Explainable and Personalized Privacy Prediction

Preetam Prabhu Srikar Dammu[1], Srinivasa Rao Chalamala[1] and Ajeet Kumar Singh[1]

[1]TCS Research, Tata Consultancy Services Ltd., India

Abstract

Due to the ever-increasing web presence of people and proliferation of image sharing on social media, it is becoming increasingly difficult for users to maintain the privacy and security of their sensitive data. Majority of the users share their images on various platforms presuming that the data would only serve its intended purpose. However, in reality, there is a significant risk of the images falling into the wrong hands and eventually being used for malignant purposes without the users' knowledge. To prevent any unwanted disclosures, it is imperative to devise effective techniques that notify the user to review their decision before any sensitive information is shared. Several methods have been proposed to execute this task, yet most have shortcomings that might make them unsuitable for end-users. In this paper, we propose a configurable privacy prediction system that addresses some of the major drawbacks of the existing methods while still achieving state-of-the art performance. The proposed solution accommodates personalization which enables the users to include their privacy preferences and tweak the system according to their requirements. Along with the predictions, the proposed system also provides user-friendly human-readable explanations.

Keywords

Privacy Prediction, Explainability, Information Leakage Prevention

1. Introduction

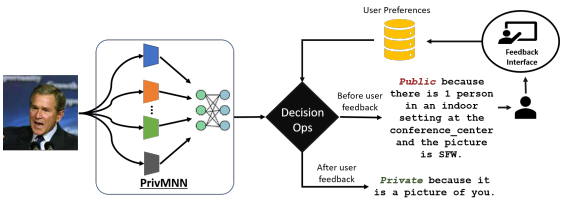

Figure 1: The goal is to devise a personalized privacy prediction system equipped with an explanation generation module and a feedback interface for the user. For example, a user image is classified as *private* (previously classified as *public*) once the user gives his feedback about the privacy of the image, as illustrated in the figure above.

In today's highly connected world, sensitive content shared on the internet without appropriate privacy settings can adversely affect all parties involved from users to corporations. For instance, malicious videos could be created using deepfakes [1] if sufficient images of the user are made available to the public, and this is a serious security concern that could damage the user's reputation [2, 3]. Even when used in legitimate tasks such as background verification of job applicants by employers, personal content which might be perceived as indecorous might harm one's career prospects [4, 5, 6].

Despite the serious concerns over mismanaged personal content, studies have shown that majority of the users fail to diligently protect the privacy of their data either due to lack of awareness or difficulties in managing privacy settings [7, 8, 9]. Even in the case of proactive users who actively manage their privacy settings, the authors of [7] demonstrate that the users' judgement of privacy risk may not accurately represent the true level of associated privacy risk. These factors project the substantial necessity of an accurate and informative privacy prediction system.

One of the major challenges in privacy prediction is the fact that privacy is highly subjective in nature. While an image can be construed as private by some, others could argue that it is public in nature. The ground truths used for training the prediction models are usually collected from multiple manual reviewers who vote based on their perception and the majority vote for a particular image is considered to be its training label [10]. Therefore, theoretically even if a model achieves 100% accuracy on the training labels, a considerable number of manual annotators would disagree with the model's predictions. This fact signifies the need of personalization in privacy prediction systems. In order to provide relevant predictions for each user based on their privacy requirements, the system needs to allow personalization. However, there are significant challenges to personalized models such as requirement of large amounts of user data, resource constraints to train and deploy them, and difficulties associated with accommodating sudden changes in preferences [11]. Our paper describes a novel approach to address these problems yet requires similar amount of data and computational resources as non-personalized

3rd International Workshop on Privacy, Security, and Trust in Computational Intelligence (PSTCI2021)

d.preetam@tcs.com (P. P. S. Dammu); chalamala.srao@tcs.com (S. R. Chalamala); ajeetk.singh1@tcs.com (A. K. Singh)

Figure 2: Explanations for images correctly classified as *Private* (a-d) and *Public* (e-h) by PrivMNN (refer section 5.1). Results discussed in further detail in section 7.1. SFW: Suitable-for-Work, NSFW: Not-Suitable-for-Work.

methods. Additionally, our method keeps up with the changing user requirements and updates its behavior instantaneously to reflect the preferences.

It is commonplace for users to overlook important details that might leak private information. For example, users may not pay much attention to what is visible in the background when taking pictures in their personal space and sensitive content such as computer screens or confidential documents might accidentally be captured in the frame. In such cases, users may assume that the prediction is erroneous unless an explanation is provided along with the prediction, bringing the unnoticed details to the users' attention. In many cases, the user may not be aware of the privacy implications of an image's contents which would lead to a difference in the user's judgement and the predicted label [7]. In such scenarios, an explanation would become imperative to convince the user about the apparent privacy violation. Otherwise, there is a risk of the user ignoring the suggested privacy status due to a lack of confidence in the automated prediction. Hence, we devise our system to generate an explanation along with each prediction.

Reusability in machine learning applications wherever applicable is paramount as building models from scratch unless required is redundant, expensive in terms of both computational resources and time, and even has environmental impacts [12]. In this work, we leverage several pretrained deep learning models that have achieved impressive results and use them efficiently in our privacy prediction system.

In this work, we introduce a novel framework that introduces multiple important characteristics to the privacy prediction process while still achieving state of the art performance. To the best of our knowledge, there does not exist any other privacy prediction method that addresses these multiple challenges simultaneously. Contributions of the proposed approach are five-fold:

1. *Personalization* realized through user feedback.
2. *Explanation* for each prediction in real-time.
3. *Configurability* by allowing modifications to the system composition.
4. *Customizable privacy settings* for enforcing elevated security constraints in user-specified conditions.
5. *Instantaneous updation* of the user's privacy preferences.

2. Related Work

Over the years, researchers have explored various machine learning techniques for evaluating the privacy of the images. These approaches employed several types of data in addition to image data, such as social group information, location information, deep tags, user tags and others. While there are advantages of training classifiers on each of these alternative data types, most of them tend to have limitations such as availability and noises. Relying solely on image contents and features derived directly from images appear to be more promising as they do not depend on the availability of additional information and as a result are more generic in nature and hence, more reliable.

Classifiers trained on deep tags in conjunction with user tags have achieved promising results [13] as the automated deep tags would mitigate the scarcity of user tags. In [14], Tran *et al.* demonstrate that employing high-level hierarchical features at object level is beneficial compared to using low-level vision features which are usually non-informative to users. Traditional computer vision techniques such as Scale-Invariant Feature Transform (SIFT) [15] and Global Image Descriptor (GIST) [16] have been used in [17, 11, 18], and the authors in [17, 19] find that SIFT along with image tags perform best for image privacy classification. In [20], Tonge *et al.* leverage multi-modal data fusing object, scene context, and image tags information and report promising results. In [21], Tonge *et al.* present impressive results by using deep visual features and tags derived from widely-known CNN architectures, and also present a detailed comparison with prior works.

Interestingly, all of the methods discussed above share a common limitation, i.e., they do not support personalization. Fewer studies have been conducted on personalized models for privacy prediction [11, 7]. In [11],

Zhong *et al.* present that personalized models are more expensive to build in terms of computational costs, space and time requirements. To address these challenges, Orekondy *et al.* [7] use clustering of user profiles and map each user to one of the representative clusters which are significantly fewer in number, and as a result, reduce the computational costs involved.

At present, there does not exist any privacy prediction method that generates intuitive human-readable explanations, to the best of our knowledge. Post-hoc interpretability methods such as LIME [22], SHAP [23], GradCAM [24] and many others can be used for image classification tasks such as privacy prediction, but most of these methods generate generic heatmaps without any contextual link to privacy. The visual cues generated by the multitude of interpretability methods in the literature are difficult to decipher even for domain experts and would not be suitable for the wider audience.

3. Problem Overview

In this section, we introduce the problem and provide a brief overview of the proposed approach.

3.1. Goal

The goal is to provide the users with a system that can be configured according to their own privacy preferences. To this end, we build a composite machine learning system to classify images as *Private* and *Public*. However, it should be noted that a standard definition of privacy does not exist, in fact, it is highly subjective as each individual might have a different interpretation of privacy. Therefore, for the predictions to be relevant to the user, it is paramount that the system incorporates the user's preferences and requirements. Additionally, we also provide cogent reasoning through explanations for the predictions to aid the users' understanding of the privacy implications and help them make an informed decisions.

3.2. Approach Overview

In our effort to capture the subjective nature of privacy, we draw parallels to how humans perceive privacy and arrive at a decision if an image is private or public. To achieve this, we adapt Modular Neural Networks (MNN) which are biologically inspired by the modularization found naturally in the human brain [25, 26]. When humans attempt to assess the sensitivity of the content present in an image, they consider several aspects and ask themselves relevant questions that would help them arrive at a logical conclusion. We model this design of thinking by allocating a dedicated module for each of the

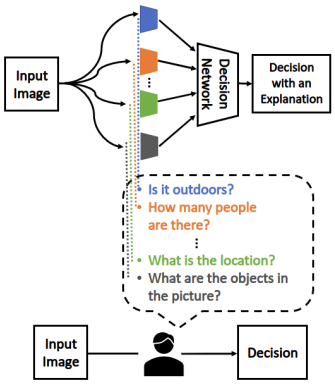

Figure 3: Modeling the design of human-thinking and decision process.

privacy-related questions in the Privacy MNN (PrivMNN) (see Figure 3). Depending on the users' perception of privacy, their preferences and requirements, these questions might vary considerably. However, research has shown that generic patterns and trends do exist with respect to privacy of images when studies are conducted at a large scale [17, 19, 18]. Therefore, we begin by training a model which captures the patterns that apply to the generic population and provide individual users with the flexibility to tweak the privacy prediction system to better suit their perspectives, requirements and personality traits. We detail this approach in Explainable and Personalized Privacy Prediction in Section 5.

4. Preliminaries

In this section, we discuss the preliminary concepts and tools that have been used in conducting this study. Pointers to relevant resources for further exploration of these topics have also been provided.

FasterRCNN [27]: In this study, we employ Faster-RCNN to perform object detection to learn the contents of an image which could provide clues on determining its privacy.

Places365-CNNs [28]: We use Places365-ResNet [28] to predict information regarding the location which is vital for determining the privacy of the image.

MTCNN [29]: In this study, we have used MTCNN to detect human faces and vehicle license plates [30].

NudeNet [31]: We use this model to determine if nudity is present in a given image.

ResNet [32]: In [21], the authors demonstrate that ResNet produces the best feature representations for privacy prediction task. We use ResNet for generating feature embeddings which are used for retrieving visually similar images.

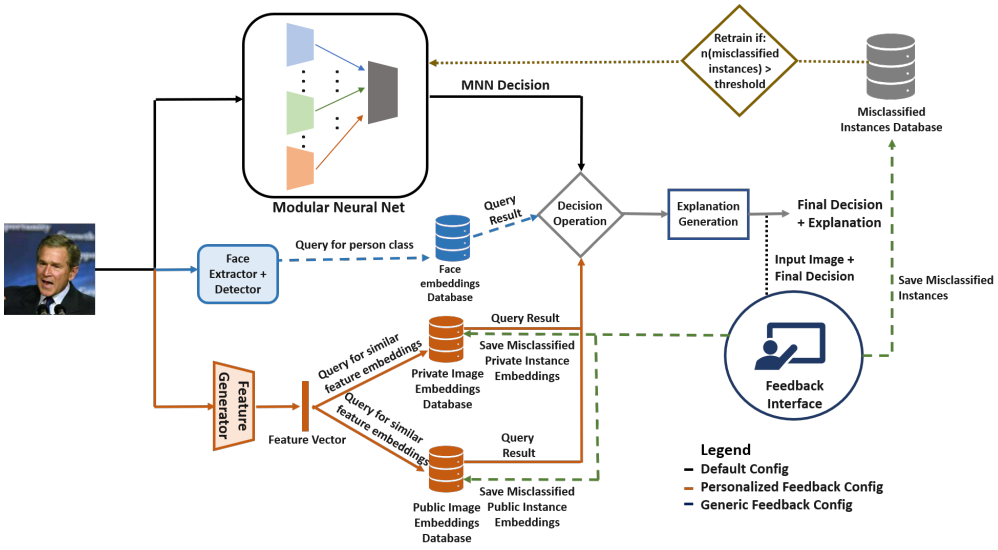

Figure 4: Workings of Explainable and Personalized Privacy Prediction System. **Note: Best viewed in color.**

FaceNet [33]: FaceNet is a deep learning model that generates high quality face embeddings. In this work, we use FaceNet to generate embeddings of faces present in a given image.

5. Explainable and Personalized Privacy Prediction

In this section, we detail the novel approach to privacy prediction that leverages the abundance of deep learning machinery. We logically combine different machine learning tools and devise a framework to deliver an explainable and personalized privacy prediction system.

The system design of our approach is largely inspired by the human decision-making process and we employ PrivMNN (5.1) that draws parallels to the questions humans would contemplate for inspecting the privacy content of an image (please refer section 3.2). To accommodate the users' differing privacy definitions and preferences, we rely on supplementary techniques in addition to PrivMNN which will be discussed in subsections 5.3 and 5.4. The role and nature of user interactions are discussed in subsection 5.2. Details of how a consolidated final prediction is arrived at by considering the outputs of all the components is discussed in subsection 5.4. The sequential flow of operations is presented in subsection 5.5. The generation of explanations for the predictions is discussed in subsection 5.6.

5.1. Privacy Prediction MNN (PrivMNN)

With the help of the composite nature of PrivMNN, we encapsulate various deep learning tasks such as scene recognition, object detection, facial detection and nudity detection into a comprehensive system that is transparent yet effective in its functioning.

In this section, we present the architectural composition of the PrivMNN. We discuss different modules and their workings in detail, and how their outputs are put together and fed to the decision network for personalized privacy predictions.

5.1.1. Modules

Each of the modules in the PrivMNN handles a subtask that addresses fundamental questions which will result in the generation of relevant information vital for assessing the privacy of an image.

Object Detection Module: This module addresses the question "what objects are present in the picture?". This information is valuable for inspecting the existence of sensitive content. For example, if a laptop is visible in the picture, then it is possible that sensitive information is on display. Pre-trained FasterRCNN model [27] is used for generating an embedding vector of size 81 indicating the presence of different objects.

Location Detection: This module addresses the questions "Is the picture taken indoors or outdoors?" and "What is the location?". Recognising the scene where the picture was taken is essential, since pictures taken in private space such as bedrooms tend to be more pri-

vate than pictures taken in public space. Pre-trained Places365-ResNet [28] model is employed for generating an embedding vector of size 367 which correspond to 365 different locations two additional dimensions for indicating if the picture is taken indoors or outdoors.

***Object Localization*:** This module tackles the question "How many people are present in the image?". If people are present in the image, it indicates that the picture contains Personally Identifiable Information (PII). It is imperative to handle PII data with additional care. MTCNN is also used for detecting license plates in later part of the study (see section 7.4). Pre-trained MTCNN model [32] is used for generating an embedding vector of size 6 to indicate the number of people present in the picture ranging from 0 to 4 persons and above 4 as many people.

Explicit Content Detection: This module answers the question "Is there any explicit content in the image?". The presence of explicit content usually indicates that the picture may not be safe for public viewing. We use pre-trained NudeNet model [31] to generate a vector of size 2 to indicate if an image is Suitable-for-Work (SFW) or Not-Suitable-for-Work (NSFW).

5.1.2. Decision Network

The decision network is responsible for consolidating the information from all the modules and making the final prediction. It takes the concatenated embedding vector of size 456 which is obtained by combining output vectors of all modules as input and performs binary classification i.e. public or private class.

The decision network in the proposed PrivMNN consists of four fully connected layers with Leaky ReLU activation function and is trained with Adam optimizer. Weighted Binary Cross-Entropy (BCE) is used to address the class imbalance problem.

5.2. Feedback Interface

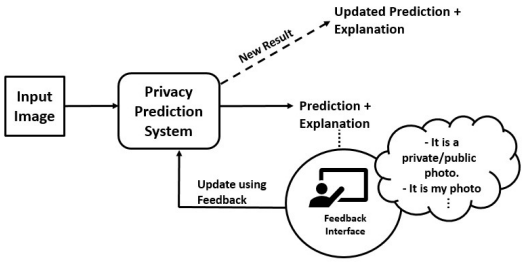

Figure 5: User Interaction Overview

Using PrivMNN alone (in default configuration), we are able to achieve high accuracy which is on par with previously reported state-of-the-art performance [21]. However, this score only reflects how well the model learns the generic privacy trends in the dataset provided by the annotators. A user whose privacy preferences deviate significantly from these generic trends might not find the predictions generated by the model to be useful. Therefore, our goal is to incorporate the users' privacy preferences by utilizing their feedback inputs.

For the feedback interface to be effective, the actions required from the user should be intuitive and user-friendly. Keeping this in mind, the user would be required to perform two simple tasks:

1. Identifying misclassified instances: The users would have the option to indicate if a prediction is contradicting their own judgement by identifying the corresponding image and providing its intended label.

2. Identifying faces of personal acquaintances: The user would have the option to identify the presence of either theirs or their personal acquaintance's face in an image, and the required privacy condition. This actions needs to be performed only once for each face.

5.3. Feature Extraction and Storage

Besides PrivMNN, we utilize two additional deep learning models to achieve our goal of personalization. These models are used for generating feature embeddings for the feedback inputs given by the user.

ResNet: Feature vectors of seemingly misclassified public and private images are extracted using pre-trained ResNet model and stored separately in respective databases for future references. Once entry has been stored, future image instances are checked for similarity with these user-rectified instances to prevent mistakes of the same nature in the future. In Figure 4, the 'Feature Generator' represents ResNet and the discussed corresponding steps are highlighted in orange.

FaceNet: FaceNet is used for generating the feature embeddings of user-identified faces which are then stored in a face embeddings database. MTCNN [29] is used in conjunction with FaceNet [33], where MTCNN is used for detecting and localizing the facial region in an image. In Figure 4, the 'Face Extractor + Detector' represents the MTCNN and FaceNet pair and the discussed corresponding steps are highlighted in blue.

5.4. Decision Operations (Ops)

In the default configuration, when the user has not provided any feedback yet, the decision operation simply passes the prediction from the PrivMNN as the final decision. Once the user starts providing feedback, the decision operation is updated accordingly to reflect the user's privacy requirements.

In this work, we make use of the traditional yet effective rule-based systems [34]. Initially, the rule-base prior to the creation of any user-specified rules or learning them from the user inputs is kept basic i.e., any query results from the databases (public, private or face embeddings) is given higher priority than the PrivMNN's predictions since these results would be based on user interactions. In cases of ambiguity or conflicting labels, the *private* class is preferred as it is the safer side to err on.

The user should not be required to provide hundreds of corrective feedback inputs before the system starts to show some improvement, which would be the case if we attempt to retrain the decision neural network to reflect these changes. Instead, a list of rules can easily be created from very few user feedback inputs or they can even be explicitly created by the user by mentioning their requirements. These rules can be put into effect at the very instant and the improvement in results can be observed in all predictions henceforth.

As the user continues to interact with the system, more complex rules can emerge. For instance, if the user marks their own pictures as *public* when outdoors in a public location like a sporting venue, but marks the pictures as *private* when located in personal space like a bedroom, a rule can be created to reflect this preference. On the contrary, users could also choose to be notified prior to sharing any images of themselves by strictly marking all of the images containing their faces as private. This design makes the system highly customizable, as any number of rules can be added or removed when not required.

5.5. Flow of Operations

In previous sections we describe every module in detail. Now we illustrate the flow of data through the system for privacy prediction. As demonstrated in Figure 4, the input is simultaneously processed by the PrivMNN, the ResNet model and the face detector. Subsequently, the query results and PrivMNN's prediction are fed to the decision operator which generates the final decision. If the user finds the result acceptable, no further actions are required. However, if the user finds the result is not appropriate according to their perspective, they could provide a feedback to the system. This feedback is used for creating customized rules in the decision ops module, and for updating the relevant databases. All misclassified instances are also stored in a separate database, and if the number of these instances is significant enough, the PrivMNN could be fine-tuned from these. However, it is unlikely that a single user would provide such a high number of misclassified instances as it is expected that the system would handle most of the user's concerns after receiving a few feedback inputs from the user.

5.6. Explanation Generation

Intuitive and human-readable explanations that require no technical expertise from the user are the most suitable ones for end-user applications, such as this one. Heatmaps and other form of explanations could be prove to be confusing for the technically uninitiated. To this effect, we generate textual explanations which provide descriptive explanations of the predictions that are easy to understand.

With abundance of information being generated by the various components of the system, it simply becomes a matter of conveying the information coherently. We employ regular expressions, a simple yet effective method, for presenting our explanations. The default explanations generated by the system are quite straight-forward and as the system evolves the explanations also change according to the scenario (see Figures 8 and 10).

6. Datasets

6.1. PicAlert! Dataset

We conduct our experiments on a subset of the PicAlert! dataset [10] used in previous studies [21, 20], which originally had 32,000 images from which 27,000 were used for training and remaining 5,000 for testing. However, few of the images are now inaccessible from the Flickr website, hence the updated dataset used in this study consists of 30,136 images out of which 25,136 were used for training and remaining 5,000 were reserved for testing. We maintain the ratio of *Public* and *Private* images to be 3:1 in each split, similar to the splits used in previous studies. The images in this dataset are labeled manually by multiple reviewers and at the least by two reviewers [10, 18]. In the occurrence of conflicting labels, the image is shown to an additional reviewer in order to reach a consensus on the final label [10, 18].

6.2. Labeled Faces in the Wild (LFW) Dataset

The Labeled Faces in the Wild (LFW) [35] dataset is a widely used benchmark dataset for face verification consisting of 13233 images of 5749 individuals. This dataset is used only for testing purposes in experiments where multiple images of the same person were required (see section 7.3).

7. Experiments and Results

For the purpose of illustrating the proposed system's configurability and its improvement in performance with user feedback, we demonstrate the system's behavior

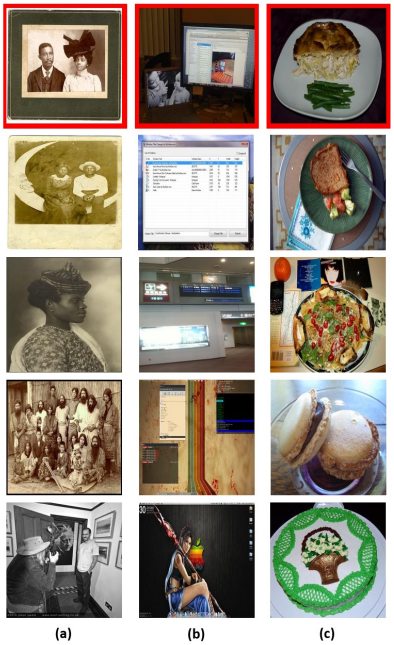

Figure 6: Examples of images falsely classified as *Public* initially. Images in first row represent corrective feedbacks and each column, from second row onwards, contains images with rectified predictions as a result of the feedback.

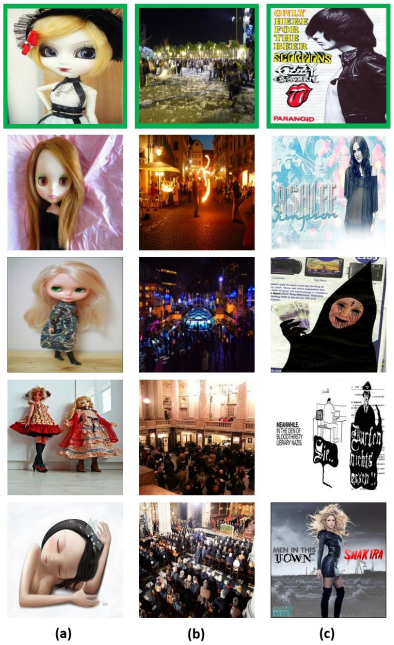

Figure 7: Examples of images falsely classified as *Private* initially. Images in first row represent corrective feedbacks and each column, from second row onwards, contains images with rectified predictions as a result of the feedback.

and prediction results in four different system configurations. The proposed method outperforms the previously reported SOTA accuracy of 87.58% [21] in *generic feedback configuration* [7.2] (See Table 1).

Method	Accuracy	Private			Public		
		F1	Prec	Rec	F1	Prec	Rec
ResNet	87.58	71.7	78.3	66.2	92.0	89.9	94.3
PCNH	83.13	62.4	70.4	56.1	89.1	86.3	92.1
UT	78.63	49.6	56.5	44.2	86.5	83.7	89.4
DT	83.78	63.1	68.8	58.4	89.6	87.6	91.7
UT+DT	84.33	67.0	70.9	63.6	89.7	88.2	91.3
Default Config	86.44	73.8	73.1	74.5	90.8	91.1	90.5
Feedback Config	**88.36**	**77.2**	**77.4**	**77.1**	**92.1**	**92.1**	**92.2**

Table 1
Performance comparison of the proposed method with prior art scores as reported in [21]. Note: All the numbers are in percentage.

Subsection 7.3 consists of experiments on LFW dataset and the concept presented in subsection 7.4 was not taken into account while labeling Flickr dataset, hence this dataset was not considered for evaluating these configurations.

7.1. Default Configuration

Configuration Settings: This is the default configuration, in which the system has not taken any feedback inputs from the user yet. The final decisions are solely based on PrivMNN predictions.

The purpose of this configuration is to demonstrate the baseline behavior of the privacy prediction system. Even in its basic configuration, the proposed system performs better than most methods proposed in the literature and it is almost on par with the SOTA model (see Table 1). However, our main objective here is not the performance alone, rather it is the incorporation of explainability and personalization to the system. In this particular baseline configuration, we focus on explainability alone and we do not utilize user inputs yet.

By examining the explanations provided for the corresponding images in Figure 2, it is evident that the explanations reveal useful insights about the model's behavior. As expected, the model almost always predicts the image to be private when humans are present in the image, unless they are situated in a public venue (Figure 2(f) and (h)). However, an interesting trend can be noticed by comparing 2 (b) and 2 (f), the picture which was determined to be NSFW was labeled private even though remaining conditions remained mostly similar (indoors,

one person and a public venue) with the one that was classified as public because it was SFW. Comparing 2(d) and 2(h) reveals the importance of location (personal space, public venue).Pictures (2 (e) and 2 (g)) which had no people present in them and were clicked outdoors were classified as public.

7.2. Generic Feedback Configuration

Configuration Settings: In this configuration, the user provides feedback inputs only on misclassified instances (see section 5.2). The performance of the system is updated with every feedback input received from the user. After 21 feedback inputs, the proposed system surpasses previously reported state-of-the-art accuracy [21]. It should be noted that the increment or reduction of performance is dependant on the user's feedback.

The purpose of this configuration is to demonstrate how the system incorporates user's preferences based on corrective feedback provided by the user on misclassified instances. In Figure 6 and 7, the images in the first row correspond to the misclassified instances identified by the user. Subsequent images shown below them in their respective columns are images which were incorrectly classified previously but are correctly classified now as a result of the corrective feedback. In Figure 6 (a), we notice that a family portrait was erroneously classified as public and when the user corrects this mistake by marking it as private, other visually similar portraits are identified as private by the updated system. A picture of a computer screen marked as private in 6 (b) resulted in multiple other images of computer screens to be classified as private as well. In 6 (c), a picture of a meal was marked as private which prompted the system to classify other pictures of food as private. Although these images of food have been labeled as private in the flickr dataset by the reviewers, it should be noted that the user might have a different view, and the proposed framework allows the users to personalise the system by incorporating their perspectives. Similarly, we observe in Figure 7 that images of dolls or album covers were initially identified as private but after receiving corrective feedback from the user, they were correctly identified as public. As the reasoning behind the predictions made based on the user feedback is different in nature, it makes more sense to use prototypical explanations [36] instead of the textual explanations. Therefore, we provide explanations for these types of decisions as shown in Figure 8.

7.3. Personalized Feedback Configuration

Configuration Settings: In this configuration, the user provides feedback by only identifying faces of personal acquaintances (see section 5.2).

Figure 8: Prototypical Explanations

This configuration demonstrates how the predictions change if the user or an acquaintance of the user is present in the image. Naturally, users are more concerned for the privacy of the images in which they or their friends or family members are present, however, the existing methods do not this into consideration while arriving at a prediction. We believe that this information is a very strong indicator of privacy and therefore cannot be overlooked and must be incorporated to the prediction process. In Figure 9, we show that identifying a single face image of a person is sufficient to treat all images of that person with higher importance with respect to privacy. By default, all images of identified persons are treated as private but more complex rules can be created based on the user's specifications.

Figure 9: Elevated privacy conditions for user's pictures.

7.4. Customizability Demo Configuration

Configuration Settings: In this configuration, we add a new module to the PrivMNN to detect license plates.

We illustrate the customizability of the proposed framework by updating the privacy prediction system with a new module in this configuration. License plates are known to be indirect PII and studies have been conducted which discuss the privacy implications of vehicle images with visible license plates [37, 38, 39]. To address this specific concern, we include a dedicated module for detecting license plates. As can be observed from Figures 2 (g) and 10, an image of cars where license plates are visible was classified as public earlier but predicted to be private after the addition of license plate recognition module. It is worthy to note that the training label for this particular image is *public* in the dataset as they have not considered this issue. Similarly, depending on the user's requirements and by leveraging domain knowl-

edge, many tailored specifications can be baked into the prediction process.

Figure 10: License plate protection

8. Conclusion and Future Work

The proposed privacy prediction system simultaneously incorporates several desirable features such as personalization, explainability, configurability, customizability and responsiveness to user's changing privacy requirements. Our approach generates explanations in a detailed manner to: (i) provide meaningful feedback, and (ii) make the user aware of the sensitive content present in the image.

The flexibility of the proposed privacy prediction system enables it to be extended in several ways, like adding new modules or using a more sophisticated decision operations. In the future, we also plan to support predictions for images which are considered to be *undecidable* in terms of privacy, by adding an additional class to the existing two classes. The modular nature of the framework allows us to do privacy prediction for other data modalities such as speech, videos and text as well.

In summary, we have proposed a privacy prediction system that addresses several drawbacks of the existing methods and outperforms them. We also demonstrate its workings through multiple configurations and examples.

References

[1] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, S. Nahavandi, Deep learning for deepfakes creation and detection: A survey, arXiv preprint arXiv:1909.11573 (2019).

[2] J. D. Cochran, S. A. Napshin, Deepfakes: awareness, concerns, and platform accountability, Cyberpsychology, Behavior, and Social Networking 24 (2021) 164–172.

[3] J. Botha, H. Pieterse, Fake news and deepfakes: A dangerous threat for 21st century information security, in: ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited, 2020, p. 57.

[4] K. A. O'Shea, Use of social media in employment: Should i fire? should i hire?, Cornell HR Review (2012).

[5] M. Madden, Privacy management on social media sites, Pew Internet Report 24 (2012) 1–20.

[6] S. Waters, J. Ackerman, Exploring privacy management on facebook: Motivations and perceived consequences of voluntary disclosure, Journal of Computer-Mediated Communication 17 (2011) 101–115.

[7] T. Orekondy, B. Schiele, M. Fritz, Towards a visual privacy advisor: Understanding and predicting privacy risks in images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[8] R. Gross, A. Acquisti, Information revelation and privacy in online social networks, in: Proceedings of the 2005 ACM workshop on Privacy in the electronic society, 2005, pp. 71–80.

[9] H. R. Lipford, A. Besmer, J. Watson, Understanding privacy settings in facebook with an audience view., UPSEC 8 (2008) 1–8.

[10] S. Zerr, S. Siersdorfer, J. Hare, Picalert! a system for privacy-aware image classification and retrieval, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 2710–2712.

[11] H. Zhong, A. C. Squicciarini, D. J. Miller, C. Caragea, A group-based personalized model for image privacy classification and labeling., in: IJCAI, volume 17, 2017, pp. 3952–3958.

[12] L. F. W. Anthony, B. Kanding, R. Selvan, Carbontracker: Tracking and predicting the carbon footprint of training deep learning models, arXiv preprint arXiv:2007.03051 (2020).

[13] A. Tonge, C. Caragea, Privacy prediction of images shared on social media sites using deep features, arXiv preprint arXiv:1510.08583 (2015).

[14] L. Tran, D. Kong, H. Jin, J. Liu, Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[15] G. Lowe, Sift-the scale invariant feature transform, Int. J 2 (2004) 2.

[16] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, C. Schmid, Evaluation of gist descriptors for web-scale image search, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 1–8.

[17] A. C. Squicciarini, D. Lin, S. Sundareswaran, J. Wede, Privacy policy inference of user-uploaded images on content sharing sites, IEEE transactions on knowledge and data engineering 27 (2014) 193–206.

[18] S. Zerr, S. Siersdorfer, J. Hare, E. Demidova, Privacy-aware image classification and search, in: Proceed-

ings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 35–44.

[19] A. Squicciarini, C. Caragea, R. Balakavi, Toward automated online photo privacy, ACM Transactions on the Web (TWEB) 11 (2017) 1–29.

[20] A. Tonge, C. Caragea, Dynamic deep multi-modal fusion for image privacy prediction, in: The World Wide Web Conference, 2019, pp. 1829–1840.

[21] A. Tonge, C. Caragea, Image privacy prediction using deep neural networks, ACM Transactions on the Web (TWEB) 14 (2020) 1–32.

[22] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[23] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[25] B. L. Happel, J. M. Murre, Design and evolution of modular neural network architectures, Neural networks 7 (1994) 985–1004.

[26] F. Azam, Biologically inspired modular neural networks, Virginia Polytechnic Institute and State University, 2000.

[27] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015) 91–99.

[28] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[29] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503.

[30] W. Wang, J. Yang, M. Chen, P. Wang, A light cnn for end-to-end car license plates detection and recognition, IEEE Access 7 (2019) 173875–173883.

[31] notAI.tech, Nudenet, https://github.com/notAI-tech/NudeNet, 2019.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[33] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[34] F. Hayes-Roth, Rule-based systems, Communications of the ACM 28 (1985) 921–932.

[35] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[36] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, C. Rudin, This looks like that: deep learning for interpretable image recognition, arXiv preprint arXiv:1806.10574 (2018).

[37] L. Du, H. Ling, Preservative license plate de-identification for privacy protection, in: 2011 International Conference on Document Analysis and Recognition, IEEE, 2011, pp. 468–472.

[38] J. Gao, L. Sun, M. Cai, Quantifying privacy vulnerability of individual mobility traces: a case study of license plate recognition data, Transportation research part C: emerging technologies 104 (2019) 78–94.

[39] J.-P. Hubaux, S. Capkun, J. Luo, The security and privacy of smart vehicles, IEEE Security & Privacy 2 (2004) 49–55.