機械学習を用いたファズデータの チェックサム及びハッシュ値の推定

石浦研究室　27016627　藤本 高史

1　はじめに

ソフトウェアの脆弱性は社会的に深刻な問題をもたらしており, リリース前に十分なテストを行うことが重要な課題になっている. ファジングは, 自動的に生成した大量の入力データによって対象プログラムをテストする, セキュリティなどの脆弱性の検出手法である. 変異ベース手法は実装が容易で, 汎用性も高いが, 入力データの棄却率が高く, 効率的にファジングを行えていないことが課題となる. この課題に対し, 文献 [1] は LSTM を用いて入力データに対するチェックサムの推定を行い, 68 %の正答率となった. しかし, 8 byte のデータに対するチェックサムのみしか対応していないため, 汎用性に欠ける課題がある.

本研究では, 8 byte 以上のデータに対するチェックサム及び様々なハッシュ値の推定を行う.

2　ニューラルネットワークによるチェックサムの推定

文献 [1] はニューラルネットワークを用いて, 8 byte の入力データに対するチェックサムの推定を行っている. 正規データからデータとチェックサムの集合を抽出し, ニューラルネットワークに学習させる. 学習済ニューラルネットワークで変異データからチェックサムを推定し, 更新する. 学習の流れを, 図 1 に示す. 初めに, 文字列とその文字列に対するチェックサムのデータを用意する. 次に, そのデータをニューラルネットワークへ入力をし, 学習を行う. そして, 学習済ニューラルネットワークに対して, 文字列を入力することにより, チェックサム及びハッシュ値を推定する. 最後に, 推定したチェックサムを, 文字列の末端に付与する. よって, 正当な入力データとして扱われる.

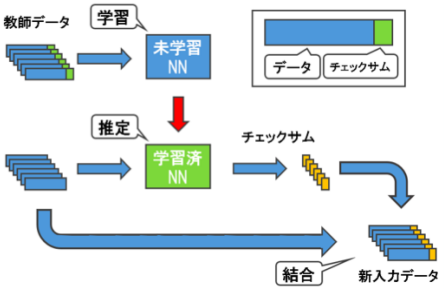

図 1 学習の流れ

3　機械学習によるチェックサム及びハッシュ値の推定

本研究では, 8 byte 以上のデータに対するチェックサム及び様々なハッシュ値の推定を行う. データとチェックサムおよびハッシュ値の位置とハッシュ関数の種類は既知であることを前提とする. 固定長ではなく, 可変長の入力データで学習を行う. 学習に関しては, Encoder・Decoder モデル [2] を利用したニューラルネットワークを構成して学習する. この時, 各ノード数などのパラメータは, どのような入力及び出力を行うかによって最適な値が異なるため, 実験を繰り返して最適な値を見つける. また, 推定する値によって計算方法は変わってくるため, 各々の値に対応した機械学習を行う.

4　実験

機械学習を用いて 8 byte 以上のランダム文字列と英文に対するチェックサム及びハッシュ値を推定する実験を行った. 提案手法を Keras を用いて Python で実装し, 機械学習を実行した. 総データ数は 10 ～ 20 万文書である. ランダム文字列は 1 ～ 64 byte の可変長の入力データである. 英文は文字列は 2 ～ 48 byte の可変長の入力データである. 学習させる文字列の長さ及びデータ数は, 推定する値によって異なる. 推定する値として, チェックサム, CRC16 [3], CRC32 [3], MD5 [4], SHA1 [5] で実験を行った.

学習時に得られた正答率と学習済ニューラルネットワークを用いて得られた正答率を, 表 1 に示す. Train Random, Train English は, 学習時に得られた正答率である. Random, English が学習済ニューラルネットワークを使用した時に得られた正答率である. 学習時に得られた正答率と学習済ニューラルネットワークを使用して得られた正答率を比較すると, 必ずしも正答率が一致しないことがわかった. 特に, ランダム文字列の方が正答率が高かったり, 英文の方が正答率が高かったりなど, ばらつきが見られた.

表 1 チェックサム及びハッシュ値の正答率

	Train Random	Train English	Random	English
cksum	52%	50%	20%	51%
CRC16	53%	47%	9%	9%
CRC32	54%	50%	12%	4%
MD5	11%	11%	12%	2%
SHA1	14%	19%	5%	11%

5　むすび

本研究では, ランダム文字列と英文からチェックサム及びハッシュ値の推定を行い, チェックサム, CRC16, CRC32, MD5, SHA1 を高精度で推定することができた. 今後の課題は, 精度向上, 他のハッシュ値の推定, 固定長による推定評価, 実装評価である.

参考文献

[1] 難波 学之: “変異ベースファジングのためのチェックサムの機械学習,” 関西学院大学理工学部情報科学科卒業論文 (Mar. 2019).

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V Le: “Sequence to sequence learning with neural networks,” in *Proc. Nueral Information Processing Systems*, pp. 3104–3112 (Sept. 2014).

[3] Andrew Tanenbaum, David Weatherroll 著, 水野忠則 訳: コンピュータネットワーク, 日経 BP(Sept. 2013).

[4] IPUSIRON 著: 暗号技術のすべて, 翔泳社 (Aug. 2017).

[5] 林 芳樹 著: 暗号理論入門, 丸善出版 (Apr. 2012).