An analysis of image filtering on WeChat Moments

Jeffrey Knockel, Lotus Ruan, Masashi Crete-Nishihata

Background

- Images increasingly used to communicate
- Image censorship understudied
- (Website blocking, text chat/posts, etc.)

WeChat Moments

- WeChat has over 1 billion active users
- Images are most frequent content on WeChat Moments
- Previous work systematically looked at text
- Known to automatically filter politically sensitive images for China-based accounts

12:13 PM

Moments

China

China
Sent
1 minute ago Delete

China

1 minute ago Delete

China
Testing an infograph of the 709 case
1 minute ago Delete

12:13

Moments

Canada

China
Sent
1 minute ago

China

1 minute ago

China
Testing an infograph of the 709 case
1 minute ago

Fido 12:46

Discover Moments

China
Sent
34 mins ago

China
Testing an infograph of the 709 case
34 mins ago

Source: https://isc.sans.edu/forums/diary/23395

新浪微博搜索敏感词列表

新浪微博搜索敏感词列表（更新中）

	敏感词	Sensitive Word	英文帖
3	1984		
4	小熊维尼		
5	歪脖子树		
6	长生不老		
7	不要脸		
8	倒行逆施		
	倒车		
	我反对		
1	无限续杯		
12	吾皇万岁		
13	千秋万代		
14	一统江湖		
15	黄袍加身		
16	昏君		
17	登基		
18	称帝		
	终身		
	年号		
1	复辟		
22	元气		
23	劝进		
24	封禅		
25	张勋		
26	蔡锷		
27	袁世凯		

Source: https://isc.sans.edu/forums/diary/23395

- Why didn’t the wavey thing evade?
- Why did the scribble evade? Does doing the scribble always evade?

- We want *effective* techniques
- We want *principles-based* techniques (based on understanding principles of how the filter works)

How we develop evasion techniques

1. Understand filter’s implementation details
    a. Modify otherwise filtered images
    b. See which modification evade filtering
2. Devise and test evasion strategies

How we develop evasion techniques

- By learning how to evade it we can learn how the filtering algorithm works
- By learning how the filtering algorithm works we can learn how to evade it

Our findings

- Two methods of filtering
- OCR-based (blacklisted keywords)
- Visual-based (blacklisted images)

“ 法輪大法好”

OCR:

“FALUN DAFA IS GOOD”

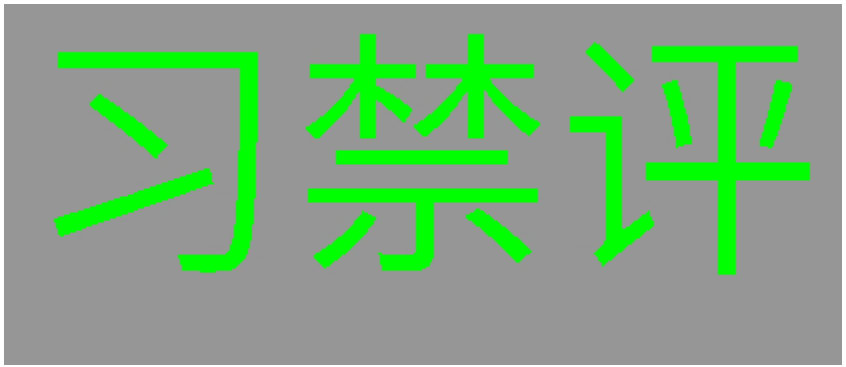
OCR performs grayscale conversion

Does WeChat use grayscale? How?

- Average
  $(r + g + b) / 3$

- Lightness
  $(\max(r + g + b) + \min(r + g + b)) / 2$

- Luminosity
  $0.299 \cdot r + 0.587 \cdot g + 0.114 \cdot b$



Background chosen to have same *luminosity* of text

If background is luminosity:

Average ✗

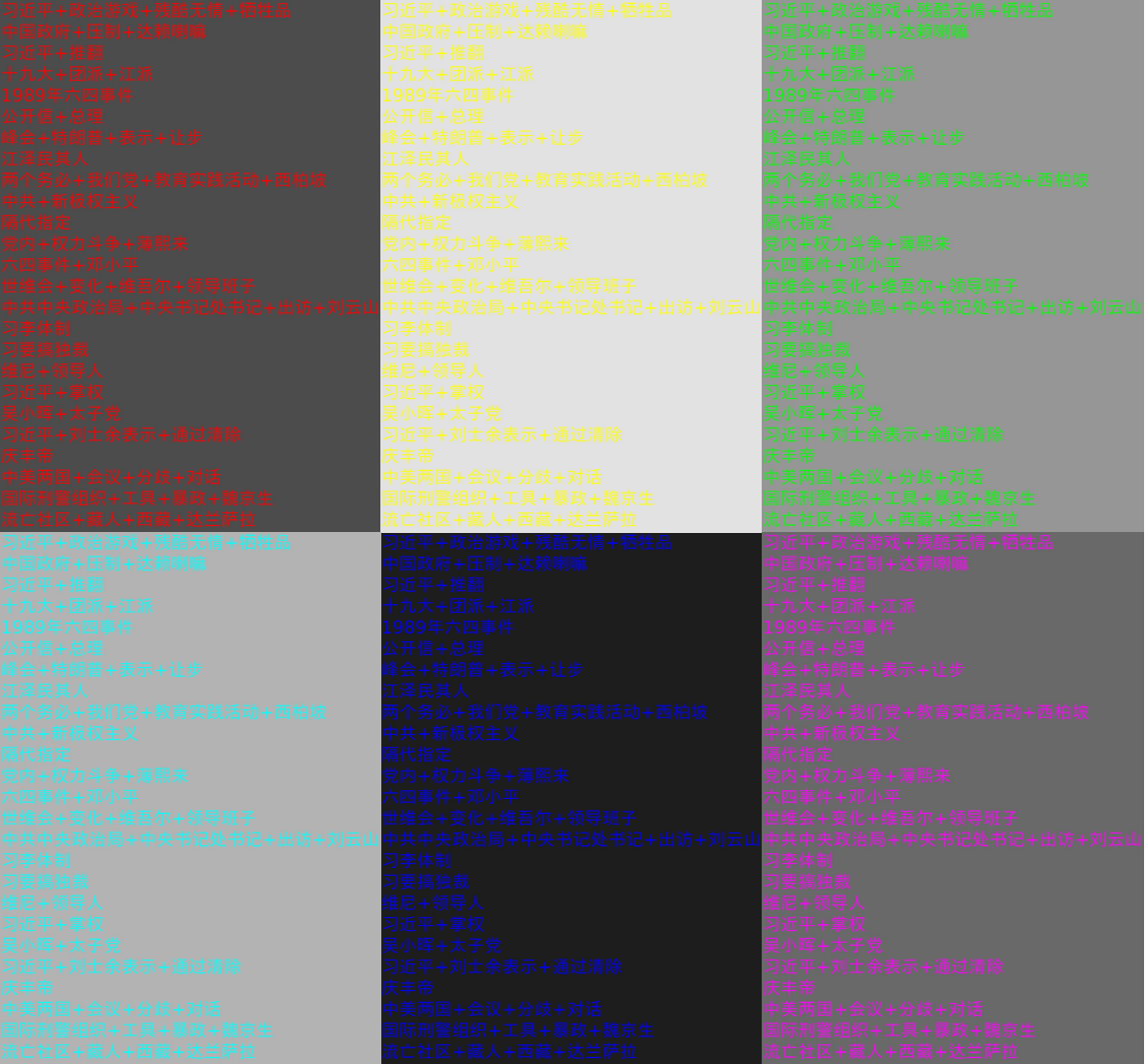
$(r + g + b) / 3$

Lightness ✗

$(\max(r + g + b) + \min(r + g + b)) / 2$

Luminosity ✓

$0.299 \cdot r + 0.587 \cdot g + 0.114 \cdot b$



Create messages where each line contains a blacklisted phrase.

Tested 6 colors…

台湾＋大陆＋谈话＋马英九

习近平权力＋写进党章＋堪比毛
赵紫阳录音回忆录

洛桑森格＋藏人行政
习近平的左右手
习主席＋皇帝
习特勒
习近平＋邓家贵

习近平思想＋修改党章＋十九大
习近平＋抗议
买官卖官＋查处＋跑官要官＋选人用人
全国人大＋告知书
习近平＋残酷无情＋政治游戏＋牺牲品
进入政治局＋十九大＋胡春华
孙文广＋政治改革＋最高领导人＋语焉不详
习近平的亲信＋陈敏尔
异曲同工＋李明哲案＋铜锣湾书店
设党主席

习近平＋政治游戏＋残酷无情＋牺牲品
中国政府＋压制＋达赖喇嘛
习近平＋推翻
十九大＋团派＋江派
1989年六四事件
公开信＋总理
峰会＋特朗普＋表示＋让步
江泽民其人
两个务必＋我们党＋教育实践活动＋西柏坡
中共＋新极权主义
隔代指定
党内＋权力斗争＋薄熙来
六四事件＋邓小平
世维会＋变化＋维吾尔＋领导班子
中共中央政治局＋中央书记处书记＋出访＋刘云山
习李体制
习要搞独裁
维尼＋领导人
习近平＋掌权
吴小晖＋太子党
习近平＋刘士余表示＋通过清除
庆丰帝
中美两国＋会议＋分歧＋对话
国际刑警组织＋工具＋暴政＋魏京生
流亡社区＋藏人＋西藏＋达兰萨拉

For each color,
vary the # of sensitive
phrases 5 times…

洛桑森格+藏人行政
习近平的左右手
习主席+皇帝
习特勒
习近平+邓家贵

江泽民其人
卡尔文森号+舰队+蒂勒森+采取行动
国家+外交政策+抵制+韬光养晦
中文网+倒数第一+自由度
北戴河+十九大

丹麦女王+大熊猫+玛格丽特+访华
修宪+连任
习近平+团派
中央纪委+主体责任+党风廉政建设+调研
黑马+十九大

郭文贵+高层
1989+民主运动
第六代领导人
习大大+向越南学习+自由迁徙
吸精瓶

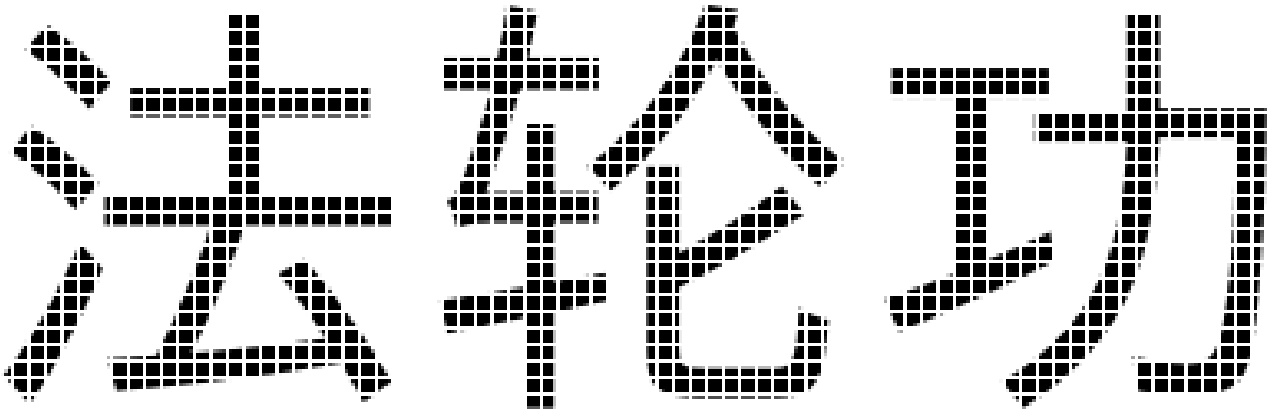
习近平+十九大+第二任期
习近平+女儿
十九大+造神运动
不同寻常+习近平+伙伴关系
全国人大+告知书

For each color and # of sensitive phrases we generated five messages…

All 150 messages evaded filtering!

OCR performs blob merging

Squares



Letters

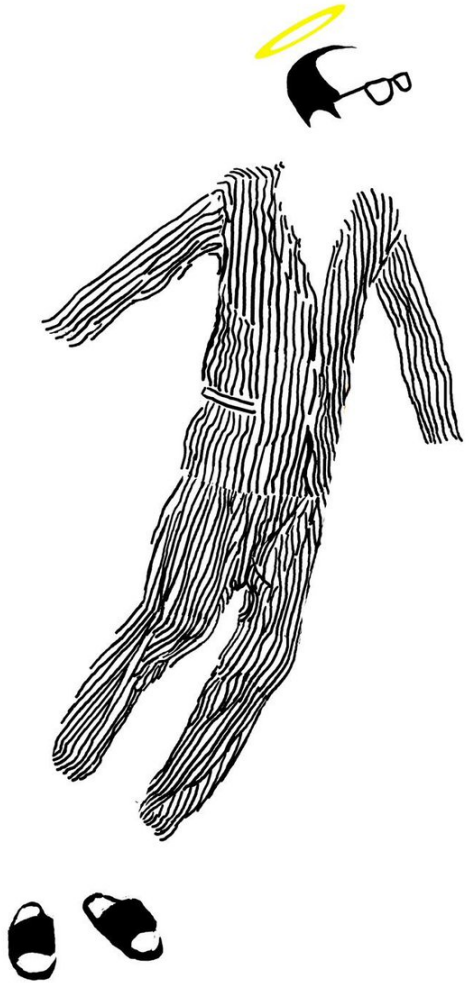
Varied the pattern (squares and letters)

Varied # of sensitive phrases 5 times

48/50 evaded filtering! ✔

Visual-based filtering

Works when image contains no text

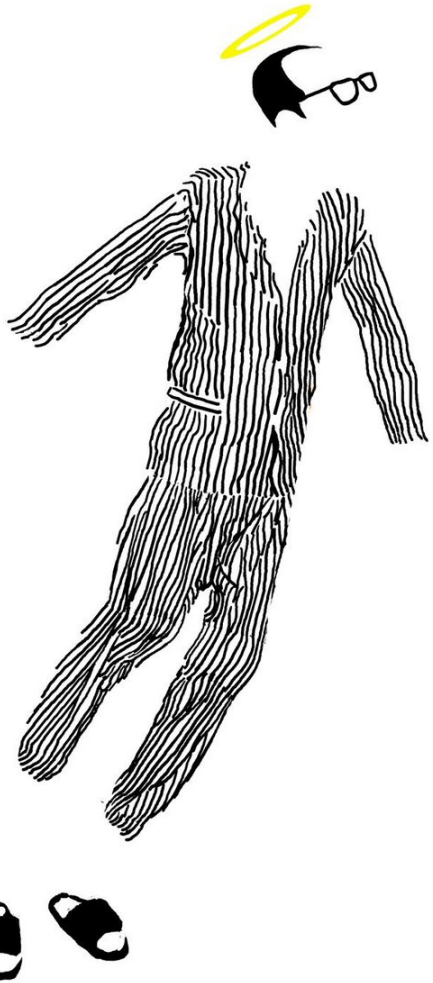


High level machine learning categorization?

Cat

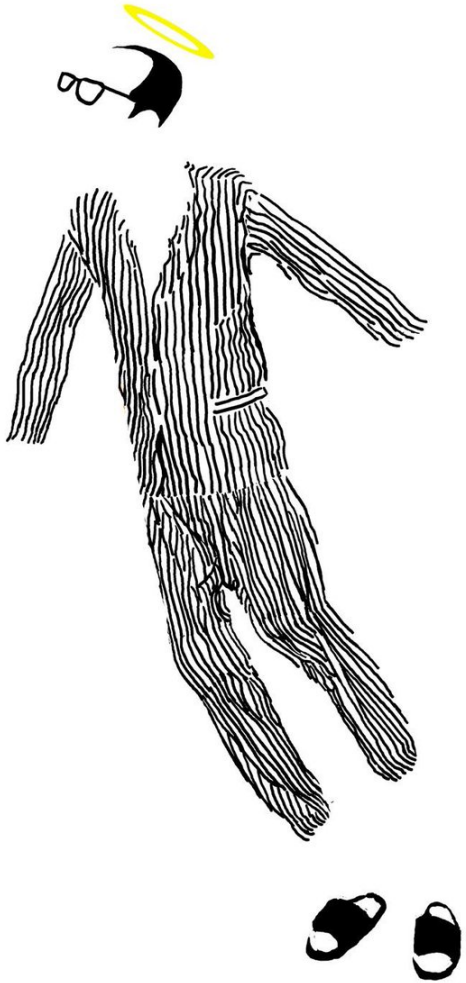
High level machine learning categorization?

Dog?



Mirroring consistently evaded filtering

So do some other simple modifications like removing/adding whitespace



High level machine learning categorization?

Training to recognize sensitive content would be difficult considering the…

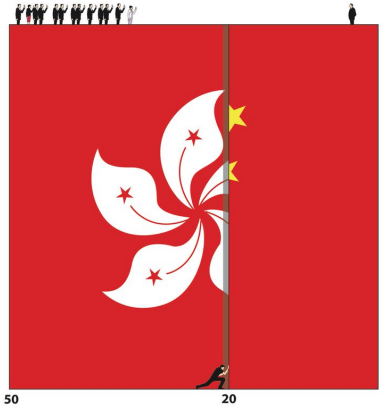
- subtlety of what makes something sensitive
- fluidity of what is considered sensitive

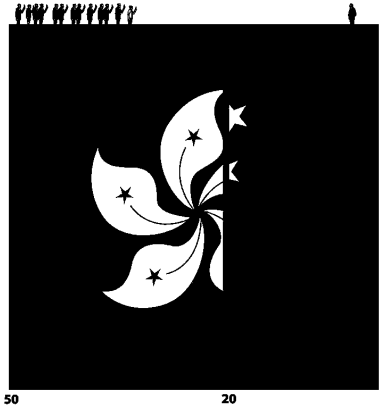
Is color important?

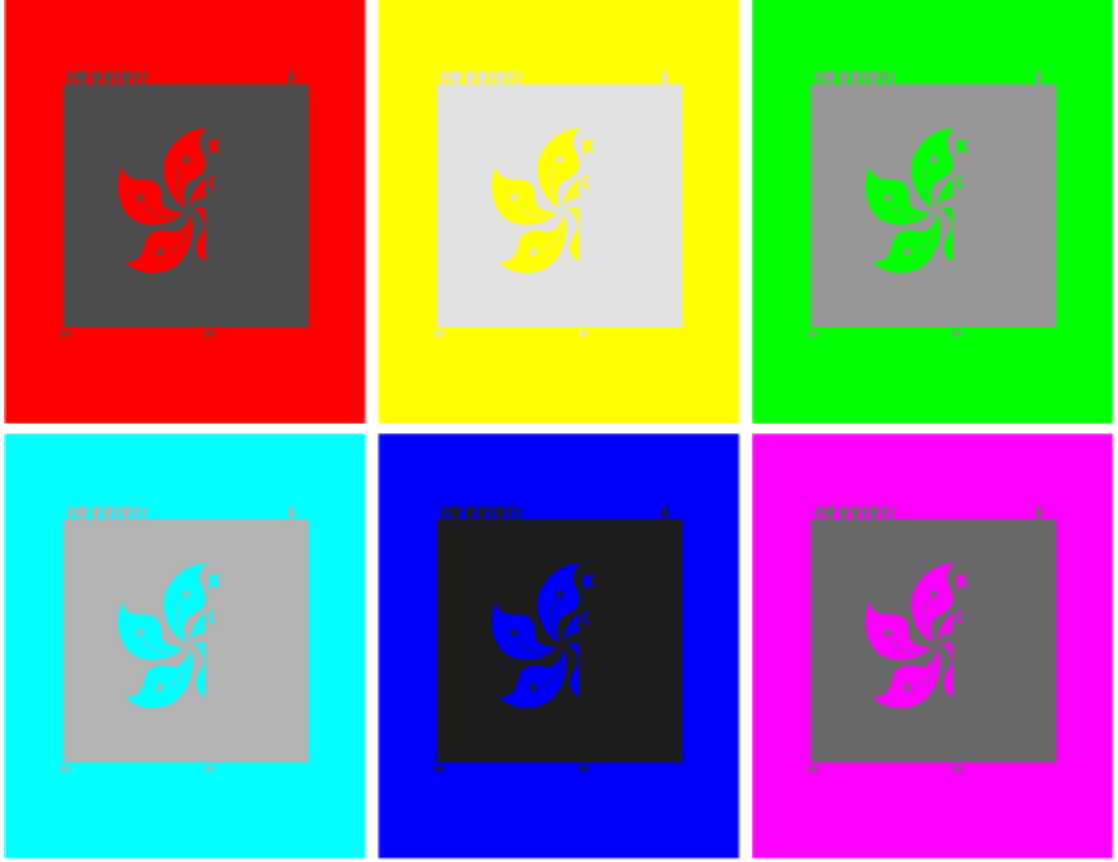
Converting images to grayscale never evaded filtering

Does it convert to grayscale? How?

Use same method we used to test OCR







Converts to grayscale using luminosity

Are edges important?

Are edges important?

Thresholding preserves edges, removes other information

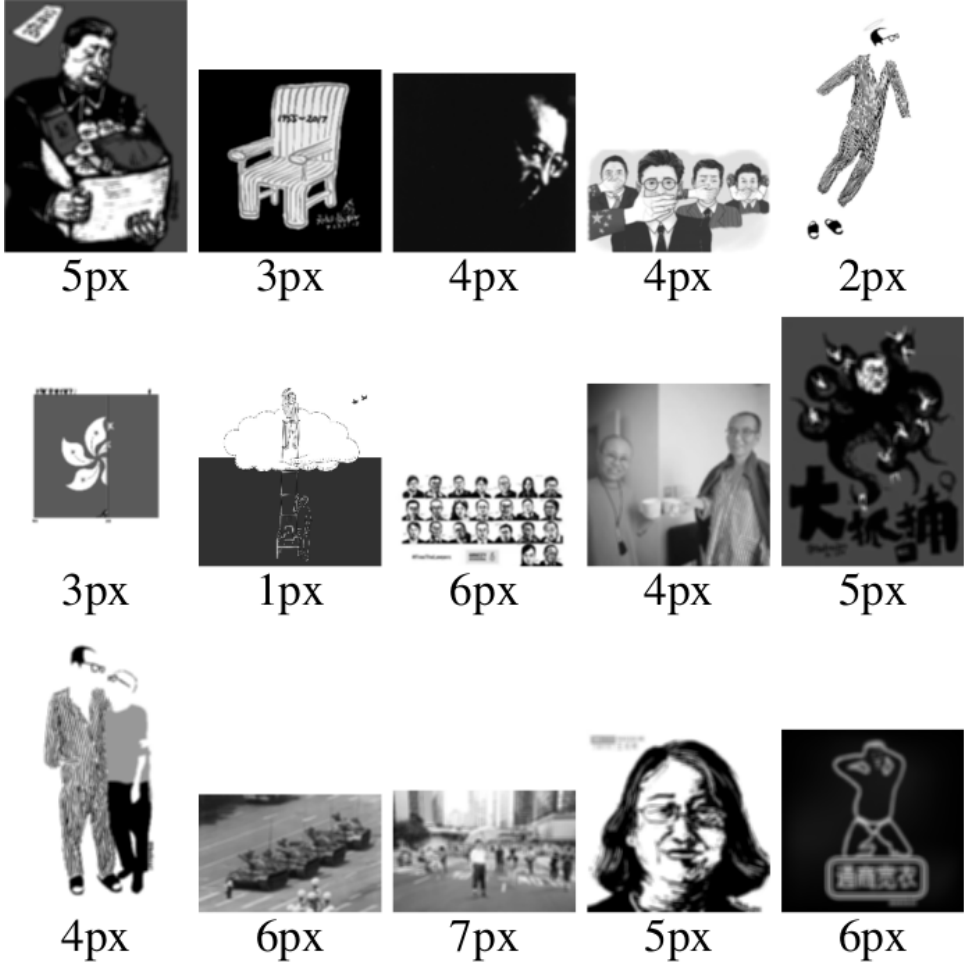
Thresholded 15 images, only 2 evaded

Are edges important?

Proportionally resized 15 images such that each image's smallest dimension(s) are 200 px.

How much can we blur before evasion?

Doesn't take much!



Largest normalized box filter kernel size

Are edges important?

How are images resized?

Hypotheses:

1. Proportionally such that their width is some value such as 100.
2. Proportionally such that their height is some value such as 100.
3. Proportionally such that their largest dimension is some value such as 100.
4. Proportionally such that their smallest dimension is some value such as 100.
5. Both dimensions are resized to some fixed size such as 100×100.

How are images resized?

Hypotheses:

5. ~~Both dimensions are resized to some fixed size such as 100×100.~~

Stretching an image evades filtering.

If space added to width but resizes by width or largest dimension, will not match

(the original)

(the original)

+

(space added to width, resized to same height)

(space added to width, resized to same width)

=

Correct hypothesis:

4. Proportionally such that their smallest dimension is some value such as 100.

Evade filtering by adding borders to the smallest dimension.

Adding surrounding content

Adding duplicate images generally evaded.
Full results are in our paper.

Conclusion

An effective image filter evasion strategy is one that modifies a sensitive image so that it…

1. no longer resembles a blacklisted image to the filter but
2. still resembles a blacklisted image to people reading it.

Evasion technique summary

- OCR-based evasion
  - By color (100%)
  - By blobs (96%)
- Visual-based evasion
  - Mirroring (100%)
  - Blurring (varies)
  - Stretching (97%)
  - Adding borders (80%)
  - Adding complex content around the image (varies)

Conclusion

We only looked at one platform, but we hope that this type of analysis provides a roadmap for looking at filtering on other platforms.

https://citizenlab.ca/2018/08/cant-picture-this-an-analysis-of-image-filtering-on-wechat-moments/

Questions?