심층 모방학습과 Grad-CAM을 이용한 특징기반 의사결정 모델의 시각화분석

배청목○, 주호택, 김경중
광주과학기술원
cmbae0307@gm.gist.ac.kr, hotaek87@gm.gist.ac.kr, kjkim@gist.ac.kr

Visual Analysis of Feature-based Decision-Making Model with Deep Reinforcement learning

Cheong-mok Bae○, Ho-Taek Joo, Kyung-Joong Kim
Gwangju Institute of Science and Technology

요 약

최근 Gradient-weighted Class Activation Map(Grad-CAM) 등의 방법을 이용한 합성곱 신경망의 활성화 시각화 방법론은 심층 학습으로 훈련된 모델의 의사결정 과정을 이해하는 것에 도움을 주었다. 그러나 일부 강화학습 환경에서는 픽셀 기반의 화면 입력을 사용하여 최적의 정책을 훈련하는 것이 어려우며 적절한 해결책을 찾는데 오랜 시간이 필요하다. 하지만 학습 도메인에 대한 지식이 있는 전문가가 존재할 경우 학습도메인에 대한 특징 정보를, 추출하고 이를 활용한 얕은 의사결정 모델을 강화학습 방법을 설계하는 것은 비교적 용이하다. 이 경우 최적의 정책을 찾는 것에 오랜 시간을 필요로 하지 않지만 화면 입력을 활용할 경우와 비교해, 특징의 중요도를 분석하는 것 이외의 시각적 형태의 분석을 하는 것이 쉽지 않다. 본 논문에서는 특징기반 의사결정 모델의 시각화 분석을 위해 합성곱 신경망(Convolutional Neural Network, CNN) 기반 심층 모방학습을 사용하는 것을 제안한다. 이를 통해 합성곱 신경망 기반의 모델을 시각적으로 분석할 때 사용할 수 있는 Grad-CAM 방법을 특징기반으로 학습된 모델에 간접적으로 적용할 수 있다.

1. 서 론

최근 설명가능한 인공지능(Explainable AI, XAI)[1]에 대한 관심이 높다. 의사결정 학습을 위해 사용하는 심층 강화학습 분야에서도 인공지능의 훈련과정을 직관적으로 이해하기 위해 CAM(Class Activation Map), Grad-CAM[2], Saliency map[3] 등의 시각화 방법을 연구하고 있다[4]. 이러한 방법들은 화면 기반의 입력(pixel-based inputs)을 사용하는 심층 강화학습에서 학습 결과를 분석하는데 효과적으로 사용할 수 있다. 즉, 인공지능이 화면상의 어떤 부분을 집중해서 의사결정을 내렸는지를 파악해 볼 수 있다.

화면기반 입력을 사용하는 것이 유용하지만, 많은 현실의 문제에서는 화면의 픽셀 정보보다는 특징(Feature) 값을 직접 파악하여 사용하는 경우가 많다. 이러한 특징값은 전문가에 의해 만들어진 전처리 프로그램을 통해 구해지거나 다양한 센서를 통해서 획득한다. 예를 들어, 자동차를 제어하기 위해 카메라 정보를 사용할 수도 있지만, 자동차의 속도, 현재 위치, 다른 차와의 거리 등을 입력으로 넣어줄 수 있으면, 학습을 빠르고 효과적으로 수행할 수 있다. 예를 들어, 전통적인 cart pole 문제를 화면기반으로 해결하는 경우와 다양한 센서값을 토대로 해결하는 것에는 학습 효율에 큰 차이가 있다[1].

적절한 특징을 입력으로 사용하여 강화학습을 수행하면, 빠른 학습을 수행할 수 있지만, 화면기반 방법에서 보여준 화면상에서의 시각적인 해석은 어려워진다. 본 논문에서는 강화학습을 이용하여 특징기반으로 학습한 모델을 시각적으로 해석하는 방법을 제안한다. 특징값을 기반으로 학습한 모델을 테스트하는 과정에서 렌더링 이미지와 수행한 행동을 수집하고, 이 데이터를 모방하는 합성곱 신경망(CNN)을 학습한다. 모방학습을 통해 얻어낸 모델에 Grad-CAM을 적용하여, 특징값을 통해 학습한 강화학습기반 모델이 행동을 결정할 때 화면의 어느 부분에 집중하였는지를 시각화한다.

2. 배경 지식

2.1 PPO (Proximal Policy Optimization)

PPO[5]는 현재 가장 진보된 강화학습 알고리즘 중 하나로, 과거에 저장된 정책과 현재 진행 중인 정책을 비교하는 것을 통해 알고리즘을 개선한다.

$$maximize\, E_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t\right] = E_t\left[r_t(\theta)A_t\right]$$

위의 수식과 같이, 현재 정책을 이전 정책으로 나눈 후 advantage value를 곱하여 학습의 방향성과 비율을 계산하고, 이를 통해 목적함수를 최대화하는 네트워크를 얻게 된다. 이 과정에서 선형 근사(first order approximation)만을 이용하여 학습의 안정성을 보장함과 동시에 낮은 계산 복잡도를 가진다.

1) https://pytorch.org/tutorials/intermediate/reinforcement_q_learning.html

이를 통해, 대표적인 기존 알고리즘들이 가졌던 Continuous Task Control에서의 성능(Deep Q-Networks, DQN), 데이터 효율성과 학습의 안정성(Vanilla Policy Gradient), 복잡한 구조와 노이즈와 매개변수를 공유하는 환경에서의 호환성 문제(Trust Region Policy Optimization, TRPO)[6]를 개선했다.

2.2 Grad-CAM

Grad-CAM[2]은 학습된 모델이 어떻게 결과 값을 도출했는지를 직관적으로 보여줄 수 있는 픽셀 기반 딥 러닝을 위한 시각화 방법으로, 딥 러닝의 결과를 인간이 이해할 수 있는 형태로 설명해준다.

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k}\left(\frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{k}_{ij}}\right)A^{k}\right)$$

위 수식과 같이, 학습된 합성곱 신경망에 대해서 k번째 특징맵 A의 클래스 c에 대한 중요도를 계산한 후. 각 특징맵을 곱하여 더한 결과에 ReLU를 취하여 k번째 특징맵 A에 대한 결과를 시각화한다.

3. 제안하는 방법

본 논문에서는 아래와 같은 강화학습, 모방학습 및 시각화 분석 절차를 제안한다.

1) 특징기반 환경에서의 강화학습: 강화학습에는 다중 퍼셉트론으로 구성된 모델을 사용하며, 이후 모방학습을 위해 정책기반(policy-based) 모델을 사용한다.

2) 학습된 모델을 통한 샘플 데이터 수집: 강화학습으로 훈련된 모델을 테스트하는 과정에서 모방학습에 필요한 샘플데이터를 수집하며, 이 때 수집하는 데이터는 RGB 또는 GRAY_SCALE로 표현된 화면 이미지 데이터, 그리고 해당 화면에서의 행동 또는 행동실행 확률 분포 데이터이다.

3) 수집된 샘플 데이터를 활용한 모방학습: 화면 관측값과 행동(또는 행동 확률 분포)을 활용하여 모방학습을 진행한다. 모방 학습의 손실함수는 행동 확률 분포에 대한 MSE(Mean Square Error) 또는 KL-Divergence(Kullback-Leibler Divergence)를 이용한다.

4) 모방학습한 모델 결과의 시각화: 모방학습된 결과를 시각화 방법을 사용하여 시각화한다. 이 과정에서는 현재 신경망 시각화 분석에 사용되는 Saliency Map, GradCAM, GradCAM++[7] 등이 적용 가능하다.

본 논문에서는, Discrete한 환경에 대해, 1)의 과정에서 정책기반 모델(Actor-Critic PPO)을 통한 강화학습을, 2), 3)의 과정에서 RGB 화면 이미지와 행동, 행동 확률분포를 수집 및 모방학습, 4)의 과정에서 Grad-CAM 방법을 통한 시각화 결과를 출력하였다.

4. 실 험

사용된 실험환경은 OpenAI Gym에서 제공하는 강화학습 환경인 LunarLander로, Discrete한 환경과 Continuous한 학습환경을 모두 제공한다. 이 환경은 화면기반으로 학습하는 것이 매우 어려운 편이다. 달 착륙선 모양의 에이전트를 지면에 안정적으로 착륙하는 것을 목적으로 하는 환경으로. 전반적인 환경의 모습은 그림 1과 같다.

센서값에 기반하여 화면상의 좌표, 가속도, 각도, 각속도, 우주선의 다리가 지면에 닿았는지에 대한 특징 정보가 주어진다. 행동 값은 상단과 좌우의 가스 분사 장치를 활성화하거나 아무것도 하지 않는 총 4가지의 행동이 있다. (Nope, LEFT, RIGHT, MAIN) 학습에 사용되는 보상 값은 여러 요인에 의해 결정이 되는데, 우선 착지 성공 여부에 따른 점수와 행동을 통해 소모한 연료에 따른 감산 등이 있으며, 달성 가능한 최고점수는 약 300점 전후이다. 에이전트는 최소한의 연료로 착지점(깃발 사이)에 착륙하는 것을 목표로 하면서도, 지면에 접지하는 착륙선 다리의 각도나 속도에 따라 성패 여부가 달라지기도 한다는 제약을 가지고 있다.

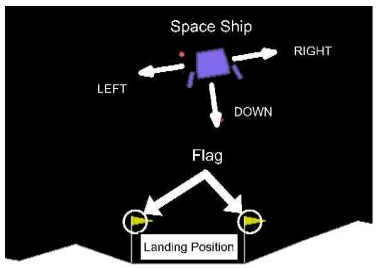

그림 1. OpenAI Gym Lunar Lander

4.1 특징기반 에이전트 학습과 이미지 데이터 수집

LunarLander 환경에 대해, PPO를 사용한 Actor-Critic 에이전트를 학습한다. 학습에 사용한 에이전트는 각각 3개의 다층 퍼셉트론으로 구성된 Actor-Critic 구조이며 각 레이어에 대해서 Hyperbolic Tangent를 취했다. 학습은 총 1천만 프레임동안 진행했으며, 매 2000프레임마다 정책을 업데이트했다. 그림 2는 학습이 진행되는 동안 얻은 보상을 보여준다.

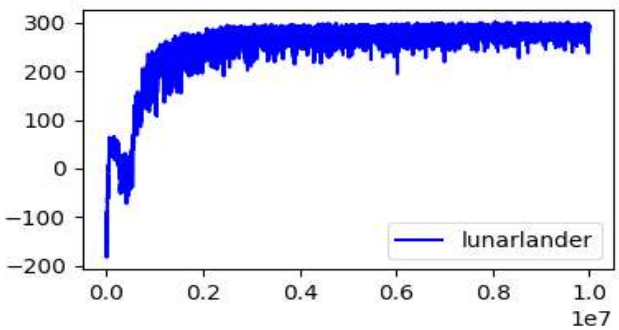

그림 2. 특징기반 LunarLander 훈련 모델의 보상

학습된 모델을 테스트하는 과정에서 렌더링된 화면 이미지와 해당 화면에서 취한 행동에 대한 데이터를 수집하여. 다음 모방학습 단계에서 사용한다. 본 논문에서는 총 20개의 에피소드에 대해서 테스트를 하여 데이터를 수집했다.

4.2 수집된 데이터를 통한 모방학습

수집된 화면 이미지와 행동 데이터를 활용해서 합성곱

신경망 모델에 대한 모방학습을 진행했다. 학습에는 앞선 단계에서 수집한 20개의 에피소드에 대한 샘플 데이터를 사용했다. 학습에 사용한 합성곱 모델은 4개의 Leaky ReLU를 취한 Convolution Layer로 구성했으며, 각 Layer의 말단에 Batch Normalization을 적용하였다. 강화학습 단계에서 출력된 행동을 모방하기 위해 행동의 분포에 대해서 행동 확률 분포에 대한 KL-Divergence와 MSE를 적용한 손실함수를 사용했으며, 각 손실함수를 사용한 모방학습 모델은 그림 3과 같이 학습이 진행됨에 따라 손실 값이 줄어듦을 확인할 수 있었다.

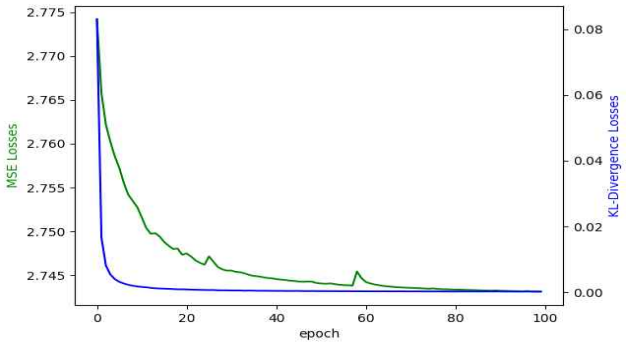

그림 3. 모방학습 진행에 따른 손실
MSE 손실(초록), KL분산 손실(파랑)

4.3 Grad-CAM 결과와 분석

이전 단계에서 학습한 모방학습에 대한 Grad-CAM 결과를 출력한다. 4개의 합성곱 레이어 중 세 번째와 네 번째 레이어가 비교적 화면상의 물체와 지면에 대한 인식을 잘하는 것으로 확인하였다. 그림 4는 세 번째 합성곱 레이어가 집중하는 부분에 대한 Grad-CAM 결과를 원본의 화면과 중첩한 것이다. 훈련한 신경망이 항상 우주선과 깃발의 주변에 집중하여 행동 값을 도출해내고 있는 것을 확인할 수 있다. 또한, 우주선이 지면에 가까워지면 지면의 평탄한 정도에 집중하는 것을 확인할 수 있다.

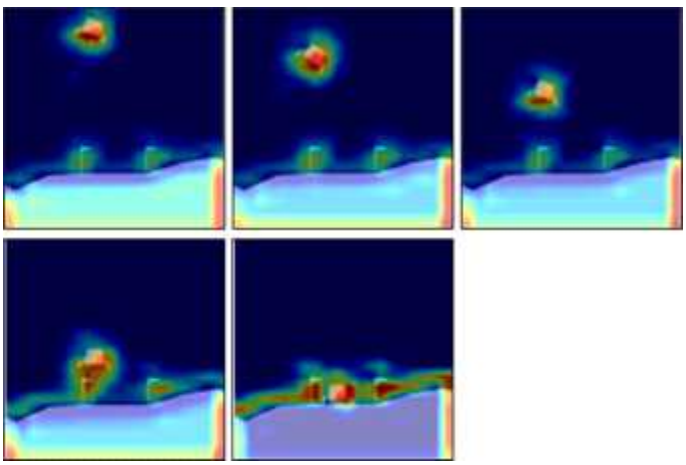

그림 4. LunarLander 모방학습 모델의 Grad-CAM 결과. 붉은색으로 표현한 부분은 네트워크가 상대적으로 집중하는 곳을 의미한다.

5. 결론 및 향후 연구

본 논문에서는 인간이 직관적으로 이해하기 힘든 특징 기반 학습환경을 화면 이미지 입력의 모방학습을 통해 Grad-CAM과 같은 시각화 방법을 적용하였다. 제안된 방법을 통해 LunarLander 환경에서 인간이 충분히 이해 가능한 시각화 결과를 이끌어 냈으며, 이를 통해, 물리 법칙 등이 적용된 환경과 같이 화면 입력만으로는 학습이 힘들거나, 학습 결과에 대해 특징분석을 제외한 설명이 어려운 환경에의 시각화분석에 대한 가능성을 보여준다. 이를 통해, 다음과 같은 몇 가지 연구에 대한 추가적인 기대를 할 수 있다.

1) *다른 특징기반 학습환경에 대한 시각화*: LunarLander와 같이 특징에 대한 이해와 분석은 가능하지만 시각적인 분석이 어려운 경우에 대해서, 모방학습을 통한 시각화에 연구를 추가적으로 진행할 수 있으며, 이를 일반적으로 적용하는 방법에 대한 연구를 기대할 수 있다. 예를 들어, OpenAI Gym환경의 BipedalWalker, CarRacing 등의 Box2D 환경이 있다.

2) 다른 신경망 시각화 방법을 적용: Grad-CAM 외에도 Saliency map, Grad-CAM++등의 방법을 통해 시각화 분석 결과를 도출할 수 있으며 이를 비교하는 연구를 기대할 수 있다.

3) *더 높은 수준의 설명 가능성 연구*: Grad-CAM은 시각화분석을 통해 에이전트가 취한 행동이 화면의 어느 부분을 집중해서 도출된 것인지를 보여주지만, 집중한 부분이 어떤 의미를 가지는지에 대해 설명하는 것은 아직 불가능하다. 이후, Object-Level RL이나 Object-oriented RL 등의 분야와의 결합을 통해 신경망이 집중한 부분과 이에 대한 추가적인 설명 가능성 연구를 기대해 볼 수 있다.

6. 감사의 글

이 논문은 2020년도 광주과학기술원의 재원으로 글로벌 선도대학 육성 사업의 지원을 받아 수행된 연구임

참고문헌

[1] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[3] Puri, Nikaash, et al. "Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution." International Conference on Learning Representations. 2019.

[4] Joo, Ho-Taek, and Kyung-Joong Kim. "Visualization of Deep Reinforcement Learning using Grad-CAM: How AI Plays Atari Games?." 2019 IEEE Conference on Games (CoG). IEEE, 2019.

[5] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[6] Schulman, John, et al. "Trust region policy optimization." International conference on machine learning. 2015.

[7] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.