The Past and Possible Future Development of Password Guessing

The Twelfth International Conference on Communications, Computation, Networks and Technologies
INNOV 2023

Authors:Ze-long Li[1]、Teng Liu[2]、Lei Li[3]

Presenter： Ze-long Li
Affiliation： [1,2]University Of Jinan
Email:202121200928@stu.ujn.edu.cn

UNIVERSITY OF JINAN 1948 济南大学

蓝剑钧新 LANJIANJUNXIN

IARIA



Presenter: Ze-long Li

A full-time graduate student from University of Jinan, majoring in computer technology, with a research focus on network and information security. Currently, he works with his supervisor in the company.

- Computer software:Electronic Document Passwords Management System.(Applying)
- Computer software:Network Security Events Warning System.(Applying)

Our research interests are mainly related to passwords, especially rainbow tables and password guessing (especially combined with deep learning).

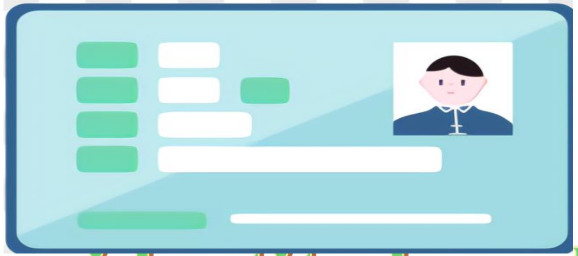
CONTENTS

01 Present situation

identity authentication

①proving your identity based on what you know

② proving your identity based on what you have



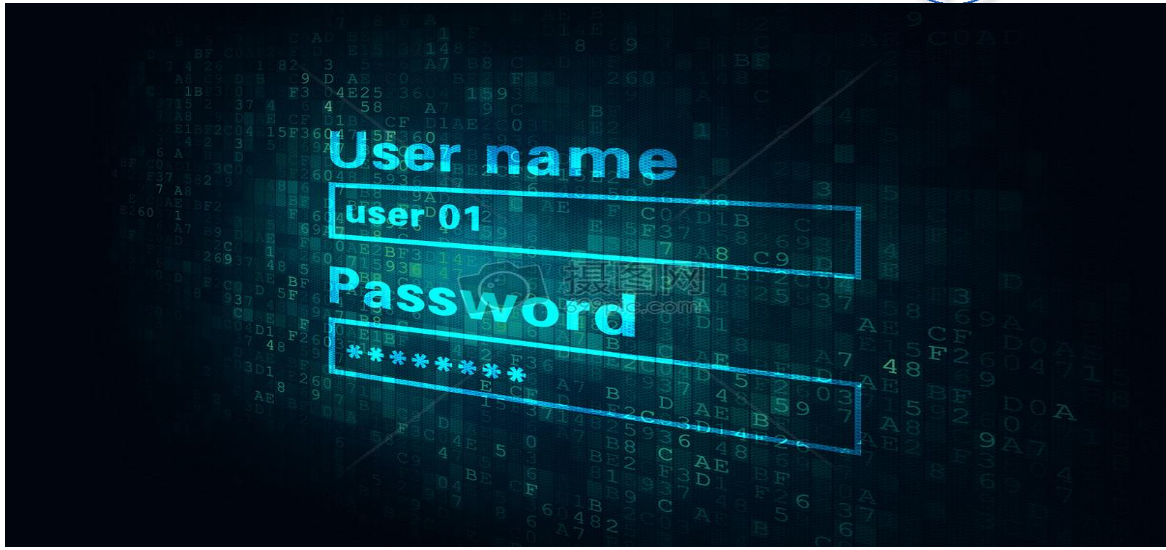
③ directly proving your identity based on unique physical characteristics

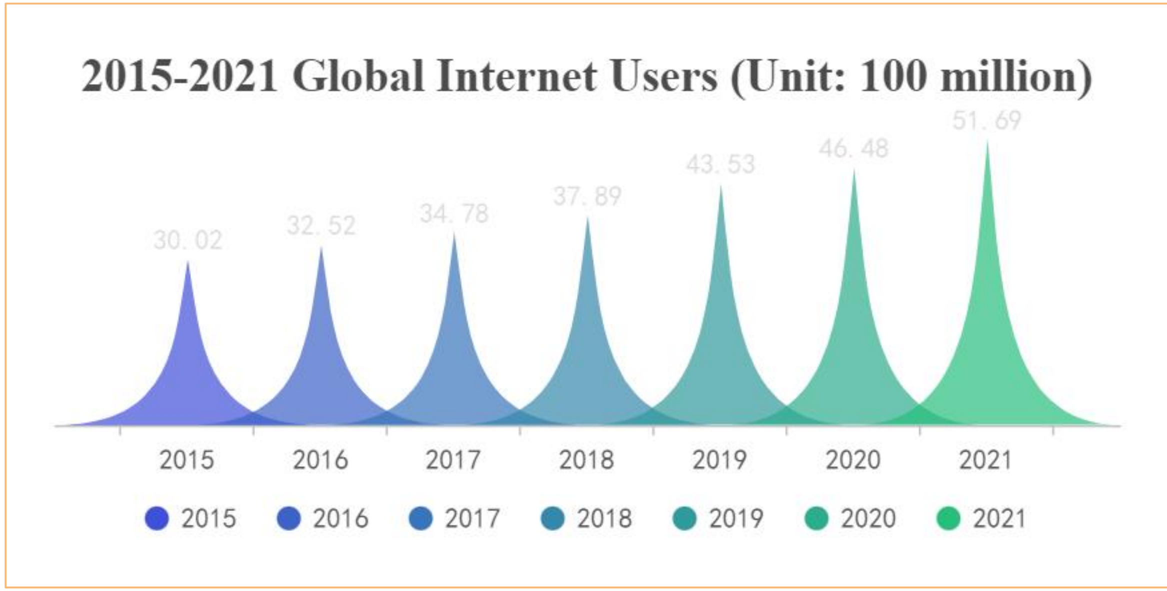
01 Present situation

According to the China Internet Network Information Center (CNNIC), by December 2022, the number of Internet users in China has reached 1.067 billion.

Password security issues have long been a concern. Therefore research on password guessing is necessary.

Passwords are still the most commonly used identity authentication method due to their strong universality, high security, and low cost.





01 Present situation

Development of password guessing(fundamental)

1979 — **Robert Morris and Ken Thompson**

Brute-Force Attack and Dictionary Attack

1980 — **Hellman**

Time-Memory Trade-Off (TMTO) method

2003 — **Oechslin**

Rainbow table

original method

01

Present situation

Development of password guessing(fundamental)

2005 Narayanan and Shmatikov

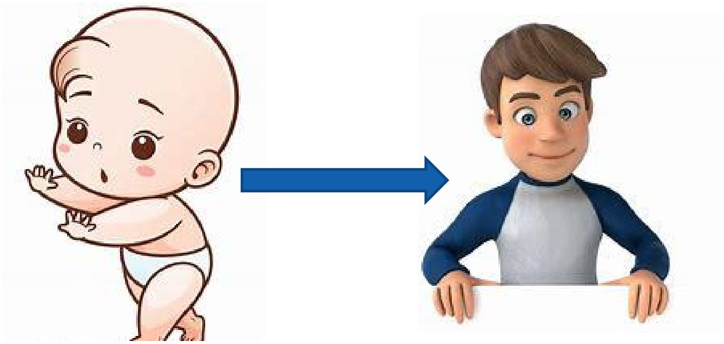
Markov model

2009 Weir et al.

Probabilistic Context Free Grammar (PCFG)

2016 Melicher's team

Recursive Neural Network(RNN)



Start considering the models

01 Present situation

LANJIANJUNXIN IARIA

Development of password guessing(fundamental)

2019 **Hitaj et al.**

Adversarial Networks (GANs)==>PassGAN

2022 **He et al.**

Transformer==>PassTrans

2022 **Sanjay et al.**

Bidirectional Generative Adversarial Network(BiGAN)==>PassMon

2023 **Rando et al.**

large language models==>PassGPT

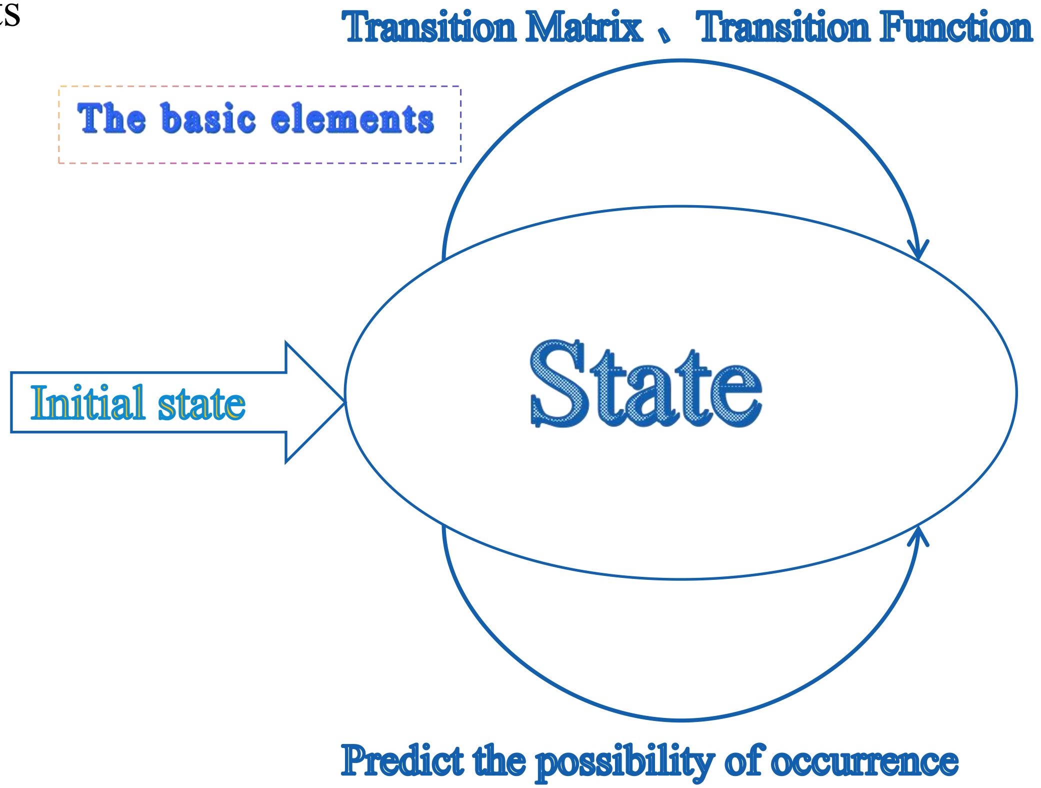
Deep learning models

02 Methods

Markov model family

Transition Probability : $P_{ij}=P(X_n=j|X_{n-1}=i)$
X represents the state at a certain moment, for example, X_n represents the state at time n.

The reason why Markov model can be used for password guessing is that a Markov model defines a probability distribution on a symbol sequence. In other words, it allows for sampling of character sequences with certain attributes.



zero-order model : $P(\alpha)=\pi_{x\in\alpha}v(x)$
first-order model : $P(x_1x_2...x_n)=v(x_1)\Pi_{i=1}^{n-1}v(x_{i+1}|x_i)$

An example: “loveyou” (first-order)
$P(loveyou)=P(lo)P(v|lo)P(e|ov)P(y|ve)P(o|ey)P(u|yo)$

02 Methods

Markov model family

single model

2005

Narayanan and Shmatikov

Since then, password guessing has entered a "new era".

2015

Dürmuth et al.

A new method, using an ordered Markov enumerator(OMEN), based on Markov model has been proposed.

2021

Guo et al.

A dynamic mechanism called the Dynamic Markov Model was proposed in 2021.

02 Methods

Markov model family

single model

Ordinary Markov	OMEN	Dynamic Markov
• The drawbacks of the Markov model are evident, as it generates a large amount of duplicate data when cracking passwords, resulting in high repetition rates and low coverage, resulting in resource waste.	• Compared with ordinary Markov, OMEN improves the speed of password guessing, which is only close to the probability of password guessing.OMEN is a deterministic algorithm.	• The purpose of the dynamic mechanism is to reduce the repetition rate and to improve coverage.And it solves the problem of OMEN always generating the same password in the same order.

02 Methods

Markov model family

Coverage

Probability	Total	Markov Model	OMEN	Dynamic Markov Model
$<10^{-12}$	310024	161	817	68
$[10^{-12},10^{-11})$	231826	1168	26988	980
$[10^{-11},10^{-10})$	334797	13507	1409272	13275
$[10^{-10},10^{-9})$	410065	95250	287433	122463
$[10^{-9},10\text{-}8)$	390731	272115	340801	350988
$[10^{-8},10^{-7})$	284911	271945	268945	284897
$[10^{-7},10^{-6})$	116095	115973	112954	116094
$[10^{-6},10^{-5})$	19827	19826	19456	19827
$[10^{-5},10^{-4})$	2701	2701	2671	2701
$[10^{-4},10^{-3})$	243	243	243	243
$[10^{-3},10^{-2})$	4	4	4	4

02

Methods

Probabilistic Context Free Grammar family

PCFG is an extension of Context Free Grammar (CFG), which is a method of Rule Based Natural Language Processing (NLP). The main function of CFG is to verify whether the input string conforms to a certain grammar G, which is similar to regular expressions, but CFG can express more complex grammars.

PCFG checks grammar structures (combinations of special characters, numbers, and alphanumeric sequences) and generates distribution probabilities, which are then used to generate candidate passwords.

02 Methods

PCFG family

2009
Weir et al.

Weir et al. [4] used only Ln, Dn and Sn (L represents letters, D represents numbers, and S represents special characters.) for the specified n-value in grammar, except for the starting symbol. They call these variables alpha variables, digit variables and special variables respectively.

1

2

Keyboard order (keyboard mode) and multi word strategy can greatly improve the PCFG. Through its development, PCFG not only allows for guessing passwords in probabilistic order, but also fully considers keyboard mode, resulting in higher cracking coverage.

2015
Houshmand et al.

2022
Guo et al.

They proposed a degenerate distribution collection method i and designed a corresponding Low Probability Generator Probabilistic Context Free Grammar (LPG-PCFG) model based on PCFG.

3

4

Due to insufficient investigation of cryptographic semantic information, Wang et al. proposed a general framework for PCFG based on semanticenhancement, named SE # PCFG.

2023
Wang et al.

02 Methods

PCFG family

Keyboard base structures VS PCFG

Passwords	PCFG	Keyboard
1234	D_4	K_4
w2w2	LDLD	K_4
ASD1234QW	$L_3D_4L_2$	$K_3D_4L_2$
Q112	LDSD	K_4

X is a password.

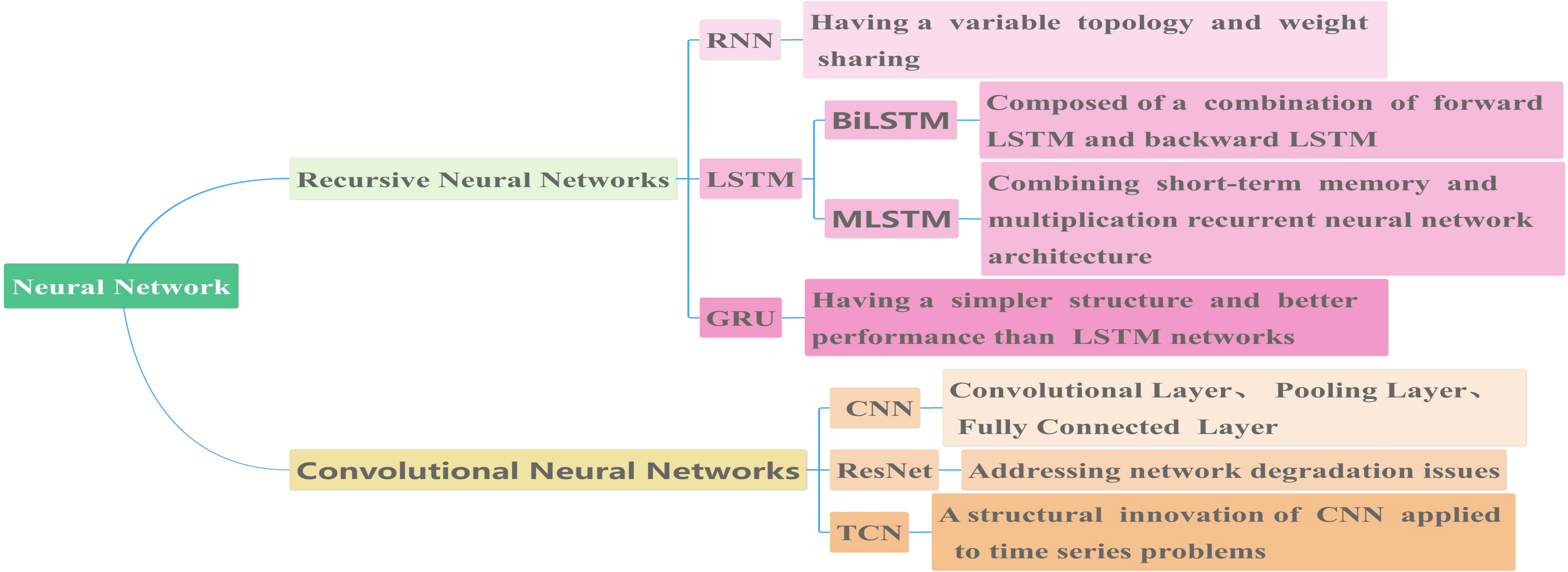
Modification rules degeneration distribusion

Rule	Adjust $p(x^+)$	Adjust $p(x^-)$
Rule1	$p(x^+)-\alpha$	$p(x^-)+\alpha/(N_s-1)$
Rule2	$p(x^+)-\alpha$	$p(x^-)+\alpha/(1-p(x^+))p(x^-)$
Rule3	$\beta p(x^+)$	$p(x^-)+(1-\beta)p(x^+)(1-p(x^+))p(x^-)$
Rule4	$\beta p(x^+)$	$p(x^-)(1-\beta)p(x^+)(1-p(x^+))p(x^-)$
Rule5	$1-\gamma(1-p(x^+))$	$\gamma p(x^-)$

02 Methods

Neural Network model family

Using Neural Networks to simulate the resistance of passwords to guessing attacks can be more effective than Markov models and PCFG. Neural network modeling uses less space than Markov models, and neural network can transfer knowledge from a task to related tasks.



02 Methods

Neural Network model family

Long Short Term Memory: LSTM

2019

Li et al.

In 2019, Li 's team also applied BiLSTM to password guessing.

2018

Zhang et al.

They combined the advantages of PCFG and neural networks to propose a password guessing method based on structural partitioning and Bidirectional Long Short-Term Memory Recursive Neural Networks, named the SPRNN model.

2016

Melicher et al.

Melicher's team used an LSTM network, which could solve the gradient explosion problem caused by long text, including memory gates, forgetting gates, and output gates. Determine which information is discarded and which information is left through three "gates".

02 Methods

Neural Network model family

Temporal Convolutional Network: TCN

2022

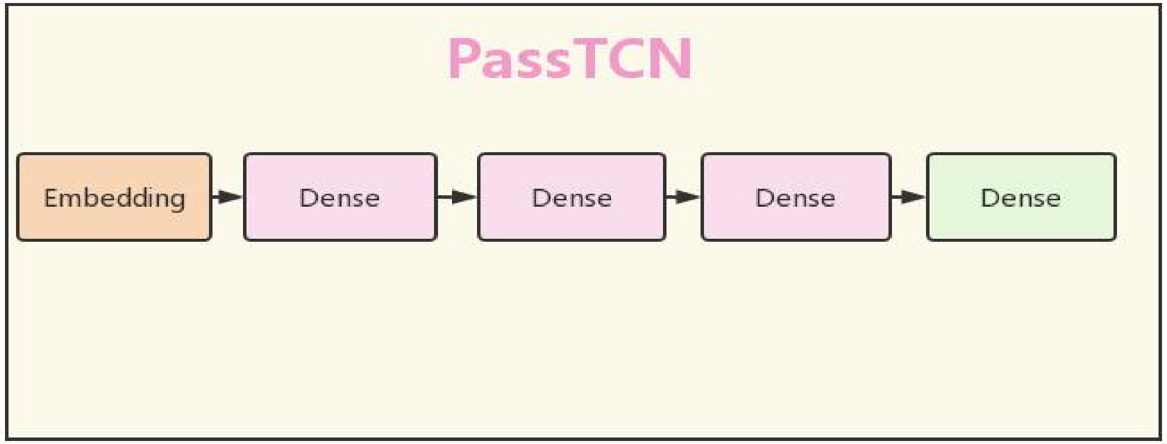
Ye et al.

TCN is an algorithm used to solve time series prediction.Ye's team proposed a password guessing model based on TCN, named PassTCN.In order to further improve the performance of password generation, a new password probability label learning method is also proposed.

2022

Chang et al.

Addressing the difficulty of selecting sequence length in traditional LSTM models for password generation, they considered user personal information and proposed a multi sequence length LSTM password guessing model.



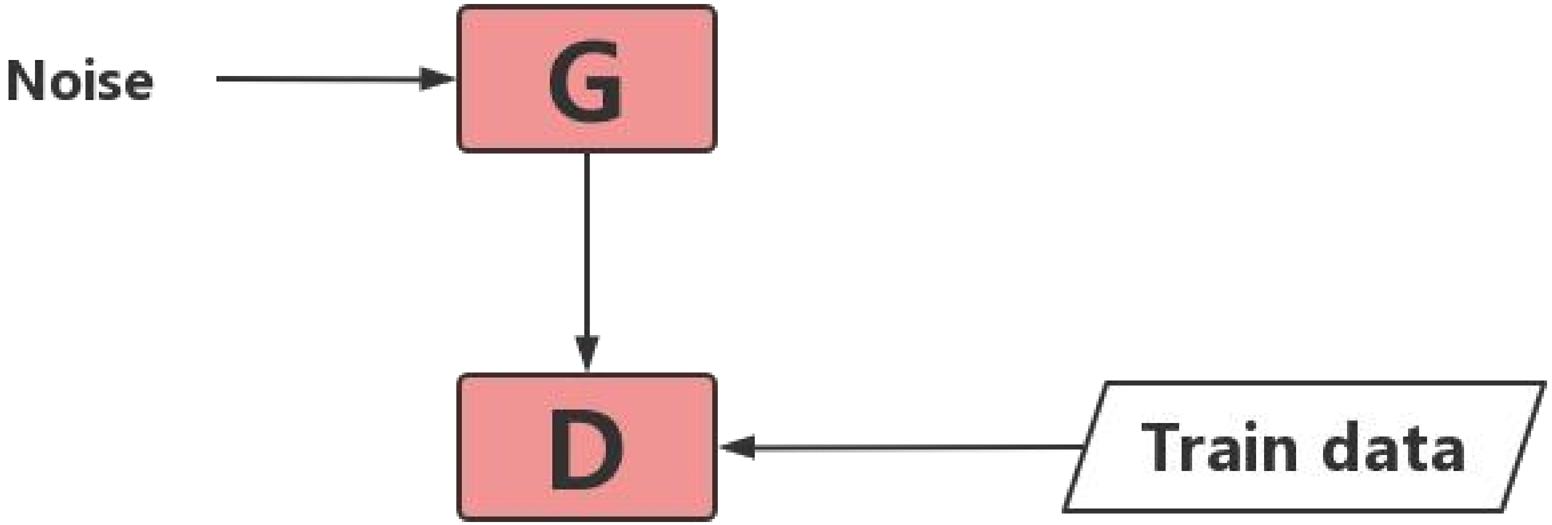
02 Methods

GAN model family

The optimization objective function of GAN: $\min_{G} \max_{D} V(D,G)$

The loss function of GAN: V(D,G)=$E_{X\sim P_{data}(x)}[logD(x)] + E_{z\sim P_{z}(z)}[log(1-D(G(z)))]$

GANs are inspired by the zero sum game theory, which consists of two parts: a generative model (G) and adiscriminant model (D).



GAN model family

Hitaj et al. 2019

PassGAN does not rely on password analysis like Markov models, PCFG, and neural networks, but instead uses GAN to automatically learn the true password distribution from publicly leaked passwords. In other words, we do not need any professional knowledge related to cryptography, and applying GAN can generate high-quality passwords for guessing.

Nam et al. 2020

Nam's team proposed a candidate password for optimizing guessing, named REDPACK using a relativistic GAN method.REDPACK effectively combines multiple generation models to generate passwords. Generator G can effectively optimize candidate password selection by selecting different models, such as OMEN, PCFG, etc.

Jiang et al. 2022

They proposed a password generation model based on ordered Markov enumeration and discriminant networks (OMECDN) for PassGAN and added gradient normalization to PassGAN.

GAN model family

Yu et al. 2022

They found that the combination of IWGAN and gradient penalty is not an ideal method to solve the shortcomings of GAN, so they added gradient normalization counting to discriminator D and named it GNPassGAN.

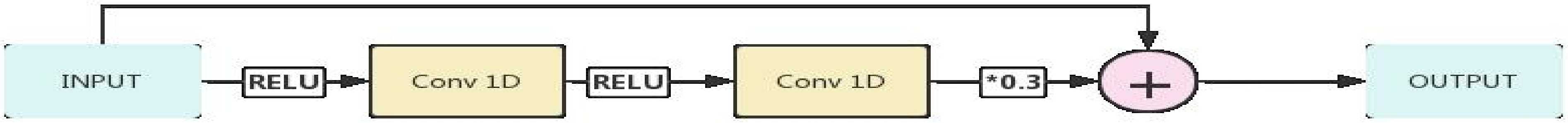
Zhou et al. 2022

They proposed a new structure based on PassGAN, which uses LSTM network for generator G and multiple convolutional layers in discriminator D, based on the non differentiability of discrete data sampling process and the impact on backpropagation. In addition, the biggest contribution is the addition of Gumbel SoftMax, named G-Pass.

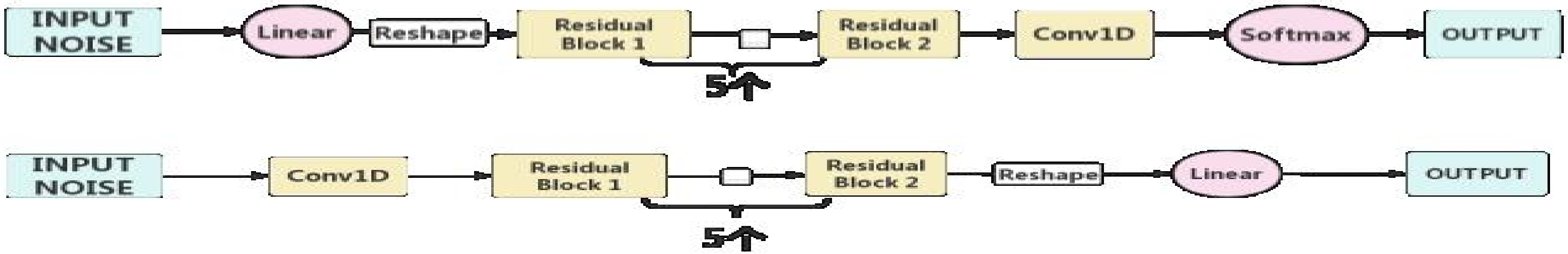
02 Methods

GAN model family

Structure



Residual Block′ s Architecture.



PassGAN′ s Architecture.Upper generator G, lower discriminator D.

02 Methods

models

Model Name	Basic Generation Model Types	Publication Year
Markov	Markov	2005
PCFG	PCFG	2009
OMEN	Markov	2015
Next Gen PCFG	PCFG	2015
FLA	RNN,LSTM	2016
PassGAN	GAN,IWGAN	2017,2019
GENPass	PCFG,LSTM	2018,2020
SPRNN	BiLSTM	2018
BiLSTM	BiLSTM	2019
REDPACK	PCFG,GAN,etc.	2020
Dynamic Markov	Dynamic Markov	2021
GNPassGAN	GAN	2022

02 Methods

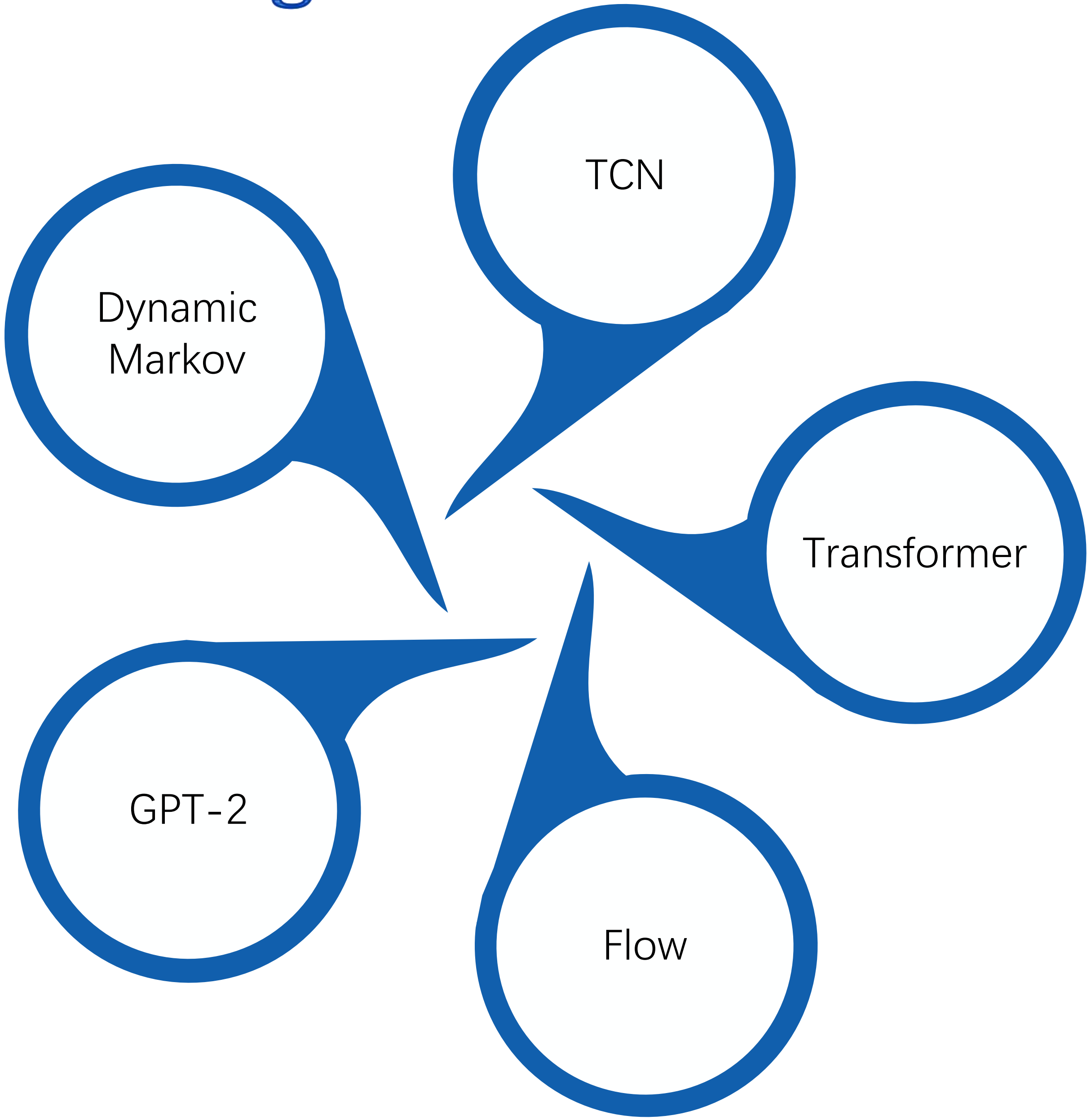
models

Model Name	Basic Generation Model Types	Publication Year
PassTCN-PPLL	TCN	2022
LPG-PCFG	PCFG	2022
G-Pass	GAN	2022
Passtrans	Transformer	2022
OMECDN	Markov,GAN	2022
PassMon	BiGAN	2022
MLSTM	MLSTM	2022
PassFlow	Flow	2021,2022
WordMarkov	Markov	2022
SE#PCFG	PCFG	2023
PassGPT	GPT-2	2023
PassTCN	TCN	2023

03 Key Technologies

03

Key Technologies



04 Future research and recommendations

1

Public password data is not easy to find. We can consider data augmentation technology (DA) to obtain more data. Note that the application of DA technology should not aimlessly expand the data, as obtaining poor data can lead to worse results. We need to clean the obtained data and eliminate bad data. According to research, some users will set their passwords based on the topic of the website. Password guessing often requires a dictionary, and we can use the DA method to obtain as many websites with the same topic as possible. Based on the special password generation strategy of website themes, a dictionary is generated for password guessing.

2

Spectral Normalization (SN) can improve the stability of discriminator D in GAN, which is also a variant of GAN. The work we are doing not only applies SN to discriminator D, but also adds SN to generator G. Multiple variants of GAN for password guessing may achieve better results. Of course, model training requires the use of optimizers.

04 Future research and recommendations

3. There are already models in the literature other than Markov models, PCFG, GAN, etc. applied to password guessing. We hope that more types of neural networks and deep learning models can be applied to password guessing.

4. Password rules cannot be ignored, as most literature does not consider password rules, and the rules required by different websites may vary. Considering the combination of password setting rules and the topic dictionary mentioned above, we hope to apply them simultaneously to password guessing.

5. We need to detect password leaks, and Honeywords is a type of bait password used to provide feedback on password leaks. As an effective method for detecting whether passwords have been cracked, how to generate Honeywords better has become a research direction.

IARIA UNIVERSITY OF JINAN 1948

蓝剑钧新 LANJIANJUNXIN

05 Conclusion

01

- A systematic review was conducted on the password guessing methods mentioned in the references, with some models providing method details.

02

- Introduce improvement methods based on the original model by class, and each method improves the original method.

03

- Discuss the limitations of password guessing and propose feasible future research directions based on new technologies.
- Mention three methods for optimizing password guessing.

The Past and Possible Future Development of Password Guessing

The Twelfth International Conference on Communications, Computation, Networks and Technologies
INNOV 2023

Authors:Ze-long Li[1]、Teng Liu[2]、Lei Li[3]

Presenter： Ze-long Li
Affiliation： [1,2]University Of Jinan
Email:202121200928@stu.ujn.edu.cn

THANK YOU FOR WATCHING

Thank you