1-2021

PrivAttNet: Predicting privacy risks in images using visual attention

Zhang CHEN
Institute for High Performance Computing

Thivya KANDAPPU
Singapore Management University, thivyak@smu.edu.sg

Vigneshwaran SUBBARAJU
Institute for High Performance Computing

2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

PrivAttNet: Predicting Privacy Risks in Images Using Visual Attention

Zhang Chen
IHPC, A*STAR
zhang_chen@ihpc.a-star.edu.sg

Thivya Kandappu
Singapore Management University
thivyak@smu.edu.sg

Vigneshwaran Subbaraju
IHPC, A*STAR
vigneshwaran_subbaraju@ihpc.a-star.edu.sg

***Abstract*—Visual privacy concerns associated with image sharing is a critical issue that need to be addressed to enable safe and lawful use of online social platforms. Users of social media platforms often suffer from no guidance in sharing sensitive images in public, and often face with social and legal consequences. Given the recent success of visual attention based deep learning methods in measuring abstract phenomena like image memorability, we are motivated to investigate whether visual attention based methods could be useful in measuring psycho-physical phenomena like "privacy sensitivity". In this paper we propose PrivAttNet – a visual attention based approach, that can be trained end-to-end to estimate the privacy sensitivity of images without explicitly detecting sensitive objects and attributes present in the image. We show that our PrivAttNet model outperforms various SOTA and baseline strategies – a 1.6 fold reduction in $L1-error$ over SOTA and 7%–10% improvement in Spearman-rank correlation between the predicted and ground truth sensitivity scores. Additionally, the attention maps from PrivAttNet are found to be useful in directing the users to the regions that are responsible for generating the privacy risk score.**

I. Introduction

The advent of resource efficient pervasive cameras, ranging from mobile phones and cameras to surveillance and wearable cameras, have enabled vast amounts of images being captured and shared in online social media platforms. Such ubiquity of these devices enables them to automatically capture and/or record a wide range sensitive information and inadvertently leaked by sharing them on social networks. To mitigate catastrophic consequences of such privacy leakage, we advocate that the users should be given means to control the privacy loss by making informed decision prior to sharing the content online based on the artefacts/attributes presented in the image.

Prior studies have proposed various computer vision techniques to detect sensitive objects and attributes in images: social relationships [25], face [26], gender [20], age [20], [5], occupation [22] and license plates [34]. Recent line of work [19] extended it to a much wider range of attributes: including (a) array of generic objects, such as credit cards, driver's license and home address and, (b) attributes that cannot be localized, e.g., age and religion. Prior works have focused on utilising traditional vision techniques, such as, pixel-wise segmentation [32] and object detection [19] techniques to detect and localise various scene elements in the images.

However, our goal of detecting a variety of sensitive attributes outlined in [19] has the following unique characteristics that may render the current techniques short-handed: (a) the images usually contain multiple objects with each of them associated with one or more labels (e.g., a driver's license can be associated with sensitive labels "face", "address" and "date of birth") (b) the multiple attributes present in an image may exhibit label correlation or inter-dependencies and (c) the multiple labels present in the image can lie anywhere, not necessarily in the foreground, hence, different parts of the image may have varying significance. To this end, we propose visual attention based privacy sensitivity estimation network to address our twin-goals: (a) given an image, predict its privacy sensitivity as a "score" based on the sensitive attributes present in the image, and (b) localise the attributes using soft heat-maps so that the user can visualise the sensitive regions.

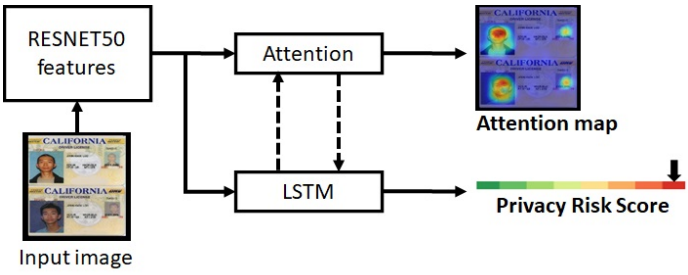


Fig. 1. Visual Attention Model To Predict Sensitivity Score of the Images

Attention mechanisms are extensively used in neural machine translations [4], image captioning [31] and visual question answering [21]. Visual attention mechanisms focus only on important regions of an image and ignore the redundancies, hence, achieving promising results on various challenging object detection tasks [17], [13]. Recently, attention mechanisms have been used to estimate memorability (the ability of human cognition to recall a visual content), by hypothesizing that salient objects can be linked with highly memorable visual content. Motivated by this, in this paper, we attempt to use a visual attention based deep neural network to predict a measure of the complex pyscho-physical phenomena of "*privacy sensitivity*" that is evoked in human subjects by images acquired in daily life. We believe that such a network will be useful in informing users about sensitive content when capturing, storing and sharing the images responsibly.

Research Questions and Contributions: We empirically investigate the following research questions:

- *Can visual attention based deep neural networks be used to measure privacy sensitivity of images?* Using a publicly available dataset of images with corresponding privacy sensitivity scores provided by human subjects, we propose and demonstrate that the use of a visual attention mechanism can enhance the ability of deep neural networks to estimate the human-provided privacy sensitivity scores of the images. We show that, the attention based PrivAttNet model outperforms state-of-the-art method proposed in [19] by 1.6 fold reduction in *L1-Error*.
- *Can attention mechanisms be used as an end-to-end trainable model to estimate sensitivity scores of the images without attempting to explicitly detecting/localising objects?* In this paper, we propose PrivAttNet , an end-to-end trainable network, that uses subjective sensitivity scores to learn "interesting" regions of the images and subsequently estimate the same for the new images. More specifically, we show that we are able to accurately estimate the privacy scores of the images – Spearman correlation between the human vs. machine estimated scores is 0.86. The attention-maps in turn can then be used to inform the user of the regions of the images that contributed to the predicted privacy risk score.

II. Related Work

A. Privacy and Computer Vision

The task of using image level features to predict whether an image would be considered *private* or *public* has been studied in the past. Zerr et al. [33] used simple image level features such as color histograms, faces, edge-direction coherence etc., along with user provided tags to perform this task using machine learning classifiers. However, user provided tags are exhibited to be noisy, hence, Tonge et al. [28], [27] proposed the use of deep features (with the advent of deep learning classifiers) and a fusion of object, scene context and image tags that are automatically generated by pre-trained models to predict whether an image is private or public. However, several tags generated by such pre-trained models are not abstract enough (e.g., religion, political affiliation, sexual orientation etc.) to capture the privacy sensitivity evoked by the images. Yu et al. [32] used convolutional neural networks for segmentation of privacy-sensitive objects in images and studied the relationship between the detected objects and privacy sensitivity. The recent work by Orekondy et al. [19] provided a large curated dataset of images (labelled VISPR) that contains 68 abstract as well as concrete attributes that are considered to be sensitive according to various privacy laws in force across the world. By conducting a human-subject study on their sensitivity towards each of those attributes, it is possible to assign a *privacy risk score* for them. Such a *risk score* can better capture the sensitivity of the users towards the content in the images than a simple binary classification into *private* vs. *public*. The authors further trained a deep learning based classifier that can perform the multi-label classification task of detecting the presence of the sensitive attributes in these images. However, hybrid CNN-RNN approaches [29] have recently shown good performance on multi-label classification tasks and therefore, it is possible that they can help in achieving better performance recognizing the multiple attributes present in the images from the VISPR dataset.

B. Private Attribute Recognition

Recognising Personally Identifiable Information (PII) in the images is an important task to understand the privacy sensitivity of the images. Ranging from detecting PIIs in the electronic documents [2] to email content [11], [6], prior works have focused on redacting attributes such as telephone numbers and address. Moving forward, various computer vision techniques are proposed to detect sensitive objects and attributes in images: social relationships [25], face [26], gender [20], age [20], [5], occupation [22] and license plates [34]. In contrast to detecting attributes, several works have focused on preserving privacy of the images: privacy enabled life-logging [14], adversarial perturbations [24] and person re-identification [16], [1]. In this paper, our primary challenge is to quantify the privacy sensitivity of the images without detecting and localising the attributes in the images.

C. Attention Networks

Visual attention based networks have been successful in producing enhanced performance in several computer vision tasks. For example, Rodriguez et al. [20] have used attention based deep neural networks for an enhanced performance in detecting the age and gender of a person appear in an image. Attention based deep learning based approaches have been extensively studied for sensory tasks such as image recognition [30], [15], [3], action recognition [23] as well as cognitive phenomena such as emotion recognition [10] etc. Recently, Fajtl et al. [9] have successfully used a CNN-RNN approach along with a soft-attention based mechanism (labelled AMNet) to measure the abstract phenomenon of image memorability. *Privacy sensitivity* is also an abstract psycho-physical phenomenon that is evoked when a person is presented with images as stimuli. Therefore, in this paper we investigate whether a similar CNN-RNN approach with a soft-attention mechanism would be successful in estimating the *privacy risk* of the content in the images provided by the VISPR dataset. However, unlike memorability estimation (where memorability is directly related to saliency), sensitivity of an image may arise due to non-salient artefacts present in the images as well.

III. Methodology

Our proposed architecture PrivAttNet is similar to the one detailed in [9] – it consists of four major components: (a) CNN network as a feature extractor, (b) a soft attention network, (c) recurrent neural network, and (d) regression network for sensitivity score estimation. In Fig. 4 we depict the architecture of our privacy sensitivity estimation model. We shall discuss each of the components in this section.

A. Dataset Description

In this paper, we make use of the publicly available dataset called *Visual Privacy (VISPR)* [19] – this is sourced from publicly available 22,167 Flickr[1] images containing 68 privacy sensitive attributes, carefully curated based on the following guidelines: (a) EU Data Protection Directive 95/46/EC, (b) US Privacy Act of 1974, and (c) personal data sharing rules from social media websites (e.g., Flickr and Twitter). The 68 privacy attributes capture a wide range of sensitivity – from *face, skin color, gender* being highest sensitive attributes to *hair color, eye color, and traditional clothing* being lowest sensitive attributes. Each image is annotated as a multi-label task. In Fig. 2, we plot the histogram of number of attributes present in each image in our training set.

B. Annotation of Privacy Scores

To quantify the sensitivity of an image, each attribute is given a likert-scale score – (1) Privacy is not violated (2) Privacy is slightly violated (3) Privacy is somewhat violated (4) Privacy is violated (5) Privacy is extremely violated, via a user study conducted in Amazon Mechanical Turk with 305 participants [19]. Fig. 2 shows the histogram of number of attributes present in a image – we see that more than 45% of the images contain at least 2 attributes in a single image, while a proportion of more than 10% of the images contain 10 or more attributes.

Each image is then given an aggregated privacy score as a *weighted sum* of individual attribute scores (hence, considering both attribute-level sensitivity and number of attributes in the image), which is then subsequently been min-max normalised across all the images in the training set.

In Fig. 3, we depict the histogram of the privacy risk scores across the images in the training set.

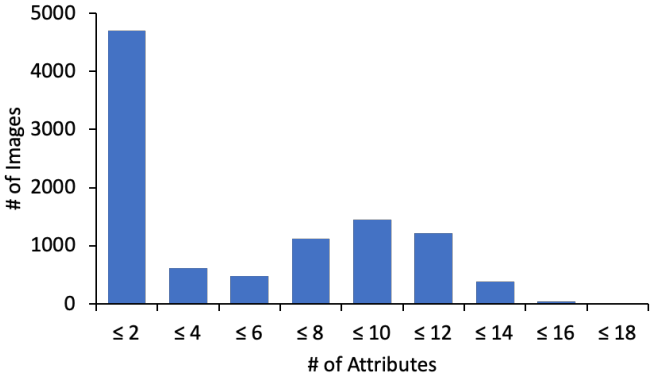


Fig. 2. Histogram of Privacy Attributes in the Training Dataset

C. Privacy Sensitivity Estimation

Our first goal is to estimate the privacy sensitivity score y of an image X based on the attributes present in it. For this task, we adopt the visual attention mechanism called AMNet proposed in [9] for memorability estimation. While AMNet hypothesizes that highly memorable visual content are linked with regions of the images that draw human attention (visual saliency), in our work, we test the efficacy of linking image regions with sensitive content (not necessarily the salient content) present in the image.

[1]https://www.flickr.com/

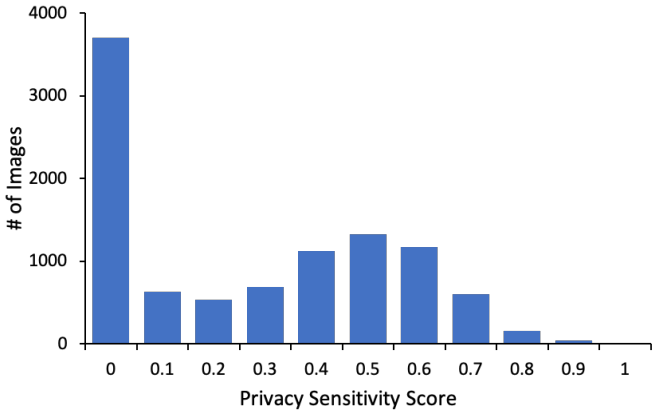


Fig. 3. Distribution of the modified privacy risk scores

Further, our task of predicting privacy score on VISPR dataset has previously been explored in [19]. The key difference is: our work estimates the sensitivity scores of the images without detecting the corresponding attributes – i.e., our training process does not involve attribute labels, instead, our model is trained using an aggregate privacy score per image.

D. Attention Mechanism

Similar to the approach of how human performs visual recognition, attention networks attend the most relevant regions of the input image [3]. We adopt the soft attention mechanism proposed in [4], [9] – instead of producing hard decision boundary, the network generates a probability weight for every visual region or element present in the image.

Our proposed PrivAttNet model aims to estimate the subjective privacy score y of an image X. In Fig. 4 we depict our proposed architecture – the first component is the deep CNN feature extractor that acts as an encoder, the second component is the visual soft-attention network, the third component is LSTM based RNN to preserve memory and the last component is the privacy score regressor.

We choose the soft-attention mechanism for the following reasons:

- Allows soft memory access – comes with the benefit that the network can be easily trained end-to-end using back propagation.
- Allows to visualize regions with potentially high sensitivity.

We use a CNN as a feature encoder of the attention model – we denote the image features extracted by a CNN with dimensions (W, H, D).

We denote the image features in step t by vector z_t: $z_t = \sum_i^L \alpha_{t,i} X_i$, where $\alpha_{t,i}$ is the attention weights/probabilities, expressed as a conditional probability on the feature vector X and LSTM hidden state of the previous step h_{t-1}. Subsequently, the attention is then represented as a softmax:

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{k=1}^{L} exp(e_t, k)} \tag{1}$$

where coefficient $e_{t,i}$ is a product of LSTM hidden state of previous step and the image feature vector: $e_{t,i} = f_{att}(X_t, h_{t-1})$. f_{att} is defined as follows:

$$f_{att}(X_t, h_{t-1}) = M_t \tanh\left(U h_{t-1} + K X_i + b\right) \tag{2}$$

where M, U *and* K *and* b are weights of the corresponding networks and biases.

Privacy score p is then computed at each step t:

$$p_t = f_p(h_t) \tag{3}$$

Fundamentally, $f_p()$ corroborates the LSTM hidden state h_t at step t to the privacy score. The final privacy score y is then computed as:

$$y = \sum_{t}^{T} p_t \tag{4}$$

The initial hidden state h_0 and memory state c_0 are derived from image features X as follows:

$$c_0 = f_{init_c}(\frac{1}{L}\sum_{i}^{L} X_i) \tag{5}$$

$$h_0 = f_{init_h}(\frac{1}{L}\sum_{i}^{L} X_i) \tag{6}$$

Where f_{init_c} and f_{init_h} are both single FC layers with tanh activation function.

E. Loss Function

Borrowing the custom-loss function from [9], we define the loss function as a combination of (a) mean squared error between the predicted vs. ground truth privacy scores and (b) a joint $l_1 - l_2$ penalty to control the network to pay attention on all the regions instead of focusing only on a few hot regions. The overall loss is the sum of two terms weighted by a parameter λ, which is chosen to be 10^{-4} experimentally [9].

$$L = (\hat{y} - y)^2 + \lambda L_\alpha \tag{7}$$

where L_α is defined as follows:

$$L_\alpha = \sum_{i}^{L} S_i^2 \tag{8}$$

where S_i is the l_1 penalty which enforces the attention module to focus more on a single region i along time dimension t:

$$S_i = 1 - \sum_{t}^{T} \alpha_{t,i} \tag{9}$$

IV. Experimental Evaluation

In this section, we present the performance of our proposed PrivAttNet model for estimating the privacy risk scores of images on the VISPR dataset [19]. We further elaborate the training process, baseline methods and evaluation metrics we considered.

A. Training Process

As stated in the section III, our PrivAttNet model is trained to minimize the loss function as a combination of (a) the mean squared error between the predicted and ground truth sensitivity scores, and (b) a joint *l1-l2* penalty to enable activations, one region at-a-time.

We adopted the same dataset split outlined in [19] – a random 45-20-35 split with 10,000 training, 4,167 validation and 8,000 test images.

B. SOTA & Baselines

In addition to the proposed PrivAttNet model, we propose 2 baseline strategies and compare against 2 SOTA models proposed in [19].

1) AP–PR & PR–CNN: Our goal of estimating the privacy scores is closely related to the models *Attribute Prediction-Based Privacy Risk* (AP-PR) and *Privacy Risk CNN* (PR-CNN) proposed and evaluated in the prior work by Orekondy et al.[19]. AP–PR first detects the presence of individual attributes, such as, face, handwriting, licence-plate etc., in the image using the fine-tuned ResNet-50 backbone network. Subsequently, AP–PR obtains the privacy risk score as the user-specific score of the most sensitive attribute in the image by using (a) the subjective privacy preferences provided by humans during the annotation user-study, and (b) attributes present in the image.

The PR–CNN on the other hand adds two additional fully connected layers to the pre-trained CNN for attribute prediction and fine-tunes the resultant network to directly predict the subjective privacy risk score of a given image.

2) PrivNet : To mimic AP-PR and PR-CNN models, we first propose PrivNet as a baseline – it first predicts the attributes present in the image and subsequently computes the user-specific privacy score of a given image. The PrivNet fundamentally differs from the SOTA methods in the following manner:

- It uses *Label-Powerset* method to transform the multi-label problem to a multi-class problem by training 1 multi-class classifier on all unique label combinations found in the training set.
- PrivNet then uses a more numerically stable Binary Cross-Entropy with Logits Loss (*BCEWithLogitsLoss*) by combining a Sigmoid layer and the BCELoss in one single class [2].

[2]https://pytorch.org/docs/master/generated/torch.nn.BCEWithLogitsLoss.html

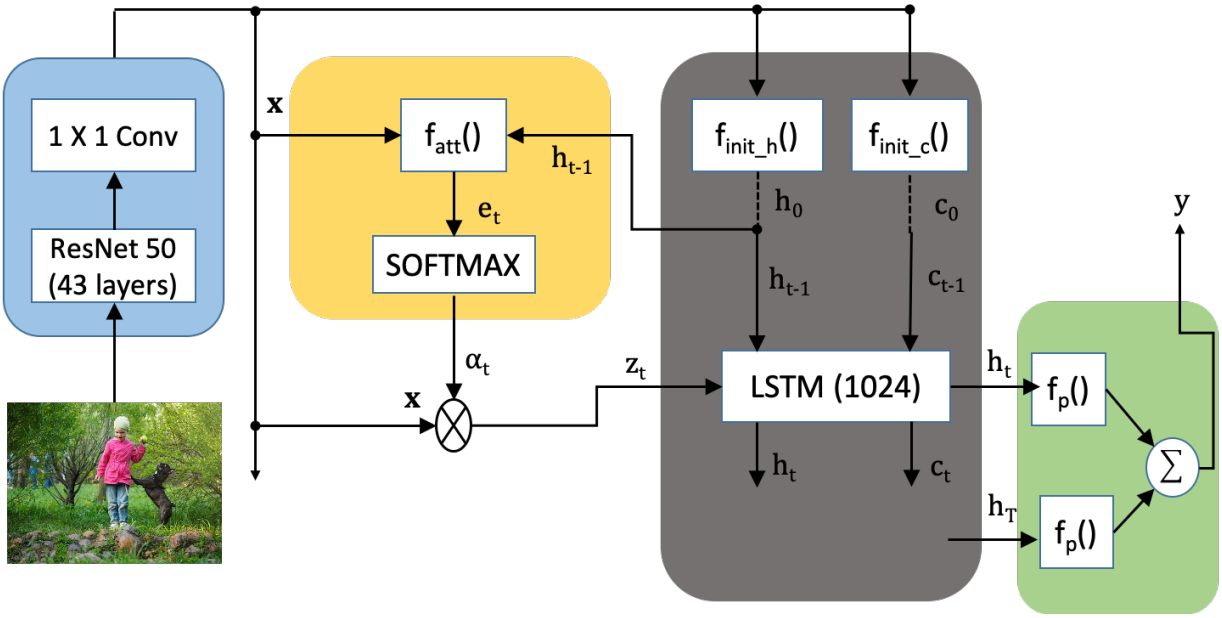


Fig. 4. PrivAttNet Model

3) PrivAttNet$_{MLC}$: To study the efficacy of soft-attention mechanisms on object detection, we propose PrivAttNet$_{MLC}$ that uses a visual attention model followed by binary classifier to detect multiple attributes.

We extend our proposed PrivAttNet by modifying the last block of our model as a multi-label classification problem. More specifically, we modify the mean square error component of our PrivAttNet loss function as a Binary cross-entropy (BCE) loss (we split our multi-label classification problem into 68 binary classification problems):

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} -[y_n.\log(y_n) + (1-y_n).\log(1-(y_n))] \tag{10}$$

C. Evaluation Metrics

To compare our proposed PrivAttNet model against SOTA and other baselines, we use the following metrics to capture various accuracy measures.

- **L1-Error:** denotes the mean value of the absolute error between the ground truth y and predicted privacy scores $\hat{y}$.
- **Correlation Coefficient:** represents the Spearman-rank and Pearson's correlation between the ground truth y and predicted privacy scores $\hat{y}$
- **C-MAP:** quantifies the mean average precision across all the unique classes available in the training set. More specifically, for the multi-label problems the C-MAP is calculated as the mean area under the precision-recall curves of individual classes.

D. Using PrivAttNet for estimating privacy risk of images and localizing sensitive attributes

In this section, we present performance evaluation of PrivAttNet . Note that our overall goal is to predict the privacy score of images and we show how accurately PrivAttNet can predict the privacy scores without detecting/localising the corresponding attributes present in the image.

1) Estimating Privacy Sensitivity Score: In Table I, we tabulate a comparison of the performance of PrivAttNet with respect to the baselines for the estimation of privacy risk in the images. In comparison with the AP-PR and PR-CNN methods given in the prior work by [19], PrivAttNet , PrivAttNet$_{MLC}$ and PrivNet are able to achieve a lower L1 error – improvement in L1 error of 59.2%, 44.7%, and 48.1% respectively. To further investigate the performance of PrivAttNet and its variants, we evaluated the correlation co-efficient between the predicted privacy risk score $\hat{y}$ and the user-provided privacy risk scores y (averaged across all 30 user profiles provided in [19]) of the images. We use the Spearman rank (ρ_s) and the Pearson correlation (ρ_s) co-efficient to quantify the monotonic relationship and show that PrivAttNet achieves $\rho_p = 0.87$ and $\rho_s = 0.84$ correlation – improvement of 10.5% and 7.7% in ρ_s over the proposed counterparts PrivAttNet$_{MLC}$ and PrivNet , respectively.

2) Localizing Sensitive Attributes: To understand the correlation between object saliency and privacy sensitivity, we use attention maps as heat maps to visualise the salient regions recognised by the PrivAttNet model. In Fig. 5, we show sample images from VISPR dataset – the top 2 rows (Fig. 5a) show the highly sensitive attributes while the bottom 2 rows (Fig. 5b and 5c) depict the attributes in small and cluttered scenes. After a thorough inspection, in Fig. 5a, we observe that heat maps highlight sharper focus on the regions

TABLE I
PERFORMANCE IN ESTIMATION OF PRIVACY RISK

Method	L1-Error	Correlation	
		ρ_p	ρ_s
AP-PR [19]	0.656	–	–
PR-CNN [19]	0.637	–	–
PrivAttNet	0.40	0.87	0.84
PrivAttNet$_{MLC}$	0.44	0.83	0.76
PrivNet	0.43	0.83	0.78

TABLE II
DETECTING PRIVACY SENSITIVE ATTRIBUTES.

Method	C-MAP	Precision	Recall	F1	Hamming loss
PrivNet	35.39	65.55	39.03	43.96	65.58
PrivAttNet$_{MLC}$	40.02	72.96	41.42	48.40	67.42

corresponding to credit card numbers, name, finger-prints, face, address etc. We then qualitatively investigate the attention maps on images where the sensitive attributes are found in not very salient parts of the images. For example Fig. 5b shows examples where the sensitive parts are found in a very small portion of the image, while Fig. 5c shows examples where sensitive attributes appear to be cluttered into various parts of the image. Even in such situations, we find that the network is able to display sharper focus on the features from the right regions.

It is worthy to note that PrivAttNet model generates false positives as well – in Fig. 6, we show a qualitative analysis of PrivAttNet . More specifically we show the instances where the generated attention maps (showed as heat maps) do not align with the sensitive attributes in the image.

3) Attribute Detection via Multi-label Classification: In this section we quantitatively measure the goodness of multi-label prediction of PrivNet and PrivAttNet$_{MLC}$ models. In Table II we show the performance by computing C-MAP, Precision, Recall, F1-score and Hamming loss. In comparison with PrivNet , the attention based PrivAttNet$_{MLC}$ is able to achieve higher precision and recall in detecting such sensitive attributes – 11.3% and 6.1% improvement in precision and recall, respectively. However, it should be noted that as evident from Table I, the performance of PrivAttNet is better than the performance of PrivAttNet$_{MLC}$ in terms of estimating the privacy risk score of images.

Since the PrivAttNet model is end-to-end trainable, it directly computes the privacy scores without detecting or localising the presence of sensitive attributes, hence, no output labels of the sensitive attributes.

The better performance of the attention-based PrivAttNet$_{MLC}$ shows that visual attention mechanisms can indeed help in detecting privacy sensitive attributes.

V. DISCUSSION

While we attempt to quantify the privacy sensitivity of the images, we identify a list of possible avenues to investigate.

A. The Correlation Between Privacy Sensitivity and Visual Saliency

Previous study in [9] primarily focused on linking memorability of an image to salient objects. However, fundamental differences between privacy sensitivity and visual saliency are not systematically studied yet.

More specifically, the influence of the attributes, such as its size, sentiment, semantics and watchable objects, on human attention, and subsequently on sensitivity should be established using empirical studies. Further, by using psychophysical fixation maps of the images and attention maps, the interplay between the sensitivity and saliency should be studied.

B. The Impact of Image Context on Privacy Sensitivity

Several works have shown that image context influences various visual cues of humans, such as, memorability [7], human attention [12], pose estimation [8], and object recognition [18]. It is important to understand the influence of image context on privacy sensitivity – the number of objects present in the image, their sentiments, and semantic categories are some of the context related features that can be considered to study their influence.

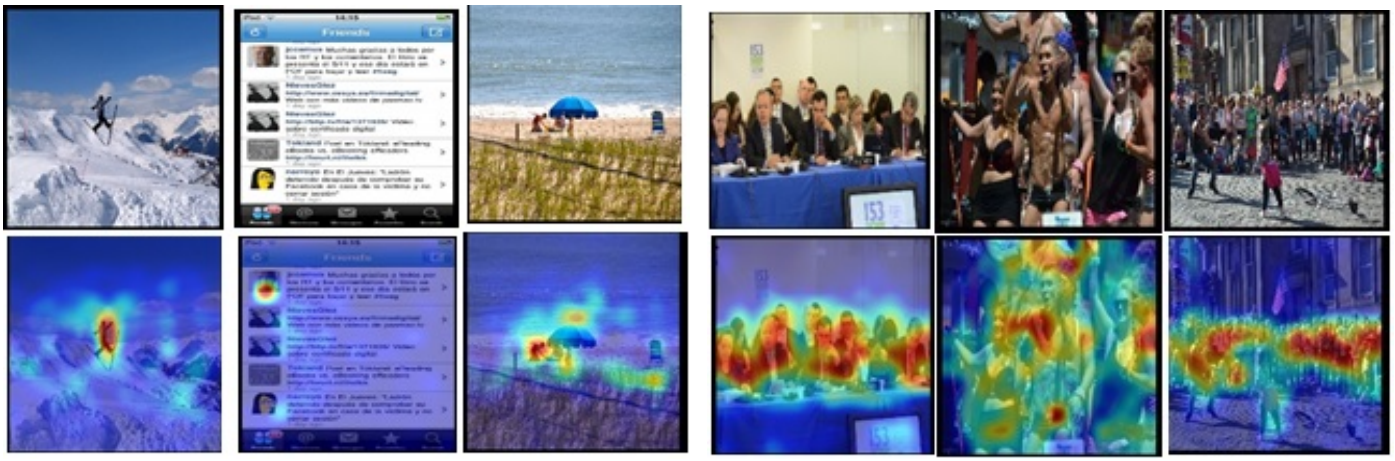
C. The Influence of Object Features on Privacy Sensitivity

Prior works have focused on various attributes of objects – such as low, mid and high level attributes – low: object texture, shape and color, mid: object size and complexity, and high: sentiment and semantics – to study the impact of such attributes on human attention. We are motivated to examine whether the conjecture holds while considering privacy sensitivity of images.

VI. CONCLUSION

In this paper we have developed and evaluated PrivAttNet – an attention based hybrid CNN-RNN approach, to automatically estimate the privacy risk of an image and inform the user about the potentially sensitive parts of the image. The performance of PrivAttNet was compared to the state-of-the-art using the publicly available VISPR dataset [19]. In contrast to the state-of-the-art (AP–PR and PR-CNN) PrivAttNet is able to be trained end-to-end to directly predict the privacy risk score of an image, by-passing the need to explicitly detect the presence of individual attributes. Our results show that PrivAttNet is able to achieve a significant 1.6 fold reduction in the *L1-error* when compared to the state-of-the-art. The privacy risk scores estimated by PrivAttNet are also found to be highly correlated with that of human assigned scores (Spearman rank correlation of 0.86). The attention-maps from PrivAttNet are found to be meaningful and can be used to highlight the regions that are responsible for generating the privacy risk score of an image. Our results show that visual attention mechanisms are indeed helpful in measuring an abstract psycho-physical phenomena like “privacy sensitivity”.

(a) Highly Sensitive



(b) Small (c) Cluttered

Fig. 5. Attention maps of PrivAttNet highlighting the sensitive regions of images.

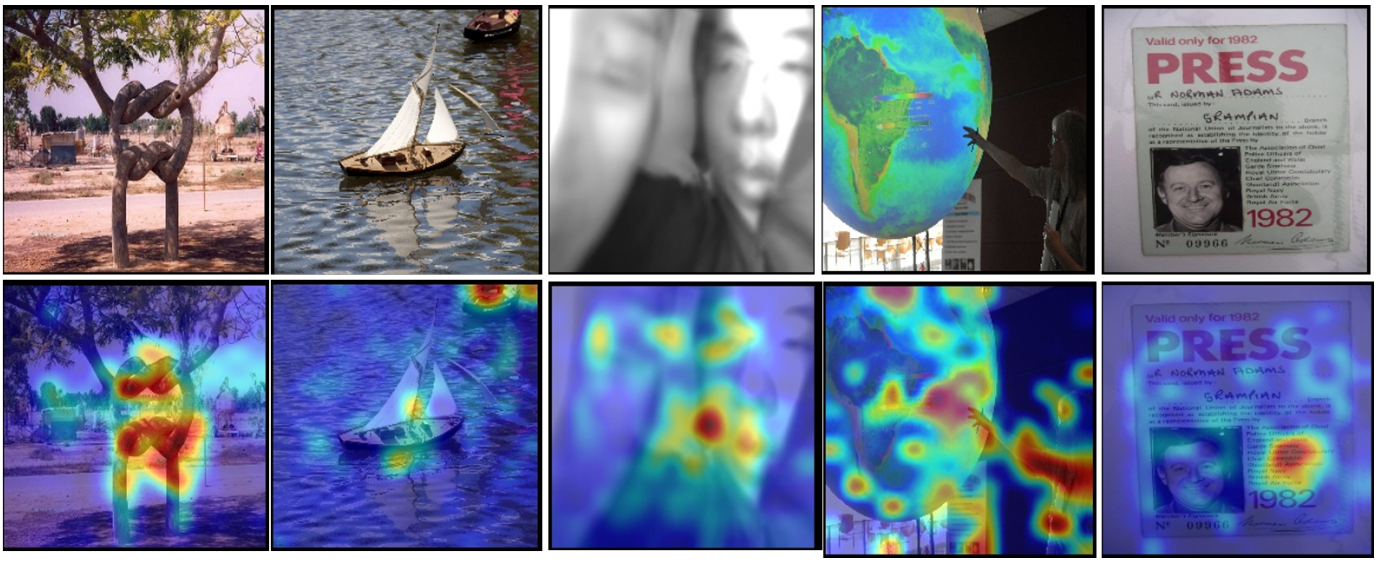


Fig. 6. False Positives By PrivAttNet – Qualitative Results

ACKNOWLEDGMENT

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3908–3916.

[2] T. Aura, T. A. Kuhn, and M. Roe, "Scanning electronic documents for personally identifiable information," in *Proceedings of the 5th ACM workshop on Privacy in electronic society*, 2006, pp. 41–50.

[3] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[5] C. Bauckhage, A. Jahanbekam, and C. Thurau, "Age recognition in the wild," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 392–395.

[6] C. Bier and J. Prior, "Detection and labeling of personal identifiable information in e-mails," in *IFIP International Information Security Conference*. Springer, 2014, pp. 351–358.

[7] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, vol. 116, pp. 165–178, 2015.

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[9] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "Amnet: Memorability estimation with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6363–6372.

[10] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 7521–7531.

[11] L. Geng, L. Korba, X. Wang, Y. Wang, H. Liu, and Y. You, "Using data mining methods to predict personally identifiable information in emails," in *International Conference on Advanced Data Mining and Applications*. Springer, 2008, pp. 272–281.

[12] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 1915–1926, 2011.

[13] K. Hara, M.-Y. Liu, O. Tuzel, and A.-m. Farahmand, "Attentional network for visual object detection," *arXiv preprint arXiv:1702.01478*, 2017.

[14] R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia, "Sensitive lifelogs: A privacy analysis of photos from wearable cameras," in *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, 2015, pp. 1645–1648.

[15] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1971–1981, 2019.

[16] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1325–1334.

[17] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[18] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, vol. 1. IEEE, 2003, pp. I–253.

[19] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3686–3695.

[20] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzàlez, "Age and gender recognition in the wild with deep attention," *Pattern Recognition*, vol. 72, pp. 563–571, 2017.

[21] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *Advances in Neural Information Processing Systems*, 2017, pp. 3664–3674.

[22] M. Shao, L. Li, and Y. Fu, "What do you do? occupation recognition in a photo via social context," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3631–3638.

[23] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[24] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5050–5059.

[25] Q. Sun, B. Schiele, and M. Fritz, "A domain based approach to social relation recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3481–3490.

[26] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

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

[28] A. K. Tonge and C. Caragea, "Image privacy prediction using deep features," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[29] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[30] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 464–472.

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[32] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1005–1016, 2016.

[33] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova, "Privacy-aware image classification and search," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 35–44.

[34] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Principal visual word discovery for automatic license plate detection," *IEEE transactions on image processing*, vol. 21, no. 9, pp. 4269–4279, 2012.