The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Uncovering Scene Context for Predicting Privacy of Online Shared Images

Ashwini Tonge,[1] **Cornelia Caragea,**[1] **Anna Squicciarini**[2]
[1]Department of Computer Science, Kansas State University
[2]College of Information Sciences and Technology, Pennsylvania State University
atonge@ksu.edu, ccaragea@ksu.edu, asquicciarini@ist.psu.edu

Abstract

With the exponential increase in the number of images that are shared online every day, the development of effective and efficient learning methods for image privacy prediction has become crucial. Prior works have used as features automatically derived object tags from images' content and manually annotated user tags. However, we believe that in addition to objects, the scene context obtained from images' content can improve the performance of privacy prediction. Hence, we propose to uncover scene-based tags from images' content using convolutional neural networks. Experimental results on a Flickr dataset show that the scene tags and object tags complement each other and yield the best performance when used in combination with user tags.

Introduction

Technology today offers innovative ways to share photos with people all around the world, making online photo sharing an incredibly popular activity for Internet users. These users share quotidian details and post pictures of their significant milestones and private events. The smartphones and other mobile devices facilitate the exchange of information virtually at any time. New privacy concerns are on the rise and mostly emerge due to users' lack of understanding that semantically rich images may reveal sensitive information. For example, a seemingly harmless photo of a student's New Year's party may accidentally reveal sensitive information about the student's location, personal habits, and colleagues.

Recently, Squicciarini et al. (2014) and Zerr et al. (2012) explored learning models for image privacy prediction using user tags and visual features and found that user tags are informative for classifying images as *private* or *public*. However, since user tags are at the sole discretion of users, they typically tend to be noisy and incomplete.Tonge and Caragea (2016) automatically derived object tags from images' content and showed that the combination of object tags and user tags outperforms each set of tags individually. Still, a manual inspection of the user tags revealed that these user tags contain both objects and scenes.

Thus, we posit that in addition to the object-centric tags, useful information about the scene context can be extracted from images' content to help better discriminate images as *private* or *public*. Figure 1 shows anecdotal evidence where only object information (shown in blue) may fail to infer correct predictions of images' privacy. Intuitively, adding scene information (shown in green italics) can help make a more accurate assessment. For example, the object tag "bed" can occur in private or public classes depending on the scene context, i.e., "home" or "bedroom" (private) vs. "furniture store" (public). To this end, we propose the extraction of scene tags to capture additional information from the visual content that is not captured by existing object tags. We explore the combination of user tags with the object, scene and object-scene tags for privacy prediction. Our results show that the combination of all three types of tags (object, scene, and user) yields better performance compared with user tags alone and the combination of user tags with scene or object tags. To our knowledge, this is the first work to uncover the scene context from the image content for privacy prediction.

(a) bed, piano, *home*

(b) persons, *sports*

Figure 1: Anecdotal evidence where only object tags fail to infer true privacy classes as private (a) and public (b).

Proposed Features

We believe that scene tags can contribute along with object tags to learn privacy characteristics of a given image, as they can help provide clues into what the image owners intended to show through the photo. Therefore, we 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 objects 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.

Object-centric Tags. To automatically obtain object tags from the visual content, we adopt an approach given in (Tonge and Caragea 2016). We use the probability distribution over 1000 object categories for the input image obtained by applying the softmax function over the last fully-

	$k = 2$					$k = 10$				
Features	Acc %	F1	Precision	Recall	#IncPred	Acc %	F1	Precision	Recall	#IncPred
UT	81.73	0.789	0.803	0.817	-	81.73	0.789	0.803	0.817	-
UT+ST	82.26	0.797	0.81	0.823	293	83.21	0.814	0.821	0.832	503
UT+OT	83.09	0.812	0.819	0.831	477	84.35	0.833	0.834	0.843	755
UT+ST+OT	**83.59**	**0.819**	**0.825**	**0.836**	**587**	**84.80**	**0.841**	**0.84**	**0.848**	**854**

Table 1: Object Tags vs. Scene Tags. The best performance is shown in bold.

connected layer of the AlexNet Convolutional Neural Network (CNN) (Krizhevsky, Sutskever, and Hinton 2012). AlexNet is pre-trained on the ImageNet dataset and obtained from the CAFFE distribution (Jia et al. 2014). We consider the top k objects of highest probabilities as *object tags*. Examples of predicted object tags for the picture in Figure 1(b) include "Maillot," "kimono," and "Tank suit."

Scene-centric Tags. Similar to object tags, we obtain the top k scenes derived from the probability distribution over 365 scene categories of the pre-trained AlexNet on the Places2 dataset. We refer to the top k predicted scenes as *scene tags*. Examples of scene tags for the picture in Figure 1(b) include "Athletic field outdoor," "Arena performance."

Experiments

Dataset and Evaluation Setting: We evaluate the quality of scene tags on a subset of Flickr images sampled from PicAlert (Zerr et al. 2012). PicAlert contains images on various subjects, which are manually labeled as *public* or *private* by external viewers. The **Train** and **Test** sets contain randomly selected $10,000$ and $22,000$ images from PicAlert, respectively. We use five different random seeds to obtain train/test splits and averaged results across the five runs. The public and private images are in the ratio of 3:1. We trained SVM models with an RBF kernel using all tag features.

Results and Observations: Tonge and Caragea (2016) showed that enriching the set of user tags with *object tags* resulted in higher performance models for image privacy prediction as compared to models that use only user tags, one reason being the sparsity of user tags. We examine the quality of automatically derived *scene tags* to understand if they bring additional information for privacy prediction.

Would scene tags obtained from the visual content bring additional information to improve privacy prediction? We examine the combination of user, scene, and object tags and compare it with that of user and object tags, proposed by Tonge and Caragea (2016) to determine if scene tags capture complementary information that is not already in the user tags and object tags. To obtain object and scene tags from CNNs, we experimented with two values of k as $k = 2$ and $k = 10$ (for the top k tags). The choice for $k = 2$ is motivated by the fact that an image may contain only a few scenes or objects, whereas the choice for $k = 10$ is consistent with prior work (Tonge and Caragea 2016) that showed best results for $k = 10$ for object tags. We also contrast the combination of user, scene, and object tags with the combination of user and scene tags and user tags alone. To encode the automatically derived scene and object tags, we use the probability of the tag obtained from the softmax layer of the corresponding CNN. The user tags are encoded using a binary representation.

Table 1 shows the performance obtained before and after adding scene tags (ST), object tags (OT) and scene + object tags (ST+OT) to the user tags (UT). We observe that adding both ST+OT to UT yields the highest performance. Particularly, models trained on the combination of all tag types yield an improvement as high as 5.2% in F1-measure over models trained on UT alone. Moreover, we note that the combination of UT+ST performs better than UT alone, but does not perform as good as the combination of OT+UT. Table 1 also shows the increase in the number of accurate predictions (denoted by #IncPred) for UT+ST, UT+OT, and UT+ST+OT over the user tags. As can be seen, the highest increase is achieved by the combination of UT+ST+OT.

Conclusion and Future work

We proposed the use of scene-centric tags (along with user tags and object tags) and showed that they can improve image privacy prediction. The results show that adding scene tags to user tags improves the performance over user tags alone. The best performance is achieved when we consider the combination of user, scene, and object tags. We conclude that scene and object tags complement each other and help boost the performance. In the future, more sophisticated methods to combine objects and scenes can be explored.

Acknowledgments. This research is supported by NSF.

References

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. MM '14, 675–678. ACM.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS'12*. 1097–1105.

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.

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

Zerr, S.; Siersdorfer, S.; Hare, J.; and Demidova, E. 2012. Privacy-aware image classification and search. In *ACM SIGIR' 12*.

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*.