RVT: Robotic View Transformer for 3D Object Manipulation

Journal	CoRL
keyword	
link	https://openreview.net/forum?id=0hPkttoGAf

Abstract

카메라 이미지 대신 3D representation을 사용하는 것이 더 좋은 성능을 가진다. 그러나 voxels 같은 3D representation은 비용이 많이 들고 이에 따라 규모 확대에 제약이 생길 수 있다.

본 논문은 정확하면서도 scalable한 RVT를 제안한다. 이 RVT의 가장 중요한 특징은 attention 메커니즘으로, 여러 view로부터의 정보를 합치고 virtual view의 카메라 입력을 re-rendering하는 데 활용한다.

Introduction

로봇 학습의 기본 목표는 제약이 없는 3D 환경에서 다양한 manipulation 작업을 해결할 수 있는 시스템을 구축하는 것이다. image-based 학습 방법이 여러 카메라의 시점에서 본 이미지를 직접 처리하는 방식으로, 이러한 방법은 다양한 pick-and-place 및 물체 재배치 작업에서 좋은 성과를 냈다. 그러나 3D 추론이 필요한 작업에서의 성공은 제한적이었다. 이를 해결하기 위해 scene을 multi-resolution voxels로 표현하는 C2FARM과 복셀을 처리하기 위해 perceiver transformer를 사용하는 PerAct와 같이 명시적인 3D 표현을 사용하는 방법이 제안됐다. 그러나 복셀 기반 방법은 image-based 추론에 비해 높은 컴퓨팅 비용이 들며, 이로 인해 scalability가 떨어진다.

이러한 문제를 해결하기 위해, 우리는 RVT를 제안한다. RVT는 기존의 SOTA 복셀 기반 방법보다 성공률과 훈련 시간 모두에서 현저한 개선을 보였다. (같은 하드웨어를 사용할 때, RVT는 PerAct의 최고 성능을 36배 더 빠른 시간에 달성하여, 훈련 시간을 14일에서 10시간으로 줄였다.)

RVT는 트랜스포머 아키텍처를 활용하는 image-based 방법으로, 장면의 여러 시점을 함께 주목하며 시점 간의 정보를 집약한다. 이를 통해 로봇의 end-effector 위치를 예측하는 데 사용되는 시점별 heatmaps와 features을 생성한다.

Related Work

Vision-based Object Manipulation

최근 들어 고차원의 시각 input이 다양한 task에 적용가능한 관찰 표현정보들을 제공하고 실세계에 적용가능해지면서 vision-based 정책이 더 주목을 받고 있다. RT-1 논문에서는 트랜스포머를 사용해 이미지 정보로부터 robot's action을 추론해냈다. RVT도 이와 비슷하게 action을 예측하기 위해 트랜스포머를 사용하지만, 추가적으로 multi-view scene representation을 구성하기 위해 depth 정보를 사용한다. (RGB-D 이미지) 정확히 말하면, point cloud를 RGB-D 이미지 셋으로 변환해서 사용한다.

Multi-Task Learning in Robotics

Transformers for Object Manipulation

attention 메커니즘을 사용해 sensory input으로부터 feature를 추출해서 policy learning을 개선할 수 있다. RVT는 큰 데이터셋을 사용하지 않고 small set of demonstrations로부터 학습한다.

Multi-View Networks in Computer Vision

Method

우리의 목표는 다양한 manipulation 작업을 완수할 수 있는 single 모델을 학습하는 것.
모델의 입력은 (1) 작업에 대한 language description, (2) 현재 시각 상태(RGB-D 카메라로부터), 그리고 (3) 현재 그리퍼 상태(열림 또는 닫힘)로 구성된다. 모델은 다음 key-frame에서의 end-effector 포즈와 그리퍼 상태로 지정된 행동을 예측해야 한다.

Rendering

RVT는 로봇 작업 공간 주위에 re-rendering된 이미지를 처리하고, 각 시점에 대한 출력을 생성한 다음 3D로 back-project하여 그리퍼 포즈 행동을 예측한다. 카메라 입력의 re-rendering은 장면의 포인트 클라우드를 먼저 재구성한 후, 로봇 기반 주변의 공간에 고정된 일련의 가상 시점에서 re-rendering하는 것으로 시작된다. (이 과정은 입력 이미지를 변환기에

공급되는 이미지와 분리시켜, 실제 카메라 배치에 의해 제한되지 않는 임의의 유용한 위치에서 재렌더링할 수 있는 능력을 포함하여 여러 이점을 제공함)

Joint Transformer

re-rendering된 이미지, task description, 그리고 그리퍼 상태는 공동 트랜스포머 모델에 의해 처리된다. text에 대해서는 사전 훈련된 CLIP 임베딩을 사용하고, 가상 이미지는 20×20 패치로 나누어 처리하며 MLP를 거쳐 이미지 토큰을 생성한다. 그리퍼 상태도 MLP를 거쳐 이미지 토큰과 concat한다. 이 과정에서 positional 임베딩을 이미지 토큰과 언어 토큰에 붙여서 positional 정보를 보존할 수 있도록 했다.

Action Prediction

모델은 8D 행동을 출력하며, 이는 6-DoF target end-effector 포즈, 1-DoF 그리퍼 상태, 그리고 저수준 모션 플래너의 충돌 허용 여부를 포함한다. RVT는 이러한 multi-view representation을 통해 2D top-down view setting에서의 이전의 전통적인 접근 방식을 확장하며, 시각 입력과 행동을 같은 공간 구조로 표현함으로써 superior sample efficiency를 갖는다.

Loss Function

RVT는 히트맵, 회전, 그리퍼 상태 및 충돌 지시기에 대해 다양한 손실을 혼합하여 사용하여 훈련된다.

Conclusions and Limitations

본 논문은 3D 객체 조작을 위한 multi-view 트랜스포머 모델인 RVT를 제안했다. RVT가 PerAct와 C2F-ARM과 같은 기존 최신 모델들을 다양한 3D 조작 작업에서 능가하며, 더 확장 가능하고 더 빠르다는 것을 발견했다. 또한 RVT가 소수의 시연만으로 실제 세계의 조작 작업에 적용될 수 있다는 것을 밝혀냈다.

RVT가 RLBench에서 최고 수준의 결과를 달성했음에도 불구하고 (62.9% 성공률), 개선의 여지는 여전히 존재한다. 저자는 가상 시점에 대한 다양한 옵션을 간략히 탐구했고, 직교 시점이 작업 전반에 걸쳐 잘 작동한다는 것을 발견했지만, 가상 시점을 더 최적화하거나 데이터에서 학습할 수 있는 방법에 대한 후속 연구를 제안했다. 또한, 기존의 view-based 방법과 비교할 때, RVT(그리고 PerAct 및 C2F-ARM과 같은 explicit 복셀 기반 방법)는 카메라에서 로봇 베이스로의 외부 캘리브레이션이 필요한데, 이런 제약을 제거하는 방법에 대해서도 탐색하는 것도 함께 제안하고 있다.