مجلة جامعة دمشق للعلوم الهندسية
المجلد 39- العدد 3 (2023) 272-280

Damascus university Journal for engineering sciences
Vol 39 No.3 (2023) : 272-280

دراسة تأثير عدد طبقات الشبكة العصبونية العميقة في تحسين مكافأة روبوت التعلم المعزز

أسامة ابراهيم[1]، د.م. سمير كرمان[2]، د.م. رؤوف حمدان[3]

[1]طالب دكتوراه في قسم هندسة الحواسيب والأتمتة- جامعة دمشق.

[2]استاذ مساعد في قسم هندسة الحواسيب والأتمتة- جامعة دمشق.

[3]مدرس في قسم هندسة الحواسيب والأتمتة- جامعة دمشق.

الملخص

تعتبر خوارزمية Q learning في التعلم المعزز احدى الخوارزميات التي تسمح للروبوت بتعلم البيئة المحيطة دون الحاجة الى عينات تدريب مسبقة بمبدأ المكافأة والعقاب للروبوت من خلال التفاعل مع البيئة. تم في هذا البحث دراسة تأثير عدد الطبقات الخفية المستخدمة في الشبكة العصبونية لتحسين مكافاة الروبوت حيث اظهرت المحاكاة انه يمكن بزياده عدد الطبقات الخفية للشبكة العصبونية العميقة المستخدمة وضبط بعض المعاملات العليا فيها زياده مكافاة الروبوت وبالتالي الحصول على افضل مسار لتحقيق الهدف.

تاريخ الإيداع: 2022/4/25
تاريخ القبول: 2022/7/4

الكلمات المفتاحية: التعلم المعزز، شبكة عصبونية عميقة، تحسين المكافأة.

Studying the effect of the number of layers of a deep neural network in improving the reward of a reinforcement learning robot

Osama Ibrahim
Dr. Eng.Samir Karaman, Dr. Eng.Raouf Hamdan
[1]PhD student in Computer and Automation Engineering Department Damascus University.
[2]Assist. prof in Computer and Automation Engineering Department, Damascus University.
[3]Lecturer in Computer and Automation Engineering Department, Damascus University.

Received: 25/4/2022
Accepted: 4/7/2022

Abstract

The Q learning algorithm in reinforcement learning is one of the algorithms that allows the robot to learn the surrounding environment without the need for prior training samples with the principle of reward and punishment for the robot through interaction with the environment. Increasing the number of hidden layers of the deep neural network used and adjusting some of the higher parameters in it can increase the reward of the robot and thus obtain the best path to achieve the goal.

keywords: Reinforcement learning, deep neural network, reward Enhancement.

1- المقدمة:

يعتبر التعلم المعزز (Reinforcement Learning) محور بحث فعّال في مجال التعلم الآلي ويتم نشر العديد من الأبحاث سنوياً في هذا المجال. قامت مجموعة صغيرة من الباحثين من شركة (Deep Mind) بنهاية عام 2013 بإصدار الورقة البحثية (Playing Atari with Deep Reinforcement Learning) [1] وبعد عدة أشهر قامت شركة google بشراء شركة (Deep Mind) بمبلغ مالي كبير نسبياً. في عام 2016 قام Alph Go برنامج شركة (Deep Mind) التابعة لشركة google بهزيمة المحترف Lee Se - dol من كوريا الجنوبية في لعبة GO المعروفة، في وقتٍ سابق من هذا العام، استُخدمت مصطلحات: الذكاء الصنعي، وتعلُّم الآلة، والتعلم العميق، في وسائل الإعلام لوصف كيف فازت شركة Deep Mind، ويُعتَبر فعلاً كلٌّ منها جزءاً من السبب في فوز Alpha Go على Lee Se-dol .

تعتبر خوارزمية Q learning في التعلم المعزز من افضل الخوارزميات المستخدمة للعمل في بيئات غير معروفة بالنسبة للعميل بحيث يتلقى العميل مكافأة في الخطوة الصحيحة أو عقوبة في الخطوة الخاطئة وبالتالي يسعى العميل إلى تراكم المكافأة في كل خطوه للوصول الى الهدف بأفضل ما يمكن، ومن هنا يختلف التعلم المعزز [2] عن الانواع الاخرى من التعلم الآلي كالتعلم بإشراف حيث يكون العميل يعرف المخرجات والمدخلات كما في عمليات التصنيف وكذلك عن التعلم بدون اشراف حيث يكون العميل يعرف بعض المدخلات ويقوم بترتيبها حسب تشابهها مثل التعرف على الصور في شبكات التواصل الاجتماعي (facebook).

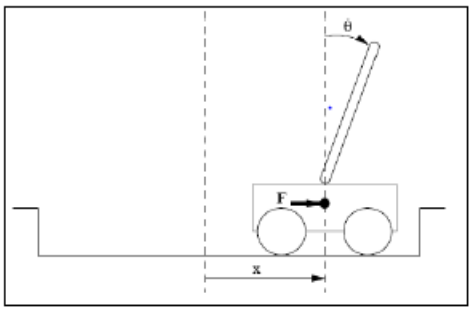
2- الأعمال السابقة:

يمكن تصنيف خوارزميات التعلم المعزز الى خوارزميات غير المستندة الى نموذج model free ومنها مقاربات مستنده الى سياسة [3] policy based مثل خوارزمية policy gradient في هذه المقاربة يتم تعلم تابع السياسة policy function هذا التابع هو طريقة ربط كل حالة والفعل المناسب لها وهناك مقاربات مستنده الى قيمة value based هنا يهدف العميل الى تحسين تابع القيمة v(s) يعرف تابع القيمة بانة التابع الذي يعطينا المكافاة المستقبلية المتوقعة الاكبر التي يمكن ان يحصل عليها العميل في حالة معينة مثل خوارزمية Q learning. اما التعلم المعزز المستند الى نموذج model based الهدف منه تحديد السياسة التي تعطينا أفضل نتائج وتحسين مكافأة العميل حيث يتم الاعتماد على نموذج للانتقالات بين الحالات مثل نموذج ماركوف لصنع القرار MDP [4] وخوارزمية مونت كارلو Monte Carlo أن الفرق الاساسي بين التعلم المعزز غير المستند الى نموذج والتعلم المعزز المستند إلى نموذج أنه في التعلم المعزز المستند الى نموذج يتم نمذجة البيئة اي صنع نموذج لتصرف البيئة المحيطة بالعميل وبالتالي تحديد السياسة هذا النمط يتطلب نمذجة بيئة جديده عند دراسة كل حالة وهذا ما يصعب القيام به دائماً، فركزنا في دراستنا على استخدام خوارزمية Q من النمط غير المستند الى نموذج [5] مع اضافة الشبكة العصبونية العميقة فنحصل على تعلم معزز عميق يعتمد بشكل كامل على الشبكة العصبونية. هناك العديد من انواع الشبكات العصبونية التي تدعم التعلم الألي منها الشبكات العصبونية التلافيفية CNN [6] التي تستخدم في مجال معالجه الصور والفيديو والشبكات العصبونية التكرارية RCC التي تدعم معالجة الصوت [7] في هذا البحث استخدمنا خوارزمية Q بالإضافة الى الشبكة العصبونية DNN .

3- ادوات البحث وطرائقه

تم استخدام مجموعة من المكتبات: TensorFlow, Keras,numpy التي تدعم التعلم العميق وتم استخدام مكتبة matplotlib لرسم المنحنيات البيانية وتم العمل في مكتبة gym حيث توفر هذه

المكتبة في python عدده نماذج لاختبار النتائج وقد اخترنا بيئة CartPole-v1 كما في الشكل (1)



الشكل (1) بيئة CartPole-v1

وتكون المعادلات التي تصف الحركة فيها على الشكل التالي:

$$2\ddot{x} + \ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta = f_x$$
$$\ddot{x}\cos\theta + \ddot{\theta} + \sin\theta = 0$$

حيث ان Θ زاوية العصا

X موضع العربة , f قوة جر العربة

4- توصيف العمل في بيئة CartPole-v1

عناصر التعلم المعزز هي العميل والبيئة التي يتحرك فيها والإجراءات التي يقوم بها والحالات والمكافآت والسياسة التي يعلم نفسة التحرك بها اعتمادا على المكافاة التي يحصل عليها وبالتالي تكون العناصر في بيئة CartPole-v1 هي:

التفاعل مع البيئة: يكون بالتحرك يمينا او يسارا

الحالات: لدينا اربع حالات: موضع العربة سرعه العربة

زاوية العصا سرعة راس العصا

المكافاة: +1 عندما تكون زاوية العصا مزاحه

اقل من 15 درجه عن وضع التوازن والعربة بعيده

اقل من 2.4 وحده عن المركز, −1 غير ذلك

السياسة : تحاول العربة التحرك يمينا او يسارا لتحافظ

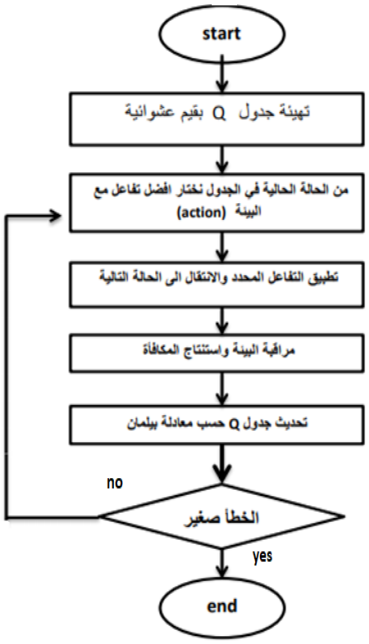
على استقرار العصا بزاوية اقل من 15 درجه.

5-خوارزمية Q Learning

تعد خوارزمية Q learning من اهم خوارزميات التعلم المعزز غير المستند الى نموذج التي تسمح للعميل بالعمل في بيئة غير معروفة بفاعلية كبيره بحيث يراكم المكافئات ليصل للهدف بأفضل طريقة الصيغة الرياضية لها بالاعتماد على معادلة بيلمان[8]

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left(R_{t+1} + \gamma \max_{at+1} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right)$$

وتكون مراحل خوارزمية Q حسب الشكل (2)

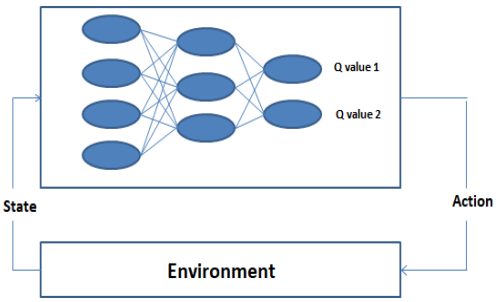


الشكل (2) مراحل حلقة Q لنظام cart pole

6-خوارزمية Deep Q Learning

يمكننا التفكير ان الشبكة العصبونية العميقة لها تأثير جيد على استخراج الميزات المعقدة وعند الدمج بين تقنيتي Deep Learning و Reinforcement Learning نحصل على Deep Q Learning حيث ان التعلم العميق Deep learning هو فرع من أفرع التعلم

الآلي يعتمد بالكامل على الشبكات العصبونية، حيث أن الشبكة العصبونية ستحاكي الدماغ البشري، لذا فإن التعلم العميق هو أيضاً نوع من أنواع محاكاة العقل البشري. وحيث ان التعلم العميق هو تعلم خاضع لإشراف ويحتاج الى تعلم مجموعة التدريب فان التعلم المعزز لا يتطلب مجموعه تدريب لإرجاع قيمة المكافاة بحيث نحصل عليها من البيئة فقط. وبالدمج بين التقنيتين نستطيع ان نقلل من الحاجه للميزات الهندسية المحددة مسبقا ونزيل التكاليف غير الضرورية .وهنا تبرز بعض السلبيات من حيث كمية البيانات الضخمة وبالتالي كلفة عالية لعملية التدريب ووقت طويل للتدريب واحتياج معالجات قوية وعليه يتمثل الفرق الرئيسي بين خوارزمية Q learning وخوارزمية Deep Q learning بحيث يتم بالخوارزمية العميقة استبدال جدول Q العادي بشبكة عصبونية عميقة تتم تهيئة مداخلها التي تمثل حالات البيئة المستخدمة بالزوج (تفاعل، قيمة Q) ويصبح نظامنا المقترح حسب الشكل (3)

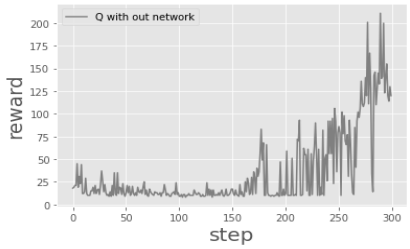


الشكل (3) المخطط الصندوقي للنظام المقترح بعد اضافة الشبكة العصبونية.

حيث ان مداخل الشبكة هي الحالات الحالية الممكنة للعميل حسب البيئة (CartPole-v1) وهي اربع حالات ومخارج الشبكة هي مخرجين تمثل التفاعلين الممكنين في البيئة اي تحرك العربة يمينا او يسارا .في المرحلة الثانية من خوارزمية Deep Q learning يتم اختيار الاجراء المناسبة في كل مرحلة حسب خوارزمية epsilon greedy التي سنتكلم عنها لاحقاً والمرحلة الاخيرة يتم تحديث اوزان الشبكة العصبونية العميقة حسب معادلة بيلمان.

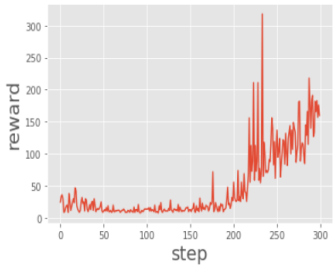
7- التطبيق العملي والنتائج

في المرحلة الاولى طبقنا خوارزمية Q learning فقط دون اضافة الشبكة العصبونية وحصلنا على تغير مكافأة العميل بعد 300 دورة تدريب كما في الشكل (4)



الشكل (4) تغير مكافاة العميل لمده 300 دوره تدريب باستخدام خوارزمية Q learning فقط

في المرحلة الثانية طبقنا النظام المقترح في الشكل (3) على بيئة Cart Pole-v1 الموجودة في مكتبة Gym في بيئة python حيث استخدمنا شبكة عصبونية مؤلفة من طبقتين خفيتين وعدد نيورونات يمثل قوى العدد 2 المتناقصة وحصلنا على تغير مكافاة العميل بعد 300 دوره تدريب كما في الشكل(5)

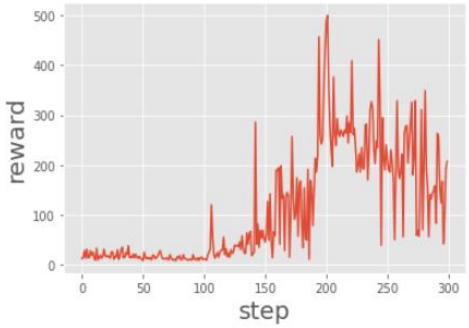


الشكل (5) تغير مكافاة العميل لمده 300 دوره تدريب بشبكة عصبونية بطبقتين خفيتين.

في الشكل (5) نلاحظ تغير مكافاة العميل لمده 300 دورة تدريب بشبكة عصبونية مؤلفه من طبقتين خفيتين وبعدد نيورونات محدد

بشكل متناقص وبأعداد من قوه العدد 2 (4 32 8) حيث 4 عدد عقد الدخل و 32 و8 عدد نيورونات الطبقتين الخفيتين والاختيار بهذا الشكل لتحسين اداء الشبكة باعتبار ان خطوط مساري المعالج وعناوين الذاكرة هي بأرقام من قوه العدد 2 حسب ملاحظات البروفيسور Andrew [9] ومن خلال الشكل (5) نلاحظ تحسن مكافاة العميل عن المرحلة السابقة نسبيا .

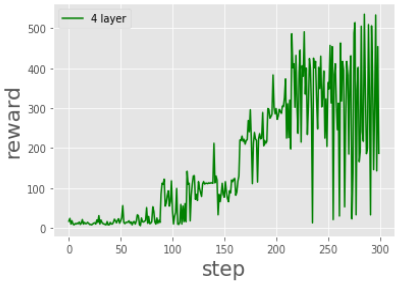
في المرحلة الثالثة استخدمنا شبكة عصبونية عميقة مؤلفة من ثلاث طبقات خفية وعدد نيورونات محدد بشكل متناقص وبأعداد تمثل قوه العدد 2 وحصلنا على تغير مكافاة العميل بعد 300دوره تدريب كما في الشكل(6)



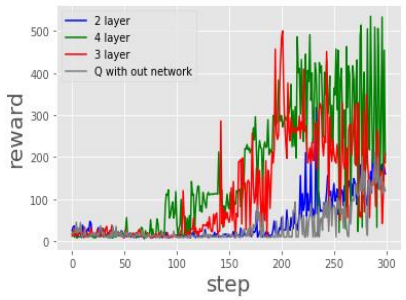
الشكل (6) تغير مكافاة العميل لمده 300 دوره تدريب بشبكة عصبونية عميقة بثلاث طبقات خفية.

في الشكل (6) استخدمنا ثلاث طبقات خفية ونيورونات مرتبة (4 64 32 16) ونلاحظ تحسن المكافاة بشكل واضح وبنسبة 50% تقريبا عن المرحلة السابقة. وهنا يجب أن نشير الى أن الانكسارات الموجودة في الخطوط البيانية للنتائج في كل مرحلة سببها تقنية epsilon greedy للخوارزمية الجشعة وهذه الميزة مستخدمة في خوارزمية Q learning بحيث ان العميل في جزء صغير من زمن دوره التدريب وهذا الجزء يمكن تحديده حسب الحاجة يقوم باختيار اجراءات عشوائية لاكتشاف إجراءات جديده وبعد ذلك يعود لاختيار الاجراء بأعلى مكافاة.

في المرحلة الرابعة استخدمنا شبكة عصبونية عميقة مؤلفة من اربع طبقات خفية وعدد نيورونات محدد بالشكل (4-64-32-16-8) وحصلنا على تغير مكافاة العميل بعد 300دوره تدريب كما في الشكل (7)



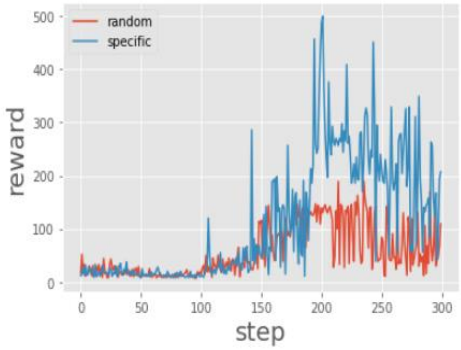
الشكل (7) تغير مكافاة العميل لمده 300 دوره تدريب بشبكة عصبونية عميقة بأربع طبقات خفية.



الشكل (8) مقارنه نتائج المراحل الاربع بعد 300 دوره تدريب.

في الشكل (8) نلاحظ مقارنة قيمة المكافأة بعد كل مرحلة من المراحل الثلاث بالإضافة الى المقارنة مع تطبيق خوارزمية Q learning من دون شبكة عصبونية ونلاحظ تحسن المكافأة بعد كل مرحلة بشكل واضح وذلك لسببين الاول لترتيب النيورونات بشكل يمثل قوى العدد 2 المتناقصة والثاني لزيادة عدد الطبقات الخفية وبالتالي استطعنا اثبات اهمية التعلم العميق في تحسين المكافاة علما انا زيادة عدد الطبقات الخفية في الشبكة لا يمكن لوحده تحسين المكافاة بالشكل المطلوب كما نرى بالشكل (9) مقارنة استخدام ثلاث طبقات خفية بالشبكة العصبونية العميقة بترتيب عشوائي لعدد

العصبونات في كل طبقة (4 14 25 77) حيث تمثل 4 عدد عقد المداخل و14 25 77 تمثل عدد نيورونات الطبقات الخفية مع المرحلة الثالثة من تطبيقنا.

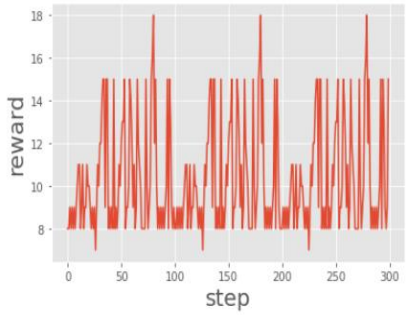


الشكل (9) مقارنة المرحلة الثالثة من تطبيقنا مع استخدام ثلاث طبقات خفية في الشبكة العصبونية العميقة بترتيب عشوائي لعدد النيورونات.

من الشكل (9) نرى اهمية ترتيب العصبونات بشكل متناقص وممثل لقوة العدد 2 لتحسين المكافاة وذلك بسبب ان عدد خطوط مساري المعالجات وعناوين الذاكرة هي من قوه العدد 2 مما يعطي افضل استخدام لموارد الحاسب.

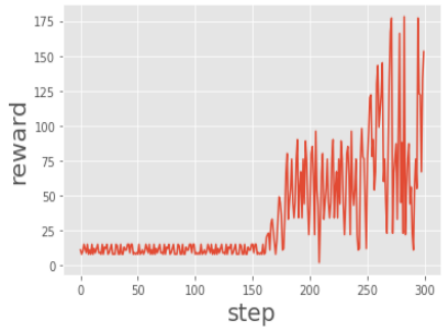
8- تأثير epsilon greedy في خوارزمية Q learning

تأتي أهمية العامل epsilon [10] في خوارزمية Q Learning من حيث التوازن بين الاستكشاف(exploration) والاستغلال (exploitation) في عمل الخوارزمية فقيام العميل باتباع مسار المكافأة الأعلى دائماً ربما يكون غير مجدي لعدم اكتشاف مسارات جديدة تؤدي الى نتائج أفضل ومكافأة اعلى ومن هنا يمكن ضبط قيمة العامل epsilon بحيث يتم تخصيص 10% من زمن تدريب الخوارزمية في كل خطوه للاستكشاف والباقي للاستغلال ونحصل عليها بضبط قيمة epsilon على قيمة بين (0-1) بحيث تكون اعظم قيمة له هي 1 تقابل 10% من زمن التدريب فعند عمل الخوارزمية بدون استكشاف اي epsilon=0 نلاحظ ثبات في قيمة المكافاة بسبب عدم اكتشاف خيارات جديده كما في الشكل (10).



الشكل (10) تغير مكافاة العميل لمده 300 دوره تدريب بشبكة عصبونية عميقة بثلاث طبقات خفية بقيمة epsilon =0.

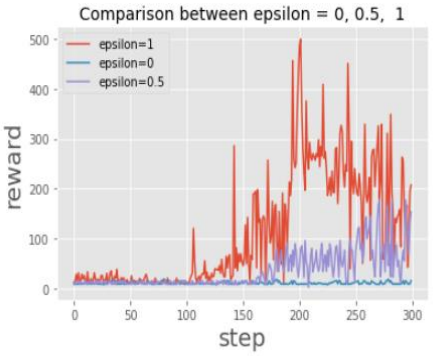
نلاحظ من خلال الشكل (10) ثبات قيمة المكافاة على قيم صغيره جدا بسبب عدم اكتشاف اجراءات جديده. من خلال استخدام epsilon=0.5 اي 5% من زمن التدريب للاستكشاف نلاحظ الاستجابة حسب الشكل(11)



الشكل (11) تغير مكافاة العميل لمده 300 دوره تدريب بشبكة عصبونية عميقة بثلاث طبقات خفية بقيمة epsilon =0.5.

من خلال الشكل (11) نلاحظ اهمية الاستكشاف في خوارزمية Q Learning من خلال زيادة المكافاة بشكل واضح. في الشكل (6) السابق نلاحظ تغير مكافاة العميل لمدة 300 دوره تدريب بشبكة عصبونية عميقة بثلاث طبقات خفية بقيمة epsilon =1 أي قيمة

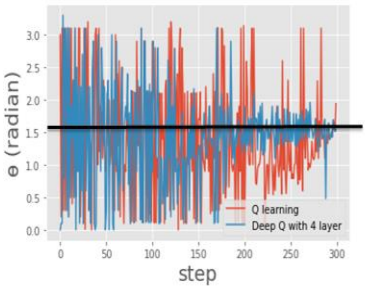
عظمى للاستكشاف اي 10% من زمن التدريب وفي الشكل (12) نلاحظ المقارنة للقيم المختلفة للعامل epsilon



الشكل (12) مقارنة قيم مختلفة للعامل epsilon من خلال تغير مكافأة العميل لمده 300 دوره تدريب بشبكة عصبونية عميقة بثلاث طبقات خفية.

من خلال الشكل (12) نلاحظ زيادة المكافاة للعميل بزيادة زمن الاستكشاف ومن هنا تتوضح اهمية تقنية epsilon greedy في خوارزمية Deep Q Learning.

في المرحلة الأخيرة نبين محاكاة لبيئة cart pole-v1 من خلال ملاحظه تغير زاوية العصا Θ مع كل دورة تدريب حسب الشكل (13) حول وضع التوازن الذي يقابل 90 درجة او 1.57 راديان.



الشكل (13) تغير زاوية العصا مع كل دورة تدريب

من الشكل (13) نلاحظ تحسن اداء النظام بعد استخدام الشبكة العصبونية العميقة بأربع طبقات خفية وثبات العصا أكثر حول وضع التوازن خاصة في المراحل الاخيرة من التدريب.

9- الاستنتاجات والتوصيات

من خلال تحليل النتائج السابقة نلاحظ تحسن أداء النظام بزيادة عدد الطبقات الخفية للشبكة العميقة واستخدام عدد نيورونات تمثل قوى العدد 2 المتناقصة الذي يسمح باستخدام أفضل لموارد الحاسب كون مساري المعالجات وعناوين الذواكر تمثل قوى العدد 2.

بقي لنا أن نذكر أنه من الممكن استخدام تقنيات أخرى لتحسين مكافأة العميل لتحقيق الهدف المطلوب منه بأفضل شكل وأفضل زمن ومن هذه التقنيات

1- استخدام شبكات عصبونية من نوع Modular MNN بحيث يمكن جمع أكثر من شبكة عصبونية ومن انواع مختلفة لتعمل معاً.

2- تطبيق النظام المقترح على عميل حقيقي وبالتالي استخدام كاميرا وبالتالي نستطيع إضاقة شبكات عصبونية ملتفة لمعالجة الصور وتحسين الاداء.

References المراجع

1. Mnih ,V., Kavukcuoglu,K., Silver,D., Graves,A., Antonoglou,L., Wierstra,D., Riedmiller,M.(2013). Playing Atari with Deep Reinforcement Learning: DeepMind Technologies. deepmind.com.
2. NAEEM ,M., RIZVI,S., CORONATO, A. (2020) .Gentle Introduction to Reinforcement Learning and Its Application in Different Fields: Digital Object Identifier 10.1109/ACCESS.
3. Aradi,S. , Becsi,T., Gaspar,P.(2018). Policy Gradient based Reinforcement Learning Approach for Autonomous Highway Driving: IEEE Conference on Control Technology and Applications (CCTA) Copenhagen, Denmark, August
4. Doltsinis, S., Ferreira, P., Lohse,N.(2014) .An MDP Model-Based Reinforcement Learning Approach for Production Station Ramp-Up Optimization: Q-Learning Analysis: IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. SEPTEMBER
5. Ronecker,M., Zhu ,Y.(2019).Deep Q-Network Based Decision Making for Autonomous Driving :2019 3rd IEEE International Conference on Robotics and Automation Sciences.
6. Saini,A. , Gupta,T., Kumar,R., Gupta ,A., Panwar, M., Mittal, A.(2017). Image based Indian Monument Recognition using Convoluted Neural Networks: 2017 International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune,
7. Zohrer,M., Pernkopf,F.(2018). Heart Sound Segmentation - An Event Detection Approach using Deep Recurrent Neural Networks: Citation information: DOI 10.1109/TBME.2018.2843258, IEEE Transactions on Biomedical Engineering.
8. Donoghue,B., Osband,I., Munos,R., Mnih,V.(2018). The Uncertainty Bellman Equation and Exploration: arXiv:1709.05380v4 [cs.AI].
9. Andrew, Ng.(2017). Machine Learning: Stanford university , http://cnx.org/content/col11500/1.4/.
10. Rao,R., Narasimhan,K.(2020) .Stage Epsilon-Greedy Exploration for Reinforcement Learning: Princeton University, Department of Computer Science.