AMBER、ROSBAG と PYTORCHで はじめるお手軽マルチモーダルロボット開発

2023/06/29 MASAYA KATAOKA

自己紹介

もともとVRがやりたくてプログラミングを始める

Oculus DK2 x Unityで遊戯王のゲームを作ろうとしていました

趣味はロボット開発

学生時代から海洋ロボットの開発、RoboCupに参加

学生時代の専門はソフトロボティクス

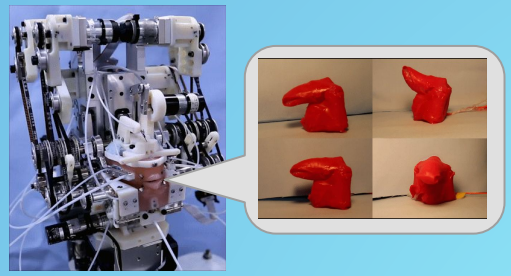
発話ロボットの舌機構制御の研究をしてました

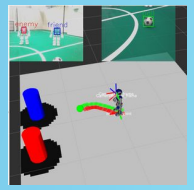
お仕事は自動運転

世界中で自動運転するためのインフラ開発をしてます

rosjpオーガナイザー

日本全国でROS関連のイベントを開いています！是非connpassからご参加ください





ロボットとマルチモーダル

ほとんどのロボットがマルチモーダル

ロボットはセンサの弱点を補い合うために複数種類のセンサを統合する

高度なタスク遂行にはマルチモーダル必須

マルチモーダルなロボットの典型例であるカチャカは障害物センサだけで3種類のセンサを搭載し、その結果を統合している

今後、ロボット技術の発展やセンサの量産による値段低下に伴いマルチモーダル化はより大きく加速すると予想される

カチャカセンサ配置図[Preferred Robotics, 2023]

マルチモーダルデータ統合の困難さ

観測の信頼度などは数値化が難しい

センサの信頼度（ex:カメラ認識結果はLiDARより優先など）を固定の順位で決め打つとエッジケースを引いたときに失敗する

うまく設計しない限り、センサを統合したことで逆に性能が劣化することも

確率ロボティクスという分野は存在するが、動的環境への対応などにはまだまだ超えるべき課題も多い

特にモダリティをまたぐとより困難に

自己位置推定で使われるカルマンフィルターは、複数のセンサ情報を正規分布という形に近似してモダリティをまたぐ

新たなモダリティ統合アルゴリズムとしての
マルチモーダルなMLモデルの登場

近藤豊さんのSNS投稿より

マルチモーダルな機械学習モデル

複数のモダリティをまたいで推論

画像 + 言語 => CLIPなど

画像 + 動画 + 言語 + 音声 => NExT-GPT

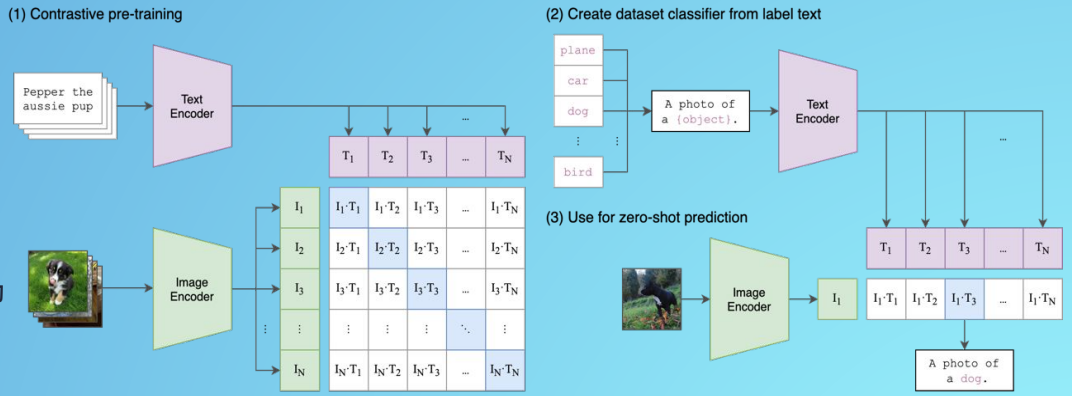
違うモダリティを共通の表現にできるものも

例えばCLIPでは、画像、テキストを512次元のベクトルに変換することが可能

このような表現を持つモデルを利用すれば自然言語の入力を物体認識に利用したりすることが可能

基盤モデル等の研究により近年大きく進展

基盤モデルとは「大量で多様なデータを用いて訓練され、様々なタスクに適応（ファインチューニングなど）できる大規模モデル」のこと



CLIPのアーキテクチャ[Open AI, 2021]

マルチモーダルな機械学習モデル

複数のモダリティをまたいで推論

画像 + 言語 => CLIPなど

画像 + 動画 + 言語 + 音声 => NExT-GPT

違うモダリティを共通の表現にできるものも

例えばCLIPでは、画像、テキストを512次元のベクトルに変換することが可能

このような表現を持つモデルを利用すれば自然言語の入力を物体認識に利用したりすることが可能

基盤モデル等の研究により近年大きく進展

基盤モデルとは「大量で多様なデータを用いて訓練され、様々なタスクに適応（ファインチューニングなど）できる大規模モデル」のこと

DALL-E 3 [Open AI, 2023]

ロボティクスへの応用 (RT-2)

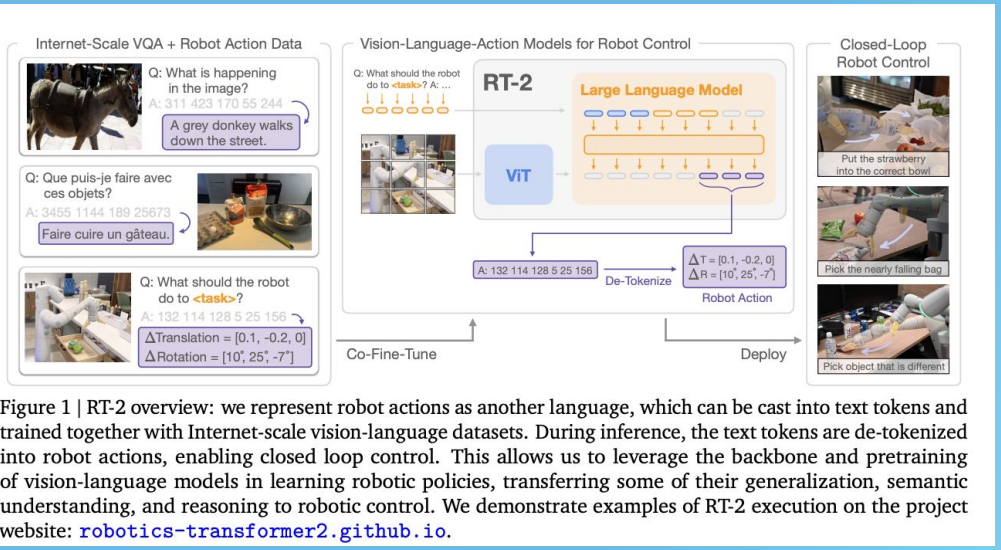


Figure 1 | RT-2 overview: we represent robot actions as another language, which can be cast into text tokens and trained together with Internet-scale vision-language datasets. During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control. This allows us to leverage the backbone and pretraining of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control. We demonstrate examples of RT-2 execution on the project website: robotics-transformer2.github.io.

RT-2 [Google, 2023]

VQAによる学習

入力された画像に対する質問に答えるタスクにロボットの行動計画の情報を混ぜてLLMをファインチューニング

マルチモーダルなLLMを用いた行動計画

マルチモーダルな情報をLLMに埋め込んで与えることで、複数のセンサ情報を統合して「この状況ではこう行動するのが望ましい」という行動計画が可能

機械学習モデルとプランナーの橋渡し

通常のLLMは入出力ともに言語であるため、プランナーに対して指示を出すことができないが、ロボットの行動をトークナイズした結果を出力して統合を可能に

ROSで独自MLモデルを作る際の困難さ

分野ごとに規格が乱立

自動運転といえばWaymo/nuScenes/KITTI etc...

画像認識に至っては多すぎて分野内ですら乱立しており、
V&L分野の流行に伴って乱立はさらに加速

アノテーションや前処理が必要

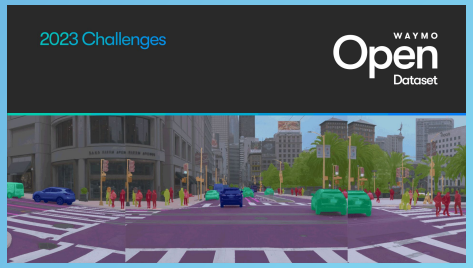
画像のアノテーションは手動でやると時間が溶ける

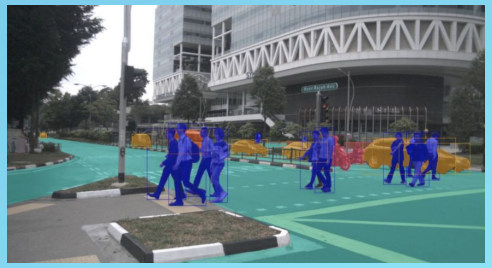
座標変換など、頻出の処理も多い

使わないデータも多い

MLモデルごとに必要なデータを読み出して使うため、色々
入ってるが画像しか使わない、などもよくある

各フォーマットに合わせてデータを読みに行くのではなく、
ROSBAGに含まれているデータを柔軟に取り出せればROSBAGそのものがデータセットに
=> ROSBAGを貯めるだけでモデルが出来上がる！





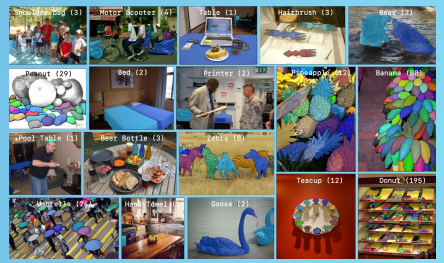


Fig. 6: Samples of annotated images in the MS COCO dataset.

AMBER: AUTOMATED ANNOTATION AND MULTIMODAL BAG EXTRACTION FOR ROBOTICS

AMBERの主な4機能

Dataset

ROSBAGからPyTorchにデータを渡す

Visualization

データを可視化し、分析する

全機能がPython APIを通して連携

PyTorchの豊富なソフトウェア資産を
活用した各種自動化

Automation

非rosbagデータをrosbagに変換

Importer

User Friendlyなツールを目指して全機能を操作できるCLIも提供

開発コンセプト

ROS 2非依存

mcapフォーマットは内部にスキーマ定義を持つためPythonで処理が完結、そのためAMBERをクラウドなどで実行するのも容易

Docker Imageも配布中

様々なデータに対応

現在は画像、点群データに対応

将来的には姿勢、音声、言語などにも対応予定

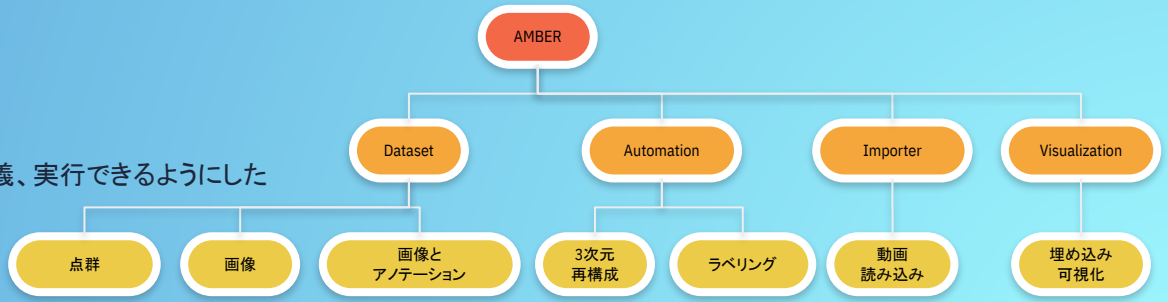
自動化とUser Friendlyの両立

すべての機能にPython APIとCLIを提供

将来的にはyamlやノードエディターで全工程を定義、実行できるようにしたい

Github Repository

Documentation



AMBERに現状実装されている機能

DATASET機能

rosbagの中身を呼んでtorch.Tensor等に変換

mcapの中身にスキーマがあるので、mcapフォーマットのrosbagさえればOK

mcapフォーマットのデータを解釈するライブラリは公式実装がyPI上で公開されているのでそれを利用

yamlでどのデータを読み出すか指定可能

例：画像とアノテーションデータを読み出す場合

```
image_topics:
  - topic_name: /image_front_left
annotation_topic: /detic_image_labeler/annotation
compressed: false
```

1.mcapに記録されたスキーマ定義を読む
2.データとそのスキーマを紐付ける
3.スキーマ定義に基づいてデータを復元
4.復元したデータをPyTorchのDatasetに入れる

AMBER

5.必要に応じてデータを変換しながら読み出す

PyTorch

AUTOMATION機能

PyTorchで動作するMLモデルによる自動化

Dataset機能によりrosbagのデータをPyTorchに入力することが可能になり、PyTorchの豊富な資産を利用可能

Deticによる自動アノテーション機能

Deticという21Kクラスを検出可能なV&Lモデルを用いて画像に対して推論、マスク、bounding box、クラスの情報が自動的にアノテーションされていく

NeRFによる三次元再構成機能

NeRF Studioを統合し、rosbagの画像データから三次元再構成を実現

marching cubes法などによるメッシュ作成も実行はできるが、現状抜けが激しく崩壊している

画像のサンプリングなどを今後改善予定

IMPORTER機能

非rosbagデータをmcap形式のrosbagに変換

動画、音声などには一般的に使われているデータ形式が存在する、それらをrosbagに変換できればそれらのデータもAMBERで利用可能になる

現在サポートしているのは動画のみ

.mp4形式の動画をrosbagに変換し、保存する

タイムスタンプなども保存される

ただし.mp4に含まれる音声データの保存は未対応

VISUALIZATION機能

埋め込み表現の可視化

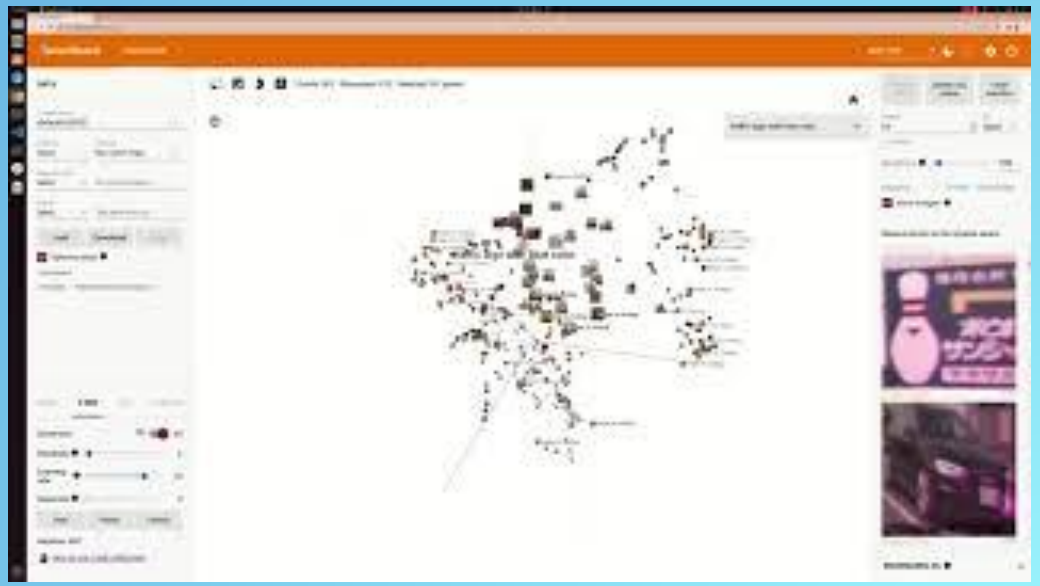
マルチモーダルなMLモデルでは「埋め込み表現」という形でデータを共通の表現に変換するものがある、得られたテンソルを更に次元圧縮すると人間も解釈しやすいデータが得られる

CLIPの埋め込み表現可視化でできること

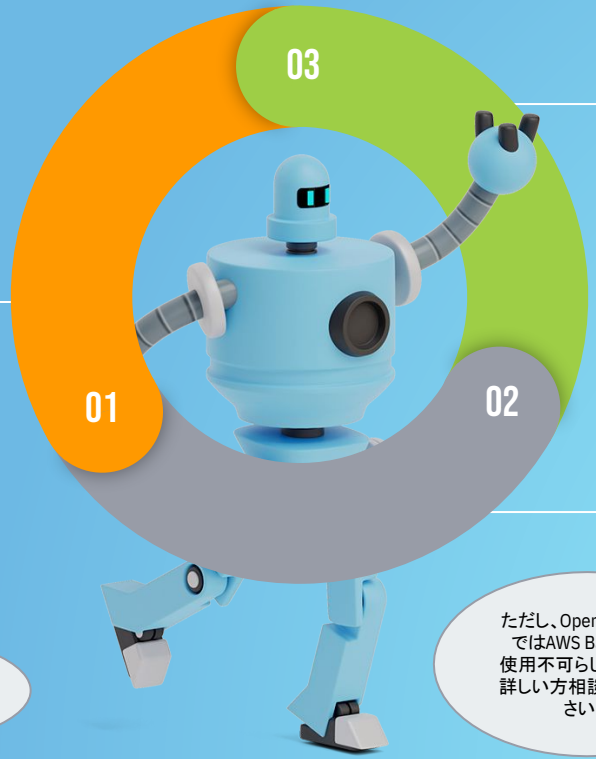
言語と画像を共通の表現に変換するので、rosbagの中からプロンプトで物体を検索

物体検出器のしきい値検討に

詳細は↓ROJP #52仙台の陣の発表資料をチェック！



AMBERを利用したロボット向けML開発基盤



データ取得

必要なデータを rosbag recordし基盤にupload

ロボットを動かして
データを貯める！

推論器のデプロイ

ML基盤で出来上がったモデルや各種パラメータを実機に対して送信する

現状はソフトウェアごと dockerでデプロイすることを検討中

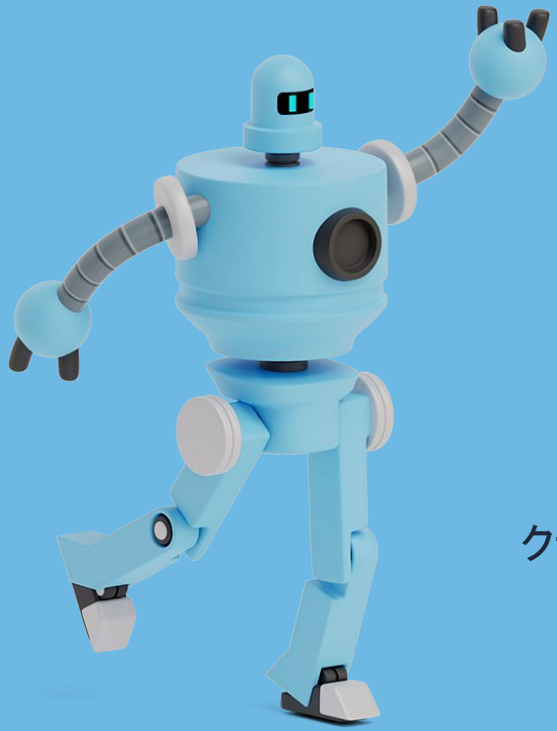
docker

データの分析、学習

現在AWSをlocalに再現するLocalStackを用いてクラウドでもオンプレでも使用可能な基盤を実装検討中

ただし、Open Source版ではAWS Batch等が使用不可らしく悩み中、詳しい方相談させてください！

まとめ



AMBERで快適ML x ROS生活

ROSのDataを直接PyTorchに

自動化ツールも装備で快適！

ROS非依存でデプロイ楽々

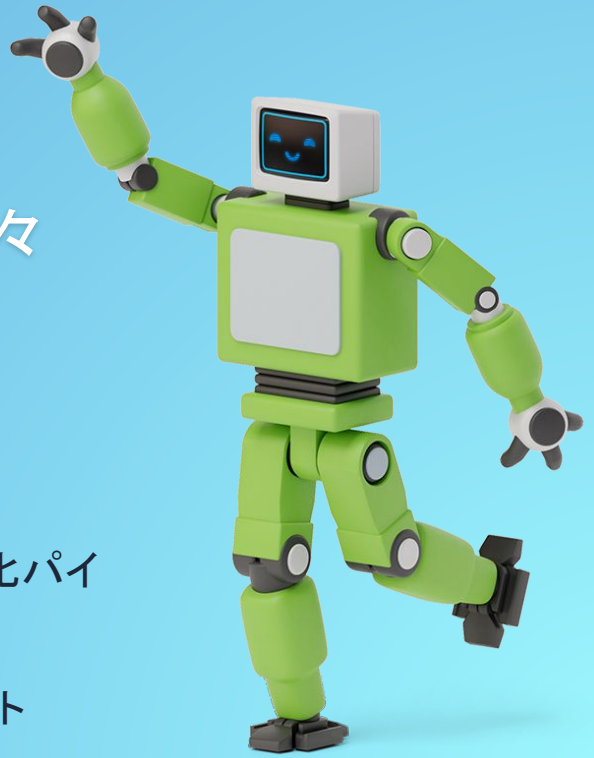
AMBERはただのPythonライブラリ

Docker Image配布もあり

今後の予定

クラウド・ローカルの区別がない自動化パイプラインの実装例提示

音声等の他のモダリティのサポート



THANK YOU FOR LISTENING

HAVE A GOOD ROBOT TIME !