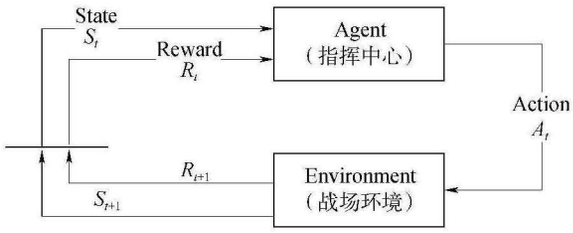

图 5　强化学习示意图

基于战场环境感知的作战决策任务具有马尔科夫性，即战场环境产生的新状态只与当前状态有关，而与历史状态无关，如公式（1）：

$$P(S_{t+1} \mid S_1, S_2, \cdots, S_t) = P(S_{t+1} \mid S_t) \tag{1}$$

一个马尔科夫随机过程由一个五元组构成，即 $< S, A, P, R, \gamma >$，其中：S 表示状态的集合；A 表示动作的集合；P 表示状态转移概率：$P^a_{SS'} = P[S_{t+1} = s' \mid S_t = s, A_t = a]$；$R$ 表示奖励函数：$R^a_s = E[R_{t+1} \mid S_t = s, A_t = a]$；$\gamma$ 为折扣系数，$\gamma \in [0,1]$，γ 越大代表越关心后续的奖励，γ 等于零代表只关心当前的奖励。

强化学习的最终目标是找到一个最优的策略π，使得累计的奖励值最大。策略π是状态为 s 下动作 a 的概率分布，如公式（2）：

$$\pi(a \mid s) = P[A_t = a \mid S_t = s] \tag{2}$$

在策略π下，无法通过当前的奖励判断策略的好坏，因此需要定义一个状态值函数 $v_\pi(s)$ 表示当前策略π的长期奖励，如公式（3）：

$$v_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right] \tag{3}$$

将 $v_\pi(s)$ 展开，如公式（4）：

$$\begin{aligned} v_\pi(s) &= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right] \\ &= E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid s_t = s] \\ &= E_\pi\left[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s\right] \end{aligned} \tag{4}$$

由于公式（5）和公式（6）：

$$E_\pi[r_{t+1} \mid s_t = s] = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P^a_{ss'} R^a_{ss'} \tag{5}$$

$$\begin{aligned} & E_\pi\left[\gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s\right] \\ &= \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P^a_{ss'} \gamma E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s'\right] \\ &= \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P^a_{ss'} \gamma v_\pi(s') \end{aligned} \tag{6}$$

推导得出公式（7）：

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P^a_{ss'}[R^a_{ss'} + \gamma v_\pi(s')] \tag{7}$$

由公式（7）可以看出状态值函数 $v_\pi(s)$ 为递归函数，在策略π下，当前状态的值函数可由下一个状态的值函数得出。公式（7）为贝尔曼方程的基本形态。同理可定义状态动作值函数 $q_\pi(s,a)$，它代表在当前状态 s 下，做出动作 a 的价值，如公式（8）：

$$q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right] \tag{8}$$

推导思路同状态值函数，可得公式（9）：

$$q_\pi(s,a) = \sum_{s' \in S} P^a_{ss'}\left[R^a_{ss'} + \gamma \sum_{a' \in A} q(s', a')\right] \tag{9}$$

称 $q_\pi(s,a)$ 为 Q 函数，智能化作战辅助决策问题可以归为一个动态规划问题，即找到使得 Q 函数值最大的最优策略 π^*。

2.2　作战辅助决策设计

2.2.1　DQN 算法流程

Q-Learning 算法是一种强化学习常用算法，主要思路是建立一个 Q 值矩阵，将每个状态动作对的 Q 函数值储存起来，然后根据 Q 函数值选取能够获得最大奖励值的动作，常用于状态空间和动作空间是离散且有限的情况。战场环境感知和指挥决策问题具有天然的复杂性，无法通过 Q-table 记录该场景下状态动作对的 Q 函数，考虑到卷积神经网络对复杂特征感知具有很好的效果，因此采取深度网络和强化学习相结合的 DQN 算法。

用一个卷积神经网络作为 Q 函数值的估计，神经网络的参数为 w，如公式（10）：

$$Q(s,a,w) = q_\pi(s,a) \tag{10}$$

利用公式（10）估计 Q 函数值存在一定的误差，引入损失函数 $L(w)$ 反映估计值和真实值之间的误差，如公式（11）：

$$L(w) = E[(\text{Target}Q - Q(s,a,w))^2] \tag{11}$$

其中 $\text{Target}Q = r + \gamma \max_{a' \in A} Q(s', a', w)$ 由目标网络 Target Net 产生，$Q(s,a,w)$ 由主网络 Main Net 产生，Target Net 和 Main Net 为结构相同、参数不同的两个卷积神经网络。指挥中心每次和战场环境交互时，根据损失函数最小原则采用梯度下降法更新 Main Net 参数，经过一定次数的迭代，将 Main Net 参数更新到 Target Net，即完成一次学习。

样本设计和奖励函数设计是深度强化学习应用于作战辅助决策设计的关键，好的样本设计为机器学习提供充足的素材，好的奖励函数设计指引最优策略的生成方向。

2.2.2　样本设计

指挥中心在过去各个时间点做出动作策略与战

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

场环境进行交互，战场环境给出下一个时间点的状态，将每一组这样的状态、动作和奖励记录下来作为指挥中心的经验，以四元组（s_t, a_t, r_t, s_{t+1}）的形式存入经验池。

在训练过程中，随机抽取小批量经验数据作为训练样本，这种方式称为经验回放机制。经验回放机制对历史数据进行备份，并且减小了数据样本之间的关联性。在过去各个时间点，指挥中心感知到的信息包含了那个时刻之前的所有知识，由于战场环境特征的复杂性和作战决策的严谨性，仅通过机器学习得到的知识还不足以做出最优的决策，指挥人员自身的作战经验也同样融入到历史样本数据中。

2.2.3 *奖励函数设计*

机器学习的目标是最大化奖励值，因此奖励函数的设计对最优决策的生成至关重要。奖励函数的设计必须能够真正表达我们的目标，而难点在于指挥中心每次与战场环境交互获取的奖励如何客观反映每个决策的优劣。

一种简单的方式是在得到阶段性作战结果（胜利或失败）的时候，根据量化的战损、战果给出奖励值，而在得到作战结果前给出的奖励均为零。这种稀疏奖励不是好的奖励设置，在海量的奖励值数据中只有个别数据包含有用信息，使得机器学习非常困难。

指挥中心做出的每个动作决策（进攻、撤退、掩护等），都应根据战场环境状态变化情况，获得相应的奖励值。这种奖励是对战场环境状态的奖励，而不是作战结果的奖励，这种方式和稀疏奖励形成了鲜明的对比，即使当前决策没有提供解决问题的完整方案，也能提供实时的、积极的反馈促进机器学习过程。

3 结束语

深度强化学习在战场态势感知和作战辅助决策中的应用是人工智能技术在解决军事决策智能化问题的一种探索。深度学习在态势感知领域，基于深度学习的战场态势大数据特征表示与挖掘技术、战场态势理解技术均有待突破，在指挥决策领域，深度强化学习的可解性有待提高，多实体协同决策技术、推理决策技术都有待提升。并且，研究表明错误的决策往往会带来非常严重的后果，战场环境的高度保密性增加了样本获取的难度，以上种种因素制约了深度强化学习在军事领域的应用。然而未来战争的智能化趋势已不可阻挡，近年来以 AlphaGo 为代表的人工智能取得的巨大成功证明了深度强化学习在面对复杂环境时的强大感知能力和决策能力，在未来智能化战争中具有巨大的发展潜力，取得了一系列应用成果，并日益成为军事领域智能化发展的技术基础与研究热点。

参考文献

[1] 金欣．指挥控制智能化现状与发展［J］．指挥信息系统与技术，2017，8（4）．

[2] 温百华．大国智能化战争理论解析［J］．指挥信息系统与技术，2020，11（6）：8-14.

[3] 郭圣明，贺筱媛．军用信息系统智能化的挑战与趋势［J］．控制理论与应用，2016，33（12）．

[4] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning［J］. Nature，2015，518（7540）：529-533.

[5] 刘祥，雷镜民，尚雷．战役级智能体训练系统［J］．指挥信息系统与技术，2020，11（3）：49-54.

[6] Lecun Y. Generalization and Network Design Strategies［C］. Connectionism in Perspective，1989.

[7] Krizhevsky A，Sutskever I，Hinton G E. ImageNet classification with deep convolutional neural networks［C］. International Conference on Neural Information Processing Systems. Curran Associates Inc，2012：1097-1105.

[8] Lecun Y，Bengio Y，Hinton G. Deep Learning［J］. Nature，2015，521（7553）：436-444.

[9] 王文竹，李智，来嘉哲，等．基于卷积神经网络的空间目标特性聚类分析研究［J］．指挥与控制学报，2020，6（2）：141-146.

[10] 曾隽芳，牟佳，刘禹．多智能体群智博弈策略轻量化问题［J］．指挥与控制学报，2020，6（4）：381-387.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net