적대적 공격을 이용한 VCM 비디오 부호화 분석

*추현곤, 임한신, 이진영, 이희경, 정원식, 서정일
한국전자통신연구원
*hyongonchoo@etri.re.kr

Analysis on Video coding for machines using Adversarial Attack

*Hyon-Gon Choo, Hanshin Lim, Jin Young Lee, Lee Hee Kyung,
Won-Sik Cheong, Jeongil Seo
Electronics and Telecommunications Research Institute

요 약

MPEG(Moving Pictures Experts Group)에서는 딥러닝을 포함한 머신 비전과 관련하여 Video for machines 란 이름의 새로운 부호화 표준에 대한 논의를 진행하고 있다. VCM 에서는 기존의 비디오 부호화와 달리 머신을 기준으로 한 비디오 부호화를 목표로 한다. 본 논문에서는 적대적 공격 모델을 이용하여 VCM 부호화에 대해서 분석을 하고자 한다. 적대적 공격 모델 관점에서 비디오 부호화의 특성에 대해서 살펴보고, 이를 고려한 부호화 개발 방향에 대해 살펴본다.

1. 서론

머신 비전은 기계(머신)에 인간이 가지고 있는 시각에 따른 판단 기능을 부여한 것으로 사람이 인지하고 판단하는 기능을 소프트웨어 또는 하드웨어 형태의 기계가 대신 처리하는 기술이다. 최근 딥러닝 기반의 기술 개발은 머신 비전의 정확도의 한계를 높여 사람이 처리할 수 있는 정확도의 범위를 뛰어넘는 계기를 마련하였으며, 기존의 산업 자동화 이외에 자율주행자동차, 스마트시티, 영상감시 등 응용 환경을 넓혀가고 있다. 더욱이 모바일 단말 및 통신 기술의 발달은 개개인이 생산하는 많은 비디오 데이터에 대한 머신 비전에 대한 서비스 요구가 점점 늘어나고 있으며, 빅데이터로 인해 넘쳐나는 데이터는 더더욱 컴퓨터를 기반한 기계의 비디오 처리의 수요는 점점 늘어가고 있는 상황이다. 이와 관련하여 멀티미디어 데이터에 대한 표준 기술을 개발하는 MPEG 에서도 딥러닝을 포함한 머신 비전과 관련하여 Video for machines(VCM)란 이름의 새로운 표준 기술에 대한 논의를 진행 중에 있다.

VCM 에서는 인지적 화질을 개선하는 것을 목표로한 기존의 비디오 부호화와 달리 머신의 성능을 기준으로 한 비디오 부호화를 목표로 한다. 이 경우, 다양한 머신에 따라서 성능을 어떻게 바라볼 것인가 하는 부분에 대해서 살펴볼 필요가 있다. 이와 관련하여 본 논문에서는 딥러닝에서의 적대적 공격 모델을 이용하여 VCM 부호화에 대해서 분석을 하고자 한다.

본 논문의 구성은 다음과 같다. 2 절에서는 적대적 공격 모델에 대해 간략히 살펴본 후, 3 절에서는 이러한 기법을 이용하 VCM 에 적용할 때 나타나는 특징에 대해 설명한다. 마지막으로 4 절에서는 본 논문에 대한 결론을 맺는다.

2. 적대적 공격 모델(Adversarial Attack Model)

딥러닝에서 적대적 공격이란 신경망을 혼란시킬 목적으로 만들어진 특수한 입력으로, 신경망으로 하여금 샘플을 잘못 분류하도록 하는 것을 의미한다. 일반적으로 인간에게 적대적 샘플은 일반 샘플과 큰 차이가 없어 보이지만, 신경망은 적대적 샘플을 올바르게 식별하지 못합니다. 이러한 특징은 예전의 워터마킹과 유사하다. 다만 워터마킹의 경우, 입력된 정보를 다시 찾아내기 위한 목적을 가지는 반면, 적대적 샘플은 입력된 정보를

통해 딥러닝 네트워크가 오동작을 일으키게 하는 특징을 가진다. 그림 1 은 [1]에서 설명한 적대적 모델에 대한 예이다.

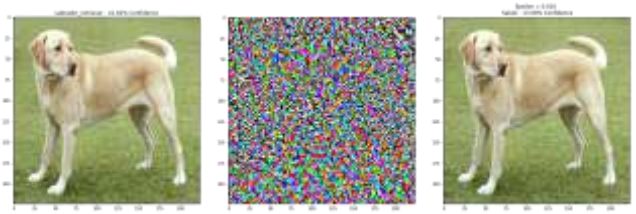

그림 1. 적대적 공격 모델의 예 [1](좌:원본, 가운데: 적대적 노이즈, 우:노이즈가 합성된 영상@ε=0.1)

그림 1 의 예에서 보는 것과 같이 사람에 눈에 띄지 않는 적당한 노이즈(Noise)를 이용하면 딥러닝 머신 비전이 제대로 된 성능을 내지 못하도록 하게한다. 이러한 노이즈를 생성하는 과정으로 Goodfellow 등은 딥러닝 네트워크의 Gradient 를 이용한 Fast gradient sign method (FGSM)를 제안하였다[1]. Fast gradient sign method 에 따른 적대적 공격 노이즈는 다음과 같이 표현이 가능하다.

$$x^{\text{adv}} = x + \epsilon \,\text{sign}(\nabla_x J\,(x, y_{\text{true}})) \quad (1)$$

(1)의 수직은 원래의 이미지에 gradient 를 더하는 방식으로 일반적인 딥러닝의 훈련 시 파라미터 최적 솔루션을 찾아가는 방식이 gradient 를 이용하여 최저점(local minima)를 찾아내는 과정으로 볼 때, FGSM 은 local minima 의 반대 방향으로 샘플의 이동으로 해석할 수 있다. 다음 장에서 적대적 공격 모델을 이용한 VCM 에 대한 해석을 살펴본다.

3. 적대적 공격 모델(Adversarial Attack Model)을 이용한 VCM 해석

머신 비전의 입장에서는 비디오 코딩도 일종의 영상 처리로 간주될 수 있다. 예를 들어 그림 1 의 팬더 영상이 부호화기에 입력될 때, 최종적으로 복호화기를 통해 재현된 영상은 부호화 및 복호화를 거쳐 나타난 에러의 합으로 표현이 가능하다. 이 때 에러는 하나의 노이즈로 표시할 수 있으며, 노이즈를 생성하기 위한 파라미터는 bitrate(R) 로 표현할 수 있다.

$$\mathrm{x}_{\text{decoded}} = x + \epsilon_c(R) \quad (2)$$

일반적으로 bitrate 가 적을수록 부호화로 인한 노이즈는 크게 되며, 노이즈의 세기가 클수록 머신 성능은 떨어진다. 그림 2 에서 비디오 부호화와 머신 비전 성능의 예를 보여준다. 그림 2 에서 데이터셋은 Cityscape 데이터셋[3]을 사용하였으며, 객체 검출은 Detectron2 에 내장된 Mask-RCNN 을 이용하였다[4]. 비디오 코딩 성능을 위해서 VTM8.2 와 HEVC 를 비교하였다.

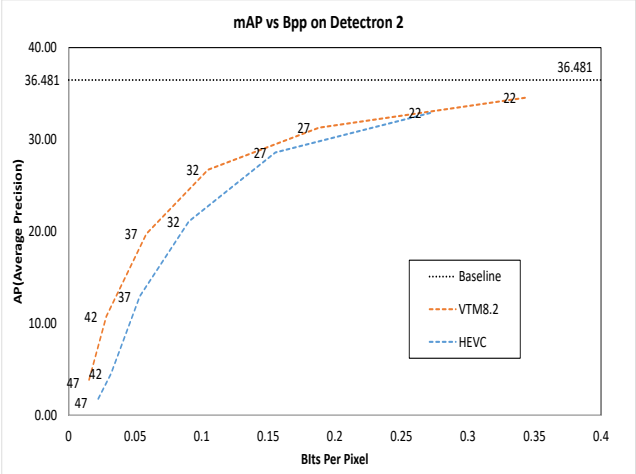

그림 2. 비디오 코딩에 따른 머신비전 성능 비교

그림 2 의 결과에 따르면 비트레이트(R)이 적어짐에 따라 머신 비전의 성능은 떨어지게 된다. 이를 다시 정리하면 머신 비전의 성능을 떨어뜨리는 노이즈를 생성하는 과정으로 볼 수 있게 되며, 비디오 코딩 자체도 머신 비전 입장에서는 하나의 노이즈 모델, 즉 공격 모델로 해석이 가능하다고 볼 수 있게 된다. 이와 같은 결과를 머신 비전 최적화 과정과 고려할 때 부호화 과정은 하나의 적대적 공격 노이즈와 유사한 노이즈 모델로 볼 수 있다. 부호화로 인한 성능의 하락을 Gradient 반대 방향으로 움직이는 벡터로 표시한다고 할 때, 벡터의 길이가 코덱의 성능을 나타낼 수 있다고 해석할 수 있다. 동일한 비트의 데이터에 대해서 머신의 성능을 많이 떨어뜨리는 코덱의 경우, 더 강력한 Attack 모델로 간주될 수 있다. 그림 2 의 결과에 따르면 HEVC 은 VTM 에 비해 더 강한 공격을 생성하는 것으로 볼 수 있으며, 일반적으로 코덱의 성능 자체는 원본 영상의 노이즈의 크기를 줄이는 방향으로 설계되는 것을 고려할 때 비디오 코덱의 성능의 좋을수록 원본 영상을 이용한 성능에 가깝게 갈 수 있다는 것을 추론할 수 있다. 또 한가지로는 고려사항으로는 같은 크기의 노이즈라고 하더라도 코덱에서 발생되는 에러(노이즈)가 실제 머신 비전 네트워크에서 Gradient 방향이 적게 포함되는 방향으로 설계하는 것도 고려해 볼 수 있다.

4. 결 론

딥러닝에서 적대적 공격이란 신경망을 혼란시킬 목적으로

만들어진 특수한 입력으로, 신경망으로 하여금 샘플을 잘못 분류하도록 하는 것을 의미한다. 본 논문에서는 적대적 공격 모델을 이용하여 VCM 부호화에 살펴보았다. 머신 비전의 입장에서는 비디오 코딩도 일종의 적대적 공격 수행하는 노이즈로 간주될 수 있으며, 동일한 비트의 데이터에 대해서 머신의 성능을 많이 떨어 뜨리는 코덱의 경우, 더 강력한 Attack 모델로 간주될 수 있음에 대해서 확인할 수 있었다. 향후 이와 관련하여 적대적 공격 모델과 VCM 부호화 성능 지표와의 연관성에 대해 연구를 진행할 예정이다.

감사의 글

본 연구 논문은 과학기술정보통신부 및 정보통신기획 평가원의 출연금으로 수행되고 있는 "기계를 위한 영상 부호화 기술(No,2020-0-00011)" 과제의 연구결과입니다.

참고문헌

[1] I. Goodfellow, Jonathon Shlens and Christian Szegedy, Explaining and Harnessing Adversarial Examples, ICLR 2015.

[2] Alexey Kurakin et.al., Adversarial Attacks and Defences Competition, NIPS '17 Competition: Building Intelligent Systems.

[3] CITYSCAPES Dataset, https://www.cityscapes-dataset.com

[5] Detectron2, https://ai.facebook.com/tools/detectron2/

[6] ISO/IEC JCT1/SC29/WG2, "Evaluation Framework for Video Coding for Machine", N78, Apr. 2021.