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

Arindam Das
Madison, USA
adas34@wisc.edu

Harsh Sahu
Madison, USA
hsahu@wisc.edu

Harshita Singh
Madison, USA
hsingh48@wisc.edu

ABSTRACT

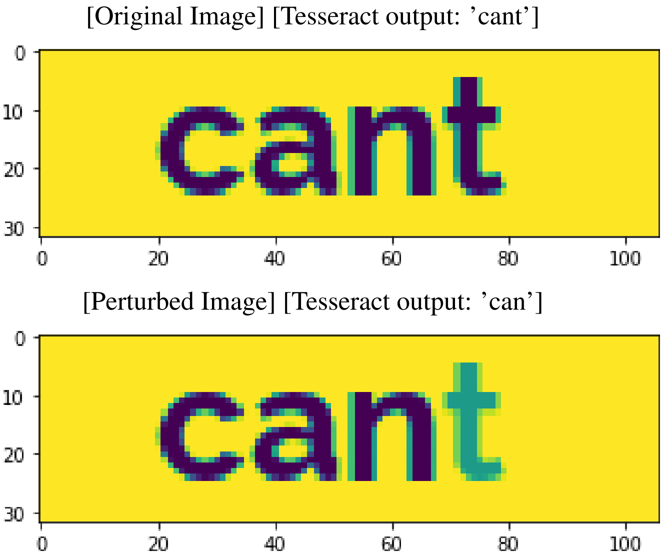
Whether it is the elections, news about a new strain of COVID or vaccine paranoia - over the past few years we’ve seen an ever evolving scope of social media to overpower the conventional news sources all over the world. According to a survey by Pew Research Services, eight out of ten adults in the US admitted to getting their news from social media platforms. No wonder, there has been an overwhelming amount of misinformation circulating on these platforms. With the influx of misinformation, social media companies have also come forward with ways to combat their spread. One such approach is using fact-checkers to verify reported images and flag them as fake/harmful if they are found to contain misinformation. Moreover, other images which are found similar to this image by image matching algorithms are also marked unsafe. Our project aims to demonstrate this approach of preventing spread of misinformation is imperfect and can be bypassed by using adversarial perturbations. Using an adversarial example, we can prevent a reportedly fake image from being hidden by tricking the fact checking algorithm into believing it is an innocuous image. We aim to study this on a popular social media platform, Instagram and demonstrate this as an outcome of our project.

INTRODUCTION

Over the past few years, one of the most frequently discussed and critiqued aspect of all social media platforms has been fake news and misinformation campaigns, which causes us to wonder as to why has this problem become so pronounced in the recent past? One of the possible reasons is the increase in consumption of content through social media platforms. Another interesting point to note highlighted in [5] by Vosoughi et al. is the observation that "falsehoods diffuse significantly farther, faster, deeper, and more broadly than the truth in all categories of information."

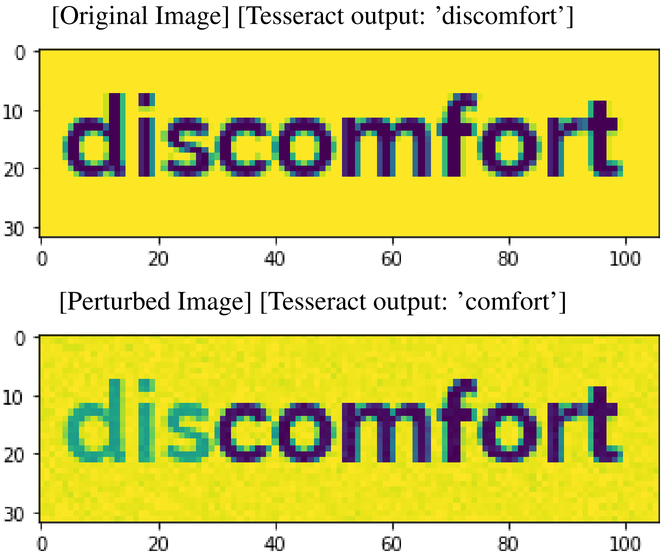
Social media platforms have definitely tried to address this problem at differing levels of priority and with this project we try to address one such platform - Instagram’s fact-checking and misinformation classification algorithm. We aim to

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235


[Original Image] [Tesseract output: ’cant’]

[Perturbed Image] [Tesseract output: ’can’]

Figure 1. Demonstration of Adversarial attack


[Original Image] [Tesseract output: ’discomfort’]

[Perturbed Image] [Tesseract output: ’comfort’]

Figure 2. Demonstration of Adversarial attack

demonstrate that the current measures in place can be tricked by introducing an adversarial perturbation to an image that otherwise would have been flagged as fake. Our goal is to highlight and demonstrate the necessity for enforcing stronger checks in the existing processes.

All the original images as well the perturbed images along with the code can be found at our github repository provided in the footnote[1]

In the following subsections, we briefly discuss the background and the techniques we are using to achieve our project's goals:

Optical Character Recognition

OCR systems transform a two-dimensional image of text, that could contain machine printed or handwritten text from its image representation into machine-readable text. Formally, Optical Character Recognition(OCR) enables us to translate documents and images into analyzable, editable and searchable data [3]. Present day platforms, like Instagram rely heavily on OCR to classify and flag images and even limit the spread of certain images that have flagged content on them. One such example is the flagging of images that contain the word "Covid" or "vaccines" on them. In an attempt to limit the spread of misinformation - all such images are currently tagged with an additional banner that redirects the users reliable information sources.

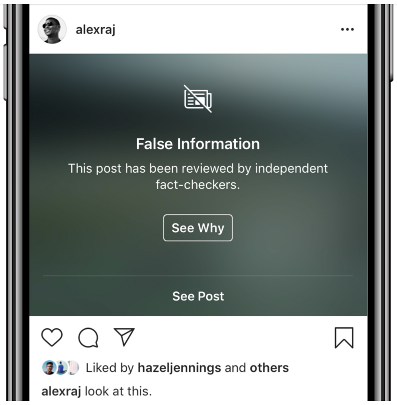
Tesseract

In the current phase of our implementation - we are using Tesseract [4], a deep learning based OCR engine developed by Google, to extract text from images. Tesseract operates in a step-by-step pipeline wherein the first step is a connected component analysis that results in the identification of "blobs". These blobs are then organized into lines and the lines are broken into words based on character spacing. This is followed by a two-step recognition process - in the first pass, the algorithm attempts to recognize each word and satisfactorily recognized words are sent to an adaptive classifier as training data. The second-pass refines the classification based on the knowledge gained from the first pass.

Successfully evading Tesseract detection is one of the milestones for our project. Since Instagram is essentially a blackbox and we aren't aware of the OCR being used, we will be implementing Tesseract as a baseline to evaluate our adversarial examples. Once we achieve satisfactory accuracy with Tesseract - we will expand our attack to Instagram.

HopSkipJump

Adversarial examples have evolved over the years with masked perturbations to images becoming barely detectable to human eyes. There are several models that have been developed to generate adversarial examples that can evade OCR detections. Some interesting works include Fast Adversarial Watermark Attack [2] - FAWA works by disguising perturbations as watermarks so as to evade human eye detection. While this is a really interesting approach, the algorithm is implemented in a white-box environment which is not feasible to our problem statement.



alexraj

False Information

This post has been reviewed by independent fact-checkers.

Liked by hazeljennings and others

alexraj look at this.

Figure 3. Current content screening on the platform.

We therefore plan to extend HopSkipJump [1] for our project. The HopSkipJump attack is based on an estimate of the gradient direction using binary information at the decision boundary. Experimentally, HopSkipJump has demonstrated requiring fewer queries than other models and can be implemented as both targeted and untargeted attack. HopSkipJump is also effective against several defense mechanisms such as defensive distillation, region based classification and adversarial training. We have implemented the HopSkipJump attack to mis-classify certain words in images as their antonyms using targeted attack.

CURRENT STATE-OF-THE-ART

What we know about Instagram's fact-checking

On December 16, 2019 - Instagram released a blog discussing about its efforts to combat misinformation on the platform. The blog listed the working of this new system as follows :

1. Any images reported as fake/misinformation are sent to third-party fact checkers to verification.

2. Once this image is fact checked, and found to be false or partly-false - the platform "reduces" its distribution by reducing its visibility. In addition, we also see a banner that covers the image as shown in Figure 1.

The interesting bit about this is, Instagram further uses "Image Matching Algorithm" to identify similar images and flag them as false.

Our Approach and existing work

With this project, we aim to perturb OCR images such that they could bypass the platform's image matching algorithm while still maintaining the legibility of the image. Keeping the content clear in the presence of a perturbation is one of the major challenges that we aim to overcome since in this use-case any perturbation that distorts the image visibly will render it useless to the adversary.

We have not found any focused work in the field addressing social media and adversarial learning so far, however- we are in the process of reviewing and searching for more literature

[1] https://github.com/SahuH/Spring2022-CS766-project-MASKquerade

in the field. There have been important works in the field of adversarial examples such as [6] that we plan to refer to while working on the project.

EVALUATION METRICS

As an outcome of our work, we expect to design an adversarial mask that can fool Instagram's image flagging mechanism and avoid detection. We hope to demonstrate using a real-life demonstration of our observations on the platform.

We will define success rate of our method as the percentage of images that are successfully able to fool Instagram after being perturbed.

PROGRESS SO-FAR AND UPDATED OBJECTIVES

We are on-track with our project goals and here's an outline of the progress we've made so far:

- Our first challenge was to determine whether we plan to proceed with a targeted or an untargeted attack. While both the attacks serve our goal of evading the algorithm's detection - we've decided to proceed with a targeted attack for our initial experiments. Our assumption is that targeted attack will make it easier for Hopskipjump to perturb the images in a particular direction.
- The next step was identifying an appropriate Adversarial algorithm to use for our process. There are several different directions we can proceed with this approach but for the given black-box requirements and the memory and query restrictions on a third-party platform - HopSkipJump satisfied most of our requirements. We have so far succesfully implemented a HopSkipJump Attack and verified in on a Tesseract engine.
- We have taken 20 images, each containing a word, and used another set of images containing their respective antonyms for the targeted attack. Using HopskipJump's pretrained model, we were successfully able to generate perturbed images where Tesseract was fooled in predicting their antonyms.
- Over the course of the next month, our goal is to fine-tune the adversarial process for our task so as to decrease the l2-error between original and perturbed image and create more legible perturbed images.
- Another important goal is to identify and consolidate a dataset of images we plan to demonstrate our attack on. We have generated some preliminary images using an ad-hoc script that we plan to refine and expand over the next week.
- Lastly, we will be demonstrating our attack through either a test Instagram account or one of our accounts depending on Instagram's flagging mechanism. In our research so far, Instagram doesn't really flag posts from newly created or sparsely followed accounts. We will look into this a little further and decide a direction.

TIMELINE

- Step 1 : Understanding the Instagram's fact-checking algorithm and its execution under ideal conditions.- 5th March - COMPLETED
- Step 2 : Creating an adversarial process - 30th March - COMPLETED
- Step 3: Refining the patch to keep the content legible to the human eye - 15th April - ON TRACK
- Step 3 : Testing the adversarial perturbation on Instagram - 10th April - TO DO
- Step 4: Testing the outcome on Instagram, tweaking and final changes. - 20th April - TO DO
- Step 5: Final presentation preparation, website creation - 28th April TO - DO

REFERENCES

[1] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. 2019. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. (2019). DOI: http://dx.doi.org/10.48550/ARXIV.1904.02144

[2] Lu Chen, Jiao Sun, and Wei Xu. 2020. FAWA: fast adversarial watermark attack on optical character recognition (OCR) systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 547–563.

[3] Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access* 8 (2020), 142642–142668. DOI: http://dx.doi.org/10.1109/ACCESS.2020.3012542

[4] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Vol. 2. 629–633. DOI: http://dx.doi.org/10.1109/ICDAR.2007.4376991

[5] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[6] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker. 2019. Adversarial examples in modern machine learning: A review. *arXiv preprint arXiv:1911.05268* (2019).