解説2
VLAモデル・ローカルLLM・シミュレーションツールのロボット開発への適用方法とヒューマノイドSIer戦略

想造技研　滝沢　一博*
*たきざわ　かずひろ：代表取締役

ロボット開発はこれまで、環境認識、動作計画、制御系の三層を個別に最適化し、それらを統合することで進化してきた。しかし、近年のAI技術の飛躍的進展により、視覚・言語・アクションを統一的に扱うVision–Language–Action（VLA）モデルが登場し、ロボットのタスク遂行能力は一段と向上している。そして、エッジAI技術の発展に伴い、従来クラウド依存だった大規模言語モデル（LLM）をローカルLLMとして産業現場へ展開し、高速かつ安全に制御を実現する動きも加速している。

加えて、NVIDIAのIsaac Sim、Isaac Labといったシミュレーションツールは物理シミュレーションだけでなく、それを利用した強化学習も実施することができ、従来のティーチングとは違うロボットの動作会得手法の確立やsim–to–realを活用することによる業務効率化につながることを期待されている。

本稿の前半では、RT–2（Robotic Transformer–2）をはじめとするVLAモデル、ローカルLLM、Isaac sim、lsaac Labとそれらの実ロボットへの適用について紹介する。後半では、世界的に著しい中国のヒューマノイドの紹介、そして、現在ヒューマノイドで後れを取る日本が採るべき戦略として、既存プラットフォームを活用した「ヒューマノイドSIer戦略」の可能性とそれを実現するために必要なスキルを提案する。

RT–2モデルの革新性

RT–2は、視覚（Vision）、言語（Language）、動作（Action）の三要素を一体的に学習・推論できる点で従来手法を大きく凌駕する。大規模Webデータとロボット操作データの両者を用いた二段階事前学習を行うことで、多様な環境・タスクへの一般化性能を獲得している。RT–2は最終的に動作トークン（Action Token）列を生成し、それはロボットの関節角度やツール操作指令として扱われる。

RT–2の革新性はVLAモデルを構築したことにある。従来のロボット制御モデルは、ビジョンプランニングや言語理解を個別に実装し統合する必要があったのに対し、RT–2のアーキテクチャでは単一のTransformerで統合的に処理可能である。その結果、未知環境での少数ショットタスク（Few–shot task）においても、最小限の追加学習で適応できる柔軟性を示した。

また、先行研究ではロボット固有の動作空間を事前定義していたが、RT–2はインターネットスケールの多種多様な動作記述を取り込むことにより、動作空間の自己拡張を可能にした。

RT–2はすでに複数のロボットプラットフォーム（ユニバーサルロボット、PR2、Spotなど）で実装されつつあり、Isaac Sim上でも動作検証が進められており、今後は、より高次のタスクへ適用範囲を拡大し、意思決定やタスクプランニングと統合されていくと予測される。

VLAモデルの動作生成への適用

VLAモデルは、視覚情報と自然言語指示を統合し、ロボットの具体的な動作計画へと変換するアーキテクチャとして注目を集めている。本章では、OpenVLAを中心にVLAモデルの技術的特徴、動

作トークン化の枠組み、および実世界での適用事例と有効性評価について詳述する。

1. OpenVLAとその技術的特徴

OpenVLAは、Llama2など既存LLMの言語理解能力と、CLIPベースの視覚エンコーダを組み合わせたオープンソースのVLAモデルである。特徴的なのは以下の点である。

(1)二重エンコーダ構成

視覚エンコーダは画像フレームを埋め込みベクトルへマッピングし、言語エンコーダは指示文をトークン化して自然言語表現を生成。それぞれのベクトルをTransformer内部で統合する。

(2)動作トークン化

ロボット操作を離散的な「動作トークン」に符号化し、言語トークン列と同等に扱うことで、自然言語処理の手法を動作生成に応用可能とした。

(3)マルチタスク学習

ピッキングやドア開閉、ナビゲーションといった複数タスクを同一モデルで学習し、タスク識別やゼロショット適応性を向上させている。

これらにより、OpenVLAは容易に他モデルやハードウェアに移植できる一方、動作精度も高く保たれている点が評価されている。

2. 動作トークン化による統一フレームワーク

動作トークン化は、連続的な関節角度系列を離散表現に変換する手法である。典型的には、以下のプロセスを経る。

(1)デモデータ収集

人手または既存制御プログラムで取得した関節角度・ツール操作シーケンスを集める。

(2)符号化モデル学習

VQ-VAEや類似の離散化モデルで、連続シーケンスを有限数のトークンに圧縮。

(3)トークン列生成

得られたトークン列が「動作言語」となり、自然言語と同様にTransformerへ入力できる。

この枠組みにより、異なるロボット種（マニピュレータ、二足歩行、四足歩行）の動作も共通フォーマットで扱え、モデルの汎用性と再利用性が大きく向上する。さらに、動作トークンにはシンタックス(速度、位置調整、グリッパー開閉など)情報を追加でき、低レベル制御から高次タスクまでカバーする階層的表現が可能となる。

3. 実世界適用での有効性評価

VLAモデルを使ったロボット操作は以下のような観点において使用価値が高まると考えられる。

(1)自然言語による直感的操作

VLAモデルはLLMを活用し、「棚から赤い箱を取って」など口語的な指示でも操作できる。プログラミングスキルは不要で、作業者・非専門家でもロボットを効率的に扱うことが可能となる。

(2)未知の物体・環境への適用

従来、未知環境には特別な学習やプログラムが必要だったが、VLAモデルでは多様なインターネットを介した最新学習モデルの常時活用、学習物体の抽象概念の学習などにより、未知のオブジェクトやタスクにも柔軟に対処できる可能性がある

(3)視覚・動作・言語の一体化

カメラ映像と作業指示を同時に理解し、人間の意図に沿って柔軟に行動するため、複雑で状況依存性の高い作業もこなせる可能性がある。

(4)効率性と学習コストの削減

VLAモデルは事前に大規模データで学習しているため、現場での追加学習や試行錯誤が圧倒的に少ない。例えばVLAモデル「CogVLA」は従来法よりも学習コストを約2.5倍削減し、推論時間も約2.8倍短縮できると発表している。

ローカルLLMとエッジコンピューティング

従来、多くのLLMは数十億～数兆パラメータを持ち、データセンターのGPUクラスタでの推論を前提としていた。しかし、産業現場やサービス現場では次の課題が顕在化する。

人と同等の反応速度で指示を解釈し動作するには、クラウド往復の通信遅延(往復100～200 ms)がロボットの動作のスムーズさに影響を与える場合がある。常時高頻度で映像・センサ情報をクラウドに送信し続けるのは帯域幅や運用コストを圧迫し、電波環境が不安定な現場では信頼性が低下する。また、機密性の高い現場情報や顧客データを外部へ送信できないケースが多く、そういった