Resolving Bias in Vision Privacy Protection Techniques

Jasmine DeHart
dehart.jasmine@ou.edu

Lisa Egede
lisaegede@ou.edu

Christan Grant
cgrant@ou.edu

School of Computer Science, University of Oklahoma, Norman, OK, USA

Abstract

With the growth and accessibility of mobile devices and the internet, the ease of posting and sharing content on social media networks (SMNs) has increased exponentially. Many users post images that contain privacy leaks regarding themselves or someone else.

To mitigate the prevalence of privacy leaks on social media, we propose a computer vision system to identify content and mitigation techniques to reduce exposure. Being mindful that the data collected accurately reflects the population of the users and evaluating the ethical aspects of sensitive content are imperative in this development process.

1. Introduction

According to Pew Research Center, 79 percent of Americans online use Facebook, 32 percent of Americans online use Instagram, and 24 percent of Americans online use Twitter [3]. Any content posted to social media networks (SMNs) can be lost to someone else even after removal of the content. Stolen visual content can then be used as a transport vector for other types of cyber-attacks or social engineering [4, 7].

In this paper we will analyze various ethical and privacy issues in relation to privacy leaks on social media, the data collection process, and the sensitivity of the images collected. In particular, we aim to analyze biases that arise in collecting data, pre-processing data and structuring the algorithm [6] for our object detection model, and privacy concerns surrounding the sensitive information that is collected (e.g. passport, licenses, credit card information).

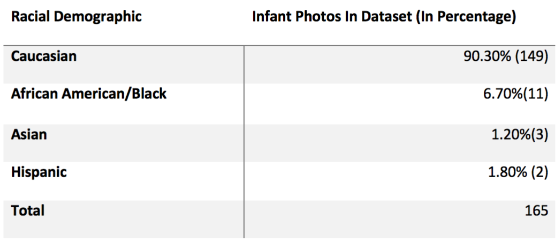
Following our use of computer vision techniques to identify "private" information, we propose and incorporate a privacy scoring metric to gauge a users privacy leaks, which computes an individuals' probable exposure regarding their visual content leaks. Private visual content exposes intimate information that can be detrimental to our finances, personal life, and reputation. Private visual content can include baby faces, credit cards, phone numbers, social security cards, house keys and others.

2. Biases in Processed Data

When collecting data to begin the computer vision process, a number of issues surrounding the idea of model transparency and ethics were taken into consideration. The sections below describe these topics in greater detail.

2.1. Data Collection and Training

In studies related to image object detection and privacy, the importance of ensuring all populations of every respective group are measured fairly and accurately is imperative to reduce bias in machine learning and computer vision models [8]. The exposure of population groups in machine learning processes is essential for organizations to make advancements in services and applications. For example, on social media networks computer vision is used to identify individuals who maybe be tagged in other photos.



Racial Demographic	Infant Photos In Dataset (In Percentage)
Caucasian	90.30% (149)
African American/Black	6.70%(11)
Asian	1.20%(3)
Hispanic	1.80% (2)
Total	165

Figure 1. The racial demographic collection of original model.

While training, we noticed that the population distribution in detecting children/babies was skewed. The model accurately identified "babies", however due to the lack of racial backgrounds the model showed biases. Based on our model at the time, and the graph in Figure 1 depicting that data, African American infants only accounted for 11 out of 165 total photos. Photos for Hispanic infants were at 3, Asian infants were at 2, and Caucasian infants accounted for 149 photos for the data set. In order to ensure that the accuracy rate was improved, the visual content collected for the category, “babies”, contained a variety of infants ranging in skin tone. Taking all of these factors into account, our original model was revamped to improve the precision and recall of the object detection model. Simply searching the term ’babies’ in Google does not give an accurate representation of all groups and populations. We wanted to incorporate as much data as possible without over-training the model to alleviate this issue.

In a real-world deployment of this system, the scope of this application will need to meet every users needs regardless of race, age, disability, or gender. The unintentional biases from web crawling data sets raises a valid concern. Intentionally working to eliminate the potential for classification inaccuracy is an important process to create a model that will distinguish privacy leaks from other objects.

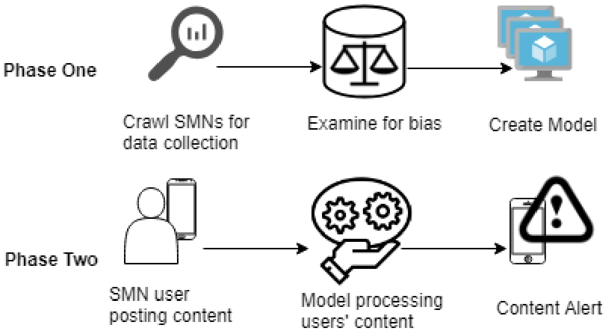


Figure 2. Flow of System Implementation

3. Resolving biases for Implementation

When developing a system that employs the computer vision algorithm, we investigated methods of transparency and ethics that could influence the user’s concept of privacy in visual content. The sections below describe these topics in greater detail.

3.1. Designing Strategies for Model transparency

Given the various categories and examples in which visual content can be exploited, we are building a visual inspecting system that will further help us understand those risks and mitigate them. Our system collects data from SMNs (via crawling on social media), and deploys the machine learning object detection techniques to identify potential privacy leaks in the content.

This system employs various methods of communication with the application users. To give the users customizability and different levels of interaction, it is important to create several modes of interaction and communication. To create this system, we are are implementing the following eight mitigation techniques [1]:

- Technique 1 - Client app (Figure 3a). A user can download a third-party application on various electronics (e.g. cellular phone, laptop) to alert the user if posting potential leaks. This third-party application will prescreen visual content (images, videos) before it can be posted on SMNs giving the users a warning if a privacy leak exists in that content.

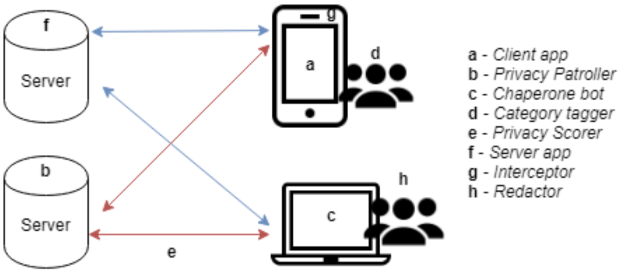


Figure 3. The location of each proposed mitigation technique.

- Technique 2 - Privacy Patroller (Figure 3b). This is a SMN crawler that will randomly look at a user’s pages, screening for privacy leaks and alerting the user of potential leaks in their content. Content will automatically be removed after 24 hours.

- Technique 3 - Chaperone bot (Figure 3c). A user can add a chaperone bot as a friend on SMNs. The chaperone bot will give the user friendly suggestionsbased on type and frequency of privacy leaks on SMNs.

- Technique 4 - Category Tagger (Figure 3d). A user can select the category that the visual content belongs to before being uploaded to SMNs. Once tagged, an automated system will check for content compliance with tag. If it does not fit the category, the user is notified of new category tag options based on the models predictions.

- Technique 5 - Privacy Scorer (Figure 3e). The user will be monitored based a privacy score. The bot will monitor the user’s content after posting. In this case, a person who has a higher privacy score will be monitored more closely than some one who has a lower score. The privacy scorer will updated the user on changes in their score and alerts about privacy leaks.

- Technique 6 - Server app (Figure 3f). The SMN will screen visual content before uploading to platform. We suggest collaboration with SMNs to provide enforcement of user compliance and techniques.

- Technique 7 - Interceptor (Figure 3g). With the SMN applications, users will agree to let the SMN intercept the camera and gallery to flag and block content that should not be selected for posting. If the user wants to query why the content was blocked, the system will give a description of the identified leak in the content.

- Technique 8 - Redactor (Figure 3h). Using the SMN application, the users will be able to use redaction techniques (*See Privacy protection*) to add secure distortions to visual content.

With our techniques, we can understand their usefulness by factoring in the frequency of engagement from users, conducting user studies of the techniques before deployment, and observing the privacy score behaviors among the users. To further help users understand and use our application, we will create a tutorial on how the system works and give the users control to override the system's decision. With these techniques, it will allow user's of all ages to engage and reflect on their visual content practices on SMNs. To make this application transparent for everyone, we hope to include features to support user's with disabilities.

3.2. Privacy protection

The privacy of user's is the number one concern for SMNs. These platforms have to create ways to protect the users from others and potentially themselves. On SMNs, content can be posted without any pre-screening or post-screening procedures, with this proposed system we are implementing several screening mitigation techniques that will secure users' information. This endeavour will help us monitor visual content from SMNs for consumer privacy and protection. This research will subsequently protect everyday users from invasions of privacy, whether the action is accidental or intentionally made. In an effort to protect users, we propose a redaction spectrum that will allow them to do so. Users may want to share images but hide parts of its' content. Web crawling systems may be collecting baby images and credit card information among other things for nefarious reasons. We propose a spectrum of techniques to obfuscate visual content from other users or machines (Figure 4).

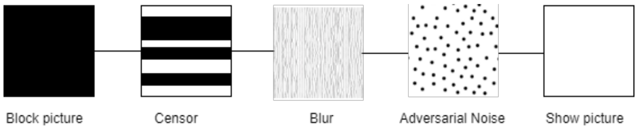


Figure 4. Spectrum of visual content redaction techniques

Once a visual content privacy leak is detected, we can handle it in various ways. The first option is to block objects in the picture. This will remove the content and/or the user's affiliation with the content from SMNs. The second option is use a censor. Censoring is essentially removing a person or object from the visual content and it will insert a blank space where the object once existed. The third option is to blur content. Blurring the content will allow the user to have some control over what is being seen without causing too much distortion. The surrounding objects will still remain visible and the leaked object will be less visible but not removing them. The fourth option is to use adversarial noise [2]. We believe that adversarial noise will be important feature added to visual content to help protect SMN users from computer attacks. By adding a few pixels, we could (1) impede their ability to learn anything from the visual content even if it is in their possession, and (2) still allow the images to be visible only to humans.

4. How Biases in Computer Vision and Data Inadvertently Affect Applications

When seeking to create applications to track and detect potential privacy leaks in photos, concerns regarding potential biases in the data collection process are to be expected. Seeking to prevent skewed data is imperative in ensuring a positive user experience, especially with the privacy detection tool used in our research. Accuracy and consistency are factors that play into improving this experience, both of which were taken into account at the beginning of the data collection process.

The ”baby” privacy category was of particular concern, being that the skin tone range of babies varies, a fact that can create unintended biases. Disproportionate representation in a data set can result in a higher chance of a particular racial demographic being flagged. While part of the goal of the privacy detector is accuracy, disproportionate predictions leave underrepresented groups at a higher risk of privacy leaks. Along with flagging accuracy, inaccuracy is of great concern. In our model, there was a consistent confusion between that of babies and adults. Along with this, identification's (eg. Permit, license, school/work) were being incorrectly predicted as babies and passports were being incorrectly predicted as licenses. Being that licenses and passports are both government issued identifications (IDs) that can lead to privacy risks, the confusion between the two is not as severe. However, the incorrect flagging of babies vs adults and IDs vs babies is an obvious error that could influence whether or not the user trusts the detection model.

5. Effects on Mitigation Techniques

Throughout the research process, various mitigation techniques are being designed to improve the process for users. Thus, inaccuracy from biases could trickle into these systems and if significant, they can cause major issues in the overall user experience. While following techniques (section 3.1) seek to resolve such biases, they can help to contribute if they are not implemented properly.

- Messages and Warnings - Frequent messages alerting users about potential security concerns regarding their posts can slowly turn into a nuisance, especially when such alerts are inaccurate.

- Interceptor - The *Interceptor's* (Section 3, Technique 7) goal is to intercept the camera and gallery to flag and block content that may pose as a privacy risk. Because errors in flagging accuracy can result in inaccurate interceptor flagging, users may experience confusion and frustration with an application that prevents them from uploading photos that do not pose as a security risk.

- Privacy Patroller - The Privacy Patroller functions is to automatically delete potential privacy leaks 24 hours after they are flagged, especially if a user is away from their account for long periods of time. While the patroller takes an extra step to ensure that all potential leaks are blocked, it can frustrate users whose content are falsely flagged.

- Privacy Scorer -Inaccurate predictions can negatively impact users privacy scores, thus resulting in increased surveillance on their account. Misrepresentation in data can result in higher privacy scores for users whose content would otherwise be considered 'safe'.

- Category Tagger - The goal for human computer interaction is to give users options and improve their experience, but users may inaccurately tag photos. Such an issue could result in the SMN failing to detect the user has selected the wrong category, which could increase the chances of a privacy leak.

6. Trade-offs for Users and Computer Models

Wearable technology is opening up many avenues for opportunity with applications, however this technology will require new techniques to preserve users privacy.This technology captures pieces of the user's life that is not normally shared through photography. Wearable technology can elicit a variety of private information to be shared like credit cards, content on computer screens, house keys, and even jeopardize the privacy of bystanders [5].

7. Discussion: What Do The Users Think?

In this study, we conducted a survey that investigated privacy perspectives of SMNs users. From this survey, we noticed several trends from the responses. Over 87% of participants agreed that visual content containing credit cards, driver's license, house keys, phone numbers, social security cards, passports and birth certificates are privacy leaks. Only 29% of participants agreed that content of babies and minors is a privacy leak. The participants ranked the dangers of privacy leaks in this order (highest threat- lowest threat): burglary, stalking, financial threat, identity theft, and lastly, explicit websites. Among these dangers, the categories with the highest threat are: identity and asset. From this survey, we began to collect more data focusing on the users concerns.

8. Conclusion

As SMNs continue to grow in popularity, they become a powerhouse for privacy leakage due to the change in social culture, development of features, and audience. This research will impact everyday users and non-users of SMNs by providing a mechanism to identify sensitive information found in visual content posted on SMN. With the improvements in understanding privacy leaks on SMNs, we can lower the amount of malicious, financial and personal, attacks made on these platforms. This research will impact everyday users and non–users of SMNs by providing a mechanism to identify sensitive information found in visual content posted on SMNs. To accomplish creating this system effectively, it is important to remain user-centered and consider the models' transparency, fairness, and accountability.

References

[1] Jasmine DeHart and Christan Grant. Visual content privacy leaks on social media networks. *arXiv preprint arXiv:1806.08471*, 2018.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2014.

[3] Shannon Greenwood, Andrew Perrin, and M Duggan. Social media update 2016. *Pew Research Center*, 11, 2016.

[4] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80. ACM, 2005.

[5] Roberto Hoyle, Robert Templeman, Denise Anthony, David Crandall, and Apu Kapadia. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pages 1645–1648. ACM, 2015.

[6] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[7] Heidi Wilcox and Maumita Bhattacharya. A framework to mitigate social engineering through social media within the enterprise. In *Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on*, pages 1039–1044. IEEE, 2016.

[8] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. *HICSS*, 2018.