ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Μελέτη και σύγκριση αλγορίθμων μηχανικής μάθησης για εκτίμηση ασύρματων καναλιών σε έξυπνες ανακλαστικές επιφάνειες και UAVs

Από

Ιωάννης Σουλιώτης

Υποβάλλεται

για την εκπλήρωση των προϋποθέσεων λήψης

Μεταπτυχιακού Διπλώματος

στην ειδίκευση «ΜΔΑ/ΠΠΣ/ΠΔ»

του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”

στο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σεπτέμβριος 2024

Επιβλέπων/Επιβλέπουσα: Μιχαήλ Φιλιππάκης

Ακαδημαϊκή Θέση:

Συγγραφέας / Author Σουλιώτης Ιωάννης

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας: Σουλιώτης Ιωάννης

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: Μελέτη και σύγκριση αλγορίθμων μηχανικής μάθησης για εκτίμηση ασύρματων καναλιών σε έξυπνες ανακλαστικές επιφάνειες και UAVs

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 24/10/2024 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων/ουσα (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς): Δρ. Φιλιππάκης, Καθηγητής του Τμήματος Ψηφιακών Συστημάτων

Μέλος Εξεταστικής Επιτροπής:Μαρία Χαλκίδη, Καθηγήτρια του Τμήματος Ψηφιακών Συστημάτων, , Πανεπιστήμιο Πειραιώς

Μέλος Εξεταστικής Επιτροπής:Δημοσθένης Κυριαζής, Καθηγητής του Τμήματος Ψηφιακών Συστημάτων, , Πανεπιστήμιο Πειραιώς

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο Σουλιώτης Ιωάννης., γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Μελέτη και σύγκριση αλγορίθμων μηχανικής μάθησης για εκτίμηση ασύρματων καναλιών σε έξυπνες ανακλαστικές επιφάνειες και UAVs», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο*

πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο/Η ΔΗΛΩΝ/ΟΥΣΑ

Ονοματεπώνυμο: Σουλιώτης Ιωάννης

Αριθμός Μητρώου: me2252

Υπογραφή:

Περιεχόμενα

1 Εισαγωγή

Τα κυψελωτά συστήματα επόμενης γενιάς θα βασίζονται αναπόφευκτα σε επικοινωνίες κυμάτων κυμάτων υψηλής συχνότητας (mmW) προκειμένου να καλύψουν την αυξανόμενη ανάγκη για ασύρματη χωρητικότητα, [1]. Ωστόσο, η επικοινωνία σε συχνότητες mmW αντιμετωπίζει πολλές προκλήσεις. Μια σημαντική πρόκληση είναι η υψηλή ευαισθησία των ζεύξεων mmW σε μπλοκάρισμα που προκαλείται από κοινά αντικείμενα, όπως δέντρα και ανθρώπινα σώματα, τα οποία μπορούν να εξασθενήσουν σοβαρά τα σήματα mmW. Η ενεργοποίηση αξιόπιστων ζεύξεων mmW που εμποδίζονται είναι επομένως ένα σημαντικό εμπόδιο που εμποδίζει την ανάπτυξη ζωνών mmW σε εμπορικές χρήσεις ευρείας κλίμακας.

Για να ξεπεραστούν αυτά τα μειονεκτήματα των mmW, έχουν προταθεί πρόσφατα ανακλαστήρες σήματος για να παρακάμπτουν τα εμπόδια και να παρατείνουν το εύρος επικοινωνίας. Ειδικότερα, με τη χρήση ανακλαστήρων, μια ζεύξη mmW μη οπτικής επαφής (NLOS) μπορεί να αντισταθμιστεί δημιουργώντας πολλαπλές, συνδεδεμένες συνδέσεις οπτικής επαφής (LOS), μειώνοντας έτσι σημαντικά την εξασθένηση του καναλιού mmW. Έχει επίσης αποδειχθεί ότι η χρήση ανακλαστήρων είναι πιο κατάλληλη για δίκτυα mmW από τους συμβατικούς σταθμούς αναμετάδοσης (Relay Stations - RS). Διαφορετικά από τους παραδοσιακούς RS που λαμβάνουν, ενισχύουν (ή αποκωδικοποιούν) και προωθούν το σήμα mmW, ένας ανακλαστήρας αντανακλά μόνο το προσπίπτον σήμα προς τον δέκτη, προκαλώντας μια συγκεκριμένη μετατόπιση φάσης. Επομένως, η ανακλαστική αναμετάδοση δεν προκαλεί πρόσθετο θόρυβο λήψης και είναι ενεργειακά αποδοτική. Λόγω της φύσης των ανακλαστικών επιφανειών, οι συνδεδεμένοι σύνδεσμοι LOS που διέρχονται από έναν ή περισσότερους ανακλαστήρες μπορούν να μοιράζονται την ίδια ζώνη συχνοτήτων, βελτιώνοντας έτσι την απόδοση του φάσματος. Συγκεκριμένα, ένας ευφυής ανακλαστήρας (Intelligent Reflector - IR) που αποτελείται από μεγάλο αριθμό παθητικών εξαρτημάτων χαμηλού κόστους μπορεί να πραγματοποιήσει διαμόρφωση δέσμης με μικρό ενεργειακό κόστος. Κάθε στοιχείο υπερύθρων αντανακλά τα προσπίπτοντα σήματα, ενώ συγκεντρώνει την ενέργεια ραδιοσυχνοτήτων (Radio Frequency - RF) από το μη ανακλώμενο κλάσμα των σημάτων για να τροφοδοτήσει το ίδιο. Προσαρμόζοντας από κοινού τις μετατοπίσεις φάσης, ένα IR μπορεί να εστιάσει το ανακλώμενο σήμα σε μια αιχμηρή δέσμη, μεγιστοποιώντας έτσι τα κέρδη απόδοσης διαμόρφωσης δέσμης.

Αρκετές πρόσφατες εργασίες έχουν μελετήσει τη χρήση των IR για τη βελτίωση της απόδοσης των κυψελοειδών δικτύων όπως στις [2] και [3] - [6]. Στο [2], παρουσιάζεται ο σχεδιασμός ενός παθητικού ανακλαστήρα και η εκτίμηση του κέρδους ανάκλασης για αστικές επικοινωνίες mmW. Στο [3], πραγματοποιούνται προσομοιώσεις και πειράματα σε εσωτερικά περιβάλλοντα για τη μέτρηση της μετάδοσης παθητικών ανακλαστήρων σε mmW. Οι συγγραφείς στο [4] βελτιστοποίησαν από κοινού τη διαμόρφωση δέσμης μετάδοσης από ένα σημείο πρόσβασης και τη διαμόρφωση ανακλαστικής δέσμης σε υπέρυθρο για να μεγιστοποιήσουν την ισχύ του λαμβανόμενου σήματος στον εξοπλισμό του χρήστη (User equipment - UE). Οι συγγραφείς στο [5] διεξήγαγαν μοντελοποίηση διάδοσης σημάτων terahertz χρησιμοποιώντας ανακλαστήρες. Παράλληλα, η εργασία στο [6] μελέτησε την ενεργειακή απόδοση των επικοινωνιών κατερχόμενης ζεύξης που υποβοηθούνται

από ανακλαστήρες. Ωστόσο, προηγούμενες εργασίες σε ανακλαστήρες mmW στα [2] και [3] επικεντρώνονται κυρίως σε πειραματικές μετρήσεις, ενώ οι εργασίες που σχετίζονται με το IR στο [4] - [6] μελετούν τις κυψελωτές επικοινωνίες σε φάσμα μη mmW. Επιπλέον, όλες οι προηγούμενες εργασίες στα [2] και [4] - [6] βασίζονται σε παθητικούς ανακλαστήρες τοποθετημένους σε σταθερή θέση, οι οποίοι δεν μπορούν να αντιμετωπίσουν τις δυναμικές αλλαγές των καναλιών mmW. Για παράδειγμα, μια απλή κίνηση του σώματος του UE μπορεί να προκαλέσει σημαντική απόφραξη mmW και να καταστήσει αναποτελεσματικό ένα στατικό IR.

Λόγω της επιρρεπούς σε απόφραξη φύσης των σημάτων mmW, οι κινητοί ανακλαστήρες είναι πιο κατάλληλοι για τη βελτίωση των επικοινωνιών mmW από τους σταθερούς ανακλαστήρες. Για παράδειγμα, μπορεί κανείς να χρησιμοποιήσει ένα UAV-carried IR (UAV-IR) που μπορεί να προσαρμόζει τη θέση του ανακλαστήρα συνεχώς, σύμφωνα με τις αλλαγές στο περιβάλλον, διατηρώντας έτσι συνδέσεις LOS τόσο με τον πομπό όσο και με τον δέκτη. Οι επικοινωνίες με τη βοήθεια UAV έχουν προσελκύσει σημαντική πρόσφατη προσοχή [7]–[9], ωστόσο, καμία προηγούμενη εργασία δεν έχει μελετήσει τη χρήση των UAV-IR. Συγκεκριμένα, ένα UAV-IR διαφέρει από ένα RS που υποστηρίζεται από UAV, λόγω της απλούστερης δομής κεραίας και της μικρότερης τροφοδοσίας, που το καθιστούν πιο κατάλληλο για ένα σενάριο πυκνής επικοινωνίας. Ενσωματωμένο σε ένα UAV, ένα IR μπορεί να βελτιώσει την αξιοπιστία των εκπομπών mmW βελτιστοποιώντας τη θέση του έξυπνα. Ωστόσο, μια τέτοια βελτιστοποίηση απαιτεί ακριβείς πληροφορίες κατάστασης καναλιού (Channel State Information - CSI) της σύνδεσης IR-UE. Λαμβάνοντας υπόψη την πιθανή κίνηση των UAV και UEs, καθώς και την επίδραση μπλοκαρίσματος του ανθρώπινου σώματος στα σήματα mmW, η τιμή του CSI σε πραγματικό χρόνο είναι δύσκολο να ληφθεί. Έτσι, για να καταστεί δυνατή η αποτελεσματική ανάπτυξη ενός UAV-IR για εκπομπές mmW, η πρόκληση της εκτίμησης CSI πρέπει να αντιμετωπιστεί σωστά.

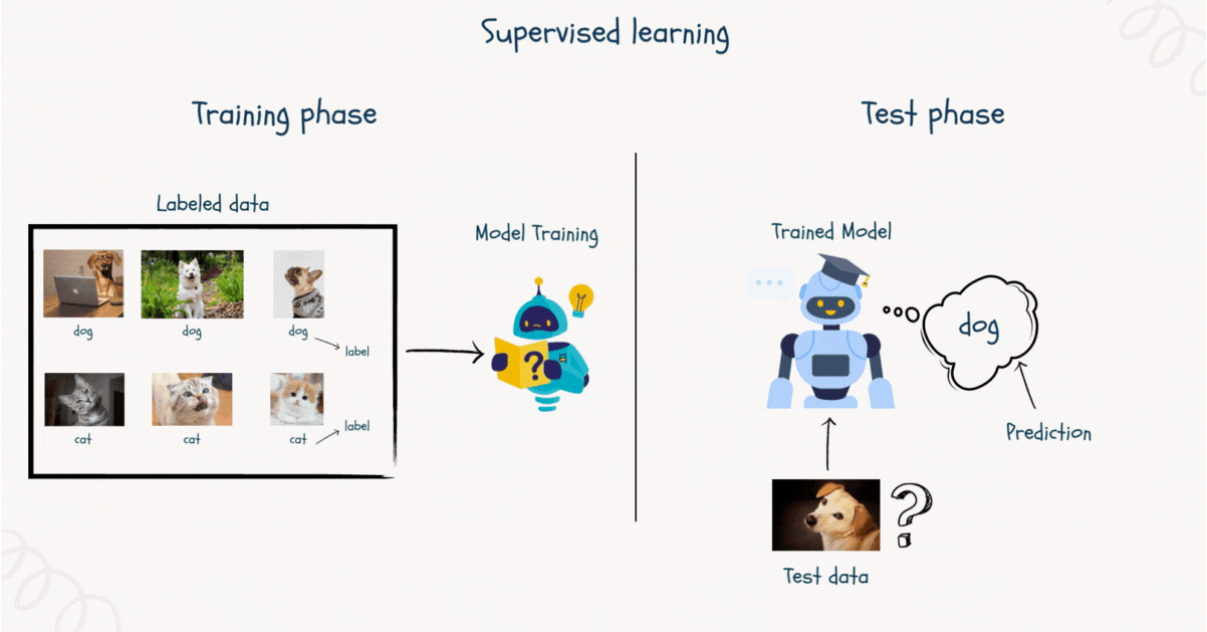
Στην παρούσα εργασία, θα γίνει μια περιγραφή της μηχανικής μάθησης, εξηγώντας το ρόλο της στις ασύρματες τηλεπικοινωνίες. Θα γίνει παρουσίαση διαφόρων αλγορίθμων που κατατάσσονται εν τέλει στους αλγορίθμους μηχανικής μάθησης για το πώς μπορούν οι αλγόριθμοι αυτοί να βοηθήσουν στην ενίσχυση του σήματος στα τηλεπικοινωνιακά κανάλια. Δεδομένου ότι για να ενισχυθεί το σήμα στις ασύρματες επικοινωνίες, πρέπει να παρθούν αποφάσεις ώστε να γίνει για παράδειγμα συγκεκριμένη εκχώρηση καναλιών, διάφοροι αλγόριθμοι όσο και η χρήση των UAVs θα συνδράμουν ώστε να επιτευχθεί η ενίσχυση αυτή πιο αποδοτικά. Θα γίνει σύγκριση των αποτελεσμάτων από διάφορους τέτοιους αλγορίθμους και θα φανεί από τη σύγκριση αυτή η σημασία της συνεισφοράς της μηχανικής μάθησης σε αυτό τον τομέα, αλλά και η βοήθεια των ανακλαστικών επιφανειών και των UAVs στην κατεύθυνση αυτή.

2 Μηχανική μάθηση

Η μηχανική μάθηση είναι ένα ισχυρό εργαλείο που επιτρέπει στους υπολογιστές να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις για νέες πληροφορίες, [14]. Όμως, δεν δημιουργείται όλη η μηχανική μάθηση ίσα. Υπάρχουν αρκετοί διαφορετικοί τύποι αλγορίθμων μηχανικής μάθησης, ο καθένας με τη δική του μοναδική προσέγγιση για την επίλυση προβλημάτων. Σε αυτό το κεφάλαιο, θα γίνει περιγραφή στους τρεις κύριους τύπους μηχανικής εκμάθησης για αρχάριους: εποπτευόμενη, μη εποπτευόμενη (ή αλλιώς μάθηση χωρίς επίβλεψη) και ενισχυτική μάθηση.

2.1 Εποπτευόμενη μάθηση – Supervised learning

Η εποπτευόμενη μάθηση είναι ο πιο βασικός και ευρέως χρησιμοποιούμενος τύπος μηχανικής μάθησης. Στην εποπτευόμενη μάθηση, ένα μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων όπου η σωστή έξοδος ή «ετικέτα» παρέχεται ήδη για κάθε είσοδο. Για παράδειγμα, ας υποθέσουμε ότι υπάρχει ένα σύνολο δεδομένων με εικόνες από γάτες και σκύλους. Οι ετικέτες για το σύνολο δεδομένων θα είναι "γάτα" και "σκύλος". Το μοντέλο μπορεί στη συνέχεια να χρησιμοποιήσει αυτές τις πληροφορίες για να κάνει προβλέψεις σχετικά με νέες φωτογραφίες γατών και σκύλων που δεν έχει δει ποτέ πριν.



Εικόνα 1: Παράδειγμα εποπτευόμενης μάθησης

Περιπτώσεις χρήσης εποπτευόμενης μάθησης

Μερικές συνήθεις περιπτώσεις χρήσης για εποπτευόμενη μάθηση είναι οι ακόλουθες:

- Ταξινόμηση εικόνας: Αναγνώριση αντικειμένων ή χαρακτηριστικών μέσα σε μια εικόνα, όπως η αναγνώριση χειρόγραφων ψηφίων ή ο προσδιορισμός εάν μια εικόνα περιέχει έναν συγκεκριμένο τύπο ζώου.
- Ταξινόμηση κειμένου: Κατηγοριοποίηση του κειμένου σε διαφορετικές κατηγορίες, όπως για παράδειγμα αν ένα email είναι ανεπιθύμητο ή όχι.

- Επεξεργασία φυσικής γλώσσας (NLP): Εργασίες όπως μετάφραση γλώσσας, σύνοψη κειμένου και ανάλυση συναισθημάτων.
- Προγνωστική μοντελοποίηση: Χρήση ιστορικών δεδομένων για την πραγματοποίηση προβλέψεων για μελλοντικά γεγονότα, όπως τιμές μετοχών ή καιρικά μοτίβα.
- Ιατρική διάγνωση: Αναγνώριση ασθενειών από ιατρικές εικόνες, ΗΚΓ ή άλλα ζωτικά σημεία
- Ανίχνευση αντικειμένων σε βίντεο και εικόνες : Προσδιορίστε και εντοπίστε αντικείμενα σε βίντεο και εικόνες.

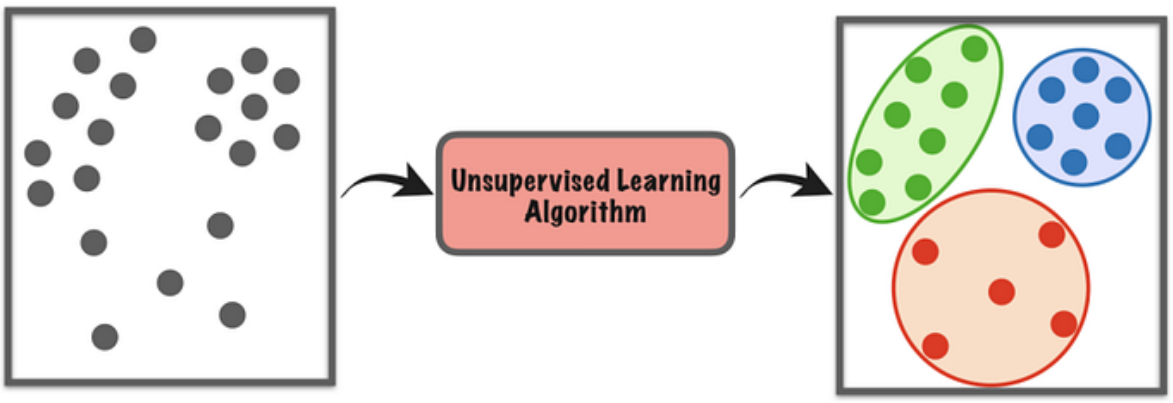
Αλγόριθμοι εποπτευόμενης μάθησης

Υπάρχουν αρκετοί δημοφιλείς αλγόριθμοι εποπτευόμενης μάθησης που χρησιμοποιούνται ευρέως στον τομέα της μηχανικής μάθησης. Μερικοί από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους αλγόριθμους περιλαμβάνουν:

- Γραμμικής παλινδρόμησης
- Logistic Regression
- Δέντρα απόφασης
- Τυχαίο Δάσος
- Υποστήριξη Vector Machines (SVM)
- k-Κοντινότεροι γείτονες (kNN)

2.2 Μη εποπτευόμενη μάθηση – Unsupervised learning

Η μάθηση χωρίς επίβλεψη, από την άλλη πλευρά, είναι όταν δίνεται στο μοντέλο ένα σύνολο δεδομένων χωρίς ετικέτες ή έξοδο. Το μοντέλο πρέπει στη συνέχεια να βρει μοτίβα και δομή μέσα στα δεδομένα από μόνο του. Ένα κοινό παράδειγμα μάθησης χωρίς επίβλεψη είναι η ομαδοποίηση, όπου ένα μοντέλο ομαδοποιεί παρόμοια σημεία δεδομένων μαζί. Φανταστείτε ότι έχετε ένα σύνολο δεδομένων με δεδομένα πελατών. Το μοντέλο θα ομαδοποιούσε τους πελάτες με βάση παρόμοια χαρακτηριστικά όπως η ηλικία, η τοποθεσία και οι συνήθειες δαπανών.



Εικόνα 2: Παράδειγμα μη εποπτευόμενης μάθησης

Περιπτώσεις μη εποπτευόμενης μάθησης

Η μάθηση χωρίς επίβλεψη χρησιμοποιείται για μια ποικιλία εργασιών, όπως:

- Ομαδοποίηση: Ομαδοποίηση παρόμοιων σημείων δεδομένων μαζί, όπως ομαδοποίηση πελατών με βάση τα μοτίβα αγορών τους.
- Μείωση διαστάσεων: Μείωση του αριθμού των χαρακτηριστικών σε ένα σύνολο δεδομένων, όπως η αναγνώριση των πιο σημαντικών χαρακτηριστικών σε ένα σύνολο εικόνων.
- Ανίχνευση ανωμαλιών: Προσδιορισμός ασυνήθιστων ή μη φυσιολογικών σημείων δεδομένων, όπως η ανίχνευση απάτης σε χρηματοοικονομικές συναλλαγές.
- Μοντέλα δημιουργίας: Δημιουργία νέων δεδομένων που είναι παρόμοια με τα δεδομένα εισόδου, όπως η δημιουργία νέων εικόνων ή κειμένου.

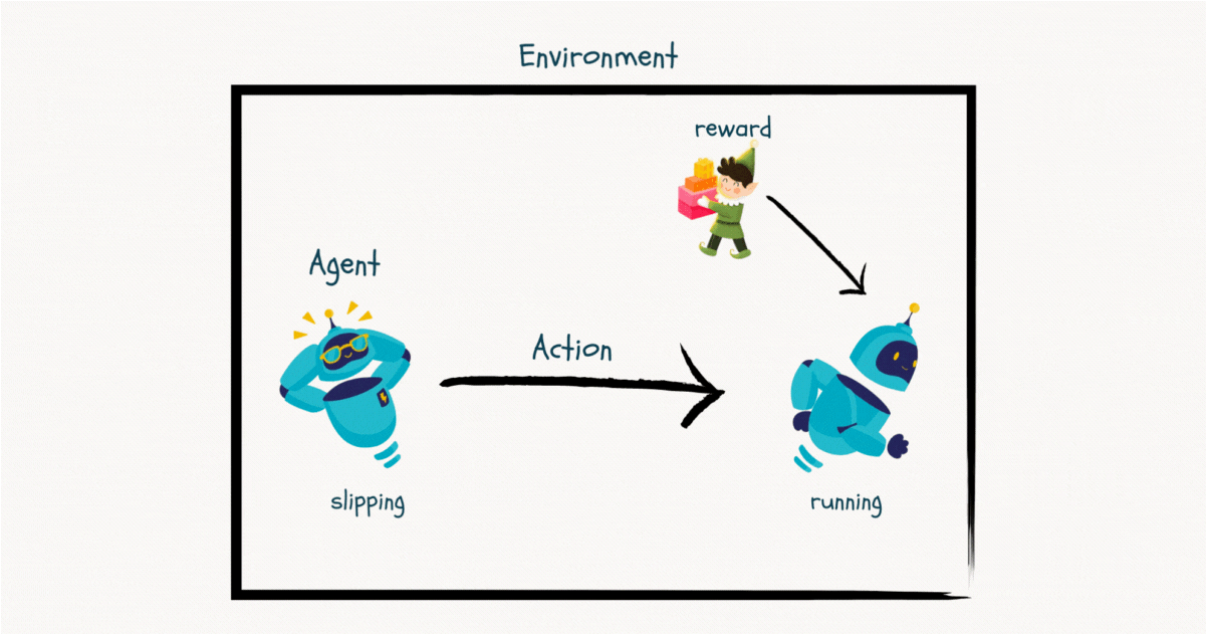
Αλγόριθμοι μη εποπτευόμενης μάθησης

Υπάρχουν διάφοροι δημοφιλείς αλγόριθμοι μάθησης χωρίς επίβλεψη:

- K-means
- Ιεραρχική Ομαδοποίηση
- PCA (Ανάλυση κύριου στοιχείου)
- t-SNE (t-Distributed Stochastic Neighbor Embedding)

2.3 Ενισχυτική μάθηση – Reinforcement learning

Η ενισχυτική μάθηση (reinforcement learning) είναι λίγο διαφορετική από την εποπτευόμενη και χωρίς επίβλεψη μάθηση. Στην ενισχυτική μάθηση, το μοντέλο μαθαίνει από τις συνέπειες των πράξεών του. Το μοντέλο λαμβάνει ανατροφοδότηση για την απόδοσή του και χρησιμοποιεί αυτές τις πληροφορίες για να προσαρμόσει τις ενέργειές του και να βελτιώσει την απόδοσή του με την πάροδο του χρόνου. Ένα κλασικό παράδειγμα ενισχυτικής μάθησης είναι η εκπαίδευση ενός μοντέλου για να παίξει ένα παιχνίδι όπως το σκάκι ή το Go. Το μοντέλο λαμβάνει ανατροφοδότηση για την απόδοσή του με τη μορφή νίκης ή απώλειας και στη συνέχεια προσαρμόζει τη στρατηγική του για να βελτιώσει τις πιθανότητές του να κερδίσει.



Εικόνα 3: Παράδειγμα ενισχυτικής μάθησης

Περιπτώσεις χρήσης Ενισχυτικής Μάθησης

Η ενισχυτική μάθηση χρησιμοποιείται σε μια ποικιλία εργασιών, όπως:

- Ρομποτική: Εκπαίδευση ρομπότ να εκτελούν συγκεκριμένες εργασίες, όπως να πιάνουν αντικείμενα ή να κινούνται μέσα σε έναν λαβύρινθο.
- Παιχνίδι: Εκπαίδευση πράκτορες να παίζουν παιχνίδια όπως σκάκι, Go και πόκερ σε υπεράνθρωπο επίπεδο.
- Αυτόνομα οχήματα: Εκπαίδευση αυτοοδηγούμενων αυτοκινήτων για πλοήγηση στην κυκλοφορία και λήψη ασφαλών αποφάσεων οδήγησης.
- Βιομηχανικός έλεγχος: Βελτιστοποίηση του ελέγχου βιομηχανικών συστημάτων, όπως σταθμοί ηλεκτροπαραγωγής ή διαδικασίες παραγωγής.
- Υγειονομική περίθαλψη: Ανάπτυξη εξατομικευμένων σχεδίων θεραπείας για ασθενείς με βάση το ιατρικό ιστορικό και την ανταπόκρισή τους στη θεραπεία.
- Οικονομικά: Ανάπτυξη στρατηγικών συναλλαγών και διαχείριση χαρτοφυλακίου.

Αλγόριθμοι Ενισχυτικής Μάθησης

Υπάρχουν διάφοροι δημοφιλείς αλγόριθμοι ενίσχυσης μάθησης:

- DQN
- PPO
- Q-Learning
- SARSA
- A3C

Συνοπτικά, η εποπτευόμενη μάθηση είναι όταν ένα μοντέλο παρέχεται με δεδομένα με ετικέτα, η μάθηση χωρίς επίβλεψη είναι όταν το μοντέλο βρίσκει μοτίβα μέσα σε δεδομένα χωρίς ετικέτα και η ενισχυτική μάθηση είναι όταν το μοντέλο μαθαίνει από τις συνέπειες των πράξεών του. Η κατανόηση των διαφορών μεταξύ αυτών των τύπων μάθησης μπορεί να βοηθήσει στην επιλογή του σωστού αλγορίθμου για το εκάστοτε πρόβλημα και να βελτιώσει την απόδοση του μοντέλου. Θα πρέπει να ληφθεί υπόψη ότι και οι τρεις τύποι μηχανικής μάθησης έχουν τη δική τους μοναδική προσέγγιση για την επίλυση προβλημάτων και ο καθένας έχει το δικό του σύνολο δυνατών και αδυναμιών.

2.4 Προγενέστερες μελέτες

Σε προγενέστερες μελέτες, ερευνητές έχουν εστιάσει στις διαφορές που έχουν ορισμένοι αλγόριθμοι ενισχυτικής μάθησης, στη χρήση της για τη βελτίωση ανερχόμενης ή κατερχόμενης ζεύξης σε δυναμικά περιβάλλοντα, καθώς και στην εύρεση προσεγγιστικών μεθόδων για τη μείωση της ισχύος πτήσης στα μη επανδρωμένα οχήματα. Πιο συγκεκριμένα, από τις σημαντικότερες βιβλιογραφίες που μελετήθηκαν, αναλύθηκε ένα νέο πλαίσιο για την αποτελεσματική ανάπτυξη ενός UAV-IR για να υποβοηθήσει τη μετάδοση κατερχόμενης ζεύξης mmW σε ένα δυναμικό περιβάλλον με κινούμενα UE. Για τη διατήρηση ενός καναλιού LOS, μια προσέγγιση ενισχυτικής μάθησης (RL), βασισμένη σε Q-learning και νευρωνικά δίκτυα, προτείνεται για τη μοντελοποίηση του περιβάλλοντος διάδοσης, έτσι ώστε η θέση και ο συντελεστής ανάκλασης του UAV-IR να μπορούν να βελτιστοποιηθούν ώστε να μεγιστοποιηθεί η ικανότητα μετάδοσης κατερχόμενης ζεύξης. Παράλληλα, προτείνεται η χρήση της συλλογής ενέργειας ραδιοσυχνοτήτων για την αυτοτροφοδότηση του IR.

Τα αποτελέσματα της προσομοίωσης δείχνουν την αποτελεσματικότητα της προτεινόμενης προσέγγισης που βασίζεται σε UAV-IR σε σύγκριση με ένα στατικό IR. Στη μελέτη αυτή είναι η πρώτη φορά που προτείνεται μια ανάπτυξη UAV-IR βασισμένη στη μάθηση για επικοινωνίες mmW με συλλογή ενέργειας ραδιοσυχνοτήτων.

Αντίστοιχα, σε άλλη σημαντική μελέτη, γίνεται εφαρμογή μετρήσεων από μαθηματικά μοντέλα και από πραγματικές μετρήσεις για τον υπολογισμό της λαμβανόμενης ισχύς σήματος σε UAV από το κυψελοειδές δίκτυο κατά τη διάρκεια της πτήσης, όπου χρησιμοποιούνται νευρωνικά δίκτυα για την εξαγωγή χαρακτηριστικών από τα κανάλια UAV ως μια σειρά μπλοκ για τη μοντελοποίηση καναλιών. Επίσης, γίνεται ανάπτυξη μιας προσέγγισης προσαρμοστικού ρυθμού μάθησης και μια νέα βελτιωμένη πρακτικής για τη βελτίωση της απόδοσης της εκπαίδευσης. Συγκεκριμένα, ένας αυτόματος κωδικοποιητής χρησιμοποιείται για τη βελτιστοποίηση μιας συγκεκριμένης παραμέτρου, ενώ οι υπόλοιπες παράμετροι εκπαιδεύονται χρησιμοποιώντας ένα νευρωνικό δίκτυο κωδικοποιητή για την πρόβλεψη του εδάφους σε διαφορετικά ύψη. Τέλος, πραγματοποιείται επαλήθευση της αποτελεσματικότητα της προτεινόμενης μεθόδου με πειραματικές μετρήσεις και συγκρίσεις με άλλα σχήματα αναφοράς. Τα αριθμητικά αποτελέσματα δείχνουν ότι το προτεινόμενο σχήμα υπερτερεί των συμβατικών αυτοκωδικοποιητών.

2.5 Σκοπός της παρούσας μελέτης και εργασίας

Σκοπός της παρούσας εργασίας είναι να παρουσιαστεί ένα αρκετά καλό θεωρητικό υπόβαθρο της μηχανικής μάθησης, περιγράφοντας τις διαφοροποιήσεις της ενισχυτικής μάθησης έναντι της εποπτευόμενης και μη εποπτευόμενης μάθησης. Στη συνέχεια, θα γίνει μια αναλυτική περιγραφή της ενισχυτικής μάθησης, τονίζοντας τις διαφορές της από τις άλλες δύο προαναφερθείσες κατηγορίες της μηχανικής μάθησης, προκειμένου να γίνει σαφές για πιο λόγο είναι προτιμητέα στις τηλεπικοινωνίες.

Έχοντας ήδη κάνει την εισαγωγή της εποπτευόμενης μάθησης, περιγράφονται στη συνέχεια οι κυριότεροι αλγόριθμοι DQN και PPO, χρησιμοποιώντας διάφορα παραδείγματα κώδικα και σχολιάζοντας λεπτομέρειες πάνω σε αυτά, προκειμένου ο αναγνώστης να μπορεί να διακρίνει καλύτερα τις διαφορές τους στη λειτουργία τους. Συνάμα, περιγράφονται οι ανακλαστικές επιφάνειες και τα μη επανδρωμένα οχήματα, ώστε να φανεί πώς μπορούν αλγόριθμοι ενισχυτικής μάθησης να εφαρμοστούν σε αυτά, ώστε να μπορέσουν να συμβάλουν στην αρτιότερη λειτουργία τους, είτε στον τομέα της κάλυψης είτε στη βελτίωση των υπηρεσιών τους. Σε κάθε ευκαιρία γίνονται προτάσεις βελτίωσης της παρούσας μελέτης, ώστε να υπάρχουν εναύσματα για περαιτέρω έρευνα και μελέτη.

Στο κεφάλαιο αυτό έγινε περιγραφή των αλγορίθμων μηχανικής μάθησης, και πιο συγκεκριμένα της εποπτευόμενης μάθησης, της μη εποπτευόμενης μάθησης καταλήγοντας στην ενισχυτική μάθηση και τις αρχές λειτουργίας τους. Στο επόμενο κεφάλαιο γίνεται ανάλυση των ανακλαστικών επιφανειών και των μη επανδρωμένων οχημάτων στις τηλεπικοινωνίες, αναφέροντας και ορισμένους τρόπους εφαρμογής τους στις τηλεπικοινωνίες. Συνακόλουθα, γίνεται μια συγκριτική μελέτη των κυριότερων αλγορίθμων ενισχυτικής μάθησης, DQN και PPO, οι οποίοι υλοποιούνται σε

κάποια παραδείγματα με τη χρήση της Python. Στη συνέχεια, παρατίθεται μια εφαρμογή νευρωνικών δικτύων για την εκτίμηση καναλιών σε τηλεπικοινωνιακά συστήματα με μη επανδρωμένα οχήματα, κλείνοντας με προτάσεις βελτίωσης για τον αναγνώστη.

3 Ανάλυση των ανακλαστικών επιφανειών και των UAVs στις τηλεπικοινωνίες

Στην ενότητα αυτή γίνεται περιγραφή των ανακλαστικών επιφανειών στις ασύρματες επικοινωνίες, οι οποίες είναι και μια πρώτη προσέγγιση για να αυξηθεί και η εμβέλειά τους. Γίνεται σαφές ότι μέσω των έξυπνων ανακλαστικών επιφανειών, τόσο τα δίκτυα 5ης γενιάς όσο και άλλα δίκτυα επωφελούνται για τη μείωση της ενέργειας που καταναλώνουν, ενώ παράλληλα αυξάνουν το εύρος της κάλυψής τους. Συνακόλουθα, στο κεφάλαιο αυτό περιγράφεται και η συμβολή των μη επανδρωμένων εναέριων οχημάτων στις τηλεπικοινωνίες, τα οποία με διάφορες μεθόδους μπορούν να αυξήσουν και αυτά την εμβέλεια ή τη χωρητικότητα ενός ασύρματου δικτύου.

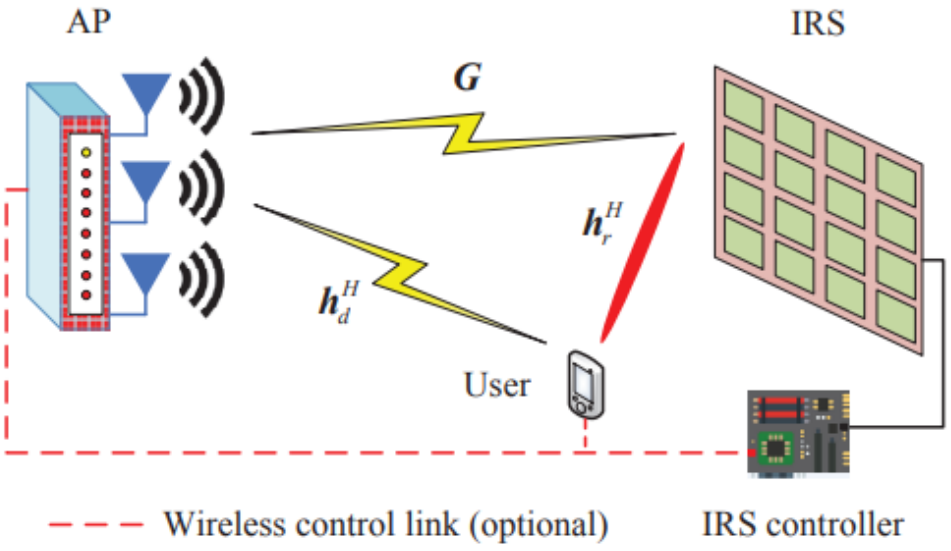
3.1 Αρχή λειτουργίας των ανακλαστικών επιφανειών στις ασύρματες επικοινωνίες

Παρόλο που έχει σημειωθεί ένα τεράστιο άλμα στην απόδοση του φάσματος των ασύρματων δικτύων τις τελευταίες δεκαετίες χάρη σε διάφορες τεχνολογικές προόδους όπως το εξαιρετικά πυκνό δίκτυο (Ultra Dense Network - UDN), η μαζική πολλαπλή έξοδος πολλαπλών εισόδων (M-MIMO) και το κύμα χιλιοστού (mmWave), η κατανάλωση ενέργειας του δικτύου και το κόστος υλικού εξακολουθούν να είναι κρίσιμα ζητήματα που αντιμετωπίζονται στην πρακτική εφαρμογή, [15]. Για παράδειγμα, τα UDN κλιμακώνουν σχεδόν γραμμικά το κύκλωμα και την κατανάλωση ενέργειας ψύξης με τον αριθμό των νέων σταθμών βάσης (Base Stations – BS), ενώ απαιτούνται δαπανηρές αλυσίδες ραδιοσυχνοτήτων (RF) και πολύπλοκες τεχνικές επεξεργασίας σήματος για αποτελεσματική επικοινωνία σε συχνότητες mmWave. Από την άλλη πλευρά, η προσθήκη υπερβολικά μεγάλου αριθμού ενεργών στοιχείων όπως BS που λειτουργούν ως αναμεταδότες μικρών κυψελών σε ασύρματα δίκτυα προκαλεί επίσης ένα πιο σοβαρό ζήτημα παρεμβολών. Ως εκ τούτου, η έρευνα για την εύρεση τόσο φασματικών όσο και ενεργειακά αποδοτικών τεχνικών με χαμηλό κόστος υλικού εξακολουθεί να είναι επιτακτική για την υλοποίηση βιώσιμων και πράσινων ασύρματων δικτύων πέμπτης γενιάς (5G) και όχι μόνο.

Η έξυπνη ανακλαστική επιφάνεια (Intelligent Reflective Surface - IRS) προτείνεται ως μια πολλά υποσχόμενη πράσινη και οικονομικά αποδοτική λύση για την επίτευξη των παραπάνω απαιτητικών στόχων. Συγκεκριμένα, το IRS είναι μια επίπεδη διάταξη που αποτελείται από ένα μεγάλο αριθμό παθητικών στοιχείων (π.χ. τυπωμένα δίπολα χαμηλού κόστους), όπου κάθε στοιχείο μπορεί να προκαλέσει μια συγκεκριμένη μετατόπιση φάσης (από έναν έξυπνο ελεγκτή) ανεξάρτητα από το προσπίπτον ηλεκτρομαγνητικό κύμα. Ως βασικό συστατικό των συμβατικών ανακλαστικών συστοιχιών, η παθητική ανακλαστική επιφάνεια έχει βρει μια ποικιλία εφαρμογών σε ραντάρ και δορυφορικές επικοινωνίες, η οποία, ωστόσο, σπάνια χρησιμοποιείται στην επίγεια ασύρματη επικοινωνία. Αυτό οφείλεται στο γεγονός ότι οι παραδοσιακές ανακλαστικές επιφάνειες έχουν μόνο σταθερούς μετατοπιστές φάσης μόλις κατασκευαστούν, οι οποίοι είναι δύσκολο να ανταποκριθούν στη δυναμική των ασύρματων δικτύων με κανάλια που μεταβάλλονται χρονικά. Ωστόσο, οι πρόσφατες εξελίξεις στα μικροηλεκτρομηχανικά συστήματα ραδιοσυχνοτήτων (MEMS) και στο μετα-υλικό (π.χ. μετα-επιφάνεια) κατέστησαν δυνατή την επαναδιαμόρφωση των ανακλώσιμων επιφανειών, ακόμη και μέσω του ελέγχου των μετατοπιστών φάσης σε

πραγματικό χρόνο. Προσαρμόζοντας έξυπνα τις μετατοπίσεις φάσης όλων των στοιχείων σε ένα IRS, τα ανακλώμενα σήματα μπορούν να προστεθούν με συνέπεια στον επιθυμητό δέκτη για να βελτιώσουν την ισχύ του λαμβανόμενου σήματος ή καταστροφικά στον μη προβλεπόμενο δέκτη για να αποφευχθούν παρεμβολές και να ενισχυθεί η ασφάλεια/απόρρητο.

Αξίζει να σημειωθεί ότι το προτεινόμενο IRS διαφέρει σημαντικά από άλλες υπάρχουσες σχετικές τεχνολογίες, όπως η αναμετάδοση ενίσχυσης και προώθησης (Amplify-And Forward - AF), η επικοινωνία backscatter και η ενεργή ευφυή επιφάνεια M-MIMO. Πρώτον, σε σύγκριση με την αναμετάδοση AF που βοηθά στη μετάδοση πηγής-προορισμού δημιουργώντας ενεργά νέα σήματα, το IRS δεν χρησιμοποιεί μια μονάδα πομπού, αλλά αντανακλά μόνο τα σήματα ραδιοσυχνοτήτων περιβάλλοντος ως παθητική συστοιχία, η οποία επομένως δεν προκαλεί πρόσθετη κατανάλωση ενέργειας. Δεύτερον, διαφορετική από την παραδοσιακή οπισθοσκέδασης της ετικέτας αναγνώρισης ραδιοσυχνοτήτων (RFID) που επικοινωνεί με τον δέκτη αντανακλώντας το προσπίπτον κύμα που αποστέλλεται από τον αναγνώστη, το IRS χρησιμοποιείται για να βελτιώσει την απόδοση της υπάρχουσας ζεύξης επικοινωνίας αντί να παρέχει οποιαδήποτε δική του πληροφορία. Ως εκ τούτου, το σήμα άμεσης διαδρομής (από τον αναγνώστη στον δέκτη) στις επικοινωνίες οπισθοσκέδασης είναι η ανεπιθύμητη παρεμβολή και ως εκ τούτου πρέπει να ακυρωθεί/κατασταλεί στον δέκτη. Ωστόσο, στις επικοινωνίες IRSenhanced, τόσο τα σήματα άμεσης διαδρομής όσο και τα σήματα διαδρομής ανάκλασης μεταφέρουν τις ίδιες χρήσιμες πληροφορίες και επομένως θα πρέπει να προστίθενται με συνέπεια στον δέκτη για να μεγιστοποιηθεί η συνολική λαμβανόμενη ισχύς. Τρίτον, το IRS διαφέρει επίσης από το ενεργό ευφυές M-MIMO που βασίζεται στην επιφάνεια λόγω των διαφορετικών αρχιτεκτονικών συστοιχιών (παθητικό έναντι ενεργού) και των μηχανισμών λειτουργίας (ανάκλαση έναντι μετάδοσης). Επιπλέον, τα IRS διαθέτουν άλλα πλεονεκτήματα, όπως χαμηλό προφίλ, ελαφριά και ομοιόμορφη γεωμετρία, που τους επιτρέπουν να προσαρτώνται/αφαιρούνται εύκολα στον τοίχο ή την οροφή, παρέχοντας έτσι υψηλή ευελιξία και ανώτερη συμβατότητα για πρακτική εφαρμογή. Για παράδειγμα, με την εγκατάσταση IRS στους τοίχους που βρίσκονται σε οπτική επαφή (LoS) ενός σημείου πρόσβασης (Access Point - AP) σε κάποιο BS, η ισχύς του σήματος και η κάλυψή του αναμένεται να βελτιωθούν σημαντικά. Όλα τα παραπάνω πλεονεκτήματα καθιστούν τα IRS μια ελκυστική λύση για βελτίωση της απόδοσης σε ασύρματα δίκτυα μελλοντικής γενιάς, ειδικά για εφαρμογές εσωτερικού χώρου με υψηλή πυκνότητα χρηστών π.χ. στάδια, εμπορικά κέντρα, εκθεσιακά κέντρα και αεροδρόμια. Ωστόσο, η έρευνα σχετικά με το σχεδιασμό IRS και τη βελτιστοποίηση απόδοσης βρίσκεται σε αρχικό στάδιο και έχει γίνει πολύ περιορισμένη εργασία σε αυτόν τον νέο τομέα.



Εικόνα 4: Ένα ασύρματο σύστημα ενισχυμένο με IRS

Σε αυτό το κεφάλαιο, εξετάζεται ένα ασύρματο σύστημα ενισχυμένο με IRS όπως φαίνεται στην Εικόνα 4, όπου ένα AP πολλαπλών κεραιών εξυπηρετεί έναν χρήστη με μία κεραία με τη βοήθεια ενός IRS (π.χ. στον τοίχο). Ένα τέτοιο σύστημα μπορεί να χρησιμοποιηθεί για τη διευκόλυνση της ασύρματης μεταφοράς πληροφοριών και/ή της μεταφοράς ενέργειας σε διάφορες εφαρμογές Internet-of-things (IoT). Δεδομένου ότι ο χρήστης λαμβάνει τα υπερτιθέμενα σήματα τόσο από τη σύνδεση AP-χρήστη (απευθείας) όσο και από τη σύνδεση IRS-χρήστη, βελτιστοποιούμε από κοινού την (ενεργητική) διαμόρφωση δέσμης εκπομπής στο AP και ανακλά παθητικά τη διαμόρφωση δέσμης από τους μετατοπιστές φάσης στο IRS για να μεγιστοποιηθεί έτσι η συνολική ισχύς σήματος που λαμβάνεται από τον χρήστη. Διαισθητικά, εάν το κανάλι της σύνδεσης AP-χρήστη είναι πολύ ισχυρότερο από αυτό της σύνδεσης AP-IRS, είναι προτιμότερο το AP να μεταδίδεται απευθείας στον χρήστη, ενώ στην αντίθετη περίπτωση, ειδικά όταν ο σύνδεσμος AP-χρήστη είναι μπλοκαρισμένο μερικώς από π.χ. εμπόδια, όπως συναντάται συχνά σε εφαρμογές εσωτερικού χώρου, το AP προσαρμόζει την κατεύθυνση διαμόρφωσης της δέσμης του προς το IRS για να αξιοποιήσει το ανακλώμενο σήμα του για να εξυπηρετήσει το χρήστη. Σε αυτήν την περίπτωση, ένας μεγάλος αριθμός έξυπνα ρυθμιζόμενων ανακλαστικών στοιχείων στο IRS μπορεί να εστιάσει την ενέργεια του σήματος σε μια ισχυρότερη δέσμη προς το χρήστη, η οποία επιτυγχάνει υψηλό κέρδος διαμόρφωσης δέσμης όπως στο M-MIMO, με χρήση μόνο μιας παθητικής συστοιχίας, εξοικονομώντας έτσι σημαντική ενέργεια. Γενικά, η διαμόρφωση δέσμης εκπομπής στο AP πρέπει να σχεδιαστεί από κοινού με τις μετατοπίσεις φάσης στο IRS με βάση όλα τα κανάλια AP-IRS, χρήστη IRS και χρήστη AP, προκειμένου να αποκομιστούν πλήρως τα κέρδη δέσμης τους. Ωστόσο, το διαμορφωμένο πρόβλημα βελτιστοποίησης φαίνεται να είναι μη κυρτό/γραμμικό και δύσκολο να λυθεί με βέλτιστο τρόπο.

Για την αντιμετώπιση της μη κυρτότητας του εξεταζόμενου προβλήματος, στο [15] έχουν προταθεί συγκεντρωτικοί αλγόριθμοι που βασίζονται στην τεχνική της ημικαθορισμένης χαλάρωσης (SDR) για να ληφθεί τόσο ένα άνω όριο απόδοσης όσο και μια κατά προσέγγιση λύση υψηλής ποιότητας. Μια τέτοια συγκεντρωτική

υλοποίηση απαιτεί τις πληροφορίες κατάστασης καθολικού καναλιού (CSI) που είναι διαθέσιμες στο IRS, και συνεπώς συνεπάγεται υπερβολικά έξοδα εκτίμησης καναλιού και ανταλλαγής σήματος στο/μεταξύ του AP και του IRS. Για τη μείωση τέτοιων γενικών εξόδων και την επίτευξη χαμηλής πολυπλοκότητας, προτείνεται περαιτέρω ένας κατανεμημένος αλγόριθμος, εμπνευσμένος από την εναλλασσόμενη βελτιστοποίηση. Η βασική ιδέα είναι ότι το AP και το IRS προσαρμόζουν ανεξάρτητα τη διαμόρφωση της δέσμης εκπομπής και τις μετατοπίσεις φάσης με εναλλασσόμενο τρόπο μέχρι να επιτευχθεί η σύγκλιση. Με την προσομοίωση φαίνεται ότι η αναλογία σήματος προς θόρυβο ζεύξης (SNR) μπορεί να βελτιωθεί σημαντικά με την ανάπτυξη του IRS σε σύγκριση με τη συμβατική εγκατάσταση χωρίς το IRS. Επιπλέον, με τον προτεινόμενο σχεδιασμό μορφοποίησης δέσμης, φαίνεται ότι το SNR λήψης κοντά στο IRS αυξάνεται με τον αριθμό των ανακλαστικών του στοιχείων N της τάξης του N^2, πράγμα που σημαίνει ότι σημαντική εξοικονόμηση ενέργειας στο AP ή SNR κέρδος ο χρήστης μπορεί να επιτευχθεί στην πράξη.

Στη συγκεκριμένη βιβλιογραφία, έχει προταθεί μια νέα προσέγγιση για τη βελτίωση της απόδοσης των ασύρματων δικτύων με την ανάπτυξη παθητικών IRS. Συγκεκριμένα, η ενεργητική δέσμη μετάδοσης στο AP και η παθητική ανακλαστική δέσμη των μετατροπέων φάσης στο IRS βελτιστοποιούνται από κοινού για να μεγιστοποιήσουν την ισχύ του σήματος που λαμβάνεται από τον χρήστη σε ένα σύστημα MISO από σημείο σε σημείο, ενισχυμένο με IRS. Αξιοποιώντας τις τεχνικές SDR και εναλλασσόμενης βελτιστοποίησης, αντίστοιχα, προτείνονται τόσο κεντρικά όσο και κατανεμημένα μοντέλα. Ειδικότερα, ο κατανεμημένος σχεδιασμός χαμηλής πολυπλοκότητας έχει αποδειχθεί ότι επιτυγχάνει σχεδόν βέλτιστη απόδοση και επομένως είναι ελκυστικός για πρακτική εφαρμογή. Τα αποτελέσματα της προσομοίωσης καταδεικνύουν επίσης τη βελτίωση του SNR και την επέκταση κάλυψης σήματος που επιτυγχάνεται με την ανάπτυξη του IRS σε σύγκριση με τη συμβατική εγκατάσταση χωρίς το IRS. Οι προτεινόμενοι συνδυασμοί διαμόρφωσης δέσμης AP και IRS αποδεικνύονται επίσης αποτελεσματικοί και κρίσιμοι για την επίτευξη βέλτιστης απόδοσης κάτω από διαφορετικές ρυθμίσεις.

3.2 Τρόποι εφαρμογής των UAVs στις τηλεπικοινωνίες

Στην ενότητα αυτή θα παρατεθούν βιβλιογραφικές πηγές και έρευνες σχετικές με την εφαρμογή των μη επανδρωμένων εναέριων οχημάτων στις τηλεπικοινωνίες.

3.2.1 Κινητά μη επανδρωμένα εναέρια οχήματα (UAV) για ενεργειακά αποδοτικές επικοινωνίες Internet of Things

Η χρήση μη επανδρωμένων εναέριων οχημάτων (UAV) ως ιπτάμενες πλατφόρμες ασύρματης επικοινωνίας έχει λάβει σημαντική προσοχή πρόσφατα, [16]. Από τη μία πλευρά, τα UAV μπορούν να χρησιμοποιηθούν ως ασύρματοι αναμεταδότες για τη βελτίωση της συνδεσιμότητας και της κάλυψης ασύρματων συσκευών γείωσης. Από την άλλη πλευρά, τα UAV μπορούν να λειτουργήσουν ως κινητοί εναέριοι σταθμοί βάσης για να παρέχουν αξιόπιστες επικοινωνίες κατερχόμενης και άνω ζεύξης για χρήστες εδάφους και να ενισχύσουν τη χωρητικότητα των ασύρματων δικτύων. Σε σύγκριση με τους επίγειους σταθμούς βάσης, το πλεονέκτημα της χρήσης εναέριων σταθμών βάσης που βασίζονται σε UAV είναι η ικανότητά τους να παρέχουν επικοινωνίες on the fly, δηλαδή ανάλογα με τις απαιτήσεις εκείνη τη στιγμή. Επιπλέον, το μεγάλο υψόμετρο των UAV τους επιτρέπει να δημιουργούν αποτελεσματικά

συνδέσμους επικοινωνίας οπτικής επαφής (LoS), μετριάζοντας έτσι το μπλοκάρισμα και τη σκίαση του σήματος. Λόγω του ρυθμιζόμενου ύψους και της κινητικότητάς τους, τα UAV μπορούν να κινηθούν προς πιθανούς χρήστες εδάφους και να δημιουργήσουν αξιόπιστες συνδέσεις με χαμηλή ισχύ μετάδοσης. Ως εκ τούτου, μπορούν να παρέχουν μια οικονομικά αποδοτική και ενεργειακά αποδοτική λύση για τη συλλογή δεδομένων από επίγειους χρήστες κινητών που είναι κατανεμημένοι σε μια γεωγραφική περιοχή με περιορισμένη επίγεια υποδομή.

Πράγματι, τα UAV μπορούν να διαδραματίσουν βασικό ρόλο στο Internet of Things (IoT), το οποίο αποτελείται από μικρές συσκευές περιορισμένης μπαταρίας, όπως αισθητήρες και οθόνες υγείας. Αυτές οι συσκευές συνήθως δεν μπορούν να εκπέμπουν σε μεγάλη απόσταση λόγω των ενεργειακών περιορισμών τους. Σε τέτοια σενάρια IoT, τα UAV μπορούν να κινηθούν δυναμικά προς τις συσκευές IoT, να συλλέξουν τα δεδομένα του IoT και να τα μεταδώσουν σε άλλες συσκευές που βρίσκονται εκτός του εύρους επικοινωνίας των πομπών. Σε αυτή την περίπτωση, τα UAV παίζουν το ρόλο των κινούμενων συσσωρευτών ή σταθμών βάσης για δίκτυα IoT. Ωστόσο, για να χρησιμοποιηθούν αποτελεσματικά τα UAV για το IoT, πρέπει να αντιμετωπιστούν αρκετές προκλήσεις, όπως η βέλτιστη ανάπτυξη, η κινητικότητα και η ενεργειακά αποδοτική χρήση των UAV.

Όπως περιγράφεται και στο [16], σε κάποιες εργασίες οι συγγραφείς ερεύνησαν τη βέλτιστη τροχιά των UAV εξοπλισμένων με πολλαπλές κεραίες για τη μεγιστοποίηση του αθροίσματος του ρυθμού στις επικοινωνίες άνω ζεύξης. Αντίστοιχα, σε κάποια άλλη εργασία, μεγιστοποιείται η απόδοση ενός συστήματος UAV που βασίζεται σε αναμετάδοση βελτιστοποιώντας από κοινού την τροχιά του UAV καθώς και την ισχύ εκπομπής πηγής/αναμεταδότη. Ωστόσο, αυτές οι εργασίες θεωρούσαν ένα μόνο UAV στα μοντέλα τους. Παράλληλα, έγινε έρευνα για τη βέλτιστη ανάπτυξη και κίνηση ενός μόνο UAV για την υποστήριξη ασύρματων επικοινωνιών κατερχόμενης ζεύξης, ενώ η εργασία στο [17] πρότεινε έναν αλγόριθμο χαμηλής πολυπλοκότητας για τη βέλτιστη ανάπτυξη πολλαπλών UAV που παρέχουν κάλυψη στους χρήστες εδάφους. Η εργασία στο [18] παρείχε μια ολοκληρωμένη ανάλυση κάλυψης κάτω ζεύξης για ένα δίκτυο στο οποίο ένας πεπερασμένος αριθμός UAV εξυπηρετεί τους χρήστες εδάφους. Στο [19], οι συγγραφείς χρησιμοποίησαν UAV για να συλλέξουν αποτελεσματικά δεδομένα και να επαναφορτίσουν την κεφαλή των συστάδων σε ένα ασύρματο δίκτυο αισθητήρων που χωρίζεται σε πολλαπλά συμπλέγματα. Ωστόσο, αυτή η εργασία περιορίζεται σε ένα στατικό δίκτυο αισθητήρων και δεν διερευνά τη βέλτιστη ανάπτυξη των UAV. Ενώ διερευνήθηκε η ενεργειακή απόδοση της μετάδοσης δεδομένων άνω ζεύξης σε δίκτυο επικοινωνίας μηχανής με μηχανή (M2M), δεν ελήφθη υπόψη η παρουσία UAV. Στην πραγματικότητα, καμία από τις προηγούμενες μελέτες δεν αντιμετώπισε το πρόβλημα της από κοινού βελτιστοποίησης της ανάπτυξης και της κινητικότητας των UAV, της συσχέτισης συσκευών και του ελέγχου ισχύος ανοδικής ζεύξης για την ενεργοποίηση αξιόπιστων και ενεργειακά αποδοτικών επικοινωνιών για συσκευές IoT. Από όσο είναι γνωστό, το [16] είναι μια από τις πρώτες ολοκληρωμένες μελέτες σχετικά με την κοινή βέλτιστη τρισδιάστατη ανάπτυξη εναέριων σταθμών βάσης, τη συσχέτιση συσκευών και τον έλεγχο ισχύος άνω ζεύξης σε ένα οικοσύστημα IoT.

Είναι εμφανές ότι στο [16] εισάγεται ένα νέο πλαίσιο για βελτιστοποιημένη ανάπτυξη και κινητικότητα πολλαπλών UAV με σκοπό την ενεργειακά αποδοτική συλλογή δεδομένων uplink από επίγειες συσκευές IoT. Ειδικότερα, μελετάται ένα δίκτυο IoT στο

οποίο οι συσκευές IoT μπορούν να είναι ενεργές σε διαφορετικές χρονικές στιγμές. Για να ελαχιστοποιηθεί η συνολική ισχύς μετάδοσης αυτών των συσκευών IoT, λαμβάνοντας υπόψη τους περιορισμούς SINR, προτείνεται μια αποτελεσματική προσέγγιση για την από κοινού και δυναμική εύρεση των θέσεων των UAV, τη συσχέτιση των συσκευών στα UAV και τη βέλτιστη ισχύ μετάδοσης ανοδικής ζεύξης. Το προτεινόμενο πλαίσιο αποτελείται από δύο βασικά βήματα. Αρχικά, δεδομένων των τοποθεσιών των συσκευών IoT, προτείνεται μια λύση για τη βελτιστοποίηση της ανάπτυξης και της σύνδεσης των UAV. Σε αυτή την περίπτωση, λύνεται το διατυπωμένο πρόβλημα αποσυνθέτοντας το σε δύο υποπροβλήματα τα οποία επιλύονται επαναληπτικά. Στο πρώτο υποπρόβλημα, δεδομένων των σταθερών θέσεων των UAV, βρίσκεται η βέλτιστη από κοινού συσχέτιση συσκευής-UAV και την ισχύ εκπομπής των συσκευών. Στο δεύτερο υποπρόβλημα, δεδομένης της συσχέτισης σταθερών συσκευών, προσδιορίζονται οι θέσεις των κοινών 3D UAV. Για αυτό το υποπρόβλημα, μετατρέπεται το μη κυρτό πρόβλημα βελτιστοποίησης συνεχούς τοποθεσίας σε κυρτή μορφή και παρέχονται λύσεις που μπορούν να λυθούν. Στη συνέχεια, ακολουθώντας τον προτεινόμενο αλγόριθμό μας, τα αποτελέσματα της επίλυσης του δεύτερου υποπροβλήματος χρησιμοποιούνται ως είσοδοι στο πρώτο υποπρόβλημα για την επόμενη επανάληψη. Εδώ, δείχνεται ότι η προτεινόμενη προσέγγισή οδηγεί σε μια αποτελεσματική λύση με λογική ακρίβεια σε σύγκριση με την παγκόσμια βέλτιστη λύση που απαιτεί σημαντικά γενικά έξοδα. Σαφώς, οι τοποθεσίες των UAV και η συσχέτιση συσκευών που λαμβάνονται σε αυτό το πρώτο βήμα θα εξαρτηθούν από τις τοποθεσίες των ενεργών συσκευών IoT.

Στο δεύτερο βήμα, αναλύεται το δίκτυο IoT σε μια χρονική περίοδο κατά την οποία αλλάζει το σύνολο των ενεργών συσκευών. Σε αυτήν την περίπτωση, παρουσιάζεται ένα πλαίσιο για τη βελτιστοποίηση της κινητικότητας των UAV επιτρέποντάς τους να ενημερώνουν δυναμικά τις τοποθεσίες τους ανάλογα με τη διαδικασία ενεργοποίησης των συσκευών που μεταβάλλονται χρονικά. Αρχικά, εξάγονται οι εκφράσεις κλειστής μορφής για τις χρονικές στιγμές (χρόνοι ενημέρωσης) στις οποίες πρέπει να κινούνται τα UAV σύμφωνα με τη διαδικασία ενεργοποίησης των συσκευών. Στη συνέχεια, χρησιμοποιώντας τα αποτελέσματα του χρόνου ενημέρωσης, εξάγεται η βέλτιστη τροχιά των 3D UAV έτσι ώστε η συνολική κίνηση των UAV κατά την ενημέρωση των τοποθεσιών τους να ελαχιστοποιείται. Τα αποτελέσματα της προσομοίωσής μας δείχνουν ότι, χρησιμοποιώντας την προτεινόμενη προσέγγιση, η συνολική ισχύς μετάδοσης των συσκευών IoT μπορεί να μειωθεί σημαντικά σε σύγκριση με μια περίπτωση στην οποία αναπτύσσονται σταθεροί εναέριοι σταθμοί βάσης. Τα αποτελέσματα επαληθεύουν επίσης τις αναλυτικές μας παραγώγους για τους χρόνους ενημέρωσης και αποκαλύπτουν μια εγγενή αντιστάθμιση μεταξύ του αριθμού των ενημερώσεων, της κινητικότητας των UAV και της ισχύος μετάδοσης των συσκευών IoT. Συγκεκριμένα, αποδεικνύεται ότι ο μεγαλύτερος αριθμός ενημερώσεων οδηγεί σε χαμηλότερες δυνάμεις μετάδοσης για τις συσκευές IoT σε βάρος της υψηλότερης κατανάλωσης ενέργειας των UAV.

3.2.2 Ανάλυση της downlink κάλυψης για ένα πεπερασμένο 3D ασύρματο δίκτυο UAV οχημάτων

Με σημαντικές προόδους στην τεχνολογία των drone, όπως η αυξημένη χωρητικότητα ωφέλιμου φορτίου, ο μεγαλύτερος μέσος χρόνος πτήσης, οι καλύτερες τεχνικές διαχείρισης ενέργειας και η δυνατότητα συλλογής ηλιακής ενέργειας, τα μη

επανδρωμένα εναέρια οχήματα (UAV) μπορούν να εξυπηρετήσουν πολλούς σκοπούς, όπως επιτήρηση, εντοπισμός και επικοινωνία, καθιστώντας τα μια ευέλικτη λύση για την αύξηση και την ενίσχυση των δυνατοτήτων των σημερινών κυψελωτών συστημάτων, [18]. Παρέχουν μια ιδιαίτερα ελκυστική λύση για την παροχή συνδεσιμότητας μετά από καταστροφές και ατυχήματα, τα οποία μπορεί να ακρωτηριάσουν εντελώς τα επίγεια δίκτυα λόγω κατεστραμμένου εξοπλισμού ή/και απώλειας ισχύος. Γενικά, τα UAV παρέχουν μια ρεαλιστική λύση σε σενάρια όπου υπάρχει προσωρινή ανάγκη για πόρους δικτύου. Αυτές θα μπορούσαν να περιλαμβάνουν καταστάσεις πρώτης ανταπόκρισης, όπως αυτή που προαναφέρθηκε, ή ακόμα και συνηθισμένα σενάρια για μη στρατιωτικούς, όπως ποδοσφαιρικούς αγώνες ή συναυλίες. Προκειμένου να παρέχεται βραχυπρόθεσμη συνδεσιμότητα σε τέτοια σενάρια, η προσωρινή ανάπτυξη UAV μπορεί να είναι ταχύτερη και πιο οικονομική σε σύγκριση με την προσωρινή εγκατάσταση συμβατικών σταθμών βάσης. Επί του παρόντος, διερευνώνται επίσης ως πιθανοί υποψήφιοι για την παροχή πανταχού παρούσας συνδεσιμότητας σε απομακρυσμένες περιοχές που στερούνται παραδοσιακής κυψελοειδούς υποδομής. Αν και δεν υπάρχει αμφιβολία για την ευελιξία ανάπτυξης και τα γενικά οφέλη των UAV, η απόδοσή τους όσον αφορά την κάλυψη και την χωρητικότητα που παρέχεται στους επίγειους χρήστες δεν είναι αρκετά κατανοητή. Αυτό ισχύει ιδιαίτερα για μια ρεαλιστική περίπτωση χρήσης πεπερασμένων δικτύων UAV, όπου έχουμε έναν δεδομένο αριθμό UAV που εξυπηρετούν χρήστες σε μια δεδομένη περιοχή (όπως μια πόλη). Σε αυτή την ενότητα, χρησιμοποιούνται εργαλεία από τη στοχαστική γεωμετρία για να γίνει εξαγωγή της κατανομής του λόγου σήματος προς παρεμβολή κατερχόμενης ζεύξης για αυτήν τη ρύθμιση, η οποία παρέχει αμέσως χρήσιμες πληροφορίες για την απόδοση κάλυψης του προκύπτοντος τρισδιάστατου δικτύου.

Οι βελτιώσεις στη χωρητικότητα ωφέλιμου φορτίου και οι παρατεταμένοι χρόνοι πτήσης επέτρεψαν την εμπορική χρήση UAV, ειδικά για τηλεπικοινωνιακούς σκοπούς. Τα δίκτυα UAV διαφέρουν σημαντικά από τα συμβατικά ασύρματα δίκτυα όσον αφορά την κινητικότητα, τους ενεργειακούς περιορισμούς, καθώς και τις συνθήκες διάδοσης. Αυτό έχει κεντρίσει το ενδιαφέρον για το σχεδιασμό πρωτοκόλλων προσανατολισμένων στην εφαρμογή για την αποτελεσματική χρήση των εναέριων δικτύων. Για παράδειγμα, ένα πρωτόκολλο που βασίζεται σε συμπλέγματα, το οποίο βελτιώνει την ανθεκτικότητα σε συχνές αστοχίες ζεύξης που προκύπτουν από την κίνηση των UAV, έχει προταθεί στο [20]. Η ευελιξία που προσφέρει η κινητικότητα των UAV έχει παρακινήσει πολλές αλγοριθμικές ερευνητικές προσπάθειες για την εύρεση αποτελεσματικών τροχιών και στρατηγικών ανάπτυξης με στόχο τη βελτιστοποίηση διαφορετικών πόρων δικτύου. Για παράδειγμα, ένας αλγόριθμος για τη βελτιστοποίηση του φάσματος ισχύος μετάδοσης και συχνότητας για αυτόνομη αυτοανάπτυξη προτάθηκε στο [21]. Ένας προσαρμοστικός αλγόριθμος για την προσαρμογή της επικεφαλίδας των UAV προτάθηκε στο [22] για τη βελτίωση της απόδοσης ανοδικής ζεύξης και την ελαχιστοποίηση της αμοιβαίας παρεμβολής. Μια προσέγγιση για τη βελτιστοποίηση του υψομέτρου των UAV για τη μεγιστοποίηση της κάλυψης στο έδαφος προτάθηκε στο [23]. Η απόδοση των UAV που λειτουργούν ως αναμεταδότες μεταξύ των επίγειων χρηστών και των σταθμών βάσης διερευνήθηκε στο [24], ενώ το πρόβλημα της αποτελεσματικής τοποθέτησης UAV με ελαφρώς διαφορετικούς στόχους μελετήθηκε στο [25], [26].

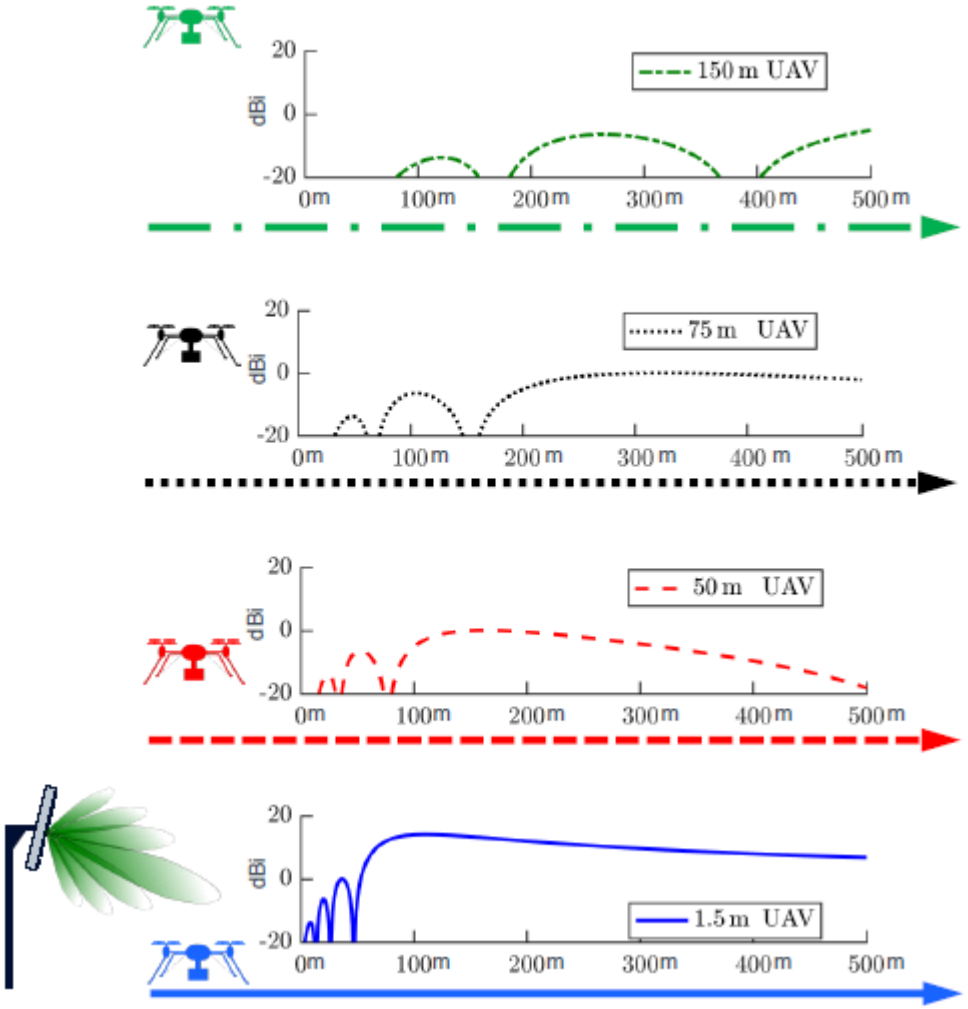
Μια άλλη κατεύθυνση έρευνας, η οποία είναι κάπως συμπληρωματική με αυτήν που συζητήθηκε παραπάνω, είναι η ανάπτυξη τεχνικών για τη ρεαλιστική ανάλυση σε επίπεδο συστήματος των δικτύων UAV. Όπως συμβαίνει στα επίγεια δίκτυα, έτσι και στα κυψελωτά δίκτυα, αυτές οι τεχνικές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη σύγκριση της απόδοσης διαφορετικών στρατηγικών ανάπτυξης και για τη συγκριτική αξιολόγηση της απόδοσής τους έναντι των τυπικών βασικών γραμμών. Στην περίπτωση των δικτύων UAV, η απόδοση σε επίπεδο συστήματος έχει κυρίως μελετηθεί μέσω επιτόπιων δοκιμών και προσομοιώσεων. Για παράδειγμα, όπως περιγράφεται και στο [18], ο χρόνος διακοπής λειτουργίας και η μέση απόδοση συγκρίθηκαν για διαφορετικούς αλγόριθμους δρομολόγησης χρησιμοποιώντας πειράματα πραγματικού κόσμου. Ενώ οι δοκιμές πεδίου ή οι προσομοιώσεις μπορούν να παρέχουν αρχικές πληροφορίες για τη συμπεριφορά του δικτύου, αυτές οι μέθοδοι συνήθως δεν είναι επεκτάσιμες όταν ο αριθμός των παραμέτρων προσομοίωσης είναι μεγάλος. Ένας τρόπος για να μειωθεί η διάσταση τέτοιων προβλημάτων είναι να εφοδιαστούν οι θέσεις των κόμβων με μια κατανομή, η οποία επιπλέον επιτρέπει τη χρήση ισχυρών εργαλείων από τη στοχαστική γεωμετρία για την εξαγωγή εύχρηστων εκφράσεων για βασικές μετρήσεις απόδοσης. Ενώ η στοχαστική γεωμετρία έχει ήδη αναδειχθεί ως προτιμώμενο εργαλείο για την ανάλυση ad hoc και κυψελοειδών δικτύων, οι δυνατότητές της δεν έχουν ακόμη αξιοποιηθεί για την ανάλυση δικτύων UAV. Μια σχετική προηγούμενη τεχνική μελετά τη συνύπαρξη ενός δικτύου επικοινωνίας συσκευής με συσκευή (D2D) με ένα μόνο UAV. Στην έρευνα που έγινε στο [18], αναπτύχθηκε το πρώτο ολοκληρωμένο μοντέλο που στοχεύει στην ανάλυση κατερχόμενης ζεύξης ενός πεπερασμένου δικτύου πολλαπλών UAV χρησιμοποιώντας εργαλεία από τη στοχαστική γεωμετρία.

Παρόλο που η ομοιογενής διαδικασία Poisson Point Process (PPP) έχει γίνει ένα κανονικό μοντέλο για τις χωρικές τοποθεσίες επίγειων σταθμών βάσης, δεν είναι αρκετά κατάλληλο για δίκτυα UAV, ειδικά όταν ένας δεδομένος αριθμός (πιθανώς μικρός) UAV αναπτύσσεται για να καλύψει μια δεδομένη πεπερασμένη περιοχή. Για τέτοια σενάρια, ένα απλό αλλά λογικό μοντέλο για τη χωρική κατανομή των UAV είναι η διαδικασία ομογενούς διωνυμικού σημείου (Binomial Point Process - BPP) [19], [18]. Τα πιο εξελιγμένα μοντέλα που ενσωματώνουν την αλληλεπίδραση μεταξύ των σημείων είναι συνήθως πολύ λιγότερο ελκυστικά. Ενώ το BPP δεν έχει ακόμη χρησιμοποιηθεί για την ανάλυση UAV, έχει λάβει σημαντική προσοχή για την ανάλυση επίγειων δικτύων με δεδομένο αριθμό κόμβων. Ωστόσο, μέχρι πρόσφατα, η ανάλυση επικεντρωνόταν σε ad hoc δίκτυα, στα οποία ένας δεδομένος αριθμός κόμβων υποτίθεται ότι κατανέμεται ομοιόμορφα τυχαία σε μια κυκλική περιοχή με τον δέκτη αναφοράς να βρίσκεται στο κέντρο του κύκλου. Η πιθανότητα διακοπής λειτουργίας αυτού του δέκτη αναφοράς προκύπτει στη συνέχεια, υποθέτοντας ότι εξυπηρετείται από έναν πομπό αναφοράς που βρίσκεται σε σταθερή απόσταση (όχι μέρος του BPP). Προκειμένου να μοντελοποιηθούν ουσιαστικά τα κυψελωτά συστήματα με αυτήν τη ρύθμιση, απαιτούνται δύο βασικές γενικεύσεις: (i) ο δέκτης αναφοράς μπορεί να βρίσκεται οπουδήποτε στην περιοχή και (ii) ο σταθμός βάσης εξυπηρέτησης για τον δέκτη αναφοράς θα επιλεγεί από το ίδιο το BPP. Για το τελευταίο, είναι λογικό να γίνει η υπόθεση ότι ο δέκτης αναφοράς εξυπηρετείται από τον πλησιέστερο σταθμό βάσης από το BPP. Η ακριβής ανάλυση αυτής της εγκατάστασης πεπερασμένου κυψελοειδούς δικτύου έγινε πολύ πρόσφατα όπως επίσης και μια κατά προσέγγιση ανάλυση μιας σχετικής εγκατάστασης. Βασιζόμενοι στις κατανομές απόστασης, στο

[18] εκτελείται ανάλυση κατερχόμενης ζεύξης για έναν χρήστη που βρίσκεται αυθαίρετα στο έδαφος που εξυπηρετείται από ένα πεπερασμένο δίκτυο UAV. Εκτός από την παροχή της πρώτης τέτοιας ανάλυσης σε επίπεδο συστήματος ενός πεπερασμένου δικτύου UAV, αρκετά ενδιάμεσα αποτελέσματα παρέχουν κατασκευές που είναι γενικότερα εφαρμόσιμες στην ανάλυση πεπερασμένων ασύρματων δικτύων.

3.2.3 5G NR Massive MIMO για αρχική πρόσβαση και επιλογή κυψέλης

Πριν από τη λήψη και τη μετάδοση δεδομένων ωφέλιμου φορτίου, τα UAV που είναι συνδεδεμένα με κυψέλη πρέπει να έχουν πρόσβαση στο δίκτυο, [27]. Για το σκοπό αυτό, τα κυψελωτά BS μεταδίδουν τακτικά μπλοκ σήματος συγχρονισμού (SSB) που διευκολύνουν την ανακάλυψή τους. Μέχρι τα δίκτυα LTE Advanced Pro, η ακτινοβολία των σημάτων SSB προσδιοριζόταν αποκλειστικά από το μοτίβο της κεραίας BS. Για το τυπικό κυψελοειδές δίκτυο με κατωφέρεια BS, αυτό σημαίνει ότι οι συσκευές που πετούν ψηλότερα από τα BS μπορούσαν να αντιληφθούν σήματα SSB μόνο μέσω των πλευρικών λοβών της κεραίας.



Εικόνα 5: Κέρδος κεραίας [dBi] έναντι απόστασης 2D BS-UAV [m] για BS που αναπτύσσεται στα 25 μέτρα και UAV διαφόρων υψών, ευθυγραμμισμένα με το οριζόντιο επίπεδο του BS.

Αυτό το φαινόμενο απεικονίζεται στην Εικόνα 5, η οποία αντιπροσωπεύει το κέρδος της κεραίας που γίνεται αντιληπτό από τα UAV που πετούν στα (από πάνω προς τα κάτω) {150, 75, 50, 1,5} m και απομακρύνονται από ένα BS με ύψος 25 m. Το BS είναι εξοπλισμένο με μια κατακόρυφη διάταξη 8 X 1 με κλίση κατά 12 μοίρες. Η εικόνα αυτή δείχνει πώς, ενώ οι κυψελωτές συσκευές κοντά στο έδαφος παρουσιάζουν γενικά λογικές και ομαλές διακυμάνσεις κέρδους κεραίας καθώς απομακρύνονται από το BS, αυτό δεν ισχύει για τα UAV, των οποίων τα κέρδη κεραίας μειώνονται δραματικά και γίνονται πολύ ακανόνιστα καθώς αυξάνεται το υψόμετρο και μόνο οι πλευρικοί λοβοί της κεραίας BS μπορούν να γίνουν αντιληπτοί.

4 Σύγκριση και περιγραφή αλγορίθμων και αποτελεσμάτων

4.1 Περιγραφή της ενισχυτικής μάθησης

4.1.1 Ορισμός

Για την καλύτερη κατανόηση της ενισχυτικής μάθησης, θα χρησιμοποιηθεί σαν παράδειγμα το γεγονός ότι ο υπολογιστής σας θέλει να παίξει σκάκι με έναν άνθρωπο, [13]. Η πρώτη ερώτηση που πρέπει να γίνει είναι η εξής:

Θα ήταν δυνατό εάν το μηχάνημα είχε εκπαιδευτεί με εποπτεία;

Θεωρητικά, ναι. Αλλά-

Υπάρχουν δύο μειονεκτήματα που πρέπει να ληφθούν υπόψη.

Αρχικά, για να προχωρήσει το παράδειγμα αυτό με την εποπτευόμενη μάθηση χρειάζεται ένα σχετικό σύνολο δεδομένων. Επιπλέον, εάν εκπαιδευτεί η μηχανή να αναπαράγει την ανθρώπινη συμπεριφορά στο παιχνίδι του σκακιού, η μηχανή δεν θα ήταν ποτέ καλύτερη από την ανθρώπινη, γιατί απλώς αναπαράγει την ίδια συμπεριφορά. Επομένως, εξ ορισμού, δεν είναι εφικτό να χρησιμοποιηθεί εποπτευόμενη μάθηση για την εκπαίδευση της μηχανής. Υπάρχει όμως τρόπος να παίξει ένας πράκτορας ένα παιχνίδι εντελώς μόνος του;

Ναι, εκεί μπαίνει στο παιχνίδι η Ενισχυτική Μάθηση. Το Reinforcement Learning είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που μαθαίνει να λύνει ένα πρόβλημα πολλαπλών επιπέδων με δοκιμή και σφάλμα. Το μηχάνημα εκπαιδεύεται σε σενάρια πραγματικής ζωής για να λαμβάνει μια σειρά αποφάσεων. Λαμβάνει είτε ανταμοιβές είτε ποινές για τις ενέργειες που εκτελεί. Στόχος του είναι να μεγιστοποιήσει τη συνολική ανταμοιβή. Με τον όρο Deep Reinforcement Learning εννοούνται πολλαπλά στρώματα τεχνητών νευρωνικών δικτύων που υπάρχουν στην αρχιτεκτονική για να αναπαράγουν τη λειτουργία ενός ανθρώπινου εγκεφάλου.

4.1.2 Σημαντικές έννοιες

Στην ενότητα αυτή, κρίνεται σκόπιμο να αναφερθούν μερικοί από τους ορισμούς που θα συναντηθούν στην Ενισχυτική Μάθηση.

Πράκτορας - Ο Πράκτορας (Agent - A) αναλαμβάνει ενέργειες που επηρεάζουν το περιβάλλον. Παραθέτοντας ένα παράδειγμα, η μηχανική εκμάθηση να παίζει σκάκι είναι ο πράκτορας.

Ενέργεια - Είναι το σύνολο όλων των πιθανών λειτουργιών/κινήσεων που μπορεί να κάνει ο πράκτορας. Ο πράκτορας αποφασίζει ποια ενέργεια θα λάβει από ένα σύνολο διακριτών ενεργειών (actions - a).

Περιβάλλον - Όλες οι ενέργειες που κάνει ο πράκτορας ενίσχυσης μάθησης επηρεάζουν άμεσα το περιβάλλον. Εδώ, η σκακιέρα είναι το περιβάλλον. Το περιβάλλον παίρνει την παρούσα κατάσταση και δράση του πράκτορα ως πληροφορία και επιστρέφει την ανταμοιβή στον πράκτορα με μια νέα κατάσταση. Για παράδειγμα, η κίνηση του bot θα έχει είτε αρνητική/θετική επίδραση σε όλο το παιχνίδι και στη διάταξη του ταμπλό. Αυτό θα αποφασίσει την επόμενη δράση και την κατάσταση του διοικητικού συμβουλίου.

Κατάσταση - Μια κατάσταση (State - S) είναι μια συγκεκριμένη κατάσταση στην οποία βρίσκεται ο πράκτορας.

Εικόνα 6: Σκακιέρα, αυτή μπορεί να είναι η κατάσταση του πράκτορα σε οποιαδήποτε ενδιάμεση στιγμή (t).

Ανταμοιβή (Reward - R) - Το περιβάλλον παρέχει ανατροφοδότηση με την οποία προσδιορίζουμε την εγκυρότητα των ενεργειών του πράκτορα σε κάθε κατάσταση. Είναι ζωτικής σημασίας στο σενάριο της Ενισχυτικής Μάθησης όπου χρειάζεται η μηχανή να μαθαίνει μόνη της και ο μόνος επικριτής που θα τη βοηθούσε στη μάθηση είναι η ανατροφοδότηση/ανταμοιβή που λαμβάνει. Για παράδειγμα, σε ένα σενάριο παιχνιδιού σκακιού συμβαίνει όταν το bot παίρνει τη θέση του κομματιού ενός αντιπάλου και αργότερα το πιάνει.

Συντελεστής έκπτωσης - Με την πάροδο του χρόνου, ο συντελεστής έκπτωσης τροποποιεί τη σημασία των κινήτρων. Δεδομένης της αβεβαιότητας του μέλλοντος, είναι καλύτερο να προσθέσουμε διακύμανση στις εκτιμήσεις της αξίας. Ο παράγοντας έκπτωσης βοηθά στη μείωση του βαθμού στον οποίο οι μελλοντικές ανταμοιβές επηρεάζουν τις εκτιμήσεις της συνάρτησης αξίας.

Πολιτική (π) - Αποφασίζει ποια ενέργεια πρέπει να κάνετε σε μια συγκεκριμένη κατάσταση για να μεγιστοποιήσετε την ανταμοιβή.

Αξία (Value - V) - Μετρά τη βελτιστοποίηση μιας συγκεκριμένης κατάστασης. Είναι οι αναμενόμενες ανταμοιβές με έκπτωση που συλλέγει ο πράκτορας ακολουθώντας τη συγκεκριμένη πολιτική.

Q-value ή **action-value** - Το Q Value είναι ένα μέτρο της συνολικής αναμενόμενης ανταμοιβής εάν ο πράκτορας (A) βρίσκεται σε κατάσταση (ες) και κάνει την ενέργεια (a) και στη συνέχεια παίζει μέχρι το τέλος του επεισοδίου σύμφωνα με κάποια πολιτική (π).

4.2 Βασική διάκριση αλγορίθμων ενισχυτικής μάθησης

Υπάρχουν δύο κύριοι τύποι αλγορίθμων Ενισχυτικής Μάθησης, οι αλγόριθμοι βασισμένοι σε μοντέλα και οι αλγόριθμοι χωρίς μοντέλα. Μια σύντομη περιγραφή βρίσκεται σε αυτή την ενότητα, ενώ στη συνέχεια ακολουθούν εκτενέστερες πληροφορίες για τις υποκατηγορίες τους.

4.2.1 Αλγόριθμοι βασισμένοι σε μοντέλα

Ο αλγόριθμος που βασίζεται σε μοντέλα χρησιμοποιεί τη συνάρτηση μετάβασης και ανταμοιβής για να εκτιμήσει τη βέλτιστη πολιτική. Χρησιμοποιούνται σε σενάρια όπου

έχουμε πλήρη γνώση του περιβάλλοντος και του πώς αντιδρά σε διαφορετικές ενέργειες.

Στην Ενισχυτική Εκμάθηση βάσει Μοντέλων, ο πράκτορας έχει πρόσβαση στο μοντέλο του περιβάλλοντος, δηλαδή, η ενέργεια που απαιτείται να εκτελεστεί για να μεταβεί από τη μια κατάσταση στην άλλη, τις πιθανότητες που συνδέονται και τις αντίστοιχες ανταμοιβές.

Επιτρέπουν στον πράκτορα ενίσχυσης μάθησης να προγραμματίσει το μέλλον με το να σκέφτεται μπροστά.

Για στατικά/σταθερά περιβάλλοντα, η Ενισχυτική Εκμάθηση βάσει Μοντέλων είναι πιο κατάλληλη.

4.2.2 Αλγόριθμοι χωρίς μοντέλα

Οι αλγόριθμοι χωρίς μοντέλα βρίσκουν τη βέλτιστη πολιτική με πολύ περιορισμένη γνώση της δυναμικής του περιβάλλοντος. Δεν κάνουν καμία λειτουργία μετάβασης/ανταμοιβής για να κρίνουν την καλύτερη πολιτική.

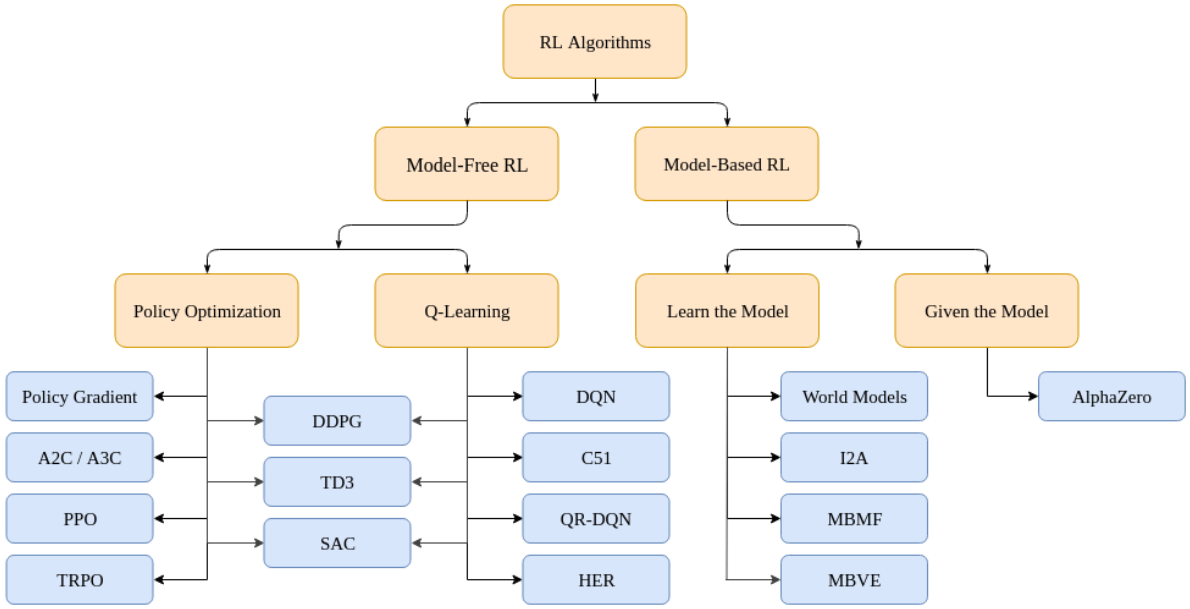
Εκτιμούν τη βέλτιστη πολιτική απευθείας από την εμπειρία, δηλαδή την αλληλεπίδραση μεταξύ πράκτορα και περιβάλλοντος χωρίς να έχουν καμία υπόδειξη της συνάρτησης ανταμοιβής.

Η Ενισχυτική Μάθηση χωρίς μοντέλα θα πρέπει να εφαρμόζεται σε σενάρια που περιλαμβάνουν ελλιπείς πληροφορίες για το περιβάλλον.

Στον πραγματικό κόσμο, δεν υπάρχει σταθερό περιβάλλον. Τα αυτοοδηγούμενα αυτοκίνητα έχουν ένα δυναμικό περιβάλλον με μεταβαλλόμενες συνθήκες κυκλοφορίας, εκτροπές διαδρομών κ.λπ. Σε τέτοια σενάρια, οι αλγόριθμοι χωρίς μοντέλα υπερτερούν των άλλων τεχνικών.

4.3 Περιγραφή αλγορίθμων ενισχυτικής μάθησης

Η ενισχυτική μάθηση (Reinforcement Learning - RL) με χρήση νευρωνικών δικτύων μπορεί να χρησιμοποιήσει διάφορους αλγόριθμους, όπως φαίνεται και στην Εικόνα 6, [11].



Εικόνα 7: Αλγόριθμοι ενισχυτικής μάθησης (Reinforcement Learning - RL)

Στην παρούσα εργασία θα εξετάσουμε ορισμένους από τους παραπάνω αλγορίθμους και πιο συγκεκριμένα τους DQN, A2C και PPO, τους οποίους και θα συγκρίνουμε μεταξύ τους.

Μια πολιτική, στην ορολογία της επαναληπτικής μάθησης, είναι μια αντιστοίχιση από χώρο δράσης σε χώρο κατάστασης. Μπορεί να φανταστεί κανείς ότι είναι οδηγίες για τον πράκτορα RL, όσον αφορά τις ενέργειες που πρέπει να κάνει με βάση την κατάσταση του περιβάλλοντος στην οποία βρίσκεται αυτήν τη στιγμή.

Το DQN, ή το δίκτυο deep-Q, είναι ένας αλγόριθμος RL χωρίς μοντέλα που μαθαίνει μέσω Q-learning και χωρίς πολιτική (off-policy). Το DQN και το Q-learning γενικά εκτελούνται χωρίς «πολιτική» ή αλλιώς off-policy (πολιτική είναι κάτι σαν τον εγκέφαλο της ενισχυτικής μάθησης), πράγμα που σημαίνει ότι κατά την εκμάθηση, κάθε ενημέρωση της πολιτικής μπορεί να χρησιμοποιεί δεδομένα που συλλέγονται σε οποιοδήποτε σημείο του χρόνου εκπαίδευσης. Αντίθετα, αλγόριθμοι όπως το PPO κάνουν χρήση «πολιτικής» ή αλλιώς on-policy, χρησιμοποιώντας μόνο δεδομένα που συλλέγονται από την πιο ενημερωμένη πολιτική και λήψη αποφάσεων με βάση αυτήν. Και οι δύο αλγόριθμοι είναι χωρίς μοντέλα, πράγμα που σημαίνει ότι ενώ εκπαιδεύονται, δεν έχουν πρόσβαση στο μοντέλο του περιβάλλοντος. Το μοντέλο του περιβάλλοντος είναι μια συνάρτηση που προβλέπει μεταβάσεις κατάστασης και ανταποκρίσεις από τις μεταβάσεις αυτές. Η Εικόνα 6 αποτελεί δίνει μια πρόχειρη ταξινόμηση των αλγορίθμων RL. Όπως είναι εμφανές, τόσο το DQN όσο και το PPO εμπίπτουν στον κλάδο της κατηγορίας της ενισχυτικής μάθησης χωρίς μοντέλα, αλλά εκεί που διαφέρουν το DQN και το PPO είναι ο τρόπος μεγιστοποίησης της απόδοσης. Το DQN χρησιμοποιεί Q-learning, ενώ το PPO υφίσταται άμεση βελτιστοποίηση πολιτικής.

4.3.1 Αλγόριθμος Deep Q-Network (DQN)

Ο Deep Q-Network (DQN) είναι ένας ισχυρός αλγόριθμος στον τομέα της ενισχυτικής μάθησης. Συνδυάζει τις αρχές των βαθιών νευρωνικών δικτύων με την Q-learning,

επιτρέποντας στους πράκτορες να μαθαίνουν βέλτιστες πολιτικές σε πολύπλοκα περιβάλλοντα. Σε αυτή της ενότητα, θα εξερευνήσουμε τις αρχές λειτουργίας του DQN, θα συζητήσουμε τις βασικές του έννοιες, θα παρέχουμε ένα παράδειγμα εφαρμογής κώδικα στην Python και θα εξετάσουμε τα πλεονεκτήματα και τους περιορισμούς του.

Ο αλγόριθμος DQN ακολουθεί μια προσέγγιση βασισμένη σε βαθύ νευρωνικό δίκτυο για την εκμάθηση και τη βελτιστοποίηση των συναρτήσεων απόφασης. Η διαδικασία λειτουργίας του μπορεί να συνοψιστεί ως εξής:

1. Αναπαράσταση κατάστασης: Μετατροπή της τρέχουσας κατάστασης του περιβάλλοντος σε μια κατάλληλη αριθμητική αναπαράσταση, όπως ακατέργαστες τιμές pixel ή προεπεξεργασμένα χαρακτηριστικά.
2. Αρχιτεκτονική νευρωνικών δικτύων: Σχεδιασμός ενός βαθιού νευρωνικού δικτύου, συνήθως ένα συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network - CNN), που λαμβάνει την κατάσταση ως είσοδο και εξάγει τιμές δράσης/απόφασης για κάθε πιθανή ενέργεια.
3. Επανάληψη εμπειρίας: Αποθήκευση των εμπειριών του πράκτορα που αποτελούνται από πλειάδες κατάστασης, δράσης, ανταμοιβής και επόμενης κατάστασης σε ένα buffer μνήμης επανάληψης.
4. Ενημέρωση Q-Learning: Δείγμα μικρών παρτίδων εμπειριών από τη μνήμη επανάληψης για να γίνει ενημέρωση των βαρών του νευρωνικού δικτύου. Η ενημέρωση πραγματοποιείται χρησιμοποιώντας τη συνάρτηση απώλειας που προέρχεται από την εξίσωση Bellman, η οποία ελαχιστοποιεί την απόκλιση μεταξύ της προβλεπόμενης και της στοχευόμενης τιμής ενέργειας.
5. Εξερεύνηση και εκμετάλλευση: Εξισορρόπηση την εξερεύνησης και την εκμετάλλευσης επιλέγοντας ενέργειες είτε άπληστα με βάση την τρέχουσα πολιτική είτε στοχαστικά για να γίνει ενθάρρυνση της εξερεύνησης.
6. Δίκτυο στόχος: Χρήση ενός ξεχωριστού δικτύου στόχου ίδιας αρχιτεκτονικής με το κύριο δίκτυο για τη σταθεροποίηση της διαδικασίας εκμάθησης. Περιοδική ενημέρωση του δικτύου προορισμού αντιγράφοντας τα βάρη από το κύριο δίκτυο.
7. Επανάληψη των βημάτων 1 έως 6: Αλληλεπίδραση με το περιβάλλον, συγκέντρωση εμπειριών, ενημέρωση του δικτύου και βελτίωση της πολιτικής επαναληπτικά μέχρι τη σύγκλιση.

Βασικές έννοιες:

1. Q-Learning: Το DQN αξιοποιεί τον αλγόριθμο Q-learning, ο οποίος στοχεύει στην εκτίμηση της βέλτιστης συνάρτησης τιμής δράσης (συνάρτηση Q) που αντιστοιχίζει τις καταστάσεις με τις αναμενόμενες μελλοντικές ανταμοιβές.
2. Εμπειρία επανάληψης: Η επανάληψη εμπειρίας βοηθά στον αποσυσχετισμό των διαδοχικών εμπειριών αποθηκεύοντάς τες σε μια προσωρινή μνήμη επανάληψης. Αυτό το buffer μνήμης λαμβάνεται τυχαία κατά τη διάρκεια της ενημέρωσης δικτύου για να διακοπεί οι χρονικές εξαρτήσεις και να σταθεροποιηθεί η εκμάθηση.

Σε αυτό το απόσπασμα κώδικα που βρίσκεται στο Παράρτημα 1, εισάγονται πρώτα οι απαιτούμενες βιβλιοθήκες, συμπεριλαμβανομένων των OpenAI Gym για το περιβάλλον, TensorFlow και Keras για το βαθύ νευρωνικό δίκτυο και τη δομή δεδομένων deque για επανάληψη εμπειρίας.

Στη συνέχεια, ορίζεται η κλάση DQNAgent, η οποία ενσωματώνει τη λειτουργικότητα του πράκτορα DQN. Αποτελείται από μεθόδους για τη δημιουργία του νευρωνικού δικτύου, την ανάμνηση εμπειριών, την επιλογή ενεργειών και την εκτέλεση ενημερώσεων δικτύου χρησιμοποιώντας την επανάληψη της εμπειρίας. Στη συνέχεια, δημιουργείται το περιβάλλον Gym και αρχικοποιείται ο παράγοντας DQN με τα κατάλληλα μεγέθη κατάστασης και ενεργειών.

Ακολούθως, ο κώδικας μπαίνει στο βρόχο εκπαίδευσης, όπου αλληλεπιδρά με το περιβάλλον, συλλέγει εμπειρίες, τις «θυμάται» και ενημερώνει περιοδικά το νευρωνικό δίκτυο του πράκτορα.

Πλεονεκτήματα:

1. Εκμάθηση βαθιάς αναπαράστασης: Το DQN αξιοποιεί βαθιά νευρωνικά δίκτυα για να μάθει αφηρημένες και υψηλών διαστάσεων αναπαραστάσεις καταστάσεων, επιτρέποντας την αποτελεσματική μάθηση σε πολύπλοκα περιβάλλοντα.
2. Αποδοτικότητα δείγματος: Εμπειρία επανάληψης και δικτύων στόχων βελτιώνουν την αποτελεσματικότητα του δείγματος επαναχρησιμοποιώντας και διασυσχετίζοντας τις εμπειρίες.
3. Γενίκευση: Το DQN μπορεί να γενικεύσει τις μαθημένες πολιτικές σε μη ορατές καταστάσεις, επιτρέποντας καλύτερη προσαρμογή και λήψη αποφάσεων σε νέα σενάρια.

Περιορισμοί:

1. Ευαισθησία υπερπαραμέτρων: Η απόδοση του DQN είναι ευαίσθητη σε ρυθμίσεις υπερπαραμέτρων, όπως ο ρυθμός εκμάθησης, ο ρυθμός εξερεύνησης και η αρχιτεκτονική δικτύου, που απαιτούν προσεκτικό συντονισμό.
2. Έλλειψη συνεχούς μάθησης: Το DQN έχει σχεδιαστεί κυρίως για μαζική μάθηση εκτός σύνδεσης και δεν χειρίζεται φυσικά σενάρια συνεχούς μάθησης στο διαδίκτυο.
3. Υπερεκτίμηση των τιμών δράσης: Η ενημέρωση Q-learning που χρησιμοποιείται στο DQN μπορεί να οδηγήσει σε υπερεκτίμηση των τιμών δράσης, επηρεάζοντας την ακρίβεια της εκμάθησης πολιτικής.

Συμπέρασμα:

Το Deep Q-Network (DQN) είναι ένας πρωτοποριακός αλγόριθμος που συνδυάζει βαθιά νευρωνικά δίκτυα με Q-learning για ενισχυτικές εργασίες μάθησης. Η ικανότητά του να μαθαίνει βέλτιστες πολιτικές σε πολύπλοκα περιβάλλοντα τον έχει καταστήσει έναν ευρέως χρησιμοποιούμενο αλγόριθμο στο πεδίο. Με τη μόχλευση του DQN, οι ερευνητές και οι επαγγελματίες μπορούν να εκπαιδεύσουν πράκτορες που μαθαίνουν από τις ακατέργαστες αισθητηριακές εισροές και λαμβάνουν αποφάσεις βασισμένες σε αναπαραστάσεις κατάστασης υψηλών διαστάσεων. Παρά τους περιορισμούς του, η εκμάθηση βαθιάς αναπαράστασης, η αποτελεσματικότητα του δείγματος και οι δυνατότητες γενίκευσης του DQN το καθιστούν πολύτιμο εργαλείο για την επίλυση ενός ευρέος φάσματος προβλημάτων ενισχυτικής μάθησης.

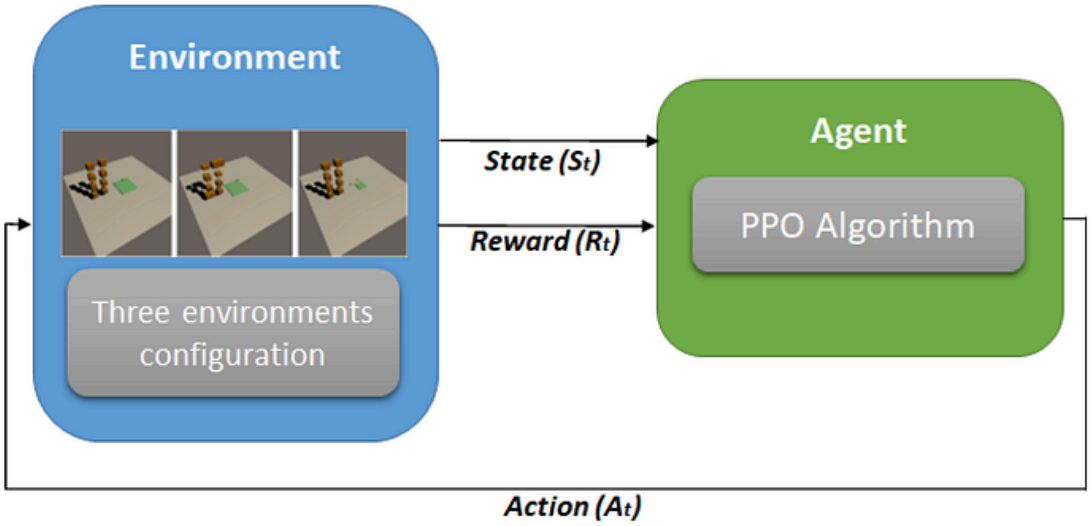
4.3.2 Βελτιστοποίησης εγγύς πολιτικής (Proximal Policy Optimization – PPO)

Ο αλγόριθμος βελτιστοποίησης εγγύς πολιτικής (Proximal Policy Optimization – PPO) είναι ένας αλγόριθμος στον τομέα της ενισχυτικής μάθησης που εκπαιδεύει τη λειτουργία λήψης αποφάσεων ενός πράκτορα για την ολοκλήρωση δύσκολων εργασιών. Το PPO αναπτύχθηκε από τον John Schulman το 2017, είχε γίνει ο προεπιλεγμένος αλγόριθμος μάθησης ενίσχυσης στην αμερικανική εταιρεία τεχνητής νοημοσύνης OpenAI, [12].

Πολλοί ειδικοί αποκάλεσαν το PPO την τελευταία λέξη της τεχνολογίας επειδή φαίνεται να επιτυγχάνει μια ισορροπία μεταξύ απόδοσης και κατανόησης. Σε σύγκριση με άλλους αλγόριθμους, τα τρία κύρια πλεονεκτήματα του PPO είναι η απλότητα, η σταθερότητα και η αποτελεσματικότητα του δείγματος. Το PPO ταξινομείται ως μέθοδος διαβάθμισης πολιτικής για την εκπαίδευση του δικτύου πολιτικής ενός πράκτορα. Το δίκτυο πολιτικών είναι η λειτουργία που χρησιμοποιεί ο πράκτορας για τη λήψη αποφάσεων. Ουσιαστικά, για να εκπαιδεύσει το σωστό δίκτυο πολιτικής, το PPO λαμβάνει μια μικρή ενημέρωση πολιτικής (μέγεθος βήματος), ώστε ο πράκτορας να μπορεί να φτάσει αξιόπιστα στη βέλτιστη λύση. Η χρήση ενός πολύ μεγάλου βήματος μπορεί να κατευθύνει την πολιτική προς την εσφαλμένη κατεύθυνση, έχοντας έτσι μικρή πιθανότητα ανάκαμψης. Αντίθετα, η χρήση ενός πολύ μικρού βήματος μειώνει τη συνολική απόδοση. Κατά συνέπεια, το PPO εφαρμόζει μια συνάρτηση που περιορίζει την ενημέρωση πολιτικής ενός πράκτορα από το να είναι πολύ μεγάλη ή πολύ μικρή.

Τι είναι το PPO;

Όταν αναφέρεται κάποιος στην αξιολόγηση ενός πράκτορα, εννοεί γενικά την αξιολόγηση της συνάρτησης πολιτικής για να μάθει πόσο καλά αποδίδει ο πράκτορας, ακολουθώντας τη δεδομένη πολιτική. Στο σημείο αυτό, οι μέθοδοι Πολιτικής Διαβάθμισης διαδραματίζουν ζωτικό ρόλο. Όταν ένας πράκτορας «μαθαίνει» και δεν γνωρίζει πραγματικά ποιες ενέργειες αποφέρουν το καλύτερο αποτέλεσμα στις αντίστοιχες καταστάσεις, το κάνει υπολογίζοντας τις κλίσεις πολιτικής. Λειτουργεί σαν μια αρχιτεκτονική νευρωνικών δικτύων, όπου η κλίση της εξόδου, δηλαδή το αρχείο καταγραφής των πιθανοτήτων των ενεργειών στη συγκεκριμένη κατάσταση, λαμβάνεται σε σχέση με τις παραμέτρους του περιβάλλοντος και η αλλαγή αντικατοπτρίζεται στην πολιτική, με βάση τις κλίσεις, όπως φαίνεται και στην Εικόνα 7.



Εικόνα 8: Αναπαράσταση της λειτουργίας του PPO

Ενώ αυτή η δοκιμασμένη μέθοδος λειτουργεί καλά, τα κύρια μειονεκτήματα αυτών των μεθόδων είναι η υπερευαισθησία τους στον συντονισμό υπερπαραμέτρων, όπως η επιλογή του μεγέθους βημάτων, ο ρυθμός εκμάθησης κ.λπ., μαζί με την κακή τους απόδοση δείγματος. Σε αντίθεση με την εποπτευόμενη μάθηση που έχει μια εγγυημένη διαδρομή προς την επιτυχία ή τη σύγκλιση με σχετικά λιγότερο συντονισμό υπερπαραμέτρων, η ενισχυτική μάθηση είναι πολύ πιο περίπλοκη με διάφορα κινούμενα μέρη που πρέπει να ληφθούν υπόψη. Το PPO στοχεύει να βρει μια ισορροπία μεταξύ σημαντικών παραγόντων όπως η ευκολία υλοποίησης, η ευκολία συντονισμού, η πολυπλοκότητα του δείγματος, η αποτελεσματικότητα του δείγματος και η προσπάθεια υπολογισμού μιας ενημέρωσης σε κάθε βήμα που ελαχιστοποιεί τη συνάρτηση κόστους, διασφαλίζοντας παράλληλα ότι η απόκλιση από την προηγούμενη πολιτική είναι σχετικά μικρή. Το PPO είναι στην πραγματικότητα, μια μέθοδος διαβάθμισης πολιτικής που μαθαίνει και από διαδικτυακά δεδομένα. Απλώς διασφαλίζει ότι η ενημερωμένη πολιτική δεν είναι πολύ διαφορετική από την παλιά πολιτική για να διασφαλίσει χαμηλή απόκλιση στην εκπαίδευση.

Σύντομη επισκόπηση του τρόπου λειτουργίας του PPO

Παρακάτω ακολουθεί μια σύντομη επισκόπηση του τρόπου λειτουργίας του PPO.

1. Μέθοδοι κλίσης πολιτικής: Το PPO βασίζεται σε μεθόδους κλίσης πολιτικής, οι οποίες βελτιστοποιούν άμεσα τη συνάρτηση πολιτικής που αντιστοιχίζει καταστάσεις σε ενέργειες. Αυτό έρχεται σε αντίθεση με τις μεθόδους που υπολογίζουν τις συναρτήσεις τιμών.
2. Αντικειμενική λειτουργία: Το PPO στοχεύει στη μεγιστοποίηση της αναμενόμενης αθροιστικής ανταμοιβής που προκύπτει από την αλληλεπίδραση με το περιβάλλον. Αυτό γίνεται συνήθως με τη μεγιστοποίηση μιας συνάρτησης υποκατάστατου στόχου που προσεγγίζει τη βελτίωση της πολιτικής.
3. Αντικείμενο αντικατάστασης με περικοπή: Ένα από τα βασικά χαρακτηριστικά του PPO είναι ο αντικειμενικός στόχος με περικοπή, ο οποίος αποτρέπει μεγάλες ενημερώσεις πολιτικής που θα μπορούσαν να οδηγήσουν σε καταστροφικά αποτελέσματα. Περικόπτοντας την αναλογία μεταξύ της πιθανότητας ενεργειών

στο πλαίσιο της νέας πολιτικής και της παλιάς πολιτικής, το PPO διασφαλίζει ότι η ενημέρωση πολιτικής παραμένει εντός ασφαλούς εύρους.

4. Multiple Epochs και Mini-Batch Updates: Το PPO συνήθως περιλαμβάνει πολλαπλές εποχές αλληλεπίδρασης με το περιβάλλον, κατά τις οποίες συλλέγονται οι τροχιές. Αυτές οι τροχιές χρησιμοποιούνται στη συνέχεια για τον υπολογισμό της συνάρτησης υποκατάστατου στόχου, η οποία βελτιστοποιείται χρησιμοποιώντας ενημερώσεις mini-batch.
5. Εκτίμηση συνάρτησης τιμής: Το PPO συχνά ενσωματώνει εκτίμηση συνάρτησης τιμής για να μειώσει τη διακύμανση στις εκτιμήσεις κλίσης. Αυτό βοηθά στη σταθεροποίηση της προπόνησης και στη βελτίωση της αποτελεσματικότητας του δείγματος.
6. Παραλληλισμός: Το PPO μπορεί να παραλληλιστεί για να επιταχύνει την εκπαίδευση συλλέγοντας τροχιές από πολλές περιπτώσεις του περιβάλλοντος ταυτόχρονα.

Αρχιτεκτονική του PPO

Η αρχιτεκτονική του PPO (Proximal Policy Optimization) αναφέρεται κυρίως στην αρχιτεκτονική νευρωνικών δικτύων που χρησιμοποιείται για την αναπαράσταση των συναρτήσεων πολιτικής και αξίας στο πλαίσιο της μάθησης βαθιάς ενίσχυσης. Ακολουθεί μια ανάλυση των τυπικών στοιχείων:

- *Δίκτυο πολιτικής:* Το δίκτυο πολιτικών είναι ένα νευρωνικό δίκτυο που παίρνει την τρέχουσα κατάσταση του περιβάλλοντος ως είσοδο και εξάγει μια κατανομή πιθανότητας σε πιθανές ενέργειες. Αυτή η κατανομή αντιπροσωπεύει την πολιτική του πράκτορα, η οποία καθορίζει τις πιθανότητες ανάληψης κάθε ενέργειας δεδομένης της τρέχουσας κατάστασης. Η αρχιτεκτονική του δικτύου πολιτικής μπορεί να ποικίλλει ανάλογα με την πολυπλοκότητα του περιβάλλοντος και την εργασία που εκτελείται. Οι συνήθεις επιλογές περιλαμβάνουν νευρωνικά δίκτυα τροφοδοσίας, επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) ή συνελικτικά νευρωνικά δίκτυα (CNN), ανάλογα με το αν ο χώρος κατάστασης είναι δομημένος (π.χ. εικόνες) ή διαδοχικός (π.χ. δεδομένα χρονοσειράς).
- *Δίκτυο αξιών:* Εκτός από το δίκτυο πολιτικής, το PPO συχνά περιλαμβάνει ένα δίκτυο αξιών. Το δίκτυο τιμών υπολογίζει την αναμενόμενη σωρευτική ανταμοιβή (τιμή) της ύπαρξης σε μια δεδομένη κατάσταση. Αυτή η εκτίμηση βοηθά στη μείωση της διακύμανσης των εκτιμήσεων της κλίσης πολιτικής και παρέχει πρόσθετο σήμα για εκμάθηση. Η αρχιτεκτονική του δικτύου τιμών μπορεί να είναι παρόμοια με το δίκτυο πολιτικής, αν και συνήθως εξάγει μια ενιαία τιμή αντί για μια κατανομή πιθανότητας.
- *Λειτουργίες ενεργοποίησης:* Τόσο στα δίκτυα πολιτικής όσο και στα δίκτυα αξίας, χρησιμοποιούνται διάφορες συναρτήσεις ενεργοποίησης για την εισαγωγή μη γραμμικότητας στο δίκτυο. Οι συνήθεις επιλογές περιλαμβάνουν διορθωμένες γραμμικές μονάδες (ReLU), σιγμοειδείς ή υπερβολικές εφαπτομένες (tanh), ανάλογα με τις απαιτήσεις της συγκεκριμένης εργασίας και της αρχιτεκτονικής δικτύου.
- *Λειτουργίες απώλειας:* Το PPO χρησιμοποιεί συναρτήσεις απώλειας για την εκπαίδευση των δικτύων πολιτικής και αξίας. Για το δίκτυο πολιτικής, η συνάρτηση απώλειας βασίζεται συνήθως στον υποκατάστατο στόχο, ο οποίος

στοχεύει στη μεγιστοποίηση της αναμενόμενης σωρευτικής ανταμοιβής υπό τον περιορισμό του μεγέθους ενημέρωσης πολιτικής. Για το δίκτυο τιμών, η συνάρτηση απώλειας βασίζεται συχνά στο μέσο τετραγωνικό σφάλμα (MSE) μεταξύ των προβλεπόμενων και των πραγματικών τιμών.

- *Αλγόριθμος βελτιστοποίησης:* Το PPO χρησιμοποιεί αλγόριθμους βελτιστοποίησης για την ενημέρωση των παραμέτρων των δικτύων πολιτικής και τιμών με βάση τις υπολογισμένες συναρτήσεις απώλειας. Για το σκοπό αυτό χρησιμοποιούνται συνήθως παραλλαγές στοχαστικής κλίσης (SGD) όπως το Adam ή το RMSProp.

Παράδειγμα: Πλοήγηση σε έναν κόσμο πλέγματος

Σε αυτό το παράδειγμα, ο πράκτορας τοποθετείται σε ένα περιβάλλον του κόσμου του πλέγματος όπου χρειάζεται να πλοηγηθεί από ένα σημείο εκκίνησης σε μια τοποθεσία στόχου αποφεύγοντας τα εμπόδια. Ο πράκτορας λαμβάνει μια θετική ανταμοιβή όταν φτάσει στο στόχο και μια αρνητική ανταμοιβή εάν συγκρουστεί με ένα εμπόδιο. Θα χρησιμοποιηθεί ο PPO για την εκπαίδευση ο πράκτορας να μάθει μια βέλτιστη πολιτική για την πλοήγηση στον κόσμο του δικτύου.

Συστατικά παραδείγματος:

1. *Χώρος κατάστασης:* Ο κόσμος του πλέγματος αναπαρίσταται ως ένα διακριτό σύνολο καταστάσεων, όπου κάθε κατάσταση αντιστοιχεί σε μια θέση στο πλέγμα.
2. *Χώρος δράσης:* Ο πράκτορας μπορεί να κάνει διακριτές ενέργειες, όπως μετακίνηση προς τα πάνω, προς τα κάτω, προς τα αριστερά ή προς τα δεξιά.
3. *Ανταμοιβές:* Ο πράκτορας λαμβάνει μια ανταμοιβή +10 κατά την επίτευξη του στόχου, -10 όταν συγκρούεται με ένα εμπόδιο και -1 για κάθε βήμα που γίνεται.

Βήματα αλγόριθμου PPO:

1. *Αρχικοποίηση πολιτικής νευρωνικού δικτύου:* Ξεκινώντας, γίνεται αρχικοποίηση ενός νευρωνικού δικτύου που αντιπροσωπεύει την πολιτική. Αυτό το δίκτυο παίρνει την τρέχουσα κατάσταση ως είσοδο και εξάγει μια κατανομή πιθανότητας σε πιθανές ενέργειες.
2. *Συλλογή τροχιών:* Ο πράκτορας αλληλεπιδρά με το περιβάλλον ακολουθώντας την τρέχουσα πολιτική του. Κατά τη διάρκεια αυτής της αλληλεπίδρασης, συλλέγει τροχιές που αποτελούνται από καταστάσεις, ενέργειες και ανταμοιβές.
3. *Υπολογισμός εκτιμήσεων πλεονεκτημάτων:* Χρησιμοποιώντας τις συλλεγόμενες τροχιές, γίνεται υπολογισμός των εκτιμήσεων των πλεονεκτημάτων για κάθε ζεύγος κατάστασης-ενέργειας. Οι εκτιμήσεις πλεονεκτημάτων αντιπροσωπεύουν πόσο καλύτερη ή χειρότερη είναι μια ενέργεια σε σύγκριση με τη μέση ενέργεια που λαμβάνεται από μια δεδομένη κατάσταση.
4. *Υπολογισμός υποκατάστατου στόχου:* Γίνεται υπολογισμός του υποκατάστατου στόχου, ο οποίος είναι συνάρτηση των παλαιών και νέων παραμέτρων του ασφαλιστηρίου και των εκτιμήσεων πλεονεκτημάτων. Αυτός ο στόχος προσεγγίζει τη βελτίωση της πολιτικής, διασφαλίζοντας ταυτόχρονα ότι η ενημέρωση πολιτικής παραμένει εντός ασφαλούς εύρους.
5. *Optimize Policy Network:* Γίνεται χρήση της στοχαστικής κλίσης κατάβασης (SGD) ή μια παραλλαγή για την ενημέρωση των παραμέτρων του δικτύου πολιτικής, ελαχιστοποιώντας τον υποκατάστατο στόχο.

Επανάληψη: Τα βήματα 2-5 επαναλαμβάνονται για πολλαπλές επαναλήψεις ή μέχρι τη σύγκλιση.

Παράδειγμα επανάληψης:

Αρχικοποίηση: Γίνεται τυχαία αρχικοποίηση του δικτύου πολιτικών.

Αλληλεπίδραση με το περιβάλλον: Ο πράκτορας ακολουθεί την τρέχουσα πολιτική του για πλοήγηση στον κόσμο του πλέγματος, συλλέγοντας τροχιές.

Εκτίμηση πλεονεκτημάτων: Χρησιμοποιώντας τις συλλεγόμενες τροχιές, υπολογίζονται τα πλεονεκτήματα για κάθε ζεύγος κατάστασης-δράσης με βάση τις ανταμοιβές που ελήφθησαν.

Υπολογισμός υποκατάστατου στόχου: Υπολογίζεται ο υποκατάστατος στόχο χρησιμοποιώντας τις εκτιμήσεις πλεονεκτημάτων και τις παλιές και νέες παραμέτρους της πολιτικής.

Optimize Policy Network: Ενημερώνονται οι παραμέτροι του δικτύου πολιτικής χρησιμοποιώντας SGD για την ελαχιστοποίηση του υποκατάστατου στόχου.

Αξιολόγηση: Αξιολογείται η ενημερωμένη πολιτική αφήνοντας τον πράκτορα να περιηγηθεί ξανά στον κόσμο του πλέγματος και να παρατηρήσει την απόδοσή του.

Επανάληψη: Τα βήματα 2-6 επαναλαμβάνονται για πολλές επαναλήψεις έως ότου η πολιτική συγκλίνει σε μια βέλτιστη λύση.

Μέσω αυτής της επαναληπτικής διαδικασίας, ο πράκτορας μαθαίνει σταδιακά μια βέλτιστη πολιτική για την πλοήγηση στον κόσμο του πλέγματος, εξισορροπώντας την εξερεύνηση και την εκμετάλλευση για να μεγιστοποιήσει τις σωρευτικές ανταμοιβές αποφεύγοντας τα εμπόδια. Το PPO εξασφαλίζει σταθερή και αποτελεσματική μάθηση περιορίζοντας τις ενημερώσεις πολιτικής και αξιοποιώντας τις εκτιμήσεις πλεονεκτημάτων για να καθοδηγήσουν τη διαδικασία εκμάθησης.

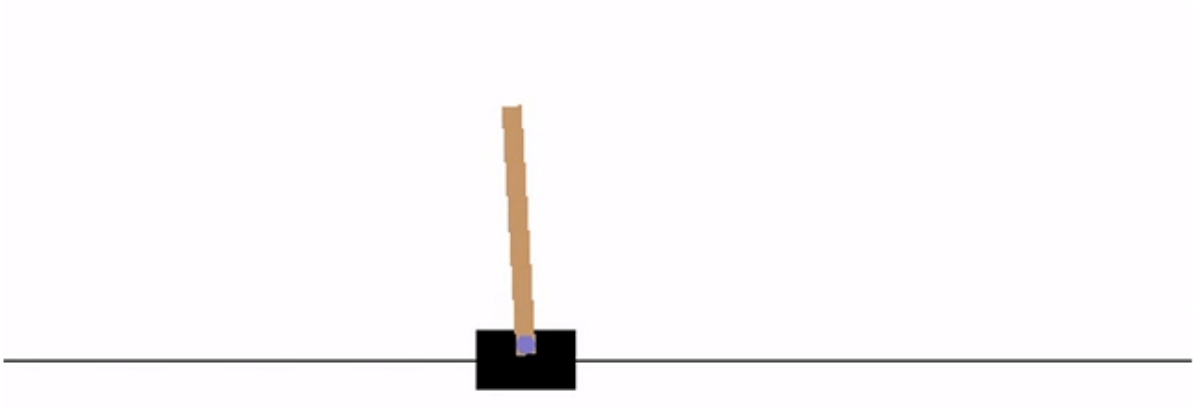
Ο κώδικας στο Παράρτημα 3 υλοποιεί τον αλγόριθμο PPO με μια απλή αρχιτεκτονική νευρωνικών δικτύων κρίσης χρησιμοποιώντας το TensorFlow. Το δίκτυο εξάγει τα αποτελέσματα της διανομής πολιτικής, ενώ το δίκτυο κριτικών εκτιμά την τιμή κατάστασης. Η συνάρτηση απώλειας συνδυάζει την απώλεια διαβάθμισης πολιτικής και την απώλεια αξίας με έναν αποκομμένο στόχο τύπου PPO. Ο βρόχος εκπαίδευσης συλλέγει τροχιές από το περιβάλλον, υπολογίζει τα πλεονεκτήματα και εκτελεί ενημερώσεις PPO για έναν καθορισμένο αριθμό εποχών. Τέλος, ο κώδικας εκπαιδεύει τον πράκτορα στο περιβάλλον CartPole από το OpenAI Gym για έναν καθορισμένο αριθμό επεισοδίων.

4.4 Σύγκριση αποτελεσμάτων

Στην ενότητα αυτή, έχοντας κάνει ήδη μια περιγραφή των δύο αλγορίθμων DQN και PPO, θα γίνει μια συγκριτική μελέτη από ένα παράδειγμα χρήσης τους. Στην πρώτη περίπτωση, στον αλγόριθμο DQN, θα μελετηθεί το περιβάλλον CartPole, το οποίο είναι και το πιο επικρατές για τη χρήση του, ενώ στη δεύτερη περίπτωση θα γίνει η περιγραφή μιας εκπαίδευσης ενός παραμετρικού δικτύου πολιτικής για την επίλυση ενός προβλήματος.

4.4.1 DQN

Στο παράδειγμα αυτό, [28], θα γίνει χρήση του PyTorch για την εκπαίδευση ενός DQN agent (DQN πράκτορα) στο περιβάλλον CartPole της βιβλιοθήκης Gymnasium. Ο πράκτορας πρέπει να αποφασίσει μεταξύ δύο ενεργειών - μετακίνησης του καροτσιού αριστερά ή δεξιά - έτσι ώστε το κοντάρι που είναι συνδεδεμένο σε αυτό να παραμείνει όρθιο. Παράδειγμα του CartPole βρίσκεται στην Εικόνα 9.



Εικόνα 9: Παράδειγμα του CarPole

Καθώς ο πράκτορας παρατηρεί την τρέχουσα κατάσταση του περιβάλλοντος και επιλέγει μια ενέργεια, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση και επίσης επιστρέφει μια ανταμοιβή που υποδεικνύει τις συνέπειες της ενέργειας. Σε αυτήν την εργασία, οι ανταμοιβές είναι +1 για κάθε σταδιακό χρονικό βήμα και το περιβάλλον τερματίζεται εάν ο πόλος πέσει πολύ μακριά ή το καλάθι απομακρυνθεί περισσότερο από 2,4 μονάδες από το κέντρο. Αυτό σημαίνει ότι τα σενάρια με καλύτερη απόδοση θα εκτελούνται για μεγαλύτερη διάρκεια, συσσωρεύοντας μεγαλύτερη απόδοση.

Το περιβάλλον CartPole έχει σχεδιαστεί έτσι ώστε οι είσοδοι στον πράκτορα να είναι 4 πραγματικές τιμές που αντιπροσωπεύουν την κατάσταση περιβάλλοντος (θέση, ταχύτητα, κ.λπ.). Λαμβάνονται αυτές οι 4 είσοδοι χωρίς καμία κλιμάκωση και περνιούνται μέσα από ένα μικρό πλήρως συνδεδεμένο δίκτυο με 2 εξόδους, μία για κάθε ενέργεια. Το δίκτυο εκπαιδεύεται να προβλέπει την αναμενόμενη τιμή για κάθε ενέργεια, δεδομένης της κατάστασης εισόδου. Στη συνέχεια επιλέγεται η ενέργεια με την υψηλότερη αναμενόμενη τιμή.

Μνήμη επανάληψης - Replay Memory

Στο παράδειγμα αυτό θα γίνει χρήση της μνήμης επανάληψης για την εκπαίδευση του DQN. Η μνήμη αυτή αποθηκεύει τις μεταβάσεις που παρατηρεί ο πράκτορας, επιτρέποντας τη χρήση αυτών των δεδομένων ξανά αργότερα. Με τη δειγματοληψία από αυτό με τυχαίο τρόπο, οι μεταβάσεις που δημιουργούν μια παρτίδα αποσυσχετίζονται. Έχει αποδειχθεί ότι αυτό σταθεροποιεί και βελτιώνει σημαντικά τη διαδικασία εκπαίδευσης DQN.

Για αυτό, χρειάζονται δύο κλάσεις:

Μετάβαση - μια πλειάδα που αντιπροσωπεύει μια μετάβαση στο περιβάλλον. Ουσιαστικά απεικονίζει το ζεύγος (κατάσταση, επόμενη δράση) στο αποτέλεσμά τους (επόμενη_κατάσταση, ανταμοιβή).

ReplayMemory - μια κυκλική προσωρινή μνήμη περιορισμένου μεγέθους που διατηρεί τις μεταβάσεις που παρατηρήθηκαν πρόσφατα. Εφαρμόζει επίσης μια μέθοδο .sample() για την επιλογή μιας τυχαίας παρτίδας (batch) μεταβάσεων για εκπαίδευση.

Περιγραφή αλγορίθμου

Το περιβάλλον είναι ντετερμινιστικό, επομένως όλες οι εξισώσεις που παρουσιάζονται εδώ διατυπώνονται επίσης ντετερμινιστικά για λόγους απλότητας. Σε εκτενέστερες μελέτες, για την ενίσχυση της μάθησης, θα μπορούσαν να συμπεριληφθούν και στοχαστικές μεταβάσεις στο περιβάλλον. Στόχος είναι η εκπαίδευση μιας πολιτικής που προσπαθεί να μεγιστοποιήσει τη σωρευτική ανταμοιβή (reward) $R_{t_0} = \sum_{t=t0}^{\infty} \gamma^{t-t_0} r_t$, όπου το R_{t_0} είναι επίσης γνωστό και ως η *επιστροφή*. Η «έκπτωση» γ είναι μια σταθερά στο διάσημα $[0, 1]$ που εξασφαλίζει ότι το άθροισμα συγκλίνει. Μια μικρή τιμή του γ κάνει τις ανταμοιβές από το αβέβαιο μακρινό μέλλον λιγότερο σημαντικές για τον πράκτορα (agent) από εκείνες στο εγγύς μέλλον για τις οποίες μπορεί να είναι αρκετά σίγουρος. Ενθαρρύνει επίσης τους πράκτορες να συλλέγουν ανταμοιβές πιο κοντά στο χρόνο από αντίστοιχες ανταμοιβές που είναι προσωρινά μακριά στο μέλλον.

Η κύρια ιδέα πίσω από το Q-learning είναι ότι αν υπήρχε μια συνάρτηση $Q^* = \text{κατάσταση} \times \text{Επόμενη δράση} \rightarrow R$, που θα μπορούσε να πει ποια θα ήταν η επιστροφή εάν επρόκειτο να γίνει μια ενέργεια σε μια δεδομένη κατάσταση, τότε θα μπορούσε εύκολα να δημιουργηθεί μια πολιτική που να μεγιστοποιεί τις ανταμοιβές:

$$\pi^*(s) = argmax_{\alpha} Q^*(s, \alpha)$$

Δίκτυο Q

Το μοντέλο θα είναι ένα νευρωνικό δίκτυο τροφοδοσίας που θα λαμβάνει τη διαφορά ανάμεσα στις τρέχουσες και τις προηγούμενες καταστάσεις. Έχει δύο εξόδους, που αντιπροσωπεύουν $Q(s, left)$ και $Q(s, right)$ (όπου s είναι η είσοδος στο δίκτυο). Στην πραγματικότητα, το δίκτυο προσπαθεί να προβλέψει την αναμενόμενη απόδοση της εκτέλεσης κάθε ενέργειας με δεδομένη την τρέχουσα είσοδο.

Εκπαίδευση μοντέλου – υπερπαράμετροι και διάφορα εργαλεία

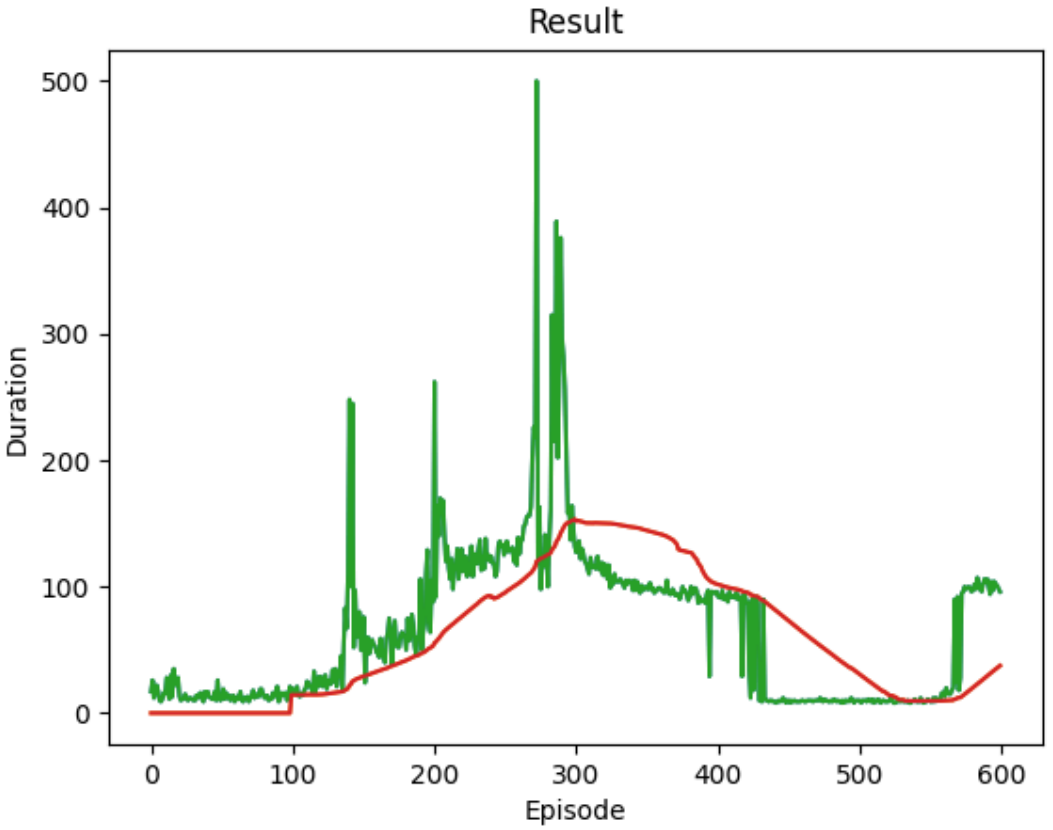
Συνάρτηση select_action: Θα επιλέξει μια ενέργεια σύμφωνα με μια άπληστη πολιτική epsilon. Με απλά λόγια, μερικές φορές θα χρησιμοποιηθεί το μοντέλο για την επιλογή της δράσης και μερικές φορές απλώς θα επιλεγεί τυχαία και ομοιόμορφα μια ενέργεια. Η πιθανότητα επιλογής μιας τυχαίας ενέργειας θα ξεκινήσει από το EPS_START και θα μειωθεί εκθετικά προς το EPS_END. Το EPS_DECAY ελέγχει το ρυθμό της επιλογής αυτής.

Συνάρτηση plot_durations: Δημιουργεί μια γραφική απεικόνιση για τη σχεδίαση της διάρκειας των επεισοδίων, μαζί με έναν μέσο όρο των τελευταίων 100 επεισοδίων (το μέτρο που χρησιμοποιείται στις επίσημες αξιολογήσεις). Η γραφική παράσταση θα βρίσκεται κάτω από το κελί που περιέχει τον κύριο βρόχο εκπαίδευσης και θα ενημερώνεται μετά από κάθε επεισόδιο.

Συνάρτηση optimize_model που εκτελεί ένα μόνο βήμα της βελτιστοποίησης. Αρχικά λαμβάνει δείγματα μιας παρτίδας, συνενώνει όλες τις εισόδους σε μία, υπολογίζει τα $Q(s_t, a_t)$ και $V(s_{t+1}) = Q(s_{t+1}, a)$ και τα συνδυάζει στις απώλειες. Εξ ορισμού ορίζεται

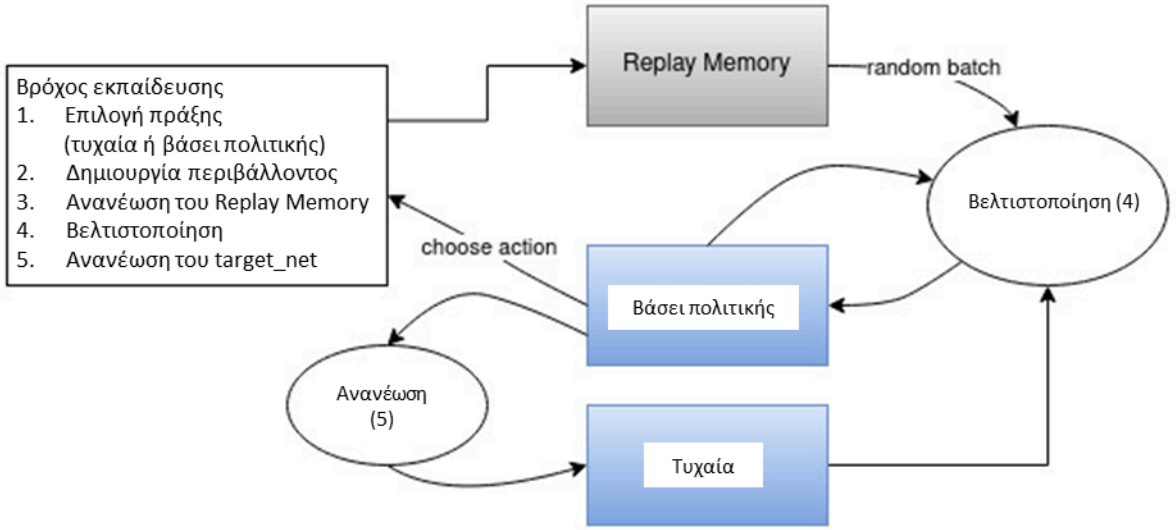
το $V(s) = 0$, αν το s είναι μια τερματική κατάσταση. Χρησιμοποιείται επίσης ένα δίκτυο προορισμού για τον υπολογισμό του $V(s_{t+1})$ για πρόσθετη σταθερότητα. Το δίκτυο προορισμού ενημερώνεται σε κάθε βήμα με μια απλή ενημέρωση που ελέγχεται από την υπερπαράμετρο TAU, η οποία είχε οριστεί προηγουμένως.

Το αποτέλεσμα εκτέλεσης του κώδικα παράγει το διάγραμμα που φαίνεται στην Εικόνα 10, το οποίο απεικονίζει πόσο χρόνο χρειάστηκε ο αλγόριθμος σε κάθε επεισόδιο.



Εικόνα 10: Διάγραμμα διάρκειας ανά επεισόδιο

Το παρακάτω διάγραμμα απεικονίζει τη συνολική προκύπτουσα ροή δεδομένων του κώδικα που βρίσκεται στο Παράρτημα 2.



Εικόνα 11: Διάγραμμα απεικόνισης της συνολικής προκύπτουσας ροής δεδομένων του κώδικα που βρίσκεται στο Παράρτημα 2.

Οι ενέργειες επιλέγονται είτε τυχαία είτε βάσει πολιτικής, λαμβάνοντας το επόμενο βήμα δείγματος από το περιβάλλον. Καταγράφονται τα αποτελέσματα στη μνήμη επανάληψης (Replay Memory) και επίσης εκτελείται το βήμα βελτιστοποίησης σε κάθε επανάληψη. Το βήμα αυτό επιλέγει μια τυχαία παρτίδα από τη μνήμη επανάληψης για να κάνει εκπαίδευση της νέας πολιτικής. Το "παλαιότερο" target_net χρησιμοποιείται επίσης στη βελτιστοποίηση για τον υπολογισμό των αναμενόμενων τιμών Q. Πραγματοποιείται μια απλή ενημέρωση των βαρών του σε κάθε βήμα.

4.4.2 PPO

Σε αυτή την ενότητα, παρουσιάζεται η χρήση της βιβλιοθήκης PyTorch και Torchrl για την εκπαίδευση ενός παραμετρικού δικτύου πολιτικής για την επίλυση της εργασίας Inverted Pendulum της βιβλιοθήκης Gymnasium, [29]. Στο παράδειγμα γίνεται χρήση του αλγορίθμου Proximal Policy Optimization (PPO). Ο PPO είναι ένας αλγόριθμος διαβάθμισης πολιτικής όπου μια παρτίδα δεδομένων συλλέγεται και καταναλώνεται απευθείας για να εκπαιδεύσει την πολιτική, ώστε να μεγιστοποιήσει την αναμενόμενη απόδοση, βάσει ορισμένων περιορισμών εγγύτητας. Πρακτικά είναι μια εξελιγμένη έκδοση του REINFORCE, του βασικού αλγόριθμου βελτιστοποίησης πολιτικής.

Ο PPO θεωρείται συνήθως ως μια γρήγορη και αποτελεσματική μέθοδος για διαδικτυακό αλγόριθμο ενίσχυσης πολιτικής. Το TorchRL παρέχει μια ενότητα απώλειας που πραγματοποιεί τους απαραίτητους υπολογισμούς, ώστε να μπορεί ο χρήστης να βασιστεί σε αυτήν την υλοποίηση και να εστιάσει στην επίλυση του προβλήματος αντί να το υλοποιεί από την αρχή κάθε φορά που επιθυμεί να εκπαιδεύσει μια πολιτική.

Ο αλγόριθμος λειτουργεί ως εξής:

1. Θα γίνει δειγματοληψία μιας παρτίδας δεδομένων παίζοντας την πολιτική στο περιβάλλον για ένα δεδομένου αριθμού βημάτων.

2. Στη συνέχεια, θα εκτελεστεί ένας δεδομένος αριθμός βημάτων βελτιστοποίησης με τυχαία υποδείγματα αυτής της παρτίδας, χρησιμοποιώντας μια αποκομμένη έκδοση της απώλειας REINFORCE.

3. Το απόκομμα θα θέσει ένα απαισιόδοξο όριο στην απώλεια: θα ευνοηθούν χαμηλότερες εκτιμήσεις απόδοσης σε σύγκριση με υψηλότερες. Ο ακριβής τύπος της απώλειας είναι:

$$L\left(s, a, \theta_k, \theta\right) = \left(\frac{\pi_\theta(s)}{\pi_{\theta_k}(s)} A^{\pi_{\theta_k}}(s, a),\ g\left(\varepsilon,\ A^{\pi_{\theta_k}}(s, a)\right)\right)$$

Υπάρχουν δύο στοιχεία σε αυτό τον τύπο της απώλειας: στο πρώτο μέρος του ελάχιστου τελεστή, υπολογίζεται απλώς μια σταθμισμένη έκδοση της απώλειας REINFORCE (για παράδειγμα, μια απώλεια REINFORCE που έχει διορθωθεί για το γεγονός ότι η τρέχουσα διαμόρφωση πολιτικής υστερεί ένα που χρησιμοποιήθηκε για τη συλλογή δεδομένων). Το δεύτερο μέρος αυτού του ελάχιστου τελεστή είναι μια παρόμοια απώλεια όπου έχει περικοπεί για τους λόγους όταν υπερέβησαν ή ήταν κάτω από ένα δεδομένο ζεύγος ορίων.

Αυτός ο τύπος διασφαλίζει ότι είτε το πλεονέκτημα είναι θετικό είτε αρνητικό, ενώ αποθαρρύνονται οι ενημερώσεις πολιτικής που θα προκαλούσαν σημαντικές αλλαγές από την προηγούμενη διαμόρφωση.

Αυτό το παράδειγμα είναι δομημένο ως εξής:

- Αρχικά, θα οριστεί ένα σύνολο υπερπαραμέτρων που θα χρησιμοποιθούν για την εκπαίδευση.
- Στη συνέχεια, θα δημιουργηθεί το περιβάλλον ή του προσομοιωτή, χρησιμοποιώντας τους μετασχηματισμούς του TorchRL.
- Στη συνέχεια, θα σχεδιαστεί το δίκτυο πολιτικής και το μοντέλο αξίας, το οποίο είναι απαραίτητο για τη συνάρτηση απώλειας. Αυτές οι μονάδες θα χρησιμοποιηθούν για τη διαμόρφωση της μονάδας απώλειας.
- Στη συνέχεια, θα δημιουργηθεί το buffer επανάληψης και το πρόγραμμα φόρτωσης δεδομένων.
- Τέλος, θα εκτελεστεί ο κύκλος προσομοίωσης και θα γίνει ανάλυση των αποτελεσμάτων.

Επίσης, σε αυτό το παράδειγμα, θα χρησιμοποιηθεί η βιβλιοθήκη TensorDict, η οποία βοηθάει στο να μπορεί ο χρήστης να εστιάσει πιο εύκολα στον ίδιο τον αλγόριθμο και όχι στην περιγραφή των δεδομένων.

Παράμετροι συλλογής δεδομένων

Κατά τη συλλογή δεδομένων, επιλέγεται πόσο μεγάλη θα είναι κάθε παρτίδα ορίζοντας μια παράμετρο frames_per_batch. Ορίζεται επίσης πόσα καρέ (όπως ο αριθμός των αλληλεπιδράσεων με τον προσομοιωτή) θα επιτρέπεται να χρησιμοποιηθούν. Γενικά, ο στόχος ενός αλγορίθμου RL είναι να μάθει να επιλύει την εργασία όσο πιο γρήγορα μπορεί όσον αφορά τις αλληλεπιδράσεις του περιβάλλοντος: όσο χαμηλότερα είναι τα total_frames τόσο το καλύτερο.

Παράμετροι PPO

Σε κάθε συλλογή δεδομένων (ή συλλογή παρτίδας) θα εκτελείται η βελτιστοποίηση σε έναν συγκεκριμένο αριθμό εποχών, κάθε φορά καταναλώνοντας ολόκληρα τα δεδομένα που μόλις αποκτήθηκαν σε έναν ένθετο βρόχο εκπαίδευσης. Εδώ, το sub_batch_size είναι διαφορετικό από το frames_per_batch που αναφέρθηκε παραπάνω: υπενθυμίζεται ότι πρόκειται για μια «παρτίδα δεδομένων» που προέρχεται από τον συλλέκτη, το μέγεθος του οποίου ορίζεται από frames_per_batch και ότι θα χωριστούν περαιτέρω σε μικρότερες δευτερεύουσες παρτίδες κατά τη διάρκεια του εσωτερικού βρόχου εκπαίδευσης. Το μέγεθος αυτών των δευτερεύουσων παρτίδων ελέγχεται από το sub_batch_size.

Ορισμός του περιβάλλοντος

Στο Reinforcement Learning, ένα περιβάλλον είναι συνήθως ο τρόπος που αναφερόμαστε σε έναν προσομοιωτή ή ένα σύστημα ελέγχου. Διάφορες βιβλιοθήκες παρέχουν περιβάλλοντα προσομοίωσης για ενισχυτική μάθηση, συμπεριλαμβανομένων των Gymnasium, της σουίτας ελέγχου DeepMind και πολλών άλλων. Ως γενική βιβλιοθήκη, στόχος της TorchRL είναι να παρέχει μια εναλλάξιμη διεπαφή σε ένα μεγάλο πάνελ προσομοιωτών RL, επιτρέποντας στο χρήστη να εναλλάσσεται εύκολα ένα περιβάλλον με ένα άλλο.

Μετασχηματισμοί

Αρκετές φορές γίνεται προσθήκη ορισμένων μετασχηματισμών σε ένα περιβάλλον για να προετοιμαστούν τα δεδομένα για την εφαρμογή της πολιτικής. Στο περιβάλλον Gym, αυτό συνήθως επιτυγχάνεται μέσω των wrappers. Το TorchRL υιοθετεί μια διαφορετική προσέγγιση, η οποία μοιάζει με άλλες pytorch βιβλιοθήκες, μέσω της χρήσης μετασχηματισμών. Για να προστεθούν μετασχηματισμοί σε ένα περιβάλλον, θα πρέπει απλώς να γίνει wrap σε ένα TransformedEnv και να προστεθεί η ακολουθία μετασχηματισμών σε αυτό. Το μετασχηματισμένο περιβάλλον θα κληρονομήσει τα στοιχεία του wrapped περιβάλλοντος και θα τα μετατρέψει ανάλογα με την ακολουθία μετασχηματισμών που περιέχει.

Πολιτική

Ο PPO χρησιμοποιεί μια στοχαστική πολιτική για να χειριστεί την εξερεύνηση. Αυτό σημαίνει ότι το νευρωνικό δίκτυο θα πρέπει να εξάγει τις παραμέτρους μιας διανομής, αντί για μια μεμονωμένη τιμή που αντιστοιχεί στην ενέργεια που έγινε.

Καθώς τα δεδομένα είναι συνεχή, χρησιμοποιείται μια κατανομή Tanh-Normal για να παραμείνει στα όρια του πεδίου ορισμού. Η TorchRL παρέχει μια τέτοια διανομή και το μόνο πράγμα που πρέπει να ληφθεί υπόψη είναι να δημιουργηθεί ένα νευρωνικό δίκτυο που θα εξάγει τον σωστό αριθμό παραμέτρων για να λειτουργήσει η πολιτική.

Δίκτυο τιμών

Το δίκτυο τιμών είναι ένα κρίσιμο συστατικό του αλγορίθμου PPO, παρόλο που δε θα χρησιμοποιηθεί κατά τον χρόνο συμπερασμάτων. Αυτή η μονάδα θα διαβάσει τις παρατηρήσεις και θα επιστρέψει μια εκτίμηση της μειωμένης απόδοσης για την ακόλουθη τροχιά. Αυτό επιτρέπει την απόσβεση της μάθησης βασιζόμενη σε κάποια εκτίμηση χρησιμότητας που μαθαίνεται αμέσως κατά τη διάρκεια της εκπαίδευσης. Το δίκτυο τιμών μοιράζεται την ίδια δομή με την πολιτική, αλλά για λόγους απλότητας του εκχωρείται το δικό του σύνολο παραμέτρων.

Συλλέκτης δεδομένων

Το TorchRL παρέχει ένα σύνολο κλάσεων DataCollector. Εν συντομία, αυτές οι κλάσεις εκτελούν τρεις λειτουργίες: επαναφέρουν ένα περιβάλλον, υπολογίζουν μια ενέργεια δεδομένης της πιο πρόσφατης παρατήρησης, εκτελούν ένα βήμα στο περιβάλλον και επαναλαμβάνουν τα δύο τελευταία βήματα έως ότου το περιβάλλον σηματοδοτήσει μια διακοπή (ή φτάσει σε κατάσταση ολοκληρωμένης).

Επιτρέπουν τον έλεγχο για το πόσα καρέ να συλλέγονται σε κάθε επανάληψη (μέσω της παραμέτρου frames_per_batch), πότε να επαναφέρεται το περιβάλλον (μέσω του ορίσματος max_frames_per_traj), σε ποια συσκευή θα πρέπει να εκτελεστεί η πολιτική κ.λπ. Έχουν επίσης σχεδιαστεί για να λειτουργούν αποτελεσματικά με ομαδικά και πολυεπεξεργασμένα περιβάλλοντα.

Ο απλούστερος συλλέκτης δεδομένων είναι ο SyncDataCollector: είναι ένας επαναλήπτης που μπορεί να χρησιμοποιηθεί για τη λήψη παρτίδων δεδομένων συγκεκριμένου μήκους και που θα σταματήσει μόλις συλλεχθεί ένας συνολικός αριθμός πλαισίων (total_frames). Άλλοι συλλέκτες δεδομένων (MultiSyncDataCollector και MultiaSyncDataCollector) θα εκτελούν τις ίδιες λειτουργίες με σύγχρονο και ασύγχρονο τρόπο σε ένα σύνολο πολυεπεξεργασμένων εργαζομένων.

Replay buffer – Buffer επανάληψης

Τα replay buffer (buffer επανάληψης) είναι ένα κοινό στοιχείο κατασκευής αλγορίθμων RL εκτός πολιτικής (off-policy RL). Σε περιβάλλοντα εντός πολιτικής (on-policy RL), ένα buffer επανάληψης ξαναγεμίζεται κάθε φορά που συλλέγεται μια παρτίδα δεδομένων και τα δεδομένα του καταναλώνονται επανειλημμένα για συγκεκριμένο αριθμό εποχών.

Τα buffer επανάληψης του TorchRL κατασκευάζονται χρησιμοποιώντας ένα κοινό ReplayBuffer το οποίο λαμβάνει ως όρισμα τα στοιχεία του buffer: έναν χώρο αποθήκευσης, έναν εγγραφέα, έναν δειγματολήπτη και πιθανώς ορισμένους μετασχηματισμούς. Μόνο η αποθήκευση (που υποδεικνύει τη χωρητικότητα buffer επανάληψης) είναι υποχρεωτική. Καθορίζεται επίσης ένα δειγματολήπτη χωρίς επανάληψη για την αποφυγή της δειγματοληψία πολλές φορές του ίδιου στοιχείου σε μια εποχή. Η χρήση ενός buffer επανάληψης για το PPO δεν είναι υποχρεωτική και θα μπορούσε απλώς να δειγματιστούν οι δευτερεύουσες παρτίδες από τη συλλεγόμενη παρτίδα, αλλά η χρήση αυτών των κατηγοριών διευκολύνει τη δημιουργία του εσωτερικού βρόχου εκπαίδευσης με έναν αναπαραγώγιμο τρόπο.

Συνάρτηση απώλειας

Η απώλεια στο PPO μπορεί να εισαχθεί απευθείας από την TorchRL για ευκολία χρησιμοποιώντας την κλάση ClipPPOLoss. Αυτός είναι ο ευκολότερος τρόπος χρήσης του PPO: κρύβει τις μαθηματικές πράξεις του PPO και τη ροή ελέγχου που συνοδεύει.

Το PPO απαιτεί να υπολογιστεί κάποια «εκτίμηση πλεονεκτημάτων». Εν ολίγοις, ένα πλεονέκτημα είναι μια τιμή που αντικατοπτρίζει μια προσδοκία σε σχέση με την τιμή απόδοσης, ενώ αντιμετωπίζει την ανταλλαγή μεροληψίας / διακύμανσης. Για να υπολογίσει κανείς το πλεονέκτημα, χρειάζεται απλώς (1) να δημιουργήσει τη μονάδα πλεονεκτήματος, η οποία χρησιμοποιεί τον τελεστή αξίας μας και (2) να περάσει κάθε παρτίδα δεδομένων μέσω αυτής πριν από κάθε εποχή. Η λειτουργική μονάδα GAE θα

ενημερώσει τον δείκτη εισόδου με νέες καταχωρήσεις "advantage" και "value_target". Το "value_target" αντιπροσωπεύει την εμπειρική τιμή που πρέπει να αντιπροσωπεύει το δίκτυο τιμών με την παρατήρηση εισόδου. Και τα δύο θα χρησιμοποιηθούν από το ClipPPOLoss για να επιστρέψει την πολιτική και τις απώλειες.

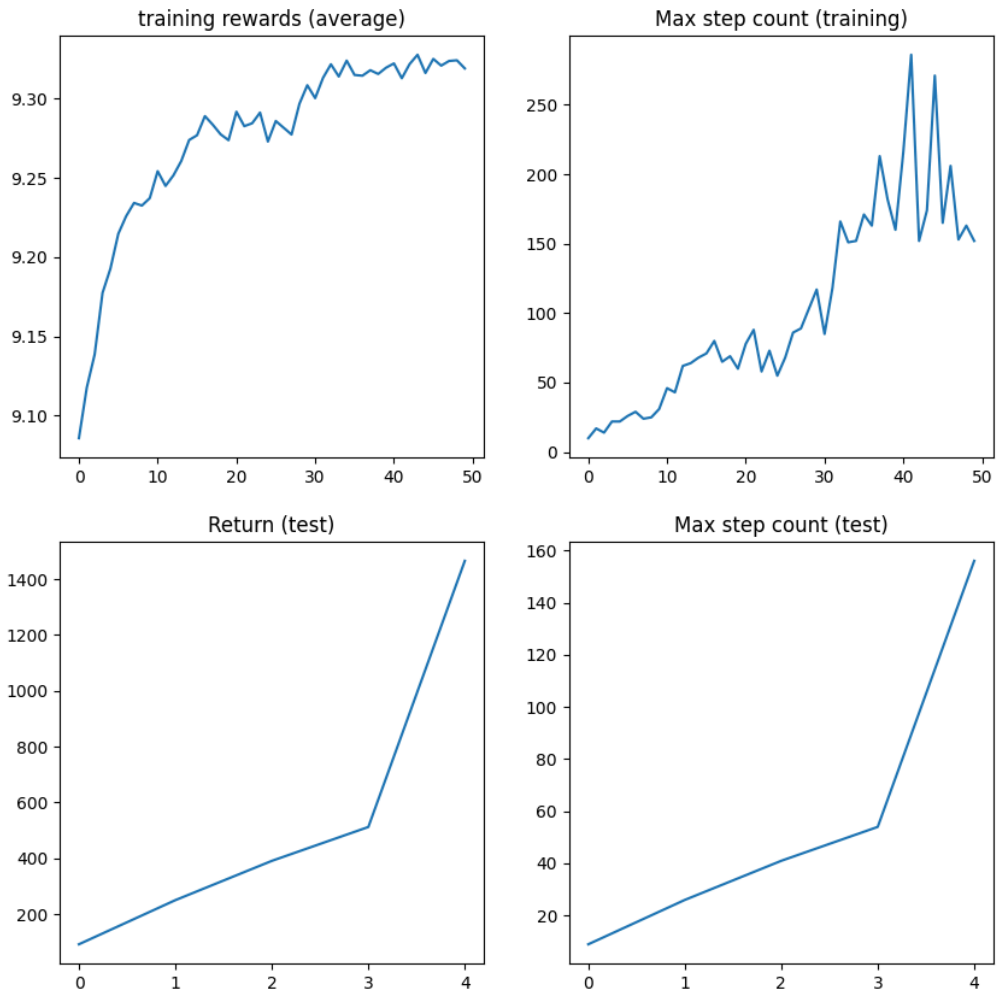
Βρόχος εκπαίδευσης

Με βάση τα παραπάνω, μπορεί να γραφτεί πλέον και ο βρόχος εκπαίδευσης που θα χρησιμοποιηθεί. Τα βήματά του περιλαμβάνουν:

- Συλλογή δεδομένων (collect data)
    - Υπολογισμός πλεονεκτήματος (compute advantage)
        - Βρόχος στις συλλεγόμενες τιμές για τον υπολογισμό των απωλειών (compute loss values)
        - Διάδοση αυτών προς τα πίσω (back propagate)
        - Βελτιστοποίηση (optimize)
        - Επανάληψη (repeat)
    - Επανάληψη (repeat)
- Επανάληψη (repeat)

Αποτελέσματα

Προτού συμπληρωθεί το ανώτατο όριο βημάτων του ενός εκατομμυρίου, ο αλγόριθμος θα πρέπει να έχει φτάσει σε μέγιστο αριθμό βημάτων 1000 βημάτων, που είναι ο μέγιστος αριθμός βημάτων πριν από την περικοπή της επανάληψης.



Εικόνα 12: Αποτελέσματα εκτέλεσης του PPO

Οι γραφικές παραστάσεις στην Εικόνα 12 οπτικοποιούν τα εξής αποτελέσματα:

α. training rewards (average): Δείχνει τη μέση ανταμοιβή (reward) που παρατηρήθηκε κατά τα βήματα εκτέλεσης του αλγορίθμου. Ο οριζόντιος άξονας είναι σε χιλιάδες και είναι εμφανές ότι στα τελευταία βήματα του αλγορίθμου, το reward είναι σαφώς μεγαλύτερο από τα πρώτα βήματα.

β. Max step count (training): Και σε αυτή τη γραφική παράσταση, ο οριζόντιος άξονας είναι σε χιλιάδες. Απεικονίζεται το μέγιστο πλήθος βημάτων που χρειάστηκαν σε κάθε στάδιο εκτέλεσης του αλγορίθμου. Όπως και στο (α), έτσι και εδώ, στα τελευταία βήματα χρειάστηκαν αρκετά παραπάνω βήματα, καθώς ο αλγόριθμος ήταν ήδη εκπαιδευμένος καλύτερα από τα πρώτα βήματα.

γ. Return (test): Στη γραφική αυτή παράσταση φαίνεται η συνολική ανταμοιβή (reward) που κρατιέται κάθε 10000 βήματα.

δ. Max step count (test): Στη γραφική αυτή παράσταση φαίνεται το συνολικό πλήθος βημάτων (steps) που έγιναν κάθε 10000 βήματα.

4.5 Εφαρμογή νευρωνικών δικτύων για την εκτίμηση καναλιών σε τηλεπικοινωνιακά συστήματα και UAV

Η μοντελοποίηση καναλιών μη επανδρωμένων εναέριων οχημάτων (UAV) από ασύρματες επικοινωνίες έχει κερδίσει μεγάλο ενδιαφέρον για ταχεία ανάπτυξη στην ασύρματη επικοινωνία. Το κανάλι UAV έχει τα δικά του διακριτικά χαρακτηριστικά σε σύγκριση με τα δορυφορικά και τα κυψελωτά δίκτυα. Πολλές προτεινόμενες τεχνικές θεωρούν και διατυπώνουν τη μοντελοποίηση καναλιών των UAV ως πρόβλημα ταξινόμησης, όπου το κλειδί είναι η εξαγωγή των διακριτικών χαρακτηριστικών του ασύρματου σήματος UAV. Για αυτό το ζήτημα, προτείνουμε ένα πλαίσιο πολλαπλών περιορισμένων μηχανών Boltzmann Gaussian-Bernoulli (GBRBM) για μείωση διαστάσεων και προ-εκπαιδευτική χρήση ενσωματωμένο σε ένα βαθύ νευρωνικό δίκτυο που βασίζεται σε αυτοκωδικοποιητή. Το αναπτυγμένο σύστημα χρησιμοποίησε μετρήσεις UAV του ήδη υπάρχοντος εμπορικού κυψελοειδούς δικτύου μιας πόλης για εκπαίδευση και επικύρωση. Για την αξιολόγησης της προτεινόμενης προσέγγισης, εκτελέστηκαν προσομοιώσεις ανίχνευσης ακτίνων σε διακριτή συχνότητα 28 GHz και χρησιμοποιήθηκαν για εκπαίδευση και επικύρωση. Τα αποτελέσματα καταδεικνύουν ότι η προτεινόμενη μέθοδος είναι ακριβής στην απόκτηση καναλιών για διάφορα σενάρια πτήσης UAV και υπερέχει των συμβατικών DNN.

4.5.1 Αρχικές επισημάνσεις

Τα μη επανδρωμένα εναέρια οχήματα χαμηλού υψομέτρου (UAV), που ονομάζονται επίσης drones, έχουν ενεργοποιήσει πολλές προσωπικές και εμπορικές εφαρμογές, όπως αεροφωτογράφηση και περιήγηση στα αξιοθέατα, παράδοση δεμάτων, έκτακτη διάσωση σε φυσικές καταστροφές, παρακολούθηση και επιτήρηση και γεωργία ακριβείας [30]. Πρόσφατα, το ενδιαφέρον για αυτήν την αναδυόμενη τεχνολογία αυξάνεται σταθερά καθώς πολλές κυβερνήσεις έχουν ήδη διευκολύνει τους κανονισμούς για τη χρήση UAV. Ως αποτέλεσμα, οι τεχνολογίες UAV αναπτύσσονται και αναπτύσσονται με πολύ γρήγορο ρυθμό σε όλο τον κόσμο για να προσφέρουν γόνιμες επιχειρηματικές ευκαιρίες και νέες κάθετες αγορές. Συγκεκριμένα, τα UAV μπορούν να χρησιμοποιηθούν ως πλατφόρμες εναέριας επικοινωνίας για τη βελτίωση της ασύρματης συνδεσιμότητας για χρήστες εδάφους και συσκευές Internet of Things (IoT) σε σκληρά περιβάλλοντα όταν τα επίγεια δίκτυα δεν είναι προσβάσιμα. Επιπλέον, οι ευφυείς πλατφόρμες UAV μπορούν να προσφέρουν σημαντική και ποικίλη συμβολή στην εξέλιξη των έξυπνων πόλεων, προσφέροντας οικονομικά αποδοτικές υπηρεσίες που κυμαίνονται από την παρακολούθηση του περιβάλλοντος έως τη διαχείριση της κυκλοφορίας.

Η ασύρματη επικοινωνία είναι μια βασική τεχνολογία ενεργοποίησης για τα UAV και η ενσωμάτωσή τους έχει τραβήξει την προσοχή τα τελευταία χρόνια. Προς αυτή την κατεύθυνση, η 3GPP έχει δραστηριοποιηθεί στον εντοπισμό των απαιτήσεων, τεχνολογιών και πρωτοκόλλων για εναέριες επικοινωνίες για να ενεργοποιηθούν τα δικτυωμένα UAV στα τρέχοντα δίκτυα μακροπρόθεσμης εξέλιξης (LTE) και 5G/B5G. Οι επικοινωνίες των UAV διαφέρουν θεμελιωδώς από τις επίγειες επικοινωνίες στο υποκείμενο κανάλι διάδοσης αέρος-εδάφους και στους εγγενείς περιορισμούς μεγέθους, βάρους και ισχύος. Τα τρισδιάστατα κινητά UAV απολαμβάνουν μεγαλύτερη

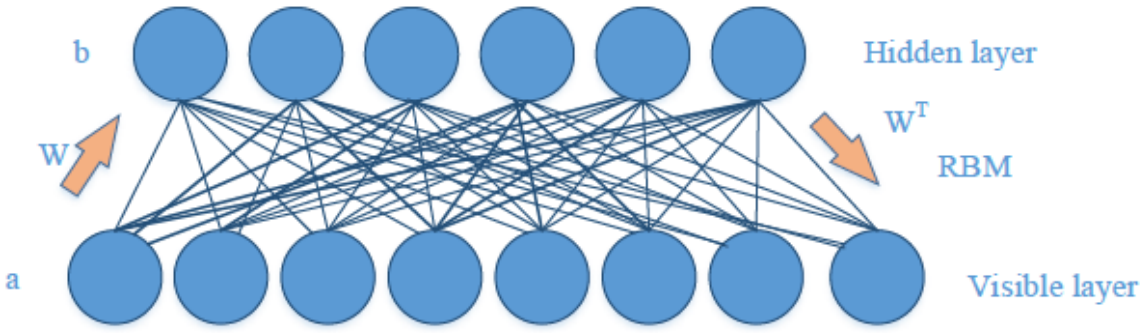
πιθανότητα επικοινωνίας οπτικής επαφής (LoS) από τους χρήστες εδάφους, κάτι που μπορεί να είναι επωφελές για την αξιοπιστία και την απόδοση ισχύος των επικοινωνιών UAV. Ωστόσο, αυτό σημαίνει επίσης ότι οι επικοινωνίες UAV μπορούν εύκολα να προκαλέσουν/υποφέρουν παρεμβολές προς/από επίγεια δίκτυα, τα οποία πρέπει να αντιμετωπιστούν προσεκτικά. Η διαθεσιμότητα του LoS εξαρτάται όχι μόνο από το περιβάλλον διάδοσης αλλά και από το ύψος, τη γωνία ανύψωσης και τις τροχιές κίνησης των UAV, οι οποίες πρέπει να αξιολογούνται από κοινού για κάθε σενάριο.

Η υλοποίηση ολοκληρωμένων UAV στο τρισδιάστατο κινητό κυψελοειδές δίκτυο εξαρτάται σε μεγάλο βαθμό από την αξιοπιστία και την αποτελεσματικότητα των καναλιών επικοινωνίας σε διάφορα λειτουργικά περιβάλλοντα και σενάρια UAV. Επιπλέον, αυτά τα κανάλια είναι ζωτικής σημασίας για το σχεδιασμό και την αξιολόγηση συνδέσεων επικοινωνίας UAV για μεταδόσεις δεδομένων ελέγχου/μη ωφέλιμου φορτίου και ωφέλιμου φορτίου σε νέα σενάρια λειτουργίας UAV, κάτι που είναι μια από τις σημαντικές προκλήσεις σε αυτό το περιβάλλον. Επιπλέον, η κινητικότητα των UAV και η χρονικά μεταβαλλόμενη τοπολογία του τρισδιάστατου δικτύου κινητής τηλεφωνίας, μαζί με τα σφάλματα εντοπισμού και την καθυστέρηση, μπορεί να περιπλέξουν την απόκτηση έγκαιρης και ακριβούς γνώσης των πληροφοριών κατάστασης καναλιού (Channel State Information – CSI). Ως εκ τούτου, η απόκτηση ακριβούς μοντελοποίησης καναλιών είναι υψίστης σημασίας για το σχεδιασμό ισχυρών και αποτελεσματικών αλγορίθμων σχηματισμού δέσμης και παρακολούθησης δέσμης, μεθόδων κατανομής πόρων, προσεγγίσεων προσαρμογής ζεύξης και τεχνικών πολλαπλών κεραιών. Ενώ γενικότερα έχουν προταθεί αρκετά στατιστικά μοντέλα καναλιών αέρος-εδάφους που αντισταθμίζουν την ακρίβεια και τη μαθηματική ελκτική ικανότητα, χρειάζεται ακόμα πιο πρακτική ανάλυση για να γεφυρωθεί αυτό το χάσμα γνώσης.

Σε μια παράλληλη λεωφόρο, η κοινότητα ασύρματης επικοινωνίας έχει δώσει σημαντική προσοχή στις τεχνικές βαθιάς μάθησης (Deep Learning - DL) λόγω της επιτυχίας τους σε διάφορες εφαρμογές, π.χ. όραση υπολογιστή, επεξεργασία φυσικής γλώσσας και αυτόματη αναγνώριση ομιλίας. Το DL είναι μια προσέγγιση μηχανικής μάθησης που βασίζεται σε νευρώνες που μπορεί να κατασκευάσει βαθιά νευρωνικά δίκτυα (DNN) με ευέλικτες δομές με βάση τις απαιτήσεις της εφαρμογής. Συγκεκριμένα, αρκετές εργασίες στην ανοιχτή βιβλιογραφία έχουν χρησιμοποιήσει μεθόδους DL για τη μοντελοποίηση καναλιών και την απόκτηση CSI. Για παράδειγμα, ένας αλγόριθμος μοντελοποίησης καναλιών που βασίζεται σε DL αναπτύχθηκε χρησιμοποιώντας ένα αποκλειστικό νευρωνικό δίκτυο που βασίζεται σε δίκτυα παραγωγής αντιπάλων που έχουν σχεδιαστεί για να μαθαίνουν τις πιθανότητες μετάβασης καναλιού από παρατηρήσεις δέκτη. Επιπλέον, σε άλλη μελέτη προτείνεται ένα σχήμα εκτίμησης καναλιών που βασίζεται σε DNN για να σχεδιάσει από κοινού τα πιλοτικά σήματα και τον εκτιμητή καναλιών για συστήματα ευρείας ζώνης μαζικής πολλαπλής εισόδου πολλαπλής εξόδου (MIMO).

Για την αντιμετώπιση των περιορισμών των υπαρχόντων συστημάτων, οι ερευνητές έχουν επικεντρωθεί στις μεθόδους (Machine Learning) ML και (Deep Learning) DL τα τελευταία χρόνια. Παρά τις τεράστιες δυνατότητές τους, οι προσεγγίσεις DL αντιμετωπίζουν σημαντικές δυσκολίες που περιορίζουν την εφαρμογή τους σε προηγμένα περιβάλλοντα επικοινωνίας. Για να αναπτυχθεί μια επαρκής

χαρτογράφηση από τα χαρακτηριστικά στα επιθυμητά αποτελέσματα, τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks – DNN) απαιτούν τεράστια σύνολα δεδομένων, τα οποία αυξάνουν τόσο την υπολογιστική πολυπλοκότητα όσο και τη δυσκολία της εκπαιδευτικής διαδικασίας. Λόγω της τρέχουσας κατάστασης των συνόλων δεδομένων επικοινωνίας, είναι αμφίβολο να δίνεται βάση αποκλειστικά σε DNN που βασίζονται σε δεδομένα ως μαύρα κουτιά και να αφήνονται όλες οι προβλέψεις στα βάρη των μοντέλων. Τα DNN που βασίζονται σε μοντέλα έχουν εξελιχθεί ως λύση στους περιορισμούς των τεχνικών που βασίζονται σε μοντέλα και στα δεδομένα. αυτά τα DNN συνδυάζουν τα πλεονεκτήματα των DNN στη μάθηση και τη χαρτογράφηση με την τεχνογνωσία στον τομέα για να μεγιστοποιήσουν τα οφέλη. Οι περισσότερες λύσεις DNN που βασίζονται σε μοντέλα εμπίπτουν σε μία από τις δύο κατηγορίες: (deep unfolded networks) βαθιά ξεδιπλωμένα δίκτυα, στα οποία τα επίπεδα DNN αναπαράγουν γύρους μιας υπάρχουσας επαναληπτικής διαδικασίας ή υβριδικά δίκτυα, στα οποία τα DNN βοηθούν τα συμβατικά μοντέλα και ενισχύουν την απόδοση. Το πρόβλημα της σπανιότητας δεδομένων έχει επίσης εξαλειφθεί παράλληλα με την πρόοδο της βιβλιογραφίας DL για τις τεχνολογίες επικοινωνίας, αφήνοντας χώρο για συστήματα DL που βασίζονται σε δεδομένα. Παρόλα αυτά, οι πολλά υποσχόμενες βελτιώσεις των τεχνικών DL για τη μοντελοποίηση καναλιών έχουν παρακινήσει τους ερευνητές να διερευνήσουν τη χρήση διαφορετικών μεθόδων εκμάθησης και προσεγγίσεων εξαγωγής χαρακτηριστικών στο πλαίσιο συστημάτων επικοινωνίας UAV με κυψέλη. Από όσο είναι γνωστό, μόνο μερικές προηγούμενες σχετικές ερευνητικές εργασίες ασχολήθηκαν με χαρακτηρισμούς καναλιών UAV χρησιμοποιώντας την ισχύ του λαμβανόμενου σήματος (RSS) στις κυψελωτές επικοινωνίες. Σε άλλη μελέτη προτείνεται ένα πλαίσιο μοντελοποίησης για τη διάδοση κυμάτων στις κινητές επικοινωνίες συνδυάζοντας πολλούς μαθητές σε μια μέθοδο εκμάθησης συνόλου για μοντελοποίηση RSS. Επιπλέον, ένα μοντέλο βασισμένο στην εκμάθηση βαθιάς ενίσχυσης (Deep Reinforcement Learning – DRL) για την εκχώρηση καναλιών και ισχύος αναπτύχθηκε για συστήματα IoT με δυνατότητα UAV, όπου ένας μόνος σταθμός βάσης UAV αναπτύσσεται για τη συλλογή δεδομένων από πολλούς κόμβους IoT. Επίσης, στη βιβλιογραφία μελετήθηκε ότι ένας αλγόριθμος εκμάθησης χρησιμοποιείται για την πρόβλεψη χαρακτηριστικών καναλιών μεταξύ UAV και χρηστών εδάφους, ο οποίος παρέχει ακριβείς πληροφορίες περιβαλλοντικής κατάστασης για αποφάσεις ανάπτυξης UAV. Επιπλέον, άλλοι ερευνητές έχουν προτείνει μεθόδους συνόλου που βασίζονται σε εποπτευόμενη εκμάθηση βάσης για την πρόβλεψη του μοντέλου καναλιού του UAV χρησιμοποιώντας τη μέθοδο ενίσχυσης ελαχίστων τετραγώνων, την πρόβλεψη σακουλών και τις μηχανές υποστήριξης διανυσμάτων (SVMs).



Εικόνα 13: Το δίκτυο ενός RBM διαγράμματος

Όσον αφορά τις μεθόδους DL, ερευνητές έχουν προτείνει ένα σχήμα DL που μπορεί να εξερευνήσει πλήρως τα χαρακτηριστικά των δεδομένων ασύρματου καναλιού και να λάβει τα βέλτιστα βάρη ως δακτυλικά αποτυπώματα, ενώ ενσωματώνει έναν άπληστο αλγόριθμο μάθησης για τη μείωση της υπολογιστικής πολυπλοκότητας. Πέρα από αυτό, η περιορισμένη μηχανή Boltzmann (Restricted Boltzmann Machine – RBM) είναι ένα παραγωγικό στοχαστικό τεχνητό νευρωνικό δίκτυο που μπορεί να μάθει μια κατανομή πιθανοτήτων σε ένα σύνολο εισόδων με τρόπο χωρίς επίβλεψη. Το RBM είναι διμερές, δηλ. δεν υπάρχουν συνδέσεις εντός του στρώματος και αποτελείται από ένα ζεύγος στρωμάτων που συνήθως αναφέρονται ως ορατές και κρυφές μονάδες, αντίστοιχα, όπως φαίνεται στην Εικόνα 13, και μπορεί να έχουν μια συμμετρική σύνδεση μεταξύ τους. Ωστόσο, η RBM έχει ορισμένους περιορισμούς λόγω του ότι ασχολείται μόνο με δυαδικά δεδομένα και για να παρακάμψει αυτό το ζήτημα, προτάθηκε η περιορισμένη μηχανή **G**aussian - **Ber**noulli **B**oltz**m**ann (GBRBM) για την επεξεργασία πραγματικών δεδομένων όπου ένας ορατός κόμβος Gauss αντικατέστησε έναν δυαδικό κόμβο για να αρχικοποιήσει DNN για εξαγωγή χαρακτηριστικών και μείωση διαστάσεων.

Το κοινό χαρακτηριστικό των προγενέστερων μελετών είναι ότι όλες ακολουθούν μεθόδους εκπαίδευσης που έχουν υψηλή διακύμανση και αργές αποκλίνουσες συμπεριφορές που απαιτούν τεράστιο όγκο δεδομένων εκπαίδευσης και εκτεταμένο χρόνο εκπαίδευσης, κάτι που είναι απαγορευτικό στα συστήματα ασύρματης επικοινωνίας. Αυτή η παρατήρηση παρακίνησε την εργασία στο [30] να καλύψει αυτό το κενό στη βιβλιογραφία και να προτείνει ένα πλαίσιο βασισμένο στο GBRBM που ενσωματώνει ένα βαθύ νευρωνικό δίκτυο βασισμένο σε αυτοκωδικοποιητή για την εκτίμηση της λαμβανόμενης ισχύος σήματος των UAV που πετούν σε μια σειρά υψών που συνδέονται σε ένα κυψελοειδές δίκτυο. Επιπλέον, προτείνεται ένας νέος αλγόριθμος που χρησιμοποιεί έναν προσαρμοστικό ρυθμό μάθησης παράλληλα με μια βελτιωμένη κλίση, η οποία επιταχύνει την εκμάθηση των κρυφών νευρώνων, σε αντίθεση με την παραδοσιακή αξιοπρεπή κλίση.

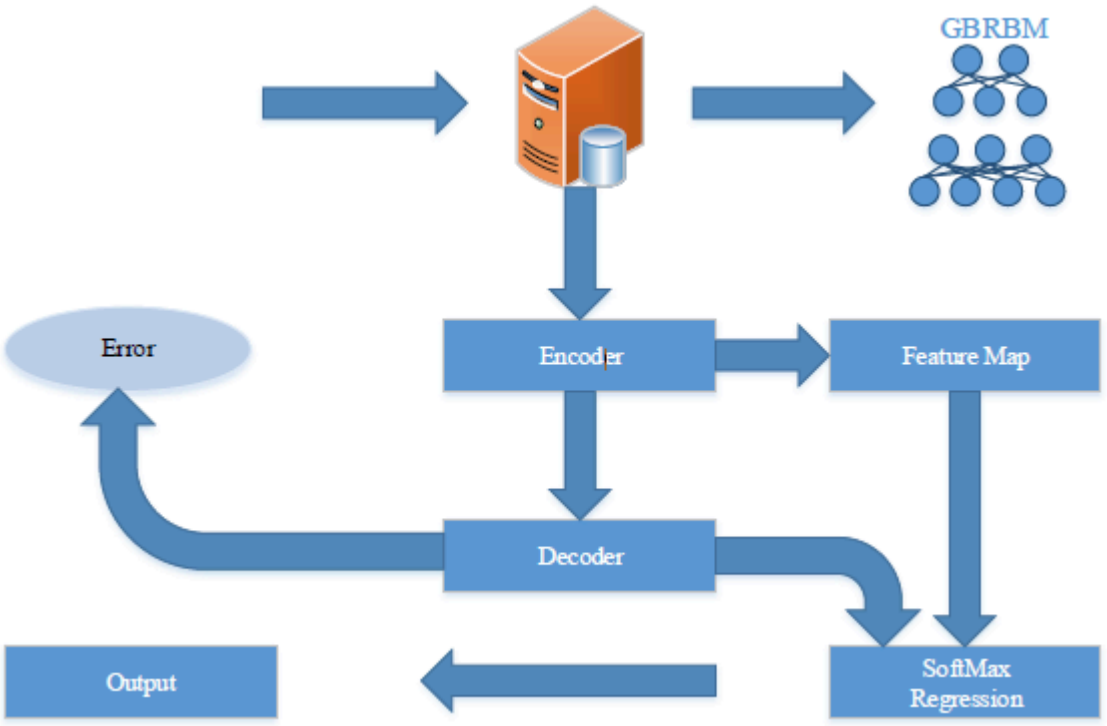
4.5.2 Μοντέλο συστήματος και Διατύπωση Προβλήματος

Οι τεχνολογίες επικοινωνιών και μεταφορών είναι σημαντικά πλεονεκτήματα για τον τρόπο ζωής των ανθρώπων. Στην κίνηση προς την ενσωμάτωση αυτών των εξελίξεων σε ένα πεδίο, ένας πομποδέκτης μπορεί να συνδεθεί σε ένα drone για να βοηθήσει μελλοντικά ασύρματα δίκτυα. Είναι σημαντικό ότι σε φυσικές καταστροφές, ειδικά όταν η υποδομή επικοινωνιών (π.χ. σταθμοί βάσης) έχει υποστεί ζημιά, το δίκτυο επικοινωνίας πρέπει να διατηρείται. Επί του παρόντος, οι υπάρχουσες μέθοδοι έχουν

περιορισμούς όσον αφορά την ευελιξία και τη διαθεσιμότητα πόρων. Η αντιμετώπιση τέτοιων προβλημάτων απαιτεί τη χρήση μη επανδρωμένων εναέριων οχημάτων (UAV) για ασύρματη εργασία λόγω των πολυλειτουργικών ιδιοτήτων και της ευελιξίας τους. Ένα drone πετά σε χαμηλό ύψος και είναι εξοπλισμένο με πομποδέκτες για να λειτουργεί με ασύρματο δίκτυο ως κυψέλη drone στην περιοχή κάλυψης. Πολλές παράμετροι καθορίζουν τη μοντελοποίηση καναλιών μεταξύ του πομποδέκτη drone και των BS, όπως το υψόμετρο, η κατευθυντικότητα της κεραίας, η τοποθεσία, η ισχύς μετάδοσης και τα χαρακτηριστικά του περιβάλλοντος. Για να γίνει διερεύνηση των επιδράσεων αυτών των παραμέτρων στη μοντελοποίηση καναλιών μεταξύ του πομποδέκτη drone και των BS, προτείνεται ένα πλαίσιο GBRBM ενσωματωμένο με ένα βαθύ νευρωνικό δίκτυο που βασίζεται σε αυτόματο κωδικοποιητή.

Διατύπωση του προβλήματος

Το GBRBM είναι γνωστό ως τυχαία πεδία Markov, (Markov random fields – MRFs) και είναι ένα μη κατευθυνόμενο πιθανοτικό γραφικό μοντέλο. Το επίπεδο εισόδου αντιπροσωπεύει τα παρατηρούμενα δεδομένα που αποτελούνται από εννέα κόμβους, καθένας από τους οποίους αντιπροσωπεύει ένα συγκεκριμένο είδος δεδομένων εισόδου, N = (N1 - N9), όπου N1 αντιπροσωπεύει το γεωγραφικό πλάτος, N2 αντιπροσωπεύει το γεωγραφικό μήκος, N3 είναι το υψόμετρο του εδάφους του drone, N4 και το N5 αντιπροσωπεύουν το γεωγραφικό πλάτος και το γεωγραφικό μήκος κυψέλης, αντίστοιχα, το n6 αντιπροσωπεύει την ανύψωση της κυψέλης, το N7 είναι το κτίριο της κυψέλης, το N8 είναι το ύψος του ιστού της κεραίας και το N9 αντιπροσωπεύει το υψόμετρο του drone. Η Εικόνα 14 απεικονίζει το διάγραμμα DNN, που είναι ενισχυμένο με το GBRBM. Αυτή η αρχιτεκτονική έχει σχεδιαστεί για να λειτουργεί συστηματικά με βάση την αρχή του προτεινόμενου αλγορίθμου. Πρώτον, τα δεδομένα συλλέγονται κατά το στάδιο της συλλογής δεδομένων σε πραγματικό χρόνο. Στη συνέχεια, τα ακατέργαστα δεδομένα θα σταλούν στο στάδιο της προεκπαίδευσης. Στη συνέχεια, τα δεδομένα αξιοποιούνται για εξαγωγή χαρακτηριστικών με χρήση βελτιωμένης GBRBM πολλαπλών μπλοκ. Στη συνέχεια, στέλνονται τα προ-εκπαιδευμένα δεδομένα στον λεπτό συντονισμό. Τέλος, τα δεδομένα εξόδου των εξαγόμενων χαρακτηριστικών ταξινομούνται χρησιμοποιώντας τη μονάδα softmax regression για να ληφθεί η έξοδος.



Εικόνα 14: Αρχιτεκτονική πλαισίου του βελτιωμένου DNN που βασίζεται σε GBRBM.

Το RBM έχει χρησιμοποιηθεί ως καθολική προσέγγιση. Ωστόσο, ο αριθμός των κρυφών κόμβων είναι πάντα περιορισμένος, γεγονός που οδηγεί στην αδυναμία μοντελοποίησης κάποιας διακριτής κατανομής πιθανοτήτων. Στην πράξη, έχει αποδειχθεί ότι δεν είναι όλοι οι κρυφοί κόμβοι ενεργοί και ορισμένοι είναι άσκοποι, συνεπώς το βάρος τους στις εξισώσεις απόφασης δεν μπορεί να ενσωματωθεί. Γι' αυτό το λόγο, εισάγεται ο προσαρμοστικό ρυθμός μάθησης (adaptive learning rate) και μια νέα βελτιωμένη κλίση (enhanced gradient) στο επόμενο μέρος.

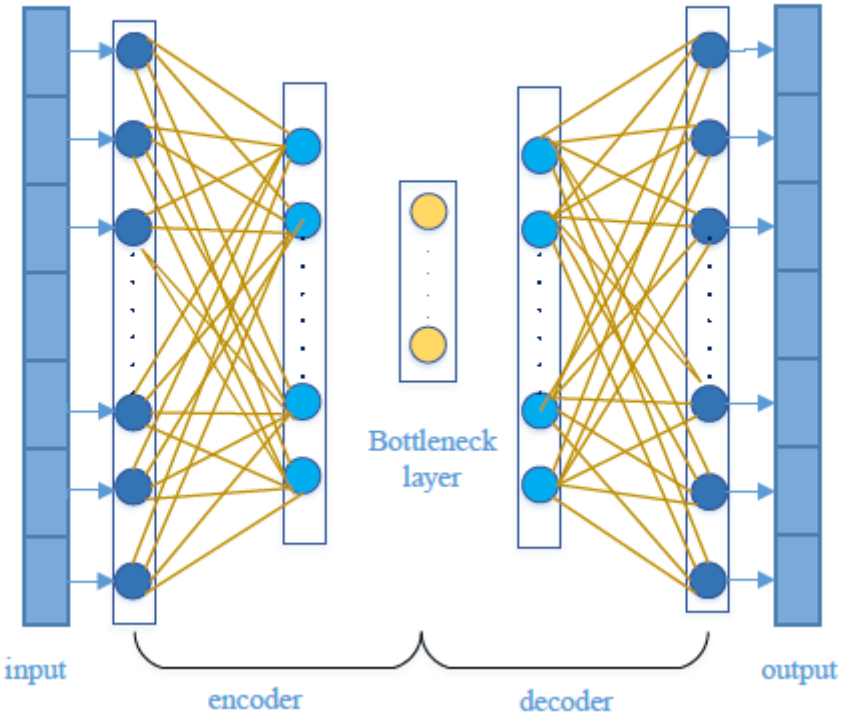
Προσαρμοστικός ρυθμός μάθησης

Με βάση τη μεγιστοποίηση της τοπικής εκτίμησης της πιθανότητας, ο ρυθμός εκμάθησης μπορεί να προσαρμοστεί αυτόματα ενώ το RBM εκπαιδεύεται χρησιμοποιώντας τη στοχαστική κλίση. Έτσι, οι παράμετροι ρυθμού εκμάθησης επιλέγονται για να μεγιστοποιηθεί η πιθανότητα των παραμέτρων GBRPM. Μπορεί λοιπόν να βρεθεί ο βέλτιστος ρυθμός μάθησης που μπορεί να μεγιστοποιήσει την πιθανότητα κάθε επανάληψης.

Βελτιωμένη κλίση

Η πρόσφατα βελτιωμένη κλίση προτάθηκε για την ενημέρωση του αμετάβλητου κανόνα των μηχανών Boltzmann για την αναπαράσταση δεδομένων. Ένας μετασχηματισμός bit-flipping εισήγαγε τη διαβάθμιση και, στη συνέχεια, ο κανόνας ενημερώνεται για να βελτιώσει τα αποτελέσματα, την εκμάθηση του RBM και να το κάνει λιγότερο ευαίσθητο στην παραλλαγή και την προετοιμασία των παραμέτρων. Η ίδια ιδέα χρησιμοποιήθηκε για την ενίσχυση της κλίσης του GDBM με τη χρήση ορατών νευρώνων Gauss ως εναλλακτική λύση στους μετασχηματισμούς που αναστρέφονται bit μετατοπίζοντας την ορατή μονάδα. Προτείνεται δηλαδή μια νέα μέθοδος για την ενίσχυση της κλίσης, στην οποία κάθε μπλοκ συνδέεται με το ανώτερο μπλοκ μέσω

ενός κρυμμένου επιπέδου, για να επιταχύνεται έτσι η προ-εκπαίδευση των GBRBM blocks.



Εικόνα 15: Δομή κωδικοποιητή και αποκωδικοποιητή σε ένα βαθύ νερωνικό δίκτυο.

Η αρχιτεκτονική νευρωνικού δικτύου βαθιάς αυτοκωδικοποιητή απεικονίζεται στην Εικόνα 15. Ο βαθύς αυτόματος κωδικοποιητής, όπως φαίνεται και στην παραπάνω εικόνα, αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή. Ο αριθμός των νευρώνων για κάθε επίπεδο του δικτύου πρέπει να καθοριστεί αφού επιλεγεί ο αριθμός των επιπέδων. Δυστυχώς, η διαδικασία βελτιστοποίησης είναι αρκετά χρονοβόρα επειδή δεν υπάρχει καθορισμένο εύρος για τον αριθμό των νευρώνων σε κάθε επίπεδο. Θα πρέπει να σημειωθεί ότι το στρώμα συμφόρησης (bottleneck layer) περιλαμβάνει τις πιο θεμελιώδεις ιδιότητες εισόδου του νευρωνικού δικτύου και ότι η απόδοση του νευρωνικού δικτύου εξαρτάται από αυτά τα χαρακτηριστικά.

4.5.3 Μετρήσεις με UAV

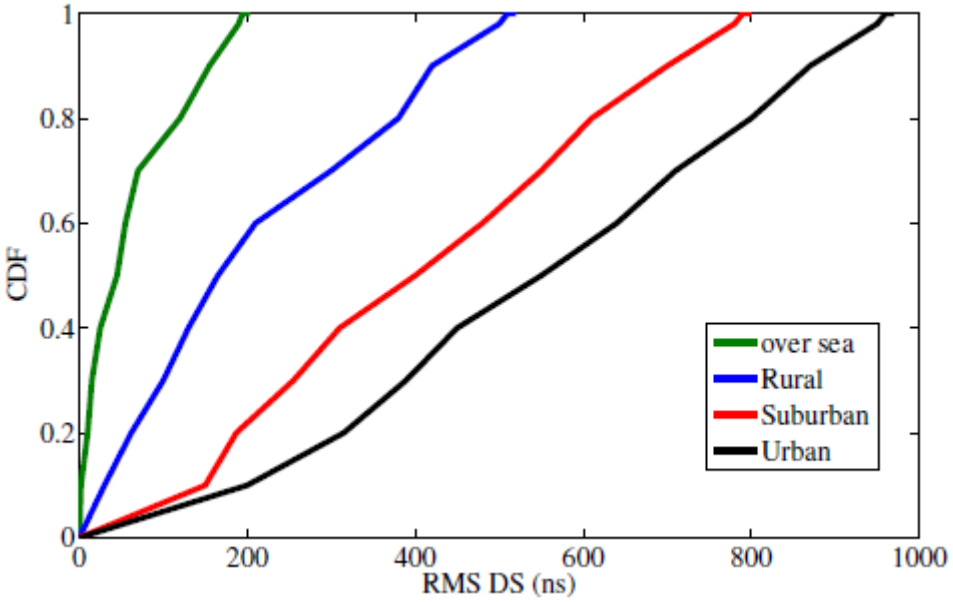
Σε αυτή την ενότητα, παρέχονται αποτελέσματα προσομοίωσης για την αξιολόγηση της απόδοσης του προτεινόμενου σχήματος πρόβλεψης CSI για συνδέσεις αέρα-εδάφους σε συστήματα επικοινωνίας UAV. Για τις επικοινωνίες UAV, οι ζώνες mmWave υπόσχονται να ανταποκριθούν στις απαιτήσεις ρυθμού δεδομένων των εφαρμογών για κινητές συσκευές υψηλής απόδοσης. Ιδιαίτερα, πετώντας σε μια προτιμώμενη θέση, τα UAV μπορούν να διατηρήσουν τη σύνδεση οπτικής επαφής (LOS) (ή τουλάχιστον μια αποδεκτή σύνδεση NLOS) με έναν επιθυμητό χρήστη. Η ανίχνευση ακτίνων προσφέρει μια ντετερμινιστική μέθοδο χαρακτηρισμού του καναλιού mmWave υπό διάφορες συνθήκες.

Δεδομένου ότι η μελέτη της συμπεριφοράς των καναλιών Air-to-Ground στις ζώνες mmWave με χρήση UAV μπορεί να είναι δύσκολη, η ανίχνευση ακτίνων παρέχει μια χρήσιμη εναλλακτική λύση. Εδώ, χρησιμοποιείται το πρόγραμμα ανίχνευσης ακτίνων RemcomWireless InSite για την προσομοίωση των κινήσεων του UAV σε πραγματικό χρόνο κατά μήκος μιας καθορισμένης τροχιάς. Οι παράμετροι προσομοίωσης ορίστηκαν ως εξής: οι διαστάσεις της περιοχής είναι 10 km επί 10 km σε μέγεθος όπως δείχνει ο Πίνακας 1. Σε όλες τις περιπτώσεις, τα UAV πετούν με ταχύτητα 15 m ανά δευτερόλεπτο από υψόμετρα 200 m (μοιάζουν με χερσαία οχήματα)· η τροχιά πτήσης του UAV είναι περίπου 2 km σε μήκος. Τόσο ο πομπός όσο και ο δέκτης χρησιμοποιούν κατακόρυφα πολωμένες διπολικές κεραίες μισού κύματος. Το κανάλι στα 28 GHz έχει ένα ημιτονοειδές για να ηχεί το κανάλι στο κέντρο της συχνότητας. Το επίπεδο ισχύος έχει ρυθμιστεί στα 30 dBm για μετάδοση.

Πίνακας 1: Παράμετροι προσομοίωσης

Σενάριο	**΄Υψη κτηρίων (m)**	**Πλήθος κτηρίων**
Αστική (Urban)	100-200	100
Προαστιακά (Suburban)	20-30	25
Αγροτική (Rural)	5-10	10
Θαλάσσιες (overseas)	-	-

Λαμβάνοντας υπόψη τις συχνότητες mmWave των 28 GHz, παρουσιάζονται οι αθροιστικές συναρτήσεις κατανομής (CDF) του RMS-DS του καναλιού πολλαπλών διαδρομών μεταξύ του πομπού GS και του UAV στην Εικόνα 16 για τέσσερα διαφορετικά σενάρια. Ο κύριος λόγος για αυτή τη συμπεριφορά είναι ότι σε μεγαλύτερα υψόμετρα UAV, το UAV κινείται πάνω από ψηλές κατασκευές και είναι σε θέση να παρατηρήσει σήματα που είναι διάσπαρτα από την πλειοψηφία των γύρω κτιρίων. Αντίθετα, τα ευρήματα σε αγροτικές και προαστιακές περιοχές καταδεικνύουν ότι σε αντίθεση με τα αστικά περιβάλλοντα, το RMS-DS στην πραγματικότητα μειώνεται καθώς αυξάνεται ο αριθμός των κτιρίων. Τα κτίρια σε αγροτικές και προαστιακές περιοχές τείνουν να είναι μικρότερα και λιγότερο πυκνοκατοικημένα από εκείνα των πόλεων. Αυτές οι διάφορες συμπεριφορές καναλιών πολλαπλών διαδρομών υποδηλώνουν ότι οι περιβαλλοντικές μεταβλητές και το ύψος του UAV μπορεί να έχουν μεγάλο αντίκτυπο στη συμπεριφορά του καναλιού και κατά συνέπεια στη σχεδίαση του δέκτη.



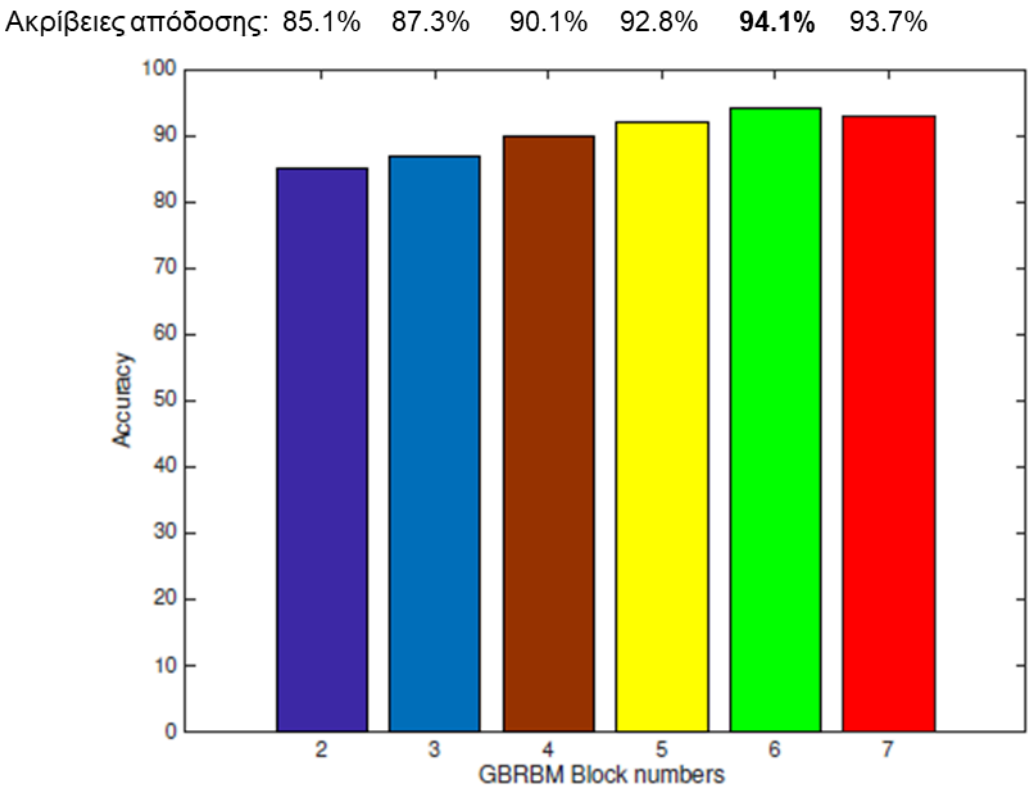
Εικόνα 16: Αθροιστικές συναρτήσεις κατανομής (CDF) του καναλιού RMS-DS στις 4 περιπτώσεις προσομοίωσης στα 28 GHz.

4.5.4 Αξιολόγηση της απόδοσης

Σε αυτήν την ενότητα, υλοποιείται το προτεινόμενο DNN που βασίζεται σε GBRBM, εκτός από άλλους αλγόριθμους μηχανικής εκμάθησης, π.χ., backpropagation ANN και SVM. Οι εξεταζόμενες παράμετροι προσομοίωσης περιλαμβάνουν το μπλοκ GBRBM, τον αριθμό των νευρώνων και των επιπέδων του δικτύου και την τιμή του προσαρμοστικού ρυθμού μάθησης και του αριθμού της εποχής. Ο συνολικός αριθμός των διανυσμάτων εκπαίδευσης είναι 710, ενώ ο αριθμός των διανυσμάτων δοκιμής είναι 177, με 201 άγνωστες μεταβλητές με 100 επαναλήψεις.

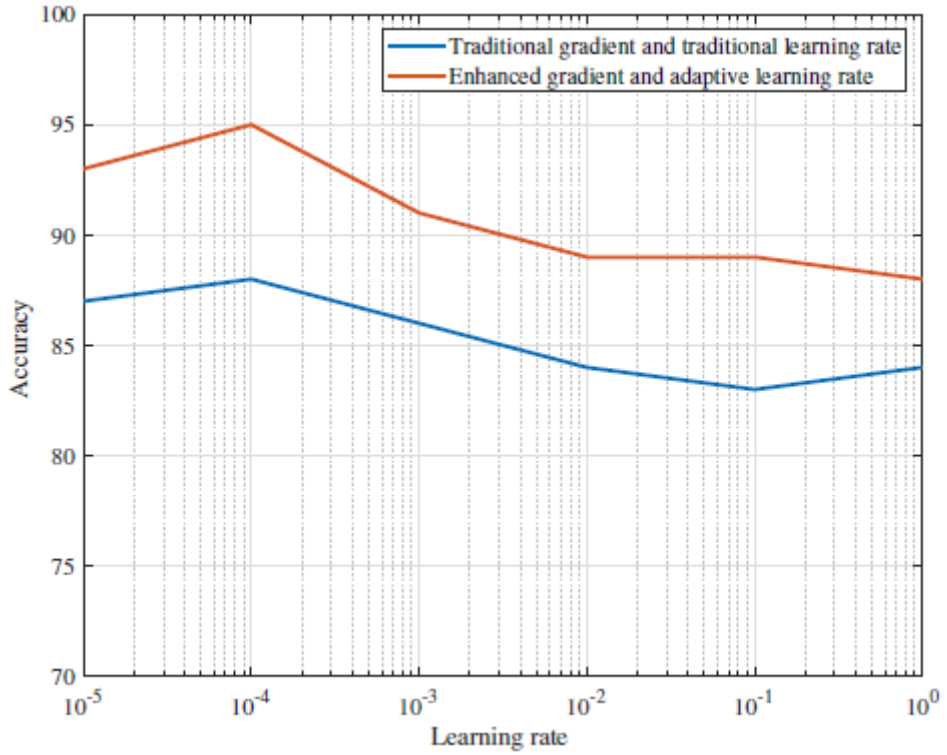
GBRBM blocks

Η Εικόνα 17 δείχνει το σφάλμα διαφοράς μεταξύ των μετρήσεων και των εκτιμώμενων τιμών χρησιμοποιώντας διαφορετικούς αριθμούς DNN που βασίζονται σε GBRBM. Οι διαφορετικοί αριθμοί GBRBM ξεκινάνε από το 2 έως το 7, ενώ το σφάλμα διαφοράς μεταξύ του εκτιμώμενου και του μετρημένου σήματος (Received Signal Strength – RSS) ποικίλλει από 5 έως 15. Μπορεί εύκολα να φανεί ότι τα καλύτερα αποτελέσματα επιτυγχάνονται όταν έξι μπλοκ DNN χρησιμοποιούνται στο προεκπαιδευτικό στάδιο. Οι ακρίβειες απόδοσης είναι 85,1%, 87,3%, 90,1%, 92,8%, 94,1% και 93,7%, αντίστοιχα.

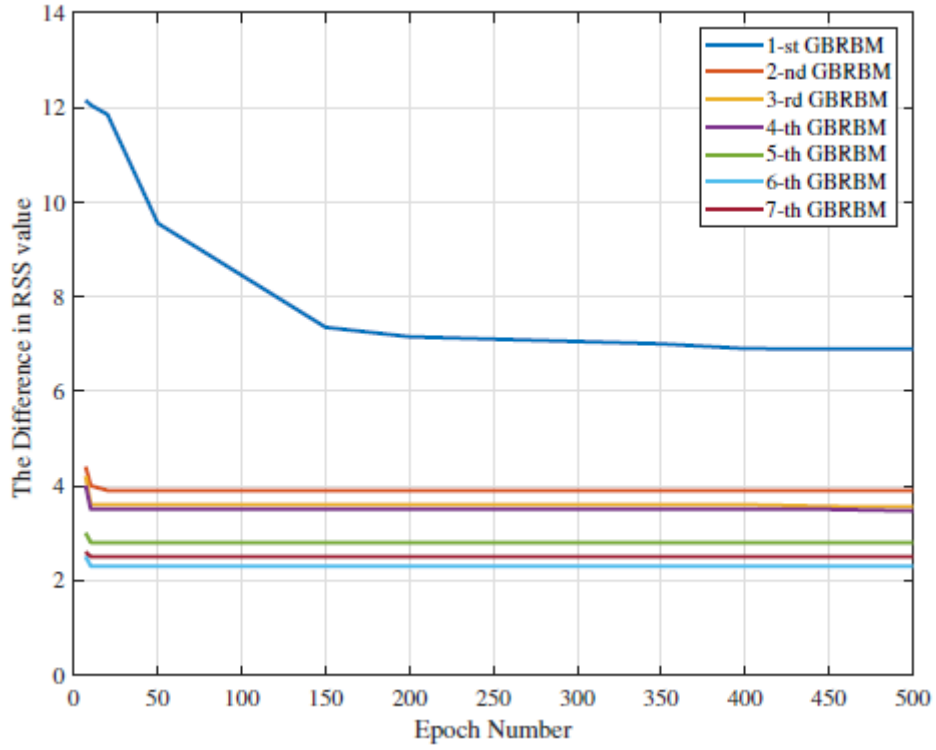


Εικόνα 17: Ακρίβειες απόδοσης του DNN σε συνάρτηση με το πλήθος των GBRBM block.

Δεδομένου ότι ένας σωστός αριθμός εποχής μπορεί να συντομεύσει την περίοδο εκπαίδευσης και τον χώρο αναζήτησης λύσεων, ο καλύτερος αριθμός εποχής είναι σημαντικός τόσο για την προ-προπονητική όσο και για την προπονητική φάση. Επομένως, ο τέλειος αριθμός εποχής θα μπορούσε να επιλεγεί με βάση την τιμή αυτής της συνάρτησης μείωσης. Για παράδειγμα, η Εικόνα 19 παρουσιάζει το σφάλμα διαφοράς σε σχέση με τον αριθμό των εποχών. Σαφώς, το σφάλμα διαφοράς είναι σταθερό όταν η προσομοίωση έφτασε γύρω στην 500η εποχή. Ως εκ τούτου, ο αριθμός εποχής του προ-προπονητικού σταδίου ορίζεται σε 250 και ο αριθμός εποχής του σταδίου εκπαίδευσης ορίζεται σε 500 εποχές.



Εικόνα 18: Ακρίβεια απόδοσης σε συνάρτηση με το ρυθμό εκμάθησης.



Εικόνα 19: Η μέση διαφορά μεταξύ του εκτιμώμενου και του αληθινού RSS στη φάση της προεκπαίδευσης.

Αφού καθοριστούν τα μπλοκ GBRBM και ο αριθμός εποχής, πρέπει να οριστεί ο αριθμός νευρώνων για κάθε επίπεδο. Η διαδικασία βελτιστοποίησης είναι αρκετά κουραστική επειδή η ποικιλία συντονισμού του αριθμού νευρώνων σε κάθε επίπεδο

είναι αυθαίρετη. Έτσι, ορίζεται εμπειρικά η ποσότητα νευρώνων και στη συνέχεια εκτελούνται πειράματα για να μεγιστοποιηθεί η ποσότητα νευρώνων του έκτου στρώματος ανάλογα με τη λειτουργία και των δύο DNN που βασίζονται σε GBRBM. Στο στάδιο της προεκπαίδευσης, η διαφορά στην τιμή RSS μεταξύ του εκτιμώμενου μέτρου και του πραγματικού μέτρου μειώνεται όσο αυξάνεται ο αριθμός της εποχής και ορίζεται στην 250η εποχή, όπως απεικονίζεται στην Εικόνα 18.

Προσαρμοστικός ρυθμός εκπαίδευσης

Η ακρίβεια στα αποτελέσματα θα επηρεαστεί πολύ αν ο ρυθμός εκμάθησης λάβει πολύ μεγάλες ή πολύ μικρές τιμές. Συγκεκριμένα, εάν το ποσοστό μάθησης είναι πολύ μικρό, η περίοδος εκπαίδευσης μεγαλώνει και είναι πιθανό να παγιδευτεί μια τοπική βέλτιστη λύση. Εφόσον υπάρχουν στάδια προεκπαίδευσης και εκπαίδευσης, πρέπει να βρεθεί ένα ατομικό βέλτιστο ποσοστό μάθησης για όλα αυτά. Η παραδοσιακή κλίση εκπαιδεύτηκε με πέντε ρυθμούς εκμάθησης (1, 0,1, 0,01, 0,001, 0,0001) για να αποδειχτεί πόσο πολύ η ταχύτητα εκμάθησης μπορεί να επηρεάσει σημαντικά τα αποτελέσματα της προπόνησης. Τα προκύπτοντα RBM έχουν τεράστια διακύμανση που καθορίζεται από την επιλογή του ρυθμού εκμάθησης. Όταν το ποσοστό μάθησης ήταν μεγάλο, τα μοντέλα αποτελεσμάτων απέτυχαν εντελώς, ενώ καλύτερα αποτελέσματα αποκτήθηκαν όταν το ποσοστό μάθησης ήταν πολύ χαμηλό. Για να ελεγχθεί ο προτεινόμενος προσαρμοστικός ρυθμός μάθησης, εκπαιδεύτηκαν τα RBM των κρυφών νευρώνων με την παραδοσιακή κλίση και τις ίδιες πέντε τιμές (1, 0,1, 0,01, 0,001, 0,0001) για να αρχικοποιηθεί ο ρυθμός εκμάθησης. Από αυτές τις πληροφορίες δοκιμής, τα αποτελέσματα είναι πιο σταθερά, η διακύμανση μεταξύ των επακόλουθων RBM είναι μικρότερη από τα αποτελέσματα που λαμβάνονται με τον ρυθμό εκμάθησης ανεξάρτητα από τον αρχικό ρυθμό εκμάθησης και όλοι οι RBM εκπαιδεύτηκαν αποτελεσματικά. Αυτά τα αποτελέσματα αποκάλυψαν ότι ο ρυθμός προσαρμοστικής μάθησης αποδίδει καλύτερα, οδηγώντας σε βελτιωμένα αποτελέσματα. Ωστόσο, ήταν ελαφρώς καλύτερο να χρησιμοποιηθεί ένας ρυθμός συνεχούς μάθησης 0,001 τόσο στα προ-προπονητικά όσο και στα προπονητικά στάδια.

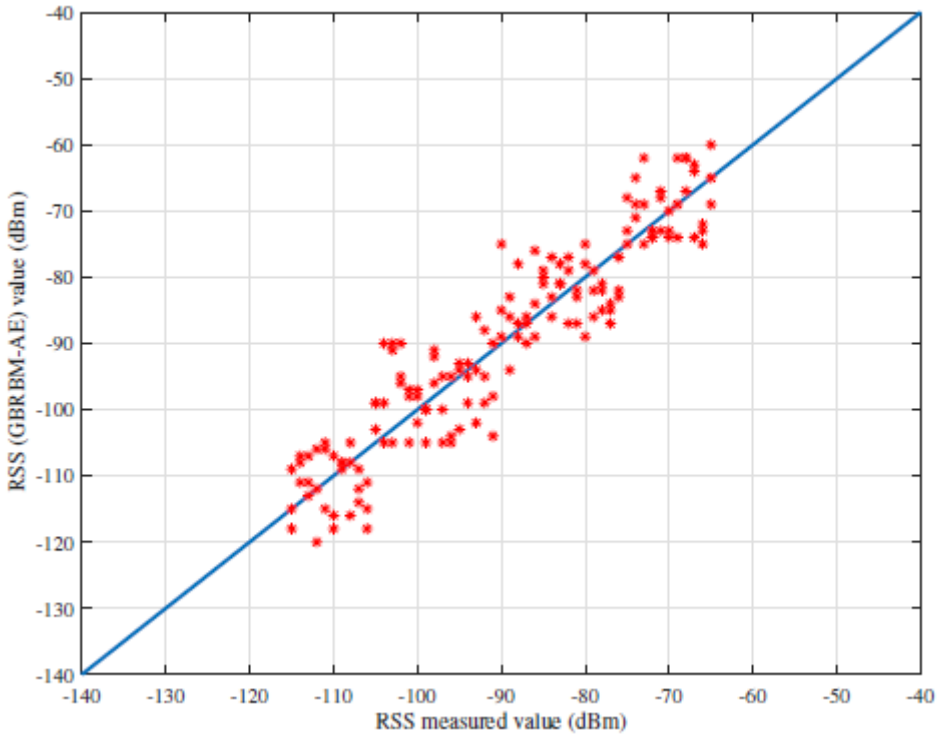
Η Εικόνα 18 δείχνει την απόδοση του προσαρμοστικού ρυθμού μάθησης κατά τη διάρκεια της μάθησης. Η διαδικασία βρήκε κατάλληλες τιμές ρυθμού μάθησης όταν χρησιμοποιήθηκε η ενισχυμένη κλίση. Συγκεκριμένα, βρέθηκαν έξι μπλοκ GBRBM για τα προεκπαιδευτικά στάδια και πέντε επίπεδα δικτύου για την εκπαίδευση στον αυτόματο κωδικοποιητή. Οι αριθμοί νευρώνων για κρυφά επίπεδα GBRBM πολλαπλών μπλοκ είναι 64, 56, 48, 32 και 16, αντίστοιχα. Τα ποσά της εποχής των φάσεων προ-προπόνησης και προπονητικής έχουν οριστεί σε 250 και 500, αντίστοιχα. Οι ταχύτητες εκμάθησης των δύο φάσεων ορίζονται εξίσου στο 0,001, όπως δείχνει ο Πίνακας 2.

Πίνακας 2: Παράμετροι προσομοίωσης.

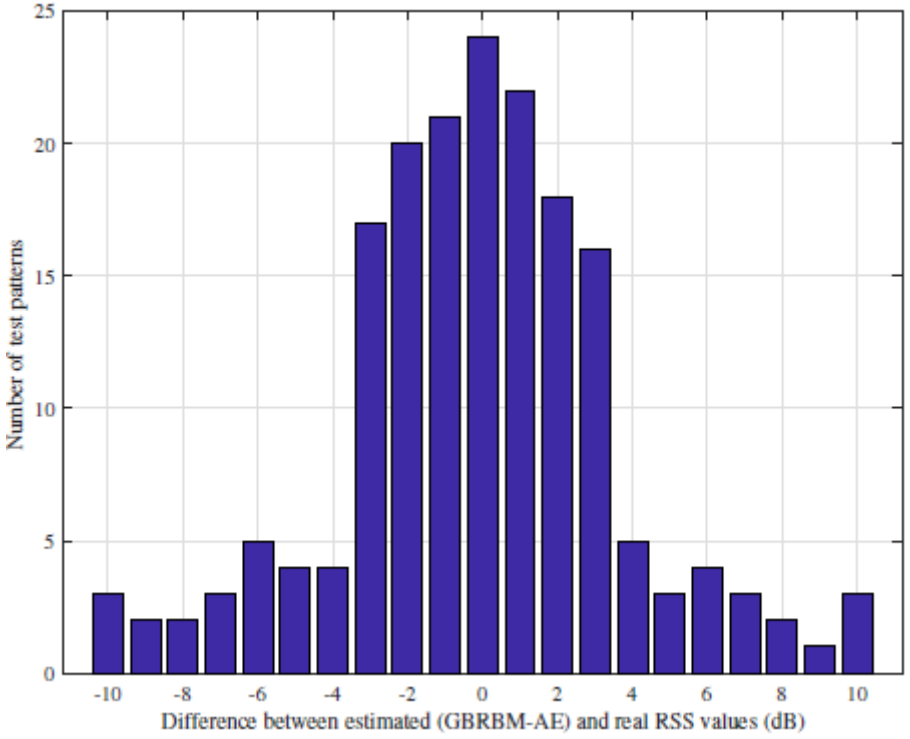
Βασικές παράμετροι	Επιλογές
Πλήθος επιπέδων	5
Πλήθος νευρώνων στο bottleneck layer	7
Πλήθος των GBRBM	5
Αριθμός εποχών στο στάδιο προεκπαίδευσης	250
Αριθμός εποχών στο στάδιο εκπαίδευσης	500

Ρυθμός εκπαίδευσης	0.001

Τα καλύτερα αποτελέσματα λήφθηκαν με το GBRBM, όπως παρουσιάζονται στην Εικόνα 20 και Εικόνα 21. Σαφώς, οι τιμές της μεθόδου GBRBM είναι τα καλύτερα αποτελέσματα σε σύγκριση με άλλους αλγόριθμους που εξετάστηκαν (SVM και ANN). Η μπλε γραμμή αντιπροσωπεύει την έξοδο GBRBM-AE, ενώ οι κόκκινες κουκκίδες απεικονίζουν πραγματικές μετρήσεις UAV. Επιπλέον, παρατηρήσαμε ότι οι προβλεπόμενες τιμές είναι επαρκώς κοντά στις τιμές μέτρησης.



Εικόνα 20: Διαφορά τιμών πραγματικής στάθμης λήψης και του GBRBM (dBm).



Εικόνα 21: Στατιστική διασπορά των διαφορών στη στάθμη λήψης του GBRBM (dB) ως προς τις πραγματικές τιμές.

Το αποτέλεσμα ακρίβειας του προτεινόμενου μοντέλου φαίνεται από την ανάλυση χρονικής πολυπλοκότητας, η οποία δείχνει επίσης πόσο χρόνο χρειάζεται για την εκπαίδευση και την επικύρωση του μοντέλου, όπως δείχνει ο Πίνακας 3.

Πίνακας 3: Χρόνοι εκπαίδευσης και επικύρωσης κάθε αλγορίθμου

Αλγόριθμος	**Χρόνος σε milliseconds**
GBRBM	2.1
ANN	7
SVM	68

4.5.5 Συμπέρασμα μελέτης

Εν κατακλείδι, στη μελέτη που περιγράφεται στο [30], προτείνεται ένα πλαίσιο που βασίζεται σε GBRBM ενσωματωμένα με ένα DNN που βασίζεται σε αυτόματο κωδικοποιητή για την εκτίμηση της λαμβανόμενης ισχύος σήματος σε ένα UAV που πετά σε ένα εύρος υψών και συνδέεται σε ένα κυψελοειδές δίκτυο. Οι πραγματικές μετρήσεις ισχύος σήματος UAV χρησιμοποιούνται για την εκπαίδευση και την επικύρωση του συστήματος. Αν και η ικανότητα των RBM να εξερευνούν τα λανθάνοντα χαρακτηριστικά με τρόπο χωρίς επίβλεψη, η εκπαίδευσή τους είναι πρόκληση καθώς η στοχαστική κλίση τείνει σε υψηλή διακύμανση και αποκλίνουσα συμπεριφορά και ο ρυθμός εκμάθησης πρέπει να ρυθμιστεί χειροκίνητα σύμφωνα με την εκπαιδευμένη δομή του RBM. Το πρόβλημα της ύπαρξης κρυμμένων νευρώνων χωρίς νόημα κατά την εκπαίδευση των RBM είναι έντονο. Για να ξεπεραστούν αυτά τα ζητήματα, προτείνεται ένας νέος αλγόριθμος που χρησιμοποιεί έναν προσαρμοστικό ρυθμό μάθησης μαζί με μια βελτιωμένη κλίση. Η ενισχυμένη κλίση χρησιμοποιείται για

την επιτάχυνση της συνολικής εκμάθησης των κρυφών νευρώνων, σε αντίθεση με την παραδοσιακή αξιοπρεπή κλίση. Επιπλέον, παρέχονται συγκρίσεις απόδοσης με αλγόριθμους SVM και ANN για να καταδειχθεί η εγκυρότητα και τα οφέλη του προτεινόμενου αλγορίθμου, όπου τα ληφθέντα αποτελέσματα αποκάλυψαν ότι ο GBRBM υπερέχει των άλλων αλγορίθμων με ισχυρή ικανότητα βελτιστοποίησης.

4.6 Προτάσεις βελτίωσης

Στα παραδείγματα που μελετήθηκαν, έγιναν πιο κατανοητά τα εξής:

1. Πώς μπορεί ένας χρήστης να δημιουργήσει και να προσαρμόσει ένα περιβάλλον με torchrl (CartPole ή InvertedDoublePendulum-v4)
2. Πώς μπορεί ένας χρήστης να γράψει ένα μοντέλο και μια συνάρτηση απώλειας
3. Πώς μπορεί ένας χρήστης να δημιουργήσει έναν τυπικό βρόχο εκπαίδευσης.
4. Πώς μπορεί ένας χρήστης να εφαρμόσει στην πράξη τους αλγορίθμους DQN και PPO.

Σαν περαιτέρω μελέτη και προτάσεις βελτίωσης, ο χρήστης μπορεί να εφαρμόσει ορισμένες ακόλουθες τροποποιήσεις. Από την άποψη της αποτελεσματικότητας, θα μπορούσε να εκτελέσει πολλές προσομοιώσεις παράλληλα για να επιταχύνει τη συλλογή δεδομένων. Κάτι τέτοιο, μπορεί να γίνει πιο εύκολα με τη χρήση του ParallelEnv και της πολυνηματικής επεξεργασίας, καθώς έτσι ο παραλληλισμός εφαρμόζεται ευκολότερα στον κώδικα.

Από την άποψη της καταγραφής, θα μπορούσε κανείς να προσθέσει έναν μετασχηματισμό torchrl.record.VideoRecorder στο περιβάλλον, για υπάρχει και οπτική αναπαράσταση των βημάτων. Η βιβλιοθήκηα torchrl.record μπορεί για παράδειγμα να χρησιμοποιηθεί, ώστε η έξοδος να είναι ένα αρχείο της μορφή mp4, το οποίο είναι αρχείο βίντεο.

5 Συμπεράσματα

Συνοψίζοντας, στην παρούσα εργασία αναλύθηκε και παρουσιάστηκαν οι αρχές της μηχανικής μάθησης καθώς και οι υποκατηγορίες οι οποίες διακρίνονται σε αυτή. Δόθηκε ιδιαίτερη έμφαση στην ενισχυτική μάθηση, καθώς αυτή χρησιμοποιείται περισσότερο στις τηλεπικοινωνίες.

Στην ανάλυση του θεωρητικού υποβάθρου, έγινε μια παρουσίαση των δύο κατηγοριών της μηχανικής μάθησης, οι οποίες είναι η εποπτευόμενη και η μη εποπτευόμενη μάθηση, ενώ ακολούθως παρουσιάστηκε η ενισχυτική μάθηση, ώστε να συμπληρωθεί έτσι η ανάλυση της μηχανικής μάθησης. Ακολούθως, έγινε ανάλυση των ανακλαστικών επιφανειών με έμφαση στα μη επανδρωμένα οχήματα, προκειμένου να γίνει πιο κατανοητό στον αναγνώστη η ανάγκη για την εφαρμογή της μηχανικής μάθησης σε αυτό τον τομέα.

Επιπρόσθετα, δεδομένου ότι η ενισχυτική μάθηση είναι αυτή που έχει τη μερίδα του λέοντος στις τηλεπικοινωνίες και στα μη επανδρωμένα οχήματα, αναλύθηκαν και συγκρίθηκαν οι πιο επικρατείς αλγόριθμοι, DQN και PPO, παραθέτοντας και τον αντίστοιχο κώδικα σε γλώσσα python. Στο ίδιο κεφάλαιο παρατίθεται και αντίστοιχες πειραματικές μελέτες από την εφαρμογή νευρωνικών δικτύων για την εκτίμηση καναλιών σε δίκτυα μη επανδρωμένων οχημάτων, προκειμένου ο αναγνώστης να αξιολογήσει καλύτερα τη συνεισφορά τους σε αυτό τον κλάδο. Φαίνεται δηλαδή από τις μελέτες και τις μετρήσεις που πραγματοποιήθηκαν ότι οι αλγόριθμοι με τη χρήση των νευρωνικών δικτύων βοηθούν σε καλύτερες τιμές σήματος λήψης και σε πολύ καλύτερο υπολογιστικό χρόνο, σε σύγκριση με άλλους αντίστοιχους αλγορίθμους. Η εργασία αυτή κλείνει με προτάσεις βελτίωσης, τόσο στις πειραματικές μελέτες όσο και στις συγκρίσεις των αλγορίθμων που πραγματοποιήθηκαν.

6 Βιβλιογραφία

[1] Q. Zhang, W. Saad and M. Bennis, "Reflections in the Sky: Millimeter Wave Communication with UAV-Carried Intelligent Reflectors," 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 2019, pp. 1-6, doi: 10.1109/GLOBECOM38437.2019.9013626.

[2] Z. Peng, L. Li, M. Wang, Z. Zhang, Q. Liu, Y. Liu, and R. Liu, “An effective coverage scheme with passive-reflectors for urban millimeterwave communication,” IEEE Antennas and Wireless Propagation Letters, vol. 15, pp. 398–401, June 2015.

[3] T. Hong, J. Yao, C. Liu, and F. Qi, “Mmwave measurement of RF reflectors for 5G green communications,” Wireless Communications and Mobile Computing, vol. 2018, May 2018.

[4] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design,” in 2018 IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, Dec 2018, pp. 1–6.

[5] M. T. Barros, R. Mullins, and S. Balasubramaniam, “Integrated terahertz communication with reflectors for 5G small-cell networks,” IEEE Transactions on Vehicular Technology, vol. 66, no. 7, pp. 5647–5657, Dec 2016.

[6] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, “Energy efficient multi-user MISO communication using low resolution large intelligent surfaces,” arXiv preprint arXiv:1809.05397, 2018.

[7] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, “A tutorial on UAVs for wireless networks: Applications, challenges, and open problems,” IEEE Communications Surveys and Tutorials, to appear, 2019.

[8] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, “Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 5, pp. 1046–1061, Mar 2017.

[9] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications,” IEEE Transactions on Wireless Communications, vol. 16, no. 11, pp. 7574–7589, Sep 2017.

[10] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[11] Liu, J. L. (2020, July 13). A clash of RL algorithms. jerrickliu.com. Retrieved July 10, 2024, from https://jerrickliu.com/2020-07-13-FourthPost/

[12] Kumar, D. K. (2024, February 21). PPO Algorithm. medium.com. Retrieved June 10, 2024, from https://medium.com/@danushidk507/ppo-algorithm-3b33195de14a

[13] Baheti, P. B. (2021, August 31). The Beginner’s Guide to Deep Reinforcement Learning. v7labs.com. Retrieved May 16, 2024, from https://www.v7labs.com/blog/deep-reinforcement-learning-guide#what-is-deep-reinforcement-learning

[14] Ben Salem, H. B. S. (2023, January 26). Supervised VS Unsupervised VS Reinforcement learning. Retrieved June 5, 2024, from https://medium.com/@bensalemh300/supervised-vs-unsupervised-vs-reinforcement-learning-a3e7bcf1dd23

[15] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," in IEEE Transactions on Wireless Communications, vol. 18, no. 11, pp. 5394-5409, Nov. 2019, doi: 10.1109/TWC.2019.2936025.

[16] M. Mozaffari, W. Saad, M. Bennis and M. Debbah, "Mobile Unmanned Aerial Vehicles (UAVs) for Energy-Efficient Internet of Things Communications," in IEEE Transactions on Wireless Communications, vol. 16, no. 11, pp. 7574-7589, Nov. 2017, doi: 10.1109/TWC.2017.2751045.

[17] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, “Placement optimization of UAV-mounted mobile base stations,” IEEE Communications Letters, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[18] V. V. Chetlur and H. S. Dhillon, “Downlink coverage analysis for a finite 3D wireless network of unmanned aerial vehicles,” IEEE Transactions on Communications, to appear, 2017.

[19] Y. Pang, Y. Zhang, Y. Gu, M. Pan, Z. Han, and P. Li, “Efficient data collection for wireless rechargeable sensor clusters in harsh terrains using UAVs,” in Proc. of IEEE Global Communications Conference (GLOBECOM), Austin, TX, USA, Dec. 2014.

[20] B. Fu and L. A. DaSilva, “A mesh in the sky: A routing protocol for airborne networks,” in Proc., IEEE MILCOM, Oct. 2

[21] H. Claussen, “Autonomous self-deployment of wireless access networks,” Bell Labs Tech. Journal, vol. 14, no. 1, pp. 55–71, Spring

[22] F. Jiang and A. Swindlehurst, “Optimization of UAV heading for the ground-to-air uplink,” IEEE Journal on Sel. Areas in Commun., vol. 30, no. 5, pp. 993–1005, Jun. 2

[23] A. Al-Hourani, S. Kandeepan, and S. Lardner, “Optimal LAP altitude for maximum coverage,” IEEE Wireless Commun. Letters, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[24] P. Zhan, K. Yu, and A. L. Swindlehurst, “Wireless relay communications with unmanned aerial vehicles: Performance and optimization,” IEEE Trans. on Aerospace and Electronic Systems, vol. 47, no. 3, pp. 2068–2085, Jul. 2011.

[25] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage,” IEEE Commun. Letters, vol. 20, no. 8, Aug. 2016.

[26] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, “Efficient 3-D placement of an aerial base station in next generation cellular networks,” in Proc., IEEE Intl. Conf. on Commun. (ICC), May

[27] G. Geraci et al., "What Will the Future of UAV Cellular Communications Be? A Flight From 5G to 6G," in IEEE Communications Surveys & Tutorials, vol. 24, no. 3, pp. 1304-1335, thirdquarter 2022, doi: 10.1109/COMST.2022.3171135.

[28] Paszke. A. P. (2024). Reinforcement Learning (DQN) Tutorial. Retrieved June 15, 2024, from https://pytorch.org/tutorials/intermediate/reinforcement_q_learning.html

[29] Moens. V. M. (2024). Reinforcement Learning (PPO) with TorchRL Tutorial. Retrieved June 11, 2024, from https://pytorch.org/tutorials/intermediate/reinforcement_ppo.html?highlight=ppo

[30] Al-Gburi, Ahmed & Abdullah, Osamah & Sarhan, Akram & Al-Hraishawi, Hayder. (2022). Channel Estimation for UAV Communication Systems Using Deep Neural Networks. Drones. 6. 10.3390/drones6110326.

7 Παραρτήματα

Παράρτημα 1: Κώδικας python για τον αλγόριθμο DQN

```
import gym
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.losses import MeanSquaredError
from collections import deque

# Define the DQN agent class
class DQNAgent:
    def __init__(self, state_size, action_size):
        self.state_size = state_size
        self.action_size = action_size
        self.memory = deque(maxlen=2000)
        self.gamma = 0.95  # Discount factor
        self.epsilon = 1.0  # Exploration rate
        self.epsilon_min = 0.01
        self.epsilon_decay = 0.995
        self.model = self._build_model()

    def _build_model(self):
        model = Sequential()
        model.add(Dense(24, input_dim=self.state_size, activation='relu'))
        model.add(Dense(24, activation='relu'))
        model.add(Dense(self.action_size, activation='linear'))
        model.compile(optimizer=Adam(), loss=MeanSquaredError())
        return model

    def remember(self, state, action, reward, next_state, done):
        self.memory.append((state, action, reward, next_state, done))

    def act(self, state):
        if np.random.rand() <= self.epsilon:
            return np.random.randint(self.action_size)
        q_values = self.model.predict(state)
        return np.argmax(q_values[0])

    def replay(self, batch_size):
        minibatch = np.array(random.sample(self.memory, batch_size))
        for state, action, reward, next_state, done in minibatch:
            target = reward
            if not done:
                target = reward + self.gamma *
np.amax(self.model.predict(next_state)[0])
            target_f = self.model.predict(state)
            target_f[0][action] = target
            self.model.fit(state, target_f, epochs=1, verbose=0)
        if self.epsilon > self.epsilon_min:
            self.epsilon *= self.epsilon_decay

# Create the environment
env = gym.make('CartPole-v1')
state_size = env.observation_space.shape[0]
action_size = env.action_space.n

# Initialize the DQN agent
```

```
agent = DQNAgent(state_size, action_size)

# Training loop
batch_size = 32
num_episodes = 1000
for episode in range(num_episodes):
    state = env.reset()
    state = np.reshape(state, [1, state_size])
    for t in range(500):
        # Render the environment (optional)
        env.render()

        # Choose an action
        action = agent.act(state)

        # Perform the action
        next_state, reward, done, _ = env.step(action)
        next_state = np.reshape(next_state, [1, state_size])

        # Remember the experience
        agent.remember(state, action, reward, next_state, done)

        # Update the state
        state = next_state

        # Check if episode is finished
        if done:
            break

        # Train the agent
        if len(agent.memory) > batch_size:
            agent.replay(batch_size)
```

Παράρτημα 2: Εκτεταμένος κώδικας python για τον αλγόριθμο DQN

```
import gymnasium as gym
import math
import random
import matplotlib
import matplotlib.pyplot as plt
from collections import namedtuple, deque
from itertools import count

import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F

env = gym.make("CartPole-v1")

# set up matplotlib
is_ipython = 'inline' in matplotlib.get_backend()
if is_ipython:
    from IPython import display

plt.ion()

# if GPU is to be used
device = torch.device(
    "cuda" if torch.cuda.is_available() else
    "mps" if torch.backends.mps.is_available() else
    "cpu"
)

Transition = namedtuple('Transition',
                        ('state', 'action', 'next_state', 'reward'))


class ReplayMemory(object):

    def __init__(self, capacity):
        self.memory = deque([], maxlen=capacity)

    def push(self, *args):
        """Save a transition"""
        self.memory.append(Transition(*args))

    def sample(self, batch_size):
        return random.sample(self.memory, batch_size)

    def __len__(self):
        return len(self.memory)


class DQN(nn.Module):

    def __init__(self, n_observations, n_actions):
        super(DQN, self).__init__()
        self.layer1 = nn.Linear(n_observations, 128)
        self.layer2 = nn.Linear(128, 128)
        self.layer3 = nn.Linear(128, n_actions)

    # Called with either one element to determine next action, or a batch
    # during optimization. Returns tensor([[left0exp,right0exp]...]).
```

```
    def forward(self, x):
        x = F.relu(self.layer1(x))
        x = F.relu(self.layer2(x))
        return self.layer3(x)


# BATCH_SIZE is the number of transitions sampled from the replay buffer
# GAMMA is the discount factor as mentioned in the previous section
# EPS_START is the starting value of epsilon
# EPS_END is the final value of epsilon
# EPS_DECAY controls the rate of exponential decay of epsilon, higher means
a slower decay
# TAU is the update rate of the target network
# LR is the learning rate of the ``AdamW`` optimizer
BATCH_SIZE = 128
GAMMA = 0.99
EPS_START = 0.9
EPS_END = 0.05
EPS_DECAY = 1000
TAU = 0.005
LR = 1e-4

# Get number of actions from gym action space
n_actions = env.action_space.n
# Get the number of state observations
state, info = env.reset()
n_observations = len(state)

policy_net = DQN(n_observations, n_actions).to(device)
target_net = DQN(n_observations, n_actions).to(device)
target_net.load_state_dict(policy_net.state_dict())

optimizer = optim.AdamW(policy_net.parameters(), lr=LR, amsgrad=True)
memory = ReplayMemory(10000)


steps_done = 0


def select_action(state):
    global steps_done
    sample = random.random()
    eps_threshold = EPS_END + (EPS_START - EPS_END) * \
        math.exp(-1. * steps_done / EPS_DECAY)
    steps_done += 1
    if sample > eps_threshold:
        with torch.no_grad():
            # t.max(1) will return the largest column value of each row.
            # second column on max result is index of where max element was
            # found, so we pick action with the larger expected reward.
            return policy_net(state).max(1).indices.view(1, 1)
    else:
        return torch.tensor([[env.action_space.sample()]], device=device,
dtype=torch.long)


episode_durations = []


def plot_durations(show_result=False):
    plt.figure(1)
```

```
    durations_t = torch.tensor(episode_durations, dtype=torch.float)
    if show_result:
        plt.title('Result')
    else:
        plt.clf()
        plt.title('Training...')
    plt.xlabel('Episode')
    plt.ylabel('Duration')
    plt.plot(durations_t.numpy())
    # Take 100 episode averages and plot them too
    if len(durations_t) >= 100:
        means = durations_t.unfold(0, 100, 1).mean(1).view(-1)
        means = torch.cat((torch.zeros(99), means))
        plt.plot(means.numpy())

    plt.pause(0.001)  # pause a bit so that plots are updated
    if is_ipython:
        if not show_result:
            display.display(plt.gcf())
            display.clear_output(wait=True)
        else:
            display.display(plt.gcf())


def optimize_model():
    if len(memory) < BATCH_SIZE:
        return
    transitions = memory.sample(BATCH_SIZE)
    # Transpose the batch (see https://stackoverflow.com/a/19343/3343043
for
    # detailed explanation). This converts batch-array of Transitions
    # to Transition of batch-arrays.
    batch = Transition(*zip(*transitions))

    # Compute a mask of non-final states and concatenate the batch elements
    # (a final state would've been the one after which simulation ended)
    non_final_mask = torch.tensor(tuple(map(lambda s: s is not None,
                                            batch.next_state)), device=device,
dtype=torch.bool)
    non_final_next_states = torch.cat([s for s in batch.next_state
                                                if s is not None])
    state_batch = torch.cat(batch.state)
    action_batch = torch.cat(batch.action)
    reward_batch = torch.cat(batch.reward)

    # Compute Q(s_t, a) - the model computes Q(s_t), then we select the
    # columns of actions taken. These are the actions which would've been
taken
    # for each batch state according to policy_net
    state_action_values = policy_net(state_batch).gather(1, action_batch)

    # Compute V(s_{t+1}) for all next states.
    # Expected values of actions for non_final_next_states are computed
based
    # on the "older" target_net; selecting their best reward with
max(1).values
    # This is merged based on the mask, such that we'll have either the
expected
    # state value or 0 in case the state was final.
    next_state_values = torch.zeros(BATCH_SIZE, device=device)
    with torch.no_grad():
```

```
        next_state_values[non_final_mask] =
target_net(non_final_next_states).max(1).values
    # Compute the expected Q values
    expected_state_action_values = (next_state_values * GAMMA) + reward_batch

    # Compute Huber loss
    criterion = nn.SmoothL1Loss()
    loss = criterion(state_action_values,
expected_state_action_values.unsqueeze(1))

    # Optimize the model
    optimizer.zero_grad()
    loss.backward()
    # In-place gradient clipping
    torch.nn.utils.clip_grad_value_(policy_net.parameters(), 100)
    optimizer.step()


if torch.cuda.is_available() or torch.backends.mps.is_available():
    num_episodes = 600
else:
    num_episodes = 50

for i_episode in range(num_episodes):
    # Initialize the environment and get its state
    state, info = env.reset()
    state = torch.tensor(state, dtype=torch.float32,
device=device).unsqueeze(0)
    for t in count():
        action = select_action(state)
        observation, reward, terminated, truncated, _ =
env.step(action.item())
        reward = torch.tensor([reward], device=device)
        done = terminated or truncated

        if terminated:
            next_state = None
        else:
            next_state = torch.tensor(observation, dtype=torch.float32,
device=device).unsqueeze(0)

        # Store the transition in memory
        memory.push(state, action, next_state, reward)

        # Move to the next state
        state = next_state

        # Perform one step of the optimization (on the policy network)
        optimize_model()

        # Soft update of the target network's weights
        # θ' ← τ θ + (1 −τ )θ'
        target_net_state_dict = target_net.state_dict()
        policy_net_state_dict = policy_net.state_dict()
        for key in policy_net_state_dict:
            target_net_state_dict[key] = policy_net_state_dict[key]*TAU +
target_net_state_dict[key]*(1-TAU)
        target_net.load_state_dict(target_net_state_dict)

        if done:
            episode_durations.append(t + 1)
```

```
            plot_durations()
            break

print('Complete')
plot_durations(show_result=True)
plt.ioff()
plt.show()
```

Παράρτημα 3: Κώδικας python για τον αλγόριθμο PPO

```
import tensorflow as tf
import numpy as np
import gym

# Environment setup
env = gym.make('CartPole-v1')
state_size = env.observation_space.shape[0]
action_size = env.action_space.n

# Hyperparameters
gamma = 0.99  # Discount factor
lr_actor = 0.001  # Actor learning rate
lr_critic = 0.001  # Critic learning rate
clip_ratio = 0.2  # PPO clip ratio
epochs = 10  # Number of optimization epochs
batch_size = 64  # Batch size for optimization

# Actor and Critic networks
class ActorCritic(tf.keras.Model):
    def __init__(self, state_size, action_size):
        super(ActorCritic, self).__init__()
        self.dense1 = tf.keras.layers.Dense(64, activation='relu')
        self.policy_logits = tf.keras.layers.Dense(action_size)
        self.dense2 = tf.keras.layers.Dense(64, activation='relu')
        self.value = tf.keras.layers.Dense(1)

    def call(self, state):
        x = self.dense1(state)
        logits = self.policy_logits(x)
        value = self.value(x)
        return logits, value

# PPO algorithm
def ppo_loss(old_logits, old_values, advantages, states, actions, returns):
    def compute_loss(logits, values, actions, returns):
        actions_onehot = tf.one_hot(actions, action_size, dtype=tf.float32)
        policy = tf.nn.softmax(logits)
        action_probs = tf.reduce_sum(actions_onehot * policy, axis=1)
        old_policy = tf.nn.softmax(old_logits)
        old_action_probs = tf.reduce_sum(actions_onehot * old_policy, axis=1)

        # Policy loss
        ratio = tf.exp(tf.math.log(action_probs + 1e-10) -
tf.math.log(old_action_probs + 1e-10))
        clipped_ratio = tf.clip_by_value(ratio, 1 - clip_ratio, 1 +
clip_ratio)
        policy_loss = -tf.reduce_mean(tf.minimum(ratio * advantages,
clipped_ratio * advantages))

        # Value loss
        value_loss = tf.reduce_mean(tf.square(values - returns))

        # Entropy bonus (optional)
        entropy_bonus = tf.reduce_mean(policy * tf.math.log(policy + 1e-10))

        total_loss = policy_loss + 0.5 * value_loss - 0.01 * entropy_bonus  #
Entropy regularization
        return total_loss
```

```
    def get_advantages(returns, values):
        advantages = returns - values
        return (advantages - tf.reduce_mean(advantages)) /
(tf.math.reduce_std(advantages) + 1e-8)

    def train_step(states, actions, returns, old_logits, old_values):
        with tf.GradientTape() as tape:
            logits, values = model(states)
            loss = compute_loss(logits, values, actions, returns)
        gradients = tape.gradient(loss, model.trainable_variables)
        optimizer.apply_gradients(zip(gradients, model.trainable_variables))
        return loss

    advantages = get_advantages(returns, old_values)
    for _ in range(epochs):
        loss = train_step(states, actions, returns, old_logits, old_values)
    return loss


# Initialize actor-critic model and optimizer
model = ActorCritic(state_size, action_size)
optimizer = tf.keras.optimizers.Adam(learning_rate=lr_actor)

# Main training loop
max_episodes = 1000
max_steps_per_episode = 1000

for episode in range(max_episodes):
    states, actions, rewards, values, returns = [], [], [], [], []
    state = env.reset()
    for step in range(max_steps_per_episode):
        state = tf.expand_dims(tf.convert_to_tensor(state), 0)
        logits, value = model(state)

        # Sample action from the policy distribution
        action = tf.random.categorical(logits, 1)[0, 0].numpy()
        next_state, reward, done, _ = env.step(action)

        states.append(state)
        actions.append(action)
        rewards.append(reward)
        values.append(value)

        state = next_state

        if done:
            returns_batch = []
            discounted_sum = 0
            for r in rewards[::-1]:
                discounted_sum = r + gamma * discounted_sum
                returns_batch.append(discounted_sum)
            returns_batch.reverse()

            states = tf.concat(states, axis=0)
            actions = np.array(actions, dtype=np.int32)
            values = tf.concat(values, axis=0)
            returns_batch = tf.convert_to_tensor(returns_batch)
            old_logits, _ = model(states)

            loss = ppo_loss(old_logits, values, returns_batch -
np.array(values), states, actions, returns_batch)
```

```
            print(f"Episode: {episode + 1}, Loss: {loss.numpy()}")
            break
```

Παράρτημα 4: Εκτεταμένος κώδικας python για τον αλγόριθμο PPO

```
import warnings
warnings.filterwarnings("ignore")
from torch import multiprocessing


from collections import defaultdict

import matplotlib.pyplot as plt
import torch
from tensordict.nn import TensorDictModule
from tensordict.nn.distributions import NormalParamExtractor
from torch import nn
from torchrl.collectors import SyncDataCollector
from torchrl.data.replay_buffers import ReplayBuffer
from torchrl.data.replay_buffers.samplers import SamplerWithoutReplacement
from torchrl.data.replay_buffers.storages import LazyTensorStorage
from torchrl.envs import (Compose, DoubleToFloat, ObservationNorm,
StepCounter,
                          TransformedEnv)
from torchrl.envs.libs.gym import GymEnv
from torchrl.envs.utils import check_env_specs, ExplorationType,
set_exploration_type
from torchrl.modules import ProbabilisticActor, TanhNormal, ValueOperator
from torchrl.objectives import ClipPPOLoss
from torchrl.objectives.value import GAE
from tqdm import tqdm

is_fork = multiprocessing.get_start_method() == "fork"
device = (
    torch.device(0)
    if torch.cuda.is_available() and not is_fork
    else torch.device("cpu")
)
num_cells = 256  # number of cells in each layer i.e. output dim.
lr = 3e-4
max_grad_norm = 1.0

frames_per_batch = 1000
# For a complete training, bring the number of frames up to 1M
total_frames = 50_000

sub_batch_size = 64  # cardinality of the sub-samples gathered from the
current data in the inner loop
num_epochs = 10  # optimization steps per batch of data collected
clip_epsilon = (
    0.2  # clip value for PPO loss: see the equation in the intro for more
context.
)
gamma = 0.99
lmbda = 0.95
entropy_eps = 1e-4

base_env = GymEnv("InvertedDoublePendulum-v4", device=device)

env = TransformedEnv(
    base_env,
    Compose(
        # normalize observations
        ObservationNorm(in_keys=["observation"]),
```

```
        DoubleToFloat(),
        StepCounter(),
    ),
)

env.transform[0].init_stats(num_iter=1000, reduce_dim=0, cat_dim=0)

print("normalization constant shape:", env.transform[0].loc.shape)

print("observation_spec:", env.observation_spec)
print("reward_spec:", env.reward_spec)
print("input_spec:", env.input_spec)
print("action_spec (as defined by input_spec):", env.action_spec)

check_env_specs(env)

rollout = env.rollout(3)
print("rollout of three steps:", rollout)
print("Shape of the rollout TensorDict:", rollout.batch_size)

actor_net = nn.Sequential(
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(2 * env.action_spec.shape[-1], device=device),
    NormalParamExtractor(),
)

policy_module = TensorDictModule(
    actor_net, in_keys=["observation"], out_keys=["loc", "scale"]
)

policy_module = ProbabilisticActor(
    module=policy_module,
    spec=env.action_spec,
    in_keys=["loc", "scale"],
    distribution_class=TanhNormal,
    distribution_kwargs={
        "min": env.action_spec.space.low,
        "max": env.action_spec.space.high,
    },
    return_log_prob=True,
    # we'll need the log-prob for the numerator of the importance weights
)

value_net = nn.Sequential(
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(num_cells, device=device),
    nn.Tanh(),
    nn.LazyLinear(1, device=device),
)

value_module = ValueOperator(
    module=value_net,
    in_keys=["observation"],
```

```
)

print("Running policy:", policy_module(env.reset()))
print("Running value:", value_module(env.reset()))

collector = SyncDataCollector(
    env,
    policy_module,
    frames_per_batch=frames_per_batch,
    total_frames=total_frames,
    split_trajs=False,
    device=device,
)

replay_buffer = ReplayBuffer(
    storage=LazyTensorStorage(max_size=frames_per_batch),
    sampler=SamplerWithoutReplacement(),
)

advantage_module = GAE(
    gamma=gamma, lmbda=lmbda, value_network=value_module, average_gae=True
)

loss_module = ClipPPOLoss(
    actor_network=policy_module,
    critic_network=value_module,
    clip_epsilon=clip_epsilon,
    entropy_bonus=bool(entropy_eps),
    entropy_coef=entropy_eps,
    # these keys match by default but we set this for completeness
    critic_coef=1.0,
    loss_critic_type="smooth_l1",
)

optim = torch.optim.Adam(loss_module.parameters(), lr)
scheduler = torch.optim.lr_scheduler.CosineAnnealingLR(
    optim, total_frames // frames_per_batch, 0.0
)

logs = defaultdict(list)
pbar = tqdm(total=total_frames)
eval_str = ""

# We iterate over the collector until it reaches the total number of frames
it was
# designed to collect:
for i, tensordict_data in enumerate(collector):
    # we now have a batch of data to work with. Let's learn something from
it.
    for _ in range(num_epochs):
        # We'll need an "advantage" signal to make PPO work.
        # We re-compute it at each epoch as its value depends on the value
        # network which is updated in the inner loop.
        advantage_module(tensordict_data)
        data_view = tensordict_data.reshape(-1)
        replay_buffer.extend(data_view.cpu())
        for _ in range(frames_per_batch // sub_batch_size):
            subdata = replay_buffer.sample(sub_batch_size)
            loss_vals = loss_module(subdata.to(device))
            loss_value = (
                loss_vals["loss_objective"]
```

```
                + loss_vals["loss_critic"]
                + loss_vals["loss_entropy"]
            )

            # Optimization: backward, grad clipping and optimization step
            loss_value.backward()
            # this is not strictly mandatory but it's good practice to keep
            # your gradient norm bounded
            torch.nn.utils.clip_grad_norm_(loss_module.parameters(),
max_grad_norm)
            optim.step()
            optim.zero_grad()

    logs["reward"].append(tensordict_data["next", "reward"].mean().item())
    pbar.update(tensordict_data.numel())
    cum_reward_str = (
        f"average reward={logs['reward'][-1]: 4.4f} (init={logs['reward'][0]:
4.4f})"
    )
    logs["step_count"].append(tensordict_data["step_count"].max().item())
    stepcount_str = f"step count (max): {logs['step_count'][-1]}"
    logs["lr"].append(optim.param_groups[0]["lr"])
    lr_str = f"lr policy: {logs['lr'][-1]: 4.4f}"
    if i % 10 == 0:
        # We evaluate the policy once every 10 batches of data.
        # Evaluation is rather simple: execute the policy without
exploration
        # (take the expected value of the action distribution) for a given
        # number of steps (1000, which is our ``env`` horizon).
        # The ``rollout`` method of the ``env`` can take a policy as
argument:
        # it will then execute this policy at each step.
        with set_exploration_type(ExplorationType.MEAN), torch.no_grad():
            # execute a rollout with the trained policy
            eval_rollout = env.rollout(1000, policy_module)
            logs["eval reward"].append(eval_rollout["next",
"reward"].mean().item())
            logs["eval reward (sum)"].append(
                eval_rollout["next", "reward"].sum().item()
            )
            logs["eval
step_count"].append(eval_rollout["step_count"].max().item())
            eval_str = (
                f"eval cumulative reward: {logs['eval reward (sum)'][-1]:
4.4f} "
                f"(init: {logs['eval reward (sum)'][0]: 4.4f}), "
                f"eval step-count: {logs['eval step_count'][-1]}"
            )
            del eval_rollout
    pbar.set_description(", ".join([eval_str, cum_reward_str, stepcount_str,
lr_str]))

    # We're also using a learning rate scheduler. Like the gradient
clipping,
    # this is a nice-to-have but nothing necessary for PPO to work.
    scheduler.step()

plt.figure(figsize=(10, 10))
plt.subplot(2, 2, 1)
plt.plot(logs["reward"])
plt.title("training rewards (average)")
```

```
plt.subplot(2, 2, 2)
plt.plot(logs["step_count"])
plt.title("Max step count (training)")
plt.subplot(2, 2, 3)
plt.plot(logs["eval reward (sum)"])
plt.title("Return (test)")
plt.subplot(2, 2, 4)
plt.plot(logs["eval step_count"])
plt.title("Max step count (test)")
plt.show()
```