ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΜΑΘΗΣΗΣ

Νευρωνικά Δίκτυα και Ευφυή Υπολογιστικά Συστήματα

Άσκηση 4: Παίζοντας Atari με Βαθιά Ενισχυτική Μάθηση

Στην τέταρτη και τελευταία άσκηση του εργαστηρίου των Νευρωνικών θα ασχοληθούμε με την ενισχυτική μάθηση και συγκεκριμένα με αλγόριθμους βαθιάς ενισχυτικής μάθησης τους οποίους θα εκπαιδεύσουμε ώστε να παίζουν κλασικά βιντεοπαιχνίδια της Atari.

Η Deepmind και τα παιχνίδια Atari

“Παίζοντας Atari με βαθιά ενισχυτική μάθηση” είναι ο τίτλος του paper[1] της Deepmind που το 2013 παρουσίασε σύστημα το οποίο πετύχαινε καλύτερες επιδόσεις από τον άνθρωπο σε μια σειρά από παιχνίδια Atari. Η Deepmind στη συνέχεια δημοσίευσε paper για πιο σύνθετα παιχνίδια και αντίστοιχα συστήματα ενισχυτικής μάθησης, όπως το AlphaGo και το AlphaZero, τα οποία έλαβαν μεγάλη δημοσιότητα και στο ευρύ κοινό. Το 2020 παρουσίασε πράκτορα ενισχυτικής μάθησης (Agent57) που έχει καλύτερες επιδόσεις από τον άνθρωπο σε όλα τα παιχνίδια της Atari.

Περιβάλλοντα, βιβλιοθήκη και notebook

Θα χρησιμοποιήσουμε τα περιβάλλοντα Atari του OpenAI gym στην εκδοχή όπου η είσοδος στο νευρωνικό είναι εικόνα ώστε να μπορέσουν να τροφοδοτηθούν συνελικτικά δίκτυα βαθιάς μάθησης.

Stable Baselines 3

Ωστόσο, δεν θα χρησιμοποιήσουμε τα περιβάλλοντα απευθείας από την OpenAI αλλά μέσα από τη βιβλιοθήκη Stable Baselines 3 (SB3). Ο λόγος είναι ότι η SB3 μας δίνει έτοιμες βολικές συναρτήσεις για την προεπεξεργασία των εικόνων, την εκπαίδευση και την καταγραφή της απόδοσης των συστημάτων. Εδώ μπορείτε να βρείτε την τεκμηρίωση της βιβλιοθήκης.

Σημειώστε ότι παρότι είναι το ίδιο project και έχουν πολλά κοινά, η Stable Baselines 3 *δεν είναι συμβατή* με την Stable Baselines. Ιδιαίτερη προσοχή όταν κοιτάτε το documentation.

Notebook εργασίας

Το notebook στο οποίο μπορείτε να βασιστείτε για όλες τις απαιτούμενες λειτουργικότητες υπάρχει σε δύο εκδόσεις, μια για Kaggle (η προτεινόμενη) και μία για Colab.

Τις τελευταίες εβδομάδες το Colab μας βγάζει συχνά μήνυμα ότι δεν είναι διαθέσιμοι πόροι GPU. Ως εκ τούτου ως default περιβάλλον που προτείνουμε για την άσκηση είναι το *Kaggle*.

Kaggle

Αυτό είναι το notebook στο Kaggle. Βεβαιωθείτε ότι αντιγράφετε το τελευταίο version 20.
Κάντε το copy και edit.
Ίσως χρειαστεί να ενεργοποιήσετε τα third party cookies στο Chrome, αν δεν εκκινεί το notebook.
ΠΡΟΣΟΧΗ: για να έχετε πρόσβαση στο Internet και στις GPU πρέπει να επιβεβαιώσετε το λογαριασμό σας με αποστολή SMS στο κινητό σας. Πηγαίντε στα Settings πάνω δεξιά.

[1] Playing Atari with Deep Reinforcement Learning. Το 2015 δημοσιεύτηκε στο *Nature* μια πιο high-level εκδοχή “Human-level control through deep reinforcement learning”

Keep-alive script by Ben10

Προκειμένου να πετύχετε πολύωρη “unattended” εκπαίδευση πρέπει να “ξεγελάσετε” το Kaggle ώστε να νομίζει ότι εκτελείτε διαδραστικά το notebook.
Μπορείτε να τρέξετε τοπικά (στον υπολογιστή σας) αυτό το python script (οδηγίες στα σχόλια).

Colab

Αυτό είναι το notebook για Colab. Κάντε ένα copy στο drive σας. Στο Colab (θεωρητικά) δεν χρειάζεται το ngrok για το tensorboard, υπάρχει ειδικό magic command.

Ομάδες - παιχνίδια

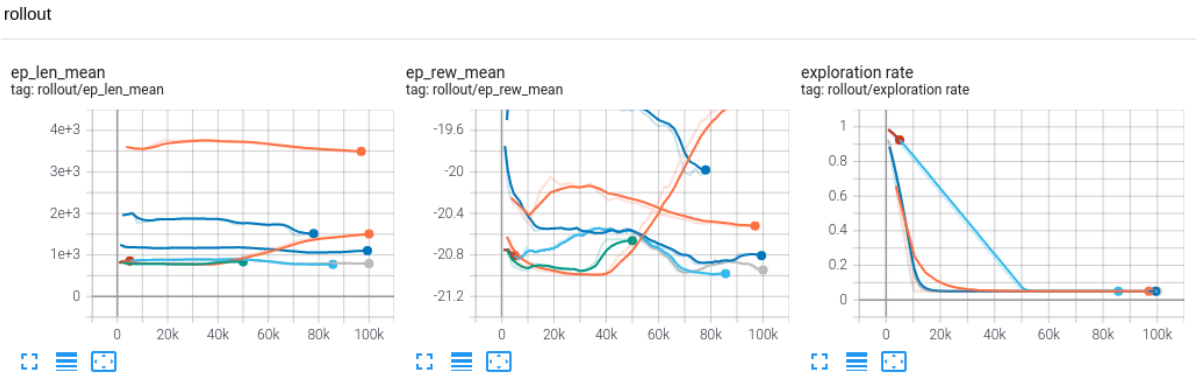
Σε κάθε ομάδα του εργαστηρίου έχει ανατεθεί ένα διαφορετικό παιχνίδι Atari. Μπορείτε να βρείτε ποιο παιχνίδι σας έχει ανατεθεί από τον πίνακα Ομάδες - Atari games.

Σχετικά με τους διαθέσιμους υπολογιστικούς πόρους

Εφόσον θεωρήσουμε ότι τρέχουμε τον κώδικά μας σε κάποιο GPU-enabled cloud (Colab, Kaggle) είναι προφανές ότι έχουμε σαφώς λιγότερους υπολογιστικούς πόρους από την Deepmind αλλά και αρκετούς επιπλέον περιορισμούς. Δεν περιμένουμε να φτάσουμε στο επίπεδο των πρακτόρων της Deepmind, ωστόσο μπορούμε να εκπαιδεύσουμε πράκτορες που σίγουρα έχουν αρχίσει να μαθαίνουν πως να παίζουν. Τα cloud notebooks δεν είναι το κατάλληλο περιβάλλον για τέτοιων απαιτήσεων εκπαίδευση και οι απαιτούμενοι χρόνοι είναι μεγάλοι, εξ ου και είναι σημαντική η δυνατότητα που έχουμε να κάνουμε save ένα μοντέλο και να συνεχίζουμε την εκπαίδευσή του σε μεταγενέστερο χρόνο.

TensorBoard

Ακριβώς επειδή έχουμε αρκετούς περιορισμούς σε υπολογιστική ισχύ και χρόνους εκτέλεσης θα μας ήταν χρήσιμο να μπορούμε να βλέπουμε από νωρίς τη συγκριτική συμπεριφορά των διαφόρων μοντέλων. Για το λόγο αυτό θα χρησιμοποιήσουμε το TensorBoard.



Το TensorBoard μας δίνει τη δυνατότητα να παρακολουθούμε απευθείας μετρικές όπως η μέση ανταμοιβή ή το μέσο μήκος επεισοδίου, κρατώντας μάλιστα με διαφορετικό χρώμα τα διάφορα προγενέστερα runs που έχουμε κάνει.
Σημειώστε ότι τα διαγράμματα είναι διαδραστικά: μπορείτε να κάνετε zoom, pan, fit στα δεδομένα, να αλλάξετε κλίμακα (log) κλπ.

ngrok

Ωστόσο, το TensorBoard είναι web εφαρμογή και οι θύρες που μπορεί να τρέξει στα clouds είναι προφανώς κλειστές. Για το λόγο αυτό, ειδικά εφόσον δουλεύουμε στο Kaggle θα κάνουμε tunneling της θύρας του tensorboard μέσα από το https. Για να το πετύχουμε θα χρησιμοποιήσουμε την υπηρεσία ngrok. Θα χρειαστεί να κάνετε απλά Sign up και στη συνέχεια στο αριστερό sidebar να μεταβείτε στο Authentication -> Your Authtoken. Από εκεί θα μπορείτε να αντιγράψετε το authentication token που χρειαζόμαστε και να το επικολλήσετε μέσα στο notebook, στο σημείο που υποδεικνύεται.

Ονοματολογία περιβαλλόντων

Η ονοματολογία στα περιβάλλοντα που αντιστοιχούν σε κάθε παιχνίδι ακολουθεί ένα συγκεκριμένο σχήμα. Έστω ότι έχουμε το παιχνίδι “Pong”. Έχουμε δύο παράγοντες που επηρεάζουν το περιβάλλον, το Frame Skip και το Sticky Actions ή Repeat action probability.

- Frame Skip: καθορίζει πόσα frames διαρκεί η κάθε ενέργεια του πράκτορα. Αν έχει “Deterministic” στο όνομα, η ενέργεια είναι η ίδια για 4 frames. Αν έχει NoFrameskip στο όνομα, τότε η ενέργεια αποφασίζεται για κάθε frame ξεχωριστά. Αν δεν έχει τίποτε από τα δύο (απλό περιβάλλον), τότε η διάρκεια της κάθε ενέργειας αποφασίζεται στοχαστικά και έχει διάρκεια 2 με 4 frames. Γενικά το Frame skip απλοποιεί και επιταχύνει το πρόβλημα.
- Sticky Actions (Repeat action probability). Το σύστημα Atari δεν έχει εγγενώς πηγές εντροπίας και είναι ντετερμινιστικό. Συνεπώς, αν ξεκινήσεις από το ίδιο φύτρο (seed) της γεννήτριας ψευδοτυχαίων και κάνεις τις ίδιες ενέργειες παίρνεις πάντα τα ίδια αποτελέσματα. Αυτό είναι πλεονέκτημα για αλγόριθμους χωρίς μάθηση που βασίζονται στη μνήμη. Ταυτόχρονα, δεν μας δίνει την μικρή στοχαστικότητα που θέλουμε στην ενισχυτική μάθηση. Ένας τρόπος που μπορούμε να προσθέσουμε στοχαστικότητα είναι με τα sticky actions: ο πράκτορας στην επόμενη κίνηση δεν θα κάνει απαραίτητα αυτό που έχει αποφασίσει αλλά με μια μικρή πιθανότητα p θα κάνει την προηγούμενη ενέργειά του (κολλημένη ενέργεια, sticky action). Αν το όνομα του περιβάλλοντος έχει v0 υπάρχει η δυνατότητα sticky action με πιθανότητα 0.25. Αν το όνομα έχει v4, δεν έχουμε sticky actions.

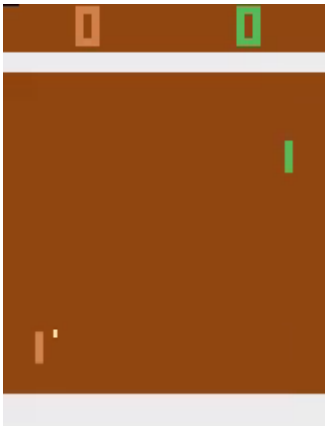
Για παράδειγμα, ο πίνακας με τις ονομασίες περιβαλλόντων για το Pong διαμορφώνεται σύμφωνα με τα παραπάνω ως εξής:

Name	**Frame Skip k**	**Repeat action probability p**
Pong-v0	2~4	0.25
Pong-v4	2~4	0
PongDeterministic-v0	4	0.25
PongDeterministic-v4	4	0
PongNoFrameskip-v0	1	0.25
PongNoFrameskip-v4	1	0

Διαβάστε περισσότερα για τον ντετερμινισμό και την στοχαστικότητα σε αυτό το άρθρο και πιο αναλυτικά στο σχετικό paper[2].

Αλγόριθμοι ενισχυτικής μάθησης

Οι αλγόριθμοι ενισχυτικής μάθησης της SB3 που μπορείτε να δοκιμάσετε με τα περιβάλλοντα Atari είναι όσοι έχουν διακριτό χώρο ενεργειών πλην του HER. Σημειώστε ότι δεν υποστηρίζουν όλοι οι αλγόριθμοι πολλαπλή επεξεργασία (n_envs > 1) η οποία οδηγεί σε ταχύτερη μάθηση.



a2c στο Pong, 200.000 timesteps

Στο αποθετήριο SB3 contrib θα βρείτε έναν ακόμα αλγόριθμο (QR-DQN) κατάλληλο για τα περιβάλλοντα Atari. Τεκμηρίωση εδώ. Προσοχή στην εγκατάσταση, να γίνει μέσω pip και όχι από το master:

```
!pip install sb3-contrib
from sb3_contrib import QRDQN
```

Παραδοτέα

Καλείστε να συγκρίνετε και βελτιώσετε την επίδοση των αλγορίθμων ενισχυτικής μάθησης στο παιχνίδι που σας αντιστοιχεί.

- Ξεκινήστε με έναν πράκτορα χωρίς καθόλου μάθηση (random agent). Ποια είναι η μέση ανταμοιβή του (score) σε ένα περιβάλλον test;
- Χρησιμοποιώντας το TensorBoard δείτε ποιοι αλγόριθμοι RL είναι πιθανότερο να έχουν καλή επίδοση. Το τελικό κριτήριο είναι η μέση ανταμοιβή στο περιβάλλον ελέγχου (test). Ξεκινήστε χωρίς στοχαστικότητα (-v4) και με Deterministic, μετά με με “απλό” περιβάλλον (που έχει στοχαστικότητα στη διάρκεια της ενέργειας) και NoFrameskip στο τέλος (το πιο αργό) . Εκτυπώστε τα συγκριτικά διαγράμματα του TensorBoard (με print screen) μεταξύ των αλγορίθμων για κάθε ένα περιβάλλον ξεχωριστά.
- Δείτε αν η σχετική απόδοση των αλγορίθμων διαφοροποιείται μεταξύ Deterministic, "απλού" και NoFrameskip. Αν δεν υπάρχει σημαντική διαφοροποίηση, μπορείτε να κρατήσετε μόνο ένα version, πιθανώς το ταχύτερο στην εκπαίδευση, και να

[2] Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents, sections 2 και 5.

προχωρήσετε με αυτό μόνο στη σύγκριση των καλύτερων αλγορίθμων (επόμενο βήμα).
- Βελτιώστε όσο είναι δυνατόν τους αλγόριθμους που ξεχωρίζουν. Μελετήστε και χρησιμοποιήστε τις διαθέσιμες παραμέτρους τους που θα βρείτε στο documentation του SB3 και SB3 contrib.
- Αν φτάσετε σε καλά reward/scores δοκιμάστε και στοχαστικότητα (-v0). Η εκπαίδευση θα πάρει περισσότερο χρόνο.
- Δείτε που βρίσκεστε από άποψη επιδόσεων (reward/score) ως προς τα state-of-the-art συστήματα (βλ. πιο κάτω).
- Στο μάθημα μιλήσαμε μόνο για αλγόριθμους value optimization όπως ο DQN. Αν ο βέλτιστος αλγόριθμος για το παιχνίδι σας είναι policy optimization ή actor critic παρουσιάστε σύντομα (σε θεωρία) τον τρόπο λειτουργίας του. Είναι πιθανό να χρειαστεί να συμβουλευτείτε τα papers που αναφέρονται στην τεκμηρίωση του SB3.
- Στο τελικό σας notebook θα πρέπει να μπορούμε να κάνουμε load έναν αποτελεσματικό πράκτορα. Μπορείτε να χρησιμοποιήσετε αυτή την τεχνική (section “φόρτωση εκπαιδευμένου μοντέλου”) για να είναι διαθέσιμο το μοντέλο σας.
- Μέσω ενός cloud ή ως αρχείο μοιράστε μας ένα video recording του gameplay του καλύτερου πράκτορά σας.
- 👾 Προαιρετικό: σημειώστε / συγκρίνετε το δικό σας σκορ στο παιχνίδι (βλ. πιο κάτω ανθρώπινος παίκτης).
- Παραδοτέα: ipynb, doc, pdf, mp4, links όπως σας βολεύει, συμπιεσμένα σε ένα αρχείο στο eclass. Μέγιστο μέγεθος αρχείου upload: 10MB.
- Ερωτήσεις - απορίες στη σχετική περιοχή συζητήσεων του μαθήματος στο e-class.

State-of-the-art επιδόσεις

Σημειώστε ότι το reward συμπίπτει με το score του παιχνιδιού. Για παράδειγμα στο Pong το χειρότερο είναι να χάσεις 21-0 (reward -21) και το καλύτερο να κερδίσεις με το ίδιο σκορ (reward 21). Μπορείτε να συγκρίνετε το σύστημά σας με τα score των state-of-the-art αλγορίθμων για το παιχνίδι σας:
- endtoend.ai
- papers with code Atari Games

Ανθρώπινος παίκτης

- Τα περιβάλλοντα Atari του OpenAI gym βασίζονται στο ALE (Atari 2600 Learning Environment).
- Με τη σειρά του το ALE βασίζεται στον Atari 2600 emulator “Stella”. Μπορείτε να τον κατεβάσετε για το λειτουργικό σας.
- Είναι παράνομο να διατίθεται ο emulator μαζί με τα παιχνίδια (ROMs). Μπορείτε να κατεβάσετε τα ROMs ξεχωριστά, για παράδειγμα από το Atari 2600 VCS ROM Collection.
- Εναλλακτικά, μπορείτε να παίξετε όλα τα Atari games **online** στο free80sarcade.