IMPROVEMENT OF DEEP REINFORCEMENT MODELS USING EXTREME LEARNING MACHINE FOR AUTONOMOUS AGENTS IN UNSTRUCTURED ENVIRONMENT

BY

NOUAR ALDAHOUL

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy (Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

MARCH 2021

ABSTRACT

Creating an autonomous agent, that gets real observations such as sensory data and images from the surrounding environment and learns optimal sequential actions, has been considered as one of the main goals of Artificial General Intelligence (AGI). Deep (Hierarchical) Reinforcement Learning (HRL/DRL) can address this objective. Traditional deep reinforcement learning methods suffer from long learning and training time resulted from the need to fine-tune the weights iteratively in the network. This research investigates the previous problem by utilizing a random weights generation approach that is based on Extreme Learning Machine. This method benefits from the randomness of input weights and least square solution in output weights calculation to reduce the training time by an order of magnitude. Hierarchical ELM (H-ELM) and Local Receptive Field ELM (LRF-ELM) are recent versions of multilayer ELM to respectively learn and extract features by hierarchical learning scheme. They have outperformed other existing deep models in terms of learning time (speed). H-ELM's architecture was found to be similar to gradient-based (GB) auto-encoder without weights fine-tuning. However, H-ELM gives higher learning speed compared to the GB autoencoder. Moreover, LRF-ELM was found as similar to Convolutional Neural Network (CNN) without weights fine-tuning. It has outperformed the traditional CNN in the term of learning time. Therefore, in this research, the proposed method, which combines RL with H-ELM or LRF-ELM, is an efficient solution to approximate the action-value function and learn an optimal policy directly from visual data (images) in a short time. In addition, this research proposed a novel method called Convolutional H-ELM (CH-ELM) which is a combination of pre-trained CNN and H-ELM. This method has outperformed either CNN or H-ELM in terms of accuracy and RMSE. The experimental results have been analyzed and evaluated in different applications such as target reaching arm, 2D maze navigation, slide puzzle game , objects sorting, and rock-paper-scissor game. The data samples have been trained and tested to investigate the robustness of the proposed systems. It was found that the proposed models can reduce the learning time by an order of magnitude in various tasks without degrading the performance. The big improvement in learning speed in the proposed method can neglect the slight drop in accuracy in few tasks compared to traditional methods. Therefore, the proposed method can balance the trade-off between learning speed and good performance. In addition, it is able to run on traditional CPUs that are available in the most of the low cost embedding systems.

خلاصة البحث

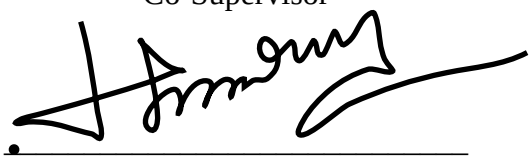
تعد عملية ايجاد عميل ذاتي القرار قادر على رصد المشاهدات الحقيقية كبيانات الحساسات و الصور من البيئة المحيطة وتعلم سلسلة من الأفعال المثالية من أهم أهداف الذكاء العام الصنعي. استطاع التعلم المعزز الهرمي (العميق) تحقيق هذا الهدف. تعاني الطرق التقليدية للتعلم المعزز العميق من طول زمن التعلم والتدريب الناتج من الحاجة إلى توليف الأوزان بشكل متكرر في الشبكة. في هذا البحث تم دراسة هذه المشكلة بالاستفادة من مفهوم توليد الأوزان العشوائية القائم على خوارزمية ELM. هذه الطريقة تستفيد من عشوائية أوزان الدخل ومن الحل القائم على المربعات الصغرى في حساب أوزان الخرج لانقاص زمن التدريب عدد من المرات. تم الاستفادة من البنية الهرمية H-ELM و حقول الاستقبال المحلي ELM-LRFs وهما إصداران حديثان لشبكة ELM متعددة الطبقات ويتم فيهما تعلم الميزات أو استخراجها عن طريق التعلم الهرمي. هذه النماذج تفوقت على نماذج التعلم العميق الموجودة مسبقاً من خلال زمن التعلم (سرعة التعلم). إن بنية H-ELM تشبه المرمز الألي القائم على هبوط الانحدار (Gradient Descent based auto encoder) ولكن بدون الحاجة إلى التوليف. ومع ذلك فإن البنية H-ELM تتمتع بسرعة تدريب أفضل مقارنة مع الأخير. كما تم استخدام حقول الاستقبال المحلي كبنية بديلة مشابهة للشبكات العصبونية الالتفافية CNN ولكن بدون الحاجة إلى توليف الأوزان وضبطها. وقد تم إثبات تفوقها على CNN من حيث سرعة التعلم. لذا فإن الطريقة المقترحة في هذا البحث والتي تعتمد على دمج التعلم المعزز مع H-ELM أو ELM-LRFs هي حل فعال لتقريب تابع قيم الأفعال (Action Value Function) و تعلم الاستراتيجية المثالية بشكل مباشر من المعطيات المرئية (الصور) كمدخل للنظام خلال زمن قصير. بالإضافة لما سبق تم في هذا البحث اقتراح طريقة جديدة تدعى CH-ELM و الذي تم فيه دمج الشبكة الالتفافية المدربة مسبقا مع الشبكة الهرمية العشوائية H-ELM و قد اثبت هذا النموذج تفوقه على كل من الشبكة الالتفافية CNN و شبكة H-ELM من حيث الدقة وجذر متوسط مربع الخطأ. تم في هذا البحث تحليل وتقييم النتائج التجريبية في تطبيقات مختلفة كتطبيق ذراع روبوتية يبحث عن الهدف وعميل في متاهة ثنائية البعد ولعبة البازل المتزحلق وفرز الأغراض المختلفة ولعبة حجر ورق مقص. تم تدريب وفحص عينات من البيانات للتأكد من متانة النظام المقترح. وجد أن النموذج المقترح قادر على انقاص زمن التعلم عدد من المرات في مهام مختلفة دون تراجع مستوى الأداء. إن التحسن الكبير في سرعة التعلم في الطريقة المقترحة سمح بإهمال التراجع الطفيف في الدقة في بعض المهام بالمقارنة مع الطرق التقليدية. لذلك فإن الطريقة المقترحة تستطيع موازنة المقايضة بين سرعة التعلم والاداء الجيد. بالإضافة الى أنها قابلة للتنفيذ على المعالجات التقليدية المتوفرة في معظم الأنظمة المضمنة منخفضة التكلفة.

APPROVAL PAGE

The thesis of Nouar AlDahoul has been approved by the following:

on behalf Zaw Zaw Htike
Supervisor

Amir Akramin bin Shafie
Co-Supervisor



Md. Raisuddin Khan
Internal Examiner

Sazali Bin Yaacob
External Examiner

Mohd Rizon Mohamed Juhari
External Examiner

Imad Fakhri Taha Alshaikhli
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Nouar AlDahoul

Signature .....Nouar AlDahoul..... Date 10/02/2021

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

IMPROVEMENT OF DEEP REINFORCEMENT MODELS USING EXTREME LEARNING MACHINE FOR AUTONOMOUS AGENTS IN UNSTRUCTURED ENVIRONMENT

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understood the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Nouar AlDahoul

……Nouar AlDahoul..…….. 10/02/2021

Signature Date

ACKNOWLEDGEMENTS

My first word of thanks goes to ALLAH Who gave me the chance and the health to accomplish this work.

I do not forget to thank Assoc professor Zaw Zaw Htike who guided me throughout the path of this research. Additionally, I also thank Prof. Md. Raisuddin Khan and Prof. Amir Akramin Shafie who supervised me after VIVA and helped to improve the quality of this work.

Finally, this study should not have seen the light without the efforts and support given to me by my mother. I give many thanks to her and dedicate this modest work to her. She was so patient with me in this journey and encouraged me to accomplish this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

a	Current Action
$E\ (\Theta)$	Loss Function
$H\ \dagger$	Moore Penrose Inverse
H	Hidden Layer Output
Q	Action Value Function
R	Immediate Reward
s	Current State
s'	Next State
T	Target
V	State Value Function
w	Input Weights
α	Learning Rate
β	Output Weights
γ	Discount Factor
λ	Regularization Coefficient
π	Policy
ϵ	probability of greedy policy
Θ_k	Network's Parameters

LIST OF ABBREVIATIONS

A3C	Asynchronous Advantage Actor-Critic
AGI	Artificial General Intelligence
AI2THOR	An Interactive The House Of inteRactions
ANN	Artificial Neural Network
AUV	Autonomous underwater vehicle
BP	Back Propagation
BRL	Bayesian Reinforcement Learning
CH-ELM	Convolutional Hierarchical Extreme Learning Machine
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DDPG	Deep Deterministic Policy Gradient
DFQ	Deep Fitted Q
DNN	Deep Neural Network
DP	Dynamic Programming
DQN	Deep Q Network
DRL	Deep Reinforcement Learning
FA	Function Approximation
FASTA	Fast Iterative Shrinkage Thresholding Algorithm
GPU	Graphical Processing Unit
HAR	Human Activity Recognition
HMM	Hidden Markov Model
HRL	Hierarchical Reinforcement Learning
ICA	Independent Component Analysis
LRF-ELM	Local Receptive Field Extreme Learning Machine
LSPI	Least Square Policy Iteration
LSTD	Least Square Temporal Difference
LSTM	Long Short-Term Memory
MC	Monte Carlo
MDP	Markov Decision Process
ML	Machine Learning
MLELM	Multi-Layer Extreme Learning Machine
MLFN	Multi-Layer Forward Network
MLP	Multilayer Perceptron
NFQ	Neural Fitted Q Iteration

OS-ELM	Online Sequential Extreme Learning Machine
PCA	Principle Component Analysis
PDF	Probability Distribution Function
PGRL	Policy Gradient Reinforcement Learning
POMDP	Partial Observable Markov Decision Process
PPO	Proximal Policy optimization
RESNET	Residual Networks
RAE- ELM	Robust AdaBoost.RT based Ensemble Extreme Learning Machine
RIWG	Random Inputs Weight Generation
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RVB LSPI	Least Squares Policy Iteration based on Random Vector Basis
RWG	Random Weight Generation
SAE	Stacked Autoencoder
SARSA	State Action Reward State Action
S-CNN	Supervised Convolutional Neural Network
SDA	Stacked Denoising Autoencoder
SGD	Stochastic Gradient Descent
SLFN	Single Layer Forward Network
SVM	Support Vector Machine
TCSC	Thyristor-Controlled Series Capacitor
TD	Temporal Difference
UAV	Unmanned Aerial Vehicle
UVFA	Universal Value Function Approximator

CHAPTER ONE

INTRODUCTION

1.1 RESEARCH BACKGROUND

Most of the robot control and planning issues are represented by uncertainties of sensing and acting devices of robots. Recently, applications in robotics have moved from a highly controlled environment to unstructured environments. Working in an unstructured environment is considered as a challenging task for autonomous agents. The unstructured environment is characterized by one or more of these factors: complexity, uncertainties and/or a high dimensional state space (Katz, Kenney, & Brock, 2008).

Creating an autonomous agent, that gets real observations such as sensory data, images, videos or audio from the environment and learns optimal sequential actions, has been considered as one of the main goals of Artificial General Intelligence (AGI) (Kuhnberger et al., 2009). Deep (Hierarchical) Learning and Reinforcement Learning (RL), two subfields of Machine Learning (ML), can address this objective. Deep learning can learn features that are relevant to discriminative action values. On the other hand, Reinforcement Learning learns to map the learned features to optimal actions (Kuhnberger et al., 2009).

Reinforcement learning (RL) is an active research area in machine learning, artificial intelligence, and neural network (Sutton & Barto, 1998; Kaelbling, Littman, & Moore, 1996). RL differs from other learning methods such as supervised learning that needs to have a pair of inputs/outputs to find the model. In RL, the agent first observes the state of the environment and generates action. The environment moves the

agent to a new state and gives it a reward which is a scalar value that evaluates the action in the current state. The idea of learning by interacting (trial-and-error) with the environment is used when there is no direct teacher but only sensory and motor connections with the surrounding (Sutton & Barto, 1998; Kaelbling, Littman, & Moore, 1996). The objective of RL is to map different situations to proper actions that maximize a reward.

RL has been used for a while to find an optimal policy or optimal series of actions in a low dimensional environment (Sutton & Barto, 1998; Kaelbling, Littman, & Moore, 1996). This policy results from the interaction between the agent and the environment by doing actions and getting positive and negative rewards. The trial – error technique formulates the RL utilizing the MDP (Markov Decision Process) model which describes the agent-environment interaction. Figure 1.1 shows the model of interaction between robot and environment in the unstructured environment. The value iteration method is one of the RL methods that use the value function, which is discounted accumulated rewards, to describe the value of the actions (Bellman, 1957). The objective is to optimize the policy by choosing actions that have maximum action values.

When the input is high dimensional data such as images, traditional RL is suffering from many problems (Bellman, 1957; Keogh & Mueen, 2017). Therefore, the method of deep reinforcement learning was found as an alternative solution. A combination of deep and reinforcement learning is the foundation of AGI (Kuhnberger et al., 2009). Deep reinforcement learning gets a stream of raw data (sensors or camera's images) and reacts to the environment with a sequence of control actions to finally achieve the desired goal. The existing deep reinforcement learning models were found to give a good performance, but they require long training time even if the algorithms run on a powerful Graphical Processing Unit. Therefore, there is still a need to have