الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08٠ ماي 5491٠ بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès

THESIS

To obtain the diploma of **Engineering Degree**
Field: **Computer Science**
Specialty: **Intelligence Artificielle et Science de Données (IASD)**

Theme

Enhancing Software Security Using Transformer-Based Language Models

---

Presented by:
Abdechakour MECHRI

Submission Date: **Sept, 2024**
In front of the jury composed of:

Dr. DIF Nassima	President
Dr. BOUSMAHA Rabab	Examiner
Dr. KHALDI Belkacem	Supervisor
Dr. FERRAG Mohamed Amine	Co-Supervisor

Academic Year : 2023/2024

Abstract

The increasing reliance on software in various domains has heightened the importance of software security, necessitating innovative approaches to protect applications from malicious attacks and other cyber threats. This thesis, titled "Enhancing Software Security Using Transformer-Based Language Models," explores the potential of transformer models to improve software security. Our primary objectives include developing algorithms that leverage transformer models for effective vulnerability detection and analysis, automating threat detection and mitigation, and evaluating the practical application of these models in real-world scenarios.

Motivated by the need to enhance the resilience of software systems against evolving cyber threats, we propose utilizing advanced machine learning techniques to understand, predict, and mitigate security vulnerabilities. Our methodology encompasses the creation of high-quality labeled datasets, and the implementation of state-of-the-art transformer-based model for detecting software vulnerability. Through this approach, we achieved fine-grained line-level detection and multi-class classification of vulnerabilities, contributing to the development of more effective and trustworthy security solutions.

Our findings demonstrate that transformer-based models significantly enhance software vulnerability detection, providing insights and methodologies for integrating these models into existing security practices. By deploying our model in a real-world scenario, we showcased its practical utility, highlighting its potential to assist analysts in safeguarding software systems. This work advances the field of cybersecurity by offering a robust and scalable solution to the persistent challenges of software security.

Keywords— Static Analysis, Vulnerability Detection, Codebase, Large Language Model, Software Security, Security, Generative Pre-trained Transformers

الملخص

لقد زادت أهمية أمان البرمجيات نتيجة الاعتماد المتزايد على البرمجيات في مختلف المجالات، مما يتطلب ابتكارات جديدة لحماية التطبيقات من الهجمات الضارة وغيرها من التهديدات السيبرانية. تستكشف هذه الأطروحة، بعنوان "تعزيز أمان البرمجيات باستخدام نماذج اللغة القائمة على المحولات"، إمكانيات استخدام نماذج المحولات لتحسين أمان البرمجيات. تشمل أهدافنا الرئيسية تطوير خوارزميات تستفيد من نماذج المحولات لاكتشاف وتحليل الثغرات الأمنية بفعالية، وأتمتة عملية اكتشاف التهديدات والتخفيف من حدتها، وتقييم التطبيق العملي لهذه النماذج في سيناريوهات العالم الحقيقي.

انطلاقًا من الحاجة إلى تعزيز مرونة أنظمة البرمجيات ضد التهديدات السيبرانية المتطورة، نقترح استخدام تقنيات التعلم الآلي المتقدمة لفهم وتوقع والتخفيف من الثغرات الأمنية. تشمل منهجيتنا إنشاء مجموعات بيانات ذات جودة عالية، وتنفيذ نموذج قائم على المحولات لاكتشاف الثغرات البرمجية. من خلال هذا النهج، حققنا اكتشافًا دقيقًا على مستوى الأسطر وتصنيفًا متعدد الفئات للثغرات، مما ساهم في تطوير حلول أمان أكثر فعالية وموثوقية.

تظهر نتائجنا أن النماذج القائمة على المحولات تعزز بشكل كبير اكتشاف الثغرات البرمجية، مما يوفر رؤى ومنهجيات لدمج هذه النماذج في ممارسات الأمان الحالية. من خلال نشر نموذجنا في سيناريو واقعي، أظهرنا فائدته العملية، مؤكدين على قدرته على مساعدة المحللين في حماية أنظمة البرمجيات. يعزز هذا العمل مجال الأمن السيبراني من خلال تقديم حل قوي وقابل للتطوير للتحديات المستمرة في أمان البرمجيات.

الكلمات المفتاحية— التحليل الثابت، اكتشاف الثغرات، قاعدة الكود، نموذج اللغة الكبيرة، أمان البرمجيات، الأمن، المحولات التوليدية المدربة مسبقًا.

Résumé

L'importance croissante de la sécurité des logiciels, rendue nécessaire par la dépendance accrue à ces derniers dans divers domaines, appelle à des approches innovantes pour protéger les applications contre les attaques malveillantes et autres menaces cybernétiques. Cette thèse, intitulée "Améliorer la sécurité des logiciels à l'aide des modèles de langage basés sur les transformateurs", explore le potentiel des modèles de transformateurs pour améliorer la sécurité des logiciels. Nos objectifs principaux incluent le développement d'algorithmes exploitant les modèles de transformateurs pour une détection et une analyse efficaces des vulnérabilités, l'automatisation de la détection et de la mitigation des menaces, et l'évaluation de l'application pratique de ces modèles dans des scénarios réels.

Motivés par la nécessité d'améliorer la résilience des systèmes logiciels face aux menaces cybernétiques évolutives, nous proposons d'utiliser des techniques avancées d'apprentissage automatique pour comprendre, prédire et atténuer les vulnérabilités de sécurité. Notre méthodologie englobe la création de jeux de données étiquetés de haute qualité, et la mise en œuvre d'un modèle basé sur les transformateurs pour la détection des vulnérabilités logicielles. Grâce à cette approche, nous avons réalisé une détection fine et au niveau des lignes ainsi qu'une classification multiclasses des vulnérabilités, contribuant au développement de solutions de sécurité plus efficaces et fiables.

Nos résultats démontrent que les modèles basés sur les transformateurs améliorent significativement la détection des vulnérabilités logicielles, fournissant des informations et des méthodologies pour intégrer ces modèles dans les pratiques de sécurité existantes. En déployant notre modèle dans un scénario réel, nous avons mis en avant son utilité pratique, soulignant son potentiel à aider les analystes à protéger les systèmes logiciels. Ce travail fait progresser le domaine de la cybersécurité en offrant une solution robuste et évolutive aux défis persistants de la sécurité des logiciels.

Keywords— Analyse statique, Détection des vulnérabilités, Code source, Grand modèle de langage, Sécurité des logiciels, Sécurité, Transformateurs génératifs pré-entraînés