퍼셉튜얼 객체기반 학습을 활용한 고수준 시각적 특징 표현 학습 연구

이동훈, 김성현, 노삼열, 장인국
한국전자통신연구원

{donghun, kim-sh, samuel, ingook}@etri.re.kr

High-level Visual Representation Learning via Perceptual Object-centric Learning

Donghun Lee, Seonghyun Kim, Samyeul Noh, Ingook Jang
Electronics and Telecommunications Research Institute

요 약

최근 연구되고 있는 표현 학습 및 비디오 예측 기술의 발전은 미래 상태의 정확한 예측을 가능하게 하고, 많은 응용 분야에서의 조작 및 제어 정책의 개선 가능성이 있음을 보여주고 있다. 하지만 실 데이터의 고유한 불확실성으로 인해 시각적인 표현들을 정확하게 학습하기 용이 하지 않다. Autoregressive 모델은 생성된 미래 프레임을 다음 미래 상태 예측에 대한 입력으로 사용한다. 이러한 방법은 Latent Vector 로부터 항상 시각적인 특징을 재구성하기 때문에 Compounding Error, 메모리 그리고 학습 시간에 취약하다. 본 논문에서는 고수준 표현을 추출하고 분리하기 위해 Perceptual network 를 Slot Attention 에 적용하여 미래 상태의 예측 정확도를 개선시키고자 한다. 사전학습된 Perceptual Network 는 각 해당 슬롯이 달성되는 각 perceptual layer 에 대해 높은 수준의 시각적 특징 표현을 얻을 수 있다. 이 높은 수준의 객체 기반 시각적 특징 표현은 현재 상태를 더 잘 이해하고 미래 상태를 정확하게 예측하는 데 도움이 될 수 있다.

Ⅰ. 서 론

표현 학습 및 비디오 예측의 최근 발전은 미래 상태에 대한 정확한 예측이 자율 주행, 드론 제어 및 엣지 장치 등 많은 애플리케이션에서 조작 및 제어 정책을 개선할 가능성이 있음을 보여주었다 [1]. 하지만 실제 데이터의 고유한 불확실성으로 인해 표현을 학습하는 것이 어렵다.

RNN 과 같은 결정론적 접근 방식은 동적 환경에 제한에 제한적인 성능을 보여준다. SVG, SV2P 및 SAVP 와 같은 autoregressive 생성된 미래 프레임을 다음 미래 상태 예측에 대한 입력으로 사용한다. 이러한 방법은 latent vector 로부터 항상 시각적인 특징을 재구성하기 때문에 compounding error, 메모리 그리고 학습 시간에 취약하다 [2]. PlaNet 과 같은 State Space Model(SSM)에 대한 최근 연구는 먼 미래 예측의 장점을 가지고 latent vector 에서 예측하기 위해 제안되었다. 그러나 이러한 접근 방식은 객체 중심 시각적인 표현을 추출하는데 제한적인 성능을 보여주고 있다.

최근 Slot Attention [3]과 같은 객체 중심 접근 방식은 시각적으로 숨겨져서 보이지 않는 부분에 대한 일반화에 좋은 성능을 보여주고 있다. 객체 중심의 접근 방식은 이미지에 대한 예측[4]뿐만 아니라 비디오에 대한 예측에도 attention 메커니즘[5]을 사용한다. 하지만 객체 중심 접근 방식은 입력에서 밀접하게 관련된 기능에 초점을 맞추고 고수준의 시각적인 특징 표현을 추출하는 데 제한된 능력을 보여준다.

본 논문에서는 고수준 표현을 추출하고 분리하기 위해 Perceptual Network 에 Slot Attention 적용하여 미래 상태의 예측 정확도를 개선시키고자 한다. 사전학습된 Perceptual Network 는 각 해당 슬롯이 달성되는 각 perceptual layer 에 대해 높은 수준의 시각적 특징 표현을 얻을 수 있다. 이 높은 수준의 객체 기반 시각적 특징 표현은 현재 상태를 더 잘 이해하고 미래 상태를 정확하게 예측하는 데 도움이 될 수 있다.

Ⅱ. 본 론

본 연구는 Slot Attention 모듈을 따르는 객체 중심의 Perceptual Network 표현학습을 위한 Perceptual Slot Attention 모듈을 제안한다. 그것은 입력 모듈, Perceptual Network 모듈, Attention 모듈 및 출력 모듈

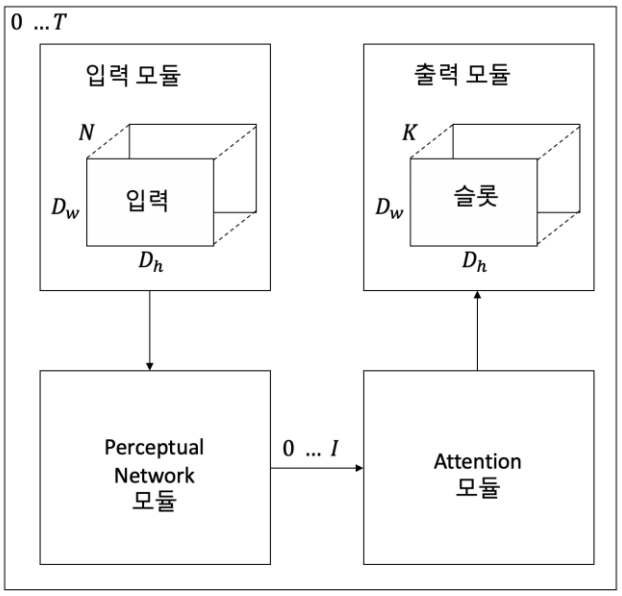

그림1. Perceptual Slot Attention

총 4 개로 구성된다.

Perceptual Slot Attention 의 목적은 그림 1 과 같이 입력의 N 벡터를 K 개의 슬롯에 매핑하는 것이다. 입력 모듈은 D 차원의 입력 벡터를 N 취한다. 본 연구에서는 정적 이미지를 입력으로 사용하였다.

Perceptual Network 모듈은 입력 이미지의 각 Perceptual layer i에 대해 고수준의 시각적 특징 표현을 생성한다. 이 방법에서는 각 Perceptual Layer i에서 고수준의 시각적 특징 표현을 위해 ImageNet 데이터셋에서 사전 훈련된 VGG16 네트워크를 활용하였다.

Attention 모듈은 Slot Attention 의 특징을 따른다. 슬롯의 각 Perceptual layer i에 대해 T 반복에 대해 Attention 메커니즘이 적용된다. Softmax 프로세스 동안 각각의 $slot$ s_i는 서로 경쟁하여 반복적으로 업데이트 한다.

출력 모듈은 N 차원의 입력이 K 차원의 고수준의 시각적 특징 표현을 포함하는 K 슬롯의 결과를 출력해 준다.

Ⅲ. 실험

본 연구의 실험에서 64 사이즈의 입력 사이즈, 256 사이즈의 RNN Layer, 128 g dimension, 10 z dimension 이 사용되었다. Batch size 100, 600 epoch, 300 iteration 동안 Adam Optimizer 를 사용하여 0.9 의 모멘텀을 사용하여 학습되었다. Posterior RNN Layer 와 Predictor RNN Layer 는 각각 1 개, 2 개가 사용되었다.

본 연구의 실험을 위해서 BAIR Push Dataset 이 사용되었으며 평가지표로 SSIM (Structural Similarity Index Map) 수치가 사용되었다. 과거의 5 개의 프레임을 입력으로 받아서, 미래의 10 개의 프레임을 예측하여 Ground Truth 와의 SSIM 수치를 비교 분석하였다. 그림 2 는 Baseline 알고리즘인 SVG-LP 와 제안하는 알고리즘의 성능을 비교하여 보여준다. 제안하는 알고리즘이 t+3 부터 더 높은 SSIM 결과값을 보여주고 있다.

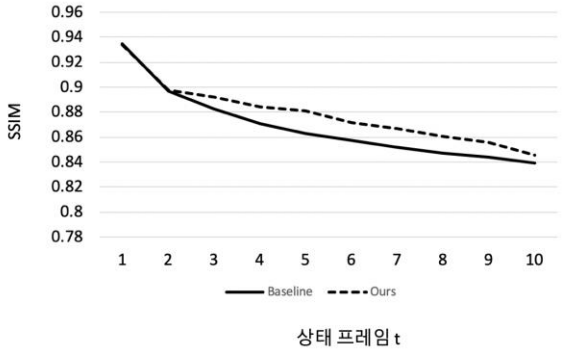

그림2. 실험 결과

Ⅳ. 결론

본 논문에서는 Perceptual Network 의 출력에 Slot Attention 을 적용하여 고수준의 시각적 특징 표현을 학습하는 방법을 제안한다. 사전학습된 Perceptual Network 는 각 해당 슬롯이 달성되는 각 Perceptual layer 에 대해 높은 수준의 시각적 특징 표현을 얻을 수 있다. 이 높은 수준의 객체 기반 시각적 특징 표현은 현재 상태를 더 잘 이해하고 미래 상태를 정확하게 예측하는 데 도움이 될 수 있다.

ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [22ZR1100, 자율적으로 연결· 제어· 진화하는 초연결 지능화 기술 연구]

참 고 문 헌

[1] Jang, Ingook, et al. "An approach to share self-taught knowledge between home IoT devices at the edge." Sensors 19.4 (2019): 833.

[2] Oprea, Sergiu, et al. "A review on deep learning techniques for video prediction." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[3] Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in Neural Information Processing Systems 33 (2020)

[4] Singh, Gautam, Fei Deng, and Sungjin Ahn. "Illiterate dall-e learns to compose." International Conference on Learning Representations. 2021.

[5] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).