Hindawi
Security and Communication Networks
Volume 2023, Article ID 5627246, 16 pages
https://doi.org/10.1155/2023/5627246

Research Article

Adaptive Sensitive Information Recognition Based on Multimodal Information Inference in Social Networks

Peiyu Ji,[1,2] Fangfang Shan,[1,2,3] Fuyang Li,[1,2] Huifang Sun,[1,2] Mengyi Wang,[1,2] and Dalong Shan[4]

[1]*School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China*
[2]*Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou 450001, China*
[3]*School of Computer Science, Zhengzhou University of Technology, Zhengzhou 450001, China*
[4]*School of Information Engineering, Henan Industry and Trade Vocational College, Zhengzhou 451191, China*

Correspondence should be addressed to Fangfang Shan; 6129@zut.edu.cn

Received 3 May 2022; Revised 12 December 2022; Accepted 11 February 2023; Published 7 July 2023

Academic Editor: Hyun-A. Park

With the advent of the multimedia era, the identification of sensitive information in social data of online social network users has become critical for maintaining the security of network community information. Currently, traditional sensitive information identification techniques in online social networks cannot acquire the full semantic knowledge of multimodal data and cannot learn cross-information between data modalities. Therefore, it is urgent to study a new multimodal deep learning model that considers semantic relationships. This paper presents an improved multimodal dual-channel reasoning mechanism (MDR), which deeply mines semantic information and implicit association relationships between modalities based on the consideration of multimodal data fusion. In addition, we propose a multimodal adaptive spatial attention mechanism (MAA) to improve the accuracy and flexibility of the decoder. We manually annotated real social data of 50 users to train and test our model. The experimental results show that the proposed method significantly outperforms simple multimodal fusion deep learning models in terms of sensitive information prediction accuracy and adaptability and verifies the feasibility and effectiveness of a multimodal deep model considering semantic strategies in social network sensitive information identification.

1. Introduction

Privacy is a significant concern for online social network users because it affects their ability to control who can access their personal information and how that information can be used. Without effective privacy protection, users may be at risk of having their personal information accessed or misused by others, which can lead to various potential harms such as identity theft, financial fraud, or online harassment. By protecting user privacy, online social networks can help ensure that users feel safe when using the network and contribute to building trust and confidence throughout the network. In addition, protecting user privacy is also important for online social networks themselves. If users feel that their personal information is not adequately protected, they may be less likely to use the network, which could reduce the network's user base and decrease its overall value. By protecting user privacy, online social networks can help maintain and expand their user base, which is critical to their continued success. Although users may be reluctant to share personal data, the inherent linkages between public data and private data often result in serious privacy breaches. The 2021 Data Security Conference has once again received much attention and various voices on data security have emerged. According to incomplete statistics, since 2015, the number of Internet black-gray industry professionals has exceeded 400,000. Although public data show that the domestic network security industry is expected to exceed 60 billion yuan in 2019, the black-gray industry has already reached a scale of 100 billion yuan. These studies show that private data are often subject to reasoning attacks, where enemies analyze a user's public data to illegally obtain

information about their private data. However, few social network users are aware of the serious dangers of privacy breaches, so in the face of explosive growth in network data, maintaining a safe environment for the dissemination of information in the network community is a domestic and international need for the network environment.

With the widespread deployment of heterogeneous networks, a large amount of high-capacity, high-diversity, high-speed, and high-accuracy data has been generated. These multimodal big data contain rich information between modalities and across modalities, which poses a great challenge to traditional sensitive information identification methods. The research on multimodal sensitive information recognition in online social networks is an important research field in the automatic recognition of sensitive information in online social networks. This can be used for various purposes, such as detecting and removing harmful content, protecting user privacy, and developing more effective auditing tools. By identifying sensitive information such as personal information, potential offensiveness or harmful content, or illegal activities, it helps prevent unauthorized parties from accessing users' personal data, such as financial information or login credentials, and using it for malicious purposes, such as identity theft or fraud. In addition, these systems can be used to monitor users' activities and any suspicious behavior on the network, such as attempting to access personal information without permission. By detecting and preventing unauthorized access to personal information, these systems can help protect user privacy and protect their data from potential threats.

Effective detection of widespread sensitive content is a critical issue. Research design of multimodal sensitive information recognition systems is a relatively effective response measure compared to systems that rely on a single mode or method, as they can detect a wider range of sensitive information. For example, systems that only use natural language processing to analyze text may miss sensitive information contained in images or videos, or that is implied rather than explicitly stated in text. By using multiple methods to analyze network content, multimodal systems can provide a more comprehensive view of sensitive information present on the network and can more effectively detect and remove it. Another advantage of multimodal systems is that they can be more robust when faced with attempts to evade detection. For example, users attempting to share personal information on the network may try to hide it in various ways, such as using abbreviations, initialisms, or slang, or by embedding it into images or other nontextual content. A multimodal system that has been trained to recognize various disguises can more effectively detect and prevent this type of sensitive information, while a system that only uses a single mode or method may be more easily fooled. In general, multimodal sensitive information recognition systems are considered an important research area because they can help make online social networks safer for all users. By automatically detecting and removing sensitive information, these systems can protect users' privacy, prevent the spread of harmful content, and make the development of more effective review tools possible.

In this paper, we attempt to learn the semantics of users' sensitive information in multimodal social network environments. We focus on the application of multimodal data interaction, feature fusion, knowledge perception, and related data mining in the field of social network privacy protection. Considering the characteristics of social networks, such as the diversity of data types posted by users, the accumulation of historical information leading to the leakage of sensitive information, and the differences in the definitions of sensitive information among different users, we propose an improved multimodal data fusion dual-channel multihop reasoning mechanism based on information content, data attributes, and user features, considering the background knowledge of users and the historical records of data published in social networks. The mechanism realizes the interaction of different modal data, explores the implicit correlations between different modalities, and determines the meaningful sensitive features in historical data. In addition, based on the user-defined sensitive list, we propose an adaptive multimodal spatial attention mechanism to generate an understanding of user-sensitive information, implement the rapid screening of implicit sensitive information, and prevent privacy information leakage caused by data association.

As shown in Figure 1, in multimodal sensitive information recognition, we consider the potential semantic dependencies in visual and textual contexts, attempt to mine the implicit correlations between multimodal data, and enhance the semantic representation of sensitive information through feature fusion and adaptive attention mechanisms. Given a user's social dynamics (including images, image descriptions, text, sensitive lists, and historical privacy settings), we can improve the accuracy of the decoder's response through iterative interaction, knowledge reasoning, and the fusion of visual and textual features. In this way, we can obtain sensitive items that may reveal the user's privacy.

The structure of this article is as follows. In Section 1, we will discuss the existing challenges in the field of identifying sensitive information in online social networks and introduce our proposed solution. Section 2 will provide an overview of previous research on privacy protection in online social networks, focusing on the problems and challenges that motivated our approach. In Section 3, we will describe the user-sensitive data leakage problem that our paper aims to solve. Section 4 will detail the method we propose for identifying sensitive information, including our approach to feature extraction, the improved multimodal semantic strategy, and the multimodal adaptive spatial attention mechanism. In Section 5, we will describe our experimental procedures and analyze the results of our experiments.

The main contributions of this thesis are summarized as follows:

(1) We propose an improved two-channel multihop reasoning mechanism for interactive reasoning of user image and text data in social networks to mine and exploit the implicit correlation between multimodal data. It breaks the semantic gap between cross-modal data and enriches the semantic representation of privacy in query text and image.

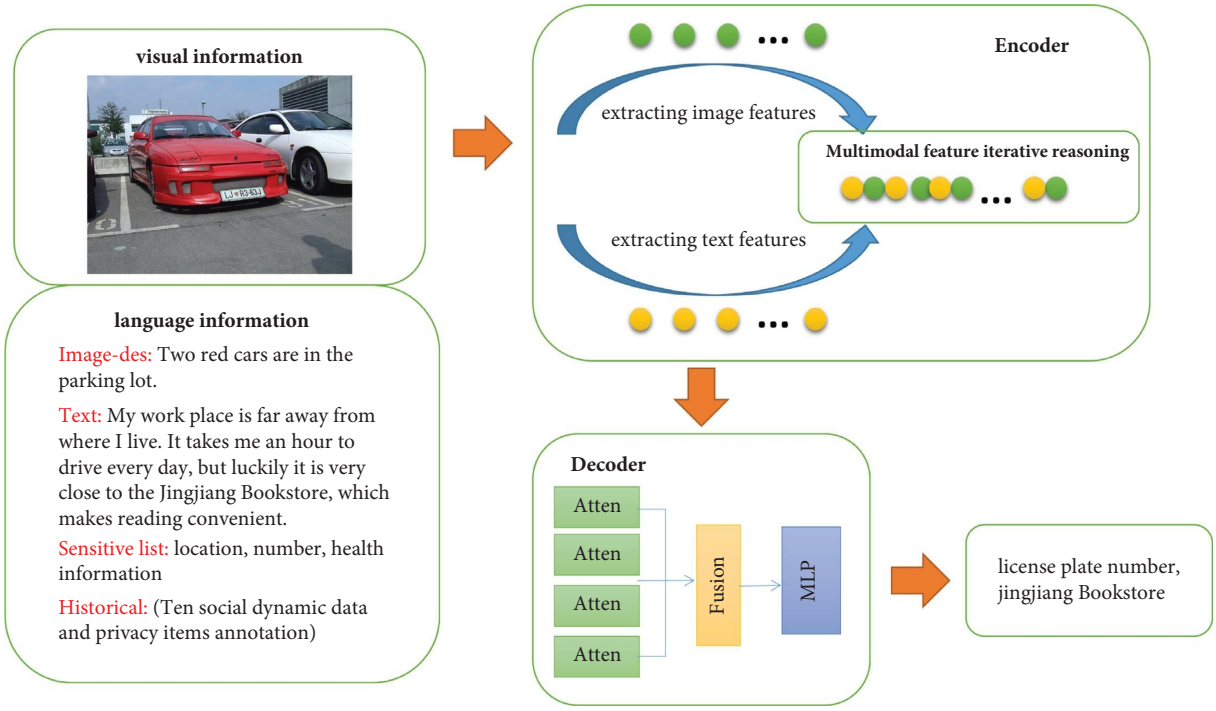

Figure 1: Diagram of a multimodal sensitive information recognition framework.

(2) User's personal sensitive preference is the difficulty of privacy protection technology. We enhance the representation of sensitive information preference by adding user-defined sensitive list and putting it into the two-channel multihop reasoning mechanism and finally realize personalized user privacy preference.

(3) We design an improved multimode spatial attention codec architecture to dynamically select the feature information that requires attention, so as to achieve accurate recognition of sensitive information.

2. Related Work

As the Internet continues to rapidly develop, a wide variety of social platforms have emerged, and users often provide personal information when using these platforms, including identification numbers, phone numbers, addresses, and health data. However, as technologies such as big data, cloud computing, and deep learning have evolved, network privacy vulnerabilities have become increasingly serious. The security environment of network communities is a major concern for both domestic and foreign users. Maintaining a secure environment for the dissemination of information in network communities has become a major challenge that needs to be addressed.

2.1. Research on Privacy Protection. Existing privacy protection methods for social networks are mostly based on anonymous algorithms for social network privacy protection models and on differential privacy-based social network privacy protection models. The former is mainly used for potential privacy attack problems, where the attacker is unable to identify themselves from a dataset consisting of multiple individual records and corresponding sensitive personal data. In order to overcome the weaknesses of traditional k-anonymity, Al-Asbahi [1] used an l-diversity method based on clustering techniques to communicate more substantial privacy protection and structural anonymity. In order to reduce the risk of sensitive information leakage or loss of a large amount of information, Lian and Chen [2] proposed a personalized (α, p, k) anonymous privacy protection algorithm. According to the sensitivity level of sensitive attributes, different anonymization methods are used for different levels of sensitive values in equivalent classes to achieve personalized privacy protection of sensitive attributes. In addition, location privacy research has a positive effect on preventing user-side write issues. Unlike traditional methods, Theodorakopoulos et al. [3] proposed location histogram dynamic privacy to focus on the efficiency of different locations being accessed. Ruan et al. [4] proposed an efficient location sharing protocol that supports location sharing between friends and strangers while protecting user privacy.

The effectiveness of a model in deep learning is proportional to the amount of available data, and large-scale massive data are indispensable. In order to enhance the availability of privacy-sensitive data in third-party infrastructure, Xu et al. [5] designed a secure computing protocol for the hybrid function encryption scheme, training deep neural networks on multiple encrypted datasets collected from multiple sources; while ensuring the accuracy of the model, the data confidentiality is improved. Despite this, data owners still worry about privacy leaks when providing sensitive data for model training. To solve this problem, Li et al. [6] proposed noninteractive privacy-preserving multiparty machine learning, providing an effective communication method for data owners. Similarly, Wang et al. [7] proposed that all sensitive data be operated in ciphertext rather than decrypted during the model training and

epidemic risk prediction stage. Lei et al. [8] considered privacy from a more granular perspective, protecting user facial features based on reversibility and reusability. In addition, data encryption for privacy protection is a common means. Idepefo et al. [9] combined blockchain technology with cryptography, hash, and consensus mechanisms. Xie et al. [10] proposed a hybrid data method based on homomorphic encryption and AES and constructed a multiclass support vector machine-based privacy-preserving medical data sharing system. However, excessively encrypted data will lead to a decrease in the accuracy of social user recommendations. Therefore, Chen et al. [11] improved on the basis of the additive secret sharing scheme and proposed a security comparison protocol and a division protocol, which strengthened the data privacy protection recommendation system.

In addition, some scholars are committed to privacy protection research in the Internet of Things [12–15], based on context privacy protection and user effective feature-based privacy protection in social networks. Chen et al. [16] proposed a privacy protection optimal nearest neighbor query (PP-OCQ) scheme that implements secure optimal nearest neighbor queries in a distributed manner without disclosing sensitive user information. Li and Zeng [17] presented a novel NRL model for generating node embeddings that can handle data incompleteness resulting from user privacy protection. Additionally, they proposed a structure-attribute enhanced matrix (SAEM) to mitigate data sparsity and developed a community-cluster informed NRL method, c2n2v, to further enhance the quality of embedding learning. Zhang et al. [18] developed a machine learning-based method to calculate malicious services and protect user data through direct and indirect trust, effectively controlling or associating leaked datasets in online social networks (OSN) and establishing a trust evaluation model in OSN. These privacy protection technologies are constantly maturing, but in the research process, the multimodality of social network information data has not been considered, and the relevant data analyzed are thin, and it is impossible to integrate multimodal data into the data.

2.2. Sensitive Information Identification. Effective identification of sensitive information is an effective way to improve privacy protection. Some scholars have studied automatic detection models of sensitive information. Heni and Gargouri [19] showed a method for identifying sensitive information in Mongo data storage, which is based on semantic rules to determine the concepts and language components that must be segmented, retrieves the attributes that are semantically corresponding to the concepts, and implements them as an expert system for automatically detecting segment candidates of attributes. Ding et al. [20] constructed a corpus to train the detection model, applied the BERT method to detect problems, and finally obtained a BERT-based automatic detection model of sensitive information. Botti-Cebriá et al. [21] proposed an auxiliary proxy to detect sensitive information according to the different categories (i.e., location, personal data, health, personal attacks, emotion, etc.) detected in the message. Liu et al. [22] trained sensitive data to establish a decision tree, which can classify and identify known data and can mark and encrypt the identified sensitive information to achieve intelligent recognition and protection of sensitive information. Kaul et al. [23] proposed a knowledge and learning-based adaptive system for sensitive information identification and processing. Gao et al. [24] used image caption technology to track the spread of image information on the network through text. Wang et al. [25] described the underlying reasoning behavior through Bayesian networks, resisting attackers' reasoning attacks on sensitive information. Petrolini et al. [26] developed a classifier that can monitor documents containing sensitive data, making it easier to identify and protect sensitive information. Bracamonte et al. [27] studied users' perceptions of monitoring sensitive information tools, quantitatively and qualitatively applying their reactions. Wu et al. [28] proposed a constraint measure to minimize the spread of sensitive information and relied on the Bandit framework to adaptively execute the spread constraint measure. Singh et al. [29] used local sampling to generate differentially private sensitive information, generating useful representations while maintaining privacy. Gao et al. [30] proposed a scheme through research that can audit the integrity of all encrypted cloud files of keywords of interest to users by only providing encrypted keywords to TPA, while unable to infer sensitive information such as files containing the keyword and the number of files containing the keyword. Neerbek [31] proposed to learn the phrases and structures that distinguish sensitive and nonsensitive documents in recursive neural networks. Unfortunately, with the rapid growth of cloud computing and remote workforce, organizations must handle a large amount of unstructured data, so automatic detection and recognition of secrets and sensitive information in structured and unstructured data is particularly important. Ahmed et al. [32] showed us the benefits of using deep learning to identify context-related sensitive information in unstructured data. Botti-Cebriá et al. [33] proposed a method for automatically monitoring sensitive information in educational social networks. Cai et al. [34] first applied three enhanced techniques in NER to Chinese sensitive information recognition based on the study of unstructured data, greatly solving the uncertainty and ambiguity of Chinese vocabulary and improving the accuracy of sensitive information recognition. However, single-mode data analysis has certain limitations in inferring the sensitive information of the current social network.

2.3. Multimodal Feature Fusion. The different modes of data dissemination make data modes diversified, so the study of multimodal fusion is gradually applied to various research fields. In the task of sentiment analysis, the importance of single modal data to the emotional result is not constant. With the extension of the time dimension, the emotional attributes of a specific natural language will be affected by the natural language data. Qi et al. [35] fully considered the long-term dependency between modes and the offset effect of nonnatural language data on natural language data, solving the long-term dependency within modes. Yan et al. [36] adopted tensor fusion network to model the interaction of multiple modes and achieve the emotion prediction of multimode features. Hu et al. [37] proposed a graph dynamic fusion module to fuse multimodal context features in the conversation. Chen et al. [38] proposed a feature fusion method based on K-means clustering and kernel canonical

correlation analysis (KCCA), which produces a higher recognition rate than existing methods (such as aware segmentation and tagging methods). Due to the inherent characteristics of each mode, it is difficult for the model to effectively use all modes when dealing with fusion mode information. Zou et al. [39] proposed the concept of main mode and used the method of main mode transformation to improve the effect of multimodal fusion. Yoon [40] proposed a cross-modal translator that can translate between three modes and can train multimodal models based on three modes using different types of heterogeneous datasets. Ghosh et al. [41] developed a multimodal multitask framework that utilizes a novel multimodal feature fusion technique and a contextuality learning module to handle emotional reasoning (ER) and accompanying emotions in conversations. Then, in the rumor detection task, Wu et al. [42] proposed a new multimodal collaborative attention network (MCAN), which combines multimodality extracted from text, spatial domain, and frequency domain (textual and visual) feature fusion as a method for detecting fake news. Experiments show that MCAN is able to learn correlations among multimodal features. Dhawan et al. [43] proposed an end-to-end trainable framework based on graph neural network (GAME-ON), which allows instant interaction between different modalities inside, and evaluated the framework on two effectiveness parameters on publicly available datasets. Azri et al. [44] proposed to use a multimodal fusion framework to evaluate message accuracy in social networks (MONITOR), which adopts supervised machine learning and utilizes all message features (text, social context, and image features) to provide interpretability for decision making. Chen et al. [45] proposed a multimodal fusion network (MFN) to integrate text and image data from social media, which uses self-attention fusion (SAF) mechanism to value for feature-level fusion. On the other hand, video captioning is a very challenging computer vision task. They used natural language sentences to automatically describe video clips. Bhooshan, R. S. et al.(2022) [46] proposed a neural structure based on discrete waveletconvolution and multimodal feature attention to generate video subtitles. Gao et al. [47] proposed a new paradigm for encrypted cloud data integrity auditing based on sensitive information privacy keywords. In this scheme, only the trusted third party auditor (TPA) possessing the encrypted keywords can audit the integrity of all encrypted cloud files containing user-relevant keywords. The scheme utilizes relationship authentication labels (RALs) to infer which files contain the keywords and how many files contain sensitive information related to those keywords. Experimental results demonstrate that the proposed scheme satisfies correctness, audit soundness, and sensitive information confidentiality. In addition, the visual question answering research that has emerged in recent years is also a research hotspot in the field of computer vision. How to fuse multimodal features extracted from images and questions is a key issue in VQA. Zhang et al. [48] designed an effective and efficient module to reason complex relationships between visual objects. They also learned a bilinear attention module to guide the attention on visual objects based on the given question. This combination of visual relationships and attention achieved a more fine-grained feature fusion. Chen et al. [49] adopted a dual-channel multihop inference mechanism to reason and fuse image features and text features to achieve cross-modal information interaction. Besides, Wang et al. [50] applied multimodal fusion to similarity user recommendation system and proposed an implicit user preference prediction method with multimodal feature fusion. Combining text and image features in user posts, the image and text features are extracted using convolutional neural network (CNN) and text CNN models, respectively, and then these features are combined as a representation of user preferences using early and late fusion methods. Finally, a list of users with the most similar preferences is suggested. Ding et al. [51] applied multimodal fusion to sarcasm detection and proposed a multimodal-based postfusion sarcasm detection method for postfusion with a three-level fusion structure and residual network model, which can better fuse the three modalities into a unified semantic space, thereby improving sarcasm detection. Xiao and Fu [52] combined visual language fusion and knowledge graph reasoning to further obtain useful information. In order to effectively detect multimodal sarcastic tweets, Xu et al. [53] constructed the decomposition and relation network (referred to as D&R Net) to model cross-modality contrast in the associated context. In this network, the decomposition network represents the commonalities and differences between images and texts, while the relation network simulates the semantic associations in the cross-modality context. Sankaran et al. [54] developed a refiner fusion network (ReFNet) that enables fusion modules to combine powerful unimodal representations with powerful multimodal representations. This approach addresses the large gap in existing multimodal fusion frameworks by ensuring that unimodal and fused representations are strongly encoded in the latent fused space.

Inspired by the field of visual dialogue, the task of identifying sensitive information on social networks is to fully understand the privacy semantics of users, recognizing privacy items not only from text history but also from visual-based information. In order to achieve our expectations, the following questions need to be considered. Firstly, to ensure the comprehensiveness of the analysis results, we use multimodal data (images and text) to decompose and integrate different pose data features, which is a daunting task. Secondly, how to make our designed reasoning mechanism similar to the visual dialogue process, constantly adjusting the final conclusion through the obtained information. Finally, since user-sensitive information varies from person to person, how to incorporate user-sensitive preferences into the information reasoning mechanism, enhance sensitive semantic information, and achieve the goal of personalized protection of user privacy.

3. Problem Description

The first problem we have to face is how to get a large and diverse set of sensitive data. This is a common challenge in the field of sensitive information recognition, as there are no

publicly available and widely accepted sensitive information datasets. In fact, all existing sensitive information recognition schemes rely on private datasets that cannot be accessed for free, which is understandable because publishing sensitive information may be illegal. To overcome this limitation, we decided to manually collect and annotate real data for the sensitive information recognition task. This enabled us to create a dataset that can be used for training and testing our model without violating any laws or regulations. Here are three examples of social updates from a user in the past month that we believe may reveal privacy and have potential risks.

As shown in Figure 2, Mike posted a picture of their new car on an online social platform with the caption "Just bought this beautiful car! Can't wait to take her for a spin!" The image clearly shows the car's manufacturer and model, as well as the license plate number. If this information is viewed by the wrong person, it could potentially be used to locate and steal the user's car.

As shown in Figure 2, Mike posted a message saying "I just got a new job at ByteDance! I'll be starting there next week, and I'm really excited." While this message may seem harmless, it could be sensitive if the user has not yet notified their current employer of their resignation plans. The information in this message, if viewed by the wrong person, could be used to harm the user's current employment situation or steal their identity.

As shown in Figure 2, Mike posted a picture of himself and his family with the caption "Having a great time on vacation in Chengdu! I can't wait to explore more of the city tomorrow." The picture shows the user and their family standing in front of a well-known tourist spot in the city, and the caption includes the name of the city they are vacationing in. In this case, the user's request to post the picture and caption on a social platform is likely to reveal sensitive information, such as their current location and the fact that their home may be unoccupied. If this information is viewed by the wrong person, it could potentially be used to locate the user and potentially break into their home while they are away on vacation.

It can be seen from the above social updates that it is always important to be cautious about the information shared on social media platforms, as it can potentially be used by others in harmful ways. If we only focus on single-modality data information, it will be difficult to fully capture the semantic information contained in the data, and it will be difficult to infer potential sensitive information. What is more troubling is that in multiple dynamic examples, we can include information related to natural persons in sentences other than those containing user-sensitive information, or even in any sentence. Therefore, it is important to look for privacy information in social updates within a certain range, which leads to the risk of false positives. In our examples, the task goal is to find data that may contain sensitive information in dynamic data within a certain range, which means that we need to reduce the false negative rate as close to zero as possible. Therefore, it is important to research and design multimodal semantic analysis strategies.

Petrolini et al. [26] introduced the concept of "sensitive topics" in their research on sensitive information, which is helpful in judging whether a sentence is sensitive information based on the analysis of its topic. Unfortunately, this unbiased approach ignores the user's personalized sensitive preferences. We have added different users' sensitive lists to solve this problem. We collected the user data and grouped the sensitive topics according to the user's sensitive list. Table 1 lists the five main sensitive items for 50 users. For these sensitive items, we searched for their hottest posts and related comments to obtain information about elements that are likely to be related to sensitive topics.

4. Architecture

In the process of protecting privacy data on social networks, understanding user-sensitive information is a critical bottleneck, which typically requires analyzing the user's historical resource data and historical access control settings to continuously adjust to determine the user's sensitive preferences. This process requires multiple adjustments and inferences. Using a multimodal data bi-channel multihop reasoning mechanism to determine user-sensitive preferences can help to use the rich potential information between multimodal privacy data to generate access control privacy permissions.

As shown in Figure 3, in this study, we focus on improving the two-channel multihop inference mechanism proposed by Chen et al. [49] for extracting sensitive information from user-posted resource data on social networks. First, we represent privacy information of users' historical texts and images with feature representations. All modal privacy feature representations are iteratively interacted through the two-channel multihop inference mechanism. After multihop interaction, the multimodal features are fused through the attention mechanism and finally input to the decoder to generate understanding of user-sensitive information.

4.1. Feature Representation. The input of this encoder is social dynamic text, sensitive list, historical privacy setting, image description, and image, and the output is language and visual pattern learning feature representation. As shown in Figure 2, the text and image are, respectively, passed through Bi-LSTM [55] and pretrained Faster R-CNN [56] to obtain the corresponding feature vector, to prepare for multimodal feature interaction reasoning.

4.1.1. Text Feature Representation. Bi-LSTM has been widely used in contextual text feature extraction, which can process historical and future information in sequence data and capture long-term dependencies in sequence arrays, thereby improving the accuracy and efficiency of sequence modeling tasks. The text input of the task includes the dynamic text D_q of user U_1, the picture description P_q, and the sensitive list L. For the convenience of processing, we combine the dynamic text and picture description to generate resource text T_q and use the pretrained Glove to vectorize the input text data.

Mike

Just got this beautiful car! I can't wait to take her for a ride!

(a)

Mike

I just got a new job at ByteDance! I'm excited to start working there next week.

(b)

Mike

What a wonderful holiday in Chengdu! I can't wait to explore more cities tomorrow

(c)

Figure 2: Sensitive social dynamics released by Mike.

Table 1: Grouping of the sensitive lists of 50 users and detection of sensitive words related to sensitive items.

Location	Salary	Religion	Sex	Health	Location
Related topics	Wage	Blasphemy	Gay	Terminal illness	Palestine
	Salary	Heresy	Lesbian	Incurable disease	West Bank
	Income	Apostasy	Bisexual	Mental illness	Gaza Strip
	Pay	Infidel	Transgender	Substance abuse	East Jerusalem
	Pay gap	Idolatry	Queer	Addiction	Golan Heights
	Gender pay gap	Sect	Genderqueer	Overdose	Crimea
	Race pay gap	Cult	Nonbinary	Suicide	Taiwan
	Pay equity	Extremism	Pansexual	Eating disorder	Tibet
	Minimum wage	Fundamentalism	Asexual	Obesity	Xinjiang
	Living wage	Conversion	Homophobia	Plastic surgery	Kashmir
	Pay discrimination	Interfaith marriage	Transphobia	Genetic disorder	North Korea
	Pay secrecy	Scripture	Biphobia	Birth defect	South Korea
	Bonus	Religious violence	Conversion therapy	Infertility	The Senkaku Islands
	Commission	Holy war	Same-sex marriage	Contagious disease	The Spratly Islands
	Overtime	Jihad	Gay rights	Pandemic	The Paracel Islands

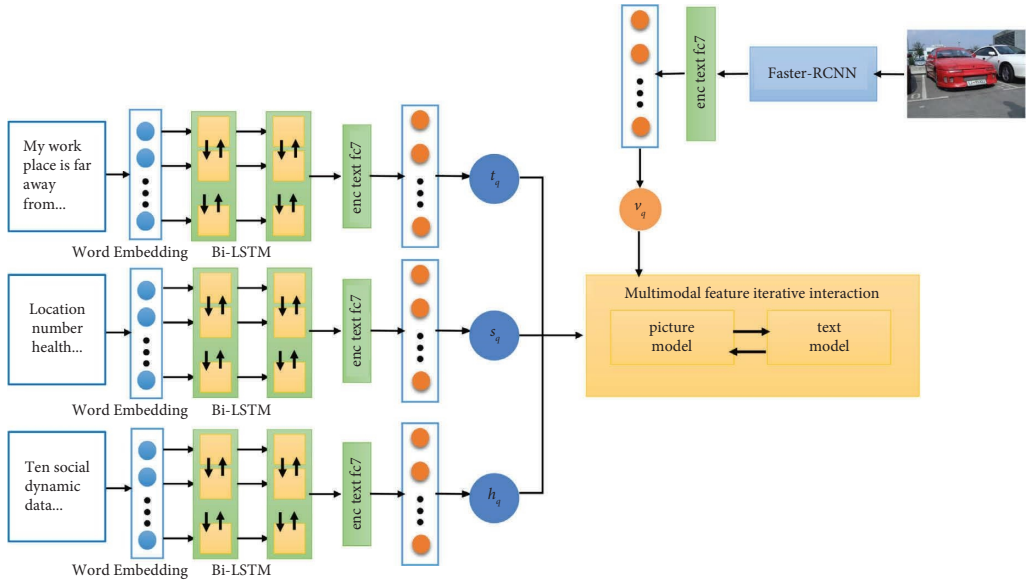

Figure 3: Feature extraction of multimodal data.

We combine dynamic text and picture descriptions to generate the resource text Tq. Then, we use pretrained Glove to vectorize the input text data, resulting in the word embeddings of the resource text, Tq = {tq1, tq2, tq3, . . ., tqm}. This allows the text vectors to contain more semantic and grammatical information. Then, we use Bi-LSTM to generate the hidden sequence $b = \{b_{q1}, b_{q2}, b_{q3} \ldots, b_{qm}\}$ and use the last hidden state as the generated resource text feature t_q, as shown in equations (1)–(3):

$$\overrightarrow{b_{qi}} = \mathrm{LSTM}_f\left(t_{qi}, b_{qi-1}\right), i \in \{0, \ldots, m-1\}, \tag{1}$$

$$\overleftarrow{b_{qi}} = \mathrm{LSTM}_b\left(t_{qi}, b_{qi+1}\right), j \in \{m-1, \ldots, 0\}, \tag{2}$$

$$t_q = \left[\overrightarrow{b_{q,m-1}}, \overleftarrow{b_{q,0}}\right]. \tag{3}$$

Historical privacy settings L and dynamic sensitive items S are embedded in the same way and combined with Bi-LSTM to generate historical privacy features $H_q = \{h_{q1}, h_{q2}, h_{q3} \ldots, h_{qn}\}$ and sensitive item features $S_q = \{s_{q1}, s_{q2}, s_{q3} \ldots, s_{qn}\}$.

4.1.2. Image Feature Representation. Faster R-CNN on ResNet-101 pretrained on Visual Genome data implements bottom-up attention to extract visual features of salient regions in input images. We took this model and pretrained it on the Visual Genome data. Specifically, we employ the Faster R-CNN framework to obtain object detection boxes in input images. Then, nonmaximum suppression is performed for each object region, and the top K ($K = 36$) detection boxes are selected, each with a feature size G ($G = 2048$). For each selected region proposal i, define v_i to be the average pooled convolutional feature for that region, so that the final representation of the input image is shown in the following equation:

$$v = \{v_1, v_2, \ldots, v_K\} \in R^{K \times G}. \tag{4}$$

This approach uses Faster R-CNN as a "hard" attention mechanism, since relatively few image regions are selected from a large number of possible configurations. In addition, we also recorded the scaled geometric features of selected image regions, denoted as $B = \{b_1, b_2, \ldots, b_K\}$, and b_i is shown in the following equation:

$$b_i = \left\{x_i, y_i, \frac{x_i}{w}, \frac{y_i}{h}, \frac{w_i}{w}, \frac{h_i}{h}\right\}, \tag{5}$$

where w_i and h_i are the coordinates, width and height of the selected region i, respectively. w and h are the width and height of the input image, respectively. These scaled geometric features will be input into our multimodal dual-channel information inference module.

4.2. Multimodal Sensitive Information Reasoning. Multimodal data contain not only intermodal information but also rich cross-modal information. In order to learn the rich intermodality and intersectionality information in multimodal sensitive information data, most of the existing multimodal deep learning models first use a deep model to capture the private features in the modality and combine the modality-specific original. The representation is transformed into a high-abstract representation of a certain global space. Then, these high-abstract representations are further concatenated into a vector, which represents a multimodal global representation. Finally, a deep model is used to model the high abstraction of the connected vectors [57]. However, the

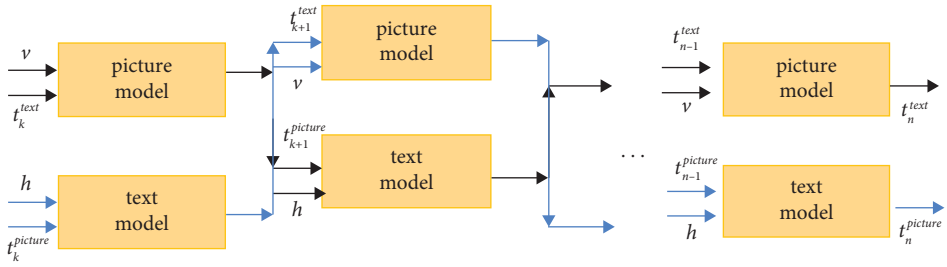

Figure 4: Dual-channel multihop reasoning module.

representation between modalities using this method is connected in a linear manner, which cannot adapt to complex relationships on multiple modalities and cannot capture full semantic knowledge of multimodal data. It can be seen that the combination of deep learning and semantic fusion strategy is an effective method to solve multimodal data fusion.

This section focuses on a new semantic fusion strategy, which is used to input multichannel sensitive information of user characteristics through a multimodal dual-channel multihop reasoning mechanism. The dual-channel multihop reasoning is used to mine the hidden semantic association between multiple modalities and jointly perform in-depth reasoning on sensitive information. This mechanism is mainly used in the field of visual dialogue and has a good effect on the question-answering mechanism.

As shown in Figure 4, the dual-channel sensitive information multihop reasoning mechanism is realized through two modules, namely, the image module and the text module. The image module fully understands sensitive semantic information through image features, and the text module fully understands sensitive semantic information from historical privacy features. The reasoning path of the image module is $I_1 \longrightarrow H_2 \longrightarrow I_3 \longrightarrow \ldots \longrightarrow I_n$, and the reasoning path of the text module is $H_1 \longrightarrow I_2 \longrightarrow H_3 \longrightarrow \ldots \longrightarrow H_n$. After the two modules are built, the output of the two modules needs to be iterated multiple times. Interaction and synchronous capture of information can not only make use of the hidden associations in text and images but also greatly enrich the understanding of sensitive semantic information.

4.2.1. Image Module Initialization. The image module is designed to enrich the semantic representation of sensitive information from images. The input of the image module is the query text $t_{picture}$ and the image feature v, and then it outputs the perceptual representation of the image privacy features. First map these feature vectors to the d_{picture} dimension vector and then use the attention mechanism to calculate the soft attention of all target detections, as shown in the following equation:

$$F = f^{t}_{\text{picture}}\left(t_{\text{picture}}\right) f^{v}_{\text{picture}}\left(v\right), \tag{6}$$

where f represents a two-layer perceptron with ReLU activation, the dimension of the input feature is d, W^s is the vector matrix of softmax activation, and o is the Hadamard product. The privacy-aware attention weight α is obtained by the following equation:

$$\alpha = \text{soft}\max\left(W^S F^S + b^S\right). \tag{7}$$

Then, the privacy-aware attention weight is applied to the image feature v, and the privacy-awareness of the image is calculated by the following equation:

$$t^{\text{out}}_{\text{track}} = \sum_{j=1}^{K} \alpha_j \times v_j. \tag{8}$$

4.2.2. Text Module Initialization. The text module is designed to enrich the semantic representation of sensitive information from historical texts. The input of the text module is the query text feature t_{text} and the historical privacy feature h, and then it outputs a query-aware representation of the text privacy features, as shown in equations (9) and (10):

$$Z = f^{t}_{\text{text}}\left(t_{\text{text}}\right) f^{\text{v}}_{\text{text}}\left(h\right), \tag{9}$$

$$\eta = \text{softmax}\left(W^Z Z^Z + b^Z\right), \tag{10}$$

where f represents a two-layer multilayer perceptron with ReLU activation, which converts the dimension of the input feature to d_{text}. W^z is a vector matrix with softmax activation, and $\circ$ is the Hadamard product. From the above equation, we obtain the attention weight η for query perception. Next, the attention weight of query perception is applied to the historical privacy feature h to calculate the query-aware representation of historical privacy, as shown in the following equation:

$$\hat{\boldsymbol{h}} = \sum_{j=1}^{T} \eta_j \times \boldsymbol{h}_j. \tag{11}$$

Next, apply $\hat{h}$ to the two-layer perceptron, with ReLU activation in the middle, and then add the representation of the sensitive list s to enhance the representation of historical text features on sensitive semantics, and finally obtain the

perceptual representation of historical privacy, as shown in equations (12) and (13):

$$g = W_u^2 \mathrm{ReLU}\left(W_u^1 \hat{h} + b_u^1\right) + b_u^2, \tag{12}$$

$$t_{\text{text}}^{\text{out}} = \mathrm{LayerNorm}(g + s). \tag{13}$$

4.2.3. Dual-Channel Multihop Reasoning. After initializing the image module and the text module, it is necessary to iteratively interact with the information of the two modules and deeply dig and utilize the implicit relationship between the image and the text. Multihop inference includes two types of multihop inference. One is the multihop inference starting from the image and ending with the image, as shown in $I_1 \longrightarrow H_2 \longrightarrow I_3 \longrightarrow \ldots \longrightarrow I_n$. The other is the multihop inference starting and ending with historical privacy, as shown in $H_1 \longrightarrow I_2 \longrightarrow H_3 \longrightarrow \ldots \longrightarrow H_n$. We implement each inference path through an image module and a text module.

Reasoning path 1 (starting and ending with the image): After initializing the image module with the image feature v input by the user and the text t_q to be queried, $t_{\text{picture}}^{\text{out}_1}$ is obtained through the calculation of the image module and then combined with the historical privacy feature h and input into the text. In the module, $t_{\text{text}}^{\text{out}_2}$ is calculated and then combined with the image feature v, and it is input into the image module to get $t_{\text{picture}}^{\text{out}_3}$. This is an interactive reasoning process, and then it iteratively proceeds in this way. Finally, the inference result $t_{\text{picture}}^{\text{out}_n}$ of the image module is obtained. The specific process is as follows:

Step 1: $\mathrm{Picture}((t_q), v) \longrightarrow t_{\text{picture}}^{\text{out}_1}$
Step 2: $\mathrm{Text}((t_{\text{picture}}^{\text{out}_1}), h) \longrightarrow t_{\text{text}}^{\text{out}_2}$
Step 3: $\mathrm{Picture}((t_{\text{text}}^{\text{out}_2}), v) \longrightarrow t_{\text{picture}}^{\text{out}_3}$

Repeat steps 1, 2, 3 iteratively...

Reasoning path 2 (starting and ending with text): After initializing the text module with the historical privacy feature h, privacy list feature s, and query feature t_q input by the user, $t_{\text{text}}^{\text{out}_1}$ is obtained through the calculation of the text module. Afterwards, the image features v are input into the image module to compute $t_{\text{picture}}^{\text{out}_2}$. Then, the historical privacy features h and the privacy list features s are combined and input into the text module to obtain $t_{\text{text}}^{\text{out}_3}$. This is an iterative process of interactive reasoning, and the computation continues in this manner. Finally, the inference result $t_{\text{text}}^{\text{out}_n}$ of the text module is obtained. The specific process is as follows:

Step 1: $\mathrm{Text}((t_q), h, s) \longrightarrow t_{\text{text}}^{\text{out}_1}$
Step 2: $\mathrm{Picture}((t_{\text{text}}^{\text{out}_1}), v) \longrightarrow t_{\text{picture}}^{\text{out}_2}$
Step 3: $\mathrm{Text}((t_{\text{picture}}^{\text{out}_2}), h, s) \longrightarrow t_{\text{text}}^{\text{out}_3}$

Repeat steps 1, 2, 3 iteratively...

Through the dual-channel multihop reasoning mechanism, the final result of multimodal feature interactive reasoning can be obtained, preparing for subsequent feature fusion.

4.3. Multimodal Fusion. Before fusing the polymorphic representations t_{picture}^n and t_{text}^n generated by the tracking module and the positioning module, we use the text feature t to be queried to enhance the representations of t_{picture}^n and t_{text}^n as follows:

$$\hat{t}_{\text{picture}}^n = f_{\text{att}}(t)\, f_{\text{att}}\left(t_{\text{picture}}^n\right), \tag{14}$$

$$\hat{t}_{\text{text}}^n = f_{\text{att}}(t) \circ f_{\text{att}}\left(t_{\text{txt}}^n\right), \tag{15}$$

where f represents a two-layer perceptron with ReLU activation. After obtaining the enhanced polymorphic representation, the representations of the two modules are fused, as shown in equations (16) and (17):

$$e = \left[W_f^1 \hat{t}_{\text{picture}}^n + b_f^1, W_f^2 \hat{t}_{\text{text}}^n + b_f^2\right], \tag{16}$$

$$\hat{e} = \tan h\left(W_f^3 e + b_f^3\right). \tag{17}$$

4.4. Multimodal Adaptive Spatial Attention Decoder. A multimodal spatial attention decoder is a neural network architecture for combining information from multiple modalities, such as audio, video, text, and image data, to make predictions or perform other tasks. It uses an attention mechanism to measure the importance of each pattern in a given context and then combines information from all patterns in a way that allows the network to make more accurate predictions.

4.4.1. Attention Mechanism. The essence of the attention mechanism is to locate the information of interest and suppress useless information. The attention mechanism in the multimodal spatial attention decoder is used to measure the importance of each modality in a given context. This means that the network can focus more on one mode than another, depending on the specific task requirements, for example, if the network is trying to recognize a word being spoken, it may focus more on audio data than visual data; once the neural network weighs the importance of each modality, it combines information from all modalities to make more accurate predictions, which may involve simply concatenating information from all modalities or may involve more complicated processing. The exact details of how a multimodal spatial attention decoder performs this fusion will depend on the specific architecture of the network.

4.4.2. Decoder. The multimodal decoder employed in this paper is an improvement on the adaptive spatial attention decoder. A recurrent neural network-based

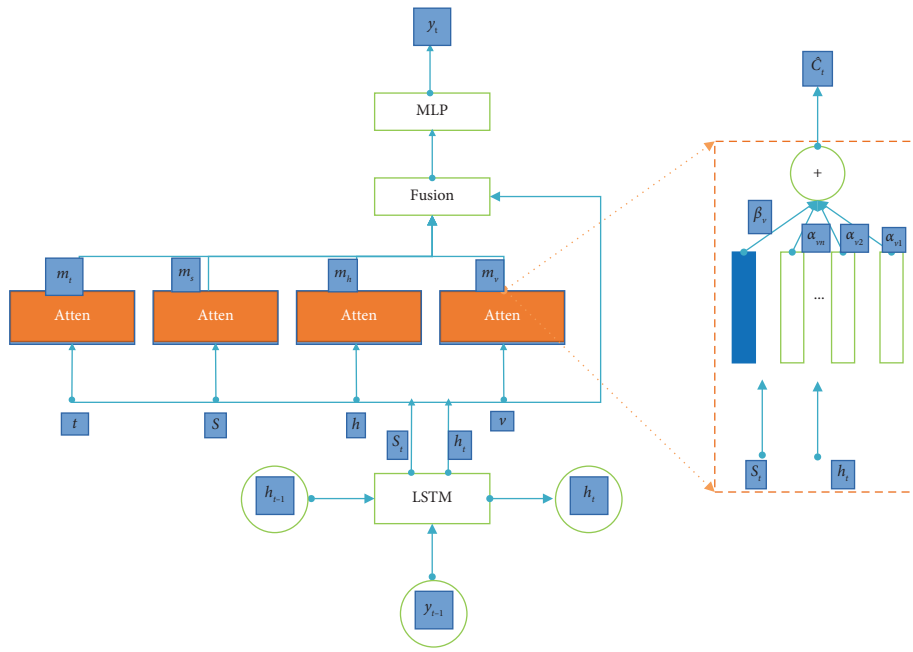

Figure 5: The improved adaptive multimodal spatial attention decoder is shown on the left, and the adaptive attention module is shown on the right.

approach is adopted that not only focuses on meaningful information but also decides as needed whether to rely on visual information or a language model to predict the next word in a sentence. The multimodal recurrent neural network can bridge the probability correlation between images and sentences, which solves the problem that new sensitive information cannot be generated when retrieving corresponding sensitive information in the sentence database based on learned image-text mapping in previous work. Unlike previous work, the recurrent neural model learns a joint distribution in semantic space given words and images. When multimodal features are present, it is possible to analyze the temporal dependencies hidden in multimodal data with the help of explicit state transitions in hidden unit calculations, using the time backpropagation algorithm to train parameters, and generate verbatim from the captured joint distribution sentence.

As shown in Figure 5, in the encoder-decoder framework, the log-likelihood of the joint probability distribution can be decomposed into ordered conditions by using the multimodal fusion feature representation and the features to be queried in the previous stage, using the chain rule, as show in the following equation:

$$\log p(y) = \sum_{t=1}^{T} \log p\left(y_t \middle| y_1, \ldots y_{t-1}, t, s, v, h\right). \tag{18}$$

Each conditional probability is modeled using a recurrent neural network, as shown in the following equation:

$$p\left(y_t \middle| y_1, \ldots, y_{t-1}, t, s, v, h\right) = f\left(h_t, \hat{c}_t\right), \tag{19}$$

where f is a 2-layer perceptron activated by ReLU and h_t is the hidden state of RNN at time t. In this paper, LSTM is used to model h_t, as shown in the following equation:

$$h_t = \text{LSTM}\left(y_{t-1}, h_{t-1}\right), \tag{20}$$

where y_{t-1} is the representation of sensitive information generated at time $t-1$.

Given the query feature t, historical privacy feature h, image feature v, privacy list s, and hidden state h_t, we input them through a single-layer perceptron with a softmax function to generate query features and sensitive list features, T round history privacy, and 4 attention distributions over K object detection features per image. The spatial attention model definition of the multimodal context vector c_t is shown in the following equation:

$$c_t = g\left(t, h, v, s, h_t\right). \tag{21}$$

The first is the historical privacy vector m_h, which is defined as follows:

$$z_t^h = W_h^h \tan h\left(W_q h + \left(W_g^h h_t\right) E^T\right), \tag{22}$$

$$\alpha_t^h = \text{softmax}\left(z_t^h\right), \tag{23}$$

where E is a vector with all elements set to 1 and W_q, W_g^h, W_h^h are learning parameters. Afterwards, the query vector m_t is obtained as follows:

$$m_h = \sum_{i=1}^{l} \alpha_{t,i}^h h_{i.} \tag{24}$$

Similar to the calculation of the query text, we obtain the participating query vector m_t, image vector m_v, and sensitive

list vector m_s, and then fuse these three context vectors to obtain the context vector c_t, as shown in the following equation:

$$c_t = \tan h\left(W_c\left[m_t \cdot m_s \cdot m_h \cdot m_v\right]\right), \tag{25}$$

where $[\cdot]$ represents the multiplication between vectors, W_e, which denotes the learnable parameters, is used to compute the vector, c_t and the vector ct is then combined with h_t to predict the next word y_{t+1}. c_t is the multimodal context vector at time t. In the attention-based framework, c_t depends on both the encoder and the decoder. At time t, the decoder will focus on specific areas of text and images according to the hidden state. In order to improve the adaptive ability, the extended LSTM is used to obtain the visual sentinel s_t, as shown in formulas (26) and (27):

$$g_t = \sigma\left(W_x x_t + W_h h_{t-1}\right), \tag{26}$$

$$s_t = g_t \circ \tan h\left(m_t\right), \tag{27}$$

where W_x, W_h are learning parameters, g_t is the gate applied to the storage unit m_t, and x_t is the LSTM input at time t.

Based on the visual sentinel s_t, the multimodal context vector $\widehat{c}_t$ is calculated by an adaptive attention model, as shown in the following equation:

$$\widehat{c}_t = \theta_t s_t + (1 - \theta_t) c_t, \tag{28}$$

where θ_t is the new sentinel gate at time t. When θ_t is 1, it means to use visual marker signal, and when θ_t is 0, it means only spatial image information is used when generating predicted words. θ_t is calculated by the attention distribution α_t on the spatial image, and the calculation process is shown in equations (29) and (30):

$$\theta_t = \widehat{\alpha}_t[k+1], \tag{29}$$

$$\widehat{\alpha}_t = \text{soft}\max\left[z_t \cdot w_h^T \tanh\left(W_s s_t + \left(W_g h_t\right)\right)\right]. \tag{30}$$

In addition, we use the encoder output $\hat{e}$ as the embedding to initialize the input of the decoder LSTM, as shown in the following equation:

$$h_0 = \text{LSTM}\left(\hat{e}, t_q\right), \tag{31}$$

where t_q is the last state of the query LSTM in the encoder and h_0 is used as the initial state of the decoder LSTM.

5. Experiment

5.1. Dataset. We evaluate our experiments with manually annotated data posted by 50 students on social platforms. Each student has 120 pieces of data, and each piece of data includes content text, images, image descriptions, sensitive lists, and historical privacy settings. There are a total of 6k pieces of such data, including 6,000 pieces of image data and 24,000 pieces of text data. In the final training dataset, there are 4800 images and 19,200 kinds of text information, and the verification set has 600 images and 2400 kinds of text information, and the experimental results are verified in the test set of 600 images and 2400 kinds of text information.

5.2. Development Platform and Environment. The experimental environment is as follows:

Nvidia-SMI: 450.80.02.

Driver version: 450.80.02.

Cuda version: 11.0.

The operating system is Windows 10.

The design language is Python 3.6.0 (64-bit).

5.3. Data Preprocessing

5.3.1. Image Preprocessing. We use the pretrained Faster R-CNN model in the Caffe framework to extract image features from a collection of 6000 images. Faster R-CNN is a state-of-the-art object detection model trained on large datasets to recognize different objects in images. By using a pretrained model, we can save a lot of time and effort compared to training a model from scratch. In order to extract image features, the Faster R-CNN model processes each image and generates a feature vector of size $36 * 2048$ for each image. This feature vector contains information about the objects in the image, their locations, and the relationship between them. As shown in Figure 6, we performed object recognition on two images and marked the objects whose recognition prediction values were greater than 50% in the images.

In Faster R-CNN, the feature maps of each layer reflect different levels of image feature information. Generally, the shallow feature maps can reflect some low-level features of the image, such as edges, corners, and textures, while the deep feature maps can reflect some high-level semantic information, such as the shape and texture of objects. These feature maps can serve as inputs for subsequent target classification and localization, helping to locate and identify targets. To better understand the information contained in each layer of feature maps, we performed a layer-by-layer feature map output analysis of the image, as shown in Figure 7.

5.3.2. Text Preprocessing. We convert all text data to lowercase, set the maximum lengths of dynamic text, image description, and sensitive list to 25 for dynamic text, 30 for image description, and 20 for sensitive list, and then construct a secondary markup vocabulary. We utilize distributed word representations with default parameter settings on preprocessed text datasets and incorporate pretrained glove models to construct the vocabulary for the dataset. We obtained word embedding features for each word in the dataset. One reason for choosing to use word embedding instead of one-hot encoding to represent words is that in one-hot encoding, when the vocabulary size is too large, insufficient text may lead to poor word features.

5.4. Results and Analysis. Our proposed model architecture consists of multiple modules, and in this experiment, we compare our work with unimodal and multimodal models

Figure 6: Example of Faster R-CNN object detection results.

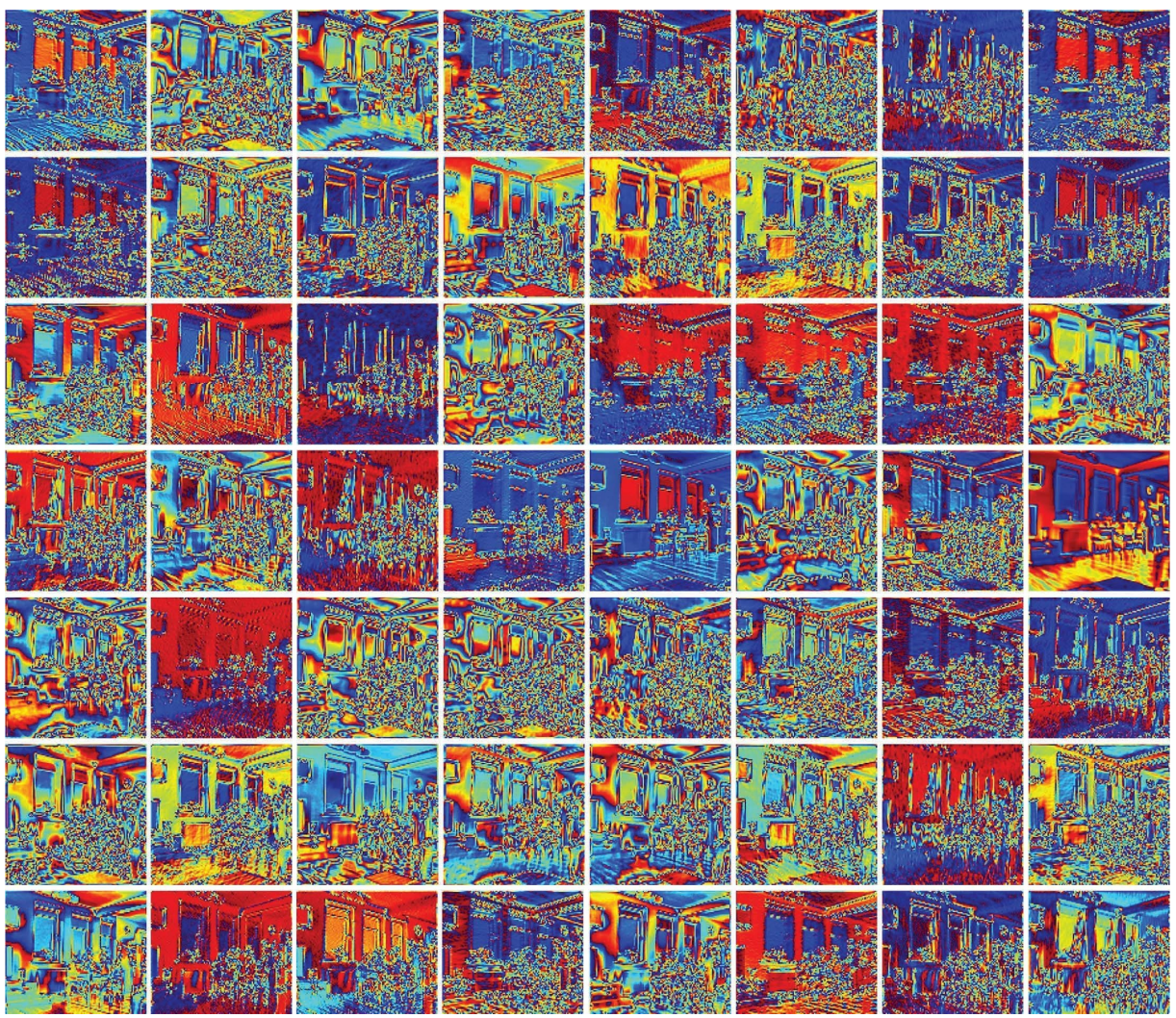

Figure 7: Visualizing feature maps.

and evaluate the impact of our designed reasoning module and multimodal spatial attention mechanism on contribution to the final prediction accuracy. We train the following comparison models on our collected real-world data and show the performance of different comparison schemes in Table 2.

Table 2: Evaluating the results of the experiment.

Model	MRR	R@1	R@5	R@10	Mean
T	21.36	15.17	26.67	35.33	17.21
V	20.50	14.33	29.17	38.33	19.32
MD + MDR	34.18	29.83	45.83	54.17	**15.20**
MD + MDR + MAA	**44.15**	**33.68**	**53.79**	**56.12**	18.46

The bold font in Table 2 shows the evaluation results of our model approach in the same data set.

(i) Single-text model (T): we embed each text as a word in the dataset and feed the feature vector into the decoder for sensitive information identification.

(ii) Single-vision model (V): ours takes only the fc7 layer feature output from the pretrained Faster R-CNN as input to our question.

(iii) Multimodal context fusion (MD + MDR): the multimodal features (MD) of the complete text T, image V, and user-sensitive preference S are taken as input to the problem, and a dual-channel multihop reasoning module (MDR) is added.

(iv) Multimodal context fusion (MD + MDR + MAA): take the complete text and image multimodal features (MD) as the input of the question and add a dual-channel multihop reasoning module (MDR) and adaptive multimodal space attention mechanism (MAA), which ultimately generates responses to sensitive information.

The experiment shows that single-modal feature analysis has limitations in outputting sensitive information, and processing multimodal data can enhance the representation of sensitive information semantics in complex social environments and relationships. Our model has good performance in online social user-sensitive information inference.

6. Conclusions

This paper improves the spatial attention decoder by proposing a multimodal adaptive spatial attention decoder. It combines a dual-channel multihop reasoning architecture to perform deep reasoning and prediction on user's historical sensitive data. This mechanism not only enables interaction between images and text but also allows for a thorough exploration and utilization of their implicit correlations. When predicting sensitive information, by paying attention to the context and context information of text and images and adaptively switching attention between visual information and language models, the flexible and accurate identification of sensitive user data is achieved, and in our study, from 50 volunteers, good results have been achieved in the data collected by the authors. Afterwards, this work will be combined with social network work access control to eliminate identified privacy items or set corresponding access rights.

In addition to the lack of privacy semantics caused by data diversity, another key challenge to protect online social network data privacy is the dynamic nature of data. Because data are constantly changing, it can be difficult to ensure that privacy is maintained over time. Traditional approaches to learning from dynamic multimodal data, such as training a new model every time the data distribution changes, can be time-consuming and impractical for online applications. Therefore, online learning and incremental learning have emerged as promising real-time learning strategies for multimodal data fusion. These methods allow new knowledge to be learned from new data without losing large amounts of historical knowledge, making them well suited to the dynamics and uncertainty of online social network data. In the following work, we will try to solve the privacy protection challenges brought about by the dynamic changes of multimodal data by designing online and incremental multimodal deep learning models.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Henan Provincial Key Laboratory of Cyberspace Situational Awareness (Project No. HNTS2022020), Key Science and Technology Project of Henan Province "Research on Machine Learning Algorithm Optimization Technology in Big Data Mining" (Project No. 222102210252), Key R&D and Promotion Project of Science and Technology of Henan Province (Project No. 212102310480), and Natural Science Foundation of Zhongyuan University of Technology (No. K2023QN018).

References

[1] R. Al-Asbahi, "Structural anonymity for privacy protection in social network," *International Journal of Scientific and Research Publications (IJSRP)*, vol. 11, no. 6, pp. 102–107, 2021.

[2] C. Lian and Z. Chen, "Anonymous privacy protection algorithm based on sensitive attribute classification," in *Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp. 222–226, IEEE, Taiyuan, China, October 2020.

[3] G. Theodorakopoulos, E. Panaousis, K. Liang, and G. Loukas, "On-the-fly privacy for location histograms," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 566–578, 2022.

[4] O. Ruan, L. Zhang, and Y. Zhang, "Location-sharing protocol for privacy protection in mobile online social networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, pp. 1–14, 2021.

[5] R. Xu, J. Joshi, and C. Li, "Nn-emd: efficiently training neural networks using encrypted multi-sourced datasets," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2807–2820, 2021.

[6] T. Li, J. Li, X. Chen, Z. Liu, W. Lou, and T. Hou, "Npmml: a framework for non-interactive privacy-preserving multiparty machine learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, 2020.

[7] F. Wang, H. Zhu, R. Lu, Y. Zheng, and H. Li, "Achieve efficient and privacy-preserving disease risk assessment over multi-outsourced vertical datasets," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1492–1504, 2020.

[8] J. Lei, Q. Pei, Y. Wang, W. Sun, and X. Liu, "PRIVFACE: fast privacy-preserving face authentication with revocable and reusable biometric credentials," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3101–3112, 2022.

[9] F. O. Idepefo, B. I. Akhigbe, O. S. Aderibigbe, and B. S. Afolabi, "Towards an architecture-based ensemble methods for online social network sensitive data privacy protection," *International Journal of Recent Contributions from Engineering Science & IT (iJES)*, vol. 9, no. 1, p. 33, 2021.

[10] B. Xie, T. Xiang, X. Liao, and J. Wu, "Achieving privacy-preserving online diagnosis with outsourced SVM in internet of medical things environment," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 4113–4126, 2022.

[11] J. Chen, L. Liu, R. Chen, W. Peng, and X. Huang, "Secrec: a privacy-preserving method for the context-aware recommendation system," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3168–3182, 2021.

[12] J. Xiong, R. Bi, Y. Tian, X. Liu, and D. Wu, "Toward lightweight, privacy-preserving cooperative object classification for connected autonomous vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2787–2801, 2022.

[13] R. Bi, Q. Chen, L. Chen, J. Xiong, and D. Wu, "A Privacy-Preserving Personalized Service Framework through Bayesian Game in Social IoT," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8891889, 13 pages, 2020.

[14] J. Xiong, R. Ma, L. Chen et al., "A Personalized Privacy Protection Framework for Mobile Crowdsensing in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4231–4241, 2020.

[15] L. Qi, B. Xia, H. Huang, Y. Zhang, and T. Zhang, "TRAC: Traceable and Revocable Access Control Scheme for mHealth in 5G-Enabled IIoT," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3437–3448, 2021.

[16] Y. Chen, H. Ku, and M. Zhang, "PP-OCQ: a distributed privacy-preserving optimal closeness query scheme for social networks," *Computer Standards & Interfaces*, vol. 74, Article ID 103484, 2021.

[17] C. T. Li and Z. Y. Zeng, "Learning effective feature representation against user privacy protection on social networks," *Applied Sciences*, vol. 10, no. 14, p. 4835, 2020.

[18] Y. Zhang, J. Tao, S. Zhang, Y. Zhang, and P. Wang, "A machine learning based approach for user privacy preservation in social networks," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1596–1607, 2021.

[19] H. Heni and F. Gargouri, "Towards an automatic detection of sensitive information in Mongo database," in *International Conference on Intelligent Systems Design and Applications*, Springer, Cham, Switzerland, pp. 138–146, 2018, December.

[20] M. Ding, X. Wang, C. Wu, K. Wang, and X. Yang, "Research on automated detection of sensitive information based on BERT Journal of Physics: Conference Series," *Journal of Physics: Conference Series*, vol. 1757, no. 1, Article ID 012088, 2021.

[21] V. Botti-Cebriá, E. D. Val, and A. García-Fornes, "Automatic detection of sensitive information in educative social networks," in *Computational Intelligence in Security for Information Systems Conference*, Springer, Cham, pp. 184–194, 2019, May.

[22] S. Liu, Z. Yang, Y. Li, and S. Wang, "Decision tree-based sensitive information identification and encrypted transmission system," *Entropy*, vol. 22, no. 2, p. 192, 2020.

[23] A. Kaul, M. Kesarwani, H. Min, and Q. Zhang, "Knowledge & Learning-Based Adaptable System for Sensitive Information Identification and Handling," in *Proceedings of the 2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*, pp. 261–271, Chicago, IL, USA, September 2021.

[24] L. Gao, X. Wu, J. Wu, X. Xie, L. Qiu, and L. Sun, "Sensitive image information recognition model of network community based on content text," in *Proceedings of the 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, pp. 47–52, IEEE, Shenzhen, China, October 2021.

[25] X. Wang, C. J. Bryan, Y. Li, R. Pan, and K. L. Ma, "Umbra: a visual analysis approach for defense construction against inference attacks on sensitive information," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 7, pp. 2776–2790, 2020.

[26] M. Petrolini, S. Cagnoni, and M. Mordonini, "Automatic detection of sensitive data using transformer-based classifiers," *Future Internet*, vol. 14, no. 8, p. 228, 2022.

[27] V. Bracamonte, W. B. Tesfay, and S. Kiyomoto, "Towards exploring user perception of a privacy sensitive information detection tool," in *Proceedings of the International Conference on Information Systems Security*, Lisbon, Portugal, June 2021.

[28] X. Wu, L. Fu, H. Long, D. Yang, and G. Chen, "Adaptive diffusion of sensitive information in online social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, 2020.

[29] A. Singh, E. Garza, A. Chopra, P. Vepakomma, V. Sharma, and R. Raskar, "Decouple-and-sample: protecting sensitive information in task agnostic data release," 2022, https://arxiv.org/abs/2203.13204.

[30] X. Gao, J. Yu, Y. Chang, H. Wang, and J. Fan, "Checking only when it is necessary: enabling integrity auditing based on the keyword with sensitive information privacy for encrypted cloud data," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 99, p. 1, 2021.

[31] J. Neerbek, "Sensitive information detection: recursive neural networks for encoding context," 2020, https://arxiv.org/abs/2008.10863.

[32] H. Ahmed, I. Traore, S. Saad, and M. Mamun, "Automated detection of unstructured context-dependent sensitive information using deep learning," *Internet of Things*, vol. 16, Article ID 100444, 2021.

[33] V. Botti-Cebriá, E. D. Val, and A. García-Fornes, *Automatic Detection of Sensitive Information in Educative Social Networks*, Springer, Cham, Switzerland, 2021.

[34] L. Cai, Y. Zhou, Y. Ding, J. Jiang, and S. H. Yang, "Utilizing lexicon-enhanced approach to sensitive information identification," in *Proceedings of the 2022 27th International Conference on Automation and Computing (ICAC)*, pp. 1–6, September 2022, Bristol, UK.

[35] Q. Qi, L. Lin, R. Zhang, and C. Xue, "MEDT: using multi-modal encoding-decoding network as in transformer for multi-modal sentiment analysis," *IEEE Access*, vol. 10, pp. 28750–28759, 2022.

[36] X. Yan, H. Xue, S. Jiang, and Z. Liu, "Multi-modal sentiment analysis using multi-tensor fusion network with cross-modal modeling," *Applied Artificial Intelligence*, vol. 36, no. 1, Article ID 2000688, 2022.

[37] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "Mm-dfn: multi-modal dynamic fusion network for emotion recognition in conversations," 2022, https://arxiv.org/abs/2203.02385.

[38] L. Chen, K. Wang, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "K-Means clustering-based kernel canonical correlation analysis for multi-modal emotion recognition in human-robot interaction," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 1, pp. 1016–1024, 2023.

[39] S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multi-modal fusion with Main Modal Transformer for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 258, Article ID 109978, 2022.

[40] Y. C. Yoon, "Can we exploit all datasets? Multi-modal emotion recognition using cross-modal translation," *IEEE Access*, vol. 10, pp. 64516–64524, 2022.

[41] S. Ghosh, G. V. Singh, A. Ekbal, and P. Bhattacharyya, "COMMA-DEER: COmmon-sense aware multi-modal multitask approach for detection of emotion and emotional reasoning in conversations," in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6978–6990, Gyeongju, Korea, October 2022.

[42] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multi-modal fusion with co-attention networks for fake news detection," in *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2560–2569, Bangkok, Thailand, August 2021.

[43] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, and P. Kumaraguru, "GAME-on: graph attention network based multi-modal fusion for fake news detection," 2022, https://arxiv.org/abs/2202.12478.

[44] A. Azri, C. Favre, N. Harbi, J. Darmont, and C. Noûs, "MONITOR: a multi-modal fusion framework to assess message veracity in social networks," in *Proceedings of the European Conference on Advances in Databases and Information Systems*, pp. 73–87, Turin, Italy, August 2021.

[45] J. Chen, Z. Wu, Z. Yang, H. Xie, and W. Liu, "Multi-modal fusion network with latent topic memory for rumor detection," in *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, July 2021.

[46] R. S. Bhooshan and K. Suresh, "A multi-modal framework for video caption generation," *IEEE Access*, vol. 10, pp. 92166–92176, 2022.

[47] X. Gao, J. Yu, Y. Chang, H. Wang, and J. Fan, "Checking only when it is necessary: enabling integrity auditing based on the keyword with sensitive information privacy for encrypted cloud data," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 3774–3789, 2022.

[48] W. Zhang, J. Yu, H. Hu, H. Hu, and Z. Qin, "Multi-modal feature fusion by relational reasoning and attention for visual question answering," *Information Fusion*, vol. 55, pp. 116–126, 2020.

[49] F. Chen, F. Meng, J. Xu, P. Li, B. Xu, and J. Zhou, "Dmrm: a dual-channel multi-hop reasoning model for visual dialog," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 7504–7511, 2020.

[50] J. H. Wang, Y. T. Wu, and L. Wang, "Predicting implicit user preferences with multi-modal feature fusion for similar user recommendation in social media," *Applied Sciences*, vol. 11, no. 3, p. 1064, 2021.

[51] N. Ding, S. W. Tian, and L. Yu, "A multi-modal fusion method for sarcasm detection based on late fusion," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8597–8616, 2022.

[52] S. Xiao and W. Fu, "Visual relationship detection with multi-modal fusion and reasoning," *Sensors*, vol. 22, no. 20, p. 7918, 2022.

[53] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multi-modal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, July 2020.

[54] S. Sankaran, D. Yang, and S. N. Lim, "Multi-modal fusion refiner networks," 2021, https://arxiv.org/abs/2104.03435.

[55] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, https://arxiv.org/abs/1508.01991.

[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[57] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multi-modal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.