Ερμηνεία Μοντέλων Βαθιάς Μηχανικής Μάθησης με Μεθόδους eXplainable AI (XAI) - Εφαρμογή στην Αντιμετώπιση Επιθέσεων DDoS

Οι επιθέσεις άρνησης παροχής υπηρεσιών (Distributed Denial of Service, DDoS) αποτελούν ένα από τα επικρατέστερα προβλήματα για τους διαχειριστές δικτυακών υποδομών. Οι επιθέσεις αυτές στοχεύουν σε δίκτυα ή/και υπηρεσίες παρεμποδίζοντας την ομαλή λειτουργία τους.

Για την αποδοτική ανίχνευση (detection) και αντιμετώπιση (mitigation) επιθέσεων DDoS έχουν μελετηθεί πολυάριθμες μέθοδοι. Ανάμεσα σε αυτές τις μεθόδους, εκείνες που βασίζονται σε μοντέλα μηχανικής μάθησης (machine learning), συμπεριλαμβανομένων μοντέλων βαθιάς μηχανικής μάθησης (deep learning) έχουν αποδειχθεί ιδιαίτερα υποσχόμενες.

Ωστόσο, τα μοντέλα machine learning αντιμετωπίζονται από τους διαχειριστές δικτύων ως black boxes, δηλαδή οι διαχειριστές αδυνατούν να κατανοήσουν την ακριβή λειτουργία τους και τα κριτήρια κατηγοριοποίησης δικτυακής κίνησης. Η προσέγγιση αυτή δυσκολεύει τη βελτίωση της ακρίβειας των μοντέλων και την παροχή εγγυήσεων για τη λειτουργία τους.

Για το σκοπό αυτό έχουν αναπτυχθεί οι τεχνικές eXplainable Artificial Intelligence (XAI) [1] που στοχεύουν στην κατανόηση των μοντέλων machine/deep learning. Οι τεχνικές αυτές εφαρμόζονται είτε για τη συνολική κατανόηση των μοντέλων, δηλαδή πώς οι παράμετροι των μοντέλων επηρεάζουν τις αποφάσεις τους (global explainability) ή την κατανόηση του πώς λαμβάνονται οι αποφάσεις ταξινόμησης (classification) για συγκεκριμένες εισόδους (local explainability) [2]. Τέτοιες μέθοδοι XAI είναι οι: LIME, SHAP, Counterfactual Explanations [2, 3].

Η διπλωματική θα διερευνήσει μεθόδους XAI για την κατανόηση μοντέλων machine learning, είτε supervised (π.χ. Multi-Layer Perceptron - MLP) ή unsupervised learning (π.χ. autoencoder), που εφαρμόζονται για την ανίχνευση ή/και την αντιμετώπιση επιθέσεων DDoS [4, 5]. Σκοπός της διπλωματικής θα είναι να διερευνήσει τη συνεισφορά διαφορετικών τύπων features και να συγκρίνει μεθόδους XAI. Για την πειραματική αξιολόγηση των μοντέλων θα χρησιμοποιηθούν δημόσια διαθέσιμα δεδομένα καλόβουλης και κακόβουλης δικτυακής κίνησης, τα οποία χρησιμοποιούνται ευρέως στη βιβλιογραφία, π.χ. [6].

Η διπλωματική μπορεί να επεκταθεί και σε συναφή πεδία έρευνας στην ασφάλεια δικτύων υπολογιστών, όπως είναι η ανίχνευση κίνησης που παράγεται από Domain Generation Algorithms (DGA's) [7] ή την ανίχνευση κακόβουλων μηνυμάτων σε κρυπτογραφημένη κίνηση, π.χ. DNS over HTTPS (DoH) [8] και QUIC [9].

[1] Explainable Artificial Intelligence, https://en.wikipedia.org/wiki/Explainable_artificial_intelligence
[2] Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/
[3] SHAP, https://github.com/slundberg/shap
[4] M. Wang et al., "An Explainable Machine Learning Framework for Intrusion Detection Systems", in IEEE Access
[5] J. Mossin Wagle, "Utilizing the SHAP Framework to Bypass Intrusion Detection Systems", https://bora.uib.no/bora-xmlui/bitstream/handle/11250/2761761/thesis_master_JonasMW-final.pdf?sequence=1&isAllowed=y
[6] DDoS Evaluation Dataset (CIC-DDoS2019), https://www.unb.ca/cic/datasets/ddos-2019.html
[7] T. Zebin, "An Explainable AI-based Intrusion Detection System for DNS over HTTPS (DoH) Attacks", in IEEE Transactions on Information Forensics and Security, June 2022
[8] G. Piras, "Explaining Machine Learning DGA Detectors from DNS Traffic Data", in arXiv preprint, August 10
[9] QUIC, https://cloudflare-quic.com/