월드 모델 기술 동향

ㅁ **박정인** / 대구경북과학기술원

요 약

인공지능 기술은 주어진 입력을 인식하고 분류하는 것을 넘어, 주변 환경이 어떻게 변화할 것인지를 예측하고 이를 기반으로 행동을 결정하는 방향으로 발전하고 있다. 이러한 지능을 구현하기 위한 핵심 기술 중 하나가 월드 모델(World Model)이다. 월드 모델은 관측된 환경의 상태와 에이전트의 행동을 바탕으로 미래 상태 또는 그 표현을 예측함으로써, 실제 환경과 반복적으로 상호작용하지 않고도 미래를 가상으로 경험하고 행동을 계획할 수 있도록 한다. 초기의 월드 모델 연구는 강화학습에서 환경의 동역학을 학습하여 데이터 효율적인 정책 학습과 계획을 수행하는 것을 주요 목적으로 하였다. 이후 잠재 공간에서 미래를 예측하는 PlaNet과 Dreamer 계열의 연구가 등장하였으며, 최근에는 대규모 영상 및 다중 데이터로부터 물리적, 공간적 변화를 학습하여 다양한 환경과 행동을 생성할 수 있는 범용 월드 모델로 그 범위가 확대되고 있다. 특히 자율주행, 로보틱스 및 embodied AI에서는 실제 환경을 대신하여 정책을 학습하고 검증할 수 있는 생성형 시뮬레이터로서 월드 모델의 중요성이 커지고 있다. 본고에서는 월드 모델의 기본 개념과 구성 요소를 살펴보고, 모델 기반 강화학습에서 시작하여 잠재 공간 예측, 표현 예측, 생성형 및 기반 월드 모델로 발전해 온 주요 기술 동향을 소개한다. 또한 자율주행 및 로보틱스를 포함한 주요 응용 분야를 살펴보고, 장시간 예측의 일관성, 물리적 정확성, 행동 제어 가능성, 공간 이해 및 평가 방법 등의 현재 한계와 향후 연구 방향에 대해 논의한다.

Ⅰ. 서 론

사람은 주변 환경을 단순히 관찰하는 것에 그치지 않고, 현재의 상황으로부터 다음에 발생할 일을 끊임없이 예측한다. 예를 들어, 사람이 움직이는 공을 잡으려고 할 때에는 공의 현재 위치뿐만 아니라 이동 방향과 속도를 통해 수 순간 뒤의 위치를 예상해야 한다. 또한 물체를 밀거나 문을 열려고 할 때에는 자신의 행동이 주변 환경에 어떠한 변화를 발생시킬지를 예측한다. 이러한 미래 예측 능력은 인식뿐만 아니라 계획, 추론 및 행동을 수행하기 위한 핵심적인 요소이다.

최근 인공지능은 영상 분류, 객체 검출 및 자연어 이해와 같이 주어진 정보를 해석하는 문제에서 뛰어난 성능을 보이고 있다. 그러나 로봇이나 자율주행 자동차와 같이 실

제 환경에서 동작하는 인공지능 시스템은 현재의 환경을 인식하는 능력만으로 충분하지 않다. 자신의 행동에 따라 환경이 어떻게 변화할지를 예측하고 여러 가능한 행동의 결과를 비교하여 적절한 행동을 선택할 수 있어야 한다. 이러한 목적으로 환경의 내부 모델(internal model)을 학습하는 기술을 월드 모델이라고 한다.

월드 모델의 개념은 오래전부터 모델 기반 강화학습(model-based reinforcement learning)의 중요한 구성 요소로 연구되어 왔으나, 현재 사용되는 의미의 '신경망 기반 월드 모델'은 [1]에서의 제안을 계기로 널리 알려졌다. 본 연구에서는 고차원의 영상 관측을 작은 잠재 표현으로 압축하고, 잠재 공간에서 시간에 따른 변화를 예측하는 모델을 학습하여 에이전트가 실제 환경뿐만 아니라 모델 내부에서 생성된 가상의 환경에서도 정책을 학습할 수 있음을 보였다. 이후 PlaNet[2], Dreamer[4-6] 등에서는 영상으로부터 직접 학습한 잠재 동역학(latent dynamics)을 이용하여 미래를 예측하고 행동을 계획하는 방법이 발전하였다. PlaNet은 영상 입력으로부터 학습된 잠재 공간에서 직접 계획을 수행함으로써 모델 기반 강화학습의 데이터 효율성을 향상시켰으며[2], Dreamer는 잠재 공간에서 생성한 가상의 미래 궤적을 이용하여 정책 자체를 학습하는 방식으로 이를 확장하였다.

한편 최근에는 월드 모델의 연구 범위가 강화학습을 위한 동역학 모델을 넘어 크게 확장되고 있다. 대규모 영상 생성 모델의 발전으로 미래 영상을 높은 품질로 생성할 수 있게 되면서, 영상 자체를 통해 실제 세계의 변화 법칙을 학습하려는 접근법이 활발하게 연구되고 있다. OpenAI의 Sora[16]는 대규모 영상 생성 모델이 물리적, 디지털 세계의 시뮬레이터로 발전할 가능성을 제시하였으며, Google DeepMind의 Genie는 행동 레이블이 없는 인터넷 영상으로부터 잠재 행동(latent action)을 학습하여 사용자의 입력에 따라 변화하는 가상 환경을 생성하였다. Genie는 110억 개의 파라미터를 가진 생성형 인터랙티브 환경 모델로 제안되었으며, 이를 저자들은 foundation world model의 한 형태로 설명하였다.

월드 모델의 응용 영역도 게임 환경에서 자율주행 및 로보틱스와 같은 실제 세계로 빠르게 확대되고 있다. GAIA-1은 영상, 텍스트 및 차량의 행동을 조건으로 미래의 주행 상황을 생성하도록 학습되었으며[11], DriveDreamer는 실제 자율주행 데이터로부터 교통 환경의 구조와 시간적 변화를 학습하였다[15]. UniSim은 영상 생성 모델을 실제 세계와 상호작용할 수 있는 시뮬레이터로 확장하여, 시뮬레이터에서 학습된 정책을 실제 로봇에 적용할 수 있는 가능성을 보였다. 최근 NVIDIA의 Cosmos와 같은 연구에서는 이러한 방향을 더욱 확장하여 범용적인 World Foundation Model을 Physical AI 학습을 위한 기반 모델로 활용하고자 한다.

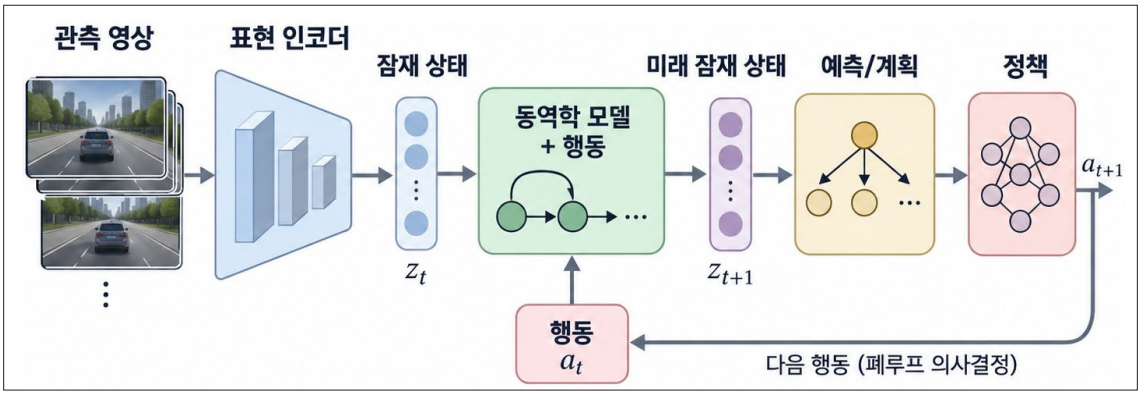
이에 따라 오늘날 월드 모델 기술의 발전 동향을 살펴보고자 한다. II장에서는 월드 모델의 기본적인 정의와 구성 요소를 살펴보고, III장에서는 모델 기반 강화학습, 잠재 공간 예측, 표현 기반 예측 및 대규모 생성형 월드 모델로 이어지는 주요 연구 흐름을 소개한다. IV장에서는 자율주행과 로보틱스 등 월드 모델의 주요 응용 분야를 살펴보고, V장에서는 현재 월드 모델이 가지고 있는 한계와 향후 연구 방향을 논의한다. 마지막으로 VI장에서 결론을 맺는다.

II. 월드 모델의 개념과 구성

월드 모델은 현재까지 관측한 환경과 에이전트의 행동을 바탕으로 미래 환경의 변화를 예측하는 모델로 정의할 수 있다. 일반적인 강화학습 환경에서는 시점 t에서의 상태를 s_t, 수행한 행동을 a_t라 할 때, 환경은 다음 상태 s_{t+1}과 보상 r_t를 생성한다. 월드 모델은 이러한 환경의 변화

$$(s_t, a_t) \rightarrow (s_{t+1}, r_t)$$

를 데이터로부터 학습한다.



<그림 1> 월드 모델의 기본 구조

그러나 실제 영상과 같은 고차원의 입력을 직접 미래 영상으로 예측하는 것은 계산량이 많으며 미래에 발생할 수 있는 모든 세부적인 변화를 정확하게 표현하기 어렵다. 따라서 최근의 월드 모델 대부분은 실제 관측값 자체가 아니라 이를 압축한 잠재 표현(latent representation)을 이용한다.

월드 모델은 크게 세 가지 구성 요소로 이해할 수 있다.

첫 번째는 환경의 표현 학습이다. 카메라 영상과 같이 복잡한 입력을 그대로 사용하는 대신 인코더를 통해 중요한 정보를 포함하는 저차원 잠재 상태로 변환한다. 좋은 잠재 표현은 물체의 형태나 위치뿐만 아니라 에이전트의 행동에 따라 변화할 수 있는 환경의 요소를 효과적으로 나타내야 한다.

두 번째는 동역학 모델(dynamics model)이다. 동역학 모델은 현재의 잠재 상태와 행동으로부터 미래 상태를 예측한다. 환경에는 불확실성이 존재하기 때문에 하나의 결정적인 미래만을 예측하기보다는 여러 가능한 미래를 표현할 수 있는 확률적 모델이 사용되기도 한다.

세 번째는 예측된 미래의 활용이다. 학습된 월드 모델은 여러 행동에 따른 미래를 실제 환경에서 수행하지 않고 가상으로 생성할 수 있다. 따라서 각 행동의 미래 결과를 비교하여 가장 적절한 행동을 선택하거나, 월드 모델이 만들어 낸 가상의 경험을 이용하여 정책을 학습할 수 있다.

초기의 월드 모델에서는 관측 영상을 다시 복원할 수 있도록 잠재 상태를 학습하는 생성형 접근법이 주로 이용되었다. 반면 최근에는 미래 영상의 모든 픽셀을 정확하게 생성하는 것보다 행동이나 추론에 필요한 정보만을 잠재 공간에서 예측하는 것이 더욱 효율적일 수 있다는 연구도 활발하다. 이러한 차이는 최근 월드 모델을 이해하는 데 중요한 두 가지 연구 방향을 형성하고 있다.

Ⅲ. 월드 모델 기술 동향

1. 모델 기반 강화학습과 잠재 동역학

초기의 모델 기반 강화학습에서는 현재 상태와 행동으로부터 다음 상태를 예측하는 모델을 학습한 후, 이 모델을 이용하여 적절한 행동을 탐색하였다. 그러나 영상과 같이 고차원의 관측을 사용하는 경우 정확한 미래 영상을 반복적으로 생성하는 것은 어렵고 작은 예측 오류가 누적되는 문제가 발생한다.

World Models[1]은 이러한 문제를 해결하기 위해 환경의 관측을 잠재 공간으로 압축하고 이 공간에서 시간적 변화를 학습하였다. 이 모델은 영상의 공간적 표현을 학습

하는 변분 오토인코더(variational autoencoder)와 잠재 상태의 시간적 변화를 예측하는 순환 신경망(recurrent neural network)으로 구성된다. 특히 실제 환경 대신 월드 모델이 생성한 가상의 환경, 즉 “dream” 내부에서 정책을 학습한 후 이를 실제 환경에 적용할 수 있다는 점을 보였다.

PlaNet[2]은 이러한 개념을 보다 체계적인 잠재 동역학 기반 계획 문제로 발전시켰다. PlaNet은 결정적 상태와 확률적 상태를 함께 이용하는 잠재 동역학 모델을 학습하고, 영상 대신 잠재 공간에서 미래 행동을 탐색한다. 이를 통해 복잡한 영상 기반 제어 문제에서 실제 환경과의 상호작용 횟수를 크게 줄일 수 있었다.

SimPLe[3]은 Atari 게임의 영상 예측 모델을 학습한 후 이 모델이 생성하는 환경에서 정책을 반복적으로 학습하는 방법을 제안하였다. 제한된 10만 번의 환경 상호작용 조건에서 당시 모델-프리 방법보다 높은 데이터 효율성을 보임으로써 영상 예측 모델이 강화학습에 효과적으로 활용될 수 있음을 보여주었다.

2. 잠재 상상을 이용한 정책 학습

월드 모델의 대표적인 발전은 Dreamer 계열의 연구에서 이루어졌다. Dreamer[4]는 실제 환경에서 정책을 반복적으로 실행하는 대신, 월드 모델의 잠재 공간에서 미래 궤적을 생성하는 잠재 상상(latent imagination)을 이용한다. 에이전트는 월드 모델이 예측한 미래 상태와 보상을 통해 정책을 업데이트하기 때문에 실제 환경과의 상호작용을 크게 줄일 수 있다.

DreamerV2[5]에서는 잠재 상태에 이산 표현(discrete representation)을 도입하여 Atari와 같이 복잡한 환경까지 적용 영역을 확장하였다. 특히 별도로 학습된 월드 모델 내부에서 정책을 학습하는 방법으로 Atari 55개 게임에서 인간 수준의 성능을 달성하였다.

이후 DreamerV3[6]는 서로 다른 관측 형태와 행동 공간 및 보상 범위를 갖는 다양한 환경을 하나의 알고리즘으로 다룰 수 있도록 학습 방법을 개선하였다. 2025년 Nature에 발표된 결과에서는 고정된 하이퍼파라미터를 사용하여 8개 도메인의 150개 이상의 과제에서 평가하였으며, Minecraft와 같은 복잡한 환경까지 월드 모델 기반 학습이 확장될 수 있음을 보여주었다.

한편 MuZero[7]는 환경 자체의 모든 요소를 정확히 복원할 필요가 없다는 또 다른 중요한 방향을 제시하였다. MuZero는 다음 영상 전체를 예측하는 대신 계획에 필요한 보상, 정책 및 가치(value)와 관련된 정보를 학습한다. 이를 통해 게임의 규칙이나 정확한 환경 모델이 주어지지 않은 상태에서도 Go, 체스, 쇼기 및 Atari에서 높은 성능을 달성하였다.

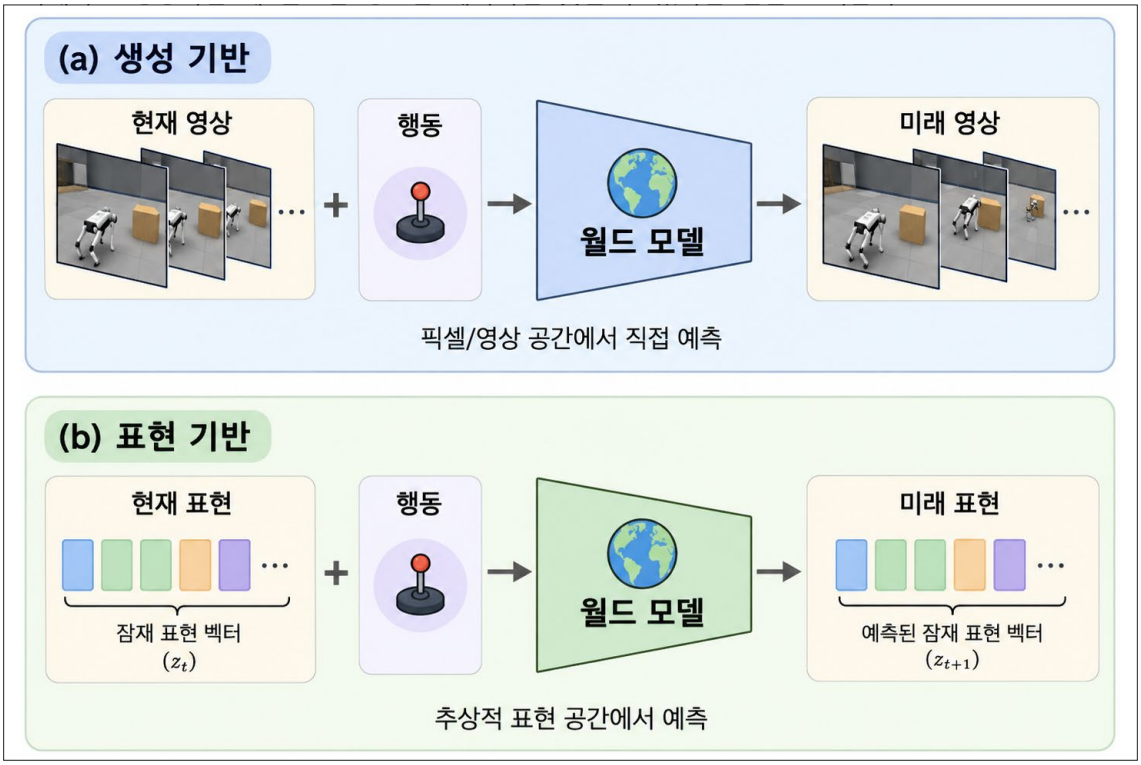
최근 TD-MPC2[8] 역시 모든 영상을 생성하는 대신 제어에 필요한 잠재 표현과 동역학을 학습하고 잠재 공간에서 Model Predictive Control을 수행한다. TD-MPC2는 104개의 연속 제어 과제에서 동일한 하이퍼파라미터를 사용하여 높은 성능을 보였으며, 하나의 317M 파라미터 모델을 80개의 서로 다른 과제와 embodiment에 학습함으로써 월드 모델의 확장 가능성을 보여주었다.

3. 픽셀 예측에서 표현 예측으로

월드 모델이 반드시 미래 영상을 생성해야 하는가에 대해서는 최근 중요한 논의가 이루어지고 있다. 영상에는 물체의 위치나 운동과 같이 미래 행동을 결정하는 데 중요한 정보뿐만 아니라 색상, 질감 및 작은 배경 변화와 같이 반드시 예측할 필요가 없는 정보도 포함되어 있기 때문이다.

I-JEPA[9]는 입력 영상 일부의 표현으로부터 보이지 않는 영역의 표현 자체를 예측하는 방식을 제안하였다. 픽셀을 직접 복원하지 않고 추상적인 표현 공간에서 예측함으로써 의미론적인 정보를 효과적으로 학습할 수 있음을 보여주었다.

이를 영상으로 확장한 V-JEPA[10]에서는 비디오의 가려진 시공간 영역에 해당하는 특징을 다른 영역으로부터



<그림 2> 월드 모델의 두 가지 예측 방식

예측한다. V-JEPA는 텍스트, 이미지 사전학습 모델 또는 픽셀 복원 없이 약 200만 개의 영상을 이용한 특징 예측만으로 학습되었으며, 동작 및 외형과 관련된 여러 영상 이해 과제에서 효과적인 표현을 학습하였다.

이러한 연구들은 월드 모델의 목적이 반드시 "미래 화면을 정확히 그리는 것"이 아니라 미래를 이해하고 행동하는 데 필요한 정보를 예측하는 것일 수 있다는 점을 보여준다.

생성 기반 모델은 사람이 직접 결과를 확인할 수 있고 가상 환경을 생성할 수 있다는 장점이 있다. 반면 표현 기반 모델은 계산량을 줄이고 과제 수행에 필요한 의미론적인 정보에 집중할 수 있다. 향후에는 두 접근법을 대립적인 관계로 보기보다, 목적에 따라 함께 사용하는 방향이 중요할 것으로 예상된다.

4. 생성형 월드 모델과 World Foundation Model

대규모 영상 생성 기술의 발전은 월드 모델 연구의 범위를 크게 변화시키고 있다. 기존의 월드 모델이 특정 환경에서 수집된 경험을 이용하여 해당 환경의 동역학을 학습하였다면, 최근 연구는 인터넷 규모의 영상과 서로 다른 환경의 데이터를 학습하여 다양한 세계를 하나의 모델로 표현하고 생성하는 범용 모델을 목표로 한다.

GAIA-1[11]은 자율주행 환경에서 이러한 가능성을 보여준 대표적인 연구이다. 영상, 텍스트 및 차량의 행동을 토큰으로 변환한 후 다음 토큰을 예측함으로써 실제와 유사한 주행 영상을 생성한다. 특히 차량의 행동을 조건으로 미래 환경의 변화를 생성할 수 있어 단순한 영상 생성 모델과 행동 조건부 월드 모델의 차이를 보여준다.

UniSim[12]은 서로 다른 종류의 데이터를 결합하여 현실 세계의 상호작용을 생성하는 범용 시뮬레이터를 학습하였다. 정적인 영상 데이터는 다양한 객체와 장면에 대한 정보를 제공하고, 로봇 데이터는 행동에 따른 변화에 대한 정보를 제공하며, 내비게이션 데이터는 공간 이동에 대한 정보를 제공한다. 이렇게 학습한 모델을 통해 고수준 언어 명령뿐 아니라 저수준 로봇 제어 입력에 따른 영상 변화를 생성하였으며, 생성된 환경에서 학습한 정책을 실제 로봇으로 전이할 수 있음을 보였다.

Genie[13]는 행동 레이블이 없는 인터넷 영상만을 이용하여 인터랙티브 환경을 학습하는 방향을 제시하였다. Genie는 시공간 비디오 토크나이저, 자기회귀 동역학 모델 및 잠재 행동 모델로 구성되어 있으며, 학습 데이터에 실제 행동 레이블이 없어도 영상의 변화로부터 행동에 대응하는 잠재 변수를 학습한다. 이를 통해 하나의 영상이나 그림으로부터 사용자가 행동을 입력하며 탐색할 수 있는 환경을 생성할 수 있다.

OpenAI의 Sora 역시 대규모 영상 생성 모델을 물리 세계의 시뮬레이터로 바라보는 관점을 제시하였다. Sora는 영상을 시공간 패치 형태로 표현한 뒤, 대규모 diffusion transformer를 학습한다. 일부 생성 결과에서는 물체의 상태 변화나 단순한 디지털 환경의 동역학이 나타났으나, 동시에 물리적으로 잘못된 상호작용이나 장시간 영상에서의 불일치도 보고되었다. 따라서 시각적으로 사실적인 영상 생성이 정확한 물리 세계 모델링과 동일하지 않다는 점도 보여준다.

최근 Cosmos[14]에서는 이러한 방향을 World Foundation Model이라는 개념으로 확장하였다. 범용적인 영상과 물리 세계의 변화를 사전 학습한 뒤 자율주행, 로봇 및 기타 Physical AI 환경에 맞게 후학습하는 방식을 제시한다. 이와 같은 접근은 언어 분야에서 범용 언어 모델을 다양한 하위 과제에 적용하는 것과 유사하게, 하나의 대규모 월드 모델을 다양한 물리적 응용에 활용하려는 시도로 볼 수 있다.

Ⅳ. 월드 모델의 주요 응용 분야

1. 자율주행

자율주행 자동차는 현재 카메라에 관측되는 객체를 검출하는 것뿐만 아니라 주변 차량과 보행자가 앞으로 어떻

<표 1> 월드 모델의 대표적인 발전 과정

구분	대표 연구	핵심 특징
초기 신경망 월드 모델	World Models [1]	압축된 잠재 공간에서의 환경 동역학 학습
잠재 공간 계획	PlaNet [2]	영상 기반 잠재 동역학 및 온라인 계획
영상 기반 MBRL	SimPLe [3]	영상 예측 모델을 이용한 Atari 정책 학습
Latent imagination	Dreamer [4-6]	가상의 잠재 궤적에서 정책 학습
과제 중심 모델	MuZero [7]	전체 환경 대신 reward/value/policy 예측
잠재 공간 MPC	TD-MPC2 [8]	제어에 필요한 implicit world model
표현 예측	I-JEPA [9] / V-JEPA [10]	픽셀 대신 의미론적 표현 예측
자율주행	GAIA-1 [11], DriveDreamer [15]	행동 조건부 미래 주행 환경 생성
범용 시뮬레이터	UniSim [12]	서로 다른 실제 데이터로 상호작용 모델링
Interactive WM	Genie [13]	비디오에서 잠재 행동과 환경 동역학 학습
World Foundation Model	Cosmos [14]	Physical AI를 위한 범용 사전학습 모델

게 움직일 것인지 예측해야 한다. 또한 차량의 가속, 감속 또는 조향 행동에 따라 미래의 교통 상황이 어떻게 변화하는지도 고려해야 한다. 따라서 자율주행은 월드 모델이 가장 직접적으로 활용될 수 있는 분야 중 하나이다.

GAIA-1은 실제 주행 영상과 텍스트 및 차량 행동으로부터 미래 주행 환경을 생성하였으며[11], DriveDreamer[15]는 diffusion model을 이용하여 실제 교통 환경의 구조적 조건과 미래 상태를 함께 모델링하였다. DriveDreamer에서 생성된 데이터는 영상 생성에만 사용되는 것이 아니라 자율주행 인식 모델을 추가로 학습하는 데이터로도 활용되었다.

월드 모델을 이용하면 실제 도로에서 경험하기 어렵거나 위험한 상황을 가상으로 생성할 수 있다는 장점도 있다. 예를 들어 갑작스럽게 도로에 진입하는 보행자, 악천후 또는 드물게 발생하는 교통 상황을 생성하여 자율주행 정책을 반복적으로 검증할 수 있다. 따라서 향후 자율주행 월드 모델은 단순한 미래 영상 생성에서 closed-loop 정책 학습 및 검증 환경으로 발전할 가능성이 크다.

2. 로보틱스 및 Embodied AI

로봇은 자신의 행동이 환경을 직접 변화시키기 때문에 월드 모델의 활용 가능성이 더욱 크다. 예를 들어 로봇이 컵을 집으려고 할 때 "손을 어느 방향으로 이동하면 어떤 결과가 발생할 것인가"를 실제로 반복 실행하기 전에 월드 모델 내부에서 예측할 수 있다면 데이터 수집에 필요한 시간과 비용을 크게 줄일 수 있다.

UniSim에서는 월드 모델이 생성한 가상의 상호작용을 이용하여 로봇 정책을 학습하고 이를 실제 로봇으로 전이하는 실험을 수행하였다. 또한 TD-MPC2와 Dreamer 계열 연구는 잠재 공간에서의 미래 예측이 다양한 제어 문제에 효과적으로 활용될 수 있음을 보여준다.

특히 실제 로봇 데이터는 인터넷 영상이나 텍스트 데이터와 비교하면 수집 비용이 매우 크다. 이에 따라 대규모 일반 영상으로부터 물리적 변화에 관한 지식을 먼저 학습하고, 소량의 로봇 행동 데이터로 행동 조건부 모델을 학습하는 방법은 향후 로봇 월드 모델의 중요한 연구 방향이 될 것으로 예상된다.

3. 에이전트 학습을 위한 가상 환경

현재 강화학습과 embodied agent 연구의 중요한 제약 중 하나는 다양한 환경을 구축하고 상호작용 데이터를 수집하는 데 많은 비용이 필요하다는 점이다.

Genie와 같은 생성형 월드 모델은 기존 게임 엔진으로 일일이 환경을 제작하는 대신 데이터로부터 새로운 환경 자체를 생성할 수 있는 가능성을 보여준다. Genie의 후속 방향은 이러한 모델을 이용하여 에이전트가 학습할 수 있는 다양하고 새로운 환경을 자동으로 제공하는 것이다. Google DeepMind가 이후 공개한 Genie 2[17]와 Genie 3[18]도 이러한 방향을 확장하여 행동으로 탐색 가능한 3차원 환경 및 실시간 인터랙티브 환경을 목표로 하고 있다. Genie 3는 텍스트 입력으로부터 720p 환경을 생성하고 24 fps로 수 분 동안 탐색 가능한 연구 시스템으로 소개되었다.

이러한 기술이 발전한다면 앞으로의 인공지능 에이전트는 사람이 구축한 제한적인 시뮬레이터에서만 학습하는 것이 아니라 월드 모델이 생성하는 무수히 많은 새로운 환경을 경험하면서 학습할 수 있을 것이다.

Ⅴ. 현재의 한계와 향후 연구 방향

최근 월드 모델은 매우 빠르게 발전하고 있지만, 실제 세계를 정확하게 이해하고 장시간 시뮬레이션하기 위해서는 아직 해결해야 할 문제가 많다.

1. 시각적 사실성에서 물리적 정확성으로

현재 대규모 영상 생성 모델은 매우 사실적인 영상을

생성할 수 있지만 영상이 자연스럽게 보인다는 사실이 물리적으로 올바른 세계를 학습했다는 것을 의미하지는 않는다.

예를 들어 물체를 들어 올렸을 때 다른 물체를 통과하거나, 물체가 갑자기 생성 또는 소멸하거나, 충돌 이후의 움직임이 실제 물리 법칙과 다르게 나타날 수 있다. Sora에서도 유리 파손이나 음식 섭취와 같은 기본적인 상호작용을 정확하게 시뮬레이션하지 못하는 사례가 보고되었다.

따라서 향후 월드 모델에서는 영상 품질 자체보다 물체의 지속성(object permanence), 인과 관계, 중력, 접촉 및 충돌과 같은 물리적 일관성을 학습하는 것이 중요하다.

2. 장시간 예측과 누적 오류

월드 모델은 자신의 예측 결과를 다시 입력으로 사용하여 다음 미래를 생성한다. 따라서 작은 오류가 반복적으로 누적되면서 시간이 길어질수록 실제 환경과 다른 상태로 발전하는 문제가 발생한다.

단기적인 영상 예측에서는 이러한 차이가 크게 나타나지 않을 수 있지만, 수십 또는 수백 단계 이후의 행동을 계획해야 하는 로봇 및 자율주행에서는 중요한 문제가 된다.

향후에는 짧은 시간 단위의 정확한 동역학뿐 아니라 장기간 유지되어야 하는 환경의 기억(memory)과 전역적인 상태(global state)를 명시적으로 관리하는 기술이 필요하다.

3. 관찰에서 행동으로의 확장

인터넷에는 막대한 양의 영상이 존재하지만 대부분의 영상에는 영상 속 사람이 실제로 어떤 행동 명령을 수행했는지에 대한 정보가 없다. 그러나 월드 모델이 에이전트의 행동 결과를 예측하려면

현재 상태 + 행동 → 미래 상태

의 관계를 학습해야 한다.

Genie는 영상에서 관측되는 변화를 이용하여 잠재 행동을 자동으로 학습함으로써 이러한 문제를 완화하였다. 그러나 잠재 행동이 실제 로봇의 세밀한 연속 제어와 직접 대응한다고 보기는 어렵다.

따라서 대규모 행동 레이블이 없는 영상에서 얻은 지식을 소량의 실제 행동 데이터와 연결하는 action grounding 기술이 중요한 연구 방향이 될 것이다.

4. 2차원 영상에서 3차원 공간 월드 모델로

현재 많은 월드 모델은 영상을 통해 환경을 표현하지만 실제 세계의 공간은 본질적으로 3차원이다. 카메라가 움직이거나 물체가 일시적으로 시야 밖으로 사라진 경우에도 해당 물체의 위치와 상태는 계속 존재한다.

따라서 미래의 월드 모델은 단순한 영상 프레임의 연속성을 넘어 장면의 3차원 구조, 객체의 위치, 카메라 움직임 및 지속적인 공간 기억을 이해해야 한다.

이는 embodied AI에서 특히 중요하다. 로봇이 방 안을 이동하면서 이전에 보았던 물체의 위치를 기억하고, 다른 시점에서 다시 관찰하더라도 동일한 물체임을 이해할 수 있어야 하기 때문이다. 향후에는 비디오 기반 월드 모델과 3차원 표현 및 공간 추론 모델의 결합이 중요한 연구 주제가 될 것으로 예상된다.

5. 결정적인 미래에서 다양한 미래로

실제 환경의 미래는 하나로 결정되지 않는다. 예를 들어 도로 옆에 서 있는 보행자는 그대로 서 있을 수도 있고 도로를 건널 수도 있다. 하나의 미래만 예측하는 월드 모델은 이러한 불확실성을 충분히 표현하지 못한다.

따라서 월드 모델은

$$p(s_{t+1:t+H} | s_t, a_{t:t+H-1})$$

와 같이 여러 가능한 미래의 분포를 모델링할 필요가 있다.

생성 모델은 자연스럽게 서로 다른 미래를 생성할 수 있다는 장점이 있지만, 각각의 미래가 얼마나 가능성이 높은지 또는 정책 결정에 어느 정도 신뢰할 수 있는지를 추정하는 문제는 여전히 남아 있다. 이에 따라 불확실성을 명시적으로 표현하는 probabilistic world model이 안전이 중요한 자율주행과 로봇 분야에서 더욱 중요해질 것으로 예상된다.

6. 범용 월드 모델과 Physical AI

언어 모델의 발전에서 중요한 변화 중 하나는 특정 과제를 위한 작은 모델을 각각 학습하는 것에서 대규모 데이터를 이용하여 하나의 foundation model을 학습한 후 여러 과제에 활용하는 방식으로 전환된 것이다.

월드 모델에서도 유사한 변화가 나타나고 있다. Cosmos는 범용 World Foundation Model을 먼저 학습한 후 Physical AI의 여러 환경에 맞게 후학습하는 방향을 제안하였다.

향후에는 하나의 월드 모델이 로봇, 자율주행, 게임 및 가상 환경에서 얻어진 지식을 공유하면서 새로운 환경의 동역학을 적은 데이터로 빠르게 학습할 가능성이 있다. 이를 위해서는 서로 다른 embodiment와 행동 공간을 하나의 표현 공간으로 통합하는 문제가 해결되어야 한다.

7. 월드 모델의 새로운 평가 방법

월드 모델의 발전을 위해서는 무엇을 “좋은 월드 모델”이라고 평가할 것인지도 중요한 문제이다.

영상 생성 모델에서는 FID 또는 FVD와 같이 생성 영상의 품질을 평가하는 지표가 사용될 수 있지만 이러한 지표만으로 물리적 정확성이나 행동에 대한 인과 관계를 평가하기는 어렵다.

향후에는 다음의 특성을 함께 평가할 필요가 있다.

- 시각적 품질: 생성된 환경이 얼마나 자연스러운가
- 시간적 일관성: 물체와 환경의 상태가 장시간 유지되는가
- 물리적 일관성: 행동의 결과가 실제 물리 법칙과 일치하는가
- 행동 제어 가능성: 동일한 상태에서 서로 다른 행동이 올바른 미래를 생성하는가
- 공간적 일관성: 시점이 변해도 3차원 환경이 일관되는가
- 정책 유용성: 월드 모델에서 학습하거나 계획한 행동이 실제 환경에서도 효과적인가

특히 로보틱스에서는 생성 영상의 품질보다 해당 월드 모델을 이용했을 때 실제 정책의 성공률이 향상되는가가 보다 직접적인 평가 기준이 될 수 있다.

VI. 결 론

본고에서는 최근 인공지능 분야에서 빠르게 발전하고 있는 월드 모델의 개념과 주요 기술 동향을 살펴보았다. 월드 모델은 초기에는 강화학습에서 환경의 상태 변화를 예측하여 데이터 효율적으로 정책을 학습하기 위한 기술로 연구되었다. World Models, PlaNet 및 Dreamer와 같은 연구를 통해 고차원의 영상 관측을 압축된 잠재 공간으로 표현하고 이 공간에서 미래 상태를 예측하거나 가상의 경험을 생성하는 기술이 발전하였다.

이후 MuZero와 TD-MPC 계열의 연구는 실제 영상의 모든 세부 정보를 복원하지 않더라도 계획과 행동에 필요한 정보만을 효과적으로 예측할 수 있음을 보여주었으며, I-JEPA와 V-JEPA에서는 이러한 관점을 확장하여 픽셀보다 추상적인 표현을 직접 예측하는 방향을 제시하였다.

최근에는 대규모 생성 모델의 발전과 함께 월드 모델의 범위가 더욱 확대되고 있다. GAIA-1과 DriveDreamer

는 실제 자율주행 환경의 미래를 생성하고, UniSim은 로봇의 상호작용을 생성하는 범용 시뮬레이터를 제안하였다. Genie는 행동 레이블이 없는 대규모 영상에서도 상호작용 가능한 환경을 학습할 수 있음을 보여주었으며, Cosmos와 같은 연구는 이러한 기술을 다양한 Physical AI 시스템에 적용할 수 있는 World Foundation Model로 확장하고 있다.

그러나 현재의 월드 모델이 시각적으로 자연스러운 미래를 생성한다고 해서 실제 세계의 물리 법칙과 인과 관계를 완전히 이해한다고 보기는 어렵다. 장시간 예측에서 발생하는 오류의 누적, 객체와 공간에 대한 지속적인 기억, 행동과 환경 변화 사이의 인과 관계, 다양한 미래에 대한 불확실성 및 3차원 공간 이해와 같은 문제는 여전히 해결되어야 한다.

앞으로 월드 모델은 단순히 미래의 영상을 생성하는 기술을 넘어, "행동을 수행했을 때 세계가 어떻게 변화할 것인가"를 내부적으로 시뮬레이션하는 범용적인 인공지능 모델로 발전할 것으로 예상된다. 이러한 기술이 충분히 발전한다면 로봇과 자율주행 자동차와 같은 Physical AI 시스템은 실제 환경에서 모든 행동을 시행착오를 통해 경험하지 않고도 가상의 미래를 이용하여 학습하고 계획할 수 있을 것이다. 따라서 월드 모델은 향후 인공지능이 실제 세계를 이해하고 상호작용하기 위한 핵심 기술 중 하나가 될 것으로 기대된다.

참 고 문 헌

[1] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in Advances in Neural Information Processing Systems, 2018.

[2] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in International Conference on Machine Learning, 2019.

[3] Ł. Kaiser, M. Babaeizadeh, P. Miłos, B. Osiński, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, et al., "Model based reinforcement learning for Atari," in International Conference on Learning Representations, 2020.

[4] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in International Conference on Learning Representations, 2020.

[5] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with discrete world models," in International Conference on Learning Representations, 2021.

[6] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse control tasks through world models," Nature, vol. 640, pp. 647-653, 2025.

[7] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, "Mastering Atari, Go, chess and shogi by planning with a learned model," Nature, vol. 588, pp. 604-609, 2020.

[8] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, robust world models for continuous control," in International Conference on Learning Representations, 2024.

[9] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[10] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," Transactions on Machine Learning Research, 2024.

[11] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "GAIA-1: A generative world model for autonomous driving," arXiv preprint arXiv:2309.17080, 2023.

[12] S. Yang, Y. Du, S. Ghasemipour, J. Tompson, L. Kaelbling, D. Schuurmans, and P. Abbeel, "Learning interactive real-world simulators," in International Conference on Learning Representations, 2024.

참 고 문 헌

[13] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, et al., “Genie: Generative interactive environments,” in International Conference on Machine Learning, 2024.
[14] NVIDIA, “Cosmos world foundation model platform for physical AI,” arXiv preprint arXiv:2501.03575, 2025.
[15] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, “DriveDreamer: Towards real-world-driven world models for autonomous driving,” in European Conference on Computer Vision, 2024.
[16] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, et al., “Video generation models as world simulators,” Technical Report, OpenAI, 2024.
[17] J. Parker-Holder et al., “Genie 2: A large-scale foundation world model,” Google DeepMind, 2024.
[18] J. Parker-Holder and S. Fruchter, “Genie 3: A new frontier for world models,” Google DeepMind, 2025.

저 자 소 개



박 정 인

- 2012년 ~ 2018년 : 연세대학교 전기전자공학부 학사
- 2018년 ~ 2024년 : 연세대학교 전기전자공학과 박사
- 2024년 ~ 2025년 : 연세대학교 박사후연구원
- 2025년 ~ 2026년 : 연세대학교 연구교수
- 2026년 ~ 현재 : 대구경북과학기술원 (DGIST) 전기전자컴퓨터공학과 조교수
- 주관심분야 : 컴퓨터비전, 멀티모달 인공지능, 로보틱스