LINKING COMMON VULNERABILITIES AND EXPOSURES TO THE MITRE ATT&CK FRAMEWORK: A SELF-DISTILLATION APPROACH

By Benjamin Ampel, Dr. Sagar Samtani, Steven Ullman, and Dr. Hsinchun Chen
Introduction: CVEs

- Harmful cyber-attacks on critical cyber-infrastructure (e.g., large servers hosting confidential data) have cost on average $7.91 million per breach, leading to over 446,000,000 exposed records containing sensitive information in 2019 (Sun et al., 2020).
  - Thus, it is imperative to build our cybersecurity knowledge base to combat new and evolving cyber-threats.

- One key piece of the cybersecurity knowledge base is the Common Vulnerability and Exposures (CVE) list, overseen by the MITRE Corporation.
  - A new CVE is created whenever a security flaw is discovered and reported to MITRE.

- However, CVEs often provide little information on how to combat the vulnerability before it is discovered in an organization’s cyber-infrastructure.
In 2018, MITRE created a new cybersecurity risk management framework (CRMF), the ATT&CK Matrix for Enterprise. This matrix aims to model the tactics, techniques, and procedures (TTP) that an attacker would take when attempting to breach cyber-infrastructure (Strom et al., 2018).

There are currently 14 tactics that an attacker may use to conduct a cyber-attack, including prominent ones like “initial access,” “defense evasion,” and “exfiltration.”

Each tactic and technique comes with a mitigation strategy (e.g., user training, account management, password policies, etc.) to assist cybersecurity analysts in protecting critical cyber-infrastructure.
Introduction: Research Approach

• Despite the tremendous benefits that both CVEs and the ATT&CK framework can provide for key cybersecurity stakeholders (e.g., analysts, educators, and managers), the two entities are currently separate.

• With over 158,000 CVEs existing as of the beginning of 2021, it would be a non-trivial task to manually link each one to the ATT&CK framework to gather mitigation strategies for every existing CVE.
Introduction: Research Approach

• In this study, we aim to develop a novel framework that leverages the CVEs and their textual descriptions currently linked to an ATT&CK tactic by prior undertakings (Hemberg et al., 2020) to link every CVE to the ATT&CK framework.
  • To achieve this goal, we draw upon state-of-the-art methodologies in deep learning-based text classification literature to guide the development of a novel cybersecurity artifact, the CVE Transformer (CVET) model.
  • To ensure the value of our proposed approach, we will rigorously evaluate our IT artifact against benchmark models found in related text classification and cybersecurity analytics literature.
Literature Review

• Three areas of literature are examined:
  1. **CVE data mining** to identify prior methodologies for studying the textual metadata in CVEs
  2. **Transformers for multi-class text classification** to review the prevailing deep learning method for text classification.
  3. **Self-distillation** to identify how to improve the internal representation of knowledge within the model to improve on state-of-art-performance.
Literature Review: CVE Data Mining

• Large undertakings have been taken to use CVEs to improve cybersecurity information systems using deep learning architectures.
  • Convolutional Neural Networks (CNN) have seen success in vulnerability severity classification (Han et al., 2017) and knowledge graph creation (Xiao et al., 2019).

• However, CNNs struggle to capture long term dependencies in textual passages (Wang et al., 2019).
  • To solve this issue, researchers have leveraged the pre-trained Transformer model known as BERT (Sun et al., 2021) to extract information from the vulnerability database ExploitDB to enhance descriptions for new CVEs.
Literature Review: CVE Data Mining

• Building a model that can effectively map CVEs to ATT&CK tactics based purely on textual descriptions requires an algorithm that can effectively represent the long text sequences found in CVE descriptions.

• The transformer model (and its extensions) is currently the state-of-art within text classification literature and has proven to be robust against adversarial attacks (Jin et al., 2020).

• We review the transformer model in depth to gain a deeper understanding of how it can assist in our target task.
Literature Review: Transformers for Multi-Class Text Classification

- Introduced in 2017, the Transformer model replaces the recurrent cells found in many prominent text classification deep learning models (e.g., BiLSTM, LSTM) with attention mechanisms (Vaswani et al., 2017).
  - While the original design incorporates an encoder-decoder structure (for machine translation tasks), multi-class text classification only requires the encoder stack.

- Transformers are often the architecture used to create massive pre-trained language models (PTLMs) (e.g., BERT and GPT-2).
  - PTLMs have achieved state-of-the-art results in text classification, generation, and masked modeling tasks (Qiu et al., 2020).
  - However, PTLMs are highly general and require intermediate steps (e.g., fine-tuning) before being used for a targeted task (Radiya-Dixit and Wang, 2020).

Figure 2. Transformer Architecture (Adapted from Vaswani et al., 2017)
Literature Review: Knowledge Distillation

• Generally, knowledge distillation combines the relational knowledge from a large, pre-trained model (teacher) and a prior untrained model (student) (Xu et al., 2020).
  • The trained student model is more generalizable to unseen data than a model without knowledge distillation.

• There are three types of knowledge distillation in a deep teacher-student network (Gou et al., 2020):
  1. **Response-Based Knowledge**: Distillation from the last output layer of the teacher model, teaching the student model to “mimic” the result.
  2. **Feature-Based Knowledge**: Distillation of the feature representation of the teacher model.
  3. **Relation-Based Knowledge**: Distillation of instance relations between data samples.

Figure 3: High Level Student – Teacher Knowledge Distillation Framework
Literature Review: Knowledge Distillation: Self-Distillation

• Self-distillation is a form of knowledge distillation where the student and teacher model are the same model.

• This form of distillation creates a student model that often outperforms the teacher model (Yang et al., 2019).
  • Theories on why this occurs include improved feature importance weighting (Furlanello et al., 2018) or enhanced regularization (Mohabi et al., 2020)
Literature Review: Knowledge Distillation: Self-Distillation

• The self-distillation architecture proposed by Xu et al. (2020) currently produces state-of-the-art results:
  • This architecture fine-tunes the seminal PTLM BERT through a self-distillation-averaged (SDA) design, where the learning strategy is:

  \[
  \mathcal{L}_\theta(x, y) = CE(BERT(x, \theta), y) + \lambda MSE\left(BERT(x, \theta), BERT(x, \bar{\theta})\right)
  \]

  • \(BERT(x, \bar{\theta})\) is the teacher model, CE is cross entropy loss, MSE is mean squared error loss, and \(\lambda\) (self-distillation weight) balances the importance of the two loss functions.

  • At each time step \(t\), \(\bar{\theta}\) is the averaged parameters of \(K\) (hyperparameter denoting the teacher size) time steps:

  \[
  \bar{\theta} = \frac{1}{K} \sum_{k=1}^{K} \theta_{t-k}
  \]
Research Gaps and Questions

• From the extant literature, we identify a couple clear gaps that we aim to cover:

  1. Many tasks have been undertaken to link CVEs to vulnerabilities, CWEs, and CAPEC, but not directly to ATT&CK.

  2. The deep learning models implemented in recent literature (e.g., CNN, BiLSTM) struggle to capture long-term dependencies in text, like the lengthy descriptions that are coupled with CVEs.

• These two gaps motivate our research questions:

  • What is the best way to create a novel link between CVEs and ATT&CK tactics by accounting for the available metadata?

  • How can we develop a novel framework that includes knowledge distillation to improve CVE to ATT&CK links?
Research Design

• To answer the posed research questions, we propose a novel framework we call CVE-Link (Figure 2).
• The CVE-Link framework is comprised of three major components: (1) Data Collection and Pre-Processing, (2) Transformer Architecture, and (3) Experiments and Evaluations. Each component is further detailed in the subsequent sections.

Figure 4. Proposed Research Design
Research Design: Data Collection

• For our research, we use the dataset provided by the BRON knowledge graph (Hemberg et al., 2020).

• The dataset successfully leverages existing knowledge to link 24,863 CVEs into 10 of the 14 ATT&CK tactics.
  • Table 2 provides a distribution of how many CVEs are in each ATT&CK tactic category.

• About 91% of our data distribution is contained within the just four Tactic categories.
  • Many ATT&CK tactics do not require specific vulnerabilities (e.g., “Resource Development” and “Command and Control”), meaning we cannot link CVEs to them.

• There are currently more than 158,000 CVEs, and our gold-standard dataset only captures a fraction of them.

<table>
<thead>
<tr>
<th>ATT&amp;CK Tactic</th>
<th>Count of CVEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defense Evasion</td>
<td>8,482</td>
</tr>
<tr>
<td>Discovery</td>
<td>6,647</td>
</tr>
<tr>
<td>Privilege</td>
<td>5,779</td>
</tr>
<tr>
<td>Escalation</td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td>1,748</td>
</tr>
<tr>
<td>Lateral Movement</td>
<td>715</td>
</tr>
<tr>
<td>Impact</td>
<td>594</td>
</tr>
<tr>
<td>Credential Access</td>
<td>427</td>
</tr>
<tr>
<td>Initial Access</td>
<td>309</td>
</tr>
<tr>
<td>Exfiltration</td>
<td>137</td>
</tr>
<tr>
<td>Execution</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24,863</strong></td>
</tr>
</tbody>
</table>

Table 2. Gold-Standard Dataset Distribution
Research Design: Pre-Processing

• To pre-process the CVE description text, stop words were removed, non-alphanumeric characters were stripped.

• The remaining text was lower-cased, lemmatized, and padded to ensure proper lengths for all inputs.
  • This sequence of pre-processing steps is common in deep learning-based text classification literature (Kamath et al., 2019).

• We used the pre-made RoBERTa tokenizer (Liu et al., 2019) to properly encode the data as an input for our self-distillation pre-trained language model architecture.
Research Design: Self-Distillation

• We utilize the RoBERTa pre-trained language model due to the high generalizability it has shown in text classification tasks (Chalkidis et al., 2020)
  • We then fine-tune the RoBERTa model on CVE descriptions to make it more effective on our target task.

• Then, we implement the self-distillation design outlined in the literature review (Xu et al., 2020).
  • In self-distillation, both the teacher and student models are the same RoBERTa model, which learns deeper latent representation of its hidden features to improve model performance.
Research Design: Benchmark Experiments

• To test the validity of our proposed approach, we will compare the results of the CVET model against prominent and state-of-the-art models in text classification literature.
  * **Classical Machine Learning:** SVM, Gradient Boosted Decision Trees, Logistic Regression, Naïve Bayes
  * **Deep Learning:** Transformer, Bi-LSTM w/ Attention, Bi-LSTM, LSTM, GRU, RNN
  * **Pre-Trained Language Models:** GPT-2, BERT, RoBERTa w/o self-distillation
  * All models will be run with 10-fold cross-validation so that accurate t-test comparisons can be made.

• All models will be evaluated with accuracy, precision, recall, and F1-score, which is the standard for multi-class text classification tasks (Thangaraj and Sivakami, 2018)
Results and Discussion: Experiment 1

Using a self distillation design to create the CVET model outperformed all deep learning and classical machine learning models in accuracy (76.93%), precision (81.88%), recall (69.49%), and F1-score (75.18%).

- All results were significant at $p < 0.001$.
Results and Discussion: Experiment 2

Using a self-distillation design to create the CVET model outperformed all PTLMs in accuracy (76.93%), precision (81.88%), recall (69.49%), and F1-score (75.18%)

- All results were significant at $p < 0.01$ or better.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Trained</td>
<td>GPT-2</td>
<td>70.21%</td>
<td>75.12%</td>
<td>62.56%</td>
<td>68.27%</td>
</tr>
<tr>
<td>Language Models</td>
<td>DistillBERT</td>
<td>72.81%</td>
<td>77.46%</td>
<td>65.56%</td>
<td>71.01%</td>
</tr>
<tr>
<td></td>
<td>MobileBERT</td>
<td>72.31%</td>
<td>78.03%</td>
<td>65.98%</td>
<td>71.50%</td>
</tr>
<tr>
<td></td>
<td>XLNet</td>
<td>74.12%</td>
<td>78.12%</td>
<td>66.56%</td>
<td>71.88%</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>73.93%</td>
<td>77.86%</td>
<td>67.41%</td>
<td>72.26%</td>
</tr>
<tr>
<td></td>
<td>RoBERTa</td>
<td>74.42%</td>
<td>79.88%</td>
<td>66.49%</td>
<td>72.57%</td>
</tr>
</tbody>
</table>

- All results were significant at $p < 0.01$ or better.

Table 5. Results of Experiment 1 Against Benchmark Pre-Trained Language Models (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$)

Figure 7. Results of Experiment 1 Against Benchmark Pre-Trained Language Models
Discussion

• Our results suggest that the self-distillation design applied to a prominent PTLM (i.e., RoBERTa) assists in model performance.

  • The self-distillation design can help bring hidden latent features created during the fine-tuning process to the surface of the model.

  • The new hidden latent features are highly targeted towards the CVE $\rightarrow$ ATT&CK framework task.
Future Directions

• The authors recognize that there can be improved novelty in several key steps of the research design:
  1. Improved NLP techniques (e.g., NER extraction, synonym/homonym generation, POS tagging) can create more novel embeddings for the work.

  2. The fine-tuning process can be improved to generate better hidden features in the PTLM, thus improving the self-distillation approach down the line.

• The model can also be extended for different types of cybersecurity risk management frameworks (e.g., NIST, CAPEC)
Conclusion

• In this study, we developed a novel self-distillation approach to automatically label CVEs with their associated ATT&CK tactic.

• The design was evaluated with a series of experiments against state-of-the-art models in classical machine learning, deep learning, and pre-trained language models.

• Results indicated that the CVET model offers a significant benefit to labeling CVEs with MITRE ATT&CK tactics over baseline non-distillation techniques.


