2024 년 봄 POSTECH 컴퓨터공학과 과제연구 연구제안서

Next action anticipation using diffusion models

학 번: 20210056
이 름: 박윤아
연구 지도교수: 조민수

연구 목적 (Problem statement)

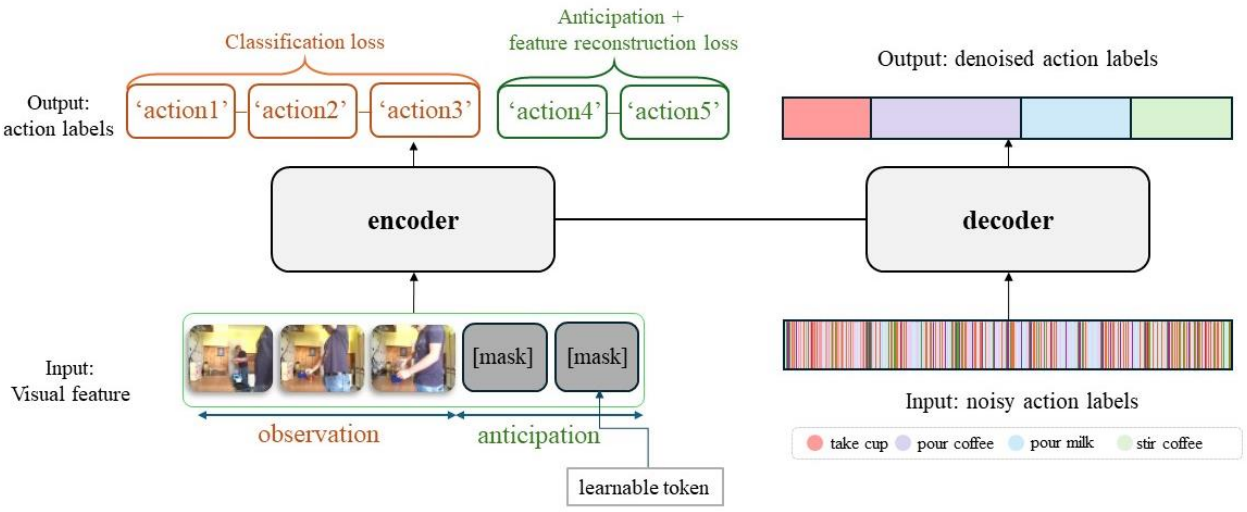
본 연구의 목적은 diffusion model 을 활용하여 next action anticipation task 의 성능을 향상시키는 것이다. Next action anticipation 이란 주어진 비디오 데이터를 관찰하여 약 10 초 간의 정보를 기반으로, 다음 1 초 후에 발생할 action 의 class 을 예측하는 작업이다. 우리는 더 높은 정확도의 action anticipation 을 위한 새로운 접근 방식으로, diffusion model 의 생성 학습 능력을 활용하여 long temporal context 를 효과적으로 모델링하는 방법을 제안한다. 이러한 future action anticipation 은 증강 현실이나 자율주행자동차 등의 AI system 에 사용된다. 예를 들어, 사용자가 요리를 하거나 가구를 조립하는 동안 AR device 는 wearable camera 로 관찰되는 정보를 바탕으로 사용자의 다음 단계를 예측하여 적절한 지원을 제공해야 한다. Action anticipation 은 사용자 경험을 향상시키고 상호작용을 개선하는 데 중요한 역할을 한다.

연구 배경 (Motivation and background)

비디오 데이터로부터 미래의 action 을 예측하는 task 는 인공지능 에이전트가 사람들과 상호작용하는 데 중요하다. 이 task 는 action 의 multi-modal distribution 를 예측하는 것뿐만 아니라, 미래의 action 을 예측하기 위해 본질적으로 과거 action 에 대한 모델링을 요구한다. 오늘날의 top performing video 모델이 동일한 test clip 에서 각각 action 인식과 예측을 다룰 때, 정확도가 42%에서 17%로 감소하는 등, action anticipation 는 어려운 task 임을 확인할 수 있다.

Diffusion model 은 데이터에 점진적으로 노이즈를 추가하고 손상시키는 forward process 와 fully random noise 로부터 iterative denoising 을 거쳐 새로운 샘플을 생성하는 reverse process 로 이루어져 있다. 이러한 reverse process 에서의 iterative denoising 은 결과를 점진적으로 개선하는 iterative refinement property 를 가지고 있다. 기존의 next action anticipation 연구들은 주로 한 번의 추론으로 행동을 예측하는 모델을 연구해왔다. 본 연구에서는 diffusion framework 의 iterative refinement 를 모델에 적용하여 예측 정확도를 향상시키고, action anticipation task 의 새로운 접근 방식을 제안하고자 한다.

연구 방법 (Research proposal)



- Overall Architecture
  본 연구에서는 diffusion action segmentation 모델을 baseline 으로 활용한다. 전반적인 모델의 구조는 다음과 같다. Encoder 에서 visual feature 를 input 으로 받아 embedding 을 한 이후, decoder 의 diffusion model 이 noisy action label 을 input 으로 받는다. Decoder 의 input 은 encoder output 인 visual embedding 에 condition 되어 결과적으로 원하는 target action 을 prediction 하게 된다.
- Encoder
  - Video 의 visual feature 를 input 으로 받는다. 이때, 예측해야 하는 미래 frame 에 해당하는 visual token 들의 경우 learnable mask embedding 으로 대체된다.
  - Input feature 는 encoder 로 들어가고, encoder layer 를 통해 video frame feature 들 사이의 relation 을 embedding 한 feature 로 encoding 된다.

- Output embedding 에 linear classifier 를 적용하여 encoder 단에서도 observation 된 visual token 에 대해서는 classification 을 하고, unobserved visual token 에 대해서는 anticipation 을 하게 된다. Anticipation 에는 class prediction 과 더불어 feature reconstruction 도 함께 적용할 예정이다.

- Decoder
  - Decoder 에서는 diffusion forward process 를 통해, one-hot action label 에 gaussian noise 가 더해진 noisy action label 을 input 으로 받게 된다.
  - Noise action label 의 input 으로 받아, decoder layer 에서는 encoder 의 output 으로 나온 visual embedding 에 condition 하여 denoising 을 하게 된다.
  - 최종적으로 decoder 에서 나온 denoised action label 을 예측 결과로 사용한다.

Head	Backbone	Init	Top-1	Top-5	Recall
RULSTM [24]	TSN	IN1k	13.1	30.8	12.5
ActionBanks [77]	TSN	IN1k	12.3	28.5	13.1
AVT-h	TSN	IN1k	13.1	28.1	13.5
AVT-h	AVT-b	IN21+1k	12.5	30.1	**13.6**
AVT-h	irCSN152	IG65M	**14.4**	**31.7**	13.2

Table 4: EK55 using only RGB modality for action anticipation. AVT performs comparably, and outperforms when combined with a backbone pretrained on large weakly labeled dataset.

- Model evaluation
  모델 평가 데이터셋은 Epic-Kitchen 55 을 사용하며, TSN backbone, Init IN1k 방법의 Top-1, Top-5, Recall metric 과 비교하여 성능을 향상시키는 것을 목표로 한다.

기대 효과 (Expected output)

Diffusion model 은 정보의 전파가 인접한 element 를 넘어 멀리 떨어진 element 까지 영향을 미치는 방식이고, 이로 인해 멀리 떨어져 있는 element 들 사이에 long-range dependency 가 형성된다. 미래 action 을 예측할 때, 사람은 일반적으로 action sequence 전체에 걸친 long-term relation 을 고려한다. 따라서 이미 관찰된 action 뿐만 아니라 미래의 가능한 action 들도 중요하다. 목표 모델은 learnable token 과 long-range dependency 를 통해, 시간이 지남에 따라 발생하는 복잡한 action 패턴 및 관계를 인식하고 예측하는데 긍정적인 발전을 보일 것으로 예상된다.

또한 추론적 접근으로도 diffusion framework 의 효과를 기대할 수 있다. Diffusion model 은 random noise 로부터 시작하는 generative model 이며, action anticipation 역시 random process 이다. 따라서 다양한 가능성이 있는 미래의 action 을 생성하는 diffusion model 은, deterministic 하게 하나의 결과를 예측하는 대신 다양한 예측을 수행한다. 동일한 observation 에서도 여러 가능성을 고려하는 diffusion model 은 action anticipation task 에 새로운 가능성을 제시하는 framework 로 발전할 것으로 기대된다.

참고 문헌

[1] GIRDHAR, Rohit; GRAUMAN, Kristen. Anticipative video transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 13505-13515.

[2] LIU, Daochang, et al. Diffusion action segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 10139-10149.

[3] FURNARI, Antonino; FARINELLA, Giovanni Maria. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: Proceedings of the IEEE/CVF International conference on computer vision. 2019. p. 6252-6261.

[4] BAO, Hangbo, et al. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

연구 추진 일정

3 월 15 일 ~ 4 월 15 일	전반적인 Model 설계 및 개발
4 월 16 일 ~ 5 월 10 일	Model 훈련과 실험
5 월 11 일 ~ 5 월 31 일	Model 성능 개선 및 결과 분석