计算机科学
COMPUTER SCIENCE

基于深度学习的活跃IPv6地址预测算法

李育强, 李林峰, 朱浩, 侯孟书

引用本文

李育强, 李林峰, 朱浩, 侯孟书.基于深度学习的活跃IPv6地址预测算法[J]. 计算机科学, 2023, 50(7): 261-269.

LI Yuqiang, LI Linfeng, ZHU Hao, HOU Mengshu. Deep Learning-based Algorithm for Active IPv6 Address Prediction [J]. Computer Science, 2023, 50(7): 261-269.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700076

基于深度学习的活跃 IPv6 地址预测算法

李育强[1]　李林峰[2]　朱　浩[1]　侯孟书[1]
1 电子科技大学信息中心　成都 611731
2 电子科技大学计算机科学与工程学院　成都 611731

摘　要　由于 IPv6 拥有庞大的地址空间,基于现有网络速度和硬件计算能力,难以实现全球 IPv6 地址扫描。通过地址生成算法来预测网络中可能出现的 IPv6 地址,随后将预测地址作为扫描的目标,可以达到 IPv6 地址快速扫描的目的。文中通过分析 IPv6 地址结构和分配方式来探索潜在的分配模式,结合已有的传统语言模型和目标生成算法,提出了一种基于深度学习的算法 6LMNS,来预测潜在的活跃 IPV6 地址。6LMNS 首先通过地址向量空间映射模型 Add2vec 来构建具有一定语义关系的 IPv6 地址词向量空间;随后基于 Transformer 构建语言训练模型 GPT-IPv6,以此来估计 IPv6 地址词向量序列的概率分布;最后引入核心采样替代传统贪心搜索解码,完成活跃地址的生成。经验证,与其他语言模型和目标生成算法相比,6LMNS 生成的地址拥有更好的多样性以及更高的活跃率。

关键词:深度学习;Word2Vec;GPT;核心采样;贪心搜索
中图法分类号　TP393

Deep Learning-based Algorithm for Active IPv6 Address Prediction

LI Yuqiang[1],LI Linfeng[2],ZHU Hao[1] and HOU Mengshu[1]
1 Information Center,University of Electronic Science and Technology of China,Chengdu 611731,China
2 School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract　The huge address space of IPv6 makes it difficult to achieve a global IPv6 address scan based on the existing network speed and hardware computing power. Fast IPv6 address scanning can be achieved by using address generation algorithms to predict the possible IPv6 addresses in the network and subsequently using the predicted addresses as the targets of scanning. This paper explores potential allocation patterns by analyzing IPv6 address structures and allocation methods,and proposes a deep learning-based algorithm 6LMNS to predict potentially active IPV6 addresses by combining existing traditional language models and target generation algorithms. 6LMNS first constructs IPv6 address word vector spaces with certain semantic relationships through the address vector space mapping model Add2vec. Subsequently,the language training model GPT-IPv6 is constructed based on Transformers to estimate the probability distribution of IPv6 address word sequences. Finally,nucleus sampling is introduced instead of traditional greedy search decoding to complete the generation of active addresses. It is verified that the addresses generated by 6LMNS have better diversity as well as higher activity rate compared with other language models and target generation algorithms.

Keywords　Deep Learning,Word2Vec,GPT,Nucleus sampling,Greedy search

1　引言

随着互联网的飞速发展,大规模 IPv6 网络部署不断涌现。IPv6 是支撑下一代工业互联网、物联网发展的关键要素和重要基础,是我国信息基础设施升级的必然要求。IPv6 庞大的地址空间,为用户行为追踪溯源、网络精细管控提供了可能。

网络扫描技术作为一种重要的网络测量方法,根据 Bou-Harb 等[1]的综述,可以分为被动测量与主动扫描[2-3]两种方式。被动扫描技术通过在互联网中建立观察点,分析经过流量的信息或者从日志中提取所需数据,能够发现实时存在的活跃 IPV6 地址,隐藏性较好,但所需的硬件成本和环境条件要求较高。主动扫描指发送特定的探测报文到指定 IP 地址后监测其回复情况,从而识别设备并收集其属性信息。全网主动扫描搜集的数据可以用于:1)网络安全方面,Kührer 等[4]通过调查网络上 NTP 放大攻击的分布态势,采取相应措施来减少具有相应漏洞的服务器数量;2)拓扑发现方面,Beverly等[5]设计了拓扑发现工具 Yarrp,以搜集到的活跃 IP

到稿日期:2022-07-07　返修日期:2022-11-08
基金项目:四川省科技计划重点研发项目(2022YFG0329)
This work was supported by the Key Technologies Research and Development Program of Sichuan Science and Technology Plan(2022YFG0329).
通信作者:李育强(yqli@uestc.edu.cn)

地址为基础，快速实现互联网拓扑发现；3）IP 地址分析方面，Plonka 等[6]对活跃 IPv6 地址从时间与空间两个角度进行分析，设计出了一种新的匿名化方法 KIP；4）在主机属性分析方面，Sarabi 等[7]应用变分自编码器（Variational Autoencoder，VAE）将收集的设备属性信息转换为低维向量，为互联网测量提供新的数据基础。

随着网络和硬件的快速发展，以及 Zmap[8]和 Masscan[9]等扫描工具的出现，对全网 IPv4 地址的主动扫描已经实现。通常情况下，扫描全球 IPV4 网络所需的时长仅为若干分钟，然而 IPv6 拥有巨大的地址空间，活跃密度较低，通过 ZMapv6 等工具扫描整个 IPV6 地址空间需要数亿年，因此对 IPV6 网络地址进行遍历式扫描在技术层面不可行。如何设计出高效的主动扫描方式来搜寻活跃 IPv6 地址是研究人员面临的一个挑战。

目前关于 IPv6 地址主动扫描有多种方法。例如，Gasser 等[10]通过对所有搜集到的活跃地址集合执行 traceroute 查询，来获取更多的属于路由器的 IPv6 地址；基于域名系统（Domain NameSystem，DNS）解析机制获取属于服务器的 IPv6 地址，利用 DNS 中的逆向域名解析机制，Strowes[11]提出了一种通过 IPv4 地址进行反查然后进行 AAAA 查询得到 IPv6 地址的方法；Fiebig 等[12]利用 ip6.arpa 域中的拒绝存在语义来减少寻找服务器 IPv6 地址的搜索范围。上述方法都不是直接的主动扫描方式，因此近年来，目标生成算法（Target Generation Algorithm，TGA）被提出以实现 IPv6 全网主动扫描。目标生成算法的原理是根据已知的活跃 IPv6 地址集，挖掘已知活跃 IPv6 地址的地址结构和分配管理等特征，分析活跃 IPv6 地址的潜在分布规律，推断其聚类区域。随后设计地址生成算法，生成对应的扫描目标地址集来缩小扫描范围，进而生成活跃度尽可能高的 IPv6 地址集，来达到 IPv6 地址主动扫描的目的。

由于 IPv6 完全由字符组成，缺少明确的语义信息与序列关系，因此难以预测活跃 IPV6 地址。虽然已经设计出各种复杂的算法，但 IPv6 网络的以下性质导致这些算法仍然面临挑战。

（1）IPv6 寻址模式

网络管理员可以自由选择 IPv6 地址分配方案，实现地址中接口标识符（IID）多种分配模式。客户端可以使用无状态地址自动配置，从而产生伪随机或 EUI-64 IID，而服务器和路由器分配地址通常是根据管理员的习惯或采用 DHCPv6 的方式进行。根据 RFC 7136 中的要求，这些模式是不透明的，因此导致算法推断困难。

（2）IPv6 别名

已有的经验表明，大规模的别名地址是未来 IPv6 扫描中必须解决的问题，其规模能够达到2^{32}以上，因为这些地址无条件地响应查询，不受设备唯一性约束。已有的算法仍需学习别名地址，导致消耗大量算力却生成低质量的地址。

基于以上问题，研究者提出利用语言模型来实现对潜在的活跃 IPV6 地址的预测。基于深度学习的方法，首先通过词向量空间映射，构建具有一定语义关系的 IPv6 向量空间；随后基于 Transformer 构建语言模型来估计词序列的概率分布，推断活跃地址的组成。

在深度生成模型中，生成解码方式是重要环节，对生成性能有显著的影响，而现有方法都使用贪心解码策略。本文基于深度学习，实现并探索了 beam search，top-k，top-p 等神经语言模型常用的生成解码方式在 IPv6 地址预测时的性能，实验结果表明 top-p 采样方法在 IPv6 地址扫描中性能最好。

本文第 2 节介绍了 IPv6 地址生成的背景，以及与目标生成算法、语言模型、地址集相关的研究现状；第 3 节介绍了活跃地址生成算法 6LMNS，包括 Add2Vec 和 GPT-IPv6 组件的整体设计，以及核心采样[13]（Nucleus）的实现细节；第 4 节给出了测试与对比结果；最后总结全文并展望未来。

2 相关工作

已有的关于利用目标生成算法来实现 IPv6 全网主动扫描的工作分为三大类：分析 IPV6 地址结构、分析地址间的相似性、设计地址生成算法。除此之外，本节还将介绍探索语义关系的相关方法、文本生成解码策略，以及地址集的获取途径。

2.1 IPv6 地址结构

IPv6 协议标准提出至今，IPv6 的地址架构及其格式还在不断演进，新的分配方式也在不断涌现。IPv6 地址按照其用途可以分为单播、组播和任播 3 种，单播地址标识唯一接口，组播地址对应标识一组接口，任播地址可以指定给多个接口，并且指定的接口通常属于不同的节点。

针对单播地址，RFC4291 定义了基本的地址架构，IPv6 地址由 128 位二进制数字组成，包括全局网络标识符、子网 ID 和一个接口标识 IID。如图 1 所示，地址分为 8 组，每组包含 4 个十六进制数字，用冒号隔开。每个十六进制数字被称为一个 nybble，IPv6 地址通常使用::来代替连续的零值组，并省略每组中的第一个零值。

	可读地址格式	常用地址格式
Fixed IID	2001:0fc8:0204:0001:0000:0000:0000:0001	2001:fc8:204:1::1
Low 64-bit Subnet	2001:0fc8:0200:0024:0000:0000:000c:0007	2001:fc8:200:24::c:7
SLAAC EUI64	2001:0fc8:0000:3256:c87a:3eff:fe6f:0c6d	2001:fc8:0:3256:c87a:3eff:fe6f:c6d
SLAAC Privacy	2001:0fc8:cab1:00ca:7c56:3254:4578:36a1	2001:fc8:cab1:ca:7c56:3254:4578:36a1

图 1 IPv6 地址样本

Fig. 1 Sample of IPv6 addresses

然而，IPv6 地址并不是简单地由无意义的数字组成。IPv6 有许多不同的寻址方案，接口标识符语义不透明，管理员可以选择使用各种标准来自定义地址类型。此外，一些 IPv6 具有 SLAAC 地址格式，64 位的 IID 通常根据 EUI-64 标准嵌入 MAC 地址，或者完全使用伪随机。参考图 1 中的样本地址，按照复杂程度递增，这些地址是：1）一个具有固定 IID 值的地址（::1）；2）一个在低 64 位有结构化数值的地址（一个以：c 区分的子网）；3）一个具有 EUI-64 的 SLAAC 地址，基于 Ethernet-MAC 的 ID（ff：fe 标志）；4）一个带有伪随机的 SLAAC 隐私地址。

2.2 地址相似性

RFC 7707 文档对 IPv6 地址空间进行了初步的探索，它记录了已知的地址分配方案，以及可能的管理员配置习惯，文档表明大多数地址都遵循特定的模式。Gasser 等[10]采用熵聚类方法，将地址寻址模式分为 6 种，分类结果与地址分配方案有很强的关联性。Coull 等[14]对网络数据中的常见数据类型构建了语义信息，将其与忽略语义信息的指标进行比较，来衡量网络地址分配方式的相似性。Ring 等[15]提出了无监督学习方法 IP2Vec，将流量的元信息作为地址的上下文，用于训练 Word2Vec[16]模型，验证了在数据集中对 IP 地址进行聚类的有效性。Plonka 等[17]首先研究了 IPv6 活跃地址在时间和空间上的潜在关系，使用多分辨率聚合图来量化地址的每一部分与分组地址的关联性。这些研究结果表明，研究人员已经发现了隐藏在活跃 IPv6 地址中的某种分配模式，这为 IPv6 地址生成的可行性奠定了基础。

2.3 目标生成算法

目前已有方法主要从信息论和概率论的思路出发进行设计，主要的生成技术如下。

Ullrich 等[18]提出了一种基于模式的 IPv6 地址生成算法，其基本思想是先对输入地址集进行概率统计，生成高位固定、低位可变的地址模式，再以该模式对低位的取值及组合进行遍历，生成 IPv6 目标地址。

Foremski 等[19]基于信息熵分析、聚类分析与贝叶斯网络建模方法提出了 Entropy/IP 算法，Zuo 等[20]提出了子网扫描目标生成算法。该类算法先根据熵值对地址进行分段，再通过 DBSCAN 聚类算法对每段的取值进行聚类，随后在不同段运行贝叶斯网络建模，提取概率统计关联性强的分段组合模式，最后根据关联性强的分段组合模式生成扫描目标地址集。

Murdock 等基于凝聚层次聚类和贪心的思想提出了 6Gen[21]，首先用 Hamming 距离来度量 IPv6 地址字符串之间的距离，通过迭代算法将 Hamming 距离近的地址依次聚集到相同的地址簇中，再在地址簇对应的密集区域内生成扫描目标地址。6Gen 发现的活跃地址数量能够达到 Entropy/IP 的数倍，但是对数据训练时的时间复杂度要求较高。

Liu 等[22]提出了 6Tree 算法，该算法先将输入的种子地址集构建成对应的树形结构，随后对所有叶节点内的地址进行扫描探测，计算各叶节点内活跃地址数量的比值，根据比值对节点进行排序，然后依次对叶节点进行扩展并更新。重复上述过程，直至扫描的地址数量达到预期值。动态扫描过程中扫描探测的所有地址构成了扫描目标地址集。

Song 等[23]在 6Tree 算法的结构基础上对分支方法进行优化，提出了 DET 算法。DET 算法在构建空间树并寻找高密度区域时在熵值最小处对节点进行分支，从而保证高密度区域聚集在同一个地址空间，然后对不同密度区域进行排序。根据排序结果，在高密度区域内依次生成扫描目标地址。

包括 Entropy/IP 和 6Gen 在内的地址生成算法都是基于人为观察和对网络数据的假设，如果过度人为干预，可能会导致算法过度依赖已有的经验，无法适应新的地址集。Cui 等[24]于 2020 年提出了 6VecLM 算法，通过将地址映射到一个向量空间来构建语义关系，并使用 Transformer 网络来对模型进行训练，从而生成目标地址。随后，他们在 2021 年又提出基于生成对抗网络（GAN）和强化学习来生成多模式目标的新型架构 6GAN[25]。6GAN 中多个生成器使用多类判别器和别名检测器进行训练，生成具有不同寻址模式类型的非别名目标地址。

2.4 词嵌入和语言模型

词嵌入是自然语言处理（Natural Language Processing，NLP）中语言模型与表征学习技术的统称。利用低维、高密度的词嵌入可以加速计算，密集的向量表征能够极大地提升模型的泛化能力，低维的向量可以抽象文本中的具体特征。同时，低维词嵌入可以发现词与词之间存在的关系。

由于 IPv6 地址语义不透明，还存在多种寻址方案，很难有效地进行模型训练。要实现高效的地址生成算法，合理挖掘地址组合的语义信息尤为关键，因此，相关研究通过构建地址词序列来生成 IPv6 语义。通过学习地址词序列模型的上下文，生成具有语义辨识度的地址向量，随后通过语言建模推测活跃地址。

为了探索词之间的语义关系，Mikolov 等提出了 Word2Vec 模型，通过训练该模型可以得到每个单词的固定长度的向量表示，向量之间的距离可以用来衡量单词之间的语义相似性。然而，即使是十六进制的 IPv6 地址，也拥有 32 个 nybbles，因此在高维空间中，构建一个高质量的地址向量极其困难。于是研究者通过模型学习来实现地址空间向量表示，并利用降维技术来获取活跃地址的聚类区域。随着词嵌入的发展，Bengio 等首次采用神经网络学习序列的联合概率函数，来代替统计语言建模，并取得了很大成功。

最近，Dai 等提出了一个完全基于 Attention 机制[26-27]的网络工作架构 Transformer[28]，该架构抛弃了传统的 CNN[29]和 RNN[30]，提高了模型的并行能力，解决了 NLP 中棘手的长期依赖问题。在此基础上，Radford 等[31]又提出了更加适用于文本生成任务的模型 GPT（Generative Pre-Training Model）。GPT 是 OpenAI 提出的第一个基于 Transformer 解码器的生成式预训练模型，首次在大规模文本语料上进行预训练，它采用单向语言模型的训练目标来进行参数的优化。

2.5 文本生成解码

利用自回归神经语言模型进行生成的过程分为两步：

1)利用语言模型生成条件概率分布;2)利用解码算法从该概率分布中选择一个字符。

贪心搜索(greedy search)最简单,直接选择每个输出的最大概率,直到出现终结符或最大句子长度。Freitag 等[32]提出的集束搜索(beam search)是一种启发式搜索算法,该算法每次都保留当前最大的 beam_num 个结果,从结果中选择概率积最大的序列,可以避免贪心搜索遗漏掉后面大概率的序列,但结果容易缺乏多样性。

在每个时间步中,top-k 采样根据其相对概率,从前 k 个可能的概率分布内的字符中进行采样。top-k 采样算法的实现直白简单,虽然 top-k 抽样比在完整的概率分布中进行完全随机采样生成的文本质量要高得多,但在不同的语境下需要使用不同大小的 K,动态地选择合适的 K 是一个难题。

为了解决这个问题,Holtzman 等[13]提出了核心采样(Nucleus Sampling),根据概率分布动态地决定采样的词空间,其采样集的大小根据每个时间步生成的概率分布动态调整。分别选取生成的 IPV6 地址,测试它们的每个地址词的条件概率。某条样本的平均采样测试结果如图 2 所示,可以发现,top-p 生成的地址每一个地址词的概率基本都大于传统 greedy search,生成的地址词的采样效果总优于传统 greedy search,另一方面也兼顾了多样性。

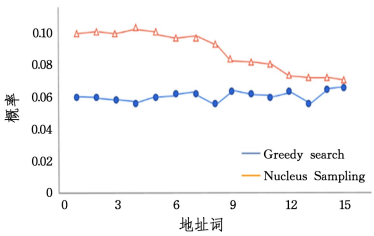


图 2 greedy search 与 Nucleus Sampling 生成地址词概率对比图

Fig. 2 Comparison of the probability of greedy search and Nucleus Sampling to generate address words

温度采样已被广泛应用于文本生成,给定求概率分布 Softmax 函数前的逻辑值 u_l 和和温度参数 t,其中 V 代表词表中所有词。在温度参数 t 的控制下,Softmax 函数会将字符 x_i 的概率估计为:

$$p(x=V_l \mid x_{1:i-1})=\frac{\exp(u_l/t)}{\sum_{l'}\exp(u_{l'}/t)}$$

其中,u_l 代表概率分布中当前字符的未归一的概率值输出,l' 代表词表大小,$p(x_i \mid x_{1:i-1})$ 代表修正后的当前字符概率值。低温度参数 t 的值可以提高生成质量,但它同样会减少多样性;相反,高温度参数 t 的值会降低生成质量,增加多样性。

2.6 地址集

类似于 IPv4,IPv6 存活地址列表(Hitlist)指一组能够大体上覆盖和代表 IPv6 网络的地址集合,具有存活性、完整性、稳定性的特点。获取 IPv6 存活地址集的方法包括:

(1)Rapid7_FDNS:FDNS 数据是 Rapid7 的 Project Sonar 公开数据的一部分,通过提取其中的 AAAA 记录,最终得到 IPv6 地址。

(2)CAIDA_Ark:IPv6 拓扑数据集作为 CAIDA Ark 平台测量数据的一部分,通过使用 Paris Traceroute 技术探测所有声明的/48 或者更短的 IPv6 前缀中的随机地址。

(3)Bitnodes:Bitnodes 通过发现网络中的所有可达节点来评估比特币网络的大小,使用 Bitnodes 提供的 API,从节点名称中提取 IPv6 地址。

(4)TUM_Responsive:德国 TUM 大学通过对 IPv6 存活地址列表的研究,提供 IPv6 Hitlist 服务。

(5)TUM_Seeds:德国 TUM 大学的 Hitlist 服务作为一个收集项目,本身也采用了不同的收集源。根据其仓库中的实际数据(包括 Alexa, CAIDA _ dnsname, CT (Certifcate Transparency), Zonefiles, Openipmap 和 Traceroute),其中前 4 个根据提取的域名查询 AAAA 记录,从而获取 IPv6 地址。Openipmap 为从 RIPE ipmap 项目中提取的 IPv6 地址,Traceroute 为对所有其他源中的地址 Traceroute 再提取所有的路由器 IPv6 地址。

2.7 设计思路

本文设计了基于深度学习的活跃地址生成算法 6LMNS,包括 Add2vec 和 GPT -IPv6 两种机制。Add2vec 将整个活动地址空间映射到语义向量空间,序列相似的地址在同一簇中。GPT-IPv6 将学习语义向量来实现 IPv6 语言建模,综合考虑多个地址序列之间的关系,通过核心采样 top-p 解码生成与地址集具有语义相似性的序列。

本文的贡献可以总结如下:

(1)探索了 IPv6 语义向量空间的构建,Add2Vec 可以有效地将活跃 IPv6 地址空间进行聚类。

(2)设计了向量空间的语言建模算法 GPT-IPv6,实现与探索了 beam search,top-k,top-p 等神经语言模型常用的生成解码方式在 IPv6 地址预测时的性能,实验结果表明 top-p 采样方法在 IPv6 地址扫描中性能最好。

(3)实验表明 6LMNS 在多个指标上优于传统语言模型和目标生成算法。

3 活跃地址生成算法

3.1 Add2Vec

本节将介绍 6LMNS 的第一个组件 Add2Vec,用于实现 IPv6 地址向量空间映射,包括地址词构建、样本生成和模型训练。

3.1.1 词构建

Add2Vec 是基于 Word2vec 思想实现的,Word2vec 是 Word Embedding 的方法之一,它是由谷歌 Mikolov 于 2013 年提出的一套新的词嵌入方法。构建有效的语义信息,首先需要先赋予 IPV6 地址新的语义。如图 3 所示,首先创建地址词来表示十六进制地址的每一个 nybble。地址中第 i 个 nybble 的值为 V_i,其中 $V\in\{0,1,\cdots,f\}$。索引 i 被定义为 S_i,其中 $S\in\{0,1,\cdots,v\}$。新表示法中第 i 个地址词由 nybble 值和索引值连接组成,为 V_iS_i(例如,第 11 个 nybble 值 4 被表示为地址词 $4a$)。词汇表即为所有通过 IPV6 地址集建立的地址词序列,目的是区分不同索引的 nybble 值,相同的 nybble 值在地址中的位置不同可能具有不同的语义性。

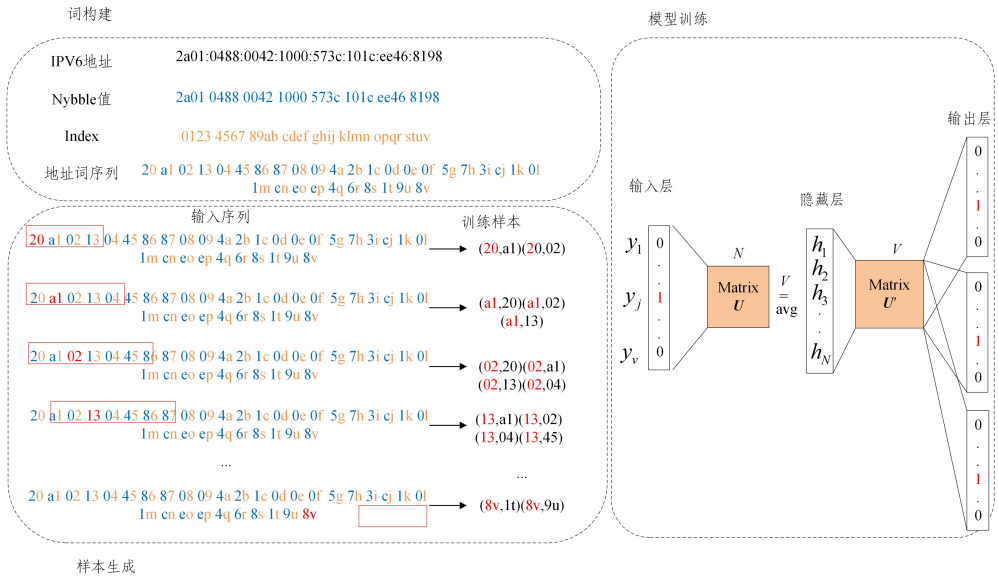


图 3　Add2Vec 整体结构

Fig. 3　Overall structure of Add2Vec

3.1.2　样本生成

在确定地址词后，按照 Mikolov 等的选词过程，选择输入词和它的上下文来生成训练样本，即输入 IPv6 的一个地址词，去预测它的上下文。如图 3 所示，对输入序列进行选词操作，当序列中的某个词被选为输入词后，输入词上下文的词被选为建立训练样本的背景词，窗口大小为 5。

3.1.3　Add2Vec 训练

如图 3 所示，Add2Vec 的整体结构由一个只有一个中间层的多层神经网络构成。由于不能将词直接输入神经网络，因此将每个地址词都表示为 one-hot 向量，记作 $\boldsymbol{u}$，即用一个词索引值位置为 1、其他位置都为 0 的向量来唯一表示这个词语，此向量的长度等于词汇表的大小，神经网络输入和输出神经元的数量等于词汇量。采用 Skip-gram 模型，将词输入神经网络，用窗口内其他词预测输入词。输出层使用一个 Softmax 分类器，它的每个结点将会输出一个 0～1 之间的值来表示一个特定的词出现在这一上下文窗口中的概率，所有这些输出层神经元结点的概率之和为 1。我们将 Add2Vec 模型的隐藏层输出作为输入地址词的词向量表示，在具体实现时，将隐藏层的维度设置为 100。在 Add2Vec 模型训练中，使用交叉熵损失函数与随机梯度上升法进行迭代优化。如图 2 所示，待 Add2Vec 模型收敛后将矩阵 $\boldsymbol{U}'$ 输出，作为实现 one-hot 编码转化为词向量的嵌入矩阵。

3.2　GPT-IPv6

通过第二个组件，即语言训练模型 GPT-IPv6，并运用地址向量来实现 IPv6 地址的生成。GPT 的整体结构如图 4 所示。

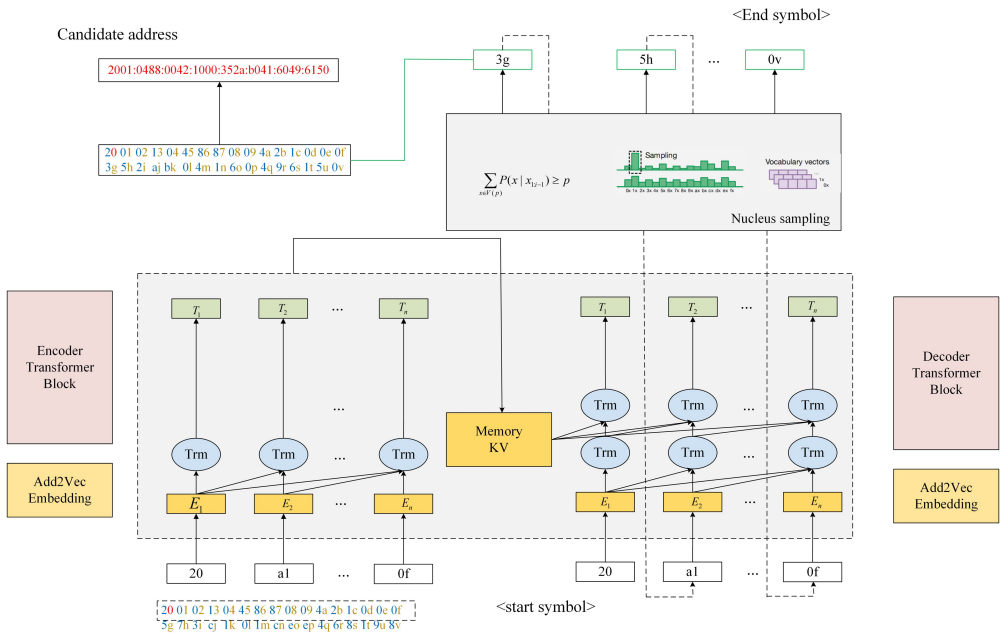


图 4　GPT-IPv6 的整体结构

Fig. 4　Overall structure of GPT-IPv6

GPT 使用 Transformer 作为特征抽取器，其特征抽取能力强于 RNN；同时在 Transformer 的基础上进行了改进，只保留了 Mask Multi-Head Attention。GPT 在进行自注意力计算时，屏蔽了来自当前计算位置右边所有单词的信息，只采用单词的上下文来进行预测。

3.2.1 语言模型

GPT 语言模型将联合概率按如下分解来计算序列 $u_{0:L}$ 的概率分布：

$$P(u_{0:L})=P(u_0)\prod_{i=1}^{L}P(u_i|u_{0:i-1})$$

其中，L 是序列长度。

对条件概率 $P(u_i|u_{0:i-1})$ 建模，训练一个 GPT 网络来处理地址词序列 $u_{0:i-1}$。GPT-IPv6 的结构如图 4 所示，输入地址词由地址集中的地址转换而来，序列前 16 个地址词的向量由预先训练的 Add2Vec 模型输出的 $\boldsymbol{U}'$ 决定，然后将其输入模型中，用于依次预测后 16 个地址词向量。

给定地址词向量编码序列 $\boldsymbol{U}=\{u_1,\cdots,u_n\}$（而非地址词 one-hot 编码序列），使用标准语言模型目标（Language Modeling Objective）来输出预测词向量序列：

$$L(U)=\sum_i L(u_i|u_{i-k},\cdots,u_{i-1};\Theta),\forall i\in[1,n]$$

$$h_0=\boldsymbol{U}$$

$$h_l=transformer\_block(h_{l-1})\ \forall l\in[1,l_{\max}]$$

$$\boldsymbol{u}_{\text{pred}}=Linear(h_n)$$

其中，$\boldsymbol{U}=(\boldsymbol{u}_{-k},\cdots,\boldsymbol{u}_{t-1})$ 是地址词向量，n 是网络层数，$\boldsymbol{U}'$ 是地址词嵌入矩阵（Embedding Matrix），Θ 是 GPT-IPv6 的模型参数。

为了完成向量空间中的地址生成任务，希望生成的词向量 $\boldsymbol{u}_{\text{pred}}$ 与目标词向量 $\boldsymbol{u}_{\text{true}}$ 具有较高的语义相似度。因此，6LMNS 模型使用余弦距离作为损失函数 L。

$$\cos(\theta)=\frac{\sum_{i=1}^{n}u_{\text{true}}^{(i)}\cdot u_{\text{pred}}^{(i)}}{\sqrt{\sum_{i=1}^{n}(u_{\text{true}}^{(i)})^2}\cdot\sqrt{\sum_{i=1}^{n}(u_{\text{pred}}^{(i)})^2}}$$

$$L=1-\cos(\theta)$$

传统语言模型会直接预测一个词的 one-hot 编码，与之不同的是，6LMNS 是预测词向量以保留向量空间的语义信息。由于训练样本是具有语义关系的地址向量，通过 Add2Vec 模型获得，因此使用最小化余弦距离函数，可使获得的预测目标与地址集具有相似的上下文结构。这种方法旨在选择向量空间中最接近的地址词，有助于发现活跃 IPV6 地址集。

在每个 epoch 生成地址词向量后，计算预测得到的词向量，以及词汇表中包含当前索引的每个词向量的余弦相似度。使用 Softmax 函数将余弦相似度 $\cos(\theta)$ 转换为单词采样概率 $p(i)$。

$$P(i)=\frac{e^{\cos(\theta)i}}{\sum_{j=1}^{C}e^{\cos(\theta)j}},i=1,\cdots,C$$

其中，C 是词汇表中具有当前索引的单词数。

3.2.2 文本生成解码

为了获得活跃率较高的地址，采用核心采样（top-p）这一随机解码策略，核心采样是对贪心解码的一种改进。top-p 采样的核心思想是根据概率分布动态地决定采样的词空间，给定 i 时刻的概率分布：$P(u|u_{1:i-1})$，它是 i 时刻每个词的选取概率，即经过 Softmax 计算得到的归一化概率，将归一化概率进行排序后，取前面若干个概率最大的词语直到累计概率大于固定阈值，再重新归一化，作为 top-p 的采样空间 $V^{(p)}$。

$$\sum_{u\in V(p)}P_{\max}(u|u_{1:i-1})\geqslant P$$

$$p'=\sum_{u\in V(p)}P_{\max}(u|u_{1:i-1})$$

$$P'(u|u_{1:i-1})=\begin{cases}P(u|u_{1:i-1})/p', & \text{if } x\in V^{(p)}\\ 0, & \text{otherwise}\end{cases}$$

核心采样策略可以兼顾活跃度与多样性，它们的累积概率密度大于预设的阈值 p。

算法 1 文本生成算法

输入：生成模型、上下文信息、生成长度

输入：生成地址

```
1. memory = encode_memory(model, context);//从 context 中编码 memory
2.    i=0;
3.    result=context
4.    while i ≤ L do
5.    logit=model(result,memory) //生成模型计算
6.    prob_distribution=Softmax(word_score) //转换为概率分布
7.    word=Nucles_Sampling(prob_distribution) //解码获得词语
8.    j=0
9.    p=0
10.      while p ≤ P do: //P 是超参数，选择前面若干个概率最大的地址词，直到累计概率超过阈值 P
11.        p+=sorted_prob_distribution[j]
12.        j+=1
13.        word=Sampling(sorted_prob_distribution[0:j])
14.   end while
15.   result=concate(result,word) //拼接到末尾
16. i=i+1
17. end while
18. return result
```

文本生成算法如算法 1 所示。解码算法接收生成模型对象、上下文信息以及生成长度作为输入，其中上下文信息包括待生成文本的前文和控制信息等。算法首先将上下文信息中的控制信息部分输入模型对象，以获得全局化记忆和本地化编码向量，产生用于记录词语重复情况的向量。完成初始化工作后算法进入循环的解码过程，该过程中首先将上下文和记忆信息输入模型以计算输出，使用 Softmax 函数将其转化为概率。得到概率分布后即可输入解码算法 Nucleus Sampling 产生特定 Nybble，并将 Nybble 词拼接在地址末尾，随后更新生成地址的长度信息。直至生成的地址达到指定长度后，完成单次循环，将生成结果返回。

4 实验评估

本节将介绍实验中使用的数据集和详细测试过程，展示 6LMNS 模型在生成地址集上的高效性。

4.1 数据集

本文实验中所使用的数据集，均来自开源地址集 IPv6

Hitlist。Gasser 等通过扫描 IPv6 公共列表，来获取并发布最新的每日活跃的 IPV6 地址。使用 IPv6 Hitlist 可以尽可能地使生成的地址拥有更高的活跃率。

4.2 评价方法

为了评估生成地址的活跃率，实验中使用 Zmapv6 工具对生成的地址执行 ICMPv6，TCP/80，TCP/443，UDP/53，UDP/443 扫描。当任何扫描方法发送的查询得到响应时，就确定该地址为活跃的。由于不同主机之间的活跃时间存在差异，为了确保结果的准确性，将对主机进行多次扫描，最后汇总扫描结果。

由于 IPv6 地址生成与普通文本生成任务不同，需要为 6LMNS 模型定义一个新的性能评价指标。在给定地址集的情况下，$N_{\text{candidate}}$ 代表生成地址的数量，N_{hit} 代表生成活跃地址的数量，N_{gen} 代表生成的地址是活跃的，且不在原始地址集中的数量。因此该模型生成地址的活跃率 r_{hit} 和有效生成率 r_{gen} 的计算公式为：

$$r_{\text{hit}}=\frac{N_{\text{hit}}}{N_{\text{candidate}}}\times 100\%$$

$$r_{\text{gen}}=\frac{N_{\text{gen}}}{N_{\text{candidate}}}\times 100\%$$

其中，r_{hit} 代表模型从地址集学习的能力，r_{gen} 强调模型生成新的活跃地址的能力。

4.3 采样策略

beam search 是一种启发式搜索算法，它每次都保留当前最大的 beam_num 个结果，从结果中选择概率积最大的序列。当 beam_num 取 1 时算法就退化为贪心解码，beam search 采样虽然比贪心采样性能更好，但生成的地址池序列重复度偏高。真实 IPv6 地址词数据集与目标生成模型不同，并不会总是选择条件概率最大的词。这些概率大的词会发生正反馈，产生循环，导致 beam search 生成的结果有大量重复，因此生成的地址去重后并不能获得高活跃率 r_{hit} 和高有效生成率 r_{gen}。

top-k 解码策略会在采样前将输出的概率分布截断，取出概率最大的 K 个词构成一个集合，然后将这个子集词的概率再归一化，最后从新的概率分布中采样词汇。但 top-k 中 k 的选择是个难题，K 过大会导致算法过多追求生成的 IPv6 地址的多样性而忽略生成的 IPv6 地址的活跃度，K 过小又会失去随机性。

而在 nucleus sampling(top-p) 下，K 的大小由 P 控制，动态地上升和下降，这与模型的置信度区域在词汇上的变化相对应。

为了评估采样策略的性能差异，对 greedy search，beam search 和 top-k，top-p 这 4 种采样策略做了对比测试。基于开源地址集 IPv6 Hitlist，利用深度学习算法 6LMNS 进行训练，最后分别利用 greedy search，beam search 和 top-k(K=8)，top-p(P=0.8) 4 种采样策略来生成地址。随后利用 ZmapV6 来统计生成地址的 r_{hit} 和 r_{gen}，结果如表 1 所列。greedy search 和 beam search 的 r_{hit} 和 r_{gen} 均小于 top-p 和 top-k。这也证实了 top-k 和 top-p 采样策略用于生成 IPv6 地址的性能优于传统基于贪心思想的采样策略以及 beam search。

表 1 采样策略性能比较

Table 1 Performance comparison of sampling strategies

(单位：%)

	r_{hit}	r_{gen}
greedy search	25.2	4.2
beam search(N=4)	27.5	4.9
top-k(K=8)	34.3	6.5
top-p(P=0.8)	35.3	6.9

生成地址的重复率也是量化生成地址质量的一个指标。如图 5 所示，本文对 greedy search，beam search，top-k 以及 top-p 生成地址的重复率做了对比测试。贪心解码由于只取概率最大的值，因此生成地址的重复率最高；改变 beam search 中 Beam 的宽度 b，b 越大，每次保留的结果数越多，生成地址的多样性越好；改变 top-k 采样中 K 的值从 2 到 12，根据其相对概率，从前 K 个可能的概率分布内的字符中进行采样，K 越大，生成地址的重复率越低；改变核心采样 P 的大小从 0.1～0.9，P 越大，相同情况下选取的地址词数量就越多，因此重复率更低。

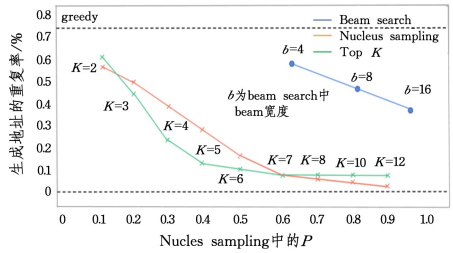


图 5 不同解码策略生成地址重复率对比图

Fig. 5 Comparison of address repetition rates generated by different decoding strategies

通过对比发现，在合理的参数范围内，beam search 生成地址的重复率远高于 top-p 和 top-k。且当 P 足够大时，top-p 采样生成的地址多样性更好，这得益于 top-p 能根据概率分布动态地决定采样的词空间，更好地将生成的地址用于拓扑发现、IP 地址分析和主机属性分析。

4.4 温度

在 6LMNS 模型中，Softmax 温度是一个关键参数，可以控制模型生成的地址质量。在选择高温 t 时，模型更倾向于随机采样，生成的地址拥有更好的多样性。6LMNS 采用改进后的核心采样 top-p，在保持低温 t 的情况下，生成的地址更接近原始地址集。图 6 给出了不同温度 t 对应的地址生成结果，温度的升高促使生成的地址更加多样化。通过测试不同 t 值的生成性能可以发现，当 t=0.01 时，模型生成的地址拥有更高的活跃率 r_{hit} 和有效生成率 r_{gen}。

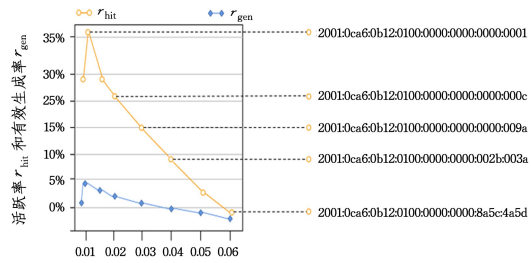


图 6 不同 Softmax 温度 t 的地址预测结果

Fig. 6 Address prediction results for different Softmax temperature t

4.5 评价结果

实验中用于比较的基准包括：

(1)传统语言模型：RNN，LSTM[36]和 GCNN 在之前的语言建模中可以获得不错的效果，通过构建模型来生成地址。

(2)传统生成算法：Entropy/IP 和 6Gen 是传统的目标生成算法，拥有较好的生成性能。Entropy/IP 采用熵信息来构建地址段或空间树，用于探索活跃的地址；而 6Gen 则使用算法分析来发掘活跃集群。本文使用 Entropy/IP 和 6Tree 的开源代码作为测试基准。

(3)深度学习算法：6VecLM 是最近提出的通过基于深度学习方法来实现地址的生成，利用 Transformer 和 Softmax 温度来构建预测地址序列的 IPv6 语言模型。将 6VecLM 开源代码作为测试基准。

(4)生成对抗网络：6GAN 是最新的基于强化学习生成式对抗网络来实现多模式目标生成。6GAN 中多个生成器使用多类判别器和别名检测器进行训练，生成具有不同寻址模式类型的非别名目标地址。

表 2 列出了基于公共数据集 IPv6 Hitlist 的所有地址生成模型的性能对比。可以看出，深度学习方法 6VecLM，6LMNS 以及生成对抗网络方法 6GAN 生成地址的活跃率明显优于常规语言模型 RNN，LSTM 和 GCNN，以及目标生成算法 Entropy/IP 和 6Gen。结果显示 6LMNS 的 r_{hit} 及 r_{gen} 均优于同样采用深度学习方法的 6VecLM，这证实了核心采样文本生成解码优于 greedy search。相比贪心解码，Nucles Sampling 每一个地址词的条件概率更大，生成地址总的活跃率更高。同时，Nucles Sampling 可以增加生成地址的多样性，降低重复率，使模型拥有更强的生成新的活跃地址的能力。

表 2 算法性能比较

Table 2 Algorithm performance comparison

类别	方法	$N_{candidate}$	N_{hit}	N_{gen}	r_{hit}	r_{gen}
常规语言模型	RNN	34604	995	851	2.88%	2.5%
	LSTM	34636	727	564	2.10%	1.6%
	GCNN	34817	787	649	2.26%	1.9%
目标生成算法	Entropy/IP	69167	8321	2540	12%	3.7%
	6Gen	67712	4612	1638	6.8%	2.4%
深度学习方法	6VecLM (greedy search)	46461	15406	2883	33.16%	6.2%
生成对抗网络	6GAN (greedy search)	58624	21107	5217	36%	8.9%
本文方法	6LMNS (top-p)	67840	23933	4694	35.3%	6.9%

6GAN 采用了基于生成对抗网络的方法，虽然生成地址时基于贪心采样解码，但因引入多类判别器和别名检测器进行训练，生成了具有不同寻址模式类型的非别名活动地址，模式判别准确率达到 96.6%，因此生成的地址拥有更高的活跃率。6GAN 从模型训练的角度出发做了优化，而本文的 6LMNS 是在解码策略上进行改进，二者并不冲突。6LMNS 所使用的核心解码策略依旧可以用于基于生成对抗网络方法的 IPv6 地址生成中。

结束语 本文针对活跃 IPV6 地址预测问题，提出了一种基于深度学习的算法 6LMNS。通过 Add2Vec 模型生成的地址向量，有效地提取了地址的基本语义信息；通过 GPT-IPv6 模型学习向量空间中的地址词序列；根据余弦相似度和核心采样解码策略完成地址生成。6LMNS 优于同样基于深度学习的算法 6VecLM，表明核心采样解码算法优于贪心思想的 greedy search 解码算法，可以提升生成地址的多样性和活跃率。同时，在后续工作中，可以引入 6GAN 的思想，加入别名检测机制，生成更高质量的地址，更好地达到 IPv6 地址主动扫描的目的。

参 考 文 献

[1] BOU-HARB E, DEBBABI M, ASSI C. Cyber Scanning: A Comprehensive Survey[J]. IEEE Communications Surveys and Tutorials, 2014, 16(3): 1496-1519.

[2] RYE E C, BEVERLY R. Discovering the IPv6 network periphery[C]// International Conference on Passive and Active Network Measurement. Cham, Springer, 2020: 3-18.

[3] BEVERLY R, DURAIRAJAN R, PLONKA D, et al. In the IP of the beholder, Strategies for active IPv6 topology discovery [C]// Proceedings of the Internet Measurement Conference 2018. 2018: 308-321.

[4] KÜHRER M, HUPPERICH T, ROSSOW C, et al. Exitfrom Hell? Reducing the Impact of Amplification DDoS Attacks [C]// Proceedings of the 23rd USENIX Security Symposium. USA, 2014: 111-125.

[5] BEVERLY R. Yarrp'ing the Internet: Randomized high-speed active topology discovery[C]// Proceedings of the 2016 Internet Measurement Conference. 2016: 413-420.

[6] PLONKA D, BERGER A. KIP, A measured approach to IPv 6 address anonymization[J]. arXiv: 1707.03900, 2017.

[7] SARABI A, LIU M. Characterizing the internet host population using deep learning, A universal and lightweight numerical embedding[C]// Proceedings of the Internet Measurement Conference 2018. 2018: 133-146.

[8] DURUMERIC Z, WUSTROW E, HALDERMAN J A. {ZMap}, Fast Internet-wide Scanning and Its Security Applications[C]// 22nd USENIX Security Symposium (USENIX Security 13). 2013: 605-620.

[9] GRAHAM R D. Masscan, Mass ip port scanner [EB/OL]. https://github.com/robertdavidgraham/masscan.

[10] GASSER O, SCHEITLE Q, GEBHAR D S, et al. Scanning the IPv6 internet, towards a comprehensive hitlist[J]. arXiv: 1607.05179, 2016.

[11] STROWES S D. Bootstrapping active IPv6 measurement with IPv4 and public DNS[J]. arXiv: 1710.08536, 2017.

[12] FIEBIG T, BORGOLTE K, HAO S, et al. In rDNS we trust, revisiting a common data-source's reliability[C]// International Conference on Passive and Active Network Measurement. Cham: Springer, 2018: 131-145.

[13] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[J]. arXiv: 1904.09751, 2019.

[14] COULL S E, MONROSE F, BAILEY M. On Measuring the Similarity of Network Hosts, Pitfalls, New Metrics, and Empirical Analyses[C]// NDSS. 2011.

[15] RING M,DALLMANN A,LANDES D,et al. Ip2vec:Learning similarities between ip addresses[C]// 2017 IEEE International Conference on Data Mining Workshops(ICDMW). IEEE,2017:657-666.

[16] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.

[17] PLONKA D,BERGER A. Temporal and spatial classification of active IPv6 addresses[C]// Proceedings of the 2015 Internet Measurement Conference. 2015:509-522.

[18] ULLRICH J,KIESEBERG P,KROMBHOLZ K,et al. On reconnaissance with IPv6:a pattern-based scanning approach [C]// 2015 10th International Conference on Availability,Reliability and Security. IEEE,2015:186-192.

[19] FOREMSKI P,PLONKA D,BERGER A. Entropy/ip:Uncovering structure in IPv6 addresses[C]// Proceedings of the 2016 Internet Measurement Conference. 2016:167-181.

[20] ZUO Z,MA Y,ZHANG P,et al. Predictional algorithm of active IPv6 address prefix[J]. Iournal on Communications, 2018, 39(S1):1-8.

[21] MURDOCK A,LI F,BRAMSEN P,et al. Target generation forinternet-wide IPv6 scanning[C]// Proceedings of the 2017 Internet Measurement Conference. 2017:242-253.

[22] LIU Z,XIONG Y,LIU X,et al. 6Tree:Efficient dynamic discovery of active addresses in the IPv6 address space[J]. Computer Networks,2019,155:31-46.

[23] SONG G,HE L,WANG Z,et al. Towards the construction of global IPv6 hitlist and efficient probing of IPv6 address space [C]// 2020 IEEE/ACM 28th International Symposium on Quality of Service(IWQoS). IEEE,2020:1-10.

[24] CUI T,XIONG G,GOU G,et al. 6veclm:Language modeling in vector space for IPv6 target generation[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham:Springer,2020:192-207.

[25] CUI T,GOU G,XIONG G,et al. 6GAN:IPv6 multi-pattern target generation via generative adversarial nets with reinforcement learning[C]// IEEE INFOCOM 2021-IEEE Conference on Computer Communications. IEEE,2021:1-10.

[26] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. Advances in Neural Information Processing Systems,2017,1706:03762.

[27] BAHDANAU D,CHO K,BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473,2014.

[28] DAI Z,YANG Z,YANG Y,et al. Transformer-xl:Attentive language models beyond a fixed-length context[J]. arXiv:1901.02860,2019.

[29] DAUPHIN Y N,FAN A,AULI M,et al. Language modeling with gated convolutional networks[C]// International Conference on Machine Learning. PMLR,2017:933-941.

[30] ZAREMBA W,SUTSKEVER I,VINYALS O. Recurrent neural network regularization[J]. arXiv:1409.2329,2014.

[31] RADFORD A,NARASIMHAN K,SALIMANS T,et al. Improving language understanding by generative pre-training[J/OL]. https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

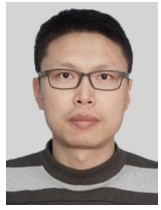
[32] FREITAG M,AL-ONAIZAN Y. Beam search strategies for neural machine translation[J]. arXiv:1702.01806,2017.

[33] VIJAYAKUMAR A,COGSWELL M,SELVARAJU R,et al. Diverse beam search for improved description of complex scenes [C]// Proceedings of the AAAI Conferenceon Artificial Intelligence. 2018.

[34] MEISTER C,VIEIRA T,COTTERELL R. If beam search is the answer,what was the question? [J]. arXiv:2010.02650, 2020.

[35] WELLECK S,KULIKOV I,ROLLER S,et al. Neural text generation with unlikelihood training[J]. arXiv:1908.04319, 2019.

[36] SUNDERMEYER M,SCHLÜTER R,NEY H. LSTM neural networks for language modeling[C]// Thirteenth Annual Conference of the International SpeechCommunication Association. 2012.



LI Yuqiang, born in 1979, master, lecturer. His main research interests include computer network and cyber security.

(责任编辑:何杨)