COSE474-2019F: Final Project Proposal
Improving Detecting DeepFake images with 어쩌고저쩌고

제한재 허환 홍주연 도재형 채병주

Abstract

이번 프로젝트에서는 deepfake image들의 feature를 파악해 classification을 진행하는 deepfake detection model을 구현하고자 한다. Dataset으로는 Google에서 공개한 Deepfake Detection Dataset을 사용할 예정이고, [여기에 목표 입력]하는 것이 우리의 목표이다.

1. Intruduction

Deepfake는 GAN과 같은 deep neural network를 사용하여 사람의 얼굴 혹은 특정 신체 부위를 합성하는 기술이다. Dataset이 충분하다면 일상적인 사진들로도 쉽게 deepfake image를 만들어낼 수 있고, 수작업으로 합성하는 것보다 훨씬 자연스러운 결과물을 생성할 수 있다. 또한, 제작 속도가 빠르기 때문에 frame 단위로 합성하여 deepfake video 역시 쉽게 만들 수 있다.

Deepfake image들은 인터넷에서 humor 혹은 meme으로 소비되기도 하지만, 악의적으로 사용되어 많은 문제가 되기도 한다. 특히, social media의 발전으로 인한 타인을 사칭하는 계정과 deepfake pornography의 양산으로 사회적으로 큰 문제가 되고 있다.

이러한 deepfake가 낳은 사회적인 문제들로 인해, 최근 deepfake detection에 대한 연구의 필요성이 대두되고 있다. Deepfake detection은 현재 computer vision 분야의 주요 issue 중 하나있로, [누구 누구] 등에 의해 연구된 바 있다. Google에서는 deepfake detection 연구를 위한 dataset을 공개하였고 (Dufour & Gully, 2019), Facebook에서도 deepfake detection challenge를 dataset 공개과 함께 진행할 예정임을 알렸다 (Schroepfer, 2019).

따라서 이번 프로젝트에서는 주어진 image가 deepfake image인지 아닌지를 판단하는 deepfake detection model을 개발하고, 그 성능을 평가해 보고자 한다.

2. Datasets

Google에서 공개한 Deepfake Detection Dataset (Dufour et al., 2019) 을 사용할 예정이다. 해당 dataset은 FaceForensics++ (Rössler et al., 2019)의 GitHub 페이지에서 내려받을 수 있다.

3. Goals

우리의 목표는 새로이 공개된 데이터셋에 대해 강하게 학습할 수 있는 모델을 생성하고 두 가지 모델 지표 중 하나를 달성하는 것이다. 첫 번째는 Sensitivity가 높은 모델을 개발하여 현실 세계에서 이미지의 위험을 성공적으로 경고하는 모델을 생성하는 것이고, 더 나아가 높은 F-Score를 달성한 쌍방향으로 신뢰가능한 모델을 생성하는 것이다.

4. Brief Schedule

- 11/01 - 공개된 데이터셋에 따른 기본적인 학습 모델
- 11/22 - 개선된 Data preprocessing 을 바탕으로 모델 성능 개선
- 12/01 - Fine-tuning 된 모델에 데이터 적용
- 12/04 - Paper work & Presentation

5. Roles

우리 팀은 총 5명으로 이루어져 있다. 모든 팀원이 모델 기획 및 구현에 참여한다. 추가로, 기타 여러 가지 업무를 다음과 같이 분담하여 진행하기로 하였다.

- 제한재 – Data preprocessing
- 도재형, 홍주연 – SOTA model & dataset research
- 채병주, 허환 – Model fine-tuning

6. Comparison with SOTA

Yuezen et al.은 DeepFake 로 생성된 비디오에서는 눈을 감고 있는 이미지에 대해 모델의 학습이 부족하여 정상적으로 눈 깜빡임을 판별하는 CNN/RNN 모델을 제안한 바 있다 (2018). David et al.의 연구에서는 RNN과 CNN에서 긍정적인 결과를 얻었지만 전체적인 접근 방식에는 약점이 있는 것을 보였다 (2018). 더 나아가 생성된 데이터를 바탕으로 학습하는것이 효율적이지 못하고 실제 이미지와 가짜 이미지를 동시에 학습이 필요하다고 제시하고 있다. 개선된 데이터셋을 바탕으로 더 기존의 학습 모델에서 더 나은 퍼포먼스를 보이는 모델을 생성하고자 한다.

References

Dufour, N. and Gully, A. Contributing data to deepfake detection research, 2019. URL https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

Dufour, N., Gully, A., Karlsson, P., Vorbyov, A. V., Leung, T., Childs, J., and Bregler, C. Deepfakes Detection Dataset by Google & Jigsaw, 2019.

Güera, D. and Delp, E. J. Deepfake video detection using recurrent neural networks. AVSS, 2018.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

Schroepfer, M. Creating a data set and a challenge for deepfakes, 2019. URL https://ai.facebook.com/blog/deepfake-detection-challenge.

Yuezen Li, Ming-Ching Chang, S. L. Exposing ai generated fake face videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018.