새로운 Robot Embodiment 와 Task 에 대한 Vision-Language-Action 모델의 일반화 성능에 관한 연구

배민지, 김동훈, 이수윤, 손진우, 남웅희, 심병효
서울대학교

{mjbae, dhkim, sylee, jinwooson, uhnam, bshim}@islab.snu.ac.kr

A Study on Generalization of Vision-Language-Action Models on Unseen Robot Embodiments and Tasks

Minji Bae, Donghoon Kim, Suyun Lee, Jinwoo Son, Unghui Nam, and Byonghyo Shim
Seoul National University

요 약

본 논문은 사전학습 데이터에 포함되지 않는 새로운 robot embodiment 와 task 에 대한 Vision-Action model 의 일반화 성능에 대해 설명한다. 사용된 Vision-Language-Action (VLA) 모델은 OpenVLA 이며 Open X-Embodiment 데이터셋으로 훈련된 모델을 사용한다. 같은 조건에서 촬영한, 길이가 다른 두 가지 데이터셋으로 각각 모델을 훈련한 후 성능을 그래프로 나타내어 데이터 양 증가에 따른 VLA 모델의 일반화 성능 향상과 전이 가능성을 시각적으로 확인한다.

.

Ⅰ. 서 론

최근 로봇 기술은 AI 와 융합하여 다양한 산업 분야에서 새로운 가능성을 열고 있으며 그 중 Vision-Language-Action (VLA) 모델은 대량의 데이터로 학습된 Vision-Language 모델의 world knowledge 를 활용할 수 있어 새로운 환경이나 물체에 대하여 높은 일반화 성능을 보인다는 점에서 최근 활발히 연구가 진행되고 있다. [1] 대표적인 VLA 모델인 OpenVLA 는 대규모 멀티 로봇 데이터셋인 Open X-Embodiment (OXE) [2]로 훈련되어 다양한 robot embodiment 와 task 에 대하여 최고 성능을 달성하였다. [3]

그러나 실제 환경에서는 OXE 에 포함되지 않은 새로운 robot embodiment 이 존재하며, 단순히 물체를 옮기는 작업 그 이상의 task 가 요구되는 상황이 많다. 따라서 환경과 물체 뿐만 아니라 새로운 embodiment 와 심층적인 task 에 대하여 모델이 얼마나 잘 적응하고 전이되는지는 중요한 문제이다. 본 논문에서는 훈련 데이터에 포함되지 않는 robot embodiment 와 단순 물건 옮기기를 넘어서는 심층적인 task 로 구성된 데이터셋에 대하여 VLA 모델의 일반화 성능에 대해 다룬다. 본 논문이 기여하는 바는 다음과 같다.

1. 훈련 데이터에 포함되지 않는 robot embodiment 와 task 로 구성된 데이터셋 수집
2. 훈련 데이터 양에 따른 모델의 일반화 성능 비교

Ⅱ. 데이터셋 구성

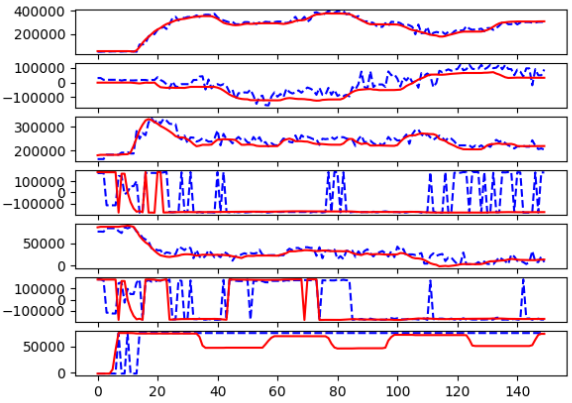
본 연구에서는 OXE 데이터셋에 포함되지 않는 robot embodiment 와 단순한 manipulation task 를 넘어선 심층적인 task 를 기반으로 독자적인 데이터를 수집하였다. 데이터는 사용 모델인 OpenVLA 의 입력 형식에 맞게 RLDS 데이터셋 형식으로 구성되어 있다..

1) Robot Embodiment: 데이터셋 수집에 사용된 robot embodiment 는 Agilex 사의 Piper robot arm 으로 OXE 데이터셋에는 포함되지 않는 robot embodiment 이다. 해당 embodiment 는 OXE 에 포함된 embodiment 와 action space 면에서 다소 차이를 보이며, 해당 embodiment 는 다른 embodiment 대비 저렴한 가격으로 향후 다양한 산업에 활용될 가능성이 높은 embodiment 이다.

2) Task: 실험에 사용된 task 는 종이컵을 옮기는 task 이다. 구체적으로, 종이컵에는 각각 숫자가 적혀 있고 해당 숫자를 보고 숫자가 낮은 순으로 컵을 옮겨서 나란히 정렬하는 task 이다. 종이컵은 찌그러지기 쉽기 때문에 종이컵을 옮기는 task 는 다른 task 대비 난이도가 높은 편이며, 숫자를 인식하여 낮은 순서대로 컵을 옮겨야 하므로 기존의 manipulation task 와는 다르게 reasoning 이 다소 필요한 task 이다. Prompt 는 'Align the cups'로 모든 데이터셋에서 동일하다.

3) 수집 절차 및 분량: 총 세 종류의 데이터셋을 수집하였다. 1 번 데이터셋은 약 10 분 분량 (20 episodes)이고 2 번 데이터셋은 약 50 분 분량 (90 episodes)이다. 두 데이터셋 모두 동일한 환경 및 동일한 task 에 대하여 수집되었으며 카메라 화각 역시 동일하다. 마지막 3 번 데이터셋은 약 10 분 분량이며 동일 환경 및 동일 task 기반이지만 카메라 화각을 달리하여 촬영된 데이터셋으로, 화각 차이에 의한 성능 변화를 관찰하기 위해 구성되었다. 1,2 번 데이터셋은 over-the-shoulder 화각으로 OXE 데이터셋에 포함된 다수의 데이터셋이 사용하는 화각이다. 3 번 데이터셋은 top-view 로 촬영되었다.

Ⅲ. 실험 구성 및 결과



[그림 1] 1 번 데이터셋 훈련 결과

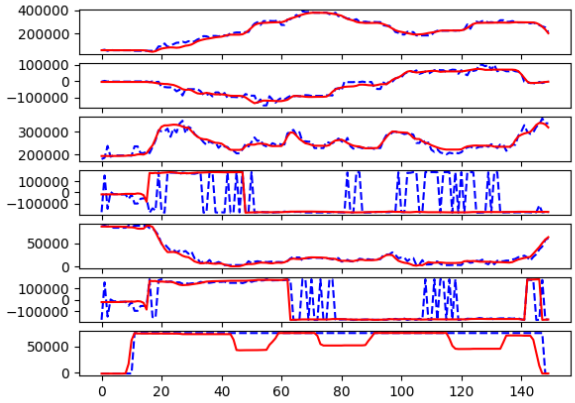
1) 실험 구성: 실험은 사전학습 된 OpenVLA 모델에 각각의 데이터셋을 LoRA tuning 후 validation set 의 ground truth 값과 모델의 아웃풋 action 과의 차이를 그래프로 시각화하여 비교하였다.

2) 실험 결과: 각 데이터셋으로 훈련한 결과는 [그림 1], [그림 2], [그림 3]에서 확인할 수 있다. 위에서부터 순서대로 x, y, z, rx, ry, rz, gripper 의 action 이다. 세 경우 모두 ground truth 경로를 따라가는 경향을 보이나 경로의 안정성 면에서 다소 차이를 보였다. 먼저 화각도 다르고 길이도 짧았던 3 번 데이터셋의 경우 경로가 안정적이지 않은 것을 확인할 수 있었다. 화각이 달라진 1 번 데이터셋의 경우 3 번 보단 안정적이나 여전히 noisy 한 경로를 보였으며, 길이가 긴 2 번 데이터셋의 경우 다른 결과 대비 안정된 경로를 보였다. 그러나 세 경우 모두 rotation 에 대해서는 눈에 띄는 오차를 보였으며 gripper 의 세밀한 제어는 전혀 학습을 하지 못했다.

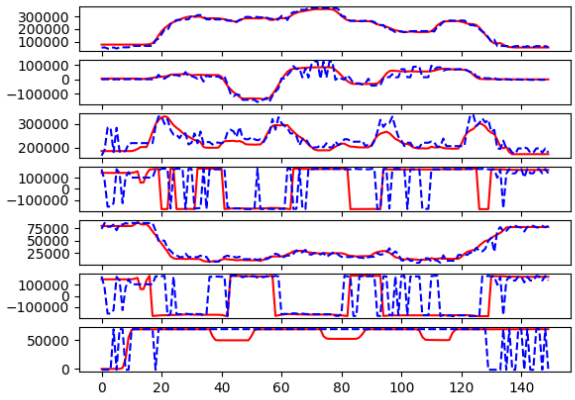
Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 새로운 robot embodiment 와 reasoning 기반 task 에 대하여 수집한 데이터로 VLA 모델을 훈련하고 일반화 성능을 확인하였다. 실험 결과, 짧은 데이터셋에서도 어느 정도의 성능을 보이나 데이터의 양이 많아질수록 안정화된 경로를 보였으며 화각 변화는 직접적인 성능 저하를 유발하였다. 그러나 동일한 화각과 1 시간 가량의 긴 데이터셋으로도 gripper 의 정밀 제어나 rotation 방향의 조절은 학습하지 못하는 것을 확인할 수 있었다.
향후 연구 방향은 다음과 같다.

1) 추가로 데이터를 수집하여 action 의 모든 값에서 안정된 성능을 보이는 데이터의 임계치를 규명하고자 한다.
2) 적은 양의 데이터로도 gripper 의 action 과 같은 세밀한 제어를 학습할 수 있는 효율적인 학습 전략을 개발하는 것을 목표로 한다.



[그림 2] 2 번 데이터셋 훈련 결과



[그림 3] 3 번 데이터셋 훈련 결과

참 고 문 헌

[1] [1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, … A. Irpan, et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” arXiv preprint arXiv:2307.15818, 2023.

[2] A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, … Z. Lin, “Open X-Embodiment: Robotic Learning Datasets and RT-X Models,” arXiv preprint arXiv:2310.08864, 2023.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, … C. Finn, “OpenVLA: An Open-Source Vision-Language-Action Model,” arXiv preprint arXiv:2406.09246, 2024. arXiv + 4 arXiv + 4 openvla.github.io + 4