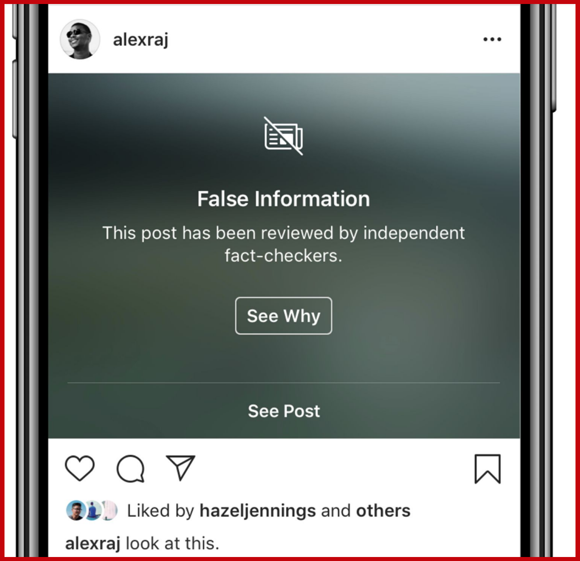
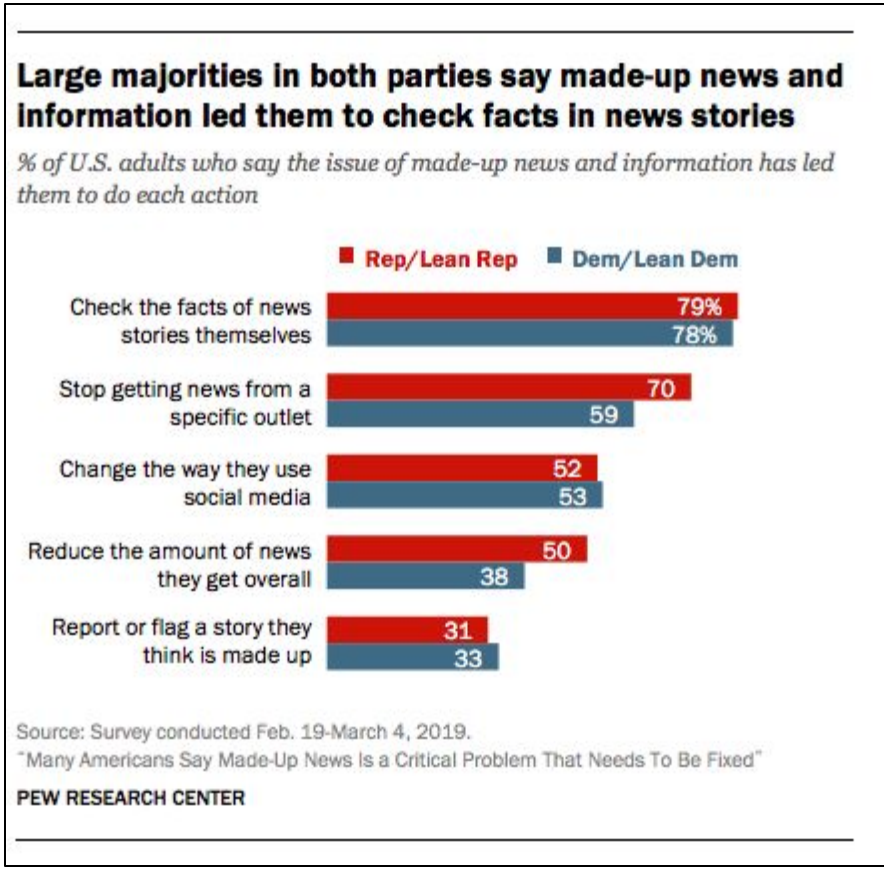
MASKuerade : Bypassing Instagram's Fact-Checking Using Adversarial Attacks

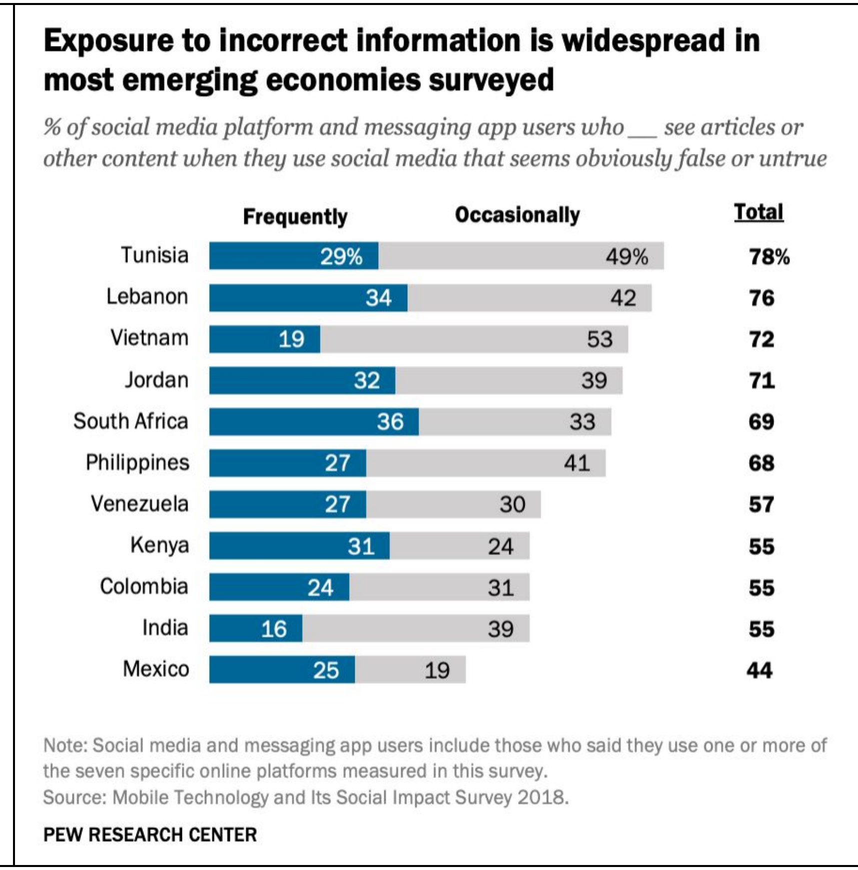


Team Members:
Harshita Singh
Arindam Das
Harsh Sahu

Background



Large majorities in both parties say made-up news and information led them to check facts in news stories
% of U.S. adults who say the issue of made-up news and information has led them to do each action
Rep/Lean Rep
Dem/Lean Dem
Check the facts of news stories themselves
79%
78%
Stop getting news from a specific outlet
70
59
Change the way they use social media
52
53
Reduce the amount of news they get overall
50
38
Report or flag a story they think is made up
31
33
Source: Survey conducted Feb. 19-March 4, 2019.
"Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed"
PEW RESEARCH CENTER

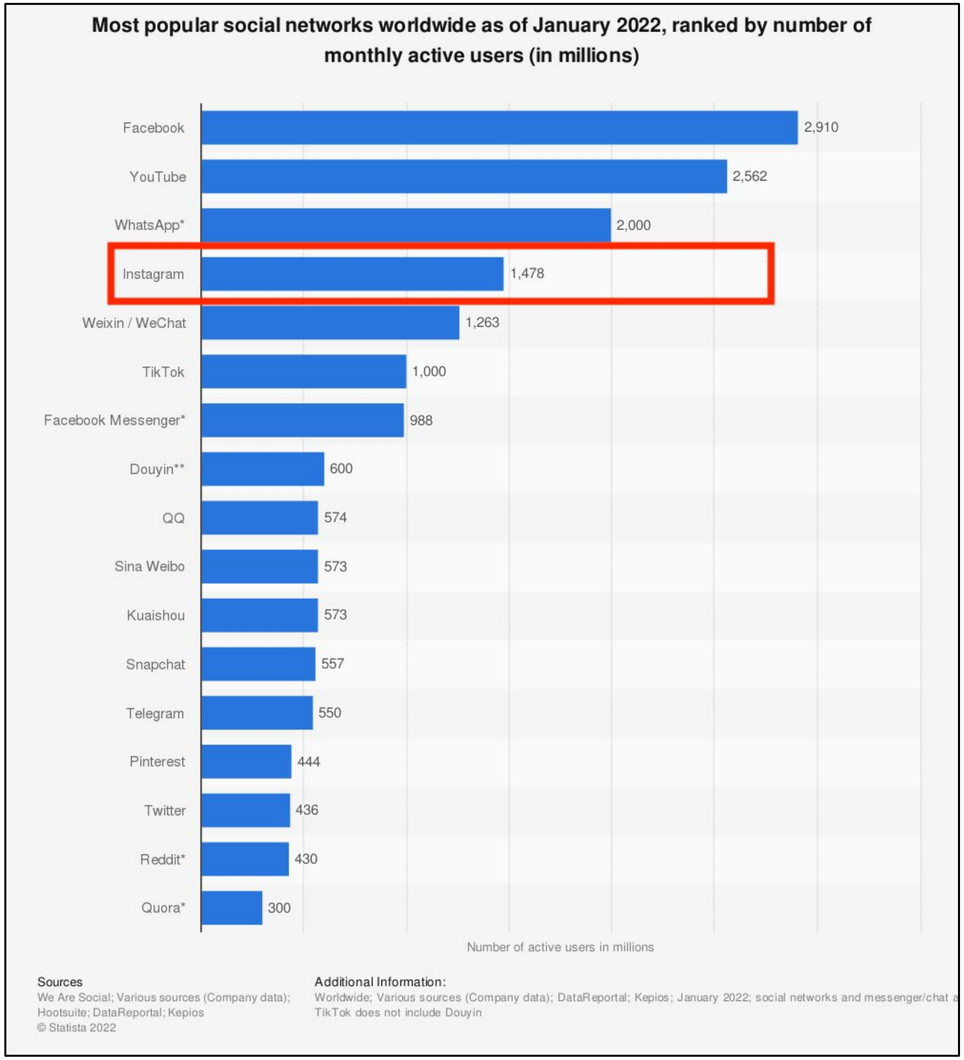


Exposure to incorrect information is widespread in most emerging economies surveyed
% of social media platform and messaging app users who __ see articles or other content when they use social media that seems obviously false or untrue
Frequently
Occasionally
Total
Tunisia 29% 49% 78%
Lebanon 34 42 76
Vietnam 19 53 72
Jordan 32 39 71
South Africa 36 33 69
Philippines 27 41 68
Venezuela 27 30 57
Kenya 31 24 55
Colombia 24 31 55
India 16 39 55
Mexico 25 19 44
Note: Social media and messaging app users include those who said they use one or more of the seven specific online platforms measured in this survey.
Source: Mobile Technology and Its Social Impact Survey 2018.
PEW RESEARCH CENTER

Background

<u>Why Instagram?</u>

- High user count
- Active user base
- Image platform
- More susceptible to the spread of misinformation



Problem Statement

Step 1: Using adversarial examples to fool State-of-the-art Optical Character Recognition Models
Step 2: Masking false-information images with adversarial masks to bypass instagram's fact-checking algorithm

Why is it important ?

1. Recent surge in misinformation regarding vaccines and the pandemic.
2. Major chunk of the teenage and adult audience uses Instagram and social media platforms, and it's still growing.
3. Our solution brings light to gaps in the current state of art.

Current State of art

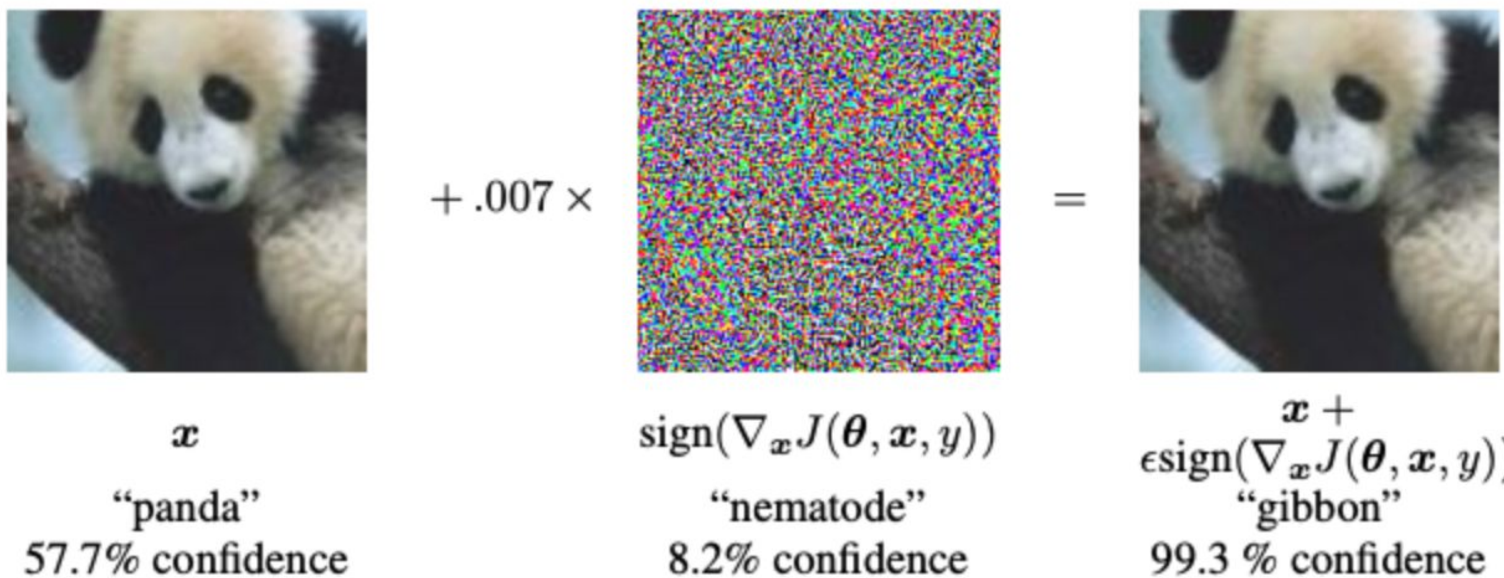
- Current solutions/workarounds focus on detection by evasion ex: V@ccination/C0V1D-19

- No known work on adversarial example detection for Instagram

- OCR Models: Tesseract, Instagram Fact-checking Algorithm

Adversarial Examples

- Specialized inputs created with the purpose of confusing neural network and misclassifying output
- Perturbation indistinguishable to human eye, not to digital eye



Attack methods and Categories

- White Box Attacks - Attacker has access to the model
- Black Box Attacks - Model is unknown
  - FAWA - Watermark perturbations
  - HopSkipJump - Our Method

HopSkipJump

- An extension of decision based attack
- Both targeted and untargeted implementations
- Requires significantly fewer model queries
- Competitive performance against defense mechanisms
- Each iteration has 3 components :
    - Iterate is pushed towards the boundary
    - Gradient direction is estimated
    - Step size is updated along the gradient direction

HopSkipJump

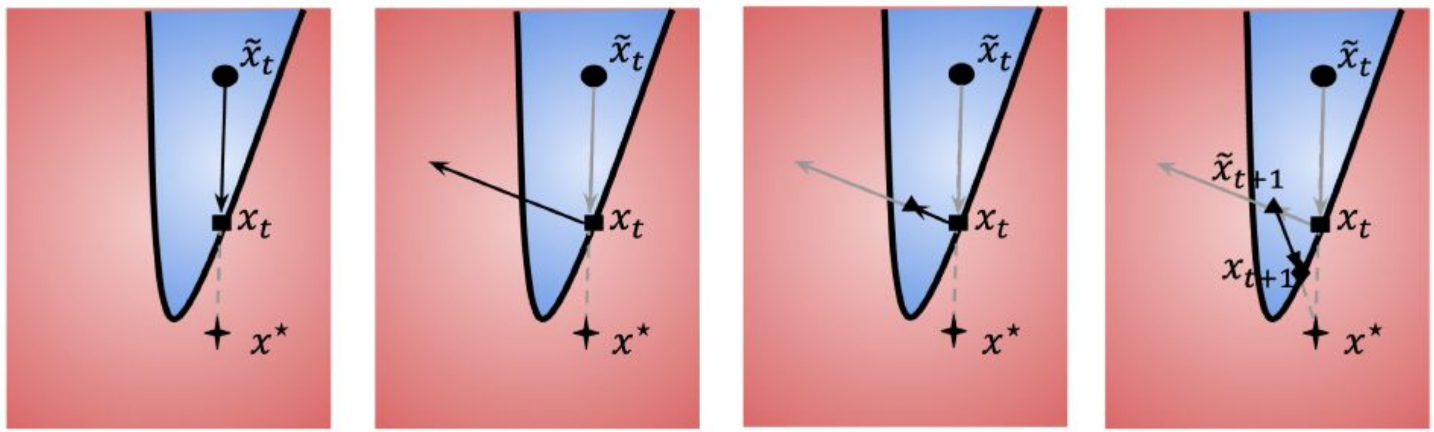
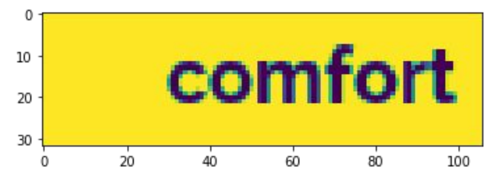


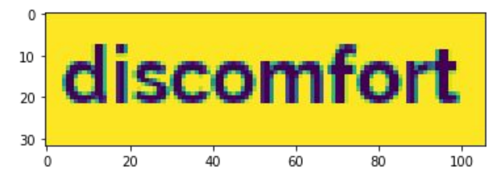
Figure 2: Intuitive explanation of HopSkipJumpAttack. (a) Perform a binary search to find the boundary, and then update $\tilde{x}_t \to x_t$. (b) Estimate the gradient at the boundary point x_t. (c) Geometric progression and then update $x_t \to \tilde{x}_{t+1}$. (d) Perform a binary search, and then update $\tilde{x}_{t+1} \to x_{t+1}$.

Ref: Chen, Jianbo, Michael I. Jordan, and Martin J. Wainwright. "Hopskipjumpattack: A query-efficient decision-based attack." *2020 ieee symposium on security and privacy (sp)*. IEEE, 2020.

Approach & Implementation

- Targeted Attack
  - The goal is to alter the output of the model to a pre-specified text. This is done by giving a *target-image* along with *input-image*
  - Experiment 1: Generated antonym Image pairs. Fooled the model to predict the antonym



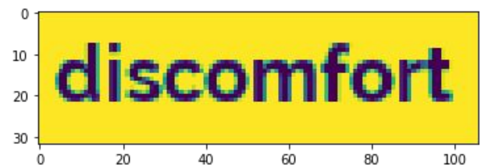


  - Experiment 2: Considered “Vaccination” images. Manually modified the *input-image* to create *target-image*

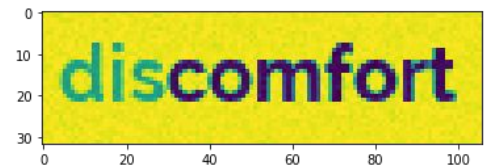
- Untargeted Attack: No *target-image* is given. The goal is to just alter the model output

Adversarial Outputs

- Targeted Attack (Experiment 1)



[Original Image] [Model output: **discomfort**]



[Perturbed Image] [Model output: **comfort**]

- Targeted Attack (Experiment 2)

[Original Image] [Model output: **VacciNation**]

[Perturbed Image] [Model output: **VacciN ion**]

- Untargeted Attack

[Original Image] [Model output: **VacciNation**]

[Perturbed Image] [Model output:]

Results & Comparison (Tesseract)

- Evaluation Metric

  Success Rate = No. of perturbed images able to fool the model/Total no. of Images considered

- Targeted Attack
  - Experiment 1:
    - Success Rate = 20/20 = 100%
  - Experiment 2:
    - Success Rate = 46/52 = **88.4%**
- Untargeted Attack
  - Success Rate = 48/52 = **92.3%**

Results & Challenges with Instagram

- Our approach (HopSkipJump) requires the OCR model to churn out predictions several no. of times, for it to able to optimize
- Instagram API restricts sending more than 25 queries in a day, severely limiting our automation capabilities
- The images are flagged in about ~30 mins after posting on an average, increasing run-time for our algorithm
- We had to resort to optimize our adversarial images over Tesseract and manually upload statuses on the platform for testing

  (Uploaded 52 statuses)

- Targeted Attack: Success Rate = 10/45 = **22.2%**
- Untargeted Attack: Success Rate = 21/48 = **43.7%**

Demonstration Video

Our takeaways and Future Work

- We are successfully able to beat Tesseract
- Untargeted attack is working better than Targeted attack

As Next steps…

- Test on other State-of-the-art OCR algorithms (Google Vision, etc)
- Test using further advancements over HopSkipJump

Thank you!

THANK YOU!