DeepFake 탐지를 통한 관련 범죄 예방

13조 21800669 정예은
21900302 박정은
22100290 박인권

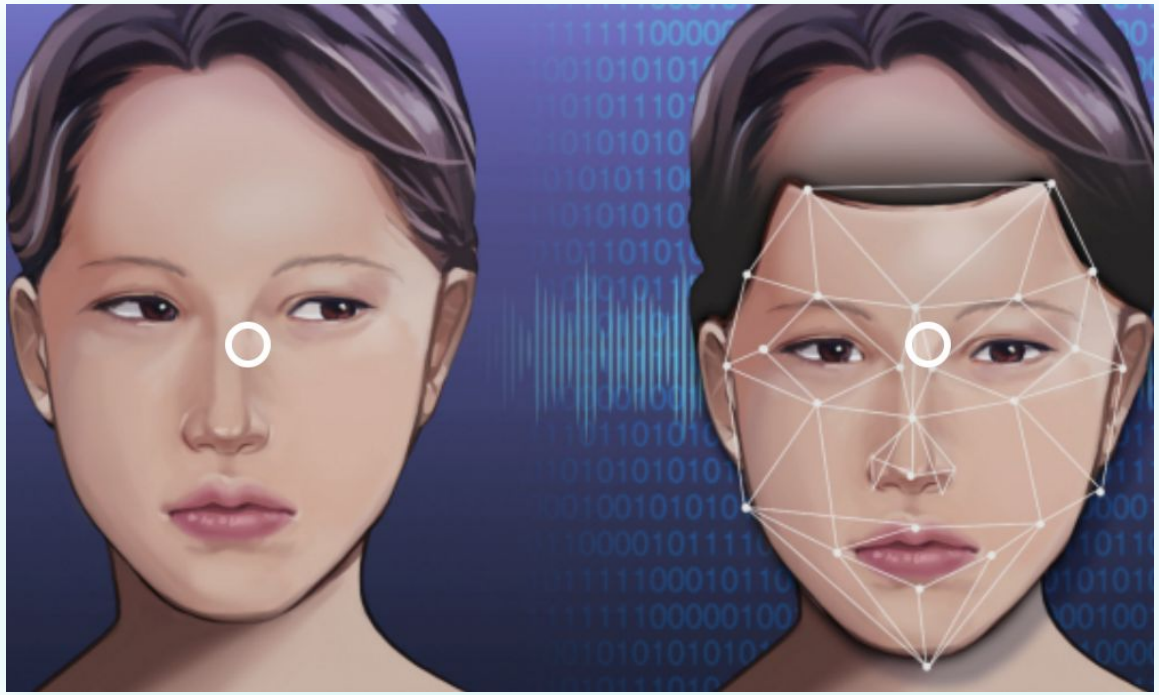
DeepFake 탐지를 통한 관련 범죄 예방

목차

딥페이크를 주제로 선정하게 된 계기

딥페이크

딥페이크'는 AI가 학습할 때 필요한 딥러닝기술과 페이크(Fake)의 합성어로 AI를 기반으로 인간 이미지를 합성하는 기술이다. AI기술이 활발하게 적용되면서 새로운 기술로 떠오른 딥페이크는 진짜와 가짜를 구분하기 어려울 정도로 기술이 발전하면서 범죄에 악용되고 있다. 고화질에 해당하는 동영상을 딥러닝하여 해당 영상에 포함된 인물의 얼굴을 프레임 단위로 합성하는 방식으로 진행된다.



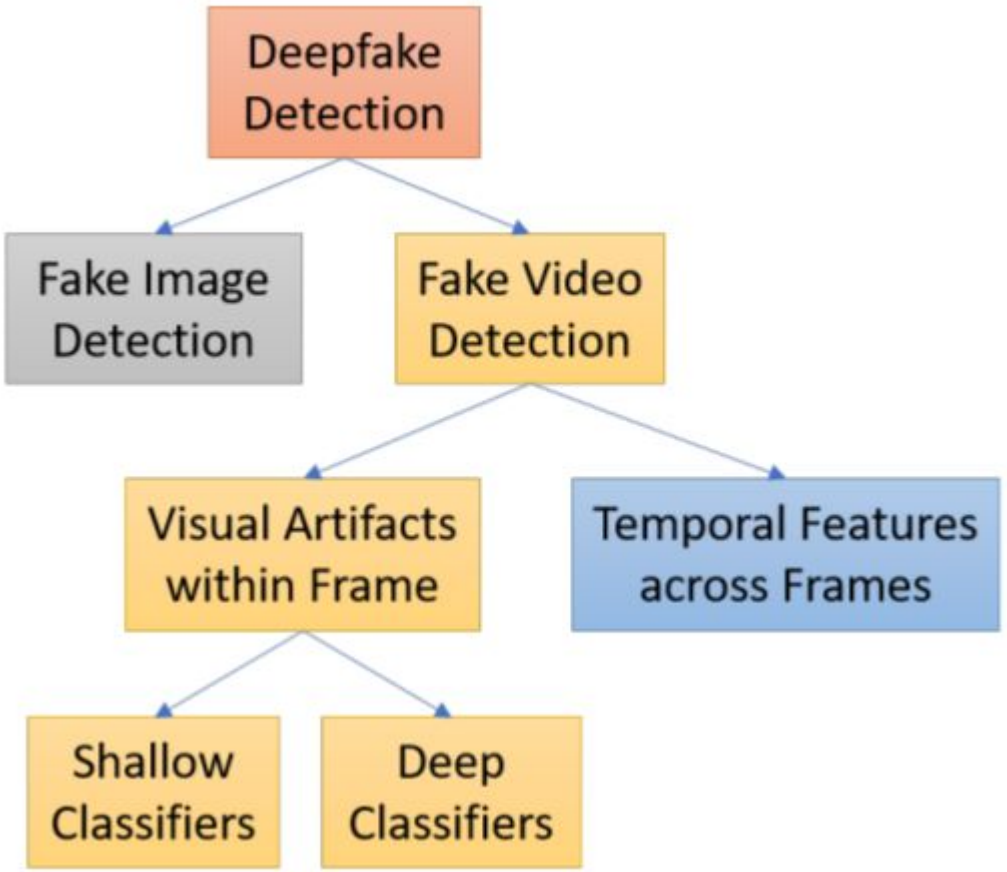
딥페이크를 주제로 선정하게 된 계기

딥페이크 범죄

경향신문에 따른 기사를 살펴보면 현재 우리나라에서 딥페이크를 활용한 불법합성물이 상당히 많이 유포되어 있고 그 양도 증가하였다는 것을 알게 되었다. 그 증가량은 무려 71%나 증가하였다고 한다. 동아일보의 기사를 살펴보면 얼굴 합성을 불법으로 운영하여서 그것으로 인해서 일반인들에게는 접근성이 낮은 딥페이크를 활용하여 성적 합성물로 사용하게 되고 있다. 심지어 딥페이크를 사람들로부터 몰래 합성한 뒤 그것으로 협박을 하고 금전을 요구하는 범죄 유형도 일어나고, 젊은 세대가 아닌 이런 디지털 기술에 취약한 시니어 계층에게도 보이스피싱과 같은 유사한 형태로 사기를 치는 범죄도 발생하고 있다고 한다.

위와 같은 문제들을 확실하게 인식하였고, 이러한 문제들에 대해서 딥페이크의 원리와 이해를 돕는 실습을 통해서 조금은 범죄 예방에 도움을 주려고 할 수 있는 방법을 찾기 위해서 주제로 선정하게 되었다.

사용된 기술



Deep Classifiers - CNN

각 프레임마다 exception net을 이용해서 classification 모델을 구축한 다음 그것으로 학습을 시켜 accuracy 측정한다.

Temporal Features across Frames - CNN+RNN

딥페이크는 매 프레임마다 얼굴의 위치를 찾아 그 위치를 미리학습 시켜놓고 피해자의 아이덴티티 정보를 이용해서 덮어 씌우는 방법을 이용한다. 이 경우 각 프레임들의 시간적인 정보는 고려하지 않고 프레임마다 매번 스와핑을 하는 것이기 때문에 인접한 프레임에 대해서 어떠한 inconsistency 가 발생할 수 있다. 즉 프레임 간의 연결이 자연스럽지 않고 어느정도 일관적이지 않은 특징들을 탐지해낸다.

사용된 기술

< System Architecture >

A. 데이터

YouTube, FaceForensics, Deep fake detection challenge dataset과 같은 서로 다른 데이터셋 소스의 동일한 양의 비디오로 구성된 혼합 데이터셋을 사용한다. 데이터 셋은 원본영상과 조작된 딥페이크 영상 반씩 포함되어 있다. 그리고 이 데이터셋을 70%의 트레이닝 데이터와 30%의 테스트 데이터로 나눈다.

사용된 기술

B. 전처리

데이터셋의 전처리는 비디오를 프레임으로 분할하는 것을 포함한다. 감지된 얼굴을 기준으로 프레임을 자른다. 프레임 수의 균일성을 유지하기 위해 데이터셋 영상의 평균이 계산되고 평균과 동일한 프레임을 포함하는 새롭게 전처리된 얼굴 데이터셋이 생성된다. 얼굴이 없는 프레임은 전처리 할 때 무시된다.

초당 30프레임의 10초짜리 영상을 처리할 때, 즉 총 300프레임은 많은 계산력을 필요로 한다. 그래서 우리는 100프레임만 사용했다.

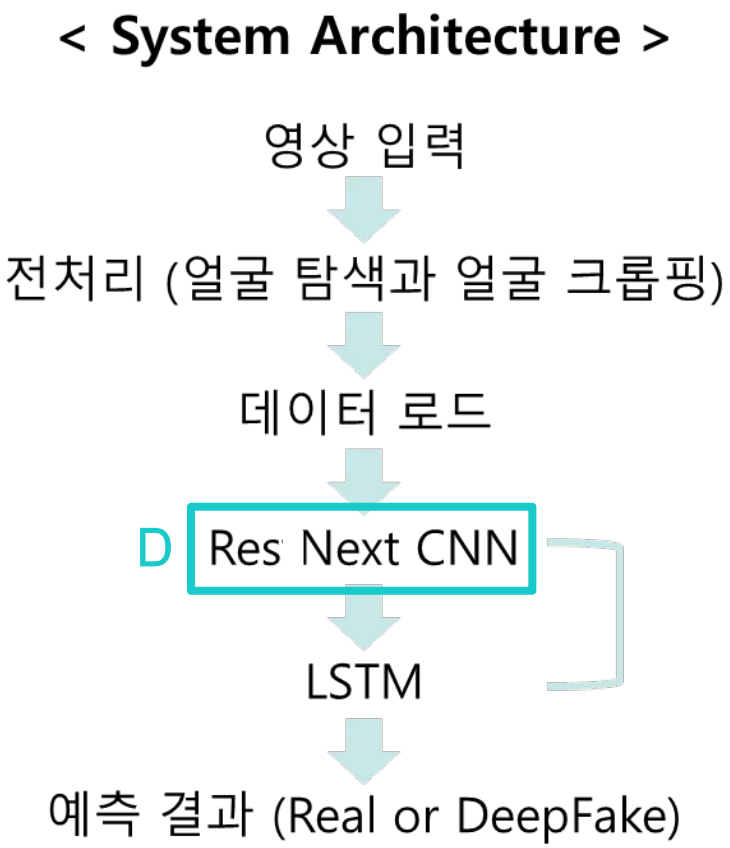
사용된 기술

< System Architecture >

C. 모델

모델은 CNN과 RNN의 일종인 LSTM으로 구성된다. 전처리된 영상에서 추출된 프레임들은 미니 배치로 train 및 test를 위해 모델로 전달된다.

사용된 기술



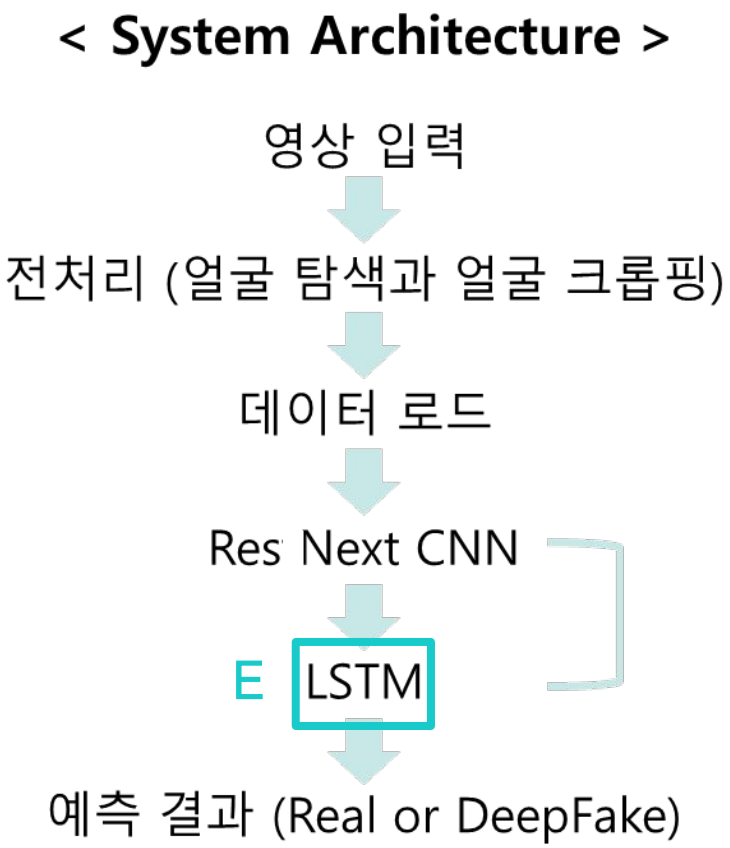
D. 특징 추출을 위한 ResNext CNN

ResNext CNN를 사용하여 특징을 추출하고 프레임의 특징을 정확하게 탐지한다.

이어서, 모델의 경사 하강법을 적절히 수렴하기 위해 필요한 레이어를 추가하고 적절한 학습 속도를 설정하여 네트워크를 미세조정한다.

마지막 풀링 레이어 이후의 2048차원 형상 벡터는 연속 LSTM 입력으로 사용된다.

사용된 기술

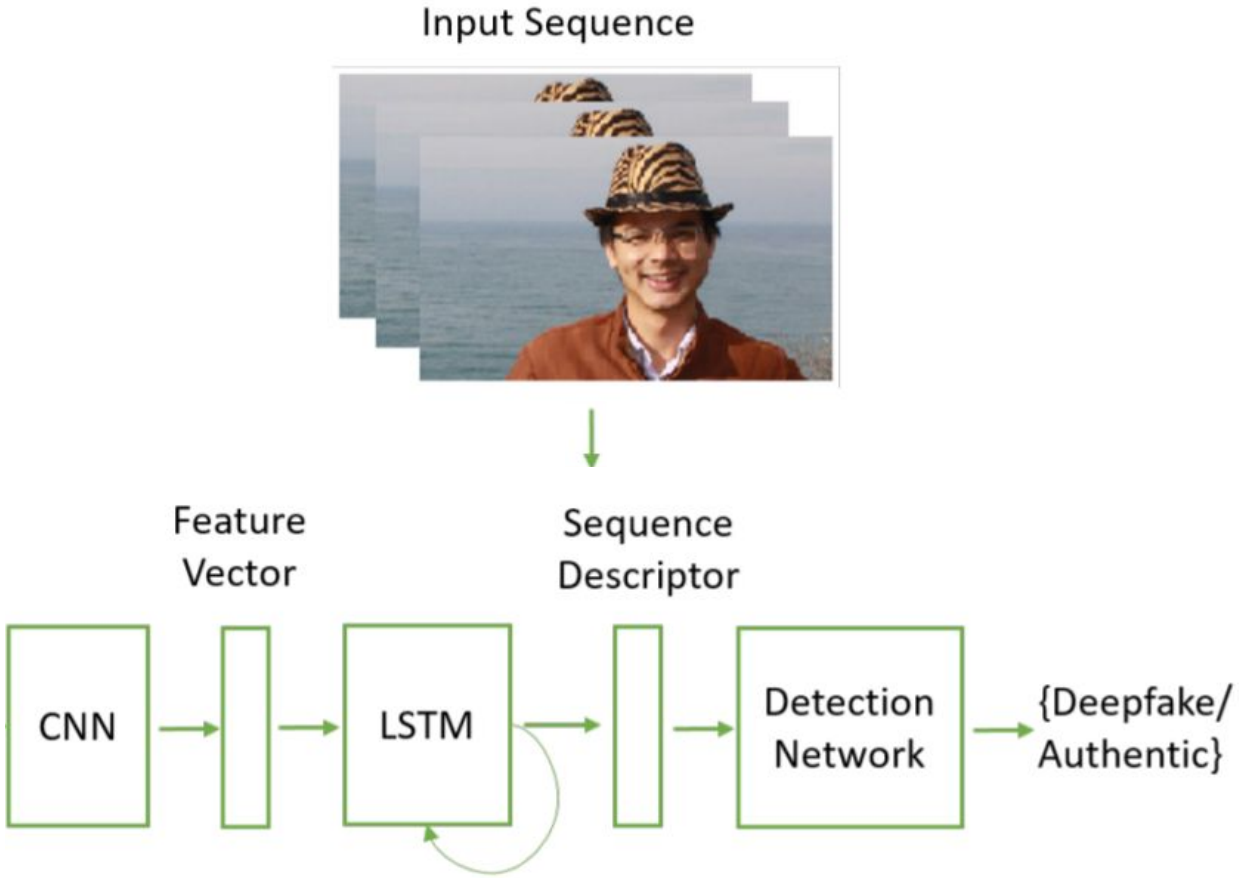


E. 시퀀스 처리를 위한 LSTM

CNN을 통해 추출한 프레임의 특징은 LSTM에 공급되어 시간적 Sequence Descriptor를 만든다. 완전히 연결된 네트워크는 마지막으로 그림에 설명된 것처럼 Sequence Descriptor를 기반으로 실제 영상에서 변조된 영상을 분류하는 데 사용된다.

LSTM은 't'초에 있는 프레임을 't-n'초에 있는 프레임과 비교해 영상의 시간적 분석이 가능하도록 프레임을 순차적으로 처리하는 방식이다. 여기서 n은 t 이전의 임의의 프레임 수일 수 있다.

사용된 기술



E. 시퀀스 처리를 위한 LSTM

CNN을 통해 추출한 프레임의 특징은 LSTM에 공급되어 시간적 Sequence Descriptor를 만든다. 완전히 연결된 네트워크는 마지막으로 그림에 설명된 것처럼 Sequence Descriptor를 기반으로 실제 영상에서 변조된 영상을 분류하는 데 사용된다.

LSTM은 't'초에 있는 프레임을 't-n'초에 있는 프레임과 비교해 영상의 시간적 분석이 가능하도록 프레임을 순차적으로 처리하는 방식이다. 여기서 n은 t 이전의 임의의 프레임 수일 수 있다.

사용된 기술

Expected Results

F. 예측

예측을 위한 훈련모델로 새로운 영상을 전달한다. 훈련된 모델의 형식에 적용하기 위해 새 영상도 전처리된다. 영상은 얼굴 크로핑 후 프레임으로 분할되고 잘린 프레임은 조작유무 감지를 위해 훈련된 모델로 바로 전달된다.

결론

모델의 정확도와 함께 영상이 딥페이크인지 진짜 비디오인지 여부를 출력한다.

기대효과

- **딥페이크를 활용한 범죄에 대해서 예방하기 위한 방향으로 사용 할 수 있을 것이다.**

딥페이크를 활용하는 원리와 작동 되는 시스템에 이해가 되어주는 실습으로 활용되어진 범죄의 분석과 예방하기 위한 해결책을 찾아볼 수 있을 것으로 기대하고 있다. 허위 정보로서 활용되는 이 딥페이크를 탐지하는 기술을 개발하거나 그것들도 활용하는 것에 있어서 더욱 쉽고 정확하게 사용하는데 도움을 줄 수 있을 것 또한 생각해 볼 수 있는 부분이다.

- **사람들에게 쉽게 접근할 수 없는 기술의 접근성을 상승 시킬 것이다.**

딥페이크라는 기술의 이해도와 설명을 통해, 그리고 실습을 살펴보는 것을 통해 일반인들도 조금은 더욱 쉽게 딥페이크라는 기술에 대해서 효과적으로 이해할 수 있을 것이라는 기대를 가질 수 있다. 사람들에게 접근성을 높일 수 있는 이 결과는 조금 더 폭 넓게, 그리고 올바르게 이 기술들을 사용할 수 있을 것을 기대할 수 있다.

감사합니다.