

Knowledge Distillation with BERT for Image Tag-Based Privacy Prediction

Chenye Zhao

University of Illinois at Chicago
Chicago, USA
czhao43@uic.edu

Cornelia Caragea

University of Illinois at Chicago
Chicago, USA
cornelia@uic.edu

Abstract

Text in the form of tags associated with online images is often informative for predicting private or sensitive content from images. When using privacy prediction systems running on social networking sites that decide whether each uploaded image should get posted or be protected, users may be reluctant to share real images that may reveal their identity, but may share image tags. In such cases, privacy-aware tags become good indicators of image privacy and can be utilized to generate privacy decisions. In this paper, our aim is to learn tag representations for images to improve tag-based image privacy prediction. To achieve this, we explore self-distillation with BERT, in which we utilize knowledge in the form of soft probability distributions (soft labels) from the teacher model to help with the training of the student model. Our approach effectively learns better tag representations with improved performance on private image identification and outperforms state-of-the-art models for this task. Moreover, we utilize the idea of knowledge distillation to improve tag representations in a semi-supervised learning task. Our semi-supervised approach with only 20% of annotated data achieves similar performance compared with its supervised learning counterpart. Last, we provide a comprehensive analysis to get a better understanding of our approach.

1 Introduction

With the rapid growth of the number of users on online social networking sites, image privacy has become a major concern (Ahern et al., 2007; Squicciarini et al., 2017). Users may accidentally disclose their sensitive information such as locations, habits or personal relationships from images that they post to their social networking sites (Squicciarini et al., 2017), which could be used in the detriment of the users (Tonge and Caragea, 2020).

Zerr et al. (2012a) defines private images as ones that belong to the private sphere (e.g., portraits, family, home) or capture sensitive contents that can not be shared with everyone on the Internet. The remaining images are considered to be public. Binary image privacy classifiers are developed (Tonge and Caragea, 2018; Yang et al., 2020; Zerr et al., 2012a) aiming to identify whether images belong to the public class or the private class. However, the access to the image content is not always allowed since users may be reluctant to share the real images (revealing user’s identity through the face, and friends, etc.) for visual content analysis. In such cases, tags attached by users to describe their images are found to be informative about the image contents and are good indicators of the privacy settings and improve the privacy prediction methods (Tonge et al., 2018). Privacy prediction models trained with image tags achieve competitive results compared with vision-based privacy prediction models (Squicciarini et al., 2017). Therefore, our goal is to learn good tag representations for images to further improve the performance of tag-based privacy prediction.

Pre-trained language models have been extensively studied in NLP communities (Howard and Ruder, 2018; Devlin et al., 2019; Peters et al., 2018; Liu et al., 2019). BERT (Devlin et al., 2019) is a pre-trained language model based on a multi-layer bidirectional Transformer, and has shown to be effective for generating universal language representations and attains state-of-the-art performance on many natural language processing tasks (Dai and Callan, 2019; Adhikari et al., 2019). In our work, we fine-tune BERT for the task of tag-based image privacy prediction to generate better tag representations. In addition, we propose to use self knowledge distillation with BERT (Hinton et al., 2015; Clark et al., 2019; Zhang and Sabuncu, 2020) to further improve the performance of tag-based

privacy prediction. Specifically, we first train a BERT (Devlin et al., 2019) teacher model for privacy prediction, and then train a BERT student model using both class labels and the pre-trained teacher model’s output probability distributions. The student model can thus learn from not only the ground truth class information, but also how to assign compatible probabilities according to various input examples. Experimental results show that knowledge distillation effectively improves tag representations and achieves boosted prediction performance.

Moreover, training a classifier often requires a large amount of annotated data. However, the annotation process is very time consuming and requires a significant human effort in many cases (Deng et al., 2009). Thus, we investigate knowledge distillation in a semi-supervised learning approach (Xie et al., 2020). To do this, we first train a BERT teacher model using limited amount of labeled image tags, and use it to annotate a large amount of unlabeled data, which is further used to train another BERT student model for privacy prediction. Experimental results show that our semi-supervised approach with BERT learns good tag representations and achieves comparable performance with its supervised counterpart with only 20% annotated image tags.

Last, we provide a comprehensive analysis for our tag-based privacy prediction. First, we perform a calibration analysis to show that models trained by improved tag representations with knowledge distillation yield better calibration (the alignment between prediction confidence and correctness likelihood (Guo et al., 2017)). Second, neural models are sensitive to small perturbation in the input and a small perturbation on the input may fool a well-trained neural network (Hsieh et al., 2019; Belinkov and Bisk, 2018; Niu et al., 2020). We analyze the robustness of our privacy classification models trained by tag representations learned with knowledge distillation against adversarial attacks. The results show that our approach shows the most robustness against adversarial attacks over compared baselines. Third, we perform a statistical analysis on the correlation between the privacy and sentiment and emotion of image tags.

2 Related Works

Knowledge distillation. Knowledge distillation is originally proposed as a model compression

method (Buciluă et al., 2006; Hinton et al., 2015). The standard knowledge distillation scheme transfers knowledge from a larger pre-trained “teacher” model to a smaller “student” model by training the student to mimic the class probability distributions generated by the teacher (Hinton et al., 2015). Recently, other works propose self-distillation (Furlanello et al., 2018; Clark et al., 2019): the teacher and the student have identical architectures, which achieve remarkable improvement on the student over the teacher. Zhang and Sabuncu (2020) experimentally demonstrate that the improvement of knowledge-distillation is correlated to the instance-level regularization on the student’s softmax outputs, meaning that by mimicking teacher’s probability distributions, instead of simply being trained to mimic one-hot class labels, the student are trained to assign compatible confidence (probabilities) according to the corresponding input examples. Meanwhile, one interesting focus of research on knowledge distillation has been on finding new applications. Chen et al. (2020) use the idea of knowledge distillation on BERT for text generation. Kim and Rush (2016) introduce knowledge distillation for sequence modeling. In contrast, we propose to utilize knowledge distillation as a tool to learn better tag representations for online images and to achieve improved performance for tag-based image privacy predictions.

Image privacy prediction. Most machine learning-based image privacy prediction models utilize images to train vision-based classification models to detect image privacy (Tran et al., 2016; Yang et al., 2020; Zerr et al., 2012b; Buschek et al., 2015). There are few works that adopt tags attached to describe images as indicators of image privacy and achieve competitive performance compared with vision-based approaches (Squicciarini et al., 2017). Further developing tag-based image privacy prediction approaches becomes a crucial direction for this task. Tonge and Caragea (2020) introduce TagCNN model based on the sentence classification CNN model (Kim, 2014) for image privacy prediction, where Word2Vec (Mikolov et al., 2013) is applied as the word embedding, and the CNN classifier is trained to predict image privacy. The bag-of-tags(BoT) model is introduced in (Tonge and Caragea, 2020) as another tag-based privacy prediction approach, where tags are embedded into multi-hot vectors similar to the bag-of-words embedding. Then a SVM classifier is trained for privacy detec-

tion. Previous tag-based works are trained using only the class labels. In contrast, we distill knowledge using BERT to utilize both hard class labels and soft probability distributions to improve tag representations and boost the performance of this task.

3 Methods

In this work, we adopt knowledge distillation with BERT to learn better tag representations for tag-based image privacy prediction. The idea behind knowledge distillation is that: soft probability distributions generated by a pre-trained image privacy prediction model carries additional privacy information compared with hard class labels. Specifically, hard labels can only reflect the class information (either private or public) of input image tags, while soft probability distributions can further reveal the confidence of the privacy classification model toward each prediction. A proper usage of such additional information, in combination with hard labels, can help learn better tag representations to boost the performance of tag-based image privacy prediction models. The goal of this work is to distill knowledge using BERT to transfer knowledge (in the form of soft probability distributions) from a strong, pre-trained BERT teacher model to a BERT student model to boost the performance of the latter for tag-based privacy prediction.

3.1 Knowledge Distillation with BERT

We first fine-tune a BERT (Devlin et al., 2019) as the teacher model for tag-based image privacy prediction. Given input image tags (x) for an online image, the teacher model generates a vector of scores, which is normalized to be the probability distribution of the two privacy classes: $P_T = \text{softmax}([p_T^{\text{pub}}(x), p_T^{\text{pri}}(x)])$. As the first teacher model is trained using hard-labels, we adopt an temperature term T (Hinton et al., 2015) to "soften" the probability distribution and avoid generating peaky probabilities: $P_T = \text{softmax}([p_T^{\text{pub}}(x), p_T^{\text{pri}}(x)]/T)$. The teacher model is then trained using the cross-entropy loss:

$$L_T = \text{CrossEntropy}(P_T, y) \quad (1)$$

After that, we perform knowledge distillation from the trained teacher model to the student model. As shown in Figure 1, our goal is to teach the student

model $\text{BERT}_{\text{student}}$ to learn from both soft probabilities (soft labels) generated by the trained teacher model, and the class labels (hard labels). Therefore in the total loss function of the student model, we need to minimize the difference between the student's predictions with both the ground truth hard label and the teacher's predictions. $\text{BERT}_{\text{teacher}}$ generates probability distributions P_T for input image tags x . The probability distribution generated by the student model $\text{BERT}_{\text{student}}$ is denoted as P_S . The training loss of the student model is the combination of the loss with the soft label P_T (L_{soft}) and the loss with the class label y (L_{hard}), where we use cross-entropy as loss functions. The above process can be denoted as:

$$L_{\text{soft}} = \text{CrossEntropy}(P_S, P_T) \quad (2)$$

$$L_{\text{hard}} = \text{CrossEntropy}(P_S, y) \quad (3)$$

$$L_S = \alpha * L_{\text{soft}} + \beta * L_{\text{hard}} \quad (4)$$

where α and β are hyperparameters.

3.2 Semi-Supervised Learning Approach with BERT

While using more labeled data improves performance, manually annotating privacy of online images is very time consuming and requires human intensive effort. This motivates us to distill knowledge with BERT in a semi-supervised manner, where a BERT teacher model is first trained using a small portion of labeled data. The trained teacher is used to annotate the large portion of unlabeled data. Next, we integrate the data annotated by the trained teacher model and the originally labeled data as the overall training set to train a student model. The trained student model becomes the next teacher model and repeats the process.

We used 50% of the whole training set D as the unlabeled set U and the rest set L is used to sample different fractions to be used as labeled data. In each experiment we randomly select $l = L * k' = D * k$, a subset of L , as the selected labeled set, which is used to train first teacher model. $k = 0.5 * k'$ is a fraction parameter ranging from [0%, 50%]. Our semi-supervised learning process can be concluded as follows:

1. Train the initial teacher model T_0 with the selected labeled set l .
2. Use the trained teacher model to annotate the unlabeled set U . Then integrate l and

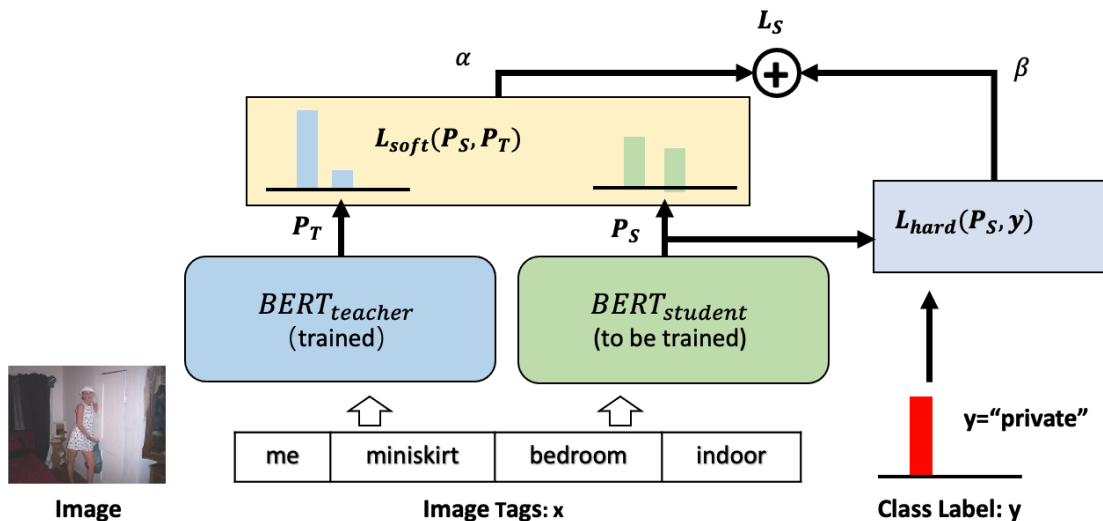


Figure 1: Illustration of knowledge distillation from BERT teacher model to BERT student model for tag-based image privacy prediction using image tags.

the annotated U as the annotated training set:
 $A = I \cup U$.

3. Train the student model S using A .
4. The student model S becomes the new teacher model T . Go back to step (2).

4 Experiments and Results

4.1 Experimental Settings

Dataset. In this work, we use a dataset of 32,000 examples from PicAlert (Zerr et al., 2012a), which, to our knowledge, is the only publicly available dataset for online image privacy prediction that captures real privacy needs of current social networks’ users. Images and tags of PicAlert are crawled from Flickr and manually annotated by external viewers, who are instructed to mark images as ”private” or ”public” following the guidance: private images are defined as images belonging to private sphere or ones you do not want to share with everyone, and the rest are public (Zerr et al., 2012a). The dataset is randomly split into train set (22000), validation set (5000) and test set (5000). The public and private images are in the ratio of 3:1 in each set. Each experiment is repeated five times using five train/validation/test splits and averaged as the final result. We delete special characters in user tags and replace tags with occurrences lower than 2 with the keyword “{UNKNOWN}” as they may bring noises to the classification model (Tonge and Caragea, 2020).

Model Configuration. All models are implemented based on python 3.6 and Pytorch 1.3.1.

For baseline models BoT and TagCNN, we apply the same hyper-parameters and network architectures suggested in (Tonge and Caragea, 2020). For BoT, we create a vector with the dimension of the vocabulary size and set 1 to the position of tags that exist in the image, and 0 otherwise (Tonge and Caragea, 2020). We fine-tune BERT with a learning rate of $2e^{-6}$ and the training batch size of 8. Hyper-parameters of BERT are selected on the validation set. We experiment different values weighting parameter (α, β) in Equation 4 and use (0.7,0.3) for BERT_{KD} as it shows the best performance. We also experiment with dynamically increase/decrease α and β along with the training process but they do not show better performance.

Research Questions. In our work, we aim to validate tag representations learned by distilling knowledge with BERT for tag-based image privacy prediction. We address the following research questions:

1. How does the performance of the BERT-based knowledge distillation approach compared with state-of-the-art models for supervised tag-based image privacy prediction?
2. What is the performance of our semi-supervised learning approach and how does it compare with its supervised learning counterparts?
3. Whether tag representations learned with knowledge distillation yield better calibration (the alignment between prediction confidence and correctness likelihood) of models?

Model	$F1_{private}$		$F1_{public}$		$F1_{overall}$	
BoT	0.613	± 0.02	0.901	± 0.004	0.831	± 0.005
TagCNN	0.629	± 0.02	0.903	± 0.005	0.839	± 0.008
BERT	0.664	± 0.014	0.906	± 0.003	0.849	± 0.005
TagCNN _{KD}	0.654	± 0.017	0.906	± 0.004	0.847	± 0.007
BERT _{KD}	0.681	± 0.01	0.907	± 0.003	0.855	± 0.005

Table 1: Results of knowledge distillation with BERT and state-of-the-art models (with standard deviation) for supervised tag-based privacy prediction.

4. Can tag representations learned with knowledge distillation show stronger robustness against some adversarial attacks toward image tags?
5. How does the privacy of images tags correlates with tag emotions?

4.2 Experimental Results

In this section, we discuss the experimental results designed to address the research questions. We measure the performance using the F1-score for each class as well as the weighted average F1-score over both classes (private and public).

4.2.1 Research Question 1

Table 1 shows the performance of the knowledge distillation approach and the state-of-art tag-based image privacy prediction models. To fine-tune BERT, we add a fully connected layer after the [CLS] token of the last BERT layer and fine-tune the whole network. For knowledge distillation, we use BERT as the teacher model to transfer knowledge to another BERT (denoted as BERT_{KD}) and TagCNN (denoted as TagCNN_{KD}), respectively. We observe that BERT_{KD} outperforms the state-of-the-art models on every compared metric especially on $F1_{private}$, yielding a large improvement upto 6.8%. We also notice that both knowledge distillation approaches effectively improve the $F1_{private}$ of corresponding student models. For example, TagCNN_{KD} improves TagCNN by 2.5%. BERT_{KD} boosts $F1_{private}$ of BERT by 1.7%, which further pushes BERT_{KD} to be the new state-of-the-art model for tag-based image privacy prediction. Such results suggest the effectiveness of our knowledge distillation approach of bringing the knowledge of the soft probability distributions generated by the teacher model to the training process of the student and achieves boosted performance. Note that, TagCNN_{KD} does not outperform its teacher, i.e., BERT. Which suggests that in our task, a more compact student model may

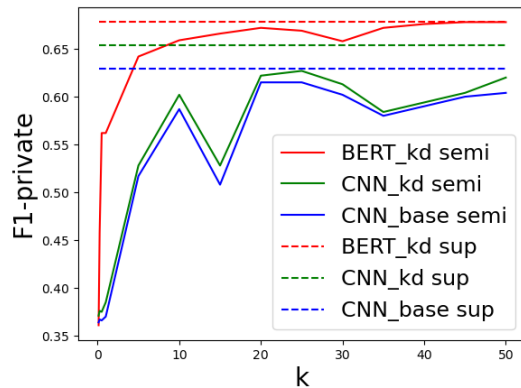


Figure 2: $F1_{private}$ for BERT_{KD} (red solid line), TagCNN_{KD} (green solid line) and TagCNN (black solid line) with the semi-supervised learning approach and their supervised learning counterparts (dashed lines with same corresponding colors) trained using varying percentage of overall training data (k).

not always outperform its teacher. Standard deviation results show that BERT-based approaches are more stable on compared metrics. Knowledge distillation can help generate more stable results.

4.2.2 Research Question 2

In our semi-supervised learning experiment, we use BERT as the initial teacher model and another BERT as the student model, denoted as BERT_{KD}^{semi}. We consider a baseline that use TagCNN as student model, denoted as TagCNN_{KD}^{semi}. Moreover, we also experiment with TagCNN playing the role of both T_0 and S , denoted as TagCNN_{base}^{semi}. Experiments are performed using different amount of labeled data L (controlled by the percentage parameter α). We randomly select α ranging from 0.25% to 100% of labeled data L . We repeat the student-teacher rotation 3 times and report the $F1_{private}$ of the student model in the last iteration. Results are shown in in Figure 2, where we observe some trends of $F1_{private}$. Firstly, BERT_{KD}^{semi} consistently show significant improvements over the two baseline approaches, yielding improvements upto 19.6%. Secondly, BERT_{KD}^{semi} at $\alpha = 20\%$ achieves comparable performance with its supervised learning counterpart BERT_{KD} trained with 100% labeled data, with only a small under-performance of 0.6%. In contrast, TagCNN_{KD}^{semi} and TagCNN_{base}^{semi} always perform much worse than their supervised learning counterparts. This encouraging result illustrates that our semi-supervised knowledge distillation approach with BERT can still be serviceable even

Model	CNN	CNN _{KD}	BERT	BERT _{KD}
ECE	2.30	2.19	11.07	8.45
ECE+TS	1.25	0.91	4.39	2.46

Table 2: Calibration results for TagCNN, TagCNN with knowledge distillation (CNN_{KD}), BERT, and BERT with knowledge distillation (BERT_{KD}). TS represents the results after using temperature scaling.

Model	$F1_{private}$	$F1_{public}$	$F1_{overall}$
Random Attack			
TagCNN	0.958	0.987	0.980
BERT	0.976	0.992	0.988
TagCNN _{KD}	0.966	0.988	0.983
BERT _{KD}	0.996	0.999	0.998
Synonym-based Attack			
TagCNN	0.875	0.961	0.939
BERT	0.939	0.980	0.970
TagCNN _{KD}	0.902	0.966	0.950
BERT _{KD}	0.962	0.986	0.980

Table 3: Results on knowledge distillation approaches with compared baselines against adversarial attacks for image tags.

when little labeled image tags are available.

4.2.3 Research Question 3

A well-calibrated classification model should be able to generate the probability of the predicted privacy class label (the confidence) which reflects its correctness likelihood (the accuracy) (Guo et al., 2017). In other words, a well-calibrated model should not only generate accurate predictions of image privacy, but should also “know what it does not know”, meaning that the model does not generate overly confident yet incorrect predictions. This is especially important for the task of tag-based image privacy prediction. For a privacy decision making system, so that if input image tags are misclassified to the wrong privacy class but with lower confidence, the system can pass the input example to the owner of the image to double-check as the privacy prediction model is not confident about its prediction (the privacy classification model “knows what it does not know”).

In this work, we first study the calibration of tag representations learned with TagCNN, TagCNN_{KD}, BERT, and BERT_{KD}. We then improve the calibration of compared models by performing post-hoc calibration for the predicted probabilities. Specifically, we adopt the temperature scaling (Desai and Durrett, 2020; Guo et al., 2017) to post-process model probabilities, where logits generated by compared models are divided by a

temperature scaling term T' , which is optimized with respect to the cross-entropy loss on the validation set. Note that temperature scaling does not affect the model’s accuracy. To evaluate the calibration of model predictions, we use the expected calibration error (ECE), which is defined as the weighted average of the difference between accuracy and confidence in m equally-partitioned confidence bins (Guo et al., 2017), where m is commonly selected to be 10. Results are shown in Table 2, where we observe that tag representations learned with knowledge distillation improves model calibration for both TagCNN and BERT: ECE scores are reduced by upto 2.62%, suggesting that soft-labels from the teacher model alleviate the overconfident issue caused by hard labels. Interestingly, we also notice that BERT exhibits higher ECE than TagCNN. This is because BERT has much higher learning capacity than TagCNN. During training, after BERT is trained to correctly classify almost all training samples, the model is able to further increase its confidence towards predictions to achieve lower training loss, while TagCNN can not perform such further optimization due to its limited learning capacity. Thus BERT achieves better prediction accuracy, but result in larger ECE (Guo et al., 2017) compared with TagCNN. We also observe that temperature scaling effectively calibrates both BERT and TagCNN with significantly reduced ECE (Guo et al., 2017).

We also plot the reliability diagrams (Guo et al., 2017; Desai and Durrett, 2020) in Figure 3 to better visualize the alignment between the accuracy and the confidence of compared models. The black dashed diagonal represents the optimal calibration when the accuracy always equals the confidence. We can observe that for both CNN and BERT, tag representations learned with knowledge distillation consistently bring model calibration closer to the optimal line. After temperature-scaling, calibration of all compared models are further optimized.

4.2.4 Research Question 4

We explore the robustness of tag representations learned with knowledge distillation model against two types of adversarial attacks (Hsieh et al., 2019). The goal of a success attack is to fool the model to give the false privacy prediction by replacing one tag in the original input image tags. In this work we consider two types of attacks.

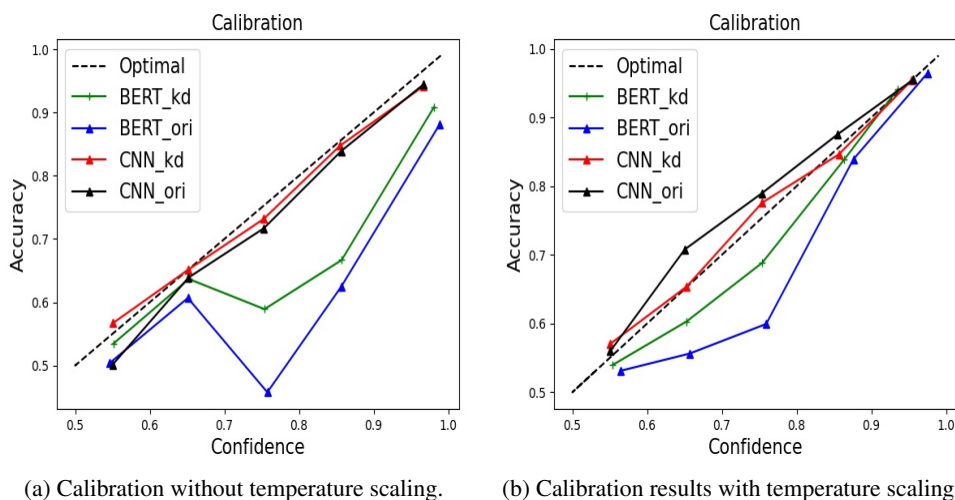


Figure 3: Calibration results of TagCNN, TagCNN_{KD}, BERT, and BERT_{KD}. Dashed line is the optimal calibration function.

Model	Trust	Surprise	Sadness	Joy	Fear	Disgust	Anticipation	Anger
Private	32.33%	4.72%	10.33%	25.06%	11.00%	8.99%	5.91%	1.65%
Public	29.42%	6.05%	9.95%	22.96%	17.00%	5.87%	6.35%	2.40%

Table 4: Emotion distributions of tags attached to private and public images.

Random Attack. This type of attack randomly select one image tag and replace it with another word that is randomly selected from the vocabulary of the dataset.

Synonym-based Attack Randomly selecting the word to replace from the vocabulary may change the meaning too much (e.g. replace "good" with "bad") which is not considered as good attacks (Hsieh et al., 2019; Niu et al., 2020). We explore the synonym-based attack: image tags are replaced by one of their synonyms. Particularly, for each image, we start by replacing the first tag with its synonyms. If none of the attacks successfully fool the model, we move to the next tag with the previous tag unchanged. This process is repeated until either the attack succeeds or all tags have been exhausted.

Experimental results addressing research question 3 are shown in Table 3. We evaluate the robustness of TagCNN, BERT, TagCNN_{KD}, and BERT_{KD} that have been well-trained for the supervised learning task in Section 4.2.1 against adversarial attacks. As suggested in (Hsieh et al., 2019), we randomly pick 100 examples from the test set that all models correctly predict, based on which we generate adversarial attacks. For random attacks, we repeat the process by 10^3 times and calculate the average as the final performance. For the synonym-based attack, all synonyms are

selected from WordNet. From Table 3 we can observe that tag representations with knowledge distillation approach improve the robustness of BERT and TagCNN against the two types of adversarial attacks, especially for the private class. Moreover, we also notice that TagCNN_{KD} does not show stronger robustness than its teacher model BERT, whereas BERT_{KD} outperforms BERT. This result further suggests the advantage of self-distillation on BERT.

4.2.5 Research Question 5

We perform analysis to study the correlation between the privacy (private or public) and the emotion of image tags. Precisely, we observe the distributions of tags with various emotions for both the private and the public class to understand whether tags with certain emotions are more often used in private or public images and the underlying reasons behind it. We use the NRC Emotion lexicon (Mohammad and Turney, 2013), a lexicon of 10,000 words, each is associated with one of the eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). In our experiment, we first randomly select the same number of private and public images (7000 for each class), and find common tags that exists in both image tags and NRC lexicon. The distribution of tags with eight emotions in both the private and the public class is shown in Table





Images				
Tags	innocent , girl, portrait kid	messy , room, indoor friends	statue, Andrew, marvel famous, poem	London, museum, wet floor, caution
Emotion	Trust	Disgust	Surprise	Fear
Privacy	Private	Private	Public	Public

Table 5: Examples of private and public images and corresponding tags associated with various emotions. Tags with specific emotions are colored in red.

Model	Positive	Negative
Private	60.71%	39.29%
Public	55.71%	44.29%

Table 6: Positive and negative emotion distributions of tags in private and public images.

4, where we observe that tags with emotions of trust, joy, and disgust are more often used to depict private images, while tags with fear and surprise emotions are more often attached to public images. Emotions of sadness, anticipation, and anger do not show obvious bias towards either privacy class.

Next, we look into some examples to better understand the underlying reasons behind such correlations. Examples of online images with tags of various emotions are shown in Table 5. We observe that the tag "innocent" with the emotion of trust is often used to depict images about children, which in many cases are considered as private images. Tags such as "messy" with the disgust emotion is often attached to images with indoor environments, which are ones that more often bias towards the private side. In contrast, tags such as "marvel" and "caution" with emotions of surprise and fear, respectively, are more often used to describe public constructions or signs, and thus are more often used in public images.

Moreover, we also analyze the distribution of tags with positive and negative sentiments for the private and the public class. Results are shown in Table 6, where we observe that tags with positive sentiments takes higher percentage in private images over public images.

5 Conclusion

In this work, we explore learning tag representations with knowledge distillation approach based

on BERT for tag-based image privacy prediction. Our approach significantly outperforms the state-of-art models for tag-based image privacy prediction. We also perform a BERT-based semi-supervised learning approach using only a small amount of annotated data, where BERT achieves comparable performance with its supervise counterpart with only 20% of labeled data provided. Moreover, we also perform calibration analysis and show that tag representations learned with knowledge distillation yield better calibration. We also study the robustness of our learned tag representations against some adversarial attacks for image tags. Our results show that our approach show stronger robustness over compared baselines against random and synonym-based attacks. Last, we analyze the correlation between the privacy and the emotion of image tags and use some examples in the PicAlert dataset (Zerr et al., 2012a) to help us understand the underlying reasons.

Our future direction is to integrate deep CNN models for image processing with BERT to develop a multi-modal image privacy prediction model with both images and tags as inputs.

Acknowledgments

This research is supported in part by NSF. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. The computing for this project was performed on AWS. We also thank our reviewers for their feedback.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

- Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 357–366.
- Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.
- Daniel Buschek, Moritz Bader, Emanuel von Zezschwitz, and Alexander De Luca. 2015. Automatic privacy classification of personal photos. In *IFIP Conference on Human-Computer Interaction*, pages 428–435. Springer.
- Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905, Online.
- Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. Bam! born-again multi-task networks for natural language understanding. In *ACL*.
- Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, Minneapolis, Minnesota. ACL.
- Tommaso Furlanello, Zachary Lipton, Michael Tschanen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. ACL.
- Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.
- Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

- Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward automated online photo privacy. *ACM TWEB*, 11(1):1–29.
- Ashwini Tonge and Cornelia Caragea. 2018. On the use of “deep” features for online image sharing. In *Companion Proceedings of the The Web Conference 2018*, pages 1317–1321.
- Ashwini Tonge and Cornelia Caragea. 2020. Image privacy prediction using deep neural networks. *ACM Trans. Web*, 14(2).
- Ashwini Tonge, Cornelia Caragea, and Anna Squicciarini. 2018. Privacy-aware tag recommendation for image sharing. In *Proceedings of the 29th on Hypertext and Social Media*, pages 52–56.
- Lam Tran, Deguang Kong, Hongxia Jin, and J. Liu. 2016. Privacy-cn: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *AAAI*.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.
- Guang Yang, Juan Cao, Zhineng Chen, Junbo Guo, and Jintao Li. 2020. Graph-based neural networks for explainable image privacy inference. *Pattern Recognition*, 105:107360.
- Sergej Zerr, Stefan Siersdorfer, and Jonathon Hare. 2012a. Picalert! a system for privacy-aware image classification and retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2710–2712.
- Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012b. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 35–44.
- Zhilu Zhang and Mert Sabuncu. 2020. Self-distillation as instance-specific label smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pages 2184–2195. Curran Associates, Inc.