Derin Öğrenme

Deep Reinforcement Learning

Emir Öztürk

Learning

Back to square one

- Canlılar çevre ile etkileşime geçerek öğrenirler
- Etkileşim genellikle sıralıdır
- Aksiyon seçimine göre sıra üretilir
- Hedef yönelimli etkileşim gerçekleştirilir.
- Hedefe ulaştığımızı belirten bir ödül
  - Mutluluk, başarma hissi
- Örnekler görmeden optimal durumu öğrenme şansı bulunmaktadır.

Öğrenme problemleri

Learning is the problem

- Supervised
  - Veri ve etiket verilir
  - Hedef veriyi etikete haritalayan fonksiyonun tespiti
    - Bu bir elmadır
- Unsupervised
  - Veri verilir
  - Hedef verileri oluşturan yapının tespit edilmesi
    - Bu iki nesne birbirine benziyor

Öğrenme problemleri

We must reinforce our defenses

- Reinforcement
  - Durum ve aksiyon ikilileri verilir
  - Hedef gelecekteki ödülü maksimize etmektir
    - Yaşamak için bu elmayı yemen gerekiyor

Reinforcement Learning'in amacı

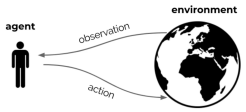
What is the purpose of life

- Problemin çözümünü bulmak
  - Otonom sistemlerde modelleme daha zor
- Adaptasyon ve canlı güncelleme
  - Generalizasyon değil, yeni veriye uygunluk
  - Yola göre düşmemeyi öğrenen robotlar

Interaction Loop

For loop, while loop and than this

- Agent
- Environment
- Action
- Observation
- Reward
- Total reward



Reward

There is no free meal

- Ödül skalar bir sayıdır
- Pozitif ya da negatif olabilir
- Negatif ödül aslında cezadır
- Hedef (G) ise ödüller toplamıdır

Reward Hypotesis

Not like guesswork

- Herhangi bir hedef, kümülatif ödüllerin toplamını maksimize etmek olarak özetlenebilir.
- Hedef, yapılan bir aksiyonun bundan sonraki her adımda kazanacağı ödüllerin toplamı olarak hesaplanır.
- Ayrıca adım sayısı arttıkça ödülün etkisi azaltılmalıdır.

Hedefin belirlenmesi

Run towards the target

- En yakın kararın en iyi seçilmesi en iyi hedefe ulaşma garantisi vermez
- Aksiyonların sonuçları ilerleyen adımlarda belli olabilir
- En yakın ödülü feda ederek uzun vadeli ödülü büyütme şansı bulunabilir
- Durumları aksiyonlara haritalamaya policy adı verilir.

Q Fonksiyonu

Double Q while learning

- $Q(S_t,A_t) = \mathbb{E}[R_t|S_t,A_t]$
- Bulunulan durumdaki aksiyonun vereceği ödülün tahmini
- Verilen Q fonksiyonunun ödülü doğru tahmin edebilmesi için
  - Policy belirlemek
- Örneğin: Policy Q fonksiyonunun bulunulan durumdaki maksimum değerini seçebilir
  - P(S) = argmaxQ(S,A)

Q Fonksiyonu

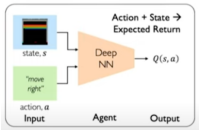
It is really hard

- Q fonksiyonu hazır bulunmamaktadır
- İlk seçenek olarak bir öğrenme algoritması
  - Değerler üzerinden Q fonksiyonunun hesaplanması
- Bir diğer seçenek
  - Policy üzerinden öğrenme gerçekleştirmek

Deep Q Network

And there is more

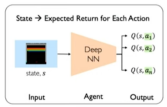
- Q değerinin hesaplanması için bir DNN eğitilir
- Örneğin bir oyun için oyunun o andaki durumu görüntü olarak verilebilir
  - Görüntü Convolution olarak verilebilir
- Bununla birlikte bir aksiyon verilip sonuç elde edilebilir.



Deep Q Network

Always game examples

- State ve Action vermek yerine yalnızca State verilebilir
- Örneğin bir oyun için oyunun o andaki durumu görüntü olarak verilebilir
  - Görüntü Convolution olarak verilebilir
- Action sayısı n olduğu durumda n adet aksiyon'un olasılık tahmini yapılabilir
  - Bunlar içerisinden Q değerini maksimize eden değer seçilir
- Sabit sonuç eldesi engellenmiş olur



Deep Q Network eğitimi

And some math

- Eğitimin gerçekleşmesi için loss'u hesaplayabilmemiz gerekir
- Aksiyona göre hedef belirlenir
- Hedefin tahmin edilen hedefe ulaşıp ulaşmadığı hesaplanabilir.

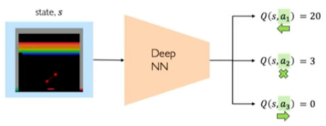
$$\mathcal{L} = \mathbb{E}\left[\left\|\overbrace{\left(r + \gamma \max_{a'} Q(s', a')\right)}^{\text{target}} - \overbrace{Q(s, a)}^{\text{predicted}}\right\|^2\right]$$ Q-Loss

---

Eğitim

We don't need no…

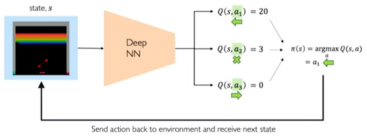
- Örneğin arkonoid için
- Input verildikten sonra Deep NN 3 farklı olasılık vermiş olsun
  - Sol
  - Sağ
  - Bekle
- Olasılıklar için Q değerleri hesaplanır



Eğitim

Decide which way

- Burada maksimum değer sola gitmek olduğu için bu seçilecektir.
- Daha sonra bu değer geriye döndürülür
- Yeni durum alınır



Q-Learning'in sorunları

Everyone has problems

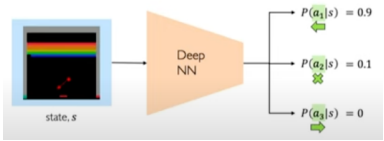
- Karmaşıklık
  - Şu an ifade edilebilen uzaylar sınırlı aksiyon ve sınırlı uzaya sahip
- Esneklik
  - Ortamın değiştiği durumda Q fonksiyonu tanımlandığı için doğru davranamayacaktır

---

Policy Gradient

And everything has gradient

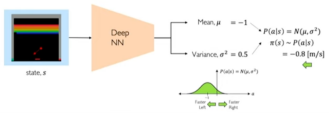
- Aksiyondan ödül hesaplamak yerine bir yöntem
- Bir durumda aksiyonu seçme olasılığını maksimize etmek



Policy avantajı

I don’t have anything to say

- Sürekli veriler tanımlanabilir
- Sol, sağ yerine belirli bir hızda sola sağa gitme tanımlanabilir
- Çıktı ortalama ve varyans olarak alınabilir
- Bu durumda yön değeri belirli bir değerde kullanılabilir



Policy training

An example

- Otonom araçlar
- Agent
  - Araç
- State
  - Kamera, lidar vs.
- Action
  - Direksiyonun çevrilmesi
- Reward
  - Gidilebilen en uzun mesafe

Policy training

And the process

- Araç başlatılır
- Araç çarpana kadar her state action ve policy saklanır
- Çarpma işleminden sonra düşük ödüllü aksiyonların olasılıkları düşürülür
- Yüksek ödüllü aksiyonların olasılıkları arttırılır

Policy training

It's impossible to try everything

- Bir sonraki denemede yüksek olasılıklı aksiyonlar seçilir
- Böylece crash olasılığı azaltılmaya başlanır
- Araçta crash ederek denemek mümkün değildir
- Üretilen mermilerin doğruluğu
  - Örneklem

VISTA

And a good finish

- Otonom araçlar için realistik sahne oluşturma