مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (٩) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (9) 2025

تصميم وتقييم منصّة تعليميّة ذكيّة قائمة على تقنيات التعلّم المعزّز

د. جعفر سـلمان *

م. أديل أسعد علي**

(تاريخ الإيداع ٢٠٢٥/١/٢٩ . قُبِل للنشر في ٢٠٢٥/٧/٣١)

□ ملخّص □

إنّ التعليم التكيفي للتعلم عبر الإنترنت يمكن أن يزيد من مكافأة التعلم ويقلل الجهد المطلوب من الطلاب والمدرسين ومصممي الدورات، يعد التعلم المعزز أداة واعدة لتطوير استراتيجيات التعليم، حيث يمكن لنماذج التعلم المعزز تعلم العلاقات المعقدة بين محاور الدورة وتفاعلات المتعلم والنتائج المحققة. يوضح هذا البحث أول نموذج تجريبي للتعلم المعزز باستخدام تقنية جدولة الأنشطة التعليمية في الزمن الحقيقي لدورة تدريبية كبيرة عبر الإنترنت حول للدارات الكهربائية وقوانين الإلكترون، لتمكين الطلاب من فهم وتحليل الدارة، وإغلاق القواطع المناسبة ووضع المكونات الكهربائية كالمقاومات والمكثفات وغيرها في المكان المناسب، لإجراء حسابات التيار الكهربائي والجهد والاستطاعة من خلال التعلم الفعال، حيث يتعلم نموذجنا تحديد سلسلة من أنشطة الدورة مع تعظيم مكافأة التعلم وتقليل عدد الإجراءات المطلوبة. باستخدام خوارزمية Q–Learning (Quality Learning) وتطبيقها على أكثر من 1800 متعلم، نبحث في كيفية تأثير تقنية الجدولة هذه على مكافأة التعلم ومعدلات التسرب والاستجابات النوعية للمتعلم، فمهمة التعلم المعزز هي تدريب الوكيل (المدرس الافتراضي) الذي يتفاعل مع بيئته (واجهة الشاشة الخاصة بالدارة الإلكترونية). يظهر الوكيل في سيناريوهات مختلفة تُعرف بالحالات عن طريق القيام بإجراءات (الوضع الصحيح للعنصر، والوضع الخاطئ للعنصر، والنقر الصحيح، والنقر الخاطئ، والنقر الخارجي) ، نبين أن نموذجنا ينتج مكافأة تعلم أفضل باستخدام أنشطة تعليمية أقل من حالة التعيين الخطيLinear، كما ينتج مكافأة تعلم مماثلة لحالة التوجيه الذاتي Self–Directed–Navigation باستخدام أنش طة تعليمية أقل ومعدلات تسرب أقل ، بينت النتائج أن النهج المقترح حقق دقة عالية حيث بلغ معدل النقر الصحيح للقاطع ووضع العناصر الإلكترونية في المكان الصحيح من قبل الوكيل 92.5% مما يدل على أن النظام قد تمّ تعلمه بشكل صحيح ودقيق بينما معدل الخطأ من قبل الوكيل هو 7.5% فقط للنقر الخاطئ على القاطع ووضع العناصر الإلكترونية في المكان الخطأ والنقر في مكان آخر.

الكلمات المفتاحية: جدولة الإجراءات، الذكاء الصنعي، التعلم عن بعد، التعلم المعزز، التعلم المعزز العميق، مكافأة التعلم، الوكيل، Q–Learning، DQN.

* أستاذ مساعد في قسم تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس – سوريا

** طالبة ماجستير في قسم تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس – سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (٩) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (9) 2025

Design and Evaluation of an Intelligent Educational Platform Based on Reinforcement Learning Techniques

Dr.Jaafar Salman*
Eng.Adeel Asaad Ali**

(Received 29/1/2025 . Accepted 31/7/2025)

□ ABSTRACT □

Adaptive online learning can enhance learning rewards and reduce the effort required from students, teachers, and course designers. Reinforcement learning is a promising tool for developing educational strategies, as reinforcement learning models can learn the complex relationships between course components, learner interactions, and achieved outcomes.

This research presents the first reinforcement learning model using real-time scheduling of educational activities for a large online course on electrical circuits and electronic laws, enabling students to understand and analyze circuits, close appropriate interrupters, and place electrical components such as resistors and capacitors in the correct locations to perform calculations of current, voltage, and power through effective learning. Our model learns to identify a series of course activities that maximize learning rewards while minimizing the number of required actions. Using Q-Learning algorithm applied to over 1800 learners, we investigate how this scheduling technique affects learning rewards, dropout rates, and qualitative learner responses. The task of reinforcement learning is to train an agent (the virtual teacher) that interacts with its environment (the electronic circuit interface). The agent appears in different scenarios defined by states by performing actions (correct placement of components, incorrect placement of components, correct clicks, incorrect clicks, external clicks). We demonstrate that our model produces better learning rewards using fewer educational activities compared to the linear assignment case. It also yields similar learning gains to self-directed navigation while utilizing fewer educational activities and achieving lower dropout rates. The results indicate that the proposed approach achieved high accuracy with a correct click rate for interrupters and proper placement of electronic components by the agent at 92.5%, demonstrating that the system has been learned correctly and accurately. Meanwhile, the error rate for incorrect clicks on interrupters and improper placement of electronic components was only 7.5%, along with clicks in other areas.

Keywords: Action Scheduling, Agent, Artificial Intelligence, Deep Reinforcement Learning, DQN, Learning Rewards, Q-Learning, Reinforcement Learning, Remote Learning.

*Assistant Professor, Information Technology Engineering Department, Faculty of Information and Communication Technology Engineering, Tartous University, Syria.
**Master Student, Information Technology Engineering Department, Faculty of Information and Communication Technology Engineering, Tartous University, Syria.

1- المقدمة

الآليات والأساليب التقليدية ليست قادرة دائماً على التغلب على الحواجز التي تفرضها قيود الزمن الأكثر صعوبة، ومن هنا بدأ التركيز على البحث والتطوير في خوارزميات الذكاء الصنعي لتطبيقات الزمن الحقيقي بشكل ملحوظ في السنوات الأخيرة.

لقد شهد العالم تطورات كبيرة في مجال التعليم الإلكتروني والتعلم عن بعد، خاصة خلال أزمة فيروس كورونا، والتي كشفت عن أهمية هذين النوعين من التعليم والفوائد المثمرة التي يقدمانها في مجموعة من الدول، خاصة تلك التي تتمتع ببنية تحتية ممتازة.

لذلك سنعمل على إنشاء منصة إلكترونية بسيطة للتعامل مع أنواع الدارات المختلفة من حيث إضافة أو حذف بعض العناصر والتحكم بحالة القواطع، ولإجراء العمل عن بعد RPW (Remote Practical Work) وتحقيق نتائج جيدة من حيث تفاعل الطلاب وحتى تسهيل توظيف المدرس، وذلك من خلال اقتراح نظام توصية يعتمد على شبكات التعلم عميق الجودة DQNs (Deep Quality–learning Networks) لتوصية وتوجيه الطلاب مسبقاً للقيام بـــ RPW وفقاً لمهاراتهم ومتابعتهم عند القيام بكل نقرة بالماوس أو لوحة المفاتيح لكل طالب.

يعد التعلم المعزز من أهم فروع الذكاء الصنعي الذي يستخدم في مجموعة من المجالات والتخصصات بما في ذلك مجال التعليم الذي سنركز عليه.

ينقسم الذكاء الصنعي إلى ثلاثة فروع:

١. التعلم الخاضع للإشراف.
٢. التعلم غير الخاضع للإشراف.
٣. التعلم المعزز.

فالتعلم المعزز هو نظام ذكاء صنعي يسمح للوكيل (عبارة عن برمجية تؤدي دور المشرف على عملية تدريس الطلاب وتوجيههم) بالتعلم من التفاعل مع البيئة التي يتم إدخاله فيها.

للتعلم المعزز ثلاثة أنواع وهي:

١) التعلم المعزز (الخالي من النماذج) Reinforcement Learning .
٢) التعلم المعزز العميق Deep Reinforcement Learning.
٣) التعلم المعزز العكسي Inverse Reinforcement Learning.

لقد شهد التعلم المعزز استخداماً ضئيلاً بشكل ملحوظ في التعليم عبر الإنترنت حيث أنّ أحد التحديات الكبيرة التي تواجهنا عندما يتعلق الأمر بجدولة الأنشطة التعليمية هو غياب المنصات التي تدعم هذه التقنية، لقد ظهرت عدة دراسات تناول الباحثون فيها العديد من الأساليب لتطوير نماذج التعلم المعزز:

إذ قام Sergio Re et al. في عام 2019 [1] باستخدام خوارزمية Q–Learning و هي خوارزمية مناسبة للتطبيقات في الزمن الحقيقي ، حيث تم اقتراح تقنية تعتمد على المضاعفات التقريبية لتقليل تعقيد أجهزة الخوارزمية وتم تنفيذ التصميم على مجموعتي التقييم Xilinx Zynq Ultrascale و MPSOC ZCU106

وجرى تقييم نتائج التنفيذ من حيت موارد الأجهزة والإنتاجية واستهلاك الطاقة وتبين أن هذا النظام مناسب لتطبيقات IOT متعددة الوكلاء.

وفي عام 2020 طوّر .Lukasz Kaiser et al [2] تقنية تعلم معزز عميق لتقديم نظام قادر على التعامل بنجاح مع مجموعة متنوعة من الألعاب الصعبة في معيارALE ، ولتحقيق ذلك تم تجريب العديد من تقنيات التنبؤ بالفيديو العشوائية، حيث تم استخدام نموذج تعلم سياسة المحاكاة Simulated (SimPLe) Policy Learning الذي يعتمد استخدام تقنيات التنبؤ بالفيديو وتدريب سياسة للعب اللعبة ضمن النموذج الذي تم تعلمه باستخدام خوارزمية DQN من خلال العديد من التكرارات لتجميع مجموعات البيانات في العديد من الألعاب، للتمكن من تشغيل اللعبة بنجاح في الزمن الحقيقي.

كما قدّم .Alex Graves et al في عام 2013 [3] أول نموذج للتعلم العميق المعزز لتعلم سياسات التحكم بنجاح مباشرة من المدخلات الحسية عالية الأبعاد باستخدام التعلم المعزز، وكان النموذج عبارة عن شبكة عصبونية تلافيفية ، تمّ تدريبها باستخدام متغير من Q–Learning والتي تكون مدخلاتها عبارة عن بكسلات أولية ومخرجاتها عبارة عن دالة قيمة لتقدير المكافآت المستقبلية، جرى تطبيقها على سبع ألعاب Atari 2600 من بيئة التعلم Arcade بدون تعديل البنية أو خوارزمية التعلم، ولكن كان التمرير الأمامي المنفصل مطلوب لحساب قيمة Q لكل إجراء.

واقترح .Yufeng Yuan et al في عام 2022 [4] نظام تعليمي غير متزامن وإجراء مقارنة منهجية بين التعلم المعزز المتسلسل وغير المتزامن باستخدام بيئات العالم الحقيقي، يتعلم هذا النظام في الزمن الحقيقي للوصول إلى الأهداف المرئية من البكسل وتتبعها في غضون ساعتين من الخبرة ويقوم بذلك مباشرة باستخدام روبوتات حقيقية، ويتعلم تماماً من نقطة الصفر، ولكن زمن دورة الإجراء كان إما طويل جداً أو قصير جداً.

وفي عام 2020 قام .Aleksander Izemski et al [5] بتوجيه عملية التعلم في منصة التعلم الإلكتروني إلى إطار عمل التعلم المعزز فأصبحت تحاكي عملية التعلم الفردي للطالب التي تنتهي بامتحان بأسئلة لم تكن معروفة من قبل حيث يتفاعل الوكيل الذي يمكن فهمه على أنه مدرس، مع البيئة من خلال تحديد المهام التي سيحاول الطالب حلها، تتميز المهمة بمهارة معينة مطلوبة لحلها (يمكن للطالب حل المهمة بشكل صحيح أو غير صحيح والذي يتضمن أيضا عدم وجود إجابة) وهو ما ينعكس في حالة البيئة، مما ينتج عنه ملاحظات للوكيل و في كلتا الحالتين، (إجابة غير صحيحة / صحيحة)، تزداد كفاءة الطالب في مهارة ما بدرجة معينة، ويمكن للطالب حل المهمة عدة مرات في عملية التعلم، ولكن مع زيادة عدد المهارات يصبح تعيين المهام بالنسبة للوكيل أكثر صعوبة.

واقترح .Markel Sanz Ausin et al في عام 2019 [6] تطبيق DRL من خلال نهج غير متصل بالإنترنت في نظام التدريس الذكي حيث تم تطبيق خوارزميات DQN وDouble DQN للحصول على استراتيجيات تعليمية تكيفية مصممة خصيصاً للطلاب الفرديين، في حالةonline تتطلب DQN مئات الملايين من التفاعلات مع البيئة لتعلم سياسة جيدة والتعميم على الحالات غير المرئية بينما يسعى هذا النظام لتعلم السياسات من مجموعات البيانات التي تحتوي على أقل من 800 سجل تفاعل بين الطالب والمعلم، بالتالي عانى النظام من بيانات التدريب المحدودة بالإضافة إلى مشكلة التخصيص.

كما قام Lu Chen et al. في عام 2017 [7] باقتراح إطار جديد للتدريس يدعى التعليم المشترك عن طريق تضمين مدرس بشري في حلقة التدريب على سياسة الحوار عبر الإنترنت لمعالجة مشكلة البداية الباردة، حيث يتم تدريب سياسة الحوار ليس فقط باستخدام مكافأة المستخدم ولكن أيضاً على مثال المعلم بالإضافة إلى الاستجابة الفورية المقدرة على مستوى الدور.

2- أهمية البحث وأهدافه

تتطلب معظم تطبيقات أنظمة الزمن الحقيقي نماذج حوسبة قوية قادرة على معالجة كمية كبيرة جداً من البيانات بأسرع ما يمكن وباستهلاك محدود للطاقة.

يتيح تطوير تقنية التعلم المعزز لـــ Q-Learning تصميم أنظمة برمجية أسرع من نظيراتها، وبالتالي إتاحة استخدامها في المشكلات التي تتطلب تلبية قيود الزمن الضيقة ومعالجة البيانات الضخمة.

كما تمكننا المنصّة المعنيّة بدراسة الدّارات الكهربائية من جمع معلومات عن الطلاب والمدرسين وتفاعلاتهم مع المحتوى التعليمي الذي سنعتمد عليه كدخل (عدد كبير من الصور في الثانية لكل نقرة بالماوس أو لوحة المفاتيح لكل طالب)، أما بالنسبة لخرج نظامنا تتجلى هذه التقنية في محاولة تجسيد نموذج المدرس الافتراضي داخل المنصة ومن ثم تدريبه بشكل مناسب باستخدام تقنية DQN لتطبيق RPW.

3- طرق البحث و مواده

تم تنفيذ هذا البحث باستخدام لغة البرمجة Python 3.8.3 ضمن بيئة العملJupyter Notebook، كما تم استخدام برنامج LabVIEW 2017، على جهاز حاسوب شخصيWindows 10, 64 bit و يمتلك وحدة معالجة مركزية Intel® Xeon® CPU E5-2603 v4 @ 1.70 GHz و وحدة معالجة رسومية NVIDIA NVS 315 ، و ذاكرة نظام 24 GB.

3-1 التعلم المعزز

هو نظام ذكاء صنعي يسمح للوكيل بالتعلم من التفاعل مع البيئة التي يتم إدخاله فيها، يستخدم هذا النهج في الحالات التي لا توجد فيها معلومات كافية حول السلوك الذي يجب أن يتخذه الوكيل للوصول إلى هدفه، أي أن الوكيل دون معرفة سابقة يتعلم من خلال التفاعل مع البيئة، يركز التعلم المعزز على دراسة الإجراءات التي تحقق أقصى قيمة للمكافأة التراكمية من البيئة و يستخدم RL عملية تعلم التجربة و الخطأ لتحقيق هدفه ، حيث تعتبر هذه الميزة الفريدة نهجاً متقدماً لبناء وكيل يحاكي البشر، و أبرز أنواع التعلم المعزز هو **التعلم المعزز العميق** نركز عليه بسبب كفاءته العالية في حل المشكلات الصعبة و من ضمنها عملية اتخاذ القرار وفي عملية التعلم تساعد المتعلم على التفاعل والاندماج في الدرس وحتى الامتحان حيث يقوم الطالب بمجموعة من الإجراءات لحل مشكلة ما[7].

3-2 Q-Learning أبرز خوارزميات التعلم المعزز وهي خوارزمية مناسبة للتطبيقات في الزمن الحقيقي، خصائصها الأساسية: الطاقة المنخفضة ، الإنتاجية العالية وموارد الأجهزة المحدودة [1]، تستخدم الجودة Q للدلالة على مدى أهمية إجراء معين.

للتعبير عن ذلك رياضياً نستخدم **معادلة بيلمان** [9] التي تعطى بالصيغة الآتية:

$$Q(s, a) = r(s, a) + \gamma \, Q\,max(s', a) \qquad (1)$$

أي أنّ قيمة Q عندما يكون الوكيل في الحالة s و ينفذ الإجراء a تساوي المكافأة الفورية r(s, a) مضافاً إليها قيمة Q الأعظمية عندما يكون الوكيل في الحالة التالية.

s هي الحالة التي يوجد فيها الوكيل.

a هو الإجراء الذي يتخذه الوكيل في حالة معينة.

r هي المكافأة التي يحصل عليها الوكيل من البيئة لقيامه بإجراء معين في حالة ما.

γ هو عامل الخصم الذي يتحكم في مدى تأثير المكافآت على الحالات التالية.

Q(s', a) تعتمد على Q(s", a) التي لها γ^2 أي تعتمد قيمة Q على قيم Q للحالات التالية كما هو موضح في المعادلة الآتية:

$$Q(s, a) = \gamma Q(s', a) + \gamma^2 Q(s'', a) + \dots + \gamma^n Q(s''^{\cdot n}, a) \quad (2)$$

بما أن هذه المعادلة تكرارية يمكننا إعطاء Q قيم كيفية عشوائية وبذلك يتكون السلوك الأمثل.

يمكن التعبير عن كل تحديث [11] بالمعادلة الآتية:

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \alpha_t(S_t, A_t)\,[R_{t+1} + \gamma\, Q_{max}(S_{t+1}, A_t) - Q_t(S_t, A_t)] \quad (3)$$

يمثل α معدل التعلم.

توضح هذه المعادلة مدى تجاهل المعلومات المكتسبة حديثاً للمعلومات السابقة.

نلاحظ أن هذه مشكلة انحدار ومع ذلك، لا نعرف الهدف أو القيمة الفعلية هنا لأننا نتعامل مع RL (هدف غير ثابت).

فيما يلي معادلة تحديث قيمة Q التي تم الحصول عليها من معادلة "**بيلمان**":

$$Q^* = R_{t+1} + \gamma\, Q_{max}(S_{t+1} + A_t) \quad (4)$$

Q^* تمثل الهدف حيث يتم التنبؤ بقيمته.

R هي المكافأة الفعلية الأمثل.

ستقوم الشبكة بتحديث ميلها باستخدام خوارزمية (Back Propagation) الانتشار الخلفي للتقارب.

3-3 شبكات التعلم عميق الجودة DQN (Deep Q-Learning Network)

تعتبر صلة الوصل بين الشبكات العصبونية العميقة و التعلم المعزز [10].

إذ تستفيد DQN من الشبكة العصبونية التلافيفية CNN لتحليل صور الدخل واستخدام هذه الشبكات العصبونية لتقريب دالة Q-value.

ومن جانب آخر، فإنّ هدف DQNs هو تقليل دالة الخطأ لشبكة CNN حيث تقوم DQNs بإنشاء وكيل ذكي يتفوق في الأداء على أفضل طرق RL.

3-4 توظيف التعلم المعزز في النظام المقترح

يكون الوكيل في الحالة الحالية S_t، حيث يتخذ الإجراء A_t، ويتفاعل مع البيئة ويستجيب لها، ويعيد إجراء للحالة التالية S_{t+1}، نظراً لحالته ومكافأته الحالية، يختار الوكيل الإجراء التالي ويكرر هذه العملية حتى يتم حل بيئته وإنهائها (عندما يحقق هدفه)، و كما ذكرنا فإنّ مهمة RL هي تدريب الوكيل (المدرس

الافتراضي) الذي يتفاعل مع بيئته (واجهة الشاشة الخاصة بالدارة الإلكترونية)، ويظهر الوكيل في سيناريوهات مختلفة تُعرف بالحالات عن طريق القيام بإجراءات (الوضع الصحيح للعنصر، والوضع الخاطئ للعنصر، والنقر الصحيح، والنقر الخاطئ، والنقر الخارجي)، ستؤدي هذه الإجراءات إلى مكافآت (يتم تعيين المكافآت بناءً على نتيجة هذه الإجراءات) إذا كان النقر صحيح أو الوضع صحيح، فإنه يستدعي مكافأة موجبة، أما النقر الخاطئ أو الوضع الخاطئ يعيد مكافأة سالبة، لكن النقر الخارجي لا يعيد أي شيء.

4- بنية النظام المقترح

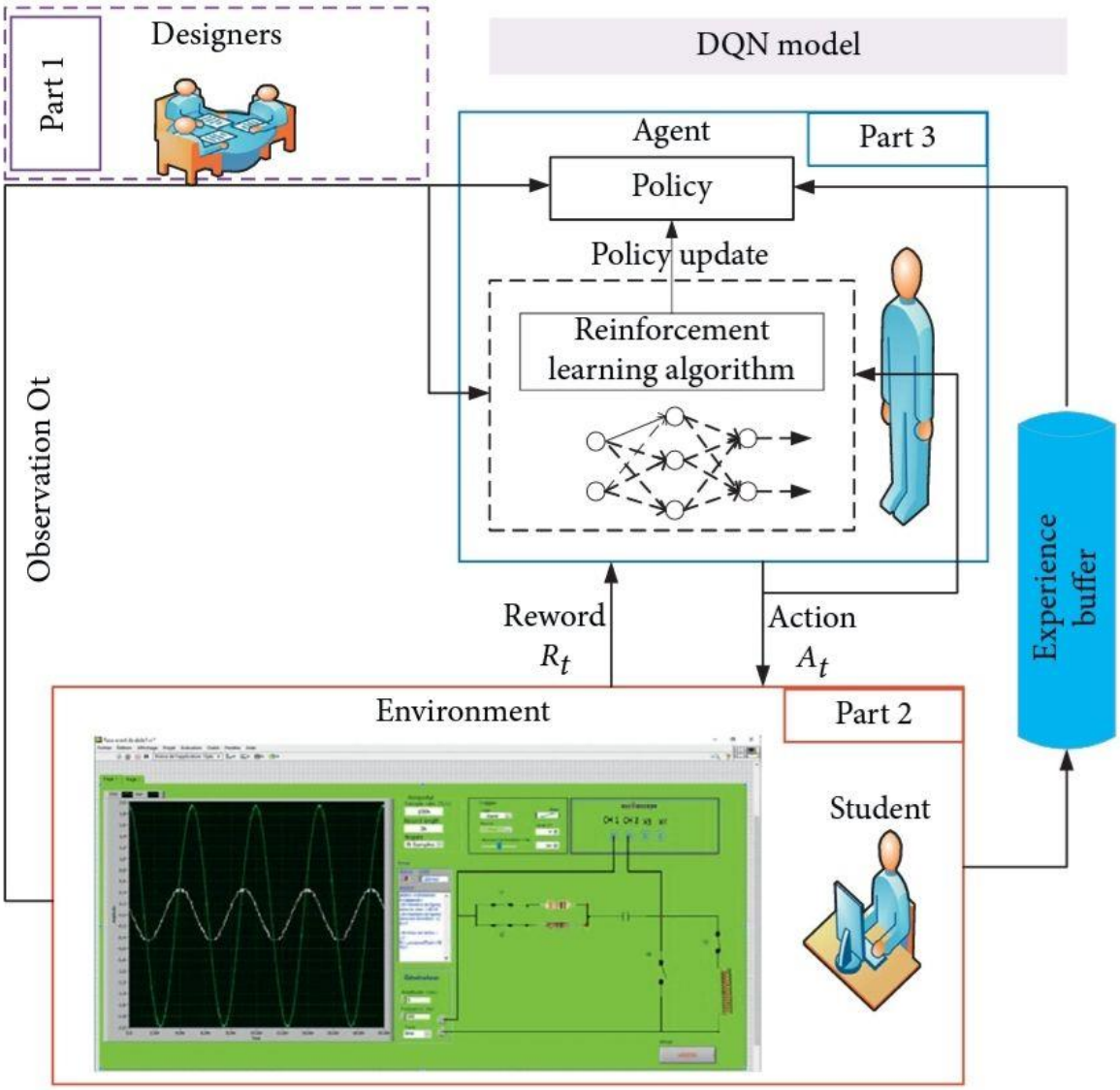
يمكن تقسيم بنية النظام المقترح إلى ثلاث مراحل:

المرحلة الأولى: هي الجزء الخاص بصنع بنية المحتوى التعليمي، وهي الجوهر الأساسي لنجاح العملية لأنه، كما هو معروف عادة في وسائل التواصل الاجتماعي والقنوات التلفزيونية، إذا كانت بنية المحتوى التعليمي المقدم للناس جذّابة، سيتفاعل الناس معه بطريقة رائعة ويوصون به أيضاً، لذلك يجب التركيز على هذا الجانب.

المرحلة الثانية: هي الجزء المتعلق بالطالب لأن هدفنا الأساسي هو مساعدة الطلاب وتوفير كافة المتطلبات التي يحتاجونها لتحقيق أهدافهم، لذلك كمصممين نقوم بالاستجابة وإظهار النتائج بأبسط الطرق، حيث يتفاعل الطالب مع بيئة بسيطة تظهر على شاشة الحاسوب الخاص به. تسمح الدارة التي تتكون من مجموعة من العناصر الكهربائية للطالب بالنقر على قاطع الدارة، على سبيل المثال، لإجراء حسابات التيار الكهربائي أو الاستطاعة.

المرحلة الثالثة: عبارة عن نموذج لمدرّس افتراضي أو مخبر أو روبوت يقوم بتوجيه الطلاب من خلال تقديم مجموعة من الاقتراحات التي تكون على شكل نصوص مكتوبة وصوت وصور أو مجموعة من الحركات لإلهام الطلاب بما سيفعلونه.

هذه المراحل موضّحة في الشكل (1):



شكل(1) بنية النظام المقترح

1-4 مكونات النظام المقترح

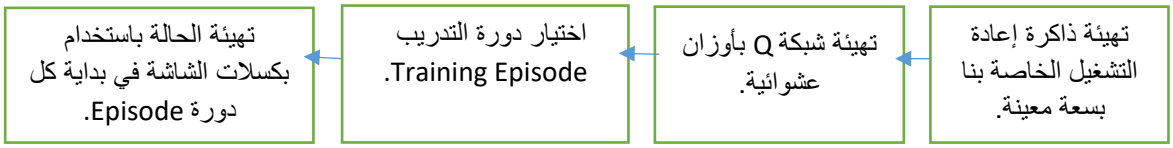
صـور الدخل: تظهر للطلاب، وتكون تعبيراتهم حول تمارين أو اختبارات أو صـور أو نصـوص مبنية على أساس محتوى النظام.

شبكات التعلم العميق: وهي الجزء المتعلق بمعالجة هذه الصور واستخراج البيانات اللازمة منها.

الإجراءات المحتملـة: وهي أهم مرحلـة بـالنســبـة للمتعلمين، يتخـذون من خلالهـا الإجراءات والقرارات التي يرونها مناســبة للحصــول على مكافأة جيدة من البيئة، والتي تكون على شــكل نقاط أو نسبة.

4-2 خوارزمية DQN مع إعادة تشغيل التجربة

هنالك عدة بارامترات تعتمد عليها هذه الخوارزمية ويجب ضبطها بشكل دقيق قبل البدء بتدريب وكيل التعلم المعزز يبينها الشكل(2):



شكل(2) مخطط صندوقي لخوارزمية DQN

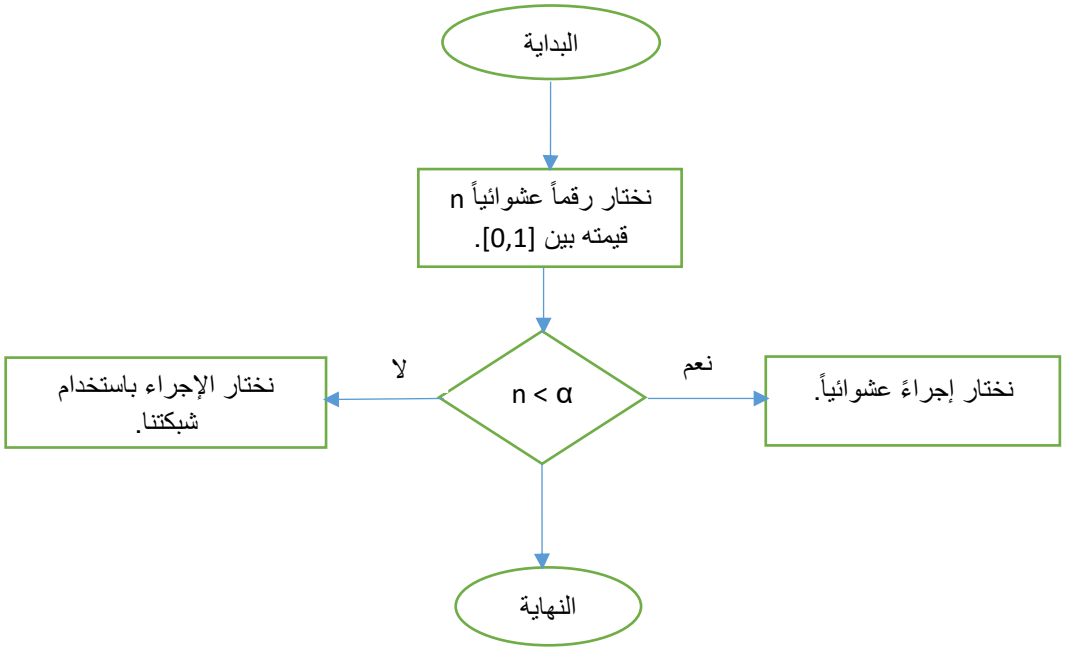
يجب القيام بخطوة المعالجة المسبقة للحصول على حالة الدخل الفعلية لكل خطوة زمنية للإجراء الذي نقوم به حالياً باحتمالية صغيرة (قمنا بمعالجة صور RGB مقاس 210 × 160 مسبقًا عن طريق تقليل حجمها إلى 84 × 84 واستخراج قناة السطوع).

4-3 الشبكة الهدف

على الرغم من أن نفس الشبكة تقوم بحساب القيمة المتوقعة وقيمة الهدف، فقد يكون هناك الكثير من التفاوت بينهما، فبدلاً من استخدام شبكة عصبونية واحدة للتعلم، يمكننا استخدام شبكتين عصبونيتين، حيث نستخدم شبكة منفصلة لتقدير الهدف لها نفس بنية تابع التقريب ولكن مع بارامترات ثابتة (محددة). و في كل تكرار، يتم إعادة توليد البارامترات من شبكة التنبؤ إلى شبكة الهدف، يؤدي هذا إلى تدريب أكثر استقراراً لأنه يحافظ على ثبات تابع الهدف [10][9].

4-4 تقدير الهدف

توضّح الخوارزمية الآتية كيفية اختيار الإجراء المناسب للوصول إلى الهدف، فالإجراء التالي الذي يجب أن يقوم به الوكيل إمّا يتم اختياره عشوائياً ضمن مجال محدد أو يتم تحديده عن طريق شبكة DQN اعتماداً على معدّل التعلم α وفق الشكل(3):



شكل (3) خوارزمية تقدير الهدف

4-5 جدولة التعزيز RS (Reinforcement scheduling)

هي خوارزمية حديثة لتعيين مواد الدورة التدريبية بناءً على حالة المعرفة الأولية للمتعلم وتاريخ التفاعل مع مواد الدورة التدريبية، يمكن لنموذجنا أن يتعلم اتخاذ قرارات تعيين الأنشطة التي ينبغي على الطلاب إجراءها في الزمن الحقيقي مع إعطاء ملاحظات حول كيفية تأثير هذه القرارات على درجات ما بعد الاختبار، ويمثل نموذجنا المعرفة المسبقة للمتعلم من خلال درجة ما قبل الاختبار وتفاعلاته مع الدورة التدريبية التي تبين المشكلات المحتملة وما إذا كانت قد تمت الإجابة عليها بشكل صحيح قبل الوصول للاختبار النهائي.

تتعلم جدولة التعزيز كيفية تعيين تسلسلات من الأنشطة التعليمية للمتعلمين عبر الإنترنت دون الحاجة إلى تسمية المهارات أو بيانات الدورة التدريبية الحالية، بالتالي هي قابلة للتعميم لإمكانية تطبيقها على مختلف أنواع الدورات باختلاف المهارات التي تتضمنها، حيث أنّ خوارزمية Q-Learning تقوم بالتمييز بين المهارات المختلفة.

5- تطوير النموذج

كان هدفنا الأســـاســـي من جدولة التعزيز هو تعظيم مكافأة التعلم مع تقليل الزمن المســـتغرق في الأنشطة التعليمية، لذلك اعتمدنا ثلاثة مبادئ تصميمية تشجع على إمكانية التعميم:

1. يجب أن يدعم النموذج الدورات التدريبية ذات الأنشطة والموضوعات التعليمية المختلفة.
2. يجب أن تكون الميزات ذات مغزى ومتسقة وغير مكررة.
3. يجب ألا يكون هناك متطلب للتسميات البشرية.

لتحقيق هذه الأهداف، اعتمدنا على التعلم المعزز لاتخاذ القرارات والتعلم من نتائجها بشكل مستمر وفي الزمن الحقيقي.

بالإضافة إلى ذلك، وضعنا **قيدين** على الإجراءات التي يتخذها وكيل التعلم المعزز:

أولاً، منعنا تكرار تعيين نفس النشاط التعليمي لنفس المتعلم؛ حيث بدا التكرار غير ضروري في دورة بها 12 نشاطاً تعليمياً فقط، وأردنا التأكد من أن طول الدورة لا يتجاوز 12 نشاطاً.

ثانياً، منعنا وكيل التعلم المعزز من تعيين الاختبار اللاحق للمتعلم حتى يكون قد أعطاه بالفعل نشاطاً تعليمياً واحداً على الأقل؛ وهذا يحمي من استكشاف مسار عديم الفائدة، وإحباط المتعلم الذي من المرجح أن يكون سبباً في فشله في التعلم.

نظراً لأهدافنا فيما يتعلق بطول الدورة، قمنا بتدريس ثلاث مهارات أساسية تتضمن فتح أو إغلاق القاطع الصحيح، ووضع العنصر في مكانه المناسب والقيام بالحسابات حسب القوانين اللازمة.

لقد أنشأنا مجموعة من اثني عشر نشاطاً تعليمياً يتم تصنيفها تلقائياً لتتناسب مع هذا المنهج، كل منها يتطلب إجابة متعددة الاختيارات، وقد تم تصميمها لتستغرق خمس دقائق، تم إنشاء أربعة أنواع من الأنشطة التعليمية لكل مهارة: تفسيرات الفيديو والأوصاف المكتوبة والأمثلة العملية وأسئلة التقييم، ولقياس مكافأة التعلم من بداية الدورة إلى نهايتها، استخدمنا اختباراً مكوناً من ست مشكلات يغطي المهارات الثلاث للدورة كاختبار مسبق واختبار لاحق متطابقين، سيتعلم وكيل التعلم المعزز تعيين أنشطة تعليمية مختلفة للمتعلمين، واستكشاف مجموعة المسارات المحتملة، واستغلال المسارات التي أظهرت نتائج جيدة لمتعلمين مماثلين في الماضي،

ولتزويد نموذج التعلم المعزز بمجموعة متسقة من الميزات في الزمن الحقيقي، سجل كل مشكلة اختبار ونشاط تعليمي درجة ثنائية 1 أو 0 بعد أن استجاب المتعلم لها، أما الأنشطة التي لم تشكل مشكلة للمتعلمين لحلها، فقد تمَّ سؤالهم بدلاً من ذلك: "هل فهمت المحتوى المقدم أعلاه؟". وكان المتعلمون يختارون ما إذا كانوا "فهموا تماماً" أو "لم يفهموا تماماً"، مما أسفر عن درجات 1 أو 0 على التوالي، وقدمت أسئلة التقييم ردود فعل بناءً على استجابات المتعلمين، ونظراً لأن مشاكل الاختبار كانت تهدف إلى أن تكون تقييمية بحتة، لم يرَ المتعلمون درجات الاختبار أو ردود الفعل حتى الانتهاء من الاختبار اللاحق.

تتضمن منصة التعلم الخاصة بنا واجهة برمجة تطبيقات API لتتبع المتعلمين والمهام التكيفية، مما يسمح لها بالتفاعل مع خوارزمية جدولة RS تتوافق مع مواصفاتها، مع كل استجابة جديدة للمتعلم ستتم إضافة درجته الثنائية إلى تتبعه، وستستخدم خوارزمية الجدولة هذه المعلومات الجديدة لتحديد النشاط التعليمي التالي مباشرة، تتكون تتبّعات المتعلمين من "إجراءات" التفاعل مع: معرف الدورة، ومعرف النشاط، ومعرف المتعلم، واختيار الإجابة، والنتيجة، والطابع الزمني.

6- إعداد التجربة

لفهم كيفية تأثير طريقة RS على المتعلمين، قمنا بمقارنتها مع طريقتين هما Self-Directed إذ طلب من المستخدمين اختيار مسارهم الخاص في الدورة ، وLinear حيث طلب من المتعلمين إكمال جميع الأنشطة التعليمية بترتيب محدد، وفي كل حالة، أعطي المتعلمون أولاً اختباراً أولياً لتقييم حالة المعرفة الأولية لديهم ثم اختباراً لاحقاً لتقييم حالة المعرفة لديهم بعد إكمال الدورة، وقد استخدمنا نفس الأسئلة لكل من الاختبار السابق والاختبار اللاحق، ومن الجدير بالذكر أن المتعلمين لم يتلقوا ملاحظات على أي من حلول الاختبار الخاصة بهم حتى اكتمال الاختبار النهائي.

وقمنا بقياس مكافأة التعلم باعتبارها الفرق في درجة المتعلم بين الاختبار اللاحق والاختبار السابق.

كما جمعنا بيانات المتعلمين طوال الدورة من خلال استبيان تضمّن: عدد الأنشطة التي أكملها المتعلم وما هي تلك الأنشطة، وما إذا كان المتعلم قد انسحب ومتى.

وفي هذا السياق، لقد حسبنا عدد الأنشطة التي أكملها كل متعلم من خلال تفاعلات المتعلم مع النظام، ولحساب معدلات التسرب، قمنا بحساب عدد المتعلمين الذين أكملوا الاختبار المسبق للدورة ولكن ليس الاختبار اللاحق.

قمنا بتعيين المشاركين في الدورة بشكل عشوائي حيث تم تقسيمهم كما يلي:

95% من المشاركين يتم توجيههم وفق شروط جدولة التعزيزRS.

2.5% منهم وفق الحالة الخطية Linear.

2.5% حسب طريقة Self-Navigation.

في كل حالة، تلقى المتعلمون تعليمات قصيرة توضح إعداد الدورة التي سيختبرونها، لقد شرحنا أهداف الدورة، وأبلغنا المتعلمين أنهم جزء من تجربة تعليمية، ومن أجل الحصول على شهادة إتمام الدورة، طُلب من جميع المتعلمين إكمال استبيان ما بعد الدورة.

التحق بالدورة 1987 شخصاً، ومن بين الأشخاص الذين تم تعيينهم عشوائيًا لكل حالة: أكمل 1830 شخصاً التسجيل في حالة جدولة التعزيز، و91 شخصاً في الحالة الخطية، و66 شخصاً في حالة self– navigation.

7- مقارنة الأداء

7-1 معدلات إتمام الدورة

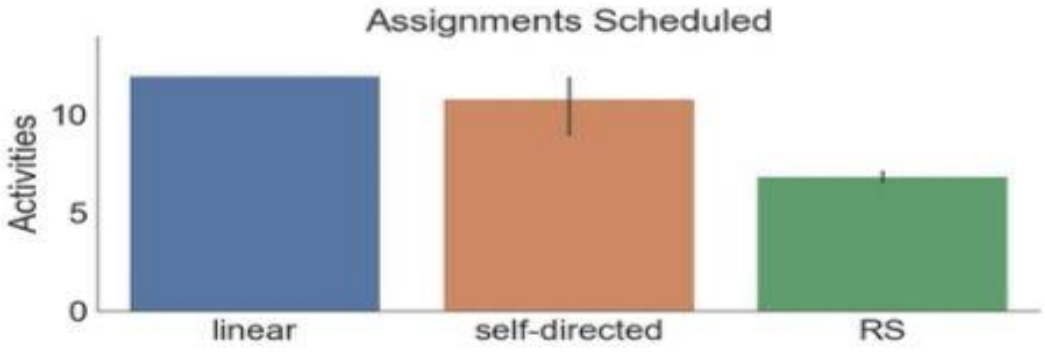
يوضّح الشكل (4) أن حالتي RS وLinear شهدتا معدلات إكمال أعلى للاختبار النهائي (أي قاموا بإتمام الدورة حتى النهاية) بشكل ملحوظ من Self–Navigation.

شكل (4) معدلات إتمام الدورة

7-2 عدد الأنشطة

خضع المتعلمون في حالة RS لعدد أقل بكثير من الأنشطة التعليمية مقارنة بالمتعلمين في الحالتين الأخريين.

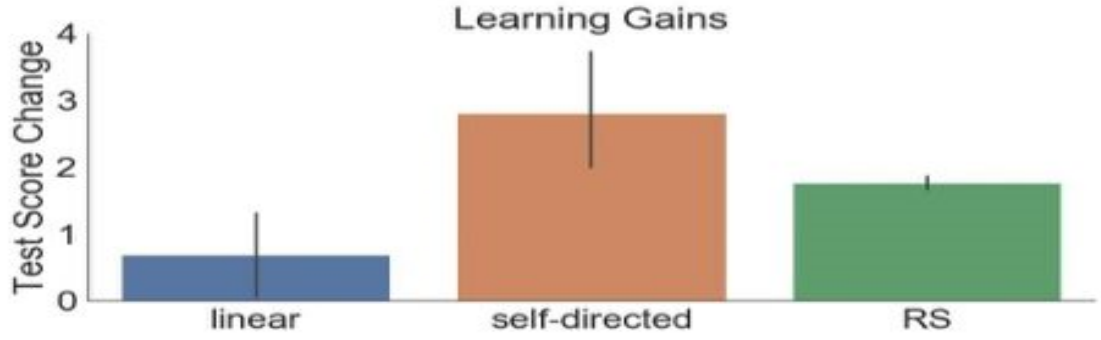
وكما هو موضّح في الشكل (5) انخفض عدد الأنشطة التعليمية المخصصة لهم من قبل RS بمرور الزمن؛ والأنشطة المذكورة هنا هي متوسط لجميع المتعلمين.



شكل (5) عدد الأنشطة

7-3 درجات الاختبار

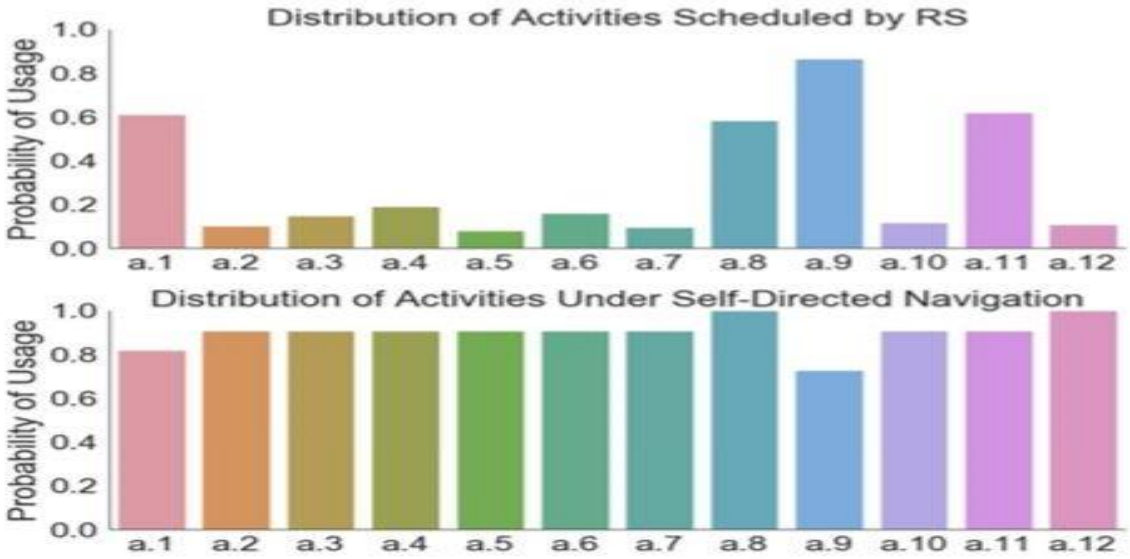
يوضّح الشكل (6) أنّ المتعلمين في حالتي RS وSelf–Directed Navigation أظهروا تحسناً كبيراً في درجات الاختبار مقارنة بالمتعلمين الذين اتبعوا حالة Linear.



شكل (6) درجات الاختبار

7-4 جدولة الأنشطة التعليمية

يبيّن الشكل (7) أنّه في نموذج RS، كان المتعلمون أكثر عرضة لرؤية الأنشطة التعليمية 1 و8 و9 و11. بينما انخرط المتعلمون وفق Self-Directed في كل نشاط في الدورة بالتردد نفسه.

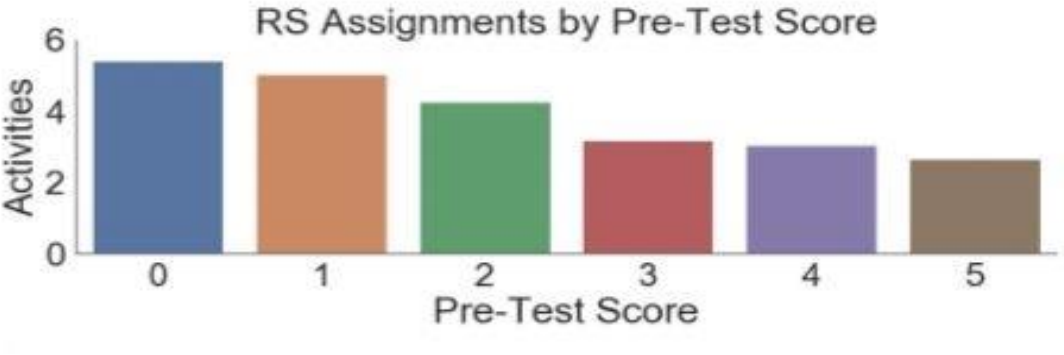


شكل (7) جدولة الأنشطة التعليمية

7-5 Pre-Test-Score

يوضّح الشكل (8) عدد الأنشطة المخصصة لآخر 200 متعلم في نموذج RS، استناداً إلى درجاتهم في الاختبارات الأولية، حيث لم يعانوا من مشكلة البداية الباردة.

وكلما كان أداء المتعلمين أفضل في الاختبارات الأولية، كان RS يخصص عدداً أقل من الأنشطة.



شكل (8) Pre-Test-Score

8- النتائج و المناقشة

8-1 وصف قاعدة البيانات و بيئة العمل

قاعدة البيانات المستخدمة في هذه الدراسة تتضمن صور للمنصة الإلكترونية من عام 2017 إلى عام 2020.

تم تقسـيم الطالب البالغ عددهم 1800 طالب إلى مجموعات مختلفة، تم جمع عدد كبير من الصـور، حيث كلنت كل نقرة يقوم بها الطالب مصـحوبة بصـورة تم التقاطها وتخزينها في ملف كل طالب، إذ تم الحصول على أكثر من 20 مليون صورة باليوم الواحد وكان ذلك كافياً لتدريب وكيلنا [8].

بيئة التعلم المعزز في حالتنا هي دائرة كهربائية تتكون من بعض القواطع والمكونات الإلكترونية التي يمكن للوكيل النقر فوق أحدها ويمكنه أيضـاً تحريك أو إضـافة أو حذف بعض المكونات، هدف التدريب هو النقر على القاطع الصحيح أو وضع المكون في مكان مناسب.

البارامترات من البيئة هي صورة الدارة وحالة القواطع، أمّا المكافأة هي 1+ للنقر على القاطع الصحيح والوضع المثالي للمكون، و 1- للنقر على القاطع الخاطئ والموضع الخطأ للمكون، و 0 لعدم النقر، أو النقر في مكان آخر، أو عدم تحريك المكونات.

نسـتخدم Critic Value Function لجعل وكيل DQN يقترب من المكافأة لذا، فإن Critic Value Function عبارة عن شبكة عصبونية عميقة لها دخل واحد (صورة) وخرج واحد (إجراء مثالي).

8-2 تدريب الوكيل

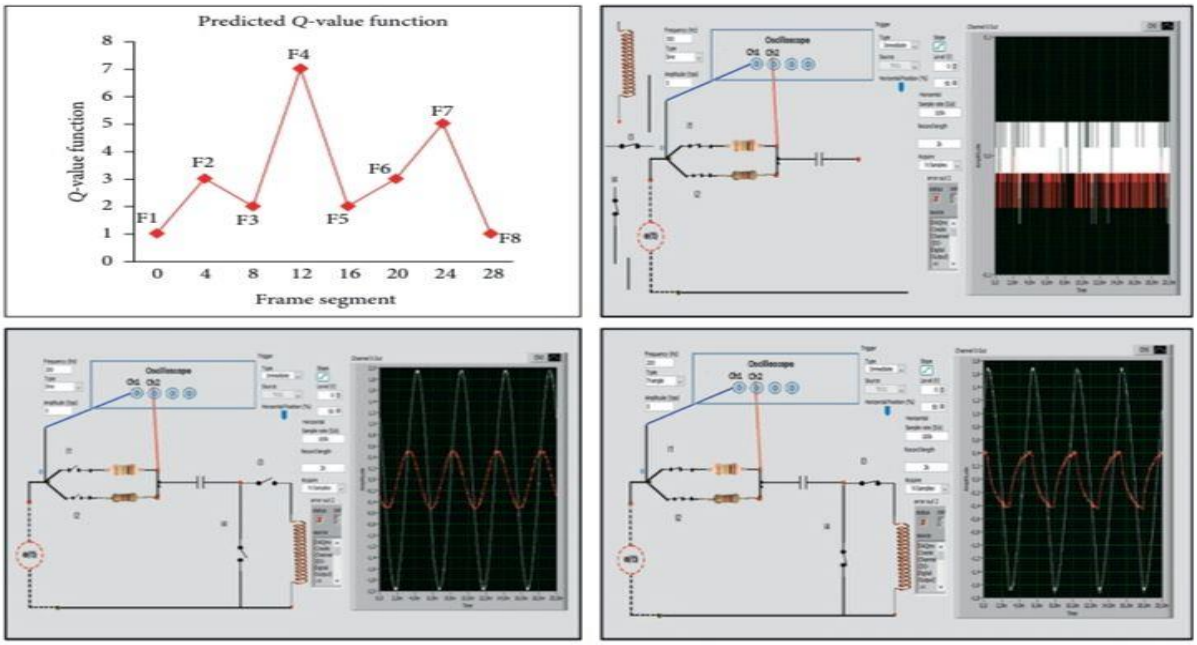
لتدريب وكيلنا، نحدد بارامترات التدريب بالشكل التالي:

نقوم بتشغيل برنامجنا 8000 episode كحد أقصى بحيث تستمر كل episode حوالي 800 خطوة زمنية كحد أقصى، ثم نوقف التدريب عندما يتلقى الوكيل متوسط مكافأة أكبر من 1500- على طول النافذة الافتراضي الذي يبلغ 15 episode متتالية، و عند ذلك، يكون الوكيل قادراً على فتح أو إغلاق القاطع الصحيح أو وضع أحد المكونات الإلكترونية في مكانه المناسب.

8-3 دالة القيمة Q-value

يوضّح الشكل (9) أن القيمة المتوقعة تبين وضع المكونات في مكانها الصحيح وذلك في النقطة F1 ثم يغلق الوكيل جميع قواطع الدارة ويبدأ التدريب، وتبلغ القيمة المتوقعة ذروتها في النقطة F4 أي يصبح لدينا فرط ملاءمة

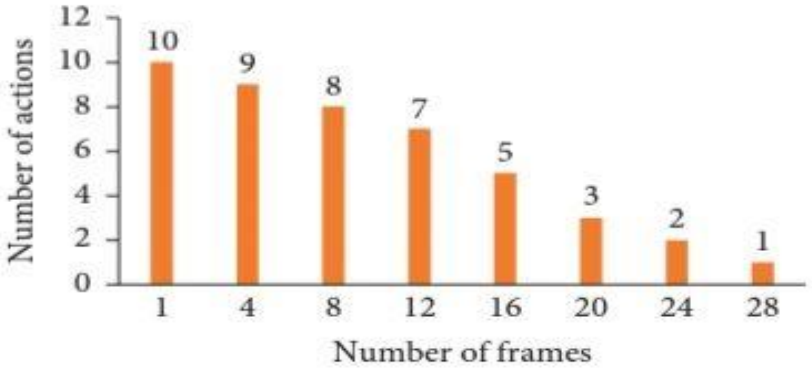
في التدريب (overfitting) ، ثم تنخفض هذه القيمة إلى قيمتها الأصلية تقريباً بعد أن يأخذ الوكيل الإجراء المناسب ليحقق الهدف في النقطة F8.



شكل (9) Q-value function

8-4 تحليل الإجراءات

يُظهر الشكل (10) عدد الإجراءات التي يجب اتخاذها في كل إطار لإيجاد الموضع الصحيح ومن الواضح أن العدد الأقل من الإجراءات يُظهر كفاءة أفضل. لذا، فإن الطريقة المقترحة فيما يقارب 93% من المرات تجد الإجراءات الصحيحة في أقل من 5 إجراءات (أنشطة) في كل إطار، لأن عدد الإجراءات الصحيحة أكبر من عدد الإجراءات الأخرى.

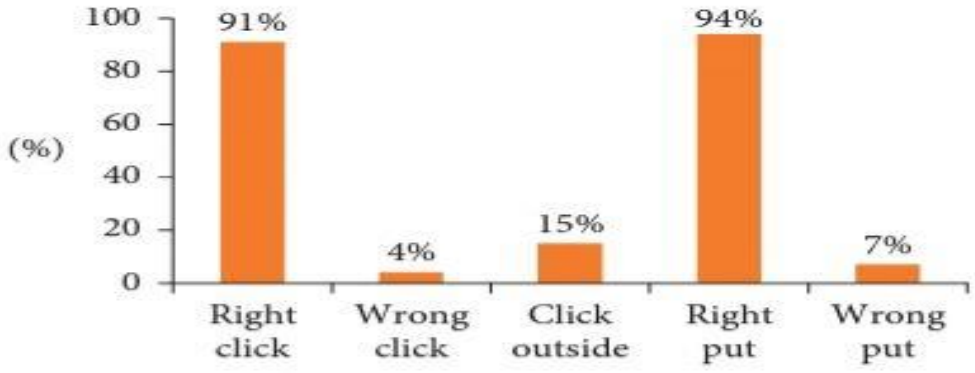


شكل (10) تحليل الإجراءات

8-5 تحقيق الهدف

نلاحظ من الشكل (11) أن معدل النقر الصحيح للقاطع ووضع العناصر الإلكترونية في المكان الصحيح من قبل الوكيل هو 92.5% مما يدل على أن النظام قد تم تعلمه بشكل صحيح ودقيق بينما معدل الخطأ من قبل الوكيل هو 7.5% فقط للنقر الخاطئ على القاطع ووضع العناصر الإلكترونية في المكان

الخطأ والضغط في مكان آخر، لذا فإن تدريب النظام بشكل جيد والحصول على نتائج مرضية يعتمد على دراسة موقف المشكلة بشكل جيد واختيار البارامترات المستخدمة في النظام بشكل صحيح وبما يتناسب مع المطلوب.



شكل (11) النسب المئوية للنتائج

9- التوصيات المستقبلية

1. سيكون من المثير للاهتمام أن نرى كيف يتصرف نموذج RL إذا وضعنا حداً لعدد الإجراءات، فإذا سمحنا لوكيل RL بإجراء مهام متكررة، سيتطلب هذا مساحة حالة أكبر لنموذجنا ويجعل وكيل RL أقل كفاءة في أخذ العينة، ولكن قد يستفيد المتعلمون من هذه التكرارات بطرق غير متوقعة.

2. كما يمكن أن نضيف مصطلحاً إلى دالة المكافأة الخاصة بنا يعاقب على عدم إكمال المتعلمين للاختبارات و ذلك لمعرفة ما إذا كان وكيل RL قادراً على إبقاء المتعلمين منخرطين في الدورة لفترة أطول ومنع بعض حالات التسرب، و سنحاول إزالة العقوبة (عامل الخصم) المفروضة على تحديد الأنشطة التعليمية الإضافية من دالة المكافأة، مما يمكن RS من تقليل عدد المهام دون عقوبة صريحة.

المراجع

[1] Sergio Re, Marco Matta, Alberto Carlo, "An Efficient Hardware Implementation of Reinforcement Learning",IEEE, vol. 7, 2019.

[2] Lukasz Kaiser , Roy H , Chelsea Finn, "Model Based Reinforcement Learning for Atari", arXiv:1903.00374v4, 2020.

[3] Alex Graves , Daan Wierstra , David Silver, "Playing Atari with Deep Reinforcement Learning", arXiv:1312.5602v1, 2013.

[4] Yufeng Yuan and Rupam Mahmood, "Asynchronous Reinforcement Learning for Real-Time Control of Physical Robots", arXiv:2203.12759v3, 2022.

[5] Aleksander Izemski , Kamil Plucinski , Mateusz Lango, "Can Reinforcement Learning Agents be E-Teachers?", Poznan, Poland, 2020.

[6] Markel Sanz Ausin , Tiffany Barnes , Min Chi, "Leveraging Deep Reinforcement Learning for Pedagogical Policy Induction in an Intelligent Tutoring System", North Carolina State University, EDM, 2019.

[7] Lu Chen , Xiang Zhou , Kai Yu, "Online Dialogue Policy Learning with Companion Teaching", Valencia, Spain, 2017.

[8] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning " Nature, vol. 518, no. 7540, pp. 529–533, 2015.

[9] A. Choudhary, "A hands-on introduction to deep Q-learning using OpenAI gym in Python ", 2018, https://www.analyticsvidhya.com/blog/2019/04/introduction-deep -q-learning-python/.

[10] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters", in Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), pp. 3207–3214, New Orleans, LA, USA, February 2018.

[11] A. Nair, "Massively parallel methods for deep reinforcement learning", arXiv:1507.04296v2, 2015.