Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών

Συστήματα Ασύρματης Επικοινωνίας

Μεταπτυχιακή Διατριβή

Βαθιά Ενίσχυση της Μάθησης στις Επικοινωνίες και τη Δικτύωση

Ιωάννης Ζιάμος

Επιβλέπων Καθηγητής

Δημοσθένης Βουγιούκας

Νοέμβριος 2021

ΛΕΥΚΗ ΣΕΛΙΔΑ

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών

Συστήματα Ασύρματης Επικοινωνίας

Μεταπτυχιακή Διατριβή

Βαθιά Ενίσχυση της Μάθησης στις Επικοινωνίες και τη Δικτύωση

Ιωάννης Ζιάμος

Επιβλέπων Καθηγητής

Δημοσθένης Βουγιούκας

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών στα Συστήματα Ασύρματης Επικοινωνίας από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών του Ανοικτού Πανεπιστημίου Κύπρου.

Νοέμβριος 2021

ΛΕΥΚΗ ΣΕΛΙΔΑ

Περίληψη

Στην παρούσα διατριβή παρουσιάζονται οι διάφορες τεχνικές που υπάρχουν στη βιβλιογραφία και πραγματεύονται τη βαθιά ενίσχυση της μάθησης (Deep Reinforcement Learning - DRL) στις επικοινωνίες και τη δικτύωση. Τα τελευταία χρόνια με τη ραγδαία τεχνολογική ανάπτυξη, τα δίκτυα, όπως αυτά των Internet of Things αλλά και των μη επανδρωμένων οχημάτων (Unmanned Aerial Vehicles – UAV), γίνονται πιο αποκεντρωμένα και αυτόνομα. Στα παραπάνω δίκτυα, οι οντότητες δικτύου, για να μεγιστοποιήσουν την απόδοσή του, πρέπει να λαμβάνουν αποφάσεις όσο το δυνατόν αυτόνομα και τοπικά, πάντοτε υπό την αβεβαιότητα του περιβάλλοντος δικτύου. Στις περιπτώσεις που τα ανωτέρω περιβάλλοντα αφορούν σε περιορισμένους χώρους, η ενίσχυση μάθησης (Reinforcement Learning) μπορεί να χρησιμοποιηθεί αποτελεσματικά για να επιτρέψει στις οντότητες να αποφασίσουν τη βέλτιστη πολιτική. Στις περιπτώσεις, όμως που ο χώρος δράσης είναι μεγάλος και τα δίκτυα πολύπλοκα, η ενίσχυση μάθησης δεν μπορεί να είναι αποτελεσματική στην εύρεση της βέλτιστης πολιτικής. Αυτό κενό έρχεται να καλύψει η DRL, που είναι ένας συνδυασμός ενίσχυσης μάθησης με βαθιά μάθηση. Η διατριβή παρουσιάζει θεμελιώδεις έννοιες της DRL, τα πιο προηγμένα μοντέλα της καθώς και επιχειρεί να εξετάσει ζητήματα δυναμικής πρόσβασης στο δίκτυο, ελέγχου ρυθμού δεδομένων, ασύρματης προσωρινής αποθήκευσης, εκφόρτωσης δεδομένων, ασφάλειας δικτύου και συντήρησης συνδεσιμότητας που είναι όλα σημαντικά για δίκτυα επόμενης γενιάς, όπως το 5G και πέραν αυτών.

Summary

This thesis presents the various techniques available in the literature that deal with Deep Reinforcement Learning (DRL) in communications and networking. In recent years, with the rapid technological development, the networks, such as those of the Internet of Things but also of the Unmanned Aerial Vehicles (UAV), are becoming more decentralized and autonomous. In the above networks, the network entities, in order to maximize its efficiency, must make decisions as autonomously and locally as possible, always under the uncertainty of the network environment. In cases where the above environments are limited, Reinforcement Learning can be used effectively to enable entities to decide on the best policy. However, in cases where the action space is large and the networks are complex, Reinforcement Learning may not be effective in finding the best policy. This gap is being filled by DRL, which is a combination of Reinforcement Learning with deep learning. The thesis introduces fundamental concepts of DRL, its most advanced models and attempts to address issues of dynamic network access, data rate control, wireless caching, data offloading, network security and connectivity preservation that are all important to next generation networks 5G and beyond.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Η ενίσχυση μάθησης (Sutton & Barto, 1998) είναι μια από τις πιο σημαντικές ερευνητικές κατευθύνσεις της μηχανικής μάθησης που έχει σημαντικές επιπτώσεις τα τελευταία 20 χρόνια στην ανάπτυξη της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI). Η ενίσχυση μάθησης είναι μια μαθησιακή διαδικασία στην οποία ένας πράκτορας μπορεί περιοδικά να λαμβάνει αποφάσεις, να παρατηρεί τα αποτελέσματα και, στη συνέχεια, να προσαρμόζει αυτόματα τη στρατηγική του για να επιτύχει τη βέλτιστη πολιτική. Ωστόσο, αυτή η μαθησιακή διαδικασία, παρόλο που αποδείχθηκε ότι αποδίδει, χρειάζεται πολύ χρόνο για να επιτύχει την καλύτερη πολιτική καθώς πρέπει να διερευνήσει και να αποκτήσει γνώση ενός ολόκληρου συστήματος, καθιστώντας το ακατάλληλο και ανεφάρμοστο σε δίκτυα μεγάλης κλίμακας. Κατά συνέπεια, οι εφαρμογές της ενίσχυσης μάθησης είναι πολύ περιορισμένες στην πράξη. Πρόσφατα, η βαθιά μάθηση (Goodfellow, Bengio, & Courville, 2016) εισήχθη ως μια νέα καινοτόμος τεχνική. Μπορεί να ξεπεράσει τους περιορισμούς της ενίσχυσης μάθησης, και επομένως να ανοίξει μια νέα εποχή για τη βελτίωσή της, ως Βαθιά Ενίσχυση Μάθησης (DRL). Το DRL αγκαλιάζει το πλεονέκτημα των βαθιών νευρωνικών δικτύων (Deep Neural Networks - DNNs) για την εκπαίδευση της μαθησιακής διαδικασίας, βελτιώνοντας έτσι την ταχύτητα εκμάθησης και την απόδοση των αλγορίθμων ενίσχυσης μάθησης. Ως αποτέλεσμα, το DRL έχει υιοθετηθεί σε πολλές εφαρμογές της ενίσχυσης μάθησης στην πράξη, όπως η ρομποτική, η υπολογιστική όραση (computer vision), η αναγνώριση ομιλίας και η επεξεργασία φυσικής γλώσσας (Goodfellow, Bengio, & Courville, 2016). Μία από τις πιο διάσημες εφαρμογές του DRL είναι το AlphaGo (BBC, 2016), ο πρώτος υπολογιστής Γκο που νίκησε επαγγελματία παίκτη Γκο επί ίσοις όροις σε ένα φυσικό μέγεθος ταμπλό 19 × 19.

Στους τομείς των επικοινωνιών και της δικτύωσης, το DRL χρησιμοποιήθηκε πρόσφατα ως αναδυόμενο εργαλείο για την αποτελεσματική αντιμετώπιση διαφόρων προβλημάτων και προκλήσεων. Συγκεκριμένα, τα σύγχρονα δίκτυα όπως το Internet of Things (IoT), τα ετερογενή δίκτυα (HetNets) και το δίκτυο μη επανδρωμένων εναέριων

οχημάτων (UAV) καθίστανται πιο αποκεντρωμένα, ad-hoc και αυτόνομα. Οι οντότητες δικτύου όπως συσκευές IoT, χρήστες κινητών συσκευών και UAV πρέπει να λάβουν τοπικές και αυτόνομες αποφάσεις όπως για παράδειγμα, πρόσβαση φάσματος, επιλογή ρυθμού δεδομένων, έλεγχος ισχύος μετάδοσης και συσχέτιση σταθμού βάσης, για την επίτευξη των στόχων διαφορετικών δικτύων, όπως μεγιστοποίηση της απόδοσης και ελαχιστοποίηση της κατανάλωσης ενέργειας. Σε αβέβαιο και στοχαστικό περιβάλλον, τα περισσότερα από τα προβλήματα λήψης αποφάσεων μπορούν να μοντελοποιηθούν από τη λεγόμενη διαδικασία αποφάσεων Markov (Markov Decision Process - MDP) (Puterman, 2014). Για την επίλυση του MDP μπορούν να υιοθετηθούν ο δυναμικός προγραμματισμός (Bertsekas, 2005), (Bellman, 2013) και προσεγγιστικοί αλγόριθμοι, καθώς και τεχνικές ενίσχυσης μάθησης. Ωστόσο, τα σύγχρονα δίκτυα είναι μεγάλης κλίμακας και περίπλοκα, και έτσι η υπολογιστική πολυπλοκότητα των τεχνικών γίνεται γρήγορα ανεξέλεγκτη. Ως αποτέλεσμα, το DRL εξελίσσεται ως εναλλακτική λύση για την αντιμετώπιση της πρόκλησης. Γενικά, οι προσεγγίσεις DRL παρέχουν τα ακόλουθα πλεονεκτήματα:

- Το DRL μπορεί να επιτύχει βελτιστοποίηση εξελιγμένων δικτύων. Έτσι, επιτρέπει στους ελεγκτές δικτύου, π.χ. σταθμούς βάσης σε σύγχρονα δίκτυα, να επιλύουν μη κυρτά και πολύπλοκα προβλήματα, π.χ. από κοινού συσχέτιση χρήστη, υπολογισμό και προγραμματισμό εκπομπών για να επιτύχουν βέλτιστες λύσεις χωρίς πλήρεις και ακριβείς πληροφορίες δικτύου.
- Το DRL επιτρέπει στις οντότητες δικτύου να μάθουν και να οικοδομήσουν γνώσεις σχετικά με το περιβάλλον επικοινωνίας και δικτύωσης. Έτσι, χρησιμοποιώντας το DRL, οι οντότητες δικτύου, π.χ. χρήστες κινητής τηλεφωνίας, μπορούν να εφαρμόσουν βέλτιστες πολιτικές, π.χ. επιλογή σταθμού βάσης, επιλογή καναλιού, απόφαση μεταβίβασης κλήσης-προσωρινής αποθήκευσης και εκφόρτωσης δεδομένων, χωρίς να γνωρίζουν το μοντέλο καναλιού και το μοτίβο κινητικότητας.
- Το DRL παρέχει αυτόνομη λήψη αποφάσεων. Με τις προσεγγίσεις DRL, οι οντότητες δικτύου μπορούν να κάνουν παρατήρηση και να εφαρμόσουν την καλύτερη πολιτική τοπικά με ελάχιστη ή χωρίς ανταλλαγή πληροφοριών μεταξύ τους. Αυτό όχι μόνο μειώνει τα γενικά έξοδα επικοινωνίας, αλλά επίσης βελτιώνει την ασφάλεια και την ευρωστία των δικτύων.
- Το DRL βελτιώνει σημαντικά την ταχύτητα μάθησης, ειδικά σε προβλήματα με μεγάλους χώρους κατάστασης και δράσης. Έτσι, σε δίκτυα μεγάλης κλίμακας, π.χ. σε

συστήματα IoT με χιλιάδες συσκευές, το DRL επιτρέπει στον ελεγκτή δικτύου ή σε πύλες IoT να ελέγχουν δυναμικά τη συσχέτιση των χρηστών, την πρόσβαση στο φάσμα και να μεταδίδουν ισχύ για έναν τεράστιο αριθμό συσκευών IoT και χρηστών κινητών.

- Αρκετά άλλα προβλήματα στις επικοινωνίες και τη δικτύωση, όπως οι επιθέσεις στον κυβερνοχώρο, η διαχείριση παρεμβολών και η εκφόρτωση δεδομένων μπορούν να μοντελοποιηθούν ως παίγνια, π.χ. το μη συνεργατικό παίγνιο. Το DRL χρησιμοποιήθηκε πρόσφατα ως αποτελεσματικό εργαλείο για την επίλυση των παιγνίων, π.χ. για την εύρεση της ισορροπίας Nash, χωρίς τις πλήρεις πληροφορίες.

Επισημαίνεται ότι, το DRL αποτελείται από δύο διαφορετικούς αλγόριθμους που είναι Deep Q-Learning (DQL) και η policy gradient (Lowe, Wu, Tamar, Harb, Abbeel, & Mordatch, 2017). Συγκεκριμένα, η DQL χρησιμοποιείται κυρίως στις εργασίες που σχετίζονται με το DRL. Επομένως, στην εργασία, χρησιμοποιούνται τα DRL και DQL εναλλακτικά για την αναφορά στους αλγόριθμους DRL.

Παρόλο που υπάρχουν ορισμένες έρευνες που σχετίζονται με τη μηχανική μάθηση, δεν εξετάζουν τις εφαρμογές του DRL στις επικοινωνίες και τη δικτύωση. Συγκεκριμένα, υπάρχουν έρευνες που πραγματεύονται με τις εφαρμογές DRL όπως (Li Y. , 2018) και (Arulkumaran, Deisenroth, Brundage, & Barath, 2017), αλλά είναι ειδικά για την υπολογιστική όραση και την επεξεργασία φυσικής γλώσσας. Επίσης, υπάρχουν έρευνες που εξετάζουν τις εφαρμογές της μηχανικής μάθησης για δικτύωση όπως (Xin, 2018), (Fadlullah, et al., 2017), (Mao, Hu, & Hao, 2018), (Chen M. , Challita, Saad, Yin, & Debbah, 2017) και (Wang, Kwasinski, Niyato, & Han, 2016). Ωστόσο, εστιάζουν κυρίως σε προσεγγίσεις βαθιάς μάθησης. Συγκεκριμένα, η έρευνα στο (Xin, 2018) ασχολείται με προσεγγίσεις βαθιάς μάθησης για την ασφάλεια στον κυβερνοχώρο του δικτύου, η έρευνα στο (Fadlullah, et al., 2017) εξετάζει προσεγγίσεις βαθιάς μάθησης για έλεγχο του δικτύου, η έρευνα στο (Mao, Hu, & Hao, 2018) παρουσιάζει προσεγγίσεις βαθιάς μάθησης για διαμόρφωση φυσικού επιπέδου, πρόσβαση στο δίκτυο / κατανομή πόρων, και δρομολόγηση δικτύου, και η έρευνα στο (Chen M. , Challita, Saad, Yin, & Debbah, 2017) παρουσιάζει προσεγγίσεις βαθιάς μάθησης για αναδυόμενα ζητήματα, όπως η προσωρινή αποθήκευση και υπολογισμός, η πολλαπλή πρόσβαση ραδιοφώνου και η διαχείριση παρεμβολών. Συνοπτικά, οι υπάρχουσες έρευνες είτε εξετάζουν εφαρμογές DRL για

όραση υπολογιστή και επεξεργασία φυσικής γλώσσας είτε πραγματεύονται εφαρμογές βαθιάς μάθησης για δικτύωση.

Τα βασικά ερευνητικά ερωτήματα της διατριβής περιλαμβάνουν τον τρόπο που η βαθιά ενίσχυση μάθησης επιτυγχάνει την πρόσβαση στο δίκτυο, τον έλεγχο ρυθμού δεδομένων, την ασύρματη προσωρινή αποθήκευση, την εκφόρτωση των δεδομένων, την ασφάλεια δικτύου και τη διατήρηση συνδεσιμότητας στις επικοινωνίες και τη δικτύωση. Ο σκοπός της εργασίας είναι αρχικά να παραθέσει τις τεχνικές DRL και τις επεκτάσεις τους που μελετώνται στη διατριβή, να απαντήσει στα παραπάνω βασικά ερευνητικά ερωτήματα με τη μελέτη των διαφόρων τεχνικών που υπάρχουν στη βιβλιογραφία και αναφέρονται σαφώς σε δίκτυα επόμενης γενιάς 5G και πέραν αυτών και τέλος να παρουσιάσει τα αποτελέσματα και τις προκλήσεις της έρευνας. Τα προσδοκώμενα αποτελέσματα είναι η παράθεση συγκριτικών πινάκων για κάθε ένα από τα ανωτέρω βασικά ερευνητικά ερωτήματα, έτσι ώστε να διεξαχθούν τα αποτελέσματα της έρευνας στα οποία φαίνεται, ανά ερώτημα, ποιοι είναι οι πιο κατάλληλοι αλγόριθμοι DRL που μπορούν να χρησιμοποιηθούν, ποια είναι τα πλεονεκτήματα – μειονεκτήματα τους, σε ποια δίκτυα μπορούν να αυτοί να εφαρμοστούν και τέλος ποιες είναι οι οντότητες του δικτύου και πως αυτές συμπεριφέρονται σε αυτό. Η μεθοδολογία που ακολουθείται για την επίτευξη του σκοπού της διατριβής είναι η βιβλιογραφική μελέτη των τεχνικών DRL στις επικοινωνίες και τη δικτύωση και αφορούν στα ερωτήματα της έρευνας, η ανάλυση των ερευνητικών δεδομένων και τέλος η παράθεση των αποτελεσμάτων. Η συνεισφορά της εργασίας που αναδεικνύει και την σπουδαιότητά της εστιάζεται σε δύο τομείς. Ο πρώτος είναι η συγκριτική παράθεση των τεχνικών DRL που υπάρχουν στη βιβλιογραφία και πραγματεύονται τα προαναφερόμενα βασικά ερευνητικά ερωτήματα για τις επικοινωνίες και τη δικτύωση, καθώς και τα αποτελέσματα που διεξάγονται από την παραπάνω μελέτη. Ο δεύτερος είναι η συγκριτική μελέτη των τεχνικών DRL για συγκεκριμένη εφαρμογή που αφορά στις ασφαλείς επικοινωνίες με χρήση UAV. Η παραπάνω αυτοτελής μελέτη παρατίθεται στο κεφάλαιο 4 και έχει ιδιαίτερο ενδιαφέρον καθώς αναφέρεται σε ένα πεδίο εφαρμογών που δεν απασχολεί μόνο την έρευνα αλλά και τη βιομηχανία, δίνοντας λύσεις σε απαιτήσεις του σήμερα για ασφαλείς επικοινωνίες με χρήση βαθιάς ενίσχυσης μάθησης.

Κεφάλαιο 2

Βαθιά Ενίσχυση της Μάθησης: Επισκόπηση

Σε αυτήν την ενότητα, παρουσιάζεται αρχικά η θεμελιώδη γνώση των διαδικασιών λήψης αποφάσεων του Markov (Markov Decision Processes – MDPs), της ενίσχυσης της μάθησης (Reinforcement Learning – RL) και των τεχνικών βαθιάς μάθησης (Deep Learning Techniques) που είναι σημαντικοί κλάδοι της θεωρίας της μηχανικής μάθησης (Machine Learning). Στη συνέχεια αναφέρεται η τεχνική Deep Reinforcement Learning (DRL) που μπορεί να αξιοποιήσει την ικανότητα της βαθιάς μάθησης (Deep Learning–DL) να βελτιώσει την αποτελεσματικότητα και την απόδοση από την άποψη του ποσοστού εκμάθησης για αλγόριθμους εκμάθησης ενίσχυσης. Στη συνέχεια, εξετάζονται προηγμένα μοντέλα DRL και οι επεκτάσεις τους.

2.1 Διαδικασία Απόφασης Markov (Markov Decision Process)

Το MDP (Puterman, 2014) είναι μια διαδικασία στοχαστικού ελέγχου διακριτού χρόνου. Το MDP παρέχει ένα μαθηματικό πλαίσιο για τη μοντελοποίηση των προβλημάτων λήψης αποφάσεων στα οποία τα αποτελέσματα είναι εν μέρει τυχαία και υπό τον έλεγχο ενός λήπτη αποφάσεων ή ενός πράκτορα. Τα MDP είναι χρήσιμα για τη μελέτη προβλημάτων βελτιστοποίησης τα οποία μπορούν να επιλυθούν με τεχνικές δυναμικού προγραμματισμού και ενισχυτικής μάθησης. Συνήθως, ένα MDP καθορίζεται από μια πλειάδα *(S, A, p, r)* όπου το *S* είναι ένα πεπερασμένο σύνολο καταστάσεων, το *A* είναι ένα πεπερασμένο σύνολο ενεργειών, το *p* είναι μια πιθανότητα μετάβασης από μια κατάσταση *s* σε μια κατάσταση *s′* μετά την εκτέλεση της ενέργειας *α*, και *r* είναι η άμεση ανταμοιβή που λαμβάνεται μετά την εκτέλεση της ενέργειας *α*. Ως *π* θεωρείται η «πολιτική» που είναι μια απεικόνιση από μια κατάσταση σε μια ενέργεια. Ο στόχος ενός MDP είναι να βρει μια βέλτιστη πολιτική για τη μεγιστοποίηση της λειτουργίας

ανταμοιβής. Η MDP μπορεί να έχει πεπερασμένο ή μη πεπερασμένο ορίζοντα. Για το MDP με πεπερασμένο ορίζοντα, μια βέλτιστη πολιτική π^*, που μεγιστοποιεί την αναμενόμενη συνολική ανταμοιβή ορίζεται από: $max_\pi E\left[\sum_{t=0}^{T} r_t(s_t, \pi(s_t))\right]$, όπου $a_t=\pi(s_t)$. Για MDP με μη πεπερασμένο ορίζοντα, ο σκοπός είναι να μεγιστοποιηθεί η αναμενόμενη συνολική προεξοφλημένη ανταμοιβή ή να μεγιστοποιηθεί η μέση ανταμοιβή. Το πρώτο ορίζεται από: $max_\pi E\left[\sum_{t=0}^{T} \gamma r_t(s_t, \pi(s_t))\right]$, ενώ το δεύτερο εκφράζεται από: $\liminf_{T\to\infty} max_\pi E\left[\sum_{t=0}^{T} r_t(s_t, \pi(s_t))\right]$, όπου $\gamma \in [0,1]$ είναι ο συντελεστής αναγωγής. Ο συντελεστής αναγωγής γ καθορίζει τη σημαντικότητα των μελλοντικών ανταμοιβών συγκρινόμενος με την τρέχουσα ανταμοιβή. Εάν γ=0, ο πράκτορας είναι «μυωπικός» (myopic), δηλαδή προσπαθεί να μεγιστοποιήσει την τρέχουσα ανταμοιβή, εάν όμως το γ προσεγγίζει το ένα, τότε ο πράκτορας αγωνίζεται για υψηλότερη μακροπρόθεσμη ανταμοιβή.

2.1.1 Partially Observable Markov Decision Process

Η κατάσταση του συστήματος στα MDPs, θεωρείται ότι είναι καταφανής από τον πράκτορα. Ωστόσο, σε πολλές περιπτώσεις ο πράκτορας μπορεί να διακρίνει μόνο ένα μέρος από την κατάσταση του συστήματος και έτσι η Partially Observable Markov Decision Processes (POMDPs) (Monahan, 1982) μπορεί να αποτελέσει το μοντέλο για προβλήματα λήψης απόφασης. Ένα τυπικό POMDP μοντέλο καθορίζεται από μια εξάδα *(S, A, p, r, Ω, O)*, όπου τα *S, A, p, r* ορίζονται όπως στα μοντέλα MDP, ενώ τα *Ω* και *O* ορίζονται ως το σύνολο των παρατηρήσεων και των πιθανοτήτων παρατήρησης, αντίστοιχα. Σε κάθε χρονική εποχή ο πράκτορας βρίσκεται σε κατάσταση *s* και επιλέγει μια ενέργεια *α*, βασισμένη στην άποψή του σχετικά με την τρέχουσα κατάσταση *s*, δηλαδή την *b(s)* και παρατηρεί την άμεση ανταμοιβή *r* και την τρέχουσα παρατήρηση *o*. Βασισμένος στην παρατήρηση *o* και στην άποψή του για την τρέχουσα κατάσταση *b(s)*, ο πράκτορας επαναπροσδιορίζει την άποψή του για τη νέα κατάσταση *s΄*, δηλαδή *b(s΄)*, ως ακολούθως:

$$b(s') = \frac{O(o|s,a,s') \sum_{s\in S} p(s'|s,a) b(s)}{\sum_{s'\in S} O(o|s,a,s') \sum_{s\in S} p(s'|s,a) b(s)} \qquad (1)$$

όπου $O(o|s,a,s')$ είναι η πιθανότητα του πράκτορα να πραγματοποιήσει παρατηρήσεις *o*, αφού λάβει μέρος η ενέργεια *α* στην κατάσταση *s* και ο πράκτορας μεταβεί στην κατάσταση *s΄*. Το $p(s'|s,a)$ ορίζεται όπως στο μοντέλο MDP, δηλαδή η πιθανότητα μετάβασης από την κατάσταση *s* στην κατάσταση *s΄*, αφού λάβει μέρος η ενέργεια *α* στην

κατάσταση *s*. Τελικά, ο πράκτορας λαμβάνει μια άμεση ανταμοιβή *r* που είναι ίση με το *r(s,a)* στο MDP. Παρόμοια με το μοντέλο MDP, ο πράκτορας στο POMDP, προσπαθεί να βρει τη βέλτιστη πολιτική π^*, με σκοπό να μεγιστοποιήσει την αναμενόμενη μακροπρόθεσμη προεξοφλημένη ανταμοιβή $\sum_{t=0}^{\infty} \gamma r_t\big(s_t, \pi^*(s_t)\big)$.

2.1.2 Markov Games

Στη θεωρία παιγνίων , ένα παίγνιο Markov ή ένα στοχαστικό παίγνιο (Shapley, 1953), είναι ένα δυναμικό παίγνιο με πιθανολογικές μεταβάσεις που πραγματοποιούνται από πολλαπλούς παίκτες, δηλαδή τους πράκτορες. Ένα τυπικό μοντέλο παιγνίου Markov ορίζεται από μια πλειάδα $\left(I, S, \{A^i\}_{i\in I}, p, \{r^i\}_{i\in I}\right)$, όπου

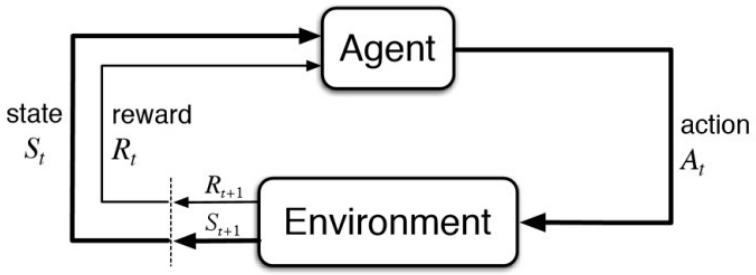
- $I \triangleq \{1, \dots, i, \dots I\}$ είναι ένα σύνολο πρακτόρων,
- $S \triangleq \{S^1, \dots, S^i, \dots S^I\}$, είναι ο συνολικός χώρος καταστάσεων όλων των πρακτόρων με S^i να είναι ο χώρος κατάστασης του πράκτορα i.
- $\{A^i\}_{i\in I}$ είναι το σύνολο των χώρων ενεργειών των πρακτόρων με A^i να εκφράζει τον χώρο ενεργειών του πράκτορα i.
- $p \triangleq S \times A^1 \times \dots \times A^I \to [0,1]$ είναι η συνάρτηση πιθανότητας μετάβασης του συστήματος
- $\{r^i\}_{i\in I}$ είναι συναρτήσεις απολαβής (payoff functions) των πρακτόρων με $r^i \triangleq S \times A^1 \times \dots \times A^I \to \mathbb{R}$, η απολαβή του πράκτορα i αποκτάται όταν εκτελούνται όλες οι ενέργειες των πρακτόρων.

Σε ένα παίγνιο Markov, οι πράκτορες ξεκινούν σε κάποια αρχική κατάσταση $s_0 \in S$. Αφού παρατηρήσουν την τρέχουσα κατάσταση, όλοι οι πράκτορες ταυτόχρονα επιλέγουν τις ενέργειές τους $\alpha = \{\alpha^1, \dots \alpha^I\}$, και θα λάβουν τις αντίστοιχες ανταμοιβές μαζί με τις καινούριες παρατηρήσεις τους. Την ίδια στιγμή, το σύστημα θα μεταβεί σε μια νέα κατάσταση $s' \in S$, με πιθανότητα $p(s'|s,a)$. Η διαδικασία επαναλαμβάνεται στη νέα κατάσταση και συνεχίζει για πεπερασμένο ή μη πεπερασμένο αριθμό σταδίων. Σ' αυτό το παίγνιο, όλοι οι πράκτορες προσπαθούν να βρουν τις βέλτιστες πολιτικές για να μεγιστοποιήσουν τις αναμενόμενες από τους ιδίους μέσες μακροπρόθεσμες ανταμοιβές, δηλαδή $\sum_{t=0}^{\infty} \gamma_i r_t^i\big(s_t, {\pi_i}^*(s_t)\big), \forall i$. Το σύνολο των βέλτιστων πολιτικών αυτού του παιγνίου, δηλαδή $\{\pi_1^*, \dots, \pi_I^*\}$ είναι γνωστή ως η ισορροπία του παιγνίου. Εάν ο αριθμός των παικτών είναι πεπερασμένος και το σύνολο των καταστάσεων και των ενεργειών

είναι και αυτά πεπερασμένα, τότε το παίγνιο Markov έχει πάντοτε μια ισορροπία του Nash κάτω από πεπερασμένο αριθμό σταδίων. Το ίδιο ισχύει για παίγνια Markov με μη πεπερασμένα στάδια, αλλά η συνολική απολαβή των πρακτόρων είναι το discounted άθροισμα. (Hu & Wellman, 2003).

2.2 Ενισχυτική Μάθηση (Reinforcement Learning)

Η ενίσχυση της μάθησης, ένας σημαντικός κλάδος της μηχανικής μάθησης, είναι ένα αποτελεσματικό εργαλείο, που χρησιμοποιείται ευρέως στη βιβλιογραφία και σχετίζεται με τα MDPs (Sutton & Barto, 1998). Σε μια διαδικασία ενίσχυσης μάθησης, ένας πράκτορας μπορεί να μάθει τη βέλτιστη πολιτική του μέσω αλληλεπίδρασης με το περιβάλλον του. Ειδικότερα, ο πράκτορας παρατηρεί την τρέχουσα κατάστασή του, και στη συνέχεια λαμβάνει μια ενέργεια, και λαμβάνει την άμεση ανταμοιβή του μαζί με τη νέα του κατάσταση όπως φαίνεται στο Σχ. 1. Οι παρατηρούμενες πληροφορίες, δηλαδή η άμεση ανταμοιβή και η νέα κατάσταση, χρησιμοποιούνται για την προσαρμογή της πολιτικής του πράκτορα και αυτή η διαδικασία θα επαναληφθεί έως ότου η πολιτική του γίνει η βέλτιστη δυνατή. Στην ενίσχυση της μάθησης, η Q-learning είναι η πιο αποτελεσματική μέθοδος και χρησιμοποιείται ευρέως στη βιβλιογραφία. Στη συνέχεια, παρουσιάζεται ο αλγόριθμο Q-learning και οι επεκτάσεις του για προηγμένα μοντέλα MDP.



Σχήμα 1. Ενισχυτική μάθηση. Πηγή: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

2.2.1 Q-Learning Algorithm

Στην MDP ο πράκτορας αναζητεί τη βέλτιστη πολιτική $\pi^*: S \to A$ για να μεγιστοποιήσει την αναμενόμενη μακροπρόθεσμη συνάρτηση ανταμοιβής για το σύστημα. Κατά συνέπεια, αρχικά ορίζεται η συνάρτηση τιμής $V^\pi: S \to \mathbb{R}$ που αντιπροσωπεύει την

αναμενόμενη τιμή που αποκτάται από την ακολουθούμενη πολιτική π, για κάθε κατάσταση $s \in S$. Η συνάρτηση τιμής V για πολιτική π, ποσοτικοποιεί την καταλληλότητα της πολιτικής σε μη πεπερασμένο ορίζοντα και προεξοφλημένη MDP, που μπορεί να εκφραστεί ως ακολούθως:

$V^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty} \gamma r_t(s_t, a_t) | s_0 = s] = E_{\pi}[r_t(s_t, a_t) + \gamma V^{\pi}(s_{t+1}) | s_0 = s]$ (2)

Για την εύρεση της βέλτιστης πολιτικής π^*, μπορεί να βρεθεί η βέλτιστη ενέργεια για κάθε κατάσταση, μέσω της βέλτιστης συνάρτησης τιμής που εκφράζεται: $V^*(s) = max_{a_t}\{E_{\pi}[r_t(s_t, a_t) + \gamma V^{\pi}(s_{t+1})]\}$.

Εάν θεωρηθεί η $Q^*(s, a) \triangleq r_t(s_t, a_t) + \gamma E_{\pi}[V^{\pi}(s_{t+1})]$, ως η βέλτιστη συνάρτηση Q για όλα τα ζεύγη κατάστασης-ενέργειας, τότε η βέλτιστη συνάρτηση τιμής μπορεί να γραφεί: $V^*(s) = max_a\{Q^*(s, a)\}$. Μ' αυτό τον τρόπο προσεγγίζεται πιο εύκολα η προσπάθεια να βρεθούν βέλτιστες τιμές για τη συνάρτηση Q, δηλαδή την $Q^*(s, a)$, για όλα τα ζεύγη κατάστασης-ενέργειας, κι αυτό μπορεί να γίνει μέσω επαναληπτικών διαδικασιών. Συγκεκριμένα, η συνάρτηση Q ενημερώνεται σύμφωνα με τον ακόλουθο κανόνα:

$Q_{t+1}(s, a) = Q_t(s, a) + a_t[r_t(s, a) + \gamma max_{a'} Q_t(s, a') - Q_t(s, a)]$. (3)

Η κεντρική ιδέα πίσω από την παραπάνω ενημέρωση είναι να βρεθεί η προσωρινή διαφορά (Temporal Difference – TD) μεταξύ της προβλεπόμενης τιμής Q, δηλαδή $r_t(s, a) + \gamma max_{a'} Q_t(s, a')$, και της τρέχουσας τιμής $Q_t(s, a)$. Στον παραπάνω κανόνα ενημέρωσης της συνάρτησης Q (3), ο ρυθμός μάθησης a_t χρησιμοποιείται για να καθορίσει την επίδραση της νέας πληροφορίας στην υφιστάμενη τιμή Q. Ο ρυθμός μάθησης επιλέγεται να είναι μια σταθερά ή μπορεί να προσαρμόζεται δυναμικά κατά τη διάρκεια της διαδικασίας μάθησης. Ωστόσο, θα πρέπει να ικανοποιεί την παρακάτω Θεώρηση 1 για να εγγυάται τη σύγκλιση με τον αλγόριθμο Q-learning.

Θεώρηση 1: Το μέγεθος βήματος a_t είναι προσδιοριστικό, μη μηδενικό και ικανοποιεί τις παρακάτω συνθήκες: $a_t \in [0,1]$, $\sum_{t=0}^{\infty} a_t = \infty$ και $\sum_{t=0}^{\infty} (a_t)^2 < \infty$.

Η προσαρμογή μεγέθους βήματος $a_t = \frac{1}{t}$ είναι ένα από τα πιο κοινά παραδείγματα που χρησιμοποιούνται στην ενίσχυση μάθησης (Dabney, 2014).

Όταν είτε όλες οι τιμές Q συγκλίνουν, είτε επιτευχθεί ένα συγκεκριμένος αριθμός επαναλήψεων, ο αλγόριθμος θα τερματιστεί. Ο αλγόριθμος, τότε, αποφέρει τη βέλτιστη πολιτική, υποδεικνύοντας τη λήψη μιας ενέργειας σε κάθε κατάσταση, έτσι ώστε η

$Q^*(s,a)$ να μεγιστοποιείται για όλες τις καταστάσεις του χώρου καταστάσεων, δηλαδή $\pi^*(s) = argmax_a Q^*(s,a)$.
Σύμφωνα με τη Θεώρηση 1, αποδεικνύεται (Watkins & Dayan, 1992) ότι, ο αλγόριθμος Q-learning συγκλίνει στη βέλτιστη ενέργεια-τιμή με πιθανότητα ένα.

Αξίζει να σημειωθεί ότι, αντίθετα με τη συνάρτηση τιμής V^π, η συνάρτηση Q είναι ένα παράδειγμα αλγορίθμου μάθησης χωρίς τη χρήση μοντέλου, στον οποίο ο πράκτορας δεν απαιτείται να γνωρίζει εκ των προτέρων τις παραμέτρους του μοντέλου του συστήματος, δηλαδή τα μοντέλα μετάβασης κατάστασης και ανταμοιβής, για να υπολογίσει τα ζεύγη τιμών κατάστασης-ενέργειας. Συγκεκριμένα η βασική ιδέα πίσω από τη συνάρτηση Q είναι να προσεγγίσει των τιμών των ζευγών κατάστασης-ενέργειας μέσω δειγμάτων που αποκτήθηκαν με την αλληλεπίδραση με το περιβάλλον. Επιπλέον, ενώ η συνάρτηση τιμής λαμβάνει τις προσδοκώμενες τιμές όλων των ενεργειών, σύμφωνα με την πολιτική π, η συνάρτηση Q επικεντρώνεται μόνο σε συγκεκριμένη ενέργεια μιας συγκεκριμένης κατάστασης. Ως αποτέλεσμα, οι αλγόριθμοι μάθησης που χρησιμοποιούν συνάρτηση Q είναι λιγότερο περίπλοκοι από αυτούς που χρησιμοποιούν συνάρτηση τιμής. Ωστόσο, από πλευράς δειγματοληψίας, η διάσταση της συνάρτησης Q είναι μεγαλύτερη από αυτής της συνάρτησης τιμής και έτσι στη συνάρτηση Q μπορεί να είναι πιο δύσκολο να ληφθούν αρκετά δείγματα δηλαδή, τα ζεύγη κατάστασης ενέργειας να μαθαίνουν. Γι' αυτό το λόγο εάν το μοντέλο του συστήματος είναι γνωστό εξ' αρχής, είναι προτιμητέα η συνάρτηση τιμής.

2.2.2 SARSA (Ένας Δικτυακός Αλγόριθμος Q-Learning)

Παρόλο που ο αλγόριθμος Q-Learning μπορεί να βρει για τον πράκτορα τη βέλτιστη πολιτική, χωρίς να απαιτεί γνώση του περιβάλλοντος, αυτός ο αλγόριθμος λειτουργεί με τρόπο που δεν απαιτεί διασύνδεση στο δίκτυο. Συγκεκριμένα ο αλγόριθμος Q-Learning μπορεί να αποκτήσει τη βέλτιστη πολιτική μόνο όταν όλες οι τιμές Q συγκλίνουν. Αντιθέτως, ο αλγόριθμος SARSA επιτρέπει στον πράκτορα να προσεγγίσει τη βέλτιστη πολιτική με σύνδεση στο δίκτυο, αποτελώντας έναν εναλλακτικό δικτυακό αλγόριθμο μάθησης.

Διαφορετικός από τον αλγόριθμο Q-Learning, ο SARSA είναι ένας δικτυακός αλγόριθμος, που δίνει τη δυνατότητα στον πράκτορα να επιλέξει τις βέλτιστες ενέργειες σε κάθε

χρονική στιγμή, σε πραγματικό χρόνο, δίχως να περιμένει τη σύγκλιση του αλγορίθμου. Στο αλγόριθμο Q-Learning, η πολιτική διαμορφώνεται σύμφωνα με τη μέγιστη ανταμοιβή των διαθέσιμων ενεργειών, ανεξάρτητα ποια πολιτική εφαρμόζεται, δηλαδή μια μέθοδος χωρίς τη χρήση δικτύου (offline). Αντιθέτως, ο αλγόριθμος SARSA αλληλεπιδρά με το περιβάλλον και διαμορφώνει την πολιτική απ' ευθείας από τις ενέργειες που λαμβάνουν μέρος, δηλαδή μια δικτυακή (online) μέθοδος. Επισημαίνεται ότι, ο αλγόριθμος SARSA διαμορφώνει τιμές Q από την πεντάδα $Q(s, a, r, s', a')$.

2.2.3 Q-Learning for Markov Games

Για να εφαρμοστεί ο αλγόριθμος Q-learning στο πλαίσιο του παιγνίου Markov, πρώτα ορίζεται η συνάρτηση Q για τον πράκτορα i ως: $Q_i(s, a^i, a^{-i})$, όπου $a^{-i} \triangleq \{a^1, \dots, a^{i-1}, a^{i+1}, \dots, a^I\}$ που δηλώνει το σύνολο των ενεργειών όλων των πρακτόρων εκτός του πράκτορα i. Έπειτα η συνάρτηση Q του Nash του πράκτορα i ορίζεται από:

$$Q_i^*(s, a^i, a^{-i}) = r^i(s, a^i, a^{-i}) + \beta \sum_{s' \in S} p(s'|s, a^i, a^{-i}) \times V^i(s', \pi_1^*, \dots, \pi_I^*), \qquad (4)$$

όπου $(\pi_1^*, \dots, \pi_I^*)$ η ισορροπία του Nash μικτής στρατηγικής, r^i είναι η άμεση ανταμοιβή στην κατάσταση s κάτω από τη μικτή ενέργεια (a^i, a^{-i}) και $V^i(s', \pi_1^*, \dots, \pi_I^*)$ είναι η συνολική προεξοφλημένη ανταμοιβή για ένα μη πεπερασμένο χρονικό ορίζοντα ξεκινώντας από την κατάσταση s', δεδομένου ότι όλοι οι πράκτορες ακολουθούν τις στρατηγικές της ισορροπίας.

Στη συνέχεια παρουσιάζεται μια πρόταση (Hu & Wellman, 2003) για ένα αλγόριθμο Q-learning με πολλούς πράκτορες για γενικού αθροίσματος παίγνια Markov, που επιτρέπουν στους πράκτορες να πραγματοποιούν ενημερώσεις βασιζόμενοι στην υπόθεση της συμπεριφοράς της ισορροπίας του Nash στις τρέχουσες τιμές Q. Συγκεκριμένα, ο πράκτορας *i* μαθαίνει τις τιμές του Q κάνοντας μια αυθαίρετη εικασία στην αρχή του παιγνίου. Σε κάθε χρονικό βήμα *t*, ο πράκτορας *i* παρατηρεί την τρέχουσα κατάσταση και κάνει μια ενέργεια a^i. Στη συνέχεια παρατηρεί την άμεση ανταμοιβή του r^i, τις ενέργειες που έλαβαν οι υπόλοιποι a^{-I} , τις άμεσες ανταμοιβές τους και την νέα κατάσταση του συστήματος s'. Έπειτα, ο πράκτορας i υπολογίζει μια ισορροπίας Nash $(\pi_1(s'), \dots, \pi_I(s'))$ για κατάσταση παιγνίου $(Q_1^t(s'), \dots, Q_I^t(s'))$ και διαμορφώνει τις τιμές Q σύμφωνα με :

$$Q_i^{t+1}(s, a^i, a^{-i}) = (1 - a_t) Q_i^t(s, a^i, a^{-i}) + a_t[r_t^i + \gamma N_t^i(s')], \qquad (5)$$

όπου $a_t \in (0,1)$ είναι ο ρυθμός μάθησης και $N_t^i(s') \triangleq Q_i^t(s') \times (\pi_1(s'), \dots, \pi_I(s'))$.

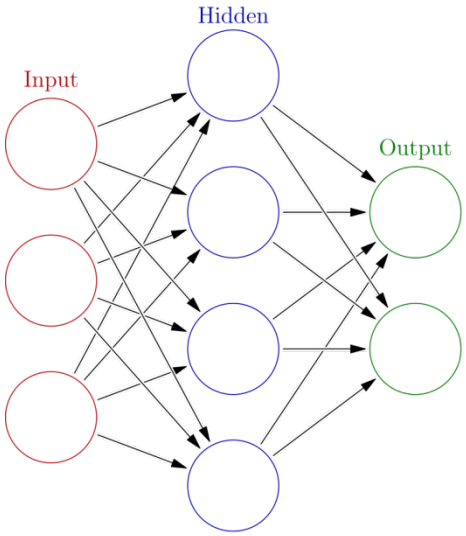
Για να μπορεί να υπολογιστεί η ισορροπίας Nash, ο πράκτορας i πρέπει να γνωρίζει $\left(Q_1^t(s'), \dots, Q_I^t(s')\right)$. Ωστόσο, η πληροφορία για τις τιμές Q των υπόλοιπων πρακτόρων δεν δίνεται, και έτσι ο πράκτορας i θα πρέπει να την αναζητήσει. Γι αυτό το λόγο, ο πράκτορας i πραγματοποιεί στην αρχή του παιγνίου μια εκτίμηση για τις τιμές Q των υπολοίπων πρακτόρων δηλαδή $Q_0^j(s, a^i, a^{-i}) = 0, \forall j, s$. Όσο το παίγνιο προχωράει, ο πράκτορας i παρατηρεί τις άμεσες ανταμοιβές και τις προηγούμενες ενέργειες των υπόλοιπων πρακτόρων. Αυτές οι πληροφορίες μπορεί, στη συνέχεια, να χρησιμοποιούνται από τον πράκτορα i να ενημερώνει τις υποθέσεις του για τις συναρτήσεις Q των υπολοίπων πρακτόρων. Ο πράκτορας i επαναδιαμορφώνει τις απόψεις του για τη συνάρτηση Q του πράκτορα j, σύμφωνα με τον ανωτέρω κανόνα (5). Αποδεικνύεται ότι, υπό υψηλά περιοριστικές υποθέσεις σχετικά με τη μορφή των παιγνίων κατάστασης κατά τη διάρκεια της μάθησης, ο προτεινόμενος αλγόριθμος πολλών πρακτόρων εγγυάται τη σύγκλιση.

2.3 Βαθιά Μάθηση (Deep Learning)

Η βαθιά μάθηση (Goodfellow, Bengio, & Courville, 2016) αποτελείται από ένα σύνολο αλγορίθμων και τεχνικών που προσπαθούν να βρουν σημαντικά γνωρίσματα των δεδομένων και να μοντελοποιήσουν τις υψηλού επιπέδου αφηρημένες έννοιές τους. Ο κύριος στόχος της βαθιάς μάθησης είναι να αποφευχθεί η μη αυτόματη περιγραφή μιας δομής δεδομένων με αυτόματη εκμάθηση από τα δεδομένα. Το όνομά της αναφέρεται στο γεγονός ότι συνήθως οποιοδήποτε νευρωνικό δίκτυο με δύο ή περισσότερα κρυμμένα στρώματα ονομάζεται βαθύ νευρωνικό δίκτυο (Deep Neural Network – DNN). Τα περισσότερα μοντέλα βαθιάς μάθησης βασίζονται σε ένα τεχνητό νευρωνικό δίκτυο (Artificial Neural Network - ANN), παρόλο που μπορούν επίσης να περιλαμβάνουν προτασιακούς τύπους ή λανθάνουσες μεταβλητές οργανωμένες σε επίπεδα στρώματος σε βαθιάς παραγωγικά μοντέλα, όπως οι κόμβοι σε Deep Belief Networks και Deep Boltzmann Machines.

Το ANN είναι ένα υπολογιστικό μη γραμμικό μοντέλο που βασίζεται στη νευρωνική δομή του εγκεφάλου που είναι σε θέση να μάθει να εκτελεί εργασίες όπως ταξινόμηση, πρόβλεψη, λήψη αποφάσεων και οπτικοποίηση. Ένα ANN αποτελείται από τεχνητούς νευρώνες και είναι οργανωμένο σε τρία διασυνδεδεμένα επίπεδα: είσοδος, κρυφό και

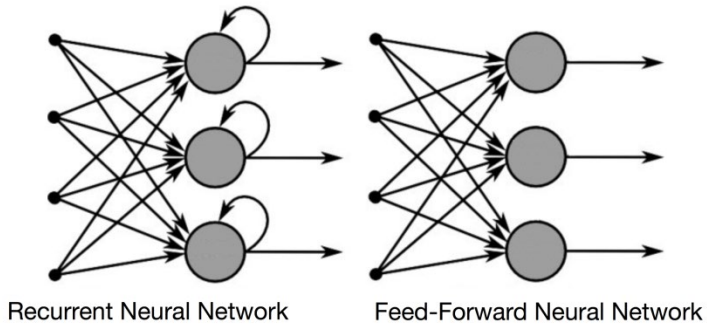
έξοδος όπως απεικονίζεται στο Σχ. 2. Το επίπεδο εισόδου περιέχει νευρώνες εισόδου που στέλνουν πληροφορίες στο κρυφό επίπεδο. Το κρυφό επίπεδο στέλνει δεδομένα στο επίπεδο εξόδου. Κάθε νευρώνας έχει σταθμισμένες εισόδους-συνάψεις, μια λειτουργία ενεργοποίησης και μία έξοδο. Οι συνάψεις είναι οι ρυθμιζόμενες παράμετροι που μετατρέπουν ένα νευρωνικό δίκτυο σε ένα παραμετροποιημένο σύστημα. Η λειτουργία ενεργοποίησης ενός κόμβου καθορίζει τις εξόδους αυτού του κόμβου δεδομένων των εισόδων. Συγκεκριμένα, η λειτουργία ενεργοποίησης θα αντιστοιχίσει τις τιμές εισόδου σε εύρη στόχων, ανάλογα με την επιλεγμένη λειτουργία ενεργοποίησης. Για παράδειγμα, η συνάρτηση λογιστικής ενεργοποίησης θα αντιστοιχίσει όλες τις εισόδους στο πεδίο πραγματικών αριθμών στην περιοχή από 0 έως 1.



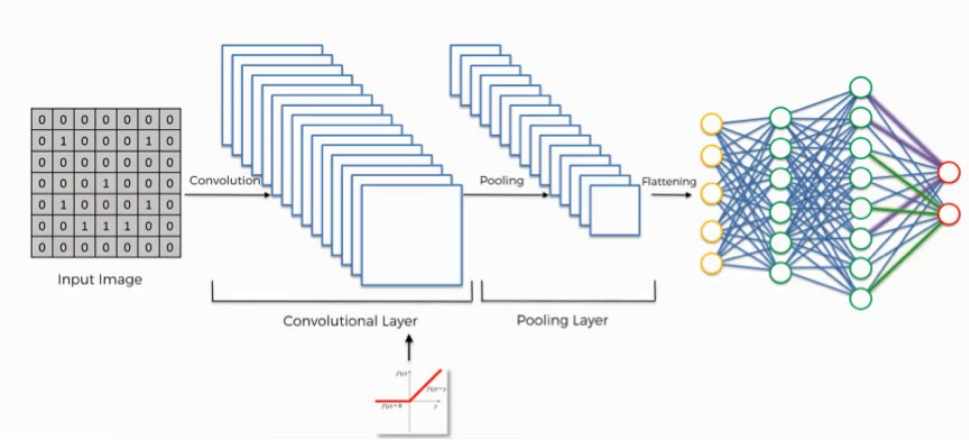
Σχήμα 2. Τεχνητά νευρωνικά δίκτυα. Πηγή:
https://en.wikipedia.org/wiki/Artificial_neural_network

Κατά τη φάση της μάθησης, τα ANN χρησιμοποιούν την ανάστροφη διάδοση ως έναν αποτελεσματικό αλγόριθμο μάθησης για να υπολογίσουν γρήγορα μια βαθμιδωτή κλίση καθόδου σε σχέση με τα βάρη. Η ανάστροφη διάδοση είναι μια ειδική περίπτωση αυτόματης διαφοροποίησης. Στο πλαίσιο της μάθησης, η ανάστροφη διάδοση χρησιμοποιείται συνήθως από τον αλγόριθμο βελτιστοποίησης καθόδου κλίσης για να ρυθμίσει τα βάρη των νευρώνων υπολογίζοντας την κλίση της συνάρτησης απώλειας. Αυτή η τεχνική ονομάζεται επίσης οπίσθια διάδοση σφαλμάτων, επειδή το σφάλμα υπολογίζεται στην έξοδο και διανέμεται προς τα πίσω μέσω των επιπέδων του δικτύου.

Ένα DNN ορίζεται ως ANN με πολλαπλά κρυφά επίπεδα. Υπάρχουν δύο τυπικά μοντέλα DNN, το νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (Feed-Forward Neural Netwok - FNN) και το επαναλαμβανόμενο νευρωνικό δίκτυο (Recurrent Neural Network - RNN). Στο FNN, οι πληροφορίες μετακινούνται σε μία μόνο κατεύθυνση, δηλαδή, από τους κόμβους εισόδου, μέσω των κρυφών κόμβων στους κόμβους εξόδου, ενώ δεν υπάρχουν κύκλοι ή βρόχοι στο δίκτυο, όπως φαίνεται στο Σχ. 3. Στα FNNs, το συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network -CNN) είναι το πιο γνωστό μοντέλο με ένα ευρύ φάσμα εφαρμογών, ιδίως στην αναγνώριση εικόνας και ομιλίας. Το CNN περιέχει ένα ή περισσότερα συνελικτικά στρώματα, συνδυασμένα ή πλήρως συνδεδεμένα και χρησιμοποιεί μια παραλλαγή πολυεπίπεδων αντιληπτών. Σε γενικές γραμμές, τα CNN έχουν δύο κύριους προσανατολισμούς, την εξαγωγή χαρακτηριστικών και την ταξινόμηση, όπως φαίνεται στο Σχ. 4. Όσον αφορά την εξαγωγή χαρακτηριστικών, αυτή λαμβάνει μέρος στα κρυφά στρώματα με σκοπό την εκτέλεση μιας σειράς συνελίξεων και λειτουργιών συνδυασμών κατά τη διάρκεια των οποίων ανιχνεύονται τα χαρακτηριστικά. Όσον αφορά δε την ταξινόμηση, πραγματοποιείται στα πλήρως συνδεδεμένα στρώματα, με την ανάθεση μιας πιθανότητας για το αντικείμενο, π.χ. στην εικόνα, σε αυτό που απαιτείται να προβλεφτεί.



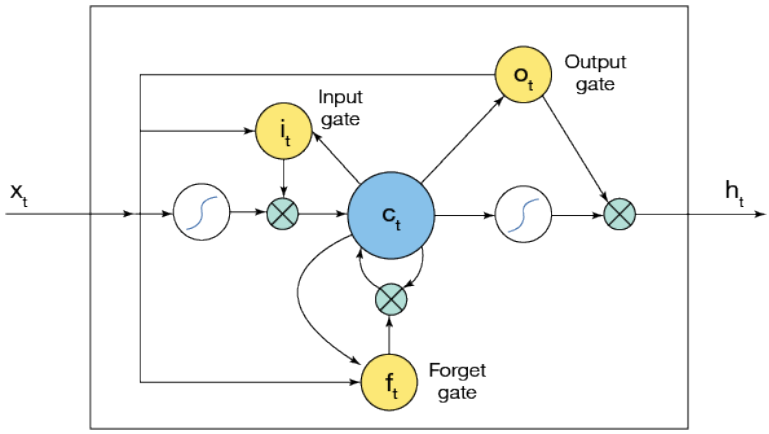
Σχήμα 3. Feed Forward Neural Network και Recurrent Neural Network
Πηγή: https://nerdthecoder.wordpress.com/2019/02/03/recurrent-neural-net/



Σχήμα 4. Convolutional Neural Network
Πηγή: https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-summary/

Σε αντίθεση με τα FNNs, το RNN είναι μια παραλλαγή ενός αναδρομικού τεχνητού νευρωνικού δικτύου στο οποίο οι συνδέσεις μεταξύ των νευρώνων κάνουν κατευθυνόμενους κύκλους. Αυτό σημαίνει ότι μια έξοδος δεν εξαρτάται μόνο από τις άμεσες εισόδους της, αλλά και από την πρότερη κατάσταση του. Τα RNN έχουν σχεδιαστεί για να χρησιμοποιούν διαδοχικά δεδομένα, όταν το τρέχον βήμα έχει κάποια σχέση με τα προηγούμενα βήματα. Αυτό καθιστά τα RNNs ιδανικά για εφαρμογές με στοιχεία χρόνου, π.χ. δεδομένα χρονοσειρών και επεξεργασία φυσικής γλώσσας. Ωστόσο, όλα τα RNNs έχουν βρόχους ανατροφοδότησης στο επίπεδο επανάληψης. Αυτό επιτρέπει στα RNNs, με την πάροδο του χρόνου, να διατηρούν πληροφορίες στη μνήμη. Ωστόσο, είναι δύσκολο να εκπαιδευτούν τυπικά RNN για την επίλυση προβλημάτων που απαιτούν μάθηση μακροχρόνιων χρονικών εξαρτήσεων. Ο λόγος είναι ότι η κλίση της συνάρτησης απώλειας μειώνεται εκθετικά με το χρόνο, γεγονός που ονομάζεται πρόβλημα εξαφάνισης διανυσμάτων κλίσης (vanishing gradient problem). Επομένως, η μακρά και βραχεία μνήμη (Long Short-Term Memory - LSTM) χρησιμοποιείται συχνά σε RNN για την αντιμετώπιση αυτού του ζητήματος. Τα LSTMs έχουν σχεδιαστεί για να μοντελοποιούν χρονικές ακολουθίες και οι μακροχρόνιες εξαρτήσεις τους είναι πιο ακριβείς από τα συμβατικά RNNs. Συγκεκριμένα, τα LSTMs παρέχουν μια λύση ενσωματώνοντας μονάδες μνήμης που επιτρέπουν στο δίκτυο να μάθει πότε να ξεχάσει τις προηγούμενες κρυφές καταστάσεις και πότε να ενημερώσει τις κρυφές καταστάσεις με νέες πληροφορίες. Συνήθως, οι μονάδες LSTM για τον έλεγχο της ροής πληροφοριών της λογιστικής συνάρτησης, υλοποιούνται σε πακέτα (μπλοκ) που έχουν τρεις ή τέσσερις πύλες, π.χ. πύλη εισόδου, πύλη λήθης (forget), πύλη εξόδου και πύλη διαμόρφωσης εισόδου, όπως

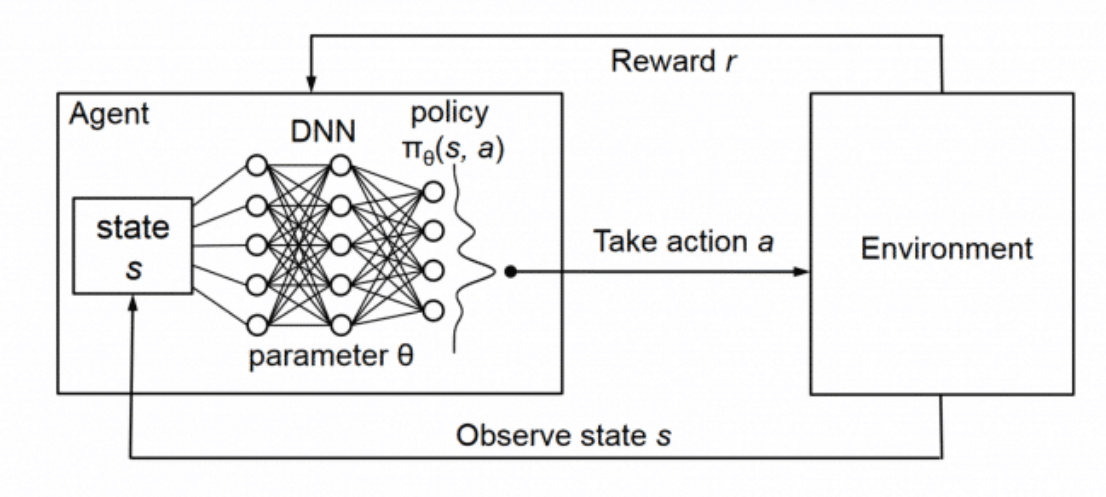
φαίνεται στο Σχ. 5 (Donahue, et al., 2015). Σε αντίθεση με το RNN, το κελί μνήμης LSTM αποτελείται από τρία στοιχεία, το προηγούμενο κελί μνήμης c_t, την τρέχουσα είσοδο x_t και την προηγούμενη κρυφή κατάσταση h_{t-1}. Εδώ, η πύλη εισόδου και η πύλη λήθης θα χρησιμοποιηθούν για να ξεχάσουν επιλεκτικά την προηγούμενη μνήμη ή να εξετάσουν την τρέχουσα είσοδο. Ομοίως, η πύλη εξόδου μαθαίνει πόση ποσότητα κελιού μνήμης θα μεταφερθεί στην κρυφή κατάσταση. Αυτά τα πρόσθετα μπλοκ επιτρέπουν στο LSTM να μάθει εξαιρετικά περίπλοκες και μακροχρόνιες χρονικές δυναμικές που το RNN δεν μπορεί να υλοποιήσει.



Σχήμα 5. Long Short Term Memory Networks
Πηγή: https://developer.ibm.com/technologies/iot/tutorials/iot-deep-learning-anomaly-detection-1/

2.4 Βαθιά Μάθηση Q (Deep Q-Learning)

Ο αλγόριθμος Q-learning μπορεί να αποκτήσει αποτελεσματικά μια βέλτιστη πολιτική όταν ο χώρος κατάστασης και ο χώρος δράσης είναι μικρός. Ωστόσο, στην πράξη, με πολύπλοκα μοντέλα συστήματος, αυτοί οι χώροι είναι συνήθως μεγάλοι. Ως αποτέλεσμα, ο αλγόριθμος Q-learning μπορεί να μην είναι σε θέση να βρει τη βέλτιστη πολιτική και γι αυτό το λόγο εισάγεται ο αλγόριθμος Deep Q-Learning (DQL). Είναι ευκολονόητο ότι, ο αλγόριθμος DQL εφαρμόζει ένα Deep Q-Network (DQN), δηλαδή ένα DNN, αντί για τον πίνακα *Q* για να αποκομίσει μια κατά προσέγγιση τιμή $Q^*(s, a)$ όπως φαίνεται στο Σχ. 6.



Σχήμα 6. Βαθιά Q-learning. Πηγή: https://www.novatec-gmbh.de/en/blog/deep-q-networks/

Όταν χρησιμοποιείται ένας μη γραμμικός προσεγγιστής συνάρτησης, η μέση ανταμοιβή που λαμβάνεται από αλγόριθμους ενίσχυσης μάθησης μπορεί να μην είναι σταθερή ή ακόμη και να αποκλίνει (Mnih, et al., Human-level control through deep reinforcement learning, 2015). Αυτό οφείλεται στο γεγονός ότι μια μικρή αλλαγή των τιμών Q μπορεί να επηρεάσει σημαντικά την πολιτική. Έτσι, η κατανομή δεδομένων και οι συσχετίσεις μεταξύ των τιμών Q και των τιμών στόχου $R + \gamma max_{a'} Q(s', a')$ ποικίλλουν. Για να αντιμετωπιστεί αυτό το ζήτημα, μπορούν να χρησιμοποιηθούν δύο μηχανισμοί, η επανάληψη εμπειρίας και ο στόχος Q-network.

- Μηχανισμός επανάληψης εμπειρίας (Experience replay mechanism): Ο αλγόριθμος αρχικοποιεί πρώτα μια μνήμη επανάληψης D, δηλαδή τη δεξαμενή μνήμης, με μεταβάσεις (s_t, a_t, r_t, s_{t+1}), δηλαδή εμπειρίες, που δημιουργούνται τυχαία, π.χ. με χρήση πολιτικής ε-greedy. Στη συνέχεια, ο αλγόριθμος επιλέγει τυχαία δείγματα, δηλαδή, μικρά σύνολα (minibatches), μεταβάσεων από το D για να εκπαιδεύσει το DNN. Οι τιμές Q που λαμβάνονται από το εκπαιδευμένο DNN θα χρησιμοποιηθούν για την απόκτηση νέων εμπειριών, δηλαδή μεταβάσεων, και αυτές οι εμπειρίες θα αποθηκευτούν στη συνέχεια στη δεξαμενή μνήμης D. Αυτός ο μηχανισμός επιτρέπει στο DNN να εκπαιδεύεται πιο αποτελεσματικά χρησιμοποιώντας παλιές και νέες εμπειρίες. Επιπλέον, με τη χρήση της επανάληψης εμπειρίας, οι μεταβάσεις είναι πιο ανεξάρτητες και πανομοιότυπα κατανεμημένες, και έτσι οι συσχετίσεις μεταξύ των παρατηρήσεων μπορούν να αφαιρεθούν.

Σταθερό δίκτυο Q-στόχου (Fixed target Q-Network): Κατά τη διαδικασία εκπαίδευσης, η τιμή Q αλλάζει. Έτσι, εάν χρησιμοποιείται ένα σύνολο τιμών που αλλάζει συνεχώς για την ενημέρωση του δικτύου Q, οι εκτιμήσεις τιμών μπορεί να είναι εκτός ελέγχου. Αυτό οδηγεί στην αποσταθεροποίηση του αλγορίθμου. Για την αντιμετώπιση αυτού του ζητήματος, το δίκτυο Q-στόχου χρησιμοποιείται για τη συχνή αλλά αργή ενημέρωση των τιμών των κύριων δικτύων Q. Με αυτόν τον τρόπο, οι συσχετίσεις μεταξύ του στόχου και των εκτιμώμενων τιμών Q μειώνονται σημαντικά, σταθεροποιώντας έτσι τον αλγόριθμο.

Τεχνική	**Αντιμετωπιζόμενο πρόβλημα**
Μη-γραμμικός προγραμματισμός	Χρησιμοποιείται για την αντιμετώπιση προβλημάτων στατικής βελτιστοποίησης, δηλαδή βελτιστοποιεί την αντικειμενική συνάρτηση για μία μόνο στιγμή. Για την αντιμετώπιση αυτής της βελτιστοποίησης, μπορεί να αναλυθεί η αντικειμενική λειτουργία και να υιοθετηθούν κατάλληλες τεχνικές. Για παράδειγμα, εάν η αντικειμενική συνάρτηση είναι τετραγωνική και οι περιορισμοί είναι γραμμικοί, μπορούν να χρησιμοποιηθούν τεχνικές τετραγωνικού προγραμματισμού.
Δυναμικός προγραμματισμός	Χρησιμοποιείται για την αντιμετώπιση σύνθετων προβλημάτων, χωρίζοντάς το σε ένα σύνολο απλούστερων υποπροβλημάτων σε πολλά βήματα, λύνοντας κάθε ένα από αυτά τα υποπροβλήματα μόνο μία φορά τη φορά και αποθηκεύοντας τις λύσεις τους στη μνήμη. Έτσι, στο μέλλον, εάν εμφανιστεί το ίδιο υποπρόβλημα, αντί να υπολογίζεται εκ νέου η λύση του, απλά αναζητείται η προηγουμένως υπολογισμένη λύση, εξοικονομώντας έτσι χρόνο υπολογισμού.
Ενίσχυση μάθησης (RL)	Πρόκειται για έναν κλάδο της μηχανικής μάθησης που χρησιμοποιείται για να βοηθήσει έναν πράκτορα να βρει τη βέλτιστη πολιτική, όταν δε διαθέτει πληροφορίες σχετικά με το περιβάλλον. Συγκεκριμένα, ο πράκτορας παρατηρεί πρώτα την τρέχουσα κατάστασή του και έπειτα λαμβάνει μια ενέργεια, ενώ στη συνέχεια λαμβάνει την άμεση ανταμοιβή του μαζί με τη νέα του κατάσταση. Οι παρατηρούμενες πληροφορίες, δηλαδή η άμεση ανταμοιβή και η νέα κατάσταση, χρησιμοποιούνται για την προσαρμογή της πολιτικής του πράκτορα. Αυτή η διαδικασία επαναλαμβάνεται έως ότου η πολιτική του πράκτορα πλησιάσει τη βέλτιστη πολιτική.
Βαθιά μάθηση (DL)	Πρόκειται για έναν κλάδο της μηχανικής μάθησης που χρησιμοποιείται για να βοηθήσει έναν πράκτορα να βρει τη βέλτιστη πολιτική όταν διατίθενται εκ των προτέρων κάποιες πληροφορίες σχετικά με το περιβάλλον. Συγκεκριμένα, ο πράκτορας θα εκπαιδεύσει το νευρωνικό δίκτυο με βάση τις ληφθείσες πληροφορίες για να βρει τις βέλτιστες παραμέτρους για το δίκτυο. Το εκπαιδευμένο νευρωνικό δίκτυο θα εφαρμοστεί στη συνέχεια στον πράκτορα για να τον βοηθήσει να λάβει αποφάσεις με ηλεκτρονικό τρόπο.
Βαθιά ενίσχυση μάθησης (DRL)	Αυτό είναι ένα προηγμένο μοντέλο τεχνικής ενίσχυσης μάθησης στο οποίο η βαθιά μάθηση χρησιμοποιείται ως αποτελεσματικό εργαλείο για τη βελτίωση του ρυθμού μάθησης για αλγόριθμους ενίσχυσης της μάθησης. Συγκεκριμένα, κατά τη διάρκεια της διαδικασίας μάθησης σε πραγματικό χρόνο, οι αποκτηθείσες εμπειρίες θα αποθηκευτούν και θα χρησιμοποιηθούν ως δεδομένα για την εκπαίδευση του νευρωνικού δικτύου. Το εκπαιδευμένο νευρωνικό δίκτυο θα χρησιμοποιηθεί στη συνέχεια για να βοηθήσει τον πράκτορα να λάβει τις βέλτιστες αποφάσεις σε πραγματικό χρόνο. Σημειώνεται ότι, σε αντίθεση με την τεχνική βαθιάς μάθησης, το νευρωνικό δίκτυο στο DRL θα εκπαιδεύεται συχνά με βάση νέες εμπειρίες που αποκτώνται κατά τη διάρκεια των αλληλεπιδράσεων με το περιβάλλον σε πραγματικό χρόνο.

Πίνακας 1. Σύγκριση μεταξύ τεχνικών βελτιστοποίησης

Η DQL συνδυάζει πλεονεκτήματα των τεχνικών ενισχυτικής και βαθιάς μάθησης, και έτσι έχει ένα ευρύ φάσμα εφαρμογών στην πράξη, όπως η ανάπτυξη παιχνιδιών (BBC, 2016), μεταφορές (Lin, Dai, Li, & Wang, 2018) και ρομποτική (Gu, Holly, Lillicrap, & Levine, 2017). Ο Πίνακας 1 συνοψίζει πόσο διαφορετικά επιλύουν προβλήματα βελτιστοποίησης

η DQL, η ενίσχυση της μάθησης, η βαθιά μάθηση και οι παραδοσιακές συνδυαστικές μέθοδοι βελτιστοποίησης.

2.5 Προχωρημένα Μοντέλα Βαθιάς Μάθησης Q (Advanced Deep Q-Learning Models)

2.5.1 Double Deep Q-Learning Models

Σε ορισμένα στοχαστικά περιβάλλοντα, ο αλγόριθμος Q-learning έχει χαμηλή απόδοση λόγω των υπερεκτιμήσεων των τιμών δράσης (Thrun & Schwartz, 1993). Αυτές οι υπερεκτιμήσεις προκύπτουν από μια θετική πόλωση (bias) που εισάγεται επειδή το Q-learning χρησιμοποιεί τη μέγιστη τιμή δράσης ως προσέγγιση για τη μέγιστη αναμενόμενη τιμή δράσης όπως φαίνεται στην εξίσωση του αλγορίθμου του Παραρτήματος Β.1:

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + a_t[r_t(s,a) + \gamma max_{a'} Q_t(s,a') - Q_t(s,a)]$$

Ο λόγος είναι ότι χρησιμοποιούνται τα ίδια δείγματα για να αποφασιστεί ποια ενέργεια είναι η καλύτερη, δηλαδή, με την υψηλότερη αναμενόμενη ανταμοιβή, και τα ίδια δείγματα χρησιμοποιούνται για την εκτίμηση αυτής της τιμής ενέργειας. Έτσι, για να ξεπεραστεί το πρόβλημα υπερεκτίμησης του αλγορίθμου Q-learning, εισάγεται μια λύση χρησιμοποιώντας δύο συναρτήσεις Q-value, δηλ. Q_1 και Q_2, για ταυτόχρονη επιλογή και αξιολόγηση τιμών ενεργείας (Hasselt, Double Q-learning, 2010). Συγκεκριμένα, η επιλογή μιας ενέργειας εξακολουθεί να οφείλεται στα δικτυακά βάρη θ_1. Αυτό σημαίνει ότι, όπως στην Q-learning, εξακολουθούμε να εκτιμούμε την αξία της greedy πολιτικής σύμφωνα με τις τρέχουσες τιμές, όπως ορίζονται από το θ_1. Ωστόσο, το δεύτερο σύνολο βαρών θ_2 χρησιμοποιείται για να αξιολογηθεί δίκαια η αξία αυτής της πολιτικής. Αυτό το δεύτερο σύνολο βαρών μπορεί να ενημερωθεί συμμετρικά αλλάζοντας τους ρόλους των θ_1 και θ_2. Με βάση αυτήν την ιδέα, αναπτύσσεται (Hasselt, Double Q-learning, 2010) το μοντέλο Double Deep Q-Learning (DDQL) (Hasselt, Guez, & Silver, Deep Reinforcement Learning with Double Q-Learning, 2016), χρησιμοποιώντας ένα Double Deep Q-Network (DDQN) με τη λειτουργία απώλειας να ενημερώνεται ως εξής:

$$\left[r_j + \gamma \hat{Q}(s_{j+1}, argmax_{aj+1} Q(s_{j+1}, a_{j+1}; \theta); \theta' - Q(s_j, a_j; \theta)\right]^2 \quad (6)$$

Σε αντίθεση με τη double Q-learning, τα βάρη του δεύτερου δικτύου θ_2 αντικαθίστανται με τα βάρη των δικτύων στόχων θ για την αξιολόγηση της τρέχουσας greedy πολιτικής,

όπως φαίνεται στην παραπάνω εξίσωση. Η ενημέρωση από το DQN στο δίκτυο στόχος παραμένει αμετάβλητη και παραμένει ως περιοδικό αντίγραφο του διασυνδεδεμένου δικτύου. Λόγω της αποτελεσματικότητας της DDQL, υπάρχουν μερικές εφαρμογές της που εισήχθησαν πρόσφατα για την αντιμετώπιση προβλημάτων πρόσβασης δυναμικού φάσματος σε πολυκάναλα ασύρματα δίκτυα (Naparstek & Cohen, 2017) και κατανομή πόρων σε ετερογενή δίκτυα (Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018).

2.5.2 Deep Q-Learning With Prioritized Experience Replay

Ο μηχανισμός επανάληψης εμπειρίας επιτρέπει στον παράγοντα ενίσχυσης μάθησης να θυμάται και να επαναχρησιμοποιεί εμπειρίες, δηλαδή μεταβάσεις, από το παρελθόν. Συγκεκριμένα, οι μεταβάσεις λαμβάνονται ομοιόμορφα από τη μνήμη επανάληψης D. Ωστόσο, αυτή η προσέγγιση απλώς επαναλαμβάνει τις μεταβάσεις στην ίδια συχνότητα με την αρχική εμπειρία του παράγοντα, ανεξάρτητα από τη σπουδαιότητά τους. Ως εκ τούτου, αναπτύχθηκε ένα πλαίσιο (Schaul, Quan, Antonoglou, & Silver, 2016) για την ιεράρχηση των εμπειριών, έτσι ώστε οι σημαντικές μεταβάσεις να αναπαράγονται πιο συχνά, και επομένως η μάθηση να γίνεται πιο αποτελεσματική. Στην ιδανική περίπτωση, εξετάζονται πιο συχνά οι μεταβάσεις, οι οποίες προσφέρουν περισσότερα στη μάθηση. Σε γενικές γραμμές, η DQL με την προτεραιοποιημένη επανάληψης εμπειρίας (Prioritized Experience Replay - PER) εξετάζει μεταβάσεις με πιθανότητα που σχετίζεται με το τελευταίο συναντημένο απόλυτο σφάλμα (Schaul, Quan, Antonoglou, & Silver, 2016). Νέες μεταβάσεις εισάγονται στην ενδιάμεση μνήμη επανάληψης με μέγιστη προτεραιότητα, αποκτώντας μια προτίμηση έναντι των πρόσφατων μεταβάσεων. Επισημαίνεται ότι, οι στοχαστικές μεταβάσεις μπορεί επίσης να προτιμούνται, ακόμα και όταν δεν έχει απομείνει μεγάλο περιθώριο μάθησης γι 'αυτές. Μέσα από πραγματικά πειράματα σε πολλά παιχνίδια Atari, αποδεικνύεται ότι η DQL με PER υπερτερεί της DQL με ομοιόμορφη επανάληψη σε 41 από τα 49 παιχνίδια. Ωστόσο, αυτή η λύση είναι κατάλληλη για εφαρμογή μόνο όταν είναι δυνατόν να βρεθούν και να οριστούν οι σημαντικές εμπειρίες στη μνήμη επανάληψης D.

2.5.3 Dueling Deep Q-Learning

Οι τιμές *Q,* δηλαδή *Q(s,α),* που χρησιμοποιούνται στον αλγόριθμο Q-learning, εκφράζουν πόσο αποτελεσματική είναι μια συγκεκριμένη ενέργεια σε μια δεδομένη κατάσταση. Η αξία μιας ενέργειας *α* σε μια δεδομένη κατάσταση *s* μπορεί να αναλυθεί σε δύο θεμελιώδεις τιμές. Η πρώτη τιμή είναι η συνάρτηση κατάστασης-τιμής *V(s)*, για να υπολογιστεί η σπουδαιότητα να βρίσκεται σε μια συγκεκριμένη κατάσταση *s*. Η δεύτερη τιμή είναι η συνάρτηση ενέργειας-τιμής *A(α)*, για να υπολογιστεί η σπουδαιότητα της επιλογής μιας ενέργειας *α* σε σύγκριση με άλλες ενέργειες. Ως αποτέλεσμα, η συνάρτηση τιμής Q μπορεί να εκφραστεί με δύο βασικές συναρτήσεις τιμής ως εξής: *Q (s, a) = V (s) + A (a).*

Σε πολλά MDPs δεν είναι απαραίτητο να εκτιμηθούν ταυτόχρονα οι τιμές ενέργειας και κατάστασης της συνάρτησης *Q(s,α).* Για παράδειγμα, σε πολλά αγωνιστικά παιχνίδια, η μετακίνηση αριστερά ή δεξιά έχει σημασία εάν και μόνο εάν ο πράκτορας συναντήσει εμπόδια ή εχθρούς. Έτσι, εισάγεται η ιδέα της χρήσης δύο ροών (Wang, Schaul, Hessel, Hasselt, Lanctot, & Freitas, 2016), δηλαδή δύο ακολουθιών, πλήρως συνδεδεμένων επιπέδων αντί να χρησιμοποιούν μία μόνο ακολουθία με πλήρως συνδεδεμένα επίπεδα για το DQN. Οι δύο ροές είναι κατασκευασμένες έτσι ώστε να είναι σε θέση να παρέχουν ξεχωριστές εκτιμήσεις σχετικά με τις συναρτήσεις τιμής ενέργειας και κατάστασης, δηλαδή, *V(s)* και *A(a).* Τέλος, οι δύο ροές συνδυάζονται για να δημιουργήσουν μία μόνο έξοδο *Q (s, a)* ως εξής:

$$Q(s,a;\boldsymbol{\alpha},\boldsymbol{\beta}) = V(s;\boldsymbol{\beta}) + \left(A(s,a;\boldsymbol{\alpha}) - \frac{\sum_{a\prime} A(s,a\prime;\boldsymbol{\alpha})}{|A|}\right) \qquad (7)$$

όπου τα $\boldsymbol{\beta}$ και $\boldsymbol{\alpha}$ είναι οι παράμετροι των δύο ρευμάτων $V(s;\boldsymbol{\beta})$ και $A(s,a';\boldsymbol{\alpha})$, αντίστοιχα. Εδώ, $|A|$ είναι ο συνολικός αριθμός ενεργειών στο χώρο δράσης A. Στη συνέχεια, η συνάρτηση απώλειας προκύπτει με τον ίδιο τρόπο με τα βάρη του αλγορίθμου δηλαδή: $\left[r_j + \gamma max_{a_{j+1}} \hat{Q}(s_{j+1}, a_{j+1}; \theta') - Q(s_j, a_j; \theta)\right]^2$. Μέσω της προσομοίωσης, οι συγγραφείς δείχνουν ότι η προτεινόμενη μονομαχία DQN μπορεί να ξεπεράσει το DDQN (Hasselt, Guez, & Silver, Deep Reinforcement Learning with Double Q-Learning, 2016) σε 50 από τα 57 εκπαιδευμένα παιχνίδια Atari. Ωστόσο, η προτεινόμενη αρχιτεκτονική μονομαχίας ωφελεί σαφώς μόνο για MDPs με μεγάλους χώρους δράσης. Για μικρούς χώρους κατάστασης, η απόδοση της μονομαχίας DQL δεν είναι ούτε τόσο καλή όσο της διπλής DQL όπως φαίνεται στα αποτελέσματα προσομοίωσης στο (Wang, Schaul, Hessel, Hasselt, Lanctot, & Freitas, 2016).

2.5.4 Asynchronous Multi-Step Deep Q-Learning

Οι περισσότερες από τις μεθόδους Q-learning, όπως η DQL και η Dueling DQL βασίζονται στη μέθοδο επανάληψης εμπειρίας. Ωστόσο, μια τέτοια μέθοδος έχει πολλά μειονεκτήματα. Για παράδειγμα, χρησιμοποιεί περισσότερους πόρους μνήμης και υπολογισμού ανά πραγματική αλληλεπίδραση και απαιτεί αλγόριθμους εκμάθησης εκτός πολιτικής που μπορούν να ενημερώνονται από δεδομένα που δημιουργούνται από παλαιότερη πολιτική. Αυτό περιορίζει τις εφαρμογές του DQL. Ως εκ τούτου, εισάγεται μια μέθοδος (Mnih, et al., 2016) που χρησιμοποιεί πολλαπλούς πράκτορες για την παράλληλη εκπαίδευση του DNN. Συγκεκριμένα, προτείνεται μια εκπαιδευτική διαδικασία η οποία χρησιμοποιεί ασύγχρονες ενημερώσεις με βαθμιδωτή κλίση καθόδου από πολλούς πράκτορες ταυτόχρονα. Αντί να εκπαιδεύεται ένας μόνο πράκτορας που αλληλεπιδρά με το περιβάλλον του, πολλοί πράκτορες αλληλεπιδρούν ταυτόχρονα με τη δική τους εκδοχή του περιβάλλοντος. Μετά από ορισμένο αριθμό χρονικών βημάτων, οι συσσωρευμένες ενημερώσεις διαβάθμισης από έναν πράκτορα εφαρμόζονται σε ένα καθολικό μοντέλο, δηλαδή στο DNN. Αυτές οι ενημερώσεις είναι ασύγχρονες και δεν κλειδώνουν. Επιπλέον, για την ανταλλαγή μεταξύ πόλωσης και διακύμανσης στη διαβάθμιση πολιτικής, υιοθετείται η μέθοδος ενημερώσεων n-step (Sutton & Barto, 1998) για να ενημερώνεται η συνάρτηση επιβράβευσης. Συγκεκριμένα, η συντετμημένη συνάρτηση ανταμοιβής n-step μπορεί να οριστεί από $r_t^{(n)} = \sum_{k=0}^{n-1} \gamma^{(k)} r_{t+k+1}$. Έτσι, η εναλλακτική απώλεια για κάθε πράκτορα θα προκύψει από:

$$\left[r_j^{(n)} + \gamma_j^{(n)} max_{a'} \hat{Q}(s_{j+n}, a'; \theta') - Q(s_j, a_j; \theta)\right]^2 \qquad (8)$$

Τα αποτελέσματα της ταχύτητας εκπαίδευσης και της ποιότητας της προτεινόμενης ασύγχρονης DQL με την εκμάθηση πολλαπλών βημάτων αναλύονται για διάφορες μεθόδους ενίσχυσης μάθησης, π.χ., 1-step Q-learning, 1-step SARSA και n-step Q-learning. Φαίνεται ότι, οι ασύγχρονες ενημερώσεις έχουν σταθεροποιητική επίδραση στις ενημερώσεις πολιτικής και τιμών. Επίσης, η προτεινόμενη μέθοδος ξεπερνά τους τρέχοντες υπερσύγχρονους αλγόριθμους στα παιχνίδια Atari, ενώ εκπαιδεύεται για το ήμισυ του χρόνου σε έναν μόνο πολυπύρηνο επεξεργαστή (CPU) αντί για μια μονάδα επεξεργασίας γραφικών (GPU). Ως αποτέλεσμα, έχουν αναπτυχθεί ορισμένες πρόσφατες εφαρμογές ασύγχρονης DQL για προβλήματα ελέγχου μεταγωγών σε ασύρματα συστήματα (Wang, Li, Xu, Tian, & Cui, 2018).

2.5.5 Distributional Deep Q-Learning

Όλες οι προαναφερθείσες μέθοδοι χρησιμοποιούν την εξίσωση Bellman για να προσεγγίσουν την αναμενόμενη αξία των μελλοντικών ανταμοιβών. Ωστόσο, εάν το περιβάλλον είναι στοχαστικό στη φύση και οι μελλοντικές ανταμοιβές ακολουθούν την πολυτροπική κατανομή, η επιλογή ενεργειών με βάση την αναμενόμενη αξία ενδέχεται να μην οδηγήσει στο βέλτιστο αποτέλεσμα. Για παράδειγμα, είναι γνωστό ότι ο αναμενόμενος χρόνος μετάδοσης ενός πακέτου σε ασύρματο δίκτυο είναι 20 λεπτά. Ωστόσο, αυτή η πληροφορία μπορεί να μην είναι τόσο σημαντική, επειδή μπορεί τις περισσότερες φορές να υπερεκτιμάται ο χρόνος μετάδοσης. Για παράδειγμα, ο αναμενόμενος χρόνος μετάδοσης υπολογίζεται με βάση τις κανονικές μεταδόσεις (χωρίς συγκρούσεις) και τις μεταδόσεις παρεμβολών (με συγκρούσεις), αν και οι μεταδόσεις παρεμβολών είναι πολύ σπάνιες, αλλά χρειάζονται πολύ χρόνο. Αυτό καθιστά τις εκτιμήσεις μη χρήσιμες για τους αλγόριθμους DQL.

Έτσι, εισάγεται μια λύση (Bellemare, Dabney, & Munos, 2017) χρησιμοποιώντας διανεμητική ενίσχυση μάθησης για να ενημερώνεται η συνάρτηση Q-value με βάση τη διανομή και όχι τις προσδοκίες της. Συγκεκριμένα, έστω $Z(s, \alpha)$ η επιστροφή που λαμβάνεται ξεκινώντας από την κατάσταση s, εκτελώντας την ενέργεια α και ακολουθώντας την τρέχουσα πολιτική, τότε $Q(s, \alpha) = E[Z(s, \alpha)]$. Εδώ, το Z αντιπροσωπεύει τη διανομή μελλοντικών ανταμοιβών, η οποία δεν είναι πλέον μια βαθμιαία ποσότητα όπως οι τιμές Q. Στη συνέχεια, λαμβάνεται η διανεμητική έκδοση της εξίσωσης Bellman ως εξής: $Z(s, \alpha) = r + \gamma Z(s', a')$. Παρόλο που η Distributional Deep Q-Learning αποδεικνύεται ότι υπερτερεί της συμβατικής DQL (Mnih, et al., Human-level control through deep reinforcement learning, 2015) σε πολλά παιχνίδια Atari 2600 (45 από τα 57 παιχνίδια), η απόδοσή της βασίζεται πολύ στη λειτουργία διανομής Z. Εάν το Z είναι καλά καθορισμένο, η απόδοση Distributional Deep Q-Learning είναι πολύ πιο σημαντική από αυτήν της DQL. Διαφορετικά, η απόδοσή της είναι χειρότερη από εκείνη της DQL.

2.5.6 Deep Q-Learning With Noisy Nets

Το Noisy Net (Fortunato, et al., 2018) είναι ένα είδος νευρωνικού δικτύου του οποίου η πόλωση (παράγων προδιάθεσης) και τα βάρη διαταράσσονται επαναληπτικά κατά τη διάρκεια της εκπαίδευσης από μια παραμετρική συνάρτηση του θορύβου. Αυτό το δίκτυο προσθέτει τον θόρυβο Gauss στα τελευταία (πλήρως συνδεδεμένα) στρώματα του δικτύου. Οι παράμετροι αυτού του θορύβου μπορούν να ρυθμιστούν από το μοντέλο κατά τη διάρκεια της εκπαίδευσης, το οποίο επιτρέπει στον πράκτορα να αποφασίσει πότε και σε ποιο ποσοστό θέλει να εισαγάγει την αβεβαιότητα στα βάρη του. Συγκεκριμένα, για να εφαρμοστεί το Noisy Net, αντικαθιστούμε πρώτα την πολιτική ε-greedy με μια συνάρτηση τυχαιοποιημένης τιμής-ενέργειας. Στη συνέχεια, τα πλήρως συνδεδεμένα επίπεδα του δικτύου τιμής παραμετροποιούνται ως Noisy Net, όπου οι παράμετροι λαμβάνονται από την κατανομή παραμέτρων θορύβου δικτύου (noisy network) μετά από κάθε βήμα επανάληψης. Για επανάληψη, το τρέχον δείγμα παραμέτρου θορύβου δικτύου διατηρείται σταθερό κατά μήκος της σειράς. Δεδομένου ότι η DQL κάνει ένα βήμα βελτιστοποίησης για κάθε βήμα δράσης, οι παράμετροι θορύβου δικτύου επαναλαμβάνονται στο δείγμα πριν από κάθε ενέργεια.

Μέσα από πειραματικά αποτελέσματα, αποδεικνύεται ότι προσθέτοντας το επίπεδο θορύβου Gauss στο DNN, η απόδοση της συμβατικής DQL (Mnih, et al., Human-level control through deep reinforcement learning, 2015), Dueling DQL (Wang, Schaul, Hessel, Hasselt, Lanctot, & Freitas, 2016) και η ασύγχρονη DQL (Mnih, et al., 2016) μπορούν να βελτιωθούν σημαντικά για ένα ευρύ φάσμα παιχνιδιών Atari . Ωστόσο, η επίδραση του θορύβου στην απόδοση των αλγορίθμων βαθιάς DQL βρίσκεται ακόμη υπό συζήτηση στη βιβλιογραφία και επομένως η ανάλυση της επίδρασης του επιπέδου θορύβου απαιτεί περαιτέρω έρευνες.

2.5.7 Rainbow Deep Q-Learning

Ο πράκτορας εκμάθησης Rainbow DQL (Hessel, et al., 2018) ενσωματώνει όλα τα πλεονεκτήματα των επτά προαναφερθεισών λύσεων, συμπεριλαμβανομένης της DQL. Συγκεκριμένα, αυτός ο αλγόριθμος καθορίζει πρώτα τη λειτουργία απώλειας με βάση την ασύγχρονη DQL πολλαπλών βημάτων και διανομής. Στη συνέχεια, συνδυάζεται η απώλεια διανομής πολλαπλών βημάτων με την double Q-learning χρησιμοποιώντας την ενέργεια greedy στο s_{t+n} που έχει επιλεγεί σύμφωνα με το δίκτυο Q ως

ανατροφοδοτούμενη δράση a_{t+n}^{*} και αξιολογεί τη δράση χρησιμοποιώντας τον δίκτυο στόχο.

Αλγόριθμος DQL	**Βασικά Χαρακτηριστικά**	**Πλεονεκτήματα**	**Μειονεκτήματα**	**Εφαρμογές**
DQL	Χρησιμοποιεί το DNN για να εκπαιδεύσει τη συνάρτηση Q-value	Απλός στην εφαρμογή με γρήγορη σύγκλιση	Υπερεκτίμηση των τιμών ενεργείας	Κατάλληλο να εφαρμόζεται σε MDPs με μικρό αριθμό ενεργειών
DDQL	Χρησιμοποιεί δύο συναρτήσεις Q-value για να επιλέξει και να αξιολογήσει ταυτόχρονα τιμές ενεργειών	Απλός στην εφαρμογή και γρηγορότερη σύγκλιση από τον DQL	Δε λαμβάνει υπόψη τα ειδικά χαρακτηριστικά των MDPs	Εφαρμόσιμο σε MDPs
Prioritized DDQL	Προτεραιοποιεί εμπειρίες στη μνήμη επανάληψης	Γρηγορότερη σύγκλιση από τον DQL και τον DDQL	Απαιτεί πληροφορίες σχετικά με σημαντικές πληροφορίες στη μνήμη επανάληψης	Ιδιαίτερα αποτελεσματικό για MDPs με προτεραιοποιημένες εμπειρίες
Dueling DDQL	Χρησιμοποιεί δύο DNNs για να υπολογίσει ταυτόχρονα τις συναρτήσεις τιμών ενέργειας και κατάστασης	Πολύ γρηγορότερη σύγκλιση από τον DQL, τον DDQL και τον Prioritized DDQL	Μεγάλη πολυπλοκότητα και μικρή αποτελεσματικότητα στις MDPs σε μικρούς χώρους κατάστασης και ενέργειας	Ιδιαίτερα αποτελεσματικό να αντιμετωπίσει MDPs με μεγάλης έκτασης χώρους δράσης και κατάστασης
Asynchronous DQL	Χρησιμοποιεί πολλαπλούς πράκτορες για να εκπαιδεύσει το DNN παράλληλα	Η ταχύτητα μάθησης είναι εξαιρετικά γρήγορη.	Μεγάλη πολυπλοκότητα με έντονες απαιτήσεις σε συσκευές υλικού για εκπαίδευση	Ιδιαίτερα αποτελεσματικό να αντιμετωπίσει MDPs με πολύ μεγάλης έκτασης χώρους δράσης και κατάστασης
Distributional DQL	Χρησιμοποιεί συνάρτηση κατανομής για να ενημερώσει τη συνάρτηση Q-value	Μεγαλύτερη ακρίβεια στην αξιολόγηση της συνάρτησης Q-value	Απαιτεί τη γνώση της κατανομής της συνάρτησης ανταμοιβής σε χώρους κατάστασης και ενέργειας	Κατάλληλο να εφαρμοστεί σε MDPs με διαθέσιμη κατανομή συνάρτησης ανταμοιβής
Noisy Nets DQL	Προσθέτει για εκπαίδευση το επίπεδο θορύβου Gauss στο DNN	Βελτίωση της αποτελεσματικότητας εξερεύνησης του περιβάλλοντος	Η αποτελεσματικότητα της προσθήκης επίπεδου θορύβου Gaussian αμφισβητείται	Ιδιαίτερα αποτελεσματικό να αντιμετωπίσει MDPs με πολύ μεγάλης έκτασης χώρους δράσης και κατάστασης
Rainbow	Συνδυάζει τα χαρακτηριστικά όλων των παραπάνω αλγορίθμων	Συνδυάζει τα παραπάνω πλεονεκτήματα	Εξαιρετικά πολύπλοκος με πολλές εκ των προτέρων απαιτήσεων σε MDPs	Κατάλληλο για MDPs μεγάλης έκτασης χώρων δράσης και κατάστασης και κάποιων ιδιοτήτων που είναι γνωστές εξαρχής.

Πίνακας 2. Σύγκριση απόδοσης μεταξύ των αλγορίθμων DQL

Στην τυπική τεχνική αναλογικής επανάληψης προτεραιότητας (Schaul, Quan, Antonoglou, & Silver, 2016), το απόλυτο σφάλμα Temporal Difference (TD) χρησιμοποιείται για να δώσει προτεραιότητα στις μεταβάσεις. Εδώ, το σφάλμα TD σε ένα χρονικό διάστημα είναι το σφάλμα στην εκτίμηση που πραγματοποιήθηκε στο χρονικό

διάστημα. Ωστόσο, στον προτεινόμενο αλγόριθμο Rainbow DQL, όλες οι παραλλαγές διανεμητικού Rainbow δίνουν προτεραιότητα στις μεταβάσεις από την απώλεια Kullbeck-Leibler (KL) επειδή αυτή η απώλεια μπορεί να είναι πιο ισχυρή σε θορυβώδες στοχαστικό περιβάλλον. Εναλλακτικά, η αρχιτεκτονική dueling στα DNNs παρουσιάζεται στο (Wang, Schaul, Hessel, Hasselt, Lanctot, & Freitas, 2016). Τέλος, το Noisy Net layer (Hessel, et al., 2018) χρησιμοποιείται για την αντικατάσταση όλων των γραμμικών επιπέδων προκειμένου να μειωθεί ο αριθμός των ανεξάρτητων μεταβλητών θορύβου. Μέσω της προσομοίωσης, αποδεικνύεται ότι αυτή είναι η πιο προηγμένη τεχνική που ξεπερνά σχεδόν όλους τους τρέχοντες αλγόριθμους DQL στη βιβλιογραφία για πάνω από 57 παιχνίδια Atari 2600.

Στον Πίνακα 2 συνοψίζονται οι αλγόριθμοι DQL και η απόδοσή τους κάτω από τις ρυθμίσεις παραμέτρων που χρησιμοποιήθηκαν στο (Hessel, et al., 2018). Όπως παρατηρήθηκε στον παραπάνω Πίνακα, όλοι οι αλγόριθμοι DQL έχουν αναπτυχθεί από το Google Deep Mind με βάση την αρχική εργασία στο (Mnih, et al., Human-level control through deep reinforcement learning, 2015). Μέχρι στιγμής, μέσω πειραματικών αποτελεσμάτων σε παιχνίδια Atari 2600, το Rainbow DQL παρουσιάζει πολύ εντυπωσιακά αποτελέσματα σε σχέση με όλους τους άλλους αλγόριθμους DQL. Ωστόσο, πρέπει να διεξαχθούν περισσότερα πειράματα σε διαφορετικούς τομείς για να επιβεβαιωθεί η πραγματική απόδοση του παραπάνω αλγορίθμου.

2.6 Βαθιά Μάθηση Q για Επεκτάσεις των Αποφάσεων Διαδικασιών Markov (Deep Q-Learning for Extensions of MDPs)

2.6.1 Deep Deterministic Policy Gradient Q-Learning for Continuous Action

Αν και ο αλγόριθμος DQL μπορεί να επιλύσει προβλήματα με χώρους καταστάσεων υψηλών διαστάσεων, μπορεί μόνο να χειριστεί διακριτούς και χαμηλών διαστάσεων χώρους δράσεων. Ωστόσο, τα συστήματα σε πολλές εφαρμογές έχουν συνεχείς, δηλαδή πραγματικές τιμές και χώρους δράσεων υψηλών διαστάσεων. Οι αλγόριθμοι DQL δεν μπορούν να εφαρμοστούν άμεσα σε συνεχείς χώρους δράσεων, καθώς βασίζονται στην επιλογή της καλύτερης δράσης που μεγιστοποιεί τη συνάρτηση Q-value. Συγκεκριμένα,

μια πλήρης αναζήτηση σε έναν συνεχή χώρο δράσης για την εύρεση της βέλτιστης δράσης είναι συχνά ανέφικτη.

Στη συνέχεια παρουσιάζεται ένας αλγόριθμος actor-critic χωρίς πολιτική και χωρίς μοντέλο, χρησιμοποιώντας προσεγγιστές βαθιάς λειτουργίας που μπορούν να μάθουν πολιτικές σε διαστατικούς χώρους συνεχούς δράσης (Lillicrap, et al., 2016). Η βασική ιδέα βασίζεται στον αλγόριθμο ντετερμινιστικής πολιτικής κλίσης (Deterministic Policy Gradient - DPG) (Silver, Lever, Heess, Degris, Wierstra, & Riedmiller, 2014). Συγκεκριμένα, ο αλγόριθμος DPG διατηρεί μια παραμετροποιημένη συνάρτηση παράγοντα $\mu(s;\theta^{\mu})$ με τον παράμετρο θ που καθορίζει την τρέχουσα πολιτική με καθοριστική αντιστοίχιση καταστάσεων σε μια συγκεκριμένη ενέργεια. Ο critic $Q(s,a)$ μαθαίνεται χρησιμοποιώντας την εξίσωση Bellman όπως στην Q-learning. Ο actor ενημερώνεται εφαρμόζοντας τον κανόνα της αλυσίδας στην αναμενόμενη επιστροφή από την αρχική διανομή σε σχέση με τις παραμέτρους του actor.

Με βάση αυτόν τον κανόνα ενημέρωσης, εισάγεται ο αλγόριθμος Deep DPG (DDPG) που μπορεί να μάθει ανταγωνιστικές πολιτικές χρησιμοποιώντας παρατηρήσεις μικρών διαστάσεων, π.χ. καρτεσιανές συντεταγμένες ή κοινές γωνίες, υπό τις ίδιες υπερπαραμέτρους και δομή δικτύου. Ο αλγόριθμος δημιουργεί ένα αντίγραφο των δικτύων actor και critic $Q'(s,a;\theta^{Q'})$ και $\mu'(s;\theta^{\mu'})$, αντίστοιχα, για τον υπολογισμό των τιμών-στόχων. Στη συνέχεια, τα βάρη αυτών των δικτύων στόχων ενημερώνονται με αργή παρακολούθηση στα δίκτυα μάθησης, δηλαδή, $\theta' \rightarrow \tau\theta + (1-\tau)\theta'$ με $\tau \ll 1$. Αυτό σημαίνει ότι οι τιμές-στόχοι περιορίζονται να αλλάζουν αργά, βελτιώνοντας σημαντικά τη σταθερότητα της μάθησης. Σημειώνεται ότι, η κύρια πρόκληση της μάθησης σε χώρους συνεχούς δράσης είναι η εξερεύνηση. Επομένως, στον προτεινόμενο αλγόριθμο, η πολιτική εξερεύνησης μ' παράγεται με την προσθήκη θορύβου ο οποίος προκύπτει με δειγματοληψία από μια διαδικασία θορύβου N στην πολιτική του actor.

2.6.2 Deep Recurrent Q-Learning για POMDPs

Για την αντιμετώπιση προβλημάτων σε μερικώς παρατηρήσιμα περιβάλλοντα με μάθηση βαθιάς ενίσχυσης, εισήχθη ένα πλαίσιο εργασίας που ονομάζεται Deep Recurrent Q-Learning (DRQN) (Hausknecht & Stone, 2015) στο οποίο χρησιμοποιήθηκε ένα στρώμα LSTM για να αντικαταστήσει το πρώτο μετά-συνελικτικό πλήρως συνδεδεμένο επίπεδο

του συμβατικού DQN. Η επαναλαμβανόμενη δομή είναι σε θέση να ενσωματώσει ένα αυθαίρετα μακρύ ιστορικό για να εκτιμήσει καλύτερα την τρέχουσα κατάσταση αντί να χρησιμοποιεί ένα ιστορικό σταθερού μήκους όπως στα DQNs. Έτσι, τα DRQNs υπολογίζουν τη συνάρτηση $Q(o_t,h_{t-1};\theta)$ αντί για $Q((s_t, \alpha_t);\theta)$, όπου θ υποδηλώνει τις παραμέτρους ολόκληρου του δικτύου, το h_{t-1} υποδηλώνει την έξοδο του επιπέδου LSTM στο προηγούμενο βήμα, δηλαδή, $h_t=LSTM(h_{t-1}, o_t)$. Το DRQN αντιστοιχεί στην απόδοση του DQN σε τυπικά προβλήματα MDP και υπερτερεί του DQN σε μερικώς παρατηρήσιμους τομείς. Όσον αφορά τη διαδικασία εκπαίδευσης, το DRQN λαμβάνει υπόψη μόνο τα συνελικτικά χαρακτηριστικά του ιστορικού παρατήρησης αντί να ενσωματώνει αναλυτικά τις ενέργειες. Μέσα από τα πειράματα, αποδεικνύεται ότι, το DRQN είναι ικανό να χειρίζεται μερική παρατηρησιμότητα και η επανάληψη παρέχει οφέλη όταν αλλάζει η ποιότητα των παρατηρήσεων κατά τη διάρκεια του χρόνου αξιολόγησης.

2.6.3 Deep SARSA Learning

Η Deep SARSA Learning είναι μια τεχνική DQL που βασίζεται στην εκμάθηση SARSA για να βοηθήσει τον πράκτορα να καθορίσει τις βέλτιστες πολιτικές μέσω διαδικτύου (Zhao, Wang, Shao, & Zhu, 2016). Σε αυτόν τον αλγόριθμο, δεδομένης της τρέχουσας κατάστασης s, το CNN χρησιμοποιείται για την απόκτηση της τρέχουσας τιμής κατάστασης-ενέργειας $Q(s,\alpha)$. Στη συνέχεια, η τρέχουσα ενέργεια α επιλέγεται από τον αλγόριθμο ε-greedy, οπότε μπορεί να παρατηρηθεί η άμεση ανταμοιβή r και η επόμενη κατάσταση s'. Για να εκτιμηθεί το τρέχον $Q(s,\alpha)$, λαμβάνεται η επόμενη τιμή κατάστασης-ενέργειας $Q(s',\alpha')$. Εδώ, όταν η επόμενη κατάσταση s' χρησιμοποιείται ως είσοδος του CNN, το $Q(s',\alpha')$ μπορεί να ληφθεί ως έξοδος. Στη συνέχεια, ένας παράγοντας ετικέτας που σχετίζεται με το $Q(s,\alpha)$ ορίζεται ως $Q(s',\alpha')$ που αντιπροσωπεύει τον παράγοντα στόχο. Οι δύο παράγοντες έχουν μόνο ένα διαφορετικό μέρος, δηλαδή, $r+\gamma Q(s',\alpha') \rightarrow Q(s,\alpha)$. Θα πρέπει να σημειωθεί ότι κατά τη διάρκεια της εκπαίδευσης και για την εκτίμηση της τρέχουσας τιμής κατάστασης-ενέργειας η επόμενη ενέργεια α' δεν είναι ποτέ greedy. Αντιθέτως, υπάρχει μια μικρή πιθανότητα να επιλεγεί μια τυχαία ενέργεια για εξερεύνηση.

2.6.4 Deep Q-Learning for Markov Games

Για να διαμορφωθούν (προβληθούν) τα προβλήματα του πραγματικού κόσμου για το δίλημμα του φυλακισμένου (Prisoner's Dilemma - PD) εισάγεται η γενική έννοια του διαδοχικού διλήμματος κρατουμένων (Sequential Prisoner's Dilemma - SPD) (Wang, Hao, Wang, & Taylor, 2018). Δεδομένου ότι το SPD είναι πιο περίπλοκο από το PD, οι υπάρχουσες προσεγγίσεις που αφορούν τη μάθηση σε παιχνίδια PD matrix δεν μπορούν να εφαρμοστούν άμεσα στο SPD. Έτσι, προτείνεται (Wang, Hao, Wang, & Taylor, 2018) μια προσέγγιση DRL πολλαπλών πρακτόρων για αμοιβαία συνεργασία σε παιχνίδια SDP. Η βαθιά ενίσχυση μάθησης πολλαπλών παραγόντων προς αμοιβαία συνεργασία αποτελείται από δύο φάσεις, τη φάση offline και τη φάση online. Η φάση offline δημιουργεί πολιτικές με διαφορετικούς βαθμούς συνεργασίας. Δεδομένου ότι ο αριθμός των πολιτικών με διαφορετικούς βαθμούς συνεργασίας είναι άπειρος, είναι υπολογιστικά ανέφικτο να εκπαιδεύονται όλες οι πολιτικές από το μηδέν. Για να αντιμετωπίσει αυτό το ζήτημα, ο αλγόριθμος εκπαιδεύει πρώτα αντιπροσωπευτικές πολιτικές χρησιμοποιώντας actor-critic έως ότου συγκλίνει, δηλαδή βασική πολιτική συνεργασίας και απόσχισης. Στη συνέχεια, ο αλγόριθμος συνθέτει το πλήρες φάσμα πολιτικών από τις παραπάνω βασικές πολιτικές. Μια άλλη εργασία είναι να ανιχνευθεί αποτελεσματικά ο βαθμός συνεργασίας του αντιπάλου. Ο αλγόριθμος χωρίζει αυτή την εργασία σε δύο βήματα. Πρώτον, ο αλγόριθμος εκπαιδεύει offline ένα δίκτυο ανίχνευσης βαθμού συνεργασίας που βασίζεται σε LSTM, το οποίο στη συνέχεια θα χρησιμοποιηθεί για ανίχνευση σε πραγματικό χρόνο κατά τη διάρκεια της φάσης online. Στη φάση online, ο πράκτορας παίζει εναντίον των αντιπάλων με μια πολιτική ελαφρώς υψηλότερου βαθμού συνεργασίας από εκείνη του αντιπάλου. Αφενός, ο αλγόριθμος προσανατολίζεται στη συνεργασία και επιδιώκει αμοιβαία συνεργασία όποτε είναι δυνατόν. Από την άλλη πλευρά, ο αλγόριθμος είναι επίσης ισχυρός ενάντια στην ατομιστική εκμετάλλευση και καταφεύγει στη στρατηγική απόσχισης όποτε είναι απαραίτητο για να αποφύγει την εκμετάλλευση.

Σε αντίθεση με ένα επαναλαμβανόμενο παιχνίδι κανονικής φόρμας με πλήρεις πληροφορίες (Wang, Hao, Wang, & Taylor, 2018), εισάγεται μια εφαρμογή DRL (Heinrich & Silver, 2016) για εκτεταμένης μορφής παιχνιδιών με ατελείς πληροφορίες. Συγκεκριμένα, το Neural Fictitious Self-Play (NFSP) είναι μια μέθοδο DRL για την εκμάθηση της κατά προσέγγισης ισορροπίας Nash των παιχνιδιών με ατελείς πληροφορίες. Το NFSP συνδυάζει FSP με προσέγγιση λειτουργιών νευρωνικού δικτύου. Ένας πράκτορας NFSP έχει δύο νευρωνικά δίκτυα. Το πρώτο δίκτυο εκπαιδεύεται με

ενίσχυση μάθησης χρησιμοποιώντας απομνημονευμένη εμπειρία του παιχνιδιού ενάντια σε συναδέλφους πράκτορες. Αυτό το δίκτυο μαθαίνει μια κατά προσέγγιση καλύτερη απάντηση από την ιστορική συμπεριφορά άλλων παραγόντων. Το δεύτερο δίκτυο εκπαιδεύεται από την εποπτευόμενη εκμάθηση από την απομνημονευμένη εμπειρία της συμπεριφοράς του ίδιου πράκτορα. Αυτό το δίκτυο μαθαίνει ένα μοντέλο με μέσο όρο τις δικές του ιστορικές στρατηγικές. Ο πράκτορας συμπεριφέρεται σύμφωνα με ένα μείγμα της μέσης στρατηγικής του και της στρατηγικής βέλτιστης απόκρισης.

Στο NSFP, όλοι οι παίκτες του παιχνιδιού ελέγχονται από ξεχωριστούς πράκτορες NFSP που μαθαίνουν από το ταυτόχρονο παιχνίδι εναντίον του άλλου, δηλαδή, το self-play. Ένας πράκτορας NFSP αλληλεπιδρά με τους συναδέλφους του και απομνημονεύει την εμπειρία του από τις μεταβάσεις παιχνιδιών και τη δική του καλύτερη συμπεριφορά απόκρισης σε δύο μνήμες, το M_{RL} και το M_{SL}. Το NFSP αντιμετωπίζει αυτές τις αναμνήσεις ως δύο ξεχωριστά σύνολα δεδομένων κατάλληλα για DRL και εποπτευόμενη ταξινόμηση, αντίστοιχα. Ο πράκτορας εκπαιδεύει ένα νευρωνικό δίκτυο, $Q(s,a;\theta^Q)$, για να προβλέψει τιμές ενέργειας από δεδομένα στην M_{RL} χρησιμοποιώντας off-policy ενίσχυση μάθησης. Το δίκτυο που προκύπτει καθορίζει τη στρατηγική βέλτιστης απόκρισης του πράκτορα, $\beta=\varepsilon\text{-greedy}(Q)$, η οποία επιλέγει μια τυχαία ενέργεια με πιθανότητα e ενώ διαφορετικά την ενέργεια που μεγιστοποιεί τις προβλεπόμενες τιμές ενέργειας. Ο πράκτορας εκπαιδεύει ένα ξεχωριστό νευρωνικό δίκτυο $\Pi(s,a;\theta^\Pi)$ για να μιμηθεί τη δική του συμπεριφορά βέλτιστης απόκρισης στο παρελθόν, χρησιμοποιώντας εποπτευόμενη ταξινόμηση στα δεδομένα στη M_{SL}. Το NFSP χρησιμοποιεί επίσης δύο τεχνικές καινοτομίες προκειμένου να διασφαλίσει τη σταθερότητα του προκύπτοντος αλγορίθμου, καθώς και να επιτρέψει την ταυτόχρονη εκμάθηση self-play. Μέσα από πειραματικά αποτελέσματα, αποδεικνύεται ότι το NFSP μπορεί να συγκλίνει σε προσέγγιση της ισορροπίας Nash σε ένα μικρό παιχνίδι πόκερ.

Κεφάλαιο 3

Χαρακτηριστικά Δικτύων με Εφαρμογή Βαθιάς Ενίσχυσης της Μάθησης

Τα σύγχρονα δίκτυα όπως το IoT γίνονται πιο αποκεντρωμένα και ad-hoc. Σε τέτοια δίκτυα, οντότητες όπως αισθητήρες και χρήστες κινητής τηλεφωνίας, πρέπει να λαμβάνουν ανεξάρτητες αποφάσεις, π.χ. επιλογές καναλιού και σταθμού βάσης, για να επιτύχουν τους δικούς τους στόχους, όπως μεγιστοποίηση της απόδοσης. Ωστόσο, αυτό είναι δύσκολο λόγω της δυναμικής και της αβεβαιότητας της κατάστασης του δικτύου. Οι αλγόριθμοι εκμάθησης όπως η DQL επιτρέπουν την εκμάθηση και τη δημιουργία γνώσεων σχετικά με τα δίκτυα, που χρησιμοποιούνται για να επιτρέψουν στις οντότητες του δικτύου να λαμβάνουν τις βέλτιστες αποφάσεις τους. Σε αυτήν την ενότητα, εξετάζονται οι εφαρμογές του DQL για τα ακόλουθα ζητήματα:

- Πρόσβαση στο δίκτυο
- Προσαρμοστικός Έλεγχος Ρυθμού Δεδομένων
- Ασύρματη Προληπτική Προσωρινή Αποθήκευση
- Δεδομένα και Υπολογισμός Εκφόρτωσης
- Ασφάλεια Δικτύου
- Διατήρηση Συνδεσιμότητας

3.1 Πρόσβαση στο Δίκτυο

Αυτή η ενότητα περιγράφει τον τρόπο χρήσης της DQL για την επίλυση της πρόσβασης φάσματος και της συσχέτισης χρηστών σε δίκτυα.

Η δυναμική πρόσβαση φάσματος επιτρέπει στους χρήστες τοπικά να επιλέγουν κανάλια για τη μεγιστοποίηση της απόδοσης τους. Ωστόσο, οι χρήστες ενδέχεται να μην έχουν

πλήρη γνώση της κατάστασης του συστήματος, π.χ. καταστάσεις καναλιού. Έτσι, η DQL μπορεί να χρησιμοποιηθεί ως ένα αποτελεσματικό εργαλείο για δυναμική πρόσβαση στο φάσμα.

Η συσχέτιση χρηστών εφαρμόζεται για να προσδιοριστεί ποιος χρήστης θα ανατεθεί σε ποιό σταθμό βάσης (BS). Τα προβλήματα κοινής σύνδεσης των χρηστών και πρόσβασης στο φάσμα μελετώνται στα (Fooladivanda & Rosenberg, 2013) και (Lin, Bao, Yu, & Liang, 2015). Ωστόσο, τα προβλήματα είναι συνήθως συνδυαστικά και μη κυρτά που απαιτούν σχεδόν ολοκληρωμένες και ακριβείς πληροφορίες δικτύου για την απόκτηση της βέλτιστης στρατηγικής. Η DQL είναι σε θέση να παρέχει κατανεμημένες λύσεις που μπορούν να χρησιμοποιηθούν αποτελεσματικά για τα ανωτέρω προβλήματα.

3.1.1 Δυναμική Πρόσβαση Φάσματος (Dynamic Spectrum Access)

Οι συγγραφείς στο (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access, 2017) προτείνουν ένα δυναμικό σχήμα πρόσβασης καναλιού ενός αισθητήρα που βασίζεται στο DQL για IoT. Σε κάθε χρονοθυρίδα, ο αισθητήρας επιλέγει ένα από τα κανάλια M για τη μετάδοση του πακέτου του. Η κατάσταση του καναλιού είναι είτε σε χαμηλή παρεμβολή, δηλαδή σε επιτυχημένη μετάδοση, είτε σε υψηλή παρεμβολή, δηλαδή σε αποτυχία μετάδοσης. Δεδομένου ότι ο αισθητήρας γνωρίζει την κατάσταση του καναλιού μόνο μετά την επιλογή του καναλιού, το πρόβλημα απόφασης βελτιστοποίησης του αισθητήρα μπορεί να διατυπωθεί ως Partially Observable Markov Decision Process (POMDP). Συγκεκριμένα, ο αισθητήρας πρέπει να επιλέγει ένα από τα κανάλια M. Ο αισθητήρας λαμβάνει μια θετική ανταμοιβή "+1" εάν το επιλεγμένο κανάλι έχει χαμηλή παρεμβολή και μια αρνητική ανταμοιβή "−1" διαφορετικά. Ο στόχος είναι να βρούμε μια βέλτιστη πολιτική που μεγιστοποιεί την αναμενόμενη συσσωρευμένη με την πάροδο του χρόνου μειωμένη ανταμοιβή του αισθητήρα. Στην πραγματικότητα, ο στόχος μπορεί να επιτευχθεί με τη μυωπική πολιτική (Zhao, Krishnamachari, & Liu, 2008). Ωστόσο, η μυωπική πολιτική απαιτεί την προηγούμενη γνώση του πίνακα μετάβασης συστήματος που είναι δύσκολο να ληφθεί. Η DQL επιτρέπει στον αισθητήρα να βρει τη βέλτιστη πολιτική από τις εμπειρίες του και έτσι μπορεί να υιοθετηθεί για την επίλυση του προβλήματος. Συγκεκριμένα, η DQL χρησιμοποιεί ένα Deep Q-Network (DQN) με επανάληψη εμπειρίας (Mnih, et al., Playing Atari with Deep Reinforcement Learning, 2013). Η είσοδος του DQN είναι μια κατάσταση

του αισθητήρα που είναι ο συνδυασμός ενεργειών και παρατηρήσεων, δηλαδή των ανταμοιβών, στις παρελθούσες χρονοθυρίδες. Η έξοδος περιλαμβάνει τιμές Q που αντιστοιχούν στις ενέργειες του αισθητήρα. Για την εξισορρόπηση της εξερεύνησης της τρέχουσας καλύτερης τιμής Q με την εξερεύνηση της καλύτερης, υιοθετείται για τον μηχανισμό επιλογής ενέργειας η πολιτική ε-greedy. Τα αποτελέσματα της προσομοίωσης με βάση τα πραγματικά δεδομένα (Govindan) δείχνουν ότι το προτεινόμενο σχήμα επιτυγχάνει μια μέση επιβράβευση 4,4 που πλησιάζει τη μυωπική πολιτική (Zhao, Krishnamachari, & Liu, 2008) με μέση ανταμοιβή 4,5. Σημειώνεται ότι, η μυωπική πολιτική απαιτεί τη γνώση του πίνακα μετάβασης συστήματος.

Η αναφορά (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access, 2017) μπορεί να θεωρηθεί πρωτοποριακή εργασία χρησιμοποιώντας το DQL για την πρόσβαση στο κανάλι. Ωστόσο, η DQL συνεχίζει να ακολουθεί την εκμαθημένη πολιτική για τις χρονοθυρίδες και σταματά να μαθαίνει μια κατάλληλη πολιτική. Τα πραγματικά περιβάλλοντα IoT είναι δυναμικά και το DQN στο DQL πρέπει να επανεκπαιδευτεί. Ένα προσαρμοστικό σχήμα DQL προτείνεται στο (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks, 2018) το οποίο αξιολογεί τη συσσωρευμένη ανταμοιβή της τρέχουσας πολιτικής για κάθε περίοδο. Όταν η επιβράβευση μειώνεται κατά ένα δεδομένο όριο, το DQN εκπαιδεύεται εκ νέου για να βρει μια νέα καλή πολιτική. Τα αποτελέσματα της προσομοίωσης (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks, 2018) δείχνουν ότι όταν αλλάζουν οι καταστάσεις των καναλιών, το προσαρμοστικό σχήμα DQL μπορεί να ανιχνεύσει την αλλαγή και να ξεκινήσει εκ νέου μάθηση για να λάβει υψηλή ανταμοιβή.

Τα μοντέλα στα (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access, 2017) και (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks, 2018) περιορίζονται σε έναν μόνο αισθητήρα. Στην περίπτωση ενός σεναρίου πολλαπλών αισθητήρων (Zhu, Song, Jiang, & Song, 2018), η κοινή επιλογή καναλιού και η προώθηση πακέτων πραγματοποιείται με τη χρήση της DQL. Στο μοντέλο ένας αισθητήρας ως αναμεταδότης προωθεί πακέτα που λαμβάνονται από τους γειτονικούς του αισθητήρες

στη δεξαμενή. Ο αισθητήρας είναι εξοπλισμένος με μία ενδιάμεση μνήμη (buffer) για την αποθήκευση των πακέτων που λαμβάνονται. Σε κάθε χρονοθυρίδα, ο αισθητήρας επιλέγει ένα σύνολο καναλιών για την προώθηση πακέτων, έτσι ώστε να μεγιστοποιήσει τη χρησιμότητά του, δηλαδή την αναλογία του αριθμού των μεταδιδόμενων πακέτων προς την ισχύ μετάδοσης. Όπως και στο (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access, 2017), το πρόβλημα του αισθητήρα μπορεί να διατυπωθεί ως Markov Decision Process (MDP). Η ενέργεια είναι να επιλεγεί ένα σύνολο καναλιών, ο αριθμός των πακέτων που μεταδίδονται στα κανάλια και ένας τρόπο διαμόρφωσης. Για να αποφευχθεί απώλεια πακέτων, η κατάσταση ορίζεται ως ο συνδυασμός της κατάστασης της ενδιάμεσης μνήμης και της κατάστασης καναλιού. Στη συνέχεια, το MDP επιλύεται από την DQL στην οποία η είσοδος είναι η κατάσταση και η έξοδος είναι η επιλογή ενέργειας. Το DQL χρησιμοποιεί τον ενωμένο αυτόματο κωδικοποιητή για να μειώσει τον τεράστιο υπολογισμό και την αποθήκευση στη φάση Q-learning. Η λειτουργία χρησιμότητας του αισθητήρα αποδεικνύεται περιορισμένη, γεγονός το οποίο μπορεί να εγγυηθεί τη σύγκλιση του αλγορίθμου. Η ανάλυση δείχνει ότι η υπολογιστική πολυπλοκότητα του προτεινόμενου αλγορίθμου είναι O (KM (J + 1)), που είναι χαμηλότερη από εκείνη του αλγορίθμου στρατηγικής προσέγγισης (Fearnley, 2010) με την υπολογιστική πολυπλοκότητα του O (KM (J +1)) (L +1) KC), όπου το K είναι ο αριθμός των buffer, το L είναι το μήκος του buffer, τα M και C αντιστοίχως είναι οι αριθμοί των καναλιών και των καταστάσεων καναλιών και το J είναι ο αριθμός των πιθανών τρόπων μετάδοσης. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο σχήμα βελτιώνει σημαντικά τη χρησιμότητα του συστήματος σε σύγκριση με το σχήμα επιλογής τυχαίας δράσης. Συγκεκριμένα, η μέση χρησιμότητα συστήματος του προτεινόμενου σχήματος είναι 0,63, ενώ αυτή που προκύπτει από την τυχαία πολιτική είναι 0,37. Ωστόσο, καθώς ο ρυθμός άφιξης των πακέτων αυξάνεται, η χρησιμότητα του συστήματος του προτεινόμενου σχήματος μειώνεται καθώς ο αισθητήρας χρειάζεται να καταναλώνει περισσότερη ισχύ για τη μετάδοση όλων των πακέτων.

Η κατανάλωση περισσότερης ισχύος οδηγεί σε κακή απόδοση του αισθητήρα λόγω του ενεργειακού του περιορισμού, δηλαδή μικρότερη διάρκεια ζωής του συστήματος IoT. Στο (Chu, Li, Liao, & Cui, 2019) διερευνάται το πρόβλημα πρόσβασης στα κανάλια στο σύστημα IoT με δυνατότητα συλλογής ενέργειας. Το μοντέλο αποτελείται από ένα BS

(πράκτορας) και αισθητήρες ενεργειακής συλλογής. Το BS ως ελεγκτής εκχωρεί κανάλια στους αισθητήρες. Ωστόσο, η αβεβαιότητα της διαθεσιμότητας ενέργειας του περιβάλλοντος στους αισθητήρες μπορεί να κάνει ανεπαρκή την κατανομή καναλιών. Για παράδειγμα, το κανάλι που εκχωρείται στον αισθητήρα με χαμηλή διαθέσιμη ενέργεια ενδέχεται να μην χρησιμοποιηθεί πλήρως, καθώς ο αισθητήρας δεν μπορεί να επικοινωνήσει αργότερα.

Επομένως, το πρόβλημα του BS είναι να προβλέπεται η κατάσταση μπαταρίας των αισθητήρων και να επιλέγονται οι κατάλληλοι αισθητήρες για την πρόσβαση στο κανάλι έτσι ώστε να μεγιστοποιείται η συνολική ταχύτητα. Για την επίλυση του προβλήματος του BS, μπορούν να υιοθετηθούν οι βέλτιστες προσεγγίσεις όπως το σύστημα κατανομής πόρων ανερχόμενη ζεύξης (Di, Xiong, Fan, Yang, & Letaief, 2017). Ωστόσο, το μοντέλο απαιτεί το BS να έχει τέλεια γνώση όλων των τυχαίων διαδικασιών. Η τέλεια γνώση ενδέχεται να μην είναι διαθέσιμη αφού οι αισθητήρες κατανέμονται τυχαία σε μια γεωγραφική περιοχή. Έτσι, η DQL χρησιμοποιείται για την επίλυση του προβλήματος του BS, δηλαδή του πράκτορα. Η DQL χρησιμοποιεί ένα DQN που αποτελείται από δύο επίπεδα νευρωνικού δικτύου που βασίζονται σε Long Short-Term Memory (LSTM). Το πρώτο επίπεδο παράγει τις προβλεπόμενες καταστάσεις μπαταρίας των αισθητήρων και το δεύτερο επίπεδο καθορίζει την πολιτική πρόσβασης καναλιών χρησιμοποιώντας τις προβλεπόμενες καταστάσεις μαζί με τις πληροφορίες κατάστασης καναλιού (Channel State Information - CSI). Ο χώρος της κατάστασης αποτελείται από (i) το ιστορικό προγραμματισμού πρόσβασης καναλιού, (ii) το ιστορικό των προβλεπόμενων πληροφοριών μπαταρίας, (iii) το ιστορικό των πραγματικών πληροφοριών μπαταρίας και (iv) το τρέχον CSI των αισθητήρων. Ο χώρος της ενέργειας περιέχει όλα τα σύνολα αισθητήρων που θα επιλεγούν για την πρόσβαση στο κανάλι και η ανταμοιβή είναι η διαφορά μεταξύ του συνολικού ρυθμού και του σφάλματος πρόβλεψης. Όπως φαίνεται στα αποτελέσματα της προσομοίωσης, το προτεινόμενο μοντέλο πλησιάζει τη βέλτιστη προσέγγιση (Di, Xiong, Fan, Yang, & Letaief, 2017) και ξεπερνά τη μυωπική πολιτική (Zhao, Krishnamachari, & Liu, 2008) ως προς το συνολικό ποσοστό. Συγκεκριμένα, τα συνολικά ποσοστά που λαμβάνονται από το προτεινόμενο μοντέλο, τη μυωπική πολιτική και τη βέλτιστη προσέγγιση είναι 6,8, 6,5 και 7,0 kbps, αντίστοιχα. Επιπλέον, το σφάλμα πρόβλεψης μπαταρίας που προέκυψε από το προτεινόμενο μοντέλο είναι σχεδόν μηδέν.

Τα παραπάνω σχήματα, π.χ. (Wang, Liu, Gomes, & Krishnamachari, Deep Reinforcement Learning for Dynamic Multichannel Access, 2017) και (Chu, Li, Liao, & Cui, 2019), επικεντρώνονται στη μεγιστοποίηση του ρυθμού. Σε συστήματα IoT όπως επικοινωνίες Vehicle-to-Vehicle (V2V), ο χρόνος αναμονής-αδράνειας πρέπει επίσης να ληφθεί υπόψη λόγω της κινητικότητας των πομπών / δεκτών V2V και των ζωτικών εφαρμογών στην ασφάλεια της κυκλοφορίας. Ένα από τα προβλήματα κάθε πομπού V2V είναι να επιλέγεται ένα κανάλι και ένα επίπεδο ισχύος μετάδοσης για να μεγιστοποιείται η χωρητικότητά του υπό περιορισμό καθυστέρησης. Δεδομένου του αποκεντρωμένου δικτύου, υιοθετείται ένα DQN για τη λήψη βέλτιστων αποφάσεων όπως προτείνεται στο (Ye & Li, 2018). Το μοντέλο αποτελείται από πομπούς V2V, δηλαδή πράκτορες, οι οποίοι μοιράζονται ένα σύνολο καναλιών. Οι ενέργειες κάθε πομπού V2V περιλαμβάνουν επιλογή καναλιών και επίπεδα ισχύος μετάδοσης. Η ανταμοιβή είναι συνάρτηση της χωρητικότητας και της καθυστέρησης του πομπού V2V. Η κατάσταση που παρατηρείται από τον πομπό V2V αποτελείται από (i) το στιγμιαίο CSI του αντίστοιχου συνδέσμου V2V, (ii) την παρεμβολή στον σύνδεσμο V2V στην προηγούμενη χρονοθυρίδα, (iii) τα κανάλια που επιλέγονται από τους γείτονες του πομπού V2V στη προηγούμενη χρονοθυρίδα, και (iv) τον εναπομείναντα χρόνο για να ανταποκριθεί στον περιορισμό καθυστέρησης. Η κατάσταση είναι επίσης ένα δεδομένο που εισάγεται στο DQN. Η έξοδος περιλαμβάνει τιμές Q που αντιστοιχούν στις ενέργειες. Όπως φαίνεται στα αποτελέσματα της προσομοίωσης, προσαρμόζοντας δυναμικά την ισχύ και την επιλογή καναλιού όταν οι συνδέσεις V2V είναι πιθανό να παραβιάζουν τον περιορισμό καθυστέρησης, το προτεινόμενο σχήμα έχει περισσότερους πομπούς V2V που πληρούν τον περιορισμό καθυστέρησης σε σύγκριση με την τυχαία κατανομή καναλιών.

Για τη μείωση του κόστους φάσματος, τα παραπάνω συστήματα IoT χρησιμοποιούν συχνά κανάλια χωρίς άδεια. Ωστόσο, αυτό μπορεί να προκαλέσει παρεμβολές σε υπάρχοντα δίκτυα, π.χ. WLAN. Για τη λύση του προβλήματος προτάθηκε (Challita, Dong, & Saad, 2017) η χρήση του DQN για να αντιμετωπιστούν από κοινού η δυναμική πρόσβαση καναλιών και η διαχείριση παρεμβολών. Το μοντέλο αποτελείται από μικρούς σταθμούς βάσης (Small Base Stations - SBS) που μοιράζονται κανάλια χωρίς άδεια σε ένα δίκτυο LTE. Σε κάθε χρονοθυρίδα, το SBS επιλέγει ένα από τα κανάλια για τη μετάδοση του πακέτου του. Ωστόσο, ενδέχεται στο επιλεγμένο κανάλι να υπάρχει κυκλοφορία WLAN, με αποτέλεσμα το SBS να έχει πρόσβαση στο επιλεγμένο κανάλι με πιθανότητα. Οι

ενέργειες του SBS περιλαμβάνουν ζεύγη επιλογής καναλιών και πιθανότητα πρόσβασης καναλιού. Το πρόβλημα του SBS είναι να προσδιοριστεί ένας παράγοντας δράσης έτσι ώστε να μεγιστοποιηθεί η συνολική του απόδοση, δηλαδή, η χρηστικότητά του, σε όλα τα κανάλια και τις χρονοθυρίδες. Το πρόβλημα κατανομής πόρων μπορεί να διατυπωθεί ως μη συνεργατικό παιχνίδι και το DQN χρησιμοποιώντας LSTM μπορεί να υιοθετηθεί για την επίλυση του παιχνιδιού. Η είσοδος του DQN είναι το ιστορικό της κυκλοφορίας των SBS και του WLAN στα κανάλια. Η έξοδος περιλαμβάνει προβλεπόμενους παράγοντες δράσης των SBS. Η λειτουργία χρησιμότητας κάθε SBS αποδεικνύεται κυρτή, και έτσι ο αλγόριθμος που βασίζεται στο DQN συγκλίνει σε μια ισορροπία Nash του παιχνιδιού. Η ανάλυση δείχνει ότι η υπολογιστική πολυπλοκότητα ανά χρονικό βήμα του προτεινόμενου σχήματος είναι $O(n^2c + n_c n_i + n_c n_0 + n_c)$, όπου n_c, n_i και n_o είναι οι αριθμοί των κυψελών μνήμης, των μονάδων εισόδου και των μονάδων εξόδου, αντίστοιχα. Τα αποτελέσματα προσομοίωσης που βασίζονται σε πραγματικά δεδομένα κίνησης (Balazinska & Castro, 2003) δείχνουν ότι το προτεινόμενο μοντέλο μπορεί να βελτιώσει τη μέση απόδοση έως και 28% σε σύγκριση με την τυπική Q-learning. Επιπλέον, η ανάπτυξη περισσότερων SBS στο δίκτυο LTE δεν επιτρέπει περισσότερο μέρος χρόνου μετάδοσης για το δίκτυο. Αυτό συνεπάγεται ότι το προτεινόμενο μοντέλο μπορεί να αποφύγει την υποβάθμιση της απόδοσης του WLAN ,αλλά απαιτεί συγχρονισμό μεταξύ των SBS και του WLAN που είναι δύσκολο σε πραγματικά δίκτυα.

Στο ίδιο πλαίσιο δικτύου κινητής τηλεφωνίας, το πρόβλημα πρόσβασης δυναμικού φάσματος για πολλούς χρήστες που μοιράζονται κανάλια K διερευνάται στο (Naparstek & Cohen, 2017). Σε μια χρονοθυρίδα, ο χρήστης επιλέγει ένα κανάλι με συγκεκριμένη πιθανότητα προσπάθειας ή επιλέγει να μην μεταδώσει καθόλου. Η κατάσταση είναι το ιστορικό των ενεργειών του χρήστη και των τοπικών παρατηρήσεών του και η στρατηγική του χρήστη σχεδιάζεται από το ιστορικό στη πιθανότητα απόπειρας. Το πρόβλημα του χρήστη είναι η εύρεση του παράγοντα των στρατηγικών, δηλαδή της πολιτικής, στις χρονοθυρίδες για τη μεγιστοποίηση του αναμενόμενου συσσωρευμένα μειωμένου ρυθμού δεδομένων του χρήστη.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Wang, Liu, Gomes, & Krishnamachari, 2017)	POMDP	DQN με χρήση FNN	Αισθητήρας	Προηγούμενες επιλογές καναλιών και παρατηρήσεις	Επιλογή καναλιού	Αποτέλεσμα +1 ή -1	IoT

(Zhu, Song, Jiang, & Song, 2018)	MDP	DQN με χρήση FNN	Αισθητήρας	Τρέχουσα κατάσταση ενδιάμεσης μνήμης και κατάσταση καναλιού	Επιλογή καναλιού, πακέτων και τρόπου διαμόρφωσης	Αναλογία αριθμού μεταδιδόμενων πακέτων προς ισχύ μετάδοσης	IoT
(Chu, Li, Liao, & Cui, 2019)	MDP	DQN με LSTM	Σταθμός Βάσης	Ιστορικό πρόσβασης καναλιού, προβλεπόμενο και πραγματικό ιστορικό πληροφοριών μπαταρίας και τρέχον CSI	Επιλογή αισθητήρα για πρόσβαση στο κανάλι	Συνολικό σφάλμα και σφάλμα πρόβλεψης	IoT
(Ye & Li, 2018)	MDP	DQN με LSTM	Πομπός οχήματος σε όχημα	Τρέχον CSI, παρελθούσες παρεμβολές, προηγούμενες επιλογές καναλιού και υπόλοιπος χρόνος για την κάλυψη των περιορισμών καθυστέρησης	Επιλογή καναλιού και ισχύος εκπομπής	Χωρητικότητα και καθυστέρηση	IoT
(Challita, Dong, & Saad, 2017)	Παίγνιο	DQN με LSTM	Μικρός Σταθμός Βάσης	Ιστορικό κυκλοφορίας μικρών σταθμών βάσης και WLAN	Επιλογή καναλιού και πιθανότητα πρόσβασης στο κανάλι	Απόδοση (Throughput)	Δίκτυο LTE
(Naparstek & Cohen, 2017)	Παίγνιο	DDQN με Dueling DQN	Κινητός Χρήστης	Προηγούμενες επιλογές καναλιών και παρατηρήσεις	Επιλογή καναλιού	Ρυθμός δεδομένων	CRN
(Liu, Hu, & Wang, 2018)	MDP	DQN με CNN	Δορυφορικό Σύστημα	Τρέχοντα τερματικά χρήστη, πίνακας κατανομής καναλιών και νεοαφιχθέν χρήστης	Επιλογή καναλιού	Αποτέλεσμα +1 ή -1	Δορυφορικό Σύστημα
(Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018)	MDP	DDQN με Dueling DQN	Κινητός Χρήστης	Καταστάσεις QoS	Επιλογή καναλιού και σταθμού βάσης	Χρησιμότητα	HetNet
(Chen, Saad, & Yin, 2017)	Παίγνιο	DQN με LSM	UAV	Κατανομή αιτήματος περιεχομένου	Επιλογή σταθμού βάσης	Χρήστες με σταθερές ουρές	Δίκτυο LTE

Πίνακας 3. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για πρόσβαση στο δίκτυο

Το παραπάνω πρόβλημα επιλύεται με την εκπαίδευση ενός DQN. Η εισαγωγή του DQN περιλαμβάνει προηγούμενες ενέργειες και τις αντίστοιχες παρατηρήσεις Η έξοδος περιλαμβάνει εκτιμώμενες τιμές Q των ενεργειών. Για να αποφευχθεί ο υπερυπολογισμός στην Q-learning, χρησιμοποιείται η Double Deep Q-Network (DDQN) (Hasselt, Double Q-learning, 2010). Επιπλέον, η dueling DQN (Wang, Schaul, Hessel, Hasselt, Lanctot, & Freitas, 2016) χρησιμοποιείται για τη βελτίωση της εκτιμώμενης τιμής Q. Στη συνέχεια, το DQN εκπαιδεύεται εκτός σύνδεσης σε σταθμό βάσης. Παρόμοια με το (Challita, Dong, & Saad, 2017), η τυχαία πρόσβαση πολλαπλών καναλιών διαμορφώνεται ως μη συνεργατικό παιχνίδι. Όπως αποδείχθηκε στο (Naparstek & Cohen, 2017), το παιχνίδι πληροί τα κριτήρια του subgame perfect Nash equilibrium. Ορισμένοι χρήστες μπορούν να συνεχίσουν να αυξάνουν την πιθανότητα προσπάθειάς τους για να αυξήσουν το ρυθμό

τους. Αυτό καθιστά το σημείο ισορροπίας ανεπαρκές, και έτσι ο χώρος στρατηγικής των χρηστών περιορίζεται για να αποφευχθεί η κατάσταση. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο μοντέλο μπορεί να επιτύχει διπλάσια απόδοση καναλιού σε σύγκριση με το slotted-Aloha (Li H. , 2010). Ο λόγος είναι ότι στο προτεινόμενο μοντέλο, κάθε χρήστης μαθαίνει μόνο από την τοπική του παρατήρηση χωρίς διαδικτυακό συντονισμό ή ανίχνευση φορέα. Ωστόσο, το προτεινόμενο σχέδιο απαιτεί την κεντρική μονάδα που μπορεί να αυξήσει την ανταλλαγή μηνυμάτων καθώς η εκπαίδευση ενημερώνεται συχνά.

Στα προαναφερθέντα μοντέλα, ο αριθμός των χρηστών είναι σταθερός σε όλες τις χρονοθυρίδες και δεν λαμβάνεται υπόψη η άφιξη νέων χρηστών. Στο (Liu, Hu, & Wang, 2018) μελετάται η κατανομή καναλιών σε νεοαφιχθέντες χρήστες σε ένα δορυφορικό σύστημα με πολλαπλές δέσμες ακτινοβολίας. Το δορυφορικό σύστημα με πολλαπλές δέσμες ακτινοβολίας δημιουργεί ένα γεωγραφικό αποτύπωμα υποδιαιρούμενο σε πολλαπλές δέσμες που παρέχουν υπηρεσίες σε επίγεια τερματικά (User Terminals - UTs). Το σύστημα διαθέτει ένα σύνολο καναλιών τα οποία εκχωρούνται, εάν υπάρχει διαθεσιμότητα στους νέους UTs, οπότε και η υπηρεσία-απαίτηση ικανοποιείται. Διαφορετικά, η υπηρεσία είναι αποκλεισμένη. Το πρόβλημα του συστήματος είναι να κατανέμει κατάλληλα τα κανάλια για να ελαχιστοποιήσει τη συνολική πιθανότητα αποκλεισμού υπηρεσίας στους νέους UTs χωρίς να προκαλέσει παρεμβολές στα υφιστάμενα UTs.

Το ανωτέρω πρόβλημα του συστήματος μπορεί να θεωρηθεί ως ένα χρονικά συσχετισμένο διαδοχικό πρόβλημα βελτιστοποίησης λήψης αποφάσεων που επιλύεται αποτελεσματικά από το DQN. Εδώ, το δορυφορικό σύστημα είναι ο πράκτορας. Η ενέργεια είναι ένας κατάλογος που δείχνει ποιο κανάλι κατανέμεται στο νεοαφιχθέν UT. Η ανταμοιβή είναι θετική όταν ικανοποιείται η νέα υπηρεσία και είναι αρνητική όταν η υπηρεσία είναι αποκλεισμένη. Η κατάσταση περιλαμβάνει το σύνολο των υφισταμένων UT, τον τρέχοντα πίνακα κατανομής καναλιών και το νέο UT που έφτασε. Σημειώνεται ότι, η κατάσταση έχει τη δυνατότητα χωρικής συσχέτισης λόγω της συνκαναλικής παρεμβολής, και ως εκ τούτου μπορεί να αναπαρασταθεί με τρόπο που μοιάζει με εικόνα, δηλαδή, μια πολυδιαστατική διάταξη (tensor) εικόνας. Επομένως, το DQN υιοθετεί το Convolutional Neural Network (CNN) για να εξαγάγει χρήσιμα χαρακτηριστικά της

κατάστασης. Τα αποτελέσματα προσομοίωσης δείχνουν ότι ο προτεινόμενος αλγόριθμος DQN συγκλίνει μετά από έναν ορισμένο αριθμό βημάτων εκμάθησης. Επίσης, με την κατανομή των διαθέσιμων καναλιών στους νέους UT, το προτεινόμενο μοντέλο μπορεί να βελτιώσει την κυκλοφορία του συστήματος έως και 24,4% σε σύγκριση με το σταθερό σύστημα κατανομής καναλιών. Ωστόσο, καθώς ο αριθμός των υφισταμένων UT αυξάνεται, ο αριθμός των διαθέσιμων καναλιών είναι χαμηλός ή ακόμη και μηδέν. Επομένως, οι αποφάσεις δυναμικής κατανομής καναλιών του προτεινόμενου σχήματος καθίστανται άνευ σημασίας και η διαφορά απόδοσης μεταξύ των δύο συστημάτων καθίσταται ασήμαντη.

3.1.2 Κοινή Συσχέτιση Χρηστών και Πρόσβαση στο Φάσμα (Joint User Association and Spectrum Access)

Τα κοινά προβλήματα συσχέτισης χρηστών και πρόσβασης στο φάσμα είναι συνήθως μη κυρτά. Για την επίλυση των προβλημάτων και την επίτευξη της βέλτιστης λύσης, αναπτύσσονται παραδοσιακές προσεγγίσεις όπως ο γραμμικός προγραμματισμός (Elsherif, Chen, Ito, & Ding, 2015). Ωστόσο, οι προσεγγίσεις απαιτούν σχεδόν ολοκληρωμένες και ακριβείς πληροφορίες δικτύου που συνήθως δεν είναι διαθέσιμες. Μπορούν να χρησιμοποιηθούν τεχνικές μάθησης όπως η Q-learning, αλλά είναι δύσκολο να επιτευχθεί μια βέλτιστη λύση λόγω των μεγάλων χώρων κατάστασης και δράσης των προβλημάτων βελτιστοποίησης των συνδέσμων. Συνδυάζοντας το βαθύ νευρωνικό δίκτυο (Deep Neural Network – DNN) με το Q-learning, το DQL χρησιμοποιείται αποτελεσματικά για την επίλυση των προβλημάτων κοινής βελτιστοποίησης όπως προτείνονται στα (Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018) και (Chen, Saad, & Yin, Liquid State Machine Learning for Resource Allocation in a Network of Cache-Enabled LTE-U UAVs, 2017).

Στη συνέχεια μελετάται η περίπτωση ενός ετερογενούς δικτύου (HetNet) το οποίο αποτελείται από πολλούς χρήστες και σταθμούς βάσης (BSs), συμπεριλαμβανομένων των σταθμών βάσης macro και femto (Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018). Τα BS μοιράζονται ένα σύνολο ορθογώνιων καναλιών, ενώ οι χρήστες βρίσκονται τυχαία στο δίκτυο. Το πρόβλημα κάθε χρήστη είναι να επιλέξει ένα BS και ένα κανάλι για να μεγιστοποιήσει το ρυθμό δεδομένων του, ενώ παράλληλα θα εξασφαλίζεται ότι ο λόγος σήματος προς παρεμβολές-συν-θορύβου (signal-to-interference-plus-noise ratio - SINR) του χρήστη είναι υψηλότερος από την ελάχιστη απαίτηση ποιότητας υπηρεσίας (Quality

of Service - QoS) . Η DQL χρησιμοποιείται για την επίλυση του προβλήματος στο οποίο κάθε χρήστης είναι πράκτορας και η κατάστασή του είναι ένας παράγοντας που περιλαμβάνει τις καταστάσεις QoS όλων των χρηστών, δηλαδή τη συνολική κατάσταση. Εδώ, η κατάσταση QoS του χρήστη αναφέρεται στο εάν το SINR του υπερβαίνει την ελάχιστη απαίτηση QoS ή όχι. Σε κάθε χρονικό διάστημα, ο χρήστης κάνει μια ενέργεια, εάν ικανοποιείται το QoS, ο χρήστης λαμβάνει την υπηρεσία ως άμεση ανταμοιβή του. Διαφορετικά, λαμβάνει μια αρνητική επιβράβευση, δηλαδή ένα κόστος επιλογής δράσης. Λαμβάνεται υπόψη ότι η αθροιστική επιβράβευση ενός χρήστη εξαρτάται από ενέργειες άλλων χρηστών και τότε το πρόβλημα του χρήστη μπορεί να οριστεί ως MDP. Παρόμοια με το (Naparstek & Cohen, 2017), το DDQN και το Dueling DQN χρησιμοποιούνται για να βρουν τη βέλτιστη πολιτική, δηλαδή τις κοινές επιλογές BS και καναλιών, ώστε ο χρήστης να μεγιστοποιήσει τη σωρευτική ανταμοιβή του. Τα αποτελέσματα της προσομοίωσης (Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018) δείχνουν ότι το προτεινόμενο σχήμα ξεπερνά την Q-learning που υλοποιήθηκε στο (Watkins & Dayan, 1992) όσον αφορά την ταχύτητα σύγκλισης και τη χωρητικότητα του συστήματος. Οι συγκρίσεις προσομοίωσης αποδεικνύουν ότι το DQN μπορεί να χρησιμοποιηθεί αποτελεσματικά για την επίλυση των πολύπλοκων προβλημάτων, όπως προβλήματα συνεργατικής βελτιστοποίησης, σε συστήματα μεγάλης κλίμακας όπως τα HetNets και IoT.

Η πρόταση (Zhao, Liang, Niyato, Pei, Wu, & Jiang, 2018) θεωρείται ότι είναι η πρώτη εργασία που χρησιμοποιεί το DQL για την επίλυση του προβλήματος κοινής συσχέτισης χρηστών και πρόσβασης στο φάσμα. Σε συνέχεια της ανωτέρω εργασίας γίνεται η πρόταση (Chen, Saad, & Yin, Liquid State Machine Learning for Resource Allocation in a Network of Cache-Enabled LTE-U UAVs, 2017) χρήσης του DQL για επίλυση του προβλήματος κοινής συσχέτισης χρηστών, πρόσβασης φάσματος και προσωρινής αποθήκευσης περιεχομένου. Το μοντέλο δικτύου είναι ένα δίκτυο LTE που αποτελείται από UAV που εξυπηρετούν χρήστες εδάφους. Τα UAV είναι εξοπλισμένα με μονάδες αποθήκευσης και μπορούν να λειτουργήσουν ως LTE-BS με δυνατότητα προσωρινής αποθήκευσης. Τα UAV έχουν πρόσβαση σε ζώνες δικτύου με και χωρίς άδεια. Τα UAV ελέγχονται από ένα διακομιστή βασιζόμενο σε cloud και οι μεταδόσεις από το cloud στα UAV πραγματοποιούνται χρησιμοποιώντας την άδεια ζώνης συχνοτήτων κινητής τηλεφωνίας. Το πρόβλημα του κάθε UAV είναι να προσδιορίσει (i) τη βέλτιστη συσχέτιση χρηστών, (ii) τους δείκτες κατανομής εύρους ζώνης στη ζώνη συχνοτήτων με άδεια, (iii)

τους δείκτες χρονοθυρίδων στη ζώνη συχνοτήτων χωρίς άδεια και (iv) ένα σύνολο δημοφιλών περιεχομένων που οι χρήστες μπορούν να ζητήσουν για να μεγιστοποιήσουν τον αριθμό των χρηστών με σταθερή ουρά, δηλαδή, χρήστες που ικανοποιούνται με καθυστέρηση μετάδοσης περιεχομένου.

Το πρόβλημα του UAV είναι συνδυαστικό και μη κυρτό και το DQL μπορεί να χρησιμοποιηθεί για την επίλυσή του. Τα UAV δεν γνωρίζουν τα αιτήματα περιεχομένου των χρηστών, και έτσι η προσέγγιση Liquid State Machine (LSM) (Maass, 2011) υιοθετείται για την πρόβλεψη της διανομής αιτημάτων περιεχομένου των χρηστών και για την εκτέλεση κατανομής πόρων. Συγκεκριμένα, η πρόβλεψη της διανομής αιτημάτων περιεχομένου εφαρμόζεται στο cloud με βάση έναν αλγόριθμο πρόβλεψης που βασίζεται σε LSM. Στη συνέχεια, λαμβάνοντας υπόψη τις διανομές αιτήσεων, κάθε UAV ως πράκτορας χρησιμοποιεί έναν αλγόριθμο εκμάθησης που βασίζεται σε LSM για να βρει τη βέλτιστη συσχέτιση χρηστών. Συγκεκριμένα, η είσοδος του αλγορίθμου εκμάθησης που βασίζεται σε LSM αποτελείται από ενέργειες, δηλαδή σχήματα συσχέτισης χρηστών UAV, που λαμβάνουν άλλα UAV και η έξοδος περιλαμβάνει τον αναμενόμενο αριθμό χρηστών με σταθερές ουρές που αντιστοιχούν σε ενέργειες που μπορεί να πραγματοποιήσει το UAV. Αφού ολοκληρωθεί η συσχέτιση χρήστη, η βέλτιστη αποθήκευση περιεχομένου προσωρινής αποθήκευσης καθορίζεται με βάση τα αποτελέσματα του (Chen, Mozaffari, Saad, Yin, Debbah, & Hong, 2017) και η βέλτιστη κατανομή φάσματος γίνεται με γραμμικό προγραμματισμό. Με βάση το Θεώρημα του Gordon (Szita, Gyenes, & Lorincz, 2006), η προτεινόμενη DQL αποδεικνύεται ότι συγκλίνει με πιθανότητα ένα. Τα αποτελέσματα της προσομοίωσης χρησιμοποιώντας τα δεδομένα αιτήματος περιεχομένου (Tyouku of China Network Video Index) δείχνουν ότι η προτεινόμενη DQL μπορεί να συγκλίνει εντός 400 επαναλήψεων. Σε σύγκριση με το Q-learning, το προτεινόμενο DQN βελτιώνει το χρόνο σύγκλισης έως και 33%. Επιπλέον, η προτεινόμενη DQL βελτιώνει σημαντικά τον αριθμό των χρηστών με σταθερές ουρές έως και 50% σε σύγκριση με το Q-learning χωρίς μνήμη cache. Στην πραγματικότητα, η ενεργειακή απόδοση είναι επίσης σημαντική για τα UAV, και συνεπώς η εφαρμογή του DQL για κοινή συσχέτιση χρηστών, πρόσβαση του φάσματος και πρόβλημα κατανομής ισχύος πρέπει να διερευνηθεί.

Στον Πίνακα 3 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για πρόσβαση στο δίκτυο.

3.2 Προσαρμοστικός Έλεγχος Ρυθμού Δεδομένων

Η δυναμική προσαρμοστική ροή μέσω HTTP (DASH) γίνεται το κυρίαρχο πρότυπο για τη ροή βίντεο (Stockhammer, 2011). Το DASH είναι σε θέση να αξιοποιήσει την υπάρχουσα υποδομή δικτύου παράδοσης περιεχομένου και είναι συμβατό με ένα πλήθος εφαρμογών από την πλευρά του πελάτη. Σε ένα γενικό σύστημα DASH τα βίντεο αποθηκεύονται σε διακομιστές ως πολλαπλά τμήματα, δηλαδή κομμάτια. Κάθε τμήμα κωδικοποιείται σε διαφορετικά επίπεδα συμπίεσης για να δημιουργεί αναπαραστάσεις με διαφορετικούς ρυθμούς bit, δηλαδή διαφορετική ποιότητα εικόνας. Σε κάθε χρονικό διάστημα, ο πελάτης επιλέγει μια αναπαράσταση, δηλαδή ένα τμήμα με συγκεκριμένο ρυθμό bit, για λήψη. Το πρόβλημα του πελάτη είναι να βρει μια βέλτιστη πολιτική που μεγιστοποιεί τη ποιότητα εμπειρίας (Quality of Experience – QoE) του, όπως η μεγιστοποίηση του μέσου ρυθμού bit και η ελαχιστοποίηση του rebuffering, δηλαδή, του χρόνου που παγώνει η αναπαραγωγή βίντεο.

Όπως παρουσιάζεται στο (Gadaleta, Chiarriotti, Rossi, & Zanella, 2017), το παραπάνω πρόβλημα μπορεί να μοντελοποιηθεί ως MDP όπου ο πράκτορας είναι ο πελάτης και η ενέργεια επιλέγει μια αναπαράσταση για λήψη. Για τη μεγιστοποίηση του QoE, η ανταμοιβή ορίζεται ως συνάρτηση της (i) οπτικής ποιότητας του βίντεο, (ii) της σταθερότητας της ποιότητας του βίντεο, (iii) του συμβάντος rebuffering και (iv) της κατάστασης ενδιάμεση μνήμης. Λαμβάνοντας υπόψη τη διαμόρφωση ανταμοιβής, η κατάσταση του πελάτη θα πρέπει να περιλαμβάνει (i) την ποιότητα βίντεο του τελευταίου τμήματος λήψης, (ii) την τρέχουσα κατάσταση προσωρινής αποθήκευσης, (iii) τον χρόνο rebuffering και (iv) τις χωρητικότητες καναλιού που παρατηρήθηκαν κατά τη λήψη τμημάτων στα προηγούμενα χρονικά διαστήματα. Το MDP μπορεί να λυθεί με τη χρήση δυναμικού προγραμματισμού, αλλά η υπολογιστική πολυπλοκότητα γίνεται γρήγορα ακατάλληλη καθώς αυξάνεται το μέγεθος του προβλήματος. Για να λυθεί το πρόβλημα υιοθετείται το DQL (Gadaleta, Chiarriotti, Rossi, & Zanella, 2017), με χρήση των δικτύων Long Short-Term Memory (LSTM), στα οποία η είσοδος είναι η κατάσταση του πελάτη και η έξοδος περιλαμβάνει τιμές Q που αντιστοιχούν στις πιθανές ενέργειες του πελάτη. Για να βελτιωθεί η απόδοση του τυπικού LSTM, προστίθενται συνδέσεις θυρίδων παρατήρησης στα δίκτυα LSTM. Τα αποτελέσματα της προσομοίωσης με βάση το σύνολο δεδομένων (Klaue, Rathke, & Wolisz, 2003) δείχνουν ότι ο προτεινόμενος αλγόριθμος DQL

μπορεί να συγκλίνει πολύ πιο γρήγορα από το Q-learning. Συγκεκριμένα, ο αλγόριθμος DQL συγκλίνει σε περίπου 3 περιόδους, ενώ το Q-learning συγκλίνει σε περίπου 180 περιόδους. Η γρήγορη σύγκλιση δείχνει ότι ο DQL μπορεί να προσφέρει αποτελεσματική λύση σε προβλήματα σε εφαρμογές πραγματικού χρόνου. Επιπλέον, η προτεινόμενη DQL βελτιώνει την ποιότητα του βίντεο και μειώνει το rebuffering καθώς είναι σε θέση να διαχειριστεί δυναμικά την ενδιάμεση μνήμη λαμβάνοντας υπόψη την κατάστασή της και την χωρητικότητα καναλιού.

Όπως παρουσιάστηκε στην ενότητα 2.6, η μέθοδος Asynchronous Advantage Actor-Critic (A3C) περιλαμβάνει δύο νευρωνικά δίκτυα, δηλαδή, το δίκτυο Actor και το δίκτυο Critic. Το δίκτυο Actor βοηθά στην επιλογή ρυθμού bit για τον πελάτη και το δίκτυο Critic βοηθά στην εκπαίδευση του δικτύου Actor. Για το δίκτυο Actor, η είσοδος είναι η κατάσταση του πελάτη και η έξοδος είναι μια πολιτική, δηλαδή μια κατανομή πιθανότητας για πιθανές ενέργειες δεδομένων καταστάσεων που ο πελάτης μπορεί να λάβει. Εδώ, η ενέργεια επιλέγει την επόμενη αναπαράσταση, δηλαδή το επόμενο τμήμα με συγκεκριμένο ρυθμό bit, για λήψη. Για το δίκτυο Critic, η είσοδος είναι η κατάσταση του πελάτη και η έξοδος είναι η αναμενόμενη συνολική ανταμοιβή που λαμβάνεται από το δίκτυο Actor όταν ακολουθείται η πολιτική του. Τα αποτελέσματα προσομοίωσης που βασίζονται στο σύνολο δεδομένων για κινητά (Riiser, Vigmostad, & Griwodz, 2013) δείχνουν ότι η προτεινόμενη DQL μπορεί να βελτιώσει το μέσο QoE έως και 25% σε σύγκριση με το σχήμα ελέγχου bitrate (Yin, Jindal, Sekar, & Sinopoli, 2015). Επίσης, έχοντας επαρκές ενδιάμεση μνήμη για να χειριστεί τις διακυμάνσεις της απόδοσης του δικτύου, το προτεινόμενο DQL μειώνει το rebuffering περίπου 32,8% σε σύγκριση με το βασικό σχήμα.

Στην πράξη, ο αλγόριθμος DQL που προτείνεται στο (Mao, Netravali, & Alizadeh, 2017) μπορεί εύκολα να αναπτυχθεί σε ένα δίκτυο πολλαπλών πελατών, καθώς το A3C είναι σε θέση να υποστηρίζει παράλληλη εκπαίδευση για πολλούς πράκτορες. Κατά συνέπεια, κάθε πελάτης, δηλαδή κάθε πράκτορας, είναι διαμορφωμένος να παρακολουθεί την ανταμοιβή του. Αρχικά, ο πελάτης στέλνει το αναγνωριστικό του σε έναν διακομιστή, που περιλαμβάνει την κατάσταση, την ενέργεια και την ανταμοιβή του. Ο διακομιστής χρησιμοποιεί τον αλγόριθμο Actor-Critic για να ενημερώσει το μοντέλο του δικτύου Actor. Στη συνέχεια, ο διακομιστής ωθεί το νεότερο μοντέλο στον πράκτορα. Αυτή η διαδικασία

ενημέρωσης μπορεί να συμβεί ασύγχρονα μεταξύ όλων των παραγόντων, οπότε βελτιώνεται η ποιότητα και επιταχύνεται την εκπαίδευση. Παρόλο που το πρόγραμμα παράλληλης εκμάθησης ενδέχεται να προκαλέσει Round-Trip Time (RTT) μεταξύ των πελατών και του διακομιστή, τα αποτελέσματα της προσομοίωσης (Mao, Netravali, & Alizadeh, 2017) δείχνουν ότι το RTT μεταξύ των πελατών και του διακομιστή μειώνει το μέσο QoE μόνο κατά 3,5%. Η υποβάθμιση της απόδοσης είναι μικρή, και έτσι το προτεινόμενο DQL μπορεί να εφαρμοστεί σε πραγματικά συστήματα δικτύου.

Η είσοδος του DQL, δηλαδή, η κατάσταση του πελάτη, περιλαμβάνει την ποιότητα βίντεο του τελευταίου τμήματος βίντεο που λαμβάνεται (Gadaleta, Chiarriotti, Rossi, & Zanella, 2017) και (Mao, Netravali, & Alizadeh, 2017). Το τμήμα βίντεο είναι ακατέργαστο και μπορεί να προκαλέσει "έκρηξη κατάστασης" στον χώρο κατάστασης (Huang, Zhang, Zhou, & Sun, 2018). Για να μειωθεί ο χώρος κατάστασης και να βελτιωθεί το QoE, προτείνεται η χρήση ενός δικτύου πρόβλεψης ποιότητας βίντεο (Huang, Zhang, Zhou, & Sun, 2018). Το δίκτυο πρόβλεψης εξάγει χρήσιμα χαρακτηριστικά από τα μη επεξεργασμένα τμήματα βίντεο χρησιμοποιώντας συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks – CNN) και αναδρομικά νευρωνικά δίκτυα (Recursive Neural Networks – RNN). Στη συνέχεια, η έξοδος του δικτύου πρόβλεψης, δηλαδή η προβλεπόμενη ποιότητα βίντεο, χρησιμοποιείται ως μία από τις εισόδους του DQL που προτείνεται στο (Mao, Netravali, & Alizadeh, 2017). Τα αποτελέσματα προσομοίωσης με βάση το σύνολο δεδομένων ευρείας ζώνης (Raw Data - Measuring Broadband America 2016) δείχνουν ότι η προτεινόμενη DQL μπορεί να βελτιώσει το μέσο QoE έως και 25% σε σύγκριση με το Google Hangout, δηλαδή μια πλατφόρμα επικοινωνίας που αναπτύχθηκε από την Google. Επιπλέον, η προτεινόμενη DQL μπορεί να μειώσει τη μέση καθυστέρηση μετάδοσης βίντεο κατά περίπου 45% λόγω του μικρού χώρου κατάστασης. Αυτό σημαίνει ότι στα σενάρια που ο χώρος κατάστασης είναι μεγάλος, το CNN θα πρέπει να χρησιμοποιείται για τη βελτίωση του QoE του χρήστη και του χρόνου σύγκλισης.

Εκτός από τα συστήματα DASH, το DQL μπορεί να χρησιμοποιηθεί αποτελεσματικά για τον έλεγχο ρυθμού σε εφαρμογές ευέλικτου χρόνου υψηλής έντασης (High Volume Flexible Time - HVFT). Οι εφαρμογές HVFT χρησιμοποιούν κυψελοειδή δίκτυα για την παροχή κυκλοφορίας IoT. Οι εφαρμογές HVFT έχουν μεγάλο όγκο κυκλοφορίας, και ο προγραμματισμός κυκλοφορίας, π.χ., έλεγχος ρυθμού δεδομένων, στις εφαρμογές HVFT

είναι απαραίτητος. Μία κοινή προσέγγιση είναι να ανατεθούν κατηγορίες στατικής προτεραιότητας ανά τύπο κυκλοφορίας και, στη συνέχεια, ο προγραμματισμός κυκλοφορίας να βασίζεται στην κατηγορία προτεραιότητας. Ωστόσο, μια τέτοια προσέγγιση δεν εξελίσσεται για να δεχθεί νέες κατηγορίες. Έτσι, μέθοδοι μάθησης όπως η DQL θα πρέπει να χρησιμοποιούνται για την παροχή προσαρμοστικών μηχανισμών ελέγχου ρυθμού (Chinchali, et al., 2018). Το μοντέλο δικτύου είναι ένα μοναδικό κελί που περιλαμβάνει ένα BS ως κεντρικό ελεγκτή και πολλούς χρήστες κινητών. Το πρόβλημα στο BS είναι να βρει μια κατάλληλη πολιτική, δηλαδή, ρυθμό δεδομένων για τους χρήστες, για να μεγιστοποιήσει την ποσότητα της μετάδοσης κίνησης HVFT, ενώ ταυτόχρονα να ελαχιστοποιήσει την υποβάθμιση των επιδόσεων στα υπάρχοντα δεδομένα κίνησης. Το πρόβλημα μπορεί να διατυπωθεί ως MDP (Chinchali, et al., 2018). Ο πράκτορας είναι το BS και η κατάσταση περιλαμβάνει την τρέχουσα κατάσταση δικτύου και τα χρήσιμα χαρακτηριστικά που έχουν εξαχθεί από καταστάσεις δικτύου στις προηγούμενες χρονικές περιόδους. Η κατάσταση δικτύου σε μια χρονοθυρίδα περιλαμβάνει (i) τη μέτρηση συμφόρησης, δηλαδή το φορτίο κίνησης της κυψέλης, στη χρονοθυρίδα, (ii) τον συνολικό αριθμό των συνδέσεων δικτύου και (iii) την αποτελεσματικότητα των κυψελών, δηλαδή την ποιότητα των κυψελών . Η ενέργεια που κάνει το BS είναι ένας συνδυασμός του ρυθμού κυκλοφορίας για τους χρήστες. Για την επίτευξη του στόχου του BS, η ανταμοιβή ορίζεται ως συνάρτηση του (i) του αθροίσματος της κίνησης HVFT, (ii) της απώλειας κίνησης σε υπάρχουσες εφαρμογές λόγω της παρουσίας της κίνησης HVFT και (iii) του ποσού των byte που εξυπηρετούνται κάτω από την επιθυμητή ελάχιστη απόδοση. Στη συνέχεια, υιοθετείται το DQL που χρησιμοποιεί τα δίκτυα Actor και Critic με το LSTM. Χρησιμοποιώντας τα πραγματικά δεδομένα δικτύου που συλλέχθηκαν στη Μελβούρνη, τα αποτελέσματα της προσομοίωσης δείχνουν ότι το προτεινόμενο σχήμα DQL αυξάνει την κίνηση HVFT έως και 2 φορές σε σύγκριση με το ευρετικό σχήμα ελέγχου. Η προτεινόμενη DQL αναμένεται επομένως να εφαρμοστεί σε σύγχρονα δίκτυα σε μεγάλες πόλεις με μεγάλη αύξηση του πληθυσμού.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Gadaleta, Chiarriotti, Rossi, & Zanella, 2017)	MDP	DQN με LSTM και συνδέσεις θυρίδων παρατήρησης	Πελάτης	Ποιότητα τελευταίου τμήματος, τρέχουσα κατάσταση ενδιάμεσης μνήμης, χρόνος rebuffering και χωρητικότητα καναλιού	Επιλογή ρυθμού bit για τμήμα	Ποιότητα βίντεο, rebuffering και κατάσταση ενδιάμεσης μνήμης	Σύστημα DASH

(Mao, Netravali, & Alizadeh, 2017)	MDP	DQN με A3C	Πελάτης	Ποιότητα τελευταίου τμήματος, τρέχουσα κατάσταση ενδιάμεσης μνήμης, χρόνος rebuffering και χωρητικότητα καναλιού	Επιλογή ρυθμού bit για τμήμα	Ποιότητα βίντεο, rebuffering και κατάσταση ενδιάμεσης μνήμης	Σύστημα DASH
(Huang, Zhang, Zhou, & Sun, 2018)	MDP	DQN με CNN και RNN	Πελάτης	Προβλεπόμενη ποιότητα βίντεο, τρέχουσα κατάσταση ενδιάμεσης μνήμης, χρόνος rebuffering και χωρητικότητα καναλιού	Επιλογή ρυθμού bit για τμήμα	Ποιότητα βίντεο, rebuffering και κατάσταση ενδιάμεσης μνήμης	Σύστημα DASH
(Chinchali, και συν., 2018)	MDP	DQN με χρήση A3C και LSTM	Σταθμός Βάσης	μέτρηση συμφόρησης, τρέχουσες συνδέσεις δικτύου και αποδοτικότητα κυψελών	Επιλογές ρυθμού κίνησης για κινητούς χρήστες	Κυκλοφορία HVFT, απώλεια κίνησης σε υπάρχουσες εφαρμογές και το ποσό των bytes που εξυπηρετούνται	Εφαρμογή HVFT
(Ferreira, και συν., 2018)	MDP	DQN με χρήση FNN	Σταθμός Βάσης	Μέτρηση του BER, απόδοσης (throughput), φασματική απόδοση, κατανάλωση ισχύος και απόδοση εκπεμπόμενης ισχύος	Ρυθμός συμβόλων, ενέργεια ανά σύμβολο, τρόπος διαμόρφωσης, αριθμός bit ανά σύμβολο και ρυθμός κωδικοποίησης	Μέτρηση του BER, απόδοσης (throughput), φασματική απόδοση, κατανάλωση ισχύος και απόδοση εκπεμπόμενης ισχύος	Διαστημικό σύστημα επικοινωνίας

Πίνακας 4. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για προσαρμοστικό έλεγχο ρυθμού δεδομένων

Στις προαναφερθείσες προσεγγίσεις, ο μέγιστος αριθμός στόχων περιορίζεται, π.χ. σε 3 (Zhang, Zheng, Li, Huang, & Yang, 2018). Το DQL μπορεί να χρησιμοποιηθεί για τον έλεγχο ρυθμού για την επίτευξη πολλαπλών στόχων σε σύνθετα συστήματα επικοινωνίας (Ferreira, et al., 2018). Το μοντέλο δικτύου είναι ένα μελλοντικό σύστημα διαστημικής επικοινωνίας που αναμένεται να λειτουργεί σε απρόβλεπτα περιβάλλοντα, π.χ. δυναμικής τροχιάς, ατμοσφαιρικού και διαστημικού καιρού και δυναμικών καναλιών. Στο σύστημα, ο πομπός πρέπει να ρυθμιστεί με διάφορες παραμέτρους μετάδοσης, π.χ. ρυθμό συμβόλων και ρυθμό κωδικοποίησης, για την επίτευξη πολλαπλών στόχων σύγκρουσης, όπως χαμηλός ρυθμός σφάλματος Bit (Bit Error Rate - BER), βελτίωση της απόδοσης, αποδοτικότητα ισχύος και φάσματος. Μπορούν να χρησιμοποιηθούν τα προσαρμοστικά σχήματα κωδικοποίησης και διαμόρφωσης (Tarchi, Corazza, & Vanelli-Coralli, 2013). Ωστόσο, οι μέθοδοι επιτρέπουν την επίτευξη μόνο περιορισμένων αριθμών στόχων. Γι αυτό το λόγο μπορούν να χρησιμοποιηθούν αλγόριθμοι εκμάθησης όπως η DQL. Ο πράκτορας είναι ο πομπός στο σύστημα. Η ενέργεια είναι ένας συνδυασμός (i) ρυθμού

συμβόλων, (ii) ενέργειας ανά σύμβολο, (iii) τρόπου διαμόρφωσης, (iv) αριθμού bits ανά σύμβολο και (v) ρυθμού κωδικοποίησης. Ο στόχος είναι η μεγιστοποίηση της απόδοσης του συστήματος. Έτσι, η ανταμοιβή ορίζεται ως συνάρτηση σταθερότητας των παραμέτρων απόδοσης, συμπεριλαμβανομένων (i) BER εκτιμώμενου στον δέκτη, (ii) απόδοσης, (iii) φασματικής απόδοσης, (iv) κατανάλωσης ισχύος και (v) μετάδοσης αποτελεσματικότητας ισχύος. Η κατάσταση είναι η απόδοση του συστήματος που μετράται από τον πομπό και έτσι η κατάσταση είναι η ανταμοιβή. Για την επίτευξη πολλαπλών στόχων, το DQL υλοποιείται χρησιμοποιώντας ένα σύνολο πολλαπλών νευρωνικών δικτύων παράλληλα. Η είσοδος του DQL είναι η τρέχουσα κατάσταση και οι συνθήκες του καναλιού, και η έξοδος είναι η προβλεπόμενη ενέργεια. Τα νευρωνικά δίκτυα εκπαιδεύονται χρησιμοποιώντας τον αλγόριθμο backpropagation Levenberg-Marquardt (Hagan & Menhaj, 1994). Τα αποτελέσματα προσομοίωσης δείχνουν ότι η προτεινόμενη DQL μπορεί να επιτύχει τη βαθμολογία πυκνότητας, δηλαδή το σταθμισμένο άθροισμα διαφορετικών στόχων, κοντά στο ιδανικό, δηλαδή την εξαντλητική προσέγγιση αναζήτησης. Αυτό υπονοεί ότι η DQL είναι σε θέση να επιλέξει σχεδόν βέλτιστες ενέργειες και να μάθει τη σχέση μεταξύ ανταμοιβών και ενεργειών δεδομένης της δυναμικής συνθήκης καναλιών.

Στον Πίνακα 4 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για προσαρμοστικό έλεγχο ρυθμού δεδομένων.

3.3 Ασύρματη Προληπτική Προσωρινή Αποθήκευση

Η ασύρματη προληπτική προσωρινή αποθήκευση έχει προσελκύσει μεγάλο ακαδημαϊκό και βιομηχανικό ενδιαφέρον. Στατιστικά, μερικά δημοφιλή περιεχόμενα συνήθως ζητούνται από πολλούς χρήστες σε σύντομο χρονικό διάστημα, το οποίο αντιπροσωπεύει το μεγαλύτερο μέρος της φόρτωσης. Επομένως, η προληπτική προσωρινή αποθήκευση δημοφιλών περιεχομένων μπορεί να αποφύγει το βαρύ φορτίο των οπισθοζευκτικών συνδέσμων. Συγκεκριμένα, αυτή η τεχνική στοχεύει στην προγενέστερη προσωρινή αποθήκευση των περιεχομένων από τους απομακρυσμένους διακομιστές περιεχομένου στις συσκευές αιχμής ή BS που είναι κοντά στους τελικούς χρήστες. Εάν τα ζητούμενα περιεχόμενα είναι ήδη προσωρινά αποθηκευμένα τοπικά, το BS μπορεί να εξυπηρετήσει άμεσα τους τελικούς χρήστες με μικρή καθυστέρηση. Διαφορετικά, το BS ζητά αυτά τα περιεχόμενα από τον αρχικό διακομιστή περιεχομένου και ενημερώνει την τοπική

προσωρινή μνήμη με βάση την πολιτική προσωρινής αποθήκευσης, η οποία είναι ένα από τα κύρια προβλήματα σχεδίασης για την ασύρματη προληπτική προσωρινή αποθήκευση.

3.3.1 Προσωρινή Αποθήκευση QoS-Aware

Η δημοτικότητα περιεχομένου είναι ο βασικός παράγοντας που χρησιμοποιείται για την επίλυση του προβλήματος προσωρινής αποθήκευσης περιεχομένου. Με μεγάλο αριθμό περιεχομένων και τη μεταβαλλόμενη δημοφιλία, η DQL είναι μια ελκυστική στρατηγική για την αντιμετώπιση αυτού του προβλήματος με υψηλών διαστάσεων χώρων καταστάσεων και ενεργειών. Στη συνέχεια παρουσιάζεται ένα σχήμα DQL για τη βελτίωση της απόδοσης προσωρινής αποθήκευσης (Zhong, Gursoy, & Velipasalar, 2018). Το μοντέλο συστήματος αποτελείται από ένα μόνο BS με σταθερό μέγεθος προσωρινής μνήμης. Για κάθε αίτημα, το BS ως πράκτορας αποφασίζει εάν θα αποθηκεύσει το περιεχόμενο που ζητείται αυτήν τη στιγμή στην κρυφή μνήμη. Εάν διατηρηθεί το νέο περιεχόμενο, το BS καθορίζει ποιο τοπικό περιεχόμενο θα αντικατασταθεί. Η κατάσταση είναι ο λειτουργικός χώρος των προσωρινά αποθηκευμένων περιεχομένων και του περιεχομένου που ζητείται την τρέχουσα στιγμή. Ο λειτουργικός χώρος αποτελείται από τον συνολικό αριθμό αιτημάτων για κάθε περιεχόμενο σε συγκεκριμένο βραχυπρόθεσμο, μεσοπρόθεσμο και μακροπρόθεσμο διάστημα. Υπάρχουν δύο τύποι ενεργειών: (i) να βρεθεί ένα ζεύγος περιεχομένων και να ανταλλάχθούν οι καταστάσεις προσωρινής μνήμης των δύο περιεχομένων και (ii) να διατηρηθούν οι καταστάσεις προσωρινής μνήμης του περιεχομένου αμετάβλητες. Ο στόχος του BS είναι η μεγιστοποίηση του μακροχρόνιου ποσοστού επίσκεψης cache, δηλαδή η ανταμοιβή.

Το σχήμα DQL (Zhong, Gursoy, & Velipasalar, 2018) εκπαιδεύει την πολιτική χρησιμοποιώντας τη μέθοδο Deep Deterministic Policy Gradient (DDPG) (Lillicrap, et al., 2016) και χρησιμοποιεί την αρχιτεκτονική Wolpertinger (Dulac-Arnold, Evans, Sunehag, & Coppin, 2015) για να μειώσει το μέγεθος του χώρου ενεργειών και να αποφύγει να χάσει μια βέλτιστη πολιτική. Η αρχιτεκτονική του Wolpertinger αποτελείται από τρία κύρια μέρη: ένα δίκτυο Actor, το K-Nearest Neighbours (K-NN) και ένα δίκτυο Critic. Το δίκτυο Actor υπάρχει για να αποφύγει ένα μεγάλο χώρο δράσης. Το δίκτυο Critic είναι να διορθώσει την απόφαση που έλαβε το δίκτυο Actor. Η μέθοδος DDPG εφαρμόζεται για την ενημέρωση των δικτύων Critic και Actor, ενώ το K-NN μπορεί να βοηθήσει στην εξερεύνηση ενός συνόλου ενεργειών για την αποφυγή κακών αποφάσεων. Τα δίκτυα

Actor και Critic εφαρμόζονται στη συνέχεια χρησιμοποιώντας Feedforward Neural Networks (FNNs). Τα αποτελέσματα της προσομοίωσης δείχνουν ότι το προτεινόμενο σχήμα DQL ξεπερνά το σχήμα πρώτος μέσα – πρώτο έξω (First-in, first-out – FIFO), όσον αφορά το ρυθμό εύστοχων αναζητήσεων μακροπρόθεσμης κρυφής μνήμης. Συγκεκριμένα, ο ανωτέρω λόγος που λαμβάνεται από το σχήμα DQL είναι 0,5, ενώ αυτός που επιτυγχάνεται με το σχήμα FIFO είναι 0,4. Οι συγκρίσεις απόδοσης καταδεικνύουν ότι το προτεινόμενο σχήμα DQL μπορεί να επιτύχει ανταγωνιστικούς ρυθμούς επιτυχίας προσωρινής αποθήκευσης, μειώνοντας αποτελεσματικά τον χρόνο εκτέλεσης. Αυτό καθιστά το προτεινόμενο πλαίσιο αποτελεσματικό και κατάλληλο για το χειρισμό δεδομένων μεγάλης κλίμακας.

Η μεγιστοποίηση του ρυθμού εύστοχων αναζητήσεων μακροπρόθεσμης κρυφής μνήμης (Zhong, Gursoy, & Velipasalar, 2018) σημαίνει ότι η προσωρινή μνήμη αποθηκεύει τα πιο δημοφιλή περιεχόμενα. Σε ένα δυναμικό περιβάλλον, τα περιεχόμενα που είναι αποθηκευμένα σε προσωρινή μνήμη πρέπει να αντικατασταθούν σύμφωνα με τα δυναμικά αιτήματα των χρηστών. Η βελτιστοποίηση της τοποθέτησης ή αντικατάστασης αποθηκευμένων περιεχομένων μελετάται (Lei, You, Dai, Vu, Yuan, & Chatzinotas, 2017) με μια μέθοδο βαθιάς μάθησης. Ο αλγόριθμος βελτιστοποίησης εκπαιδεύεται από έναν DNN αρχικά και στη συνέχεια χρησιμοποιείται για προσωρινή αποθήκευση σε πραγματικό χρόνο ή προγραμματισμό με ελάχιστη καθυστέρηση. Για να βρεθούν οι χρόνοι λήξης της προσωρινής μνήμης , δηλαδή το Time-To-Live (TTL), για δυναμικά μεταβαλλόμενα αιτήματα σε δίκτυα παράδοσης περιεχομένου, προτείνεται μια βέλτιστη πολιτική προσωρινής αποθήκευσης (Schaarschmidt, Gessert, Dalibard, & Yoneki, 2016). Το σύστημα περιλαμβάνει διακομιστή βάσης δεδομένων cloud και πολλές φορητές συσκευές που μπορούν να εκδίδουν ερωτήματα και να ενημερώνουν καταχωρήσεις σε μία βάση δεδομένων. Τα αποτελέσματα του ερωτήματος μπορούν να αποθηκευτούν προσωρινά στην προσωρινή μνήμη για ένα καθορισμένο χρονικό διάστημα σε κρυφές μνήμες που ελέγχονται από το διακομιστή. Όλα τα προσωρινά αποθηκευμένα ερωτήματα θα καταστούν μη έγκυρα εάν έχει ενημερωθεί μία από τις αποθηκευμένες εγγραφές. Ένα μεγάλο TTL θα επιβαρύνει τις δυνατότητες προσωρινής αποθήκευσης, ενώ ένα μικρό TTL αυξάνει σημαντικά τις καθυστερήσεις εάν ο διακομιστής βάσης δεδομένων είναι φυσικά απομακρυσμένος.

Σε αντίθεση με την προσέγγιση DDPG (Zhong, Gursoy, & Velipasalar, 2018), προτείνεται (Schaarschmidt, Gessert, Dalibard, & Yoneki, 2016) η χρήση των κανονικοποιημένων λειτουργιών πλεονεκτήματος (Normalized Advantage Functions - NAFs) για συνεχές DQL σχήμα για να υπολογιστεί η βέλτιστη διάρκεια λήξης της προσωρινής μνήμης. Το βασικό πρόβλημα στη συνεχή DQL είναι να επιλεγεί μια ενέργεια που μεγιστοποιεί τη συνάρτηση Q, αποφεύγοντας παράλληλα την εκτέλεση μιας δαπανηρής αριθμητικής βελτιστοποίησης σε κάθε βήμα. Η χρήση των NAF αποτρέπει ένα δεύτερο δίκτυο Actor που πρέπει να εκπαιδευτεί ξεχωριστά. Αντ 'αυτού, ένα μόνο νευρωνικό δίκτυο χρησιμοποιείται για την έξοδο τόσο μιας συνάρτησης αξίας όσο και ενός όρου πλεονεκτήματος. Ο πράκτορας DQL στη βάση δεδομένων cloud χρησιμοποιεί κωδικοποίηση του ίδιου του ερωτήματος και τους ρυθμούς απώλειας ερωτήματος, ως καταστάσεις του συστήματος, κάτι που επιτρέπει μια ευκολότερη γενίκευση. Η ανταμοιβή του συστήματος είναι γραμμικά ανάλογη με το τρέχον φορτίο, δηλαδή τον αριθμό των προσωρινά αποθηκευμένων ερωτημάτων διαιρούμενο με τη συνολική χωρητικότητα. Αυτή η συνάρτηση ανταμοιβής μπορεί να ενθαρρύνει μεγαλύτερα TTLs όταν αποθηκεύονται λιγότερα ερωτήματα και μικρότερα TTL όταν το φορτίο είναι κοντά στην χωρητικότητα του συστήματος. Λαμβάνοντας υπόψη τις ατελείς μετρήσεις για τις ανταμοιβές και τις επόμενες καταστάσεις στο χρόνο εκτέλεσης, εισάγεται η προσέγγιση Delayed Experience Injection (DEI) που επιτρέπει στον πράκτορα DQL να παρακολουθεί τις ατελείς μεταβάσεις όταν οι μετρήσεις δεν είναι άμεσα διαθέσιμες. Στη συνέχεια αξιολογείται ο αλγόριθμος εκμάθησης από το έλεγχο επίδοσης Yahoo! εξυπηρέτηση cloud με προσαρμοσμένους φόρτους εργασίας στο Web (Cooper, Silberstein, Tam, Ramakrishnan, & Sears, 2010). Τα αποτελέσματα της προσομοίωσης επιβεβαιώνουν ότι η μαθησιακή προσέγγιση που βασίζεται στα NAF και το DEI ξεπερνά έναν στατιστικό εκτιμητή.

3.3.2 Έλεγχος Κοινής Προσωρινής Αποθήκευσης και Μετάδοσης

Οι πολιτικές προσωρινής αποθήκευσης καθορίζουν τον τόπο αποτελεσματικής αποθήκευσης και να ανάκτησης του ζητούμενο περιεχομένου, π.χ., μαθαίνοντας τις δημοτικότητα του περιεχομένου (Zhong, Gursoy, & Velipasalar, 2018) και τον χρόνο λήξης της προσωρινής μνήμης (Schaarschmidt, Gessert, Dalibard, & Yoneki, 2016). Μια άλλη σημαντική πτυχή του σχεδιασμού προσωρινής αποθήκευσης είναι ο έλεγχος μετάδοσης της παράδοσης περιεχομένου από τις κρυφές μνήμες στους τελικούς χρήστες, ειδικά για

ασύρματα συστήματα με δυναμικές συνθήκες καναλιού. Για να αποφευχθούν αμοιβαίες παρεμβολές σε ασύρματα δίκτυα πολλαπλών χρηστών, ο έλεγχος μετάδοσης αποφασίζει για τα αποθηκευμένα περιεχόμενα που μπορούν να μεταδοθούν ταυτόχρονα, καθώς και για τις πιο κατάλληλες παραμέτρους ελέγχου, π.χ. ισχύς μετάδοσης, προκωδικοποίηση, ρυθμός δεδομένων και κατανομή καναλιών. Ως εκ τούτου, απαιτείται ο από κοινού σχεδιασμός της προσωρινής αποθήκευσης και του ελέγχου μετάδοσης για να είναι δυνατή η αποτελεσματική παράδοση περιεχομένου σε ασύρματα δίκτυα πολλαπλών χρηστών.

Πρόσφατα, προτείνονται ορισμένες προσεγγίσεις, π.χ. (Deghel, Bastug, Assaad, & Debbah, 2015) για την κοινή προσωρινή αποθήκευση και ευθυγράμμιση παρεμβολών σε ασύρματα συστήματα. Ωστόσο, οι περισσότερες από τις προσεγγίσεις υποθέτουν ότι οι πληροφορίες κατάστασης καναλιού είναι αμετάβλητες που μπορεί να μην συμβαίνει σε δυναμικά ασύρματα συστήματα. Η εφαρμογή του πλαισίου DQL (He & Hu, 2017), (He, Liang, Yu, Zhao, & Yin, Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach, 2017), (He, et al., 2017) στην κοινή προσωρινή αποθήκευση και ευθυγράμμιση παρεμβολών δύναται να αντιμετωπίσει το πρόβλημα των αμοιβαίων παρεμβολών σε ασύρματα δίκτυα πολλαπλών χρηστών. Έστω ένα σύστημα MIMO με περιορισμένη οπισθοζευκτική χωρητικότητα και τις κρυφές μνήμες στον πομπό. Ο κωδικοποιημένος σχεδιασμός για ευθυγράμμιση παρεμβολών απαιτεί καθολικό Channel State Information (CSI) σε κάθε πομπό. Ένας κεντρικός προγραμματιστής είναι υπεύθυνος για τη συλλογή της κατάστασης CSI και της προσωρινής μνήμης από κάθε χρήστη μέσω της οπισθοζεύξης, προγραμματίζοντας τη μετάδοσης των χρηστών και βελτιστοποιώντας την κατανομή πόρων. Ενεργοποιώντας την προσωρινή αποθήκευση περιεχομένου σε μεμονωμένους πομπούς, μπορεί να μειωθεί η ζήτηση για μεταφορά δεδομένων και, επομένως, να εξοικονομηθεί περισσότερη οπισθοζευκτική χωρητικότητα για ενημέρωση και κοινή χρήση CSI σε πραγματικό χρόνο. Η χρήση της προσέγγισης που βασίζεται σε DQL στον κεντρικό προγραμματιστή μπορεί να μειώσει τη σαφή ζήτηση για CSI και την υπολογιστική πολυπλοκότητα στη βελτιστοποίηση του πίνακα, ειδικά με τις πολυποίκιλλες συνθήκες καναλιού. Ο πράκτορας DQL εφαρμόζει το DNN για να προσεγγίσει τη συνάρτηση Q με επανάληψη εμπειρίας στην εκπαίδευση. Για να γίνει η διαδικασία εκμάθησης πιο σταθερή, η παράμετρος του στόχου δικτύου Q ενημερώνεται

από το δίκτυο Q σε τακτά χρονικά διαστήματα. Οι πληροφορίες που συλλέγονται συγκεντρώνονται σε μια κατάσταση συστήματος και αποστέλλονται στον πράκτορα DQL, ο οποίος τροφοδοτεί μια βέλτιστη ενέργεια για την τρέχουσα στιγμή. Η ενέργεια δείχνει ποιοι χρήστες θα είναι ενεργοί και την κατανομή των πόρων μεταξύ ενεργών χρηστών. Η ανταμοιβή του συστήματος αντιπροσωπεύει τη συνολική απόδοση πολλαπλών χρηστών. Ένα DQN βασισμένο στο CNN υιοθετείται και αξιολογείται σε πιο πρακτικές συνθήκες με ατελές ή καθυστερημένο CSI (He, et al., 2017). Τα αποτελέσματα προσομοίωσης δείχνουν ότι η απόδοση του συστήματος MIMO βελτιώνεται σημαντικά σε σύγκριση με το βασικό σχήμα (Deghel, Bastug, Assaad, & Debbah, 2015) όσον αφορά τη συνολική απόδοση και την ενεργειακή απόδοση. Συγκεκριμένα, σε SNR = 15 dB, ο συνολικός ρυθμός που λαμβάνεται από το προτεινόμενο σχήμα DQN είναι 240 Mbps, ενώ αυτός που λαμβάνεται από το βασικό σχήμα είναι 200 Mbps.

Η διαχείριση παρεμβολών αποτελεί σημαντική απαίτηση των ασύρματων συστημάτων. Το QoS που σχετίζεται με την εφαρμογή ή η εμπειρία χρήστη είναι επίσης βασική μέτρηση. Σε αντίθεση με τα (He & Hu, 2017), (He, Liang, Yu, Zhao, & Yin, Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach, 2017), (He, et al., 2017), στο (He, Wang, Huang, Miyazaki, Wang, & Guo, 2020) προτείνεται μια προσέγγιση DQL για τη μεγιστοποίηση της ποιότητας εμπειρίας (QoE) των συσκευών IoT, βελτιστοποιώντας από κοινού την κατανομή cache και το ρυθμό μετάδοσης σε ασύρματα δίκτυα που επικεντρώνονται στο περιεχόμενο. Η κατάσταση του συστήματος καθορίζεται από τις συνθήκες προσωρινής αποθήκευσης των κόμβων, π.χ. από τις πληροφορίες υπηρεσίας και τα αποθηκευμένα περιεχόμενα, καθώς και από τους ρυθμούς μετάδοσης των αποθηκευμένων περιεχομένων. Ο σκοπός του πράκτορα DQL είναι να ελαχιστοποιεί συνεχώς το κόστος δικτύου ή να μεγιστοποιεί το QoE. Το προτεινόμενο πλαίσιο DQL ενισχύεται περαιτέρω με τη χρήση των Prioritized Experience Replay (PER) και Double Deep Q-Network (DDQN). Το PER επαναλαμβάνει σημαντικές μεταβάσεις πιο συχνά, ώστε το DQN να μπορεί να μαθαίνει από τα δείγματα πιο αποτελεσματικά. Η χρήση του DDQN μπορεί να σταθεροποιήσει τη μάθηση παρέχοντας δύο λειτουργίες αξίας σε ξεχωριστά νευρωνικά δίκτυα, με αποτέλεσμα να μην επιτρέπεται η υπερεκτίμηση του DQN με τον αυξανόμενο αριθμό ενεργειών. Αυτά τα δύο νευρωνικά δίκτυα δεν αποσυνδέονται εντελώς, καθώς το δίκτυο στόχος είναι ένα περιοδικό αντίγραφο του δικτύου εκτίμησης. Ένας διακριτός

προσομοιωτής ccnSim (Wu, Li, & Xie, 2013) χρησιμοποιείται για τη μοντελοποίηση της συμπεριφοράς προσωρινής αποθήκευσης σε διάφορες δομές γραφημάτων. Το ίχνος δεδομένων εξόδου του προσομοιωτή εισάγεται στη συνέχεια στο MATLAB και χρησιμοποιείται για την αξιολόγηση του αλγορίθμου εκμάθησης. Όπως φαίνεται στα αποτελέσματα προσομοίωσης, το προτεινόμενο πλαίσιο DQL μπορεί να επιτύχει μια τιμή QoE 4 που είναι διπλάσια από εκείνη του τυπικού σχήματος δοκιμής διείσδυσης. Επιπλέον, η υπολογιστική πολυπλοκότητα του DDQN είναι O (lg n) που είναι χαμηλότερη από εκείνη του τυπικού σχήματος δοκιμής διείσδυσης με την υπολογιστική πολυπλοκότητα του O (lsn^3), όπου n είναι ο αριθμός των κεντρικών υπολογιστικών κόμβων περιεχομένου, s είναι το αριθμός κόμβων υπηρεσίας και l είναι ο αριθμός των τιμών ταχύτητας μετάδοσης.

Το QoE μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει την αντίληψη των χρηστών για τις υπηρεσίες εικονικής πραγματικότητας (Virtual Reality - VR). Το (Chen, Saad, & Yin, Echo-Liquid State Deep Learning for 360° Content Transmission and Caching in Wireless VR Networks With Cellular-Connected UAVs, 2019) πραγματεύεται με την κοινή στρατηγική αποθήκευσης και αποθήκευσης περιεχομένου σε ένα ασύρματο δίκτυο VR, όπου τα UAV καταγράφουν βίντεο σε ζωντανά παιχνίδια και τα μεταδίδουν σε μικρούς κυψέλες BS που εξυπηρετούν τους χρήστες VR. Η μετάδοση περιεχομένου VR από τα UAV σε BS πραγματοποιείται με τους συνδέσμους ανάστροφης ζεύξης millimeter wave (mmWave). Τα BS μπορούν επίσης να αποθηκεύουν προσωρινά τα δημοφιλή περιεχόμενα που μπορεί να ζητούνται συχνά από τους τελικούς χρήστες. Το πρόβλημα κοινής προσωρινής αποθήκευσης και μετάδοσης περιεχομένου διατυπώνεται ως βελτιστοποίηση για τη μεγιστοποίηση της αξιοπιστίας των χρηστών, δηλαδή, την πιθανότητα της καθυστέρησης μετάδοσης περιεχομένου να ικανοποιεί τον στόχο στιγμιαίας καθυστέρησης. Η μεγιστοποίηση περιλαμβάνει τον έλεγχο της μορφής μετάδοσης, της συσχέτισης των χρηστών, της σειράς και της μορφής των προσωρινών αποθηκευμένων περιεχομένων. Ένα πλαίσιο DQL που συνδυάζει το Liquid State Machine (LSM) και το Echo State Network (ESN) προτείνεται για κάθε BS για να βρει τις βέλτιστες στρατηγικές μετάδοσης και προσωρινής αποθήκευσης. Ως τυχαία δημιουργημένο Spiking Neural Network (SNN) (Chen M. , Challita, Saad, Yin, & Debbah, 2017), το LSM μπορεί να αποθηκεύει πληροφορίες με την πάροδο του χρόνου σχετικά με το περιβάλλον του δικτύου και να προσαρμόζει την πολιτική συσχέτισης των χρηστών, τα περιεχόμενα και τις μορφές που έχουν

αποθηκευτεί στην κρυφή μνήμη σύμφωνα με τα αιτήματα περιεχομένου των χρηστών. Χρησιμοποιήθηκε στο (Chen, Saad, & Yin, Liquid State Machine Learning for Resource Allocation in a Network of Cache-Enabled LTE-U UAVs, 2017) για να προβλέψει τη διανομή αιτημάτων περιεχομένου των χρηστών, ενώ έχει περιορισμένες μόνο πληροφορίες σχετικά με το δίκτυο και τους διαφορετικούς χρήστες. Το συμβατικό LSM χρησιμοποιεί Feedforward Neural Networks (FNNs) ως λειτουργία εξόδου, η οποία απαιτεί υψηλή πολυπλοκότητα στην εκπαίδευση λόγω του υπολογισμού των βαθμίδων για όλους τους νευρώνες. Αντίθετα, το προτεινόμενο πλαίσιο DQL χρησιμοποιεί ένα Echo State Network (ESN) ως λειτουργία εξόδου, η οποία χρησιμοποιεί ιστορικές πληροφορίες για να βρει τη σχέση μεταξύ της αξιοπιστίας, της προσωρινής αποθήκευσης και της μετάδοσης περιεχομένου των χρηστών. Έχει επίσης χαμηλότερη πολυπλοκότητα στην εκπαίδευση και καλύτερη μνήμη για πληροφορίες δικτύου. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο πλαίσιο DQL μπορεί να αποφέρει κέρδος 25,4% από την άποψη της αξιοπιστίας των χρηστών σε σύγκριση με το επίπεδο αναφοράς Q-learning.

3.3.3 Κοινή Προσωρινή Αποθήκευση, Δικτύωση και Υπολογισμός

Ο έλεγχος προσωρινής αποθήκευσης και μετάδοσης θα εμπλακεί περισσότερο σε ένα ετερογενές δίκτυο (HetNet) που ενσωματώνει διαφορετικές τεχνολογίες επικοινωνίας, π.χ. κυψελοειδές σύστημα, δίκτυο συσκευή σε συσκευή, δίκτυο οχημάτων και δικτυακά UAV, για την υποστήριξη διαφόρων απαιτήσεων εφαρμογής. Η ετερογένεια του δικτύου εγείρει το πρόβλημα του περίπλοκου σχεδιασμού του συστήματος που πρέπει να αντιμετωπίσει προκλήσεις όπως αμοιβαίες παρεμβολές, διαφοροποιημένη παροχή QoS και κατανομή πόρων, σε ένα ενοποιημένο πλαίσιο. Αυτό απαιτεί κοινή βελτιστοποίηση πολύ περισσότερο από το του κοινό έλεγχο προσωρινής αποθήκευσης και μετάδοσης.

Στο (He, Zhang, & Zhang, A Big Data Deep Reinforcement Learning Approach to Next Generation Green Wireless Networks, 2017) προτείνεται ένα πλαίσιο DQL για ενεργειακά αποδοτική κατανομή πόρων σε πράσινα ασύρματα δίκτυα, λαμβάνοντας υπόψη από κοινού τις συνδέσεις μεταξύ δικτύωσης, προσωρινής αποθήκευσης στο δίκτυο και υπολογισμού. Το σύστημα αποτελείται από ένα Software-Defined Network (SDN) με πολλαπλά εικονικά δίκτυα και χρήστες κινητών που ζητούν αρχεία βίντεο κατά παραγγελία που απαιτούν ένα ορισμένο ποσό υπολογιστικού πόρου είτε στο διακομιστή

περιεχομένου είτε σε τοπικές συσκευές. Σε κάθε εικονικό δίκτυο, ένας εξουσιοδοτημένος χρήστης εκδίδει ένα αίτημα για λήψη αρχείων από ένα σύνολο διαθέσιμων μικρών σταθμών βάσης (Small Base Stations – SBS) στη γειτονική του περιοχή. Τα ασύρματα κανάλια μεταξύ κάθε χρήστη κινητής τηλεφωνίας και των SBS χαρακτηρίζονται ως κανάλια Markov Finite-State (FSMC). Οι καταστάσεις είναι η διαθέσιμη χωρητικότητα προσωρινής μνήμης στα SBS, οι συνθήκες καναλιού μεταξύ χρηστών κινητών και SBS, η υπολογιστική ικανότητα των διακομιστών περιεχομένου και των χρηστών κινητών. Ο πράκτορας DQL σε κάθε SBS αποφασίζει μια συσχέτιση μεταξύ κάθε χρήστη κινητού και SBS, πού θα εκτελέσει την υπολογιστική εργασία και πώς να προγραμματίσει τις μεταδόσεις SBS για την παράδοση των απαιτούμενων δεδομένων. Ο στόχος είναι να ελαχιστοποιηθεί η συνολική κατανάλωση ενέργειας του συστήματος από την αποθήκευση δεδομένων, την ασύρματη μετάδοση και τον υπολογισμό. Τα αποτελέσματα προσομοίωσης δείχνουν ότι η συνολική κατανάλωση ενέργειας σε διαφορετικά σενάρια δοκιμών είναι πολύ υψηλή στην αρχή της μαθησιακής διαδικασίας και σταδιακά μειώνεται κατά μια σταθερή τιμή όταν η μάθηση συγκλίνει. Επιπλέον, η κατανάλωση ενέργειας του ενοποιημένου πλαισίου DRL που εξετάζει την προσωρινή αποθήκευση, τη δικτύωση και τον υπολογιστή είναι σημαντικά χαμηλότερη από εκείνη άλλων πλαισίων DRL που εστιάζουν μόνο σε μέρος των μεταβλητών ελέγχου.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Zhong, Gursoy, & Velipasalar, 2018)	MDP	DQN με χρήση δράστη-κριτή DDPG	Σταθμός Βάσης	Προσωρινά αποθηκευμένα περιεχόμενα και ζητούμενο περιεχόμενο	Αντικατάσταση του επιλεγμένου περιεχόμενου ή όχι	Ρυθμός bit προσωρινής μνήμης (βαθμός 1 ή 0)	CRN
(He, Zhang, & Zhang, 2017)	MDP	DQN με χρήση FNN	Σταθμός Βάσης	Καταστάσεις καναλιού και υπολογιστικές δυνατότητες	Συσχέτιση χρηστών, υπολογιστική μονάδα, παράδοση περιεχομένου	Κατανάλωση ενέργειας	CRN
(Schaarschmidt, Gessert, Dalibard, & Yoneki, 2016)	MDP	DQN με χρήση NAFs	Βάση δεδομένων Cloud	Κωδικοποίηση ερωτήματος, ρυθμός απώλειας προσωρινής μνήμης ερωτήματος	Χρόνοι λήξης προσωρινής μνήμης	Ρυθμός bit προσωρινής μνήμης, χρήση CDN	Βάση δεδομένων Cloud
(He & Hu, 2017) (He, Liang, Yu, Zhao, & Yin, 2017)	MDP	DQN με χρήση FNN	Κεντρικός προγραμματιστής	Συντελεστές καναλιού, κατάσταση προσωρινής μνήμης	Ενεργοί χρήστες και κατανομή πόρων	Απόδοση δικτύου	Σύστημα MU MIMO

(He, και συν., 2017)	MDP	DQN με χρήση CNN	Κεντρικός προγραμματιστής	Συντελεστές καναλιού, κατάσταση προσωρινής μνήμης	Ενεργοί χρήστες και κατανομή πόρων	Απόδοση δικτύου	Σύστημα MU MIMO
(He, Wang, Huang, Miyazaki, Wang, & Guo, 2020)	MDP	DDQN	Πάροχος υπηρεσιών	Καταστάσεις κόμβων κρυφής μνήμης, ρυθμοί μετάδοσης κομματιών περιεχομένου	Το περιεχόμενο κομματιών για προσωρινή αποθήκευση και κατάργηση	Κόστος δικτύου, QoE	Κεντρικό περιεχόμενο IoT
(Chen, Saad, & Yin, 2019)	MDP	DQN με χρήση LSM και ESN	Σταθμός Βάσης	Αίτημα ιστορικού περιεχομένου	Συσχέτιση χρηστών, προσωρινά αποθηκευμένα περιεχόμενα και μορφές	Αξιοπιστία	Σύστημα κυψελωτής τηλεφωνίας
(He, Yu, Zhao, Yin, & Boukerche, 2017) (He, Yu, Zhao, Leung, & Yin, 2017)	MDP	DQN με χρήση CNN	Πάροχος υπηρεσιών	Διαθέσιμο σταθμό βάσης MEC και προσωρινή μνήμη	Συσχέτιση χρηστών, ασύρματη προσωρινή αποθήκευση και εκφόρτωση δεδομένων	Σύνθετα έσοδα	Ad hoc δίκτυο οχημάτων
(He, και συν., 2017)	MDP	DQN με χρήση FNN	Πάροχος υπηρεσιών	Διαθέσιμο σταθμό βάσης MEC και προσωρινή μνήμη	Συσχέτιση χρηστών, ασύρματη προσωρινή αποθήκευση και εκφόρτωση δεδομένων	Σύνθετα έσοδα	Ad hoc δίκτυο οχημάτων
(He, Zhao, & Yin, 2017)	MDP	DDQN και Dueling DQN	Πάροχος υπηρεσιών	Διαθέσιμο σταθμό βάσης MEC και προσωρινή μνήμη	Συσχέτιση χρηστών, ασύρματη προσωρινή αποθήκευση και εκφόρτωση δεδομένων	Σύνθετα έσοδα	Ad hoc δίκτυο οχημάτων
(He, Yu, Zhao, & Yin, 2018)	MDP	DQN με χρήση CNN	Σταθμός Βάσης	Καταστάσεις καναλιού, υπολογιστικές δυνατότητες, δείκτης περιεχομένου / έκδοσης και η τιμή εμπιστοσύνης	Συσχέτιση χρηστών, ασύρματη προσωρινή αποθήκευση και εκφόρτωση δεδομένων	Έσοδα	Κινητό κοινωνικό δίκτυο

Πίνακας 5. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για ασύρματη προσωρινή αποθήκευση

Το σχέδιο DQL που προτάθηκε στο (He, Zhang, & Zhang, A Big Data Deep Reinforcement Learning Approach to Next Generation Green Wireless Networks, 2017) έχει εφαρμοστεί για τη βελτίωση της απόδοσης των Vehicular Ad hoc NETworks (VANETs) στο (He, et al., 2017), (He, Yu, Zhao, Yin, & Boukerche, Deep Reinforcement Learning (DRL)-based Resource Management in Software-Defined and Virtualized Vehicular Ad Hoc Networks, 2017), (He, Zhao, & Yin, Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach, 2017). Το μοντέλο δικτύου περιλαμβάνει πολλαπλούς BS, Road Side Units (RSUs), διακομιστές MEC και διακομιστές περιεχομένου. Όλες οι συσκευές ελέγχονται από έναν πάροχο εικονικού δικτύου κινητής τηλεφωνίας. Τα οχήματα ζητούν περιεχόμενο βίντεο που μπορεί να αποθηκευτεί προσωρινά στα BSs ή να ανακτηθεί από απομακρυσμένους διακομιστές περιεχομένου.

Στο (He, et al., 2017) διατυπώνεται το πρόβλημα κατανομής πόρων ως από κοινού βελτιστοποίηση της προσωρινής αποθήκευσης, της δικτύωσης και της πληροφορικής, π.χ. συμπίεση και κωδικοποίηση λειτουργιών του περιεχομένου βίντεο. Οι καταστάσεις συστήματος περιλαμβάνουν το CSI από κάθε BS, την υπολογιστική ικανότητα και το μέγεθος κρυφής μνήμης κάθε διακομιστή MEC / περιεχομένου. Ο διαχειριστής δικτύου τροφοδοτεί την κατάσταση του συστήματος με το DQN που βασίζεται στο FNN και λαμβάνει τη βέλτιστη πολιτική που καθορίζει την κατανομή πόρων για κάθε όχημα. Η Q-learning ενισχύεται χρησιμοποιώντας CNNs στο DQN, για την εκμετάλλευση χωρικών συσχετισμών στη μάθηση (He, Yu, Zhao, Yin, & Boukerche, Deep Reinforcement Learning (DRL)-based Resource Management in Software-Defined and Virtualized Vehicular Ad Hoc Networks, 2017). Αυτό καθιστά δυνατή την εξαγωγή λειτουργιών υψηλού επιπέδου από ακατέργαστα δεδομένα εισόδου. Για τη βελτίωση της σταθερότητας και της απόδοσης της συνήθους μεθόδου DQN έχουν εισαχθεί δύο σχήματα (He, Zhao, & Yin, Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach, 2017). Πρώτον, το DDQN έχει σχεδιαστεί για να αποφεύγει την υπερβολική εκτίμηση της τιμής Q στο συνηθισμένο DQN, ως εκ τούτου, η ενέργεια μπορεί να αποσυνδεθεί από τη δημιουργία της τιμής Q-στόχου. Αυτό καθιστά τη διαδικασία εκμάθησης πιο γρήγορη και πιο αξιόπιστη. Δεύτερον, η προσέγγιση DDQN είναι επίσης ενταγμένη στο σχεδιασμό με την ιδέα ότι δεν είναι πάντα απαραίτητο να εκτιμηθεί η ανταμοιβή κάνοντας κάποια ενέργεια. Η κατάσταση – ενέργεια τιμή-Q στην DDQN διασπάται σε μια συνάρτηση τιμής που αντιπροσωπεύει την ανταμοιβή στην τρέχουσα κατάσταση και στη συνάρτηση πλεονεκτήματος που μετρά τη σχετική σημασία μιας συγκεκριμένης ενέργειας σε σύγκριση με άλλες ενέργειες. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο σχήμα DQL ξεπερνά το υπάρχον στατικό σχήμα ως προς τη συνολική χρησιμότητα. Συγκεκριμένα, η συνολικό ωφελιμότητα που λαμβάνεται από το σχήμα DQL είναι 8000, ενώ αυτή που λαμβάνεται από το υπάρχον στατικό σχήμα είναι 5000.

Λαμβάνοντας υπόψη τον τεράστιο χώρο δράσης και την υψηλή πολυπλοκότητα, λόγω της κινητικότητας του οχήματος και τις προθεσμίες καθυστέρησης εξυπηρέτησης T_d, προτείνεται ένα πλαίσιο DQN πολλαπλής κλίμακας (Tan & Hu, 2018) για την ελαχιστοποίηση του κόστους του συστήματος με τον κοινό σχεδιασμό επικοινωνίας, προσωρινής αποθήκευσης και υπολογισμού στο VANET. Ο σχεδιασμός πολιτικής

ευθύνεται για περιορισμένη χωρητικότητα αποθήκευσης και υπολογιστικούς πόρους στα οχήματα και τις RSUs. Η μικρή χρονική κλίμακα DQN είναι για κάθε χρονοθυρίδα και στοχεύει στη μεγιστοποίηση της ακριβούς άμεσης ανταμοιβής. Επιπλέον, η μεγάλη χρονική κλίμακα DQN έχει σχεδιαστεί για κάθε χρονοθυρίδα T_d εντός της προθεσμίας καθυστέρησης υπηρεσίας και χρησιμοποιείται για την εκτίμηση της ανταμοιβής λαμβάνοντας υπόψη την κινητικότητα του οχήματος σε μεγάλο χρονικό διάστημα. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο πλαίσιο μπορεί να μειώσει το κόστος έως και 30% σε σύγκριση με το σχήμα τυχαίας κατανομής πόρων.

Το προαναφερθέν πλαίσιο DQL για VANET, π.χ. (He, et al., 2017), (He, Yu, Zhao, Yin, & Boukerche, Deep Reinforcement Learning (DRL)-based Resource Management in Software-Defined and Virtualized Vehicular Ad Hoc Networks, 2017), (He, Zhao, & Yin, Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach, 2017), έχει επίσης γενικευτεί σε εφαρμογές έξυπνων πόλεων (He, Yu, Zhao, Leung, & Yin, Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach, 2017), κάτι που απαιτεί δυναμικό συντονισμό δικτύωσης, προσωρινής αποθήκευσης και υπολογισμού για την κάλυψη διαφορετικών απαιτήσεων συντήρησης. Μέσω του Network Function Virtualization (NFV) (Han, Gopalakrishnan, Ji, & Lee, 215), το φυσικό ασύρματο δίκτυο σε έξυπνες πόλεις μπορεί να διαιρεθεί λογικά σε πολλά εικονικά από τον χειριστή του δικτύου, ο οποίος είναι υπεύθυνος για τον τεμαχισμό δικτύου και τον προγραμματισμό πόρων, καθώς και για την κατανομή χωρητικότητας προσωρινής αποθήκευσης και υπολογισμού . Οι περιπτώσεις χρήσης σε έξυπνες πόλεις παρουσιάζονται στα (He, Yu, Zhao, & Yin, Secure Social Networks in 5G Systems with Mobile Edge Computing, Caching, and Device-to-Device Communications, 2018), (He, Liang, Yu, & Han, Trust-Based Social Networks with Computing, Caching and Communications: A Deep Reinforcement Learning Approach, 2020), που εφαρμόζουν το γενικευμένο πλαίσιο DQL για τη βελτίωση της ασφάλειας και της αποτελεσματικότητας για την ανταλλαγή δεδομένων, την κοινή χρήση και την παράδοση σε κινητά κοινωνικά δίκτυα μέσω της κατανομής πόρων και βελτιστοποίησης της κατανομής MEC, προσωρινής αποθήκευσης και D2D (Device-to-Device) δικτύωσης.

Στον Πίνακα 5 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για ασύρματη προσωρινή αποθήκευση.

3.4 Δεδομένα και Υπολογισμός Εκφόρτωσης

Με περιορισμένη υπολογιστική ικανότητα, μνήμη και ενέργεια, συσκευές IoT όπως αισθητήρες και φορητές συσκευές στέκονται εμπόδιο για την υποστήριξη προηγμένων εφαρμογών όπως διαδραστικά διαδικτυακά παιχνίδια και αναγνώριση προσώπων. Για να αντιμετωπιστεί το παραπάνω πρόβλημα, γίνεται μεταφορά των υπολογιστικών εργασιών από τις συσκευές IoT σε κοντινούς διακομιστές MEC, συνδεδυασμένους με BS, σημεία πρόσβασης (Access Points - AP), ακόμη και με γειτονικούς χρήστες κινητών (Mobile Users - MUs). Ως αποτέλεσμα, η εκφόρτωση δεδομένων και υπολογισμών μπορεί ενδεχομένως να μειώσουν την καθυστέρηση επεξεργασίας, να εξοικονομήσουν ενέργεια της μπαταρίας και ακόμη και να ενισχύσουν την ασφάλεια για εφαρμογές IoT που απαιτούν εντατικούς υπολογισμούς. Ωστόσο, το κρίσιμο πρόβλημα στην εκφόρτωση υπολογισμών είναι ο προσδιορισμός του ρυθμού εκφόρτωσης, δηλαδή το ποσό του υπολογιστικού φόρτου εργασίας και η επιλογή του διακομιστή MEC από όλους τους διαθέσιμους διακομιστές. Εάν ο επιλεγμένος διακομιστής MEC αντιμετωπίζει βαρύ φόρτο εργασίας και υποβαθμισμένες συνθήκες καναλιού, ενδέχεται να χρειαστεί ακόμη περισσότερος χρόνος για την εκφόρτωση δεδομένων από τις συσκευές IoT και τη λήψη των αποτελεσμάτων από το διακομιστή MEC. Ως εκ τούτου, ο σχεδιασμός μιας πολιτικής εκφόρτωσης πρέπει να λαμβάνει υπόψη τις εκάστοτε συνθήκες καναλιού, την κινητικότητα του χρήστη, την τροφοδοσία, τον υπολογισμό του φόρτου εργασίας και τις υπολογιστικές δυνατότητες διαφόρων διακομιστών MEC. Παρακάτω παρατίθενται ορισμένες βέλτιστες προσεγγίσεις εκφόρτωσης π.χ. ο αλγόριθμος εκφόρτωσης που βασίζεται στον δυναμικό προγραμματισμό και ο ευρετικός αλγόριθμος εκφόρτωσης (Zhang, Gu, Liu, Yamori, & Tanaka, 2018). Ωστόσο, οι προσεγγίσεις υποθέτουν ότι τα πρότυπα κινητικότητας των χρηστών κινητής τηλεφωνίας δίδονται εκ των προτέρων. Χωρίς να είναι γνωστό εκ των προτέρων το μοτίβο κινητικότητας, το DQL μπορεί να χρησιμοποιηθεί για κάθε χρήστη για να βρεθεί η βέλτιστη πολιτική εκφόρτωσης βασισμένη σε προηγούμενες εμπειρίες όπως οι προσεγγίσεις που προτείνονται στα (Zhang, Liu, Gu, Yamori, & Tanaka, 2018) και (Ji, Hui, Tiejun, & Yueming, 2018). Στο (Zhang, Liu, Gu, Yamori, & Tanaka, 2018) επικεντρώνεται η προσπάθεια στην ελαχιστοποίηση του κόστους και της κατανάλωσης ενέργειας του χρήστη κινητής

τηλεφωνίας, εκφορτώνοντας την κίνηση του δικτύου στο WLAN. Κάθε χρήστης κινητής τηλεφωνίας μπορεί να έχει πρόσβαση είτε στο δίκτυο κινητής τηλεφωνίας, είτε στο δωρεάν WLAN, αλλά με διαφορετικό χρηματικό κόστος. Ο χρήστης κινητής τηλεφωνίας πρέπει επίσης να χρεωθεί αναλόγως εάν η μετάδοση δεδομένων δεν ολοκληρωθεί πριν από την προθεσμία. Η απόφαση εκφόρτωσης δεδομένων του χρήστη κινητής τηλεφωνίας μπορεί να μοντελοποιηθεί ως MDP. Η κατάσταση του συστήματος περιλαμβάνει την τοποθεσία του χρήστη του κινητού και το μέγεθος του εναπομείναντος αρχείου όλων των ροών δεδομένων. Ο χρήστης θα επιλέξει τη μετάδοση δεδομένων είτε μέσω WLAN είτε μέσω του δικτύου κινητής τηλεφωνίας, και θα αποφασίζει πώς να κατανέμεται η χωρητικότητα του καναλιού σε ταυτόχρονες ροές. Στο DQL τα Convolutional Neural Networks (CNN) χρησιμοποιούνται για να προβλέψουν μια συνεχή τιμή των εναπομεινάντων δεδομένων του χρήστη. Τα αποτελέσματα προσομοίωσης αποκαλύπτουν ότι το σχήμα που βασίζεται στο DQN ξεπερνά γενικά τον αλγόριθμο δυναμικού προγραμματισμού για το MDP όσον αφορά το κόστος και την κατανάλωση ενέργειας. Συγκεκριμένα, το σύστημα που βασίζεται σε DQL μπορεί να μειώσει την κατανάλωση ενέργειας έως 500 Joules σε σύγκριση με αυτόν του δυναμικού αλγορίθμου προγραμματισμού. Ο λόγος είναι ότι το DQN μπορεί να μάθει από την εμπειρία, ενώ ο αλγόριθμος δυναμικού προγραμματισμού δεν μπορεί να αποκτήσει τη βέλτιστη πολιτική με λανθασμένη πιθανότητα μετάβασης.

Η κατανομή περιορισμένων υπολογιστικών πόρων στο διακομιστή MEC είναι κρίσιμη για την ελαχιστοποίηση κόστους και ενέργειας. Στο (Ji, Hui, Tiejun, & Yueming, 2018) παρουσιάζεται ένα σύστημα κινητής τηλεφωνίας με δυνατότητα MEC, στο οποίο πολλοί χρήστες μπορούν να μεταφέρουν τις υπολογιστικές τους εργασίες μέσω ασύρματων καναλιών σε έναν διακομιστή MEC, που είναι τοποθετημένος δίπλα σε ένα BS. Κάθε χρήστης έχει μια εντατική υπολογιστική εργασία, που χαρακτηρίζεται από τους απαιτούμενους υπολογιστικούς πόρους, τους κύκλους CPU και τη μέγιστη ανεκτή καθυστέρηση. Η χωρητικότητα του διακομιστή MEC είναι περιορισμένη για την κάλυψη του φόρτου εργασιών όλων των χρηστών κινητών. Η κατανομή του εύρους ζώνης μεταξύ των διαφορετικών χρηστών επηρεάζει επίσης τη συνολική απόδοση καθυστέρησης και την κατανάλωση ενέργειας. Το DQL χρησιμοποιείται για την ελαχιστοποίηση του κόστους καθυστέρησης και της κατανάλωσης ενέργειας για όλους τους χρήστες κινητών συσκευών, βελτιστοποιώντας από κοινού την απόφαση εκφόρτωσης και την κατανομή

υπολογιστικών πόρων. Οι καταστάσεις συστήματος περιλαμβάνουν το άθροισμα του κόστους ολόκληρου του συστήματος και τη διαθέσιμη υπολογιστική χωρητικότητα του διακομιστή MEC. Η ενέργεια του BS είναι να προσδιορίσει την κατανομή πόρων και την απόφαση εκφόρτωσης για κάθε χρήστη. Για να περιοριστεί το μέγεθος του χώρου δράσης, πραγματοποιείται ένα βήμα προ-ταξινόμησης για να ελεγχθεί η εφικτότητα του σύνολο των ενεργειών των χρηστών. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο σχήμα μπορεί να μειώσει το συνολικό κόστος έως και 55% σε σύγκριση με τις στρατηγικές στατικής κατανομής.

Σε αντίθεση με το (Ji, Hui, Tiejun, & Yueming, 2018), πολλαπλά BS σε ένα εξαιρετικά πυκνό δίκτυο εξετάζονται στα (Chen, Zhang, Wu, Mao, Ji, & Bennis, Performance Optimization in Mobile-Edge Computing via Deep Reinforcement Learning, 2018) και (Chen, Zhang, Wu, Mao, Ji, & Bennis, Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning, 2019), με στόχο την ελαχιστοποίηση του μακροπρόθεσμου κόστους καθυστέρησης στην εκφόρτωση υπολογισμών. Όλες οι υπολογιστικές εργασίες εκφορτώνονται στον κοινόχρηστο διακομιστή MEC μέσω διαφορετικών BSs. Εκτός από την κατανομή των υπολογιστικών πόρων και τον έλεγχο μετάδοσης, η πολιτική εκφόρτωσης πρέπει επίσης να βελτιστοποιήσει τη σχέση μεταξύ των χρηστών κινητών και των BS. Με δυναμικές συνθήκες δικτύου, η λήψη αποφάσεων των χρηστών κινητής τηλεφωνίας μπορεί να διατυπωθεί ως MDP. Οι καταστάσεις συστήματος είναι οι συνθήκες καναλιού μεταξύ του χρήστη κινητής τηλεφωνίας και των BS, οι καταστάσεις ενέργειας και οι ουρές εργασιών. Η συνάρτηση κόστους ορίζεται ως ένα σταθμισμένο άθροισμα της καθυστέρησης εκτέλεσης, της καθυστέρησης μεταβίβασης των κλήσεων και του κόστους απόρριψης υπολογιστικών εργασιών. Στο (Chen, Zhang, Wu, Mao, Ji, & Bennis, Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning, 2019) προτείνεται αρχικά ένας αλγόριθμος DQL που βασίζεται σε DDQN για να βρεθεί η βέλτιστη πολιτική εκφόρτωσης χωρίς να είναι γνωστή η δυναμική του δικτύου. Αξιοποιώντας την αθροιστική δομή της συνάρτησης χρησιμότητας, η διάσπαση της συνάρτησης Q σε συνδυασμό με το DDQN οδηγεί περαιτέρω σε έναν νέο διαδικτυακό αλγόριθμο DRL που βασίζεται σε SARSA. Τα αριθμητικά πειράματα δείχνουν ότι ο νέος αλγόριθμος επιτυγχάνει μια σημαντική βελτίωση στην απόδοση υπολογιστικής εκφόρτωσης σε σύγκριση με τις βασικές

πολιτικές, π.χ. τον αλγόριθμο DQL που βασίζεται στο DQN και κάποιες ευρετικές στρατηγικές εκφόρτωσης χωρίς εκμάθηση. Η υψηλή πυκνότητα των μικρών σταθμών βάσης (Small Base Stations – SBS) μπορεί να χαλαρώσει την πίεση της εκφόρτωσης δεδομένων σε ώρες αιχμής, αλλά καταναλώνεται μεγάλη ποσότητα ενέργειας σε χρόνο εκτός αιχμής. Στα (Ye & Zhang, 2020), (Li, Gao, Lv, & Lu, 2018) και (Liu, Krishnamachari, Zhou, & Niu, 2018) προτείνεται μια στρατηγική βασισμένη σε DQL για τον έλεγχο της ενεργοποίησης/απενεργοποίησης διαφορετικών SBS για την ελαχιστοποίηση της κατανάλωσης ενέργειας χωρίς να διακυβεύεται η ποιότητα της παροχής υπηρεσίας εκφόρτωσης. Συγκεκριμένα, στο (Ye & Zhang, 2020), το πλαίσιο απόφασης on / off χρησιμοποιεί ένα σχήμα DQL για να προσεγγίσει τις συναρτήσεις πολιτικής και αξίας με μια μέθοδο Actor-Critic. Η ανταμοιβή του πράκτορα DQL ορίζεται ως συνάρτηση κόστους που σχετίζεται με την κατανάλωση ενέργειας, την υποβάθμιση QoS και το κόστος εναλλαγής των SBS. Η προσέγγιση DDPG χρησιμοποιείται επίσης μαζί με ένα σχέδιο ενίσχυσης ενεργειών για την επιτάχυνση της διαδικασίας εκμάθησης. Μέσα από εκτεταμένες αριθμητικές προσομοιώσεις, το προτεινόμενο σχήμα αποδεικνύεται ότι υπερτερεί σε μεγάλο βαθμό έναντι άλλων βασικών μεθόδων τόσο από πλευράς ενεργειακής όσο και υπολογιστικής απόδοσης.

Με παρόμοιο μοντέλο με αυτό στο (Chen, Zhang, Wu, Mao, Ji, & Bennis, Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning, 2019), η εκφόρτωση υπολογισμών βρίσκει μια κατάλληλη εφαρμογή για ανίχνευση κακόβουλου λογισμικού που βασίζεται στο cloud (Wan, Sheng, Li, Xiao, & Du, 2017). Στο (Xiao, Wan, Dai, Du, Chen, & Guizani, 2018) παρουσιάζεται μια ανασκόπηση των μοντέλων απειλών και των λύσεων που βασίζονται σε ενισχυτική μάθηση (Reinforcement Learning – RL) για την ασφάλεια και την προστασία της ιδιωτικότητας για εκφόρτωση και προσωρινή αποθήκευση στην κινητή τηλεφωνία. Με περιορισμένη κατανάλωση ενέργειας, υπολογιστικούς πόρους και χωρητικότητα καναλιού, οι χρήστες κινητών συσκευών δεν μπορούν πάντα να ενημερώσουν την τοπική βάση δεδομένων κακόβουλου λογισμικού και να επεξεργαστούν όλα τα δεδομένα εφαρμογών εγκαίρως και επομένως είναι ευάλωτοι σε επιθέσεις μηδενικής ημέρας (zero-day) (Shamili, Bauckhage, & Alpcan, 2010). Αξιοποιώντας τον απομακρυσμένο διακομιστή MEC, όλοι οι χρήστες κινητών συσκευών μπορούν να εκφορτώσουν τα δεδομένα των εφαρμογών τους και τις εργασίες ανίχνευσης μέσω διαφορετικών BS στον

διακομιστή ασφαλείας MEC, που διαθέτει μεγαλύτερη και πιο εξελιγμένη βάση δεδομένων κακόβουλου λογισμικού, περισσότερες υπολογιστικές δυνατότητες και ισχυρές υπηρεσίες ασφαλείας. Αυτό μπορεί να μοντελοποιηθεί από ένα δυναμικό παιχνίδι ανίχνευσης κακόβουλου λογισμικού στο οποίο πολλοί χρήστες κινητών αλληλεπιδρούν μεταξύ τους στον ανταγωνισμό πόρων, π.χ., η κατανομή της χωρητικότητας των ασύρματων καναλιών και οι υπολογιστικές δυνατότητες του διακομιστή ασφάλειας MEC. Για κάθε χρήστη κινητής συσκευής προτείνεται ένα σχήμα DQL για να ενημερώσει για τον ρυθμό εκφόρτωσης δεδομένων του τον διακομιστή MEC / ασφαλείας. Οι καταστάσεις συστήματος περιλαμβάνουν την κατάσταση καναλιού και το μέγεθος των καταγραφών εφαρμογών. Ο στόχος είναι να βελτιστοποιηθεί η ακρίβεια ανίχνευσης του διακομιστή ασφαλείας, ο οποίος ορίζεται ως κοίλη συνάρτηση στο συνολικό ποσό δειγμάτων κακόβουλου λογισμικού. Η τιμή Q υπολογίζεται χρησιμοποιώντας ένα CNN στο πλαίσιο DQL. Προτείνεται επίσης η τεχνική Q-learning hotbooting που παρέχει καλύτερη αρχικοποίηση για την Q-learning αξιοποιώντας τις εμπειρίες εκφόρτωσης σε παρόμοια σενάρια. Μπορεί να εξοικονομήσει χρόνο εξερεύνησης στο αρχικό στάδιο και να επιταχύνει την ταχύτητα εκμάθησης σε σύγκριση με έναν τυπικό αλγόριθμο Q-learning με αρχικοποίηση all-zero της τιμής Q (Li, Liu, Li, & Xiao, 2015). Το προτεινόμενο σχήμα DQL όχι μόνο βελτιώνει την ταχύτητα και την ακρίβεια ανίχνευσης, αλλά επίσης αυξάνει τη διάρκεια ζωής της μπαταρίας των χρηστών κινητών. Τα αποτελέσματα της προσομοίωσης αποκαλύπτουν ότι, σε σύγκριση με τα hotbooting Q-learning και τα τυπικά σχήματα Q-learning, η ανίχνευση κακόβουλου λογισμικού που βασίζεται σε DQL έχει ταχύτερο ρυθμό εκμάθησης, υψηλότερη ακρίβεια και χαμηλότερη καθυστέρηση εντοπισμού. Για παράδειγμα, η καθυστέρηση ανίχνευσης του προτεινόμενου σχήματος DQL μειώνεται κατά 24,6% και 35,3%, αντίστοιχα στη χρονοθυρίδα 2000, σε σύγκριση με εκείνες του hotbooting Q-learning και τα τυπικά σχήματα Q-learning.

Στο (Min, Xiao, Chen, Cheng, Wu, & Zhuang, 2019) επιδιώκεται ο σχεδιασμός βέλτιστου πολιτικής εκφόρτωσης για συσκευές IoT με δυνατότητες συλλογής ενέργειας. Το σύστημα αποτελείται από πολλούς διακομιστές MEC, όπως BS και AP, με διαφορετικές υπολογιστικές και επικοινωνιακές δυνατότητες. Οι συσκευές IoT είναι εξοπλισμένες με αποθήκες και συλλέκτες ενέργειας. Μπορούν να εκτελέσουν υπολογιστικές εργασίες τοπικά και να εκφορτώσουν τις εργασίες στους διακομιστές MEC. Η απόφαση εκφόρτωσης των συσκευών IoT μπορεί να διατυπωθεί ως MDP. Οι καταστάσεις

συστήματος περιλαμβάνουν την κατάσταση της μπαταρίας, την χωρητικότητα του καναλιού και την προβλεπόμενη μελλοντική ποσότητα της συλλεγόμενης ενέργειας. Η συσκευή IoT αξιολογεί την ανταμοιβή με βάση τη συνολική καθυστέρηση, την κατανάλωση ενέργειας, τη ζημιά λόγω απόρριψης εργασιών και τα κέρδη κοινής χρήσης δεδομένων σε κάθε χρονικό διάστημα. Παρόμοια με το (Wan, Sheng, Li, Xiao, & Du, 2017), στο (Min, Xiao, Chen, Cheng, Wu, & Zhuang, 2019) ενισχύεται η Q-learning με την τεχνική hotbooting για να εξοικονομηθεί ο χρόνος τυχαίας εξερεύνησης στην αρχή της μάθησης. Προτείνεται επίσης ένα γρήγορο DQL σχήμα εκφόρτωσης που χρησιμοποιεί τη τεχνική hotbooting για να αρχικοποιήσει το CNN και να επιταχύνει την ταχύτητα εκμάθησης. Στο (Quan, Wang, & Ren, 2018) αντιμετωπίζονται τα BS με δυνατότητα MEC ως διαφορετικά φυσικά μηχανήματα που αποτελούν μέρος των πόρων cloud. Το cloud βελτιστοποιεί την εκφόρτωση υπολογισμών των κινητών χρηστών σε διαφορετικές εικονικές μηχανές που ενυπάρχουν σε φυσικές μηχανές. Ένας αλγόριθμος DQL δύο επιπέδων προτείνεται για το πρόβλημα εκφόρτωσης για τη μεγιστοποίηση της χρήσης των πόρων cloud. Η κατάσταση του συστήματος σχετίζεται με τον χρόνο αναμονής κάθε υπολογιστικής εργασίας και τον αριθμό των εικονικών μηχανών. Το πρώτο επίπεδο εφαρμόζεται από ένα πλαίσιο DQL που βασίζεται στο CNN για να εκτιμήσει ένα βέλτιστο σύμπλεγμα (cluster) για κάθε υπολογιστική εργασία. Δημιουργούνται διαφορετικά σύνολα φυσικών μηχανών με βάση τον αλγόριθμο K-NN. Το δεύτερο επίπεδο καθορίζει τη βέλτιστη φυσική μηχανή εξυπηρέτησης εντός του συμπλέγματος με τη μέθοδο Q-learning.

Τα προαναφερθέντα έργα εστιάζουν σε εκφόρτωση δεδομένων ή υπολογισμών σε κυψελοειδή συστήματα μέσω BS σε απομακρυσμένους διακομιστές MEC, π.χ. (Zhang, Liu, Gu, Yamori, & Tanaka, 2018), (Ji, Hui, Tiejun, & Yueming, 2018), (Chen, Zhang, Wu, Mao, Ji, & Bennis, Performance Optimization in Mobile-Edge Computing via Deep Reinforcement Learning, 2018), (Chen, Zhang, Wu, Mao, Ji, & Bennis, Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning, 2019), (Wan, Sheng, Li, Xiao, & Du, 2017), (Min, Xiao, Chen, Cheng, Wu, & Zhuang, 2019), (Quan, Wang, & Ren, 2018). Στα (Le & Tham, A deep reinforcement learning based offloading scheme in ad-hoc mobile clouds, 2018) και (Le & Tham, Quality of Service Aware Computation Offloading in an Ad-Hoc Mobile Cloud, 2018), μελετάται η εκφόρτωση υπολογισμών με γνώμονα τη QoS σε ένα ad-hoc δίκτυο κινητής τηλεφωνίας. Κάνοντας μια συγκεκριμένη πληρωμή, ο χρήστης κινητής τηλεφωνίας μπορεί να

εκφορτώσει τις υπολογιστικές του εργασίες σε κοντινούς χρήστες κινητής τηλεφωνίας που αποτελούν ένα κινητό cloudlet. Κάθε χρήστης κινητής τηλεφωνίας έχει μια ουρά πρώτος μέσα πρώτος έξω (First In First Out – FIFO) με περιορισμένο μέγεθος ενδιάμεσης μνήμης για την αποθήκευση των εργασιών άφιξης που φθάνουν ως διαδικασία Poisson. Ο χρήστης κινητής τηλεφωνίας επιλέγει κοντινά cloudlets εντός του εύρους επικοινωνίας Device to Device (D2D) για εργασίες εκφόρτωσης. Η απόφαση εκφόρτωσης εξαρτάται από τις καταστάσεις που περιλαμβάνουν τον αριθμό των υπολειπόμενων εργασιών, την ποιότητα των συνδέσεων μεταξύ χρηστών κινητής τηλεφωνίας και του cloudlet και τη διαθεσιμότητα των πόρων του cloudlet. Ο στόχος είναι η μεγιστοποίηση μιας σύνθετης συνάρτησης χρησιμότητας, με την επιφύλαξη των απαιτήσεων QoS του χρήστη κινητού, π.χ. κατανάλωση ενέργειας και καθυστέρηση επεξεργασίας. Η συνάρτηση χρησιμότητας είναι αρχικά μια αυξανόμενη συνάρτηση του συνολικού αριθμού εργασιών που έχουν επεξεργαστεί τοπικά ή εξ αποστάσεως από τα cloudlets. Συνδέεται επίσης με τα οφέλη του χρήστη, όπως η ενεργειακή απόδοση και η πληρωμή για την εκφόρτωση των εργασιών. Αυτό το πρόβλημα διατυπώνεται ως MDP και επιλύεται με γραμμικό προγραμματισμό και προσεγγίσεις Q-learning, βασιζόμενη στη διαθεσιμότητα πληροφοριών σχετικά με τις πιθανότητες μετάβασης κατάστασης. Αυτή η εργασία ενισχύεται περαιτέρω με τη χρήση του DNN ή του DQN για να βρεθεί αποτελεσματικότερα η στρατηγική λήψης αποφάσεων. Ένα παρόμοιο μοντέλο μελετάται στο (Yu, Wang, & Langar, 2017), όπου η εκφόρτωση υπολογισμών διαμορφώνεται ως MDP για την ελαχιστοποίηση του κόστους της. Η λύση στο MDP μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός DNN μέσω εποπτευόμενης μάθησης. Το καλά εκπαιδευμένο DNN εφαρμόζεται στη συνέχεια σε αόρατες συνθήκες δικτύου για τη λήψη αποφάσεων σε πραγματικό χρόνο. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι η χρήση της βαθιάς εποπτευόμενης μάθησης επιτυγχάνει σημαντικό κέρδος απόδοσης με σημαντική ακρίβεια και εξοικονόμηση κόστους.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Zhang, Liu, Gu, Yamori, & Tanaka, 2018)	MDP	DQN με χρήση CNN	Κινητός χρήστης	Θέση του χρήστη και εναπομείναν μέγεθος αρχείου	Αδράνεια, μετάδοση μέσω WLAN ή κυψελοειδούς δικτύου	Συνολικός ρυθμός δεδομένων	Κυψελωτό σύστημα
(Ji, Hui, Tiejun, & Yueming, 2018)	MDP	DQN με χρήση FNN	Σταθμός Βάσης	Άθροισμα κόστους και υπολογιστική χωρητικότητα του διακομιστή MEC	Απόφαση εκφόρτωσης δεδομένων και κατανομής πόρων	Άθροισμα κόστους καθυστέρησης και κατανάλωσης ενέργειας	Κυψελωτό σύστημα

(Chen, Zhang, Wu, Mao, Ji, & Bennis, 2018)	MDP	DQN με χρήση FNN	Κινητός χρήστης	Ποιότητα καναλιού, καταστάσεις ενέργειας και ουρές εργασιών	Εκφόρτωση δεδομένων και κατανομή πόρων	Μακροπρόθεσμη συνάρτηση κόστους	Κυψελωτό σύστημα
(Chen, Zhang, Wu, Mao, Ji, & Bennis, 2019)	MDP	DDQN, SARSA	Κινητός χρήστης	Ποιότητα καναλιού, καταστάσεις ενέργειας και ουρές εργασιών	Απόφαση εκφόρτωσης δεδομένων και υπολογιστική κατανομή πόρων	Μακροπρόθεσμη συνάρτηση κόστους	Κυψελωτό σύστημα
(Wan, Sheng, Li, Xiao, & Du, 2017)	Παίγνιο	DQN με χρήση CNN, hotbooting Q-learning	Κινητός χρήστης	Καταστάσεις καναλιού, μέγεθος ιχνών εφαρμογών	Ρυθμός εκφόρτωσης δεδομένων	Βοηθητικό πρόγραμμα που σχετίζεται με την ακρίβεια ανίχνευσης, την ταχύτητα απόκρισης και το κόστος μετάδοσης	Κυψελωτό σύστημα
(Tang, Zhou, Zhang, Jia, & Zhao, 2019)	MDP	DDQN	Κόμβος fog	Καθυστέρηση, θέση περιέκτη και κατανομή πόρων	Επόμενη θέση του περιέκτη	Σύνθετο βοηθητικό πρόγραμμα που σχετίζεται με καθυστέρηση, κατανάλωση ενέργειας και κόστος μετανάστευσης	Fog computing

Πίνακας 6. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για εκφόρτωση δεδομένων

Η εκφόρτωση δεδομένων και υπολογισμών χρησιμοποιούνται επίσης στον fog computing. Η εφαρμογή για κινητά που απαιτεί ένα σύνολο δεδομένων και υπολογιστικών πόρων μπορεί να φιλοξενηθεί σε ένα κοντέινερ, π.χ. εικονική μηχανή ενός κόμβου fog. Με την κινητικότητα του χρήστη, το κοντέινερ πρέπει να μετεγκατασταθεί ή να εκφορτωθεί σε άλλους κόμβους και να ενοποιηθεί δυναμικά. Με τη μετεγκατάσταση κοντέινερ, ορισμένοι κόμβοι με χαμηλή χρήση πόρων μπορούν να απενεργοποιηθούν για να μειωθεί η κατανάλωση ενέργειας. Στο (Tang, Zhou, Zhang, Jia, & Zhao, 2019) μοντελοποιείται η μετεγκατάσταση κοντέινερ ως πολυδιάστατο MDP, το οποίο επιλύεται από το DQL. Οι καταστάσεις του συστήματος αποτελούνται από την καθυστέρηση, την κατανάλωση ενέργειας και το κόστος μετεγκατάστασης. Η ενέργεια περιλαμβάνει την πολιτική επιλογής που επιλέγει τα κοντέινερ προς μετεγκατάσταση από κάθε κόμβο προέλευσης και την πολιτική κατανομής που καθορίζει τον κόμβο προορισμού κάθε κοντέινερ. Ο χώρος δράσης μπορεί να βελτιστοποιηθεί για πιο αποτελεσματική εξερεύνηση διαιρώντας τους κόμβους fog σε ομάδες, κανονικής χρήσης και υπερχρησιμοποίησης. Με την απενεργοποίηση των κόμβων υποαξιοποίησης, όλα τα κοντέινέρ τους θα μεταφερθούν σε άλλους κόμβους για τη μείωση της κατανάλωσης ενέργειας. Η εκπαιδευτική διαδικασία βελτιστοποιείται επίσης χρησιμοποιώντας DDQN και Prioritized Experience Replay (PER) που εκχωρεί διαφορετικές προτεραιότητες στις μεταβάσεις στη μνήμη εμπειρίας. Αυτό βοηθά τον παράγοντα DQL να αποδίδει καλύτερα

σε κάθε κόμβο fog, όσον αφορά την μεγαλύτερη ταχύτητα εκμάθησης και περισσότερης σταθερότητας. Η ανάλυση δείχνει ότι το προτεινόμενο σχήμα μπορεί να εκτελεστεί σε πολυωνυμικό χρόνο. Τα αποτελέσματα της προσομοίωσης αποκαλύπτουν ότι το σχήμα DQL επιτυγχάνει γρήγορη λήψη αποφάσεων και ξεπερνά σημαντικά τις υπάρχουσες βασικές προσεγγίσεις όσον αφορά την καθυστέρηση, την κατανάλωση ενέργειας και το κόστος μετεγκατάστασης.

Στον Πίνακα 6 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για εκφόρτωση δεδομένων.

3.5 Ασφάλεια Δικτύου

Αυτή η ενότητα περιγράφει τις εφαρμογές της DQL για την αντιμετώπιση της επίθεσης παρεμβολών και κυβερνοφυσικής.

3.5.1 Επίθεση Παρεμβολών

Το μοντέλο δικτύου που μελετάται (Han, Xiao, & Poor, 2017) είναι ένα Cognitive Radio Network (CRN) που αποτελείται από έναν Δευτερεύοντα Χρήστη (Secondary User - SU), πολλούς Κύριους Χρήστες (Primary Users - PU) και πολλούς παρεμβολείς (jammers). Το δίκτυο διαθέτει ένα σύνολο καναλιών συχνότητας για μετάβαση. Σε κάθε χρονική στιγμή, κάθε παρεμβολέας μπορεί αυθαίρετα να επιλέξει ένα από τα κανάλια για να στείλει το σήμα παρεμβολής του, ενώ το SU, δηλαδή, ο πράκτορας, πρέπει να επιλέξει μια σωστή ενέργεια με βάση την τρέχουσα κατάσταση του SU. Η ενέργεια είναι (i) να επιλεγεί ένα από τα κανάλια για να αποσταλούν τα σήματά του ή (ii) να εγκαταλειφθεί η περιοχή για τη σύνδεση με άλλο BS. Οι παρεμβολείς θεωρείται ότι αποφεύγουν να προκαλέσουν παρεμβολές στα PU. Η τρέχουσα κατάσταση του SU αποτελείται από τον αριθμό των PU και το διακριτό SINR του σήματος SU στην τελευταία χρονική στιγμή. Ο στόχος του SU είναι να μεγιστοποιήσει την αναμενόμενη μειωμένη χρησιμότητα με την πάροδο του χρόνου. Σημειώνεται ότι όταν το SU επιλέγει να εγκαταλείψει την περιοχή για να συνδεθεί με άλλο BS, πληρώνει ένα κόστος κινητικότητας. Έτσι, η χρησιμότητα ορίζεται ως συνάρτηση του SINR του σήματος SU και του κόστους κινητικότητας. Δεδομένου ότι ο αριθμός των καναλιών συχνότητας μπορεί να είναι μεγάλος που οδηγεί σε ένα μεγάλο σύνολο ενεργειών, το CNN χρησιμοποιείται για την DQL για να μάθει γρήγορα τη βέλτιστη πολιτική. Όπως φαίνεται στα αποτελέσματα της προσομοίωσης, η προτεινόμενη DQL έχει

ταχύτερη ταχύτητα σύγκλισης από αυτήν του αλγορίθμου Q-learning. Συγκεκριμένα, η χρησιμότητα του SU αυξάνεται από 2,73 στην αρχή σε 3,39 στην χρονική υποδοχή 1000 που είναι 8,3% υψηλότερη από αυτήν του αλγορίθμου Q-learning. Επιπλέον, λαμβάνοντας υπόψη το σενάριο με δύο παρεμβολείς, το προτεινόμενο DQL ξεπερνά τη μέθοδο μεταπήδησης συχνότητας σε σχέση με το SINR και το κόστος κινητικότητας.

Το μοντέλο στο (Han, Xiao, & Poor, 2017) περιορίζεται σε δύο παρεμβολείς. Καθώς ο αριθμός των παρεμβολών στο δίκτυο αυξάνεται, το προτεινόμενο σχήμα ενδέχεται να μην είναι αποτελεσματικό. Ο λόγος είναι ότι γίνεται δύσκολο για το SU να βρει κατάλληλες ενέργειες όταν αυξάνεται ο αριθμός των μπλοκαρισμένων καναλιών. Μια κατάλληλη λύση, όπως προτείνεται στο (Xiao, Jiang, Wan, Su, & Tang, 2018), επιτρέπει στον δέκτη του SU να εγκαταλείψει την τρέχουσα θέση του. Δεδομένου ότι η αποχώρηση επιβαρύνεται με το κόστος κινητικότητας, ο παραλήπτης, δηλ. ο πράκτορας, χρειάζεται μια βέλτιστη πολιτική, δηλαδή να παραμένει ή να αποχωρεί από την τρέχουσα τοποθεσία, για να μεγιστοποιήσει τη χρησιμότητά του. Σε αυτό το σενάριο, το DQL που βασίζεται στο CNN μπορεί να χρησιμοποιηθεί για τον δέκτη για να βρει τη βέλτιστη ενέργεια για να μεγιστοποιήσει την αναμενόμενη χρησιμότητά του. Εδώ, η χρησιμότητα και η κατάσταση του δέκτη καθορίζονται ουσιαστικά παρόμοια με εκείνη του πράκτορα στο (Han, Xiao, & Poor, 2017). Συγκεκριμένα, η κατάσταση περιλαμβάνει το διακριτό SINR του σήματος που μετράται από τον δέκτη στην τελευταία χρονική υποδοχή. Τα αποτελέσματα προσομοίωσης δείχνουν ότι η προτεινόμενη DQL συγκλίνει σε SINR και τιμές χρησιμότητας που είναι υψηλότερες από αυτές που λαμβάνονται από την Q-learning και τυχαία σχήματα. Συγκεκριμένα, η τιμή SINR που λαμβάνεται από την προτεινόμενη DQL είναι 3,4, ενώ εκείνες που λαμβάνονται από τα Q-learning και τυχαία σχήματα είναι 3,3 και 2,8, αντίστοιχα.

Οι παραπάνω προσεγγίσεις, δηλ. στα (Han, Xiao, & Poor, 2017) και (Xiao, Jiang, Wan, Su, & Tang, 2018), καθορίζουν καταστάσεις των παραγόντων με βάση τις αρχικές τιμές SINR των σημάτων. Σε πρακτικά ασύρματα περιβάλλοντα, ο αριθμός των τιμών SINR μπορεί να είναι μεγάλος και ακόμη και άπειρο. Επιπλέον, το ακατέργαστο SINR μπορεί να είναι ανακριβές και θορυβώδες. Για να αντιμετωπιστεί η πρόκληση του άπειρου αριθμού καταστάσεων, η DQL μπορεί να χρησιμοποιήσει ένα Recursive Convolutional Neural Network (RCNN) (Liu, Xu, Jia, Wu, & Anpalagan, 2018). Χρησιμοποιώντας το

προεπεξεργασμένο επίπεδο και τα αναδρομικά συνελικτικά επίπεδα, το RCNN είναι σε θέση να απομακρύνει τον θόρυβο από το περιβάλλον του δικτύου και να εξαγάγει χρήσιμα χαρακτηριστικά του SINR, όπως διακριτές τιμές δείγματος φάσματος μεγαλύτερες από ένα όριο θορύβου, μειώνοντας έτσι την υπολογιστική πολυπλοκότητα. Το μοντέλο του δικτύου και η διατύπωση του προβλήματος που εξετάζονται στο (Liu, Xu, Jia, Wu, & Anpalagan, 2018) είναι παρόμοια με αυτά του (Han, Xiao, & Poor, 2017). Ωστόσο, αντί να χρησιμοποιεί απευθείας το ακατέργαστο SINR, η κατάσταση του SU είναι τα εξαγόμενα χαρακτηριστικά του SINR. Επίσης, η ενέργεια του SU περιλαμβάνει μόνο τη λήψη συχνότητας. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι το προτεινόμενο DQL που βασίζεται στο RCNN μπορεί να συγκλίνει τόσο σε σταθερά όσο και σε δυναμικά σενάρια παρεμβολών, ενώ το Q-learning δεν μπορεί να συγκλίνει σε δυναμικά σενάρια. Επιπλέον, η προτεινόμενη DQL μπορεί να επιτύχει τη μέση απόδοση που πλησιάζει εκείνη του βέλτιστου σχήματος, δηλ. Ένα αντιπαρεμβολικό σχήμα με πλήρως γνωστές ενέργειες παρεμβολής.

Στο (Chen, Li, Xu, & Xiao, 2018) προτείνεται η χρήση του DQL για να βρεθεί μια βέλτιστη πολιτική ελέγχου ισχύος για την αντιμετώπιση της παρεμβολής. Το μοντέλο είναι ένα δίκτυο IoT που περιλαμβάνει συσκευές IoT και έναν παρεμβολέα. Ο παρεμβολέας μπορεί να παρατηρήσει τις επικοινωνίες του πομπού και να επιλέγει μια στρατηγική παρεμβολής για τη μείωση του SINR στον δέκτη. Έτσι, ο πομπός επιλέγει μια ενέργεια, δηλαδή, επίπεδο ισχύος μετάδοσης, για να μεγιστοποιήσει τη χρησιμότητά του. Εδώ, η χρησιμότητα είναι η διαφορά μεταξύ του SINR και του κόστους κατανάλωσης ενέργειας λόγω της μετάδοσης. Σημειώστε ότι η επιλογή της ισχύος μετάδοσης επηρεάζει τη μελλοντική στρατηγική παρεμβολής, και έτσι η αλληλεπίδραση μεταξύ του πομπού και του παρεμβολέα μπορεί να διατυπωθεί ως MDP. Ο πομπός είναι ο πράκτορας και η κατάσταση το SINR που μετράται στον δέκτη την τελευταία χρονική υποδοχή. Το DQN που χρησιμοποιεί το CNN στη συνέχεια υιοθετείται για να βρει μια βέλτιστη πολιτική ελέγχου ισχύος για τον πομπό για τη μεγιστοποίηση της αναμενόμενης συσσωρευμένης μειωμένης ανταμοιβής, δηλαδή της χρησιμότητας, με την πάροδο του χρόνου. Τα αποτελέσματα προσομοίωσης δείχνουν ότι το προτεινόμενο DQL μπορεί να βελτιώσει τη χρησιμότητα του πομπού έως και 17,7% σε σύγκριση με το Q-learning. Επίσης, η προτεινόμενη DQL μειώνει τη χρησιμότητα του παρεμβολέα περίπου 18,1% σε σύγκριση με το Q-learning. Επιπλέον, η προτεινόμενη DQL έχει ταχύτερη ταχύτητα σύγκλισης από

αυτήν της Q-learning. Συγκεκριμένα, η προτεινόμενη DQL συγκλίνει στην υποδοχή χρόνου 210, ενώ η Q-learning συγκλίνει στην υποδοχή χρόνου 240.

Τα αποτελέσματα της προσομοίωσης στο (Xiao, Xie, Min, & Zhuang, 2018) δείχνουν ότι το προτεινόμενο DQL μπορεί να βελτιώσει τη χρησιμότητα του UAV έως και 13% σε σύγκριση με το βασικό σχήμα (Bowling & Veloso, 2002) που χρησιμοποιεί το Win or Learn Faster-Policy Hill Climbing (WoLF-PHC) για να αποτρέψει την επίθεση. Επίσης, ο ασφαλής ρυθμός του UAV, δηλαδή η πιθανότητα επίθεσης του UAV, που λαμβάνεται από την προτεινόμενη DQL είναι 7% υψηλότερη από εκείνη της γραμμής βάσης. Ωστόσο, η προτεινόμενη DQL έχει υψηλότερη υπολογιστική πολυπλοκότητα και χρειάζεται περισσότερο χρόνο για να ληφθεί απόφαση σε σύγκριση με το WoLH-PHC. Έτσι, η προτεινόμενη DQL εφαρμόζεται μόνο σε ένα σύστημα UAV. Για μελλοντική εργασία, πρέπει να ληφθούν υπόψη σενάρια με πολλαπλά UAV. Σε ένα τέτοιο σενάριο, αναμένεται περισσότερη υπολογιστική επιβάρυνση και μπορούν να εφαρμοστούν αλγόριθμοι DQL πολλαπλών παραγόντων.

3.5.2 Κυβερνοφυσική επίθεση

Σε αυτόνομα συστήματα όπως τα ITS, ο εισβολέας μπορεί να επιδιώξει τη μετάδοση ψευδών δεδομένων σε πληροφορίες που μεταδίδονται από τους αισθητήρες στα AV. Τα AV που λαμβάνουν τις ψευδείς πληροφορίες μπορεί να εκτιμήσουν ανακριβώς την ασφαλή απόσταση μεταξύ τους. Αυτό αυξάνει τον κίνδυνο ατυχημάτων. Οι αλγόριθμοι ασφάλειας επικοινωνίας των οχημάτων, π.χ. (Chen, Kar, & Moura, 2018), μπορούν να χρησιμοποιηθούν για την ελαχιστοποίηση της απόκλισης απόστασης. Ωστόσο, οι ενέργειες του επιτιθέμενου σε αυτούς τους αλγόριθμους θεωρείται ότι είναι σταθερές και μπορεί να μην εφαρμόζονται σε πρακτικά συστήματα. Συνεπώς μπορεί να χρησιμοποιηθεί η DQL που επιτρέπει στα AVs να μάθουν τις βέλτιστες ενέργειες με βάση τις ποικίλες παρελθούσες παρατηρήσεις των ενεργειών του επιτιθέμενου.

Η πρώτη εργασία που χρησιμοποιεί το DQL για την κυβερνοφυσική επίθεση σε ένα ITS μπορεί να βρεθεί στο (Ferdowsi, Challita, Saad, & Mandayam, 2018). Το σύστημα είναι ένα μοντέλο car-following (Brackstone & McDonald, 1999) της General Motors,. Στο μοντέλο,

κάθε AV ελέγχει την ταχύτητά του με βάση τις πληροφορίες μέτρησης που λαμβάνονται από τους πιο κοντινούς έξυπνους αισθητήρες δρόμου. Ο επιτιθέμενος προσπαθεί να εισάγει ψευδή δεδομένα, ως πληροφορίες μέτρησης. Ωστόσο, ο επιτιθέμενος δεν μπορεί να επηρεάσει εξίσου τους διαφορετικούς αισθητήρες λόγω του περιορισμού των πόρων του. Έτσι, το AV μπορεί να επιλέξει τις λιγότερο ψευδείς μετρήσεις επιλέγοντας έναν παράγοντα μέτρησης βαρών. Ο στόχος του επιτιθέμενου είναι να μεγιστοποιήσει την απόκλιση, δηλαδή, τη χρησιμότητα, από την ασφαλή απόσταση μεταξύ του AV και του κοντινού AV, ενώ ο στόχος του AV είναι η ελαχιστοποίηση της απόκλισης. Η αλληλεπίδραση μεταξύ του επιτιθέμενου και του AV μπορεί να μοντελοποιηθεί ως παιχνίδι μηδενικού αθροίσματος. Στο (Ferdowsi, Challita, Saad, & Mandayam, 2018) φαίνεται ότι το DQL μπορεί να χρησιμοποιηθεί για να βρει τις στρατηγικές ισορροπίας. Συγκεκριμένα, η ενέργεια του AV είναι η επιλογή ενός παράγοντα βάρους. Η κατάστασή του περιλαμβάνει τις παρελθούσες ενέργειες, δηλαδή, τους παράγοντες βάρους και τις τιμές παρεκκλίσεων. Δεδομένου ότι οι ενέργειες και οι αποκλίσεις έχουν συνεχείς τιμές, ο χώρος κατάστασης είναι άπειρος. Έτσι, οι μονάδες LSTM που είναι σε θέση να εξαγάγουν χρήσιμα χαρακτηριστικά υιοθετούνται για το DQL για τη μείωση του χώρου κατάστασης. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι χρησιμοποιώντας τις προηγούμενες ενέργειες και αποκλίσεις για την εκμάθηση της ενέργειας του επιτιθέμενου, το προτεινόμενο σχήμα DQL μπορεί να εγγυηθεί μια χαμηλότερη απόκλιση σταθερής κατάστασης από το σύστημα βασισμένο στο φίλτρο Kalmar (Chen, Kar, & Moura, 2018) Επιπλέον, με τη χρήση των μονάδων LSTM, τα αποτελέσματα δείχνουν ότι το προτεινόμενο σχήμα DQL μπορεί να συγκλίνει πολύ πιο γρήγορα από το βασικό σχήμα.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Han, Xiao, & Poor, 2017)	Παίγνιο	DQN με χρήση CNN	Δευτερεύων χρήστης	Αριθμός πρωτευόντων χρηστών και σηματοθορυβικός λόγος σήματος	Επιλογή καναλιού και απόφαση αποχώρησης	Σηματοθορυβικός λόγος και κόστος κινητικότητας	CRN
(Xiao, Jiang, Wan, Su, & Tang, 2018)	Παίγνιο	DQN με χρήση CNN	Μετατροπέας λήψης	Σηματοθορυβικός λόγος σήματος	Αποφάσεις παραμονής ή αποχώρησης	Σηματοθορυβικός λόγος και κόστος κινητικότητας	Υποβρύχιο ακουστικό δίκτυο
(Liu, Xu, Jia, Wu, & Anpalagan, 2018)	MDP	DQN με χρήση RCNN	Δευτερεύων χρήστης	Σηματοθορυβικός λόγος σήματος	Επιλογή καναλιού	Σηματοθορυβικός λόγος και κόστος κινητικότητας	CRN
(Chen, Li, Xu, & Xiao, 2018)	MDP	DQN με χρήση CNN	IoT συσκευή εκπομπής	Σηματοθορυβικός λόγος σήματος	Επιλογή καναλιού	Σηματοθορυβικός λόγος και κόστος κατανάλωσης ενέργειας	IoT
(Ferdowsi, Challita, Saad, &	Παίγνιο	DQN με χρήση	Αυτόνομο όχημα	Τιμές απόκλισης	Επιλογή μέτρησης βάρους	Ασφαλής απόκλιση απόστασης	ITS

Mandayam , 2018)		μονάδων LSTM					
(Ferdowsi & Saad, 2019)	Παίγνιο	DQN με χρήση μονάδων LSTM	Cloud	Ενέργειες επίθεσης σε συσκευές IoT	Επιλογή ομάδας συσκευών IoT	Τιμές δεδομένων συσκευών IoT	IoT

Πίνακας 7. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για ασφάλεια δικτύου

Μια άλλη εργασία που χρησιμοποιεί το LSTM για την εξαγωγή χρήσιμων χαρακτηριστικών από τις πληροφορίες μέτρησης για τον εντοπισμό της κυβερνοφυσικής επίθεσης προτείνεται στο (Ferdowsi & Saad, Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in the Internet of Things, 2018). Το μοντέλο είναι ένα σύστημα IoT που περιλαμβάνει ένα cloud και ένα σύνολο συσκευών IoT. Οι συσκευές IoT παράγουν σήματα και μεταδίδουν τα σήματα στο cloud. Το cloud χρησιμοποιεί τα ληφθέντα σήματα για εκτίμηση και έλεγχο της λειτουργίας των συσκευών IoT. Ένας επιτιθέμενος μπορεί να εξαπολύσει την κυβερνοφυσική επίθεση επηρεάζοντας τα σήματα εξόδου των συσκευών IoT που προκαλούν σφάλματα ελέγχου στο cloud και υποβαθμίζουν την απόδοση του συστήματος IoT. Για να εντοπίσει την επίθεση, το cloud χρησιμοποιεί μονάδες LSTM για την εξαγωγή στοχαστικών χαρακτηριστικών ή δακτυλικών αποτυπωμάτων όπως η επιπεδότητα, η λοξότητα και η κύρτωση των σημάτων των συσκευών IoT. Το cloud στέλνει τα αποτυπώματα πίσω στις συσκευές IoT και οι συσκευές IoT ενσωματώνουν, δηλ. υδατογράφημα, τα αποτυπώματα εντός των σημάτων. Το cloud χρησιμοποιεί τα δακτυλικά αποτυπώματα για τον έλεγχο ταυτότητας των σημάτων των συσκευών IoT για τον εντοπισμό της επίθεσης. Η υπολογιστική πολυπλοκότητα της προτεινόμενης μεθόδου ελέγχου ταυτότητας σήματος είναι O (df_s^i), όπου d είναι η καθυστέρηση που το cloud επικυρώνει οποιοδήποτε σήμα IoT, και το f_s^i είναι ο ρυθμός δειγματοληψίας της συσκευής IoT (i).

Ο αλγόριθμος που προτείνεται στο (Ferdowsi & Saad, Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in the Internet of Things, 2018) ονομάζεται επίσης δυναμικό υδατογράφημα (Satchidanandan & Kumar, 2017) ο οποίος είναι ικανός να ανιχνεύσει την κυβερνοφυσική επίθεση και να αποτρέψει τις επιθέσεις λαθρακρόασης. Ωστόσο, ο αλγόριθμος απαιτεί μεγάλους υπολογιστικούς πόρους στο cloud για τον έλεγχο ταυτότητας σήματος συσκευής IoT. Κατά συνέπεια, το cloud μπορεί να πιστοποιήσει μόνο έναν περιορισμένο αριθμό ευάλωτων συσκευών IoT. Το cloud μπορεί να επιλέξει τις ευάλωτες συσκευές IoT παρατηρώντας την κατάσταση ασφαλείας

τους. Ωστόσο, αυτό μπορεί να μην είναι πρακτικό, καθώς οι συσκευές IoT ενδέχεται να μην αναφέρουν την κατάσταση ασφαλείας τους. Έτσι, στο (Ferdowsi & Saad, Deep Learning for Signal Authentication and Security in Massive Internet-of-Things Systems, 2019)προτείνεται τη χρήση της DQL που επιτρέπει στο cloud να αποφασίσει ποιες συσκευές IoT θα πιστοποιηθούν. Δεδομένου ότι οι συσκευές IoT με πιο πολύτιμα δεδομένα είναι πιθανό να προσβληθούν, η ανταμοιβή ορίζεται ως συνάρτηση των τιμών δεδομένων των συσκευών IoT. Η κατάσταση του cloud περιλαμβάνει ενέργειες επίθεσης του εισβολέα στις συσκευές IoT τα τελευταία χρονικά διαστήματα. Οι ενέργειες του εισβολέα στις συσκευές IoT μπορούν να αποκτηθούν χρησιμοποιώντας τον αλγόριθμο δυναμικής υδατογράφησης στο (Ferdowsi & Saad, Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in the Internet of Things, 2018). Στη συνέχεια, η DQL χρησιμοποιεί μια μονάδα LSTM για να βρει τη βέλτιστη πολιτική. Η είσοδος της μονάδας LSTM είναι η κατάσταση του cloud και η έξοδος περιλαμβάνει πιθανότητες επίθεσης στις συσκευές IoT. Χρησιμοποιώντας ένα πραγματικό σύνολο δεδομένων από τα επιταχυνσιόμετρα, τα αποτελέσματα της προσομοίωσης δείχνουν ότι η προτεινόμενη DQL μπορεί να βελτιώσει τη χρησιμότητα του cloud έως και 30% σε σύγκριση με την περίπτωση στην οποία το cloud επιλέγει τις συσκευές IoT με την ίδια πιθανότητα.

Στον Πίνακα 7 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για ασφάλεια δικτύου.

3.6 Διατήρηση Συνδεσιμότητας

Συστήματα πολλαπλών ρομπότ, όπως συνεργατικά δίκτυα πολλαπλών UAV, έχουν εφαρμοστεί ευρέως σε πολλούς τομείς όπως στις ένοπλες δυνάμεις, π.χ. για την ανίχνευση εχθρών. Στο συνεργατικό σύστημα πολλαπλών ρομπότ, απαιτείται η σύνδεση μεταξύ των ρομπότ, π.χ. UAV ώστε να είναι δυνατή η επικοινωνία και η ανταλλαγή πληροφοριών. Για την αντιμετώπιση του προβλήματος διατήρησης συνδεσιμότητας, χρησιμοποιείται ο αλγόριθμος Τεχνητού Δυνητικού Πεδίου (Artificial Potential Field - APF) (Vadakkepat, Tan, & Ming-Liang, 2000) Ωστόσο, ο αλγόριθμος δεν μπορεί να υιοθετηθεί άμεσα όταν τα ρομπότ πραγματοποιούν αποστολές σε δυναμικά και σύνθετα περιβάλλοντα. Η DQL που επιτρέπει σε κάθε ρομπότ να λαμβάνει δυναμικές αποφάσεις με βάση τη δική του κατάσταση μπορεί να εφαρμοστεί αποτελεσματικά για τη διατήρηση της

συνδεσιμότητας στο σύστημα πολλαπλών ρομπότ. Μια τέτοια προσέγγιση προτείνεται στο (Huang, Wang, & Yi, Deep Q-Learning to Preserve Connectivity in Multi-robot Systems, 2017).

Το μοντέλο στο (Huang, Wang, & Yi, Deep Q-Learning to Preserve Connectivity in Multi-robot Systems, 2017) αποτελείται από δύο ρομπότ ή UAV, δηλαδή, ένα ρομπότ επικεφαλής και ένα ρομπότ ακόλουθος. Στο μοντέλο, ένας κεντρικός έλεγχος, δηλ., ένα επίγειο BS, ρυθμίζει την ταχύτητα του ακόλουθου έτσι ώστε ο ακόλουθος να παραμένει στο εύρος επικοινωνίας του επικεφαλή ανά πάσα στιγμή. Το πρόβλημα διατήρησης της συνδεσιμότητας μπορεί έτσι να διατυπωθεί ως MDP. Ο πράκτορας είναι το BS, και οι καταστάσεις είναι η σχετική θέση και η ταχύτητα του επικεφαλή σε σχέση με τον ακόλουθο. Ο χώρος δράσης αποτελείται από πιθανές τιμές ταχύτητας του ακόλουθου. Η πραγματοποίηση μιας ενέργειας επιστρέφει μια ανταμοιβή που είναι +1 εάν ο ακόλουθος βρίσκεται στο εύρος του ηγέτη και −1 διαφορετικά. Χρησιμοποιείται ένα DQN που χρησιμοποιεί FNN το οποίο επιτρέπει στο BS να βρει μια βέλτιστη πολιτική για τη μεγιστοποίηση της αναμενόμενης εκπτωτικής αθροιστικής ανταμοιβής. Η είσοδος του DQN περιλαμβάνει τις καταστάσεις των δύο ρομπότ και η έξοδος είναι ο χώρος δράσης του ακόλουθου. Τα αποτελέσματα προσομοίωσης δείχνουν ότι για διαφορετικές τοποθεσίες του επικεφαλή και του ακολούθου, η βαθμολογία που λαμβάνεται από το προτεινόμενο σχήμα είναι πάντα 100, ενώ η βαθμολογία της μεθόδου APF μπορεί περιστασιακά να είναι μικρότερη από 100. Αυτό σημαίνει ότι το προτεινόμενο σχήμα επιτυγχάνει καλύτερη σύνδεση μεταξύ τα δύο ρομπότ από αυτό της μεθόδου APF. Ωστόσο, πρέπει να διερευνηθεί ένα γενικό σενάριο με περισσότερους από έναν επικεφαλή και έναν ακόλουθο.

Λαμβάνοντας υπόψη το γενικό σενάριο, στο (Huang, Wang, & Yi, A deep reinforcement learning approach to preserve connectivity for multi-robot systems, 2017) αντιμετωπίζεται η διατήρηση της συνδεσιμότητας μεταξύ πολλαπλών επικεφαλών και πολλαπλών ακολούθων. Το ρομποτικό σύστημα είναι σίγουρα συνδεδεμένο εάν δύο ρομπότ είναι συνδεδεμένα μέσω απευθείας συνδέσμου ή συνδέσμου πολλαπλών μεταπηδήσεων. Για να εκφραστεί η συνδεσιμότητα σε ένα τέτοιο σύστημα, εισάγεται η έννοια της αλγεβρικής συνδεσιμότητας (algebraic connectivity) (Poonawala, Satici, Eckert, & Spong, 2015), η οποία είναι η δεύτερη μικρότερη ιδιοτιμή ενός πίνακα Laplace.

Το ρομποτικό σύστημα συνδέεται εάν η αλγεβρική συνδεσιμότητα του συστήματος είναι θετική. Έτσι, το πρόβλημα είναι να προσαρμόσουμε την ταχύτητα των ακολούθων έτσι ώστε η αλγεβρική συνδεσιμότητα να είναι θετική με την πάροδο του χρόνου. Αυτό το πρόβλημα μπορεί να διατυπωθεί ως MDP στο οποίο ο παράγοντας είναι το επίγειο BS, η κατάσταση είναι ένας συνδυασμός των καταστάσεων όλων των ρομπότ, η ενέργεια είναι ένα σύνολο πιθανών τιμών ταχύτητας για τους ακολούθους. Η ανταμοιβή είναι +1 εάν η αλγεβρική συνδεσιμότητα του συστήματος αυξάνεται ή διατηρείται, ενώ γίνεται ποινή −1 εάν η αλγεβρική συνδεσιμότητα μειωθεί. Παρόμοια με το (Huang, Wang, & Yi, Deep Q-Learning to Preserve Connectivity in Multi-robot Systems, 2017), υιοθετείται ένα DQN. Λόγω του μεγάλου χώρου δράσης των ακολούθων, χρησιμοποιείται το νευρωνικό δίκτυο Actor-Critic (Mnih, et al., 2016). Τα αποτελέσματα της προσομοίωσης δείχνουν ότι οι ακόλουθοι ακολουθούν πάντα την κίνηση των ηγετών, ακόμη και αν η πορεία των ηγετών αλλάζει δυναμικά. Αυτό καταδεικνύει την ικανότητα του DQN να αντιμετωπίζει το πρόβλημα διατήρησης συνδεσιμότητας για συστήματα πολλαπλών ρομπότ. Ωστόσο, το προτεινόμενο DQN απαιτεί περισσότερο χρόνο για σύγκλιση από αυτόν στο (Huang, Wang, & Yi, Deep Q-Learning to Preserve Connectivity in Multi-robot Systems, 2017) λόγω της παρουσίας περισσότερων ακολούθων.

Τα προτεινόμενα σχήματα στα (Huang, Wang, & Yi, Deep Q-Learning to Preserve Connectivity in Multi-robot Systems, 2017) και (Huang, Wang, & Yi, A deep reinforcement learning approach to preserve connectivity for multi-robot systems, 2017) δεν λαμβάνουν υπόψη την ελάχιστη απόσταση μεταξύ των επικεφαλής και των ακολούθων. Οι επικεφαλής και οι ακόλουθοι μπορούν να συγκρουστούν μεταξύ τους εάν η απόσταση μεταξύ τους είναι πολύ μικρή. Έτσι, το BS πρέπει να εγγυηθεί την ελάχιστη απόσταση μεταξύ τους. Μία λύση είναι να υπάρχει η ελάχιστη απόσταση στην ανταμοιβή όπως προτείνεται στο (Wang C. , Wang, Zhang, & Zhang, 2017). Ειδικότερα, εάν ο επικεφαλής είναι πολύ κοντά στον ακόλουθο, η ανταμοιβή του συστήματος τιμωρείται λόγω της ελάχιστης απόστασης. Ο αλγόριθμος DQL που προτείνεται στο (Huang, Wang, & Yi, A deep reinforcement learning approach to preserve connectivity for multi-robot systems, 2017) στη συνέχεια χρησιμοποιείται έτσι ώστε το BS να βρίσκει σωστές ενέργειες, π.χ., στροφή αριστερά και δεξιά, για μεγιστοποίηση της αθροιστικής ανταμοιβής.

Όταν τα BS είναι πυκνά αναπτυγμένα, οι UAV ή οι χρήστες κινητών συσκευών πρέπει να εκτελούν συχνές μεταγωγές κλήσεων για να διατηρήσουν τη συνδεσιμότητα. Οι συχνές μεταγωγές κλήσεων αυξάνουν τα γενικά έξοδα επικοινωνίας και την κατανάλωση ενέργειας των χρηστών κινητής τηλεφωνίας και διακόπτουν τη ροή δεδομένων. Επομένως, είναι απαραίτητο να διατηρηθεί ένα κατάλληλος ρυθμός μεταγωγών κλήσεων. Στο (Wang, Li, Xu, Tian, & Cui, 2018) αντιμετωπίζεται το πρόβλημα της απόφασης μεταγωγής σε ένα εξαιρετικά πυκνό δίκτυο. Το μοντέλο δικτύου αποτελείται από πολλούς χρήστες κινητών συσκευών, μικρούς σταθμός βάσεως (Small Base Stations – SBS) και έναν κεντρικό ελεγκτή. Σε κάθε χρονικό διάστημα, ο χρήστης πρέπει να επιλέξει το SBS που θα συνδεθεί. Η διαδικασία λήψης αποφάσεων μεταγωγών μπορεί να μοντελοποιηθεί ως MDP, και η DQL υιοθετείται για να βρεθεί μια βέλτιστη πολιτική μεταγωγών για κάθε χρήστη, έτσι ώστε να ελαχιστοποιηθεί ο αριθμός των μεταγωγών διασφαλίζοντας παράλληλα συγκεκριμένη απόδοση. Η κατάσταση του χρήστη, δηλαδή, του πράκτορα, περιλαμβάνει την ποιότητα σήματος αναφοράς που λαμβάνεται από υποψήφια SBS και την τελευταία ενέργεια του χρήστη. Η ανταμοιβή ορίζεται ως η διαφορά μεταξύ του ρυθμού δεδομένων του χρήστη και της κατανάλωσης ενέργειας για τη διαδικασία μεταγωγής. Δεδομένης της υψηλής πυκνότητας των χρηστών, η DQL που χρησιμοποιεί Asynchronous Advantage Actor-Critic (A3C) και LSTM υιοθετείται για να βρει τη βέλτιστη πολιτική σε σύντομο χρόνο εκμάθησης. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι η προτεινόμενη DQL μπορεί να επιτύχει υψηλότερη απόδοση και χαμηλότερο ρυθμό παράδοσης από εκείνες του αλγόριθμου upper confidence bandit (Shen, Tekin, & van der Schaar, 2016) με παρόμοιο χρόνο εκμάθησης. Συγκεκριμένα, η απόδοση και ο ρυθμός παράδοσης της DQL είναι 0,7 bit/s/Hz και 0,0003, αντίστοιχα, ενώ εκείνες του αλγορίθμου άνω ζώνης συχνοτήτων είναι 0,67 bit/s/Hz και 0,00049, αντίστοιχα.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**	**Δίκτυα**
(Huang, Wang, & Yi, 2017)	MDP	DQN με χρήση FNN	Σταθμός βάσης εδάφους	Σχετικές θέσεις και η ταχύτητα των ρομπότ	Απόφαση ταχύτητας	Βαθμός +1 ή -1	Σύστημα ρομπότ
(Huang, Wang, & Yi, 2017)	MDP	DQN με A3C	Σταθμός βάσης εδάφους	Σχετικές θέσεις και η ταχύτητα των ρομπότ	Απόφαση ταχύτητας	Βαθμός +1 ή -1	Σύστημα ρομπότ
(Wang C. , Wang, Zhang, & Zhang, 2017)	MDP	DQN με A3C	Σταθμός βάσης εδάφους	Πληροφορίες αποστάσεων μεταξύ των ρομπότ	Απόφαση για στροφή αριστερά ή δεξιά	Βαθμός +1 ή -1	Σύστημα ρομπότ

(Wang, Li, Xu, Tian, & Cui, 2018)	MDP	DQN με χρήση A3C και LSTM	Κινητοί χρήστες	Η ποιότητα του λαμβανόμενου σήματος αναφοράς και η τελευταία ενέργεια	Επιλογή να εξυπηρετήσει μικρούς σταθμούς βάσης	Ρυθμός δεδομένων και κατανάλωση ενέργειας	Εξαιρετικά πυκνό δίκτυο
(Faris & Brian, 2019)	MDP	DQN με χρήση CNN	Σταθμός βάσης μέσης εμβέλειας	Ο αριθμός των ενεργών συναγερμών	Ενεργοποίηση της διαφορικότητας εκπομπής και αλλαγή του αζιμουθίου της κεραίας	Βαθμός -1, 0, +1 και +5	Αυτοοργανωμένο δίκτυο

Πίνακας 8. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DQL για συντήρηση συνδεσιμότητας

Για να ενισχυθεί η αξιοπιστία της επικοινωνίας μεταξύ των SBS και των χρηστών κινητής τηλεφωνίας, τα SBS πρέπει να είναι σε θέση να χειρίζονται αυτόματα σφάλματα και αποτυχίες δικτύου ως αυτοθεραπεία. Το DQL μπορεί να εφαρμοστεί όπως προτείνεται στο (Mismar & Evans, 2018) για να κάνει τις βέλτιστες ρυθμίσεις παραμέτρων με βάση την παρατήρηση της απόδοσης του δικτύου. Το μοντέλο είναι το δίκτυο 5G συμπεριλαμβανομένου ενός μεσαίας εμβέλειας σταθμού βάσης (Medium Range Base Station – MBS). Το MBS ως πράκτορας πρέπει να χειριστεί σφάλματα δικτύου, όπως σφάλματα διαφορισμού μετάδοσης και αλλαγή αζιμουθίου κεραίας, π.χ. λόγω ανέμου. Αυτά τα σφάλματα παρουσιάζονται ως η κατάσταση του MBS που είναι ο αριθμός των ενεργών συναγερμών. Με βάση τους συναγερμούς, το MBS μπορεί να προβεί σε ενέργειες όπως (i) ενεργοποίηση του διαφορισμού μετάδοσης και (ii) ρύθμιση του αζιμούθιου της κεραίας στην προεπιλεγμένη τιμή. Η ανταμοιβή που λαμβάνει το MBS είναι οι βαθμολογίες, π.χ. −1, 0 και +1, ανάλογα με τον αριθμό των σφαλμάτων που συμβαίνουν. Το DQL χρησιμοποιείται για να βρεθεί η βέλτιστη πολιτική. Τα αποτελέσματα προσομοίωσης δείχνουν ότι η προτεινόμενη DQL μπορεί να επιτύχει απόδοση δικτύου κοντά σε αυτήν της self-healing που βασίζεται στην oracle, δηλ. the upper performance bound, αλλά εμφανίζει λιγότερα μηνύματα σφάλματος passing overhead. Συγκεκριμένα, η πολυπλοκότητα μετάδοσης μηνυμάτων της προτεινόμενης DQL είναι O (N) και αυτή της self-healing που βασίζεται στην oracle είναι O (N^2), όπου N είναι ο αριθμός των SBS στο δίκτυο.

Στον Πίνακα 8 συνοψίζονται οι προσεγγίσεις που χρησιμοποιούν DQL για συντήρηση συνδεσιμότητας.

Κεφάλαιο 4

Βαθιά Ενίσχυση της Μάθησης για Ασφαλείς UAV Επικοινωνίες

Σε αυτήν την ενότητα, παρουσιάζεται αρχικά ο ρόλος που παίζουν τα UAV στις σύγχρονες επικοινωνίες και στη δικτύωση και τα πλεονεκτήματα της χρήσης τους σε σχέση με τα παραδοσιακά συστήματα επίγειας επικοινωνίας. Στη συνέχεια παρατίθεται η συμβολή της τεχνητής νοημοσύνης στις εν λόγω επικοινωνίες, καθώς και οι λύσεις που προσφέρει η μηχανική μάθηση στις UAV επικοινωνίες. Τέλος αναλύονται οι τεχνικές βαθιάς ενίσχυσης της μάθησης για ασφαλείς UAV επικοινωνίες, ως η πιο προηγμένη και αποτελεσματική μέθοδος μηχανικής μάθησης.

4.1 Αξιοποίηση των UAV στις Επικοινωνίες και στη Δικτύωση

Η επικοινωνία με μη επανδρωμένα αεροσκάφη (UAV) παίζει σημαντικό ρόλο στα δίκτυα επόμενης γενιάς με δυνατότητα τεχνητής νοημοσύνης (AI) (Tang, Kawamoto, Kato, & Liu, 2020). Η αξιοποίηση των UAV είναι επίσης μια πολλά υποσχόμενη τεχνική για την παροχή αποτελεσματικής και αξιόπιστης ασύρματης επικοινωνίας σε ορισμένα απαιτητικά σενάρια, όπως σε απομακρυσμένες περιοχές και σε καταστάσεις έκτακτης ανάγκης, λόγω της ευελιξίας και του εύκολου ελέγχου (Koch, Mancuso, West, & Bestavros, 2019). Ειδικά, όταν οι σταθεροί επίγειοι σταθμοί βάσης (BS) καταστρέφονται από φυσικές καταστροφές, τα UAV μπορούν να χρησιμεύσουν ως εναέρια BS για την παροχή επικοινωνίας έκτακτης ανάγκης στους επίγειους χρήστες (Zhao, et al., 2019). Σε ορισμένες περιοχές hotspot, όπως ένα αθλητικό ή μουσικό γεγονός μεγάλης κλίμακας, τα ιπτάμενα UAV μπορούν να παρέχουν έκτακτες και κατ' απαίτηση υπηρεσίες για να αντιμετωπιστεί αποτελεσματικά η τοπικά έντονη κίνηση στις επιβαρυμένες κυψέλες (Cheng, Zhang, Yunfei, Zhao, Yu, & Leung, 2018). Μια άλλη εφαρμογή των UAV είναι να παρέχουν υπηρεσίες ιπτάμενων αναμεταδοτών για τη δημιουργία της σύνδεσης μεταξύ

δύο απομακρυσμένων επικοινωνιακών κόμβων (Cheng, Gui, Zhao, Chen, Tang, & Sari, 2019). Τα δεδομένα μπορούν να μεταδοθούν σε μεγάλες αποστάσεις στα επίγεια BS, μέσω πολλαπλών UAV αναμεταδόσεων. Ο αριθμός των αναπηδήσεων και η καλύτερη θέση των αναμεταδιδόμενων UAV συζητήθηκαν μέσω μαθηματικών εξισώσεων (Wang H. , Wang, Ding, Chen, Li, & Han, 2018), (Chen, Zhao, Ding, & Alouini, 2018). Εκτός από το να εκμεταλλεύονται ως ιπτάμενα BS ή αναμεταδότες για κυψελωτή επικοινωνία, τα UAV μπορεί επίσης να χρησιμοποιηθούν για τη συλλογή δεδομένων για εφαρμογές Internet of Things (IoT) (Ding, Wu, Zhang, Lin, Tsiftsis, & Yao, 2018), (Yang, Zheng, Bian, Song, & Han, 2018).

Σε σύγκριση με τις παραδοσιακές επίγειες επικοινωνίες, στις οποίες παρουσιάζονται τα φαινόμενα υποβάθμισης του H/M σήματος, λόγω σκίασης και εξασθένισης πολλαπλών διαδρομών, ένα δίκτυο με τη χρήση UAV είναι σαφώς πιο αποδοτικό, λόγω της καλύτερης διάδοσης του H/M κύματος, διαθέτοντας οπτική επαφή (LoS) με τους επίγειους κόμβους (Zhong, Yao, & Xu, 2019). Ωστόσο, οι ανωτέρω διασυνδέσεις αέρος-εδάφους (Air to Ground-A2G) των UAV με τους επίγειους σταθμούς, επιτρέπουν τις υποκλοπές από τους επίγειους υποκλοπείς (Ground Eavesdroppers-GEs), γεγονός που αποτελεί απειλή για το δίκτυο. Κατά συνέπεια, η εγγύηση της ασφάλειας των επικοινωνιών UAV είναι επείγουσα και απαραίτητη. Τα τελευταία χρόνια, η ασφάλεια φυσικού επιπέδου (Physical Layer Security-PLS) έχει αναγνωριστεί ως μια πολλά υποσχόμενη τεχνική για την προστασία των επικοινωνιών με τη βοήθεια UAV από τους ύποπτους υποκλοπείς, αξιοποιώντας τα χαρακτηριστικά του ασύρματου καναλιού ως εναλλακτική ή συμπληρωματική λύση στην κρυπτογράφηση των επικοινωνιών. Η παρούσα εργασία, εστιάζεται στις ανωτέρω μεθόδους ασφάλειας φυσικού επιπέδου για την αντιμετώπιση φαινομένων παρεμβολών (jamming), υποκλοπών (eavesdropping) και παραποιήσεων πληροφοριών (spoofing) σε συστήματα UAV που ενσωματώνονται με τους μηχανισμούς ασφαλείας του ανώτερου στρώματος, αντί να τις αντικαθιστούν. Για παράδειγμα, οι παραδοσιακές μέθοδοι κρυπτογράφησης ανώτερου επιπέδου δεν μπορούν να αντιμετωπίσουν όλες τις προκλήσεις υποκλοπής στα συστήματα UAV, λόγω της δυναμικής τοπολογίας του δικτύου, του ενεργειακού περιορισμού των UAV, καθώς και της αυξανόμενης υπολογιστικής ικανότητας των υποκλοπών (Yilmaz & Arslan, 2015). Για το σκοπό αυτό, τα πρωτόκολλα κρυπτογράφησης φυσικού επιπέδου είναι αποτελεσματικά για την

προστασία των συστημάτων UAV από την υποκλοπή (Yilmaz & Arslan, 2015), (Shiu, Chang, Wu, Huang, & Chen, 2011).

4.2 Τεχνικές Μηχανικής Εκμάθησης για ασφαλείς UAV Επικοινωνίες

Η ενσωμάτωση τεχνικών τεχνητής νοημοσύνης (AI) και μηχανικής μάθησης (ML) σε ασύρματα δίκτυα μπορεί να αξιοποιήσει τη νοημοσύνη για την αντιμετώπιση διαφόρων ζητημάτων. Έτσι, ο συνδυασμός AI/ML και UAV φαίνεται να συσχετίζεται έντονα σε διαφορετικούς κλάδους και εφαρμογές και σε όλα τα επίπεδα του δικτύου, υποσχόμενος πρωτοφανή κέρδη απόδοσης και μείωση της πολυπλοκότητας.

Η τεχνητή νοημοσύνη έχει θεωρηθεί ως η επιστήμη της εκπαίδευσης μηχανών για την εκτέλεση ανθρώπινων καθηκόντων. Υπάρχουν πολλές εφαρμογές στις οποίες έχει εμπλακεί η AI, όπως ρομποτικά οχήματα, αναγνώριση ομιλίας, αυτόματη μετάφραση και πρόσφατα ασύρματες επικοινωνίες. Επιπλέον, ένα συγκεκριμένο υποσύνολο της τεχνητής νοημοσύνης είναι οι τεχνικές που χρησιμοποιούνται για την εκπαίδευση μηχανών για το πώς να μάθουν, η οποία προέρχεται από ένα νέο πλαίσιο γνωστό ως ML. Σε αυτό το πλαίσιο, το ML μπορεί να παρέχει λύσεις σε σενάρια όπου ένας τεράστιος αριθμός συσκευών απαιτεί ταυτόχρονα πρόσβαση στους πόρους του δικτύου με δυναμικό, ετερογενή και απρόβλεπτο τρόπο, π.χ. σε επικοινωνίες IoT. Υπό αυτή την έννοια, η έξυπνη διαχείριση θα πρέπει να πραγματοποιείται σε ολόκληρο το δίκτυο προκειμένου να ανταποκριθεί στις διάφορες έντονες απαιτήσεις αυτού του νέου τύπου υπηρεσιών. Ο σκοπός είναι η προσαρμοστική και σε πραγματικό χρόνο διαχείριση των πόρων του δικτύου με τον βέλτιστο τρόπο. Ως εκ τούτου, οι αλγόριθμοι ML έχουν προταθεί ως μια αποτελεσματική προσέγγιση για την αντιμετώπιση όλων αυτών των αντιφατικών προκλήσεων που προέρχονται από το σύστημα IoT. Γενικά, το ML βασίζεται στο πλαίσιο αναγνώρισης προτύπων και η κύρια ιδέα του είναι να εκμεταλλεύεται τη συσχέτιση μεταξύ ενός συνόλου δεδομένων και/ή προηγούμενων ακολουθιών καλής δράσης για προσαρμογή στις περιβαλλοντικές αλλαγές χωρίς κανενός είδους ανθρώπινη παρέμβαση. Σαφώς, το πλεονέκτημα που προσφέρει το πλαίσιο ML στη λειτουργία ασύρματων δικτύων είναι ότι επιτρέπει στα στοιχεία του δικτύου να παρακολουθούν, να μαθαίνουν και να προβλέπουν διάφορες παραμέτρους που σχετίζονται με την

επικοινωνία, όπως η συμπεριφορά του ασύρματου καναλιού, τα μοτίβα επισκεψιμότητας, το περιβάλλον χρήστη και οι τοποθεσίες συσκευών. Το ML ταξινομείται σε διάφορες κατηγορίες, όπως η εποπτευόμενη μάθηση (Supervised learning), η ημι-εποπτευόμενη μάθηση (Semi-supervised learning), η χωρίς επίβλεψη μάθηση (Unsupervised learning) και η ενισχυτική μάθηση (Reinforcement Learning-RL) (Alpaydin, 2014).

- Επίβλεψη μάθησης: Στην εποπτευόμενη μάθηση, οι αλγόριθμοι χρησιμοποιούν σύνολα δεδομένων, στα οποία είναι διαθέσιμα τόσο η είσοδος όσο και η επιθυμητή έξοδος. Επομένως, αυτού του είδους οι αλγόριθμοι μπορούν να χρησιμοποιηθούν μόνο σε σενάρια όπου υπάρχουν αρκετά διαθέσιμα επισημασμένα δεδομένα για εκμετάλλευση.

- Μάθηση χωρίς επίβλεψη: Οι αλγόριθμοι μάθησης χωρίς επίβλεψη απαιτούν επίσης τη διάθεση δεδομένων για εκπαίδευση, τα οποία, ωστόσο, δεν περιλαμβάνουν επισημασμένη έξοδο. Επομένως, σε αυτόν τον τύπο μάθησης, η ομαδοποίηση ή η ανακάλυψη προτύπων πραγματοποιείται στα διαθέσιμα δεδομένα.

- Ημι-εποπτευόμενη μάθηση: Έχει ακολουθηθεί μια ενδιάμεση προσέγγιση σχετικά με τη φύση των διαθέσιμων δεδομένων με τους αλγόριθμους ημι-εποπτευόμενης μάθησης. Σε αυτόν τον τύπο μάθησης, τόσο τα επισημασμένα όσο και τα μη επισημασμένα δεδομένα αξιοποιούνται για την εκπαίδευση.

- Ενίσχυση της μάθησης: Στο RL, τα προβλήματα επιλύονται χρησιμοποιώντας μια σειρά ενεργειών που χρησιμοποιούν τον κανόνα δοκιμής και σφάλματος. Επομένως, η κύρια ιδέα αυτού του τύπου μάθησης είναι ριζικά διαφορετική σε σύγκριση με τις προηγούμενες που αναφέρθηκαν, οι οποίες εκμεταλλεύονται ιστορικά δεδομένα. Αντ' αυτού, οι αλγόριθμοι RL εκπαιδεύονται από τις προηγούμενες αποφάσεις για την επίλυση του προβλήματος. Οι αλγόριθμοι RL χρησιμοποιούνται σε διάφορα σενάρια στον τομέα της βελτιστοποίησης ασυρμάτων δικτύων.

Επιπλέον, μια συγκεκριμένη κατηγορία ML είναι η βαθιά μάθηση (DL). Στο DL, έχουν χρησιμοποιηθεί πολλαπλά στρώματα για την κατασκευή ενός τεχνητού νευρωνικού δικτύου, το οποίο είναι σε θέση να λαμβάνει έξυπνες αποφάσεις χωρίς κανενός είδους

ανθρώπινη παρέμβαση. Οι αλγόριθμοι DL μπορούν να εφαρμοστούν όταν απαιτείται περιορισμένη χειροκίνητη παρέμβαση, με κόστος τις υψηλότερες υπολογιστικές απαιτήσεις.

Οι μέθοδοι AI, ML έχουν χρησιμοποιηθεί ευρέως σε διάφορα σενάρια ασύρματων επικοινωνιών για τη βελτίωση πολλών παραμέτρων του δικτύου. Η παράμετρος που απασχολεί τη συγκεκριμένη ενότητα είναι η ασφάλεια των UAV επικοινωνιών. Παρακάτω αναφέρονται οι τεχνικές ML για ασφαλείς UAV επικοινωνίες στο φυσικό επίπεδο ασφάλειας (PLS).

Στην αναφορά (Li, Xu, Xia, & Zhao, 2018), ένα UAV χρησιμοποιείται για να εκτελέσει μια έξυπνη επίθεση σε ζεύγος πομπού - δέκτη. Το UAV είναι ικανό να ακούει τη μετάδοση του πομπού, ενώ παράγει ένα σήμα παραποίησης πληροφορίας (spoofing) ή παρεμβολής (jamming) για να μειώσει την ποιότητα της λήψης στο δέκτη. Προκειμένου να αξιολογηθεί η επίδραση των πρακτικών υποθέσεων, θεωρείται ατελής η εκτίμηση καναλιού λόγω περιορισμένων πιλοτικών σημάτων. Σε αυτή τη βάση, διαμορφώνεται ένα ασφαλές παιχνίδι επικοινωνίας και τα σφάλματα εκτίμησης καναλιού καθώς και οι έξυπνες επιθέσεις UAV αντιμετωπίζονται από κοινού μέσω της Q-learning. Αξιοποιώντας τα δεδομένα του ιστορικού μετάδοσης, πραγματοποιείται προσαρμογή της ισχύος μετάδοσης για να ενισχυθεί το απόρρητο της μετάδοσης. Επιπλέον, αναπτύχθηκε η στρατηγική ισορροπίας του Nash (Nash Equilibrium-NE) του μη συνεργάσιμου παιχνιδιού βάσει των σφαλμάτων εκτίμησης καναλιού, μεγιστοποιώντας τη λειτουργία χρησιμότητας της πηγής και μετριάζοντας ταυτόχρονα την επίδραση των έξυπνων επιθέσεων UAV. Τα αποτελέσματα προσομοίωσης έδειξαν ότι η προτεινόμενη στρατηγική βασισμένη στην εκμάθηση Q ενισχύει την απόδοση του PLS, μειώνοντας το ποσοστό επίθεσης ανεξάρτητα από το σφάλμα εκτίμησης καναλιού. Παρ 'όλα αυτά, για να διασφαλιστεί η σύγκλιση του αλγορίθμου εκμάθησης, απαιτείται γνώση του χώρου δράσης κάθε παίκτη.

Το θέμα των έξυπνων επιθέσεων σε ad hoc δίκτυα ενισχυμένων με UAV (VANETs) μελετάται στην αναφορά (Xiao, et al., 2018). Εδώ, ένα UAV λειτουργεί ως αναμεταδότης για την προώθηση των μηνυμάτων μιας μονάδας σε όχημα (On-Board Unit-OBU) σε μια μονάδα παρακείμενη στο δρόμο (Road Side Unit-RSU) όταν το τελευταίο αντιμετωπίζει

σοβαρές παρεμβολές ή παρενοχλήσεις. Η αλληλεπίδραση UAV-έξυπνου παρεμβολέα οδηγεί σε ένα παιχνίδι αντιπαρεμβολικής UAV αναμετάδοσης, όπου το UAV καθορίζει την απόφαση αναμετάδοσης του προς μια διαφορετική RSU και, ταυτόχρονα, ο παρεμβολέας παρατηρεί αυτή τη στρατηγική και επιλέγει ένα κατάλληλο επίπεδο ισχύος παρεμβολής. Η ισορροπία του Nash αυτού του παιχνιδιού προέρχεται επιδεικνύοντας την εξάρτηση της βέλτιστης στρατηγικής αναμετάδοσης από το κόστος μετάδοσης και το μοντέλο καναλιού αέρος-εδάφους. Για την επιλογή της βέλτιστης στρατηγικής αναμετάδοσης ενάντια στην παρεμβολή, σχεδιάστηκε μια προσέγγιση εκμάθησης με βάση την Policy Hill Climbing (PHC) χωρίς να απαιτείται γνώση του μοντέλου καναλιού UAV και της στρατηγικής παρεμβολής. Τα αποτελέσματα προσομοίωσης αποκαλύπτουν ότι η αναμετάδοση βασισμένη σε PHC μπορεί να μειώσει την απόδοση του ποσοστού σφάλματος bit του VANET σε σύγκριση με ένα σχήμα που βασίζεται στην Q-learning.

Περαιτέρω έρευνα σχετικά με τις πτυχές ασφάλειας των συνυπάρχων VANET και UAV έχει διεξαχθεί στην αναφορά (Xiao, Xie, Min, & Zhuang, 2018). Αναλυτικότερα, η θεωρία προοπτικής (Prospect Theory-PT) χρησιμοποιήθηκε για τη διαμόρφωση ενός υποκειμενικού παιχνιδιού έξυπνης επίθεσης στη μετάδοση UAV, στο οποίο ένας έξυπνος εισβολέας επιλέγει τον τύπο της επίθεσής του ανάμεσα από παρεμβολή, παραποίηση και υποκλοπή χωρίς να γνωρίζει την ακρίβεια ανίχνευσης επίθεσης του συστήματος UAV. Ταυτόχρονα, η ισχύς εκπομπής του UAV σε διαφορετικά κανάλια είναι κατάλληλα ρυθμισμένη για να αντισταθεί σε αυτές τις έξυπνες επιθέσεις. Οι συντελεστές στάθμισης πιθανοτήτων και αξίας χρησιμοποιούνται από το PT προκειμένου να μοντελοποιήσουν τις υποκειμενικές διαδικασίες λήψης αποφάσεων ενσωματώνοντας το γεγονός ότι οι άνθρωποι τείνουν να αποφεύγουν κινδύνους λόγω ζημιών. Στη συνέχεια, προτείνεται η κατανομή ισχύος UAV που βασίζεται σε RL για τη διασφάλιση της μετάδοσης από έξυπνες επιθέσεις χωρίς να γνωρίζουμε το μοντέλο επίθεσης και το μοντέλο καναλιού. Η αξιολόγηση της απόδοσης έδειξε ότι η προτεινόμενη μέθοδος κατανομής ισχύος UAV μπορεί να μειώσει τον αποτέλεσμα των έξυπνων επιθέσεων και να αυξήσει την ικανότητα μυστικότητας της μετάδοσης UAV κατά 16% και τη χρησιμότητα κατά 22%, σε σύγκριση με το σχέδιο βασισμένο στην εκπαίδευση Q.

Η προστασία από επιθέσεις παραποίησης πληροφορίας (spoofing attacks) του συστήματος παγκόσμιου εντοπισμού (GPS) σε δίκτυα UAV περιγράφεται στην αναφορά

(Manesh, Kenney, Hu, Devabhaktuni, & Kaabouch, 2019). Καθώς τα πλαστά σήματα μπορούν να μπερδέψουν τα UAV και τους ελεγκτές εναέριας κυκλοφορίας, ο μετριασμός των επιπτώσεων αυτού του τύπου επίθεσης είναι ζωτικής σημασίας για τα συστήματα UAV. Έτσι, αναπτύχθηκε μια εποπτευόμενη προσέγγιση ML βασισμένη στο Artificial Neural Network (ANN) για τον εντοπισμό των σημάτων παραποίησης πληροφορίας (spoofing) GPS. Προκειμένου να ταξινομηθούν τα σήματα GPS και να εκπαιδευτεί το ANN, απαιτούνται χαρακτηριστικά όπως το ψευδο εύρος, η μετατόπιση Doppler και το SNR. Ένα σημαντικό χαρακτηριστικό αυτής της λύσης είναι η δυνατότητα εφαρμογής της στα τρέχοντα συστήματα UAV, καθώς δεν απαιτεί καμία τροποποίηση του εξοπλισμού GPS. Τα αποτελέσματα έδειξαν ότι η λύση που βασίζεται σε ANN παρουσιάζει υψηλή πιθανότητα ανίχνευσης πλαστών σημάτων και μειωμένη πιθανότητα ψευδούς συναγερμού, σε σύγκριση με τη Stochastic Gradient Descent (SGD), αναζητώντας το ελάχιστο χρησιμοποιώντας συμβατική επαναληπτική επίλυση Newton-Raphson, με βάση ένα υποσύνολο δειγμάτων και την εκτίμηση της προσαρμοστικής ροπής (Adam), που χρησιμοποιεί δυναμικό ρυθμό μάθησης ανά παράμετρο για τον υπολογισμό του ελάχιστου της συνάρτησης κόστους.

Στη συνέχεια, η δυνατότητα χρήσης μη επίβλεψης μάθησης για την ανίχνευση ενεργών υποκλοπών σε δίκτυα που βασίζονται σε αναμετάδοση που αξιοποιούν UAV διερευνήθηκε στην αναφορά (Hoang, Nguyen, & Duong, 2019), όπου η uplink ήταν υπεύθυνη για τη διαδικασία αυθεντικοποίησης. Η ομαδοποίηση one-class Support Vector Machine (SVM) και k-means αξιοποιήθηκε για τον εντοπισμό πιθανών επιθέσεων κατά την αυθεντικοποίηση. Για τη δημιουργία συνόλων δεδομένων κατάρτισης για τα μοντέλα ML και τη διευκόλυνση της διαδικασίας κατάρτισης, χρησιμοποιήθηκαν τα ασύρματα σήματα και η στατιστική γνώση του Channel State Information (CSI). Σύμφωνα με τα αποτελέσματα, το one-class SVM ήταν πιο σταθερό από το k-means. Ωστόσο, το k-means είναι προτιμότερο, υπό την προϋπόθεση ότι η εκπεμπόμενη ισχύς του υποκλοπέα έχει αρκετά υψηλή τιμή.

Στον Πίνακα 9 συνοψίζονται οι πέντε ανωτέρω τεχνικές, που χρησιμοποιούν ML για ασφαλείς UAV επικοινωνίες (Bithas, Michailidis, Nomikos, Vouyioukas, & Kanatas, 2019).

Αναφορά	**Στόχος ασφάλειας**	**Λύση ML**

(Li, Xu, Xia, & Zhao, 2018)	Αντιμετώπιση υποκλοπών (eavesdropping)	Q-learning
(Xiao, et al., 2018)	Αντιμετώπιση παρεμβολών (jamming)	PHC-based learning
(Xiao, Xie, Min, & Zhuang, 2018)	Αντιμετώπιση παρεμβολών, υποκλοπών επιθέσεων παραποίησης πληροφορίας (spoofing)	RL
(Manesh, Kenney, Hu, Devabhaktuni, & Kaabouch, 2019)	Προστασία από επιθέσεις παραποίησης πληροφορίας (spoofing) σήματος GPS	ANN-supervised learning
(Hoang, Nguyen, & Duong, 2019)	Ανίχνευση υποκλοπών	One-class SVM and k-means

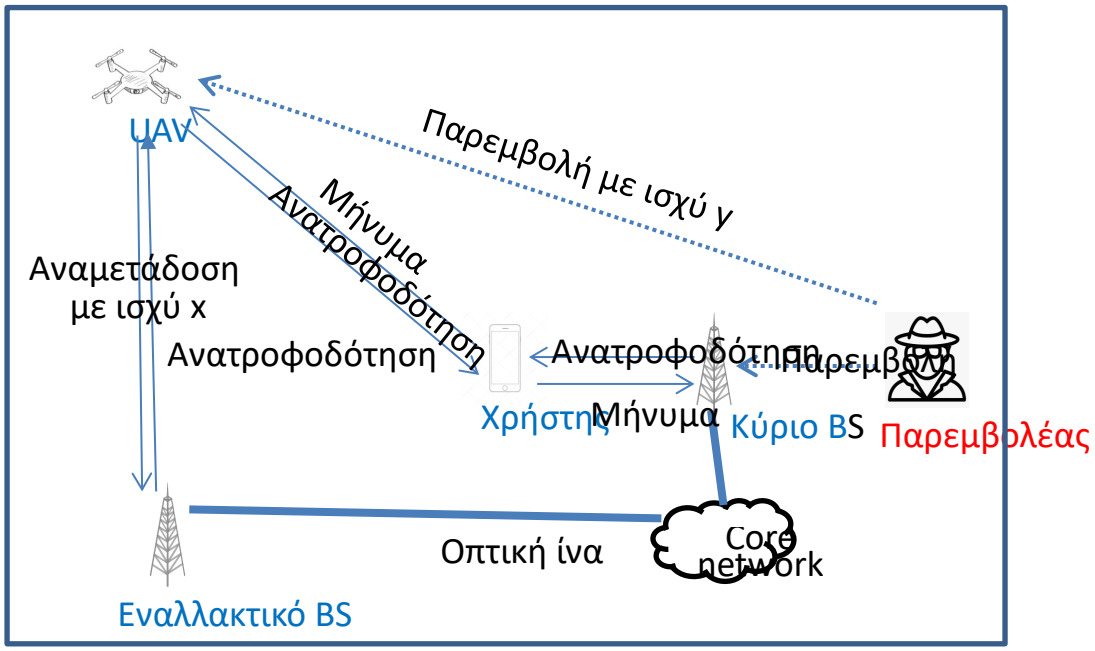
Πίνακας 9. Σύνοψη των προσεγγίσεων που χρησιμοποιούν ML για ασφαλείς UAV επικοινωνίες

4.3 Τεχνικές Βαθιάς Ενίσχυσης της Μάθησης για ασφαλείς UAV Επικοινωνίες.

Η βαθιά ενίσχυση της μάθησης, ως το πιο εξελιγμένο υποπεδίο της ML έχει προσελκύσει μεγάλο ενδιαφέρον για την επίλυση προβλημάτων βελτιστοποίησης υψηλής πολυπλοκότητας που δεν μπορούν να αντιμετωπίσουν οι συμβατικές μαθηματικές προσεγγίσεις. Η συγκεκριμένη ενότητα αναφέρεται στις τεχνικές DRL για ασφαλείς UAV επικοινωνίες.

Η πρώτη τεχνική (Lu, Xiao, Dai, & Dai, 2020) αναφέρεται σε ένα μοντέλο που αποτελείται από ένα UAV, σε ρόλο αναμεταδότη, έναν παρεμβολέα, έναν χρήστη κινητού και το BS που εξυπηρετεί. Ο χρήστης κινητής τηλεφωνίας μεταδίδει μηνύματα στον διακομιστή του μέσω της υπηρεσίας BS. Σε περίπτωση που το BS που εξυπηρετεί παρεμβάλλεται, το UAV, που βρίσκεται μακριά από την περιοχή των παρεμβολών, βοηθά τον κινητό χρήστη να μεταδώσει τα μηνύματα στον διακομιστή μέσω ενός εφεδρικού BS (Σχήμα 7). Συγκεκριμένα, ανάλογα με τις τιμές Signal to Interference plus Noise Ratio (SINR) και Bit Error Rate (BER) που αποστέλλονται από το BS που εξυπηρετεί, το UAV ως πράκτορας αποφασίζει το επίπεδο ισχύος του αναμεταδότη για να μεγιστοποιήσει τη χρησιμότητά του, δηλαδή τη διαφορά μεταξύ του SINR και του κόστους αναμετάδοσης. Το επίπεδο ισχύος αναμετάδοσης μπορεί να θεωρηθεί ότι είναι οι ενέργειες του UAV, ενώ οι SINR και BER είναι οι καταστάσεις του. Ως εκ τούτου, η επόμενη κατάσταση που παρατηρείται από το UAV είναι ανεξάρτητη από όλες τις προηγούμενες καταστάσεις και ενέργειες. Το

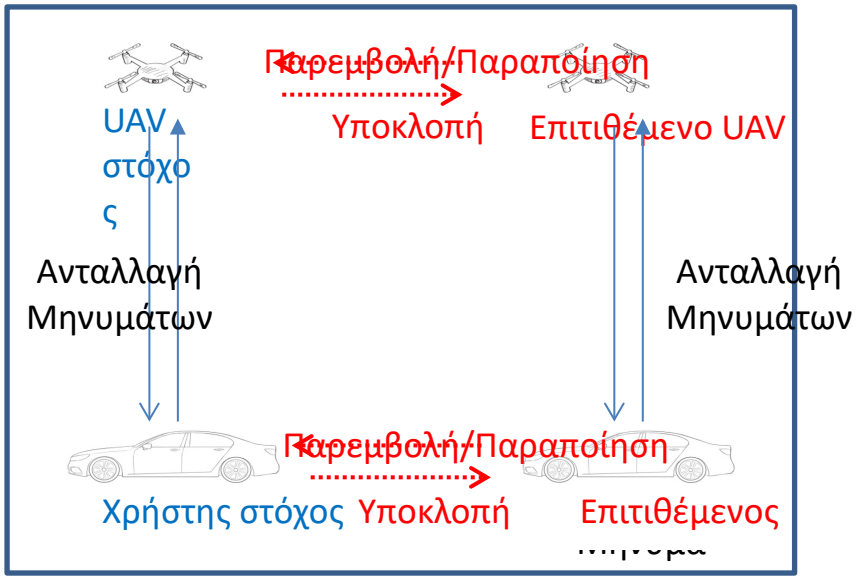
πρόβλημα διατυπώνεται ως MDP. Για να επιτευχθεί γρήγορα η βέλτιστη πολιτική αναμετάδοσης για το UAV, υιοθετείται το DQL που βασίζεται στο CNN. Τα αποτελέσματα της προσομοίωσης (Lu, Xiao, Dai, & Dai, 2020) δείχνουν ότι το προτεινόμενο σχήμα DQL παίρνει μόνο 200 χρονικά διαστήματα για να συγκλίνει στη βέλτιστη πολιτική, η οποία είναι 83,3% μικρότερη από αυτήν του σχήματος αναμετάδοσης που βασίζεται στην Q-learning (Xiao, et al., 2018). Επιπλέον, το προτεινόμενο σχήμα DQL μειώνει το BER του χρήστη κατά 46,6% σε σύγκριση με το σχήμα αναμεταδότη UAV που βασίζεται στην αναρρίχηση (Lv, Xiao, Hu, Wang, Hu, & Sun, 2017).



Σχήμα 7. Σύστημα κυψελωτής τηλεφωνίας με συμβολή UAV εναντίων επιθέσεων παρεμβολών

Το ανωτέρω μοντέλο που προτείνεται στο (Lu, Xiao, Dai, & Dai, 2020) προϋποθέτει ότι ο αναμεταδότης UAV απέχει αρκετά από την περιοχή παρεμβολής. Ωστόσο, ο παρεμβολέας μπορεί να χρησιμοποιήσει ένα UAV, που βρίσκεται κοντά στον αναμεταδότη UAV για να το παρεμβάλει. Σε ένα τέτοιο σενάριο, το DQL μπορεί να χρησιμοποιηθεί για την αντιμετώπιση της επίθεσης (Xiao, Xie, Min, & Zhuang, 2018). Στη δεύτερη αυτή τεχνική το μοντέλο βασίζεται στη ασφάλεια φυσικού επιπέδου και αποτελείται από ένα UAV και έναν εισβολέα (Σχήμα 8). Ο επιτιθέμενος θεωρείται ότι είναι «πιο έξυπνος» από ότι στο μοντέλο του (Lu, Xiao, Dai, & Dai, 2020). Αυτό σημαίνει ότι ο εισβολέας μπορεί να παρατηρήσει κανάλια που χρησιμοποιεί το UAV για να επικοινωνήσει με το BS στα προηγούμενα χρονικά διαστήματα και στη συνέχεια να επιλέξει τα επίπεδα ισχύος παρεμβολής στα κανάλια προορισμού. Επομένως, το UAV πρέπει να βρει μια πολιτική

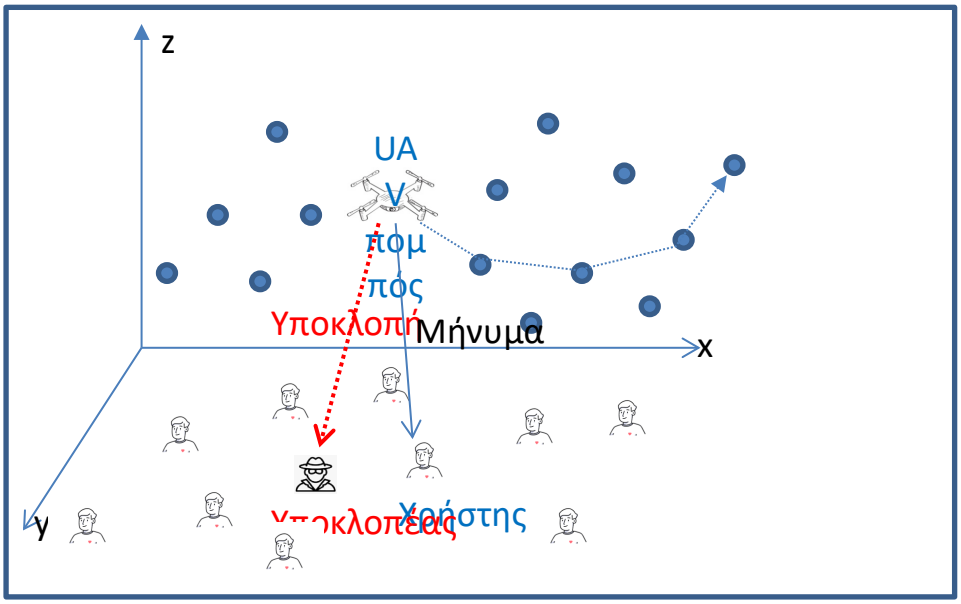
κατανομής ισχύος, δηλαδή να μεταδίδει επίπεδα ισχύος στα κανάλια, για να μεγιστοποιήσει τη χωρητικότητα της ασφαλής επικοινωνίας UAV-BS. Παρόμοια με το (Lu, Xiao, Dai, & Dai, 2020), χρησιμοποιείται το DQL που βασίζεται στο CNN, το οποίο επιτρέπει στο UAV να επιλέξει τις ενέργειές του, δηλαδή, να μεταδίδει επίπεδα ισχύος στα κανάλια, με βάση την κατάστασή του, δηλαδή το επίπεδο ισχύος παρεμβολής του εισβολέα στην τελευταία ώρα. Η ανταμοιβή είναι η διαφορά μεταξύ της χωρητικότητας ασφαλής επικοινωνίας των UAV και BS και του κόστους κατανάλωσης ενέργειας. Τα αποτελέσματα της προσομοίωσης στο (Xiao, Xie, Min, & Zhuang, 2018) δείχνουν ότι το προτεινόμενο DQL μπορεί να βελτιώσει τη χρησιμότητα του UAV έως και 13% σε σύγκριση με το βασικό σχήμα (Bowling & Veloso, 2002) που χρησιμοποιεί το Win or Learn Faster-Policy Hill Climbing (WoLF-PHC) για να αποτρέψει την επίθεση. Επίσης, ο ασφαλής ρυθμός του UAV, δηλαδή η πιθανότητα επίθεσης του UAV, που λαμβάνεται από την προτεινόμενη DQL είναι 7% υψηλότερη από εκείνη της γραμμής βάσης. Ωστόσο, η προτεινόμενη DQL έχει υψηλότερη υπολογιστική πολυπλοκότητα και χρειάζεται περισσότερο χρόνο για να ληφθεί απόφαση σε σύγκριση με το WoLH-PHC. Έτσι, η προτεινόμενη DQL εφαρμόζεται μόνο σε ένα σύστημα UAV.



Σχήμα 8. Σύστημα κυψελωτής τηλεφωνίας με συμβολή UAV εναντίων επιθέσεων παρεμβολών

Η τρίτη τεχνική (Jing, Jia, Lv, & Wan, 2021) αναφέρεται σε ένα σύστημα Mobile Edge Computing (MEC) που υποστηρίζεται από UAV και όπου υφίσταται υποκλοπέας (Σχήμα 9). Για να αντιμετωπιστεί το πρόβλημα της ασφαλούς επικοινωνίας μεταξύ του UAV και των χρηστών, χρησιμοποιείται ένας αλγόριθμος ασφαλείας βασισμένος στην DRL, που επιτρέπει στο UAV να βρει τη βέλτιστη στρατηγική πτήσης για να μεγιστοποιήσει το μέσο

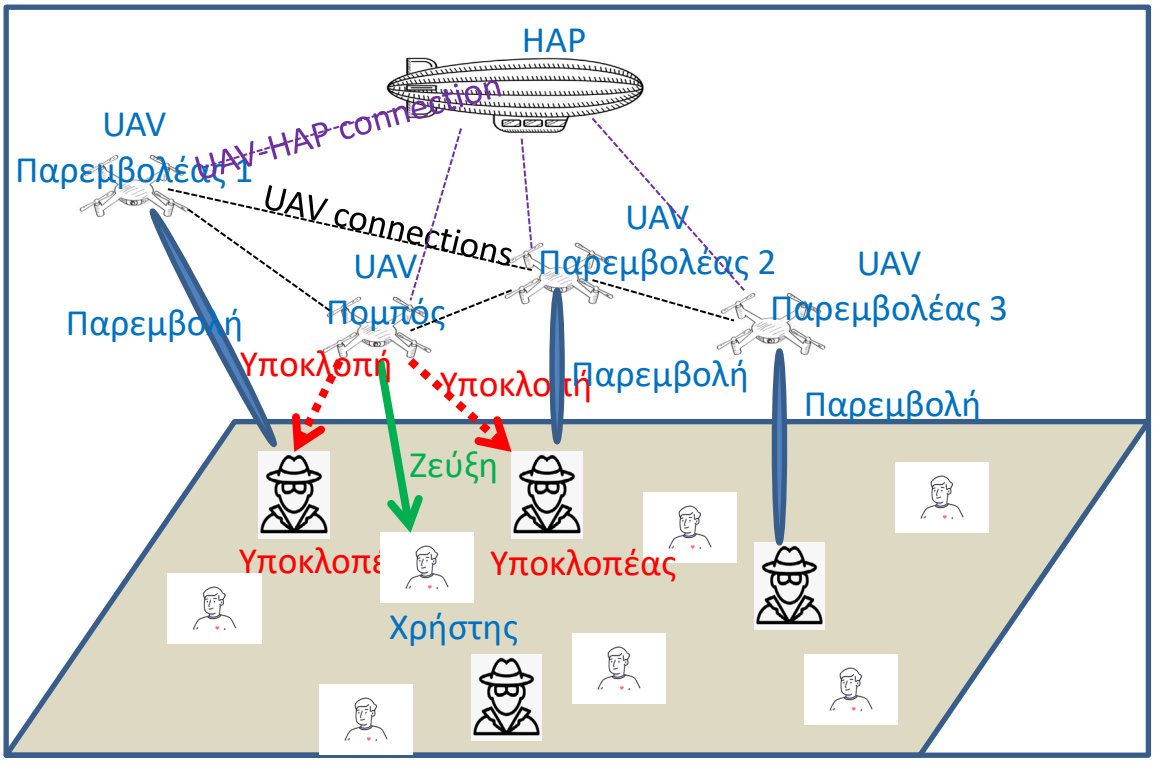
ποσοστό μυστικότητας των χρηστών που εξυπηρετεί. Η διαδικασία μεγιστοποίησης του μέσου ποσοστού μυστικότητας διαμορφώνεται ως διαδικασία απόφασης Markov (MDP) χωρίς πιθανότητα μετάβασης. Τα βασικά στοιχεία της MDP είναι οι καταστάσεις, που αναφέρονται στη θέση και το επίπεδο της μπαταρίας του UAV σε μια δεδομένη στιγμή, η ενέργεια που αναφέρεται στην ικανότητα του UAV να εξυπηρετεί τον χρήστη από συγκεκριμένη θέση, δεδομένη στιγμή και τέλος η ανταμοιβή που είναι το ποσοστό μυστικότητας κατά τη διάρκεια της εξυπηρέτησης. Το UAV χρησιμοποιεί τον προτεινόμενο αλγόριθμο για να αλλάξει τη θέση του μέσω online εκμάθησης RL και εκπαίδευσης εκτός σύνδεσης σε βαθύ νευρωνικό δίκτυο (DNN) για να βρει τη βέλτιστη στρατηγική πτήσης, ώστε να μεγιστοποιήσει το μέσο ποσοστό μυστικότητας. Τέλος, ο προτεινόμενος αλγόριθμος συγκρίνεται με τους παραδοσιακούς αλγορίθμους και τα αποτελέσματα προσομοίωσης δείχνουν ότι ο προτεινόμενος αλγόριθμος μπορεί να βελτιώσει αποτελεσματικά το μέσο ποσοστό μυστικότητας όταν το UAV εξυπηρετεί τους χρήστες και έχει ταχύτερο ρυθμό σύγκλισης από τον αλγόριθμο Q-Learning.



Σχήμα 9. Μοντέλο συστήματος MEC που υποστηρίζεται από UAV

Σε αυτή την τεχνική (Zhang, Zhuang, Gao, Wang, & Han, 2020) ερευνάται ένας μηχανισμός συνεργασίας πολλαπλών UAV για ασφαλείς επικοινωνίες, όπου ο πομπός UAV κινείται για να εξυπηρετήσει τους πολλαπλούς χρήστες εδάφους (Ground Users-GU), ενώ οι παρεμβολείς των UAV στέλνουν τα 3D σήματα παρεμβολής στους υποκλοπείς εδάφους (Ground Eavesdroppers-GEs) για την προστασία του πομπού UAV από την υποκλοπή (Σχήμα 10). Η τρισδιάστατη παρεμβολή εγγυάται ότι οι GU δεν θα παρεμβάλλονται από τα σήματα παρεμβολής. Είναι δύσκολο να γίνει ένας κοινός σχεδιασμός τροχιάς και

έλεγχος ισχύος για μια ομάδα UAV χωρίς κεντρικό έλεγχο. Για το σκοπό αυτό, προτείνεται μια προσέγγιση εκμάθησης βαθιάς ενίσχυσης πολλαπλών πρακτόρων για την επίτευξη του μέγιστου ποσοστού ασφαλούς ρυθμού, σχεδιάζοντας τη δυναμική τροχιά κάθε UAV. Η προτεινόμενη τεχνική βαθιάς ντετερμινιστικής πολιτικής πολλαπλών παραγόντων (Multi-Agent Deep Deterministic Policy Gradient-MADDPG) είναι η κεντρική εκπαίδευση σε πλατφόρμες μεγάλου υψομέτρου (High Altitude Platforms-HAP) και η κατανεμημένη εκτέλεση σε κάθε UAV, η οποία επιτρέπει την πλήρως κατανεμημένη συνεργασία μεταξύ των UAV. Στο ανωτέρω σύστημα που προσεγγίζεται ως παίγνιο Markov, θεωρούνται ως πράκτορες κάθε ένα UAV, είτε παίζει το ρόλο του πομπού, είτε του παρεμβολέα με το συνολικό αριθμό των UAV παρεμβολέων να είναι μικρότερος των GE. Ο χώρος των ενεργειών περιλαμβάνει τρία στοιχεία, το πρώτο αφορά στην κατεύθυνση πτήσης του UAV, το δεύτερο στο επίπεδο ισχύος εκπομπής και το τρίτο στο επίπεδο ισχύος της παρεμβολής των UAV. Ο χώρος της κατάστασης αποτελείται από τρία μέρη, τη θέση όλων των πρακτόρων, την ισχύ εκπομπής ή παρεμβολής ανάλογα με τον ρόλο των UAV και το ποσοστό μυστικότητας συγκεκριμένου χρήστη. Ως χώρος ανταμοιβής ορίζεται για τα UAV που έχουν ρόλο πομπού η διαφορά μεταξύ του ποσοστού μυστικότητας και της ποινής ισχύος εκπομπής, ενώ για τα UAV που έχουν ρόλο παρεμβολέα η διαφορά μεταξύ του ποσοστού μυστικότητας και της ποινής ισχύος παρεμβολής. Τέλος, τα αποτελέσματα της προσομοίωσης δείχνουν ότι η προτεινόμενη μέθοδος μπορεί να λύσει αποτελεσματικά το πρόβλημα σχεδιασμού τροχιάς συνεργασίας πολλαπλών UAV σε σενάρια ασφαλούς επικοινωνίας



Σχήμα 10. Ασφαλές επικοινωνιακό σύστημα με χρήση UAV

Στην πέμπτη τεχνική (Zhang, Mou, Gao, Jiang, Ding, & Han, 2020), σε ένα σενάριο παρόμοιο με αυτό της ανωτέρω τέταρτης τεχνικής (Σχήμα 10), προτείνεται μια συνεργατική προσέγγιση παρεμβολών, επιτρέποντας τους UAV παρεμβολείς να βοηθήσουν τον UAV πομπό να αμυνθεί έναντι των GE. Πιο συγκεκριμένα, ο πομπός UAV στέλνει τις εμπιστευτικές πληροφορίες σε GU και οι συσκευές παρεμβολών UAV στέλνουν τα σήματα τεχνητού θορύβου στα GE με τρισδιάστατη διαμόρφωση δέσμης. Προτείνεται μια προσέγγιση εκμάθησης βαθιάς ενίσχυσης πολλαπλών παραγόντων (MADRL), δηλαδή βαθιά ντετερμινιστική κλίση πολλαπλών παραγόντων (MADDPG) για τη μεγιστοποίηση της ικανότητας ασφάλειας, βελτιστοποιώντας από κοινού την τροχιά των UAV, την ισχύ μετάδοσης από τον πομπό UAV και την ισχύ παρεμβολής από τους UAV παρεμβολείς. Το παίγνιο Markov για το συγκεκριμένο σενάριο διαθέτει ως πράκτορες το κάθε ένα UAV, ο κάθε πράκτορας παρακολουθεί τη δική του κατάσταση και εκτελεί ενέργειες σύμφωνα με τη δική του πολιτική και στη συνέχεια λαμβάνει την ανταμοιβή από το περιβάλλον και μεταβαίνει στη καινούρια κατάσταση. Η κατάσταση του πράκτορα περιλαμβάνει τη θέση του πράκτορα και τον αύξων αριθμό του επίγειου χρήστη (GU), ο οποίος ανά συγκεκριμένα χρονικά slot θα αλλάζει για να μεταβεί διαδοχικά σε όλους. Η ενέργεια κάθε πράκτορα περιλαμβάνει το διάνυσμα της ταχύτητας στο ορθογώνιο σύστημα αξόνων και τη ισχύος του σήματος του πράκτορα, ανάλογα με το ρόλο πομπού ή παρεμβολέα. Η

ανταμοιβή περιλαμβάνει την ποινή των ορίων του χάρτη, το ποσοστό μυστικότητας, την ποινή ισχύος και την ανταμοιβή απόστασης. Ο αλγόριθμος MADDPG υιοθετεί κεντρική εκπαίδευση και κατανεμημένη εκτέλεση. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι η μέθοδος MADRL μπορεί να πραγματοποιήσει τον σχεδιασμό της κοινής τροχιάς των UAV και να επιτύχει καλή απόδοση. Για τη βελτίωση της αποτελεσματικότητας και της σύγκλισης της μάθησης, προτείνεται μια μέθοδο συνεχούς δράσης MADDPG (CAA-MADDPG), όπου ο πράκτορας μαθαίνει να δίνει προσοχή στις ενέργειες και τις παρατηρήσεις άλλων παραγόντων που είναι πιο συναφείς με αυτό. Από τα αποτελέσματα της προσομοίωσης, η απόδοση ανταμοιβής του CAA-MADDPG είναι καλύτερη από την MADDPG χωρίς προσοχή.

Αναφορά	**Μοντέλο**	**Αλγόριθμος εκμάθησης**	**Πράκτορας**	**Καταστάσεις**	**Ενέργειες**	**Ανταμοιβές**
(Lu, Xiao, Dai, & Dai, 2020)	MDP	DQN με χρήση CNN	Αναμεταδότης UAV	Σηματοθορυβικός λόγος σήματος και ρυθμός σφάλματος bit	Ισχύς αναμετάδοσης	Σηματοθορυβικός λόγος και κόστος αναμετάδοσης
(Xiao, Xie, Min, & Zhuang, 2018)	MDP	DQN με χρήση CNN	Πομπός UAV	Ισχύς παρεμβολής	Ισχύς εκπομπής	Ικανότητα απορρήτου και κόστος κατανάλωσης ενέργειας
(Jing, Jia, Lv, & Wan, 2021)	MDP	DQL	Πομπός UAV	Θέση και επίπεδο της μπαταρίας του UAV	Ικανότητα του UAV να εξυπηρετεί τον χρήστη από συγκεκριμένη θέση δεδομένη στιγμή	Ποσοστό μυστικότητας
(Zhang, Zhuang, Gao, Wang, & Han, 2020)	MDP	MADDPG	Πομπός UAV Παρεμβολέας UAV	Θέση όλων των πρακτόρων, Ισχύς εκπομπής ή παρεμβολής, Ποσοστό μυστικότητας	Κατεύθυνση πτήσης του UAV, Επίπεδο ισχύος εκπομπής-παρεμβολής	Ποσοστό μυστικότητας
(Zhang, Mou, Gao, Jiang, Ding, & Han, 2020)	MDP	CAA-MADDPG	Πομπός UAV Παρεμβολέας UAV	Θέση του πράκτορα και τον αύξων αριθμό του επίγειου χρήστη (GU)	Διάνυσμα της ταχύτητας στο ορθογώνιο σύστημα αξόνων Επίπεδο ισχύος εκπομπής-παρεμβολής	Ποινή των ορίων του χάρτη, Ποσοστό μυστικότητας, Ποινή ισχύος, Ανταμοιβή απόστασης

Πίνακας 10. Σύνοψη των προσεγγίσεων που χρησιμοποιούν DRL για ασφαλείς UAV επικοινωνίες

Στον Πίνακα 10 συνοψίζονται οι πέντε ανωτέρω τεχνικές, που χρησιμοποιούν DQL για ασφαλείς UAV επικοινωνίες. Συμπερασματικά φαίνεται η κάθε μια από τις εξεταζόμενες τεχνικές να ανταποκρίνεται στις προκλήσεις ασφαλών UAV επικοινωνιών σύμφωνα με το εκάστοτε συγκεκριμένο σενάριο στο οποίο καλούνται να λειτουργήσουν, με καλύτερα

αποτελέσματα σε σχέση με άλλους αλγορίθμους. Επίσης διαφαίνεται ότι οι τεχνικές με την πάροδο του χρόνου βελτιώνονται και τείνουν να συγκλίνουν σε ένα ρεαλιστικό επικοινωνιακό σύστημα, όπου δραστηριοποιούνται επίγειοι χρήστες, επίγειοι υποκλοπείς, αναμεταδότες UAV, παρεμβολείς UAV, HAP. Από τις πέντε παραπάνω τεχνικές, οι τρεις πρώτες διαθέτουν πιο απλό σενάριο με ελάχιστο αριθμό πρακτόρων UAV και GE ή παρεμβολέων. Οι δύο τελευταίες διαθέτουν παρόμοιο και πιο πολύπλοκο σενάριο με πολλαπλούς πράκτορες, επιτιθέμενους υποκλοπείς, καθώς και πιο έξυπνο τρόπο αντιμετώπισης των υποκλοπών, κάνοντας χρήση παρεμβολών από τα UAV στοχευμένα στα GE. Μια μελλοντική πρόκληση θα ήταν να περιλαμβάνονται στο σύστημα και επίγειοι ή εναέριοι (UAV) παρεμβολείς που να συνδράμουν στην προσπάθεια των υποκλοπέων GE.

Κεφάλαιο 5

Συμπεράσματα-Προκλήσεις

Αυτή η εργασία παρουσίασε μια έρευνα για τις εφαρμογές της βαθιάς ενίσχυσης της μάθησης στις επικοινωνίες και τη δικτύωση. Αρχικά μετά από μια σύντομη εισαγωγή στο κεφάλαιο 2, παρουσιάστηκαν βασικά στοιχεία της ενίσχυσης μάθησης, της βαθιάς μάθησης και της DQL, ενώ στη συνέχεια παρατέθηκαν διάφορες προηγμένες τεχνικές DQL και οι επεκτάσεις τους. Διαπιστώθηκε, επίσης ότι, μπορούν να χρησιμοποιηθούν διαφορετικές τεχνικές DQL για την επίλυση διαφορετικών προβλημάτων σε διαφορετικά σενάρια δικτύου.

Στο κεφάλαιο 3 εξετάστηκαν εφαρμογές DQL δυναμική πρόσβαση στο δίκτυο, προσαρμοστικό ρυθμό ελέγχου, ασύρματη προσωρινή αποθήκευση, εκφόρτωση δεδομένων, ασφάλεια δικτύου και διατήρηση της συνδεσιμότητας Οι αναθεωρημένες προσεγγίσεις συνοψίζονται στους Πίνακες 3 έως 8. Παρατηρείται ότι τα προβλήματα διαμορφώνονται ως MDP.

Για τη δυναμική πρόσβαση στο δίκτυο και τον προσαρμοστικό έλεγχο ρυθμού δεδομένων, οι προσεγγίσεις DQL για τα συστήματα IoT και DASH λαμβάνουν περισσότερη προσοχή από άλλα δίκτυα. Τα μελλοντικά δίκτυα, π.χ. δίκτυα 5G, περιλαμβάνουν πολλαπλές οντότητες δικτύου με πολλαπλούς αντικρουόμενους στόχους, π.χ. έσοδα από τον πάροχο σε σχέση με τη μεγιστοποίηση των χρηστών.

Για την ασύρματη προσωρινή αποθήκευση και εκφόρτωση δεδομένων παρατηρείται ότι το πλαίσιο DQL για προσωρινή αποθήκευση είναι συνήθως συγκεντρωτικό και ως επί το πλείστον υλοποιείται στον ελεγκτή δικτύου, π.χ. στο BS, στον πάροχο υπηρεσιών και στον κεντρικό προγραμματιστή, ο οποίος είναι πιο ισχυρός στη συλλογή πληροφοριών και στον σχεδιασμό πολιτικών πολλαπλών επιπέδων. Αντίθετα, οι τελικοί χρήστες έχουν περισσότερο έλεγχο στις λήψεις αποφάσεων εκφόρτωσης, και ως εκ τούτου

παρατηρείται πιο δημοφιλή εφαρμογή του παράγοντα DQL σε τοπικές συσκευές, π.χ. σε χρήστες κινητών συσκευών, συσκευές IoT και κόμβους fog. Αν και η ενοποίηση της δικτύωσης, της προσωρινής αποθήκευσης, των δεδομένων και της εκφόρτωσης υπολογισμών σε ένα ενοποιημένο πλαίσιο DQL είναι πολλά υποσχόμενη για μεγιστοποίηση της απόδοσης του δικτύου, αντιμετωπίζονται πολλές προκλήσεις στο σχεδιασμό εξαιρετικά σταθερών και ταχέων συγκλινόντων αλγορίθμων μάθησης, λόγω υπερβολικής καθυστέρησης και μη συγχρονισμένης συλλογής πληροφοριών από διαφορετικές οντότητες δικτύου .

Για την ασφάλεια του δικτύου και τη διατήρηση της συνδεσιμότητας παρατηρείται ότι το CNN χρησιμοποιείται κυρίως για το DQL για την ενίσχυση της ασφάλειας του δικτύου. Επιπλέον, οι προσεγγίσεις DQL για το ανώνυμο σύστημα όπως τα συστήματα ρομπότ και τα ITS λαμβάνουν περισσότερη προσοχή από άλλα δίκτυα. Ωστόσο, οι εφαρμογές της DQL για την κυβερνοφυσική ασφάλεια είναι σχετικά λίγες και πρέπει να διερευνηθούν.

Τέλος στο κεφάλαιο 4 εξετάστηκαν οι εφαρμογές DRL για ασφαλείς επικοινωνίες στα UAV. Αρχικά έγινε μια αναφορά στην αξιοποίηση των UAV στις επικοινωνίες και στη δικτύωση, στη συνέχεια αναφέρθηκαν οι τεχνικές μηχανικής εκμάθησης για ασφαλείς UAV επικοινωνίες των οποίων οι προσεγγίσεις συνοψίζονται στον Πίνακα 9 και τέλος παρατέθηκαν οι πλέον προηγμένες τεχνικές βαθιάς ενίσχυσης μάθησης για ασφαλείς UAV επικοινωνίες, οι αναθεωρημένες προσεγγίσεις των οποίων συνοψίζονται στον Πίνακα 10. Διαφαίνεται ότι, οι τεχνικές με την πάροδο του χρόνου βελτιώνονται, γίνονται πιο πολύπλοκες και τείνουν να ανταποκρίνονται στις ρεαλιστικές απαιτήσεις ασφάλειας επικοινωνιών με χρήση UAV.

Κατά τη διάρκεια της συγγραφής της εν λόγω εργασίας έχουν προκύψει οι παρακάτω προκλήσεις που αφορούν στην βαθιά ενίσχυση της μάθησης στις επικοινωνίες και τη δικτύωση:

- Προσδιορισμός κατάστασης στα δίκτυα υψηλής πυκνότητας: Οι προσεγγίσεις DRL, επιτρέπουν στους χρήστες να βρουν μια βέλτιστη πολιτική πρόσβασης χωρίς να έχουν πλήρεις και/ή ακριβείς πληροφορίες δικτύου. Ωστόσο, οι προσεγγίσεις DRL απαιτούν συχνά από τους χρήστες να αναφέρουν τις τοπικές καταστάσεις τους σε κάθε

χρονική περίοδο. Για να παρατηρήσει την τοπική κατάσταση, ο χρήστης πρέπει να παρακολουθεί τους δείκτες ισχύος λαμβανόμενου σήματος (Received Signal Strength Indicators-RSSI) από τα γειτονικά BS και, στη συνέχεια, συνδέεται προσωρινά στο BS με το μέγιστο RSSI. Ωστόσο, τα μελλοντικά δίκτυα θα αναπτύξουν υψηλή πυκνότητα των BS και τα RSSI από διαφορετικά BS ενδέχεται να μην είναι διαφορετικά. Επομένως, είναι δύσκολο για τους χρήστες να προσδιορίσουν το προσωρινό BS (Cao, Lu, Wen, Lei, & Hu, 2018).

- Γνώση των πληροφοριών καναλιού του παρεμβολέα- υποκλοπέα: Η προσέγγιση DRL για την ασύρματη ασφάλεια όπως προτείνεται στο (Xiao, Xie, Min, & Zhuang, 2018) επιτρέπει στο UAV να βρει τα βέλτιστα επίπεδα ισχύος μετάδοσης για να μεγιστοποιήσει την ικανότητα ασφαλείας του UAV και του BS. Ωστόσο, για να διαμορφωθεί η ανταμοιβή του UAV, απαιτείται άριστη γνώση των πληροφοριών καναλιών των παρεμβολών. Αυτό είναι δύσκολο και μάλιστα αδύνατο στην πράξη.

- Multi-Agent DRL σε δυναμικά HetNets: Οι περισσότερες από τις υπάρχουσες εργασίες επικεντρώνονται στις προσαρμογές του πλαισίου DRL για μεμονωμένες οντότητες δικτύου, με βάση τις τοπικά παρατηρούμενες ή ανταλλασσόμενες πληροφορίες δικτύου. Αν το περιβάλλον δικτύου είναι σχετικά στατικό εξασφαλίζονται συγκλίνοντα αποτελέσματα μάθησης και σταθερές πολιτικές. Αντιθέτως σε ένα δυναμικό ετερογενές δίκτυο 5G, το οποίο αποτελείται από συσκευές/δίκτυα IoT με ταχέως μεταβαλλόμενες απαιτήσεις υπηρεσιών και συνθήκες δικτύωσης, οι πράκτορες DQL για μεμονωμένες οντότητες πρέπει να είναι ευέλικτοι στην αλλαγή των συνθηκών δικτύου. Αυτό συνεπάγεται μείωση της κατάστασης και των χώρων δράσης στη μάθηση, η οποία ωστόσο μπορεί να θέσει σε κίνδυνο την απόδοση της σύγκλισης πολιτικής. Οι αλληλεπιδράσεις μεταξύ πολλών πρακτόρων περιπλέκουν επίσης το περιβάλλον του δικτύου και προκαλούν σημαντική αύξηση του χώρου καταστάσεων, κάτι που αναπόφευκτα επιβραδύνει τους αλγόριθμους εκμάθησης.

- Εκπαίδευση και αξιολόγηση της απόδοσης του πλαισίου DRL: Το πλαίσιο DRL απαιτεί μεγάλο όγκο δεδομένων για την αξιολόγηση τόσο της εκπαίδευσης όσο και της απόδοσης. Στα ασύρματα συστήματα, τέτοια δεδομένα δεν είναι εύκολα προσβάσιμα καθώς σπάνια έχουμε σχετικές δεξαμενές δεδομένων. Οι περισσότερες από τις υπάρχουσες εργασίες βασίζονται σε σύνολο δεδομένων προσομοιώσεων, το οποίο δεν

ανταποκρίνεται σε ρεαλιστικά συστήματα. Το σύνολο δεδομένων προσομοιώσεων δημιουργείται συνήθως από ένα συγκεκριμένο στοχαστικό μοντέλο, το οποίο είναι μια απλοποίηση του πραγματικού συστήματος και μπορεί να παραβλέπει τα κρυφά μοτίβα. Ως εκ τούτου, απαιτείται ένας πιο αποτελεσματικός τρόπος για τη δημιουργία δεδομένων προσομοίωσης για να διασφαλιστεί ότι η εκπαίδευση και η αξιολόγηση απόδοσης του πλαισίου DRL είναι πιο συνεπής με τα ρεαλιστικά συστήματα.

Παράρτημα Α

Πίνακας Συντμήσεων

A3C	Asynchronous Advantage Actor-Critic
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Access Point
APF	Artificial Potential Field
AV	Autonomous Vehicle
BBU	BaseBand Unit
BER	Bit Error Rate
BS	Base Station
CAA-MADDPG	Continuous Action Attention-Multi-Agent Deep Deterministic Policy Gradient
CNN	Convolutional Neural Network
CRN	Cognitive Radio Network
CSI	Channel State Information
D2D	Device to Device
DASH	Dynamic Adaptive Streaming over HTTP
DDPG	Deep Deterministic Policy Gradient
DDQN	Double DQN
DEI	Delayed Experience Injection
DL	Deep Learning
DNN	Deep Neural Network
DoS	Denial of Service
DPG	Deterministic Policy Gradient
DQL	Deep Q-Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DRQN	Deep Recurrent Q-Learning
ESN	Echo State Network
FIFO	First In First Out

FNN	Feedforward Neural Network
FSMC	Finite-State Markov Channel
GE	Ground Eavesdropper
GPS	Global Positioning System
GU	Ground User
HAP	High-Altitude Platform
HVFT	High Volume Flexible Time
IoT	internet of Things
ITS	Intelligent Transportation System
K-NN	K-Nearest Neighbours
LoS	Line of Sight
LSM	Liquid State Machine
LSTM	Long Short Term Memory
LTE	Long Term Evolution
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MADRL	Multi-Agent Deep Reinforcement Learning
MBS	Medium range Base Station
MDP	Markov Decision Process
MEC	Mobile Edge Computing
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MU	Mobile User
NAF	Normalized Advantage Function
NE	Nash Equilibrium
NFSP	Neural Fictitious Self-Play
NFV	Network Function Virtualization
OBU	On-Board Unit
PD	Prisoner's Dilemma
PER	Prioritized Experience Replay
PHC	Policy Hill Climbing
PLS	Physical Layer Security
POMDP	Partially Observable MDP
PT	Prospect Theory
PU	Primary User
QoE	Quality of Experience

QoS	Quality of Service
RCNN	Recursive Convolutional Neural Network
RDPG	Recurrent Deterministic Policy Gradient
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RRH	Remote Radio Head
RSSI	Received Signal Strength Indicators
RSU	Road Side Unit
RTT	Round Trip Time
SBS	Small Base Station
SDN	Software-Defined Network
SGD	Stochastic Gradient Descent
SINR	Signal to Interference plus Noise Ratio
SNN	Spiking Neural Network
SNR	Signal to Noise Ratio
SPD	Sequential Prisoner's Dilemma
SU	Secondary User
SVN	Support Vector Machine
TD	Temporal Difference
TTL	Time To Live
UAN	Uderwater Acoustic Network
UAV	Unmanned Aerial Vehicle
UDN	Ultra-Density Network
UT	User Terminal
V2V	Vehicle to Vehicle
VANET	Vehicular Ad hoc Network
VR	Virtual Reality
WLAN	Wireless Local Area Network
WoLF-PHC	Win or Learn Faster-Policy Hill Climbing
H/M	Ηλεκτρομαγνητικό

Βιβλιογραφία

Alpaydin, E. (2014). *Introduction to Machine Learning.* Cambridge, MA, USA: MIT Press.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Barath, A. A. (2017). A Brief Summary of Deep Reinforcement Learning. *IEEE Signal Process Magazine , 34*, 26-38.

Balazinska, M., & Castro, P. (2003, February 19). *IBM Watson research center*. Ανάκτηση από https://crawdad.org/ibm/watson/20030219/.

BBC. (2016, Jan 27). *Google Achieves AI 'Breakthrough' by Beating Go Champion*. Ανάκτηση από BBC news: www.bbc.com/news/technology-35420579

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A Distributional Perspective on Reinforcement Learning. *Proceedings of the 34th International Conference on Machine Learning* (σσ. 449-458). Proceedings of Machine Learning Research.

Bellman, R. (2013). *Dynamic Programming.* Mineola, NY, USA: Courier Corporation.

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control* (Τόμ. 1). Belmont, MA, USA: Athena Scientific.

Bithas, P. S., Michailidis, E. T., Nomikos, N., Vouyioukas, D., & Kanatas, A. G. (2019). A Survey on Machine-Learning Techniques for UAV-Based Communications. *Sensors* .

Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence , 136* (2), 215-250.

Brackstone, M., & McDonald, M. (1999). Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour , 2* (4), 181-196.

Challita, U., Dong, L., & Saad, W. (2017, December 15). Proactive Resource Management for LTE in Unlicensed Spectrum: A Deep Learning Perspective. *arXiv.org arXiv:1702.07031v2* .

Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2017, October 9). Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks. *arXiv.org arXiv:1710.02913* .

Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2017, Oct 9). Machine Learning for Wireless Networks with Artificial Intelligence: A Tutorial on Neural Networks. arXiv preprint arXiv:1710.02913. Ανάκτηση από arXiv.org.

Chen, M., Mozaffari, M., Saad, W., Yin, C., Debbah, M., & Hong, C. S. (2017). Caching in the Sky: Proactive Deployment of Cache-Enabled Unmanned Aerial Vehicles for Optimized Quality-of-Experience. *IEEE Journal on Selected Areas in Communications , 35* (5), 1046 - 1061.

Chen, M., Saad, W., & Yin, C. (2019). Echo-Liquid State Deep Learning for 360° Content Transmission and Caching in Wireless VR Networks With Cellular-Connected UAVs. *IEEE Transactions on Communications , 67* (9), 6386 - 6400.

Chen, M., Saad, W., & Yin, C. (2017). Liquid State Machine Learning for Resource Allocation in a Network of Cache-Enabled LTE-U UAVs. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference.* Singapore: IEEE.

Chen, X., Zhang, H., Wu, C., Mao, S., Ji, Y., & Bennis, M. (2019). Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning. *IEEE Internet of Things Journal , 6* (3), 4005 - 4018.

Chen, X., Zhang, H., Wu, C., Mao, S., Ji, Y., & Bennis, M. (2018). Performance Optimization in Mobile-Edge Computing via Deep Reinforcement Learning. *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall).* Chicago, IL, USA: IEEE.

Chen, Y., Kar, S., & Moura, J. M. (2018). Cyber-Physical Attacks With Control Objectives. *IEEE Transactions on Automatic Control , 63* (5), 1418 - 1425.

Chen, Y., Li, Y., Xu, D., & Xiao, L. (2018). DQN-Based Power Control for IoT Transmission against Jamming. *2018 IEEE 87th Vehicular Technology Conference (VTC Spring).* Porto, Portugal: IEEE.

Chen, Y., Zhao, N., Ding, Z., & Alouini, M. (2018). "Multiple UAVs as relays:Multi-hop single link versus multiple dual-hop links. *IEEE Trans. Wireless Commun* , 6348-6359.

Cheng, F., Gui, G., Zhao, N., Chen, Y., Tang, J., & Sari, H. (2019). UAV-relaying assisted secure transmission with caching. *IEEE Trans. Commun.* , 3140-3153.

Cheng, F., Zhang, S., Yunfei, C., Zhao, N., Yu, F. R., & Leung, V. C. (2018). UAV Trajectory Optimization for Data Offloading at the Edge of Multiple Cells. *IEEE Transactions on Vehicular Technology* , 6732-6736.

Chinchali, S., Hu, P., Chu, T., Sharma, M., Bansal, M., Misra, R., και συν. (2018). Cellular Network Traffic Scheduling With Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31.*

Chu, M., Li, H., Liao, X., & Cui, S. (2019). Reinforcement Learning-Based Multiaccess Control and Battery Prediction With Energy Harvesting in IoT Systems. *IEEE Internet of Things Journal , 6* (2), 2009 - 2020.

Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., & Sears, R. (2010). Benchmarking cloud serving systems with YCSB. *Proceedings of the 1st ACM symposium on Cloud computing* (σσ. 143-154). SoCC '10.

Dabney, W. C. (2014). *Adaptive step-sizes for reinforcement learning.* Amherst: University Massachusetts.

Deghel, M., Bastug, E., Assaad, M., & Debbah, M. (2015). On the benefits of edge caching for MIMO interference alignment. *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).* Stockholm, Sweden: IEEE.

Di, X., Xiong, K., Fan, P., Yang, H.-C., & Letaief, K. B. (2017). Optimal Resource Allocation in Wireless Powered Communication Networks With User Cooperation. *IEEE Transactions on Wireless Communications , 16* (12), 7936 - 7949.

Ding, G., Wu, Q., Zhang, L., Lin, Y., Tsiftsis, T. A., & Yao, Y. D. (2018). An amateur drone surveillance system based on the cognitive Internet of Things. *IEEE Commun. Mag.* , 29-35.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., και συν. (2015). Lomg-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (σσ. 2625-2634). Boston: Computer Vision Foundation.

Dulac-Arnold, G., Evans, R., Sunehag, P., & Coppin, B. (2015). Reinforcement Learning in Large Discrete Action Spaces. *CoRR* .

Elsherif, A. R., Chen, W.-P., Ito, A., & Ding, Z. (2015). Resource Allocation and Inter-Cell Interference Management for Dual-Access Small Cells. *IEEE Journal on Selected Areas in Communications , 33* (6), 1082 - 1096.

Fadlullah, Z. M., Tang, F., Mao, B., Kato, N., Akashi, O., Inoue, T., και συν. (2017). State-of-the-art Deep Learning:Evolving Machine Intelligence toward Tomorrow's Intelligent Network Traffic Control Systems. *IEEE Communications Surveys and Tutorials , 19* (4), 2432-2455.

Fearnley, J. (2010, August). Strategy Iteration Algorithms for Games and Markov Decision Processes. *A Thesis Submitted for the Degree of PhD at the University of Warwick* . Warwick, uk: The University of Warwick.

Ferdowsi, A., & Saad, W. (2019). Deep Learning for Signal Authentication and Security in Massive Internet-of-Things Systems. *IEEE Transactions on Communications , 67* (2), 1371-1387.

Ferdowsi, A., & Saad, W. (2018). Deep Learning-Based Dynamic Watermarking for Secure Signal Authentication in the Internet of Things. *2018 IEEE International Conference on Communications (ICC).* Kansas City, MO, USA: IEEE.

Ferdowsi, A., Challita, U., Saad, W., & Mandayam, N. B. (2018). Robust Deep Reinforcement Learning for Security and Safety in Autonomous Vehicle Systems. *2018 21st International Conference on Intelligent Transportation Systems (ITSC).* Maui, HI, USA: IEEE.

Ferreira, P. V., Paffenroth, R., Wyglinski, A. M., Hackett, T. M., Bilen, S. G., Reinhart, R. C., και συν. (2018). Multiobjective Reinforcement Learning for Cognitive Satellite Communications Using Deep Neural Network Ensembles. *IEEE Journal on Selected Areas in Communications , 36* (5), 1030 - 1041.

Fooladivanda, D., & Rosenberg, C. (2013). Joint Resource Allocation and User Association for Heterogeneous Wireless Cellular Networks. *IEEE Transactions on Wireless Communications , 12* (1), 248 - 257.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Hessel, M., Osband, I., και συν. (2018). Noisy Networks For Exploration. *6th International Conference on Learning Representations.* Vancouver: ICLR 2018 Conference.

Gadaleta, M., Chiarriotti, F., Rossi, M., & Zanella, A. (2017). D-DASH: A Deep Q-Learning Framework for DASH Video Streaming. *IEEE Transactions on Cognitive Communications and Networking , 3* (4), 703 - 718.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, U.K.: MIT Press.

Govindan, R. (n.d.). *Tutornet: A Low Power Wireless IoT Testbed*. Ανάκτηση από http://anrg.usc.edu/www/tutornet/.

Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *IEEE International Conference on Robotics and Automation (ICRA)*, (σσ. 3389-3396). Singapore.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks , 5* (6), 989 - 993.

Han, B., Gopalakrishnan, V., Ji, L., & Lee, S. (215). Network function virtualization: Challenges and opportunities for innovations. *IEEE Communications Magazine , 53* (2), 90-97.

Han, G., Xiao, L., & Poor, H. V. (2017). Two-dimensional anti-jamming communication based on deep reinforcement learning. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, LA, USA: IEEE.

Hasselt, H. v. (2010). Double Q-learning. *Proceedings of the 23rd International Conference on Neural Information. 2*, σσ. 2613-2621. NIPS'10.

Hasselt, H. v., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-Learning. *Thirtieth AAAI Conference on Artificial Intelligence* , *30*, σσ. 2094-2100.

Hausknecht, M., & Stone, P. (2015, July). Deep Recurrent Q-Learning for Partially Observable MDPs. *arXiv.org eprint arXiv:1507.06527* .

He, X., Wang, K., Huang, H., Miyazaki, T., Wang, Y., & Guo, S. (2020). Green Resource Allocation Based on Deep Reinforcement Learning in Content-Centric IoT. *IEEE Transactions on Emerging Topics in Computing , 8* (3), 781-796.

He, Y., & Hu, S. (2017, June 27). Cache-enabled Wireless Networks with Opportunistic Interference Alignment. *arXiv.org arXiv:1706.09024v1* .

He, Y., Liang, C., Yu, F. R., & Han, Z. (2020). Trust-Based Social Networks with Computing, Caching and Communications: A Deep Reinforcement Learning Approach. *IEEE Transactions on Network Science and Engineering , 7* (1), 66-79.

He, Y., Liang, C., Yu, F. R., Zhao, N., & Yin, H. (2017). Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach. *2017 IEEE International Conference on Communications (ICC).* Paris, France: IEEE.

He, Y., Liang, C., Zhang, Z., Yu, F. R., Zhao, N., Yin, H., και συν. (2017). Resource Allocation in Software-Defined and Information-Centric Vehicular Networks with Mobile Edge Computing. *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall).* Toronto, ON, Canada: IEEE.

He, Y., Yu, F. R., Zhao, N., & Yin, H. (2018). Secure Social Networks in 5G Systems with Mobile Edge Computing, Caching, and Device-to-Device Communications. *IEEE Wireless Communications , 25* (3), 103-109.

He, Y., Yu, F. R., Zhao, N., Leung, V. C., & Yin, H. (2017). Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach. *IEEE Communications Magazine , 55* (12), 31-37.

He, Y., Yu, F. R., Zhao, N., Yin, H., & Boukerche, A. (2017). Deep Reinforcement Learning (DRL)-based Resource Management in Software-Defined and Virtualized Vehicular Ad Hoc Networks. *Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications* (σσ. 47-54). DIVANet '17.

He, Y., Zhang, Z., & Zhang, Y. (2017). A Big Data Deep Reinforcement Learning Approach to Next Generation Green Wireless Networks. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference.* Singapore: IEEE.

He, Y., Zhang, Z., Yu, F. R., Zhao, N., Yin, H., Leung, V., και συν. (2017). Deep-Reinforcement-Learning-Based Optimization for Cache-Enabled Opportunistic Interference Alignment Wireless Networks. *IEEE Transactions on Vehicular Technology , 66* (11), 10433 - 10445.

He, Y., Zhao, N., & Yin, H. (2017). Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology , 67* (1), 44-55.

Heinrich, J., & Silver, D. (2016, June 28). Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *arXiv.org arXiv:1603.01121v2* .

Hessel, M., Modayil, J., Hasselt, H. v., Schaul, T., Ostrovski, G., Dabney, W., και συν. (2018). Rainbow: Combining Improvements in Deep Reinforcement Learning. *Thirty-Second AAAI Conference on Artificial Intelligence .* New Orleans: AAAI Press.

Hoang, T. M., Nguyen, N. M., & Duong, T. Q. (2019). Detection of eavesdropping attack in UAV-aided wireless systems:Unsupervised learning with one-class SVM and k-means clustering. *IEEE Wirel. Commun. Lett.*

Hu, J., & Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of machine learning research , 4*, 1039-1069.

Huang, T., Zhang, R.-X., Zhou, C., & Sun, L. (2018). QARC: Video Quality Aware Rate Control for Real-Time Video Streaming based on Deep Reinforcement Learning. *Proceedings of the 26th ACM international conference on Multimedia* (σσ. 1208-1216). MM '18.

Huang, W., Wang, Y., & Yi, X. (2017). A deep reinforcement learning approach to preserve connectivity for multi-robot systems. *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).* Shanghai, China: IEEE.

Huang, W., Wang, Y., & Yi, X. (2017). Deep Q-Learning to Preserve Connectivity in Multi-robot Systems. *Proceedings of the 9th International Conference on Signal Processing Systems* (σσ. 45-50). ICSPS 2017.

Ji, L., Hui, G., Tiejun, L., & Yueming, L. (2018). Deep reinforcement learning based computation offloading and resource allocation for MEC. *2018 IEEE Wireless Communications and Networking Conference (WCNC).* Barcelona, Spain: IEEE.

Jing, L., Jia, X., Lv, Y., & Wan, N. (2021). *IAEAC.* IEEE.

Klaue, J., Rathke, B., & Wolisz, A. (2003). EvalVid—A framework for video transmission and quality evaluations. *In Proceedings of the International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, (σσ. 255-272). Urbana, Ilinois, USA.

Koch, W., Mancuso, R., West, R., & Bestavros, A. (2019). Reinforcement learning for UAV attitude control. *ACM Trans. Cyber-Phys.Syst.* , 1-21.

Le, D. V., & Tham, C.-K. (2018). A deep reinforcement learning based offloading scheme in ad-hoc mobile clouds. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* Honolulu, HI, USA: IEEE.

Le, D. V., & Tham, C.-K. (2018). Quality of Service Aware Computation Offloading in an Ad-Hoc Mobile Cloud. *IEEE Transactions on Vehicular Technology , 67* (9), 8890 - 8904.

Lei, L., You, L., Dai, G., Vu, T. X., Yuan, D., & Chatzinotas, S. (2017). A deep learning approach for optimizing content delivering in cache-enabled HetNet. *2017 International Symposium on Wireless Communication Systems (ISWCS).* Bologna, Italy: IEEE.

Li, C., Xu, Y., Xia, J., & Zhao, J. (2018). Protecting secure communication under UAV smart attack with imperfect channel estimation. *IEEE Access* , 76395-76401.

Li, H. (2010, May 23). Multiagent Q-Learning for Aloha-Like Spectrum Access in Cognitive Radio Systems. *EURASIP Journal on Wireless Communications and Networking* .

Li, H., Gao, H., Lv, T., & Lu, Y. (2018). Deep Q-Learning Based Dynamic Resource Allocation for Self-Powered Ultra-Dense Networks. *2018 IEEE International Conference on Communications Workshops (ICC Workshops).* Kansas City, MO, USA: IEEE.

Li, Y. (2018, November 26). *Deep Reinforcement Learning.* Ανάκτηση από arXiv.org: https://arxiv.org/abs/1701.07274

Li, Y., Liu, J., Li, Q., & Xiao, L. (2015). Mobile cloud offloading for malware detections with learning. *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* Hong Kong, China: IEEE.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., και συν. (2016, February 29). Continuous control with deep reinforcement learning. *arXiv.org arXiv:1509.02971 v5* .

Lin, Y., Bao, W., Yu, W., & Liang, B. (2015). Optimizing User Association and Spectrum Allocation in HetNets: A Utility Perspective. *IEEE Journal on Selected Areas in Communications , 33* (6), 1025 - 1039.

Lin, Y., Dai, X., Li, L., & Wang, F.-Y. (2018, Aug). An Efficient Deep Reinforcement Learning Model for Urban Traffic Control. *arXiv preprint arXiv:1808.01876* , 1-10.

Liu, J., Krishnamachari, B., Zhou, S., & Niu, Z. (2018). DeepNap: Data-Driven Base Station Sleeping Operations Through Deep Reinforcement Learning. *IEEE Internet of Things Journal , 5* (6), 4273 - 4282.

Liu, S., Hu, X., & Wang, W. (2018). Deep Reinforcement Learning Based Dynamic Channel Allocation Algorithm in Multibeam Satellite Systems. *IEEE Access , 6*, 15733 - 15742.

Liu, X., Xu, Y., Jia, L., Wu, Q., & Anpalagan, A. (2018). Anti-Jamming Communications Using Spectrum Waterfall: A Deep Reinforcement Learning Approach. *IEEE Communications Letters , 22* (5), 998-1001.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., & Mordatch, I. (2017). *Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.* Ανάκτηση από arXiv.org: https://arxiv.org/abs/1706.02275

Lu, X., Xiao, L., Dai, C., & Dai, H. (2020). UAV-Aided Cellular Communications with Deep Reinforcement Learning Against Jamming. *IEEE Wireless Communications , 27* (4), 48-53.

Lv, S., Xiao, L., Hu, Q., Wang, X., Hu, C., & Sun, L. (2017). Anti-Jamming Power Control Game in Unmanned Aerial Vehicle Networks. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference.* Singapore: IEEE.

Maass, W. (2011). Liquid State Machines: Motivation, Theory, and Applications. *World Scientific* , 275-296.

Manesh, M. R., Kenney, J., Hu, W. C., Devabhaktuni, V. K., & Kaabouch, N. (2019). Detection of GPS spoofing attacks on unmanned aerial systems. *IEEE Cons. Commun. Netw. Conf. (CCNC)* (σσ. 1-6). IEEE.

Mao, H., Netravali, R., & Alizadeh, M. (2017). Neural Adaptive Video Streaming with Pensieve. *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (σσ. 197-210). SIGCOMM '17.

Mao, Q., Hu, F., & Hao, Q. (2018). Deep Learning for Intelligent Wireless Networks:A Comprehensive Survey. *IEEE Communications Surveys and Tutorials , 20* (4), 2595-2621.

Min, M., Xiao, L., Chen, Y., Cheng, P., Wu, D., & Zhuang, W. (2019). Learning-Based Computation Offloading for IoT Devices With Energy Harvesting. *IEEE Transactions on Vehicular Technology , 68* (2), 1930-1941.

Mismar, F. B., & Evans, B. L. (2018). Deep Q-Learning for Self-Organizing Networks Fault Management and Radio Performance Improvement. *2018 52nd Asilomar Conference on Signals, Systems, and Computers.* Pacific Grove, CA, USA: IEEE.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., και συν. (2016). Asynchronous Methods for Deep Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning* (σσ. 1928-1937). Proceedings of Machine Learning Research.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., και συν. (2013, December 19). Playing Atari with Deep Reinforcement Learning. *arXiv.org arXiv:1312.5602* .

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., και συν. (2015). Human-level control through deep reinforcement learning. *Nature , 518*, 529-533.

Monahan, G. E. (1982). State of the art - A survey of partially observable Markov decision process: Theory, models and algorithms. *Management Science , 28* (1), 1-16.

Naparstek, O., & Cohen, K. (2017). Deep Multi-User Reinforcement Learning for Dynamic Spectrum Access in Multichannel Wireless Networks. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference.* Singapore: IEEE.

Poonawala, H. A., Satici, A. C., Eckert, H., & Spong, M. W. (2015). Collision-Free Formation Control with Decentralized Connectivity Preservation for Nonholonomic-Wheeled Mobile Robots. *IEEE Transactions on Control of Network Systems , 2* (2), 122-130.

Puterman, M. L. (2014). *Markov Decision Process.* New York, NY, USA: Wiley.

Quan, L., Wang, Z., & Ren, F. (2018). A Novel Two-Layered Reinforcement Learning for Task Offloading with Tradeoff between Physical Machine Utilization Rate and Delay. *Future Internet , 10* (7), 60.

Raw Data - Measuring Broadband America 2016. (n.d.). Ανάκτηση από Federal Communication Commission: https://www.fcc.gov/reports-research/reports/measuring-broadband-america/raw-data-measuring-broadband-america-2016

Riiser, H., Vigmostad, P., & Griwodz, G. (2013). Commute path bandwidth traces from 3G networks: analysis and applications. *Proceedings of the 4th ACM Multimedia Systems Conference* (σσ. 114-118). MMSys '13.

Satchidanandan, B., & Kumar, P. R. (2017). Dynamic Watermarking: Active Defense of Networked Cyber–Physical Systems. *Proceedings of the IEEE , 105* (2), 219-240.

Schaarschmidt, M., Gessert, F., Dalibard, V., & Yoneki, E. (2016, October 31). Learning Runtime Parameters in Computer Systems with Delayed Experience Injection. *arXiv.org arXiv:1610.09903v1* .

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016, February 25). Prioritized Experience Replay. *arXiv.org arXiv:1511.05952* .

Shamili, A. S., Bauckhage, C., & Alpcan, T. (2010). Malware Detection on Mobile Devices Using Distributed Machine Learning. *2010 20th International Conference on Pattern Recognition.* Istanbul, Turkey: IEEE.

Shapley, L. S. (1953). Stochastic games. *Proceedings of National Academy Sciences , 39* (10), 1095-1100.

Shen, C., Tekin, C., & van der Schaar, M. (2016). A Non-Stochastic Learning Approach to Energy Efficient Mobility Management. *IEEE Journal on Selected Areas in Communications , 34* (12), 3854 - 3868.

Shiu, Y. S., Chang, S. Y., Wu, H. C., Huang, C. H., & Chen, H. H. (2011). Physical layer security in wireless networks: A tutorial. *IEEE Wireless Communication* , 66-74.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic Policy Gradient Algorithms. *Proceedings of the 31st International Conference on Machine Learning. 32*, σσ. 387-395. PMLR.

Stockhammer, T. (2011). Dynamic adaptive streaming over HTTP: standards and design principles. *Proceedings of the second annual ACM conference on Multimedia systems* (σσ. 133-144). MMSys '11.

Sutton, S. R., & Barto, A. G. (1998). *Reinforcement Learning:An Introduction.* Cambridge, U.K.: MIT Press.

Szita, I., Gyenes, V., & Lorincz, A. (2006). Reinforcement Learning with Echo State Networks. *International Conference on Artificial Neural Networks*, (σσ. 830-839).

Tan, L. T., & Hu, R. Q. (2018). Mobility-Aware Edge Caching and Computing in Vehicle Networks: A Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology , 67* (11), 10190 - 10203.

Tang, F., Kawamoto, Y., Kato, N., & Liu, J. (2020). Future intelligent and secure vehicular network towards 6G: Machine-learning approaches. *IEEE* , 292-307.

Tang, Z., Zhou, X., Zhang, F., Jia, W., & Zhao, W. (2019). Migration Modeling and Learning Algorithms for Containers in Fog Computing. *IEEE Transactions on Services Computing , 12* (5), 712-725.

Tarchi, D., Corazza, G. E., & Vanelli-Coralli, A. (2013). Adaptive coding and modulation techniques for next generation hand-held mobile satellite communications. *2013 IEEE International Conference on Communications (ICC).* Budapest, Hungary: IEEE.

Thrun, S., & Schwartz, A. (1993). *Issues in using function approximation for reinforcement learning.* Hillsdale, New Jersey, US: Lawrence Erlbaum Asssociates.

Tyouku of China Network Video Index. (n.d.). Ανάκτηση από http://index.youku.com/

Vadakkepat, P., Tan, K. C., & Ming-Liang, W. (2000). Evolutionary artificial potential fields and their application in real time robot path planning. *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512).* La Jolla, CA, USA: IEEE.

Wan, X., Sheng, G., Li, Y., Xiao, L., & Du, X. (2017). Reinforcement Learning Based Mobile Offloading for Cloud-Based Malware Detection. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference.* Singapore: IEEE.

Wang, C., Wang, J., Zhang, X., & Zhang, X. (2017). Autonomous navigation of UAV in large-scale unknown complex environment with deep reinforcement learning. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* Montreal, QC, Canada: IEEE.

Wang, H., Wang, J., Ding, G., Chen, J., Li, Y., & Han, Z. (2018). Spectrum Sharing Planning for Full-Duplex UAV Relaying Systems With Underlaid D2D Communications. *IEEE Journal on Selected Areas in Communications* , 1986-1999.

Wang, S., Liu, H., Gomes, P. H., & Krishnamachari, B. (2017). Deep Reinforcement Learning for Dynamic Multichannel Access. *2017-International Conference on Computer Network and Communication technologies.* ICNC.

Wang, S., Liu, H., Gomes, P. H., & Krishnamachari, B. (2018). Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks. *IEEE Transactions on Cognitive Communications and Networking , 4* (2), 257 - 265.

Wang, W., Hao, J., Wang, Y., & Taylor, M. (2018, March 1). Towards Cooperation in Sequential Prisoner's Dilemmas: a Deep Multiagent Reinforcement Learning Approach. *arXiv.org arXiv:1803.00162* .

Wang, W., Kwasinski, A., Niyato, D., & Han, Z. (2016). A Survey on Applications of Model-Free Strategy Learning in Cognitive Wireless Networks. *IEEE Communications Surveys and Tutorials , 18* (3), 1717-1757.

Wang, Z., Li, L., Xu, Y., Tian, H., & Cui, S. (2018). Handover Control Optimization via Asynchronous Multi-User Deep Reinforcement Learning. *2018 IEEE International Conference on Communications (ICC).* Kansas City: IEEE.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. (2016). Dueling Network Architectures for Deep Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning. 48.* Proceedings of Machine Learning Research.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning , 8* (3-4), 279-292.

Wu, Q., Li, Z., & Xie, G. (2013). CodingCache: multipath-aware CCN cache with network coding. *Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking* (σσ. 41-42). ICN '13.

Xiao, L., Jiang, D., Wan, X., Su, W., & Tang, Y. (2018). Anti-Jamming Underwater Transmission With Mobility and Learning. *IEEE Communications Letters , 22* (3), 542 - 545.

Xiao, L., Lu, X., Xu, D., Tang, Y., Wang, L., Zhuang, και συν. (2018). UAV Relay in VANETs Against Smart Jamming With Reinforcement Learning. *IEEE Transactions on Vehicular Technology , 67* (5), 4087-4097.

Xiao, L., Wan, X., Dai, C., Du, X., Chen, X., & Guizani, M. (2018). Security in Mobile Edge Caching with Reinforcement Learning. *IEEE Wireless Communications , 25* (3), 116-122.

Xiao, L., Xie, C., Min, M., & Zhuang, W. (2018). User-Centric View of Unmanned Aerial Vehicle Transmission Against Smart Attacks. *IEEE Transactions on Vehicular Technology , 67* (4), 3420-3430.

Xin, Y. (2018). Machine Learning and Deep learning Methods for Cybersecurity. *IEEE Access , 6*, 35365-35381.

Yang, Y., Zheng, Z., Bian, K., Song, L., & Han, Z. (2018). Real-time profiling of fine-grained air quality index distribution using UAV sensing. *IEEE Internet Things J.* , 186-198.

Ye, H., & Li, G. Y. (2018). Deep Reinforcement Learning for Resource Allocation in V2V Communications. *2018 IEEE International Conference on Communications (ICC).* Kansas City, MO, USA: IEEE.

Ye, J., & Zhang, Y.-J. A. (2020). DRAG: Deep Reinforcement Learning Based Base Station Activation in Heterogeneous Networks. *IEEE Transactions on Mobile Computing , 19* (9), 2076 - 2087.

Yilmaz, M. H., & Arslan, H. (2015). A survey: Spoofing attacks in physical layer security. *Proc. IEEE Local Comput. Netw. Conf. Workshops* (σσ. 812-817). Clearwater Beach, Florida, USA: IEEE.

Yin, X., Jindal, A., Sekar, V., & Sinopoli, B. (2015). A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication* (σσ. 325-338). SIGCOMM '15.

Yu, S., Wang, X., & Langar, R. (2017). Computation offloading for mobile edge computing: A deep learning approach. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC).* Montreal, QC, Canada: IEEE.

Zhang, C., Gu, B., Liu, Z., Yamori, K., & Tanaka, Y. (2018). Cost- and Energy-Aware Multi-Flow Mobile Data Offloading Using Markov Decision Process. *IEICE Transactions on Communications , E101-B* (3), 657-666.

Zhang, C., Liu, Z., Gu, B., Yamori, K., & Tanaka, Y. (2018). A Deep Reinforcement Learning Based Approach for Cost- and Energy-Aware Multi-Flow Mobile Data Offloading. *IEICE Transactions on Communications , E101-B* (7), 1625-1634.

Zhang, Y., Mou, Z., Gao, F., Jiang, J., Ding, R., & Han, Z. (2020). UAV-Enabled Secure Communications by Multi-Agent Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology , 69* (10), 11599-11611.

Zhang, Y., Zhuang, Z., Gao, F., Wang, J., & Han, Z. (2020). Multi-Agent Deep Reinforcement Learning for Secure UAV Communications. *2020 IEEE Wireless Communications and Networking Conference (WCNC).* Seoul, Korea: IEEE.

Zhang, Z., Zheng, Y., Li, C., Huang, Y., & Yang, L. (2018). Cache-Enabled Adaptive Bit Rate Streaming via Deep Self-Transfer Reinforcement Learning. *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP).* Hangzhou, China: IEEE.

Zhao, D., Wang, H., Shao, K., & Zhu, Y. (2016). Deep reinforcement learning with experience replay based on SARSA. *2016 IEEE Symposium Series on Computational Intelligence (SSCI).* Athens: IEEE.

Zhao, N., Liang, Y.-C., Niyato, D., Pei, Y., Wu, M., & Jiang, Y. (2018, December). Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Networks. *2018 IEEE Global Communications Conference (GLOBECOM)* .

Zhao, N., Lu, W., Sheng, M., Chen, Y., Tang, J., Yu, F. R., και συν. (2019). UAV-assisted emergency networks in disasters. *IEEE Wireless Communications* , 45-51.

Zhao, Q., Krishnamachari, B., & Liu, K. (2008). On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *IEEE Transactions on Wireless Communications , 7* (12), 5431 - 5440.

Zhong, C., Gursoy, M. C., & Velipasalar, S. (2018). A deep reinforcement learning-based framework for content caching. *2018 52nd Annual Conference on Information Sciences and Systems (CISS).* Princeton, NJ, USA: IEEE.

Zhong, C., Yao, J., & Xu, J. (2019). Secure UAV communication with cooperative jamming and trajectory control. *IEEE Commun. Lett.* , 286-289.

Zhu, J., Song, Y., Jiang, D., & Song, H. (2018). A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things. *IEEE Internet of Things Journal , 5* (4), 2375 - 2385.