ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Βαθιά ενίσχυση της μάθησης στις επικοινωνίες και στη δικτύωση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΓΕΩΡΓΑΛΟΠΟΥΛΟΥ

Επιβλέπων : ΔΗΜΟΣΘΕΝΗΣ ΒΟΥΓΙΟΥΚΑΣ, Καθηγητής τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων (Μ.Π.Ε.Σ.)

Μέλη εξεταστικής επιτροπής: ΔΗΜΟΣΘΕΝΗΣ ΒΟΥΓΙΟΥΚΑΣ, ΔΗΜΗΤΡΙΟΣ ΣΚΟΥΤΑΣ, ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΑΛΙΑΤΣΟΣ

Σάμος, Απρίλιος 2022

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΓΕΩΡΓΑΛΟΠΟΥΛΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

Λίστα Σχημάτων

Λίστα Πινάκων

Ακρωνύμια

ADC	Analog to Digital Converter
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASIC	Application Specific Integrated Circuits
AV	Autonomous Vehicle
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BLER	BLock Error Rate
BS	Base Station
CNN	Convolutional Neural Network
CRN	Cognitive Radio Network
DAC	Digital to Analog Converter
DDPG	Deep Deterministic Policy Gradient
DDQL	Double Deep Q-Learning
DDQN	Double Deep Q-Network
DL	Deep Learning
DNN	Deep Neural Network
DPG	Deterministic Policy Gradient
DPN	Data Processing Networks
DQL	Deep Q-Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DRQL	Deep Recurrent Q-Learning
DRQN	Deep Recurrent Q-Network
FA	False Alarm
FCL	Fully Connected Layer
FNN	Feedforward Neural Network
FPGA	Field Programmable Gate Arrays
GDR	Generalized Data Representation
GPS	Guided Policy Search
GPU	Graphics Processing Unit
HAPS	High Altitude Platform System
HetNets	Heterogeneous Networks
IoT	Internet of Things
ITS	Intelligent Transportation System
KPI	Key Performance Indicators
LEO	Low Earth Orbit
LFU	Least Frequently Used

LIDAR	LIght Detection And Ranging
LRU	Least Recently Used
LSTM	Long Short-Term Memory
MD	Missed Detection
MDP	Markov Decision Process
MEC	Mobile Edge Caching
MIMO	Multi-Input Multi-Output
ML	Machine Learning
MPC	Model Predictive Control
MSE	Mean Squared Error
NAF	Normalized Advantage Functions
NFSP	Neural Fictitious Self-Play
NLOS	Non-line Of Sight
PER	Prioritized Experience Replay
PID	Proportional Integral Derivative
PU	Primary User
RAN-Core	RadioAccessNetwork-Core
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SARSA	State Action Reward State Action
SDN	Software Defined Network
SDP	Session Description Protocol
SDR	Software Defined Radio
SER	Symbol Error Rate
SINR	Signal to Interference & Noise Ratio
SPD	Sequential Prisoner's Dilemma
SU	Secondary User
TD	Temporal Difference
UAV	Unmanned Aerial Vehicles
WoLF-PHC	Win or Learn Fast- Policy Hill Climbing

Περίληψη

Τα σύγχρονα δίκτυα, π.χ. του Διαδικτύου των Πραγμάτων (IoT) και των Μη επανδρωμένων Εναέριων Οχημάτων (UAV), καθίστανται πιο αποκεντρωμένα και αυτόνομα. Σε τέτοια δίκτυα, οι φορείς δικτύου πρέπει να λαμβάνουν αποφάσεις τοπικά για να μεγιστοποιούν την απόδοση του δικτύου υπό την αβεβαιότητα του περιβάλλοντος. Η ενίσχυση της μάθησης (RL) έχει χρησιμοποιηθεί αποτελεσματικά για να δώσει τη δυνατότητα στις οντότητες του δικτύου να αποκτήσουν τη βέλτιστη πολιτική, συμπεριλαμβανομένων π.χ. αποφάσεων ή ενεργειών, δεδομένης της κατάστασής τους όταν οι χώροι κατάστασης και δράσης είναι περιορισμένοι. Προς αυτή την κατεύθυνση στην παρούσα διπλωματική εργασία περιγράφονται οι διάφορες τεχνικές που υπάρχουν στη βιβλιογραφία και πραγματεύονται τη βαθιά ενίσχυση της μάθησης στις επικοινωνίες και τη δικτύωση. Η καινοτομία της συγκεκριμένης διπλωματικής εργασίας έγκειται στη μελέτη της δυναμικής πρόσβασης στο δίκτυο, στον έλεγχο του ρυθμού δεδομένων, στην ασύρματη προσωρινή αποθήκευση, στην εκφόρτωση δεδομένων, στην ασφάλεια δικτύου και στη συντήρηση συνδεσιμότητας, τα οποία είναι όλα σημαντικά για δίκτυα επόμενης γενιάς, όπως 5G και πέραν τούτου, με χρήση εφαρμογών βαθιάς ενίσχυσης της μάθησης.

Λέξεις Κλειδιά: *Βαθιά Μάθηση, Επικοινωνίες, Δικτύωση, 5G*

Abstract

Modern networks, e.g. Internet of Things (IoT) and Unmanned Aerial Vehicles (UAV) networks are becoming more decentralized and autonomous. In such networks, network operators must make decisions locally to maximize network performance under environmental uncertainty. Reinforcement learning (RL) has been used effectively to enable network entities to obtain the best policy, including e.g. decisions or actions, given their status when space for action and situation is limited. Towards this direction, the present dissertation describes the various techniques available in the literature and deals with the deep enhancement of learning in communications and networking. The innovation of this dissertation lies in the study of dynamic network access, data rate control, wireless caching, data downloading, network security and connectivity maintenance, all of which are important for next-generation networks such as 5G and beyond, using deep learning enhancement applications.

Keywords: *Deep Learning*, Next-Generation Networks, 5G

1

Εισαγωγή

1.1 Δίκτυα Επικοινωνιών Νέας Γενιάς

Με την ανάπτυξη των συστημάτων 5G σε πλήρη εξέλιξη, η έρευνα έχει πια επικεντρωθεί στα κινητά συστήματα 6G (Patwary etal., 2020; Letaief etal., 2019). Συνεχίζοντας την παράδοση μιας νέας γενιάς κυψελοειδούς συστήματος μία φορά κάθε δέκα χρόνια περίπου, υπάρχει η προσδοκία ότι ένα σύστημα 6G θα τυποποιηθεί με εφαρμογές από το 2030. Δεδομένου ότι συχνά απαιτούνται περισσότερα από δέκα χρόνια ώστε μια νέα τεχνολογία να δει το φως της ημέρας, ήρθε η ώρα να ξεκινήσει η έρευνα για νέα τεχνολογικά στοιχεία για το 6G.

Είναι ουσιαστικό να δημιουργηθεί ένα όραμα για μελλοντικές επικοινωνίες με σκοπό την παροχή καθοδήγησης όσον αφορά την έρευνα. Γίνεται προσπάθεια να δημιουργηθεί μια ευρεία εικόνα των αναγκών και των τεχνολογιών επικοινωνίας στο χρονικό πλαίσιο του 6G. Είναι πιθανό ότι ορισμένες από αυτές τις απαιτήσεις ενδέχεται να μπορούν ήδη να ικανοποιούνται με την ενσωμάτωση νέων τεχνολογιών στο πλαίσιο του 5G. Σε γενικές γραμμές, αναμένεται να διερευνηθεί ως μέρος της εξέλιξης 5G η εισαγωγή τυχόν τροποποιήσεων που μπορούν να εισαχθούν με τρόπο συμβατό προς τα πίσω με λογικό κόστος εντός του πλαισίου 5G για την κάλυψη νέων απαιτήσεων απόδοσης. Από την άλλη πλευρά, οι τροποποιήσεις, οι οποίες αποτελούν θεμελιώδη αλλαγή και είναι ασυμβίβαστες με το υπάρχον πλαίσιο 5G ή μπορούν να

ενσωματωθούν μόνο με υψηλό κόστος στο δίκτυο ή τις συσκευές, θα αποτελέσουν μέρος της επόμενης γενιάς.

Εκτός από την ενισχυμένη ευρυζωνική σύνδεση για κινητές συσκευές για τους καταναλωτές, το 5G αναμένεται ευρέως ότι θα επιτρέψει την Τέταρτη Βιομηχανική Επανάσταση, ή Βιομηχανία 4.0 (Industry 4.0), μέσω της ψηφιοποίησης και της συνδεσιμότητας όλων των μεγάλων και μικρών πραγμάτων. Τα ψηφιακά δίδυμα διαφόρων αντικειμένων που δημιουργούνται σε άκρα νεφών θα αποτελέσουν το βασικό θεμέλιο του μελλοντικού ψηφιακού κόσμου. Οι ψηφιακοί δίδυμοι κόσμοι τόσο των φυσικών όσο και των βιολογικών οντοτήτων θα είναι μια βασική πλατφόρμα για τις νέες ψηφιακές υπηρεσίες του μέλλοντος. Η υλοποίηση ενός ολοκληρωμένου ψηφιακού κόσμου, που αποτελεί μια πλήρη και αληθινή αναπαράσταση του φυσικού κόσμου σε κάθε χωρική και χρονική στιγμή, θα απαιτήσει τεράστια ποσότητα χωρητικότητας σε χαμηλή καθυστέρηση. Η ψηφιοποίηση θα ανοίξει επίσης το δρόμο για τη δημιουργία νέων εικονικών κόσμων με ψηφιακές αναπαραστάσεις φανταστικών αντικειμένων που μπορούν να αναμειχθούν με τον ψηφιακό δίδυμο κόσμο ώστε να δημιουργηθεί ένας υπερ-φυσικός κόσμος μικτής πραγματικότητας. Καθώς τα έξυπνα ρολόγια και οι μετρητές καρδιακών παλμών μετατρέπονται σε επιδερμικά, στοματικά, εμφυτεύματα σώματος, σκελετούς πανοπλιών και ανιχνευτές εγκεφαλικής δραστηριότητας, η βιολογία των ανθρώπων θα χαρτογραφηθεί με ακρίβεια κάθε στιγμή και θα ενσωματωθεί στον ψηφιακό και εικονικό κόσμο, επιτρέποντας νέες υπεράνθρωπες δυνατότητες. Οι διεπαφές χρήστη επαυξημένης πραγματικότητας θα επιτρέψουν τον αποτελεσματικό και διαισθητικό ανθρώπινο έλεγχο όλων αυτών των κόσμων, είτε είναι φυσικός, είτε εικονικός, είτε βιολογικός.

Συνεπώς, η συνδεσιμότητα του μέλλοντος αφορά τη δυνατότητα απρόσκοπτης ενσωμάτωσης αυτών των διαφορετικών κόσμων, με σκοπό τη δημιουργία μιας ενοποιημένης εμπειρίας για τους ανθρώπους, αλλιώς θα πρέπει να θεωρείται ότι δημιουργείται ένα διαδίκτυο από cyborgs. Κατά την εξέταση ενός τέτοιου μέλλοντος, εμφανίζονται νέα σημαντικά θέματα, επιπλέον των νέων αναγκών επικοινωνίας όπως: (i) τελικές συσκευές που εκτείνονται από ενιαίες οντότητες σε μια συλλογή πολλαπλών τοπικών οντοτήτων που δρουν από κοινού για τη δημιουργία της νέας διεπαφής ανθρώπου-μηχανής. (ii) πανταχού παρόν καθολικός υπολογισμός που κατανέμεται μεταξύ των πολλαπλών τοπικών συσκευών και του νέφους, (iii) συστήματα γνώσης που αποθηκεύουν, επεξεργάζονται και μετατρέπουν δεδομένα σε ενεργή γνώση και (iv) ανίχνευση και ενεργοποίηση ακριβείας για τον έλεγχο του φυσικού κόσμου. Αρκετές δημοσιεύσεις έχουν ήδη υποστηρίξει τις απόψεις τους για το 6G (Strinati etal., 2019; Letaief etal.,

2019). Εμφανίζεται πια μια μοναδική και ευρύτερη προοπτική εστιάζοντας όχι μόνο στις τεχνολογίες, αλλά και στον ανθρώπινο μετασχηματισμό που αναμένεται να επιτευχθεί στην εποχή του 6G, ο οποίος βοηθά στο να παρέχεται μια εικόνα των απαιτήσεων απόδοσης και των αρχών σχεδιασμού για το 6G. Η άποψη για τους τεχνολογικούς μετασχηματισμούς ξεκινά από εκεί που βρίσκονται τα σημερινά συστήματα 5G βλέποντας τον τρόπο με τον οποίο αυτά εξελίσσονται και, εν συνεχεία, σε αυτό που μπορεί να διαφέρει πια θεμελιωδώς σε σχέση με το σημερινό. Επίσης αναφέρονται οι μετασχηματισμοί οι οποίοι είναι πιθανόν να συμβούν με τη φύση της τυποποίησης που απαιτείται σε έναν κόσμο με ανοιχτές πλατφόρμες.

1.2 Χρησιμότητα Δικτύων Νέας Γενιάς

Τα σύγχρονα δίκτυα, π.χ. τα δίκτυα του Διαδικτύου των Πραγμάτων (Internet of Things - IoT) και των Μη επανδρωμένων Εναέριων Οχημάτων (Unmanned Aerial Vehicles - UAV), καθίστανται πιο αποκεντρωμένα και αυτόνομα. Σε τέτοια δίκτυα, οι φορείς δικτύου πρέπει να λαμβάνουν αποφάσεις τοπικά για να μεγιστοποιούν την απόδοση του δικτύου υπό την αβεβαιότητα του περιβάλλοντος. Η μηχανική μάθηση (Machine Learning - ML) καθιστά δυνατή τη δημιουργία μοντέλων που μπορούν να αναλύουν ταχύτερα και με μεγαλύτερη ακρίβεια μεγαλύτερα και πολυπλοκότερα δεδομένα.

Ο αυξανόμενος όγκος και η πολυπλοκότητα όσον αφορά την προέλευση των δεδομένων, η οικονομική αποθήκευσή τους, αλλά και η ισχυρότερη και χαμηλότερου κόστους υπολογιστική επεξεργασία είναι παράγοντες που συνηγορούν στο να είναι η μηχανική μάθηση πιο δημοφιλής από οποιαδήποτε άλλη στιγμή.

Η ενισχυτική μάθηση (Reinforcement Learning - RL) έχει χρησιμοποιηθεί αποτελεσματικά για να δώσει τη δυνατότητα στις οντότητες του δικτύου να αποκτήσουν τη βέλτιστη πολιτική, συμπεριλαμβανομένων π.χ. αποφάσεων ή ενεργειών, δεδομένης της κατάστασής τους, όταν οι χώροι κατάστασης και δράσης είναι περιορισμένοι.

Έχοντας αυτά ως δεδομένα η παρούσα διπλωματική εργασία παρουσιάζει και αναλύει τις διάφορες τεχνικές που αναγράφονται στη βιβλιογραφία και πραγματεύονται τη βαθιά ενίσχυση της μάθησης, τόσο στον τομέα των επικοινωνιών, όσο και της δικτύωσης. Η εν λόγω διπλωματική εργασία αποτελεί καινοτομία όσον αφορά τη μελέτη της δυναμικής πρόσβασης στο δίκτυο, στον έλεγχο του ρυθμού δεδομένων, στην ασύρματη προσωρινή αποθήκευση, στην εκφόρτωση δεδομένων, στην ασφάλεια δικτύου και στη συντήρηση συνδεσιμότητας, στοιχείων με μεγάλη

σημασία για τα δίκτυα επόμενης γενιάς, όπως το 5G και το 6G, στηριζόμενη στη χρήση εφαρμογών βαθιάς μηχανικής και ενισχυτικής μάθησης.

Πώς θα είναι η ζωή και η ψηφιακή κοινωνία κατά τη δεκαετία του 2030; Αρχικά αναφέρονται οι συσκευές με τις οποίες θα μπορούν οι άνθρωποι να συνδέονται στο δίκτυο. Ενώ το smartphone και το tablet θα εξακολουθούν να υπάρχουν, είναι πιθανό να αναπτυχθούν νέες διεπαφές ανθρώπου-μηχανής ικανές να διευκολύνουν σημαντικά στην κατανάλωση και τον έλεγχο των πληροφοριών.

Συγκεκριμένα αναμένονται τα ακόλουθα:

- Οι φορητές συσκευές, όπως τα ακουστικά και οι συσκευές που είναι ενσωματωμένες στον ανθρώπινο ρουχισμό, θα καταστούν συνήθεις και τα επιθέματα δέρματος, καθώς και τα βιο-εμφυτεύματα ίσως δεν θα είναι πια τόσο σπάνια. Ενδεχομένως ο άνθρωπος να εξαρτάται από νέους αισθητήρες εγκεφάλου για την ενεργοποίηση μηχανών. Θα υπάρχουν πολλαπλοί ρουχισμοί που θα φέρει και οι οποίοι θα λειτουργούν άψογα μεταξύ τους παρέχοντας φυσικές και διαισθητικές διεπαφές.
- Η πληκτρολόγηση στην οθόνη αφής πιθανότατα θα είναι ξεπερασμένη. Η χειρονομία και η ομιλία σε όποιες συσκευές χρησιμοποιούνται θα γίνει μάλλον ο κανόνας.
- Οι χρησιμοποιούμενες συσκευές θα έχουν πλήρη επίγνωση του πλαισίου και το δίκτυο θα γίνεται όλο και πιο εξελιγμένο στην πρόβλεψη των ανθρωπίνων αναγκών. Αυτή η επίγνωση του περιβάλλοντος σε συνδυασμό με νέες διεπαφές ανθρώπου-μηχανής θα κάνει την ανθρώπινη αλληλεπίδραση με τον φυσικό και ψηφιακό κόσμο πολύ πιο διαισθητική και αποτελεσματική.

Η υπολογιστική που θα απαιτείται για αυτές τις συσκευές πιθανότατα δεν θα βρίσκεται εξ ολοκλήρου στις ίδιες τις συσκευές, εξαιτίας παραμέτρων μορφής και ισχύος μπαταρίας. Αντίθετα, ίσως χρειαστεί να βασίζεται σε τοπικά διαθέσιμους υπολογιστικούς πόρους για την ολοκλήρωση εργασιών, πέρα από το νέφος. Τα δίκτυα συνεπώς θα παίξουν σημαντικό ρόλο στη διεπαφή του αύριο μεταξύ ανθρώπου και μηχανής.

1.3 Απαιτήσεις και Βασικά Μέτρα Απόδοσης Δικτύων Νέας Γενιάς

Η σειρά νέων περιπτώσεων χρήσης που αναμένεται έως το 2030 και μετά θα παίξει το ρόλο οδηγού για τις νέες απαιτήσεις του 6G. Οι βασικοί δείκτες απόδοσης 5G (Key Performance

Indicators - KPI) του ρυθμού δεδομένων της απόδοσης/χωρητικότητας, της καθυστέρησης, της αξιοπιστίας, της κλίμακας και της ευελιξίας θα συνεχίσουν να είναι σημαντικά μέτρα όσον αφορά την απόδοση του 6G. Διάφορα νέα χαρακτηριστικά θα γίνουν επίσης σημαντικά για το 6G. Ακολουθεί ομαδοποίηση των απαιτήσεων για 6G σε έξι κατηγορίες - τρεις κατηγορίες με KPI παρόμοιες με το 5G και τρεις νέες κατηγορίες:

- Ο εντοπισμός και η ανίχνευση χρησιμοποιώντας το δίκτυο επικοινωνίας θα είναι ένα σημαντικό χαρακτηριστικό του 6G. Η πιστότητα και η ακρίβεια προσδιορίζονται ως τα αντίστοιχα μέτρα απόδοσης για τον εντοπισμό και την ανίχνευση, αντίστοιχα. Αναμένεται ότι θα επιτευχθεί ακρίβεια σε επίπεδο εκατοστού. Η ακρίβεια ανίχνευσης αντικειμένων μπορεί να μετρηθεί με βάση πιθανότητες χαμένης ανίχνευσης (Missed Detection - MD) και ψευδούς συναγερμού (False Alarm - FA) και σφάλματα εκτίμησης παραμέτρων.
- Το δίκτυο θα κατασκευαστεί με κατανεμημένες τεχνικές AI (Artificial Intelligence) και ML ενσωματωμένες σε διάφορους κόμβους, ενώ το πόσο γρήγορα προσαρμόζονται στις νέες συνθήκες στο δίκτυο είναι ένα σημαντικό μέτρο. Ο αυτοματισμός δικτύου θα είναι ο κανόνας και, ως εκ τούτου, το πόσο κοντά είναι ένα δίκτυο στην ολοκλήρωση της αυτοματοποίησης με μηδενική χειροκίνητη παρέμβαση θα αποτελεί ένα επιπλέον κριτήριο.
- Τέλος, αναμένεται μια μεγάλη επανάσταση στην τελική συσκευή στο χρονικό πλαίσιο του 6G. Ως εκ τούτου, εισάγονται μερικά χαρακτηριστικά σε μια κατηγορία συσκευών που επισημαίνουν τις κύριες αναμενόμενες μεταβάσεις. Αρχικά, θεωρείται ότι η τελική συσκευή θα εξελιχθεί σε πολλά σενάρια για να είναι ένα δίκτυο συσκευών ή ένα υποδίκτυο. Ως παράδειγμα μπορεί να θεωρηθεί ένα δίκτυο περιοχής μηχανής ή ένα δίκτυο περιοχής ρομπότ που περιλαμβάνει σύνδεση πολλαπλών τμημάτων ενός μηχανήματος, όπως ένας ελεγκτής και οι μονάδες κίνησής του. Ένα άλλο χαρακτηριστικό της συσκευής στο χρονικό πλαίσιο του 6G θα είναι ότι οι διεπαφές θα γίνουν πολύ πιο διαισθητικές, με πρόσβαση μέσω χειρονομιών και όχι πληκτρολόγησης,. Επίσης μια άλλη δυνατότητα για μια συγκεκριμένη κατηγορία συσκευών θα είναι η εξαιρετικά χαμηλή κατανάλωση και δυνητικά η χωρίς μπαταρία λειτουργία τους, βασισμένη στο δίκτυο για την τροφοδοσία της συσκευής.

1.4 Βασικές Αρχές Σχεδιασμού Δικτύων Νέας Γενιάς

Σε κάθε γενιά έως το 5G οι τρεις θεμελιώδεις διαστάσεις του φάσματος, η φασματική απόδοση και η επαναχρησιμοποίηση του χώρου υπαγόρευαν τον τρόπο με τον οποίο είναι δυνατή η αύξηση της χωρητικότητας. Το ίδιο θα συνεχίσει να ισχύει και για το 6G. Η τεχνολογία RF (Radio Frequency) θα εξελιχθεί σε ισχύ και θα χρησιμοποιήσει οικονομικά το φάσμα σε ακόμη υψηλότερες ζώνες. Υπάρχει η ευκαιρία τουλάχιστον δεκαπλάσιας αύξησης της ποσότητας του φάσματος με μετάβαση σε ζώνες συχνοτήτων terahertz. Η φασματική απόδοση θα βελτιωθεί με τη χρήση μαζικών MIMO (Multi-Input Multi-Output) πολλαπλών χρηστών όχι μόνο σε εκατοστομετρικά κύματα (cmWave), αλλά και σε χιλιοστομετρικά (mmWave), καθώς πραγματοποιείται η μετάβαση από αναλογική σε υβριδική/ψηφιακή δέσμη σε αυτές τις χαμηλότερες ζώνες mmWave. Καθώς το κόστος της μαζικής MIMO μειώνεται, ακόμη μεγαλύτερες συστοιχίες μπορούν να αναπτυχθούν για να αυξήσουν περαιτέρω την φασματική απόδοση. Η πυκνότητα δικτύου αναμφίβολα θα συνεχίσει να αυξάνεται, όχι μόνο για λόγους χωρητικότητας, αλλά και για να παρέχεται αυξημένη κάλυψη σε ζώνες υψηλότερης συχνότητας, σε υψηλότερους ρυθμούς δεδομένων και με μεγαλύτερη αξιοπιστία, ενώ θα εφαρμοστεί μια πιο διαδεδομένη πρόσβαση στο φάσμα. Η κοινή χρήση μεταξύ χειριστών, ακόμη και αδειοδοτημένου φάσματος που τροφοδοτείται από ασύρματη καθορισμένη από λογισμικό επικοινωνία (Software Defined Radio - SDR) και AI/ML, θα επιτρέψει πολύ υψηλότερη επαναχρησιμοποίηση φάσματος. Η αποτελεσματική επαναχρησιμοποίηση φάσματος είναι ιδιαίτερα σημαντική στις χαμηλότερες ζώνες, καθώς αυτές διαθέτουν καλές ιδιότητες διάδοσης εκτός οπτικής (Non Line Of Sight - NLOS) και οι πόροι φάσματος σε αυτές τις ζώνες είναι λιγοστοί.

Το 6G θα διαφέρει θεμελιωδώς από τις προηγούμενες γενιές στο ότι τρεις νέες θεμελιώδεις διαστάσεις θα παίξουν σημαντικό ρόλο σε συνδυασμό με τις τρείς παραδοσιακές διαστάσεις. Αυτές οι διαστάσεις αντιπροσωπεύουν τους θεμελιώδεις πόρους των δεδομένων, υπολογιστικής ικανότητας και ενέργειας. Όπως είναι γνωστό, οι τεχνικές AI/ML βασίζονται σε δεδομένα και όποιος έχει πρόσβαση σε μεγάλους όγκους δεδομένων συγκεκριμένων τομέων θα είναι επιτυχής στην εφαρμογή αυτών των τεχνικών. Η εφαρμογή AI/ML στο σχεδιασμό συστημάτων 6G θα είναι θεμελιώδης και παρόμοια με διάφορους άλλους τομείς, ενώ τα δεδομένα δικτύου και αισθητήρων θα γίνουν θεμελιώδεις πόροι που θα αξιοποιηθούν για τη βελτίωση της απόδοσης του συστήματος. Παρόλο που η υπολογιστική ισχύς ήταν πάντα ένας σημαντικός πόρος για τα κυψελοειδή συστήματα, οι δύο κύριες τάσεις δείχνουν πια προς την κατεύθυνση του να καταστεί περιορισμένος πόρος, ενώ ιδιαίτερη σημασία αποκτά το πώς θα αξιοποιηθεί κάτι τέτοιο

στο 6G. Η πρώτη τάση είναι ο αναδυόμενος κορεσμός στον αριθμό των τρανζίστορ που μπορούν να συσκευαστούν σε μονάδα όγκου, γεγονός που περιορίζει την υπολογιστική ισχύ των συσκευών. Η δεύτερη τάση αφορά στο ότι θα υιοθετηθούν πολλαπλές τελικές συσκευές ώστε να αυξηθούν οι ικανότητες ανίχνευσης του ανθρώπου, όπως γυαλιά, ακουστικά και άλλος φερόμενος εξοπλισμός, ο οποίος λόγω των πολύ μικρών μορφολογικών παραγόντων θα παρουσιάζει περιορισμένη υπολογιστική ικανότητα. Η τρέχουσα προσέγγιση της εκφόρτωσης υπολογιστών στο νέφος είναι απίθανο να είναι επαρκής όσον αφορά την κάλυψη των σύγχρονων αναγκών υπολογισμού στις διάφορες συσκευές. Η αξιοποίηση των υπολογιστών που είναι διαθέσιμοι στην τοπική περιοχή, αλλά διαφορετικοί από τις συσκευές, θα είναι ένα νέο ζήτημα στο χρονικό πλαίσιο 6G. Με αυτή την έννοια η υπολογιστική αντιμετωπίζεται ως μια άλλη ουσιαστική διάσταση που οδηγεί στο σχεδιασμό του νέου συστήματος επικοινωνίας. Τέλος, η διαθέσιμη ενέργεια σε κάθε στοιχείο του δικτύου θα καθορίσει τις επιτεύξιμες επιδόσεις, κάτι που κυμαίνεται από σχεδόν μηδενική ενέργεια σε ορισμένους τύπους συσκευών έως όρια τροφοδοσίας σε ραδιοφωνικούς σταθμούς βάσης και περιορισμούς ισχύος σε κέντρα δεδομένων. Επιπλέον, οι λύσεις για την κλιματική αλλαγή θα αποτελέσουν ένα σημαντικό επίκεντρο παντού στον κόσμο μέχρι τη δεκαετία του 2030 και η αυξανόμενη κατανάλωση ενέργειας δικτύων και συσκευών θα εξεταστεί σε μεγάλο βαθμό. Έτσι η ενέργεια γίνεται μια άλλη σημαντική διάσταση για το σχεδιασμό του 6G.

1.5 Βασικές Τεχνολογίες Δικτύων Νέας Γενιάς

Μια νέα γενιά χαρακτηρίζεται τελικά από τον αριθμό των νέων βασικών τεχνολογιών που διαμορφώνουν το σύστημα επικοινωνίας. Οι πραγματικά θεμελιώδεις νέες τεχνολογίες χρειάζονται συνήθως μια δεκαετία ή περισσότερο προκειμένου να υλοποιηθούν στην πράξη. Λαμβάνοντας υπόψη αυτό, οι πραγματικά καινοτόμες τεχνολογίες που απαρτίζουν το 6G πρέπει να είναι ερευνητικές έννοιες. Συνεχίζοντας το θέμα του «έξι» για το 6G, εντοπίστηκαν έξι νέες πιθανές τεχνολογικές μετατροπές που αναμένεται ότι θα αποτελέσουν μέρος της διαμόρφωσης του συστήματος 6G: (i) ο σχεδιασμός και η βελτιστοποίηση διεπαφών αέρα με οδηγό AI/ML, (ii) η επέκταση σε νέες ζώνες φάσματος και νέες μεθόδους γνωστικής κατανομής φάσματος, (iii) η ενσωμάτωση των δυνατοτήτων εντοπισμού και ανίχνευσης στον ορισμό του συστήματος, (iv) η επίτευξη ακραίων απαιτήσεων απόδοσης όσον αφορά την καθυστέρηση και την αξιοπιστία, (v) νέα παραδείγματα αρχιτεκτονικής δικτύου που περιλαμβάνουν υποδίκτυα και σύγκλιση RAN-Core (Radio Access Network-Core) και (vi) νέα συστήματα ασφάλειας και απορρήτου. Κάθε ένα από αυτά περιγράφεται παρακάτω:

Από τις έξι βασικές τεχνολογίες για το 6G παραλείπονται ορισμένες αναδυόμενες τεχνολογίες, οι οποίες δεν είναι ακόμη μέρος της τρέχουσας προδιαγραφής 5G, ωστόσο ακόμη θεωρούνται τμήμα της 5G εποχής. Τέτοια παραδείγματα αποτελούν η πλήρης διπλή επικοινωνία, η επικοινωνία χωρίς κελιά (Wang etal., 2019), τα μη επανδρωμένα εναέρια οχήματα (UAV) (Fotouhi etal., 2019), τα συστήματα πλατφόρμας υψηλού υψομέτρου (High Altitude Platform System - HAPS) (Mohammed etal., 2011) και η δορυφορική επικοινωνία με χαμηλή τροχιά γύρω από τη Γη (Low Earth Orbit - LEO) (Di etal., 2018). Στο άλλο άκρο του χρονικού άξονα για την εποχή 6G, έχουν αποκλειστεί πιο μακροπρόθεσμες αναδυόμενες τεχνολογίες, όπως η κβαντική επικοινωνία, η μοριακή επικοινωνία και η επικοινωνία ορατού φωτός (Gupta&Andrews, 2019), ώστε να αποτελέσουν βασικές τεχνολογίες του 6G. Αυτές οι τεχνολογίες τελικώς είτε θεωρούνται ειδικές λύσεις σε σύντομο χρονικό διάστημα είτε πιο γενικευμένες στο χρονικό πλαίσιο πέραν των δικτύων 6G.

1.6 Ο Ρόλος της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης

Οι τεχνικές AI/ML, ειδικά η βαθιά μάθηση, έχουν προχωρήσει ραγδαία την τελευταία δεκαετία και είναι πλέον εκ των ων ουκ άνευ σε πολλούς τομείς που περιλαμβάνουν ταξινόμηση εικόνας και όραση υπολογιστή από κοινωνικά δίκτυα έως ασφάλεια. Εφαρμόζονται σε προβληματικές περιοχές όπου σημαντικές ποσότητες δεδομένων είναι άμεσα διαθέσιμες για εκπαίδευση. Η ενισχυτική μάθηση αρχίζει να εφαρμόζεται σε μια ποικιλία εφαρμογών ρομποτικού ελέγχου μετά από διάφορες επιδείξεις της ικανότητάς της σε περιβάλλοντα παιχνιδιών, όπως το AlphaGo.

Πρόσφατα έχει διερευνηθεί πολύ η εφαρμογή τεχνικών βαθιάς εκμάθησης σε ασύρματα συστήματα. Τα επόμενα χρόνια αναμένεται ότι οι τεχνικές AI/ML θα εφαρμοστούν σε συστήματα 5G με τουλάχιστον τρεις διαφορετικούς τρόπους. Πρώτον, έχουν τη δυνατότητα να αντικαταστήσουν μερικούς από τους αλγορίθμους Layer 1 και Layer 2 που βασίζονται σε μοντέλα, όπως εκτίμηση καναλιού, ανίχνευση προοιμίου, εξίσωση και προγραμματισμό χρηστών, είτε επειδή έχουν καλύτερη απόδοση είτε επειδή είναι λιγότερο πολύπλοκοι. Δεύτερον, είναι πιθανό να εφαρμοστούν εκτενώς στη βελτιστοποίηση ανάπτυξης, όπως για παράδειγμα για τη διαμόρφωση ενός βέλτιστου υποσυνόλου από δέσμες με τις οποίες θα καλύπτεται η περιοχή κάλυψης, λαμβάνοντας υπόψη τα μοτίβα κίνησης κυψελών. Δεδομένης της πολυπλοκότητας των συστημάτων 5G όσον αφορά τον τεράστιο αριθμό παραμέτρων που πρέπει να διαμορφωθούν κατά τη στιγμή της ανάπτυξης, οι τεχνικές AI/ML θα διαδραματίσουν σημαντικό ρόλο στο όραμα μηδενικής βελτιστοποίησης ανθρώπινου δικτύου αφής. Τέλος, αναμένονται κάποιες άλλες

περιπτώσεις χρήσης, όπως ο εντοπισμός τελικών συσκευών με τεχνολογία 5G, ώστε να αξιοποιηθούν οι τεχνικές εκμάθησης για μεγαλύτερη ακρίβεια. Στην εργασία των (Wang etal., 2020), παρουσιάστηκε μια ολοκληρωμένη ανασκόπηση των πιθανών εφαρμογών AI/ML σε συστήματα 5G και μετά. Ομοίως, οι (Nawaz etal., 2019) στην εργασία τους παρείχαν μια μεγάλη περίληψη εφαρμογών που ταξινομούνται ανάλογα με τον τύπο της τεχνικής μάθησης, εκτός από μια επισκόπηση του κβαντικού υπολογισμού και των επικοινωνιών.

Εκτός από τη χρήση AI/ML στο RAN, το AI/ML θα καταστεί απαραίτητο για την αυτοματοποίηση του δικτύου 5G από άκρο σε άκρο που αντιμετωπίζει την πολυπλοκότητα της ενορχήστρωσης σε πολλούς τομείς και επίπεδα δικτύου, κάτι που θα επιτρέψει τη δυναμική προσαρμογή των πόρων του δικτύου και του νέφους, σύμφωνα με τις μεταβαλλόμενες απαιτήσεις, την ταχεία ανάπτυξη νέων υπηρεσιών και το γρήγορο μετριασμό των βλαβών, ενώ θα μειώσει σημαντικά τις λειτουργικές δαπάνες.

Τα συστήματα 6G ιδεατά θα χρησιμοποιούν AI/ML με πιο θεμελιώδη τρόπο από την αντίστοιχη 5G προσέγγιση. Αναμένεται να γίνει η μετάβαση από την τεχνητή νοημοσύνη ως ενίσχυση στην τεχνητή νοημοσύνη ως θεμέλιο για το σχεδιασμό και τη βελτιστοποίηση της διεπαφής αέρα - αυτο-βελτιστοποίηση πομπών και δεκτών, τη χρήση γνωστικού φάσματος και την επίγνωση περιβάλλοντος.

1.6.1 Πομποί και δέκτες Αυτοβελτιστοποίησης

Η συνεχιζόμενη έρευνα έχει δείξει ότι τα συστήματα βαθιάς εκμάθησης μπορούν να μάθουν να επικοινωνούν πάνω από στατικούς συνδέσμους πιο αποτελεσματικά από τους σχεδιασμούς συστημάτων που βασίζονται σε μοντέλα (O'Shea etal., 2018). Δεν απαιτείται ρητός σχεδιασμός κυματομορφής, σχηματισμού ή σημάτων αναφοράς. Μέσα από εκτεταμένη εκπαίδευση, ένα ενιαίο δίκτυο βαθιάς εκμάθησης στον πομπό και ένα στον δέκτη μαθαίνουν να επιλέγουν τον καλύτερο σχεδιασμό για αυτές τις παραμέτρους. Ενώ μια τέτοια προσέγγιση μάθησης από άκρο σε άκρο μπορεί να είναι ανέφικτη για πολύπλοκα, δυναμικά μεταβαλλόμενα περιβάλλοντα πολλαπλών χρηστών, το πλαίσιο επικοινωνίας 6G θα σχεδιαστεί με τέτοιο τρόπο ώστε να επιτρέπει την εκμάθηση στον τομέα να κάνει κάποιες επιλογές σχεδιασμού. Αυτό θα επιτρέψει τη βελτιστοποίηση των χαρακτηριστικών της διασύνδεσης αέρα, με βάση την επιλογή του φάσματος, του περιβάλλοντος, του υλικού που αναπτύσσεται και των στοχευόμενων απαιτήσεων. Μια σημαντική αλλαγή θα είναι η συμπερίληψη των δυνατοτήτων του υλικού στη βελτιστοποίηση του πλαισίου επικοινωνίας. Στην τρέχουσα προσέγγιση, η διασύνδεση αέρα έχει σχεδιαστεί

λαμβάνοντας υπόψη ορισμένα πρακτικά όρια στην εφαρμογή. Όμως μετά τη φάση σχεδιασμού, αναμένεται ότι όλες οι υλοποιήσεις θα διαθέτουν το απαιτούμενο υλικό για τον επιλεγμένο σχεδιασμό διεπαφής αέρα. Στο μέλλον, αναμένεται ότι η διασύνδεση αέρα θα μπορεί να προσαρμοστεί στις δυνατότητες του υλικού. Για παράδειγμα, μια συγκεκριμένη υλοποίηση μπορεί να έχει περιορισμένο αριθμό αναλύσεων σε ψηφιακή μετατροπή (Analog to Digital Converter - ADC) ή ψηφιακή σε αναλογική μετατροπή (Digital to Analog Converter - DAC), οι οποίες μπορούν να ληφθούν υπόψη από τα συστήματα εκμάθησης για τον προσδιορισμό της βέλτιστης επιλογής σηματοδότησης.

1.6.2 Χρήση Γνωστικού Φάσματος

Το φάσμα χαμηλών συχνοτήτων θα εξακολουθήσει να είναι υψίστης σημασίας για την κάλυψη ευρείας περιοχής, λόγω των ανώτερων ιδιοτήτων διάδοσης στο NLOS σε σύγκριση με τις ζώνες υψηλότερης συχνότητας. Κατά την επόμενη δεκαετία σημαντικές ποσότητες νέου φάσματος θα διατεθούν στο 5G και τις εξελίξεις του και αυτό είναι πιθανό να οδηγήσει σε σχεδόν εξάντληση του φάσματος σε ζώνες κάτω των 6 GHz. Έτσι, στο χρονικό πλαίσιο του 6G θα απαιτηθούν νέες μέθοδοι χρήσης φάσματος, ακόμη και εντός του αδειοδοτημένου καθεστώτος φάσματος, για να επιτραπεί καλύτερη τοπική πρόσβαση στο φάσμα και συνύπαρξη με άλλους χρήστες. Οι φορείς εκμετάλλευσης μπορεί να χρειαστεί να μοιραστούν φάσμα μεταξύ τους, αλλά και με άλλα ιδιωτικά αποκλειστικά δίκτυα. Ακόμη και μέσα σε έναν μόνο φορέα πολλές γενιές τεχνολογιών θα συνυπάρχουν και θα μοιράζονται φάσμα. Με την πρόοδο στην τεχνολογία της ασύρματης επικοινωνίας που επιτρέπει τη λειτουργία πολλαπλών ζωνών και τις τεχνικές εκμάθησης, όπως η βαθιά ενισχυτική μάθηση, η αποτελεσματική αυτόνομη κοινή χρήση φάσματος μπορεί να ανακουφίσει τα κύρια εμπόδια που προκύπτουν από την κοινή χρήση του (Tilghman,2019). Με την αυξανόμενη χρήση προηγμένων τεχνικών διαμόρφωσης δέσμης και πυκνότητας η χρήση του φάσματος γίνεται ιδιαίτερα τοπική, διευκολύνοντας περισσότερη επαναχρησιμοποίηση φάσματος και ως εκ τούτου επιτρέποντας διάφορες μορφές συνύπαρξης μεταξύ των γνωστικών συστημάτων κοινής χρήσης που θα είναι ιδιαίτερα επωφελείς.

1.6.3 Ευαισθητοποίηση Περιβάλλοντος

Μια άλλη σημαντική εξέλιξη που αναμένεται στο χρονικό πλαίσιο του 6G είναι η απρόσκοπτη ενσωμάτωση της ευαισθητοποίησης για το περιβάλλον, τα πρότυπα κυκλοφορίας, τα πρότυπα κινητικότητας και θέσης στη βελτιστοποίηση των συστημάτων επικοινωνίας που υποβοηθούνται από τις νέες τεχνικές AI/ML (Korpi etal., 2020). Για παράδειγμα, σε περιβάλλοντα όπως

εργοστασιακά δάπεδα, οι βιντεοκάμερες θα μπορούν να καταγράφουν την παρουσία και την κίνηση διαφόρων μηχανών και συσκευών που μπορούν να υποβληθούν σε επεξεργασία σε πραγματικό χρόνο μέσω δικτύων βαθιάς εκμάθησης για να προβλέψουν αλλαγές στο περιβάλλον διάδοσης, το οποίο με τη σειρά του θα μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση της επικοινωνίας. Ουσιαστικά, οι νέες τεχνικές απόκτησης και επεξεργασίας δεδομένων που ενσωματώνονται στο σύστημα επικοινωνίας μπορούν να μειώσουν την τυχαιότητα στους συνδέσμους επικοινωνίας. Τα μακροπρόθεσμα μοτίβα κινητικότητας μπορούν να προκύψουν σε εσωτερικούς και εξωτερικούς χώρους οι οποίοι μπορούν εν συνεχεία να χρησιμοποιηθούν για τη βελτιστοποίηση της εμπειρίας εξυπηρέτησης, δημιουργώντας συνδεσιμότητα με την κατάλληλη τεχνολογία την κατάλληλη στιγμή. Ένα άλλο σημαντικό στοιχείο των μελλοντικών συστημάτων είναι η χρήση ψηφιακά ελεγχόμενων παθητικών στοιχείων, όπως οι μετα-επιφάνειες μεγάλης κλίμακας (DiRenzo etal., 2019) που είναι πιθανό να κατανεμηθούν ευκαιριακά, ειδικά σε εσωτερικούς χώρους, ωστόσο απαιτούνται νέες μέθοδοι για να αξιοποιηθεί κάτι τέτοιο στη βελτίωση των επικοινωνιών. Ο προσδιορισμός του βέλτιστου ελέγχου αυτών των στοιχείων χρησιμοποιώντας μεθόδους βελτιστοποίησης βάσει μοντέλου μπορεί να είναι δύσκολος με συνέπεια να καθίσταται δύσκολη η διατύπωση της ακριβούς διάδοσης του σήματος που περιλαμβάνει τα συλλογικά αποτελέσματα, τα οποία με τη σειρά τους εξαρτώνται από τον τρόπο ελέγχου τους. Οι τεχνικές AI/ML πιθανότατα θα χρησιμοποιηθούν για την επίλυση τέτοιων πολύπλοκων προβλημάτων στην εποχή του 6G. Η σημασιολογική γνώση υψηλότερου επιπέδου για τον τρόπο με τον οποίο χρησιμοποιείται η επικοινωνία, όπως για παράδειγμα αν πρόκειται για έλεγχο ρομπότ ή επαυξημένη πραγματικότητα σε εργοστάσιο ή για παιχνίδια, μπορεί να αντληθεί από τα μοτίβα κίνησης και τα χαρακτηριστικά της συσκευής, ενώ μπορούν να παρέχονται αυτόματα οι κατάλληλες υπηρεσίες. Η ακριβής εξατομίκευση των υπηρεσιών έως τα χαμηλότερα επίπεδα επικοινωνίας επιτυγχάνεται μέσω τεχνικών εκμάθησης. Μεταβαίνοντας από το AI για 5G στο AI καθαρά για 6G αναμένεται ότι θα χρησιμοποιηθούν διάφορες μορφές μάθησης για την υλοποίηση τέτοιου τύπου εφαρμογών. Η μεταφορά και η ομοσπονδιακή μάθηση θα παίξουν κρίσιμο ρόλο. Τα συστήματα θα πρέπει πρώτα να εκπαιδευτούν εκτός σύνδεσης σε περιβάλλοντα προσομοίωσης σε τέτοιο βαθμό, ώστε να μπορούν να δημιουργηθούν βασικές επικοινωνίες και στη συνέχεια να εκπαιδευτούν εντός του τομέα για τη βελτιστοποίηση της απόδοσης. Με τον τρόπο αυτό θα υπάρξει μεταφορά της μάθησης από την προσομοίωση στο περιβάλλον πεδίου. Οι συσκευές και η υποδομή δικτύου πρέπει να μάθουν από κοινού να ενσωματώνουν λειτουργίες από άκρο σε άκρο και εδώ ακριβώς είναι το σημείο που η ομοσπονδιακή μάθηση θα παίξει σημαντικό ρόλο. Αντί να μοιράζονται τα μεγάλα σύνολα δεδομένων μεταξύ διαφόρων συσκευών και δικτύου, θα μοιράζονται τα μοντέλα. Στα υψηλότερα στρώματα η ενισχυτική βαθιά μάθηση θα

είναι απαραίτητη για τη βελτιστοποίηση της κατανομής των πόρων και τον έλεγχο διαφόρων παραμέτρων. Η ιεραρχική και πολλαπλών παραγόντων ενίσχυση μάθησης θα πρέπει να χρησιμοποιηθεί σε διαφορετικούς κόμβους.

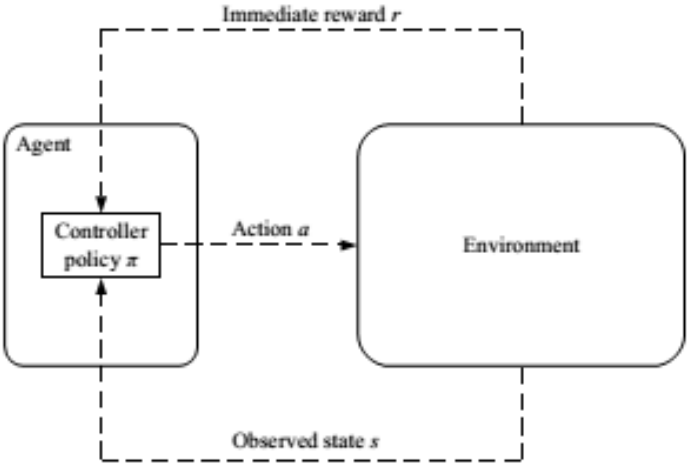
1.7 Δομή Εργασίας

Στο κεφάλαιο 1 παρουσιάζονται και αναλύονται οι βασικές απαιτήσεις των δικτύων νέας γενιάς καθώς και οι βασικές αρχές σχεδιασμού των δικτύων αυτών. Αναπτύσσονται οι κυριότερες τεχνολογίες των Δικτύων νέας γενιάς και αναπτύσσεται ιδιαιτέρως ο ρόλος της τεχνικής νοημοσύνης και της μηχανικής μάθησης. Το κεφάλαιο 2 ασχολείται με την ενισχυτική μάθηση ως αρκετά σημαντικός κλάδος της μηχανικής μάθησης και αναλύεται ο αλγόριθμος Q-learning. Στο κεφάλαιο 3 παρουσιάζεται και αναλύεται η βαθιά μάθηση η οποία στόχο έχει την αποφυγή της μη αυτόματης περιγραφής μιας δομής δεδομένων με αυτόματη εκμάθηση από τα δεδομένα. Παρουσιάζεται και πάλι ο αλγόριθμος Q-learning για την βαθιά μάθηση και γίνεται αναφορά στα μοντέλα βαθιάς Q-learning. Το κεφάλαιο 4 αναφέρεται στις εφαρμογές των DRL/DQL, τόσο στο κομμάτι των επικοινωνιών, όσο και της δικτύωσης και συγκεκριμένα σε εφαρμογές που σχετίζονται με επιθέσεις παρεμβολών και ασφάλεια στον κυβερνοχώρο, ενώ παρουσιάζει τις στρατηγικές για συστήματα επικοινωνιών που βασίζονται σε βαθιά μάθηση. Στο κεφάλαιο 5 παρουσιάζεται μία βιβλιογραφική ανασκόπηση εφαρμογών με μεθόδους ενισχυτικής βαθιάς μάθησης που αφορούν σε μη επανδρωμένα αεροσκάφη. Τέλος, στο 6ο κεφάλαιο θα γίνει παρουσίαση των συμπερασμάτων που θα προκύψουν από το σύνολο της διπλωματικής μας εργασίας.

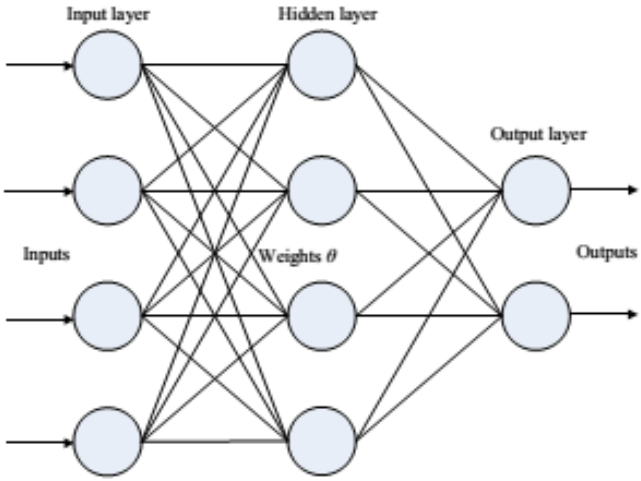
2

Ενισχυτική Μάθηση

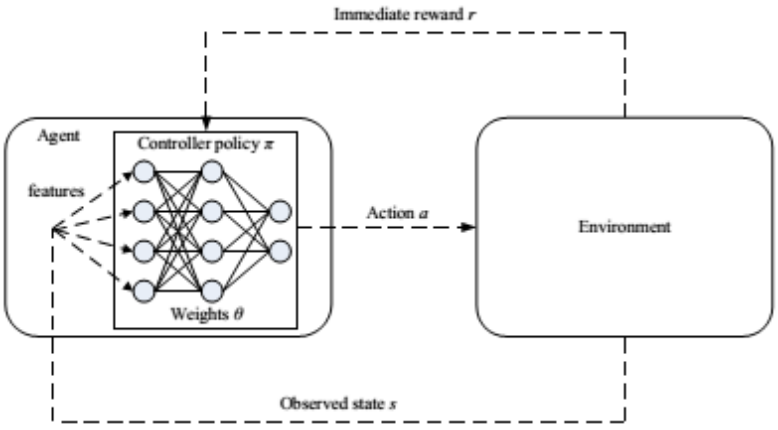
Η ενίσχυση της μάθησης, ένας σημαντικός κλάδος της μηχανικής μάθησης, είναι ένα αποτελεσματικό εργαλείο και χρησιμοποιείται ευρέως στη βιβλιογραφία για την αντιμετώπιση των MDPs (Markov Decision Process) (Sutton&Barto, 1998). Σε μια διαδικασία ενίσχυσης μάθησης, ένας παράγοντας μπορεί να μάθει τη βέλτιστη πολιτική του μέσω αλληλεπίδρασης με το περιβάλλον του. Συγκεκριμένα, ο παράγοντας παρατηρεί πρώτα την τρέχουσα κατάστασή του και έπειτα επιτελεί μια ενέργεια και λαμβάνει την άμεση ανταμοιβή του μαζί με τη νέα του κατάσταση, όπως φαίνεται στο παρακάτω σχήμα (Σχήμα 1).



Σχήμα 1. Μάθηση Ενίσχυσης



Σχήμα 2. Τεχνητό νευρωνικό δίκτυο



Σχήμα 3. Βαθιά μάθηση Q

Σχήμα: (1) Μάθηση ενίσχυσης, (2) Τεχνητό νευρωνικό δίκτυο και (3) Βαθιά μάθηση Q

Οι παρατηρούμενες πληροφορίες, δηλαδή η άμεση ανταμοιβή και η νέα κατάσταση, χρησιμοποιούνται για την προσαρμογή της πολιτικής του παράγοντα και αυτή η διαδικασία θα επαναληφθεί έως ότου η πολιτική του παράγοντα πλησιάσει τη βέλτιστη πολιτική. Στην ενίσχυση της μάθησης, η Q-learning είναι η πιο αποτελεσματική μέθοδος και χρησιμοποιείται ευρέως στη βιβλιογραφία. Στη συνέχεια, θα συζητηθεί ο αλγόριθμος Q-learning και οι επεκτάσεις του για προηγμένα μοντέλα MDP.

2.1 Αλγόριθμος Q-Learning

Σε ένα MDP στόχος είναι η εύρεση της βέλτιστης πολιτικής π^*: S → A ώστε ο παράγοντας να ελαχιστοποιεί το συνολικό κόστος για το σύστημα. Κατά συνέπεια, ορίζεται πρώτα η συνάρτηση τιμής $V^{\pi}: S \to R$ που αντιπροσωπεύει την αναμενόμενη τιμή που λαμβάνεται ακολουθώντας την πολιτική π από κάθε κατάσταση s ∈ S. Η συνάρτηση τιμής V για την πολιτική π ποσοτικοποιεί το καλώς έχειν της πολιτικής μέσω ενός άπειρου ορίζοντα και μείωση MDP που μπορεί να εκφραστεί ως εξής:

$$\begin{aligned}\mathcal{V}^{\pi}(s) &= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma r_t(s_t, a_t) \mid s_0 = s\right] \\ &= \mathbb{E}_{\pi}[r_t(s_t, a_t) + \gamma \mathcal{V}^{\pi}(s_{t+1}) \mid s_0 = s]\end{aligned}$$

Δεδομένου ότι στόχος είναι η εύρεση της βέλτιστης πολιτικής π^*, μια βέλτιστη ενέργεια σε κάθε κατάσταση μπορεί να βρεθεί μέσω της λειτουργίας βέλτιστης τιμής που εκφράζεται με $\mathcal{V}^*(s) = \max_{a_t}\{\mathbb{E}_{\pi}[r_t(s_t, a_t) + \gamma \mathcal{V}^{\pi}(s_{t+1})]\}$.

Αν $Q^*(s,a) \triangleq r_t(s_t, a_t) + \gamma \mathbb{E}_{\pi}[\mathcal{V}^{\pi}(s_{t+1})]$ η βέλτιστη συνάρτηση Q για όλα τα ζεύγη κατάστασης-δράσης, τότε η συνάρτηση βέλτιστης τιμής μπορεί να γραφτεί $\mathcal{V}^*(S) = \max_{a}\{Q^*(s,a)\}$. Τώρα το πρόβλημα μειώνεται στο να βρεθούν οι βέλτιστες τιμές της συνάρτησης δηλαδή Q δηλαδή $Q^*(s,a)$ για όλα τα ζεύγη κατάστασης-δράσης, και αυτό μπορεί να γίνει μέσω επαναληπτικών διαδικασιών. Συγκεκριμένα, η συνάρτηση Q ενημερώνεται σύμφωνα με τον ακόλουθο κανόνα:

$$\begin{aligned}Q_{t+1}(s,a) = Q_t(s,a) &+ \\ \alpha_t &\left[r_t(s,a) + \gamma \max_{a'} Q_t(s,a') - Q_t(s,a)\right]\end{aligned}$$

Η βασική ιδέα πίσω από αυτήν την ενημέρωση είναι η εύρεση της χρονικής διαφοράς (Temporal Difference - TD) μεταξύ της προβλεπόμενης τιμής Q, δηλαδή $r_t(s,a) + \gamma \max_{a'} Q_t(s,a')$ και της τρέχουσας τιμής της, δηλαδή $Q_t(s,a)$. Στην προηγούμενη σχέση ο ρυθμός εκμάθησης α_t χρησιμοποιείται για τον προσδιορισμό της επίδρασης των νέων πληροφοριών στην υπάρχουσα τιμή Q. Ο ρυθμός εκμάθησης μπορεί να επιλεγεί ως σταθερό ή μπορεί να προσαρμοστεί δυναμικά κατά τη διάρκεια της μαθησιακής διαδικασίας. Ωστόσο πρέπει να ικανοποιεί την Υπόθεση 1 για να εγγυηθεί τη σύγκλιση για τον αλγόριθμο εκμάθησης Q.

Αν και ο αλγόριθμος Q-learning μπορεί να βρει τη βέλτιστη πολιτική για τον παράγοντα χωρίς να χρειάζεται γνώση για το περιβάλλον, αυτός ο αλγόριθμος λειτουργεί με τρόπο χωρίς σύνδεση. Παρουσιάζεται ένας εναλλακτικός αλγόριθμος μάθησης στο διαδίκτυο, ο SARSA (State Action Reward State Action), ο οποίος επιτρέπει στον παράγοντα να προσεγγίσει τη βέλτιστη πολιτική μέσω διαδικτύου.

Διαφορετικός από τον αλγόριθμο Q-learning, ο SARSA είναι ένας διαδικτυακός αλγόριθμος που επιτρέπει στον πράκτορα να επιλέγει τις βέλτιστες ενέργειες σε κάθε βήμα σε πραγματικό χρόνο, χωρίς να περιμένει έως ότου ο αλγόριθμος συγκλίνει. Στον αλγόριθμο εκμάθησης Q, η πολιτική ενημερώνεται σύμφωνα με τη μέγιστη ανταμοιβή των διαθέσιμων ενεργειών ανεξάρτητα από την πολιτική που εφαρμόζεται, δηλαδή μια μέθοδο εκτός πολιτικής. Αντίθετα, ο αλγόριθμος SARSA αλληλεπιδρά με το περιβάλλον και ενημερώνει την πολιτική απευθείας από τις ενέργειες που πραγματοποιήθηκαν, δηλαδή, μια μέθοδος εντός-πολιτικής. Πρέπει να σημειωθεί ότι ο αλγόριθμος SARSA ενημερώνει τις τιμές Q από το πενταπλάσιο Q (s, α, r, s', α').

Για να εφαρμοστεί ο αλγόριθμος Q-learning στο περιβάλλον παιγνίου Markov, καθορίζεται πρώτα η συνάρτηση Q για τον παράγοντα i από το $Q_i\left(s, a^i, a^{-i}\right)$ όπου $a^{-i} \triangleq \{\alpha^1, \dots, \alpha^{i-1}, \alpha^{i+1} \dots, a^I\}$ δηλώνει το σύνολο ενεργειών όλων των παραγόντων εκτός από το i. Στη συνέχεια, η συνάρτηση Nash Q του παράγοντα i ορίζεται από:

$$Q_i^*\left(s, a^i, \mathbf{a}^{-i}\right) = r^i\left(s, a^i, \mathbf{a}^{-i}\right) + \beta \sum_{s' \in \mathcal{S}} p\left(s' \mid s, a^i, \mathbf{a}^{-i}\right) \mathcal{V}^i(s', \pi_1^*, \dots, \pi_I^*)$$

Όπου $(\pi_1^*, \dots \pi_I^*)$ είναι η κοινή στρατηγική ισορροπίας Nash, $r^i(s, a^i, \mathbf{a}^{-i})$ είναι η άμεση ανταμοιβή του παράγοντα i υπό την κοινή δράση $(a^i, \mathbf{a}^{-i})$ και $V^i(s', \pi_1^*, \dots, \pi_I^*)$ είναι η συνολική μειωμένη ανταμοιβή για έναν άπειρο χρονικό ορίζοντα ξεκινώντας από την κατάσταση s', δεδομένου ότι όλοι οι παράγοντες ακολουθούν τις στρατηγικές ισορροπίας.

Οι (Hu&Wellman, 2003) πρότειναν έναν αλγόριθμο Q-learning πολλαπλών πρακτόρων για παίγνια Markov γενικού αθροίσματος που επιτρέπει στους παράγοντες να πραγματοποιούν ενημερώσεις με βάση την υπόθεση συμπεριφοράς ισορροπίας Nash έναντι των τρεχουσών τιμών Q. Συγκεκριμένα, ο παράγοντας i θα μάθει τις τιμές Q σχηματίζοντας μια αυθαίρετη εικασία από την ώρα έναρξης του παιγνίου. Σε κάθε χρονικό βήμα t, ο παράγοντας i παρατηρεί την τρέχουσα κατάσταση και αναλαμβάνει μία δράση α^i . Στη συνέχεια, παρατηρεί την άμεση ανταμοιβή του r^i, τις ενέργειες που λαμβάνουν άλλοι a^{-i}, τις άμεσες ανταμοιβές των άλλων και τη νέα κατάσταση του συστήματος s'. Μετά από αυτό, ο παράγοντας υπολογίζει μια ισορροπία Nash (π_1 (s'),., π_I(s')) για την κατάσταση παιγνίου $(Q_1^t(s'), \dots, Q_I^t(s'))$ και ενημερώνει τις τιμές Q σύμφωνα με:

$$Q_i^{t+1}(s, a^i, \mathbf{a}^{-i}) = (1 - \alpha_t)Q_i^t(s, a^i, \mathbf{a}^{-i}) + \alpha_t\left[r_t^i + \gamma \mathcal{N}_t^i(s')\right]$$

όπου $\alpha_t \in (0, 1)$ είναι ο ρυθμός εκμάθησης και $\mathcal{N}_t^i(s') \triangleq Q_i^t(s')x\pi_1(s')x \dots x\pi_I(s')$

Για να υπολογιστεί η ισορροπία Nash, ο παράγοντας i πρέπει να γνωρίζει $(Q_1^t(s'), \dots, Q_I^t(s')$.

Ωστόσο, οι πληροφορίες σχετικά με τις τιμές Q άλλων παραγόντων δεν δίνονται, και ως εκ τούτου ο πράκτορας i πρέπει να μάθει αυτές τις πληροφορίες επίσης. Για να γίνει αυτό, ο πράκτορας i θα ορίσει εκτιμήσεις σχετικά με τις τιμές Q των άλλων στην αρχή του παιγνίου, π.χ. $Q_0^j(s, a^i, \mathbf{a}^{-i}) = 0, ¥\, j, s$. Καθώς προχωρά το παιχνίδι, ο παράγοντας i παρατηρεί τις άμεσες ανταμοιβές και τις προηγούμενες ενέργειες άλλων παραγόντων. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για την ενημέρωση των εικασιών του παράγοντα i στις λειτουργίες Q άλλων παραγόντων. Ο παράγοντας i ενημερώνει τις πεποιθήσεις του σχετικά με τη λειτουργία Q του παράγοντα j, σύμφωνα με τον ίδιο κανόνα ενημέρωσης στην προηγούμενη σχέση. Στη συνέχεια, αποδεικνύεται ότι κάτω από ορισμένες εξαιρετικά περιοριστικές παραδοχές σχετικά με τη μορφή των καταστάσεων των παιγνίων κατά τη διάρκεια της μάθησης, ο προτεινόμενος αλγόριθμος Q-learning πολλαπλών παραγόντων είναι εγγυημένος για την επίτευξη σύγκλισης.

3

Βαθιά Μάθηση

3.1 Η Έννοια της Βαθιάς Μάθησης

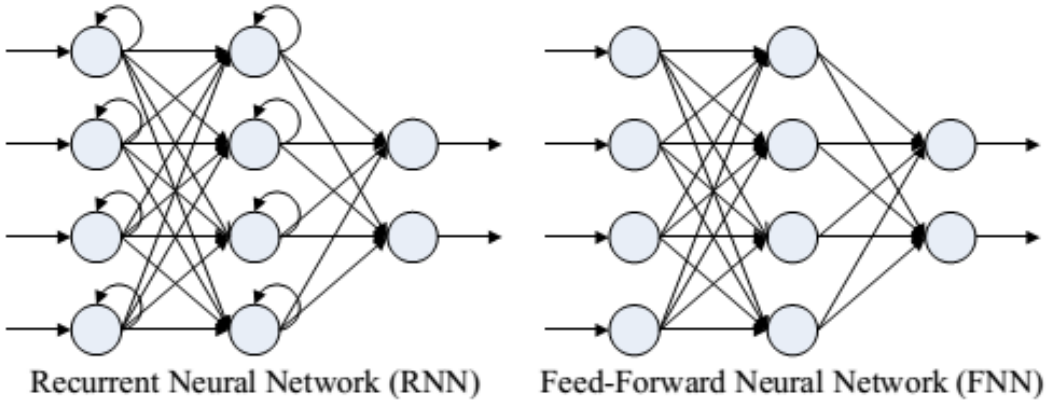
Η βαθιά μάθηση (Goodfellow etal., 2016) αποτελείται από ένα σύνολο αλγορίθμων και τεχνικών που προσπαθούν να βρουν σημαντικά χαρακτηριστικά των δεδομένων και να μοντελοποιήσουν τις υψηλού επιπέδου αφαιρέσεις τους. Ο κύριος στόχος της βαθιάς μάθησης είναι να αποφευχθεί η μη αυτόματη περιγραφή μιας δομής δεδομένων (όπως χειρόγραφες δυνατότητες) με αυτόματη εκμάθηση από τα δεδομένα. Το όνομα αναφέρεται στο γεγονός ότι συνήθως οποιοδήποτε νευρωνικό δίκτυο με δύο ή περισσότερα κρυμμένα επίπεδα ονομάζεται 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 στο επίπεδο σε μοντέλα βαθιάς δημιουργίας, όπως οι κόμβοι στα DeepBeliefNetworks και DeepBoltzmannMachines.

Το ANN είναι ένα υπολογιστικό μη γραμμικό μοντέλο βασισμένο 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: είσοδο, κρυφό και έξοδο, όπως απεικονίζεται στο προηγούμενο σχήμα (Σχήμα 2). Το στρώμα εισόδου περιέχει νευρώνες εισόδου που στέλνουν πληροφορίες στο κρυφό επίπεδο. Αυτό στέλνει δεδομένα στο επίπεδο εξόδου. Κάθε νευρώνας

έχει σταθμισμένες εισόδους (συνάψεις), συνάρτηση ενεργοποίησης (καθορίζει την έξοδο που δίνει μια είσοδος) και μία έξοδο. Οι συνάψεις είναι οι ρυθμιζόμενες παράμετροι που μετατρέπουν ένα νευρωνικό δίκτυο σε ένα παραμετροποιημένο σύστημα.

Κατά τη διάρκεια της φάσης εκμάθησης, τα ANN χρησιμοποιούν ανάστροφη-αναδιάδοση ως έναν αποτελεσματικό αλγόριθμο εκμάθησης για να υπολογίσουν γρήγορα μια κλίση σε σχέση με τα βάρη. Η ανάστροφη-αναδιάδοση είναι μια ειδική περίπτωση αυτόματης διαφοροποίησης. Στο πλαίσιο της μάθησης η ανάστροφη-αναδιάδοση 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 μερικές φορές προς τα πίσω διάδοση σφαλμάτων, επειδή το σφάλμα υπολογίζεται στην έξοδο και κατανέμεται πίσω μέσω των επιπέδων δικτύου.

Το DNN ορίζεται ως ANN με πολλαπλά κρυμμένα επίπεδα. Υπάρχουν δύο τυπικά μοντέλα DNN, το νευρωνικό δίκτυο πρόσθιας τροφοδότησης (Feedforward Neural Network - 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 και δεν υπάρχουν κύκλοι ή βρόχοι στο δίκτυο, όπως φαίνεται στο σχήμα που ακολουθεί (Σχήμα 4). Στα FNNs, το συμβατικό 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, ομαδοποιημένα ή πλήρως συνδεδεμένα και χρησιμοποιεί μια παραλλαγή πολυεπίπεδων αντιληπτών. Τα συνελικτικά στρώματα χρησιμοποιούν μια λειτουργία συνέλιξης στην είσοδο που περνά το αποτέλεσμα στο επόμενο επίπεδο. Αυτή η λειτουργία επιτρέπει στο δίκτυο να είναι βαθύτερο με πολύ λιγότερες παραμέτρους.



Σχήμα 4. RNN-CNN

Σε αντίθεση με τα FNN, το RNN είναι μια παραλλαγή ενός αναδρομικού τεχνητού νευρωνικού δικτύου στο οποίο οι συνδέσεις μεταξύ των νευρώνων σχηματίζουν κατευθυνόμενους κύκλους. Αυτό σημαίνει ότι μια έξοδος εξαρτάται όχι μόνο από τις άμεσες εισόδους της, αλλά και από την κατάσταση νευρώνων του προηγούμενου περαιτέρω βήματος. Τα RNN έχουν σχεδιαστεί για να χρησιμοποιούν διαδοχικά δεδομένα, όταν το τρέχον βήμα έχει κάποια σχέση με τα προηγούμενα βήματα. Αυτό καθιστά τα RNN ιδανικά για εφαρμογές με χρονική συνιστώσα, π.χ. δεδομένα χρονοσειρών και επεξεργασία φυσικής γλώσσας. Ωστόσο, όλα τα RNN έχουν βρόχους ανατροφοδότησης στο επαναλαμβανόμενο επίπεδο. Αυτό επιτρέπει στα RNN να διατηρούν πληροφορίες στη μνήμη με την πάροδο του χρόνου. Ωστόσο, μπορεί να είναι δύσκολο να εκπαιδευτεί τυπικά ένα RNN για την επίλυση προβλημάτων που απαιτούν μάθηση μακροπρόθεσμων χρονικών εξαρτήσεων. Ο λόγος είναι ότι η κλίση της συνάρτησης απώλειας μειώνεται εκθετικά με το χρόνο, το οποίο ονομάζεται πρόβλημα διαβάθμισης εξαφάνισης. Επομένως η μακροχρόνια βραχυπρόθεσμη μνήμη (Long Short-Term Memory - LSTM) χρησιμοποιείται συχνά σε RNN για την αντιμετώπιση αυτού του ζητήματος. Η LSTM έχει σχεδιαστεί για να μοντελοποιεί χρονικές ακολουθίες και οι εξαρτήσεις μεγάλου εύρους είναι ακριβέστερες από τις συμβατικές RNN. Η LSTM δεν χρησιμοποιεί μια συνάρτηση ενεργοποίησης στα επαναλαμβανόμενα στοιχεία του, οι αποθηκευμένες τιμές δεν τροποποιούνται και η κλίση δεν τείνει να εξαφανιστεί κατά τη διάρκεια της εκπαίδευσης. Συνήθως οι μονάδες LSTM υλοποιούνται σε «μπλοκ» με αρκετές μονάδες. Αυτά τα μπλοκ έχουν τρεις ή τέσσερις «πύλες», π.χ. πύλη εισόδου, πύλη ξεχασμού, πύλη εξόδου, που ελέγχουν τη ροή πληροφοριών με βάση τη λογιστική συνάρτηση.

3.2 Βαθιά Q- Μάθηση

Ο αλγόριθμος Q-learning μπορεί να αποκτήσει αποτελεσματικά μια βέλτιστη πολιτική, όταν ο χώρος κατάστασης και ο χώρος δράσης είναι μικροί. 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. Έτσι εισάγεται ο αλγόριθμος DQL (Deep Q-Learning) για την αντιμετώπιση αυτού του ελλείμματος. Διαισθητικά 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 σε προηγούμενο σχήμα περί βαθιάς μάθησης (Σχήμα 3).

Όπως αναφέρεται στην εργασία των (Mnih etal., 2015), η μέση ανταμοιβή που λαμβάνεται από αλγόριθμους ενισχυτικής μάθησης μπορεί να μην είναι σταθερή ή ακόμη και να αποκλίνει, όταν

χρησιμοποιείται ένας μη γραμμικός προσεγγιστής συνάρτησης. 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 $R + \gamma \max_{\alpha'} Q(s', a')$ ποικίλλουν. Για να αντιμετωπιστεί αυτό το ζήτημα, μπορούν να χρησιμοποιηθούν δύο μηχανισμοί: η επανάληψη εμπειρίας και ο στόχος Q-network.

3.2.1 Μηχανισμός επανάληψης εμπειρίας

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. μέσω της χρήσης ϵ-άπληστης πολιτικής. Στη συνέχεια, ο αλγόριθμος επιλέγει τυχαία δείγματα, δηλαδή, μίνι-παρτίδες, 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 ταυτόσημα κατανεμημένες και έτσι οι συσχετίσεις μεταξύ των παρατηρήσεων μπορούν να αφαιρεθούν.

3.2.2 Q δίκτυο σταθερού στόχου

Κατά τη διαδικασία εκπαίδευσης η τιμή Q θα αλλάξει. Έτσι οι εκτιμήσεις τιμών μπορεί να είναι εκτός ελέγχου, εάν χρησιμοποιείται ένα συνεχώς μεταβαλλόμενο σύνολο τιμών για την ενημέρωση του Q-δικτύου. Αυτό οδηγεί στην αποσταθεροποίηση του αλγορίθμου. Για την αντιμετώπιση αυτού του ζητήματος το Q-δίκτυο στόχος χρησιμοποιείται για την ενημέρωση συχνά, αλλά αργά, των τιμών των κύριων δικτύων Q. Με αυτόν τον τρόπο οι συσχετίσεις μεταξύ του στόχου και των εκτιμώμενων Q-τιμών μειώνονται σημαντικά σταθεροποιώντας έτσι τον αλγόριθμο.

Η DQL κληρονομεί και προωθεί τα πλεονεκτήματα, τόσο των τεχνικών ενίσχυσης όσο και της βαθιάς μάθησης, αποκτώντας ένα ευρύ φάσμα εφαρμογών στην πράξη, όπως η ανάπτυξη παιγνίων, μεταφοράς και ρομποτικής.

3.3 Προηγμένα Μοντέλα Βαθιάς Q-Μάθησης

3.3.1 Διπλή βαθιά μάθηση Q

Σε ορισμένα στοχαστικά περιβάλλοντα ο αλγόριθμος εκμάθησης Q αποδίδει άσχημα, λόγω των μεγάλων υπερβολικών εκτιμήσεων των τιμών δράσης (Thrun&Schwartz, 1993). Αυτές οι υπερεκτιμήσεις προκύπτουν από μια θετική προκατάληψη που εισάγεται επειδή η Q-μάθηση χρησιμοποιεί τη μέγιστη τιμή δράσης ως προσέγγιση για τη μέγιστη αναμενόμενη τιμή δράσης, όπως φαίνεται στην ακόλουθη εξίσωση:

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

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 επίσης για την εκτίμηση αυτής της τιμής δράσης.

Έτσι για να ξεπεραστεί το πρόβλημα υπερεκτίμησης του αλγορίθμου Q- learning εισήχθη μια λύση χρησιμοποιώντας δύο συναρτήσεις Q-τιμών, δηλαδή Q_1 και Q_2, για ταυτόχρονη επιλογή και αξιολόγηση των τιμών δράσης μέσω της συνάρτησης απώλειας ως ακολουθεί:

$$\left[r_j + \gamma Q_2\left(s_{j+1}, \arg\max_{a_{j+1}} Q_1(s_{j+1}, a_{j+1}; \boldsymbol{\theta}_1); \boldsymbol{\theta}_2\right) - Q_1(s_j, a_j; \boldsymbol{\theta}_1)\right]^2$$

Η επιλογή μιας ενέργειας στο argmax εξακολουθεί να οφείλεται στα διαδικτυακά βάρη θ_1. Αυτό σημαίνει ότι, όπως και στο Q-learning εξακολουθείται να εκτιμάται η αξία της άπληστης 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. Οπότε αναπτύχθηκε το μοντέλο Double Deep Q-Learning (DDQL) (VanHasselt etal., 2016) χρησιμοποιώντας ένα Double Deep Q-Network (DDQN) με τη συνάρτηση απώλειας να ενημερώνεται ως εξής:

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

Σε αντίθεση με τη διπλή Q-μάθηση, τα βάρη του δεύτερου δικτύου θ_2 αντικαθίστανται με τα βάρη των δικτύων στόχων θ΄ για την αξιολόγηση της τρέχουσας άπληστης πολιτικής, όπως φαίνεται

στην προηγούμενη εξίσωση. Η ενημέρωση στο δίκτυο προορισμού παραμένει αμετάβλητη από το DQN και παραμένει περιοδικό αντίγραφο του διαδικτυακού δικτύου. Λόγω της αποτελεσματικότητας της DDQL, υπάρχουν μερικές εφαρμογές της DDQL που παρουσιάστηκαν πρόσφατα για την αντιμετώπιση προβλημάτων πρόσβασης δυναμικού φάσματος σε πολυκάναλα ασύρματα δίκτυα και την κατανομή πόρων σε ετερογενή δίκτυα.

3.3.2 Βαθιά Q-Μάθηση με Προτεραιότητα Επανάληψης Εμπειρίας

Ο μηχανισμός επανάληψης εμπειρίας επιτρέπει στον παράγοντα εκμάθησης ενίσχυσης 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 σημασία τους. Ως εκ τούτου, οι (Schaul etal., 2015) ανέπτυξαν ένα πλαίσιο για την ιεράρχηση των εμπειριών, έτσι ώστε να αναπαράγονται πιο συχνά σημαντικές μεταβάσεις και επομένως να μαθαίνουν πιο αποτελεσματικά. Στην ιδανική περίπτωση επιθυμητό θα ήταν να δοκιμαστούν συχνότερα αυτές οι μεταβάσεις από τις οποίες υπάρχουν πολλά για μάθηση. Ως διακομιστής μεσολάβησης για το μαθησιακό δυναμικό, η προτεινόμενη προτεραιότητα της επανάληψης εμπειρίας (Prioritized Experience Replay - PER) δειγματίζει μεταβάσεις με πιθανότητα p_t σε σχέση με το τελευταίο απόλυτο σφάλμα που ορίστηκε ως εξής:

$$p_t \propto \left| r_j + \gamma \max_a \hat{Q}(s_{j+1}, a'; \boldsymbol{\theta}') - Q(s_j, a_j; \boldsymbol{\theta}) \right|^{\omega}$$

όπου ω είναι μια υπερ-παράμετρος που καθορίζει το σχήμα της κατανομής. Νέες μεταβάσεις εισάγονται στο buffer επανάληψης με μέγιστη προτεραιότητα, παρέχοντας προκατάληψη προς τις πρόσφατες μεταβάσεις. Πρέπει να σημειωθεί ότι οι στοχαστικές μεταβάσεις μπορεί επίσης να προτιμούνται. Μέσα από πραγματικά πειράματα σε πολλά παιχνίδια Atari, οι συγγραφείς αποδεικνύουν 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 μπορούμε να βρούμε και να ορίσουμε τις σημαντικές εμπειρίες στη μνήμη επανάληψης D.

3.3.3 DuelingDeep Q-Learning

Οι τιμές Q, δηλαδή Q(s,a), που χρησιμοποιούνται στον αλγόριθμο Q-learning είναι να εκφράσουν πόσο καλό είναι να πραγματοποιηθεί μια συγκεκριμένη ενέργεια σε μια δεδομένη κατάσταση. Η αξία μιας ενέργειας a σε μια δεδομένη κατάσταση s μπορεί πραγματικά να αποσυντεθεί σε δύο θεμελιώδεις τιμές. Η πρώτη τιμή είναι η συνάρτηση τιμής κατάστασης, δηλαδή V(s), για την εκτίμηση της σημασίας του να είναι σε μια συγκεκριμένη κατάσταση s. Η δεύτερη τιμή είναι η

συνάρτηση τιμής-ενέργειας, δηλαδή A(a), για την εκτίμηση της σημασίας της επιλογής μιας ενέργειας a σε σύγκριση με άλλες ενέργειες. Ως αποτέλεσμα, η συνάρτηση Q-value μπορεί να εκφραστεί με δύο βασικές συναρτήσεις τιμής ως εξής: Q (s,a) = V(s) + A(a).

Σε πολλά MDP δεν είναι απαραίτητο να εκτιμηθούν ταυτόχρονα και οι δύο τιμές, δηλαδή οι τιμές δράσης και κατάστασης της συνάρτησης Q(s, a). Για παράδειγμα, σε πολλά αγωνιστικά παιχνίδια, η μετακίνηση αριστερά ή δεξιά έχει σημασία, εάν και μόνο εάν ο παράγοντας συναντήσει τα εμπόδια ή τους εχθρούς. Εμπνευσμένοι από αυτήν την ιδέα οι συγγραφείς στο (Wang etal., 2016) εισάγουν την ιδέα χρήσης δύο ροών, 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 δράσης και κατάστασης, δηλαδή, V(s) και A(a). Τέλος, οι δύο ροές συνδυάζονται για να δημιουργήσουν μία μόνο έξοδο Q (s, a) ως εξής:

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

όπου β και α είναι οι παράμετροι των δύο ροών V(s; β) και A (s,α'; **α**), αντίστοιχα. Εδώ |A| είναι ο συνολικός αριθμός ενεργειών στο χώρο δράσης A. Στη συνέχεια η συνάρτηση απώλειας προκύπτει ως εξής:

$$\left[r_j + \gamma \max_{a_{j+1}} \hat{Q}\left(s_{j+1}, a_{j+1}; \boldsymbol{\alpha}', \boldsymbol{\beta}'\right) - Q\left(s_j, a_j; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right]^2$$

Μέσω της προσομοίωσης, οι (VanHasselt etal., 2016) έδειξαν ότι η προτεινόμενη «μονομαχία» DQN μπορεί να ξεπεράσει το DDQN σε 50 από τα 57 εκπαιδευμένα παιχνίδια Atari. Ωστόσο, η προτεινόμενη αρχιτεκτονική μονομαχίας ωφελεί σαφώς μόνο για MDP με μεγάλους χώρους δράσης. Για μικρούς χώρους κατάστασης η απόδοση της μονομαχίας DQL δεν είναι τόσο καλή όσο της διπλής DQL, όπως φαίνεται στα αποτελέσματα προσομοίωσης στην εργασία των (Wang etal., 2016).

3.3.4 Ασύγχρονη Βαθιά Q-Μάθηση πολλαπλών βημάτων

Οι περισσότερες από τις μεθόδους Q-learning, όπως DQL και 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, οι συγγραφείς στο (Mnih etal., 2016) εισάγουν μια μέθοδο που χρησιμοποιεί πολλαπλούς παράγοντες για να εκπαιδεύσει το DNN παράλληλα. Συγκεκριμένα οι (Nguyen etal., 2018) στην εργασία τους πρότειναν μια εκπαιδευτική διαδικασία που χρησιμοποιεί ασύγχρονες διαβαθμισμένες ενημερώσεις από πολλούς παράγοντες ταυτόχρονα. Αντί να εκπαιδεύουν έναν μόνο παράγοντα που αλληλεπιδρά με το περιβάλλον του, πολλοί παράγοντες αλληλεπιδρούν ταυτόχρονα με τη δική τους εκδοχή του περιβάλλοντος. Μετά από ένα ορισμένο αριθμό χρονικών βημάτων, οι συσσωρευμένες ενημερώσεις διαβάθμισης από έναν παράγοντα εφαρμόζονται σε ένα καθολικό μοντέλο, το DNN. Αυτές οι ενημερώσεις είναι ασύγχρονες και δεν κλειδώνουν. Επιπλέον, για να αντισταθμιστεί η προκατάληψη και η διακύμανση της διαβάθμισης πολιτικής, υιοθετήθηκε η μέθοδος ενημερώσεων n-βημάτων για να ενημερώσουν τη συνάρτηση ανταμοιβής. Συγκεκριμένα η συντετμημένη συνάρτηση ανταμοιβής n-βημάτων μπορεί να οριστεί από: $r_t^{(n)} = \sum_{k=0}^{n-1} \gamma^{(k)} r_{t+k+1}$

Έτσι η εναλλακτική απώλεια για κάθε παράγοντα θα προκύψει από:

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

Τα αποτελέσματα της ταχύτητας εκπαίδευσης και της ποιότητας της προτεινόμενης ασύγχρονης DQL με την εκμάθηση πολλαπλών βημάτων αναλύονται για διάφορες μεθόδους εκμάθησης ενίσχυσης, π.χ. Q-learning 1 βήματος, SARSA 1-βήματος και Q-learning n-βημάτων. Δείχνουν ότι οι ασύγχρονες ενημερώσεις έχουν σταθεροποιητική επίδραση στις ενημερώσεις πολιτικής και τιμών. Επίσης η προτεινόμενη μέθοδος ξεπερνά τους τρέχοντες «τελευταίας λέξης» αλγόριθμους στα παιχνίδια Atari, ενώ εκπαιδεύεται για το ήμισυ του χρόνου σε έναν μόνο πολυπύρηνο επεξεργαστή CPU (Central Processing Unit), αντί για GPU (Graphics Processing Unit). Ως αποτέλεσμα έχουν αναπτυχθεί μερικές πρόσφατες εφαρμογές ασύγχρονης DQL για προβλήματα ελέγχου παράδοσης σε ασύρματα συστήματα (Wang etal., 2018).

3.3.5 Κατανεμημένη βαθιά Q-μάθηση

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, γνωρίζουμε ότι ο αναμενόμενος χρόνος μετάδοσης ενός πακέτου σε ασύρματο δίκτυο είναι 20 λεπτά. Ωστόσο, αυτές οι πληροφορίες μπορεί να μην είναι τόσο σημαντικές,

επειδή μπορεί να υπερεκτιμούν τον χρόνο μετάδοσης τις περισσότερες φορές. 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 να συμβούν. Στη συνέχεια, η εκτίμηση για την αναμενόμενη μετάδοση υπερεκτιμάται τις περισσότερες φορές. Αυτό καθιστά τις εκτιμήσεις μη χρήσιμες για τους αλγόριθμους DQL.

Μια λύση χρησιμοποιώντας κατανεμημένη ενισχυτική εκμάθηση για να ενημερώσουν τη συνάρτηση Q-value με βάση την κατανομή και όχι τις προσδοκίες της. Συγκεκριμένα, έστω Z (s, α) είναι η επιστροφή που λαμβάνεται ξεκινώντας από την κατάσταση s, εκτελώντας την ενέργεια α και ακολουθώντας την τρέχουσα πολιτική, τότε Q (s, α) = E [Z (s, α)]. Εδώ, το Z αντιπροσωπεύει τη κατανομή μελλοντικών ανταμοιβών, η οποία δεν είναι πλέον μια βαθμιαία ποσότητα όπως το Q-values. Στη συνέχεια, περιγράφεται η έκδοση της εξίσωσης Bellman ως εξής: Z (s, α) = r + γZ (s', α'). Για παράδειγμα, εάν χρησιμοποιηθεί το DQN και εξαχθεί μια εμπειρία (s, α, r, s') από το buffer επανάληψης, τότε το δείγμα της κατανομής στόχου είναι:

$$Z(s,a) = r + \gamma Z(s', a^*)$$

Με $a^* = arg\max_{a'} Q(s,a')$ Παρόλο που η προτεινόμενη κατανεμημένη βαθιά Q-μάθηση αποδεικνύεται ότι υπερτερεί της συμβατικής DQL (Mnih etal., 2015) σε πολλά παιχνίδια Atari 2600 (45 από τα 57 παιχνίδια) η απόδοσή της βασίζεται πολύ στη λειτουργία κατανομής Z. Εάν η Z είναι καλά καθορισμένη, η απόδοση κατανεμημένης βαθιάς Q-μάθησης είναι πολύ πιο σημαντική από εκείνη της DQL. Διαφορετικά η απόδοσή του είναι ακόμη χειρότερη από αυτήν της DQL.

3.3.6 Βαθιά Q-μάθηση με NoisyNet

Στην εργασία τους οι (Fortunato etal., 2018) εισήγαγαν το NoisyNet, έναν τύπο νευρωνικού δικτύου του οποίου η μεροληψία και τα βάρη διαταράσσονται επαναληπτικά κατά τη διάρκεια της εκπαίδευσης από μια παραμετρική λειτουργία του θορύβου. Αυτό το δίκτυο προσθέτει βασικά τον θόρυβο Gauss στα τελευταία (πλήρως συνδεδεμένα) επίπεδα του δικτύου. Οι παράμετροι αυτού του θορύβου μπορούν να ρυθμιστούν από το μοντέλο κατά τη διάρκεια της εκμάθησης, το οποίο επιτρέπει στον παράγοντα να αποφασίσει πότε και σε ποιο ποσοστό θέλει να εισαγάγει την αβεβαιότητα στα βάρη του. Ειδικότερα, για να εφαρμοστεί το θορυβώδες δίκτυο, αντικαθίσταται πρώτα η ϵ-greedy πολιτική με μια συνάρτηση τυχαιοποιημένης τιμής-ενέργειας. Στη συνέχεια, τα πλήρως συνδεδεμένα επίπεδα του δικτύου τιμών παραμετροποιούνται ως θορυβώδες δίκτυο, όπου οι παράμετροι λαμβάνονται από τη θορυβώδη κατανομή παραμέτρων δικτύου μετά από κάθε

βήμα επανάληψης. Για επανάληψη, το τρέχον θορυβώδες δείγμα παραμέτρου δικτύου διατηρείται σταθερό κατά μήκος της παρτίδας. Δεδομένου ότι η DQL κάνει ένα βήμα βελτιστοποίησης για κάθε βήμα δράσης, οι θορυβώδεις παράμετροι δικτύου επαναλαμβάνονται στο δείγμα πριν από κάθε ενέργεια.

Μετά από αυτό, η συνάρτηση απώλειας μπορεί να διατυπωθεί ως εξής:

$$\mathcal{L} = \mathbb{E}\left[\mathbb{E}_{(s,a,r,s')\sim \mathrm{D}}\left[r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s', a', \epsilon'; \boldsymbol{\theta}') - Q(s, a, \epsilon; \boldsymbol{\theta})\right]\right]$$

όπου οι εξωτερικές και εσωτερικές προσδοκίες είναι σε σχέση με τις κατανομές των μεταβλητών θορύβου ϵ και ϵ' για τις λειτουργίες θορυβώδους τιμής $Q(s', \widehat{a', \epsilon}'; \boldsymbol{\theta}')$ και $Q(s, a, \epsilon, \boldsymbol{\theta})$ αντίστοιχα.

Μέσα από πειραματικά αποτελέσματα, οι συγγραφείς καταδεικνύουν ότι προσθέτοντας το επίπεδο θορύβου Gauss στο DNN, η απόδοση της συμβατικής DQL (Mnih etal., 2015), Dueling DQL (Wang etal., 2016) και της ασύγχρονης DQL (Mnih etal., 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.

3.3.7 Rainbow Deep Q-learning

Στην εργασία των (Nguyen etal., 2018) οι συγγραφείς προτείνουν μια λύση που ενσωματώνει όλα τα πλεονεκτήματα των επτά προαναφερθεισών λύσεων (συμπεριλαμβανομένης της DQL) σε έναν μοναδικό παράγοντα εκμάθησης, που ονομάζεται Rainbow DQL. Συγκεκριμένα, αυτός ο αλγόριθμος καθορίζει πρώτα τη συνάρτηση απώλειας με βάση την ασύγχρονη DQL πολλαπλών βημάτων και διανομής. Στη συνέχεια, οι (Nguyen etal., 2018) συνδύασαν την κατανεμημένη απώλεια πολλαπλών βημάτων με τη διπλή Q-learning χρησιμοποιώντας την άπληστη ενέργεια στο s_{t+n} που επιλέγεται σύμφωνα με το δίκτυο Q ως δράση εκκίνησης α^*_{t+n} και αξιολόγησαν τη δράση χρησιμοποιώντας τον στόχο δίκτυο.

Στην τυπική τεχνική αναλογικής επανάληψης προτεραιότητας (Schaul etal., 2015), το απόλυτο σφάλμα TD χρησιμοποιείται για την προτεραιότητα των μεταβάσεων. Εδώ το σφάλμα TD σε ένα χρονικό διάστημα είναι το σφάλμα στην εκτίμηση που έγινε στη χρονική υποδοχή. Ωστόσο, στον προτεινόμενο αλγόριθμο Rainbow DQL, όλες οι παραλλαγές 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 μονομαχίας των DNN παρουσιάζεται στην εργασία των (Wang etal., 2016). Τέλος, το επίπεδο NoisyNet (Hessel

etal., 2018) χρησιμοποιείται για την αντικατάσταση όλων των γραμμικών επιπέδων προκειμένου να μειωθεί ο αριθμός των ανεξάρτητων μεταβλητών θορύβου. Μέσω της προσομοίωσης, φαίνεται ότι αυτή είναι η πιο προηγμένη τεχνική που ξεπερνά σχεδόν όλους τους τρέχοντες αλγόριθμους DQL στη βιβλιογραφία με πάνω από 57 παιχνίδια Atari 2600.

3.3.8 Βαθιά Q-Learning για Επεκτάσεις MDPs

Βαθιά ντετερμινιστική πολιτική Gradient Q-Learning για Συνεχή δράση.
Αν και ο αλγόριθμος DQL μπορεί να λύσει προβλήματα με χώρους καταστάσεων υψηλής διάστασης, μπορεί να χειριστεί διακριτούς και χαμηλών διαστάσεων χώρους κατάστασης. Ωστόσο, τα συστήματα σε πολλές εφαρμογές έχουν συνεχείς, δηλαδή πραγματικές τιμές και χώρους δράσης υψηλών διαστάσεων. Οι αλγόριθμοι DQL δεν μπορούν να εφαρμοστούν άμεσα σε συνεχείς ενέργειες, καθώς βασίζονται στην επιλογή της καλύτερης ενέργειας που μεγιστοποιεί τη συνάρτηση Q-value. Συγκεκριμένα, μια πλήρης αναζήτηση σε έναν χώρο συνεχούς δράσης για την εύρεση της βέλτιστης δράσης είναι συχνά ανέφικτη.

Στην εργασία τους οι (Lillicrap etal., 2015) εισήγαγαν έναν μοντέλο ελεύθερο από πολιτική αλγόριθμου δράσης-κριτικής χρησιμοποιώντας προσεγγιστές βαθιάς λειτουργίας που μπορούν να μάθουν πολιτικές σε χώρους διαστάσεων συνεχούς δράσης. Η βασική ιδέα βασίζεται στον αλγόριθμο ντετερμινιστικής διαβάθμισης πολιτικής (Deterministic Policy Gradient - DPG). Συγκεκριμένα, ο αλγόριθμος DPG διατηρεί μια παραμετροποιημένη συνάρτηση δράσης μ(s; $\boldsymbol{\theta}^{\mu}$) με τον φορέα παραμέτρου $\boldsymbol{\theta}$ που καθορίζει την τρέχουσα πολιτική με καθοριστική αντιστοίχιση καταστάσεων σε μια συγκεκριμένη ενέργεια. Η κριτική Q (s,α) μαθαίνεται χρησιμοποιώντας την εξίσωση Bellman όπως στην Q-learning. Ο δρών ενημερώνεται εφαρμόζοντας τον κανόνα της αλυσίδας στην αναμενόμενη επιστροφή από την αρχή διανομής J σε σχέση με τις παραμέτρους του δρώντος ως εξής:

$$\nabla_{\boldsymbol{\theta}^{\mu}} J \approx \mathbb{E}_{s_t \sim \rho^{\beta}}\left[\nabla_{\boldsymbol{\theta}^{\mu}} Q(s, a; \boldsymbol{\theta}^{Q})|_{s=s_t, a=\mu(s_t|\boldsymbol{\theta}^{\mu})}\right]$$
$$\approx \mathbb{E}_{s_t \sim \rho^{\beta}}\left[\nabla_a Q(s, a; \boldsymbol{\theta}^{Q})|_{s=s_t, a=\mu(s_t)} \nabla_{\boldsymbol{\theta}\mu} \mu(s; \boldsymbol{\theta}^{\mu})\big|_{s=s_t}\right]$$

Με βάση αυτόν τον κανόνα ενημέρωσης, εισάγεται στη συνέχεια ο αλγόριθμος DDPG (Deep Deterministic Policy Gradient) που μπορεί να μάθει ανταγωνιστικές πολιτικές χρησιμοποιώντας παρατηρήσεις χαμηλών διαστάσεων (π.χ. καρτεσιανές συντεταγμένες ή κοινές γωνίες) υπό τις ίδιες υπερ-παραμέτρους και δομή δικτύου. Ο αλγόριθμος δημιουργεί ένα αντίγραφο των δικτύων δρώντων και κριτικών

$Q'(s, a; \theta^{Q'})$ και $\mu'(s; \theta^{\mu'})$ αντίστοιχα, για τον υπολογισμό των τιμών-στόχων. Τα βάρη αυτών των δικτύων στόχων στη συνέχεια ενημερώνονται με αργή παρακολούθηση στα εκπαιδευμένα

δίκτυα, δηλαδή $\theta' \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}'$με $\tau \ll 1$. Αυτό σημαίνει ότι οι τιμές-στόχοι περιορίζονται να αλλάζουν αργά, βελτιώνοντας σημαντικά τη σταθερότητα της μάθησης. Πρέπει να σημειωθεί ότι μια σημαντική πρόκληση της μάθησης σε χώρους συνεχούς δράσης είναι η εξερεύνηση. Επομένως, μια πολιτική εξερεύνησης μ΄ κατασκευάζεται με την προσθήκη θορύβου που έχει ληφθεί ως δείγμα από μια διαδικασία θορύβου N στην πολιτική του δρώντος.

Βαθιά επαναλαμβανόμενη Q-μάθηση για POMDPs

Για την αντιμετώπιση προβλημάτων με μερικώς παρατηρήσιμα περιβάλλοντα με ενίσχυση βαθιάς μάθησης, οι (Hausknecht&Stone, 2015) πρότειναν ένα πλαίσιο που ονομάζεται Deep Recurrent Q-Learning (DRQL) στο οποίο χρησιμοποιήθηκε ένα επίπεδο LSTM για να αντικαταστήσει το πρώτο μετα-συνελικτικό πλήρως συνδεδεμένο επίπεδο του συμβατικού DQN. Η επαναλαμβανόμενη δομή είναι σε θέση να ενσωματώσει ένα αυθαίρετα μακρύ ιστορικό για καλύτερη εκτίμηση της τρέχουσας κατάστασης, αντί να χρησιμοποιεί ένα ιστορικό σταθερού μήκους όπως στα DQN. Έτσι, τα Deep Recurrent Q-Network (DRQN) υπολογίζουν τη συνάρτηση $Q(o_t, h_{t-1}; \boldsymbol{\theta})$ αντί $Q(s_t, a_t); \boldsymbol{\theta})$ όπου $\boldsymbol{\theta}$ 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 ξεπερνά το DQN σε μερικώς παρατηρήσιμους τομείς. Όσον αφορά τη διαδικασία εκπαίδευσης, το DRQN λαμβάνει υπόψη μόνο τα συνελικτικά χαρακτηριστικά του ιστορικού παρατήρησης, αντί να ενσωματώνει ρητά τις δράσεις. Μέσα από τα πειράματα οι συγγραφείς αποδεικνύουν 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.

3.3.9 Βαθιά μάθηση SARSA

Στην εργασία τους οι (Zhao etal., 2016) εισήγαγαν μια τεχνική DQL που βασίζεται στην εκμάθηση SARSA για να βοηθήσουν τον παράγοντα να καθορίσει τις βέλτιστες πολιτικές μέσω διαδικτύου. Δεδομένης της τρέχουσας κατάστασης s, χρησιμοποιείται ένα CNN για τη λήψη της τρέχουσας τιμής κατάστασης-δράσης Q (s, α). Στη συνέχεια, η τρέχουσα ενέργεια α επιλέγεται από τον αλγόριθμο ϵ-greedy. Μετά από αυτό, μπορεί να παρατηρηθεί η άμεση ανταμοιβή r και η επόμενη κατάσταση s'. Για να εκτιμηθεί το τρέχον Q (s,α), λαμβάνεται η επόμενη τιμή κατάστασης-δράσης Q (s', α'). Εδώ, όταν η επόμενη κατάσταση s' χρησιμοποιείται ως είσοδος του CNN, μπορεί να ληφθεί ως έξοδος Q (s', α'). Στη συνέχεια, ο φορέας σήμανσης που σχετίζεται με το Q (s, a) ορίζεται ως Q (s', α') που αντιπροσωπεύει τον φορέα στόχο. Οι δύο φορείς έχουν μόνο μία διαφορετική συνιστώσα, δηλαδή, $r + \gamma Q(s', a') \rightarrow Q(s, a)$. Θα πρέπει να σημειωθεί ότι κατά τη φάση της εκπαίδευσης, η επόμενη δράση a' για την εκτίμηση της

τρέχουσας τιμής κατάστασης δεν είναι άπληστη. Αντιθέτως, υπάρχει μια μικρή πιθανότητα να επιλεγεί μια τυχαία δράση για εξερεύνηση.

3.3.10 Βαθιά Q-Learning για Παίγνια Markov

Οι (Wang etal., 2018) εισήγαγαν τη γενική έννοια του διαδοχικού διλήμματος κρατουμένων (Sequential Prisoner's Dilemma - SPD) για να διαμορφώσουν τα προβλήματα του πραγματικού κόσμου για το δίλημμα των κρατουμένων. 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. Για το λόγο αυτό προτάθηκε μια προσέγγιση βαθιάς ενισχυτικής μάθησης (Deep Reinforcement Learning – DRL) πολλαπλών παραγόντων για αμοιβαία συνεργασία σε παιχνίδια SDP (Session Description Protocol). Η βαθιά ενισχυτική μάθηση πολλαπλών παραγόντων προς την αμοιβαία συνεργασία αποτελείται από δύο φάσεις, την online και την offline. Η φάση εκτός σύνδεσης 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 δράση-κριτική έως ότου συγκλίνει. Δηλαδή: Πολιτική συνεργασίας και αφαιρετικής βάσης, Δεύτερον, ο αλγόριθμος συνθέτει το πλήρες φάσμα πολιτικών από τις παραπάνω πολιτικές βάσης. Ένα άλλο καθήκον είναι η αποτελεσματική ανίχνευση του βαθμού συνεργασίας του αντιπάλου. Ο αλγόριθμος χωρίζει αυτήν την εργασία σε δύο βήματα. Πρώτον, ο αλγόριθμος εκπαιδεύει ένα δίκτυο ανίχνευσης βαθμού συνεργασίας που βασίζεται σε LSTM εκτός σύνδεσης, το οποίο στη συνέχεια θα χρησιμοποιηθεί για ανίχνευση σε πραγματικό χρόνο κατά τη διάρκεια online φάσης. Στην φάση αυτή, ο παράγοντας παίζει εναντίον των αντιπάλων παλινδρομώντας με μια πολιτική ελαφρώς υψηλότερου βαθμού συνεργασίας από εκείνη του αντιπάλου. Αφενός, διαισθητικά ο αλγόριθμος προσανατολίζεται στη συνεργασία και επιδιώκει αμοιβαία συνεργασία, όποτε είναι δυνατόν. Από την άλλη πλευρά, ο αλγόριθμος είναι επίσης ισχυρός έναντι της εγωιστικής εκμετάλλευσης και καταφεύγει σε στρατηγική αφαίρεσης για να αποφύγει την εκμετάλλευση, όποτε είναι απαραίτητο.

Σε αντίθεση με τους (Wang etal., 2018) που θεώρησαν ένα επαναλαμβανόμενο κανονικό παιχνίδι με πλήρεις πληροφορίες, οι (Heinrich&Silver, 2016) παρουσίασαν μια εφαρμογή DRL για εκτεταμένα παιχνίδια μορφής με ατελείς πληροφορίες. Συγκεκριμένα, οι συγγραφείς του (Heinrich&Silver, 2016) εισάγουν το Neural Fictitious Self-Play (NFSP), μια μέθοδο DRL για την εκμάθηση της ισορροπίας Nash κατά προσέγγιση των ατελών παιχνιδιών πληροφοριών. Το NFSP συνδυάζει FSP με προσέγγιση λειτουργιών νευρωνικού δικτύου. Ένας παράγοντας NFSP

έχει δύο νευρωνικά δίκτυα. Το πρώτο δίκτυο εκπαιδεύεται με ενίσχυση που μαθαίνει από την απομνημονευμένη εμπειρία του παιχνιδιού ενάντια σε συναδέλφους παράγοντες. Αυτό το δίκτυο μαθαίνει μια κατά προσέγγιση καλύτερη απόκριση στην ιστορική συμπεριφορά άλλων παραγόντων. Το δεύτερο δίκτυο εκπαιδεύεται από εποπτευόμενη εκμάθηση από απομνημονευμένη εμπειρία της συμπεριφοράς του παράγοντα. Αυτό το δίκτυο μαθαίνει ένα μοντέλο που έχει μέσο όρο έναντι των ιστορικών στρατηγικών του παράγοντα, ο οποίος συμπεριφέρεται σύμφωνα με ένα μείγμα της μέσης στρατηγικής του και της στρατηγικής βέλτιστης απόκρισης.

Στο NSFP, όλοι οι παίκτες του παιχνιδιού ελέγχονται από ξεχωριστούς παράγοντες NFSP που μαθαίνουν από το ταυτόχρονο παιχνίδι εναντίον των υπολοίπων, δηλαδή, το αυτο-παιχνίδι. Ένας παράγοντας NFSP αλληλεπιδρά με τους συναδέλφους του και απομνημονεύει την εμπειρία του από τις μεταβάσεις παιχνιδιών και τη δική του καλύτερη συμπεριφορά απόκρισης σε δύο αναμνήσεις, την M_{RL} και την M_{SL}. Το NFSP αντιμετωπίζει αυτές τις αναμνήσεις ως δύο ξεχωριστά σύνολα δεδομένων κατάλληλα για DRL και εποπτευόμενη ταξινόμηση, αντίστοιχα. Ο παράγοντας εκπαιδεύει ένα νευρωνικό δίκτυο, $Q(s, a; \theta^Q)$ για να προβλέψει τιμές δράσης από δεδομένα σε M_{RL} χρησιμοποιώντας εκμάθηση ενίσχυσης εκτός πολιτικής. Το προκύπτον δίκτυο ορίζει το η στρατηγική βέλτιστης απόκρισης του παράγοντα, β = ϵ-greedy (Q), η οποία επιλέγει μια τυχαία ενέργεια με πιθανότητα ϵ και διαφορετικά επιλέγει την ενέργεια που μεγιστοποιεί τις προβλεπόμενες τιμές δράσης. Ο παράγοντας 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 στο 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, οι συγγραφείς δείχνουν ότι το NFSP μπορεί να συγκλίνει σε προσέγγιση της ισορροπίας Nash σε ένα μικρό παιχνίδι πόκερ.

4

Εφαρμογές των DRL/DQL στις Επικοινωνίες και στη Δικτύωση

Προσφάτως η βαθιά μάθηση (Deep Learning - DL) εισήχθη ως μια νέα πρωτοποριακή τεχνική, η οποία είναι ικανή να ξεπεράσει τους περιορισμούς της ενισχυτικής μάθησης και με αυτόν τον τρόπο να ανοίξει μια νέα εποχή για την ανάπτυξη της ενισχυτικής μάθησης, δηλαδή τη βαθιά ενισχυτική μάθηση (DRL). Η DRL ενστερνίζεται το πλεονέκτημα των βαθιά νευρωνικών δικτύων (DNNs) για την εκπαίδευση της διαδικασίας της μάθησης, βελτιώνοντας έτσι την ταχύτητα εκμάθησης και την απόδοση των αλγορίθμων ενίσχυσης μάθησης. Ως αποτέλεσμα, η DRL έχει υιοθετηθεί σε πολυάριθμες εφαρμογές ενισχυτικής μάθησης, όπως η ρομποτική, η υπολογιστική όραση, η αναγνώριση ομιλίας και η επεξεργασία φυσικής γλώσσας.

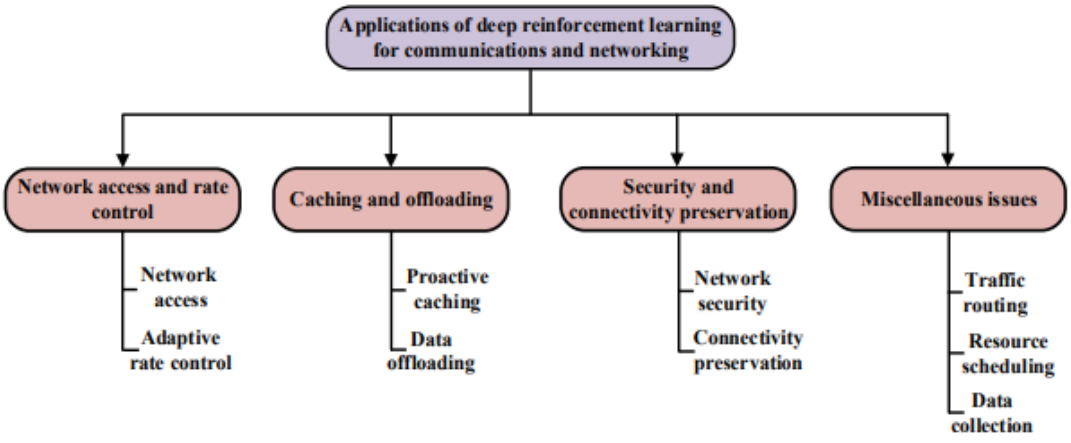
Όσον αφορά τους τομείς των επικοινωνιών και της δικτύωσης, η DRL έχει χρησιμοποιηθεί πρόσφατα ως αναδυόμενο εργαλείο για την αποτελεσματική αντιμετώπιση διαφόρων προβλημάτων και προκλήσεων. Ειδικότερα, τα σύγχρονα δίκτυα, όπως το Διαδίκτυο των Πραγμάτων (IoT), τα ετερογενή δίκτυα (Heterogeneous Networks - HetNets) και το δίκτυο των μη επανδρωμένων εναέριων οχημάτων (UAV), γίνονται πιο αποκεντρωμένα και αυτόνομα. Οι οντότητες δικτύου όπως οι συσκευές IoT, οι χρήστες κινητών τηλεφώνων και τα UAVs πρέπει να λαμβάνουν τοπικές και αυτόνομες αποφάσεις, π.χ. πρόσβαση φάσματος, επιλογή ρυθμού δεδομένων, έλεγχος ισχύος μετάδοσης και συσχέτιση σταθμών βάσης, για να επιτύχουν τους στόχους διαφορετικών δικτύων, όπως π.χ. μεγιστοποίηση και ελαχιστοποίηση της κατανάλωσης ενέργειας. Σε αβέβαια και στοχαστικά περιβάλλοντα, τα περισσότερα από τα προβλήματα λήψης

αποφάσεων μπορούν να διαμορφωθούν μέσω της λεγόμενης διαδικασίας αποφάσεων Markov (MDP) (Puterman, 2014). Ο δυναμικός προγραμματισμός και άλλοι αλγόριθμοι, όπως η επανάληψη τιμών, καθώς και τεχνικές ενίσχυσης εκμάθησης μπορούν να υιοθετηθούν για την επίλυση της MDP. Ωστόσο, λόγω του ότι τα σύγχρονα δίκτυα είναι μεγάλης κλίμακας και περίπλοκα, η υπολογιστική πολυπλοκότητα των τεχνικών γίνεται γρήγορα αδιαχείριστη. Ως αποτέλεσμα, η DRL έχει αναπτυχθεί ως μια εναλλακτική λύση για να ξεπεραστεί αυτή η πρόκληση. Γενικά, οι προσεγγίσεις DRL παρέχουν τα ακόλουθα πλεονεκτήματα:

- Η DRL μπορεί να βρει τη λύση εξελιγμένων βελτιστοποιήσεων δικτύου. Έτσι, επιτρέπει στους ελεγκτές δικτύου, π.χ. σταθμούς βάσης στα σύγχρονα δίκτυα, να επιλύουν πολύπλοκα προβλήματα, π.χ. συσχέτιση κοινών χρηστών, υπολογισμούς και χρονοδιάγραμμα μετάδοσης, ούτως ώστε να μπορούν να επιτύχουν τις βέλτιστες λύσεις χωρίς να απαιτούνται πλήρεις και ακριβείς πληροφορίες δικτύου.
- Η DRL επιτρέπει στις οντότητες του δικτύου να αναπτύξουν γνώσεις σχετικά με το περιβάλλον επικοινωνίας και δικτύωσης. Έτσι, χρησιμοποιώντας την DRL, οι οντότητες δικτύου, π.χ. ένας χρήστης κινητής τηλεφωνίας, μπορούν να μάθουν βέλτιστες πολιτικές, π.χ. επιλογή σταθμού βάσης, επιλογή καναλιού, απόφαση παράδοσης, προσωρινή αποθήκευση και εκφόρτωση αποφάσεων, χωρίς να γνωρίζουν το μοντέλο καναλιού και το πρότυπο κινητικότητας.
- Η DRL παρέχει αυτόνομη λήψη αποφάσεων. Με τις προσεγγίσεις DRL οι οντότητες του δικτύου μπορούν να παρατηρήσουν και να αποκτήσουν την καλύτερη πολιτική τοπικά με ελάχιστη ή χωρίς ανταλλαγή πληροφοριών μεταξύ τους. Αυτό όχι μόνο μειώνει τα γενικά έξοδα επικοινωνίας, αλλά βελτιώνει επίσης την ασφάλεια και την ευρωστία των δικτύων.
- Η DRL βελτιώνει σημαντικά την ταχύτητα εκμάθησης, ειδικά στα προβλήματα με μεγάλους χώρους κατάστασης και δράσης. Έτσι, σε δίκτυα μεγάλης κλίμακας, π.χ. συστήματα IoT με χιλιάδες συσκευές, η DRL επιτρέπει στον ελεγκτή δικτύου ή στις πύλες IoT να ελέγχουν δυναμικά τη συσχέτιση χρηστών, την πρόσβαση στο φάσμα και τη μετάδοση ισχύος για έναν τεράστιο αριθμό συσκευών IoT και χρηστών κινητών τηλεφώνων.
- Αρκετά άλλα προβλήματα στις επικοινωνίες και τη δικτύωση, όπως οι φυσικές επιθέσεις στον κυβερνοχώρο, η διαχείριση παρεμβολών και η εκφόρτωση δεδομένων μπορούν να διαμορφωθούν ως παίγνια, π.χ., το μη συνεργατικό παίγνιο (Non-cooperative game). Η DRL έχει χρησιμοποιηθεί πρόσφατα ως ένα αποτελεσματικό εργαλείο για την επίλυση παιγνίων, π.χ. για την εύρεση της ισορροπίας Nash (Nash equilibrium).

Αν και υπάρχουν κάποιες έρευνες που σχετίζονται με την DRL, δεν επικεντρώνονται στις επικοινωνίες και τη δικτύωση. Όπως για παράδειγμα, οι έρευνες των συγγραφέων (Li, 2017) και (Arulkumaran etal., 2017) που αφορούν εφαρμογές της DRL για την υπολογιστική όραση και την επεξεργασία φυσικής γλώσσας. Επίσης υπάρχουν έρευνες που σχετίζονται με τη χρήση της βαθιάς μάθησης μόνο για δικτύωση. Για παράδειγμα, η έρευνα των (Chen etal., 2019) εστιάζει στη μηχανική μάθηση για ασύρματα δίκτυα, αλλά όχι και στις προσεγγίσεις DRL. Σύμφωνα με τους συγγραφείς (Luong etal., 2019), δεν υπάρχει έρευνα που να συζητά συγκεκριμένα τις εφαρμογές της DRL στις επικοινωνίες και τη δικτύωση. Για λόγους ευκολίας, ταξινομούν τις σχετικές εφαρμογές με βάση τα ζητήματα στις επικοινωνίες και τη δικτύωση, όπως φαίνεται στο πιο κάτω σχήμα (Σχήμα 5). Τα κύρια ζητήματα περιλαμβάνουν πρόσβαση στο δίκτυο, έλεγχο ρυθμού δεδομένων, ασύρματη προσωρινή αποθήκευση, εκφόρτωση δεδομένων, ασφάλεια δικτύου, διατήρηση συνδεσιμότητας, δρομολόγηση κυκλοφορίας και συλλογή δεδομένων.

Επιπλέον, οι συγγραφείς (Luong etal., 2019) αναφέρονται σε εφαρμογές της DQL για τη δυναμική πρόσβαση στο δίκτυο και τον προσαρμοστικό έλεγχο ρυθμού. Παρατηρούν ότι τα προβλήματα διαμορφώνονται ως επί το πλείστων ως MDP. Επιπλέον, οι προσεγγίσεις DQL για τα συστήματα IoT και DASH λαμβάνουν μεγαλύτερη προσοχή σε σχέση με άλλα δίκτυα. Τα μελλοντικά δίκτυα, π.χ. τα δίκτυα 5G, περιλαμβάνουν πολλαπλές οντότητες δικτύου με πολλαπλούς αντικρουόμενους στόχους, π.χ. τα έσοδα του παρόχου έναντι της μεγιστοποίησης της χρησιμότητας των χρηστών. Αυτό θέτει μια σειρά από προκλήσεις στους παραδοσιακούς μηχανισμούς διαχείρισης πόρων που χρήζουν μια πιο εις βάθος διερεύνηση.



Σχήμα 5. Ταξινόμηση των εφαρμογών της βαθιάς ενισχυτικής μάθησης για επικοινωνίες και δικτύωση. Πηγή: (Luong etal., 2019).

Στη συνέχεια αναλύονται οι εφαρμογές της DQL για την αντιμετώπιση της επίθεσης παρεμβολών. Μία πρωτοποριακή έρευνα που χρησιμοποιεί την DQL για την αντιμετώπιση της παρεμβολής είναι αυτή των (Han etal., 2017). Σε αυτήν το μοντέλο δικτύου είναι ένα Γνωστικό Ραδιοφωνικό Δίκτυο (Cognitive Radio Network - CRN) που αποτελείται από έναν δευτερεύοντα χρήστη (Secondary User - SU), πολλαπλούς κύριους χρήστες (Primary Users - PUs) και πολλούς παρεμβολείς. Επιπλέον το δίκτυο διαθέτει ένα σύνολο καναλιών συχνότητας για μεταπήδηση συχνότητας. Σε κάθε χρονοθυρίδα κάθε ένας από τους παρεμβολείς μπορεί αυθαίρετα να επιλέξει ένα από τα κανάλια για να στείλει το σήμα εμπλοκής του και στη συνέχεια ο SU, δηλαδή ο πράκτορας λογισμικού, πρέπει να επιλέξει μια σωστή ενέργεια με βάση την τρέχουσα κατάστασή του. Η ενέργεια αυτή μπορεί να είναι (1) επιλογή ενός από τα κανάλια για αποστολή των σημάτων του ή (2) έξοδος από την περιοχή για σύνδεση σε άλλο σταθμό βάσης (Base Station - BS). Οι παρεμβολές θεωρείται ότι αποφεύγουν την πρόκληση παρεμποδίσεων στους PUs και έτσι η τρέχουσα κατάσταση του SU αποτελείται από τον αριθμό των PUs και τη διακριτική αναλογία σήματος προς παρεμβολές και θόρυβο (Signal to Interference & Noise Ratio – SINR) του σήματος SU στην τελευταία χρονική θυρίδα. Ο στόχος του SU είναι να μεγιστοποιήσει την αναμενόμενη μείωση της χρησιμότητας σε χρονοθυρίδες. Πρέπει επίσης να σημειωθεί ότι, όταν ο SU επιλέξει να εγκαταλείψει την περιοχή για να συνδεθεί σε έναν άλλο BS, ξοδεύει ένα κόστος μετακίνησης. Έτσι, η χρησιμότητα ορίζεται ως συνάρτηση της SINR όσον αφορά το σήμα του SU και του κόστους της κινητικότητας. Δεδομένου ότι ο αριθμός των καναλιών συχνότητας μπορεί να είναι μεγάλος (πράγμα που οδηγεί σε ένα μεγάλο σύνολο ενεργειών), χρησιμοποιείται ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN) για την DQL, ούτως ώστε να αφομοιώσει γρήγορα τη βέλτιστη πολιτική. Επιπλέον, λαμβάνοντας υπόψη ένα σενάριο με δύο παρεμβολείς, η προτεινόμενη DQL υπερέχει της μεθόδου αναπήδησης συχνότητας, όσον αφορά την SINR και το κόστος κινητικότητας.

Επιπροσθέτως, το μοντέλο στην έρευνα των (Han etal., 2017) περιορίζεται σε δύο παρεμβολείς. Καθώς ο αριθμός των παρεμβολών στο δίκτυο αυξάνεται, το προτεινόμενο σχήμα ενδέχεται να μην είναι αποτελεσματικό στον ίδιο βαθμό. Ο λόγος που συμβαίνει αυτό είναι ότι γίνεται δύσκολο για τον SU να εντοπίσει βέλτιστες ενέργειες όταν αυξάνεται ο αριθμός των μπλοκαρισμένων καναλιών. Μια κατάλληλη λύση, όπως προτείνεται στην έρευνα των (Xiao etal., 2018), επιτρέπει στον δέκτη του SU να εγκαταλείψει την τρέχουσα θέση του. Εφόσον αυτή η αναχώρηση συνεπάγεται το κόστος κινητικότητας, ο παραλήπτης, δηλαδή ο πράκτορας λογισμικού, χρειάζεται μια βέλτιστη πολιτική, δηλαδή να παραμείνει στην τρέχουσα τοποθεσία ή να αποχωρήσει, για να μεγιστοποιήσει τη χρησιμότητά του. Σε αυτό το σενάριο, μπορεί να

χρησιμοποιηθεί η DQL που βασίζεται σε ένα CNN, ώστε ο δέκτης να εντοπίσει τη βέλτιστη ενέργεια προκειμένου να μεγιστοποιήσει την αναμενόμενη χρησιμότητά του. Σε αυτή τη περίπτωση η χρησιμότητα και η κατάσταση του δέκτη ορίζονται παρόμοια με αυτή του πράκτορα λογισμικού στην έρευνα των (Han etal., 2017). Πιο συγκεκριμένα, η κατάσταση περιλαμβάνει τη διακριτοποιημένη SINR του σήματος που μετρήθηκε από τον δέκτη, στην τελευταία χρονική θυρίδα.

Από την άλλη πλευρά, οι συγγραφείς (Chen etal., 2018) προτείνουν τη χρήση της DQL για την εύρεση μιας βέλτιστης πολιτικής ελέγχου ισχύος κατά της παρεμβολής. Εδώ το μοντέλο είναι ένα δίκτυο IoT που περιλαμβάνει συσκευές IoT και ένα παρεμβολέα. Ο παρεμβολέας παρατηρεί τις επικοινωνίες του πομπού και επιλέγει μια στρατηγική εμπλοκής για να μειώσει την SINR στον δέκτη. Έτσι, ο πομπός επιλέγει μια ενέργεια, δηλαδή το επίπεδο ισχύος μετάδοσης, για να μεγιστοποιήσει τη χρησιμότητά του. Στη προκειμένη περίπτωση, η χρησιμότητα είναι η διαφορά μεταξύ της SINR και του κόστους κατανάλωσης ενέργειας λόγω της μετάδοσης. Η επιλογή της ισχύος μετάδοσης επηρεάζει τη μελλοντική στρατηγική εμπλοκής και έτσι η αλληλεπίδραση μεταξύ του πομπού και του παρεμβολέα μπορεί να διαμορφωθεί ως MDP. Ο πομπός αποτελεί τον πράκτορα λογισμικού, ενώ η κατάσταση είναι η SINR που μετράται στον δέκτη του πομπού στην τελευταία χρονική θυρίδα. Στη συνέχεια το DQN που χρησιμοποιεί το CNN, υιοθετείται για να βρεθεί μια βέλτιστη πολιτική ελέγχου ισχύος για τον πομπό, προκειμένου να μεγιστοποιήσει την αναμενόμενη συσσωρευμένη και μειωμένη χρησιμότητα σε χρονικές χρονοθυρίδες. Τα αποτελέσματα αυτής της προσομοίωσης δείχνουν ότι η προτεινόμενη DQL μπορεί να βελτιώσει τη χρησιμότητα του πομπού έως και 17,7% σε σύγκριση με τον αλγόριθμο της ενισχυτικής μάθησης (Q-learning). Επίσης, η προτεινόμενη DQL μειώνει τη χρησιμότητα του παρεμβολέα κατά περίπου 18,1% σε σχέση με την Q-learning.

Προκειμένου να αποτρέψει τις παρατηρήσεις των επικοινωνιών από τον παρεμβολέα, ο πομπός μπορεί να αλλάξει τη στρατηγική επικοινωνίας του, π.χ. χρησιμοποιώντας ηλεκτρονόμους που βρίσκονται μακριά από την περιοχή εμπλοκής. Οι ηλεκτρονόμοι μπορεί να είναι UAV όπως προτείνεται στην έρευνα των (Lue tal., 2020), με το μοντέλο να αποτελείται από ένα UAV, δηλαδή, έναν ηλεκτρονόμο, έναν παρεμβολέα, έναν χρήστη κινητού τηλεφώνου και τον BS που εξυπηρετεί. Ο χρήστης κινητού τηλεφώνου μεταδίδει μηνύματα στον διακομιστή του μέσω του BS εξυπηρέτησης. Σε περίπτωση που ο BS είναι πολύ μπλοκαρισμένος, το UAV βοηθά τον χρήστη να αναμεταδίδει τα μηνύματα στον διακομιστή μέσω ενός εφεδρικού BS. Ειδικότερα, ανάλογα με τις τιμές της SINR και του ρυθμού σφάλματος Bit (Bit Error Rate - BER) που αποστέλλονται από το BS, το UAV ως πράκτορας λογισμικού, αποφασίζει το επίπεδο ισχύος του

ηλεκτρονόμου για να μεγιστοποιήσει τη χρησιμότητά του, δηλαδή τη διαφορά μεταξύ της SINR και του κόστους του ηλεκτρονόμου. Το επίπεδο ισχύος του ηλεκτρονόμου μπορεί να θεωρηθεί ως οι ενέργειες του UAV, ενώ τα SINR και BER ως οι καταστάσεις του. Ως εκ τούτου, η επόμενη κατάσταση που παρατηρείται από το UAV είναι ανεξάρτητη από όλες τις προηγούμενες καταστάσεις και ενέργειες και το πρόβλημα διατυπώνεται ως μία MDP. Τέλος και για να επιτευχθεί γρήγορα η βέλτιστη πολιτική αναμετάδοσης για το UAV, υιοθετείται στη συνέχεια η DQL που βασίζεται στο CNN. Τα αποτελέσματα της προσομοίωσης στην έρευνα των (Lu etal., 2020) δείχνουν ότι το προτεινόμενο σχήμα DQL χρειάζεται μόνο 200 χρονικές χρονοθυρίδες για να συγκλίνει στη βέλτιστη πολιτική, που είναι κατά 83,3% λιγότερη από αυτή του σχήματος αναμετάδοσης που βασίζεται στην ενισχυτική μάθηση. Ακόμα, το προτεινόμενο σχήμα DQL μειώνει τον BER του χρήστη κατά 46,6% σε σύγκριση με το σχέδιο ενός ηλεκτρονόμου UAV.

Επιπρόσθετα, τα αποτελέσματα της προσομοίωσης στην έρευνα των (Xiao etal., 2017) δείχνουν ότι η προτεινόμενη DQL μπορεί να βελτιώσει τη χρησιμότητα του UAV έως και 13% σε σύγκριση με το βασικό σχήμα της έρευνας των (Bowling&Veloso, 2002), το οποίο χρησιμοποιεί τον αλγόριθμο WoLF-PHC (Win or Learn Fast - Policy Hill Climbing) για να αποτρέψει επιθέσεις. Επίσης, ο ασφαλής ρυθμός του UAV, δηλαδή η πιθανότητα επίθεσης στο UAV που προκύπτει από την προτεινόμενη DQL, είναι κατά 7% υψηλότερος από αυτόν της γραμμής βάσης. Ωστόσο, η προτεινόμενη DQL εφαρμόζεται μόνο σε ένα σύστημα ενός UAV. Για μελλοντικές μελέτες, πρέπει να ληφθούν υπόψη σενάρια με πολλαπλά UAVs, καθώς σε ένα τέτοιο σενάριο αναμένεται περισσότερη υπολογιστική επιβάρυνση και μπορούν να εφαρμοστούν αλγόριθμοι DQL πολλαπλών πρακτόρων.

Τέλος, οι συγγραφείς (Luong etal., 2019) αποτυπώνουν μέσω του Πίνακα 1 παρακάτω, διάφορες εφαρμογές της DQL για την ασφάλεια του δικτύου. Παρατηρούν ότι το CNN χρησιμοποιείται κυρίως για την DQL με σκοπό τη βελτίωση της ασφάλειας του δικτύου. Επιπλέον, οι προσεγγίσεις DQL για ένα ανώνυμο σύστημα, όπως τα συστήματα ρομπότ και το ευφυές σύστημα μεταφορών (Intelligent Transportation System – ITS), λαμβάνουν μεγαλύτερη προσοχή σε σχέση με άλλα δίκτυα. Ωστόσο, οι εφαρμογές της DQL για την κυβερνο-φυσική ασφάλεια είναι σχετικά λίγες και πρέπει να διερευνηθούν περαιτέρω. Στην επόμενη υποενότητα παρουσιάζεται μια πιο ειδικευμένη προσέγγιση στην εφαρμογή της DRL, σε σχέση με την ασφάλεια στον Κυβερνοχώρο και αποτυπώνεται η αρχιτεκτονική ενός αλγορίθμου DRL για τον έλεγχο ενός αυτόνομου οχήματος.

Πίνακας 1. Σύνοψη προσεγγίσεων που χρησιμοποιούν DQL για ασφάλεια δικτύου. Πηγή: (Luong etal., 2019).

Μοντέλο	Αλγόριθμοι μάθησης	Πράκτορας λογισμικού	Καταστάσεις	Δράσεις	Ανταμοιβές	Δίκτυα
Παίγνιο	DQN πουχρησιμοποιεί CNN	SU	Αριθμός PUs και σήματος SINR	Επιλογή καναλιού και απόφαση αποχώρησης	SINR και κόστος κινητικότητας	CRN
Παίγνιο	DQN πουχρησιμοποιεί CNN	Μετατροπέας λήψης	Σήμα SINR	Αποφάσεις παραμονής και αποχώρησης	SINR και κόστοςκινητικότητας	Υποβρύχιο ακουστικόδίκτυο
MDP	DQN πουχρησιμοποιεί CNN	Συσκευήμετάδοσης IoT	Σήμα SINR	Επιλογή καναλιού	SINR και κόστος κατανάλωσης ενέργειας	IoT
MDP	DQN πουχρησιμοποιεί CNN	Ηλεκτρονόμος UAV	Σήμα SINR και BER	Ισχύς ηλεκτρονόμου	Κόστος SINR και ηλεκτρονόμου	UAV
MDP	DQN πουχρησιμοποιεί CNN	Μετάδοση UAV	Ισχύς παρεμβολής	Μετάδοση ισχύος	Χωρητικότητα μυστικότητας και κόστος κατανάλωσης ενέργειας	UAV

Εφαρμογή της DRL για την ασφάλεια στον Κυβερνοχώρο

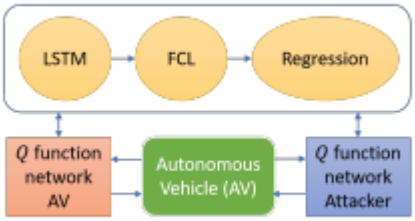
Έχουν προταθεί κατά καιρούς διάφοροι μέθοδοι βαθιάς ενισχυτικής μάθησης (DRL), για την αντιμετώπιση των ολοένα και αυξανόμενων κυβερνοεπιθέσεων. Με την ενσωμάτωση της βαθιάς μάθησης στην παραδοσιακή RL, η DRL είναι εξαιρετικά ικανή στο να επιλύει πολύπλοκα, δυναμικά και υψηλών διαστάσεων προβλήματα άμυνας στον κυβερνοχώρο. Ειδικότερα, η έρευνα των συγγραφέων (Nguyen&Reddi, 2019) παρουσιάζει μια ανάλυση των προσεγγίσεων DRL που αναπτύχθηκαν για την ασφάλεια στον κυβερνοχώρο. Μέσω αυτής αγγίζουν διάφορες πτυχές ζωτικής σημασίας, συμπεριλαμβανομένων μεθόδων ασφαλείας που βασίζονται σε DRL για κυβερνοφυσικά συστήματα, τεχνικών αυτόνομης ανίχνευσης εισβολής και προσομοιώσεων θεωρίας παιγνίων που βασίζονται σε DRL πολλαπλών παραγόντων για αμυντικές στρατηγικές έναντι επιθέσεων στον κυβερνοχώρο. Δίνονται επίσης εκτενείς συζητήσεις και μελλοντικές

ερευνητικές κατευθύνσεις σχετικά με την ασφάλεια στον κυβερνοχώρο που βασίζεται σε DRL. Αυτή η ολοκληρωμένη ανασκόπηση παρέχει τα θεμέλια και διευκολύνει μελλοντικές μελέτες σχετικά με τη διερεύνηση των δυνατοτήτων των αναδυόμενων DRL για την αντιμετώπιση ολοένα και πιο περίπλοκων προβλημάτων ασφάλειας στον κυβερνοχώρο.

Ένα πλαίσιο κατανομής πόρων που βασίζεται σε DRL και το οποίο ενσωματώνει δυνατότητες δικτύωσης, προσωρινής αποθήκευσης και υπολογισμού για εφαρμογές έξυπνων πόλεων προτείνεται στην έρευνα των (He etal., 2017). Ο αλγόριθμος DRL χρησιμοποιείται για την επίλυση αυτού του προβλήματος, επειδή περιλαμβάνει έναν μεγάλο χώρο κατάστασης, ο οποίος αποτελείται από τη δυναμική μεταβαλλόμενη κατάσταση των σταθμών βάσης, τους διακομιστές προσωρινής αποθήκευσης ακρών κινητής τηλεφωνίας (Mobile Edge Caching - MEC) και την κρυφή μνήμη. Το πλαίσιο έχει αναπτυχθεί με βάση την αρχή προγραμματιζόμενου ελέγχου ενός δικτύου που καθορίζεται από λογισμικό (Software Defined Network – SDN) και την ικανότητα αποθήκευσης στην κρυφή μνήμη της πληροφοριοκεντρικής δικτύωσης. Εναλλακτικά, οι συγγραφείς (Zhu etal., 2018) διερεύνησαν τις πολιτικές MEC χρησιμοποιώντας την έννοια της επίγνωσης του περιβάλλοντος που αντιπροσωπεύει τις πληροφορίες περιβάλλοντος του χρήστη και τα στατιστικά μοτίβων επισκεψιμότητας. Η χρήση τεχνολογιών τεχνητής νοημοσύνης στα άκρα του δικτύου κινητής τηλεφωνίας υποστηρίζεται για την έξυπνη εκμετάλλευση του λειτουργικού περιβάλλοντος και τη λήψη των σωστών αποφάσεων σχετικά με το τι, πού και πώς να αποθηκεύει ο εκάστοτε χρήστης το κατάλληλο περιεχόμενο. Για να αυξηθεί η απόδοση της προσωρινής αποθήκευσης, χρησιμοποιείται μια προσέγγιση DRL, δηλαδή ο αλγόριθμος ασύγχρονου πλεονεκτήματος παίκτη-κριτικού (actor-critic algorithm), για την εύρεση μιας βέλτιστης πολιτικής με στόχο τη μεγιστοποίηση της κίνησης εκφόρτωσης. Τέλος, οι εφαρμογές της DRL στον κυβερνοχώρο και σε διάφορα περιβάλλοντα κατηγοριοποιούνται γενικά σε δύο προοπτικές: την βελτιστοποίηση και την ενίσχυση των δυνατοτήτων επικοινωνίας και δικτύωσης των εφαρμογών IoT, καθώς και την άμυνα έναντι επιθέσεων στον κυβερνοχώρο. Η έρευνα των (Nguyen&Reddi, 2019) εστιάζει στο δεύτερο σκέλος όπου χρησιμοποιούνται μέθοδοι DRL για την επίλυση προβλημάτων ασφάλειας στον κυβερνοχώρο με την παρουσία επιθέσεων ή απειλών.

Έχει προταθεί σε ποικίλες έρευνες ένας μεγάλος αριθμός εφαρμογών της RL σε διάφορες πτυχές της ασφάλειας στον κυβερνοχώρο, που κυμαίνονται από το απόρρητο δεδομένων έως την προστασία κρίσιμων υποδομών. Ωστόσο, τα μειονεκτήματα της παραδοσιακής RL έχουν περιορίσει την ικανότητά της να επιλύει περίπλοκα και μεγάλης κλίμακας προβλήματα ασφάλειας στον κυβερνοχώρο. Τα τελευταία χρόνια, ο αυξανόμενος αριθμός συνδεδεμένων συσκευών IoT έχει οδηγήσει σε σημαντική αύξηση του αριθμού των περιπτώσεων κυβερνοεπιθέσεων καθώς και

στην πολυπλοκότητά τους. Η εμφάνιση της βαθιάς μάθησης και η ενσωμάτωσή της με την RL έχουν δημιουργήσει μια κατηγορία μεθόδων DRL που είναι σε θέση να ανιχνεύουν και να καταπολεμούν εξελιγμένους τύπους επιθέσεων στον κυβερνοχώρο, όπως η έγχυση πλαστών δεδομένων σε κυβερνοφυσικά συστήματα (Akazaki etal., 2018), η επίθεση εξαπάτησης σε αυτόνομα συστήματα (Gupta&Yang, 2018), κατανεμημένες επιθέσεις άρνησης παροχής υπηρεσίας (Malialis&Kudenko, 2015), εισβολές σε κεντρικούς υπολογιστές ή δίκτυα (Lopez-Martin etal., 2020), πλαστογράφηση (Xiao etal., 2016), κακόβουλο λογισμικό (Wan etal., 2017), επιθέσεις σε περιβάλλοντα δικτύωσης σειριακού τύπου (Han etal., 2018) και ούτω καθεξής. Περιλαμβάνονται δηλαδή έρευνες που αποσκοπούν στις σύγχρονες λύσεις DRL για την ασφάλεια στον κυβερνοχώρο, οι οποίες κυμαίνονται από μεθόδους άμυνας για κυβερνοφυσικά συστήματα έως αυτόνομες προσεγγίσεις ανίχνευσης εισβολής και λύσεις βασισμένες στη θεωρία παιγνίων.



Σχήμα 6. Η αρχιτεκτονική του αντιπάλου αλγορίθμου DRL για ισχυρό έλεγχο αυτόνομου οχήματος (AV). Ένα βαθύ νευρωνικό δίκτυο (DNN) αποτελείται από μια μακροπρόθεσμη βραχυπρόθεσμη μνήμη (LSTM), ένα πλήρως συνδεδεμένο επίπεδο (FCL) και παλινδρόμηση (Regression) και χρησιμοποιείται για την εκμάθηση μακροπρόθεσμων εξαρτήσεων μέσα σε μεγάλα σύνολα δεδομένων, τα οποία περιέχουν τα αποτελέσματα των παικτών προηγούμενων αλληλεπιδράσεων. Πηγή: (Nguyen&Reddi, 2019).

Στρατηγικές για συστήματα επικοινωνίας που βασίζονται σε DL

Ο αυτόματος κωδικοποιητής που βασίζεται σε βαθιά μάθηση (DL) είναι μια πολλά υποσχόμενη αρχιτεκτονική για την υλοποίηση συστημάτων επικοινωνίας από άκρο σε άκρο. Ένα θεμελιώδες πρόβλημα τέτοιων συστημάτων είναι ο τρόπος με τον οποίο μπορεί να αυξηθεί ο ρυθμός μετάδοσης. Προτείνονται δύο νέα συστήματα για την αντιμετώπιση του ζητήματος του περιορισμένου ρυθμού δεδομένων: το σύστημα προσαρμοστικής μετάδοσης και το σύστημα γενικευμένης αναπαράστασης δεδομένων (Generalized Data Representation - GDR). Στο πρώτο σύστημα, έχει σχεδιαστεί μια προσαρμοστική μετάδοση, ούτως ώστε να επιλέγει τα διανύσματα μετάδοσης για τη μεγιστοποίηση του ρυθμού δεδομένων υπό διαφορετικές συνθήκες καναλιού. Το ποσοστό εσφαλμένων μπλοκ (Block Error Rate - BLER) του πρώτου συστήματος είναι κατά 80% χαμηλότερο από αυτό του συμβατικού συστήματος διανύσματος one-hot (one-hotvector). Αυτό σημαίνει ότι μπορεί να επιτευχθεί ένας υψηλότερος ρυθμός μετάδοσης δεδομένων με το προσαρμοστικό σχήμα μετάδοσης. Στο δεύτερο σύστημα, η GDR αντικαθιστά τη συμβατική

αναπαράσταση ενός one-hot. Το σύστημα GDR μπορεί να επιτύχει υψηλότερο ρυθμό δεδομένων σε σχέση με το συμβατικό σύστημα διανύσματος one-hot, με μια συγκρίσιμη απόδοση ενός BLER. Επιπλέον, το κοινό καθεστώς των δύο προτεινόμενων συστημάτων μπορεί να δημιουργήσει περαιτέρω οφέλη. Η επίδραση του λόγου σήματος προς θόρυβο (SNR) αναλύεται για αυτά τα συστήματα επικοινωνίας που βασίζονται σε DL και τα αριθμητικά αποτελέσματα δείχνουν ότι η εκπαίδευση του αυτόματου κωδικοποιητή χρησιμοποιώντας ένα σύνολο δεδομένων με διάφορες τιμές SNR, μπορεί να επιτύχει ισχυρή απόδοση BLER υπό διαφορετικές συνθήκες καναλιού.

Για να μπορέσει να ικανοποιήσει την αυξανόμενη ζήτηση για διάφορες εφαρμογές και υπηρεσίες επικοινωνίας, το δίκτυο επόμενης γενιάς πρέπει να προσφέρει βελτιωμένη ευρυζωνική σύνδεση κινητής τηλεφωνίας, εξαιρετικά αξιόπιστες και χαμηλής καθυστέρησης επικοινωνίες και τεράστια οικοσυστήματα Διαδικτύου των Πραγμάτων (IoT). Ένα πρωταρχικό μέλημα είναι να αντιμετωπιστεί η εκθετική αύξηση του αριθμού των εξοπλισμών των χρηστών και της χωρητικότητας της κίνησης στα μελλοντικά συστήματα επικοινωνίας. Ως εκ τούτου, έχουν προταθεί αρκετές υποσχόμενες τεχνολογίες, οι οποίες περιλαμβάνουν μαζικές μεταδόσεις πολλαπλών εισόδων και πολλαπλών εξόδων (massive multi-input and multi-output - MIMO), επικοινωνίες κυμάτων χιλιοστών, εξαιρετικά πυκνά δίκτυα και μη ορθογώνια πολλαπλή πρόσβαση. Για αυτά τα συμβατικά συστήματα επικοινωνίας υπάρχουν ορισμένοι περιορισμοί, όπως μη διαθέσιμες πληροφορίες κατάστασης καναλιού σε σύνθετο σενάριο μετάδοσης, υψηλή πολυπλοκότητα στην επεξεργασία μεγάλων δεδομένων και υποβέλτιστη απόδοση που προκαλείται από τη συμβατική δομή μπλοκ. Για αυτούς τους λόγους και με τη σημαντική ανάπτυξη της βαθιάς μάθησης, μερικοί ερευνητές όπως οι (Yu&Deng, 2010) και (Jiang etal., 2016) έχουν εφαρμόσει τη μηχανική μάθηση (ML) και ειδικότερα τις τεχνολογίες DL, προκειμένου να σχεδιάσουν συστήματα επικοινωνίας για οφέλη που δεν μπορούν να αποκτηθούν χρησιμοποιώντας τις συμβατικές προσεγγίσεις.

Ως μια πολλά υποσχόμενη τεχνική, η βαθιά μάθηση υλοποιεί συστήματα επικοινωνίας χρησιμοποιώντας βαθιά νευρωνικά δίκτυα (DNNs). Όντας διαφορετικό από το συμβατικό σύστημα επικοινωνίας που αποτελείται από πολλαπλά ανεξάρτητα μπλοκ (π.χ. κωδικοποίηση πηγής/καναλιού, διαμόρφωση, εκτίμηση καναλιού, εξισορρόπηση), το σύστημα επικοινωνίας που βασίζεται σε DL μπορεί να βελτιστοποιήσει από κοινού τον πομπό και τον δέκτη για καλύτερη απόδοση από άκρο σε άκρο και χωρίς δομή μπλοκ. Ο σχεδιασμός συστήματος που βασίζεται σε DL είναι πολλά υποσχόμενος για τους ακόλουθους λόγους: (1) Ένα σύστημα επικοινωνίας που βασίζεται σε DL μπορεί να βελτιστοποιηθεί για απόδοση από άκρο σε άκρο χρησιμοποιώντας

DNNs, τα οποία διαφέρουν θεμελιωδώς από τη δομή μπλοκ στα συμβατικά συστήματα επικοινωνίας. (2) Ένα σύστημα επικοινωνίας που βασίζεται σε DL μπορεί να βελτιστοποιηθεί για ένα πρακτικό σύστημα σε οποιονδήποτε τύπο καναλιού, χωρίς να απαιτείται ένα μαθηματικό μοντέλο με δυνατότητα μεταφοράς. Αυτό περιλαμβάνει τα μοντέλα καναλιών που λαμβάνουν υπόψη διαφορετικά σενάρια μετάδοσης και μη γραμμικότητες. (3) Οι αλγόριθμοι DL μπορούν να παρέχουν πιο γρήγορη ταχύτητα επεξεργασίας συγκριτικά με τους συμβατικούς αλγόριθμους επικοινωνίας, καθώς η εκτέλεση των NNs μπορεί να είναι παράλληλη σε σύγχρονες αρχιτεκτονικές και μπορεί να υλοποιηθεί χρησιμοποιώντας τύπους δεδομένων χαμηλής ακρίβειας.

Ελκυσμένοι από αυτά τα πλεονεκτήματα, αρκετοί ερευνητές όπως οι (O'shea&Hoydis, 2017) και (Dorner etal., 2017) έχουν παρουσιάσει πληθώρα μελετών σχετικά με τις επικοινωνίες που βασίζονται σε DL και την επεξεργασία σήματος με χρήση εργαλείων και υλικού τελευταίας τεχνολογίας. Η μέθοδος DL χρησιμοποιείται για την αντιμετώπιση ορισμένων προκλήσεων στα υπάρχοντα συστήματα επικοινωνίας. Για παράδειγμα, ο αλγόριθμος διάδοσης πεποιθήσεων που βασίζεται σε DL χρησιμοποιήθηκε αρχικά για τη βελτίωση των επιδόσεων της αποκωδικοποίησης καναλιών, όπου ελήφθησαν χαμηλή πολυπλοκότητα και σχεδόν βέλτιστη απόδοση αποκωδικοποιητή. Περίπου την ίδια εποχή, αναπτύχθηκε ο αυτόματος κωδικοποιητής για να αντιμετωπίσει το πρόβλημα της εκμάθησης ενός αποτελεσματικού φυσικού στρώματος. Στη θεωρία DL, ένας αυτόματος κωδικοποιητής περιγράφει ένα βαθύ NN για να βρει μια χαμηλών διαστάσεων αναπαράσταση της εισόδου του, σε ένα συγκεκριμένο ενδιάμεσο στρώμα που επιτρέπει την ανακατασκευή στην έξοδο με ελάχιστο ποσοστό σφάλματος. Το σύστημα επικοινωνίας που βασίζεται σε DL μπορεί να αναπαρασταθεί και να υλοποιηθεί από έναν αυτόματο κωδικοποιητή που εκπαιδεύεται, χρησιμοποιώντας το σύνολο δεδομένων εκτός σύνδεσης. Στη συνέχεια, ο εκπαιδευμένος αυτόματος κωδικοποιητής μπορεί να εφαρμοστεί απευθείας σε πρακτικά συστήματα που είναι συνδεδεμένα στο δίκτυο. Ένα σύστημα επικοινωνίας που βασίζεται σε DL (ερμηνεύεται ως αυτόματος κωδικοποιητής), εκτελεί μια εργασία ανακατασκευής από άκρο σε άκρο που βελτιστοποιεί από κοινού τον πομπό και τον δέκτη και μαθαίνει την κωδικοποίηση σήματος. Για την αντιμετώπιση των προκλήσεων του συγχρονισμού πλαισίων, προτάθηκε ένας αυτόματος κωδικοποιητής προκειμένου να αντιπροσωπεύει ένα πλήρες σύστημα επικοινωνίας. Από την άλλη πλευρά, μία συγκρίσιμη απόδοση μπορεί να επιτευχθεί ακόμη και χωρίς εκτεταμένο συντονισμό υπερπαραμέτρων. Πιο πρόσφατα, χρησιμοποιήθηκε ένας αλγόριθμος βασισμένος σε DL για την επίλυση προβλημάτων ανατροφοδότησης σε πληροφορίες κατάστασης καναλιού και εκτίμησης καναλιών σε τεράστια συστήματα MIMO. Ο συγκεκριμένος αλγόριθμος ξεπερνά σε επίδοση τους σύγχρονους αλγόριθμους που βασίζονται σε ανίχνευση

συμπίεσης (Wen etal., 2018). Στη συνέχεια αναλύεται ένας αυτόματος κωδικοποιητής που βασίζεται σε DL για ένα σύστημα επικοινωνίας από άκρο σε άκρο.

Αυτόματος κωδικοποιητής για συστήματα επικοινωνίας από άκρο σε άκρο

Εδώ, παρουσιάζεται ένα σύστημα επικοινωνίας που βασίζεται σε DL, το οποίο αναπαρίσταται ως ένας αυτόματος κωδικοποιητής που αποτελείται από πομπό, κανάλι και δέκτη με την αντίστοιχη δομή NN. Ο αυτόματος κωδικοποιητής περιγράφει ένα βαθύ NN που εφαρμόζει μάθηση χωρίς επίβλεψη, προκειμένου να ανακατασκευάσει την είσοδο στην έξοδο. Στον πομπό, ένα μήνυμα $s \in \{1,2,\dots,M\}$ μετατρέπεται πρώτα σε διάνυσμα $\mathrm{s} \in \mathbb{R}^M$ μετά την επεξεργασία διανυσματικής έκφρασης, όπου π.χ. $M \in \{4,8,16,32,64\}$. Για παράδειγμα, εάν μεταδίδεται το μήνυμα $s = 2$, η αντίστοιχη διανυσματική έκφραση είναι ένα ενιαίο διάνυσμα one-hot με $\mathrm{s} = [0,1,0,\dots,0]^T$ σε ένα συμβατικό σύστημα επικοινωνίας που βασίζεται σε DL. Στη συνέχεια, τα πολλαπλά πυκνά στρώματα, συμπεριλαμβανομένου ενός στρώματος ανορθωμένης γραμμικής μονάδας (Rectified Linear Unit - ReLU) και ενός γραμμικού στρώματος, εφαρμόζουν τον μετασχηματισμό $f_t: \mathbb{R}^M \mapsto \mathbb{R}^n$ για την παραγωγή του μεταδιδόμενου σήματος για n διακριτές χρήσεις καναλιών. Τέλος, το επίπεδο κανονικοποίησης διασφαλίζει τον περιορισμό ισχύος του μεταδιδόμενου σήματος $\mathrm{x} = [x_1,\dots,x_n]^T$ ως $\mathbb{E}\{x_j^2\} \leq 1 (j = 1,\dots,n)$, όπου το $\mathbb{E}\{\cdot\}$ υποδηλώνει προσδοκία. Το κανάλι εκπομπής υλοποιείται από ένα στρώμα θορύβου, με την έξοδο του να είναι το λαμβανόμενο σήμα y που δίνεται μέσω της εξίσωσης

$$y = x + n \qquad (Εξίσωση\ 1)$$

όπου $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ υποδηλώνει ένα μηδενικό μέσο και προσθετικό διάνυσμα λευκού Gaussian θορύβου (Additive White Gaussian Noise - AWGN) όπου κάθε στοιχείο έχει διακύμανση $\sigma^2 = (2RE_b/N_0)^{-1}$ και όπου R είναι ο ρυθμός δεδομένων, E_b είναι η ενέργεια ανά bit, και το N_0 υποδηλώνει τη φασματική πυκνότητα ισχύος θορύβου. Σημειωτέον, δεν υπάρχει σύνθετη λειτουργία στις υπάρχουσες αρχιτεκτονικές NN και ο μιγαδικός αριθμός (Complex number) αντιπροσωπεύεται από δύο πραγματικούς αριθμούς. Συνεπώς, υποτίθεται ότι όλοι οι συντελεστές καναλιού έχουν πραγματικές τιμές. Επιπλέον, το σύστημα επικοινωνίας που αντιπροσωπεύεται από τον αυτόματο κωδικοποιητή είναι κατάλληλο για κάθε τύπο καναλιού χωρίς μαθηματικό μοντέλο [2]. Δηλαδή, ο αυτόματος κωδικοποιητής μπορεί να εφαρμοστεί σε οποιονδήποτε τύπο μοντέλου καναλιού, εφόσον είναι διαθέσιμα πραγματικά σύνολα δεδομένων για εκπαίδευση και εκμάθηση.

Στον δέκτη, το λαμβανόμενο σήμα $\mathbf{y}$ διέρχεται από το στρώμα ReLU [3] για να πραγματοποιήσει τον μετασχηματισμό $f_r: \mathbb{R}^n \mapsto \mathbb{R}^M$. Το τελευταίο στρώμα του δέκτη έχει μια ενεργοποίηση

softmax, η οποία είναι μια γενίκευση της λογιστικής συνάρτησης που συμπιέζει ένα διάνυσμα διαστάσεων M-αυθαίρετων πραγματικών τιμών σε ένα διάνυσμα πιθανότητας M-διαστάσεων $\mathrm{p} = [p_1, \dots, p_M]^T$ όπου κάθε στοιχείο $p_i (i = 1,2, \dots, M)$ βρίσκεται στην περιοχή $(0,1]$ και όλα τα στοιχεία αθροίζονται σε ένα [4]. Για το συμβατικό σχήμα αυτόματου κωδικοποιητή, το εκτιμώμενο μήνυμα $\hat{s}$ προκύπτει από τον δείκτη του στοιχείου που έχει τη μεγαλύτερη πιθανότητα στο διάνυσμα**p**. Εδώ, το BLER των συστημάτων επικοινωνίας που βασίζονται σε DL ορίζεται ως

$$\text{BLER} = \frac{1}{M} \sum_{\mathrm{g}} \Pr(\hat{s} \neq s) . \qquad (Εξίσωση 2)$$

Το BLER ισούται με το ποσοστό σφάλματος συμβόλων (Symbol Error Rate - SER) του συστήματος επικοινωνίας που βασίζεται σε DL.

Τέλος, το σύστημα επικοινωνίας που βασίζεται σε αυτόματο κωδικοποιητή μπορεί να εκπαιδευτεί εκτός σύνδεσης χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων εκπαίδευσης, ενώ η επαναληπτική διαδικασία εκπαίδευσης εξαρτάται από την τιμή της συνάρτησης απώλειας σε κάθε επανάληψη. Οι πιο κοινές συναρτήσεις απώλειας είναι το μέσο τετράγωνο σφάλμα (Mean Squared Error – MSE) και η κατηγορική διασταυρούμενη εντροπία. Αυτές οι συναρτήσεις απώλειας καθορίζονται από την έκφραση του διανύσματος **s** και το διάνυσμα πιθανότητας **p**. Οι παράμετροι εκπαίδευσης του αυτόματου κωδικοποιητή παράγονται για την ελαχιστοποίηση της λειτουργίας απώλειας, ενώ ο εκπαιδευμένος αυτόματος κωδικοποιητής με τις σταθερές παραμέτρους NN εφαρμόζεται σε πρακτικά σενάρια επικοινωνίας στο διαδίκτυο.

Αξιοποίηση της DRL σε ασύρματα δίκτυα επεξεργασίας δεδομένων

Η συλλογή, η ροή και η επεξεργασία δεδομένων αποτελούν βασικές λειτουργίες για τα σύγχρονα ασύρματα δίκτυα λόγω της ταχείας ανάπτυξης σε τομείς όπως το Διαδίκτυο των Πραγμάτων (IoT), τα δίκτυα αισθητήρων, η ηλεκτρονική ανάλυση δεδομένων και το πρότυπο κατανεμημένων υπολογιστών Edgecomputing. Σε τέτοιες εφαρμογές, τεράστιοι όγκοι δεδομένων που συλλέγονται (και αποθηκεύονται) από το δίκτυο πρέπει να μεταφερθούν σε πολλαπλές μονάδες επεξεργασίας για τον έγκαιρο υπολογισμό των επιθυμητών αποτελεσμάτων. Ένα παράδειγμα αυτού του τύπου εφαρμογής είναι η έξυπνη μεταφορά, όπου μεγάλοι όγκοι δεδομένων αισθητήρων και βίντεο καταγράφονται από διάφορα σημεία παρακολούθησης και στη συνέχεια μεταφέρονται κατ' απαίτηση (μέσω ασύρματων δικτύων) σε κατανεμημένους κόμβους υπολογιστών, για εφαρμογές που κυμαίνονται από την αναγνώριση οχήματος έως την ανάλυση της κυκλοφορίας. Λόγω της συνεχώς αυξανόμενης κίνησης στα Δίκτυα Επεξεργασίας Δεδομένων (Data Processing Networks - DPN), βελτιώνουν συχνά την απόδοση αποθηκεύοντας δεδομένα σε κρυφή μνήμη και σε

κόμβους, κοντά σε υπολογιστικές μονάδες και παραδίδοντας γρήγορα αυτά τα δεδομένα προς κατανάλωση.

Ωστόσο, ο σχεδιασμός των πολιτικών προσωρινής αποθήκευσης επικεντρώνεται κυρίως στα πιθανά κέρδη στην απόδοση του δικτύου, π.χ., αναλογία επιτυχίας της προσωρινής μνήμης και καθυστέρηση λήψης, ενώ παραβλέπεται ο αντίκτυπος στην επεξεργασία και την κατανάλωση δεδομένων. Αυτές οι πολιτικές περιλαμβάνουν τις στρατηγικές του λιγότερο συχνά χρησιμοποιήσιμου (Least Frequently Used - LFU) και του πιο δημοφιλούς αντικειμένου (Most Popular Object), οι οποίες επιτυγχάνουν τόσο μια υψηλή αναλογία επιτυχίας της προσωρινής μνήμης, όσο και τις στρατηγικές LRU (Least Recently Used), qLRU και kLRU που χρησιμοποιούν ένα πρόσφατο αίτημα για ενημέρωση προσωρινής μνήμης. Όμως, για τα δίκτυα επεξεργασίας δεδομένων, συλλέγεται και καταναλώνεται ταυτόχρονα μια ροή δεδομένων. Οι υπάρχουσες πολιτικές αποθήκευσης κρυφής μνήμης που λαμβάνουν υπόψη μόνο την απόδοση του δικτύου κατά την παράδοση δεδομένων, υπολείπονται στην αντιμετώπιση του χρόνου επεξεργασίας πλειάδας από άκρο σε άκρο, ο οποίος εξαρτάται τόσο από την παράδοση όσο και από την κατανάλωση δεδομένων. Στην πράξη, όταν το διαθέσιμο εύρος ζώνης δικτύου δεν μπορεί να υποστηρίξει πλήρως όλες τις ανάγκες επεξεργασίας δεδομένων, οι καθυστερήσεις κατά την επεξεργασία δεδομένων γίνονται αναπόφευκτες. Ως αποτέλεσμα, ο σχεδιασμός των πολιτικών προσωρινής αποθήκευσης θα πρέπει να λαμβάνει υπόψη τις ανάγκες κατανάλωσης δεδομένων, για τον μετριασμό των καθυστερήσεων επεξεργασίας και την εκπλήρωση των απαιτήσεων έγκαιρης ηλεκτρονικής επεξεργασίας.

Στην έρευνα των συγγραφέων (Wang etal., 2019) προτείνεται ένα νέο πλαίσιο, το DeepChunk, το οποίο αξιοποιεί τη βαθιά ενισχυτική μάθηση για προσωρινή αποθήκευση, σε ένα Δίκτυο Επεξεργασίας Δεδομένων (DPN). Η βασική ιδέα είναι ότι οι πολιτικές κρυφής μνήμης πρέπει να βελτιστοποιούνται για (1) την απόδοση του δικτύου κατά την παράδοση δεδομένων και (2) για την αποτελεσματικότητα επεξεργασίας κατά την κατανάλωση δεδομένων. Δεδομένου ότι η κλασική μέτρηση αναλογίας επισκέψεων δεν λαμβάνει υπόψη τα μερικά αρχεία στην κρυφή μνήμη, οι συγγραφείς χρησιμοποιούν την έννοια της αναλογίας τμημάτων επισκέψεων που εξηγεί την ύπαρξη μερικών αρχείων στη κρυφή μνήμη.

5

Βιβλιογραφική Ανασκόπηση Εφαρμογών μη Επανδρωμένων Αεροσκαφών με Μεθόδους DRL

Αρχικά, ένα μη επανδρωμένο αεροσκάφος (UAV) ή αλλιώς drone, ονομάζεται το κάθε είδος ιπτάμενου οχήματος που δεν έχει χειριστή, πλήρωμα ή επιβάτες, αλλά μπορεί να πραγματοποιεί πτήσεις είτε αυτόνομα είτε μέσω τηλεκατεύθυνσης. Σε έρευνά τους, οι συγγραφείς (Li etal., 2019) επικεντρώνονται στην τροχιά αυτού του τύπου αεροσκαφών. Ο συνεχής έλεγχος της τροχιάς των UAV σταθερής πτέρυγας είναι περίπλοκος, όταν εξετάζεται η κρυφή δυναμική. Λόγω των πολλαπλών βαθμών ελευθερίας των UAV, οι μεθοδολογίες παρακολούθησης που βασίζονται στη συμβατική θεωρία ελέγχου, όπως το Αναλογικό-Ολοκληρωτικό-Παράγωγο (Proportional Integral Derivative - PID), έχουν περιορισμούς στο χρόνο απόκρισης και την ευρωστία προσαρμογής, ενώ σε μια προσέγγιση βασισμένη σε μοντέλο που υπολογίζει τη δύναμη και τις ροπές με βάση το ρεύμα του UAV, η κατάσταση είναι περίπλοκη και άκαμπτη. Λόγω των παραπάνω δυσκολιών, οι συγγραφείς (Li etal., 2019), παρουσιάζουν ένα πλαίσιο ενισχυμένης μάθησης χρησιμοποιώντας την τεχνική ηθοποιού-κριτικού (actor-critic) που ελέγχει την τροχιά των UAV μέσω ενός συνόλου επιθυμητών σημείων πορείας. Με αυτό τον τρόπο δημιουργείται ένα βαθύ νευρωνικό δίκτυο με σκοπό να γίνει γνωστή η βέλτιστη πολιτική παρακολούθησης και παράλληλα αναπτύσσεται ενισχυτική μάθηση για τη βελτιστοποίηση του προκύπτοντος συστήματος παρακολούθησης. Έπειτα, τα πειραματικά αποτελέσματα που προκύπτουν από την προτεινόμενη προσέγγισή τους, δείχνουν ότι μπορεί να επιτύχει 58,14% μικρότερο σφάλμα θέσης, 21,77% λιγότερη κατανάλωση ισχύος συστήματος και 9,23% ταχύτερη επίτευξη από την αρχική γραμμή. Τέλος, και με βάση τη

θεωρία του ηθοποιού-δικτύου (Actor–network theory), το δίκτυο ηθοποιού αποτελείται μόνο από γραμμικές λειτουργίες, επομένως η επιτάχυνση υλικού που βασίζεται σε συστοιχίες πύλης προγραμματιζόμενου πεδίου (Field Programmable Gate Arrays - FPGA) μπορεί να σχεδιαστεί εύκολα για ενεργειακά αποδοτικό έλεγχο σε πραγματικό χρόνο.

Πρόσφατα, οι εφαρμογές των UAV έχουν χρησιμοποιηθεί ευρέως σε πολυάριθμες εφαρμογές του πραγματικού κόσμου όπου οι ανθρώπινες δραστηριότητες είναι περιορισμένες. Με την αύξηση του όγκου δεδομένων και των απαιτήσεων ακρίβειας για πρακτικές εφαρμογές, οι αξιόπιστες λειτουργίες των UAV, δηλαδή η σταθερή αυτόνομη καθοδήγηση και έλεγχος, θεωρήθηκαν ως μία από τις πιο κρίσιμες. Οι αποτελεσματικοί αλγόριθμοι παρακολούθησης επιτρέπουν μια ομαλή τροχιά και συνεπώς χαμηλότερη κατανάλωση ισχύος/ενέργειας του συστήματος κατά τη διάρκεια της πτήσης. Παραδοσιακά, ο μηχανισμός ελέγχου PID είναι η πιο σύγχρονη επιλογή για το βιομηχανικό σύστημα παρακολούθησης τροχιών UAV. Ως αποτέλεσμα, οι ελεγκτές PID είναι εύκολο να εφαρμοστούν στο FPGA και προσφέρουν επάρκεια για πολλά προβλήματα ελέγχου. Επιπλέον, λειτουργούν καλά σε περιπτώσεις όπου η δυναμική της διαδικασίας είναι καλοήθης και οι απαιτήσεις της απόδοσης μέτριες. Ωστόσο, ο ελεγκτής PID δεν μπορεί να αντιμετωπίσει αποτελεσματικά τις διαδικασίες με μεγάλη χρονική καθυστέρηση και έχει κακή απόδοση ως προς την παρακολούθηση προβλημάτων που απαιτούν επιθετικές δυναμικές διαμορφώσεις, συμπεριλαμβανομένης της αβέβαιης αντιστάθμισης εσωτερικών διαταραχών και την ανάκτηση ανισορροπιών. Για την αντιμετώπιση αυτού του προβλήματος, έχουν εφαρμοστεί για ορισμένες εφαρμογές, τροποποιημένα μοντέλα PID που βελτιώνουν την απόδοση. Εν τω μεταξύ, ο έλεγχος της σταθερότητας των UAV, αποτελεί μια μεγάλη πρόκληση σε περιπτώσεις που χρησιμοποιούνται πλατφόρμες χαμηλού κόστους ισχύος. Ο κύριος λόγος είναι ότι είναι δύσκολο να αποκτηθεί ένα μαθηματικό μοντέλο υψηλής πιστότητας ενός UAV, το οποίο έχει υπολειτουργικό σύστημα με μη γραμμική δυναμική. Για τη βελτίωση της σταθερότητας και του ελέγχου σε πραγματικό χρόνο, εισάγονται βαθιά νευρωνικά δίκτυα (DNNs) ενσωματωμένα σε διαφορετικές πλατφόρμες υλικού. Μέσω εκπαίδευσης μεγάλων δεδομένων, το σύστημα ελέγχου που βασίζεται σε DNN, επιτυγχάνει προσαρμοστικότητα και στιβαρότητα που εγγυώνται τη σταθερότητα της πτήσης. Επιπλέον, οι ελεγκτές είναι σε θέση να ακολουθήσουν την επιθυμητή τροχιά με την ανοχή σε απροσδόκητες διαταραχές. Παρόμοια με τους ελεγκτές PID, ο ελεγκτής που βασίζεται σε DNN εκτιμά τις ενέργειες ελέγχου με βάση την εμπειρία των προηγούμενων πτήσεων για τη μείωση των στιγμιαίων ατελειών παρακολούθησης. Λόγω του ότι κανένας από αυτούς τους ελεγκτές δεν εξετάζει το πώς ακριβώς η επιλεγμένη ενέργεια θα επηρεάσει τις επακόλουθες ανταμοιβές, είναι πολύ πιθανό να παράγουν μη βέλτιστες λύσεις.

Η ενισχυτική μάθηση (RL) παρέχει ένα μαθηματικό πλαίσιο για τη μάθηση ή τη δημιουργία πολιτικών, που χαρτογραφούν τις καταστάσεις σε ενέργειες, με στόχο τη μεγιστοποίηση μιας σωρευτικής ανταμοιβής. Σε αντίθεση με την εποπτευόμενη μάθηση, στην RL ο πράκτορας (δηλ. ο μαθητής) μαθαίνει την πολιτική λήψης αποφάσεων μέσω αλληλεπιδράσεων με το περιβάλλον. Ο στόχος του πράκτορα είναι να μεγιστοποιήσει τη σωρευτική μακροπρόθεσμη ανταμοιβή, πραγματοποιώντας την κατάλληλη δράση σε κάθε βήμα ανάλογα με την τρέχουσα κατάσταση του περιβάλλοντος. Πρέπει όμως να λάβει υπόψη του και την αντιστάθμιση μεταξύ εξερευνήσεων και εκμεταλλεύσεων. Η χρήση του αλγορίθμου Q-learning είναι μία από τις στρατηγικές της RL χωρίς μοντέλο, που αποθηκεύει ζεύγη πεπερασμένης κατάστασης δράσης και αντίστοιχες τιμές Q σε πίνακα αναζήτησης. Επίσης, έχει εφαρμοστεί για θερμική και ενεργειακή διαχείριση σε αυτόνομα υπολογιστικά συστήματα. Ο συνδυασμός της συμβατικής Q-learning και του βαθιού νευρωνικού δικτύου, παρέχει μια σημαντική ανακάλυψη στη βαθιά ενισχυτική μάθηση (DRL). Ωστόσο, το νευρωνικό δίκτυο του αλγορίθμου DQN, πρέπει να συγκεντρώνει αρκετά δείγματα τιμών. Επίσης, τα δεδομένα που απαιτούνται για την εκπαίδευσή του, μπορούν είτε να προέρχονται από προσομοίωση βάσει μοντέλου είτε από πραγματική μέτρηση. Έχοντας αρχικά αναπτυχθεί από τη θυγατρική εταιρεία DeepMind, η DRL παρέχει μια πολλά υποσχόμενη προσαρμοστική τεχνική για τη διαχείριση δεδομένων, στο χειρισμό μεγάλου χώρου με πολύπλοκα προβλήματα ελέγχου. Η βαθιά ενισχυτική μάθηση της τεχνικής ηθοποιού-κριτικού, έχει καταφέρει να ξεπεράσει τις δυσκολίες στις πολιτικές ελέγχου εκμάθησης συστημάτων. Αυτό επιτεύχθηκε με συνεχή κατάσταση και χώρο δράσης, παρέχοντας έτσι μια πιθανή λύση για αποτελεσματικό έλεγχο αποστολών σε αυτόνομα UAV σε πραγματικό χρόνο. Στη συνέχεια αναλύεται η δομή της βαθιάς ενισχυτικής μάθησης και αναφέρονται ορισμένες παράμετροι της βαθιάς μάθησης, όπως είναι ο έλεγχος κίνησης και ο προγραμματισμός και επίγνωση της κατάστασης.

5.1 Αξιοποίηση της DRL για τον έλεγχο των UAVs

Τα μη επανδρωμένα εναέρια οχήματα (UAVs) μπορούν να χρησιμοποιηθούν ως εναέριοι σταθμοί βάσης (BSs) για να βελτιώσουν τόσο την κάλυψη όσο και την απόδοση των δικτύων επικοινωνίας σε διάφορα σενάρια, όπως επικοινωνίες έκτακτης ανάγκης και πρόσβαση στο δίκτυο για απομακρυσμένες περιοχές. Όταν τα δίκτυα επικοινωνίας διακόπτονται από μια ολέθρια φυσική καταστροφή, τα κινητά UAVs μπορούν να αναπτυχθούν γρήγορα, ούτως ώστε να δημιουργήσουν αποτελεσματικές συνδέσεις επικοινωνίας για τους χρήστες εδάφους, με τελικό στόχο την παράδοση πακέτων.

Προκειμένου να παρέχεται μία μακροπρόθεσμη αποτελεσματική επικοινωνιακή κάλυψη, τα UAVs με υψηλό βαθμό κινητικότητας, πρέπει να λειτουργούν αυτόνομα ως ομάδα. Σε ένα τέτοιο δίκτυο UAV, τα UAVs μπορούν να λειτουργήσουν ως BSs για να παρέχουν συνδέσμους επικοινωνίας για χρήστες εδάφους που χρησιμοποιούν τρέχουσες ασύρματες τεχνολογίες, όπως WiFi ή LTE. Ένα ή ένας μικρός αριθμός UAVs έχει συνδέσεις μεγάλων αποστάσεων (όπως δορυφορικές συνδέσεις) με εξωτερικά δίκτυα (όπως το Διαδίκτυο), τα οποία ονομάζονται πύλες. Αυτό το έργο είναι αρκετά δύσκολο, επειδή τα UAVs έχουν πολύ περιορισμένο εύρος επικοινωνίας και ενεργειακούς πόρους. Επίσης, ένα δίκτυο UAV έχει συνήθως πολύ περιορισμένο αριθμό πυλών. Πρώτον, λόγω του περιορισμένου εύρους επικοινωνίας και του σχετικά υψηλού κόστους (πολλές χιλιάδες δολάρια για κάθε εμπορικό UAV), είναι αδύνατο να υπάρχουν επαρκή UAVs για την συνεχή κάλυψη μιας μεγάλης περιοχής στόχου. Επομένως, τα UAVs πρέπει να μετακινούνται για να διασφαλίσουν ότι κάθε περιοχή καλύπτεται για ένα εύλογο χρονικό διάστημα. Επιπλέον, η επιείκεια είναι κρίσιμη για την κάλυψη της επικοινωνίας, καθώς δεν είναι επιθυμητό να καλύπτονται (τις περισσότερες φορές) συγκεκριμένες περιοχές, ενώ στις υπόλοιπες να μην παρέχεται το ανάλογο εύρος κάλυψης. Δεύτερον, λόγω των περιορισμένων ενεργειακών πόρων, ένα UAV δεν μπορεί να συνεχίσει να πετά για μεγάλο χρονικό διάστημα. Επομένως, πρέπει να λειτουργεί με έναν ενεργειακά αποδοτικό τρόπο για να παρατείνεται η διάρκεια ζωής του δικτύου. Επιπροσθέτως, λόγω του πολύ περιορισμένου αριθμού πυλών, ένα δίκτυο UAV πρέπει να διατηρείται συνεχώς συνδεδεμένο. Διαφορετικά, οι χρήστες γείωσης που σχετίζονται με έναν αποσυνδεδεμένο κόμβο (ο οποίος δεν αποτελεί πύλη), μπορούν να χάσουν τις συνδέσεις τους με το εξωτερικό δίκτυο.

Για την αντιμετώπιση των παραπάνω ζητημάτων, οι συγγραφείς (Liu etal., 2018) προτείνουν την αξιοποίηση της αναδυόμενης βαθιάς ενισχυτικής μάθησης (DRL), η οποία έχει πρόσφατα αποδειχθεί ότι προσφέρει ανώτερη απόδοση σε μερικές εργασίες παιγνίων. Θεωρούν επίσης ότι η DRL παρέχει μια πολλά υποσχόμενη λύση, επειδή μπορεί να χειριστεί έναν περίπλοκο χώρο κατάστασης και ένα περιβάλλον μεταβαλλόμενου χρόνου, χρησιμοποιώντας παράλληλα ισχυρά βαθιά νευρωνικά δίκτυα (DNNs) για να καθοδηγήσει τη λήψη αποφάσεων. Έχει αποδειχθεί ότι τα εν λόγω δίκτυα προσφέρουν κορυφαία απόδοση σε αρκετές μαθησιακές εργασίες με περιορισμένες έως και μηδενικές γνώσεις τομέα. Ωστόσο, δεν είναι εύκολο να λυθεί το πρόβλημα ελέγχου UAV χρησιμοποιώντας την DRL. Η βασική τεχνική της DRL, δηλαδή η βαθιά εκμάθηση Q (DQL), χρησιμοποιεί ένα βαθύ δίκτυο Q (DQN) για να εκτιμήσει την τιμή Q για κάθε ζεύγος κατάστασης-ενέργειας, το οποίο μπορεί να χειριστεί μόνο έναν πολύ περιορισμένο χώρο δράσης. Το πρόβλημα ελέγχου στην προκειμένη περίπτωση είναι ένα πρόβλημα συνεχούς ελέγχου με

απεριόριστο χώρο δράσης. Η συνήθως χρησιμοποιούμενη μέθοδος για συνεχή έλεγχο είναι η μέθοδος παίκτη-κριτικού (actor-critic method) (Sutton&Barto, 1998). Οι (Liu etal., 2018) επιλέγουν λοιπόν να χρησιμοποιήσουν μια υπερσύγχρονη μέθοδο παίκτη-κριτικού που ονομάζεται βαθιά ντετερμινιστική κλίση πολιτικής (DDPG), ως σημείο εκκίνησης για το σχέδιό τους. Σε αυτή τη περίπτωση, το πρόβλημα ελέγχου είναι πιο περίπλοκο σε σχέση με τα περισσότερα άλλα προβλήματα ελέγχου, καθώς περιλαμβάνει πολλαπλούς στόχους (δηλ. κάλυψη, επιείκεια και κατανάλωση ενέργειας) και περιορισμό στη συνδεσιμότητα δικτύου. Παρόλο που η DRL έχει σημειώσει αξιοσημείωτες επιτυχίες σε μερικές εργασίες παιγνίου, παραμένει άγνωστο εάν μπορεί να καταφέρει κάτι ανάλογο σε εργασίες ελέγχου που αφορούν πολύπλοκα δίκτυα επικοινωνίας. Αυτό συμβαίνει διότι αυτά τα δίκτυα έχουν συνήθως αρκετά διαφορετικούς στόχους, περιορισμούς και καταστάσεις, καθώς και χώρους δράσης.

Στην ενισχυτική μάθηση ένας πράκτορας ορίζεται να αλληλεπιδρά με ένα περιβάλλον αναζητώντας να βρει την καλύτερη δράση για κάθε κατάσταση σε οποιοδήποτε βήμα στο χρόνο. Ο πράκτορας πρέπει να εξισορροπήσει την εξερεύνηση και την εκμετάλλευση του κρατικού χώρου, προκειμένου να βρει τη βέλτιστη πολιτική που μεγιστοποιεί τη συσσωρευμένη ανταμοιβή από την αλληλεπίδραση με το περιβάλλον. Σε αυτό το πλαίσιο, ένας πράκτορας τροποποιεί τη συμπεριφορά ή την πολιτική του με επίγνωση των καταστάσεων, των ενεργειών και των ανταμοιβών για κάθε βήμα. Η ενισχυτική μάθηση συνθέτει μια διαδικασία βελτιστοποίησης σε ολόκληρο τον χώρο της κατάστασης, προκειμένου να μεγιστοποιηθεί η συσσωρευμένη ανταμοιβή. Τα ρομποτικά προβλήματα συχνά βασίζονται σε εργασίες με χρονική δομή. Επομένως, αυτοί οι τύποι προβλημάτων είναι κατάλληλοι για επίλυση μέσω ενός πλαισίου ενισχυτικής μάθησης. Η τυπική θεωρία ενισχυτικής μάθησης δηλώνει ότι ένας πράκτορας είναι σε θέση να αποκτήσει μια πολιτική, η οποία αντιστοιχεί σε κάθε κατάσταση $s \in \mathbb{S}$ σε μία δράση $a \in \mathbb{A}$, όπου $\mathbb{S}$ είναι ο χώρος κατάστασης (πιθανές καταστάσεις του πράκτορα στο περιβάλλον) και $\mathbb{A}$ είναι ο πεπερασμένος χώρος δράσης. Η εσωτερική δυναμική του παράγοντα αντιπροσωπεύεται από το μοντέλο πιθανότητας μετάβασης $p(s_{t+1} \mid s_t, a_t)$, τη στιγμή t. Η πολιτική μπορεί να είναι στοχαστική $\pi(a \mid s)$, με πιθανότητα να σχετίζεται με κάθε πιθανή ενέργεια ή ντετερμινιστική $\pi(s)$. Σε κάθε χρονικό βήμα, η πολιτική καθορίζει τη δράση που θα επιλεγεί και η ανταμοιβή $r(s_t, a_t)$ παρατηρείται από το περιβάλλον. Ο στόχος του πράκτορα είναι να μεγιστοποιήσει τη συσσωρευμένη έκπτωση ανταμοιβής $R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$ από κατάσταση σε χρόνο t, σε χρόνο $T (T = \infty$ για άπειρα προβλήματα ορίζοντα). Ο συντελεστής έκπτωσης γ ορίζεται για να κατανέμει διαφορετικά βάρη για τις μελλοντικές ανταμοιβές. Για μια συγκεκριμένη πολιτική, η συνάρτηση αξίας V^{π} στην παρακάτω εξίσωση (1), είναι μια αναπαράσταση της προσδοκίας της

συσσωρευμένης προεξοφλημένης ανταμοιβής R_t για κάθε κατάσταση $s \in \mathbb{S}$ (υποθέτοντας μια ντετερμινιστική πολιτική $\pi(s_t)$):

$$V^{\pi}(s_t) = \mathbb{E}[R_t \mid s_t, a_t = \pi(s_t)] \quad (Εξίσωση\ 1)$$

Ένα ισοδύναμο της συνάρτησης τιμής, αντιπροσωπεύεται από τη συνάρτηση ενέργειας-τιμής Q^{π} στην παρακάτω εξίσωση (2), για κάθε ζεύγος δράσης-κατάστασης (s_t, a_t):

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} \mid s_t, a_t) V^{\pi}(s_{t+1}) \quad (Εξίσωση\ 2)$$

Η βέλτιστη πολιτική π^* είναι αυτή που μεγιστοποιεί τη συνάρτηση τιμής (ή ισοδύναμα τη συνάρτηση τιμής ενέργειας), όπως στην ακόλουθη εξίσωση (3):

$$\pi^* = \arg\max_{\pi} \quad V^{\pi}(s_t) \quad (Εξίσωση\ 3)$$

Ένα γενικό πρόβλημα στις πραγματικές ρομποτικές εφαρμογές είναι ότι η κατάσταση και οι χώροι δράσης είναι συχνά συνεχείς χώροι. Μια συνεχής κατάσταση ή/και χώρος δράσης μπορεί να κάνει το πρόβλημα βελτιστοποίησης δυσεπίλυτο, λόγω του συντριπτικού συνόλου διαφορετικών καταστάσεων ή/και ενεργειών. Ως γενικό πλαίσιο για την αναπαράσταση, οι μέθοδοι ενίσχυσης της μάθησης (ή ενισχυτικής μάθησης) ενισχύονται μέσω της βαθιάς μάθησης για να βοηθήσουν τον σχεδιασμό για την αναπαράσταση χαρακτηριστικών, η οποία είναι γνωστή ως βαθιά ενισχυτική μάθηση. Η ενισχυτική μάθηση και ο βέλτιστος έλεγχος, στοχεύουν στην εύρεση της βέλτιστης πολιτικής π^* μέσω πολλών μεθόδων. Η βέλτιστη λύση μπορεί να αναζητηθεί σε αυτό το αρχικό πρόβλημα ή μέσω της διπλής διατύπωσης V^*, Q^*, η οποία μπορεί να αποτελεί τον στόχο βελτιστοποίησης. Στην έρευνα των (Carrio etal., 2017), οι μέθοδοι εκμάθησης βαθιάς ενίσχυσης χωρίζονται σε δύο κύριες κατηγορίες: τη συνάρτηση αξίας και τις μεθόδους αναζήτησης πολιτικής.

Μέθοδοι συνάρτησης αξίας. Αυτές οι μέθοδοι επιδιώκουν να βρουν τη βέλτιστη V^*, Q^*, από την οποία προέρχεται άμεσα η βέλτιστη πολιτική π^* στην εξίσωση (4). Οι προσεγγίσεις μάθησης βασίζονται στη βελτιστοποίηση της συνάρτησης τιμής δράσης Q, με βάση την εξίσωση βελτιστοποίησης Bellman. Για Q (βλ. εξίσωση (5)):

$$\pi^* = \arg\max_{a_t} \quad Q^*(s_t, a_t) \quad (Εξίσωση\ 4)$$

$$Q^*(s_t, a_t) = \mathbb{E}\left[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})\right] \quad (Εξίσωση\ 5)$$

Η μέθοδος DQN υπολογίζει τη συνάρτηση τιμής ενέργειας (βλ. εξίσωση (6)) μέσω ενός μοντέλου συνδυαστικού νευρωνικού δικτύου (CNN) με ένα σύνολο βαρών θ ως $Q^*(s,a) \approx Q(s,a;\theta)$:

$$Q_i^*(s_t,a_t) = y_i \\ = \mathbb{E}\left[r(s_t,a_t) + \gamma \underset{a_{t+1}}{max} Q(s_{t+1},a_{t+1};\theta_{i-1}) \mid s_t,a_t\right]. \quad (Εξίσωση\ 6)$$

Το CNN μπορεί να εκπαιδευτεί ελαχιστοποιώντας μια ακολουθία συναρτήσεων απώλειας $L_i(\theta_i)$, που βελτιστοποιούνται σε κάθε επανάληψη i, όπως φαίνεται στην ακόλουθη εξίσωση (7):

$$L_i(\theta_i) = \mathbb{E}\left[\left(y_i - Q(s_t,a_t;\theta_i)\right)^2\right] \quad (Εξίσωση\ 7)$$

Η κατάσταση s του αλγορίθμου DQN, είναι η ακατέργαστη εικόνα και έχει δοκιμαστεί ευρέως με παιχνίδια Atari, σύμφωνα με τη μελέτη των (31). Ο DQN δεν έχει σχεδιαστεί για συνεχείς εργασίες, με αποτέλεσμα αυτή η μέθοδος να αντιμετωπίζει δυσκολίες στην προσέγγιση ορισμένων προβλημάτων ρομποτικής που επιλύθηκαν προηγουμένως με συνεχή έλεγχο. Η συνεχής Q-learning με απλοποιημένες λειτουργίες πλεονεκτήματος (Normalized Advantage Functions - NAF) ξεπερνά αυτό το ζήτημα με τη χρήση ενός νευρωνικού δικτύου. Το δίκτυο αυτό βγάζει ξεχωριστά μια συνάρτηση τιμής $V(x)$ και έναν όρο πλεονεκτήματος $A(x,u)$, ο οποίος παραμετροποιείται ως τετραγωνική συνάρτηση με μη γραμμικά χαρακτηριστικά. Αυτές οι δύο συναρτήσεις συνθέτουν το τελικό $Q(x,u \mid \theta^Q)$ που δίνεται από την ακόλουθη εξίσωση (8):

$$Q(x,u \mid \theta^Q) = A(x,u \mid \theta^A) + V(x \mid \theta^V) \quad (Εξίσωση\ 8)$$

με x να είναι η κατάσταση, u να είναι η δράση, και θ^Q, θ^A και θ^V να είναι τα σύνολα βαρών των συναρτήσεων Q, A και Vαντίστοιχα. Αυτή η αναπαράσταση επιτρέπει την απλοποίηση πιο συνηθισμένων αλγορίθμων τύπου ηθοποιού-κριτικού, διατηρώντας παράλληλα τα οφέλη της προσέγγισης μη γραμμικής συνάρτησης τιμών. Οι NAF ισχύουν για εργασίες συνεχούς ελέγχου και εκμεταλλεύονται εκπαιδευμένα μοντέλα για να προσεγγίσουν την τυπική συνάρτηση τιμής χωρίς μοντέλο.

Μέθοδοι αναζήτησης πολιτικής. Οι μέθοδοι ενισχυτικής μάθησης που βασίζονται σε πολιτικές, στοχεύουν στην άμεση αναζήτηση της βέλτιστης πολιτικής π^*, η οποία παρέχει ένα εφικτό πλαίσιο για συνεχή έλεγχο. Η μέθοδος βαθιάς ντετερμινιστικής πολιτικής βαθμίδας (DDPG) βασίζεται στο πρότυπο της τεχνικής ηθοποιού-κριτικού, με δύο νευρωνικά δίκτυα να προσεγγίζουν μια άπληστη ντετερμινιστική πολιτική (ηθοποιός) και μία Q λειτουργία (κριτικός). Το δίκτυο ηθοποιών ενημερώνεται εφαρμόζοντας τον κανόνα αλυσίδας στην αναμενόμενη απόδοση από την αρχική διανομή J, σε σχέση με τις παραμέτρους του ηθοποιού (βλ. εξίσωση (9)):

$$\nabla_{\theta_\mu} J = \mathbb{E}_{s_t-\rho^\beta}\left[\nabla_{\theta_\mu} Q(s,a \mid \theta^Q)\Big|_{s=s_i,\lambda=\mu(s_t|\theta')}\right] \quad (Εξίσωση\ 9)$$

Η μέθοδος DDPG μαθαίνει με μέσο συντελεστή, 20 φορές λιγότερα βήματα εμπειρίας συγκριτικά με τη μέθοδο DQN. Τόσο η DDPG, όσο και η DQN απαιτούν μεγάλα σύνολα δεδομένων δειγμάτων, καθώς είναι αλγόριθμοι χωρίς μοντέλα. Όσον αφορά τη μέθοδο της καθοδηγούμενης αναζήτησης πολιτικής (Guided Policy Search – GPS) που βασίζεται σε DNN (DNN-based), μαθαίνει να χαρτογραφεί από τις πολλαπλές ακατέργαστες οπτικές πληροφορίες και τις καταστάσεις άρθρωσης απευθείας στις ροπές των αρθρώσεων. Σε σύγκριση με τα προηγούμενα έργα, κατάφερε να εκτελέσει έλεγχο υψηλής διάστασης, ακόμη και από ατελή δεδομένα αισθητήρων. Η DNN-based GPS έχει εφαρμοστεί ευρέως στον ρομποτικό έλεγχο, από χειρισμό έως εργασίες πλοήγησης.

5.2 Βαθιά μάθηση για προγραμματισμό και επίγνωση της κατάστασης

Αρκετές εξελίξεις βαθιάς μάθησης έχουν αναφερθεί για εργασίες που σχετίζονται με τον προγραμματισμό UAV και την επίγνωση της κατάστασης. Οι εργασίες προγραμματισμού αναφέρονται στη δημιουργία λύσεων για πολύπλοκα προβλήματα, χωρίς να χρειάζεται να γίνει κωδικοποίηση του μοντέλου περιβάλλοντος ή των δεξιοτήτων ή τις στρατηγικές του ρομπότ σε έναν αντιδραστικό ελεγκτή. Ο προγραμματισμός απαιτείται παρουσία αδόμητων, δυναμικών περιβαλλόντων ή όταν υπάρχει ποικιλία στο εύρος ή/και στις εργασίες του ρομπότ. Οι τυπικές εργασίες περιλαμβάνουν σχεδιασμό διαδρομής, κίνησης, πλοήγησης ή χειραγώγησης. Οι εργασίες επίγνωσης της κατάστασης επιτρέπουν στα ρομπότ να έχουν γνώση για τη δική τους κατάσταση και την κατάσταση του περιβάλλοντός τους. Μερικά παραδείγματα τέτοιου είδους εργασιών είναι η εκτίμηση της κατάστασης του ρομπότ, ο αυτοπροσδιορισμός και η χαρτογράφηση.

Προγραμματισμός. Ο προγραμματισμός διαδρομής για συνεργατικές αποστολές αναζήτησης και διάσωσης, ο οποίος λειτουργεί με εξερεύνηση βασισμένη στη βαθιά μάθηση, παρουσιάζεται στη μελέτη των συγγραφέων (Delmerico etal., 2017). Αυτή η εργασία, όπου δηλαδή ένα UAV εξερευνά και χαρτογραφεί το περιβάλλον προσπαθώντας να βρει μια διαδρομή, μέσω της οποίας μπορεί να περάσει ένα ρομπότ εδάφους, εστιάζει στην ελαχιστοποίηση του συνολικού χρόνου ανάπτυξης (δηλαδή, εξερεύνηση και διάβαση διαδρομής). Προκειμένου να χαρτογραφηθεί το έδαφος και να βρεθεί μια διαδρομή, προτείνεται ένα CNN για ταξινόμηση του εδάφους. Όμως, αντί να χρησιμοποιηθεί ένα ήδη προπονημένο CNN, η εκπαίδευση πραγματοποιείται εκείνη τη στιγμή, επιτρέποντας με αυτό το τρόπο την εκπαίδευση του ταξινομητή στο υπάρχον έδαφος του σημείου της καταστροφής. Ωστόσο, ο χρόνος που χρειάζεται το εκάστοτε μοντέλο για να προπονηθεί, είναι περίπου 15 λεπτά.

Επίγνωση της κατάστασης. Σύμφωνα με τη μελέτη των συγγραφέων (Taisho etal., 2015), ο εντοπισμός διασταυρούμενων εικόνων επιτυγχάνεται με τη βοήθεια της βαθιάς μάθησης. Παρόλο που η εργασία τους παρουσιάζεται ως λύση για τον εντοπισμό UAV, δεν χρησιμοποιήθηκαν UAV για τη συλλογή εικόνων. Τα πειράματα που πραγματοποιήθηκαν βασίστηκαν μόνο σε εικόνες εδάφους. Η προσέγγιση της μελέτης τους βασίζεται στην εξόρυξη μιας βιβλιοθήκης ακατέργαστων δεδομένων εικόνας, για την εύρεση οπτικών χαρακτηριστικών του πλησιέστερου γείτονα, τα οποία στη συνέχεια ταιριάζουν με τα χαρακτηριστικά που εξάγονται από μια εικόνα ερωτήματος εισόδου. Έπειτα χρησιμοποιείται ένα προκαταρκτικό CNN για την εξαγωγή χαρακτηριστικών με σκοπό την αντιστοίχιση σκοπών επαλήθευσης, και παρόλο που αυτή η προσέγγιση λέγεται ότι έχει χαμηλή υπολογιστική πολυπλοκότητα, οι συγγραφείς δεν παρέχουν λεπτομέρειες σχετικά με το χρόνο ανάκτησης.

Στην έρευνα των (Lin etal., 2015) οι εικόνες ερωτήματος επιπέδου αντιστοιχίζονται σε μια βάση δεδομένων αναφοράς εναέριων εικόνων. Η βαθιά μάθηση εφαρμόζεται σε αυτή τη περίπτωση για να μειώσει τις ευρείες διακυμάνσεις της βασικής γραμμής και της εμφάνισης μεταξύ επιπέδων εδάφους και αεροφωτογραφιών. Επιπλέον, προτείνεται μια δομή δικτύου που βασίζεται σε ζεύγη, έτσι ώστε να μάθει βαθιές αναπαραστάσεις από δεδομένα, με σκοπό τη διάκριση αντιστοιχισμένων και ασύγκριτων ζευγών εικόνων εγκάρσιας προβολής. Παρόλο που η διαδικασία κατάρτισης στα αναφερόμενα πειράματα διήρκεσε 4 ημέρες, η χρήση γρήγορων αλγορίθμων επέτρεψε την αντιστοίχιση σε πραγματικό χρόνο και σε κλίμακα πόλης. Ο κύριος περιορισμός της προσέγγισης των συγγραφέων (Lin etal., 2015) είναι η ανάγκη να εκτιμηθεί η κλίμακα, ο προσανατολισμός και το κυρίαρχο βάθος κατά τη διάρκεια των δοκιμών για ερωτήματα επιπέδου. Σε μία άλλη έρευνα, οι συγγραφείς (Aznar etal., 2016) προτείνουν ένα CNN, ούτως ώστε να δημιουργήσει ενέργειες ελέγχου (οι επιτρεπόμενες στροφές για ένα UAV), δεδομένης μιας εικόνας που έχει ληφθεί επί του σκάφους και ενός παγκόσμιου σχεδίου κίνησης. Αυτό το παγκόσμιο σχέδιο κίνησης υποδεικνύει τις ενέργειες που πρέπει να ληφθούν, δεδομένης μιας θέσης στο χάρτη μέσω μιας πιθανής συνάρτησης. Ο σκοπός του CNN είναι να μάθει την απεικόνιση από εικόνες σε ενέργειες που εξαρτώνται από τη θέση. Η διαδικασία αυτή θα ήταν ισοδύναμη με την εκτέλεση καταχώρισης εικόνας και με τη δημιουργία των ενεργειών ελέγχου δεδομένων, του παγκόσμιου σχεδίου κίνησης. Αλλά σε αυτή την έρευνα των (Aznar etal., 2016) η συμπεριφορά μαθαίνεται και κωδικοποιείται αποτελεσματικά σε ένα CNN, επιδεικνύοντας ανώτερα αποτελέσματα από τις κλασικές τεχνικές καταγραφής εικόνας. Ωστόσο, λόγω του ότι δεν πραγματοποιήθηκαν δοκιμές σε πραγματικό UAV και, κατ' επέκταση, δεν παρέχονται

πληροφορίες σχετικά με το χρόνο εκτέλεσης, η ανάπτυξη για μια πραγματική εφαρμογή UAV μπορεί να παρουσιάσει επιπλοκές.

Τέλος, όπως φαίνεται από τις μελέτες που παρουσιάστηκαν προηγουμένως, οι εξελίξεις στον προγραμματισμό και την επίγνωση της κατάστασης με τη βαθιά μάθηση για τα UAV, βρίσκονται ακόμα σε ένα στοιχειώδες στάδιο. Η προσέγγιση σχεδιασμού διαδρομής που παρουσιάζεται περιορίζεται σε μικρής κλίμακας περιοχές καταστροφών. Όσον αφορά τις διαφορετικές προσεγγίσεις εντοπισμού και χαρτογράφησης, παραμένουν αργές και έχουν μικρή ακρίβεια για πραγματικές εφαρμογές UAV.

5.3 Βαθιά μάθηση για έλεγχο κίνησης

Τεχνικές βαθιάς μάθησης για τον έλεγχο της κίνησης, έχουν χρησιμοποιηθεί πρόσφατα σε πολλές επιστημονικές έρευνες. Ο κλασικός έλεγχος έχει λύσει διάφορα προβλήματα ρομποτικού ελέγχου με ακριβή και αναλυτικό τρόπο, επιτρέποντας στα ρομπότ να εκτελούν πολύπλοκους ελιγμούς. Παρ' όλα αυτά, η τυπική θεωρία ελέγχου λύνει το πρόβλημα μόνο για μια συγκεκριμένη περίπτωση και για ένα κατά προσέγγιση μοντέλο ρομπότ. Κατ' επέκταση, δε μπορεί να προσαρμοστεί εύκολα στις αλλαγές που εφαρμόζονται στο μοντέλο ρομπότ ή/και σε εχθρικά περιβάλλοντα (π.χ. μια προπέλα σε UAV που καταστρέφεται, άνεμος και βροχή). Σε αυτό το πλαίσιο, η μάθηση εκ πείρας είναι ένα ζήτημα σημαντικής σημασίας που μπορεί να ξεπεράσει πολλούς αναφερόμενους περιορισμούς. Ως βασικό πλεονέκτημα, οι μέθοδοι βαθιάς μάθησης είναι σε θέση να γενικεύσουν σωστά χρησιμοποιώντας ορισμένα σύνολα επισημασμένων δεδομένων εισόδου. Η βαθιά μάθηση επιτρέπει την εξαγωγή ενός μοτίβου από ακατέργαστες εισόδους, όπως π.χ. εικόνες και δεδομένα αισθητήρων LIDAR (LIght Detection And Ranging), που μπορούν να οδηγήσουν σε σωστή συμπεριφορά ακόμη και σε άγνωστες καταστάσεις. Όσον αφορά την εργασία εσωτερικής πλοήγησης UAV, οι πρόσφατες εξελίξεις οδήγησαν σε μια επιτυχημένη εφαρμογή των CNN προκειμένου να αντιστοιχίσουν εικόνες σε οδηγίες συμπεριφοράς υψηλού επιπέδου (π.χ. στρίψε αριστερά, στρίψε δεξιά, περίστρεψε αριστερά και περίστρεψε δεξιά). Στην έρευνα των (Sadeghi&Levine, 2016) η συνάρτηση Q εκτιμάται μέσω ενός CNN, το οποίο εκπαιδεύεται στην προσομοίωση και δοκιμάζεται επιτυχώς σε πραγματικά πειράματα. Ωστόσο, στην έρευνα των (Kim&Chen, 2015) οι ενέργειες αντιστοιχίζονται απευθείας από ακατέργαστες εικόνες. Σε όλες τις αναφερόμενες μεθόδους, το μαθημένο μοντέλο 'τρέχει' εκτός λειτουργίας, αξιοποιώντας συνήθως μια μονάδα επεξεργασίας γραφικών (GPU) σε έναν εξωτερικό φορητό υπολογιστή. Όσον αφορά την πλοήγηση με UAV σε μη δομημένο περιβάλλον, ορισμένες μελέτες έχουν επικεντρωθεί σε ακατάστατα φυσικά σενάρια, όπως πυκνά δάση ή μονοπάτια. Πιο

συγκεκριμένα, στην έρευνα των (Giusti etal., 2015) εκπαιδεύτηκε ένα μοντέλο DNN για να αντιστοιχίσει την εικόνα σε πιθανότητες δράσης (στρίψε αριστερά, συνέχισε ευθεία ή στρίψε δεξιά) με ένα τελικό στρώμα softmax και δοκιμάστηκε επί του σκάφους μέσω επεξεργαστή ODROID-U3.

Σε άλλη έρευνα, όπως αυτή των συγγραφέων (Shah etal., 2016), οι πλωτές περιοχές προβλέπονται από μια εικόνα ανισότητας με τη μορφή έως τριών οριακών πλαισίων. Το κέντρο του μεγαλύτερου πλαισίου ορίου που βρέθηκε επιλέγεται ως το επόμενο σημείο. Χρησιμοποιώντας αυτήν τη στρατηγική οι πτήσεις UAV εκτελούνται με επιτυχία, έχοντας όμως ως κύριο μειονέκτημα, την απαίτηση αποστολής των εικόνων ανισότητας σε μια συσκευή υποδοχής, όπου γίνονται όλοι οι υπολογισμοί. Ολόκληρος ο αγωγός για την οριζόντια μεταφορά UAV, τη δημιουργία χαρτών ανισότητας και την επιλογή σημείων διαρκεί περίπου 1,3 δευτερόλεπτα, γεγονός που καθιστά την πλοήγηση αρκετά αργή για πραγματικές εφαρμογές. Από την άλλη πλευρά, ο έλεγχος κίνησης χαμηλού επιπέδου είναι αρκετά δύσκολος, καθώς η αντιμετώπιση συνεχών και πολλαπλών μεταβλητών χώρων δράσης μπορεί να γίνει ένα δυσεπίλυτο πρόβλημα. Παρ' όλα αυτά, πρόσφατες έρευνες έχουν προτείνει νέες μεθόδους για την εκμάθηση πολιτικών ελέγχου χαμηλού επιπέδου από ατελή δεδομένα αισθητήρων στην προσομοίωση. Σε μία από αυτές, και πιο συγκεκριμένα σε αυτή των συγγραφέων (Zhang etal., 2016), χρησιμοποιήθηκε ένα μοντέλο προγνωστικού ελέγχου (Model Predictive Control - MPC) για τη δημιουργία δεδομένων κατά την προπόνηση, προκειμένου να εκπαιδεύσει μια πολιτική DNN. Σε αυτή τη πολιτική επιτρέπεται η πρόσβαση μόνο σε ακατέργαστες παρατηρήσεις από τους ενσωματωμένους αισθητήρες UAV. Κατά το χρόνο δοκιμής το UAV μπόρεσε να ακολουθήσει μια τροχιά χωρίς εμπόδια ακόμη και σε άγνωστες καταστάσεις. (Η χρήση του MPC, σε συνδυασμό με την ενισχυτική μάθηση, αναλύεται περισσότερο παρακάτω).

Συνοψίζοντας, οι τεχνικές βαθιάς εκμάθησης για ρομποτικό έλεγχο κίνησης, μπορούν να προσφέρουν αυξανόμενα οφέλη, προκειμένου να συναχθούν πολύπλοκες συμπεριφορές από ακατέργαστα δεδομένα παρατήρησης. Οι προσεγγίσεις βαθιάς μάθησης έχουν τη δυνατότητα και τη προοπτική γενίκευσης, αλλά πρέπει να αντιμετωπίσουν τους περιορισμούς των τρεχουσών μεθόδων. Επομένως, πρέπει να ξεπεραστούν οι δυσκολίες των συνεχών χώρων κατάστασης και δράσης, καθώς και ζητήματα που σχετίζονται με την αποτελεσματικότητα των δειγμάτων. Επιπλέον, τα νέα μοντέλα βαθιάς εκμάθησης απαιτούν τη χρήση GPU για να λειτουργούν σε πραγματικό χρόνο. Σε αυτό το πλαίσιο, οι ενσωματωμένες GPU, οι προγραμματιζόμενες συστοιχίες πύλης πεδίου (FPGA) ή τα ολοκληρωμένα κυκλώματα ειδικών εφαρμογών

(Application Specific Integrated Circuits - ASIC) είναι ένα ζήτημα σημαντικής σημασίας που θα πρέπει να λάβουν υπόψη οι κατασκευαστές υλικού.

5.4 Μοντέλο προγνωστικού ελέγχου

Το μοντέλο προγνωστικού ελέγχου (MPC) είναι μια αποτελεσματική μέθοδος για τον έλεγχο ρομποτικών συστημάτων, ιδιαίτερα των αυτόνομων εναέριων οχημάτων όπως π.χ. τα τετρακόπτερα. Όμως η εφαρμογή του μπορεί να είναι υπολογιστικά απαιτητική καθώς απαιτεί εκτίμηση της κατάστασης του συστήματος. Ειδικότερα σε περιπτώσεις που υπάρχουν πολύπλοκα και αδόμητα περιβάλλοντα, αυτή η εκτίμηση μπορεί να είναι αρκετά προκλητική. Η ενισχυτική μάθηση μπορεί σε πρώτη φάση να αποφύγει την ανάγκη για ρητή εκτίμηση της κατάστασης και να αποκτήσει μια πολιτική που να αντιστοιχίζει άμεσα τις ενδείξεις των αισθητήρων σε ενέργειες. Ωστόσο είναι δύσκολο να εφαρμοστεί σε ασταθή συστήματα και ως αποτέλεσμα, μπορεί να αποδειχθεί καταστροφική κατά τη διάρκεια της εκπαίδευσης. Για το λόγο αυτό, οι συγγραφείς (Zhang etal., 2016) προτείνουν τον συνδυασμό του MPC με την ενισχυμένη μάθηση στο πλαίσιο της αναζήτησης καθοδηγούμενης πολιτικής, όπου το MPC χρησιμοποιείται για τη δημιουργία δεδομένων κατά τη διάρκεια της εκπαίδευσης, κάτω από πλήρεις κρατικές παρατηρήσεις που παρέχονται από ένα εργαλείο εκπαιδευτικού περιβάλλοντος. Στη συνέχεια, αυτά τα δεδομένα χρησιμοποιούνται για την εκπαίδευση μιας πολιτικής για τα νευρωνικά δίκτυα, η οποία επιτρέπεται να έχει πρόσβαση μόνο στις ακατέργαστες παρατηρήσεις από τους αισθητήρες του οχήματος. Μετά από αυτή την εκπαίδευση, η πολιτική νευρωνικών δικτύων είναι πλέον σε θέση να ελέγχει με επιτυχία το ρομπότ, ακόμα και χωρίς να γνωρίζει την πλήρη κατάσταση και με μόλις ένα μέρος του υπολογιστικού κόστους του MPC. Τέλος, οι συγγραφείς (Zhang etal., 2016) αξιολογούν τη μέθοδό τους μαθαίνοντας πολιτικές αποφυγής εμποδίων για ένα προσομοιωμένο τετρακόπτερο, χρησιμοποιώντας προσομοιωμένους αισθητήρες επί του σκάφους και χωρίς ρητή εκτίμηση κατάστασης τη στιγμή της δοκιμής.

6

Συμπεράσματα

Ο αυξανόμενος όγκος και οι πολλαπλές κατηγορίες διαθέσιμων δεδομένων, η υπολογιστική επεξεργασία που είναι φθηνότερη και ισχυρότερη, αλλά και η οικονομικά προσιτή αποθήκευση των δεδομένων, αποτελούν παράγοντες που καθιστούν τη μηχανική μάθηση πιο δημοφιλή από ποτέ. Χάρη σε αυτή είναι δυνατή η δημιουργία γρήγορων και αυτόματων μοντέλων που μπορούν να αναλύουν μεγαλύτερα και πιο περίπλοκα δεδομένα, αλλά και να παρέχουν ταχύτερα, πιο ακριβή αποτελέσματα, ακόμη και σε πολύ μεγάλη κλίμακα.

Οι τεχνικές AI και ML και ειδικά η βαθιά μάθηση, είναι πλέον εκ των ων ουκ άνευ σε πολλούς τομείς που περιλαμβάνουν ταξινόμηση εικόνας και όραση υπολογιστή, από κοινωνικά δίκτυα έως ασφάλεια. Εφαρμόζονται σε προβληματικές περιοχές, όπου σημαντικές ποσότητες δεδομένων είναι άμεσα διαθέσιμες για εκπαίδευση. Η ενισχυτική μάθηση αρχίζει να εφαρμόζεται σε μια ποικιλία εφαρμογών ρομποτικού ελέγχου μετά από διάφορες επιδείξεις της ικανότητάς της σε περιβάλλοντα παιχνιδιών.

Η συνεχιζόμενη έρευνα έχει δείξει ότι τα συστήματα βαθιάς εκμάθησης μπορούν να μάθουν να επικοινωνούν πάνω από στατικούς συνδέσμους πιο αποτελεσματικά από τους σχεδιασμούς συστημάτων που βασίζονται σε μοντέλα.

Οι περισσότερες βιομηχανίες που επεξεργάζονται μεγάλες ποσότητες δεδομένων, αναγνωρίζουν την αξία της τεχνολογίας της μηχανικής μάθησης. Με τη συλλογή πληροφοριών από αυτά τα δεδομένα που πραγματοποιείται συχνά σε πραγματικό χρόνο, οι οργανισμοί μπορούν να εργάζονται πιο αποτελεσματικά, ή να αποκτούν πλεονέκτημα έναντι των ανταγωνιστών τους.

Τόσο η βαθιά, όσο και η ενισχυτική μάθηση συνδέονται σε μεγάλο βαθμό με την υπολογιστική ισχύ της τεχνητής νοημοσύνης (AI). Είναι αυτόνομες λειτουργίες μηχανικής μάθησης που ανοίγουν τον δρόμο για τους υπολογιστές να δημιουργήσουν τις δικές τους αρχές για να βρουν λύσεις. Αυτά τα δύο είδη μάθησης μπορεί επίσης να συνυπάρχουν σε πολλά προγράμματα. Γενικά, η βαθιά μάθηση χρησιμοποιεί τρέχοντα δεδομένα, ενώ η ενισχυτική μάθηση χρησιμοποιεί τη μέθοδο δοκιμής και λάθους για τον υπολογισμό των προβλέψεων.

Η ενισχυτική μάθηση (RL) χρησιμοποιείται συχνά στη ρομποτική, σε παιχνίδια και στην πλοήγηση. Στην ενισχυτική μάθηση ο αλγόριθμος ανακαλύπτει μέσω δοκιμής και σφάλματος ποιες ενέργειες αποδίδουν τα καλύτερα αποτελέσματα. Αυτός ο τύπος μάθησης έχει τρία βασικά συστατικά: τον παράγοντα (τον μαθητή ή τον υπεύθυνο λήψης αποφάσεων), το περιβάλλον (οτιδήποτε αλληλοεπιδράει με τον παράγοντα) και τις ενέργειες (τι μπορεί να κάνει ο παράγοντας). Στόχος είναι ο παράγοντας να επιλέξει τις ενέργειες που θα αποδώσουν το βέλτιστο δυνατό αποτέλεσμα για ένα δεδομένο χρονικό διάστημα, κάτι που θα πετύχει εφόσον ακολουθήσει την καλύτερη δυνατή πολιτική.

Η βαθιά μάθηση είναι σε θέση να εκτελέσει τη συμπεριφορά-στόχο αναλύοντας υπάρχοντα δεδομένα και εφαρμόζοντας όσα έμαθαν σε ένα νέο σύνολο πληροφοριών. Από την άλλη πλευρά, η ενισχυτική μάθηση είναι σε θέση να αλλάξει την απόκρισή της προσαρμόζοντας τη συνεχή ανατροφοδότηση.

Η βαθιά μάθηση προσφέρει σημαντικά πλεονεκτήματα:

- Σε τομείς όπως οι επικοινωνίες και η δικτύωση, όπου η διαχείριση των παρεμβολών, οι κυβερνο-επιθέσεις και η εκφόρτωση δεδομένων μπορούν αν αντιμετωπισθούν ως παίγνια, ενώ η ενισχυτική μάθηση είναι συχνά το εργαλείο που χρησιμοποιείται για την επίλυση παιγνίων.
- Παρέχει αυτόνομη λήψη αποφάσεων, με συνέπεια την μείωση των εξόδων, αλλά και την βελτίωση της ασφάλειας και της ευρωστίας του δικτύου.
- Βρίσκει τη λύση εξελιγμένων βελτιστοποιήσεων δικτύου (χωρίς να απαιτούνται ακριβείς και πλήρεις πληροφορίες δικτύου).
- Δίνει τη δυνατότητα στις οντότητες του δικτύου να αναπτύξουν γνώσεις, όσον αφορά το περιβάλλον της δικτύωσης και τέλος,

- Βελτιώνει την ταχύτητα εκμάθησης.

Ειδικά σε τομείς, όπως η ασφάλεια στον κυβερνοχώρο, έχουν προταθεί διάφορες μέθοδοι βαθιάς ενισχυτικής μάθησης (DRL) η οποία είναι ιδιαιτέρως ικανή στην επίλυση πολύπλοκων, δυναμικών και υψηλών διαστάσεων προβλημάτων που σχετίζονται με την άμυνα στον κυβερνοχώρο. Εδώ πρέπει να σημειωθεί ότι οι εφαρμογές της DRL όσον αφορά τον κυβερνοχώρο διακρίνονται, αφενός μεν σε εφαρμογές βελτιστοποίησης και ενίσχυσης των δυνατοτήτων επικοινωνίας και δικτύωσης των εφαρμογών IoT, αφετέρου δε σε εφαρμογές άμυνας έναντι των κυβερνοεπιθέσεων.

Όσον αφορά τα συστήματα επικοινωνίας η DL αποτελεί μία πολλά υποσχόμενη τεχνική, επειδή ένα τέτοιο σύστημα:

1. Βελτιστοποιείται από άκρο σε άκρο χρησιμοποιώντας βαθιά νευρωνικά δίκτυα.
2. Βελτιστοποιείται για οιονδήποτε τύπο καναλιού χωρίς τη χρήση μαθηματικού μοντέλου με δυνατότητα μεταφοράς.
3. Οι αλγόριθμοι DL παρέχουν μεγαλύτερη ταχύτητα επεξεργασίας.

Η βαθιά μάθηση χρησιμοποιείται με μεγάλη επιτυχία σε ασύρματα δίκτυα επεξεργασίας δεδομένων, όπου η συλλογή, η ροή και η επεξεργασία των δεδομένων αποτελούν βασικές λειτουργίες για τα σύγχρονα δίκτυα.

Η ενισχυτική μάθηση (RL) παρέχει ένα μαθηματικό πλαίσιο για τη μάθηση ή τη δημιουργία πολιτικών, που χαρτογραφούν τις καταστάσεις σε ενέργειες, με στόχο τη μεγιστοποίηση μιας σωρευτικής ανταμοιβής. Στην RL ο πράκτορας μαθαίνει την πολιτική λήψης αποφάσεων μέσω αλληλεπιδράσεων με το περιβάλλον. Ο συνδυασμός της συμβατικής Q-learning και του βαθιού νευρωνικού δικτύου, παρέχει μια σημαντική ανακάλυψη στη βαθιά ενισχυτική μάθηση (DRL). Η DRL αποτελεί μία πολλά υποσχόμενη τεχνική, όσον αφορά τη διαχείριση δεδομένων, στο χειρισμό μεγάλου χώρου με προβλήματα ελέγχου, επειδή μπορεί να χειριστεί έναν περίπλοκο χώρο κατάστασης και ένα περιβάλλον μεταβαλλόμενου χρόνου, χρησιμοποιώντας παράλληλα ισχυρά βαθιά νευρωνικά δίκτυα (DNNs) για να καθοδηγήσει τη λήψη αποφάσεων. Η ενισχυτική μάθηση συνθέτει μια διαδικασία βελτιστοποίησης σε ολόκληρο τον χώρο της κατάστασης, προκειμένου να μεγιστοποιηθεί η συσσωρευμένη ανταμοιβή.

Τεχνικές βαθιάς μάθησης έχουν χρησιμοποιηθεί πρόσφατα σε πολλές επιστημονικές έρευνες για τον έλεγχο της κίνησης, επειδή επιτρέπουν την εξαγωγή ενός μοτίβου από ακατέργαστες εισόδους, όπως π.χ. εικόνες και δεδομένα αισθητήρων LIDAR, που μπορούν να οδηγήσουν σε σωστή συμπεριφορά ακόμη και σε άγνωστες καταστάσεις.

Εν κατακλείδι, οι τεχνικές βαθιάς εκμάθησης για ρομποτικό έλεγχο κίνησης, μπορούν να προσφέρουν αυξανόμενα οφέλη προκειμένου να συναχθούν πολύπλοκες συμπεριφορές από ακατέργαστα δεδομένα παρατήρησης. Οι προσεγγίσεις βαθιάς μάθησης έχουν τη δυνατότητα και τη προοπτική γενίκευσης, ωστόσο θα πρέπει να αντιμετωπίσουν τους περιορισμούς των τρεχουσών μεθόδων, συνεπώς θα πρέπει να ξεπεραστούν οι δυσκολίες των συνεχών χώρων κατάστασης και δράσης, καθώς και όποια ζητήματα σχετίζονται με την αποτελεσματικότητα των δειγμάτων. Επιπλέον, τα νέα μοντέλα βαθιάς εκμάθησης απαιτούν τη χρήση GPU για να λειτουργούν σε πραγματικό χρόνο. Σε αυτό το πλαίσιο, οι ενσωματωμένες GPU, οι προγραμματιζόμενες συστοιχίες πύλης πεδίου (FPGA) ή τα ολοκληρωμένα κυκλώματα ειδικών εφαρμογών (ASIC) είναι ένα ζήτημα σημαντικής σημασίας που θα πρέπει να λάβουν υπόψη οι κατασκευαστές υλικού.

Το μοντέλο προγνωστικού ελέγχου αποτελεί μια αποτελεσματική μέθοδο για τον έλεγχο ρομποτικών συστημάτων, ιδιαίτερα των αυτόνομων εναέριων οχημάτων, ωστόσο η εφαρμογή του και ειδικά σε περιβάλλοντα πολύπλοκα και αδόμητα μπορεί να αποβεί υπολογιστικά απαιτητική. Η ενισχυτική μάθηση έχει το πλεονέκτημα ότι μπορεί αρχικά να αποφεύγει την ρητή αποτύπωση της κατάστασης, ωστόσο σε ασταθή συστήματα μπορεί να είναι καταστροφική για την εκπαίδευση για το λόγο αυτό πρέπει να συνδυάζεται με το MPC.

Βιβλιογραφία

Akazaki, T., Liu, S., Yamagata, Y., Duan, Y., &Hao, J. (2018, July). Falsification of cyber-physical systems using deep reinforcement learning. In *International Symposium on Formal Methods* (pp. 456-465). Springer, Cham.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, *34*(6), 26-38.

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

Carrio, A., Sampedro, C., Rodriguez-Ramos, A., &Campoy, P. (2017). A review of deep learning methods and applications for unmanned aerial vehicles. *Journal of Sensors*, *2017*.

Challita, U., Dong, L., & Saad, W. (2018). Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective. *IEEE transactions on wireless communications*, *17*(7), 4674-4689.

Chen, M., Challita, U., Saad, W., Yin, C., &Debbah, M. (2019). Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*, *21*(4), 3039-3071.

Chen, M., Saad, W., & Yin, C. (2017, December). Liquid state machine learning for resource allocation in a network of cache-enabled LTE-U UAVs. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1-6). IEEE.

Chen, Y., Li, Y., Xu, D., & Xiao, L. (2018, June). DQN-based power control for IoT transmission against jamming. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)* (pp. 1-5). IEEE.

Chinchali, S., Hu, P., Chu, T., Sharma, M., Bansal, M., Misra, R., ... &Katti, S. (2018, April). Cellular network traffic scheduling with deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*.

Chu, M., Li, H., Liao, X., & Cui, S. (2018). Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems. *IEEE Internet of Things Journal*, *6*(2), 2009-2020.

Delmerico, J., Mueggler, E., Nitsch, J., &Scaramuzza, D. (2017). Active autonomous aerial exploration for ground robot path planning. *IEEE Robotics and Automation Letters*, *2*(2), 664-671.

Di Renzo, M., Debbah, M., Phan-Huy, D. T., Zappone, A., Alouini, M. S., Yuen, C., ... & Fink, M. (2019). Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come. *EURASIP Journal on Wireless Communications and Networking*, *2019*(1), 1-20.

Di, B., Zhang, H., Song, L., Li, Y., & Li, G. Y. (2018). Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading. *IEEE Transactions on Wireless Communications*, *18*(1), 47-62.

Dörner, S., Cammerer, S., Hoydis, J., & Ten Brink, S. (2017). Deep learning based communication over the air. *IEEE Journal of Selected Topics in Signal Processing*, *12*(1), 132-143.

F. Aznar, M. Pujol, and R. Rizo, "Visual Navigation for UAV with Map References Using ConvNets," in Advances in Artificial Intelligence, vol. 9868 of Lecture Notes in Computer Science, pp. 13–22, Springer, 2016.

Ferreira, P. V. R., Paffenroth, R., Wyglinski, A. M., Hackett, T. M., Bilén, S. G., Reinhart, R. C., & Mortensen, D. J. (2018). Multiobjective reinforcement learning for cognitive satellite communications using deep neural network ensembles. *IEEE Journal on Selected Areas in Communications*, *36*(5), 1030-1041.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., ... & Legg, S. (2017). Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*.

Fotouhi, A., Qiang, H., Ding, M., Hassan, M., Giordano, L. G., Garcia-Rodriguez, A., & Yuan, J. (2019). Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges. *IEEE Communications Surveys & Tutorials*, *21*(4), 3417-3442.

Gadaleta, M., Chiariotti, 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.

Giusti, A., Guzzi, J., Cireşan, D. C., He, F. L., Rodríguez, J. P., Fontana, F., ... & Gambardella, L. M. (2015). A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, *1*(2), 661-667.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gupta, A. K., & Andrews, J. G. (2019). Comments on "coverage analysis of multiuser visible light communication networks". *IEEE Transactions on Wireless Communications*, *18*(9), 4605-4606.

Gupta, A., & Yang, Z. (2018). Adversarial reinforcement learning for observer design in autonomous systems under cyber attacks. *arXiv preprint arXiv:1809.06784*.

Han, G., Xiao, L., & Poor, H. V. (2017, March). Two-dimensional anti-jamming communication based on deep reinforcement learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2087-2091).IEEE.

Han, Y., Rubinstein, B. I., Abraham, T., Alpcan, T., De Vel, O., Erfani, S., ... & Montague, P. (2018, October). Reinforcement learning for autonomous defence in software-defined

networking. In *International Conference on Decision and Game Theory for Security* (pp. 145-165). Springer, Cham.

Hausknecht, M., & Stone, P. (2015, September). Deep recurrent q-learning for partially observable mdps. In *2015 aaai fall symposium series*.

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.

Heinrich, J., & Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., ... & Silver, D. (2018, April). Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*.

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

Huang, T., Zhang, R. X., Zhou, C., & Sun, L. (2018, October). Qarc: Video quality aware rate control for real-time video streaming based on deep reinforcement learning. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 1208-1216).

Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K. C., &Hanzo, L. (2016). Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, *24*(2), 98-105.

Kim, D. K., & Chen, T. (2015). Deep neural network for real-time autonomous indoor navigation. *arXiv preprint arXiv:1511.04668*.

Korpi, D., Yli-Opas, P., Jaramillo, M. R., &Uusitalo, M. A. (2020, March). Visual detection-based blockage prediction for beyond 5G wireless systems. In *2020 2nd 6G Wireless Summit (6G SUMMIT)* (pp. 1-5). IEEE.

Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y. J. A. (2019). The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine*, *57*(8), 84-90.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Li, Y., Li, H., Li, Z., Fang, H., Sanyal, A. K., Wang, Y., &Qiu, Q. (2019, August). Fast and accurate trajectory tracking for unmanned aerial vehicles based on deep reinforcement learning. In *2019 IEEE 25th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)* (pp. 1-9). IEEE.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... &Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lin, T. Y., Cui, Y., Belongie, S., & Hays, J. (2015). Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5007-5015).

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.

Liu, C. H., Chen, Z., Tang, J., Xu, J., & Piao, C. (2018). Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach. *IEEE Journal on Selected Areas in Communications*, *36*(9), 2059-2070.

Liu, S., Hu, X., & Wang, W. (2018). Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems. *IEEE Access*, *6*, 15733-15742.

Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2020). Application of deep reinforcement learning to intrusion detection for supervised problems. *Expert Systems with Applications*, *141*, 112963.

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.

Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y. C., & Kim, D. I. (2019). Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, *21*(4), 3133-3174.

Maass, W. (2011). Liquid state machines: motivation, theory, and applications. *Computability in context: computation and logic in the real world*, 275-296.

Malialis, K., &Kudenko, D. (2015). Distributed response to network intrusions using multiagent reinforcement learning. *Engineering Applications of Artificial Intelligence*, *41*, 270-284.

Mao, H., Netravali, R., & Alizadeh, M. (2017, August). Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (pp. 197-210).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... &Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928-1937). PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529-533.

Mohammed, A., Mehmood, A., Pavlidou, F. N., &Mohorcic, M. (2011). The role of high-altitude platforms (HAPs) in the global wireless connectivity. *Proceedings of the IEEE*, *99*(11), 1939-1953.

Monahan, G. E. (1982). State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. *Management science*, *28*(1), 1-16.

Naparstek, O., & Cohen, K. (2017, December). Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1-7). IEEE.

Nawaz, S. J., Sharma, S. K., Wyne, S., Patwary, M. N., &Asaduzzaman, M. (2019). Quantum machine learning for 6G communication networks: State-of-the-art and vision for the future. *IEEE Access*, *7*, 46317-46350.

Nguyen, T. T., &Reddi, V. J. (2019). Deep reinforcement learning for cyber security. *arXiv preprint arXiv:1906.05799*.

O'shea, T., &Hoydis, J. (2017). An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, *3*(4), 563-575.

O'Shea, T. J., Roy, T., West, N., & Hilburn, B. C. (2018, September). Physical layer communications system design over-the-air using adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 529-532). IEEE.

Patwary, M. N., Nawaz, S. J., Rahman, M. A., Sharma, S. K., Rashid, M. M., & Barnes, S. J. (2020). The potential short-and long-term disruptions and transformative impacts of 5G and beyond wireless networks: Lessons learnt from the development of a 5G testbed environment. *IEEE Access*, *8*, 11352-11379.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Sadeghi, F., & Levine, S. (2016). Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Shah, U., Khawad, R., & Krishna, K. M. (2016, December). Deepfly: Towards complete autonomous navigation of mavs with monocular camera. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing* (pp. 1-8).

Stockhammer, T. (2011, February). Dynamic adaptive streaming over HTTP-- standards and design principles. In *Proceedings of the second annual ACM conference on Multimedia systems* (pp. 133-144).

Strinati, E. C., Barbarossa, S., Gonzalez-Jimenez, J. L., Ktenas, D., Cassiau, N., Maret, L., &Dehos, C. (2019). 6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication. *IEEE Vehicular Technology Magazine*, *14*(3), 42-50.

Sutton, R. S., &Barto, A. G. (1998). Reinforcement learning: an introduction MIT Press. *Cambridge, MA*, *22447*.

Sutton, R. S., &Barto, A. G. (1998). Reinforcement learning: an introduction MIT Press. *Cambridge, MA*, *22447*.

Szita, I., Gyenes, V., &Lőrincz, A. (2006, September). Reinforcement learning with echo state networks. In *International Conference on Artificial Neural Networks* (pp. 830-839). Springer, Berlin, Heidelberg.

Taisho, T., Enfu, L., Kanji, T., &Naotoshi, S. (2015, December). Mining visual experience for fast cross-view UAV localization. In *2015 IEEE/SICE International Symposium on System Integration (SII)* (pp. 375-380). IEEE.

Thrun, S., & Schwartz, A. (1993, December). Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School* (pp. 255-263).

Tilghman, P. (2019). Will rule the airwaves: A DARPA grand challenge seeks autonomous radios to manage the wireless spectrum. *IEEE Spectrum*, *56*(6), 28-33.

Van Hasselt, H., Guez, A., & Silver, D. (2016, March). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

Wan, X., Sheng, G., Li, Y., Xiao, L., & Du, X. (2017, December). Reinforcement learning based mobile offloading for cloud-based malware detection. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1-6). IEEE.

Wang, D., Wang, M., Zhu, P., Li, J., Wang, J., & You, X. (2019). Performance of network-assisted full-duplex for cell-free massive MIMO. *IEEE Transactions on Communications*, *68*(3), 1464-1478.

Wang, J., Jiang, C., Zhang, H., Ren, Y., Chen, K. C., &Hanzo, L. (2020). Thirty years of machine learning: The road to Pareto-optimal wireless networks. *IEEE Communications Surveys & Tutorials*, *22*(3), 1472-1514.

Wang, S., Liu, H., Gomes, P. H., &Krishnamachari, B. (2017, January). Deep reinforcement learning for dynamic multichannel access. In *International Conference on Computing, Networking and Communications (ICNC)* (pp. 257-265).

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). Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*.

Wang, Y., Li, Y., Lan, T., & Aggarwal, V. (2019). Deepchunk: Deep q-learning for chunk-based caching in wireless data processing networks. *IEEE Transactions on Cognitive Communications and Networking*, *5*(4), 1034-1045.

Wang, Z., Li, L., Xu, Y., Tian, H., & Cui, S. (2018). Handover control in wireless systems via asynchronous multiuser deep reinforcement learning. *IEEE Internet of Things Journal*, *5*(6), 4296-4307.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. (2016, June). Dueling network architectures for deep reinforcement learning. In *International conference on machine learning* (pp. 1995-2003). PMLR.

Wen, C. K., Shih, W. T., &Jin, S. (2018). Deep learning for massive MIMO CSI feedback. *IEEE Wireless Communications Letters*, *7*(5), 748-751.

Xiao, L., Li, Y., Han, G., Liu, G., & Zhuang, W. (2016). PHY-layer spoofing detection with reinforcement learning in wireless networks. *IEEE Transactions on Vehicular Technology*, *65*(12), 10037-10047.

Xiao, L., Wan, X., Su, W., & Tang, Y. (2018). Anti-jamming underwater transmission with mobility and learning. *IEEE Communications Letters*, *22*(3), 542-545.

Xiao, L., Xie, C., Min, M., & Zhuang, W. (2017). User-centric view of unmanned aerial vehicle transmission against smart attacks. *IEEE Transactions on Vehicular Technology*, *67*(4), 3420-3430.

Ye, H., & Li, G. Y. (2018, May). Deep reinforcement learning for resource allocation in V2V communications. In *2018 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

Yu, D., & Deng, L. (2010). Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine*, *28*(1), 145-154.

Zhang, T., Kahn, G., Levine, S., &Abbeel, P. (2016, May). Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search. In *2016 IEEE international conference on robotics and automation (ICRA)* (pp. 528-535). IEEE.

Zhao, D., Wang, H., Shao, K., & Zhu, Y. (2016, December). Deep reinforcement learning with experience replay based on SARSA. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-6). IEEE.

Zhao, N., Liang, Y. C., Niyato, D., Pei, Y., Wu, M., & Jiang, Y. (2019). Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, *18*(11), 5141-5152.

Zhu, H., Cao, Y., Wang, W., Jiang, T., &Jin, S. (2018). Deep reinforcement learning for mobile edge caching: Review, new features, and open issues. *IEEE Network*, *32*(6), 50-57.

Zhu, J., Song, Y., Jiang, D., & Song, H. (2017). A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things. *IEEE Internet of Things Journal*, *5*(4), 2375-2385.