Sensitive Data Leakage Detection Based on Vision-Language Models

1st Ran Dubin
Department of Computer Science
Ariel University, Ariel, Israel
rand@ariel.ac.il

***Abstract*—The identification of sensitive information in text and images is a critical challenge in privacy and data security. This project aims to explore and evaluate existing methods for detecting sensitive content in multimodal contexts, particularly combining text and visual data. The research focuses on analyzing various detection approaches to understand their strengths, limitations, and practical applicability in real-world scenarios.**

A key aspect of this project is to address the challenges posed by current datasets, including their limitations and how these affect the performance of detection algorithms. To tackle these issues, the project will create a novel dataset that integrates sensitive text and image examples. Additionally, the study will evaluate state-of-the-art techniques and investigate the potential of Vision-Language Models (VLMs) for multimodal analysis. This work aims to enhance the understanding of sensitive information detection and contribute to the development of more robust tools for safeguarding privacy, while identifying key areas for future research.

I. Introduction

In today's digital age, the spread of personal and confidential data has created significant challenges in safeguarding sensitive information. Sensitive data, which includes personal identifiers, financial records, medical histories, and proprietary information, is constantly at risk of being exposed or leaked. Detecting and protecting sensitive information is crucial in both research and industry. As data leakage becomes increasingly common, developing accurate and efficient ways to identify sensitive information is essential.

From an industrial perspective, the threat of sensitive data leakage has broad implications for privacy, security, and regulatory compliance. Industries such as finance, healthcare, and social media are under increasing pressure to protect customer data from breaches. Current solutions in the industry are primarily driven by rule-based systems or standard machine learning models that analyze text data. However, these approaches often struggle to detect sensitive content in complex, multimodal formats—such as images combined with text—leading to gaps in security. Furthermore, industrial applications often face issues related to scalability, real-time detection, and ensuring compliance with strict privacy regulations like GDPR and HIPAA.

On the academic side, researchers have developed a variety of methodologies to address these challenges. From machine learning techniques to deep learning models, significant progress has been made in detecting sensitive information. Recent advances in Vision-Language Models (VLMs), which combine computer vision and natural language processing, promise better analysis of image and text data. However, academic research often faces constraints related to the availability of large, labeled datasets, particularly those that contain both sensitive text and images, which can limit the effectiveness of detection models.

This project seeks to bridge the gap between academic advancements and industrial needs by developing a comprehensive dataset that combines text and image data related to sensitive information, and by evaluating cutting-edge multimodal detection techniques. This work aims to enhance both academic understanding and practical industry tools for protecting privacy in the digital world.

II. Related Work

The detection of sensitive data leakage in Vision-Language Models (VLMs) is an emerging field that intersects privacy concerns with multimodal machine learning systems. Below is a detailed summary of various studies that explore sensitive data leakage, including methodologies for detection, the datasets used, and the key findings.

[1] **"Privacy-Aware Visual Language Models"**: PRIVBENCH and PRIVBENCH-HARD are benchmarks for testing privacy in Vision-Language Models (VLMs), supported by a fine-tuning dataset, PRIVTUNE. Using LoRA on TinyLLaVa and MiniGPT-v2, models achieved up to 96% AUC-ROC on privacy datasets with just 150 fine-tuning samples. The dataset includes private (e.g., passports) and public (e.g., landscapes) images, with text generated by GPT-4V. This approach enhances privacy detection without sacrificing overall performance.

[2] **"Extracting Training Data from Document-Based VQA Models"**: Document-Based VQA models can leak sensitive information like PII. Counterfactual analysis separates memorization from understanding, and Extraction Blocking (EB) prevents answers without enough context. EB reduces privacy risks and improves adaptability. The study highlights the need for privacy-aware training in VQA systems.

[3] **"Private Attribute Inference from Images with Vision-Language Models"**: VLMs can infer private details like income and location from images using techniques like advanced prompts and zooming. The VIP dataset shows models like GPT4-V achieving up to 77.6% accuracy in data infer-

ence, raising concerns about online privacy. Recommendations include better safeguards and user education.

[4] **"Analyzing Leakage of Personally Identifiable Information in Language Models"**: Language models like GPT-2 can leak PII even with measures like differential privacy (DP). Larger models and redundant PII increase risks. DP-SGD helps but doesn't fully prevent leakage. The study calls for improved privacy techniques to balance safety and utility.

[5] **"Sensitive Data Detection in Structured Datasets Using Large Language Models"**: LLMs outperform traditional methods in detecting sensitive data in structured formats like tables. Using CASSED, LLMs handle diverse data types effectively without much fine-tuning. Tested on datasets like UCI ML Students, this approach aids privacy compliance and scalable data management in multilingual environments.

III. Contribution

The primary goal of this research is to advance the field of sensitive data detection by leveraging the capabilities of Vision-Language Models (VLMs) to analyze multimodal content. This project contributes by addressing the following objectives:

1) **Creation of a Novel Dataset:** Build a detailed dataset combining examples of sensitive text and images. This dataset will support strong multimodal analysis and help fill the gap in high-quality labeled data needed for training and testing detection models.
2) **Evaluation of State-of-the-Art Techniques:** Review and compare current detection methods to see how well they identify sensitive information in both text and images. The analysis will highlight their strengths, weaknesses, and where they can be improved.
3) **Exploration of Vision-Language Models (VLMs):** Explore how Vision-Language Models (VLMs) can improve the detection of sensitive data by examining how well they combine and understand both images and text at the same time.
4) **Addressing Dataset and Model Challenges:** Suggest ways to improve datasets, such as handling the challenges of combining text and images, reducing mistakes in detection (false positives and negatives), and making the models work better across different areas and types of data.
5) **Advancing Practical Applications:** Contribute to the development of tools that are fast and easy to use in real-world settings, focusing on needs like real-time detection, following regulations, and protecting privacy.

By achieving these goals, this research aims to bridge the gap between academic advancements in sensitive data detection and practical industrial applications, driving innovation in privacy protection and data security.

IV. Background

What is a Vision-Language Model?

A Vision-Language Model (VLM) is an advanced machine learning model designed to process and understand both visual and textual data in an integrated manner, bridging computer vision (processing images or video) and natural language processing (understanding and generating text). Examples of VLMs include CLIP, BLIP, and GPT-4V, which are capable of tasks such as captioning, answering questions about images, and detecting relationships between text and visual content.

What are the Challenges in Sensitive Data Detection?

- **Complexity of Multimodal Data** – Unlike single-modal data, multimodal content combines visual and textual information, making it harder to define what constitutes "sensitive" information.
- **Unstructured Data** – Sensitive information in unstructured formats (e.g., free text, images, or videos) is harder to analyze than structured data.
- **Diversity of Sensitive Information** – Sensitive information can include Personally Identifiable Information (PII), financial data, health records, trade secrets, etc.
- **Large Volumes of Data** – Organizations often handle vast amounts of data, making real-time detection challenging.
- **Advanced Evasion Techniques** – Attackers use tricks like misspelling words, hiding information using special codes, or breaking data into smaller pieces to avoid being detected by security systems.
- **Balancing Detection and Privacy** – If detection systems are too strict, they might invade user privacy or collect too much personal data.
- **High False Positives/Negatives** – Machine learning models may flag non-sensitive data as sensitive or fail to identify truly sensitive data.
- **Transferability** – Models trained on one domain may not generalize well to other domains.
- **Computational Costs** – Detecting sensitive data, especially in high volumes, requires significant computational resources.
- **Lack of High-Quality Training Data** – uilding robust models demands extensive labeled datasets, which can be hard to obtain for sensitive data.
- **Dynamic Regulations** – Rules about handling sensitive data, like GDPR in Europe or HIPAA in the U.S., vary by location and keep changing over time.

Stages of Vision-Language Models:

1) **Input Encoding** –
   Visual Input: Images or videos are processed using computer vision techniques, often through a Convolutional Neural Network (CNN), Vision Transformer (ViT), or similar architecture to extract features such as objects, scenes, or patterns.
   Textual Input: Text is broken down into smaller pieces (tokens) and processed using models like transformers (e.g., BERT, GPT). These models help capture the meaning and structure of the text.

2) **Cross-Modal Alignment** – A key feature of Vision-Language Models (VLMs) is connecting visual and text data. This is done by creating a shared space where images and text are represented together, allowing the model to understand and work with both types of data at the same time.
3) **Contextual Understanding** – The model uses attention mechanisms to focus on important parts of the data, both within each type (like text or images) and between them, to better understand their connection and context.
   For example: The model finds connections between objects in an image (like specific areas or regions) and the text describing them.
4) **Task-Specific Processing** – Depending on the task (like image captioning, visual question answering, or sensitive data leakage detection), the model uses specific methods designed for that purpose:
   Image Captioning: Creates descriptive sentences for images by analyzing their key features.
   Visual Question Answering: Understands the question and looks at the image to figure out the right answer.
   Data Leakage Detection: Spots sensitive information by connecting visual details (like text in images) with specific concepts related to the topic.
5) **Reasoning and Decision-Making** – Models often use multi-layer transformers to understand connections and make decisions.
   For example: They check if a visual element matches a specific idea or concept in the text.
6) **Output Generation** – The model's output depends on the task:
   Text: For tasks like generating captions or answering questions.
   Labels: For identifying categories in classification tasks.
   Confidence Scores: For showing how certain the model is in detection tasks.
7) **Learning and Feedback** –
   Training: VLMs use large-scale datasets with paired image-text data (e.g., COCO, Flickr30k).
   Fine-tuning: Specialized datasets are used to improve the model's performance for specific tasks.
   Loss Functions: Techniques like contrastive loss, cross-entropy, or other task-specific methods are used to improve how well the model aligns data and makes accurate predictions.

Open-Source Frameworks for Vision-Language Models:

- **LLaVA: Large Language and Vision Assistant**
- **Qwen-VL: A Versatile Vision-Language Model**
- **CogVLM: Visual Expert for Pretrained Language Models**
- **Phi-3-Vision**

For the frameworks discussed in this section click here.

Limitations:

- **Shallow Semantic Understanding:** Vision-Language Models (VLMs) can have difficulty understanding subtle connections between objects in an image and the text, such as recognizing sarcasm or cultural references in combined visual and text content.
- **Unclear Descriptions:** VLMs can struggle with vague or unclear text, especially when the text doesn't clearly explain what's in the image.
- **Domain-Specific Challenges:** VLMs often perform poorly in areas they weren't specifically trained for, like medical imaging or other specialized tasks.
- **Overfitting:** These models might depend too much on patterns in their training data instead of learning to handle new, unseen situations effectively.
- **Data Quality Problems:** Many VLMs are trained on messy web data, which can include incorrect or inappropriate matches between text and images.
- **High Resource Requirements:** Training and improving these models requires a lot of computing power, which makes it hard for many researchers or organizations to use them.
- **Temporal Reasoning:** They struggle with temporal aspects in videos, such as predicting sequences of events.
- **Black-Box Nature:** It's hard to understand how these models make decisions, which makes it difficult to trust or explain their results in important areas like healthcare or self-driving systems.
- **One-Directional Understanding:** Many VLMs work better when translating from images to text (e.g., generating captions) than vice versa (e.g., generating images from text with high semantic fidelity).
- **Non-Verbal Signals:** These models struggle to understand non-verbal signals like gestures, emotions, or creative expressions like abstract art, which require deeper personal interpretation.
- **Vulnerability to Tricks:** Even small changes to the input image or text can confuse these models, making them less reliable for important tasks.
- **Scalability Issues:** Processing large datasets with privacy-sensitive content.
- **Accuracy Trade-offs:** Balancing high detection accuracy with low false positives.

REFERENCES

[1] L. Samson, N. Barazani, S. Ghebreab, and Y. M. Asano, "Privacy-Aware Visual Language Models," arXiv preprint arXiv:2405.17423, 2024.

[2] F. Pinto, N. Rauschmayr, F. Tramer, P. Torr, and F. Tombari, "Extracting Training Data from Document-Based VQA Models" Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, PMLR 235, 2024.

[3] B. Tömekçe, M. Vero, R. Staab, and M. Vechev, "Private Attribute Inference from Images with Vision-Language Models," 38th Conference on Neural Information Processing Systems (NeurIPS), Vienna, Austria, 2024.

[4] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguélin, "Analyzing Leakage of Personally Identifiable Information in Language Models," IEEE Security and Privacy (S&P), 2023.

[5] A. Anand, “Sensitive Data Detection in Structured Datasets Using Large Language Models,” TechRxiv preprint, 2024.
[6] Md. Faruk Alam, “Top Open-Source Vision-Language Models (VLMs),” LinkedIn Pulse, 2024. Available: https://www.linkedin.com/pulse/top-open-source-vision-language-models-vlms-md-faruk-alam-rugvc/. [Accessed: Nov. 25, 2024].