Identifying Private Content for Online Image Sharing

Ashwini Tonge
Department of Computer Science
Kansas State University
atonge@ksu.edu

Abstract I present the outline of my dissertation work, Identifying Private Content for Online Image Sharing. In my dissertation, I explore learning models to predict appropriate binary privacy settings (i.e., private, public) for images, before they are shared online. Specifically, I investigate textual features (user-annotated tags and automatically derived tags), and visual semantic features that are transferred from various layers of Convolutional Neural Network (CNN). Experimental results show that the learning models based on the proposed features outperform strong baseline models for this task on the Flickr dataset of thousands of images.

Introduction

Online image sharing through social networking sites such as Facebook, Flickr, Foursquare, and Instagram is on the rise, and so is the sharing of sensitive images, which can lead to potential threats to users' privacy. Many users share private images about themselves, their family and friends, but they rarely change the default privacy settings, which could jeopardize their privacy (Zerr et al. 2012). For example, it is common to take photos at cocktail parties and upload them on social networking sites without much hesitation. These pictures can potentially reveal users' personal and social habits and may be used in the detriment of the photos' owner. Even though current social networking sites allow users to set their privacy preferences, it is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings.

Given the enormous and growing amounts of images shared online, the development of automatic approaches that can accurately predict binary privacy settings (i.e., private, public) for these images are required to avoid a possible loss of the user's privacy. An online user's privacy is recognized as a concern for social networking sites by researchers as well. For example, the Director of AI Research at Facebook, Yann LeCun [1] urged the development of a digital assistant to warn people before uploading sensitive content (or embarrassing photos), helping them avoid regrets later.

Several studies explored binary prediction models of image privacy based on user tags and image content features such as SIFT (Scale Invariant Feature Transform) and RGB (Red Green Blue) (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2014). These studies found that users tags are informative and perform better than image content features such as SIFT. More recently, due to the success of object recognition from images using CNN (Krizhevsky, Sutskever, and Hinton 2012), researchers started to investigate learning models of image privacy based on CNN (Tran et al. 2016). Tran et al. proposed privacy framework that combines features obtained from the last fully-connected layers (of 24 neurons) of two CNNs: one that extracts convolutional features, and another that derives object features corresponding to the features extracted from 48 neurons.

However, unlike prior works, I show that the features controlling the distinct attributes of the objects obtained through the higher number of neurons and using the strengths of very deep CNNs, can improve the privacy prediction performance. Intuitively, the objects present in images significantly impact images' privacy. I also show that uncovering the scene context from the image content, in addition to object features, further improves the performance. Moreover, I evaluate the models on a large set of images sampled from the PicAlert dataset (Zerr et al. 2012) and show that the models trained on the proposed features can infer accurate privacy settings for a diverse set of image subjects.

Contributions

In my dissertation, I formulate the problem of identifying private content for online image sharing as follows:

Problem Formulation: Given an image to be uploaded online, classify it into one of two classes: *private* or *public*, i.e., consisting of private or public content, respectively.

This research is motivated by the fact that, increasingly, online users' privacy is routinely compromised by using social and content sharing applications. Identifying sensitive content is inherently difficult because it requires the tool to have an in-depth understanding of the visual content of the image. Moreover, the problem is very subjective, and users are generally reluctant to give full access to their private images (but only access to the images' tags) for the image content analysis, which can hinder the personalized privacy prediction using visual features. Hence, I aim to carefully identify features derived from the multi-modal information of the image that can adequately understand the image con-

[1] https://www.wired.com/2014/12/fb/all/1

tent and predict the prevalent privacy and sharing needs of users' uploaded images. The models trained on these features can enable users to better manage their participation in online image sharing systems by making it easier for regular users to control the amount of personal information shared through images, and thus reduce the escalating privacy risks. Moreover, the proposed tags can also provide the relevant cues for privacy-aware image retrieval (Zerr et al. 2012) and can become an essential tool for surfacing hidden content of the deep Web without exposing sensitive details. In my research, I propose to derive image tags, and visual content features by leveraging CNN architectures, which are used in conjunction with machine learning classifiers to identify sensitive content accurately.

Feature Extraction

The features used in the classification are described below.

Deep features: I used the AlexNet CNN architecture (Krizhevsky, Sutskever, and Hinton 2012) to extract deep visual features and deep image tags for all images in the PicAlert dataset (Zerr et al. 2012) that are labeled as *private* or *public*. AlexNet implements an eight-layer network that is pre-trained on a subset of the ImageNet dataset (Russakovsky et al. 2015). The first five layers of AlexNet interleave convolution and pooling, whereas the remaining three layers are fully-connected (FC). The convolution layers represent high-level features, whereas the FC layers give the non-linear combination of the features in the layers below.

I extract deep visual features from the last three FC layers, and the "prob" layer that produces a probability distribution over 1000 object categories for the input image. Since, not all images on social networking sites have tags or the set of tags is very sparse, I automatically derive tags (deep tags) for images based on their visual content. For deep tags, the top K object categories are predicted from the probability distribution extracted from the CNN (Tonge and Caragea 2016).

Very Deep CNN features: The choice of the CNN architecture used in previous works and features is limited to AlexNet. Given the strengths of deeper CNN architectures for object recognition, features transferred from the deep layers of the very deep CNNs provide finer clues for the image privacy prediction task, which can improve its performance. I explore very deep CNN architectures, i.e., GoogLeNet and VGG, and carefully identify very deep visual semantic features obtained from the various layers of these CNNs to adequately infer the privacy for online shared images. I extract features from the pre-trained and fine-tuned CNNs on privacy data to get more dataset specific features.

Semantic features: As discussed earlier, scene features can contribute along with object features to learn privacy characteristics of a given image as they can help provide clues into what the image posters intended to show through the photo. Therefore, I employ two types of semantic features for privacy prediction based on: (1) objects stream, pre-trained on a large scale object dataset (ImageNet) (Russakovsky et al. 2015), to capture the object information depicted in the image; and (2) scene stream, pre-trained on a large scale scene dataset (Places2) (Zhou et al. 2016), to obtain the pattern about scene context of the image (Tonge, Caragea, and Squicciarini 2018).

Privacy-aware User Tags: Image tags have become very important for indexing, sharing, and searching applications. As these tags are at the sole discretion of the users, they tend to be noisy and incomplete. Thus, I propose privacy-aware tag recommendation algorithm, that aims at improving the quality of user annotations while also preserving the images' original sharing settings. These improved set of tags can help to improve the privacy prediction performance.

Multimodal feature fusion: Finally, I investigate the meta-classifiers trained on multimodal information obtained using all features to combine the strengths of tags, semantic (object and scene) and generic features (irrespective to object and scene). This work is currently under development.

Conclusion and Future work

In my dissertation work, I explore AI technology, i.e., deep features derived from multimodal information obtained using various layers of CNN networks for image privacy classification. The result of the classification task is expected to aid other very practical applications. Consider, for example, a law enforcement agent who needs to review digital evidence on a suspected equipment to detect sensitive content in images and videos, e.g., child pornography. The learning models developed here can be used to filter or narrow down the number of images and videos having sensitive or private content before other more sophisticated approaches can be applied to the data. In future, with the help of these features, it would be interesting to explore learning models for personalized image privacy prediction with varying degree of sensitivity.

Acknowledgments I would like to thank my advisor Dr. Cornelia Caragea for guidance and Dr. Anna Squicciarini for helpful discussions. This research is supported by NSF.

References

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 1–42.

Squicciarini, A. C.; Caragea, C.; and Balakavi, R. 2014. Analyzing images' privacy for the modern web. HT '14, 136–147. ACM.

Tonge, A., and Caragea, C. 2016. Image privacy prediction using deep features. In *AAAI' 16*, 4266–4267.

Tonge, A.; Caragea, C.; and Squicciarini, A. 2018. Uncovering scene context for predicting privacy of online shared images. In *AAAI' 18*.

Tran, L.; Kong, D.; Jin, H.; and Liu, J. 2016. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *AAAI '16*, 1317–1323.

Zerr, S.; Siersdorfer, S.; Hare, J.; and Demidova, E. 2012. Privacy-aware image classification and search. In *SIGIR*. NY, USA: ACM.

Zhou, B.; Khosla, A.; Lapedriza, A.; Torralba, A.; and Oliva, A. 2016. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.