Dynamically Identifying Deep Multimodal Features for Image Privacy Prediction

Ashwini Tonge,[1] **Cornelia Caragea**[2]
[1]Department of Computer Science, Kansas State University
[2]Department of Computer Science, University of Illinois at Chicago
atonge@ksu.edu, cornelia@uic.edu

Abstract

With millions of images shared online, privacy concerns are on the rise. In this paper, we propose an approach to image privacy prediction by dynamically identifying powerful features corresponding to objects, scene context, and image tags derived from Convolutional Neural Networks for each test image. Specifically, our approach identifies the set of most "competent" features on the fly, according to each test image whose privacy has to be predicted. Experimental results on thousands of Flickr images show that our approach predicts the sensitive (or private) content more accurately than the models trained on each individual feature set (object, scene, and tags alone) or their combination.

Introduction

Technology today offers innovative ways to share photos with people all around the world, making online photo sharing incredibly popular among Internet users. These users document daily details about their whereabouts through images and post pictures of their significant milestones and private events, e.g., family photos and cocktail parties. Privacy concerns are on the rise, and mostly emerge due to users' lack of understanding that semantically rich images may reveal sensitive information (Zerr et al. 2012). For example, a seemingly harmless photo of a birthday party may unintentionally reveal sensitive information about a person's location, personal habits, and friends. Thus, recently, researchers started to explore machine learning and deep learning models to predict images as either public or private (Tran et al. 2016; Tonge and Caragea 2016; Tonge, Caragea, and Squicciarini 2018; Tonge and Caragea 2018; Squicciarini, Caragea, and Balakavi 2014; Zerr et al. 2012). These studies mainly rely on objects, scenes, and user tag features as well as the combination of these features to achieve better performance than using individual features.

We conjecture that combining the object, scene and user tag features does not always help to identify sensitive content for images that are shared online. Consider the image given in Figure 1, which contains private content (someone's bedroom). A learning model trained on the combination of all features (object, scene and user tags) yields a very low probability (0.21) for the private class. However, the scene context given in the image (bedroom) proved to be sufficient to capture the sensitive content of the image and obtains a high probability of 0.62 for the private class. We believe that for a specific image, a certain type of feature (e.g., only scene) or a combination of features (e.g., object and scene) may be sufficient to accurately identify its sensitive content. Thus, in this paper, we propose to dynamically identify the smallest set of features derived from various Convolutional Neural Networks (CNNs) for target images that can adequately predict the class of an image as *private* or *public*. Our results show that the proposed approach identifies images' sensitive content more accurately than the individual feature sets (object, scene, and tags) or their combination. Our major contribution is to dynamically identify the most relevant semantic features by taking into account the type of features to be used for a specific type of input image.

	bed, studio dining table music, speakers	object: 0.5, **scene:** 0.62 tags: 0.29, all: 0.21

Figure 1: Anecdotal evidence of a private image and its tags.

Proposed Approach

We propose an approach for image privacy prediction that effectively identifies the most powerful or "competent" object, scene, and tag features for a target image whose privacy has to be predicted. Our features are derived from Object-CNN, Scene-CNN, and Tag-CNN, respectively. To predict the privacy of an image, we consider three stages:

(1) **Identify two neighborhoods for a target image**: Our approach assumes the existence of three datasets, denoted as $\mathcal{DS}1$, $\mathcal{DS}2$, and *Test*. $\mathcal{DS}1$ and $\mathcal{DS}2$ contain images that are labeled as *public* or *private*, with $\mathcal{DS}1$ being used for model training and $\mathcal{DS}2$ being used for neighborhoods' estimation. The *Test* dataset contains target images for which we aim to predict privacy. Given a target image I, we estimate two neighborhoods for I: Visual similarity based ($\mathcal{N_S}^I$) and privacy profile based ($\mathcal{N_P}^I$) neighborhoods. For $\mathcal{N_S}^I$, we use visual content features to calculate the similarity between the images. Specifically, we concatenate object & scene features and use them to identify the top k_v nearest neighbors from $\mathcal{DS}2$. For $\mathcal{N_P}^I$, we consider k_p most similar images

	Overall				Private			Public		
Features	Accuracy (%)	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
MPP	**86.30**	**0.858**	**0.857**	**0.863**	**0.681**	0.753	**0.622**	**0.913**	**0.890**	0.937
object	84.99	0.838	0.843	0.85	0.616	**0.772**	0.513	0.907	0.864	**0.953**
scene	84.45	0.833	0.836	0.844	0.606	0.749	0.51	0.903	0.863	0.947
Tags	83.03	0.826	0.823	0.83	0.612	0.662	0.57	0.891	0.873	0.91
concat	83.04	0.822	0.821	0.83	0.592	0.682	0.524	0.893	0.863	0.925

Table 1: Proposed approach (MPP) vs. Features.

that are identified by calculating the similarity between the privacy profiles of I and the images in $\mathcal{DS}2$. We define the privacy profile for an image as a vector of posterior privacy probabilities obtained by the models trained on all feature sets (object, scene and tags) using $\mathcal{DS}1$. We use these privacy profiles to bring sensitive content closer irrespective of their disparate visual content (e.g., different types of bedroom images). Also, we consider two different parameters k_v and k_p because the competence of a model is dependent on the neighborhood and estimating the appropriate number of neighbors for respective neighborhoods reduces the noise.

(2) **Competence estimation:** In this stage, we estimate the "competence" of all the models trained on the three types of features. To determine the competence of a model (e.g., scene), we consider images from both neighborhoods for a target image and determine which models classify these images correctly. The correctness of a model (e.g., scene) provides that the most competent model is selected for a particular type of images (e.g., home, bedrooms).

(3) **Feature selection:** Last, for a given target image, we first check the agreement on the privacy label between the models trained on all the feature sets. If not all models agree, then we estimate the competence of all the models and identify the subset of most competent models for the target image. Finally, we form the union of models trained on the subset of most competent features and take the majority vote to predict the privacy of the given target image.

Experiments and Results

Dataset and Evaluation Setting. We evaluated our approach on a subset of Flickr images sampled from the PicAlert dataset (Zerr et al. 2012). PicAlert consists of images on various subjects, which are manually labeled as *public* or *private* by external viewers. Private images belong to the private sphere (e.g., self-portraits, family, friends, home) or contain information that one would not share with everyone else (such as private documents). The remaining images are labeled as public. We split the dataset in $\mathcal{DS}1$, $\mathcal{DS}2$ and *Test* sets of $15,000$, $10,000$ and $7,000$ images, respectively. We used five different random seeds to generate the splits and averaged the results across the five different runs. The public and private images are in the ratio of 3:1 in all three sets. We used the Calibrated linear Support Vector Machine (SVM) to predict more accurate probability outputs. We used the best values of the parameters, i.e., $k_v = 700$ and $k_p = 100$, estimated over $\mathcal{DS}2$ dataset using 3-fold cross-validation.

Results and Observations. We compare the performance obtained by the proposed approach, Multimodality-based Privacy Prediction (MPP) with that of models trained on: (1) object, (2) scene, (3) image tags, and (4) their concatenation (concat). Table 1 shows the performance obtained by MPP and various features. We observe that MPP achieves the highest performance as compared to the models trained on individual features. We get the overall F1-score of 0.858, 0.838, 0.833, 0.826 and 0.822 using MPP, object, scene, tags and concatenation of the features, respectively. We also show the class-specific privacy prediction performance in Table 1 to identify which features identify the private class effectively since sharing private images on the Web with everyone is not desirable. MPP yields the best F1-score and recall for the private class. Precisely, F1-score for the private class improves from 0.592 (concat) to 0.681 (MPP), yielding a highest increase of $\approx 12\%$. Similarly, for recall, we obtain an increase of 11% over the object and scene features. With a paired T-test, the increase in F1-score (private) over all the feature sets are statistically significant for p-values < 0.05.

Conclusions and Future Work

In this work, we proposed an approach to adequately predict the class of a target image as *private* or *public* by dynamically identifying the subset of most competent features, derived from CNN architectures, corresponding to multimodal features of images. Precisely, our approach identifies powerful features corresponding to objects, scene context, and image tags that are derived from CNNs. Experimental results show that our approach predicts the private content more accurately than the models trained on individual feature sets. In the future, it will be interesting to study a dynamic multi-modal approach in a personalized setting.

Acknowledgments. This research is supported by NSF.

References

Squicciarini, A.; 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*.

Tonge, A., and Caragea, C. 2018. On the use of "deep" features for online image sharing. In *The Web Conference Companion*.

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*.

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