MASKuerade : Implementing adversarial mask on misinformative images to bypass Instagram's fact-checking

Arindam Das
adas34@wisc.edu

Harsh Sahu
hsahu@wisc.edu

Harshita Singh
hsingh48@wisc.edu

February 25, 2022

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.

1 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 [VRA18] 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 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.

2 Current State-Of-The-Art

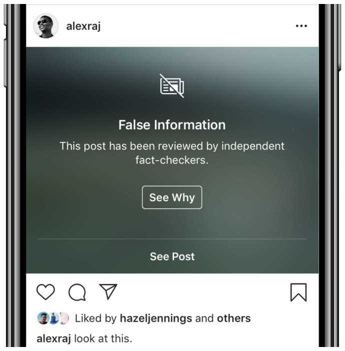
2.1 What we know about Instagram's fact-checking

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.



alexraj

False Information

This post has been reviewed by independent fact-checkers.

Liked by **hazeljennings** and **others**

alexraj look at this.

Figure 1: Current content screening on the platform.

2.2 Our Approach and existing work

With this project, we aim to develop an adversarial mask that 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 in the field. There have been important works in the field of adversarial examples such as [WXDdB19] that we plan to refer to while working on the project.

3 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.

4 Timeline

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

References

[VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[WXDdB19] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker. Adversarial examples in modern machine learning: A review. *arXiv preprint arXiv:1911.05268*, 2019.