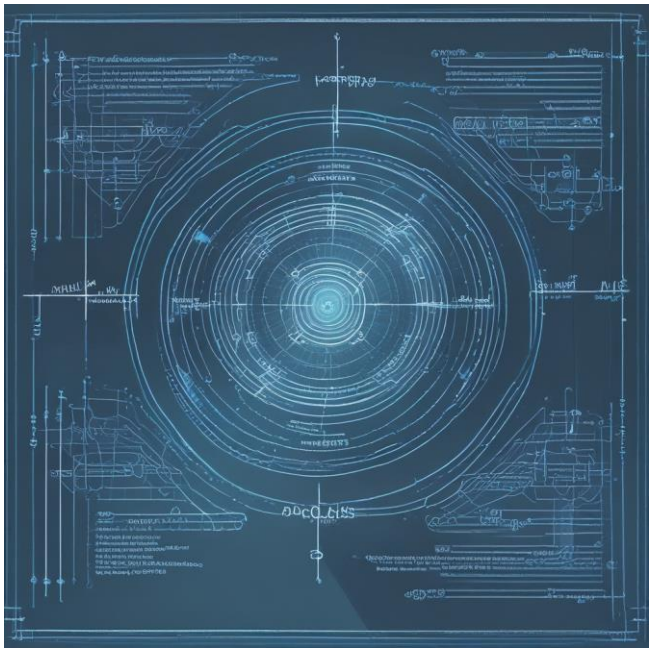

GCSP THEME PAPER

Theme: Security

ROLE OF EMERGING TECH IN SECURITY

Exploring the Future of Security with Emerging Technologies like Machine Learning and Quantum Computing.

Aryan Keluskar

For FSE 150

Introduction

This paper is an analysis of three research projects that investigate the future of security in this age of constant digital innovation. Since this paper is being written for the GCSP theme of security, the research topics will include a rather inclusive definition of security, extending beyond the traditional notion of data security. As we continue to rely on digital platforms for various aspects of our lives, it is important that one should understand and address these security challenges. The increasing complexity of security threats, new forms of malware, and the ever-growing amount of digital content requires a comprehensive and proactive approach to ensure the security and integrity of digital systems in our lives and in our society.

With the increasing reliance on digital platforms for communication, young social media users are exposed to potential online risks, harassment, and inappropriate content. Therefore, the first research project will evaluate the recent work in utilizing the emerging technology of Large Language Models in safeguarding young children in their online conversations. As AI and LLMs become more prevalent, concerns about the security of ML models deployed with publicly accessible query interfaces have emerged because their sensitive training data is just one prompt away. The second research project in this paper investigates vulnerabilities and suggests countermeasures related to such attacks. On the physics end however, a new problem arises from the world that no one can see yet feel disastrous effects of. The rapid development of quantum computing poses a significant threat to the security of cryptography for a few reasons. This implies all our data could have been rendered decrypted, however research in the third research project show that there is significant development in finding quantum-immune encryption algorithms.

Research Project 1: Novel Machine Learning to Safeguard Children Online.

First research project addresses a critical issue of our time, building safer message inboxes for young social media users. Machine learning is increasingly becoming a pivotal technology in enhancing security and user experience on social media platforms, one such instance being detecting unsafe conversations within Instagram Direct Messages for its young users. A research titled "A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth"[1] was conducted to investigate the development of a system that can automatically protect young social media users from potential online risks, harassment and exposure to inappropriate content. This research contributes to a safer online environment for vulnerable populations. It was conducted in a collaboration between researchers from Boston University, Georgia Institute of Technology, and Vanderbilt University.

The core purpose of the research project was to test multiple feature sets by going through the metadata, picking on the linguistic cues of the messages, and analyzing image features to differentiate explicit content[2]. This would then be ultimately used to train classifiers and detect risky conversations. By analyzing patterns in the language and context of content on the social media platform, the models learnt to identify potentially offensive content automatically. Once identified, such messages and comments can be hidden or removed, thereby preventing them from reaching the innocent audience of young children. Before the advent of such large language models, social media platforms primarily relied on user reports and manual moderation to detect and remove unsafe conversations. This approach had multiple limitations. Manual moderation is a reactive approach that requires immediate attention once a report is made. This can lead to delays in addressing harmful content, as moderators must review and assess each report before taking action. This also implies giving up sensitive and personal data to a human. Automated risk-detection models

can identify and flag offensive content in real-time, while preventing the personal messages from ever being seen by any individual. Bias can also be a concern in the implementation of manually moderated inboxes. For instance, moderators may unintentionally introduce bias due to their personal beliefs or values, leading to false positives or negatives. To battle such limitations, this research was conducted.

This research consists of several stages, including data collection, model development, and system evaluation. The researchers collected anonymized and de-identified data from Instagram Direct Messages for users aged 13-17, who provided consent for the study. For text analysis, the researchers used sentiment analysis, topic modeling, and named entity recognition. Most of these techniques were made possible with Natural Language Processing libraries like NLTK, SpaCy, and Gensim which allowed the models to understand the context and meaning of the text to identify potential risks. LLMs are a type of machine learning model that can learn the patterns and structures of language by analyzing large amounts of text data. By incorporating LLMs into this research's text analysis process, the researchers aimed to improve the models' ability to understand the context and meaning of the text messages, thereby enhancing the accuracy of potential risk identification.

Sentiment analysis, in particular, plays a crucial role in detecting the emotional tone of the messages, which can be indicative of potential risks or harmful behaviors. It has become a significant security advantage in social media due to its ability to analyze and learn individual reactions to deduce trends and user needs, both subjectively and objectively.

Figure 1.1 – Sentiment Analysis on text using tokenization [3]

For analyzing the images sent in Instagram DMs, the researchers used convolutional neural networks (CNNs). The CNNs were trained to identify potentially inappropriate content that include explicit or harmful images, which could pose a risk to young users. The researchers evaluated the performance of the multimodal detection system using metrics like accuracy, precision, recall, and F1 score. In the end, researchers used user study tools such as Qualtrics and SurveyMonkey to conduct user surveys and gather feedback.

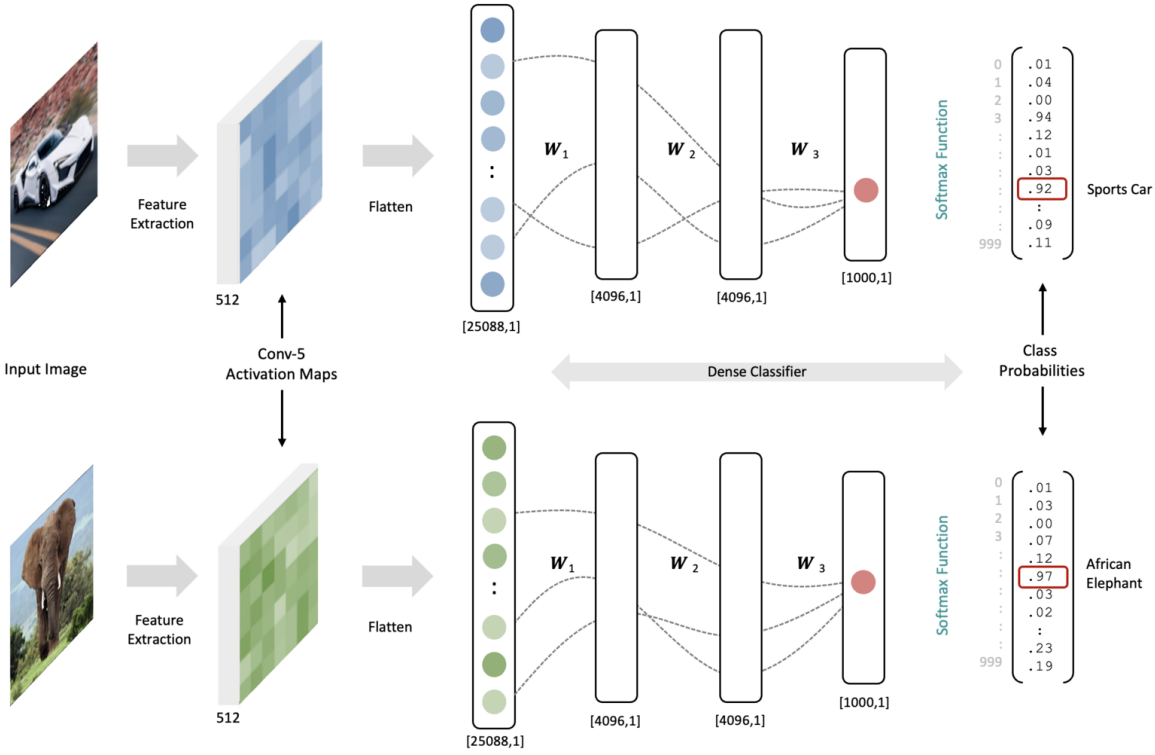

Figure 1.2 – How a CNN detects features from an image, and then uses probability to find matching elements in a new image. [4]

Researchers found that features based on metadata information work well when used to develop risk classifiers. They also found that the contextual information provided by text and images is crucial to develop accurate classifiers. Through a combination of both metadata traits and linguistic cues, the program achieved a staggering 85% accuracy to identify risky conversations. The project's findings and outcomes are contributing significantly to the protection of young users on social media by offering a much-needed insight into developing

protection systems to mitigate risks associated with online interactions. The GC Theme of Security revolves around maintaining a holistic security of an individual, which certainly includes digital security. While this model excels at preventing scams in one's inbox, its primary contribution is in securing the emotional well-being of young children by preventing inappropriate content in their personal inbox.

Research Project 2: Preventing Model Extraction Attacks.

In today's world powered by LLMs and AI Chatbots, it is certainly scary to imagine being able to access the sensitive data behind the friendly chat window. Well, researchers at The University of North Carolina at Chapel Hill and Cornell University expose ways to break into the chatbot model in their paper titled 'Stealing Machine Learning Models via Prediction APIs'[5]. The research fills a gap in existing work by focusing on the security of ML models deployed with publicly accessible query interfaces which contain sensitive training data of commercial value. Since most of the AI developers and startups tend to overlook penetration testing to speed up development, there are traces of vulnerabilities left. Compared to existing work and solutions, this research project provides a more comprehensive understanding of the vulnerabilities and countermeasures related to Model Extraction attacks. Such attacks are used to extract the parameters from a target model, that can be used to steal proprietary ML models, which can have significant commercial value[6]. They can also be used to compromise the privacy of the data used in the target model. The goal of the research project is to highlight the steps for careful ML model deployment and suggests countermeasures.

The researchers selected popular model classes, like logistic regression, neural networks, and decision trees, to demonstrate the feasibility of model extraction attacks. The researchers started by conducting background research on ML-as-a-service offerings, which allow users to train models on potentially sensitive data and charge others to access the fine-

tuned model on a pay-per-query basis. They identified the practices of accepting partial feature vectors as inputs and including confidence values with predictions, which are common in ML-as-a-service offerings. The researchers then identified the vulnerabilities in these platforms, which allow for model extraction attacks. They demonstrated that even the well-known countermeasure of omitting confidence values from model outputs still admits potentially harmful model extraction attacks. With techniques such as adding small perturbations to the input data, causing the model to make a wrong prediction, and using the confidence values provided by the model to estimate the gradient of the loss function, they were able to extract the target information from the ML models. The researchers demonstrated these attacks against the online services of BigML and Amazon Machine Learning, showcasing the potential wide-spread implications of the possibility of such attacks.

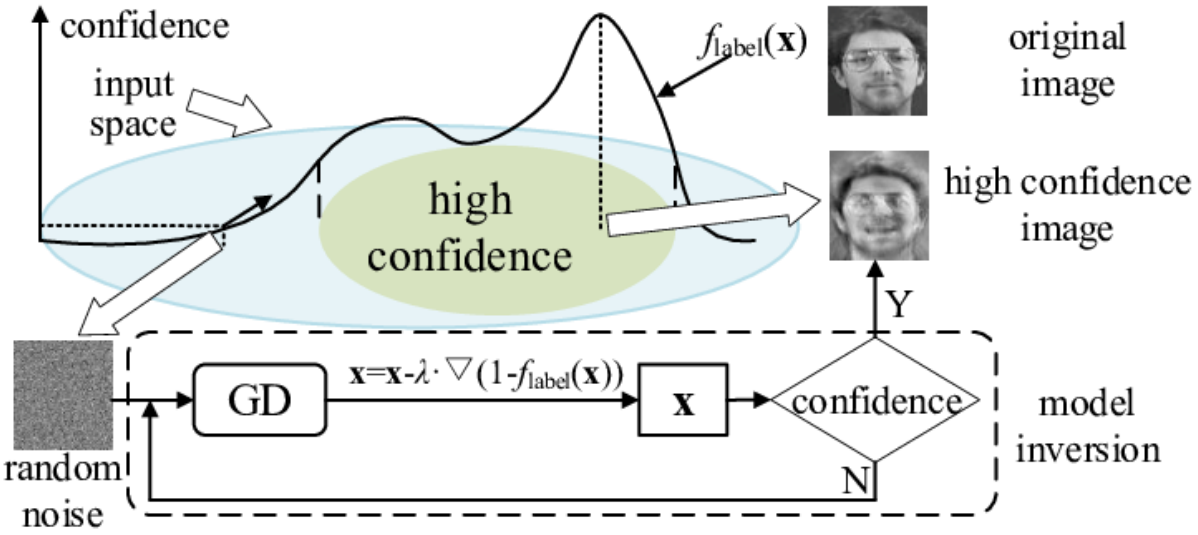

Figure 2.1 – Reconstructing an image from the training data using model extraction attack [7]

To conclude this research project, researchers provided countermeasures to prevent such model extraction attacks. One potential defense mechanism against such attacks is the rounding of confidence scores to a fixed precision. Researchers have shown that limiting precision in this manner can have a significant impact on the success of model extraction

attacks. Similarly, decision tree pathfinding attacks become less successful when confidence scores are rounded instead of precisely reported, as the latter increases the likelihood of node collisions and decreases the attack's success rate. Differential privacy has been explored as mechanisms for protecting ML training data privacy. By applying DP directly to model parameters, the research project finds that a query does not allow an adversary to distinguish between closely neighboring model parameters, preventing unnecessary data leakage.

Research Project 3: Saving our data from a widespread encryption failure.

Looking into the future again from a different lens reveals a problem imported from physics. The rapid development of quantum computing poses an unimaginable threat to the security of cryptography. Today, we rely on the computational hardness of mathematical problems to encrypt sensitive data, but quantum computers can efficiently break this because they have the ability to search for a needle in a haystack within seconds. Our secure world was built upon encryption schemes like RSA and AES, but decrypting them means posing a risk to the confidentiality and integrity of data in online banking, e-commerce, digital signatures, and blockchain transactions [9]. In fact, multiple countries are already storing military data to decrypt them at a later stage with a quantum computer [10]. In response to this threat, researchers are working on quantum-resistant cryptography. One such research project was conducted at ASU, titled ‘Implementation and Analysis of Quantum Homomorphic Encryption’[8]. Its focus is on the technique involving homomorphic encryption, a type of encryption that allows computations to be performed on encrypted data without decrypting it, minimizing the possibility of free data to float freely on public internet. It was conducted by SenSIP Center at School of ECEE, Arizona State University. Homomorphic Encryption continues to show promising potential, yet its quantum-resistant version remains under-researched. Therefore, this research project aims to bridge that gap and provide much-needed and well-researched information into encryption of the future.

The quantum homomorphic encryption scheme was implemented on a quantum computer built using a quantum teleportation circuit. The researchers designed quantum circuits to perform homomorphic encryption on the teleportation circuit. Researchers propose a private-key quantum homomorphic encryption scheme that uses the centralizer of a subgroup of operations. The quantum data is then encoded on bosons of distinct species in distinct spatial modes, and the quantum computations are manipulations of these bosons in a manner independent of their species. For context, Bosons are one of the two quantum fundamental particles, which act as “force carriers” between particles.

The quantum homomorphic encryption scheme was then tested on the teleportation circuit, and the results were analyzed to evaluate the scheme's performance, security, and feasibility for practical applications. The researchers demonstrate that a particular instance of this data encoding can hide up to a constant fraction of the information encrypted[11]. This highlights that implementing a protocol (like we have TCP/IP for the internet today) can hide large amounts of information, while maintain its encoding. The research project is expected to contribute to implementation of quantum homomorphic encryption, which will strengthen our security in the age of quantum-computers. Such an implementation can be used in various applications, such as protection of data on the cloud, analyzing sensitive user without exposing it, and ultimately ensuring data privacy in an era when all of our present encryption schemes have broken.

Conclusion

The three research projects analyzed in this paper highlight significant societal challenges and impacts in the development and implementation of security solutions in the ever innovative digital age. The first project addresses the challenge of safeguarding young children in their online conversations by utilizing large language models (LLMs) to detect unsafe conversations. By leveraging the power of machines to understand human language, we are effectively combating the social challenge of children being exposed to inappropriate content online. In the United States and many other countries, it is illegal to show indecent media to someone under the age of 18. Therefore, this is not only good for their mental well-being, but also upholds the legal responsibility of these social media platforms. The challenge lies in the economic cost of running these models all the time, which decreases company's profits. However, a more crucial challenge lies in the risk of false positives leading to the censorship of harmless content, or the potential for biased moderation based on the values and beliefs inherited by LLMs from its training data.

The second project focuses on the vulnerabilities and countermeasures related to model extraction attacks on models deployed with publicly accessible query interfaces. One negative impact is the potential for increased surveillance. The societal challenge here is the potential theft of sensitive training data and proprietary models, which can have significant economic loss and compromise the privacy of the users' data being used to train the model. The countermeasures suggested in this research project may lead to increased monitoring on the AI websites and may lead to tracking of individuals' online activities. Some models are be too complex or too large to implement the suggested countermeasures, making them vulnerable to model extraction attacks. Finally, the suggested countermeasures may have unintended consequences. For example, limiting precision in confidence scores could lead to false positives or false negatives, which could result in lowered accuracy of the model. This

lowered accuracy could increase the bias of a model, or even lead to the model being unusably inaccurate for certain tasks.

The third project tries to navigate encryption through a sea of quantum computers. The societal challenge here is the potential for all data to be rendered decrypted by quantum computers, posing a risk to the confidentiality and integrity of data in various applications. Therefore, the research project works to implement and evaluate new encryption schemes. Since quantum-resistant cryptography is still in its very early stages of development, it is a bit too early to fully understand its impacts. Another negative may be that using a new encryption scheme may have undiscovered consequences, especially when done on a technology of which we have limited knowledge presently. This approach is expected to be computationally expensive, and all the encrypted data that is stored up till now is already on the risk of decryption in the near future. Therefore, this paper does take a step towards preventing any privacy risks in the future, however it does not mitigate the risk of present data being decrypted and misused.

References and Related Works

[1] S. Ali et al., “Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth Shiza Ali et al. ACM Reference Format,” doi: https://doi.org/10.1145/3579608.

[2] “Machine Learning Can Help to Flag Risky Messages on Instagram While Preserving Users’ Privacy,” drexel.edu, Apr. 17, 2023. https://drexel.edu/news/archive/2023/April/metadata-instagram-risky-conversation-detection (accessed Mar. 23, 2024).

[3] “NLTK Sentiment Analysis Tutorial: Text Mining & Analysis in Python,” www.datacamp.com. https://www.datacamp.com/tutorial/text-analytics-beginners-nltk

[4] LearnOpenCV, “Understanding Convolutional Neural Networks: A Complete Guide,” learnopencv.com, Jan. 18, 2023. https://learnopencv.com/understanding-convolutional-neural-networks-cnn/

[5] W. Melicher et al., “Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks,” 2016. Available: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_melicher.pdf

[6] A. Yan, T. Huang, L. Ke, X. Liu, Q. Chen, and C. Dong, “Explanation leaks: Explanation-guided model extraction attacks,” vol. 632, pp. 269–284, Mar. 2023, doi: https://doi.org/10.1016/j.ins.2023.03.020.

[7] J. Zhang, C. Li, J. Ye, and G. Qu, ‘Privacy Threats and Protection in Machine Learning’, 09 2020.

[8] M. Yarter, G. Uehara, and A. Spanias, “Implementation and Analysis of Quantum Homomorphic Encryption,” Jul. 2022, doi: https://doi.org/10.1109/iisa56318.2022.9904399.

[9] R. Trent, “The Challenges and Benefits of Quantum Computing Security,” Rod’s Blog, Feb. 19, 2024. https://rodtrent.substack.com/p/the-challenges-and-benefits-of-quantum

[10] J. Duran, “Guest Post: Harvest Now, Decrypt Later? The Truth Behind This Common Quantum Theory,” The Quantum Insider, Feb. 07, 2023. https://thequantuminsider.com/2023/02/07/guest-post-harvest-now-decrypt-later-the-truth-behind-this-common-quantum-theory/

[11] S.-H. Tan, J. A. Kettlewell, Y. Ouyang, L. Chen, and J. F. Fitzsimons, “A quantum approach to homomorphic encryption,” Scientific Reports, vol. 6, no. 1, p. 33467, Sep. 2016, doi: https://doi.org/10.1038/srep33467.