Εργασία Βαθιά Μάθηση και Ανάλυση Πολυμεσικών Δεδομένων

Βαθιά Ενισχυτική Μάθηση

Ημερομηνία: 13/05/2025

Δήμος Δημήτριος Πλακωτάρης ΑΕΜ:180

Κώδικας (git repo): https://github.com/dp-ops/DRL_Doom

Εισαγωγή

Στην παρούσα εργασία υλοποιείται ένα σύστημα βασισμένο στη βαθιά ενισχυτική μάθηση (Deep Reinforcement Learning - DRL) το οποίο έχει ως στόχο την εκπαίδευση ενός πράκτορα (agent) να αλληλεπιδρά και να επιβιώνει σε ένα εικονικό, δυναμικό και εχθρικό περιβάλλον πρώτου προσώπου. Συγκεκριμένα, ο πράκτορας εκπαιδεύεται στο περιβάλλον του δημοφιλούς παιχνιδιού Doom, μέσω της πλατφόρμας ViZDoom, η οποία προσφέρει τη δυνατότητα πειραματισμού με αλγορίθμους ενισχυτικής μάθησης σε τρισδιάστατα περιβάλλοντα προσομοίωσης πραγματικού χρόνου.

Σκοπός της εργασίας είναι να μελετηθεί η ικανότητα μοντέλων βαθιάς ενισχυτικής μάθησης να λαμβάνουν αποφάσεις με βάση ακολουθίες οπτικών δεδομένων και να προσαρμόζουν τη συμπεριφορά τους σε σύνθετα περιβάλλοντα μεταβαλλόμενης δυσκολίας. Πιο συγκεκριμένα, η εκπαίδευση του πράκτορα πραγματοποιείται σε τρία διαφορετικά περιβάλλοντα του Doom, σχεδιασμένα σε επίπεδα αυξανόμενης δυσκολίας. Κάθε περιβάλλον θέτει διαφορετικές προκλήσεις και στόχους στον πράκτορα, ξεκινώντας από απλές δοκιμές κίνησης και επιβίωσης και καταλήγοντας σε σενάρια επιβίωσης με πολλαπλούς αντιπάλους και περιορισμένους πόρους.

Για την εκπαίδευση του πράκτορα επιλέχθηκαν δύο διαφορετικές πολιτικές ενισχυτικής μάθησης, οι οποίες υλοποιούνται με τη χρήση συνελικτικών νευρωνικών δικτύων (CNNPolicy). Οι πολιτικές που εξετάστηκαν ήταν οι Deep Q-Network (DQN) και Proximal Policy Optimization (PPO). Και οι δύο αλγόριθμοι ανήκουν στην κατηγορία της βαθιάς ενισχυτικής μάθησης, ωστόσο διαφοροποιούνται σημαντικά στον τρόπο με τον οποίο ενημερώνουν την πολιτική του πράκτορα και διαχειρίζονται την εξερεύνηση του περιβάλλοντος.

Η εργασία εστιάζει στη σύγκριση των δύο πολιτικών τόσο σε επίπεδο τελικής απόδοσης του πράκτορα όσο και σε χαρακτηριστικά όπως ο ρυθμός εκμάθησης, η σταθερότητα κατά την εκπαίδευση και η συμπεριφορά του πράκτορα σε σενάρια που δεν έχει ξανασυναντήσει. Ο συνδυασμός CNNPolicy με PPO και DQN επιτρέπει στον πράκτορα να λαμβάνει αποφάσεις βάσει της ανάλυσης των καρέ του παιχνιδιού, αξιοποιώντας τα βαθιά συνελικτικά δίκτυα για την εξαγωγή κρίσιμων οπτικών χαρακτηριστικών.

Ο κώδικας της εργασίας, ο οποίος είναι διαθέσιμος στον σύνδεσμο στην αρχή του εγγράφου, περιλαμβάνει όλα τα σκριπς για την προετοιμασία του περιβάλλοντος, την εκπαίδευση και αξιολόγηση του πράκτορα, καθώς και την αποθήκευση και ανάλυση των αποτελεσμάτων.

Υλικά και Μέθοδοι

Περιβάλλον και επίπεδα

Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκε η πλατφόρμα ViZDoom, μια προσαρμοσμένη έκδοση του κλασικού παιχνιδιού Doom, σχεδιασμένη για εφαρμογές ενισχυτικής μάθησης. Η πλατφόρμα αυτή επιτρέπει την αλληλεπίδραση πράκτορα-περιβάλλοντος μέσω απευθείας ανάγνωσης της εικόνας της οθόνης (screen buffer) και παροχής ενεργειών σε πραγματικό χρόνο.

Η κατασκευή των περιβαλλόντων υλοποιήθηκε μέσω της κλάσης VizDoomGym, η οποία επεκτείνει την κλάση Env της βιβλιοθήκης Gymnasium. Κάθε περιβάλλον βασίζεται σε διαφορετικό αρχείο διαμόρφωσης (.cfg), το οποίο καθορίζει τους κανόνες του σεναρίου, τους στόχους, τις διαθέσιμες ενέργειες και τις παραμέτρους του παιχνιδιού. Τα αρχεία διαμόρφωσης αντλήθηκαν έτοιμα από την επίσημη συλλογή σεναρίων του ViZDoom και δεν δημιουργήθηκαν εξ αρχής.

Κοινό χαρακτηριστικό όλων των περιβαλλόντων είναι ότι το μοντέλο εισόδου αποτελείται από εικόνες μεγέθους 160×100 pixels, μετασχηματισμένες σε αποχρώσεις του γκρι για μείωση της υπολογιστικής πολυπλοκότητας και τυποποίηση των δεδομένων εισόδου. Ο πράκτορας μπορεί να επιλέξει ενέργειες από έναν διακριτό χώρο ενεργειών (Discrete space), ο οποίος διαφοροποιείται ανά περιβάλλον ανάλογα με τις απαιτήσεις του εκάστοτε σεναρίου.

Αναλυτικότερα, στα περιβάλλοντα basic και defend_the_center, ο πράκτορας διαθέτει τρεις δυνατές ενέργειες:

- Move Left
- Move Right
- Shoot

Αντίθετα, στο πιο απαιτητικό σενάριο deadly_corridor, ο χώρος ενεργειών επεκτείνεται σε επτά επιλογές, οι οποίες περιλαμβάνουν κινήσεις εμπρός, πίσω, αριστερά, δεξιά, καθώς και ενέργειες εστίασης όπλου και χρήσης αντικειμένων.

Τα τρία σενάρια που αξιοποιήθηκαν στην εργασία περιγράφονται παρακάτω:

- Basic:
  Ένα εισαγωγικό σενάριο, όπου ο πράκτορας βρίσκεται σε ένα δωμάτιο και καλείται να πυροβολήσει έναν μόνο αντίπαλο. Ο στόχος είναι η γρήγορη εξουδετέρωση του αντιπάλου και η αποφυγή τραυματισμού του χαρακτήρα. Αποτελεί το απλούστερο περιβάλλον, χρήσιμο για τη βασική εκμάθηση της σχέσης μεταξύ εικόνας-ανταμοιβής-ενέργειας.
- Defend The Center:
  Σε αυτό το σενάριο, ο πράκτορας τοποθετείται στο κέντρο ενός κυκλικού χώρου και περιβάλλεται από πολλαπλούς εχθρούς οι οποίοι εμφανίζονται σταδιακά από διάφορες κατευθύνσεις. Ο πράκτορας πρέπει να κινηθεί αριστερά ή δεξιά και να πυροβολεί για να αμυνθεί όσο το δυνατόν περισσότερο. Το σενάριο εισάγει στοιχεία χρονικής διαχείρισης και σταδιακά αυξανόμενης δυσκολίας, καθώς οι εχθροί εμφανίζονται πιο γρήγορα και σε μεγαλύτερους αριθμούς.
- Deadly Corridor:
  Το πιο απαιτητικό σενάριο της εργασίας, όπου ο πράκτορας κινείται σε έναν στενό διάδρομο γεμάτο εχθρούς και εμπόδια. Ο στόχος είναι να φτάσει στο τέλος του διαδρόμου επιβιώνοντας και συλλέγοντας διαθέσιμα πυρομαχικά. Η ανταμοιβή διαμορφώνεται όχι μόνο από τις ενέργειες σκοποβολής, αλλά και από την αποφυγή ζημιάς, την εξουδετέρωση αντιπάλων και τη συλλογή αντικειμένων. Το σενάριο αυτό απαιτεί πιο σύνθετη στρατηγική και διαχείριση κινήσεων, ενισχύοντας την ανάγκη για αποτελεσματική εξερεύνηση και αξιοποίηση του χώρου ενεργειών.

Σε όλα τα παραπάνω περιβάλλοντα, η διαδικασία λήψης παρατήρησης και απόδοσης ανταμοιβής (reward shaping) υλοποιείται με προσαρμοσμένο τρόπο ανά σενάριο. Στο deadly_corridor, για παράδειγμα, η ανταμοιβή τροποποιείται επιπλέον με βάση τη ζημιά που δέχεται ή αποτρέπει ο πράκτορας, την πρόοδο στο διάδρομο και την ποσότητα των πυρομαχικών, ώστε να ενθαρρύνεται στρατηγική συμπεριφορά και όχι απλή τυχαία κίνηση.

Η αρχιτεκτονική των περιβαλλόντων, σε συνδυασμό με τις διαφορετικές προκλήσεις και βαθμούς δυσκολίας που αυτά θέτουν, επιτρέπουν την αξιολόγηση και σύγκριση διαφορετικών αλγορίθμων ενισχυτικής μάθησης σε περιβάλλοντα αυξανόμενης πολυπλοκότητας.

Αλγόριθμοι και Εκπαίδευση

Για την εκπαίδευση του πράκτορα στα τρία σενάρια του ViZDoom, εφαρμόστηκαν δύο διαφορετικές προσεγγίσεις βαθιάς ενισχυτικής μάθησης: οι αλγόριθμοι DQN (Deep Q-Network) και PPO (Proximal Policy Optimization). Η επιλογή αυτών των δύο έγινε ώστε να συγκριθούν δύο ευρέως χρησιμοποιούμενες μεθοδολογίες, μία value-based και μία policy-based, σε περιβάλλοντα αυξανόμενης πολυπλοκότητας.

Η πολιτική PPO αποτελεί μία πιο πρόσφατη και εξελιγμένη προσέγγιση, η οποία βασίζεται στη βελτιστοποίηση της πολιτικής μέσω σταδιακών και σταθερών ενημερώσεων. Η PPO περιορίζει δραστικά τις υπερβολικές αλλαγές στην πολιτική από βήμα σε βήμα, γεγονός που συμβάλλει σε σταθερότερη και ασφαλέστερη εκπαίδευση, ειδικά σε περιβάλλοντα με υψηλή στοχαστικότητα και πολύπλοκες δυναμικές, όπως τα προχωρημένα σενάρια του Doom. Η επιλογή της PPO πραγματοποιήθηκε με σκοπό τη διερεύνηση των διαφορών στην απόδοση και τη σταθερότητα των δύο αλγορίθμων σε διαφορετικά επίπεδα δυσκολίας.

Ο αλγόριθμος PPO ανήκει στις policy-based μεθόδους και μαθαίνει απευθείας την πολιτική (policy π(a|s)) που καθορίζει την πιθανότητα επιλογής κάθε ενέργειας σε κάθε κατάσταση. Ο PPO διακρίνεται για τη σταθερότητα και την αποτελεσματικότητά του σε πιο σύνθετα και δυναμικά περιβάλλοντα. Η χρήση του έγινε σε όλα τα σενάρια, δηλαδή στο basic, στο defend the center και κρίθηκε απαραίτητη στο σενάριο deadly corridor, λόγω της πολυπλοκότητας του χώρου ενεργειών και της ανάγκης για συνεχόμενη προσαρμογή της πολιτικής του πράκτορα καθώς εξερευνά και αντιμετωπίζει ποικίλες καταστάσεις. Για την εκπαίδευση στο πιο απαιτητικό σενάριο έγινε προσπάθεια με curriculum learning όπου ουσιαστικά η εκπαίδευση του πράκτορα έγινε σε σενάρια αυξανόμενης δυσκολίας έτσι ώστε να μάθει ο agent σε βήματα να λύνει το πρόβλημα. Η άλλη χρήση της πολιτικής έγινε με στεγνή εκπαίδευση στην ενδιάμεση δυσκολία για να δούμε και μόνη της η πολιτική πως μαθαίνει.

Αντίθετα, η πολιτική DQN βασίζεται στην προσέγγιση εκτίμησης της τιμής Q (Q-value) για κάθε διαθέσιμη ενέργεια, σε κάθε κατάσταση. Το μοντέλο εκπαιδεύεται να προβλέπει τη μελλοντική ανταμοιβή που θα λάβει ο πράκτορας επιλέγοντας μια συγκεκριμένη ενέργεια σε μια συγκεκριμένη κατάσταση και σταδιακά προσαρμόζει την πολιτική του ώστε να μεγιστοποιεί το άθροισμα των αναμενόμενων ανταμοιβών. Ο αλγόριθμος αυτός είναι αποδοτικός σε περιβάλλοντα όπου ο χώρος των ενεργειών είναι διακριτός και σχετικά περιορισμένος, όπως συμβαίνει σε αρκετά σενάρια του Doom.

Ο αλγόριθμος DQN αποτελεί μια προσέγγιση που στηρίζεται στη μάθηση της βέλτιστης συνάρτησης αξίας κατάστασης-ενέργειας Q(s, a). Με χρήση νευρωνικού δικτύου, εκτιμάται η αναμενόμενη συνολική ανταμοιβή για κάθε πιθανή ενέργεια, και σε κάθε βήμα ο πράκτορας επιλέγει την ενέργεια με τη μέγιστη εκτιμώμενη τιμή. Το DQN χρησιμοποιήθηκε στο σενάριο deadly_corridor, καθώς είναι αποτελεσματικό σε περιβάλλοντα με διακριτές ενέργειες και περιορισμένη πολυπλοκότητα, όπου η εκτίμηση της Q-συνάρτησης είναι σταθερή και αξιόπιστη. Ο λόγος που χρησιμοποιήθηκε σε αυτό το σενάριο ήταν για να γίνει η σύγκριση της αποτελεσματικότητας των δύο στρατηγικών σε πιο ένα πιο απαιτητικό περιβάλλον και να γίνει η σύγκριση στην αποτελεσματικότητα τους.

Κοινό στοιχείο και των δύο αλγορίθμων ήταν η χρήση CNN policy, δηλαδή συνελικτικών νευρωνικών δικτύων ως δομή εισόδου, για την επεξεργασία της εικόνας του παιχνιδιού. Η επιλογή CNN policy είναι ιδανική για περιβάλλοντα όπως το ViZDoom, όπου οι παρατηρήσεις

προέρχονται από εικόνες και απαιτείται εξαγωγή χωρικών χαρακτηριστικών (π.χ. θέσεις εχθρών, αντικειμένων) πριν από τη λήψη απόφασης.

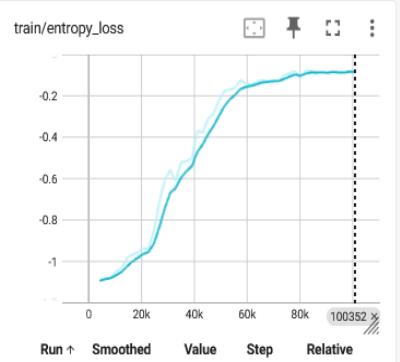
Τα συνελικτικά δίκτυα επιτρέπουν στον πράκτορα να αναγνωρίζει οπτικά μοτίβα και να εντοπίζει σημαντικές περιοχές στην οθόνη, ενισχύοντας έτσι την ποιότητα των εκτιμήσεων Q στο DQN και των πιθανοτήτων ενεργειών στο PPO. Επιπλέον, η χρήση κοινής CNN αρχιτεκτονικής εξασφαλίζει άμεση συγκρισιμότητα των αποτελεσμάτων μεταξύ των δύο αλγορίθμων, καθώς η διαφορά προκύπτει αποκλειστικά από τον τρόπο εκμάθησης και ενημέρωσης της πολιτικής.

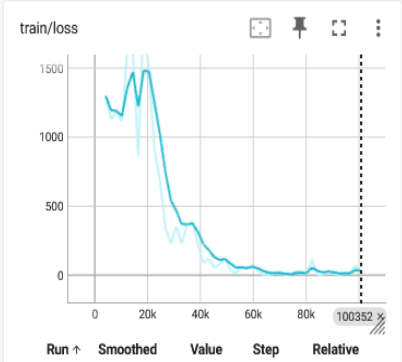
Αποτελέσματα και Συζήτηση

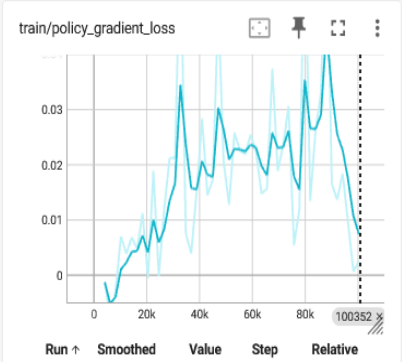
Basic environment:

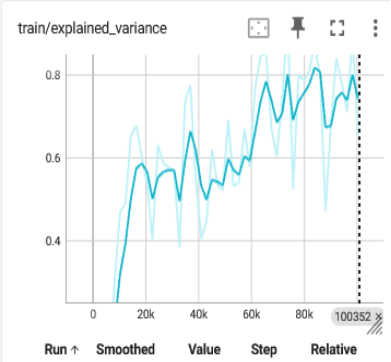
Το βασικό περιβάλλον που εκπαιδεύεται ο agent στο Doom είναι το basic environment, στο οποίο ο πράκτορα μπορεί αν κουνηθεί αριστερά δεξιά και να πυροβολεί. Το περιβάλλον μοιάζει όπως είναι παρακάτω.

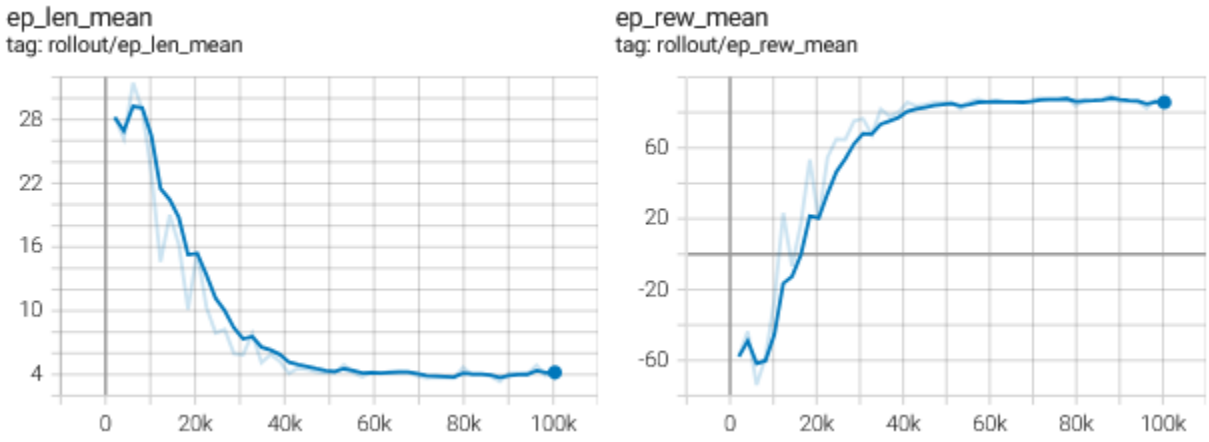
Ο πράκτορας εκπαιδευτικέ με ρυθμό εκπαίδευσης 0.0001, και βήματα εποχής 2048, για συνολικό αριθμό βημάτων 100 χιλιάδες. Η πολιτική που ακολούθησε ήταν PPO με συνελικτικά, και τα αποτελέσματα τις εκπαίδευσης φαίνονται ακολούθως.











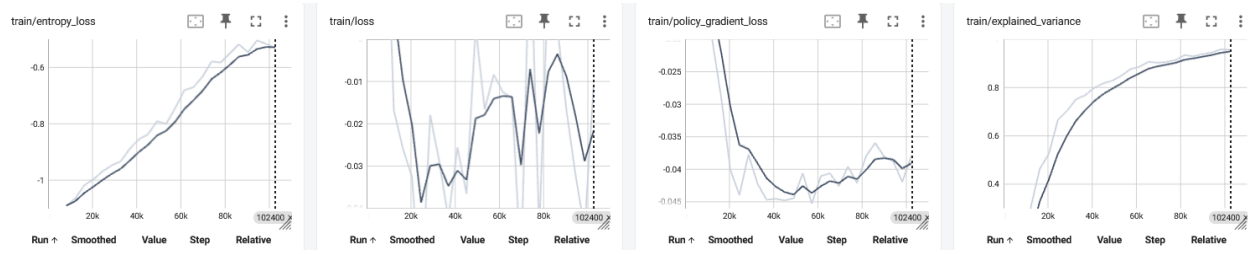
Όπως είναι ορατό ο πράκτορας βρίσκει την βέλτιστη πολιτική λύσης του προβλήματος, μεγιστοποιώντας το αποδεκτό reward ανά επεισόδιο και μειώνοντας την μέση διάρκεια του επεισοδίου. Επίσης το θετικό και αυξανόμενο επί τον πλείστων της συνάρτησης του policy gradient loss δείχνει επίσης την καλή επίδοση του πράκτορα και ότι αυτό κατά την διάρκεια της εκπαίδευσης μαθαίνει.

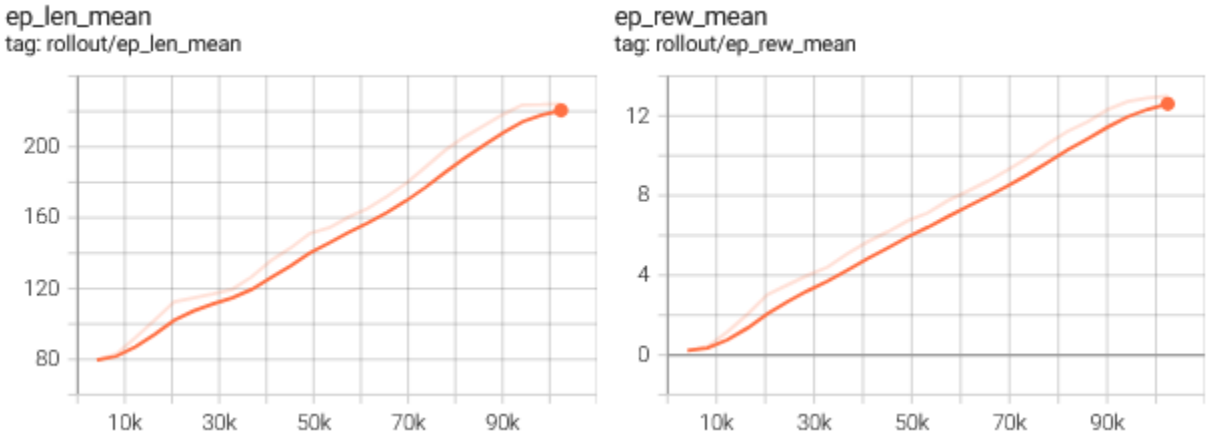
Defend the center environment:

Το επόμενο περιβάλλον που εκπαιδεύεται ο agent στο Doom είναι το defend the center σενάριο, στο οποίο ο πράκτορα μπορεί αν κουνηθεί αριστερά δεξιά και να πυροβολεί, όπως και πριν. Βασική διαφορά του περιβάλλοντος αυτού είναι ότι ο πράκτορας πρέπει αν επιβιώσει όσο το δυνατόν περισσότερο, με του εχθρούς να το προσεγγίζουν από κάθε κατεύθυνση. Το περιβάλλον μοιάζει όπως είναι παρακάτω.

Ο πράκτορας εκπαιδευτικέ με ρυθμό εκπαίδευσης 0.0001, και βήματα εποχής 4096 λόγο των περισσότερων frames του κάθε επεισοδίου, για συνολικό αριθμό βημάτων 100 χιλιάδες. Η

πολιτική που ακολούθησε ήταν PPO με συνελικτικά, και τα αποτελέσματα τις εκπαίδευσης φαίνονται ακολούθως.





Και εδώ βλέπου ενθαρρυντικά αποτελέσματα για τον πράκτορα. Φαίνεται με την πάροδο του χρόνου ότι αυξάνεται και η μέση ανταμοιβή και η μέση διάρκεια του επεισοδίου πράγμα που δηλώνει ότι ο πράκτορας καταφέρνει να επιβιώσεις περισσότερο μέσα στο επεισόδιο. Επίσης και εδώ φαίνεται αυξανόμενο entropy loss και φθίνον policy gradient loss (αλλά όχι σταθερό μηδέν) που δηλώνουν ότι ο πράκτορας εκπαιδεύεται. Τα μέσα fps κατά την εκπαίδευση είναι 33.

Deadly corridor environment:

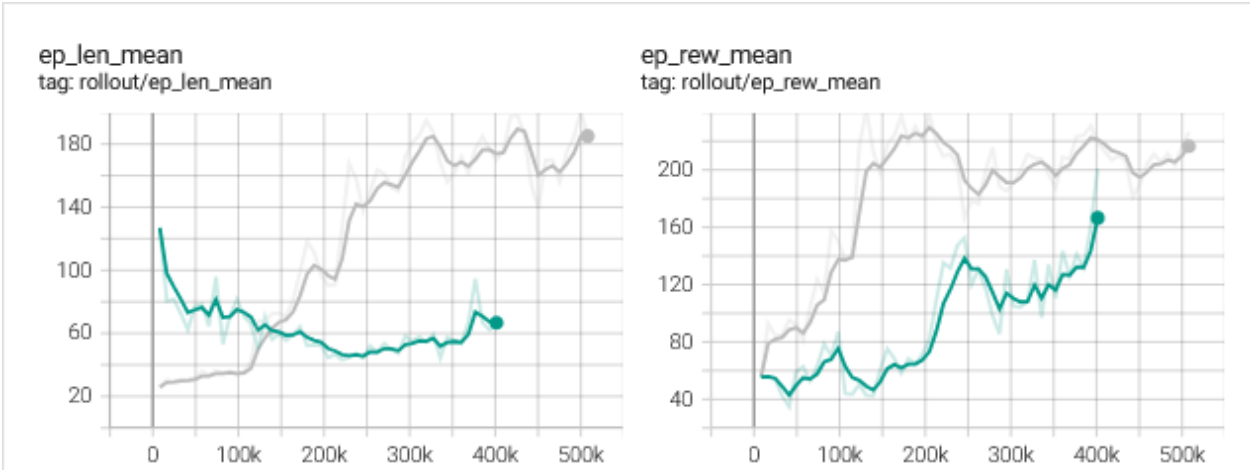
Στο τελευταίο και πιο απαιτητικό σενάριο έγινε σύγκριση δύο διαφορετικών πολιτικών εκμάθησης του πράκτορα, αυτή της Proximal Policy Optimization (PPO) και της Deep Q Networks (DQN).

PPO

Για την PPO δοκιμάστηκαν δύο διαφορετικές προσεγγίσεις μάθησης. Η πρώτη ήταν απλή εκπαίδευση του αλγορίθμου στο σενάριο ενδιάμεσης δυσκολίας και η δεύτερη ήταν αυτή του curriculum learning με reward shaping που ήταν εκμάθηση αυξανόμενης δυσκολίας του ίδιου σεναρίου για πιο καθοδηγούμενη εκπαίδευση του πράκτορα.

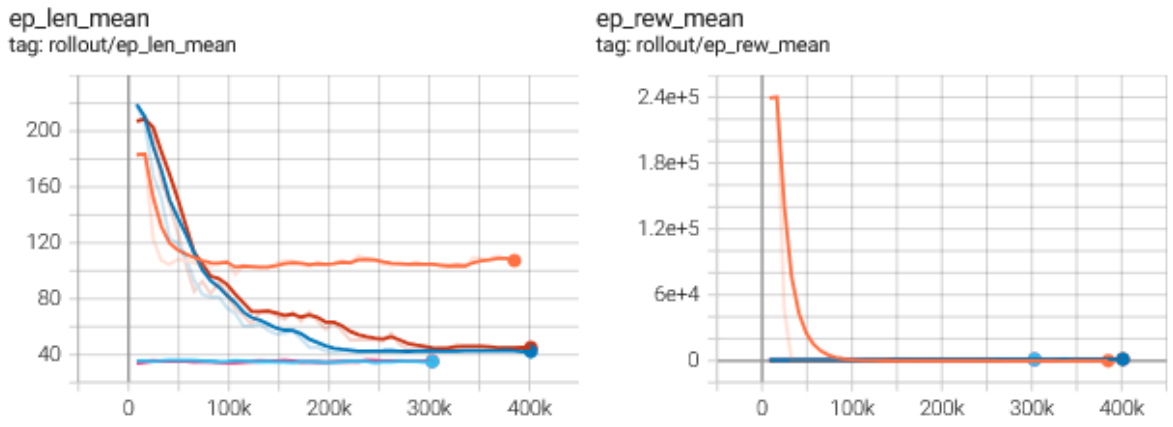
Το σενάριο του deadly corridor περιβάλλοντος μοιάζει κάπως έτσι:

Για τον πρώτο τρόπο εκπαίδευσης έχουμε ρυθμό εκπαίδευσης 0.00001, και βήματα εποχής 8192 λόγο των περισσότερων frames του κάθε επεισοδίου, clip range να είναι 0.1, gamma στο 0.95 και gae_lamda στα 0.9. Έγιναν δύο δοκιμές με την σταθερή ενδιάμεση δυσκολία, αυτές τις δύο να έχουν συνολικό αριθμό βημάτων η καθεμία από 300 και 500 χιλιάδες.



Και εδώ έχουμε ενθαρρυντικά αποτελέσματα με του πράκτορες ξεκάθαρα να μαθαίνουν και να αυξάνουν την ανταμοιβή που λαμβάνουν με το πέρας των βημάτων. Ο λόγος που ο μεγαλύτερος διάρκειας εκπαίδευσης πράκτορας έχει μεγαλύτερης διάρκειας μέσου επεισοδίου είναι διότι του δίνεται μεγαλύτερη ανταμοιβή στο να προσπαθεί να σκοτώνει τους αντιπάλους τους (θα γίνει σχολιασμός στα συμπεράσματα). Και εδώ φαίνεται ότι μαθαίνει ο πράκτορας με αυξανόμενο entropy loss και μεταβλητό αλλά φθηνών loss και policy gradient loss.

Ο πράκτορας με τον δεύτερο τρόπο εκπαίδευσης εκπαιδευτικέ με ρυθμό εκπαίδευσης 0.00001, και βήματα εποχής 8192 λόγο των περισσότερων frames του κάθε επεισοδίου, clip range να είναι 0.1, gamma στο 0.95 και gae_lamda στα 0.9. Το σύνολο των βημάτων για το κάθε στάδιο της εκπαίδευσης στο δεύτερο σενάριο ήταν από 300 χιλιάδες μέχρι 500 χιλιάδες.



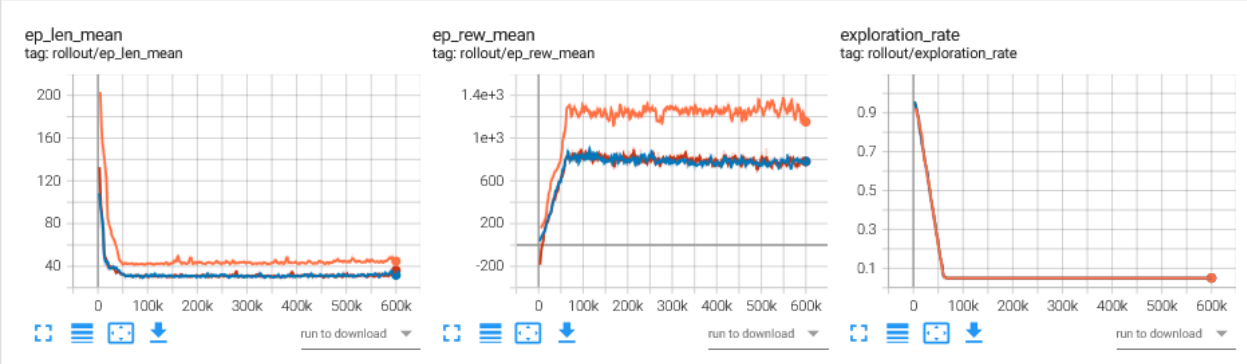
Το PPO_1 έως PPO_6 δείχνει τα σενάρια αυξανόμενης δυσκολίας. Εδώ φαίνεται ότι ο πράκτορας μαθαίνει στην αρχή και μεταφέρει αυτή την γνώση και στα επόμενα σενάρια δυσκολίας. Στο διάγραμμα reward mean φαίνεται ότι η μέση ανταμοιβή τείνει στο μηδέν μετά από κάποιον αριθμό επαναλήψεων. Αυτό δεν ισχύει, στην πραγματικότητα η μέση ανταμοιβή είναι της τάξης των 700

εκεί που φαίνεται μηδέν και ο λόγος είναι ότι στην αρχή της διερεύνησης της πολιτικής ο πράκτορας σύλλεξε μεγάλες ανταμοιβές που κάνουν το διάγραμμα να μοιάζει έτσι. Όσο αφορά τον curriculum learning, φαίνεται ότι για το πρώτο ιδιαίτερα επίπεδο δυσκολίας και για τα επόμενα δύο άλλα λίγο λιγότερο, ο πράκτορας μαθαίνει και δοκιμάζει πράγματα, αναζητώντας και πετυχαίνοντας την βέλτιστη πολιτική για την δυσκολία και τροποποιώντας την όταν η δυσκολία αυξάνεται. Αυτό είναι κυρίως ενδεικτικό από το entropy loss που αυξάνεται μέχρι να προσεγγίσει το μηδέν στις πρώτες τρεις. Όσο για τις τελευταίες δύο και πιο δύσκολες, βλέπουμε ότι υιοθετούν την πολίτη που αναπτύχθηκε από τις προγενέστερες τους με σταθερό gradient loss, ωστόσο συνεχίζουν να μαθαίνουν στην δυσκολία αυτή, πράγμα που φαίνεται στο μεταβλητό gradient loss.

DQN

Με την πολιτική DQN εφαρμόστηκαν στο ίδιο σενάριο δύο διαφορετικές εκδοχές ανταμοιβών, μία που εστιάζει στο hitcount που είναι και αυτή στο παρακάτω διάγραμμα με το μεγαλύτερο μέσο reward επεισοδίου και μία που εστιάζει στο damage_count. Η πρώτη, δηλαδή το hitcount, είναι αυτή που χρησιμοποιήθηκε και στο μεγαλύτερο κομμάτι του PPO και ουσιαστικά είναι ανταμοιβή που δέχεται ο πράκτορας για το damage που δέχονται οι αντίπαλοι, ανεξαρτήτου προέλευσης (δηλαδή είτε από τον χαρακτήρα είτε από τους ίδιους τους αντιπάλους), ενώ το damage count επικεντρώνεται στο damage που κάνει ο πράκτορας μόνο στους αντιπάλους. Η διάκριση αυτή έγινε για να δούμε αν θα επικεντρώνεται ο πράκτορας στο να «καθαρίζει» την πίστα, η να την τελειώνει το συντομότερο δυνατό.

Ο πράκτορας εκπαιδευτικέ με ρυθμό εκπαίδευσης 0.00001, gamma στο 0.95 και buffer size στα 100 χιλιάδες. Το σύνολο των βημάτων για το κάθε στάδιο της εκπαίδευσης ήταν 600 χιλιάδες σε όλες τις εκτελέσεις.



Εδώ όπως και αναμενόμενο λόγο της φύσης της πολιτικής του πράκτορα, βλέπου ότι πετυχαίνει σε πολύ λίγα βήματα την βέλτιστη μέση επιβράβευση και μικρότερη μέση διάρκειας επεισοδίου. Αυτή η συνθήκη είναι κάπως αναμενόμενη καθώς όταν υπολογίζονται η συνάρτηση ανταμοιβής για την κάθε συνθήκη που βρίσκεται ο πράκτορα, από κει και πέρα η κινήσεις του και οι ανταμοιβές του βασίζονται μόνο στο μεγαλύτερο reward και έχουν σταθερή ντετερμινιστική φύση

Συμπεράσματα

Στα πρώτα δύο σενάρια του περιβάλλοντος ViZDoom, όπου αξιολογήθηκε αποκλειστικά η πολιτική PPO χωρίς σύγκριση με τον DQN, ο πράκτορας πέτυχε σταθερή και ταχέως βελτιούμενη απόδοση. Συγκεκριμένα, στο πρώτο σενάριο βασικής επιβίωσης, η PPO προσέγγισε πολύ γρήγορα τη μέγιστη δυνατή βαθμολογία, διατηρώντας υψηλή συνέπεια μεταξύ των επεισοδίων. Στο δεύτερο σενάριο, όπου προστεθήκαν εχθροί και αυξημένος κίνδυνος, ο πράκτορας PPO κατάφερε να αναπροσαρμόσει τη στρατηγική του με επιτυχία, επιτυγχάνοντας εντυπωσιακή αύξηση στην απόδοση και αποδεικνύοντας την ικανότητα του αλγορίθμου να διαχειρίζεται σύνθετες συνθήκες, ακόμα και χωρίς προηγούμενη εμπειρία σε αντίστοιχα σενάρια.

Η παρούσα εργασία στη συνέχεια διερεύνησε την αποτελεσματικότητα δύο αλγορίθμων βαθιάς ενισχυτικής μάθησης, των PPO και DQN, σε ένα δυναμικό και απαιτητικό περιβάλλον πρώτου προσώπου. Τα αποτελέσματα των πειραμάτων ανέδειξαν σημαντικές διαφορές τόσο ως προς την απόδοση όσο και ως προς τη σταθερότητα και την ικανότητα προσαρμογής κάθε πολιτικής σε μεταβαλλόμενα περιβάλλοντα.

Ο αλγόριθμος DQN παρουσίασε ανώτερη συνολική απόδοση στα περισσότερα περιβάλλοντα, με υψηλότερη μέση βαθμολογία και μικρότερη διακύμανση στις επιδόσεις του πράκτορα κατά την εκπαίδευση. Αυτό αποδίδεται στον τρόπο με τον οποίο ο DQN διαχειρίζεται την ενημέρωση της πολιτικής μέσω εμπειριών και πίνακα Q, επιτρέποντας ταχύτερη σταθεροποίηση στις αποφάσεις. Αντιθέτως, ο PPO εμφάνισε μεγαλύτερη ευαισθησία στις παραμέτρους εκπαίδευσης και απαιτούσε περισσότερα επεισόδια για να σταθεροποιήσει τη συμπεριφορά του, ιδίως σε σενάρια αυξημένης δυσκολίας.

Ένα ακόμη σημαντικό εύρημα αφορά τον ρυθμό εκμάθησης. Ο DQN πέτυχε υψηλές επιδόσεις σε σημαντικά λιγότερα επεισόδια σε σχέση με τον PPO, υποδεικνύοντας μεγαλύτερη δειγματική αποδοτικότητα. Η διαφορά αυτή αποκτά ιδιαίτερη σημασία σε περιβάλλοντα πραγματικού χρόνου, όπου η ταχύτητα εκμάθησης αποτελεί κρίσιμο παράγοντα.

Παρά τα θετικά αποτελέσματα, αξίζει να σημειωθούν και οι περιορισμοί της μελέτης. Αρχικά και για τις δύο συγκρίσιμε πολιτικές οι αλγόριθμοι στο σενάριο αγνωούν εντελώς τους αντιπάλους που προσφέρουν ανταμοιβή, και κατευθύνονται ολοταχώς προς το σημείο που τελειώνει το επεισόδιο, καθώς αυτός είναι ένας ασφαλής τρόπος να μαζέψουν ανταμοιβή. Επιπροσθέτως, η επιλογή μόνο δύο πολιτικών και τριών περιβαλλόντων περιορίζει τη δυνατότητα γενίκευσης των συμπερασμάτων σε ευρύτερες κατηγορίες προβλημάτων. Μελλοντικές επεκτάσεις θα μπορούσαν να περιλαμβάνουν τη δοκιμή επιπλέον αλγορίθμων, όπως οι A2C και SAC, καθώς και την εφαρμογή των πολιτικών σε πιο σύνθετα και ρεαλιστικά σενάρια. Επιπλέον, η ενσωμάτωση διαφορετικών τύπων αισθητηριακών δεδομένων (ήχος, depth maps κ.λπ.) θα μπορούσε να εμπλουτίσει τη λήψη αποφάσεων του πράκτορα.

Συνοψίζοντας, τα πειραματικά αποτελέσματα ανέδειξαν την δύναμη και τις ποιοτικές διαφορές των δύο αλγορίθμων βαθιάς ενισχυτικής μάθησης. Ο PPO από την μία αναζητώντας και λόγο της

διερεύνησης της βέλτιστης επιλογής, είναι πιο ασταθής στην αρχή και αργεί να έρθει σε σταθερή και θετική πολιτική. Όσο για τον DQN τα αποτελέσματα δείχνουν την υπεροχή του σε σενάρια υψηλής πολυπλοκότητας και μεταβαλλόμενων συνθηκών, καταδεικνύοντας τη σημασία της επιλογής κατάλληλης πολιτικής ενισχυτικής μάθησης ανάλογα με τις απαιτήσεις του εκάστοτε προβλήματος. Η εργασία αυτή συμβάλλει στην κατανόηση των πρακτικών διαφορών μεταξύ PPO και DQN και θέτει τη βάση για περαιτέρω πειραματισμό σε περιβάλλοντα βαθιάς ενισχυτικής μάθησης.