 深層強化学習の動向 | 論文PickUp

はじめに

機械学習の動向のうち，基本的には強化学習(RL(*Reinforcement Learning*))を対象にここ半年程度のトピックをまとめる．

新技術

OpenAIから文書から画像を生成するAIの第2世代発表 : DALL·E 2

https://openai.com/dall-e-2/

- 強化学習ではなく，教師なし機械学習の一種，敵対的生成ネットワーク(GAN)の応用であるが技術的発展性があると思うので紹介．正直技術的な詳細は把握できておらず，ベンチマーク的な意味合いで取り上げる．
- 事例は上記URLから紹介動画を参照するのが良いと思うが，2021に発表のあった初代DALL・Eの進化版の理解で自然言語の意味を元に，その意味に沿った現実には存在しない画像を生成するというもの．

https://cdn.openai.com/papers/dall-e-2.pdf

Googleの汎用AI : PaLM(Pathways Language Model)

- こちらも強化学習の話題ではない．まだ全容が公開されている訳ではなさそうだが，22/04/04に(PaLM)Pathways Language Modelを発表している．
- 2020年にOpenAIの自然言語処理(NLP)文書作成AIであるGPT-3が1750億パラメータという膨大なパラメータを半ば力技で学習させて構築されたという発表があったと記憶している．自然言語処理モデル「GPT-3」の紹介
- 蛇足だが，GPTについてはOpenAIからライセンス供与を受けているMicrosoftからスピンアウトしたAIベンチャーであるRINNA社が日本語モデルを作成 している．マイナーな言語である日本語は機械学習の取り組み事例が少なく，英語/中国語と比較して発展途上と聞く．
- 本題であるGoogleの"Pathway"についても技術的な理解はできていない．特徴として，これまでのものモーダルなニューラルネットワークから，マルチモーダルに取り組んだものである，と聞き及ぶ．
- ここでのマルチモーダルの意味について補足すると，これまでのニューラルネットはひとつのタスクしか学習できなかったがこれを複数タスクをひとつのモデルとして学習できるようにしたもの理解する．
- 前述のGPT-3を上回る5400億パラメータで学習させることで現在のSOTA(State-of-the-art)の座に居るらしい．

応用研究

今年の2月のネイチャー誌に掲載された論文で面白いものがあるので下記紹介.

強化学習を用いたトカマク型核融合炉の制御

(doi: 10.1038/s41586-021-04301-9)

- Google(alphabet)傘下の英DeepMind社とのEPFLスイスプラズマセンターが発表した論文で，トカマク型核融合炉の核融合反応に伴うプラズマを高磁場内に閉じ込める制御を，深層強化学習を用いて実施したという研究.
- トピックとしては強化学習としては一般的ないわゆるActor-Critic法によるエージェントを用いて，比較的高速(10KHz)な制御周期でプラズマ抑制のための磁場制御を行っているという内容. これについては，以前にPMSMモータの電流制御を対象に机上で同じ構成で試したことがあり，逆説的にトライした方法論の裏付けとなると感想を持った.
- また，学習のアプローチとしても実機(核融合炉)を用いた学習ではなく(当たり前か)，完全にオフラインのモデルベースでの学習結果をゼロショット転移，すなわち修正なしで実機の制御に移行して機能することを確認したと述べられている. モデルベースシミュレーションの精度が高く，現実世界で生じる様々な誤差を含めて学習しておかないと容易に実現できるものではないと推測するので単純にすごいと思う.

- なお，バックアップとして問題があれば従来制御に切り替えるとあり，従来制御については下記の様に記述されている．

> **この時変・非線形・多変量制御問題に対する従来のアプローチは、まず逆問題を解いてフィードフォワードコイルの電流と電圧のセットをあらかじめ計算することである。次に、プラズマの垂直位置を安定化させ、半径方向の位置とプラズマ電流を制御するために、独立した1入力1出力のPIDコントローラのセットが設計される**が、これらはすべて、相互に干渉しないように設計されなければならない。ほとんどの制御アーキテクチャは、プラズマ形状の外部制御ループによってさらに強化され、フィードフォワードコイル電流を変調するためにプラズマ平衡のリアルタイム推定を実装することが含まれる。制御器は線形化されたモデルダイナミクスに基づいて設計され、時間的に変化する制御目標に追従するためにゲインスケジューリングが必要である。これらのコントローラは通常効果的であるが、目標とするプラズマ形状が変化するたびに、平衡推定のための複雑なリアルタイム計算とともに、かなりの工学的努力、設計努力、専門知識を必要とする。

- 上記のふるまいを強化学習によって得ていると理解できる．
- また，従来の制御では困難であった，プラズマを任意の形状に形成維持するということも報酬を変更することで対応できたという報告もある．
- PMSMの制御に強化学習を適応した際に，強化学習が勝手に弱め磁束制御を発見・習得していったことと似たふるまいで非常に興味深いと思うところであった．

💡 Degrave, J., Felici, F., Buchli, J. *et al.* Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602,** 414–419 (2022).

Neural networks overtake humans in *Gran Turismo* racing game

(doi: 10.1038/s41586-021-04357-7)

- 機械学習専門のニュース以外でも話題になったので，概要は知られているかもしれないが，表題の通りPlayStationのメジャーなタイトルであるグランツーリスモというレースシミュレータでの車両ドライバを強化学習エージェントが行い，eスポーツの世界大会(グランツーリスモが日本の国体の競技になっているのをここで初めて知った!)ランカーと複数同時に走行して勝利したという内容かと思う．
- 深層強化学習のアルゴリズムはどうやら本稿のオリジナルで，QR-SAC(quantile regression soft actor-critic)法とある．全くコースを走れないような初期の状態から数時間の学習(実機で行っているので実時間)でコースを周回できるようになり，9日でほぼすべての人間ドライバーより早いラップタイムを記録した，と述べられている．
- 得られた特性の評価として論文内の以下の記述が興味深かい．

> “レースエージェントの開発にディープRLを用いる利点は、エンジニアがレースで勝つために必要なスキルをいつどのように実行するかをプログラムする必要がなく、適切な条件にさらされさえすれば、エージェントは試行錯誤によって正しい行動を学習することである。GT Sophyは、コーナーでの追い越し、スリップストリームの有効利用、後続車のドラフトの乱れ、ブロック、緊急時の対応など、様々なタイプの操作を行うことができることが確認できました。”

- 最終的に単独走行のラップタイムに加え，トップドライバとの混合走行でも勝利したとある．しかしながら，ラインのブロックやスリップストリームの使い方など戦略面では少し単純なふるまいをした，と結論されている．
- 所感として面白いと感じるのは，単純に予測できない人間ドライバの戦略的な走行を連続的に判断してより良い行動を取る制御を学習することができたというところ．本稿の要旨とは異なるが，現実世界で現実車両を用いた自律走行レースは行われているがまだ目立った結果を残せていないと記憶する．仮想世界でも，人間のトップドライバとエージェントが直接干渉することで得られる学習量と質がこの結果をもたらしたということだと思われ，自動運転の一般化へのアプローチへの大きなヒントとなると感じた．

💡 Wurman, P.R., Barrett, S., Kawamoto, K. *et al.* Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602,** 223–228 (2022).

自動運転分野(AutonomusVehicles)

カルパシーさんがしばし休養

- OpenAIの創設者で現在Teslaのディープラーニング部門を統括しているAndrej Karpathy氏，Tesla自前のディープラーニング用スーパーコンピュータ郡であるDojoが稼働するまでしばらく休養，というニュース.
- 断片的な情報として，同社の Full Self Driving(beta!!)のニューラルネットワークプログラム(スタックと呼ばれる)がこれまで用途ごとに個別にあったものを統合する大規模な書き換えが終わり，大規模な学習を行う準備中と聞く．大きく性能向上する見込みと例のTWTRおじさんは言っているが...
- また，カルパシーさんが戻ってきたら同じ様な着想でOptimusに基礎教育をするらしい.