Sensitive Data Leakage Detection Based on Vision-Language Models

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

I. 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”**: One of the most directly relevant works is “Privacy-Aware Visual Language Models”, which introduces dedicated benchmarks such as PRIVBENCH and PRIVBENCH-HARD for evaluating the privacy capabilities of Vision-Language Models (VLMs). The study demonstrates that VLMs like MiniGPT-v2 and TinyLLaVa can be fine-tuned using a privacy-specific dataset (PRIVTUNE) to accurately distinguish between sensitive and non-sensitive visual-textual inputs, such as passports versus public scenes. This aligns closely with the goals of our project, which also focuses on training VLMs to detect sensitive information in multimodal content. Their approach validates the potential of fine-tuning VLMs for privacy-focused tasks and highlights methodologies for building effective privacy benchmarks, both of which inform and support our own work.

[2] **”Extracting Training Data from Document-Based VQA Models”**: The study “Extracting Training Data from Document-Based VQA Models” investigates how Vision-Language Models (VLMs) used in Visual Question Answering (VQA) systems can unintentionally leak sensitive information memorized during training. The authors propose techniques such as counterfactual analysis and Extraction Blocking (EB) to reduce these risks by ensuring that answers are only generated when sufficient contextual evidence exists. Although the focus is on preventing leakage at the model behavior level, the work directly relates to our project, which aims to proactively detect sensitive information in multimodal inputs. Both studies operate in the domain of document-based VLMs and address the broader issue of privacy in models that process visual-textual data, reinforcing the importance of developing safeguards around sensitive data in real-world systems.

[3] **”Private Attribute Inference from Images with Vision-Language Models”**: One closely related study is “Private Attribute Inference from Images with Vision-Language Models”, which highlights a critical privacy concern in modern multimodal systems. The authors demonstrate that Vision-Language Models (VLMs), such as GPT-4V, can infer sensitive personal attributes like income level or geographic location from seemingly non-sensitive images. By using techniques such as advanced prompting and zooming into image details, the models achieved high inference accuracy, raising serious concerns about unintentional privacy violations. This work is highly relevant to our project, as it underscores the need not only to detect explicitly sensitive content but also to consider how VLMs may leak sensitive information through visual or contextual inference. Our project aims to address this issue by building detection models that can proactively identify and flag content that poses a risk of sensitive data exposure, either directly or indirectly.

[4] **”Analyzing Leakage of Personally Identifiable Information in Language Models”**: Another relevant study is “Analyzing Leakage of Personally Identifiable Information in Language Models”, which examines how large language models (LLMs) like GPT-2 can unintentionally reveal sensitive data, such as names or phone numbers, even when trained with privacy-preserving techniques like Differential Privacy (DP). The research shows that larger models and repeated exposure to PII in training data significantly increase the risk of leakage. While this work focuses on text-only models, it highlights an important concern that extends to Vision-Language Models (VLMs) as well — especially when dealing with multimodal content where sensitive information may be encoded in both text and images. This paper reinforces the need for robust detection and mitigation strategies, supporting our project’s goal of proactively identifying and preventing sensitive data leakage in VLMs.

II. Dataset

To support the training and evaluation of Vision-Language Models (VLMs) for sensitive data detection, we created a comprehensive synthetic dataset consisting of labeled multimodal examples. The dataset is divided into two main categories: sensitive and non-sensitive information, with a total of approximately 300,000 samples.

1) **Data Sources:**
Due to the privacy-sensitive nature of the task, all data was generated synthetically using custom scripts and publicly available tools such as Faker. The goal was to simulate realistic content commonly found in personal and financial documents without using any

real user information. Images were created to resemble actual documents (e.g., bank statements, phone bills, medical letters, credit cards), and paired with text strings containing sensitive or non-sensitive information.

2) **Dataset Composition:**

Sensitive Information Dataset

- 20,000 images of bank emails containing PIN codes for credit cards.
- 20,000 images of credit cards (with names, card numbers, expiration dates).
- 20,000 images of medical appointment letters containing personal medical records.
- 20,000 images of phone bills, including customer details and billing data.

Non-Sensitive Information Dataset

- 20,000 images of generic or template-style letters.
- 20,000 images of job interview invitations or administrative emails.
- 20,000 images of online reviews.

3) **Preprocessing and Labeling:**
The sensitive images were normalized to a consistent resolution for model compatibility, and minor augmentations (e.g., noise, font variation, slight distortions) were applied to increase diversity and robustness.
The non-sensitive images have not yet undergone this process, as we initially attempted to train the model using only sensitive images. Later, we realized that the model would need to be trained on both sensitive and non-sensitive images together. In the future, when we train the model to do so, each image will be labeled as either "sensitive" or "non-sensitive".
The textual part of each example, extracted from the JSON files, is stored in a natural format that reflects how such data would appear in real-world scenarios (e.g., "SSN: 966-73-1658", "Account Number: 0499704"). This format enables direct use in Vision-Language Models without the need for further caption engineering.

4) **Usage:**
The dataset was used to fine-tune and evaluate Vision-Language Models, beginning with OpenAI's CLIP model. It is structured to support image-text alignment and supervised classification.

III. Results

This stage of the project focused on validating the dataset and conducting the first round of fine-tuning using the BLIP Vision-Language Model. While the model's outputs are still limited in quality, the process revealed promising trends, demonstrated the dataset's effectiveness, and offered clear insights for future development.

1) **Fine-Tuning Process:**
We fine-tuned the Salesforce/blip-image-captioning-base model using a balanced subset of 72,000 image-text examples across four sensitive categories: credit card, medical letter, phone bill, and PIN code notifications. Validation was performed on an additional 8,000 examples.
Key training configuration details:

- Training method: LoRA (Low-Rank Adaptation of weights)
- Total training steps: 54,000
- Epochs completed: 3
- Trainable parameters: 737K (out of 248M total)
- Average training loss after 3 epochs: 4.28 (down from 7.57 initially)
- Gradients: Showed high early variance and later stabilized (peak $\sim$ 29.0, ending $\sim$ 5–10)

The consistent downward trend in training loss over time indicates that the model was learning signal from the data and adapting gradually to the sensitive document domain.

2) **Caption Comparison - Base vs. Fine-Tuned:**
After training, we performed a qualitative comparison between the base model and the fine-tuned model. For a test image showing a credit card:

- Base Model Output: "a black credit card sitting on top of a wooden table"
  – Descriptive, but lacks awareness of sensitivity.
- Fine-Tuned Model Output: "this image contains sensitive data"
  – More privacy-aware and aligned with the detection objective.

Another test output included:
"this image contains credit card".
– indicating that the model is beginning to understand and identify sensitive entities directly, even if the phrasing remains rough.

3) **Interpretation:**
While the outputs are not yet production-ready, the following trends were observed:

- The model has shifted from generic visual descriptions to privacy-aware phrasing
- It can now recognize sensitive concepts like "credit card" or "sensitive data"
- Captions are still grammatically weak, but structurally aligned with the task

These findings validate the model's learning path and suggest that further training and prompt-based fine-tuning can improve the quality and specificity of detection.

IV. Contribution

This research advances the field of sensitive data detection by leveraging Vision-Language Models (VLMs) to analyze multimodal content, focusing on practical and scalable solutions for identifying sensitive information embedded in text and images. The main contributions of this work include:

1) **Development of a Large-Scale Synthetic Dataset:** We created a high-quality dataset consisting of 140,000 image-text pairs, carefully designed to include both sensitive (e.g., credit cards, medical letters, phone bills) and non-sensitive (e.g., reviews online, job interview invitations) examples. Each item is labeled and annotated for supervised learning with VLMs.
2) **Evaluation of Vision-Language Models for Sensitivity Detection:** We selected CLIP as the base Vision-Language Model and began fine-tuning it on our dataset to evaluate its ability to distinguish between sensitive and non-sensitive multimodal content.
3) **Benchmarking Detection Challenges:** Through early experiments and dataset analysis, we identified key issues such as visual-textual imbalance, annotation noise, and high false positive rates. These insights guide our ongoing refinement of both dataset and model pipeline.
4) **Designing Multimodal Detection Frameworks:** We constructed a training-ready data pipeline that pairs realistic synthetic images with their corresponding structured text annotations. Each annotation includes sensitive content (e.g., credit card numbers, medical records, phone bills) in a human-readable format. This format enables direct input into Vision-Language Models (VLMs), supporting efficient multimodal processing for sensitive data detection.

By addressing both academic and applied aspects of sensitive data detection in multimodal systems, this project contributes a novel dataset, a fine-tuned VLM pipeline, and an analysis framework that collectively push forward the capabilities of privacy-aware AI systems.

V. Summary

The project has successfully laid the foundation for detecting sensitive information using Vision-Language Models (VLMs). A high-quality synthetic dataset was created, featuring over 140,000 realistic examples of sensitive and non-sensitive documents. These were annotated with structured text and used to train and evaluate a fine-tuned BLIP model.

Through a carefully managed training process using LoRA, the model completed 3 epochs over 54,000 steps. The training loss decreased from 7.57 to 4.28, indicating consistent learning. The model evolved from basic visual captioning to generating outputs that reflect awareness of sensitive content. Examples include identifying documents as containing “sensitive data” or explicitly mentioning the presence of a “credit card.”

These early successes confirm that the model is learning to understand privacy-related context, though it still requires refinement in grammar and specificity. The results validate the potential of VLMs for multimodal privacy analysis and highlight the effectiveness of the dataset.

What Remains to Be Completed:

- **Continue Fine-Tuning:** Train for additional epochs, improve regularization, and refine hyperparameters.
- **Expand Dataset Usage:** Include both sensitive and non-sensitive samples for binary classification.
- **Integrate a Classification Head:** Add a dedicated component to label images as sensitive/non-sensitive.
- **Evaluate with Formal Metrics:** Use BLEU and ROUGE for caption quality, and Precision/Recall/F1 for classification tasks.

VI. Github

https://github.com/Yuval10Dahan/Final-Project-CS

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] 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].