****

机器人科学与工程学院

硕士研究生学位论文中期报告

|  |  |
| --- | --- |
| 学号： | 2002015 |
| 姓名： | 邵亚东 |
| 导师： | 贾子熙 |
| 学科类别： | □工学□全日制专业学位 |
| 学科/工程领域： | 机器人科学与工程 |
| 研究方向： | 强化学习 |
|  |  |
| 论文题目： | 基于强化学习的合金参数优化 |
|  | 研究 |
| 完成学分： | 29 |

**东北大学研究生院**

**2022年月日**

**排版格式**

一、中期检查报告首页各项按要求填写。

二、除原有标题外，中期检查报告正文部分使用非加粗宋体、小四号字；如需设置小标题，使用加粗宋体、小四号字。

三、段落首行缩进2字符。

四、撰写时，将各部分标题下括号中的说明文字删除。

五、如需打印，请A4纸正反面打印。

**一、论文工作总体完成情况**

**1. 开题报告中的论文主要研究内容**

（严格按照开题报告中的主要研究内容撰写）

研究构想与思路

总体框架见图1-1:

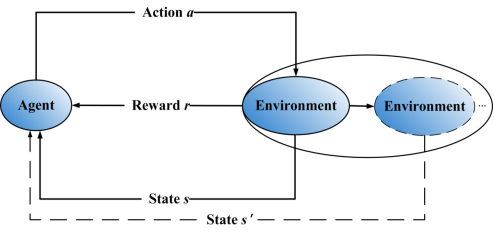


图1-1总体框架

DDPG结合了之前获得成功的DQN结构,提高了Actor-Critic的稳定性和收敛性。为了体现DQN的思想，每种神经网络都需要再细分为两个，Actor有估计网络和现实网络，估计网络用来输出实时的动作,供actor在现实中实行。而现实网络则是用来更新价值网络系统的。Critic这边也有现实网络和估计网络，他们都在输出这个状态的价值，而输入端却有不同，状态现实网络这边会拿着从动作现实网络来的动作加上状态的观测值加以分析，而状态估计网络则是拿着当时Actor施加的动作当作输入。DQN使用神经网络来近似值函数，即神经网络的输入是state s,输出是Q(s,a),∀a∈A。通过神经网络计算出值函数后，DQN使用ϵ−greedy策略来输出action（第四部分中介绍）。值函数网络与ϵ−greedy策略之间的联系是这样的：首先环境会给出一个obs，智能体根据值函数网络得到关于这个obs的所有Q(s,a)，然后利用ϵ−greedy选择action并做出决策，环境接收到此action后会给出一个奖励Rew及下一个obs。这是一个step。此时我们根据Rew去更新值函数网络的参数。接着进入下一个step。如此循环下去，直到我们训练出了一个好的值函数网络。DQN网络结构如图1-2。

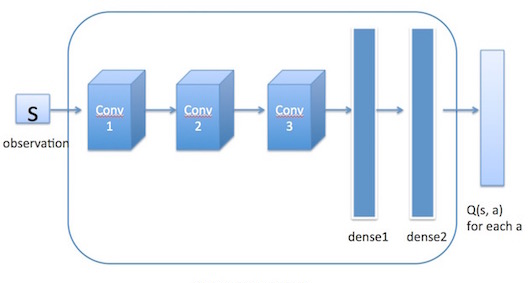


图1-2 DQN网络

本课题广泛参考了当前国内外研究的最新进展，采用改进DDPG算法对pybullet中的机器人模型进行步态训练，Agent中DDPG网络模型如图1-3。

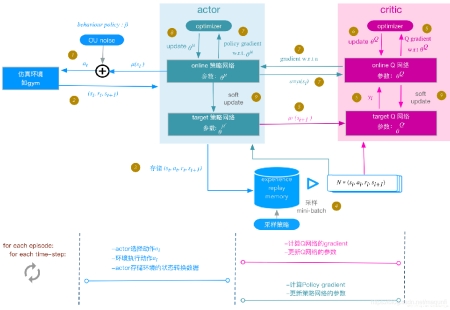


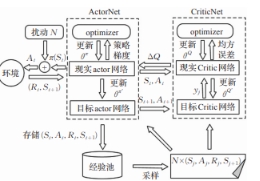
图1-3 DDPG网络结构

DDPG算法原理：

利用智能体与环境交互所得的奖惩信息来指导智 能体的行为的策略，叫做强化学习。DDPG是强化学习 中基于 Actor-Critic 的算法，它的核心思想是利用 Actor 网络用来生成智能体的行为策略，Critic 网络来评判动 作好坏，并指导动作的更新方向。DDPG结构中包含一个参数为 θπ 的 Actor网络和一个参数为 θ Q 的 Critic 网络来分别计算确定性策略 a = π(s|θπ ) 和动作 价值函数 Q(s,a|θ Q ) 。由于单个网络学习过程并不稳定，因此

|  |
| --- |
|  |
|  |  |

借鉴了 DQN 固定目标网络的成功经验，将 Actor 网络和 Critic 网络又各自细分为一个现实网络和一个目标网络。现实网络和目标网络结构相同，目标网 络参数以一定频率由现实网络参数软更新。DDPG结构如图1-4

图1-4 DDPG结构

以往的实践证明，如果只使用单个"Q神经网络"的算法，学习过程很不稳定，因为Q网络的参数在频繁gradient update的同时，又用于计算Q网络和策略网络的gradient.基于此，DDPG分别为策略网络、Q网络各创建两个神经网络拷贝,一个叫做online，一个叫做target。

策略网络：

*online*:*μ*(*s*∣*θμ*):*gradient updateθμ*

*target*:*μ*′(*s*∣*θμ*′):*soft* *update* *θμ*′

Q网络：

*online*:*Q*(*s*,*a*∣*θQ*):*gradien updateθQ*

*target*:*Q*′(*s*,*a*∣*θQ*′):*soft* *update* *θQ*′​

在训练完一个mini-batch的数据之后，通过SGA/SGD算法更新online网络的参数，然后再通过soft update算法更新 target 网络的参数。soft update是一种running average的算法：

*soft* *update*:*τ*一般取值0.001​

*θQ*′←*τθQ*+(1−*τ*)*θQ*′*θμ*′←*τθμ*+(1−*τ*)*θμ*′​

主要研究内容及拟解决的关键技术：

多足机器人是腿足机器人领域的研究热点和重要的研究方向，相对于轮式、履带式其他类型的移动方式的机器人而言，具有更高灵活性和环境适应性，且更贴近人类的特征，具有重要的研究价值。其中步行稳定性是双足机器人未来能应用于实际环境的基础和关键。自主式移动机器人可以工作在复杂非结构化环境中,无需人工干预,对环境无特定要求,具有高度自规划和自适应能力,这是一种有目的地自主式移动和完成任务的智能系统.在未知环境中,移动机器人需要拥有快速的学习能力,以提高对环境的适应能力,解决在自身不完备知识情况下的复杂问题.强化学习具有通过自我学习来逐渐获取智能行为的特征,在自主式移动机器人应用中受到广泛关注.强化学习通过不断试错寻求从状态到动作相适应的映射关系,最终获得最优化的动作状态策略集.

传统的双足机器人控制方法一般会采用模块化控制器，但模块化控制器设计难度较大，并且需要耗费大量时间进行参数调试。而基于数据动的强化学习能够自主进行学习优化，以实现最优的控制效果，可以有效解决模块化控制器的难点。考虑到双足机器人模型的复杂度和动态规划对实时性的要求，需要对DDPG进行优化。本文将改进的DDPG用于机器人控制中，本文的主要研宄问题如下：

1，双足机器人稳定自然动态行走；

2，双足机器人训练时间过长问题;

3，双足机器人行走速度问题。

为了解决以上问题，对DDPG的网络进行改进，采用RBF神经网络，以及双层经验池结构优化收敛速度。

RBF神经网络是一种三层神经网络，其包括输入层、隐层、输出层。从输入空间到隐层空间的变换是非线性的，而从隐层空间到输出层空间变换是线性的。1985年，Powell提出了多变量插值的径向基函数（RBF）方法。径向基函数是一个取值仅仅依赖于离原点距离的实值函数，也就是Φ（x）=Φ(‖x‖),或者还可以是到任意一点c的距离，c点称为中心点，也就是Φ（x，c）=Φ(‖x-c‖)。任意一个满足Φ（x）=Φ(‖x‖)特性的函数Φ都叫做径向基函数，标准的一般使用欧氏距离（也叫做欧式径向基函数），尽管其他距离函数也是可以的。最常用的径向基函数是高斯核函数 ,形式为k(||x-xc||)=exp{-||x-xc||^2/(2\*σ)^2)}其中x\_c为核函数中心,σ为函数的宽度参数,控制了函数的径向作用范围。

RBF网络的基本思想是：用RBF作为隐单元的“基”构成隐含层空间，这样就可以将输入矢量直接映射到隐空间，而不需要通过权连接。当RBF的中心点确定以后，这种映射关系也就确定了。而隐含层空间到输出空间的映射是线性的，即网络的输出是隐单元输出的线性加权和，此处的权即为网络可调参数。其中，隐含层的作用是把向量从低维度的p映射到高维度的h，这样低维度线性不可分的情况到高维度就可以变得线性可分了，主要就是核函数的思想。这样，网络由输入到输出的映射是非线性的，而网络输出对可调参数而言却又是线性的。网络的权就可由线性方程组直接解出，从而大大加快学习速度并避免局部极小问题。

径向基神经网络的激活函数可表示为公式2-1。

C:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps5.pngC:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps6.png 2-1

其中xp为第p个输入样本，ci为第i个中心点，h为隐含层的结点数，n是输出的样本数或分类数。径向基神经网络的结构可得到网络的输出见公式2-2。

C:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps7.jpgC:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps8.jpg

采用最小二乘的损失函数表示见公式2-3。

C:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps9.jpgC:\Users\Lenovo\AppData\Local\Temp\ksohtml9988\wps10.jpg

RBF神经网络结构简单，可以极大减少迭代中需要调整的参数数量，可以提高学习收敛速度。相比较其他神经网络，RBF网络更加适合实时性比较高的场景。RBF网络如图1-5。

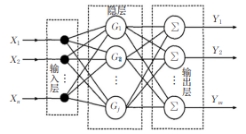


图1-5 RBF神经网络

优质的经验池样本对训练的速度至关重要，所以我打算设置了双经验池，优化经验池与非优化经验池，在不同的训练批次下，采用不同的经验池概率不一样，由于优化经验池充满时间比较长，所以前期使用非优质资源池，当随着训练批次增加，使用优化经验池的概率将会增加，在经验池存储数据时应该对存储的数据进行过滤筛选，把奖励值最大的几条数据放入优化经验池中。双层经验池流程如图1-6。

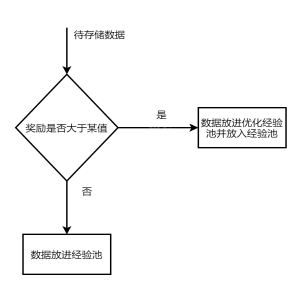


图1-6双经验池流程图

**2. 论文主要研究内容调整情况**

论文内容进行了调整，由原来的基于强化学习的双足机器人步态训练更改为基于强化学习的合金材料的参数优化研究，两个方向采用同种方法，应用背景更改，算法一致。

调整理由：

在此背景下，能更好的验证算法。

研究背景

主要研究内容为，对于mg-ln-nd 合金计算的一些未知参数优化。Pourbaix图和Tafe方程是其基本理论对于电化学腐蚀的热力学和动力学对腐蚀[1]的发展有巨大的世界性影响。Pourbaix图是电化学反应的热力学相图，绘制了电化学反应的电极电位与溶液pH值的反应;它在决定时非常有用金属与水溶液接触时的稳定化学形态。Tafe方程显示了电极电位与电极电位之间的关系电流密度;得到腐蚀速率(CRs)非常有用。然而，电化学腐蚀的热力学和动力学之间的关系非常弱。他们甚至不同意这在一定程度上阻碍了腐蚀的发展。这个问题可以归结为两个方面。一方面，腐蚀是一个溶解-电离-扩散-沉积(DIDD)过程金属/溶液(M/S)接口。当金属原子溶解形成时在阳离子中，不同类型的金属原子表现出不同的溶解速率，导致M/S界面的阳离子浓度不同。随后，不同浓度的阳离子导致不同水平的水解，导致M/S界面pH值发生显著变化。因此，过饱和度和临界表面能腐蚀产物的成核改变，诱发不同的腐蚀垢形成行为。溶解、电离和沉积过程是复杂且相互依赖的，它们的相互作用通常导致不同的腐蚀行为金属。然而，Tafel方程过于强调金属溶解和不够重视电离/水解和阳离子的沉积。定性的有源-无源极化曲线考虑了沉积过程(无源膜的形成)，但忽略了电离过程。因此，得到了有源-无源极化曲线未能在M/S的沉积中给出定量描述接口，并且到目前为止还不可能完全用数学公式来建模。根据Pourbaix图，在适当的pH值下区，Fe和Cr均表现为被动行为。然而,动能结果表明，Cr表现为被动行为，Fe则不表现为被动行为被动行为。Pourbaix图解也表明被动式奥氏体不锈钢的膜由Fe2O3, Cr2O3和NiO组成 许多已发表的结果表明，NiO很少被检测到XPS无源膜。而且，内层和外层是被动的薄膜分别富集Fe2O3和Cr2O3，也不存在基于Pourbaix图和Tafel方程的合理解释。另一方面，腐蚀是一个复杂的过程，受一个因素的影响 十几个环境因素和材料性能，如温度，pH值，腐蚀性气体的分压，浓度和类型 溶液中的离子，应力，流体，合金元素和微观结构材料。相互竞争和相互依存的关系上述影响因素之间，就像一个复杂的网络，给定量的确定造成了很大的障碍电化学腐蚀热力学与动力学的关系。对于一个复杂系统的建模，思路可以是分两个步骤描述。首先，将复杂的系统分成几个部分具有简洁物理意义的独立单位。其次，用腐蚀逻辑处理各单元之间的关系的过程。降维是处理复杂问题的有力工具系统受许多影响因素的影响。考虑DIDD在M/S界面中，影响腐蚀的因素可能是分为M/S界面阳离子浓度、pH值三个方面M/S界面，沉积行为。因此，本研究的目的是提出一个DIDD模型，以找到一个连接电化学热力学和动力学的桥梁腐蚀。然后，应用该模型来更好地理解CO220钢和3Cr钢的腐蚀。这一选择主要基于两点首先，CO2在水中[7]有二级电离。第二，CRs总是不遵循Arrhenius方程[8]。此外，在未来的研究中，该模型也将被应用于更好地理解一系列传统的和具有挑战性的腐蚀问题如H2S-CO2腐蚀(阴极过程)的机理与复杂的平衡和沉积竞争)，铬的影响不锈钢腐蚀行为的含量(八分之一定律)Cr含量)，从阳极溶解到钝化和反式钝化行为(极化的定量计算曲线)，盐雾试验得到的CR与现场的关系暴露试验(腐蚀寿命预测)和高不锈钢、铝合金、镁合金的腐蚀性能。如图2-1

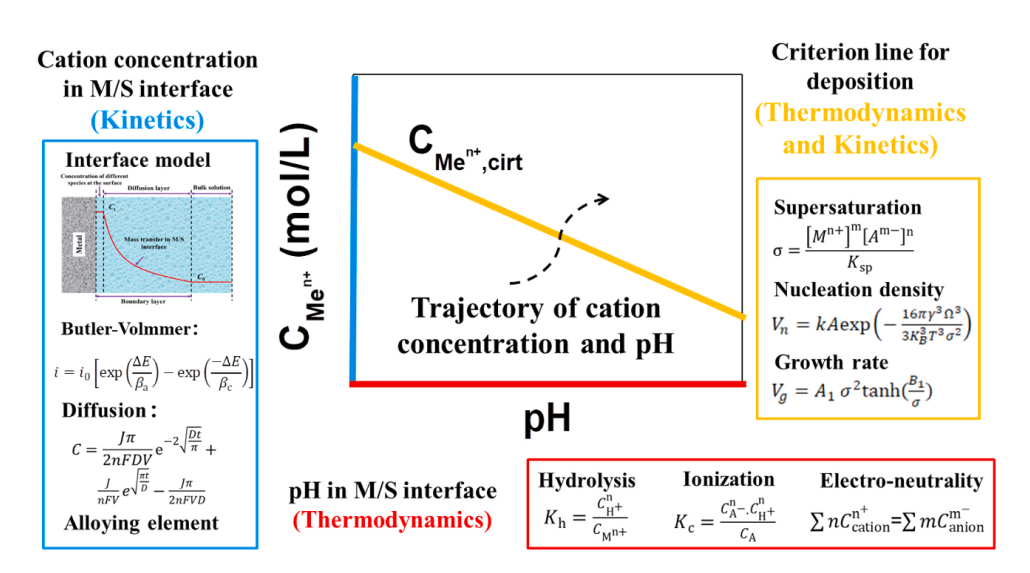


图 2-1

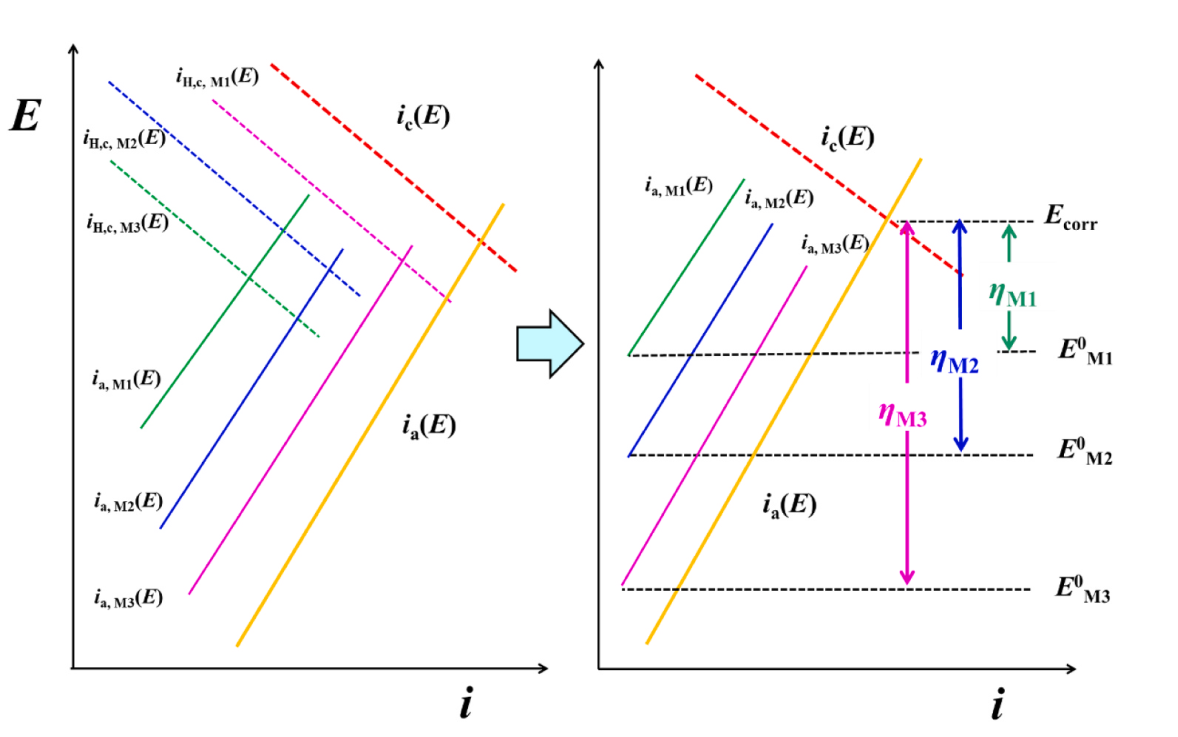


图 2-2

根据溶解-电离-扩散-沉积过程 在M/S界面中，十几个影响腐蚀的因素可以归纳为三个方面:M/S界面阳离子浓度、pH值 M/S界面，沉积行为。DIDD模型见 图2-1所示。纵坐标为M/S界面中离子浓度，代表金属阳极溶解动力学，为M/S界面中的pH值 界面代表溶液的热力学平衡，充当横坐标。该图使用溶解度积计算 腐蚀产物Ksp是腐蚀产物沉积(CLD)的判据线 腐蚀的规模。根据M/S界面中阳离子浓度和pH (tcp)的轨迹函数，确定了成核和生长速率的腐蚀尺度，即，如果TFCP超过CLD, a腐蚀垢可能沉积在合金表面并显示出来不同的地层行为，导致不同的保护性能。否则，合金将表现出阳极溶解行为导致耐蚀性差。同时在整个材料腐蚀过程中，有很多未知参数极大的影响了性能，只能进行人工调整，所以本文中，我采用强化学习的方法对参数进行优化，使之性能达到最佳。本文拟采用启发式算法和深度强化学习算法，以及启发式算法和强化学习算法相结合的方式对此研究背景进行训练研究。

（包括论文主要研究内容是否调整、调整的理由及依据、调整后的主要研究内容）

**3. 目前论文主要研究内容的总体完成情况**

**二、论文具体研究进展与所取得的阶段性成果**

本研究现在阶段已基本完成了启发式算法,模拟退火算法，深度强化学习算法DDPG算法。但是模拟环境尚未完成。

（包括论文研究工作的具体进展、取得的成果、与开题报告所定的研究内容和进展是否相符等内容）

**三、后续论文研究内容与研究计划**

**四、目前存在的问题与解决方法**

**五、论文的预期成果**

完成论文

导师签字：