Privacy-Preserving Visual Content Tagging using Graph Transformer Networks

Xuan-Son Vu[1], Duc-Trong Le[2], Christoffer Edlund[3], Lili Jiang[4], Hoang D. Nguyen[5]
[1,4]Department of Computing Science, Umeå University, Sweden
[2]Uni. of Engineering and Technology, Vietnam National University, Vietnam
[3]Corporate Research, Sartorius AG, Umeå, Sweden
[5]School of Computing Science, University of Glasgow, Singapore
{sonvx,lili.jiang}@cs.umu.se;trongld@vnu.edu.vn
christoffer.edlund@sartorius.com;Harry.Nguyen@glasgow.ac.uk

ABSTRACT

With the rapid growth of Internet media, content tagging has become an important topic with many multimedia understanding applications, including efficient organisation and search. Nevertheless, existing visual tagging approaches are susceptible to inherent privacy risks in which private information may be exposed unintentionally. The use of anonymisation and privacy-protection methods is desirable, but with the expense of task performance. Therefore, this paper proposes an end-to-end framework (SGTN) using Graph Transformer and Convolutional Networks to significantly improve classification and privacy preservation of visual data. Especially, we employ several mechanisms such as differential privacy based graph construction and noise-induced graph transformation to protect the privacy of knowledge graphs. Our approach unveils new state-of-the-art on MS-COCO dataset in various semi-supervised settings. In addition, we showcase a real experiment in the education domain to address the automation of sensitive document tagging. Experimental results show that our approach achieves an excellent balance of model accuracy and privacy preservation on both public and private datasets. Codes are available at https://github.com/ReML-AI/sgtn.

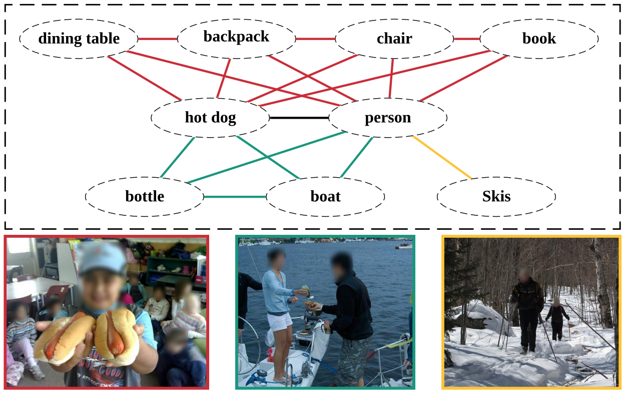
KEYWORDS

privacy-preservation, visual tagging, graph-transformer

ACM Reference Format:
Xuan-Son Vu, Duc-Trong Le, Christoffer Edlund, Lili Jiang, Hoang D. Nguyen. 2020. Privacy-Preserving Visual Content Tagging using Graph Transformer Networks. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414047

1 INTRODUCTION

The advent of smartphones and cloud services has led to the growth explosion of multimedia contents with the intertwinement of different types of information. Therefore, content tagging has become an increasingly important task in multimedia, computer vision, and information retrieval [30]. In 2015, one trillion photos were captured among a massive pool of multimedia documents [16]. As a result, it is imperative to automatically annotate visual objects with comprehensive textual semantics for accurate and efficient multimedia understanding and sharing. Nevertheless, this automated document annotation process is prone to inherent privacy risks; because the use of visual information typically conveys sensitive data to a certain degree. For example, personal information such as faces and license plates may be accidentally exposed in Web media.



backpack, hot dog, book, person, chair, dining table

hot dog, person, boat, bottle

person, skis

Figure 1: Knowledge graph built using object labels to model inter-object correlations. The graph typically depicts both common nodes (e.g., hot dog, dining table, and chair) and uncommon data patterns (e.g., hot dog and boat). Local correlations based on data-driven adjacency construction hence is susceptible to privacy attacks such as re-identification and link retrieval.

The key motivation of this paper is to develop an approach for visual content tagging, which has to be aware of privacy preservation with state-of-the-art performance. The early strategies for visual content tagging, including Scale-Invariant Feature Transform (SIFT) [24] or Histogram of Oriented Gradients (HOG) [9], are typically limited by hand-crafted concept representation. With the recent advancement in deep learning, multi-label classification using neural networks has been effectively used for image tagging [37] to achieve much better performance. Nonetheless, privacy issues need to be addressed at different levels, including sensitive visual information, associated multimedia semantics, and deep learning regime.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3414047

First, visual understanding tasks such as image tagging, facial recognition, or visual search entail the learning of patterns and representations, in which input data privacy plays a vital role in personal data protection. There have been many privacy incidents documented in the literature [10], in which the authors used a hill-climbing algorithm on the output probabilities of a computer-vision classifier to reveal individual faces from the training data. It is, therefore, intriguing to investigate a deep learning approach to perform multi-label tagging effectively on privacy-protected visual data. We apply a General Data Protection Regulation (GDPR) compliant method to obfuscate sensitive information such faces and plate numbers in images. This paper describes a multi-label visual classification to assign textual tags to censored inputs.

Second, as objects are typically co-occurred in visual data, the use of inter-object correlations in classification tasks has been explored to improve significant performance in visual classification tasks [4, 40]. We posit that local knowledge can be derived from data observations including label semantics or multimedia content semantics (e.g., optical character recognition); whereas, global knowledge can be drawn from publicly available corpora (e.g., Wikidata [35] or Common Crawl [5]). The local knowledge is often useful for knowledge graph construction and machine learning; however, it is prone to the disclosure of private data patterns. Figure 1 raises an interesting observation, in which uncommon correlations hint to a potential privacy breach. The co-occurrence of *person, chair, dining table*, and *book* may appear together in an intuitive way. On the other hand, *person, hot dog*, and *boat* is less observable in a dataset; hence, such a relationship may lead to re-identification of concerned objects. Furthermore, the combination of local correlations such as *person, skies*, and *hot dog* also enables the possibility of privacy attacks. Therefore, we propose several techniques including noise-added mechanism and differential privacy approach to protecting the use of inter-relationships among tagged objects.

Third, modelling the object dependencies, hence, is the core challenge in multi-label classification problems. One of the early approaches developed by Wang et al. [38] combined convolutional neural networks (CNN) with recurrent neural networks (RNN) [32] to learn the semantic relevance and dependency of multiple labels in order to boost the classification performance. Nevertheless, this approach is prone to the high computational cost and the sub-optimal reciprocity between visual and semantic information. In reality, objects are inter-connected which reflect as the network nature of object label dependencies. Kipf et al. [18] proposed semi-supervised learning on network data using graph convolutional network (GCN) unveiled spectral graph convolutions for classification tasks. The graph-based approach was adopted with visual data by Chen et al. [4] to get the state-of-the-art performance for multi-label image recognition. Furthermore, Li et al. [20] and [40] proposed several topological and architectural changes to enhance the learning capabilities with minor performance improvements. We propose a novel privacy-preserving graph transformer networks to achieve novel performance with our privacy-preserving mechanisms.

We apply our framework on the COCO dataset (MS-COCO) and an EU Education dataset (EDU-MM). Automating the task of classifying contents on arrival has a potential impact on saving thousands of labour hours and makes it more efficient for information processing. In education, application documents from students are very sensitive (e.g., passport, education records, education transcripts). Given the main task is building a good multi-label image classifier, one could argue that it did not necessary have to be aware of privacy. However, any algorithms running on personal data should be aware of the case, where the adversary observes outputs from the model to infer side knowledge regarding user information in the training data (e.g., membership attack [23]). In general, the same requirements would exist in other parties such as in hospital, finance department, and the like. Therefore, the requirement for having a kind of model that performs effectively the task and be aware of privacy preservation is in high demand.

Compared with existing visual content tagging studies, our proposed SGTN has the following contributions:

- We develop SGTN, a privacy-preserving visual tagging framework that leverages global knowledge to perform the visual tagging task with new state-of-the-art performances. Meanwhile, it uses less local information of the task to preserve user privacy by avoiding the use of sensitive information (e.g., faces, passport numbers, vehicle license plates).
- We propose two approaches to construct graph information from label embeddings with privacy guarantee under differential privacy theorem. These constructed graphs help SGTN avoid to use private sensitive information from local data.
- We evaluate the effectiveness of SGTN with comprehensive experiments on a public bench-marking dataset - i.e., MS-COCO, and a real-world education dataset with personal sensitive information.

The remainder of this paper is structured as follows. In Section 2, we discuss related work in visual classification, privacy-preserving graph. Section 3 presents our proposed neural architecture to address the issue that our education partner faced in the reality. In Section 4, we evaluate to show that SGTN performs effectively not only on private dataset EDU-MM but also on MS-COCO- i.e., the public benchmark dataset, and achieves new state-of-the-art results. Lastly, we conclude this paper in Section 5.

2 RELATED WORK

Privacy preservation is a complex topic and has been studied for decades. Among all requirements for privacy preservation, *the right to be left alone* is the most essential requirement. It is “the capacity of an individual or group to stop opinion about themselves from becoming known to people other than those they give the information to” [15]. To fulfil this requirement, to protect data donors from re-identification problem, any algorithms that run on personal data, must not give adversaries any chance to infer any side information by observing outputs of the algorithms. The techniques of anonymization [3] and sanitization [39] have been widely applied. Differential privacy later emerged as the key privacy guarantee by providing rigorous, statistical guarantees against any inference from an adversary [6]. Differential privacy has been applied in many research on different types of data including images [1, 42], network [26], text [29, 46], and general neural network architectures [28]. Therefore, it raises a potential need to consider differential privacy in algorithms that learn from personal data.

With the increasing use of graph-based techniques in multimedia research, privacy-preserving graph aims to create or modify graphs for privacy control based on graph statistics such as nodes, edge distribution, distance, subgraphs etc. The big challenge is its high sensitivity due to graph features (e.g., cluster coefficient). The survey [48] investigates a few studies on anonymisation techniques for privacy preserving publishing of social network data, especially graph modification approaches. They categorised the graph modification methods into three sub-categories: the optimisation configuration based approach [41], perturbation based modification approach [22], and greedy graph modification approach [47]. [41] generates privacy-preserving graphs for releasing by calibrating noise based on smooth sensitivity. They developed private dK-graph generation models that enforce rigorous differential privacy while preserving utility. [22] makes a trade-off of protection of sensitive weights of network links and some global structure utilities (e.g, the shortest path length) by applying two perturbation strategies on social network data. The authors in [47] addressed the l-diversity problem in social network data where they associated each vertex with some non-sensitive attributes and some sensitive attributes.

Multimedia tagging has been recognised as an interesting problem in computer vision research. With the rapid development of the Internet, online media is typically created with multiple tags to supplement visual data with semantic information. Early solutions for such classification task were developed based on the combinations of single-label classifications, which decomposed the task into multiple sub-problems for learning. Tsoumakas et al. [33] defined the multi-label nature of datasets and proposed the use of multiple classifiers. However, this approach ignored the inter-object correlations among various labels in visual data. Label co-occurrence dependencies were recognised as essential in multi-label classification problems [43]. Kipf et al. [18] proposed the encoding of graph structures using Graph Convolutional Networks (GCN) to learn representations for multi-label image classification [18]. Chen et al. (2019) employed this spectral graph convolution approach to model object label relationships for recognising multiple objects in images [4]. Knowledge such as semantic label embeddings and data-driven adjacency matrix have also effectively employed perform multi-label image tagging.

3 APPROACH

Visual content tagging is to generate descriptive textual comprehension on visual data. In computer vision, visual data often conveys meaningful relationships, where objects appear to be in correlated patterns. Recognising these patterns, therefore, lay the foundation for improving the tagging performance. Nevertheless, the exploit of object correlations is susceptible to privacy issues as such information may reflect the true nature or habitat of concerned objects.

We propose a novel approach that captures concurrently visual features and correlated semantic associations among objects under the privacy-preserving constraint. Inspired by Wang et al. [37], the visual content tagging task is formed as a multi-label classification problem. We develop an end-to-end privacy-preserving learning framework, which employs various neural network components to classify anonymised data inputs. Specifically, convolutional neural networks are utilised to extract visual features whilst graph transformer and graph convolutional networks are to exploit semantic and topological knowledge graphs of inter-correlated tags (i.e., labels). Next, we will thoroughly describe each component

3.1 Learning Architecture

Figure 2 illustrates the network architecture of our proposed model named SGTN for the multi-label classification task on a set of C tags. It is built upon three main components namely: (1) a graph transformer network (GTN), (2) a graph convolutional network (GCN); and a convolutional neural network (CNN).

Firstly, various inter-correlation views between labels, i.e., local and global knowledge, are transformed into privacy-preserved graphs in the form of a tensor $\mathbb{A}$ of multiple adjacency matrices (subsection 3.3). The tensor is fed into the graph transformer component (subsection 3.2) to leverage the most important connections, which are expressed via the representative adjacency matrix $\hat{A} \in \mathbb{R}^{C \times C}$:

$$\hat{A} = GTN(\mathbb{A}) \tag{1}$$

Subsequently, the matrix $\hat{A}$ is aggregated with a pre-trained embedding $\mathbb{E}$ (e.g., Glove) in the graph convolutional network component [18] to produce the privacy-preserving representation $\mathcal{W} \in \mathbb{R}^{C \times D}$ of the local and global information as follows:

$$\mathcal{W} = GCN(\hat{A}, \mathbb{E}) \tag{2}$$

Finally, $\mathcal{W}$ is fused with the visual representation extracted $\mathcal{F} \in \mathbb{R}^{D}$ from the convolution neural network component to generate tag prediction scores as: $\hat{y} = \mathcal{W}^T \mathcal{F}$.

The objective function is defined as follows:

$$\mathcal{L} = -\frac{1}{C} \sum_{c=1}^{C} y_c \log(\sigma(\hat{y}_c)) + (1 - y_c) \log(1 - \sigma(\hat{y}_c)) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, and y is the ground-truth vector.

3.2 Graph Transformer Network

The advantage of topological information is verified in improving the multi-label classification performance [4, 40]. Using a data-driven correlation matrix, the correlation among nodes is leveraged to favour the prediction of correlative labels. In these approaches, usefulness and privacy are but a screen away, especially for the case that the connectivity is exploited to violate people's privacy. Instead of using the data-driven matrix directly, Li et al. [20] construct the correlation matrix based on a global knowledge, i.e., pre-trained semantic embeddings of labels. Inspired by this idea, we seek to build the matrix by aggregating multiple pre-trained embeddings via Graph Transformer Networks [45].

Let us denote $\mathcal{E}$ as the set of pre-trained embeddings. For each embedding $\mathbf{E} \in \mathbb{R}^{C \times D_{\mathbf{E}}}$, we build the respective similarity matrix $S \in \mathbb{R}^{C \times C}$ with $S_{ij} = \cos(\mathbf{E}_i, \mathbf{E}_j)$; and an adjacency matrix $A \in \mathbb{R}^{C \times C}$, where $A_{ij} = 1$ if $S_{ij} \geq \tau$, the different of the mean and standard deviation of S's values, 0 otherwise. Subsequently, A is normalised as follows:

$$A_{ij} = \alpha * A_{ij} / \mathcal{D}_i \tag{4}$$

where $\mathcal{D}$ is the degree matrix ($\mathcal{D}_i = \sum_k A_{ik}$), and ϱ is α is 0.25.

The adjacency tensor $\mathbb{A} \in \mathbb{R}^{K \times C \times C}$ consists of K adjacency matrices, in which $\mathbb{A}_1$ is the identity matrix I, and the remaining is constructed as Eq(4) from the respective $(K - 1)$ embeddings.

Following to Yun et al. [45], the two softly chosen adjacency matrices $Q_1, Q_2 \in \mathbb{R}^{C \times C}$ are computed via two 1×1 convolutions

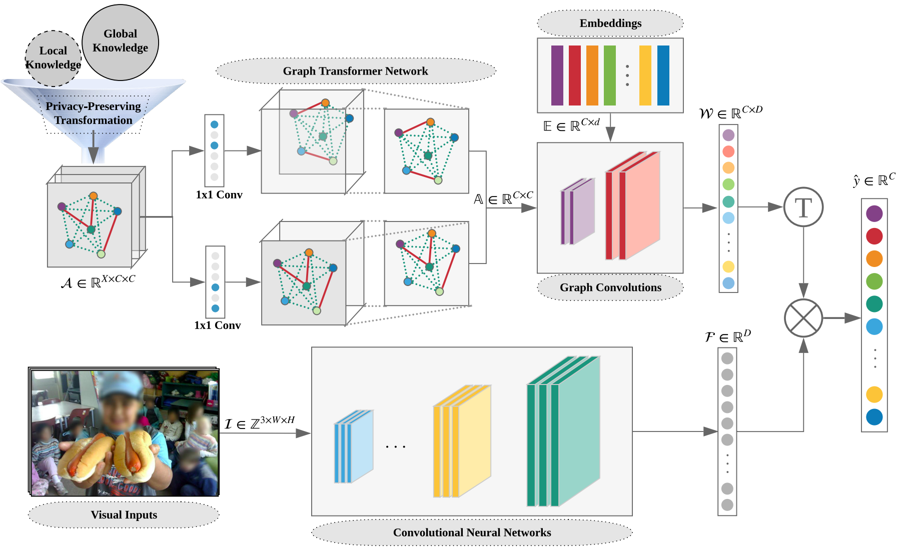

Local Knowledge
Global Knowledge
Privacy-Preserving Transformation
$\mathcal{A} \in \mathbb{R}^{X\times C\times C}$
Graph Transformer Network
1x1 Conv
1x1 Conv
Embeddings
$\mathbb{E} \in \mathbb{R}^{C\times d}$
$\mathbb{A} \in \mathbb{R}^{C\times C}$
Graph Convolutions
$\mathcal{W} \in \mathbb{R}^{C\times D}$
T
$\hat{y} \in \mathbb{R}^{C}$
$\mathcal{F} \in \mathbb{R}^{D}$
$\mathcal{I} \in \mathbb{Z}^{3\times W\times H}$
Visual Inputs
Convolutional Neural Networks

Figure 2: The network architecture of SGTN. It consists of (1) a graph transformer, (2) a graph convolutional network; (3) a convolution neural network (e.g., ResNeXt-50). The graph transformer enables global knowledge information by processing multiple adjacency matrices detailed in Figure 3, to enhance and guide the learning process for the visual classification task.

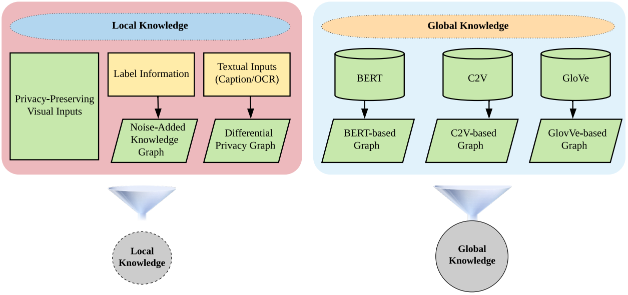


Figure 3: Local and global knowledge inputs of SGTN

as follows:

$$Q_1 = \psi(\mathbb{A}, \text{softmax}(W_\psi^1)) \tag{5}$$

$$Q_2 = \psi(\mathbb{A}, \text{softmax}(W_\psi^2)) \tag{6}$$

where ψ is the convolution layer, and $W_\psi^1, W_\psi^2 \in \mathbb{R}^{1\times 1\times K}$ are learning parameters. The final transformed matrix $\hat{A} \in \mathbb{R}^{C\times C}$ is by:

$$\hat{A} = \eta(Q_1 Q_2 + I) \tag{7}$$

where $\eta(A) = \mathcal{D}^{\frac{-1}{2}} A \mathcal{D}^{\frac{-1}{2}}$ is the Laplacian normalisation [18].

3.3 Privacy Preservation

The above classification model successfully discriminates between different classes using categorical information. However, user data is not directly protected within the model. For example, to differentiate a car from a motorbike, the model may memorise the numbers on the license plates of vehicles. Therefore, anonymising sensitive visual content is desirable, but with the expense of classification performance. Motivated by the challenge to achieve the trade off between privacy preservation and model accuracy, we present to apply **privacy-guaranteed label embeddings** to mask sensitive links (using differential privacy) to preserve privacy. Moreover, to leverage the local correlation information of the task without privacy leakage, we propose a **privacy-guaranteed graph construction** to leverage non-sensitive local knowledge for maintaining classification performance.

Label embeddings

To protect user privacy, we apply differentially private representations based on dpUGC [36]. The main intuition behind dpUGC is that, when the embedding is trained on sensitive text corpus, it injects noise to the word vectors to guarantee privacy at the highest level. Especially to address the common out-of- vocabulary (OOV) issue (i.e., a certain word might be missing from the pre-trained embeddings), dpUGC proposes character-level differential private embeddings. Thus, by applying dpUGC on the captions of MS-COCO dataset and the extracted texts of EDU-MM, we learn the differential private embeddings (dp-embeddings) for label representation of each dataset accordingly.

Let us denote the label set $C = \{l_1, l_2, \ldots, l_C\}$, which each label l_i might consist of multiple words $\{w_1, w_2, \ldots, w_k\}$. The representation of l_i is inferred as the mean vector of these word embedding vectors. Obviously, vec_{l_i} is also differential private due to any operation on the output of differentially private vectors (i.e., word-level vectors), its output is also differentially private [6].

Character-level dp-embeddings: As mentioned above that the out-of-vocabulary (OOV) issue is a common problem. In the case of EDU-MM dataset, it is simply because of the extracted text corpus is small and in multiple languages, hence, there is no representation for certain words in label names can be found after the training using dpUGC. Therefore, we introduce a character-level dp-embeddings to address the issue. Based on word-level embeddings,

Algorithm 1 Laplace Mechanism [8] for generating a differentially private adjacency matrix.

```
Require: Adjacency matrix $A \in \mathbb{R}^{C \times C}$, noise-level $\epsilon$.
Ensure: return a differentially private adjacency matrix $A_{dp}$
function DP_ADJACENCY_MATRIX($A$, $\epsilon$)
Initialize a zero matrix $A_{dp} \in \mathbb{R}^{C \times C}$
$\Delta = GS(A)$ ▷ Calculate the global sensitivity
for all $i \leftarrow 0, \ldots, C$ do
    for all $j \leftarrow 0, \ldots, C$ do
        $\tilde{y_i} \sim Lap(\frac{\Delta}{\epsilon})$ ▷ Get the noise based on the $\epsilon$ and sensitivity from the Laplace distribution
        $y_i = A(i, j)$
        $y_i \leftarrow y_i + \tilde{y_i}$ ▷ Add the noise to the weight.
        $A_{dp}(i, j) = y_i$
    end for
end for
return $A_{dp}$
end function
```

character-level embeddings can be easily calculated by averaging all vectors where a character occurred. Afterwards, vectors of missing words in a certain labels are calculated based on character-level dp-embeddings. Similarly to the word-level embedding, the averaging vector based on character-level embeddings also preserves the differentially private property.

Privacy preservation for graph construction

Most of data-driven methods try to learn as much information as possible from the data, which is the main cause of privacy leakage. Hence, we investigate into a different approach - i.e., leveraging global information to guide the optimisation process. The adjacency matrix in ML.GCN [4]'s variants is basically a graph to model the correlation between labels in the task. However, it might reveal sensitive information from the training data in case of unique links. Therefore, we propose Algorithm 1 to mask sensitive links in the adjacency matrix by injecting Laplace noise. Its effectiveness is further proofed in our experiments.

4 EXPERIMENTS

This section describes our experimental procedure, including implementation details and benchmarking metrics. A large number of experiments are investigated and we report the relevant empirical results on two datasets: MS-COCO (public) and EDU-MM (private).

4.1 Experiment Settings

The multi-label property has been seen in many publicly available datasets such as Microsoft COCO [21] or Fashion550K [14]. In this study, we seek to provide a fair comparison to the current state-of-the-art (e.g., ML.GCN [4]); thus, MS-COCO and EDU-MM datasets are selected for evaluation. asdfasdf

- MS-COCO dataset has been recognised as an important benchmark datasets with multiple features such as object segmentation, recognition in context, and captions. It consists of 82,783 training, 40,504 validation, and 40,775 test images. We tested on two versions of COCO dataset: (1) regular one without anonymization (i.e., MS-COCO) and (2) PP-MS-COCO- an anonymized version of the MS-COCO dataset, in which images having faces and license plates of vehicles are blurred using detection algorithms.

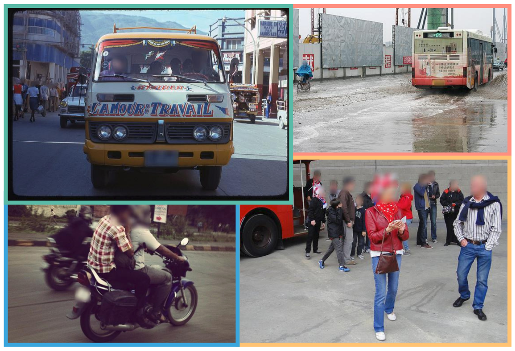


Figure 4: Examples of anonymised images, where faces and license plates were blurred in PP-MS-COCO.

- EDU-MM dataset: the education dataset from an education partner consists of 130,362 images in 23 different categories of document types. The used documents came from applications submitted by students applying for postgraduate programmes in an EU country. It contains a great variety of documents, ranging from ID documents to academic merits, curriculum vitae (CV), professional certification, and proof of proficiency in languages. The proof of proficiency in languages is often in the form of proofs of passing language tests, such as the International English Language Testing System (IELTS). The documents are protected under the General Data Protection Regulation (GDPR) and cannot be made public or shared. Therefore, all experiments were performed within the originated infrastructure of the education partner. We split the EDU-MM dataset into subsets of 20% for testing and 80% for training (using stratified selection on labels [25]). In numbers, it has 104,290 images for training, 26,072 images for testing.

Preprocessing

Table 1: Data statistics of PP-MS-COCO created from MS-COCO by removing sensitive visual contents (e.g., faces).

Dataset	Set	#images	Anonymised Ratio
PP-MS-COCO	Training	82,783	52,043 (62.9%)
PP-MS-COCO	Validation	40,504	25,299 (62.5%)

- MS-COCO: Removing sensitive visual features from images: face and id numbers (e.g., id on passport or plate number of vehicles) via pre-trained models provided by [34].
- EDU-MM dataset: In order to retrieve text features from documents, the Optical character Recognition (OCR) program called Tesseract [31] is used together with some preprocessing of the image, such as thresholding to reduce noise. These extracted texts are then being used to train a differentially private embedding.

Pre-trained embeddings for label representation

There is a number of pre-trained embeddings which were trained on public corpus such as Wikipedia or Common Crawl (commoncrawl.org). These text corpuses capture the semantic meaning of the global knowledge. Here we investigated into four different models including (1) GloVe, (2) Bert, (3) Char2Vec, and (4) dpUGC.

- **GloVe** [27] stands for "Global Vectors", it captures both global statistics and local statistics of a corpus, in order to learn word vectors. GloVe has been used in ML.GCN [4], therefore, we also use it to extract label embeddings for our proposed model.
- **Char2Vec** [17] is a neural language model, which relies only on character-level inputs. It employs a convolutional neural network (CNN) [19] and a highway network over characters. Then the output is given to a long short-term memory (LSTM) [13] recurrent neural network language model (RNN-LM). After training on a large text corpus, it has the ability to deal with the texts containing abbreviations, slang, words with unusual symbols and the like. In this work, the Char2Vec model was trained on English Wikipedia corpus with embedding dimension of 300.
- **dpUGC** [36] is a differentially private word embedding (dp-embedding) used for learning word representation of sensitive datasets such as medical records, or in this case are recognised texts from document images (e.g., education records, passport) of student applications.
- **BERT** [7] makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. The Transformer encoder reads the entire sequence of words simultaneously, therefore, it allows the model to learn the context of a word based on all of its surroundings. Here we use BERT_Base pre-trained model. To get the label embeddings, for a given label, we average all vectors of its subwords from the last layer provided by Akbik et al. [2]. Regarding Bert-Finetune, we reload the pre-trained weights of Bert-Base, and add a softmax layer for the text classification task on 80 categories of the COCO dataset. Then we run the finetune for 4 epochs to have a fine-tuned language model (i.e., Bert-Ftune) specifically for the MS-COCO dataset. It is noted that we only use captions in the training data of the MS-COCO dataset for this fine-tuning process. Our tendency in this work is to avoid the use of data-driven information, which is Bert-Finetune model in this case. Therefore, Bert-Ftune is only used as a comparison to see the differences in the signals of multiple adjacency matrices based on different language models.

Implementation

Our proposed SGTN framework is developed using PyTorch (version 1.3.1). We employ a ResNeXt-50 backbone [12] for visual feature extraction with a semi-weakly supervised pre-trained model on ImageNet [44]. The concentration of visual presentations amounts to a tensor $\mathcal{F}$ of 2048 features.

For data augmentation, we adopt the same approach from Chen et al. [4] and Wang et al. [40] as follows. Firstly, all input images are resized to 512 × 512 and randomly cropped regions of 448 × 448 with random horizontal flips. SGD optimiser is used with the momentum of 0.9. Weight decay is 10^{-4}. The learning rate is 0.03 for all datasets. For all experiments, we only run 80 epochs in total without fine tuning learning rate. The experiments were run on an Nvidia Titan RTX 24GB and Tesla V100 32GB for MS-COCO and EDU-MM datasets, respectively. It is noted that, the experimental results can also be reproduced on less memory GPUs. The two given GPUs were used because of their availability, not because of their high memory capacity. In fact, our proposed model has less trainable parameters in comparison to ML.GCN [4].

Evaluation metrics: this paper employs the mean average precision (mAP), average per-class precision (CP), recall (CR), per-class F1 (CF1), average overall precision (OP), overall recall (OR), and the overall F1 (OF1) for benchmarking with the most recent state-of-the-art models [4, 40].

5 RESULTS AND DISCUSSIONS

This section presents our comparisons with the existing state-of-the-arts on MS-COCO to show the effectiveness of the proposed approach for the multi-label classification task. We then present the performance that the proposed model was applied to solve the given issue of the education partner in *an anonymous* European country (i.e., EDU-MM dataset).

Classification performance

We tested our approach with several settings as shown in Table 4. Our Graph Transformer and Convolutional Networks work as desired to produce significant results on the MS-COCO dataset. In the original datasets, the tendency of using global knowledge has superior impact compared to the utilisation of local correlations. The noisy-induced graph transformation has shown some advantages over other models. Most importantly, our differential privacy graph construction (based on dpUGC) has achieved significant results in comparison to other settings.

In details, our approach outperformed the state-of-the-art techniques of multi-label image classification. Table 2 demonstrates the significant improvements of 9.3% and 4.2% compared to the baseline and ML.GCN respectively.

Comparison of ML.GCN and SGTN on PP-MS-COCO. In Table 2, it is obvious that the precision has been improved while the recall has been decreased due to the lack of local knowledge; It hints that by removing sensitive visual information from the data, the model was forced to learn other information (e.g., size and shape of objects, instead of detailed but sensitive features). However, due to the lacks of sensitive but unique features (e.g., license plates), it has lower recall.

Performance in comparison on both PP-MS-COCO and MS-COCO datasets. For privacy-preserving, we propose the use of global knowledge; therefore, it is a clear trend that the recall has been much improved while the precision has been decreased due to the lack of local knowledge. In numbers, it is actually in reverse: precision gets higher and recall gets lower, see Table 2. This observation supports our novel idea to reduce uncommon inter-object links, which would potentially lead to privacy breach.

Performance on EDU-MM dataset. For automated document classification, we applied our model on EDU-MM. In both original and anonymised datasets, we observe the adequate improvements compared to ML-GCN. It is important to note that our model is lighter and does not use the data-driven local correlations. The private information in our graph convolutional networks, therefore, is preserved with multiple privacy preservation mechanisms.

Privacy preservation

Taking privacy preservation strategies under consideration, we reveal the following findings with qualitative analysis.

Table 2: Performance comparisons on MS-COCO. SGTN outperforms baselines with large margins. PP denotes the use of anonymised MS-COCO dataset.

Method	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN [38]	61.2	-	-	-	-	-	-
SRN [49]	77.1	81.6	65.4	71.2	82.7	69.9	75.8
Baseline(ResNet101) [12]	77.3	80.2	66.7	72.8	83.9	70.8	76.8
Multi-Evidence [11]	–	80.4	70.2	74.9	85.2	72.5	78.4
ML-GCN [4]	82.4	**84.4**	71.4	77.4	**85.8**	74.5	**79.8**
SGTN	**86.6**	77.2	**82.2**	**79.6**	76.0	**82.6**	77.2
ML-GCN [4] (PP)	80.3	84.6	68.1	75.5	85.2	72.4	78.3
SGTN (PP)	**85.6**	**85.3**	**75.3**	**79.9**	**85.3**	**78.7**	**81.8**

Table 3: Performance comparisons on EDU-MM. PP denotes the use of anonymised version of EDU-MM dataset, in which faces, ID numbers were censored to protect user privacy.

Method	mAP	CP	CR	CF1	OP	OR	OF1
ML-GCN [4]	66.06	**73.91**	58.32	65.19	**78.03**	65.33	71.12
SGTN	**66.70**	73.89	**61.59**	**67.18**	74.07	**70.63**	**72.31**
ML-GCN [4] (PP)	66.52	**74.51**	57.08	64.64	**80.25**	64.72	**71.65**
SGTN (PP)	**66.60**	67.55	**61.86**	64.58	74.54	**68.51**	71.40

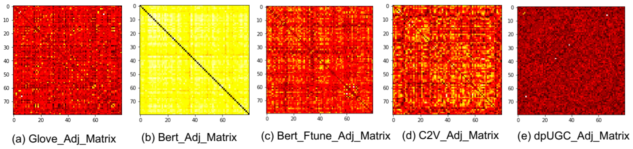
Here global knowledge is considered as public knowledge which does not contain personal information, since the models (Glove, Bert, C2V) were trained on, e.g., Wikidata [35] or Common Crawl [5]. In Table 4, experiment#2 clearly shows that using the global knowledge, SGTN can achieve better performance than ML.GCN (as shown in Table 2) in terms of mAP scores from 4.14% to 5.19% for MS-COCO and PP-MS-COCO respectively.

Given the fact that, one only takes the use of a privacy-guaranteed information when it can help the task achieve better performance. Otherwise, one might decide to not use the information at all. In Table 4, experiment#4 actually shows that, the performance of SGTN is the highest among different settings on both MS-COCO and PP-MS-COCO datasets. The experiment shows that the use of local knowledge with privacy guarantee is a good strategy for incorporating sensitive information to boost the performance. Because in many downstream tasks, global knowledge from public corpora might not always exist (e.g., medical data of patients).

Performance between privacy guaranteed adjacency matrix (dpUGC-based) versus noisy adjacency matrix. Table 4 shows the comparison results between experiment#3 and experiment#4. With privacy guarantee at the level of ($\epsilon = 0.125, \delta = 0.81$)-dp, SGTN has the best performance in comparison to others, including the noisy setting in experiment#3. However, the noisy setting has its own benefit in the case of private text corpus does not exist. Then Algorithm 1 can be applied to protect privacy for the adjacency matrix, while maintaining a good performance.

Investigation to different adjacency matrices

SGTN enables global knowledge being the guidance for performing the downstream task via graph transformer. Therefore, we investigate into the adjacency matrices to see the similarity of signals between adjacency matrices created using different language models. Figure 5 shows the heatmap of 5 different adjacency matrices.



(a) Glove_Adj_Matrix (b) Bert_Adj_Matrix (c) Bert_Ftune_Adj_Matrix (d) C2V_Adj_Matrix (e) dpUGC_Adj_Matrix

Figure 5: Heatmap of adjacency matrices for MS-COCO based on different pre-trained embeddings. Bert_Ftune is a fine-tuned variant of the pre-trained Bert model on the text classification task with MS-COCO image captions. The Bert_Ftune-based adjacency matrix is included as a reference only, and not used for the learning process of SGTN due to the use of local information of the task.

The Bert_Ftune_Adj_Matrix is used as a representative standard for using local knowledge from the training data. Here we have some interesting findings by observing the signals:

- Given the fact that different pre-trained word embeddings were trained on different public corpus, the according adjacency matrices between them are significantly different. By introducing the graph transformer and the graph convolutional network in SGTN, we can incorporate these signals to guide the learning process of the task.
- The adjacency matrices (a), (c), and (d) of GloVe, Bert_Ftune, and C2V possess similar signals. Here the global knowledge preserved in the adjacency matrices from Bert and C2V is in fact, similar to the local knowledge, i.e., the adjacency matric from Bert_Ftune.
- The pre-trained embedding of dpUGC preserves good tradeoff signals from the training data while guaranteeing data privacy at ($\epsilon = 0.125, \delta = 0.81$)-dp. In fact, using dpUGC helps boost the performance of the task ranked highest among all settings.

Performance analysis

Figure 6 shows the results in comparison of our proposed approach to ML.GCN on MS-COCO and PP-MS-COCO. It presents the effectiveness of SGTN in terms of leveraging global knowledge to classify anonymised images. We have the following insights.

(1) The improvement of SGTN in comparison to ML.GCN is more significant on MS-COCO than that of PP-MS-COCO. Especially, on the PP-MS-COCO, the degradation is higher. It suggests that when the sensitive visual features were censored, it affects the precision of the model. However, in general, overall performance of SGTN is higher thanks to the global knowledge embedded in multiple adjacency matrices (empowered by label embeddings).

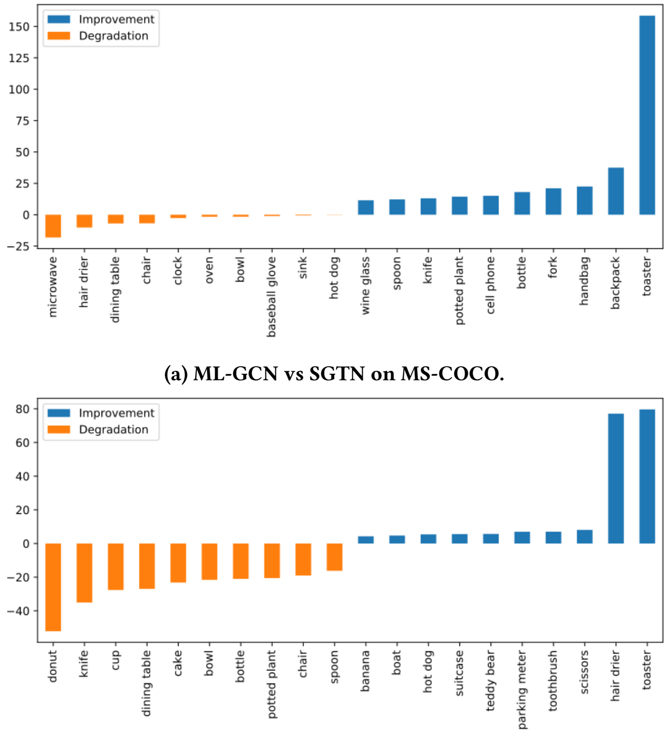
(2) Regarding the top of degradation labels, SGTN can help to reduce the sensitivity of labels that are highly related to sensitive visual features. In fact, *baseball glove* is no-longer on the degradation list, *chair* also reduces the rank from top-4 to top-9. For the case of *donut*, statistic of the dataset reflects that *donut* has a strong correlation with *person.* When the sensitive visual features got censored (i.e., faces), it reduces the accuracy on the label *person* and its related labels, which include *donut* and most of the labels in the degradation list.

The above insights clearly shows that, in general, SGTN gets better performance. However, when sensitive features got censored, it affects the performance of relevant labels (e.g., *donut*)

Last but not least, we explore the patterns of different models on the PP-MS-COCO data to understand the correlation between the performance of

Table 4: The performance comparison of SGTN on various label embeddings based on four different pre-trained models including GloVe, Bert, C2V, dpUGC. Noisy denotes the adjacency matrix construction based on the proposed Algorithm 1.

Experiment#	Adjacency Matrices in A						mAP	
	Local Knowledge			Global Knowledge				
	Frequency-based	dpUGC-based	Noisy	Glove-based	Bert-based	C2V-based	PP-MS-COCO	MS-COCO
1	✓		✓	✓	✓	✓	85.54	86.47
2				✓	✓	✓	85.49	86.54
3			✓	✓	✓	✓	85.57	86.60
4		✓	✓	✓	✓	✓	**85.58**	**86.64**



(a) ML-GCN vs SGTN on MS-COCO.

(b) ML-GCN vs SGTN on anonymised MS-COCO (PP-MS-COCO).

Figure 6: Per-class improvement or degradation of F1 between ML-GCN and SGTN on MS-COCO (a) and PP-MS-COCO (b). The top-10 improved classes from our SGTN are indicated as blue, and the top-10 degraded classes as orange.

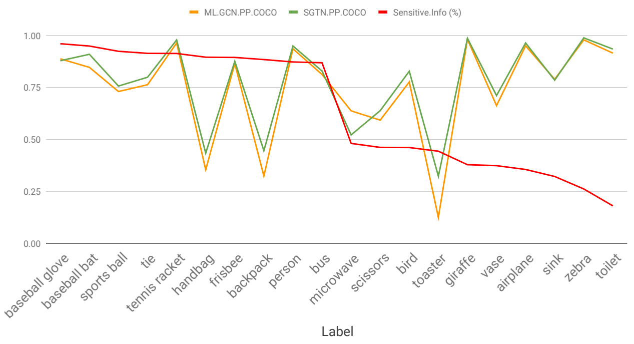


Figure 7: Per-class comparison of F1 between ML-GCN and SGTN on PP-MS-COCO. For visibility, only the top-10 of the most (and the least) sensitive visual labels are shown.

ML.GCN versus SGTN according to the amount of sensitive visual features. Figure 7 visualises the differences in performance of ML.GCN and SGTN in corresponding to the amount of sensitive visual features being censored in PP-MS-COCO dataset. The first 10 labels have the highest number of censored objects, and the last 10 labels have the least number of censored objects in percentage (%). In general, for the both cases, the improvement of SGTN outweighs the degradation of some labels, thereby leading to state-of-the-art performance.

6 CONCLUSION

This paper presents SGTN, a privacy preserving multi-label classification model for visual tagging task by applying the techniques of graph transformer and convolutional neural network. SGTN is designed to incorporate privacy-conscious knowledge to perform the downstream tasks with high performance, and meanwhile prevent privacy breach by avoiding using the sensitive knowledge from the data of the task itself.

SGTN showcases a new approach in dealing with several datasets. It effectively performs better on both censored multimedia data (MS-COCO and EDU-MM) by leveraging global knowledge into the learning process. Moreover, the proposed algorithm for constructing the dp-adjacency matrix is very efficient, which can guide the model to avoid using private relationships between labels in the downstream data. In the case that global knowledge is not available for specific reason such as the case of EDU-MM dataset, the dpUGC based graph construction is an advantage in helping the task to boost the performance. We conducted extensive experimental studies on a benchmark dataset (i.e., MS-COCO) and a real education dataset. The results show our proposed SGTN outperforms the state-of-the-art approaches with various settings.

By introducing SGTN we enable a new way of applying visual tagging tasks in multimedia data. For instance, it can be used for processing audio tagging tasks with the use of spectrogram images and the transcript of speech content. Especially, for the case of sensitive data such as medical records and medical imaging tasks, SGTN can be applied without the need to modify its architecture.

7 ACKNOWLEDGEMENT

This work is partially supported by the Federated Database project from the Umeå University, Sweden. The authors also thank the ITS organisation for the support on the EDU-MM data.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. Brendan McMahan, I. Mironov, K. Talwar, and L. Zhang. 2016. Deep Learning with Differential Privacy. *ArXiv e-prints* (July 2016).

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1638–1649.

[3] Roberto J Bayardo and Rakesh Agrawal. 2005. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International conference on data engineering*. 217–228.

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5177–5186.

[5] Common Crawl. 2018. Common crawl. *URl: http://commoncrawl.org* (2018).

[6] Dwork Cynthia. 2006. Differential Privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*. 1–12.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2009), 1627–1645.

[10] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.

[11] Weifeng Ge, Sibei Yang, and Yizhou Yu. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1277–1286.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa. 2017. Multi-Label Fashion Image Classification with Minimal Human Supervision. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*. 2261–2267.

[15] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. 2016. Big data privacy: a technological perspective and review. *Journal of Big Data* 3, 1 (2016), 25.

[16] Gerald C Kane and Alexandra Pear. 2016. The rise of visual content online. *MIT Sloan Management Review* (2016).

[17] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2741–2749.

[18] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*. 1–14.

[19] Yann LeCun and Yoshua Bengio. 1998. *Convolutional Networks for Images, Speech, and Time Series*. MIT Press, Cambridge, MA, USA, 255–258.

[20] Qing Li, Xiaojiang Peng, Yu Qiao, and Qiang Peng. 2019. Learning Category Correlations for Multi-label Image Recognition with Graph Networks. (2019). arXiv:1909.13005 http://arxiv.org/abs/1909.13005

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[22] L. Liu, J. Wang, J. Liu, and J. Zhang. 2008. Privacy preserving in social networks against sensitive edge disclosure. *Technical Report Technical Report CMIDA-HiPSCCS 006-08* (2008).

[23] David Lorenzi and Jaideep Vaidya. 2011. Identifying a critical threat to privacy through automatic image classification. In *Proceedings of the first ACM conference on Data and application security and privacy*. 157–168.

[24] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[25] Hoang D. Nguyen, Xuan-Son Vu, Quoc-Tuan Truong, and Duc-Trong Le. 2020. Reinforced Data Sampling for Model Diversification. arXiv:cs.LG/2006.07100

[26] Hiep H Nguyen, Abdessamad Imine, and Michaël Rusinowitch. 2016. Detecting communities under differential privacy. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. 83–93.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[28] NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. 2017. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *Proceedings of the 2017 IEEE International Conference on Data Mining*. IEEE, 385–394.

[29] Vadim Popov, Mikhail Kudinov, Irina Piontkovskaya, Petr Vytovtov, and Alex Nevidomsky. 2018. Distributed Fine-tuning of Language Models on Private Data. In *Proceedings of the International Conference on Learning Representations*.

[30] Jialie Shen, Meng Wang, Shuicheng Yan, and Xian-Sheng Hua. 2011. Multimedia tagging: past, present and future. In *Proceedings of the 19th ACM international conference on Multimedia*. 639–640.

[31] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*. 629–633.

[32] Son N. Tran, Qing Zhang, Anthony Nguyen, Xuan-Son Vu, and Son Ngo. 2018. Improving Recurrent Neural Networks with Predictive Propagation for Sequence Labelling. In *Neural Information Processing*. Springer International Publishing, 452–462.

[33] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13.

[34] Understand.AI. 2019. Understand.AI Anonymizer. https://github.com/understand-ai/anonymizer. commit 2fc7ab3.

[35] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Journal of Communications of the ACM* 57, 10 (2014), 78–85.

[36] Xuan-Son Vu, Son N. Tran, and Lili Jiang. 2019. dpUGC: Learn differentially private representation for user generated contents. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing*. 1–16.

[37] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.

[38] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.

[39] Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K. Reiter, and Zheng Dong. 2009. Privacy-preserving Genomic Computation Through Program Specialization. In *Proceedings of the 16th ACM conference on Computer and communications security*. 338–347.

[40] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2019. Multi-Label Classification with Label Graph Superimposing. (2019). arXiv:1911.09243 http://arxiv.org/abs/1911.09243

[41] Y. Wang and X. Wu. 2013. Preserving differential privacy in degree-correlation based graph generation. *Transactions on Data Privacy* 6, 2 (2013), 127–145.

[42] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. 2018. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision*. 606–624.

[43] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. 2011. Correlative multi-label multi-instance image annotation. In *Proceedings of the 2011 IEEE International Conference on Computer Vision*. 6–13.

[44] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. (2019). arXiv:1905.00546 http://arxiv.org/abs/1905.00546

[45] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph Transformer Networks. In *Proceedings of Advances in Neural Information Processing Systems*. 11960–11970.

[46] Ye Zhang, Nan Ding, and Radu Soricut. 2018. SHAPED: Shared-Private Encoder-Decoder for Text Style Adaptation. (2018), 1528–1538.

[47] Bin Zhou and Jian Pei. 2011. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and information systems* 28, 1 (2011), 47–77.

[48] Bin Zhou, Jian Pei, and WoShun Luk. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter* 10, 2 (2008), 12–22.

[49] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5513–5522.