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.

Introduction: MITRE ATT&CK

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.

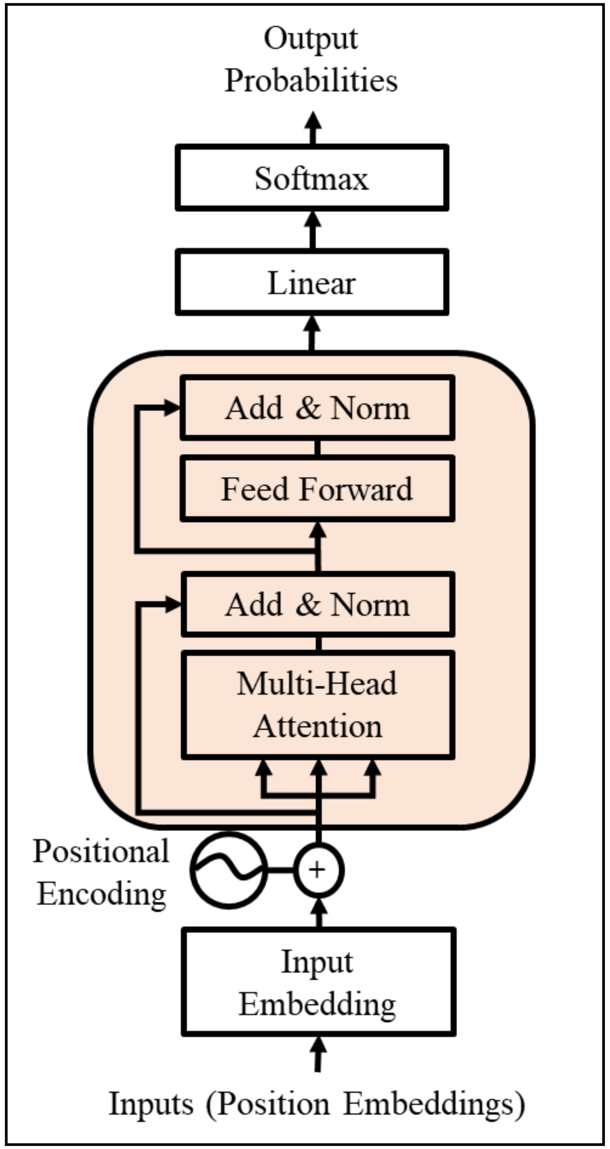

Figure 2. Transformer Architecture (Adapted from Vaswani et al., 2017)

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).

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.

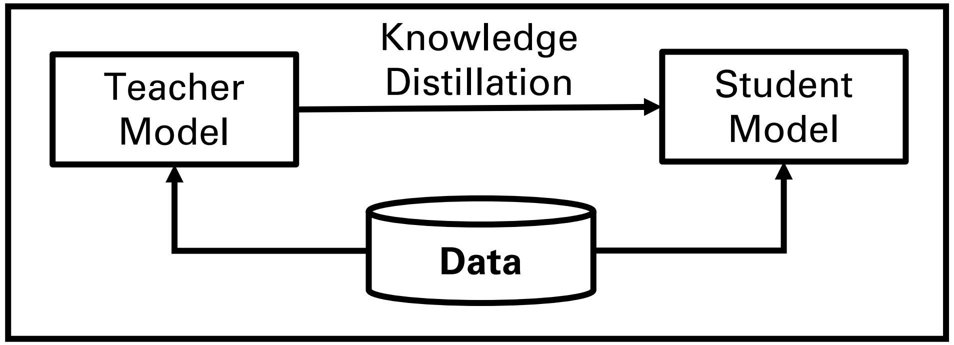

Figure 3: High Level Student – Teacher Knowledge Distillation Framework

- 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.

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 λ (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.

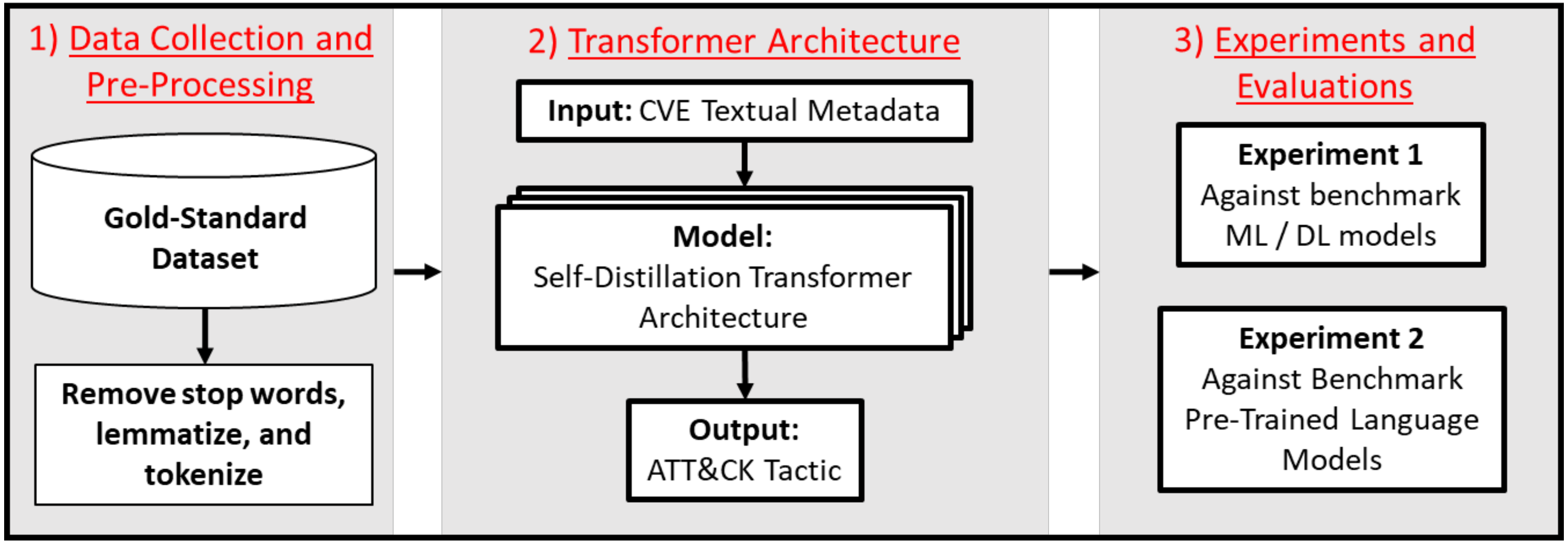

Figure 4. Proposed Research Design

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.

ATT&CK Tactic	**Count of CVEs**
Defense Evasion	8,482
Discovery	6,647
Privilege Escalation	5,779
Collection	1,748
Lateral Movement	715
Impact	594
Credential Access	427
Initial Access	309
Exfiltration	137
Execution	25
Total	**24,863**

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

Type	Model	Accuracy	Precision	Recall	F1-score
Classical Machine Learning	Random Forest	63.70% ***	15.83% ***	17.67% ***	16.70% ***
	SVM	65.70% ***	51.23% ***	46.34% ***	48.66% ***
	Naive Bayes	67.30% ***	34.22% ***	23.92% ***	28.16% ***
	Logistic Regression	67.10% ***	31.65% ***	24.12% ***	27.38% ***
Deep Learning	RNN	68.45% ***	69.66% ***	67.30% ***	68.46% ***
	GRU	70.90% ***	72.55% ***	69.19% ***	70.83% ***
	LSTM	72.75% ***	74.14% ***	71.89% ***	73.00% ***
	BiLSTM	72.55% ***	73.71% ***	71.71% ***	72.70% ***
	BiLSTM with Attention	71.41% ***	72.52% ***	70.32% ***	71.40% ***
	Transformer	72.45% ***	74.49% ***	70.82% ***	72.61% ***
Self-Distillation	CVET	76.93%	81.88%	69.49%	75.18%

Table 4. Results of Experiment 1 Against Benchmark ML / DL Models (*: p<0.05, **: p<0.01, ***: p<0.001)

- 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 $p < 0.001$.

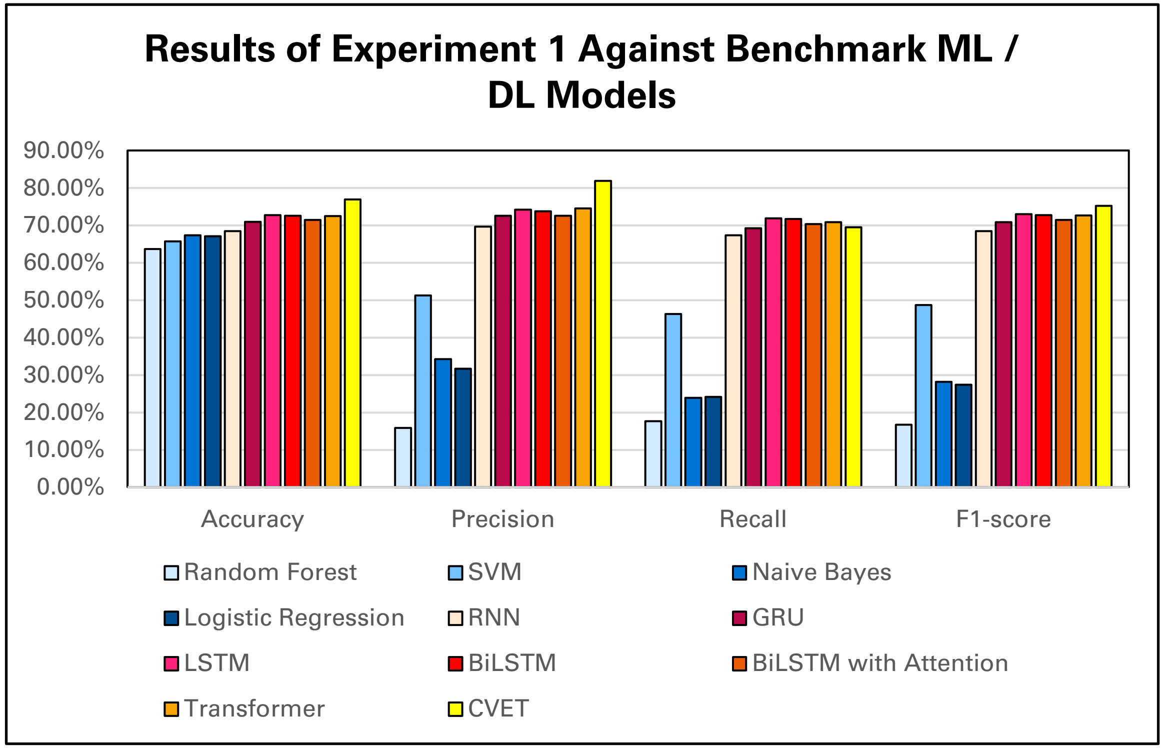

Figure 6. Results of Experiment 1 Against Benchmark ML / DL Models

Results and Discussion: Experiment 2

Type	Model	Accuracy	Precision	Recall	F1-score
Pre-Trained Language Models	GPT-2	70.21% ***	75.12% ***	62.56% ***	68.27% ***
	DistillBERT	72.81% ***	77.46% ***	65.56% ***	71.01% ***
	MobileBERT	72.31% ***	78.03% ***	65.98% ***	71.50% ***
	XLNet	74.12% ***	78.12% ***	66.56% ***	71.88% ***
	BERT	73.93% ***	77.86% ***	67.41% **	72.26% ***
	RoBERTa	74.42% **	79.88% **	66.49% **	72.57% **
Self-Distillation	CVET	76.93%	81.88%	69.49%	75.18%

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

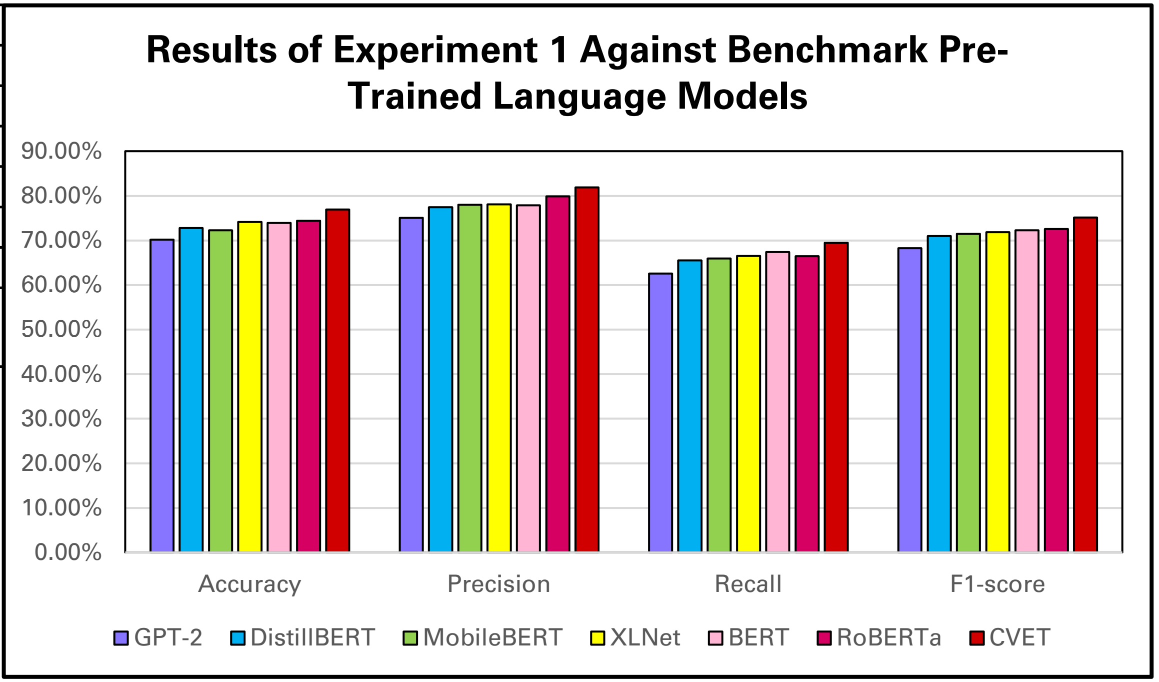

Figure 7. Results of Experiment 1 Against Benchmark Pre-Trained Language Models

- 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.

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 → 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.

References

- Al-Shaer, R., Ahmed, M., & Al-Shaer, E. (2018). Statistical Learning of APT TTP Chains from MITRE ATT&CK. In *Proceedings RSA Conference* (pp. 1-2).
- Chen, Q., Bao, L., Li, L., Xia, X., & Cai, L. (2018). Categorizing and Predicting Invalid Vulnerabilities on Common Vulnerabilities and Exposures. 2018 25th Asia-Pacific Software Engineering Conference (APSEC), 2018-Decem, 345–354. https://doi.org/10.1109/APSEC.2018.00049
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Mlm. http://arxiv.org/abs/1810.04805
- Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.
- Enoch, S. Y., Huang, Z., Moon, C. Y., Lee, D., Ahn, M. K., & Kim, D. S. (2020). HARMer: Cyber-Attacks Automation and Evaluation. IEEE Access, 8, 129397-129414.
- Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018, July). Born again neural networks. In International Conference on Machine Learning (pp. 1607-1616). PMLR.
- Gong, X., Xing, Z., Li, X., Feng, Z., & Han, Z. (2019). Joint Prediction of Multiple Vulnerability Characteristics Through Multi-Task Learning. 2019 24th International Conference on Engineering of Complex Computer Systems (ICECCS), 2019-Novem(December 2018), 31–40. https://doi.org/10.1109/ICECCS.2019.00011
- Guo, H., Xing, Z., & Li, X. (2020). Predicting Missing Information of Vulnerability Reports. Companion Proceedings of the Web Conference 2020, 81–82. https://doi.org/10.1145/3366424.3382707
- Han, Z., Li, X., Xing, Z., Liu, H., & Feng, Z. (2017). Learning to Predict Severity of Software Vulnerability Using Only Vulnerability Description. 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), 125–136. https://doi.org/10.1109/ICSME.2017.52
- Hemberg, E., Kelly, J., Shlapentokh-Rothman, M., Reinstadler, B., Xu, K., Rutar, N., & O'Reilly, U.-M. (2020). BRON -- Linking Attack Tactics, Techniques, and Patterns with Defensive Weaknesses, Vulnerabilities and Affected Platform Configurations. ArXiv. http://arxiv.org/abs/2010.00533
- Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8018–8025. https://doi.org/10.1609/aaai.v34i05.6311

References

- Kamath, U., Liu, J., & Whitaker, J. (2019). Deep Learning for NLP and Speech Recognition. In Deep Learning for NLP and Speech Recognition. Springer International Publishing. https://doi.org/10.1007/978-3-030-14596-5
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Mobahi, H., Farajtabar, M., & Bartlett, P. L. (2020). Self-distillation amplifies regularization in hilbert space. arXiv preprint arXiv:2002.05715.
- Mell, P., Scarfone, K., & Romanosky, S. (2006). Common Vulnerability Scoring System. IEEE Security and Privacy Magazine, 4(6), 85–89. https://doi.org/10.1109/MSP.2006.145
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 1-26.
- Strom, B. E., Miller, D. P., Nickels, K. C., Pennington, A. G., & Thomas, C. B. (2018). MITRE ATT&CKTM : Design and Philosophy. July.
- Sun, H., Xu, M., & Zhao, P. (2020). Modeling Malicious Hacking Data Breach Risks. North American Actuarial Journal, 0(0), 1–19. https://doi.org/10.1080/10920277.2020.1752255
- Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117–135. https://doi.org/10.28945/4066
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 5999–6009. http://arxiv.org/abs/1706.03762
- Wang, R., Li, Z., Cao, J., Chen, T., & Wang, L. (2019). Convolutional Recurrent Neural Networks for Text Classification. 2019 International Joint Conference on Neural Networks (IJCNN), 2018, 1–6. https://doi.org/10.1109/IJCNN.2019.8852406
- Xiao, H., Xing, Z., Li, X., & Guo, H. (2019). Embedding and predicting software security entity relationships: A knowledge graph-based approach. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11955 LNCS. Springer International Publishing. https://doi.org/10.1007/978-3-030-36718-3_5
- Xu, Y., Qiu, X., Zhou, L., & Huang, X. (2020). Improving bert fine-tuning via self-ensemble and self-distillation. arXiv preprint arXiv:2002.10345.
- Yang, C., Xie, L., Qiao, S., & Yuille, A. L. (2019, July). Training deep neural networks in generations: A more tolerant teacher educates better students. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 5628-5635).