Towards KB Embedding in Malware Detection

Daniel Trizna and Martin Homola

Comenius University in Bratislava, Mlynská dolina, 84248 Bratislava, Slovakia

The problem of *malware detection* is to classify logged data about processes and their behaviour and to identify which processes are potentially dangerous (malware). Historically, large number of diverse methods have been applied on this problem, including some AI methods such as machine learning [3].

Knowledge base (KB) embedding is a neural-symbolic approach to represent knowledge base in a low-dimensional vector space keeping as much essential information as possible. In recent years there have been many approaches to achieve such representation ranging from simpler attempts like TransE [2] to more sophisticated ones like Cone embedding [5] and Sphere embedding [4].

We propose to apply KB embedding to the problem of malware detection. We will work with the EMBER dataset [1] and its ontological representation developed by Švec et al. [6, 7]. The main advantages we hope this approach could give us are the following:

- the approach is combinable with standard machine learning algorithms once we have good embedding;
- the embedding could reveal hidden connections in our dataset;
- the embedding may possibly improve detection results due to incorporated symbolic knowledge.

Acknowledgents. This work was partially supported by projects ORBIS, funded by Slovak SRDA agency under contract No APVV-19-0220, and by TAILOR, funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Anderson, H.S., Roth, P.: EMBER: An open dataset for training static PE malware machine learning models. CoRR abs/1804.04637, arXiv.org, https://arxiv.org/abs/1804.04637 (2018)
2. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data p. 2787–2795 (2013)
3. Husák, M., Komárková, J., Bou-Harb, E., Celeda, P.: Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun. Surv. Tutorials **21**(1), 640–660 (2019)
4. Kulmanov, M., Liu-Wei, W., Yan, Y., Hoehndorf, R.: EL embeddings: Geometric construction of models for the description logic EL ++. CoRR abs/1902.10499, arXiv.org, https://arxiv.org/abs/1902.10499 (2019)
5. Özçep, Ö.L., Leemhuis, M., Wolter, D.: Cone semantics for logics with negation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 1820–1826 (2020)

6. Švec, P., Balogh, S., Homola, M.: Experimental evaluation of description logic concept learning algorithms for static malware detection. In: Proceedings of the 7th International Conference on Information Systems Security and Privacy, ICISSP 2021, February 11-13. pp. 792–799. SCITEPRESS (2021)
7. Švec, P., Balogh, S., Homola, M., Kľuka, J.: Ontological representation of the EMBER dataset. In: Procs. of AKMIS 2022, The 2nd workshop on Application of Knowledge Methods in Information Security (to appear) (2022)