Newest machine learning password guessing techniques

Andrius Chaževskas, Igoris Belovas, Virginijus Marcinkevičius

E-mails: andrius.chazevskas@mif.stud.vu.lt, igoris.belovas@mif.vu.lt, virginijus.marcinkevicius@mif.vu.lt

Vilnius University, Institute of Data Science and Digital Technologies

Abstract

Password guessing is essential for forensic encrypted data examination. The analysis of leaked password databases shows that users tend to use easy-to-remember passwords. It means that the passwords usually exhibit a logical structure; they are not just random character sets. Modern automated password guessing strategies relying on machine learning and natural language processing try to exploit this defect. This poster presents primary and latest password guessing approaches based on the analysis of password structures and content.

The problem and motivation

Forensic information technology (IT) examinations have been carried out in Lithuania since 1995. The forensic IT experts perform investigations of digital information following the tasks assigned by the courts and pre-trial investigation institutions. The problem of encrypted digital information directly affects the timing and quality of IT examinations. In terms of encrypted data examination, the most important are the technical characteristics of the hardware used by laboratories for password guessing and the strategies of password guessing attacks selected by forensic experts. The most common password guessing attacks are dictionary and brute-force. The main drawback of the brute-force attack is the size of a set of all possible password candidates, which grows exponentially with the length of the password. The laboratories can evaluate the possibility of applying the brute force attack using the formula (1). Where A – the alphabet size, M – password's length, N – password guess attempts per second.

$$MaxTime = \frac{A^M}{N}. \qquad (1)$$

Our and others' studies show that users tend to set their passwords predictably, favoring short strings, names, places, animals, ordinary and diminutive words typical of English and national language dictionaries. Figure 1 shows one hundred most common Lithuanian passwords (leaked from one of the Lithuanian social service providers) classified by different compositions, regarding their structures and meaning. This information is used to develop various password guessing techniques.

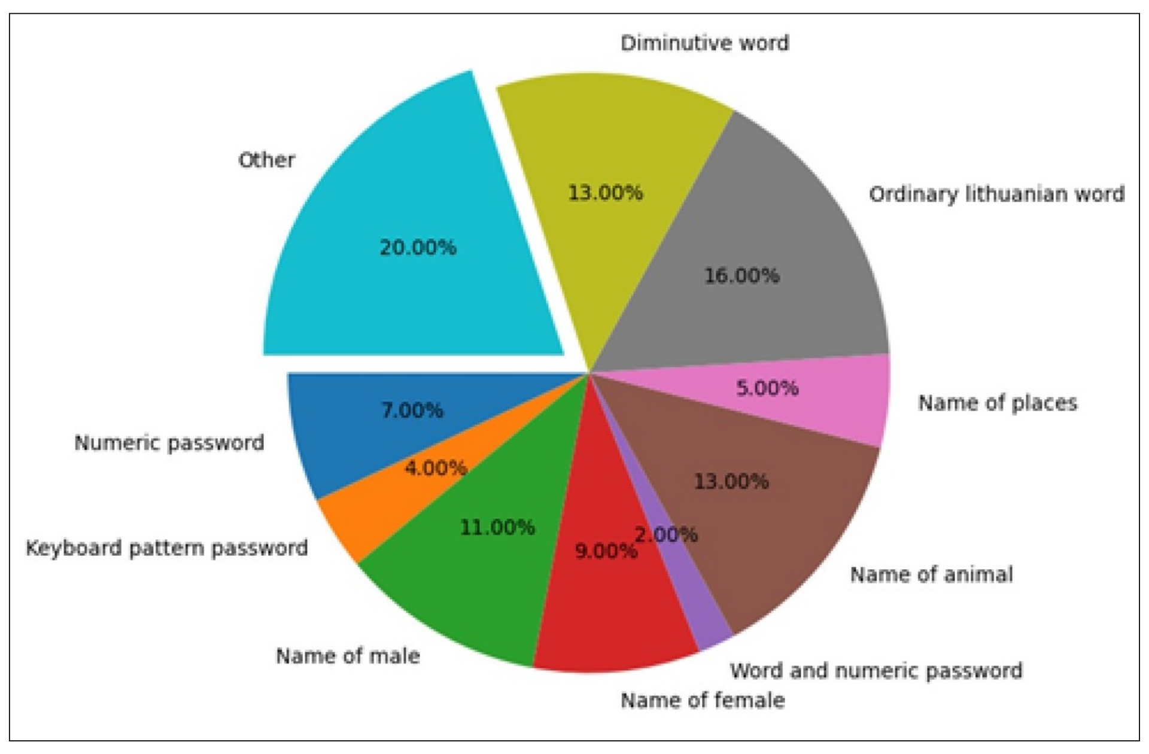

Figure 1: The percentage distribution of Lithuanian passwords.

Main password guessing approaches

The rule-based method is a popular strategy in digital forensics field and is widely used by IT experts. The most commonly used password-cracking tools are Hashcat and John the Ripper. The main idea of rule-based methods is combining training lists with mangling rules (appending characters, reversing the items, capitalizing the first letter, and so on) and transforming the original passwords into new candidate passwords.

Markov based methods, where users construct passwords calculating the password's probability through the connection between characters from left to right. The n-gram model is trained during the training, and the frequency of each letter appeared after the sub-string of length n is counted. During the generation stage, the probability of a probable password is calculated according to the Markov chain, and then the candidate passwords are generated.

Probabilistic Context-Free Grammar-Based Methods analyze the password structures as grammars and divide passwords into different segment types according to their character composition. Using password examples, you can create grammar rules that are used to generate new passwords in a password guessing strategy. When parsing the training set, we denote alphabet symbols as L, digit symbols as D, and special characters as S. For example the password @password123 would define the simple structure $S_1L_8D_3$.

Table 1: Example of probabilistic context-free grammar.

From left	to	right	Probability
S	->	$L_4S_1L_4$	1
L_4	->	pass	0.25
L_4	->	word	0.25
L_4	->	love	0.5
S_1	->	!	0.5
S_1	->	@	0.5

The neural network-based methods with deep learning models are recently used to construct passwords. The models could be divided into probabilistic that generate candidate passwords according to their probability and generative models that randomly generate candidates in batches. Figure 2 shows password generation model based on LSTM.

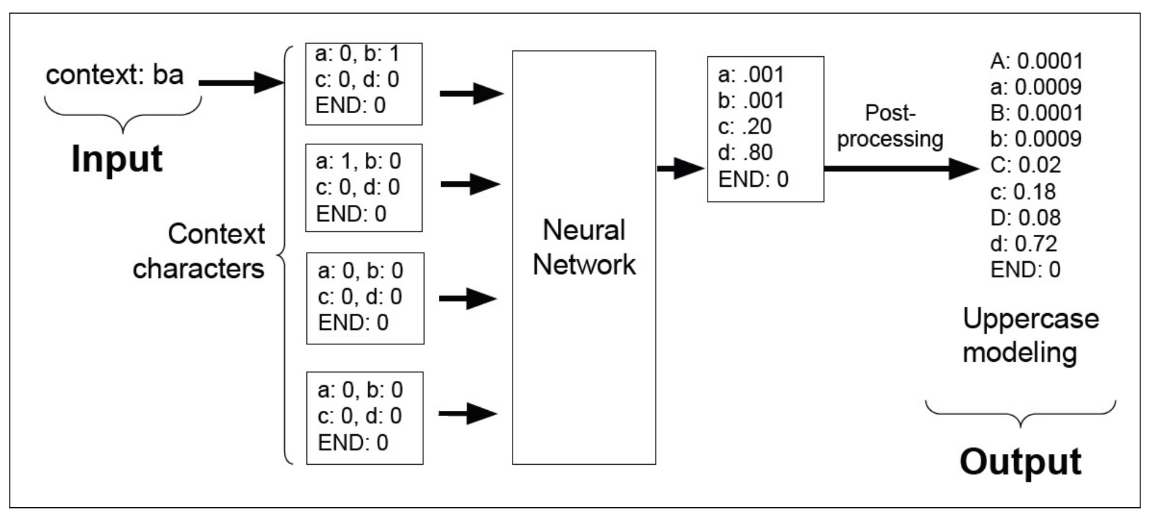

Figure 2: LSTM model.

Generative adversarial network-based password guessing model (PassGAN) consists of one generational network which 'learns' and tries to find out the statistical structure of the data, seeking to create new, statistically similar examples. The second network is used for testing (error detection), which tries to detect which samples are generated by the generation network and which are original. Training ends when the testing network cannot distinguish the origin of the meaning.

New password guessing approaches

A targeted password attacking model (TG-SPSR) synthesizes PCFG and Markov chain models' advantages and proposes a targeted password attacking model based on structure partition and string reorganization. The basic structure of passwords is divided into diffrent segments using the PCFG algorithm. At the same time, the Markov model is used to model the strings in each segment to generate new strings to achieve the model's accuracy and generalization ability. This model mines the password features from the structural level and the character level (which is more consistent with the human habit of constructing passwords). Figure 3 shows the schema of the model.

A Chinese syllables and Neural Network-based password generation method (CSNN) is proposed for Chinese password sets. This method treats Chinese Syllables as integral elements and uses them to parse and process passwords. The processed passwords are trained in Long Short-Term Memory Neural Network, and the trained model is used to generate password dictionaries (guessing sets). Chinese Syllables to process the training set and obtain a password structure frequency table Σ. Neural Network is used to train the processed training set and obtain the model M. Frequency table Σ works with M to generate password guessing sets. Figure 4 shows the model implementation.

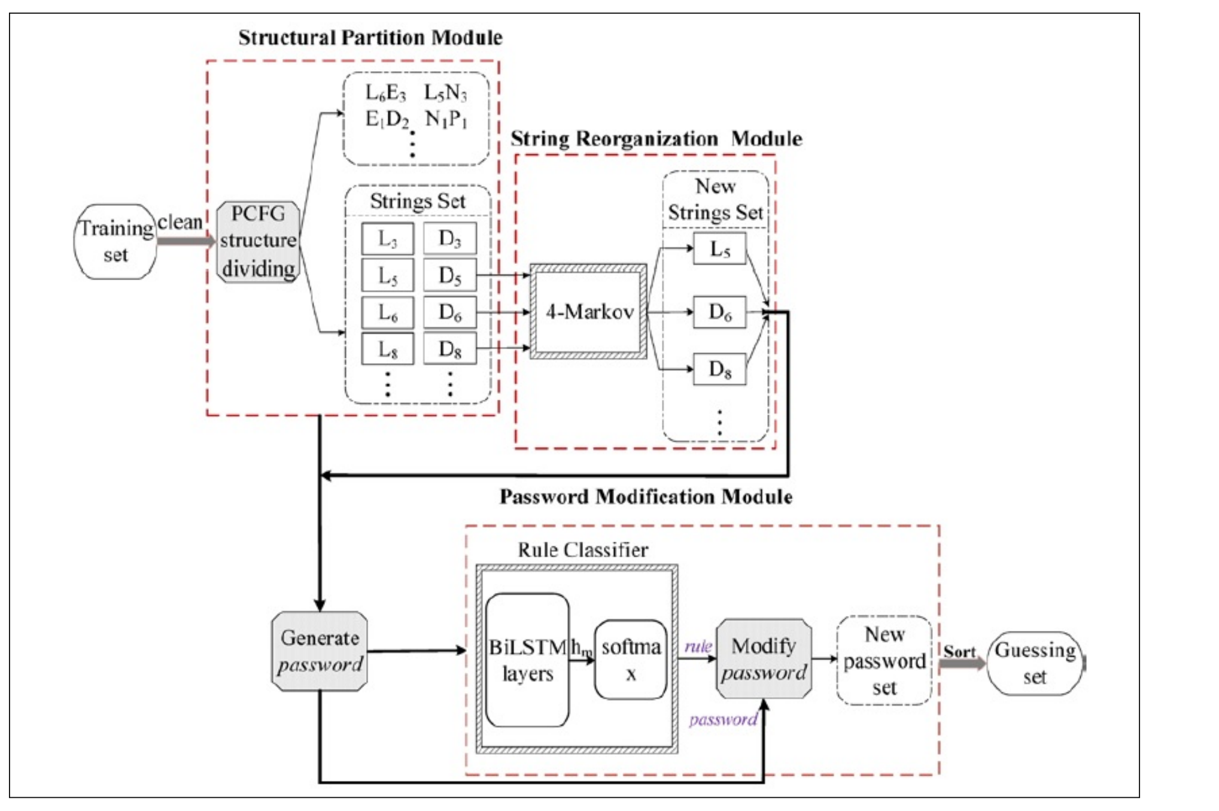

Figure 3: TG-SPRS model.

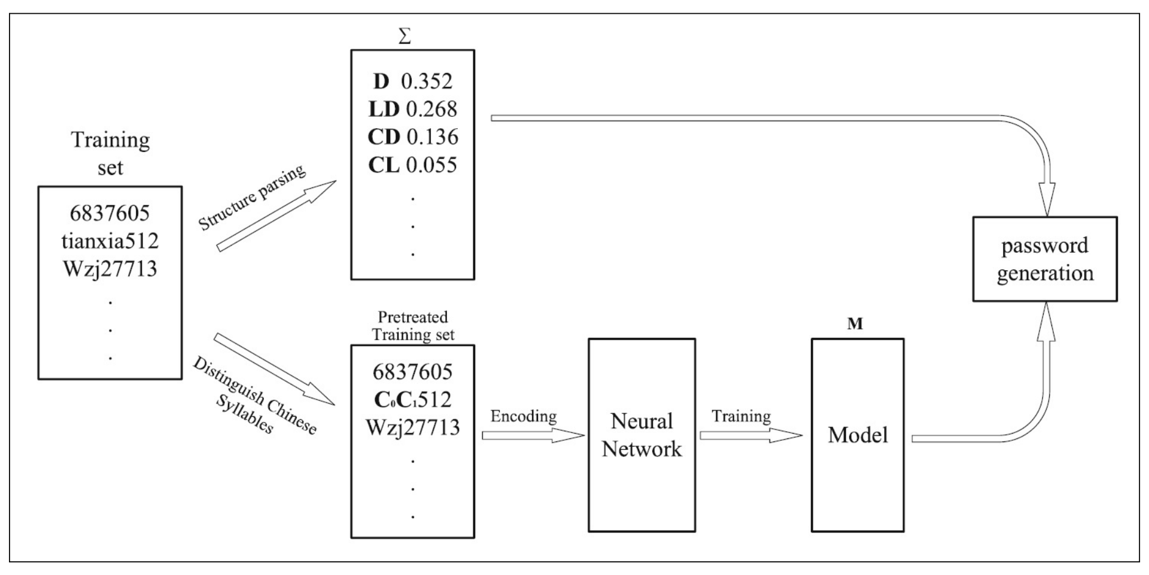

Figure 4: CSNN model.

A multi-source deep learning password guessing model (GENPass) can generate a wordlist of passwords based on several sources and improve cross-site attacks' performance. The model consists of several generators (PL) based on Probabilistic Context-Free Grammar (PCFG) and Long short-term memory (LSTM), and a classifier. Generators create passwords from leaked datasets. The task of the classifier and the discriminator is to make sure the output does not belong to a specific dataset so that the output is believed to be general to all training datasets. Figure 5 shows the diagram of the GENPass model.

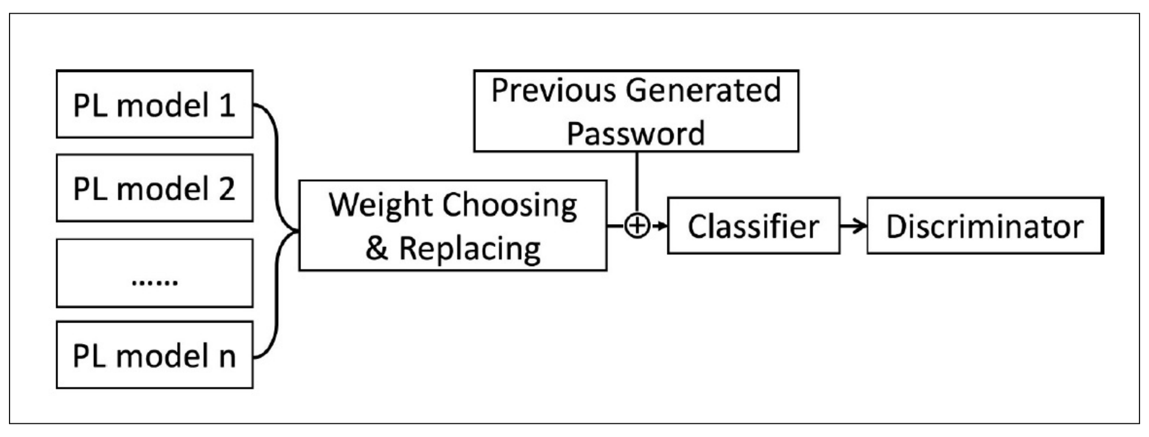

Figure 5: GENPass model.

Summary and conclusions

Despite all well-known security weaknesses, passwords are still one of the most common authentication and digital information encryption solutions. Due to its importance password guessing is an active field of study conducting researches both for offensive and defensive purposes. In this poster, we have briefly discussed the main password guessing solutions and presented some new methods based on password content analysis, which could be successfully applied for encrypted information forensic examinations field.

References

[1] M. Zhang, Q. Zhang, W. Liu, X. Hu and J. Wei, "TG-SPSR: A Systematic Targeted Password Attacking Model," KSII Transactions on Internet and Information Systems, vol. 13, no. 5, pp. 2674-2697, 2019. DOI: 10.3837/tiis.2019.05.024.

[2] H. Xian, Y. Zhang, D. Wang, Z. Li, and Y. He, "CSNN: Password Set Security Evaluation Method Based on Chinese Syllables and Neural Network," Dianzi Yu Xinxi Xuebao/Journal Electron. Inf. Technol., vol. 42, no. 8, pp. 1862–1871, 2020, doi: 10.11999/JEIT7 190856.

[3] Z. Xia, P. Yi, Y. Liu, B. Jiang, W. Wang and T. Zhu, "GENPass: A Multi-Source Deep Learning Model for Password Guessing," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1323-1332, May 2020, doi: 10.1109/TMM.2019.2940877.

12th Data Analysis Methods for Software Systems Conference, 2-4 December 2021, Druskininkai, Lithuania