November 2021

Automatic Image Redaction

Andrew Cheng

Automatic Image Redaction

ABSTRACT

Large numbers of images are uploaded and shared over the internet, e.g., via email, social networks, messaging applications, image sharing applications or websites, etc. Such images can sometimes include potentially sensitive information. This disclosure describes techniques to automatically identify and redact potentially sensitive information from images or other media. A tool that implements content detection and redaction can be integrated at various points in applications that are used to share images. The redaction is under user control and can be performed at the point of capture, at the point of upload, etc. Automatic detection and user-controlled redaction of images can prevent leakage of potentially sensitive information.

KEYWORDS

- Image redaction
- Document redaction
- Text recognition
- Sensitive information
- Personally identifiable information (PII)

BACKGROUND

Large numbers of images are uploaded and shared over the internet, e.g., via email, social networks, messaging applications, image sharing applications or websites, etc. Images uploaded by a user can sometimes include potentially sensitive information such as the user's name, date of birth, address, or other personally identifiable information (PII). For example, after receiving vaccines, some users may share selfies holding their vaccine cards which display their full name and birthday. As another example, illustrated in Fig. 1, proud parents may share photos of their

young children holding signs celebrating their first day of school. Such signs can include information such as the child's name, their school's name, their teacher's name, grade, age, height, weight, favorite color, favorite food, etc. One or more such pieces of information may be inappropriate to share or upload due to potential for abuse. For example, such information can serve as answers to security questions needed to access a financial or web account, and the exposure of such information to malicious actors can enable exploits or predation.

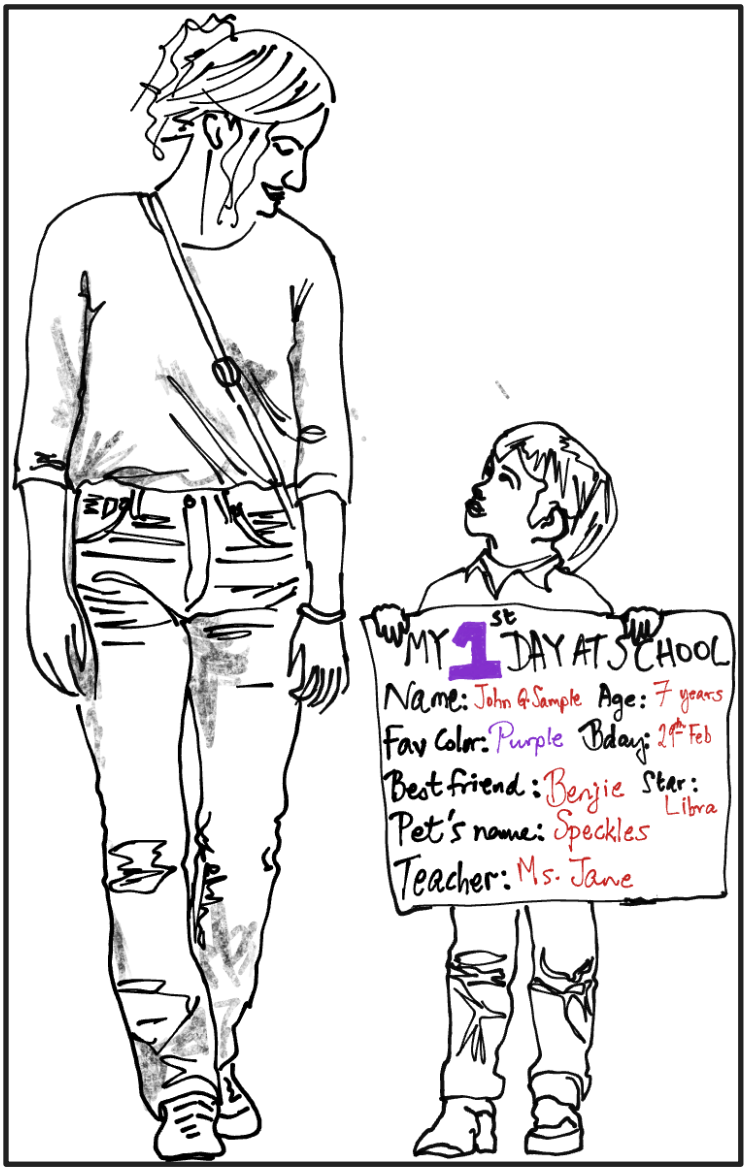

Fig. 1: Image that includes user information - first day at school

While some users simply don't think of the ramifications of sharing such an image when uploading or sharing images, some may be aware, but nevertheless do not modify the image to remove such information due to the inconvenience of doing so. For example, scrubbing information from a photo may require the user to (a) save the photo; (b) open it in an editing app to edit it; (c) save the modified photo; and (d) access it via the image upload sharing or app. The number of steps and the complexity of performing each action can be a barrier to removing information from images.

While some software applications can recognize text in an image, such applications do not determine whether the text is potentially sensitive information. For example, the word 'Joy' in an image can be a person's name (suitable for removal) and also a state of mind (not user information suitable for removal). Another example is that of a person holding a sign that states 'My trip to a strawberry field' (not PII) versus a sign that states 'my favorite food is strawberry' (potentially sensitive information). Software tools that can automatically blur content such as human faces, vehicle license plates, etc. exist but do not specifically recognize sensitive information in an image.

DESCRIPTION

This disclosure describes techniques to automatically identify and redact (e.g., blur) certain types of information, e.g., sensitive information, personally identifiable information (PII), or other type of information from images or other visual media. The tool can be integrated at various points in image capture, social media, media-sharing, or messaging applications, e.g., at the point of capture of an image, at the point of upload of an image, etc. The techniques help mitigate against leakage of information that is present in photos.

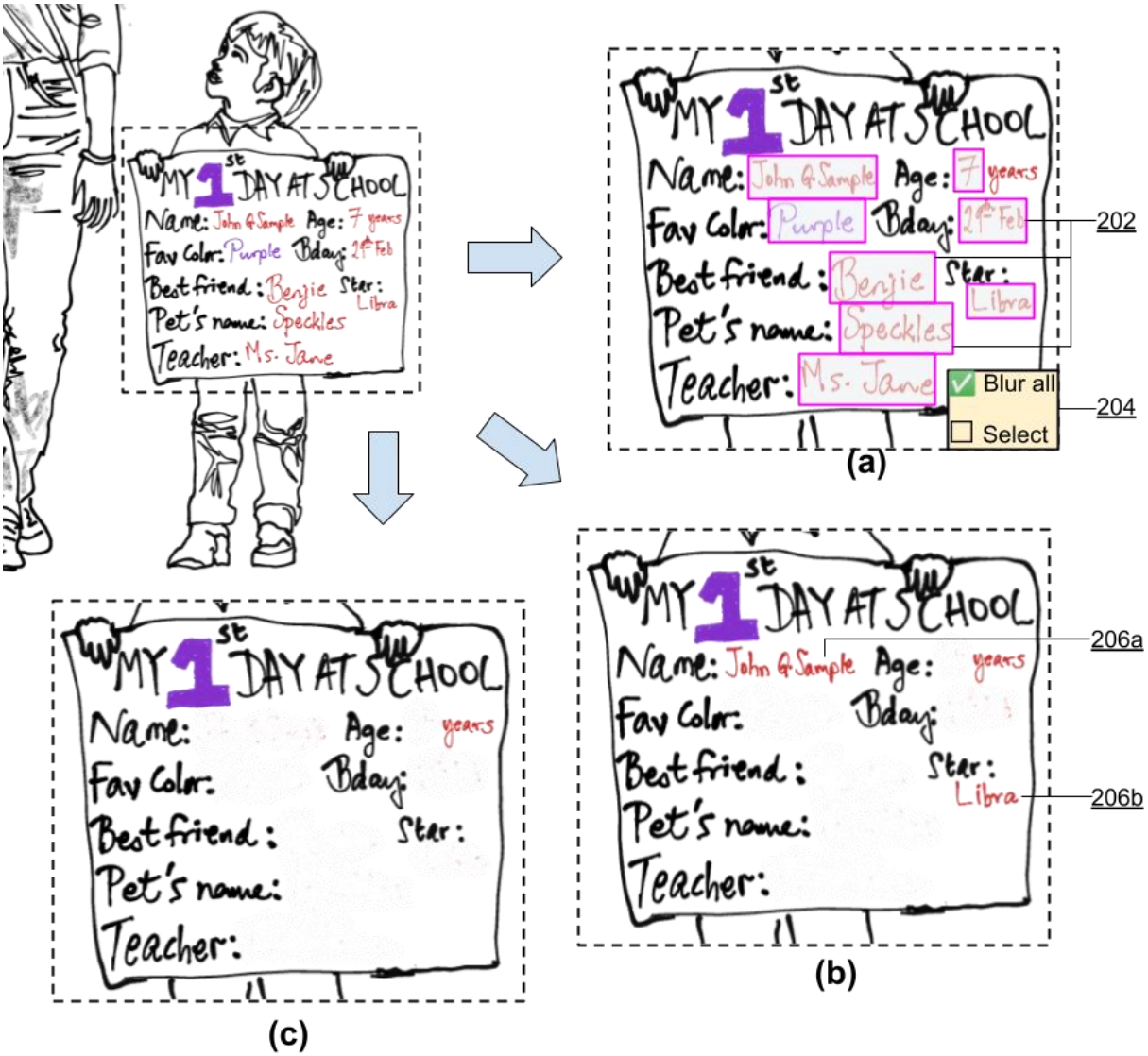

Fig. 2: Automatic information redaction: (a) Content is automatically selected and offered to the user for redaction; (b) With user permission, sensitive content is automatically redacted; (c) All potentially sensitive information is automatically redacted.

Fig. 2 illustrates techniques of automatic redaction of information from images. In Fig. 2(a), potentially sensitive (202) is automatically detected, selected, and highlighted. A selection mechanism (204) enables the user to select (e.g., with a finger or stylus) the highlighted information for redaction. In Fig. 2(b), sensitive information is automatically detected, and with user permission, information in predetermined categories such as birthday, pet's name, etc. is automatically redacted. Other sensitive information such as name, star sign, etc. (206a-b) is left

unredacted. The user is provided with options to redact such content as well, similar to the depiction of Fig. 2(a).

In Fig. 2(c), potentially sensitive information is automatically detected and redacted, if the tool is configured such by the user. The user setting can specify a level of automatic redaction, e.g., auto-highlight and manual-select (Fig. 2a); auto-redact only certain categories of information (Fig. 2b); auto-redact all potentially sensitive information (Fig. 2c); etc. Other possibilities include the complete redaction of text, whether it is sensitive information or not (in this case, even the text 'MY 1ST DAY AT SCHOOL' is redacted, along with all other text on the chart.

Different options are suitable for users with different preferences - the manual option (Fig. 2a) may be selected by users that desire full control over image edits resulting from redaction; for other users, a pop-up warning can be issued if potentially sensitive information is detected in images in the context of an upload or share event related to the information. The pop-up warning can enable the user to select individual portions of the image to be redacted. The options of automatic redaction of some or all text (Fig. 2b-c) are convenient, fast, and provide user control over the redaction process. If auto-redaction results in an unsatisfactory image, the user can use the manual option (Fig. 2a) to edit the image.

Information in images can be detected using machine learning techniques. For example, a machine learning model can be trained with images that include text labeled as positive and negative examples of potentially sensitive information. Different options can be provided for content redaction, e.g., blurring, pixelating, replacing with a solid color (e.g., black, red, white), etc. For ease of use and low friction, the described redaction techniques can be integrated at one or more points in the image creation and sharing process, for example, at the point of capture (e.g., via a camera app); at a point of image recognition (e.g., via an image description app); at

the point of upload (e.g., via a photo-sharing, file-sharing, social-media app, etc.); at the point of transmittal (e.g., via a messaging app); etc.

Further to the descriptions above, a user is provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein may enable the collection of user information (e.g., information about a user's images, a user's name or other information, or a user's preferences), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user. Thus, the user has control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to automatically identify and redact potentially sensitive information from images or other media. A tool that implements content detection and redaction can be integrated at various points in applications that are used to share images, e.g., social media, media-sharing, or messaging apps. The redaction is under user control and can be performed at the point of capture, at the point of upload, etc. Automatic detection and user-controlled redaction of images can prevent leakage of potentially sensitive information.

REFERENCES

[1] “Redacting sensitive data from images” available online at https://cloud.google.com/dlp/docs/redacting-sensitive-data-images

[2] Xue, Hanyu, Bo Liu, Ming Din, Li Song, and Tianqing Zhu. “Hiding private information in images from AI.” In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1-6. IEEE, 2020.