색상 차이와 신경망을 이용한 위조 이미지 검출

(Fake Image Detection Using Color Disparities and Neural Networks)

이 정 한, 박 한 훈
(Jeonghan Lee, Hanhoon Park)

부경대학교 전자공학과

Abstract : In the field of deep learning, generative adversarial network (GAN) is in the spotlight as a representative image generation model, and it has already reached a point where it is difficult to classify real and fake (i.e., GAN-generated) images with the naked eye. However, based on the generation principle of GAN, a previous method could classify real and fake images with high accuracy and robustness by extracting useful features using hand-crafted filters from their chromatic components. Inspired by the previous method, we also attempt to classify real and fake images in chromatic domains. However, we use convolutional neural networks (CNN) to extract features from images. To be specific, we try to use the transfer learning with pre-trained CNN models. Also, we try to train the PeleeNet, a type of deep CNNs, with or without a pre-processing high pass filter. The CNN-based methods are compared with the previous method. For experiments, we prepared four image datasets consisting of images generated with different image contexts and different GAN models. Extensive experimental results showed that CNN-based methods are not accurate for those whose generation GAN model and contexts are different from the training images, unlike the previous method showed high classification accuracy. However, we found that the previous method could be further improved by using the luminance components together with the chromatic components.

Keywords : Fake image identification, Generative adversarial networks, Color disparities, Transfer learning, CNN

Ⅰ. 서론

인공지능 및 딥러닝 기술의 급격한 발전으로 인해 다양한 분야에서 적용하려는 시도가 늘고 있다. 특히, 영상 생성 신경망의 일종인 GAN(generative adversarial network)은 기법과 분야에 따라 다양하게 세분화되어 과거와는 달리 육안으로 실제(real) 영상과 생성된(GAN-generated) 위조(fake) 영상을 구분하기 힘들게 되었다. 기술의 발전에 따른 다양한 범죄 발생의 우려도 함께 늘고 있다.

최근 Li 등은 입력 영상을 HSV, YCbCr 색공간으로 변환한 후, 다양한 형태의 필터를 설계하여 색차(color disparity) 정보를 추출함으로써 위조 영상을 높은 정확도로 식별할 수 있음을 증명하였다[1]. 본 논문은 새로운 데이터셋에 대해 Li 방법의 성능을 검증하고, 색차 정보를 추출하기 위해 직접 필터를 설계하는 것 대신 CNN을 사용하는 방법을 제안하고, 성능을 분석한다. 구체적으로, 각 방법의 학습 때와 다른 영상 생성 모델과 영상 컨텍스트에 대한 위조 영상 식별의 정확도 및 유연성을 살펴본다.

Ⅱ. GAN 기반 위조 영상 생성

GAN은 주어진 데이터를 학습하여 같은 분포를 가지는 표본을 생산하는 것을 목표로 한다[2]. 기본적으로 GAN 모델은 두 개의 네트워크, 생성기(generator)와 식별기(discriminator)로 구성되어 있다. 생성기가 주어진 데이터를 합성하여 표본을

생성하면, 식별기가 합성 표본과 실제 데이터를 분류하는 형태로 구성되어 있다. 최대한 결과를 개선하고 학습 데이터의 양과 품질을 높임으로써, 최종적으론 실존하는 데이터와 극도로 유사한 샘플을 생성하는 것이 GAN의 이상적인 목표이다.

본 논문은 가장 보편적이고 성능이 우수한 모델인 PGGAN[3]과 StyleGAN2[4]를 사용해 생성된 위조 영상을 식별하는 데 초점을 둔다.

Ⅲ. 위조 영상 분류

일반적인 GAN의 생성기는 무작위 잠재(latent) 벡터를 컨볼루션 계층을 통해 점진적으로 크기를 확장하고, 생성기 마지막 계층에서 채널 크기가 3인 영상을 생성한다. 실제 영상과는 달리 색 공간으로 디지털화하면서 이전에 존재하지 않던 생성 영상의 본질적인 속성 상관관계들이 생성과정을 통해 발생하기 때문에 이러한 정보를 활용함으로써 높은 정확도로 분류가 가능하다.

합성곱 신경망(CNN)은 영상에 합성곱 필터를 통과시켜가며 각 요소를 계산하고 중요한 특징값을 추출하여 값을 도출한다. 따라서 CNN은 2차원 영상의 공간정보를 유지한 채 분류 패턴을 파악하는 점에 있어 유리하며 동일한 방식을 반복하여 엄청난 학습량을 확보할 수 있다.

본 논문에서는 GAN에 의해 생성된 위조 영상을 검출하기 위해 CNN을 활용하는 방법으로, 사전학습된 CNN 모델(ResNeXt101, AlexNet, SqueezeNet, VGG-19)[5]을 사용한 전이학습(transfer learning) 방법과 PeleeNet[6]을 학습하는 방법을 사용하였다.

Ⅳ. 실험

1. 실험환경

1.1 데이터셋

실험은 LSUN 데이터셋[7]에 포함된 cat과 church_outdoor 영상을 사용했으며, 각각 학습 데이터 25,000장/테스트 데이터 2,000장씩 임의추출하였고 크기는 256x256로 조절하였다. PGGAN과 StyleGAN2를 사용하여 cat, church_outdoor 영상을 생성한 영상을 실제 영상과 동일하게 256x256로 크기를 조절하고 학습 데이터 25,000장/테스트 데이터 2,000장씩 임의추출하였다.

1.2 영상 특징 추출

Li 방법에서 제안된 필터[1]를 사용하여 입력 영상으로부터 컬러 채널당 75개의 특징값을 추출하였다. 기존 연구에서 H, S, Cb, Cr 4개의 색도 채널을 사용하는 것이 가장 효과적이라고 했으나, RGB, HSV, YCbCr 등의 3개 채널을 사용하거나 H, S, V, Y, Cb, Cr 6개 채널을 사용했을 때의 성능을 추가적으로 살펴보았다.

또한, 전이학습을 활용하는 방법으로, Pytorch에서 제공하는 사전학습 모델인 ResNeXt101_32x8d, AlexNet, SqueezeNet 1.1, VGG-19을 각각 사용하여 특징 벡터를 추출하였다.

마지막으로, PeleeNet를 사용하여 특징 벡터를 추출하였는데, 사전 학습되지 않은 CNN 모델이므로 위조 영상 검출을 위해 분류기와 함께 효과적인 특징 벡터를 추출하도록 학습되었다. 위조 영상 검출에 유리한 높은 주파수 성분을 추출하도록 전처리 과정으로 고주파 통과 필터(HPF)를 사용하는 방법에 대한 성능을 살펴보았다.

1.3 분류기

Li 방법[1]와 달리 본 논문에서는 분류기로 신경망을 사용했다. 각각의 방법이 추출한 특징 벡터를 입력으로, 출력층의 2개의 노드는 실제 영상과 위조 영상에 대한 확률값을 출력하도록 학습되었다.

모든 학습에서 epoch 값은 50, 배치 크기는 64, 학습률은 0.001, 최적화 알고리즘은 Adam을 사용하였다.

2. 실험 결과

표 1에서 보는 것처럼, 학습과 테스트 데이터의 영상 생성 모델 및 영상 컨택스트가 같으면 모든 방법이 높은 정확도로 위조 영상을 검출할 수 있었다. 일부 AlexNet이나 VGG를 사용한 전이학습과 RGB 색공간에서 Li 방법을 사용하여 영상 특징을 추출하는 경우 상대적으로 정확도가 낮은 결과를 보였다.

표 2, 3, 4에서 보는 것처럼, 학습과 테스트 데이터의 영상 생성 모델 또는 영상 컨택스트가 다르면, 일부 영상 생성 모델 차이에 대해서는 전처리 필터로 HPF를 사용하는 PeleeNet이 위조 영상 검출에 성공했으나, 대부분의 경우 CNN을 사용하여 영상 특징을 추출하는 방법은 위조 영상 검출에 실패하였다. 반면, Li 방법을 사용하여 영상 특징을 추출한 경우 사용된 색공간에 따라 정확도 차이는 있었으나 대부분 높은 정확도로 위조 영상 검출이 가능하였다. 한 가지 중요한 결과로, 기존 연구에서는 H, S, Cb, Cr 4개의 색도 채널을 사용하는 것이

가장 효과적이라고 했지만, V나 Y 채널을 추가적으로 사용할 경우 정확도는 향상되었으며 결과적으로 H, S, Cb, Cr에 Y, V 채널을 합쳐서 6개의 채널을 함께 사용하는 것이 가장 높은 정확도로 위조 영상을 검출하였다.

표 1. 학습/테스트 생성 모델과 영상 컨택스트가 동일한 경우

Accuracy : Average(Real/Fake)[%]

Train / Test	Style(church) / Style(church)	Style(cat) / Style(cat)	PG(church) / PG(church)	PG(cat) / PG(cat)
Transfer(ResNeXt)	99.50(99.50/99.50)	99.42(98.95/99.90)	95.88(94.45/97.30)	99.08(99.25/98.90)
Transfer(AlexNet)	97.48(97.30/97.65)	96.40(95.80/97.00)	73.75(65.25/82.25)	73.95(75.25/72.65)
Transfer(SqueezeNet)	99.75(99.90/99.60)	98.20(97.90/98.50)	99.65(99.75/99.55)	97.98(97.30/98.65)
Transfer(VGG-19)	98.02(98.35/97.70)	96.70(95.30/98.10)	84.70(85.15/84.25)	91.72(95.05/88.40)
PeleeNet+HPF	100(100/100)	100(100/100)	99.72(99.45/100)	99.93(100/99.95)
PeleeNet	99.90(99.80/100)	99.58(99.80/99.35)	99.82(100/99.65)	99.92(99.90/99.95)
H,S,Cb,Cr	100(100/100)	99.93(100/99.95)	99.98(99.95/100)	100(100/100)
R,G,B	99.95(100/99.90)	99.88(99.85/99.90)	98.75(98.40/99.10)	99.70(99.60/99.80)
H,S,V	100(100/100)	99.93(100/99.95)	99.95(100/99.90)	99.80(99.80/99.80)
Y,Cb,Cr	100(100/100)	100(100/100)	100(100/100)	100(100/100)
H,S,V,Y,Cb,Cr	100(100/100)	100(100/100)	100(100/100)	100(100/100)

표 2. 학습/테스트 생성 모델이 다른 경우

Accuracy : Average(Real/Fake)[%]

Train / Test	Style(church) / PG(church)	Style(cat) / PG(cat)	PG(church) / Style(church)	PG(cat) / Style(cat)
Transfer(ResNeXt)	52.72(99.50/5.95)	54.65(98.95/10.35)	62.80(94.45/31.15)	55.25(99.25/11.25)
Transfer(AlexNet)	52.45(97.30/7.60)	52.32(95.80/8.85)	68.15(65.25/71.05)	54.30(75.25/33.35)
Transfer(SqueezeNet)	53.68(99.90/7.45)	61.58(97.90/25.25)	50.62(99.75/1.50)	63.22(97.30/29.15)
Transfer(VGG-19)	52.80(98.35/7.25)	53.32(95.30/11.35)	69.95(85.15/54.75)	52.65(95.05/10.25)
PeleeNet+HPF	53.88(100/17.75)	84.05(100/68.10)	98.38(99.45/97.30)	50.00(100/0)
PeleeNet	57.38(99.80/14.95)	55.85(99.80/11.90)	53.37(100/6.75)	49.98(99.90/0.05)
H,S,Cb,Cr	96.78(100/93.55)	86.68(100/73.35)	99.98(99.95/100)	100(100/100)
R,G,B	85.42(99.55/71.30)	98.30(99.35/96.75)	99.20(98.40/100)	99.80(99.60/100)
H,S,V	96.15(100/92.30)	81.70(100/63.40)	99.82(100/99.65)	99.85(99.80/99.90)
Y,Cb,Cr	98.38(100/96.75)	90.50(100/81.00)	100(100/100)	100(100/100)
H,S,V,Y,Cb,Cr	97.80(100/95.60)	91.35(100/82.70)	100(100/100)	100(100/100)

표 3. 학습/테스트 영상 컨택스트가 다른 경우

Accuracy : Average(Real/Fake)[%]

Train / Test	Style(church) / Style(cat)	Style(cat) / Style(church)	PG(church) / PG(cat)	PG(cat) / PG(church)
Transfer(ResNeXt)	54.42(99.40/9.45)	51.12(99.95/2.30)	83.75(94.35/73.15)	59.68(99.65/19.75)
Transfer(AlexNet)	50.02(100/0.05)	50.20(99.25/1.15)	52.88(74.85/30.90)	55.82(67.05/44.60)
Transfer(SqueezeNet)	50.25(100/0.50)	64.48(99.35/29.60)	59.95(26.65/93.25)	77.98(98.75/57.20)
Transfer(VGG-19)	50.05(100/0.10)	50.10(99.90/0.30)	61.15(93.00/29.30)	57.75(90.90/24.60)
PeleeNet+HPF	49.68(99.35/0)	50.00(100/0)	61.65(41.75/81.55)	51.00(99.95/2.05)
PeleeNet	49.98(99.80/0.15)	50.00(100/0)	49.92(99.80/0.05)	56.05(97.90/14.20)
H,S,Cb,Cr	99.75(99.90/99.60)	99.90(100/99.80)	98.62(98.80/98.45)	98.68(97.35/100)
R,G,B	61.55(100/23.10)	94.18(93.05/95.30)	95.38(99.15/91.60)	87.12(87.30/86.95)
H,S,V	95.12(90.60/99.65)	88.75(100/77.50)	87.88(79.25/96.50)	97.48(99.75/95.20)
Y,Cb,Cr	99.80(100/99.60)	99.98(100/99.95)	98.95(99.75/98.15)	99.60(99.20/100)
H,S,V,Y,Cb,Cr	99.85(99.90/99.80)	99.88(100/99.75)	98.55(99.20/97.90)	99.88(99.75/100)

Accuracy : Average(Real/Fake)[%]

Train / Test	Style(church) / PG(cat)	Style(cat) / PG(church)	PG(church) / Style(cat)	PG(cat) / Style(church)
Transfer(ResNeXt)	50.12(99.40/0.85)	50.32(99.95/0.70)	48.65(94.35/2.95)	50.05(99.65/0.45)
Transfer(AlexNet)	50.00(100/0)	50.38(99.25/1.50)	48.45(74.85/22.05)	46.08(67.05/25.10)
Transfer(SqueezeNet)	50.02(100/0.05)	57.12(99.35/14.90)	59.12(26.65/91.60)	50.72(98.75/2.70)
Transfer(VGG-19)	0.5(100/0)	50.18(99.90/0.45)	47.75(93.00/2.50)	48.65(90.90/6.40)
PeleeNet+HPF	49.68(99.35/0)	50.00(100/0)	46.88(41.75/52.00)	49.98(99.95/0)
PeleeNet	49.98(99.80/0.15)	50.00(100/0)	49.90(99.80/0)	49.05(97.90/0.20)
H,S,Cb,Cr	87.30(99.90/74.70)	86.85(100/73.70)	99.40(98.80/100)	98.68(97.35/100)
R,G,B	79.20(99.55/58.85)	83.18(93.05/73.30)	92.02(99.15/84.90)	93.00(87.30/98.70)
H,S,V	83.68(90.60/76.75)	66.02(100/32.05)	89.55(79.25/99.85)	98.32(99.75/96.90)
Y,Cb,Cr	93.35(100/86.70)	88.55(100/77.10)	99.75(99.75/99.75)	99.60(99.20/100)
H,S,V,Y,Cb,Cr	89.02(99.90/78.15)	86.50(100/73.00)	99.52(99.20/99.85)	99.88(99.75/100)

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1045749).

참고문헌

[1] H. Li, B. Li, S. Tan, and J. Huang, “Identification of deep network generated images using dispairites in color components,” Signal Processing, vol. 174, 2020.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Proc. of NeurIPS, vol. 2, pp. 2672-2680, 2014.

[3] T. Karras, T. Aila, S, Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” Proc. of ICLR, 2018.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” Proc. of CVPR, pp. 8110-8119, 2020.

[5] http://pytorch.org/vision/stable/models.html

[6] R. Wang, X. Li, and C. Ling, “Pelee: a real-time object detection system on mobile devices,” Proc. of NeurIPS, pp. 1967-1976, 2018.

[7] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN: construction of a large-scale image dataset using deep learning with humans in the loop,” CoRR abs/1506.03365, 2015.