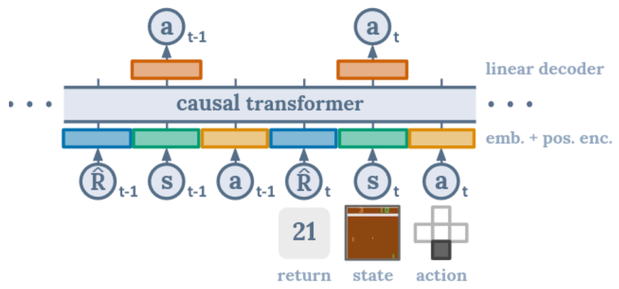
DPT: Pretrained Decision Transformers

المرة لي دازت هضرنا على العاب و كيفاش عووعلي قدرات تخلي المستعمل يخلق لعبة ديالو من صورة او رسمة، دابا تخيل وأنت كتدرب برنامج حاسوبي باش يلعب مثلا لعبة فيديو، لكن بدل يد التحكم او clavier، كتستعمل وكيل تعلم التعزيز (Reinforcement Learning) لي كيستعمل أحدث التقنيات. الوكلاء التقليدية لتعلم التعزيز كيتعلمو عن طريق التجربة والخطأ، كيبحثو في البيئة ويتلقاو المكافآت على الأفعال الجيدة (التعزيز الإيجابي) والعقوبات على الأفعال السيئة (التعزيز السلبي). هاد النهج يكون بطيء وكيطلب الكثير من البيانات او data، خاصة مع الألعاب المعقدة لي فيها فضاءات حالة عالية الأبعاد.

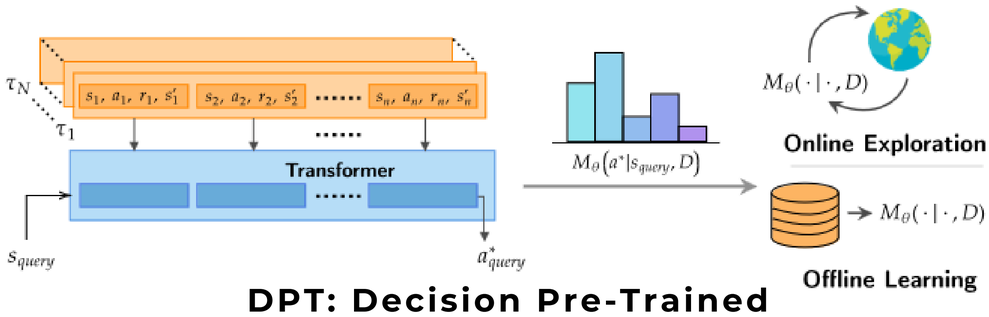
بوجود التقنيات الجديدة، كيوجد حلا ممكنا. في السنوات الأخيرة، عرفنا تقدما هائلا في مجال التعلم الآلي، خاصة في مجال التعلم بالتعزيز(Reinforcement Learning). ما يميز النهج الجديد، المسمى بـ "Decision-Pretrained Transformer (DPT)"، هو قدرته على تطوير استراتيجيات قرار متعمقة بدون الحاجة إلى كميات ضخمة من البيانات. هاد الشي يعني أن البرنامج قادر على تعلم اللعبة بسرعة أكبر ودقة أكثر، حتى مع الألعاب المعقدة والمتطلبات العالية

هاد النهج الجديد لي سميتو (DPT). DPT كيستفاد من قوة supervised learning وTransformers،هو نوع معيّن من Neural Networks. خلال **التّدريب** **المُشرف**(supervised) نموذج DPT كيتعرض لمجموعة ضخمة من البيانات : الحالة-العمل-المكافأة(State, Action, Reward) من مهام RL متنوعة. هاد لتنوّع في المجموعة كيسمح للنموذج يتعلم استراتيجيات و قرارات قابلة للتّعميم على بيئات مختلفة.



Decision Transformers

- الا بغينا ندويو عل كل مفهوم بوحدو كاين:
  - التدريب المُسبق Pretraining: الهندسة المعمارية لTransformer ضمن DPT تعالج هذه الثلاثيات الحالة-العمل-المكافأة بكفاءة. هذا التدريب المسبق يعطي النموذج فهما أساسيا للهياكل المكافأة واختيار الإجراءات الأمثل عبر مختلف السيناريوهات.
  - المحوّلات القرارية (Decision Transformers): على عكس الأساليب التقليدية لتعلم التعزيز اللي كيتركزو بشكل حصري على الحالة الحالية (على سبيل المثال، موقع اللاعب، مواقع الأعداء)، DPT كيستعمل Decision Transformers. هاد النوع من Transformers مصمم خصيصا باش يأخذ بعين الاعتبار ليس فقط الحالة الحالية بل كذلك تسلسل الحالات والإجراءات السابقة (السياق Context). هاد المعلومات السياقية مهمة لاتخاذ القرارات في المستقبل.



DPT: Decision Pre-Trained Transformer

هاد الطريقة كتحاول تقويم أداء النموذج على أساس قدرته على اتخاذ القرارات الصحيحة في سياقات جديدة ومتغيرة. بمعنى آخر، DPT ماكتكتفيش بمجرد تطبيق ما دربات عليه، بل كتعتمد على هاد **transformers** لتوجيه القرارات الأمثل في الظروف الجديدة. هاد الشي كيجعل النموذج أكثر قابلية للتكيف والتطبيق في مجموعة متنوعة من الألعاب والمهام، مما كيزيد من أدائه وفاعليته.

المفهوم الأساسي هنا هو **السياق** (Context)او In-Context Learning ICL. السياق يشير لي اللي حصل قبل فاللعبة – الحركات لي درت، والعراقيل لي واجهت، والمكافآت لي حصلت عليها. بواسطة تذكر هادشي، DPT قادر يحسن من قراراتو فالوضعية الحالية.
على سبيل المثال، إذا كنت مطارد من طرف عدو في اطار لعبة، DPT قادر يستغل معرفتو باستراتيجيات الهروب (تعلمها من ألعاب أخرى) باش يتجنب الخطر.

هاد التجميع ديال التدريب المسبق، Decision Transformers، والوعي ب Context كيمكن لـ DPT يتعلم الألعاب الجديدة بشكل أسرع بزاف من الوكلاء التقليديين لRL. هو ماكيحتاجش لتمرين كثير وكيقدر يتكيف استراتيجيتو بناء على اللعبة.

- التكيف في السياقIn Context Adaptation: منين كيواجه مهمة جديدة في التعلم بالتعزيز، DPT يستفيد من معرفته المسبقة والسياق (التفاعلات السابقة) باش يتنبأ بالإجراء الأمثل في تلك الوضعية المعينة. هذا يسمح لـ DPT بالتكيف بسرعة مع البيئات غير المعروفة وتحقيق كفاءة الحاجة لي كتطلب كمية أقل من data مقارنة بالأساليب التقليدية ديال RL.

نظريا تتعمق الدراسة في الخصائص التقنية لDPT عبر استكشاف العلاقة بينه وبين Posterior Sampling، وهي طريقة معروفة بكفاءتها في Reinforcement Learning من زمان. Posterior Sampling ترتبط بقدرة النموذج على استنتاج توزيعات Distributions الاحتمالات الممكنة للنتائج المستقبلية بناءً على المعرفة السابقة والبيانات المتاحة. يعتبر هذا الاتصال مهما بزاف، حيث كيوفر لنموذج DPT واحد أساسا نظري لقدرته على Adaptation مع البيئات المختلفة واستخدام البيانات بكفاءة أكبر.
هاد الارتباط بين DPT وعمليات Baysian Framework كيساهم في توجيه البحث ل فهم أعمق لكيفية استعمال البيانات بكفاءة فعمليات التعلم، وكيفاش يمكن نحسنو أداء النماذج Models في نفس الوقت. هاد الاتصال كيوجه الابتكارات التقنية للي تطور نماذج أقوى وأكثر كفاءة في ميدان التعلم الآلي وRL.

واحد ملاحظة من الضروري الاعتراف بأن DPT هو مجال بحث نسبيا جديد حيت عاد خرج البحث العام لي داز، ومن الضروري إجراء مزيد من البحوث لفهم فعاليته بشكل كامل عبر مشاكل ديال Reinforcement Learning.

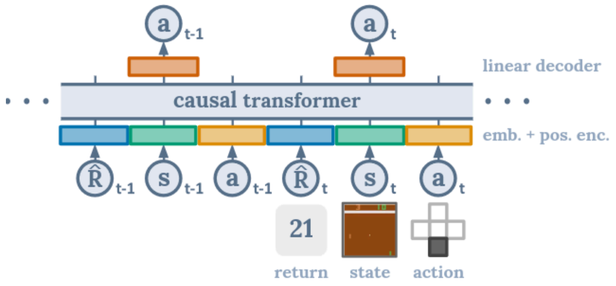
References:
- Decision Transformer: Reinforcement Learning via Sequence Modeling: https://arxiv.org/pdf/2106.01345
- Supervised Pretraining Can Learn In-Context Reinforcement Learning: https://arxiv.org/pdf/2306.14892

DPT: Pretrained Decision Transformers

تخيل أنك تقوم بتدريب برنامج كمبيوتر لتشغيل لعبة فيديو، ولكن بدلاً من وحدة التحكم، فإنك تستخدم عامل التعلم المعزز (RL) المتطور. يتعلم عملاء RL عادة عن طريق التجربة والخطأ، واستكشاف البيئة والحصول على مكافآت على الأفعال الجيدة (التعزيز الإيجابي) وعقوبات على الأفعال السيئة (التعزيز السلبي). يمكن أن يكون هذا الأسلوب بطيئًا ومكثفًا للبيانات، خاصة بالنسبة للألعاب المعقدة ذات مساحات عالية الأبعاد.

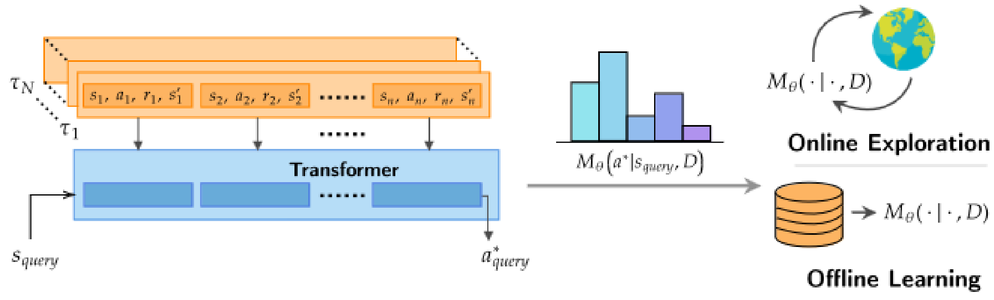
بفضل التقنيات الجديدة، يمكننا الآن أن نجد حلاً ممكناً. في السنوات الأخيرة، شهدنا تقدماً هائلاً في مجال التعلم الآلي، خاصةً في مجال التعلم بالتعزيز. ما يميز النهج الجديد، المعروف بـ "Decision-Pretrained Transformer (DPT)"، هو قدرته على تطوير استراتيجيات قرار عميقة بدون الحاجة إلى كميات ضخمة من البيانات. هذا يعني أن البرنامج قادر على تعلم اللعبة بشكل أسرع وأدق، حتى مع الألعاب المعقدة والمتطلبات العالية.

هذا النهج الجديد الذي يُعرف باسم DPT يستفيد من قوة التعلم بالإشراف (supervised learning) والمحولات (Transformers)، وهو نوع معين من الشبكات العصبية (Neural Networks). خلال عملية التدريب بالإشراف، يتعرض نموذج DPT لمجموعة ضخمة من البيانات المتمثلة في الحالة، العمل، والمكافأة (State, Action, Reward) من مهام التعلم بالتعزيز المتنوعة. هذا التنوع في مجموعة البيانات يسمح للنموذج بتعلم استراتيجيات وقرارات يمكن تعميمها على بيئات مختلفة.



Decision Transformers

- إذا أردنا التفصيل في كل مفهوم على حدة:
  - **التدريب المُسبق (Pretraining)**:هندسة المحولات (Transformers) ضمن نموذج DPT تمكن من معاجة هذه الثلاثيات الحالة-العمل-المكافأة (State-Action-Reward) بكفاءة. يعطي هذا التدريب المُسبق النموذج فهمًا أساسيًا لهياكل المكافأة واختيار الإجراءات الأمثل عبر مختلف السيناريوهات.
  - **المحوّلات القرارية (Decision Transformers)**: على عكس الأساليب التقليدية لتعلم التعزيز التي تركز بشكل حصري على الحالة الحالية (على سبيل المثال، موقع اللاعب، مواقع الأعداء)، يستخدم DPT المحوّلات القرارية. هذا النوع من المحوّلات مصمم خصيصًا ليأخذ بعين الاعتبار ليس فقط الحالة الحالية بل أيضًا تسلسل الحالات والإجراءات السابقة (السياق). هذه المعلومات السياقية مهمة لاتخاذ القرارات في المستقبل



DPT: Decision Pre-Trained Transformers

هذه الطريقة تحاول تقييم أداء النموذج استنادًا إلى قدرته على اتخاذ القرارات الصحيحة في سياقات جديدة ومتغيرة. بمعنى آخر، فإن DPT لا يكتفي بمجرد تطبيق ما تم تدريبه عليه، بل يعتمد على هذه المحوّلات القرارية لتوجيه القرارات الأمثل في الظروف الجديدة.
هذا الأسلوب يجعل النموذج أكثر قابلية للتكيف المرونة في مجموعة متنوعة من الألعاب والمهام، مما يزيد من أدائه وفاعليته. فيفضل استخدام المحوّلات القرارية، يمكن لـ DPT تحديث وتعديل استراتيجياته واتخاذ القرارات الأمثل بناءً على السياقات الجديدة التي يواجهها، مما يعزز قدرته على التكيف وتحسين أدائه في بيئات مختلفة ومتغيرة.
المفهوم الأساسي هنا هو السياق (Context) أو التعلم في السياق (In-Context Learning - ICL). يشير السياق إلى كل ما حدث مسبقًا في اللعبة، مثل الحركات التي قمت بها، والعقبات التي واجهتها، والمكافآت التي حصلت عليها. من خلال تذكر هذه الأحداث، يمكن لـ DPT تحسين قراراته في الوضعية الحالية.
على سبيل المثال، إذا كنت تُطارَد من قِبَل عدو، يمكن لـ DPT استخدام معرفته باستراتيجيات الهروب التي تعلمها من ألعاب أخرى لتجنب الخطر.
هذا التجميع للتدريب المسبق، المحوّلات القرارية، والوعي بالسياق يمكن أن يساعد DPT على تعلم الألعاب الجديدة بشكل أسرع بكثير من الوكلاء التقليديين لتعلم التعزيز. فعندما يتم تدريب DPT بشكل جيد مسبقًا، يكون لديه فهم أساسي للهياكل والأنماط المشتركة في ألعاب التعلم بالتعزيز.

تكييف السياق In Context Adaptation: عندما يواجه نموذج DPT مهمة جديدة في التعلم بالتعزيز، يستفيد من معرفته المسبقة والسياق (التفاعلات السابقة) ليتنبأ بالإجراء الأمثل في تلك الحالة المحددة. يتيح ذلك لـ DPT التكيف بسرعة مع البيئات غير المألوفة وتحقيق كفاءة باستخدام كمية أقل من البيانات مقارنة بالأساليب التقليدية في تعلم التعزيز.

من الناحية النظرية، يشمل التعمق في الدراسة للخصائص التقنية لـ **DPT** استكشاف العلاقة بينه وبين **Posterior Sampling**، وهي طريقة معروفة بكفاءتها في التعلم بالتعزيز منذ فترة طويلة. ترتبط Posterior Sampling بقدرة النموذج على استنتاج توزيعات احتمالية لنتائج المستقبل بناءً على المعرفة السابقة والبيانات المتاحة. يُعتبر هذا الارتباط مهمًا لأنه يوفر لـ DPT أساسًا نظريًا لتكيفه مع بيئات متنوعة واستخدام البيانات بكفاءة أكبر.
يسهم هذا الاتصال بين DPT و Bayesian posterior sampling في توجيه البحث نحو فهم أعمق لاستخدام البيانات بكفاءة في عمليات التعلم، وكيفية تحسين أداء النماذج في نفس الوقت. يوجه هذا الارتباط الابتكارات التقنية التي تُطوِّر نماذجًا أقوى وأكثر كفاءة في مجالات التعلم الآلي والتعلم بالتعزيز(RL).

المصادر :

- Decision Transformer: Reinforcement Learning via Sequence Modeling: https://arxiv.org/pdf/2106.01345
- Supervised Pretraining Can Learn In-Context Reinforcement Learning: https://arxiv.org/pdf/2306.14892

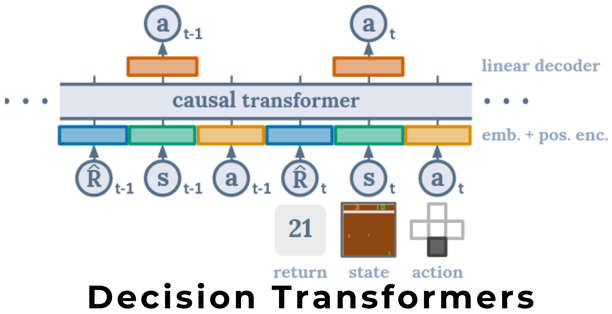
Lase9 f l'AI

30st April 2024

DPT: Pretrained Decision Transformers

Imagine you're training a computer program to play a video game, but instead of a controller, you're using a cutting-edge reinforcement learning (RL) agent. Traditional RL agents learn by trial and error, exploring the environment and receiving rewards for good actions (positive reinforcement) and penalties for bad ones (negative reinforcement). This approach can be slow and data-intensive, especially for complex games with high-dimensional state spaces.

Here's where a new approach called Decision-Pretrained Transformer (DPT) comes in. DPT leverages the power of **supervised pre-training** and transformers, a specific type of neural network architecture. During pre-training, a DPT model is exposed to a vast dataset of state-action-reward triplets from diverse RL tasks. This diverse dataset allows the model to learn generalizable decision-making strategies across different environments.

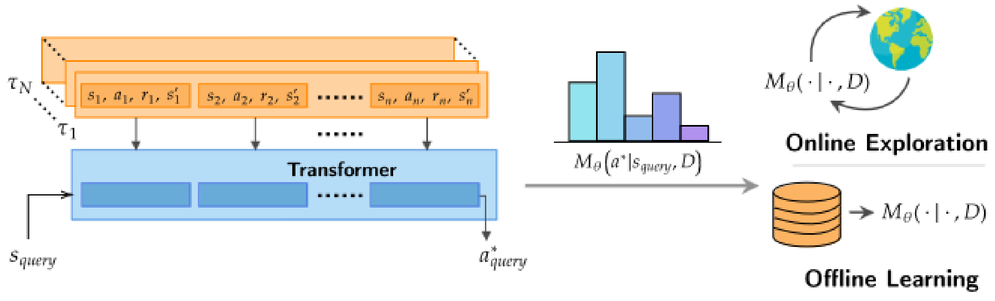


Decision Transformers

Technical Deep Dive:

- Pretraining: The transformer architecture within DPT efficiently processes these state-action-reward triplets. This pre-training essentially imbues the model with a foundational understanding of reward structures and optimal action selection across various scenarios.
- Decision Transformers: Unlike traditional RL methods that focus solely on the current state (e.g., player position, enemy locations), DPT utilizes decision transformers. These transformers are specifically designed to consider not just the current state but also the sequence of past states and actions (context). This contextual information is crucial for informed decision-making.

DPT: Decision Pre-Trained Transformer



Now, when faced with a new game, DPT doesn't have to start from scratch. It uses its pre-trained knowledge to grasp the basic rules and goals. But DPT has another trick up its sleeve: **decision transformers**. These are special AI models specifically designed for games or similar tasks where making the right choice is crucial. Imagine a decision transformer as the part of the student's brain that analyzes the situation, remembers past experiences, and picks the best move.

The key concept here is **context**. Context refers to what happened earlier in the game – the moves you've made, the obstacles you've faced, and the rewards you've earned. By remembering this history, DPT can make smarter decisions in the current situation.

For example, if you've been collecting coins throughout the game, DPT might prioritize reaching another coin area based on the context. On the other hand, if you're being chased by an enemy, DPT can leverage its knowledge of escape strategies (learned from other games) to navigate away from danger.

- In-Context Adaptation: When presented with a new RL task, DPT leverages its pre-trained knowledge and the provided context (past interactions) to predict the optimal action in that specific situation. This allows DPT to adapt quickly to unseen environments and potentially achieve sample efficiency, requiring less data compared to traditional RL methods.

Delving more on the technical properties, the paper introducing DPT also explores the theoretical foundation by demonstrating connections to Bayesian posterior sampling, a method known for its sample efficiency in RL. This connection provides a theoretical justification for the potential efficiency gains of DPT.
Overall, DPT offers a promising approach to in-context RL by combining supervised pre-training with decision transformers. This approach has the potential to significantly improve the efficiency and adaptability of RL agents in various tasks.

Note: It's important to acknowledge that DPT is a relatively new research area, and further investigation is needed to fully understand its effectiveness across different RL problems.

References:

- Decision Transformer: Reinforcement Learning via Sequence Modeling: https://arxiv.org/pdf/2106.01345
- Supervised Pretraining Can Learn In-Context Reinforcement Learning: https://arxiv.org/pdf/2306.14892