IPv6 Network Measurement

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Gaopeng Gou
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences

www.iie.ac.cn

Recent research on Ipv6 measurement

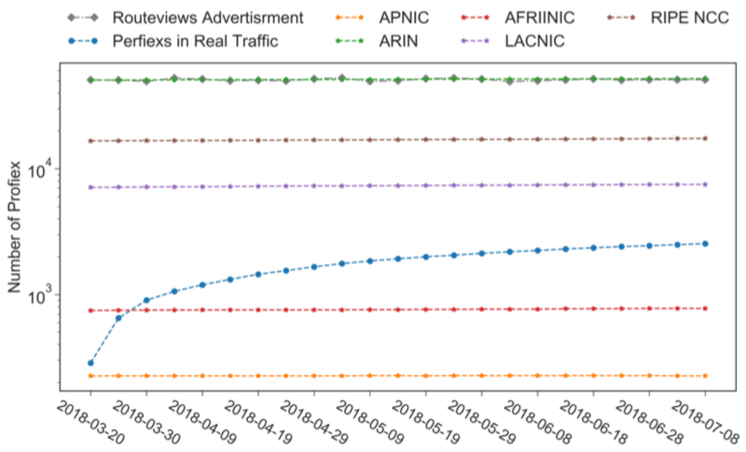
- IPv6 Deployment
  - ✓ A Comprehensive Study of Accelerating IPv6 Deployment
- IPv6 Target Generation
  - ✓ IPv66GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation
  - ✓ 6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

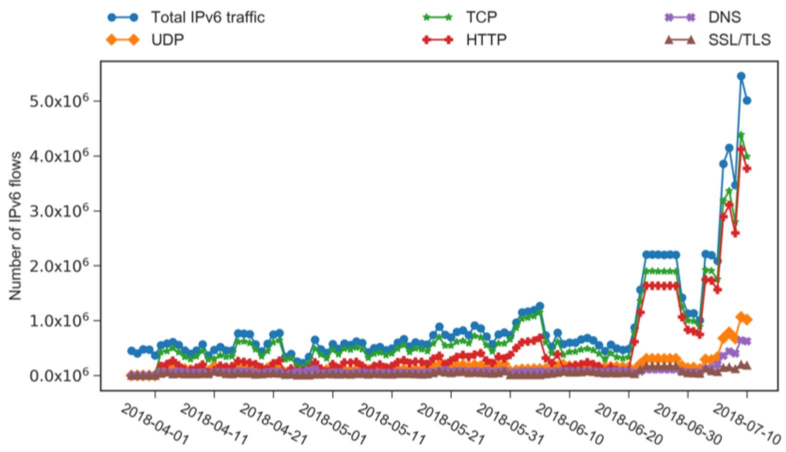
Background

2002 – 2014
- IPv6 traffic has increased over 400% per year. Czyz et al.

2016
- The increase rate of traffic volume and advertised prefixes are 130% and 65%. Pickard et al.

2017
- China circulated an Action Plan for Promoting Scale Deployment of Internet Protocol Version 6 (IPv6).
  - 10-fold IPv6 traffic volume and 3-fold IPv6 prefixes increase
  - Problems behind outbreak of traffic and performance improvement events
  - A broader approach combined with multiple means



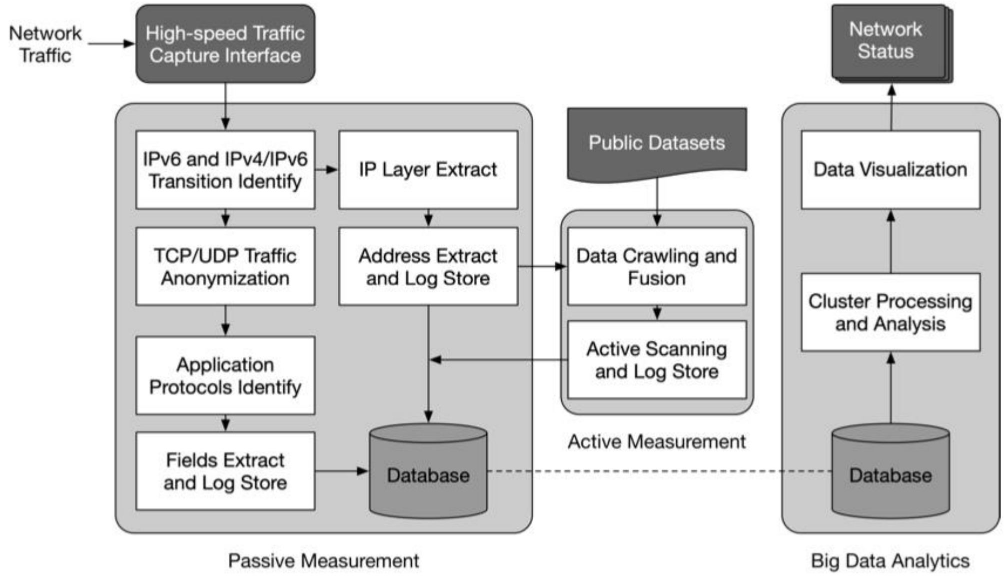


Methodology

Passive Measurement
- On the network named China Unicom from March to July in 2018

Active Measurement
- IPv6 TCP SYN scan for the Alexa Top sites and active IPv6 addresses

Big Data Analytics
- Selectively analyze from one or more standards.



Methodology

Normal Status (Category N)
- Contrast global public datasets and IPv4 real traffic

Accelerating Status (Category A)
- Captured during the accelerating deployment

Dataset	Time Peroid	Scale	Category	Collection
MAXMIND GeoLite2 City	10 July 2018	global Geo-IP dataset	N	Public
RIR Address Allocations	20 March - 10 July in 2018	565 daily allocation snapshots		
Routing: Route Views	20 March - 10 July in 2018	113 BGP table snapshots		
Google IPv6 Client Adoption	20 March - 10 July in 2018	daily global samples		
Verisign TLD Zone Files	20 March - 10 July in 2018	daily snapshots of A and AAAA records (.com & .net)		
Alexa Top Sites	10 July 2018	Top 10K global sites, 1,696 Chinese sites in Top 1M		
CAIDA IPv6 Day and Launch Day	8 June 2011 and 6 June 2012	1TB pcap files of anonymized passive traffic	A	
IPv6 Real Traffic Dataset	20 March - 10 July in 2018	170 million flows captured in 5 months	A	Passive Measurement
IPv4 Real Traffic Dataset	20 March - 30 April in 2018	35 million flows captured in 1 month	N	

A Comprehensive Study of Accelerating IPv6 Deployment

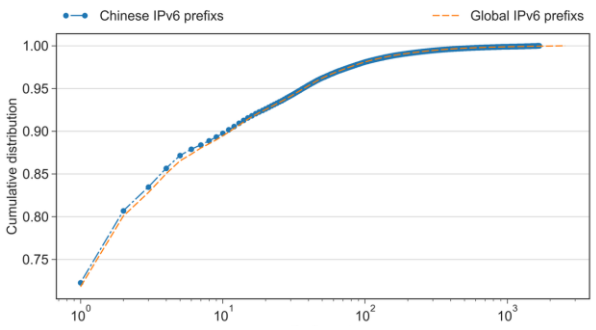
Network Status

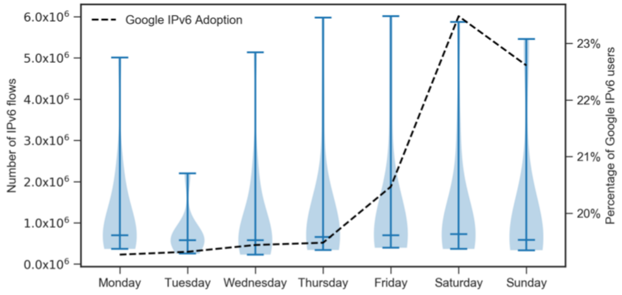
Address Distribution
- Uneven geographical distribution
- Excessive concentration of prefixes

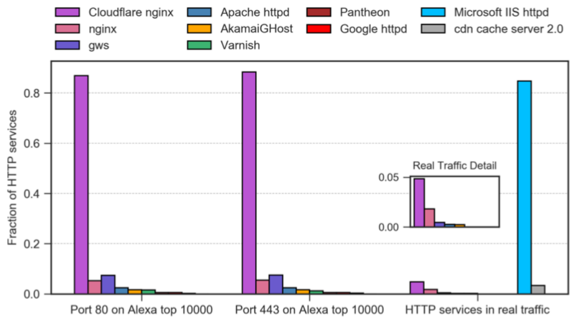
Traffic Trend
- Explosive growth of IPv6 active addresses and traffic volume
- IPv6 traffic usage is different from the traditional website access services

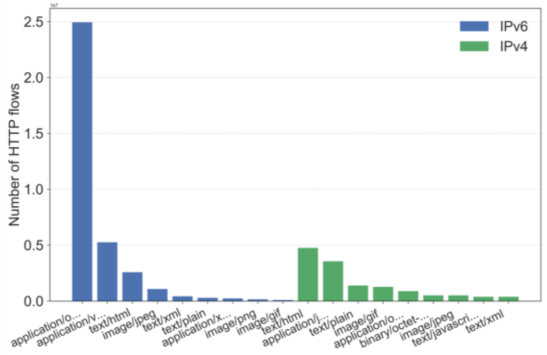
Service Deployment
- Serious lack of IPv6 website and secure service deployment
- Huge shift on the server software deployment on IPv6 websites

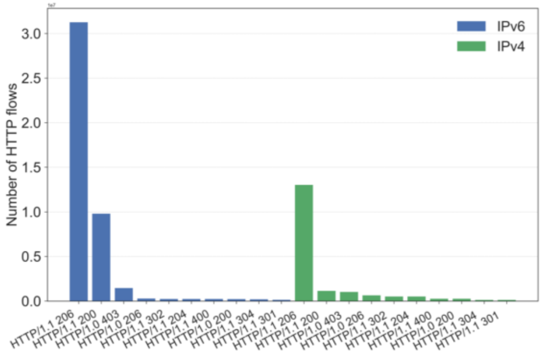
Protocols
- Content Type and Response line in HTTP traffic is different from IPv4











Issues

User Privacy Threat
- Insufficient encrypted application usage
- low content encryption rates

Inappropriate Access Methods
- Excessive IPv4/IPv6 transition usage

Common Nature
- Accelerated deployment of protocol growth is similar
- Imbalance between HTTP and SSL/TLS traffic growth

Future Work

Conclusion
- the current accelerating status is unbalanced and unstable
- accelerating status exposes unresolved issues
- The improvement of network performance conflicts with the challenge of network security and stability.

Key
- IPv6 websites and network services construction
- Security web application for IPv6 users

Recent research on Ipv6 measurement

- □ IPv6 Deployment
  - ✓ A Comprehensive Study of Accelerating IPv6 Deployment
- □ IPv6 Target Generation
  - ✓ IPv66GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation
  - ✓ 6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

PAKDD 2020

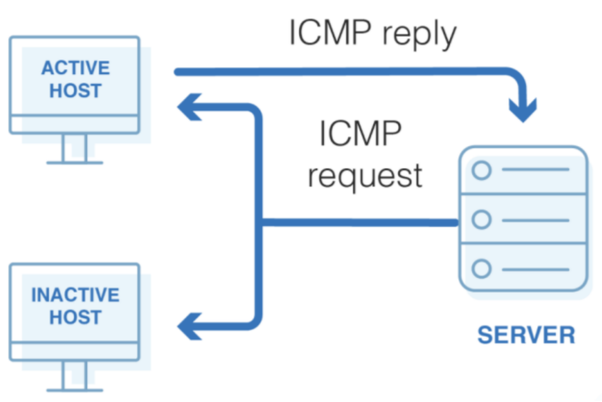
IPv6 Scanning

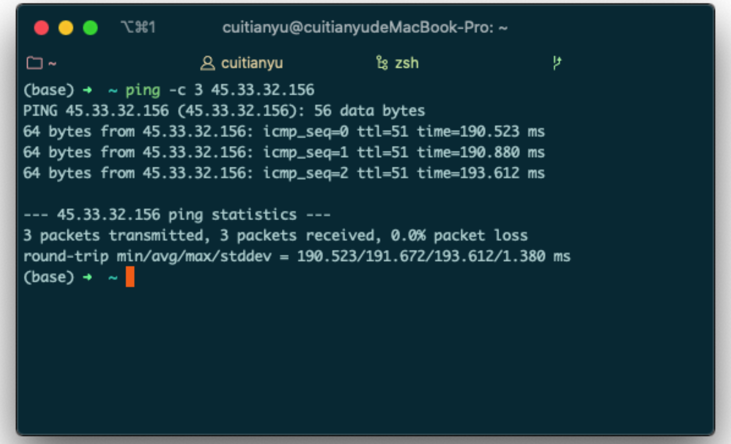
What is Network Scanning

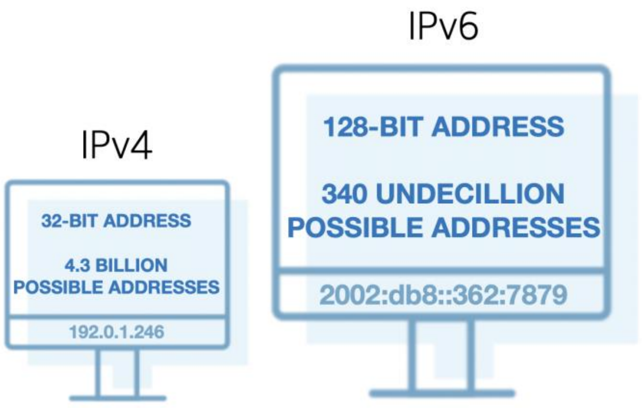
- A method to determine all active devices on the network.
- The system sends a ping to each device on the network and awaits a response by using a protocol (e.g. ICMP).
- Global IPv4 scanning has fundamentally enhanced the ability of researchers to conduct wide-ranging assessments of Internet services.

Challenge

- IPv6 has a 128-bit address space.
- Using a brute-force approach to probe the entire network space of IPv6 is completely infeasible.







6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

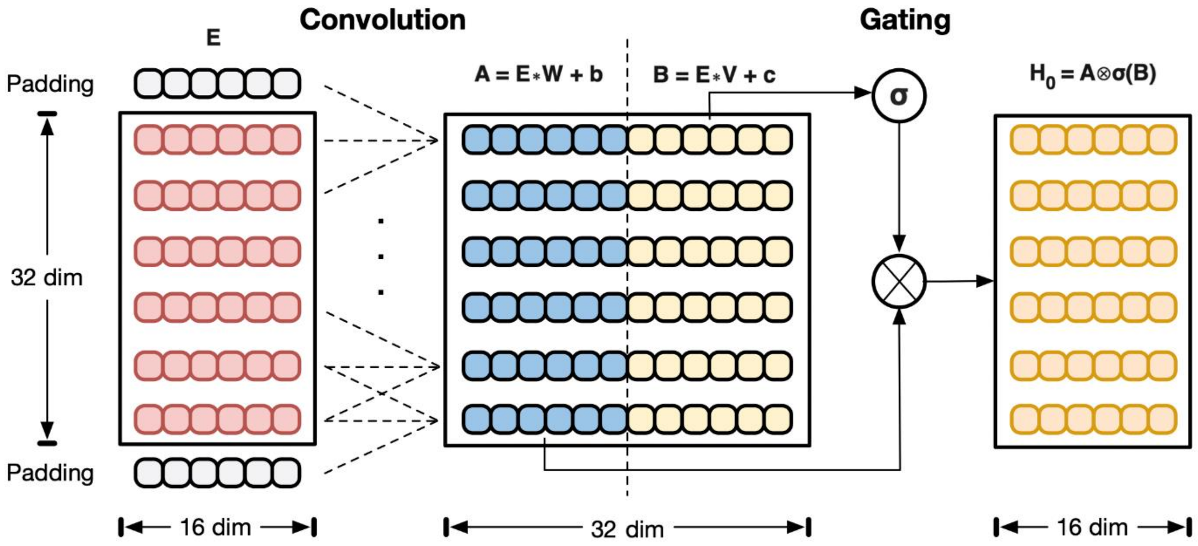
PAKDD 2020

- Replace target generation algorithms with deep learning architecture – 6GCVAE
- Analyze address structure – Gated convolutional networks
- Generate candidate sets – VAE architecture
- Solve the multiple addressing scheme problem – Seed classification

6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

PAKDD 2020

Gated Convolutional Networks



Why gated convolutional networks

- Convolution - focus on structure and nybble relationships.
- Gating - monitor the important nybbles in an address.

Convolution

- Convert hexadecimal address to 32 × 16 One-Hot Input E. (Address nybbles → 32; 0-f alphabet → 16)
- 32 3 × 16 convolution kernels output 2 32 × 16 vectors A and B.

Gating

- Vector B with sigmoid function as the gate to control the output of vector A.
- Hidden layer H :

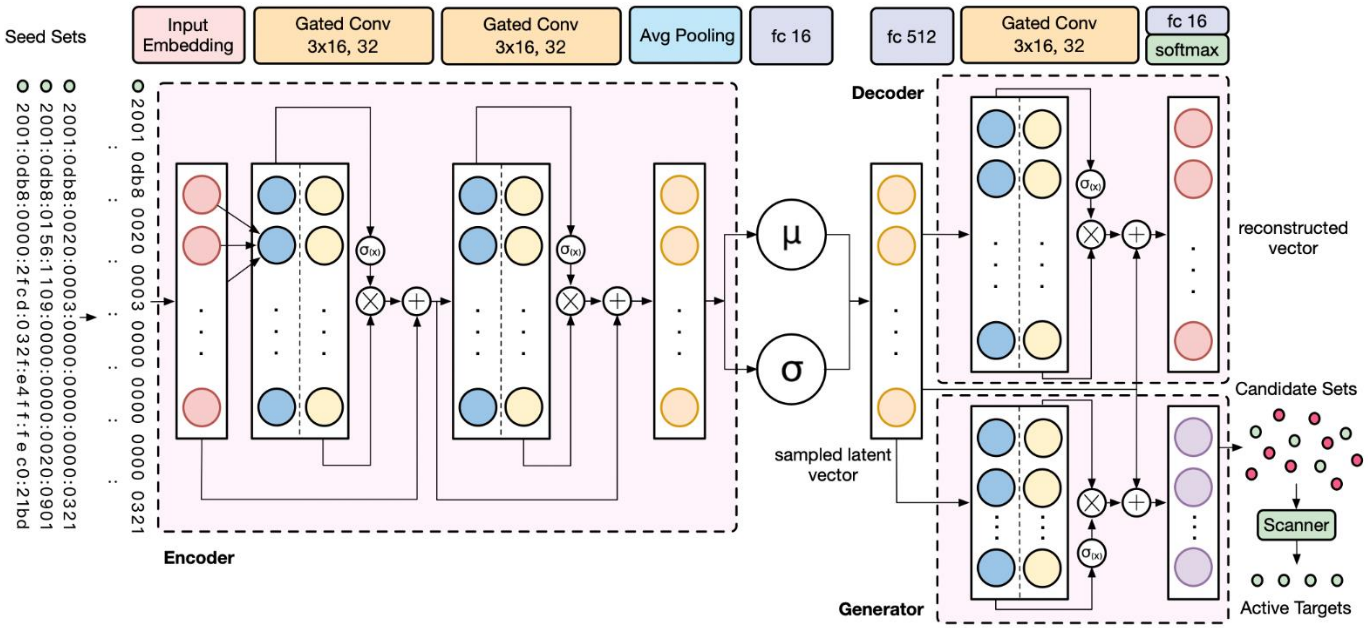
$$H_i = A \otimes \sigma B$$

- σ is the sigmoid function. ⊗ is the element-wise product between matrices.

6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

PAKDD 2020

6GCVAE



Encoder
- Two gated convolutional layers and an average pooling layer with a residual connection between each layer.
- Two fully connected layers to train the mean μ and the log variance log σ^2.

Decoder
- A gated convolutional layer, a fully connected layer, and a softmax activation.
- Generate a reconstructed vector for calculating reconstruction error.

6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

PAKDD 2020

Seed Classification

Early classification of seeds with different structural patterns can help to improve model performance.

Manual Classification

- Fixed IID - a unique consecutive 0 in address
- Low 64-bit subnet - two or more consecutive 0 segments in address
- SLAAC EUI-64 - the 23rd-26th nybbles fffe flag
- SLAAC privacy - pseudo-random address

Unsupervised Clustering

- Calculate entropy value $H(X_i)$ on each nybble.
- Build fingerprint F_b^a by using the nybble entropy.

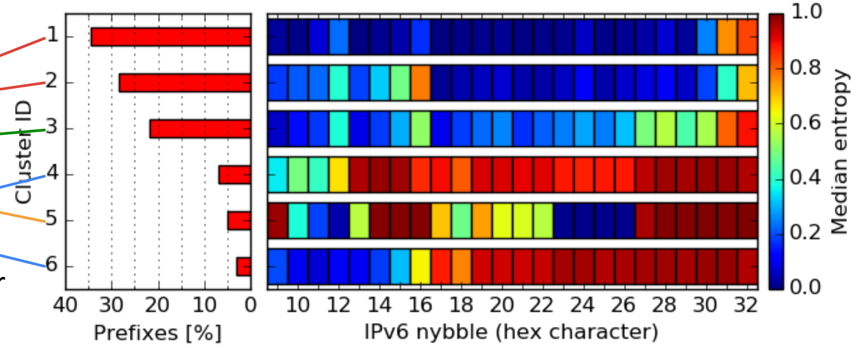
$$F_b^a = (H(X_a), ..., H(X_i), ..., H(X_b))$$

$$H(X_i) = -\frac{1}{4} \sum_{x \in \Omega} P(x_i) \cdot log\ P(x_i)$$

- Perform entropy clustering by using the fingerprint.

Category	Feature	Seeds	Percentage
Fixed IID	The last 16 nybbles have a consecutive 0	1,208,117	38.26%
Low 64-bit Subnet	The last 16 nybbles have more consecutive 0	1,062,093	33.64%
SLAAC EUI-64	The 23-26th nybbles is fffe.	279,458	8.85%
SLAAC Privacy	Entropy value of the last 16 nybbles > 0.8	608,007	19.25%
Total	-	3,157,675	100%

similar



6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

PAKDD 2020

Dataset and Evaluation Method

Dataset

- IPv6 Hitlist - Public dataset. [1]
- CERN IPv6 2018 - Passively collected address sets under the China Science and Technology Network from March to July 2018.

Dataset	Seeds	Period	Collection Method
IPv6 Hitlist	3,157,675	October 14, 2019	Public
CERN IPv6 2018	90,010	March 2018 - July 2018	Passive measurement

Evaluation Method

- Zmapv6 tool. [2]
- ICMPv6, TCP/80, TCP/443, UDP/53, UDP/443 scanning.
- 3 days continuously scanning to ensure the accuracy.

Evaluation Metric

- $N_{candidate}$ - Number of the generated candidate set
- N_{hit} - Number of generated active addresses
- N_{new} - Active addresses which are not in the seed set

$$r_{hit} = \frac{N_{hit}}{N_{candidate}} \times 100\% \qquad r_{gen} = \frac{N_{new}}{N_{candidate}} \times 100\%$$

- r_{hit} - learning ability to learn from the seed set.
- r_{gen} - generation ability to generate new active addresses.

[1] Gasser, O., Scheitle, Q., Foremski, P., Lone, Q., Korczyński, M., Strowes, S.D., Hendriks, L., Carle, G.: Clusters in the expanse: Understanding and unbiasing ipv6 hitlists. In: Proceedings of the Internet Measurement Conference 2018. pp. 364–378. ACM (2018)
[2] IPv6 Hitlist. https://ipv6hitlist.github.io/

6GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation

Experimental Results

Seed Classification Results

- Manual Classification - Fixed IID
- Unsupervised Clustering - Cluster 2
- Seed Classification can improve model performance.

Model Performance

- 5 Conventional VAE models - GRU VAE is the best but not competent for the task.
- Target generation algorithm - Entropy/IP performs better than conventional VAE models.
- 6GCVAE reaches the best performance.

Seed Classification	Category	$N_{candidate}$	N_{hit}	N_{new}	r_{hit}	r_{gen}
None	IPv6 Hitlist	756,658	14,894	9,685	1.97%	1.28%
Manual Classification	Fixed IID	412,181	32,589	**17,933**	7.91%	**4.35%**
	Low 64-bit Subnet	901,222	7,092	5,450	0.79%	0.61%
	SLAAC EUI-64	981,204	1,299	1,263	0.13%	0.13%
	SLAAC Privacy	999,920	13,351	13,351	1.34%	1.34%
Unsupervised Clustering	Cluster 1	526,542	25,235	12,364	4.79%	2.35%
	Cluster 2	450,919	57,245	**35,508**	12.70%	**7.87%**
	Cluster 3	759,617	5,273	2,404	0.69%	0.32%
	Cluster 4	985,390	6,605	6,309	0.67%	0.64%
	Cluster 5	832,917	1,748	845	0.21%	0.10%
	Cluster 6	968,178	1,193	994	0.12%	0.10%

Model	$N_{candidate}$	N_{hit}	N_{new}	r_{hit}	r_{gen}
FNN VAE	1,000,000	68	68	0.007%	0.007%
RNN VAE	498,509	3,009	2,085	0.604%	0.418%
Convolutional VAE	595,475	4,432	2,856	0.744%	0.480%
LSTM VAE	478,660	4,464	3,203	0.933%	0.669%
GRU VAE	525,134	5,694	4,548	1.084%	0.866%
Entropy/IP	593,795	15,244	5,402	2.570%	0.910%
6GCVAE	756,658	14,894	9,685	1.970%	1.280%
6GCVAE with Manual Classification	557,653	28,957	15,870	5.193%	2.846%
6GCVAE with Unsupervised Clustering	571,330	54,915	**31,376**	9.611%	**5.492%**

Recent research on Ipv6 measurement

- IPv6 Deployment
  - ✓ A Comprehensive Study of Accelerating IPv6 Deployment
- IPv6 Target Generation
  - ✓ IPv66GCVAE: Gated Convolutional Variational Autoencoder for IPv6 Target Generation
  - ✓ 6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

Target Generation Challenge

Challenge 1 - Missing semantics

Target Generation can take place of traditional IPv6 scanning. - phrase

2001:0db8:0106:0001:0000:0000:0000:0003 - what relationship ?

- IPv6 address entirely consists of digits.
- Inability to infer active addresses using sequence relationships.

Challenge 2 - Complexity of address composition

	Human-readable Text Format	Commonly Used Address Format
Fixed IID	2001:0db8:0106:0001:0000:0000:0000:0003	2001:db8:106:1::3
Low 64-bit Subnet	2001:0db8:0100:0015:0000:0000:000a:0005	2001:db8:100:15::a:5
SLAAC EUI64	2001:0db8:0000:4144:f816:3eff:fe57:0e6d	2001:db8:0:4144:f816:3eff:fe57:e6d
SLAAC Privacy	2001:0db8:fbd0:0021:7c61:2880:3148:36e1	2001:db8:fbd0:21:7c61:2880:3148:36e1

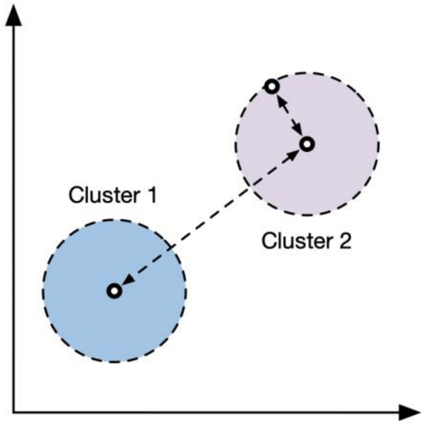
- 2001:0db8:0106:0001:????:????:????:???? - How to determine ?
- Multiple IPv6 schemes cause difficulty in algorithmic inferences.

6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

IPv6 Address Space

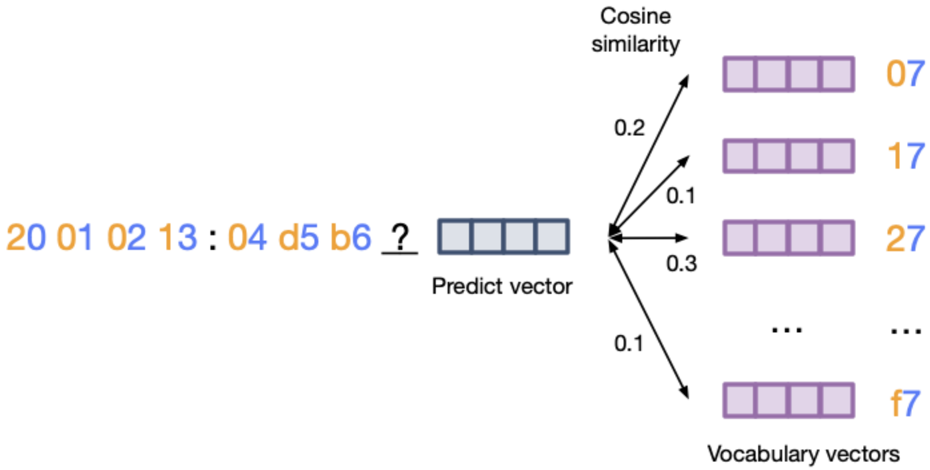
128-bit address space →



- Map IPv6 address space into semantic vector space.
- The distance between vectors can be defined as the relationship between addresses.
- Similar addresses will be clustering, which helps to find the target clustering area.

IPv6 Semantic

- Build address words and train word vectors.
- Language modeling to predict address vectors.
- cosine similarity as the next word probability for predicting address located in the target clusters with similar semantic.



6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

IPv62Vec

Word Building

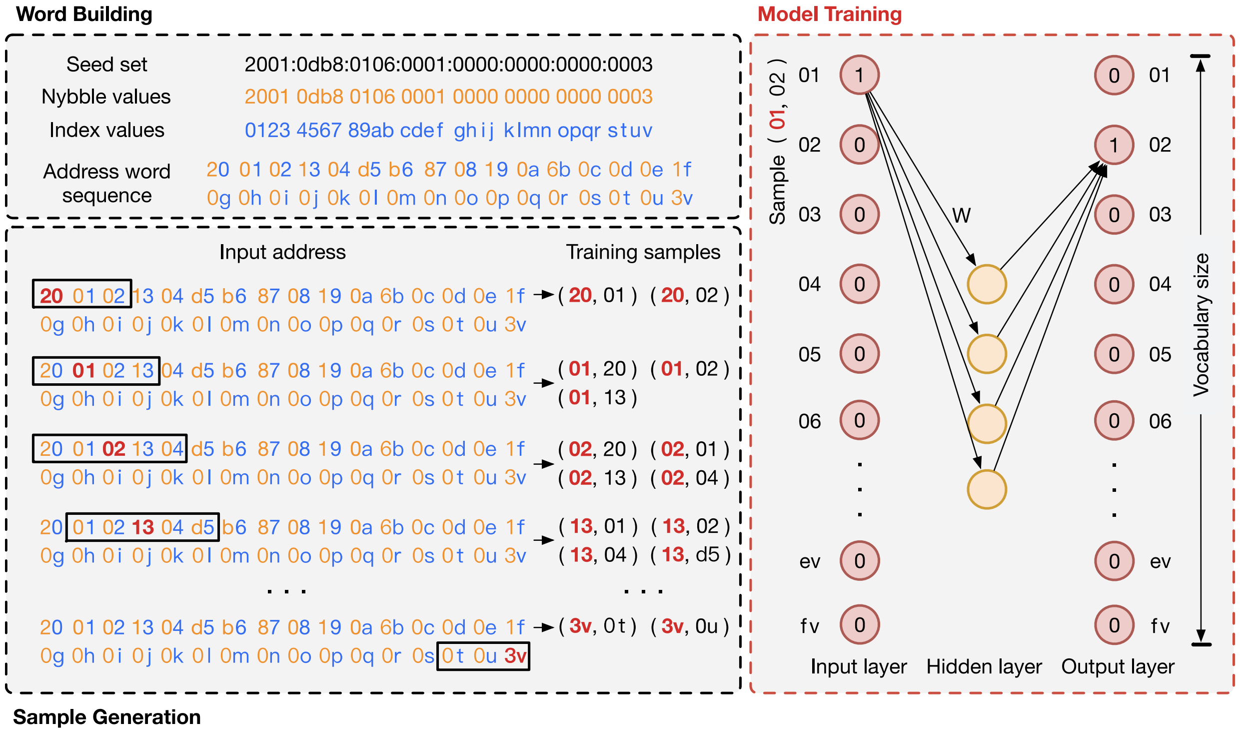
- The new address word - $V_i S_i$

Sample Generation

- The training samples are generated from the corresponding combinations of input words and context words.

Model Training

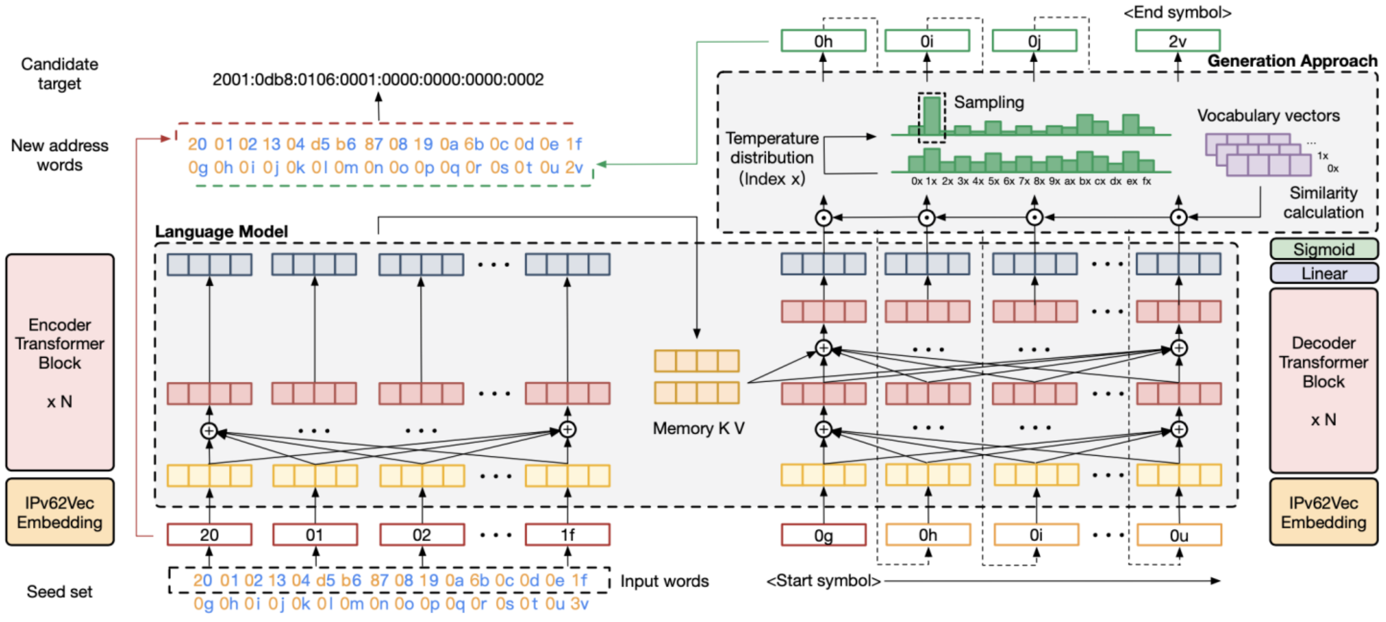
- The neural network is fed with the input word and tries to predict the probability of the context word.
- The final hidden layer result is the vector representation of the input word.



6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

Transformer–IPv6



Language Modeling

- The first 16 words of the sequence are inputted in the Transformer encoder to predict the last 16 words.
- The last 16 words use the mask method to select the current input of the Transformer decoder.

Architecture

- Sigmoid activation function
- Transformer block layers n = 6
- Attention head number = 10

Why Transformer

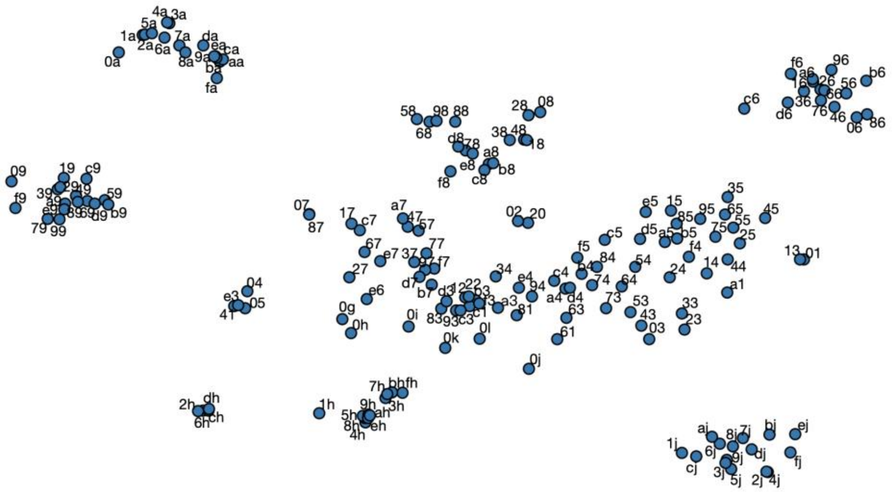
- Attention mechanism helps address consider critical parts of the sequence.
- Multi-head attention mechanism helps generative tasks by observing more address word combinations.

6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

IPv6 Vector Space

Word Vector Space



- Most address words are clustered according to their index attributes
- Index values 0-7 are close, which indicates that the network prefixes often have similar structures.

Address Vector Space

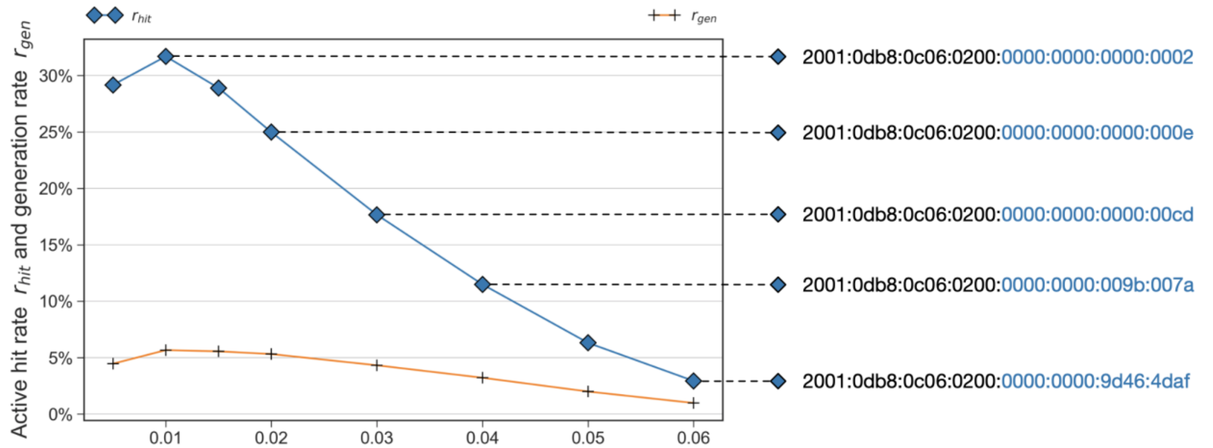
- IPv62Vec successfully divided addresses into several classes.
- Addresses under the same class perform a high similarity.

6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

ECML-PKDD 2020

Temperature

Sampling Strategy



Example of address with different temperature

$$Pr(i) = \frac{e^{\log P(i)^{1/t}}}{\sum_{j=1}^{C} e^{\log P(j)^{1/t}}}, \quad i = 1, ..., C$$

- Low temperature t - sample greedily and the generated address is more close to the seed set.
- High temperature t - sample randomly and the generated address contains more creative sequences.
- The model keeps the highest active hit rate r_{hit} and active generation rate r_{gen} when t = 0.01.

6VecLM: Language Modeling in Vector Space for IPv6 Target Generation

Evaluation Results

Baselines

- Prior paradigms of language models - RNN, LSTM, GCNN.
- Target generation algorithms - Entropy/IP, 6Gen.
- Adding IPv62Vec and generation approach - RNN, LSTM, GCNN.

Experimental Results

Category	Model	$N_{candidate}$	N_{hit}	N_{gen}	r_{hit}	r_{gen}
Conventional language model	RNN [17]	34,604	995	851	2.88%	2.46%
	LSTM [14]	34,636	727	564	2.10%	1.63%
	GCNN [6]	34,817	787	649	2.26%	1.86%
Target generation algorithm	Entropy/IP [11]	69,167	8,321	2,540	12.03%	3.67%
	6Gen [19]	67,712	4,612	1,638	6.81%	2.42%
Adding IPv62Vec and generation approach	RNN [17]	44,242	12,133	2,409	27.42%	5.44%
	LSTM [14]	61,950	10,640	2,019	17.18%	3.26%
	GCNN [6]	52,046	11,360	2,146	21.83%	4.12%
Our approach	6VecLM	46,461	**15,406**	**2,883**	**33.16%**	**6.21%**

- By adding generation approach and IPv62Vec mechanism, language models can reach a not bad performance.
- 6VecLM outperforms all the baselines in the experiment.

THANK YOU FOR ATTENTION

Gaopeng Gou
gougaopeng@iie.ac.cn