TRANSDREAMER: REINFORCEMENT LEARNING WITH TRANSFORMER WORLD MODELS

Chang Chen , Yi-Fu Wu , Jaesik Yoon , Sungjin Ahn, Rutgers University & KAIST

https://doi.org/10.48550/arXiv.2202.09481

MAKALE RAPORU

YUSUF AYKUT

1. Giriş

Pekiştirmeli öğrenme ajanın (örneğin bir robot veya ortamla etkileşime giren oyuncu denilebilir.) yazılımsal ya da fiziksel olarak tanımlanmış bir ortamla etkileşime girerek deneme-yanılma yoluyla öğrenme gerçekleştirmesini sağlayan bir yöntemdir. Örneğin bir ajan bir oyunu nasıl oynaması gerektiğini, sürekli başarı gösterek yahut hata yaparak öğrenebilir. Lakin oyun aslında önceden simüle edilmiş bir ortam sağlamaktadır. Gerçek hayata dair bir olguyu, bir oyunmuş gibi öğrenmek kolay değildir. Bunun için Model-Based RL denilen bir yöntem kullanılır. Ajan doğrudan gerçek dünyadaki verilerle öğrenmek yerine, gerçek dünyanın temsiliyeti olan bir simülasyon üzerinden öğrenme gerçekleştirir.

Bu makalede TransDreamer adında transformer-based MBRL ajanı tanıtılmaktadır. Daha önce tanıtılmış Dream adlı bir çerçevenin özelliklerini kalıtır, ancak transformer'ın faydalarını elde etmeyi amaçlar. Bunun bazı sebepleri vardır: RNN'ler ardışık verilerle çalışmak için tasarlanmıştır ancak uzun vadeli bağımlılıkların bulunduğu durumlarda (yani çok önce gerçekleşen bir durumun, çok sonra gerçekleşecek bir durumu etkilemesi) ya da hafızadaki bilgilere dayalı çıkarımda bulunulması gereken durumlarda RNN'ler etkili değillerdir. Bunu çözmek için RNN'ler yerine transformer tabanlı bir model kullanılması önerilmiştir.

Ayrıca makalede TSSM(Transformer State-Space Model) tanıtılmıştır. Aslında bu transformer tabanlı bir modelde durum uzayının nasıl tanımlanacağıyla ilgilidir, Durum uzay mimarisi bir sistemin durum geçişlerini matematiksel olarak modeller. Burada amaç transformer'ların dikkat mekanizması ve uzun vadeli bağımlılıklardaki yetenklerinden durum-uzayı tanımlaması yapılırken faydalanmaktır.

Makalede tanıtılan model, daha önce geliştirilmiş Dream modelinin üzerine kuruludur. Bu model kısmi gözlemlenebilir Markov karar süreci mantığıyla çalışmaktadır (MDP kesikli zamanlı olasılıksal ve stokastik bir kontrol sürecidir.). Burada modelin optimalliğe yakınsaması sırasında 3 tane aşamayı döngüsel olarak çalıştırmaktadır:
1) world model learning , 2) policy learning , 3) environment interaction.

2 "World model"i öğrenmek

Dünya modeli öğrenme, ajanın çevrenin dinamiklerini anlamasını ve tahmin etmesini sağlayan bir süreçtir. Bu model, ajanın mevcut durumundan (state) ve gerçekleştirdiği eylemden (action) yola çıkarak bir sonraki durumu, ödülü ve görevin bitip bitmediğini öngörür. Dünya modeli stokastik(rastgele ve önceden belirlenemyen durumlar) ve deterministik durumları aynı anda modelleyen bir yapıdır. Bu modelde ajan ile politikayı (amacımıza yönelik optimal politikayı) öğreniyoruz, yani en iyi eylemleri seçmeyi öğreniyoruz. Bu modelin matematiksel bir ifadeyle anlamaya çalışırsak,

$h_t = f(h_{t-1}, z_{t-1}, a_{t-1})$ (1)

(1) Denkleminde t anındaki deterministik durum h_t, bu denklemi deterministik şekilde güncelleyen f fonksiyonunun; deterministik durumu anlatan h_{t-1} , stokastik durumları anlatan z_{t-1}, ve ajanın aldığı a_{t-1} aksiyonu ile t anındaki h durumunu modeller. Yani stokastik durumları da hesaba katan bir dünya modellemesi yapılmıştır.

$z_t \sim p(z_t|h_t,\ x_t)$ (2)

(2) Denkleminde t anındaki stokastik durumun yine t anındaki deterministik duruma ve gözleme (x_t) bağlı olduğunu anlıyoruz. Yani çevrenin "belirsiz" kısımları yine "belirli" kısımlara bağlıdır. Örnek verecek olursak bir simülasyonda, ormanda mantar görme olasılığımız(yani z) şehre göre daha fazladır ancak her iki durumda da mantar görme durumu rastgeledir.

3 Dreamer'ın Politika Öğrenimi

Ajanın belirli bir miktar iterasyondan sonra Dreamer, politikayı ($\pi_\phi(a_t|\ s_t)$; bir s durumunda hangi a aksiyonu alınmalıdır?) günceller. Ancak burada politika optimizasyonu, gerçek dünya ile etkileşime girerek yapılmaz, öğrenilen dünya modeli üzerinden gelecekteki olası senaryolar "hayal edilir", bu yapılırken actor-critic bir model kullanılır.

Actor, bir sinir ağı ile modellenir ve belirli bir durumda s_t hangi eylemin a_t seçileceğini belirleyen bir olasılık dağılımı üretir: $\pi(a_t \mid s_t)$. Critic ise Critic, başka bir sinir ağı ile durumu $s_t$bir skalar değere $V(s_t)$ eşler. Bu, o durumdan itibaren beklenen toplam ödülü temsil eder. Actor eylemleri seçer, critic bu eylemleri değerlendirir ve geri bildirim sağlar. (TD error, bu geri bildirimlerin bir türüdür ve genel olarak bu modellerde kullanılmaktadır.)

Dolayısıyla Dreamer'da politika öğreniminin iki aşamasından bahsedebiliriz. Birisi politika'nın kendisi ($\pi_\phi(a_t \mid s_t)$) diğeri değer fonksiyonu (ya da *value model* $v_\psi(S_t)$). Dreamer'da hedef fonksiyonun politika parametreleri (yani ϕ) açısından gradyan hesaplanarak politika güncellenir. Bu değer fonksiyonuna doğrudan geri yayılım uygulanarak yapılır. Bu pekiştirmeli öğrenme alanında kullanılan REINFORCE politika gradyanı yönteminden daha iyidir çünkü 'gürültü'yü (variance) azaltır.

4. TRANSDREAMER

Transformer'lar, geçmiş durumlara doğrudan ulaşarak, bunların arasındaki kompleks ilişkiyi öğrenebilir, böylece uzun vadeli bağımlılıklar gerektiren problemlerde başarı sağlarlar.

4.1 Transformer State Space Model (TSSM)

TSSM'nin tasarımında 3 tane beklentimiz var:

1. Geçmiş durumlara direk erişebilmeli.
2. Eğitim sırasında durumları(state) paralel olarak güncelleyebilmeli.
3. ajan gelecekteki adımları sırayla hayal edebilmeli.
4. Model, çevrenin belirsizliklerini (stokastik doğasını) yakalayabilmek için rastgele gizli değişkenler içermeli.

RSSM modelinde geçmiş durumlara doğrudan ulaşılmaz, bunlar $h_t$içinde sıkıştırılır ve dolaylı olarak ulaşılmış olunur. Her h_t durumu, h_{t-1} durumuna bağlıdır ($h_t = f(h_{t-1}, z_{t-1}, a_{t-1})$ denkleminden bunu biliyoruz.), bu da paralel güncellemeyi olanaklı kılmaz. Tam tersi TSSM'de geçmiş durumlara doğrudan ulaşma, paralel hesaplama ve dikkat mekanizması politika öğrenimini, "world model" için daha başarılı kılar.

Burada vurgulanması gereke şey, modelin sıralı değil bütün durumları paralel olarak çalıştırmasıdır. Ancak bunu yapabilmesinde RSSM için çelişkili bir durum mevcuttur: h_t'yi üretirken yine önceki h durumuna ihtiyacımız bulunmaktadır. Dolayısıyla durumların paralel işlenmesi mümkün değildir. Bunu gerçekleştirmek için TSSM'de miyopik representasyon modeli olarak adlandırılan bir yaklaşım kullanılmıştır. Yeni modelde $q(zt \mid ht, xt)$ yerine $zt \sim q(zt|xt)$ kullanılmıştır. Yani modelin h_t'ye olan bağımlılığı kaldırılarak bahsettiğimiz tutarsızlık giderilmiştir. Makalede yapılan testlerle ilgili verilerde orijinal Dreamer ile benzer metrikler verdiği gözlenmektedir.

Transformer modelleri ile RL entegrasyonundaki bir sorun, ödül dönütünün seyrekliğinin yol açtığı istikrarsızlıktır. Seyrek ödüller ajanın hangi eylemi doğru yaptığını anlamasını zorlaştırır. Bu sorun ajanın eğitimi sırasında transformer parametrelerinin sabit tutularak çözülmüş.

Ajanın eğitilmesi için kullanılan tek sinyal, ajanın aldığı aksiyonlar sonrasında aldığı ödüldür. *Prioritized Replay* mekanizması ajanın ödülleri parameterize ederek yüksek ödüllü olanları öncelikendirmesini sağlamaktadır.

Transformatörlerin yüksek bellek talepleri nedeniyle her imajinasyon *replay buffer'*a(geçmiş durumların saklandığı yer) eklenmesi mümkün değildir. Dolayısıyla her durumda imajinasyon gerçekleştirmek yerine daha küçük bir alt küme seçilir ve bu küme yapılacak imajinasyonların başlangıç durumu kabul edilir. Buna rağmen performansın orijinal Dreamer ile karşılaştırılabilir ya da daha iyi olduğu söylenmektedir.

5. Deneyler

Ajan, kısmen gözlemlenebilir ortamlarda belirli bir renk sırasına göre topları toplamalıdır. TransDreamer, hem 2D hem de 3D versiyonlarda Dreamer'a kıyasla daha yüksek başarı oranı ve ödül toplama performansı sergilemiştir. Örneğin, 4 toplu 2D görevde TransDreamer'ın başarı oranı %23 iken, Dreamer'ınki yalnızca %7 olmuştur. Top sayısı arttıkça (4 → 6) veya toplar arası mesafe uzadıkça, TransDreamer'ın performansı Dreamer'a göre daha az düşüş göstermiştir. Bu, transformer tabanlı modelin uzun vadeli bağımlılıkları daha iyi yakaladığını kanıtlamaktadır.

TransDreamer'ın dünya modeli (TSSM), gelecek görüntüleri ve ödülleri Dreamer'ın modeline (RSSM) kıyasla daha doğru tahmin etmiştir. Özellikle, uzun bağlam pencereleri (örn. 80 adım) verildiğinde, ödül tahmin doğruluğu belirgin şekilde artmıştır. TransDreamer'ın "hayal ettiği" sahnelerde topların renkleri ve konumları daha netken, Dreamer bulanık ve hatalı tahminler üretmiştir.

TransDreamer, Dreamer ile benzer nihai performansa ulaşmış ancak öğrenme hızı daha yavaş olmuştur. Örneğin, "Cheetah Run" görevinde TransDreamer daha hızlı yakınsama göstermiştir. Pong gibi oyunlarda ise iki model de benzer skorlar elde etmiş, ancak TransDreamer'ın dünya modelinin daha stabil olduğu gözlemlenmiştir.

6. Sonuç

Transformer'ların paralel işleme ve dikkat mekanizması, dinamiklerin modellenmesinde RNN'lerden daha etkili olmuştur ve uzun vadeli bellek ve karmaşık nedensellik gerektiren görevlerde Dreamer'ı açık ara geride bırakmıştır. Özetle, TransDreamer, model tabanlı pekiştirmeli öğrenmede transformer'ların potansiyelini ortaya koyarak, özellikle bellek odaklı görevlerde yeni bir standart belirlemiştir.