Invatare automata pentru asistarea deciziei

(Reinforcement Learning)

- Intro -

Luciana Morogan

Academia Tehnica Militara

May 20, 2025

Cuprins

Admin

- Admin - curs

Bibliografie

- *"Reinforcement Learning"*, second edition: An Introduction, Richard S. Sutton, Andrew G. Barto - Google Books
- *"Algorithms for Reinforcement Learning"*, Csaba Szepesvari, online
- *Introduction to Reinforcement Learning with David Silver* (DeepMind) https://www.deepmind.com/learning-resources/introduction-to-reinforcement-learning-with-david-silver
- *UCL Course on RL - David Silver* https://www.davidsilver.uk/teaching/
- *DeepMind x UCL RL Lecture Series* Youtube
- Emma Brunskill 2023 Lecture Materials https://web.stanford.edu/class/cs234/modules.html

Continut curs RL - wannabe

RL

- Bazele RL (Saptamana 1)
  - Intro in RL - curs curent
  - Procese Markov de decizie
  - Planificare prin programare dinamica
  - Predictii de tip model-free
  - Control de tip model-free
- RL aplicat (Saptamana 2)
  - Aproximarea functiei valoare
  - Metode "policy gradient"
  - Integrarea invatarii si planificarii
  - Explorare si exploatare
  - Studiu de caz

Exemple de aplicatii ce folosesc RL

ChatGPT

- **Intro.**

Generative AI is taking off, and along with it excitement and hype about the potential of the technology

Exercitiu. Sa ne gandim la *Gen AI* ca la o tehnologie de uz general (in *en.* suna mai bine :) *general-purpose technology*)!

$\Longrightarrow$ GPT = generative pretrained transformer

Dar... *Gen AI* suna ca si Gen X, Gen Y sau Gen Z care refera grupuri specifice, ca... *Generatie*

Din nou, *en.*: This abbreviation suggests that all of us who are alive today are part of *Generation AI*

Ce este ChatGPT? - OpenAI lanseaza GPT-4 +...

- Model de limba multimodal care foloseste DL pentru a prezice cuvintele dintr-o propozitie
- Genereaza text remarcabil de fluent si poate raspunde la imagini, solicitari bazate pe cuvinte, deep research, multiple reasoning (o4-mini, o4-mini-high, o3), research preview GPT-4.5, Sora video generation, Codex agent (model de limbaj AI antrenat pe cantitati mari de cod sursa, specializat in intelegerea si generarea de cod)

- Cea mai mare parte a tehnologiei din ChatGPT nu este noua!
  - ChatGPT - versiune ajustata a GPT-3.5 (familie de modele mari de limba pe care OpenAI a lansat-o cu luni inainte de chatbot)
  - GPT-3.5 - versiune actualizata a GPT-3 din 2020
  - Anterioar GPT-3.5 - InstructGPT, ianuarie 2022
  - Niciuna dintre aceste versiuni nu a fost prezentata public

De ce atatea update-uri ale ChatGPT?

- S-a folosit tehnica numita *antrenament adversarial* pentru a impiedica ChatGPT sa-i lase pe utilizatori sa-i induca un comportament rau (*jailbreaking*)
  - Chatbot vs. chatbot: un adversar tip chatbot ataca un alt chatbot generand text pentru a-l forta sa invinga constrangerile obisnuite si sa produca raspunsuri nedorite (inclusiv bias)
  - Atacurile de succes sunt adaugate in datele de antrenare ale ChatGPT

Si care este legatura cu RL???

John Schulman (cofounder of OpenAI): ChatGPT was trained in a very similar way to InstructGPT, using a technique called **reinforcement learning from human feedback (RLHF)**. This is ChatGPT's secret sauce. The basic idea is to take a large language model with a tendency to spit out anything it wants - in this case, GPT-3.5 - and tune it by teaching it what kinds of responses human users actually prefer.

RLHF - invatare prin consolidare din preferintele umane

RLHF

- Tehnica care antreneaza un "model de recompensa" direct din feedbackul uman si utilizeaza modelul ca functie de recompensa pentru a optimiza *politica* unui agent (folosind alg. precum *Proximal Policy Optimization*)
- Feedbackul uman este colectat cerand oamenilor sa clasifice cazurile de comportament al agentului
- Exp.:
  - NLP: ChatGPT si InstructGPT (OpenAI), Sparrow (DeepMind), Bing Chat (Microsoft, ianuarie 2023), Copilot, Gemini, Claude AI (de la *Anthropic* - dezvolta conceptul de *Constitutional AI*)
  - Alte domenii: dezvoltarea de roboti pentru jocuri video (exp. OpenAI si DeepMind au instruit agenti sa joace Atari pe baza preferintelor umane)

Drawbacks/Issues

Prin instruirea ChatGPT cu RLHF: modelul a invatat automat comportamentul de refuz $\Longleftrightarrow$ Refuza o multitudine de solicitari

Modelul este inca foarte partinitor. Si DA, ChatGPT este foarte bun in a refuza cererile proaste, dar este, de asemenea, destul de usor sa de pacalit a.i. sa nu refuze

Biggest concern: modelului ii place sa fabrice lucruri/halucineze

Echipa a incercat prinda cele mai problematice exemple de ceea ce poate produce ChatGPT pentru versiunile viitoare: de la cantece despre dragostea lui Dumnezeu pentru preotii violatori si pana la coduri malware care fura numere de carduri

Si in sfarsit... de ce RL si nu si Transformers?

- Amandoua sunt... notoriously data-hungry
- **Transformers**: Vincent Micheli si echipa (Universitatea din Geneva) au antrenat un sistem bazat pe transformeri pentru a simula jocurile Atari folosind o cantitate mica de gameplay. Apoi au folosit simularea pentru a antrena un agent RL, IRIS, pentru a depasi performanta umana

Key insight

Un transformer exceleaza in a prezice urmatorul element dintr-o secventa $\Longrightarrow$ Pt. rezultatul unui joc video: poate invata sa estimeze o recompensa pentru apasarea butonului jucatorului si sa prezica jetoane care reprezinta urmatorul cadru video.
Apoi, un autoencoder poate invata sa reconstruiasca cadrul.

DAR impreuna formeaza un simulator de joc care poate ajuta un agent RL sa invete sa joace

Si in sfarsit... de ce RL si nu si Transformers?

How it works for each Atari game: intr-un ciclu

- un agent RL a jucat pentru o perioada scurta fara a invata
- un sistem invata din cadrele de joc si apasarea butonului agentului pentru a simula jocul
- agentul invata din simulare

Timp total: Aproximativ doua ore, 100.000 cadre si apasari de butoane asociate - per joc

Si in sfarsit... de ce RL si nu si Transformers?

Fluxul sistemului:

- Agentul (CNN + LSTM) joaca jocul timp de 200 de cadre. Primeste input un cadru si raspunde apasand un buton (init aleator). Nu primeste recompense si deci nu a invatain timpul jocului
- Autoencoderul invata sa codifice un cadru intr-un set de jetoane si sa-l reconstruiasca din jetoane
- Transformerul invata sa estimeze recompensa pentru ultima apasare de buton si sa genereze jetoane care reprez. urm. cadru si daca cadrul actual incheie jocul
- Deci cu simbolurile pentru urm. cadru, autoencoderul genereaza imaginea. Prin imagine, agentul invata sa aleaga butonul de apasare care ii maximizeaza recompensa
- Ciclul se repeta: agentul joaca generand noi cadre si apasari de butoane pentru a antrena autoencoderul si transformerul. La randul lor, iesirile autoencoderului si ale transformatorului antreneaza agentul

Si in sfarsit... de ce RL si nu si Transformers?

Rezultate

Agentul depaseste scorul mediu uman in 10 jocuri (inclusiv Pong)
Agentul depaseste ultimele abordari care includ cautarea anticipata (un agent alege apasarile de butoane pe baza cadrelor prezise in plus fata de cadrele anterioare) in 6 jocuri si cele fara cautare anticipata in 13 jocuri.
Drawback: Functioneaza ok cu jocurile care nu implica schimbari bruste in mediul de joc (ex. trecere la alt nivel)

De ce conteaza?

Transformerii au fost folositi in RL ca agenti, nu ca modele de mediu
Aici, este model de mediu: invata sa simuleze un joc sau un mediu
PLUS: Astfel de abordare $\Longrightarrow$ simulatoare de inalta performanta, eficiente pentru esantionare

Si in sfarsit... de ce RL si nu si Transformers?

Concluzie

Succesul initial al modelelor de joc Atari a fost interesant deoarece abordarea prin RL nu a necesitat construirea sau utilizarea unui model de joc

Introducere in RL

Introducere in RL

Step 1

RL Course by David Silver - Lecture 1: Introduction to Reinforcement Learning

- Slideuri: https://www.davidsilver.uk/wp-content/uploads/2020/03/intro_RL.pdf
- Video: https://www.youtube.com/watch?v=2pWv7GOvuf0