파운데이션 모델을 활용한 객체기반 표현 학습 일반화에 관한 실험적 연구

이동훈, 장인국, 송순용, 배희철
한국전자통신연구원

{donghun, ingook, soony, hessed}@etri.re.kr

An Experimental Study on Object-Centric Representation Learning for Generalization Using Foundation Model

Donghun Lee, Ingook Jang, Soonyong Song, Heechul Bae
Electronics and Telecommunications Research Institute

요 약

분포 외(out-of-distribution, OOD) 환경에서의 객체의 시각적인 특징을 학습하는 visual scene reconstruction 연구 분야는 다양한 상황에서의 예측을 하는데 활용될 수 있지만 아직 제한적인 성능을 보여주고 있다. 최근 비전 분야에서도 foundation 모델이 우수한 성능을 보여주고 있으며 특히 image segmentation 분야에서도 뛰어난 성능을 보여주고 있다. 본 연구에서는 visual foundation model 중의 하나인 SAM 모델에서 도출된 세그멘테이션 결과를 활용하여 visual scene reconstruction 성능을 향상시키는 객체 기반의 표현학습 기술을 제안한다. 제안되는 연구에서는 scene representation 학습을 위한 객체 기반 표현학습 중 unsupervised feature 추출을 위한 간단하지만 효과적인 fine-tuning 방법을 소개한다. 또한, 실험을 통해 OOD 환경에서 비지도 표현 학습의 성능을 향상시켰으며, 그 결과 단일 객체 OOD 시나리오에서 성능이 개선되었음을 보여주었다.

Ⅰ. 서 론

최근 딥러닝을 통해 컴퓨터 비전의 성능은 크게 향상되었지만, 이는 주로 훈련 데이터와 테스트 데이터가 동일한 분포에서 나올 때를 가정한 것이다. 하지만 여전히 실제 환경에서, 특히 객체가 비정상적인 속성을 가지고 있거나 어려운 조건에서 나타날 때 인간 수준의 시각적 인식을 따라가지 못한다. [1] 이러한 분포 외의(OOD; out-of-distribution) 시나리오에서의 부족함은 자율 주행, 제어 및 조작과 같은 여러가지 응용 분야에서 주요한 도전 과제 중 하나이다.

컴퓨터 비전에서 visual scene representation learning 은 시각적인 특징점을 추출하여 다양한 어플리케이션에 활용할 수 있는 중요한 연구 분야이다. 복잡한 작업에서는 단일 표현에 의존하는 것이 비효율적이며, 이는 객체 간의 관계를 추출하기 어렵게 만든다. 인간은 장면을 구성적으로 인식할 수 있기 때문에 시각적 장면을 이해하는 데 능숙하다. 장면을 구성 요소로 나누어 각각의 시각적 개념에 대응하는 영역으로 표현을 추출하면 시각적 장면의 이해가 향상되며, 이는 인간의 이해 방식과 일치한다. 또한, 객체 기반 표현 학습 방식을 통해 OOD 환경에서의 robust 하고 신뢰할 수 있는 예측이 가능하다.

최근 대규모 데이터셋으로 사전 학습된 파운데이션 모델은 zero shot 및 few shot 등의 일반화에서 놀라운 성과를 보여주고 있다. Meta AI 의 Segment Anything Model (SAM)[2]은 이미지 세분화에서 뛰어난 성능을 보이며 컴퓨터 비전 연구에서 유용한 응용 가능성을 제시하고 있다. 하지만 SAM 을 활용한 OOD 일반화에 대한 연구는 아직 제한적이다.

본 연구에서는 파운데이션 모델 SAM 을 활용하여 scene representation learning 의 일반화 성능을 빠르게 수렴하고 보다 정확한 모델 생성 방법을 제시하고자 한다. 이를 통해 OOD 환경에서 robust 하고 신뢰할 수 있는 모델을 구현할 수 있다.

Ⅱ. 본 론

본 연구는 segmentation foundation 모델 SAM 에서 생성된 결과를 Slot Attention[3] 모델에 적용하여 scene representation learning 을 진행하였다.

Visual Scene Reconstruction 을 위한 표현 학습은 시각적 장면의 입력 이미지 $x \in R^{N \times C}$ 와 재구성된 이미지 $x' \in R^{N \times C}$ 를 사용하는 것을 의미한다. 여기서 N 은 각 입력 이미지의 픽셀 수로, 일반적으로 높이와 너비의 곱이다. D_{inputs} 는 입력 이미지의 차원으로, 일반적으로 RGB 값을 나타낸다. 함수 $f_x: R^{N \times D_{inputs}} \rightarrow R^{N \times D_{enc}}$ 은 차원이 $N \times D_{inputs}$인 입력 데이터를 세분화 정보를 포함하는 인코더 공간 $N \times D_{enc}$ 으로 매핑한다. 이 세그멘테이션 정보는 세그멘테이션 파운데이션 모델인 SAM 모델에서 얻어지며, zero shot 세그멘테이션 기법이 사용된다. 함수 $g_x : R^{N \times D_{enc}} \rightarrow R^{N \times D_{inputs}}$ 는 인코더 공간에서 인코딩된 표현 x를 $N \times D_{inputs}$ 차원의 공간으로 매핑하며, 여기서 K 는 슬롯 표현의 개수를 나타낸다.

기존의 Slot Attention 의 접근 방법은 각 슬롯을 무작위로 초기화를 시켰지만, 본 연구에서는 각 슬롯이 비지도 학습을 통해 특징을 학습하고, 반복적인 attention mechanism 을 통해 입력 데이터와 슬롯을 맵핑한다.

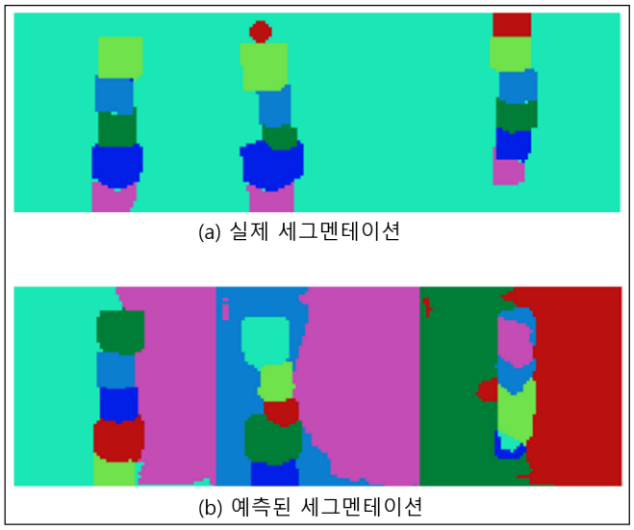

그림 1. Objects Room Dataset 에서의 실험 결과

Ⅲ. 실험 및 결과

본 연구의 실험은 벤치마크 알고리즘과 데이터셋을 제공하는 object centric library[4]를 사용하여 수행되었다. 이 실험에서 Objects Room 데이터셋이 활용되었으며, Slot Attention 알고리즘이 벤치마크 알고리즘으로 사용되었다. 데이터셋은 8,000 개의 학습 세트, 1,000 개의 테스트 세트, 1,000 개의 검증 세트로 나누어졌다.

실험 중 하나의 객체에 분포 이동을 유도하여 객체의 색상, 모양 또는 크기와 같은 요소가 변하는 분포 외(out-of-distribution, OOD) 환경이 생성되었다. 성능 측정 metric 으로는 Adjusted Rand Index (ARI)가 사용되었으며, 세그멘테이션 정확도 평가는 재구성된 이미지와 실제 이미지 간의 비교를 통해 이루어진다. 그림 1 을 통해 실제 세그멘테이션의 ground truth 와 제안된 알고리즘의 결과를 확인할 수 있다.

알고리즘	ARI
Slot Attention	0.775
Ours	0.853

표 1. 실험결과

표 1 에서 실험 결과를 확인할 수 있으며 제안하는 알고리즘이 베이스라인 대비 10% 정도 향상된 ARI 성능을 보여주는 것을 확인 할 수 있다.

Ⅳ. 결론

본 연구에서는 세그멘테이션 foundation model SAM 에서 얻은 zero shot 결과를 활용하여 OOD 상황에서의 scene representation learning 의 visual scene reconstruction 성능을 향상시키는 간단하지만 효과적인 객체 기반의 표현학습 방법을 제안한다. 본 연구를 확장하여 보다 복잡한 OOD 환경에서 높은 성능을 얻기 위해 학습 방법에 활용될 수 있다.

ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [24ZR1100, 자율적으로 연결· 제어· 진화하는 초연결 지능화 기술 연구]

참 고 문 헌

[1] Zhao, Bingchen, et al. "Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[2] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[3] Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in neural information processing systems 33 (2020): 11525-11538.

[4] Dittadi, Andrea, et al. "Generalization and robustness implications in object-centric learning." arXiv preprint arXiv:2107.00637 (2021).