Lecture 24: Image forensics
Announcements

• Final project guidelines are on webpage.
• Sign up for a presentation time slot here.
• PS9 (on NeRF) will be released by tomorrow.
  • Shorter than usual (to give you time for the project).
Fake images in the news

How misinformation helped spark an attempted coup in Gabon

Analysis by Sarah Cohen
Video reporter
February 13, 2020 at 3:00 a.m. EST

Paparazzi Photos Were the Scourge of Celebrities. Now, It's AI.
Researchers say advancements in artificial intelligence could be used to stoke misinformation about public figures. A recent image had even experts fooled.

#ALIBONGO
#GABON
#DEEPEFAKE
#LIAR
#ALIBONGO

The suspicious video that helped spark an attempted coup in Gabon
8:24

Gabon’s president was ill. He had not been seen in public for months. A week after his first video address, there was an attempted coup. (Video: Sarah Cohen/The Washington Post)
“Catholic Pope Francis wearing Balenciaga puffy jacket in drill rap music video, throwing up gang signs with hands, taken using a Canon EOS R camera with a 50mm f/1.8 lens, f/2.2 aperture, shutter speed 1/200s, ISO 100 and natural light, Full Body, Hyper Realistic Photography, Cinematic, Cinema, Hyperdetail, UHD, Color Correction, hdr, color grading, hyper realistic CG animation --ar 4:5 --upbeta --q 2 --v 5.”
But image manipulation also has a long history

Abraham Lincoln?  John C. Calhoun
But image manipulation also has a long history

From Forrest Gump, 1994
Malicious image manipulation
Malicious image manipulation
Malicious image manipulation

Fonda Speaks To Vietnam Veterans At Anti-War Rally

Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans. As activist and former Vietnam Vet, John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)
his associates simply found photos of athletes on the Internet and either used those photos or used software such as PhotoShop to insert the applicants’ faces onto the bodies of legitimate athletes. For example, as set forth in greater detail below, CW-1 explained to McGLASHAN that he would create a falsified athletic profile for McGLASHAN’s son, something he told McGLASHAN he had “already done … a million times,” and which would involve him using “Photoshop and stuff” to deceive university admissions officers.

FBI affidavit on 2019 college admissions scandal
Malicious image manipulation
Malicious image manipulation
New image manipulation methods are emerging every day

- ProGAN (2018)
- StyleGAN2 (2019)
- StyleGAN3 (2020)
- DALL-E 2 (2021)
- Midjourney (2023)
- Midjourney (2024)
The challenge of fake image detection

Training data
Images from known methods
- ProGAN
- StyleGAN2

Test data
Images from future methods
- StyleGAN3
- DALL-E 2
- Midjourney
- github.com/.../mycoolgenerator

2018 2019 2020 2021 2022 2023 2024
Hard to directly use supervised learning!
Strategy #1: physical models
Self-consistent lighting direction

[Johnson and Farid, 2005]
Specular reflections

[Johnson and Farid, 2007]
Strategy #2: low-level imaging properties
JPEG artifacts

- Cameras vary in how they do JPEG compression.
- When you quantize a floating point numbers:
  - Some do \texttt{round()}, others do \texttt{floor()} or \texttt{ceil()}
  - If a photo seems to have both kinds of quantization, it’s probably a fake: e.g., a composite from images taken by different cameras!

[Agarwal and Farid, “JPEG Dimples”, 2017]
Detecting duplicated image regions

- Traditional inpainting methods copy-and-paste image patches.
- Detect near-duplicated patches.
- But sensitive to postprocessing operations, like compression.

← amount of JPEG compression

[Popescu and Farid, 2004]
Strategy #3: learned anomaly detection

Instead of hand-crafting cues, can we learn to detect “anomalous” images, and flag suspicious images?

Predicting metadata consistency

CameraMake: Apple
CameraModel: iPhone 4s
ColorSpace: sRGB
ExifImageLength: 2448
ExifImageWidth: 3264
Flash: Flash did not fire
FocalLength: 107/2
WhiteBalance: Auto
ExposureTime: 1/2208
...

Same white balance?

Different

CameraMake: NIKON CORPORATION
CameraModel: NIKON D90
ColorSpace: sRGB
ExifImageLength: 2848
ExifImageWidth: 4288
Flash: Flash did not fire
FocalLength: 18/796
WhiteBalance: Auto
ExposureTime: 1/30
...
Input

Photo source: TheOnion.com
Another approach: learning joint embeddings
Learning Joint Embeddings

Make: NIKON
Model: NIKON D3200
Flash: Fired
Exposure Time: 1/500
Focal Length: 30.0mm
Exposure Program: Aperture
Components Configuration: YCbCr

“Make: NIKON, Model: NIKON D3200, Flash: Fired, Exposure Time: 1/500, Focal Length: 30.0mm, Exposure Program: Aperture, Components Configuration: YCbCr, Scene Capture Type: Standard, ...”
Learning Joint Embeddings

“Make: NIKON, Model: NIKON D3200, Flash: Fired, Exposure Time: 1/500, Focal Length: 30.0mm, Exposure Program: Aperture, Components Configuration: YCbCr, Scene Capture Type: Standard, …”
Learning Joint Embeddings

Patch Encoder

EXIF Text Encoder

"Make: NIKON, Model: NIKON D3200, Flash: Fired, Exposure Time: 1/500, Focal Length: 30.0mm, Exposure Program: Aperture, Components Configuration: YCbCr, Scene Capture Type: Standard, ... "

I₁, I₂, I₃, I₄, I₅

I₁·T₁, I₂·T₁, I₃·T₁, I₄·T₁, I₅·T₁

I₂·T₂, I₂·T₂, I₃·T₂, I₄·T₂, I₅·T₂

I₃·T₃, I₃·T₃, I₃·T₃, I₃·T₃, I₅·T₃

I₄·T₄, I₄·T₄, I₄·T₄, I₄·T₄, I₅·T₄

I₅·T₅, I₅·T₅, I₅·T₅, I₅·T₅, I₅·T₅
Linear classification evaluation

Radial distortion

Image manipulation
Linear classification evaluation

Radial distortion estimation (Dresden dataset)

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>ImageNet</th>
<th>CLIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc. predicting quantized $\kappa_1$</td>
<td>31%</td>
<td>23%</td>
<td>21%</td>
</tr>
</tbody>
</table>

Image splice detection (CASIA I dataset)

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>ImageNet</th>
<th>CLIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real vs. fake accuracy</td>
<td>75%</td>
<td>69%</td>
<td>71%</td>
</tr>
</tbody>
</table>
Video forensics as anomaly detection

Training

\[ f \rightarrow \text{Density } p_\theta \]

Testing

Real (inlier)

Fake (outlier)

e.g., “deepfake” videos

[Feng, Chen, Owens, 2023]
Data representation

Raw pixels? Just as hard as generation!
Instead: self-supervised feature space.

Source: [Ho et al., 2020]
Data representation

Synchronization model of [Chen et al., 2021]
Learn the distribution for self-supervised features

Self-supervised feature learning

Audio-visual Synchronization Model

Discrete time delays
Distribution over delays
Feature activations
Learn the distribution for self-supervised features

Self-supervised feature learning

Audio-visual Synchronization Model

\[-1 \quad +1 \quad \cdots \quad -2\]
Discrete time delays

\[\text{Prob.} \quad \text{Distribution over delays}\]

Feature activations

Audio-visual anomaly detection

Autoregressive Prediction

\[\mathcal{L}\]
Likelihood

Target features

Predicted features

Time

···
Learn the distribution for self-supervised features

**Self-supervised feature learning**

Audio-visual Synchronization Model

-1 +1 ... -2
Discrete time delays

V A
Distribution over delays

Feature activations

**Audio-visual anomaly detection**

Time

Autoregressive Prediction

\[ \mathcal{L} \]
Predicted features

Target features

Likelihood
Learn the distribution for self-supervised features

**Stage #1:** Learning audio-visual synchronization feature sets:

\[ S(i, j) = \frac{\exp(\phi(V_i, A_j))}{\sum_{k=i-\tau}^{i+\tau} \exp(\phi(V_i, A_k))} \]

**Stage #2:** Learning autoregressive model on self-supervised audio-visual feature sets:

\[ p_\theta(x_1, x_2, \cdots, x_N) = \prod_{i=0}^{N-1} p_\theta(x_{i+1}|x_1, \cdots, x_i) \]
Results

FakeAVCeleb [Khalid et al., 2021]

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception</td>
<td>85.5</td>
<td>85.3</td>
</tr>
<tr>
<td>LipForensics</td>
<td>89.4</td>
<td>91.1</td>
</tr>
<tr>
<td>AD DFD</td>
<td>88.8</td>
<td>88.1</td>
</tr>
<tr>
<td>FTCN</td>
<td>92.3</td>
<td>93.1</td>
</tr>
<tr>
<td>RealForensics</td>
<td>95.3</td>
<td>97.1</td>
</tr>
<tr>
<td>Ours</td>
<td>94.2</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Robustness to postprocessing

Limitation: only works for out-of-sync lip motions (not face swaps)
Strategy #4: supervised learning
Make **random** fakes by scripting Photoshop.

```python
def make_random_fakes():
    detect and crop face;
    open Photoshop;
    open Face-Aware Liquify;
    move mouth keypoint 1;
    ...
    save(warped image);
```

Photoshop Face-Aware Liquify tutorial. Source: https://youtu.be/5Qqv_C6iVvQ?t=86
Warp detector
Real-or-fake classification

<table>
<thead>
<tr>
<th></th>
<th>% accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>99.4%</td>
</tr>
<tr>
<td>Human</td>
<td>53.5%</td>
</tr>
</tbody>
</table>
What moved where?

Manipulated Image → Dilated ResNet → Warp Prediction
What moved where?
Manipulated Photo
Suggested “Undo”
Manipulated vs. Original
Undo vs. Original
Manipulated Photo
Warp Prediction
Suggested “Undo”
Suggested “Undo”
Similar approaches for “deepfakes”

Create lots of deepfake videos, then learn to detect them.

[Rossler et al., “FaceForensics++”, 2019]
New challenges on the horizon

Celeb-DF: A New Dataset for DeepFake Forensics

Yuezun Li¹, Xin Yang¹, Pu Sun², Honggang Qi² and Siwei Lyu¹

¹University at Albany, State University of New York, USA
²University of Chinese Academy of Sciences, China

[Li et al., “Celeb-DF”, 2020]
The forensics generalization problem

New architectures & datasets

StyleGAN2 [Karras 2019]

New models

Cascaded refinement networks [Chen & Koltun 2017]

Lots of potential issues for “universal” detector: dataset bias, domain adaptation, etc.
CNN-generated images are surprisingly easy to spot... for now

Sheng-Yu Wang  Oliver Wang  Richard Zhang  Andrew Owens  Alexei Efros

https://peterwang512.github.io/CNNDetection
Dataset of CNN-generated fakes

GANs
- ProGAN (Karras 2018)
- StyleGAN (Karras 2018)
- BigGAN (Brock 2019)
- CycleGAN (Zhu 2017)
- StarGAN (Choi 2018)
- GauGAN (Park 2019)

Perceptual loss
- Cascaded refinement (Chen 2017)
- IMLE (Li 2019)

Low-level vision
- Seeing in the dark (Chen 2018)
- Super-resolution (Dai 2019)

Deep fakes
- Faceswap (Anonymous 2018, Rossler 2019)
Dataset of CNN-generated fakes

GANs
- ProGAN (Karras 2018)
- StyleGAN (Karras 2018)
- BigGAN (Brock 2019)
- CycleGAN (Zhu 2017)
- StarGAN (Choi 2018)
- GauGAN (Park 2019)

Perceptual loss
- Cascaded refinement (Chen 2017)
- IMLE (Li 2019)

Low-level vision
- Seeing in the dark (Chen 2018)
- Super-resolution (Dai 2019)

Deep fakes
- Faceswap (Anonymous 2018)
- Rossler (2019)
How well do classifiers generalize?

- Train with 720K images from 20 LSUN categories
- JPEG + Blurring data augmentation

Real images

ProGAN

Real vs. fake?
How well do classifiers generalize?

Synthesized images from another CNN

Real “target” images

ProGAN detector

Real vs. fake?
How well do classifiers generalize?

Images the CNN **actually** makes

Images the CNN **should** make

ProGAN detector

Real vs. fake?
Surprising amounts of generalization

Average precision

<table>
<thead>
<tr>
<th>Method</th>
<th>100</th>
<th>100</th>
<th>99</th>
<th>99</th>
<th>98</th>
<th>97</th>
<th>95</th>
<th>93</th>
<th>88</th>
<th>66</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StyleGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GauGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CycleGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StarGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seeing dark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BigGAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep fake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Super-res.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Generalization to other CNNs: no preprocessing

Average precision

<table>
<thead>
<tr>
<th>Model</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProGAN</td>
<td>100</td>
</tr>
<tr>
<td>IMLE</td>
<td>90</td>
</tr>
<tr>
<td>CRN</td>
<td>94</td>
</tr>
<tr>
<td>StyleGAN</td>
<td>96</td>
</tr>
<tr>
<td>GauGAN</td>
<td>67</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>84</td>
</tr>
<tr>
<td>StarGAN</td>
<td>100</td>
</tr>
<tr>
<td>Seeing dark</td>
<td>96</td>
</tr>
<tr>
<td>BigGAN</td>
<td>72</td>
</tr>
<tr>
<td>Deep fake</td>
<td>98</td>
</tr>
<tr>
<td>Super-res.</td>
<td>94</td>
</tr>
</tbody>
</table>
Generalization example

Detection accuracy: 93% AP

"Out-of-distribution" dataset:

- StyleGAN faces
- 1024x1024 JPEGs
- Use minimal preprocessing: take 224x224 center crop

http://whichfaceisreal.com [West and Bergstrom 2019]
A model trained on a model from 2019 (ProGAN) generalizes to a (similar) model in 2021 (StyleGAN3)
Implications

• Suggests CNN-generated images have common artifacts
• These artifacts can be detected with a simple classifier!
• But what are these artifacts?

Example from literature: checkerboard/aliasing artifacts [Xu Zhang et al. 2019]
Need online “open world” detection

Source: [Epstein et al., “Online Detection of AI-Generated Images”, 2023],
See also [Girish et al., “Towards discovery and attribution of open-world gan generated images”, 2021]
What’s real and what’s fake?

[“The suspicious video that helped spark an attempted coup in Gabon” Washington Post. 2020]
https://www.youtube.com/watch?v=F5yzKs4z1dc
Challenges on the horizon

- Lots of ways to make fake images.
- If we know what methods were used, there’s a good chance we can succeed.
- But it’s hard to capture all of them!
- False positives are still a huge problem.
- So are postprocessing operations, like cropping and compression.
- Need methods that can handle unseen models.
- Alternative approaches: watermarking, signatures, etc.
Open-ended discussion

• How susceptible are people to fake images?
• Is there any hope of detecting “most” fake images?
• Under what situations might it be important and/or feasible?
• How do we deal with false positives?