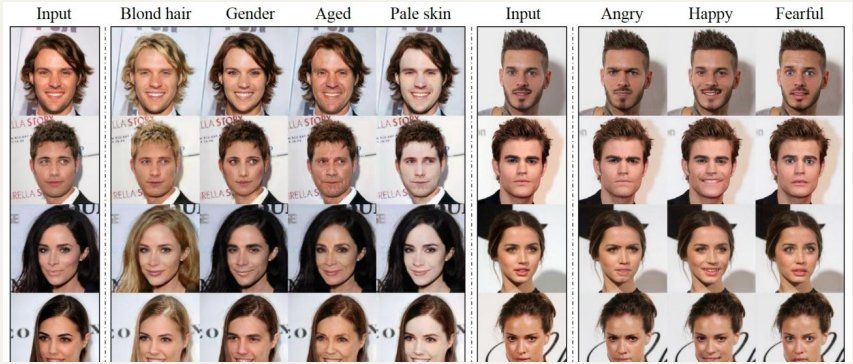
Identification of Deep Network Generated Images Using Disparities in Color Components

발표자 : 문경환

Abstract

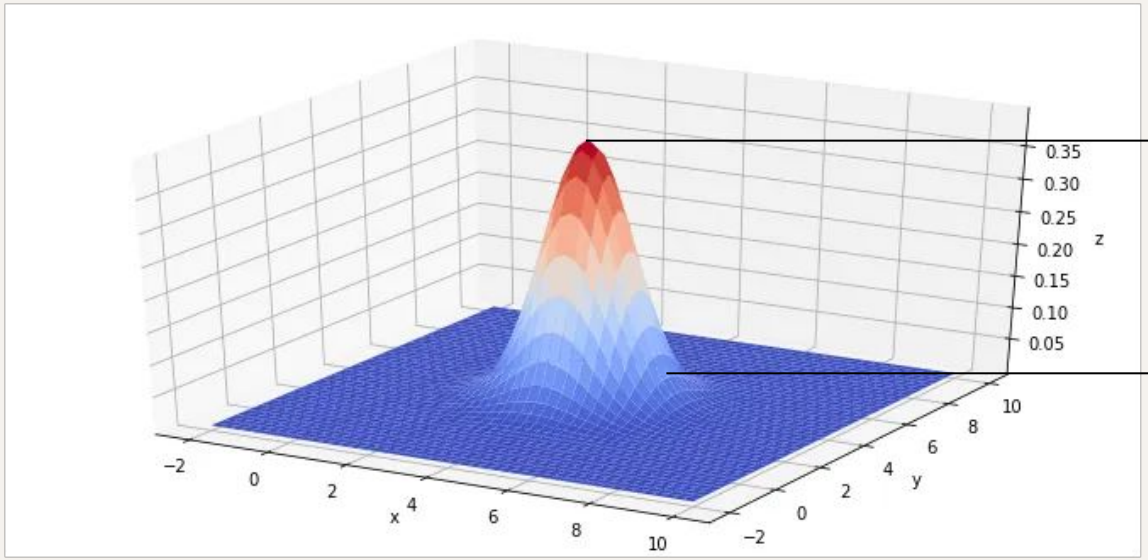
- DNG 이미지와 진짜 이미지 간의 차이를 각 color components에 따라 구함.
- training-testing data가 이미지 source나 생성 모델, 진짜 이미지만으로의 detection에서 match 되었나 mismatch 되었나와 같은 몇몇 detection situation을 평가함.
- GAN 모델이 무엇인지 알 수 없을 때에 좋은 성능을 뽑아냄.

출처 : Konrad Weber, medium



Generative model

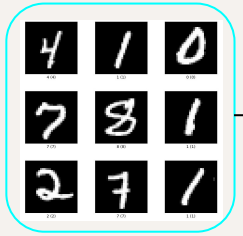
다변수 확률분포

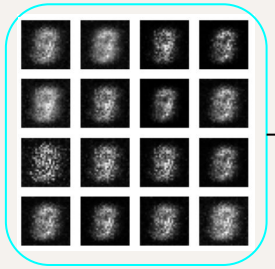


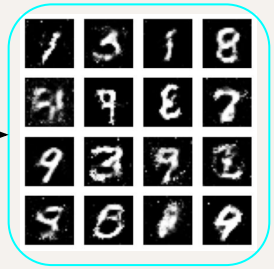
출처 : Aly Shmahell, medium

출처 : analyticsvidhya

GAN







x^*_{start}

x^*_{end}

x

x^*

D 판별자

분류 오차

G 생성자

반복 훈련

z

반복 훈련

Loss Function

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Analysis from the Perspective of Color

The generation pipeline DNG images

- DNG 이미지를 진짜 이미지와 구분하기 위해서는 GAN이 이미지를 생성하고 남은 결함을 찾아야 함.
- GAN 모델 중 생성자의 마지막 layer에서, 몇몇 feature map이 3-channel(R, G, B)의 tensor으로 변화함.
- 반면 카메라로 찍은 진짜 이미지의 pixel은 본질적으로 다른 방식으로 correlated 됨.
- 즉 진짜 이미지와 DNG 이미지는 본질적으로 다를 수 있다는 것임.

Analysis from the Perspective of Color

Discernibility of color component

- GAN은 이미지를 RGB 공간에서 생성하기에 다른 color space의 properties에는 주의를 덜 기울이게 됨.
- 따라서 본 논문에서는 이미지를 RGB, HSV, YCbCr 등의 공간에서 분석함.
- 분석을 간소화하기 위해 horizontal한 인접 픽셀에 관해서만 분석을 진행함.

Analysis from the Perspective of Color

$$r_i^c = \frac{\sum_{j=1}^{m}\sum_{k=1}^{n-1}\left(\mathbf{I}_{j,k}^c - \bar{\mathbf{I}}^c\right)\left(\mathbf{I}_{j,k+1}^c - \bar{\mathbf{I}}^c\right)}{\sqrt{\sum_{j=1}^{m}\sum_{k=1}^{n-1}\left(\mathbf{I}_{j,k}^c - \bar{\mathbf{I}}^c\right)^2 \sum_{j=1}^{m}\sum_{k=1}^{n-1}\left(\mathbf{I}_{j,k+1}^c - \bar{\mathbf{I}}^c\right)^2}},$$

Color component에 따른 인접한 pixel 값들의 연관성

- 본 논문에서는 color component I^c 에 인접한 pixel 사이의 상관계수를 계산함.
- 본 논문에서는 분석을 간소화하기 위해 horizontal 인접 pixel에 관해 분석을 진행.
- I : i 번째 이미지
- hat(I^c) : I^c 의 평균값
- m, n : 이미지의 height, width
- r_i^c : 인접한 pixel 값들의 연관성
  - 이 값이 크면 클수록 I^c의 인접한 pixel 값들의 correlation이 높음.

Analysis from the Perspective of Color

$$d_{\chi^2}(\mathbb{H}^c_{\mathrm{DNG}}, \mathbb{H}^c_{\mathrm{Real}}) = \frac{1}{2}\sum_x \frac{(\mathbb{H}^c_{\mathrm{DNG}}(x) - \mathbb{H}^c_{\mathrm{Real}}(x))^2}{\mathbb{H}^c_{\mathrm{DNG}}(x) + \mathbb{H}^c_{\mathrm{Real}}(x)},$$

Color component에 따른 인접한 pixel 값들의 연관성

- H^c_{DNG} : 진짜 이미지의 히스토그램
- X : bin index
- 본 논문은 DNG 이미지의 set에서, 각 이미지당 r_i^c를 계산한고, r_i^c의 히스토그램인 H^c_{DNG}를 구축함.
- Chi-square distance를 통해 두 히스토그램의 유사도를 판단함.
- 이 metric의 값이 커질수록 DNG 이미지와 진짜 이미지의 식별가능성이 더욱 커짐.

Analysis from the Perspective of Color

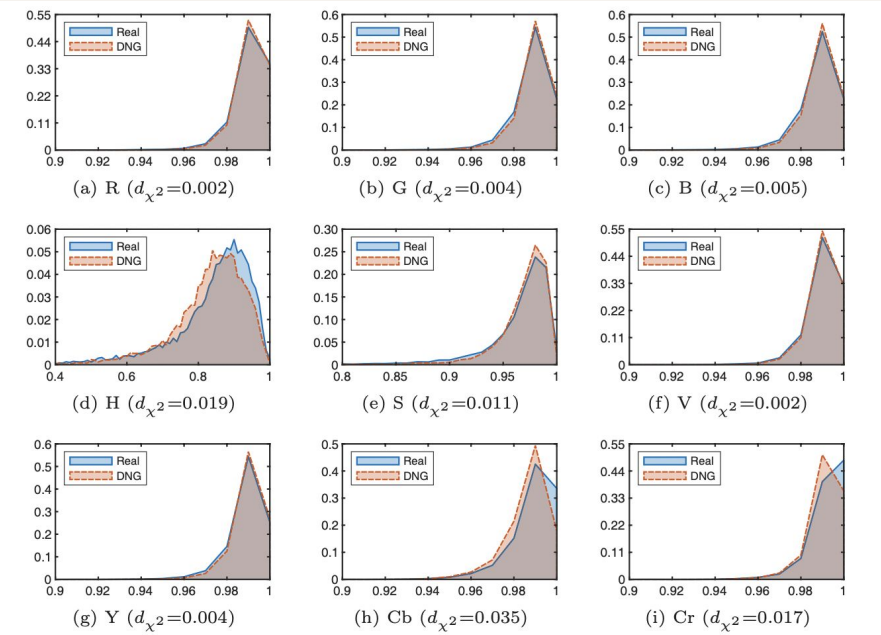


Figure 1: The histograms $\mathbb{H}^c_{\text{DNG}}$ (red) and $\mathbb{H}^c_{\text{Real}}$ (blue) for different color components. The values of $d_{\chi^2}(\mathbb{H}^c_{\text{DNG}}, \mathbb{H}^c_{\text{Real}})$ are included in the sub-captions.

- 진짜 이미지와 DNG 이미지의 각 color component별 히스토그램을 표현.
- 특정 color component에서 진짜 이미지와 DNG 이미지의 히스토그램이 차이가 많이 남을 확인할 수 있음.
- 하지만 이미지의 color component에서 항상 확정적으로 content를 구분 가능하지 않음.

Analysis from the Perspective of Color

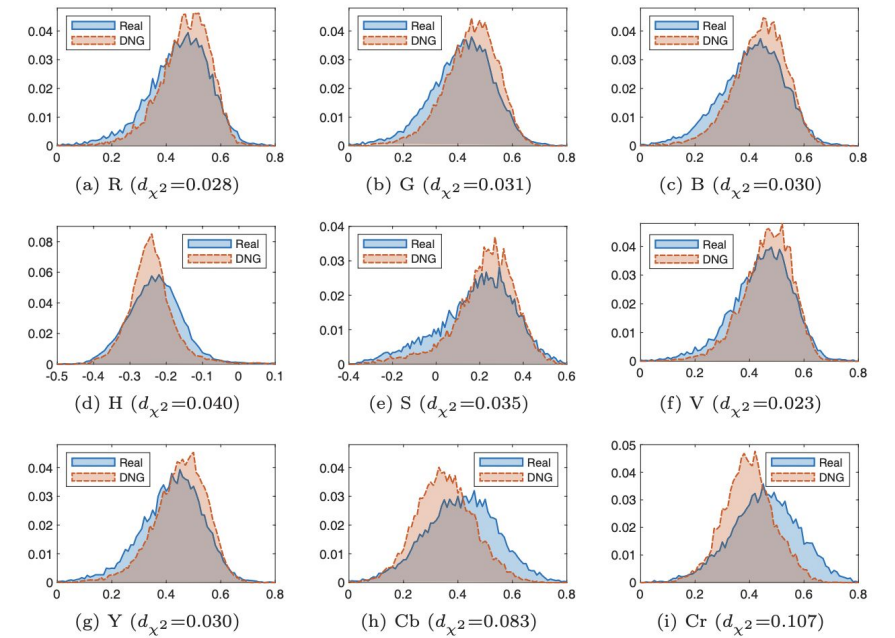


Figure 2: The histograms $\mathbb{H}^c_{\text{DNG}}$ (red) and $\mathbb{H}^c_{\text{Real}}$ (blue) for different color components in the residual domain. The values of $d_{\chi^2}(\mathbb{H}^c_{\text{DNG}}, \mathbb{H}^c_{\text{Real}})$ are included in the sub-captions.

$$\mathbf{R}^c_{j,k} = \mathbf{I}^c_{j,k} - \mathbf{I}^c_{j,k+1},\ c \in \{\text{R}, \text{G}, \text{B}, \text{H}, \text{S}, \text{V}, \text{Y}, \text{Cb}, \text{Cr}\}.$$

- 본 논문에서는 1차 미분 operator을 예로 적용해 이미지의 residual을 구함.
- $\mathbf{I}^c$: 이미지 I의 c번째 구성요소
- $\mathbf{R}^c$: I에 대응되는 residual
  - 본 논문에서는 horizontal 차이면 고려
- 진짜 이미지와 DNG 이미지는 residual domain에서 더욱 구분하기가 용이함을 확인할 수 있음.
- Residual 값 중 일정 범위를 벗어난 값을 특정 값으로 초기화해 residual 안의 distinct element를 줄임

Detection Scenarios and Strategies

Matched training-testing data

- 이 케이스에서 DNG 이미지는 동일한 진짜 이미지로 학습된 하나의 모델에서 생성됨.

Mismatched training-testing data

- 이 케이스에서 DNG 이미지는 다른 source로부터 생성됨.
    - 결과가 좋을수록 실제 상황에서 판별을 잘 진행할 확률이 높음.

Model-unaware case

- DNG 이미지가 어떠한 모델에서 생성되었는지 알 수 없는 경우를 대비하여 판별자를 진짜 이미지로만 학습시키고, 테스트 이미지가 진짜인지 가짜인지 판별함.

Overall framework

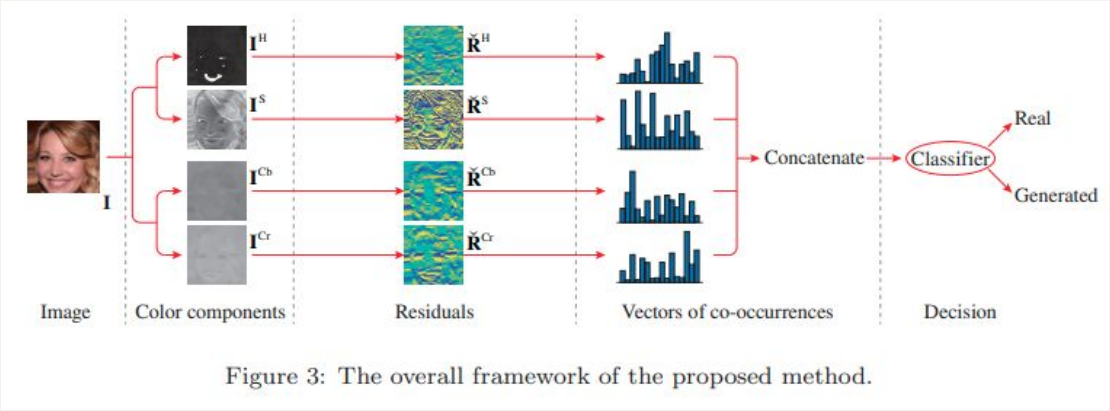


Figure 3: The overall framework of the proposed method.

- Residual 공간에서 도출된 co-occurrences vector을 모두 합침.
- 마지막 부분의 classifier을 학습시켜 진짜 이미지와 DNG 이미지를 구분할 수 있도록 함.

Performance

Table 6: Detection accuracies (%) obtained by **one-class classification**.

v	Train dataset	Method	$\mathcal{R}_{\text{F-LR}} + \mathcal{R}_{\text{B-LR}}$ Best	Worst	Average
0.10	$\mathcal{R}_{\text{F-LR}} + \mathcal{R}_{\text{B-LR}}$	Sub-SRM [41]	88.99	100.0	12.56	65.94
		SRM [54]	89.81	100.0	2.15	53.32
		CoALBP+LPQ [25]	89.64	100.0	13.49	67.80
		Proposed	89.90	100.0	99.85	**99.94**
0.05	$\mathcal{R}_{\text{F-LR}} + \mathcal{R}_{\text{B-LR}}$	Sub-SRM [41]	93.76	100.0	8.00	62.82
		SRM [54]	94.93	100.0	0.55	51.51
		CoALBP+LPQ [25]	94.66	100.0	4.35	59.45
		Proposed	94.91	100.0	99.19	**99.69**
v	Train dataset	Method	$\mathcal{R}_{\text{F-HR}} + \mathcal{R}_{\text{B-HR}}$ Best	Worst	Average
0.10	$\mathcal{R}_{\text{F-HR}} + \mathcal{R}_{\text{B-HR}}$	Sub-SRM [41]	89.65	99.95	4.44	46.38
		SRM [54]	90.15	77.44	5.67	30.71
		CoALBP+LPQ [25]	90.04	67.62	1.33	27.04
		Proposed	89.86	98.09	4.89	**49.82**
0.05	$\mathcal{R}_{\text{F-HR}} + \mathcal{R}_{\text{B-HR}}$	Sub-SRM [41]	94.29	99.81	1.92	**41.26**
		SRM [54]	95.03	64.14	2.56	22.66
		CoALBP+LPQ [25]	95.20	50.51	0.45	17.60
		Proposed	94.78	94.57	2.02	39.49

† $\mathcal{R}_{\text{F-LR}} + \mathcal{R}_{\text{B-LR}}$ denotes the combination of LR real image datasets $\mathcal{R}_{\text{F-LR}}$ and $\mathcal{R}_{\text{B-LR}}$, and $\mathcal{R}_{\text{F-HR}} + \mathcal{R}_{\text{B-HR}}$ denotes the combination of HR real image datasets $\mathcal{R}_{\text{F-HR}}$ and $\mathcal{R}_{\text{B-HR}}$.

Table 4: Classification results (%) for **mismatched image sources** (different semantic types).

Method	FPR Average (Best/Worst)	FNR Average (Best/Worst)	ACC Average (Best/Worst)
Sub-SRM [41]	23.72 (0.00 / 99.96)	17.54 (0.00 / 84.88)	79.37 (99.94 / 50.02)
SRM [54]	23.04 (0.00 / 99.58)	17.77 (0.00 / 99.00)	79.59 (100.0 / 50.21)
CoALBP+LPQ [25]	28.08 (0.00 / 98.77)	12.86 (0.00 / 88.28)	79.53 (99.99 / 49.74)
Sat-Cues [27]	42.72 (16.11 / 62.53)	45.54 (16.64 / 78.97)	55.87 (64.15 / 50.46)
VGG-16 [59]	19.81 (0.01 / 79.47)	61.26 (3.74 / 99.98)	59.46 (91.05 / 49.99)
ResNet_v2-50 [60]	4.30 (0.00 / 23.40)	95.19 (75.28 / 99.99)	50.26 (51.34 / 49.64)
Mo *et al.* [29]	17.45 (0.02 / 87.48)	50.01 (0.08 / 99.99)	66.27 (99.95 / 48.48)
CGFace [30]	16.77 (0.07 / 44.16)	63.87 (0.82 / 99.85)	59.68 (81.97 / 45.96)
TS-CDNN [31]	11.71 (0.01 / 28.37)	79.40 (49.92 / 99.97)	54.45 (72.13 / 50.02)
Proposed	28.33 (0.00 / 96.29)	9.45 (0.00 / 41.53)	**81.11** (99.99 / 51.85)

Table 5: Classification results (%) for **mismatched GAN models** in training and testing.

Method	FPR Average (Best/Worst)	FNR Average (Best/Worst)	ACC Average (Best/Worst)
Sub-SRM [41]	0.06 (0.00 / 0.82)	39.45 (0.00 / 100.0)	80.25 (100.0 / 50.00)
SRM [54]	0.02 (0.00 / 0.19)	35.11 (0.00 / 100.0)	82.44 (100.0 / 50.00)
CoALBP+LPQ [25]	0.01 (0.00 / 0.08)	22.05 (0.00 / 100.0)	88.97 (100.0 / 49.99)
Sat-Cues [27]	26.79 (17.77 / 42.76)	40.18 (17.19 / 63.48)	66.52 (78.43 / 48.82)
VGG-16 [59]	0.34 (0.01 / 1.32)	77.09 (0.35 / 99.94)	61.29 (99.77 / 49.92)
ResNet_v2-50 [60]	0.54 (0.01 / 1.99)	82.09 (12.82 / 100.0)	58.68 (93.56 / 49.64)
Mo *et al.* [29]	0.32 (0.00 / 3.51)	58.04 (0.00 / 99.99)	70.82 (99.95 / 48.39)
CGFace [30]	6.87 (0.04 / 37.81)	71.18 (0.20 / 99.91)	60.98 (99.84 / 48.54)
TS-CDNN [31]	6.96 (0.04 / 22.32)	84.19 (55.68 / 99.96)	54.42 (68.16 / 46.81)
Proposed	0.14 (0.00 / 1.96)	16.12 (0.00 / 100.0)	**91.87** (100.0 / 50.00)

Q&A