Learning to Detect Fake Face Images in the Wild

Abstract

- Binary classifier을 학습시켜 딥페이크 이미지를 구분하는 것은 어려운 task이다.
- 따라서 본 논문에서는 contrasive loss를 사용해 딥페이크 이미지를 검출하기 위해 classifier과 연결하였다.

Introduction

Image generation model이 발전함에 따라 보안 문제들이 대두되었다. 따라서 이러한 이미지들이 위조 이미지인지 진짜인지 구분하는 것이 중요해졌다.

본 논문에서는 이 문제들을 해결하기 위한 효과적인 방안으로 deep neural network 인 deep forgery discriminator(DeepFD)을 제안하였다.

새로운 GAN 모델의 가짜 이미지들은 기존의 데이터를 통해 학습 시킨 discriminator 모델에서 구분하기 어려운 경우가 많다. 이러한 문제를 해결하기 위해 다른 GAN에서 생성한 이미지를 contrastive loss를 사용해 공통적으로 구분되는 특징을 학습한다.

본 논문의 contribution은 다음과 같다.

- 본 논문의 저자들은 contrasive loss에 기반한 discriminator을 제안하였다. 이를 통해 어떠한 GAN 모델에서 생성된 이미지라도 구분해 낼 수 있도록 하였다.
- 본 논문에서 제안된 DeepFD는 가짜 이미지의 비현실적 detail을 localize 할 때 사용 가능하다.

The Proposed Deep Forgery Discriminator

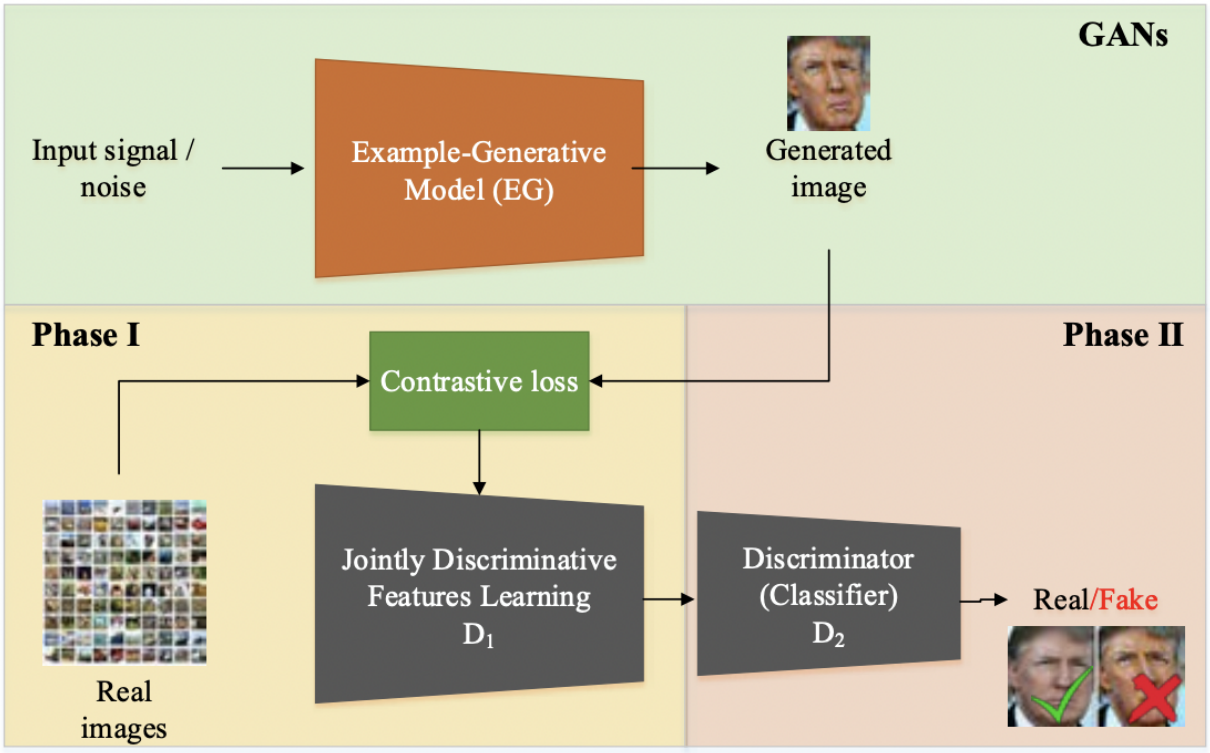

Fig. 1 The flowchart of the proposed deep forgery discriminator (DeepFD). Example-generative model (one or more than one models) will continuously synthesize training samples for the proposed DeepFD training.

상기된 이미지는 DeepFD의 flowchart를 나타낸다. 해당 flowchart에서는 2가지 단계의 학습이 진행된다. 먼저 contrastive loss에 기초하여 discriminative feature인 D_1 을 학습하기 위해 여러 GAN에서 만들어진 딥페이크 이미지와 진짜 이미지들을 모은다. 그 후, discriminator인 D_2 는 가짜 이미지를 구분하기 위해 D_1 과 concatenate 될 것이다. 테스트 단계에서 D_1 와 D_2 를 통해 이미지가 진짜인지 가짜인지 구분하는 것은 간단하다.

D_1 과 D_2 의 네트워크 아키텍쳐의 디테일은 다음과 같다.

TABLE I
NETWORK STRUCTURES IN JOINTLY DISCRIMINATIVE FEATURE LEARNING (D_1) AND CLASSIFIER TRAINING (D_2)

Layers	***D1***	D2
1	Conv.layer, kernel=7*7, stride=4, channel=96	Conv. layer, kernel=3*3, channel = 2
2	Residual block *2, channel=96	Global average pooling
3	Residual block *2, channel=128	Fully connected layer, neurons=2 Softmax layer
4	Residual block *2, channel=256	
5	Fully connected layer, neurons=128 Softmax layer	

Classifier D_2 는 D_1 의 4번째 layer에 concatenate 된다.

M 개의 GAN에서 얻은 학습 데이터셋을 $X_{fake} = [x_{i=1}^{k=1}, x_{i=2}^{k=1}, ..., x_{i=N_1}^{k=1}, ..., x_{i=N_M}^{k=M}]$ 이라고 가정하자(N_k 의 생성된 이미지).

진짜 이미지와 가짜 이미지가 포함된 학습 이미지 개수의 합은 $N_T = N_r + N_f = N_r + \sum_{k=1}^{M} N_k$ 이다. 레이블 $y = [y_1, y_2, ..., y_{N_T}]$ 은 진짜($y = 1$)와 가짜($y = 0$)로 이루어진다. Pairwise information 을 learning architecture로 통합하기 위해, $C(N_T, 2)$ **pair(covariance)** 이 pairwise information인 $P = [p_{i=0,j=0}, p_{i=0,j=1}, ..., p_{i=0,j=N_r}, ..., p_{i=N_f,j=N_r}]$ 에 있다.

Jointly Discriminative Feature Learning

Feature representation을 $R_i = D(x_i)$ 라고 가정한다. Paired input images를 통해 jointly discriminative feature learning의 목적은 similarity function을 최소화 하는 것이다. Similarity function은 다음과 같다.

$$\mathrm{E}_W(\mathbf{x}_1, \mathbf{x}_2) = \|\mathrm{D}_1(\mathbf{x}_1) - \mathrm{D}_1(\mathbf{x}_2)\|,$$

- D_1 을 사용해 paired input image로부터 feature을 뽑아낸다.
- $E_W(x_1, x_2)$ 값을 바로 최소화 하는 것은 feature representation $D_1(x_i)$ 가 constant mapping 되게 할 수 있다. 이는 feature representation이 쓸모 없게 만들 수 있다. 따라서 본 논문에서는 다음과 같은 contrastive loss를 제안한다.

$$L(W, (P, \mathbf{x}_1, \mathbf{x}_2))$$

$$= \frac{1}{2}\left(p_{ij}(\mathrm{E}_W)^2 + (1 - p_{ij})(\max(0, m - \mathrm{E}_W)^2\right),$$

- E_W, m : Predefined marginal value
- p_{ij} : pair of images

Classifier Training

Jointly discriminative feature representation이 학습되면, 여러 알고리즘을 통해 이미지들을 분류할 수 있다. 본 논문에서는 convoutional layer과 fully connected layer을 네트워크 D_1 에 바로 concatenate 한다.

Classifier의 loss function은 cross-entropy loss로 정의된다. ⬇️

$$L_C(\mathbf{x}_i, \mathrm{y}_i) = -\sum_{i}^{N_T}\left(\mathrm{D}_2\left(\mathrm{D}_1(\mathbf{x}_i)\right)\log\ \mathrm{y}_i\right).$$

Classifier은 back-propagation으로 쉽게 학습이 가능하다.

Conclusion

- Contrastive loss는 다른 GAN에서 생성된 가짜 이미지의 joint discriminative features를 잡아내는 것에 사용되기 좋다.

- 본 논문에서 제안된 딥페이크 이미지 구분 방안이 좋은 성능을 내는 것을 알 수 있었다.

개선 방안

- Discriminator인 D_1, D_2 의 구조를 손 본다면 더욱 좋은 결과를 얻을 수도 있지 않을까 생각한다.
- Color space 의 각 color component의 residual domain을 비교하여 더욱 확실한 진짜 이미지와의 차이를 확인한다.