Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Image Privacy Prediction Using Deep Features

Ashwini Tonge and **Cornelia Caragea**
Computer Science and Engineering
University of North Texas
AshwiniTonge@my.unt.edu, ccaragea@unt.edu

Abstract

Online image sharing in social media sites such as Facebook, Flickr, and Instagram can lead to unwanted disclosure and privacy violations, when privacy settings are used inappropriately. With the exponential increase in the number of images that are shared online, the development of *effective* and *efficient* prediction methods for image privacy settings are highly needed. In this study, we explore deep visual features and deep image tags for image privacy prediction. The results of our experiments show that models trained on deep visual features outperform those trained on SIFT and GIST. The results also show that deep image tags combined with user tags perform best among all tested features.

Introduction

The rapid increase in multi-media sharing through social networking sites such as Facebook, Flickr, and Instagram can cause potential threats to users' privacy. Many users are quick to share private images about themselves, their family and friends, without thinking much about the consequences of an unwanted disclosure of these images. Moreover, social networking sites allow users to tag other people, which can reveal private information of all users in a particular image (Ahern et al. 2007). Users rarely change default privacy settings, which could jeopardize their privacy (Zerr et al. 2012).

Current social networking sites do not assist users in making privacy decisions for images that they share online. Manually assigning privacy settings to each image each time can be cumbersome. To avoid a possible loss of users' privacy, it has become critical to develop automated approaches that can accurately predict the privacy settings for shared images.

Several studies have started to explore classification models of image privacy using image tags and image content features such as SIFT (or Scale Invariant Feature Transform) and RGB (or Red Green Blue) (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2014) and found that image tags are very informative for classifying images as *private* or *public*. However, given large datasets of labeled images, e.g., the ImageNet dataset (Russakovsky et al. 2015), deep neural networks are now able to learn powerful deep features (Jia et al. 2014) that go beyond SIFT and RGB, and work well in many image analysis tasks.

In this study, we explore an approach to image privacy prediction that uses deep visual features and deep tags for predicting an image as *private* or *public*. Specifically, we investigate two deep feature representations corresponding to the output of two layers of an eight-layer deep neural network pre-trained on ImageNet (Russakovsky et al. 2015). We further investigate deep image tags, which correspond to the top ranked categories derived from the probability distribution over categories obtained from the last layer of the deep neural network via the softmax function. The results of our experiments on Flickr images show that models trained on deep visual features outperform those trained using the combination of SIFT and global descriptor features (GIST) (Oliva and Torralba 2001). Moreover, deep image tags combined with user tags yield better performing models compared with those based only on user tags.

Privacy Prediction Using Deep Features

The privacy of an image can be determined by the image content and its description. We extract visual features and tags for differentiating between private and public settings.

Deep Visual Features: A deep neural network takes one or more blob as input and produces one or more blob as output. Layers in the network are responsible for forward pass and backward pass. Forward pass takes inputs and generates the outputs. Backward pass takes gradients with respect to the output and computes the gradient with respect to the parameters and to the inputs, which are consecutively back-propagated to the previous layers (Jia et al. 2014). In the convolutional neural network architecture, features are extracted from images through each layer in a feed-forward fashion. The architecture consists of eight layers with weights; the first five layers are convolutional and the remaining three are fully-connected (FC). The last two fully connected layers are referred as FC_7 and FC_8. We used the output of FC_7 and FC_8 as *deep visual features* for images. The final output layer is referred as "Prob" and is obtained from the output of FC_8 via a 1000-way softmax function, which produces a probability distribution over the 1000 object categories for an input image.

Deep Tag Features: We extract deep tags for images based on their visual content. Specifically, for an image, we

Features	Accuracy	F1-Measure	Precision	Recall
Test ($PiCalert_{783}$)				
FC_7	81.23%	0.805	0.804	0.812
FC_8	**82.63%**	**0.823**	**0.822**	**0.826**
SIFT + GIST	72.67%	0.661	0.672	0.727
User Tags	79.82%	0.782	0.786	0.798
Deep Tags	80.59%	0.801	0.799	0.806
User+Deep Tags	**83.14%**	**0.827**	**0.826**	**0.831**

Table 1: Results for visual features.

predict top k object categories, corresponding to the highest k probabilities from the probability distribution over categories, i.e., the "Prob" layer of the deep neural network. The k predicted categories are used as tags to describe an image.

Experiments and Results

We evaluate our approach on Flickr images sampled from the PiCalert dataset (Zerr et al. 2012). PiCalert consists of Flickr images on various subjects, which are manually labeled as *private* or *public* by external viewers. *Private* images disclose sensitive information about a user, e.g., images with people, family pictures, etc., whereas *public* images generally depict scenery, objects, animals, etc., which do not provide any personal information about a user. Previous works on this task used SIFT and user tags for privacy prediction (Zerr et al. 2012; Squicciarini, Caragea, and Balakavi 2014). Our results indicate that deep visual features and deep tags outperform these previous works.

We sampled $4,700$ images from PiCalert and used them for our privacy prediction task. The public and private images are in the ratio of 3:1. We divided these images into two subsets, **Train** and **Test**, using 6-fold stratified sampling. **Train** consists of five folds randomly selected, whereas **Test** consists of the remaining fold. The number of images in **Train** and **Test** are $3,917$ and 783, respectively. In all experiments, we used the Support Vector Machine (SVM) classifier implemented in Weka and chose the hyper-parameters (i.e., the C parameter and the kernel in SVM) using 5-fold cross-validation on **Train**. We experimented with different C values, and two kernels, linear and RBF.

Deep Visual Features vs. SIFT and GIST. We first compare the deep features, FC_7 and FC_8 with the combination of SIFT and GIST features.

For SIFT, we constructed a vocabulary of 128 visual words. Other vocabulary sizes, e.g., 500, 1000, did not yield improvement in performance on **Train** using 5-fold cross-validation. For GIST, we followed the steps: (1) convolve the image with 32 Gabor filters at 4 scale and 8 orientations, which produces 32 feature maps; (2) divide the feature map into a 4×4 grid and average feature values of each cell; (3) concatenate these 16 averaged values for 32 feature maps, which result in a feature vector of 512 (16×32) length.

For the deep visual features, we used an already trained deep convolutional neural network implemented in CAFFE (Jia et al. 2014), which is an open-source framework for deep neural networks. CAFFE implements an eight-layer network pre-trained on a subset of the ImageNet dataset (Russakovsky et al. 2015), which consists of more than one million images annotated with 1000 object categories. We resized images in both **Train** and **Test** to the CAFFE convolutional neural net compatible size of 227×227 and encoded each image using the two deep feature representations corresponding to the output of the layers FC_7 and FC_8.

Table 1 shows the performance of SVM using FC_7 and FC_8 deep features, and SIFT+GIST features, on **Test** (first three lines). We do not show the performance of SIFT and GIST independently since they perform worse than their combination. As can be seen from Table 1, both deep visual features FC_7 and FC_8 outperform SIFT + GIST, and FC_8 performs better than FC_7, in terms of all measures.

Deep Tags vs. User Tags. Next, we contrast the performance of SVM on user tags, deep tags, and the combination of user tags and deep tags. For deep tags, we consider top $k = 10$ object categories as tags. We tried different k values for the deep tags and achieved good results with $k = 10$. Since tags generally appear only once per image, we used Boolean features, i.e., 1 if a tag is present in an image and 0 otherwise. Table 1 shows the results obtained from the experiments for tag features on the **Test** set (last three lines). As can be seen from the table, deep tags perform only slightly better than user tags, but their combination outperforms both user tags and deep tags. A closed look at both tag types revealed that user tags typically consist of general terms, whereas deep tags consist of specific terms, which could explain the improved performance of user + deep tags.

Conclusion and Future work

We tackled the image privacy prediction task and showed experimentally that models trained on deep visual features and the combination of user tags and deep tags yield remarkable improvements in performance over models trained on features such as SIFT, GIST and user tags. In future, it would be interesting to explore the combination of deep visual features with deep tags, using ensemble of classifiers.

Acknowledgments We would like to thank Adrian Silvescu and Raymond J. Mooney for helpful discussions. This research is supported in part by the NSF award #1421970.

References

Ahern, S.; Eckles, D.; Good, N. S.; King, S.; Naaman, M.; and Nair, R. 2007. Over-exposed?: Privacy patterns and considerations in online and mobile photo sharing. In *SIGCHI'07*.

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. In *Intl. Conf. on Multimedia*.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision* 42(3):145–175.

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. *International Journal of Computer Vision (IJCV)* 1–42.

Squicciarini, A. C.; Caragea, C.; and Balakavi, R. 2014. Analyzing images' privacy for the modern web. In *HT 2014*.

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