UNIVERSITY OF CALIFORNIA
Los Angeles

Efficient Latent Semantic Extraction from Cross Domain Data with Declarative Language

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Mingda Li

2020

ABSTRACT OF THE DISSERTATION

Efficient Latent Semantic Extraction from Cross Domain Data with Declarative Language

by

Mingda Li
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2020
Professor Junghoo Cho, Co-Chair
Professor Carlo Zaniolo, Co-Chair

With large amounts of data continuously generated by intelligence devices, efficient analysis of huge data collections to unearth valuable insights has become one of the most elusive challenges for both academia and industry. The key elements to establishing a scalable analyzing framework should involve (1) an intuitive interface to describe the desired outcome, (2) a well-crafted model that integrates all available information sources to derive the optimal outcome and (3) an efficient algorithm that performs the data integration and extraction within a reasonable amount of time. In this dissertation, we address these challenges by proposing (1) a cross-language interface for a succinct expression of recursive queries, (2) a domain specific neural network model that can incorporate information of multiple modality, and (3) a sample efficient training method that can be used even for extremely-large output-class classifiers.

Our contributions in this thesis are thus threefold: First, for the ubiquitous recursive queries in advanced data analytics, on top of BigDatalog and Apache Spark, we design a succinct and expressive analytics tool encapsulating the functionality and classical algorithms of Datalog, a quintessential logic programming language. We provide the Logical Library (LLib), a Spark MLlib-like high-level API supporting a wide range of recursive algorithms and the Logical DataFrame (LFrame), an extension to Spark DataFrame supporting both relational and logical operations. The LLib and LFrame enable smooth collaborations between logical applications and other Spark libraries and cross-language logical programming in Scala, Java, or Python. Second, we utilize

variants of recurrent neural network (RNN) to incorporate some enlightening sequential information overlooked by the conventional works in two different domains including Spoken Language Understanding (SLU) and Internet Embedding (IE). In SLU, we address the problem caused by solely relying on the first best interpretation (hypothesis) of an audio command through a series of new architectures comprising bidirectional LSTM and pooling layers to jointly utilize the other hypotheses' texts or embedding vectors, which are neglected but with valuable information missed by the first best hypothesis. In IE, we propose the DIP, an extension of RNN, to build up the internet coordinate system with the IP address sequences, which are also unnoticed in conventional distance-based internet embedding algorithms but encode structural information of the network. Both DIP and the integration of all hypotheses bring significant performance improvements for the corresponding downstream tasks. Finally, we investigate the training algorithm for multi-class classifiers with a large output-class size, which is common in deep neural networks and typically implemented as a softmax final layer with one output neuron per each class. To avoid expensive computing the intractable normalizing constant of softmax for each training data point, we analyze the well-known negative sampling and improve it to the amplified negative sampling algorithm, which gains much higher performance with lower training cost.

The dissertation of Mingda Li is approved.

Yizhou Sun

Yingnian Wu

Carlo Zaniolo, Committee Co-Chair

Junghoo Cho, Committee Co-Chair

University of California, Los Angeles

2020

To my mother, father and girlfriend.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisors, Professor Junghoo Cho and Professor Carlo Zaniolo. Professor Junghoo has been an exceptional advisor. He is my first mentor in the natural language processing field. Over the years, he gave me extremely visionary advice, generous support and timely feedbacks. His research foresight and ambition greatly motivated me. For each of my projects, he enthusiastically provided insightful suggestions and related articles. For each of my papers, he patiently pointed out logic flaws and unclear descriptions. I also feel so fortunate to have Professor Carlo as my co-advisor. His professional expertise and brilliant ideas guided me to resolve many challenging problems. He spared his valuable time teaching me how to come up an idea to satisfy urgent needs in academy, establish a system to implement the idea, and to illustrate findings with concrete examples.

I would also thank Professor Yizhou Sun and Professor Yingnian Wu for agreeing to be on my committee and providing their valuable time and advice on my research.

I would thank Tyson Condie for offering me the precious opportunity to join UCLA CS department as a PhD student and offering funding support to my research. During the early years, he taught me the way of thinking as a system researcher. He showed me the frontier of distributed system research. He is quite smart and experienced. I could always get innovative thoughts while discussing with Tyson.

It is a great honour to work in ScAi lab. I would thank the whole UCLA ScAi group, especially Matteo Interlandi, Youfu Li, Zijun Xue, Jiaqi Gu, Jin Wang, Muhao Chen, Manoj Reddy, Ling Ding, Xuelu Chen, and Shi Gao who helped me at various times. Chapter Four is a version of the paper collaborated with Zijun Xue and my advisor Junghoo Cho, which is in preparation for publication.

I also want to thank all staff members of UCLA Computer Science Department, especially Joseph Brown, who helped me a lot for various questions I have ever met.

I have ever done four marvelous internships at Teradata, Nec Laboratories, AWS RedShift and

Alexa AI, where I learned a lot from many remarkable people. I want to thank my mentors in Nec Laboratories, Cristian Lumezanu and Bo Zong for their guidance on IP embedding and spoofing detection projects. Chapter Two consists of a version of the IP embedding paper (Li et al., 2018a) and a version of the spoofing detection paper, which is in preparation for publication. I want to thank Chengwei Su, Weitong Ruan, Xinyue Liu, Luca Soldaini, Wael Hamza and Fan Yang for their assistance and participation in the spoken language understanding project. Chapter Three is a version of the spoken language understanding project's paper (Li et al., 2020). I also want to thank Yi Xia, Emiran Curtmola, Chainani Naresh, Gaurav Saxena for their generous support.

Lastly but most importantly, I own my parents and my girlfriend a debt of gratitude. Without their unconditional love and unbounded support, I cannot remain true to my original aspiration, keep pressing ahead and get through all those hard times.

VITA

2011-2015	B.S. Computer Science and Technology Harbin Institute of Technology Harbin, China
2015–2016	Research Assistant ScAi Laboratory University of California, Los Angeles Los Angeles, California
2016–2020	Teaching Assistant Computer Science Department University of California, Los Angeles Los Angeles, California
2016	Research Intern Teradata Los Angeles, US
2017	Research Intern NEC Lab Princeton, US
2018	SDE Intern Amazon AWS Redshift Palo Alto, US
2019	Applied Scientist Intern Amazon Alexa AI Boston, US

PUBLICATIONS

Mingda Li, Weitong Ruan, Xinyue Liu, Luca Soldaini, Wael Hamza, Chengwei Su. "Improving Spoken Language Understanding By Exploiting ASR N-best Hypotheses." ArXiv 2020. To be submitted to *SLT 2021*

Mingda Li*, Zijun Xue*, Junghoo Cho. "Amplified Negative Sampling: Sample-Efficient Training for a Large-Class Classifier." In review by *KDD 2020*

Mingda Li, Jin Wang, Youfu Li, Carlo Zaniolo. "LLib and LFrame: An Expressive Multi-Language Datalog Interface." To be submitted to *ICLP 2020*

Mingda Li, Cristian Lumezanu, Bo Zong, Haifeng Chen. "Learning-based spoofing detection." In review by the *EuroS&P WTMC 2020*

Jin Wang, Chunbin Lin, Mingda Li, Carlo Zaniolo. "An Efficient Sliding Window Approach for Approximate Entity Extraction with Synonyms." Accepted by *EDBT 2019*

Mingda Li, Cristian Lumezanu, Bo Zong, Haifeng Chen. "Deep Learning IP Network Representations" **BEST PAPER AWARD** of *ACM SIGCOMM Big-DAMA 2018*

Mingda Li, Cristian Lumezanu, Bo Zong, Haifeng Chen. "Learning IP Network Representations." Accepted by *ACM SIGCOMM CCR 2018*

Youfu Li, Mingda Li, Ling Ding, Matteo Interlandi . "RIOS: Runtime Integrated Optimizer for Spark." Accepted by *SOCC 2018*

CHAPTER 1

Introduction

1.1 Motivations

In the big data era, we have witnessed the rising demand of efficiently and conveniently extracting insights from large-scale data sets for decision making in different domains. The demand has driven researchers to propose various neural network-based algorithms revolutionizing many fields, ranging from image processing (He et al., 2016), natural language processing (Devlin et al., 2018) to speech recognition (Amodei et al., 2016), etc. In addition, the big data analytics and machine learning platforms like PyTorch (Paszke et al., 2019), Tensorflow (Abadi et al., 2016) and Apache Spark (Zaharia et al.) are continuously built up in open source and commercial markets, to provide maximal flexibility and speed while implementing the analyzing pipeline with existing or user-defined algorithms.

Although the huge success has been achieved, for the advanced analytics on scalable data sets, there are still some challenges and ongoing efforts to tackle them, including:

- **Excluded data.** While great efforts have been dedicated to assimilate massive amounts of data for analysing algorithms, only a tip of the big data iceberg has been utilized during analysing. To make full use of the assimilated data, researchers design novel architectures of algorithms like the bi-directional RNN (Graves et al., 2013) to exploit the backward information, and involve more tasks for training like the masked language model and next sentence prediction mentioned in Bert (Devlin et al., 2018).

  In two of the fastest growing areas in computer science, the internet embedding and spoken language understanding, we find some enlightening information not considered in the existing analyzing pipeline, which makes the performance of corresponding downstream

tasks non-optimal. Currently, while embedding the internet structure, only a single source of structural data among IP addresses like hop counts (Eriksson et al., 2008, 2009) is utilized to build up the network coordinate system. We realize the IP address of the host in internet, a sequence of bits, could provide a coarse indication of the location of the host, but is overlooked by current internet embedding techniques. As for understanding the speech, only the recognition result (hypothesis) of an input speech with highest ASR confidence score (Tur and De Mori, 2011) or reranking score (Peng et al., 2013; Morbini et al., 2012) is relied and transferred to the natural language understanding (NLU) module for domain or intent classification. We find the first best hypothesis can be noisy, while the other hypotheses can be more similar to the ground-truth transcription of the speech. Driven by the above findings, we would like to propose novel frameworks involving the unnoticed information for a better performance in downstream tasks.

- **Expensive computing.** Among all the analysing algorithms, the neural networks with numerous layers and parameters are recently widely utilized in different domains due to their strength at learning features at different levels of data abstraction for a better decision making. However, the training could be quite computationally expensive especially when the final layer is a softmax layer for large-class classification (e.g. word embedding, graph embedding), since the cost of standard training algorithm is proportional to the output class size (e.g. size of words in a dictionary, nodes in a graph). To tackle the issue, a number of techniques like, negative sampling (Mikolov et al., 2013a), hierarchical softmax (Morin and Bengio, 2005), adaptive softmax (Bengio, 2008; Rawat et al., 2019) , are developed.

  Negative sampling is one of the most popular techniques utilized in practice due to its simplicity and efficiency. Experimental observations show that a larger negative sample size can achieve a better downstream task performance. However, the training cost is more expensive for a larger sample size. Could we get the best of the two worlds? This dream motivates us to analyze the technique and further amplify the negative sampling to get a higher performance with unchanged or even smaller sample size.

- **Expressive power.** Recursions are ubiquitous in advanced data analytics, such as graph an-

alytics, data mining algorithms. An efficient development of a complicated recursive algorithm on the well-known data analytics platforms like Apache Spark requires deep understanding of the algorithm and platform's libraries. To simplify the development, a renaissance of interest has been brought to Datalog, a declarative logic programming language, for its succinct expression of recursions. Numerous of Datalog systems including DeALS (Yang et al.), BigDatalog (Shkapsky et al., 2016), RaSQL (Gu et al., 2019) are built up but there is still space for improvement on usability and interoperability.

Most of the conventional Datalog systems adopt the Datalog syntax, which requires a deep understanding of logical programming. Recently, this is realized by RaSQL, which proposes a simple extension of SQL syntax to improve the usability. However, for a wider audience from data science community, should it be better to provide a data scientists' familiar cross-language API encapsulating a wide range of Datalog algorithms like Spark MLlib? For a flexible developing, is it possible to support the logical operations within a DataFrame-like data structure? These two questions stimulate us for a better Datalog-based data analytics tool design.

1.2 Contributions

All the contributions of this thesis center around the aforementioned motivations and are summarized as follows:

- We realize and experimentally show the structural information contained in IP addresses for internet embedding and propose a deep learning based framework, DIP, to utilize the information. To the best of our knowledge, DIP is the first framework to predict distance (or hop count) to arbitrary IPs (even unknown hosts i.e., not contained in training data) based only on the value of their IP address and routable prefix without any other domain knowledge.

- We further explore the impact of deep learning in the network security area by using network embeddings to learn IP maps for spoofing detection. We combine the DIP with the hop count

filtering, a well-known map-based spoofing detection mechanism, which maps each IP to a hop count value for one target IP address and is restricted to the detection on specific targets. The new framework can help any Internet server detect packets spoofed with any IP address without any additional measurements to that IP.

- We pioneer the spoken language understanding research on jointly utilizing all hypotheses from ASR module. We investigate the exact matching between hypotheses and the ground-truth transcription of the input speech, which reveals the value contained in the $2^{nd} - n^{th}$ best hypotheses. To involve more than one hypothesis during NLU, we introduce a series of simple yet efficient models and significantly improve the SLU system robustness to the noises from ASR module.

- We take the effort to efficiently train the high computational cost large-class classifiers. We propose a new sample-efficient training algorithm, amplified negative sampling (ANS). We theoretically and experimentally demonstrate that the ANS leads to the higher-accuracy model of a larger sample size without paying its high computational cost.

- Finally, we design a cross-language (Python, Scala, or Java) Datalog programming interface with two important components, LLib and LFrame. LLib is a high-level logical library, providing the encapsulation of a wide range of Datalog algorithms. LFrame is an extension to DataFrame with the functionality of basic logical operations like definitions of recursive rules. With running examples, we show the simplified development of recursive applications and flexible collaborations between LLib or LFrame and existing Spark libraries.

1.3 Thesis Outline

In this dissertation, we mainly address three of the most important elements of the efficient scalable data analyzing. We firstly explore how to make full use of all information while extracting semantics and insights for decision making. More specifically, in Chapter 2: (a) We propose the DIP, a variant of RNN, for establishing the network coordinate system via embedding the IP address sequence. Compared to the coordinate system built on single source of structural data like

latency or hop count, we demonstrate the superiority of DIP for unknown IPs; (b) We show the spoofing detection, one of a wide range of problems the DIP framework can be applicable to, and the benefits from DIP to significantly reduce the cost of achieving complete IP maps. In Chapter 3, we explore the value of unused ASR interpretations and numerous ways to integrate them in the spoken language understanding system.

Secondly, we explore how a large-class classifier can be efficiently trained through a sample-efficient training algorithm in Chapter 4. In the chapter, the intuition, theoretical and experimental analysis of the developed algorithm, amplified negative sampling, are discussed. Then, we explore how to design a succinct interface for advanced data analytics with superiority on recursion expressions in Chapter 5. We develop a high-level Datalog library, LLib, for simplified end-to-end recursive application development with existing Datalog algorithms and a data structure, LFrame, for flexibly defining the logic of a new recursive application. We finally conclude the dissertation and discuss avenues for future work in Chapter 6.

CHAPTER 2

Extracting Latent Information from Unused IP Addresses

In this chapter, we discuss IP addresses (IPs), the unused data in internet embedding domain by illustrating (1) the reasons why IPs could reveal the structural information of nodes in the Internet, (2) one approach to incorporate the IPs in the internet embedding and (3) one application which the new embedding framework can be utilized in.

We firstly present DIP, a deep learning based framework to learn structural properties of the Internet, such as node clustering or distance between nodes, from the IP addresses. Existing embedding-based approaches use linear algorithms on a single source of data, such as latency or hop count information, to approximate the position of a node in the Internet. In contrast, DIP computes low-dimensional representations of nodes that preserve structural properties and non-linear relationships across multiple, heterogeneous sources of structural information, such as IP, routing, and distance information. Using a large real-world data set, we show that DIP learns representations that preserve the real-world clustering of the associated nodes and predicts distance between them more than 30% better than a mean-based approach. Furthermore, DIP accurately imputes hop count distance to unknown hosts (*i.e.*, not used in training) given only their IP addresses and routable prefixes. Our framework is extensible to new data sources and applicable to a wide range of problems in network monitoring and security.

Then, we consider an important topic, spoofing defense, based on the full knowledge of the internet structural properties. Map-based IP spoofing defenses associate source IPs to immutable structural properties of the Internet, such as paths, hop counts, or neighbors to a target, and filter out packets whose header information does not match the maps. Although accurate, existing methods lack sufficient coverage. Network maps on AS border routers do not detect spoofed packets that traverse unprotected networks. Host maps at the edge protect only against spoofed packets with

source IPs known by the host. We propose to learn IP maps by constructing a structural model (or embedding) of the Internet from a limited number of measurements. We study the feasibility of learned IP maps by combining hop count filtering, a known map-based spoofing defense mechanism that maps IPs to hop counts to a target, and DIP, a deep learning based learning algorithm that computes vector representations of IPs that preserve hop count distance between them. Using a large data set of hop counts between Internet hosts, we show that learned maps can detect packets spoofed with almost *any* IP address and traversing *any* path using an embedding generated from only a few thousand IP addresses and hop counts between them. In addition, our embeddings are general: an Internet model trained for a set of hosts can be used by any other host to generate new IP maps with little loss in accuracy.

2.1 Deep Learning IP Network Representations

2.1.1 Introduction

The ability to map, analyze, and understand the structure of the Internet helps network management and operations by revealing opportunities for improvement or potential design flaws. For example, accurately predicting the closest server is critical in peer selection and load balancing (Miao et al., 2017). Knowing how remote IPs are clustered can help diagnose anomalous events such as spoofing attacks (Jin et al., 2003). A holistic view of the network and its structure is essential towards achieving the vision of self-driving networks (Feamster and Rexford, 2017).

Most previous attempts to uncover the Internet structure relying on active probing from multiple vantage points using tools such as *traceroute* and *ping* (Levchenko et al., 2017; Spring et al., 2003). Such techniques provide fine-grained introspection (*i.e.*, can measure specific properties in specific parts of the network, such as the latency of a path) but pose a significant cost in terms of network overhead.

In contrast, embedding-based approaches use fewer, strategic measurements (Dabek et al., 2004; Eriksson et al., 2008) or passive observations on network traffic (Eriksson et al., 2007) to learn vector representations for the network end-hosts in a low-dimensional space. The representa-

tions approximate the positions of hosts in the Internet and are used to recover structural network properties, such as distance between nodes or clustering of nodes. However, the complexity of the Internet and the sparse input data make it difficult to compute accurate representations. Oftentimes, embedding approaches rely on additional data sources, which cannot be easily used in the embedding process, to refine and tune the final embeddings. For example, several embedding methods build representations based on distance-based metrics, such as latency or hop count, and then refine (or even replace) the final representations using additional probes or static information such as AS membership or routing information (Dabek et al., 2004; Eriksson et al., 2009).

The emergence of deep learning as a powerful tool to extract hidden features in data calls for revisiting the problem of learning network representations through embedding. In particular, deep learning techniques provide two key benefits. First, they allow multiple heterogeneous sources of information as input, thereby identifying more accurately the relationships between multiple sources of data that jointly contribute to a specific structural property (Karpathy and Fei-Fei, 2015; Wang et al., 2018; Mikolov et al., 2013b). Second, deep neuron networks are extensible and can easily incorporate additional sources of information by attaching more neurons, network layers, or network branches (Wang et al., 2018). One can start with a model trained on the original components and re-train it using only the newly added parts or data sources (Erhan et al., 2010). This makes it easier for network operators to deploy, apply, or update neural network based models.

We propose DIP, a deep learning based framework to learn the structure of the Internet. DIP is a ten-layer neural network[1] that computes a low-dimensional vector representation for any node[2] in the Internet *given only its IP address and routable prefix*. DIP preserves both local and global structure: clustered nodes have similar representations and the distance between two representations approximates the hop count between the associated nodes.

We train our neural network using three heterogeneous data sets: hop count distances between Internet nodes, the 32 bits IP address and inter-domain routable prefix information for each node.

[1]To avoid confusion and unless explicitly stated otherwise, we use *network* or *Internet* to refer to the physical IP network and *neural network* to refer to the neural network we design to learn the structural properties of the physical network

[2]We use *node* or *(end-)host* to denote any computer connected to the Internet and assigned an IP address.

A key insight to train DIP is to *first compute representations based on the IP and routing information*, thereby recovering structural information hidden in the IP values, and refine them using a distance-based optimization. As the size of the routable prefix varies by IP, we first normalize the IP and prefix data by representing an IPv4 address on 64, rather than, 32 bits. To capture structural information encoded in the IP address value, we feed the eight bytes of the normalized IP sequentially at each of the first eight layers of the neural network. We then use the last two layers to get the hop count matrix and optimize the embedding distance prediction. With a trained DIP, we can estimate distance between *any* two Internet hosts as long as we have their IPs, even if they are not part of the training data.

Results on large real-world data sets of hop counts between thousands of IP addresses and 95 geographically distributed servers show that we can predict hop count distance between known hosts (*i.e.*, whose IP address were used in the training) with an absolute error of around 2 hops and over 30% better than a mean-based method. We infer the distance between unknown IPs (*i.e.*, not appearing in training data) with a small loss in accuracy compared to known IPs. In addition, the representations learned by DIP preserve the real-world clustering of the associated hosts. The accuracy of our model increases when we increase the training data set.

While our results are preliminary, they offer us a glimpse of the power of deep learning in recovering structural properties of the Internet from sparse data. DIP is the first framework that can estimate accurately the distance to any Internet host given only its IP address and routable prefix without any distance data.

2.1.2 Background and Related Work

What is structure? Many properties can make up the structure of the Internet: connectivity between IPs, routers, or networks; distance-based metrics such as hop count or latency; similarity-based metrics such as the set of one's neighbors in the connectivity graph; path-based properties such as the sequence of routers on a path. Here, we focus on two specific properties that define both the local and the global structure of a network: clustering of end-hosts and hop count based distance between end-hosts.

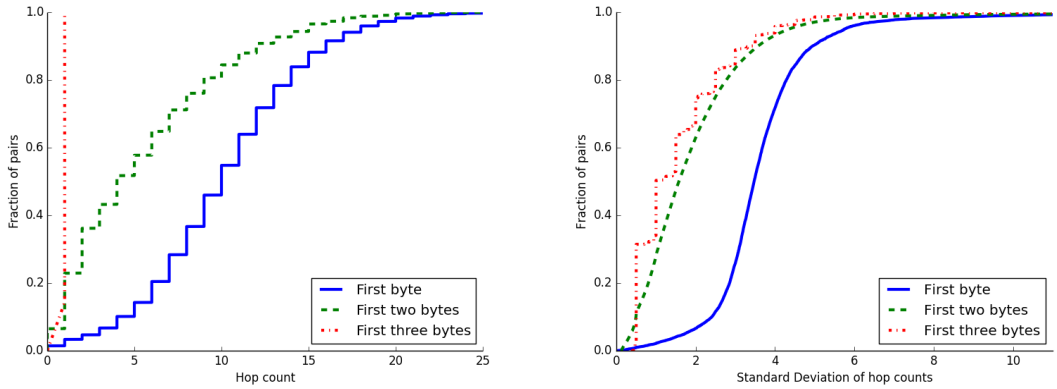


Figure 2.1: Cumulative distribution of (left) hop counts between pairs of host-server IPs that share the first, first two, or first three bytes, and (right) standard deviation of hop count distribution among groups of IPs sharing the first, first two, or first three bytes. The more similar two IPs are, the closer they are and the more similar their distances to the same third IP are.

Network coordinate systems learn vector representations for participating nodes such that the position of the node in the embedding approximates its position in the Internet. Most coordinate systems build embeddings using a single source of structural data: latency measurements among nodes or to predetermined landmark servers (Dabek et al., 2004; Ng and Zhang, 2002; Costa et al., 2004; Pyxida; Zhao et al., 2011) or hop count information from passive traffic observations (Eriksson et al., 2008, 2009). Latency and hop count data is often sparse and cannot always be accurately embedded in metric spaces. To overcome these issues, several approaches use out-of-band information, such as location (Dabek et al., 2004) or routing (Eriksson et al., 2009) data, or perform active measurements (Eriksson et al., 2008) to impute the missing data and detect clusters or distances. Unlike them, we propose to train our embedding jointly using distances, routing information and host IP values, thereby learning hidden structural features encoded in a node's IPv4 address. With a trained model, we are able to embed and find the hop count to any IP, without the participation of its host.

Deep neural networks consist of multiple layers of interconnected neurons (LeCun et al., 2015). A neuron aggregates multiple input values using local weights and biases, applies an activation function, and produces one or more numerical values as output. Given a training task, one can define an objective function to evaluate the output of the entire neural network, *e.g.* , prediction error. Using gradient-based back-propagation algorithms to optimize the objective function (Kingma

and Ba, 2014), neural networks automatically tune the weights and biases of each neuron to achieve a better performance.

2.1.3 Learning Network Representations

DIP learns an embedding model that accurately reflects the structure of the Internet, *i.e.*, preserves node clustering and distances between nodes. The goal of learning is to minimize the prediction error for the distance between any two nodes. The learned model is defined by the structure of the neural network and the final values for the weights and biases of each neuron. Next, we describe the data used in learning and how we construct the neural network.

2.1.3.1 Data Sources

IP addresses and routing information. The IP address of a host provides a coarse indication of the location of the host in the Internet. To make routing scalable and fast, IP addresses are assigned hierarchically and divided into a network (or routable) part and a host (or local) part. The routable part, usually expressed by an integer representing the number of bits (also called prefix), tells routers how to route the packet through the core of the Internet towards the destination network. Intuitively, IPs with the same routable prefix share a path towards them through the Internet core and are more likely to be close to each other.

Hop counts. The hop count between two hosts represents the number of routers on the default path between hosts. We use hop count, rather than latency, to measure the distance between two hosts, as it can be easily extracted from the TTL value of a network packet (Jin et al., 2003), without active measurements. In Section 2.1.5, we discuss how to extend the model using latency measurements. Our hop count matrix is asymmetric and very sparse; it does not contain hop counts between all IPs.

2.1.3.2 IP Transformation

The key idea of our work is to use both local (IPs and routing information) and global (hop counts) structural information to guide the embedding of network nodes. By utilizing deep learning for embedding, we can identify and use hidden features encoded in the IP address of a given node. We perform several transformations on the input, guided by observations on real network data.

IP normalization. Because the routable information is tied to an IP address, we combine the IP and prefix values when feeding them to the neural network. To keep the size of the input constant and independent on the prefix size, we generate a *normalized IP address* for each regular IP. The process of normalization is depicted in Figure 2.2. We divide each IP into the network and the host parts. We pad the end of the network part and the beginning of the host part with zero to obtain two four-byte values. We concatenate the values and get the eight bytes normalized IP. Further, for easier processing, we represent each byte of the input in one-hot vector format (256 dimensions), *e.g.* , a one and the rest are 0s, where the 1's position is the value of the byte (0 to 255).

Sequential feeding. IP addresses are assigned hierarchically and encode structural information of the network. To better understand how the hierarchical assignment affects node clustering, we perform two experiments on a data set of hop counts between 95 geographically distributed servers and ten million IP addresses of end hosts. Section 2.1.4.1 describes the data in more detail.

First, we group all pairs of host-server IPs according to whether they share (within the pair) the first byte, first two bytes, or first three bytes. We show the all-to-all hop counts between pairs in each of the three groups in Figure 2.1(left). The more similar two IP addresses are, the closer they are in terms of number of hops. Second, we group separately hosts and servers according to whether they share the first one, two, or three bytes and generate the hop count distribution for each pair of host-server groups that share the same prefix. We present the standard deviation for each pair in Figure 2.1(right). The smaller the standard deviation is, the more similar the distances are. This means that the more similar two IPs are, the more likely they have the same hop count to another node.

As shown in Figure 2.1, an IP address can help learn node representations that capture the network structure. The more bytes of an IP address we know, the better we can constrain the

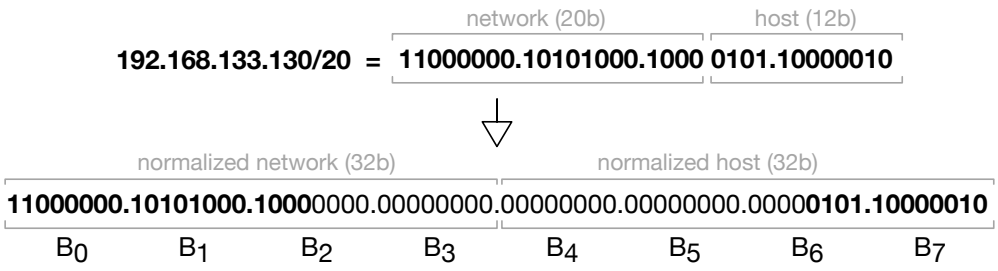


Figure 2.2: Generating a normalized IP address for 192.168.133.130/20.

representation we assign to it. In addition, the more significant bytes of an IP address have a higher influence on the position of the associated host relative to other hosts. Therefore, the key is to capture the sequential correlation among the bytes of an IP address.

2.1.3.3 Network Construction

Driven by the insight gained in the previous section, we develop DIP, a deep neural network that computes vector representations of network hosts based on their IP addresses and the hop counts to other hosts. The design of DIP, depicted in figure 2.3, is similar to that of a recurrent neural network (Mikolov et al., 2010), where new data is processed in the context provided by previous data (*e.g.* , like processing natural language). We explain the details below. Even though the figure and our explanation refer to the input as one-hot vectors (*e.g.* , a normalized IP is represented as a vector of size 8x256=2,048), in reality the inputs are matrices (*i.e.*, the number of IP addresses times 2,048). Because our hop count data (see Section 2.1.4.1) is between separate end-hosts (sources) and servers (destinations) and because distances in the Internet are not always symmetric, we choose to feed the source and destination IPs separately in the neural network.

Intermediate IP representation. As mentioned earlier, to get the most out of the format and value of an IP address towards building a representative embedding for its host, we should treat each byte separately. The more significant bytes can provide a context for how to interpret the less significant bytes. Thus, we choose to input each byte of the normalized IP (a 256-dimension one-hot vector $B_{256\times 1}^{i\in\{0,\ldots,7\}}$) separately at each layer of the network. The input of layer i is the con-

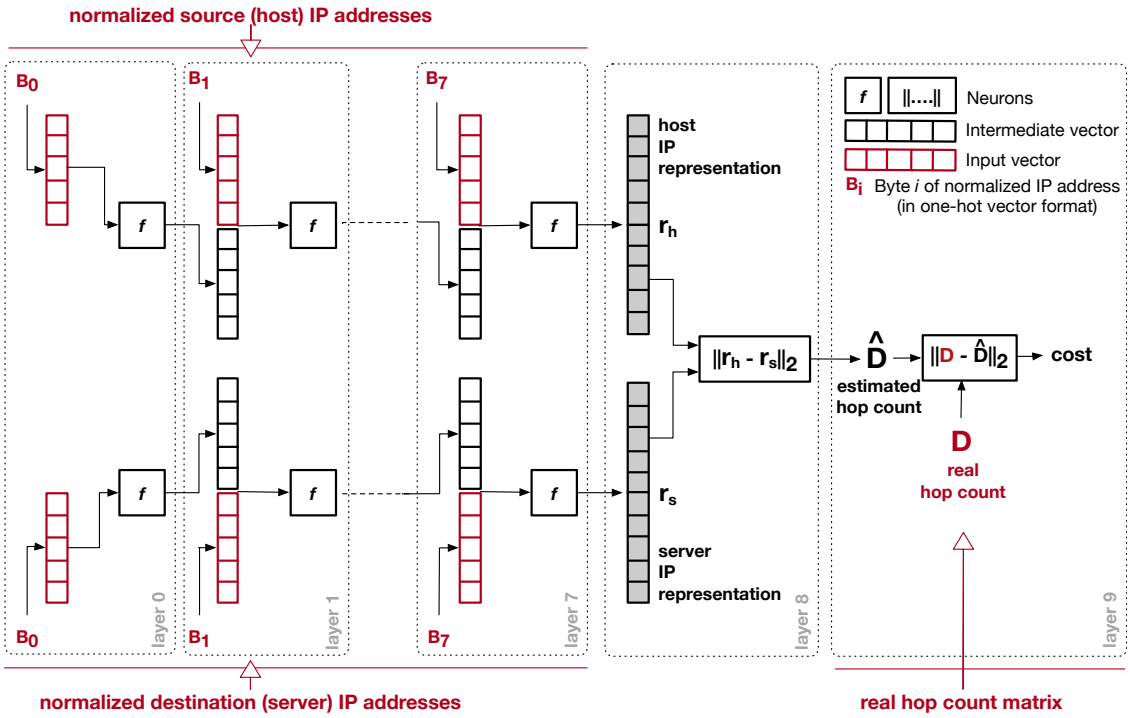


Figure 2.3: The neural network used for training our embedding model. The first eight layers receive the normalized IP addresses as input and compute the IP representations. The ninth layer estimates the hop count between two IP addresses and the tenth layer measures the model error. Elements in red are input. For simplicity we depict the input as one-dimensional vectors (one normalized IP); in reality, all inputs are matrices.

catenation of byte i with the output of the previous layer (except for the first layer). This

$$Input = \begin{cases} i = 0 & B^{i=0}_{256\times 1} \\ i \in \{1,...,7\} & concat(f^{i-1}_{d\times 1}, B^{i}_{256\times 1}) \end{cases} \tag{2.1}$$

where d is the dimension of the final IP representation and *concat* represents the vector concatenation operation.

At each layer, the activation function f is given by:

$$f^i = \begin{cases} i = 0 & softsign(w_{d\times 256}^{i=0} \times B_{256\times 1}^{i=0} + b_{d\times 1}^{i=0}) \\ i \in \{1,...,7\} & softsign(w_{d\times(256+d)}^{i\in\{1,...,7\}} \times concat(B_{256\times 1}^{i\in\{1,...,7\}}, f_{d\times 1}^{i-1}) + b_{d\times 1}^{i\in\{1,...,7\}}) \end{cases} \tag{2.2}$$

where $w_{d\times(256+d)}^{i\in\{0,...,7\}}$ are weights and $b_{d\times 1}^{i\in\{0,...,7\}}$) are biases; the *softsign* function is $f(x) = \frac{1}{1+|x|}$. Initially, we assign random values to all weights and zeros to all biases. We employ *softsign* as the activation function for the ease of training, as *softsign* is more robust to saturation compared to other popular activation functions, such as *sigmoid* and *tanh*.

Intermediate distance estimation. We use the first eight layers of the neural network to process each of the eight bytes of the input normalized IP address. The output of the eighth layer is the intermediate vector representation for each IP address in the input data. We then use the last two layers to estimate how good the representation is. First we compute the estimated hop counts given by the current representation using an Euclidean distance. Given two matrices $H_{h\times d}$ and $S_{s\times d}$ storing the intermediate representations for the h hosts and s servers separately, the estimated distance matrix is:

$$Dist_{h\times s} = Euclidean(H_{h\times d}, S_{s\times d}) \tag{2.3}$$

Error reduction. Finally, we compare the estimated hop counts with the real hop counts matrix $D_{h\times s}$ to compute the cost as the mean difference of hop-counts. As the real hop count matrix is sparse, we compare only the valid entries:

$$Cost = \frac{\sum_{i=1}^{h}\sum_{j=1}^{s} W^{(i,j)}(||r_{d\times 1}^{H_{i\in\{1,..,h\}}} - r_{d\times 1}^{S_{j\in\{1,..,s\}}}|| - D^{(i,j)})}{count\,of\,non-zero\,D^{(i,j)}} \tag{2.4}$$

$D^{(i,j)}$ represents the value of the element at i^{th} row and j^{th} column in matrix D. $r_{d\times 1}^{H_{i\in\{1,..,h\}}}$ and $r_{d\times 1}^{S_{j\in\{1,..,s\}}}$ are rows in the matrices $H_{h\times d}$ and $S_{s\times d}$, and correspond to the representation of a host or

server in the embedding space. W is a binary (0-1) matrix whose elements are defined as:

$$W^{(i,j)} = \begin{cases} 0 & D^{(i,j)} == 0 \\ 1 & D^{(i,j)} \neq 0 \end{cases} \tag{2.5}$$

To minimize the cost, we utilize the Adam algorithm, a gradient descent based back-propagation method (Kingma and Ba, 2014), which is able to automatically tune the learning rate during the training process.

2.1.4 Evaluation

2.1.4.1 Data and Methodology

We use a large data set of network hop counts from the Ark project (Ark IPv4). The data contains hop count information from 95 geographically distributed servers to ten million IP addresses that cover all routable prefixes in the Internet. We use data collected by Ark during Jun 2015. For each IP in the data, we look up the routable prefix and normalize it using the steps in Section 2.1.3.2. Due to the cost of monitoring a large number of IPs, not all servers have hop counts for all ten million IPs. Our hop count matrix is incomplete and contains valid entries for only 29% of the pairs. We extract IP prefix information from Routeviews data (RouteViews) and use a default value of 24 for missing prefixes.

We build a prototype for DIP using TensorFlow. We train the neural network using several smaller data sets obtained by randomly sampling 1,000, 10,000, and 100,000 IPs from the original data and keeping only the hop counts to them. Sampling increases the sparsity of the data: less than 15% of the entries in the smaller data sets are valid. We also vary the number of servers and the dimensionality of the embedding space. Intuitively, having fewer IPs or servers may not provide sufficient constraints to learn accurate representations and lead to an underfit model. Increasing the number of dimensions can reveal more hidden features, invisible at lower dimensions, but may lead to overfitting. Each training session has 2,500 iterations, *i.e.*, passes through the neural network to update the weights and biases. We use a GPU server with four 3.5GHz quad-core Intel

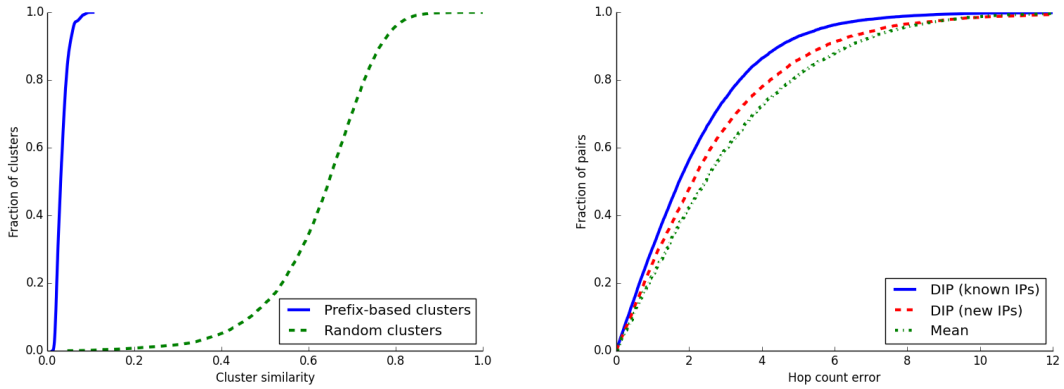


Figure 2.4: (left) Cumulative distribution of cluster similarity, computed using IP vector representations, for prefix-based and end-host random clusters; (right) Cumulative distributions of absolute distance estimation errors for DIP and *mean*. DIP representations preserve real-world prefix-based clustering and predict distances accurately.

Xeon processors and 128GB of RAM. We generate testing sets by randomly sampling the original data and preserving the previously trained parameters for embedding arbitrary IP via its address (*e.g.* , the weights and biases of each byte/layer).

2.1.4.2 Embedding Accuracy

Clustering. First, we assess how well DIP preserves the clustering of hosts in the original IP space. For this, we group all IPs first according to their routable prefix and then at random. For each cluster we compute an embedding similarity metric, defined as the ratio between the average distance between all pairs of IP representations in the cluster and the maximum distance across all clusters. The lower the similarity value, the closer to each other the IPs of a cluster are in the embedding space. Figure 2.4(left) shows the similarity distribution for prefix-based and random clusters. Each IP representations is a 140-dimensional vector and computed after training the network using 10,000 IP addresses and 95 servers. Our embedding preserves the clustering of the original IP space well.

Distance prediction. To assess the quality of distance prediction, we first look at previous embedding mechanisms. Network coordinate approaches (Dabek et al., 2004; Ng and Zhang, 2002) are not directly comparable as they embed latencies between strategically chosen pairs of nodes,

	DIP		Mean	
	known IPs	**new IPs**	**known IPs**	**new IPs**
Number of IPs				
1,000	2.16 (1.86)	2.89 (2.38)	2.99 (2.44)	3.27 (2.51)
10,000	2.15 (1.79)	2.68 (2.34)	3.00 (2.40)	3.04 (2.43)
100,000	2.06 (1.76)	2.29 (2.00)	2.98 (2.40)	2.97 (2.40)
Number of servers				
12	2.79 (2.25)	3.03 (2.47)	3.21 (2.54)	3.28 (2.62)
24	2.39(2.14)	2.60 (2.25)	2.90 (2.36)	2.97 (2.42)
48	2.34 (2.05)	2.75 (2.27)	2.99 (2.39)	3.02 (2.40)
95	2.15 (1.79)	2.68 (2.34)	3.00 (2.40)	3.04 (2.43)
Embedding dimension				
110	2.28 (1.93)	2.80 (2.39)	3.00 (2.42)	3.04 (2.43)
140	2.15 (1.79)	2.68 (2.34)	3.00 (2.42)	3.04 (2.43)
170	2.19 (1.86)	2.72 (2.33)	3.00 (2.42)	3.04 (2.43)

Table 2.1: Absolute mean error (standard deviation between brackets) of distance prediction of DIP and *mean*, for both known and new IPs, when varying the number of IPs, the number of servers and the embedding dimension. The default values are 10,000 IPs, 95 servers, and 140 dimensions.

while we rely on hop count information from passively observed traffic. Eriksson *et al.* (Eriksson et al., 2009) propose a matrix factorization based algorithm to predict hop count information but first build baseline representations of the monitoring servers using an all-to-all hop count matrix. We lack complete hop count information among servers and build our embedding directly from incomplete server-to-host distances. Therefore, we compare against a *mean estimation approach*, where we predict a host-to-server distance as the mean of all valid distances to the same server.

We look at how well our embedding estimates the hop count value between a host and a server. We first consider only the IP addresses used in the training process (*i.e.*, *known IPs*). For this, we train a model using 90% of all host-server pairs and use the remaining 10% for testing. Figure 2.4 (right) compares the absolute error between estimated and real hop count for DIP and *mean* for the 10,000 IPs data set on 140 dimensions. DIP predicts distances with a mean absolute error of around two hops (23% mean relative error) and reduces the error of the *mean* estimation by almost 30%. Table 2.1 presents the average absolute error and standard deviation for hop count estimation for embeddings trained with different number of hosts, servers, or dimensions. As expected, increasing the the number of IPs, servers, or the dimensionality reduces the absolute error. We also trained models with parameters outside the ranges presented in the table but found no improvement.

New IPs. An important feature of DIP is its ability to impute hop count values to arbitrary

nodes based on their IP address. *New IPs* are IP addresses not used in the training process and that DIP has never seen before. Figure 2.4 (right) and Table 2.1 show that DIP approximates distance to new IPs with high accuracy. The distance prediction error is only around half a hop more than that for known IPs. To the best of our knowledge, DIP is the first framework to predict hop counts to arbitrary hosts based only on the value of their IP address and routable prefix and without any other domain knowledge.

2.1.5 Discussion: Limitations and Opportunities

We discuss several future applications and directions of using deep learning to understand and capture the Internet structure.

Extensions. An important benefit of using neural networks for learning the structure of the Internet is that they can be extended easily for other data sources. Similarly to previous embedding approaches (Dabek et al., 2004; Ng and Zhang, 2002), we could use latency measurements instead of, or in addition to, hop counts. This would require simply changing the cost estimation part of the neural network (last two layers). While gathering latency measurements is expensive as it introduces traffic into the network, the ability of our approach to work with sparse data can limit the cost necessary to obtain the measurements.

Furthermore, AS membership information could help find more accurate representations as many ASes cover limited areas in the network and provide a coarse indication of locality (Eriksson et al., 2009). To add AS membership information, we could either extend the shape of our input vectors (by adding two bytes for AS number) or adapt the cost estimation layers to use AS data in estimating error, similarly to Eriksson *et al.* (Eriksson et al., 2009).

Applications. Building a model that accurately predicts structural properties of the Internet has several applications. Knowing the distance to remote IPs can help selecting a load balancing server or an overlay peer more efficiently and without having to perform expensive measurements. Understanding how nodes are clustered can make the transmission of video or large files faster by using close-by CDN nodes. DIP can be a passive defense mechanism against IP spoofing attacks, where malicious users change the source IP of attack packets to avoid identification and subvert

authentication. By comparing the predicted distance according to the spoofed source IP to the real distance (extracted from a packet's TTL field), one could verify whether the packet is spoofed or not (Jin et al., 2003). This idea inspires our next research work introduced in the next section (Learning IP Maps for Network Spoofing Detection).

Limitations and future work. Our current approach uses structural information embedded in the value of IP addresses, routing data, and distances between nodes, but does not consider the actual physical links between nodes on the Internet (*i.e.*, the Internet physical topology). Adding topology would further constrain the embedding, since it is well known that the Internet is not a metric space and latency or hop count distances cannot always be embedded in metric spaces (Dabek et al., 2004). We plan to extend our framework using graph embedding algorithms to take advantage of physical topology information (Wang et al., 2016a).

While our preliminary experiments focused on accuracy, the performance of building an embedding model is equally important. Training a model with 100,000 addresses and 95 servers on our 16-core GPU server takes a few hours, indicating that we may need to train models incrementally when resources are constrained (Bruzzone and Prieto, 1999). For example, in a live deployment, we envision reconstructing our model every few days to capture the changes in topology triggered by the dynamic Internet. We are currently studying ways to incrementally add or update models without rebuilding from scratch.

Because our data is sparse, not even the best embedding may be able to recover all structural properties. While we show that our results are reasonably accurate, even when we have less than 15% of all distances available, getting more data is clearly helpful (Eriksson et al., 2009). We plan to use active monitoring techniques (*e.g.* , *traceroute*) to collect more information for the training phase. Knowing the IPs and connectivity of routers in the network would make the training data set richer and constrain the representation of end-hosts further.

2.2 Learning IP Maps for Network Spoofing Detection

2.2.1 Introduction

Spoofing the source IP address of network packets is a mechanism frequently used in denial-of-service attacks (Chen et al., 2008; Spoofing-ps). IP spoofing can both hide the true source of the attack, thereby subverting IP-based firewalls and authentication mechanisms, and force legitimate hosts to redirect malicious traffic, thereby amplifying the effects of the attack (DDoS incident). In 2018 alone, IP spoofing was the vehicle for several high-profile, terabyte-size DDoS attacks (Cloudflare; DDoS incident).

There are two types of defenses against spoofing attacks. Active methods encrypt connections (Frankel and Krishnan, 2011), probe suspicious IPs (Feng et al., 2005), or mark legitimate packets with unique identifiers (Bremler-Barr and Levy, 2005; Yaar et al., 2006). They come at the expense of bandwidth, latency, or computation overhead (Shue et al., 2005). Passive (or map-based) defenses construct explicit offline maps between IPs and immutable structural network properties, such as paths (Park and Lee, 2001; Duan et al., 2008), neighboring networks (Ferguson and Senie, 2000; Killalea, 2000; Baker and Savola, 2004), or hop counts (Wang et al., 2007) and filter packets whose header information does not match the maps.

Passive map-based defenses have little detection overhead and are less intrusive than active defenses, but pose a significant construction cost. Measuring the network properties associated with each IP is a long and tedious process that requires intrusively probing other hosts, querying routing tables, or passively waiting to receive sufficient traffic (Jin et al.). In addition, as network properties are different for every vantage point, maps computed for a location cannot be easily transferred to a different location and must be recomputed.

We propose to *learn a structural model (or embedding) of the Internet and use the model to predict, rather than explicitly compute, IP maps*. Learning instead of measuring network properties could significantly reduce the cost of achieving complete host IP maps while making them more general. First, training a network embedding requires only a small set of ground truth information, such as distances between IPs (*e.g.* , hop counts, latencies) (Dabek et al., 2004; Lumezanu et al.,

2007; Eriksson et al., 2009). This reduces the amount of data necessary *a priori* for constructing a good map, thereby reducing the construction cost. Second, network embeddings are general and can be used to predict associations between *any* IP and its local network properties, even when the IP was not part of the training.

We study the feasibility of using network embeddings to learn IP maps for spoofing detection. Towards this goal, we present and build a learning-based spoofing detector that combines DIP, our previous deep learning based network embedding algorithm (Li et al., 2018a,b) and hop count filtering, and a popular host-based spoofing detection mechanism (Jin et al.). Hop count filtering compares the hop count information of packets arriving at a target server to the known hop count value from the packets' IP source to the target. It considers a packet spoofed if the two values do not match. Rather than build an explicit map of known IP-to-hop-count associations, we use DIP to learn a network embedding that preserves hop count distance between IPs. Using the embedding, we predict with high accuracy the hop count between any two IPs and identify when a packet is spoofed.

We analyze the coverage, accuracy, and cost of our learning-based detector. We show that learning, instead of explicitly computing IP maps, dramatically increases the detection coverage over spoofed sources and targets. Our learning-based detector needs hop count information from only around 1,000 IPs to several targets to build a model that can help detect spoofing from most of the Internet. Although, not as accurate as the original hop count filtering in detecting packets spoofed with previously known IPs (*i.e.*, to which the hop count is known), ours is the first map-based detector to identify spoofing with unknown IPs (*i.e.*, for which the hop count to the target is not known). It can also be used to detect spoofing to new targets, not part of the training process, with only a small loss in accuracy. In addition, learning an embedding is fast: even with almost three million hop counts from 100,000 IPs, it takes less than an hour to learn an embedding that can predict maps for any IP address.

This work brings two contributions. First, we introduce a framework to help *any* Internet server detect packets spoofed with *any* IP address without any additional measurements to that IP. This represents a major shift from previous map-based spoofing detection mechanisms, who are either restricted to specific targets or to detect packets spoofed with known IPs. Although our method

does not yet achieve by itself the accuracy necessary for a real-world deployment, it nevertheless shows that learning, rather than measuring, network properties could be an important piece in the spoofing detection puzzle.

Second, our work explores the impact of deep learning in the network and security operations decision making. As AI moves beyond just being the word *du jour* and increasingly becomes an integral part of the network, it is important to understand the trade-offs between its cost and benefits (Sommer and Paxson, 2010). It is precisely such a trade-off that we analyze in our work: while learning IP maps for spoofing detection can ease the task of administrators and help build complete maps much faster, it may decrease the detection accuracy. Understanding this balance can lead to better and more protected networks.

2.2.2 Towards Learned Maps for Spoofing Detection

Map-based spoofing defense mechanisms associate IP addresses with immutable network properties that are difficult to modify by attackers. Packets that carry information (*i.e.*, in headers) not matching the map are dropped. Unlike active spoofing defense methods that encrypt connections (Frankel and Krishnan, 2011) or probe suspicious IPs (Feng et al., 2005), map-based approaches are less intrusive and do not introduce additional traffic. To be effective, map-based detection must provide *coverage*: given a random attacker, protect any target against attack packets spoofed with any IP address.

Network maps. Network-based maps reside on AS border routers and pair IPs to the incoming network interface (Baker and Savola, 2004; Duan et al., 2008) or with keys inserted in the packet by an upstream trusted party (Bremler-Barr and Levy, 2005). Such maps have the potential to provide complete coverage, if installed pervasively by *all* ASes. However, their deployment has been slow. Recent measurements performed by CAIDA (Spoofing-state) show that around 25% of all ASes allow spoofed traffic to traverse them. Until all networks deploy network-based detection, the Internet is susceptible to spoofing. A likely cause for the incomplete coverage is the misaligned economic incentives (Lone et al., 2017): the cost to deploy network maps is high and supported by each AS, while the benefit is incremental and spread to the entire Internet.

Host maps. Host-based detection aligns the incentives by pushing the onus of detection towards the edge of the network. Hosts[3], rather than routers, build and maintain IP maps and benefit immediately from them. Most notably, several solutions build maps between IPs and hop count values to the target, relying on hop count data as a measure of the network topology (Wang et al., 2007; Jin et al.). Hop counts can be easily derived from the IP TTL field, whose value depends on the network infrastructure (*i.e.*, every router decrements the value before forwarding the packet) and cannot be easily forged by an attacker.

By pushing map computation to the edge, host-based maps decentralize the spoofing detection process and may see decreased coverage. First, their *detection is limited* to packets spoofed with IPs to which the target knows the hop count. They fail in detecting packets spoofed with IPs unknown[4] to the target. As attackers rely on surprise and obfuscation, they often use random spoofed addresses that can be from anywhere, even from non-routable ranges (Cloudflare). Increasing coverage by building complete maps is difficult. Passively collecting hop counts from incoming traffic is not likely to provide a large coverage (Jin et al.), while actively probing all IPs is expensive, intrusive, and time-consuming (Ark IPv4). Second, host maps are *location specific*. A map associating IPs with hop counts to a target is only useful for that target and cannot be transferred to other hosts, even if they are part of the same network.

Learned host maps. We propose to *learn* IP-to-hop-count mappings, instead of extracting them from incoming packets. A natural solution is to first learn representations of IPs in a vector space and then estimate the distance (*i.e.*, hop count) between IPs as the distance between their representations. Network embedding methods can easily compute IP representations in Euclidean spaces from a limited set of latency or hop count information (Dabek et al., 2004; Ng and Zhang, 2002; Costa et al., 2004; Pyxida; Eriksson et al., 2009). However, they are limited to the hosts part of the training set and cannot generate representations or hop counts for other hosts. This means that they cannot be used to estimate hop counts for, and thereby detect packets spoofed with, IP

[3]Throughout the chapter, we interchangeably refer to the server that deploys host-based maps as host, end-host, target, destination, or victim.

[4]We consider an IP address *unknown* to a target if the target does not know the hop count to it. Existing host-based maps cannot detect packets spoofed with an unknown IP address.

addresses unknown to a target.

The emergence of learning frameworks that preserve structural network properties such as distances or clustering among nodes offers an alternative to build explicit host-based IP maps from passive or active measurements (Li et al., 2018a; Wang et al., 2016b). Deep learning based embeddings use several data sets, *i.e.*, distances, routing information and host IP values, to learn IP representations. Because they learn hidden structural features encoded in a node's IP address, they could generate representations for any node, given only its IP address (Li et al., 2018a). In addition, deep learning algorithms are easily extensible and adaptable when new or updated data is available.

Figure 2.5 presents a visual representation of the benefits of learning-based spoofing detection. Unlike network maps, which detect only spoofed packets traversing protected ASes (left diagram), or explicit host maps, which detect only packets spoofed with IP sources known by the host (middle diagram), learned host maps can detect packets spoofed with unknown IPs (right diagram). The host uses a learned structural model of the Internet to estimate a specific network property, *e.g.* hop count information, associated with the source IP of an incoming packet and decide whether the packet is spoofed or not. In the next section, we describe a learning-based spoofing detector combining hop count filtering with the DIP network embedding framework.

2.2.3 Learning-Based Spoofing Detection

To illustrate the benefits of a learning-based spoofing detector, we propose a prototype detection framework based on the hop count filtering method first introduced by Wang *et al.* (Jin et al.). Our prototype consists of two main components: an offline IP map learning module (to learn IP representations and estimate hop counts between IPs) and a spoofing detector (to detect spoofed traffic). We describe both modules next and study their effectiveness in the next section.

2.2.3.1 IP Map Learning

In a previous work, we introduced DIP, a deep learning framework that uses hop count information between network host to learn an embedding of the Internet (Li et al., 2018a). Using a

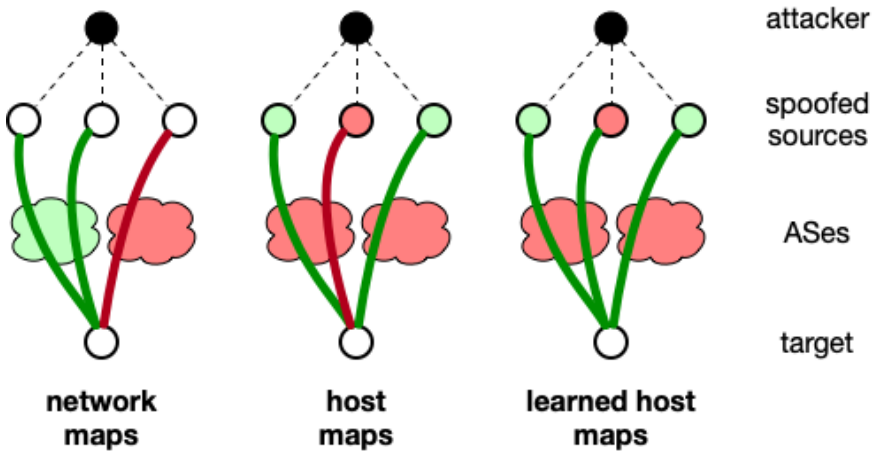


Figure 2.5: Map-based spoofing detection. Network maps detect only spoofed packets traversing protected ASes (colored in green). Host maps detect only packets spoofed with IPs present in the map (also colored in green). Learned host maps have the ability to detect all packets because they learn missing map entries. The paths in the diagrams indicate the apparent source of the packet (the spoofed source). In reality, all packets originate from the attacker.

small data set of IP addresses, their prefix information, and hop counts among them, DIP computes a vector representation of each IP in a high-dimensional Euclidean space. The algorithm ensures that representations reflect hop counts between hosts: the Euclidean distance between two representations estimates the hop count between the associated IPs. We briefly describe DIP below and refer the reader to the work of Li et al. (2018a,b) for more details.

The training data for DIP includes the IP addresses and hop counts among them. The IP addresses are hierarchical sequences which contain two parts, i.e. network (routable, routing prefix) and host (local) part. For example, for a 32-bit IPv4 address 192.167.2.17/24, the number 24 means the first 24 bits form the routable prefix, which can indicate the identification of the subnet. The rest 8 bits will be the local information for this IP address within the subnet. The hop count between two IPs represents the total number of intermediate devices (e.g. routers) that a data package needs to pass. We use it as the distance between two IPs and the training objective is to estimate the hop count between two IPs. The utilized hop count matrix is actually very sparse and asymmetric. During training, we will not pre-filling any value to the matrix to avoid introducing noises. We will measure the loss only when the hop count information exists.

The DIP is designed to exploit the above mentioned information efficiently with a variant of

recurrent neural networks to capture the structure encoded in the value of an IP address (*i.e.*, network part and host part provide an implicit hierarchy of IPs). According to the definition of the network and host part, we can find it is actually more significant to use the network part expressing the structural position. However, for IP addresses, their network parts can have different sizes. The network part's size can be any value from 0 to 32 (IPv4). Similarly, the number of bits contained in the host part varies. Hence, before feeding in the IP addresses, we will normalize them by: 1) Divide the 32 bits into network and host parts; 2) For the n bits in network parts, we pad in $32-n$ 0s at the end; 3) For the $32-n$ bits in the host parts, we pad in n 0s at the beginning. Then, we can get two four-byte normalized values.

With the two four-byte values, we can convert their each byte's numerical value to a one-hot vector (256 dimensions). The reason to use one-hot vector is we think there is no relationship like natural ordering or neighboring between different values of byte. For example, there are three hosts with the second byte equalling 156, 192, 15. We cannot easily say that the first two hosts are nearer to each other compared to the third because the difference between 156 and 192 is smaller. So we would like to regard the byte value as categorical data and use one-hot vector to represent.

By two experiments result shown in Figure 1 of Li et al. (2018a), we find the more bytes for an IP address, the better we can constrain the representation of the IP. This shows the sequential information contained by the IP addresses. So, in each layer of DIP, we feed the one-hot vector of the eight bytes from the normalized IP separately and use the structure similar to Recurrent Neural Network to capture the sequential information. Suppose the 8 bytes are $B_i, i \in 0,...,7$, where each B is actually converted to a one-hot vector of 256 dimensions. The input for the first layer of DIP will be B_0. Via the formula:

$$h_0 = activate(W_0 \times B_0 + bias_0) \tag{2.6}$$

we can get the hidden output of the first layer. The activate function can be either $softsign$, $tanh$ or $sigmoid$. In our experiments, we use $softsign$ for an easy and efficient training.

Then, for the following layers, the hidden output of the last layer will participate into the input

of the current layer.

$$h_i = activate(W_i \times [B_i, h_{i-1}] + bias_i), i \in 1, .., 7 \tag{2.7}$$

The $[B_i, h_{i-1}]$ means to concatenate the current input and the last hidden state. Then, by the h_7, we can add another feedforward layer to convert the h_7 to h_8 and estimate the hop count by:

$$hop - count = Euclidean(h_{8,source}, h_{8,destination}), \tag{2.8}$$

which utilizes the source and destination IPs' hidden representation for estimation. During learning, the trainable parameters include the W_i, $bias_i$ where $i \in 0, ..., 7$ and the final layer's parameters to convert h_7 to h_8. Since we think the first 4 bytes, which is the normalized network part of the IP address, are more significant for structural information. We will assign a higher initial value for the W_i when $i \in 0, ..., 3$ and a smaller value for the W_i when $i \in 4, .., 7$. This will enhance the influence of the first four layers and weaken the influence of the second four layers. The learning will be supervised: at each iteration of the algorithm, the representations are refined towards minimizing the error between estimated hop counts (computed from the Euclidean distance between two representations) and real hop counts. All the previous mentioned formulas and steps are simplified version for the processes described in Li et al. (2018a). If you need more detailed information, please check the Li et al. (2018a).

The advantage of DIP compared to traditional network embedding approaches is that it incorporates structural information encoded in the value of an IP address towards learning their position in a vector space. Learning from the value of an IP address is critical for our purpose. The learned model can be used to generate representations for *any IP*, even if it was not part of the training set or we do not have distance information to it. This, in turn, increases the coverage of our spoofing detection. We are not limited to the information from IP maps anymore; missing associations between IPs and hop counts can be easily estimated from the learned model. In addition, embedding models are general: they can be transferred to and immediately used by other hosts, without additional retraining, thereby reducing the overall cost of achieving global detection.

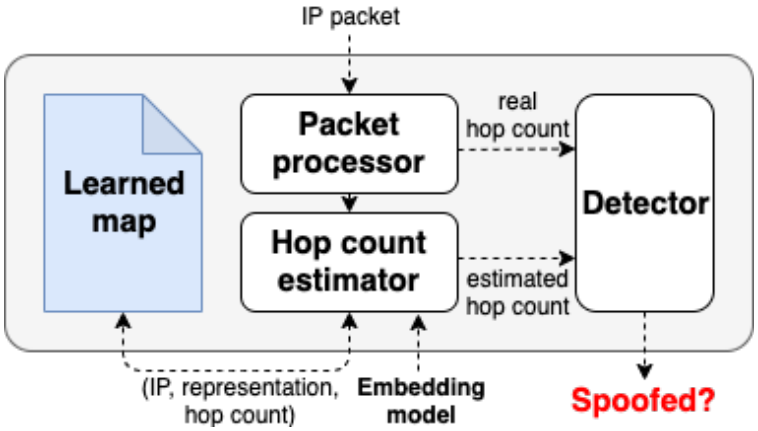


Figure 2.6: The learning-based spoofing detector uses an embedding of the Internet to estimate hop count information to any IP address and detect when the IP is used as the spoofed source of an attack packet.

Notwithstanding its advantages, DIP suffers from the same limitation as previous network embedding methods (Dabek et al., 2004; Ng and Zhang, 2002; Costa et al., 2004; Pyxida; Eriksson et al., 2009): Internet hop counts do not form a metric space (*e.g.* , they violate the triangle inequality (Lumezanu et al., 2007)) and cannot be embedded accurately into an Euclidean space. DIP is able to predict distances between Internet nodes with a mean absolute error of 2.06 hops, after being trained on distances between around 15,000 pairs of nodes. The error increase is small (from 2.06 hops to 2.29 hops) when we consider distances to IPs not used in the training, and therefore unknown to the targets. Detecting spoofed packets based on imprecise hop count information may lead to incorrect decisions. To minimize the number of missed spoofed packets, we must allow a margin of error when comparing the estimated and real hop counts. In Section 2.2.4, we show that spoofing detection still works when the hop count estimations are not exact and discuss how to select the detection threshold.

2.2.3.2 Spoofing Detector

The spoofing detector is deployed on a potential target and consists of the learned map, packet processor, hop count extractor, and detector (Figure 2.6). The packet processor extracts the source IP and TTL values from an incoming packet. It then computes the hop count information from the TTL, using the algorithm described by Wang *et al.* (Wang et al., 2007). The hop count estimator

Method	Description
network maps	Complete coverage if all ASes deploy network maps; currently 25% of ASes are unprotected (Spoofing-state); cost of building maps is high (Lone et al., 2017).
host maps	Coverage limited to the IPs with known hop count to the host: potentially covering packets spoofed with 95% of all IPs (Jin et al.), if map contains *all* IP addresses; maps cannot be transferred and need to be explicitly computed by each host; cost of building maps is high (Jin et al.; Ark IPv4).
learned host maps	Limited by the accuracy of embedding; covering packets spoofed with around 70% of all IPs, even when the IP is unknown to the host; embedding models can be transferred and do not need to retrained; cost of building maps is low.

Table 2.2: Map-based spoofing detection methods.

uses an existing embedding model (*i.e.*, built offline using IP and hop count data from legitimate packets) to compute a representation for the packet's source IP and estimate the hop count to the target. The detector compares the real and estimated hop counts to the target. If the difference between them is greater that a pre-specified threshold, we consider the packet spoofed and raise an alert. To save future computation, all estimated hop counts and IP representations are saved into a cached learned map. When packets from an IP present in the map arrive at the target, we use information from the map without recomputing the hop count value.

2.2.4 Evaluation

We evaluate the feasibility of learning-based spoofing detection as follows. First, we analyze what coverage we can achieve compared to detectors based on exact maps, such as hop count filtering. Second, we simulate spoofing and measure how well we detect spoofed packets under various scenarios. Finally, we look at the cost of building a detector. Table 2.2 contains a summary of our findings.

Data. We use a large data set of network hop counts collected by the Ark project (Ark IPv4) during June 2015. The data contains hop count information from 96 geographically distributed servers to ten million IP addresses that cover all routable prefixes in the Internet. Because moni-

toring a large number of IPs is costly, not all servers have hop counts for all ten million IPs. Our hop count matrix is incomplete and contains entries for only 29% of the pairs. Furthermore, due to privacy issues, the last eight bits of each IP were zeroed out after collection. Although this makes the set of possible spoofed sources more general, it also helps us learn more robust models, as there are likely more servers monitoring the same /24 prefix than the same single IP address, *i.e.*, the number of hop counts per IP is higher.

It is possible that the hop counts collected by Ark are to routers on the path, rather than the end host itself. This can happen when routers, and not the target, reply to the monitoring packets. As our detection algorithm is intended for the edge of the network, training with router IPs and hop counts may bias the results. It is difficult to determine precisely which response comes from a router. To minimize the potential bias, we use a simple heuristic based on the fact that many router IPs are not part of advertised BGP prefixes. Using this heuristic, we filter out around 1.2% of the IPs in our data.

Methodology. We build a prototype spoofing detector using TensorFlow and Python, reimplementing the learning in DIP (Li et al., 2018a) and hop count filtering (Jin et al.). We train the neural network offline using several smaller data sets obtained by randomly sampling 1,000, 10,000, and 100,000 IPs from the original data and keeping only the hop counts to them. Sampling increases the sparsity of the data: less than 15% of the entries in the smaller data sets are valid. We train our spoofing detector for 1,000 iterations on a server with four 3.5GHz quad-core Intel Xeon processors and 128GB of RAM.

Assumptions. We assume a single attacker, randomly placed in the Internet, with no knowledge about network structure around the spoofing target or the spoofed IP. This means that the attacker cannot use information about hop count to the target to avoid detection. We also assume that the hop count information is static and does not change. In Section 2.2.5, we discuss what happens when we relax these assumptions.

Goals and non-goals. Our main goal is to evaluate the feasibility of learned maps for detecting spoofed network packets and to inform future decisions of security operators or researchers. We use an existing embedding algorithm, DIP (Li et al., 2018a) to learn hop count information between

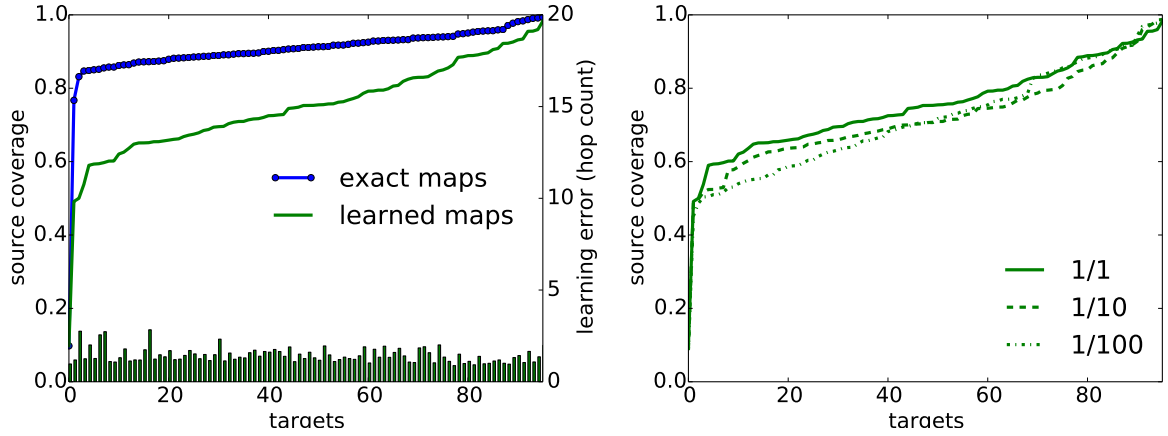


Figure 2.7: Coverage for (left) exact and learned maps for 1,000 source IP addresses, and (right) learned maps for the same sources used in training (labeled "1/1"), ten times as many sources ("1/10"), and a hundred times as many sources ("1/100"). The bars represent the error of a learning-based spoofing detector, for each target, in hop counts. Learning-based spoofing detectors adapt well to new sources with little loss of coverage.

IPs, and study its interaction with hop count filtering (Jin et al.). A non-goal is to evaluate the DIP embedding accuracy. For a detailed study, we refer the reader to the DIP chapter (Li et al., 2018a), which analyzes the embedding accuracy when varying various learning parameters and in contrast to other distance prediction mechanisms.

2.2.4.1 Coverage

The limited range of hop count values (*i.e.*, 0 to 255) means that no hop-count-based spoofing detector can achieve perfect detection. To understand the limits of detection, we define the *coverage* of each target as the expected fraction of source IPs that can be unambiguously identified when spoofed by a random attacker. In other words, the coverage represents the probability that a packet spoofed with a random IP by a random attacker is detected by the target.

Exact maps. Before studying learning-based maps, we analyze the boundaries of what exact hop count based maps can achieve. Hop count filtering (Jin et al.) identifies packets as spoofed when the packet hop count does not match the known hop count associated with the source address. If the attacker spoofs a packet with an IP address that has the same hop count to the target as the

attacker IP, hop count filtering cannot detect the attack. To compute the coverage of exact maps, we repeatedly select a source IP at random as the attacker for each target, calculate the fraction of IPs that share the same hop count to the target as the attacker and subtract the value from 1. Figure 2.7(left) shows the distribution of coverage for all targets in our data set. We consider 1,000 IPs; results for other data sets are similar. Focus only on the line labeled *exact maps* for now. Most targets have a coverage of at least 0.8.

An important observation is that exact maps are not necessarily limited to detecting traffic spoofed with IPs that have communicated with the target in the past. They are efficient in detecting attacks spoofed with any IP, as long as the target knows the hop count to the IP. Oftentimes, IPs that are part of the same prefix, especially for smaller prefixes, share the same hop count to a target. Our data set already contains at most one IP per /24 prefix to ensure it does not underestimate the performance of exact maps.

Learned maps. As described in Section 2.2.3, learning maps introduces errors in estimating hop count values. Any learning-based spoofing detector needs to account for those errors in identifying spoofing. First, for each target, we compute the average error between real and estimated hop counts from all sources to the target. We use hop counts from 1,000 IPs to the 96 servers to learn the embedding. Then, similarly to above, we select an IP at random from the 1,000 IPs in the training set as the attacker and compute the fraction of IPs whose hop count is within the estimation error of the hop count from attacker to target. We subtract this value from 1 to obtain the coverage for the learned map of each target. This captures the probability that learned maps detect packets spoofed with a random IP from a random attacker. Note that here we select an IP *from the training set* as the attacker to compute source coverage for known IPs and compare with exact maps. Below, we re-do the analysis for unknown IPs, that are not from the training set. Figure 2.7(left) shows that the coverage of learned maps is lower than that of exact maps when considering the same set of source IPs. On the average, a target could detect spoofed packets only 70% of the time, compared to 90% of the time with exact maps.

The power of learned maps lies in estimating hop counts to IPs that are *not* in the training set. We increase the number of source IPs to which we estimate hop counts to 10,000 and 100,000 (10, respectively 100 times more than the training set) and show the coverage in Figure 2.7(right). Even

when we increase the number of IPs 10- or 100-fold (lines labeled "1/10", respectively "1/100") compared to the training set, the coverage does not decrease much. In comparison, the coverage of exact maps for any unknown IPs *is 0*. Thus, learned maps offer an immense benefit when protecting against spoofing carried with IPs not known to the target.

Summary. IP-to-hop-count maps learned via deep learning embedding are less accurate than exact maps extracted from incoming traffic, when evaluated on IPs to which the host knows the hop count. However, learned maps dramatically outperform exact maps on unknown IPs.

2.2.4.2 Accuracy

We define the sensitivity of the detector as the fraction of detected spoofed packets out of all spoofed packets, and the specificity as the fraction of correctly identified legitimate packets out of all legitimate packets. Sensitivity (also known as recall or true positive rate) represents the probability of detecting a spoofed packet and specificity (also known as selectivity or true negative rate) captures the probability of not raising a false alert. Unlike coverage, which gives a measure of how well the detector can do for a randomly spoofed IP at any target, sensitivity and specificity measure the accuracy of the detector in a realistic scenario, given both legitimate and spoofed packets.

We set up the experiment as follows. We use the 10,000 IP data set for both training and detection. We simulate packets arriving from each of the 10,000 IPs at random and introduce spoofing with an average rate of 0.01 (one spoofed packet for every 100 packets). For each spoofed packet, we generate a fake hop count value at random from a normal distribution with a mean of 15 and standard deviation of 5. We chose these values as they match the distribution of hop counts in our data set. We measure sensitivity and specificity as averages across all targets in a specific experiment.

Varying threshold. Figure 2.8 (left) shows the specificity and sensitivity as we vary the detection threshold. For a threshold of two hops, which is the average training estimation error, we obtain an overall sensitivity of around 0.75, which is consistent with the expected coverage from Figure 2.7 (left). However, a high sensitivity also results in a low specificity: while we detect most

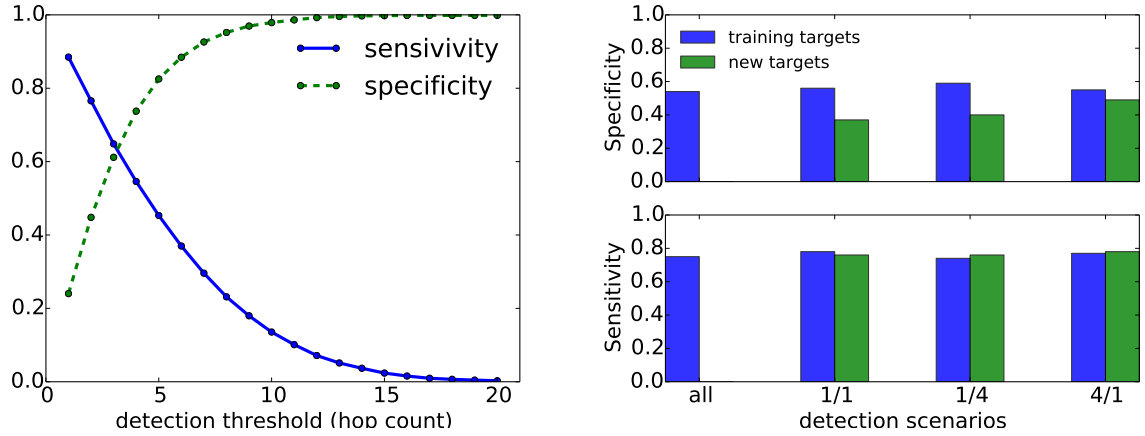


Figure 2.8: Detector accuracy under various scenarios. (left) We run detection on the same targets used in training and vary the detection threshold; increasing the threshold reduces sensitivity and improves specificity; (right) We perform detection using both training targets and new targets not used in the training process; we set the detection threshold to two hops (the average estimation error of the model); as we vary the ratio between training and detection targets, the sensitivity for the new targets is comparable to that of the training targets, while the specificity is lower; "all" indicates that we perform training and detection on all targets, therefore there are no new targets.

spoofed packets, we do so at the expense of tagging the majority of legitimate packets as spoofed. Modifying the detection threshold adjusts the trade-off: we find that the best balance between sensitivity and specificity is when the threshold is just above three hops.

Summary. Learning maps can trade-off high accuracy in detecting spoofed traffic with a high rate of false positives. This indicates that learning maps could be an important part of a bigger framework for spoofing detection, which could process the false alarms faster or dynamically adjust the detection threshold to control their rate.

2.2.4.3 Model Transfer

To understand whether our learned model is general and can be transferred to new hosts, we split the data set into two disjoint sets, such that the targets in each set are disjoint. We train a model using only the targets in the first set and then run detection on both sets. Our goal is to understand how the detector performance for a target changes when we use a model learned for other targets. We consider three different splits: the number of targets in each set is the same ("1/1" split), the

number of training targets is four times larger ("4/1" split) and, the training targets are four times fewer ("1/4" split). We set the detection threshold at two hops and compute the average sensitivity and specificity for spoofing detection on each set across all splits.

Figure 2.8 (right) presents the results. The sensitivity of targets not part of the training process is comparable to that of those targets used in training, showing that our models are transferable to new targets with little loss in accuracy. This reduces the cost of global deployment as new targets do not need to gather training data and build their own models. An interesting observation from Figure 2.8 is that the specificity of new targets is lower than for training targets: raising false alerts is more likely on hosts not part of the training.

Summary. Maps learned at a single location are transferable to other servers in the Internet with little loss in accuracy. This is because maps are learned based on structural network properties that are the same from every vantage point and do not depend on specific locations.

2.2.4.4 Performance

Training a representative embedding does not require significant resources. Learning a model with hop counts from 100,000 IPs to 96 servers takes about one hour on a server with four 3.5GHz quad-core Intel Xeon processors and 128GB of RAM (Table 2.3). Each model takes less than 5MB of memory and thus is easily transferable over the network. Detection is fast as well. Using the same machine used for training, and given an IP packet, we are able to decide whether the packet is spoofed or not in under a millisecond. While running the detector at line speed on a network gateway instead of a powerful GPU machine may reduce these numbers, recent deep learning platforms, such as Net2Vec (Gonzalez et al., 2017), that are able to capture and process packets at 60Gbps can make deployment easier and faster.

2.2.5 Discussion: Limitations and Opportunities

Accuracy trade-off. Our detector trades off accuracy for generality. Low specificity may be unacceptable for many operators, who would still need to sift through alerts to identify legitimate packets incorrectly identified as spoofed. In practice, we envision that our approach works in

Number of source IPs	1,000	10,000	100,000
Time to train (min)	1	6	58
Size of model (MB)	4.5	4.5	4.5

Table 2.3: Performance of training for data sets containing hop count information from 96 servers to 1,000, 10,000, and 100,000 IPs. Each data set is sparse, with only about 15% of all entries available.

conjunction with network-based maps and host maps built from passive measurements to offer a comprehensive and cost-efficient spoofing detection solution.

Non-metric Internet. Paths between Internet hosts are dictated by routing policies and AS peering agreements, and are not always shortest in terms of hop counts. Estimating the hop count between two hosts as the Euclidean distance between their embeddings does not capture the intricacies of Internet routing and may introduce errors. We are working on reducing these errors in two ways. First, we plan to introduce AS membership as an additional data set in the deep learning framework. Knowing to which AS a host belongs to can help constrain its possible positions in the embedding space. Second, we plan to add IP addresses and hop counts to routers on the path between a source and a destination. While a perfect metric embedding of Internet hop counts is impossible (Lumezanu et al., 2007), we can reduce the learning error and better predict anomalies.

Furthermore, we are exploring non-metric embeddings. It is possible to learn structural embeddings of the Internet that estimate other network properties helpful in spoofing detection. For example, graph embeddings preserve local and remote network structure features such as first- and second-order neighbors (Wang et al., 2016b). We are investigating how to use such mechanisms to devise spoofing detectors that avoid the shortcomings of hop based spoofing detection.

Dynamic Internet. The dynamic nature of the Internet with frequent misconfigurations, outages, or policy changes means that hop counts or IP addresses may also change (Cunha et al., 2011; Padmanabhan et al., 2016). If this happens, we may need to retrain the embedding model to reflect the updated values. A practical deployment would passively listen to incoming legitimate traffic, or selectively probe IPs from learned maps, and update existing maps to reflect new hop count values. We are currently working on how to re-train our model to learn a more accurate hop count estimation in the presence of new hop count data.

Asymmetric Internet. Internet routing may not always be symmetric (John et al., 2010). Due to ISP routing practices, the direct and reverse route between two nodes may be different, resulting in potentially different hop count measurements between the same pair of nodes. Because we compute hop count information from TTL values decremented exclusively by routers on the *direct* path between source IPs and targets, our detection is not impaired by the routing asymmetry.

Bootstrapping. Learning a representative structural model of the Internet requires hop count measurements from many vantage points. Not all enterprises have access to many geographically distributed monitors to collect hop count data for the initial model training. There are two possibilities to generate learned maps in such scenarios. First, one could use third-party measurement data, such as that provided by CAIDA (Ark IPv4), to train a descriptive model and generate learned maps. Second, an enterprise could deploy a model already trained in another location. As we showed in Section 2.2.4.3, our models are transferable with little loss in accuracy.

Attack types. We assumed a spoofing attack carried by a random attacker without any knowledge of the network topology. In reality, some attackers may be able to obtain information about the network that could help subvert the defense. For example, if an attacker controls multiple bots, it can send the attack from the bots that have a more popular hop count value to the target. This maximizes the likelihood of passing by our filter as there are more IP addresses with which to spoof the source. Furthermore, if the attacker learns the hop count between the spoofed IP and the target, it can also spoof the TTL field of the attack packets, inserting a value that corresponds to the learned hop count. In this case, no hop count filtering can detect the attack. To learn the hop count between two arbitrary hosts, one can use DIP (Li et al., 2018a), the same framework we employ, or the algorithm described by Barford *et al.* (Eriksson et al., 2009). However, both mechanisms require coordination and control of multiple geographically distributed servers, available only to more complex attackers.

2.3 Conclusion

We used deep learning to learn vector representations for nodes in the Internet based on their IP address, routing information, and a sparse hop count distance matrix. Deep learning helps uncover

hidden features in the input data and recover structural properties of the Internet, such as node clusters or distances between nodes. Our experiments on a large real-world data set show that our embeddings can recover most distances, even to arbitrary hosts, with two hops absolute error, even when the training data is sparse. Using the deep learning framework in network spoofing detection, we found the learning-based network coordinate system (DIP) could dramatically increase the detection coverage. For any Internet server's received packets spoofed with any IP address, the DIP-based spoofing detection mechanism could determine whether the packets are spoofed or not with acceptable accuracy.

CHAPTER 3

Extracting Latent Information from Unused Speech Interpretations

In this chapter, we describe the second to n-th best interpretations of a speech, the unused data in spoken language understanding (SLU) domain by introducing (1) reasons why the second to n-th best interpretations of a speech could help understanding, (2) approaches to include those interpretations in SLU systems and (3) downstream tasks which could benefit from incorporating those interpretations.

In a modern SLU system, the natural language understanding (NLU) module takes interpretations of a speech from the automatic speech recognition (ASR) module as the input. The NLU module usually uses the first best interpretation of a given speech in downstream tasks such as domain and intent classification. However, the ASR module might misrecognize some speeches and the first best interpretation could be erroneous and noisy. Solely relying on the first best interpretation could make the performance of downstream tasks non-optimal especially when the ASR model does not perform well. To address this issue, we introduce a series of simple yet efficient models in Li et al. (Li et al., 2020) for improving the understanding of semantics of the input speeches by collectively exploiting the n-best speech interpretations from the ASR module.

3.1 Introduction

Currently, voice-controlled smart devices are widely used in multiple areas to fulfill various tasks, e.g. playing music, acquiring weather information and booking tickets. The SLU system employs several modules to enable the understanding of the semantics of the input speeches. When there is an incoming speech, the ASR module picks it up and attempts to transcribe the speech.

An ASR model could generate multiple interpretations for most speeches, which can be ranked by their associated confidence scores. Among the n-best hypotheses, the top-1 hypothesis is usually transformed to the NLU module for downstream tasks such as domain classification, intent classification and named entity recognition (slot tagging). Multi-domain NLU modules are usually designed hierarchically (Tur and De Mori, 2011). For one incoming utterance, NLU modules will firstly classify the utterance as one of many possible domains and the further analysis on intent classification and slot tagging will be domain-specific.

In spite of impressive development on the current SLU pipeline, the interpretation of speech could still contain errors. Sometimes the top-1 recognition hypothesis of ASR module is ungrammatical or implausible and far from the ground-truth transcription (Peng et al., 2013; Jyothi et al., 2012). Among those cases, we find one interpretation exact matching with or more similar to transcription can be included in the remaining hypotheses ($2^{nd} - n^{th}$).

To illustrate the value of the $2^{nd} - n^{th}$ hypotheses, we count the frequency of exact matching and more similar (smaller edit distance compared to the 1^{st} hypothesis) to transcription for different positions of the n-best hypotheses list. Table 3.1 exhibits the results. For the explored dataset, we only collect the top 5 interpretations for each utterance ($n = 5$). Notably, when the correct recognition exists among the 5 best hypotheses, 50% of the time (sum of the first row's percentages) it occurs among the $2^{nd} - 5^{th}$ positions. Moreover, as shown by the second row in Table 3.1, compared to the top recognition hypothesis, the other hypotheses can sometimes be more similar to the transcription.

Table 3.1: Spoken recognition quality distribution of the n best hypotheses.

n Best Rank Position	2^{nd}	3^{rd}	4^{th}	5^{th}
Match	19%	14%	10%	7%
Prob (better than 1^{st} best)	22%	17%	16%	15%

Over the past few years, we have observed the success of reranking the n-best hypotheses (Peng et al., 2013; Charniak and Johnson, 2005; Morbini et al., 2012; Dikici et al., 2012; Sak et al., 2011b, 2010, 2011a; Collins et al., 2005; Chan and Woodland, 2004) before feeding the best interpretation to the NLU module. These approaches propose the reranking framework by involving morphological, lexical or syntactic features (Sak et al., 2011a; Collins et al., 2005; Chan

Table 3.2: Motivating example: comparison of ASR *n*-Best hypotheses with the corresponding transcription.

Transcription	1^{st} best	2^{nd} best	3^{rd} best
play muse	play news	**play muse**	play mus
track on bose	check on bowls	check on bose	**track on bose**
harry porter	how **porter**	how patter	**harry** power

and Woodland, 2004), speech recognition features like confidence score (Peng et al., 2013; Morbini et al., 2012), and other features like number of tokens, rank position (Peng et al., 2013). They are effective to select the best from the hypotheses list and reduce the word error rate (WER) (Oba et al., 2007) of speech recognition.

Those reranking models could benefit the first two cases in Table 3.2 when there is an utterance matching with transcription. However, in other cases like the third row, it is hard to integrate the fragmented information in multiple hypotheses.

This chapter proposes various methods integrating *n*-best hypotheses to tackle the problem. To the best of our knowledge, this is the first study that attempts to collectively exploit the *n*-best speech interpretations in the SLU system. This chapter serves as the basis of our *n*-best-hypotheses-based SLU system, focusing on the methods of integration for the hypotheses. Since further improvements of the integration framework require considerable setup and descriptions, where jointly optimized tasks (e.g. transcription reconstruction) trained with multiple ways (multitask (Caruana, 1997), multistage learning (Gong et al., 2013)) and more features (confidence score, rank position, etc.) are involved, we leave those to a subsequent article.

This chapter is organized as follows. Section 3.2 introduces the Baseline, Oracle and Direct models. Section 3.3 describes proposed ways to integrate *n*-best hypotheses during training. The experimental setup and results are described in Section 3.4. Section 3.5 contains conclusions and future work.

3.2 Baseline, Oracle and Direct Models

3.2.1 Baseline and Oracle

The preliminary architecture is shown in Fig. 3.1. For a given transcribed utterance, it is firstly encoded with Byte Pair Encoding (BPE) (Sennrich et al., 2015), a compression algorithm splitting words to fundamental subword units (*pairs of bytes* or *BP*s) and reducing the embedded vocabulary size. Then we use a BiLSTM (Schuster and Paliwal, 1997) encoder and the output state of the BiLSTM is regarded as a vector representation for this utterance. Finally, a fully connected Feed-forward Neural Network (FNN) followed by a softmax layer, labeled as a multilayer perceptron (MLP) module, is used to perform the domain/intent classification task based on the vector.

Figure 3.1: Baseline pipeline for domain or intent classification.

For convenience, we simplify the whole process in Fig.3.1 as a mapping *BM* (Baseline Mapping) from the input utterance S to an estimated tag's probability $p(\tilde{t})$, where $p(\tilde{t}) \leftarrow BM(S)$. The *Baseline* is trained on transcription and evaluated on ASR 1^{st} best hypothesis ($S =$ ASR 1^{st} best). The *Oracle* is trained on transcription and evaluated on transcription ($S =$ Transcription). We name it Oracle simply because we assume that hypotheses are noisy versions of transcription.

3.2.2 Direct Models

Besides the Baseline and Oracle, where only ASR 1-best[1] hypothesis is considered, we also perform experiments to utilize ASR n-best hypotheses during evaluation. The models evaluating with n-bests and a BM (pre-trained on transcription) are called *Direct Models* (in Fig. 3.2):

- *Majority Vote.* We apply the BM model on each hypothesis independently and combine the predictions by picking the majority predicted label, i.e. Music.

[1] We use ASR *n-best hypotheses* or n-bests to denote the top n interpretations of a speech, and the *1,5-best* standing for the top 1 or 5 hypotheses.

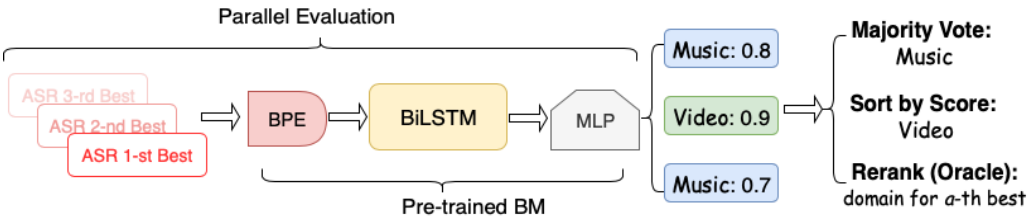


Figure 3.2: Direct models evaluation pipeline.

- *Sort by Score.* After parallel evaluation on all hypotheses, sort the prediction by the corresponding confidence score and choose the one with the highest score, i.e. Video.
- *Rerank (Oracle).* Since the current rerank models (e.g., (Peng et al., 2013; Charniak and Johnson, 2005; Morbini et al., 2012)) attempt to select the hypothesis most similar to transcription, we propose the Rerank (Oracle), which picks the hypothesis with the smallest edit distance to transcription (assume it is the a-th best) during evaluation and uses its corresponding prediction.

3.3 Integration of N-Best Hypotheses

All the above mentioned models apply the BM trained on one interpretation (transcription). Their abilities to take advantage of multiple interpretations are actually not trained. As a further step, we propose multiple ways to integrate the n-best hypotheses during training. The explored methods can be divided into two groups as shown in Fig. 3.3. Let $H_1, H_2, ..., H_n$ denote all the hypotheses from ASR and $bp_{H_k,i} \in BPs$ denotes the i-th pair of bytes (BP) in the k^{th} best hypothesis. The model parameters associated with the two possible ways both contain: embedding e_{bp} for pairs of bytes, BiLSTM parameters θ and MLP parameters W, b.

3.3.1 Hypothesized Text Concatenation

The basic integration method (*Combined Sentence*) concatenates the n-best hypothesized text. We separate hypotheses with a special delimiter (<SEP>). We assume BPE totally produces m

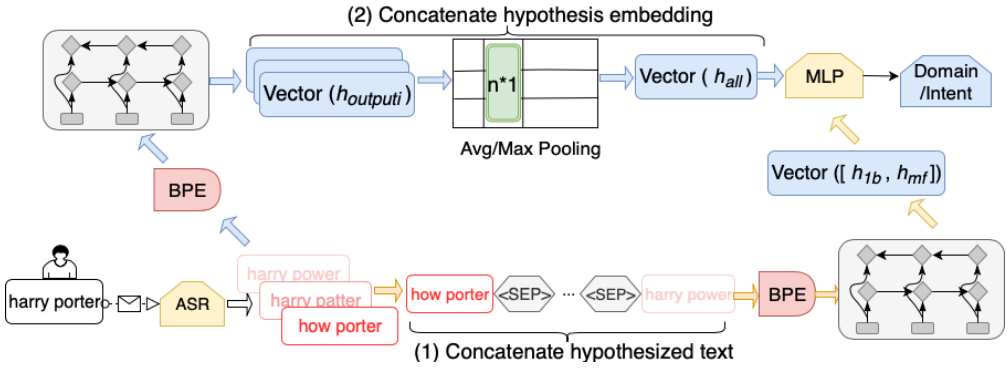


Figure 3.3: Integration of n-best hypotheses with two possible ways: 1) concatenate hypothesized text and 2) concatenate hypothesis embedding.

BPs (delimiters are not split during encoding). Suppose the n^{th} hypothesis has j pairs. The entire model can be formulated as:

$$(h_1, ..., h_m) \leftarrow BiLSTM_\theta(bp_{H_1,1}, ..., bp_{<sep>}, ..., bp_{H_n,j}) \tag{3.1}$$

$$p(\tilde{t}) = \sigma(W[h_{1b}, h_{mf}] + b) \tag{3.2}$$

In Eqn. 3.1, the connected hypotheses and separators are encoded via BiLSTM to a sequence of hidden state vectors. Each hidden state vector, e.g. h_1, is the concatenation of forward h_{1f} and backward h_{1b} states. The concatenation of the last state of the forward and backward LSTM forms the output vector of BiLSTM (concatenation denoted as $[,]$). Then, in Eqn. 3.2, the MLP module defines the probability of a specific tag (domain or intent) $\tilde{t}$ as the normalized activation (σ) output after linear transformation of the output vector.

3.3.2 Hypothesis Embedding Concatenation

The concatenation of hypothesized text leverages the n-best list by transferring information among hypotheses in an embedding framework, BiLSTM. However, since all the layers have access to both the preceding and subsequent information, the embedding among n-bests will influence each other, which confuses the embedding and makes the whole framework sensitive to the noise

in hypotheses.

As the second group of integration approaches, we develop models, *PoolingAvg/Max*, on the concatenation of hypothesis embedding, which isolate the embedding process among hypotheses and summarize the features by a pooling layer. For each hypothesis (e.g., i^{th} best in Eqn. 3.3 with j pairs of bytes), we could get a sequence of hidden states from BiLSTM and obtain its final output state by concatenating the first and last hidden state (h_{output_i} in Eqn. 3.4). Then, we stack all the output states vertically as shown in Eqn. 3.5. Note that in the real data, we will not always have a fixed size of hypotheses list. For a list with r ($< n$) interpretations, we get the embedding for each of them and pad with the embedding of the first best hypothesis until a fixed size n. When $r \geq n$, we only stack the top n embeddings. We employ h_{output_1} for padding to enhance the influence of the top 1 hypothesis, which is more reliable. Finally, one unified representation could be achieved via *Pooling* (Max/Avg pooling with n by 1 sliding window and stride 1) on the concatenation and one score could be produced per possible tag for the given task.

$$(h_{H_i,1},...,h_{H_i,j}) \leftarrow BiLSTM_\theta(bp_{H_i,1},...,bp_{H_i,j}) \tag{3.3}$$

$$h_{output_i} = [h_{H_i,1b},\ h_{H_i,jf}] \tag{3.4}$$

$$h_{outputs} = \left\{ \begin{array}{ll} \left\{ \begin{array}{l} h_{output_1} \\ ... \\ h_{output_r} \end{array} \right\} & r-bests \\ \left\{ \begin{array}{l} h_{output_1} \\ ... \end{array} \right\} & Padding\ with\ h_{output_1} \end{array} \right\} \tag{3.5}$$

$$h_{all} = Pooling(h_{outputs}) \tag{3.6}$$

$$p(\tilde{t}) = \sigma(Wh_{all} + b) \tag{3.7}$$

3.4 Experiments

3.4.1 Dataset

We conduct our experiments on $\sim$ 8.7M annotated anonymised user utterances. They are annotated and derived from requests across 23 domains.

3.4.2 Performance on Entire Test Set

Table 3.3 shows the relative error reduction (RErr)[2] of Baseline, Oracle and our proposed models on the entire test set ($\sim$ 300K utterances) for multi-class domain classification. We can see among all the direct methods, predicting based on the hypothesis most similar to the transcription (Rerank (Oracle)) is the best.

Table 3.3: Micro and Macro F1 score for multi-class domain classification.

Category	Model	RErr(%)
Baseline		**0.00**
Integration	PoolingAvg	**14.29**
	PoolingMax	13.20
	Combined Sentence	11.67
Direct	Sort by Score	1.85
	Majority Vote	1.64
	Rerank (Oracle)	**3.71**
Oracle		27.04

As for the other models attempting to integrate the *n*-bests during training, PoolingAvg gets the highest relative improvement, 14.29%. It as well turns out that all the integration methods outperform direct models drastically. This shows that having access to *n*-best hypotheses during training is crucial for the quality of the predicted semantics.

Table 3.4: Performance comparison for the subset ($\sim$ 19%) where ASR first best disagrees with transcription.

Category	Model	RErr(%)
Baseline		**0.00**
Integration	PoolingAvg	24.67
	PoolingMax	26.23
	Combined Sentence	19.23
Direct	Sort by Score	9.95
	Majority Vote	7.59
	Rerank (Oracle)	7.25
Oracle		53.02

Table 3.5: Performance comparison for the subset ($\sim$ 81%) where ASR first best agrees with transcription.

Category	Model	RErr(%)
Baseline		**0.00**
Integration	PoolingAvg	3.56
	PoolingMax	-0.38
	Combined Sentence	4.50
Direct	Sort by Score	-8.269
	Majority Vote	-3.19
	Rerank (Oracle)	0.00
Oracle		0.00

3.4.3 Performance Comparison among Various Subsets

To further detect the reason for improvements, we split the test set into two parts based on whether ASR first best agrees with transcription (ignore difference with wake words of hypotheses) and evaluate separately. Comparing Table 3.4 and Table 3.5, obviously the benefits of using multiple hypotheses are mainly gained when ASR 1^{st} best disagrees with the transcription. When ASR 1^{st} best agrees with transcription, the proposed integration models can also keep the performance. Under that condition, we can still improve a little (3.56%) because, by introducing multiple ASR hypotheses, we could have more information and when the transcription/ASR 1^{st} best does

[2]The RErr for a model m is calculated by comparing the relative difference between $100\% - MicroF1_m$ and $100\% - MicroF1_{\text{Baseline}}$.

not appear in the training set's transcriptions, its n-bests list may have similar hypotheses included in the training set's n-bests. Then, our integration model trained on n-best hypotheses as well has clue to predict. The series of comparisons reveal that our approaches integrating the hypotheses are robust to the ASR errors and whenever the ASR model makes mistakes, we can outperform more significantly.

3.4.4 Improvements on Different Domains and Different Numbers of Hypotheses

Among all the 23 domains, we choose 8 popular domains for further comparisons between the Baseline and the best model of Table 3.3, PoolingAvg. Fig. 3.4 exhibits the results. PoolingAvg consistently improves the accuracy for all 8 domains.

In the previous experiments, the number of utilized hypotheses for each utterance during evaluation is five, which means we use the top 5 interpretations when the size of ASR recognition list is not smaller than 5 and use all the interpretations otherwise. Changing the number of hypotheses while evaluation, Fig. 3.5 shows a monotonic increase with the access to more hypotheses for the PoolingAvg and PoolingMax (Sort by Score is shown because it is the best achievable direct model while the Rerank (Oracle) is not realistic). The growth becomes gentle after four hypotheses are leveraged.

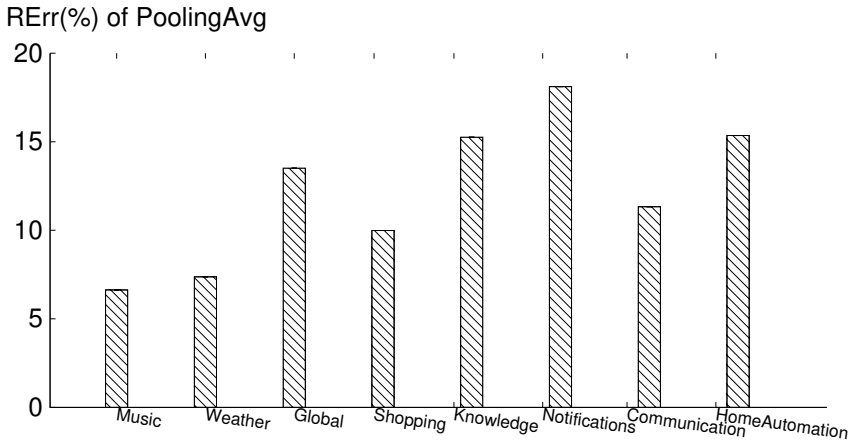


Figure 3.4: Improvements on important domains.

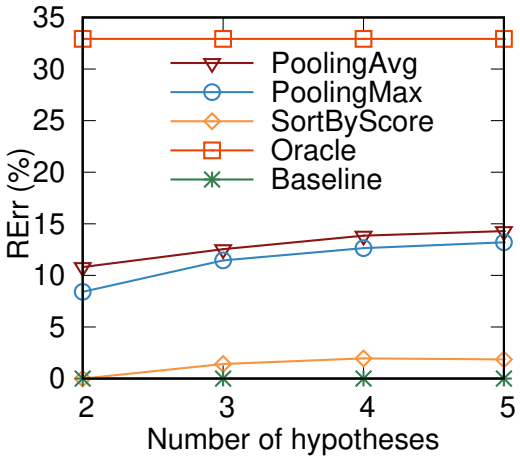


Figure 3.5: The influence of different amount of hypotheses.

3.4.5 Intent Classification

Table 3.6: Intent classification for three important domains.

Domain	Metric	Shopping	Knowledge	Communication
Baseline	RErr (%)	0.0	0.0	0.0
Oracle		47.63	40.28	32.89
PoolingAvg		**25.55**	**25.00**	**11.92**

Since another downstream task, intent classification, is similar to domain classification, we just show the best model in domain classification, PoolingAvg, on domain-specific intent classification for three popular domains due to space limit. As Table 3.6 shows, the margins of using multiple hypotheses with PoolingAvg are significant as well.

3.5 Conclusion

This chapter improves the SLU system robustness to ASR errors by integrating n-best hypotheses in different ways, e.g. the aggregation of predictions from hypotheses or the concatenation of hypothesis text or embedding. We can achieve significant classification accuracy improvements over production-quality baselines on domain and intent classifications, 14% to 25% relative gains. The improvement is more significant for a subset of testing data where ASR first best is different

from transcription. We also observe that with more hypotheses utilized, the performance can be further improved. In the future, we aim to employ additional features (e.g. confidence scores for hypotheses or tokens) to integrate n-bests more efficiently, where we can train a function f to obtain a weight for each hypothesis embedding before pooling. In addition, since more improvements are from the disagree part, which indicates the integration model is more helpful for low-quality hypotheses. We want to discuss, if the quality of hypotheses gets improved by a better ASR recognizing algorithm or some cleaning techniques on hypotheses (our datasets are raw recognition results from ASR without further cleaning), the change of the benefits brought by our integration model and how to improve the design for this condition. Another direction is using deep learning framework to embed the word lattice (Liu et al., 2014) or confusion network (Hakkani-Tür et al., 2006; Tur et al., 2002), which can provide a compact representation of multiple hypotheses and more information like times, in the SLU system.

CHAPTER 4

Efficient Training with Amplified Negative Sampling

In this chapter, we discuss the training of expensive multi-class classifiers with a large output-class size, which are widely utilized in different domains including word embedding, graph embedding and the previously introduced Internet embedding and SLU domains.

We propose a new sample-efficient training method, called *amplified negative sampling*, for efficient training large output-class multi-class classifiers. Our method jumps out of the framework of softmax approximation and directly approaches the optimum convergence point of the learning algorithm. Our proposed method is based on our analysis of the well-known negative sampling technique (Mikolov et al., 2013c) and is designed for (1) higher performance in the general task and (2) lower training computational cost. Our experiments on real-world datasets demonstrate that our proposed method leads to sampling cost savings with performance boost compared to the standard technique.

4.1 Introduction

In this work,we provide a novel sample-efficient method for training a multi-class classifier $C : X \rightarrow Y$ (X: input features, Y: output class labels) when the output-class size is large, say, $|Y| = 50,000$. Typically, when a classifier is modeled as a neural network, the final layer is implemented as a softmax layer with *one output neuron per each output class label* $y \in Y$, making it prohibitively expensive to train even for a reasonably large output-class size. To address this computational challenge, a number of techniques have been proposed, such as hierarchical softmax (Morin and Bengio, 2005), negative sampling (Mikolov et al., 2013a), adaptive softmax (Bengio, 2008) and its variants (Rawat et al., 2019; Blanc and Rendle, 2017; Grave et al., 2017; Chen et al., 2015;

Bai et al., 2017). Due to its simplicity and efficiency, negative sampling is one of the most popular techniques used in practice (Mikolov et al., 2013c; Wang et al., 2017; Grover and Leskovec, 2016a; Barkan and Koenigstein, 2016). In particular, it is widely utilized in many embedding frameworks, such as word embedding (Mikolov et al., 2013c), graph embedding (Grover and Leskovec, 2016a; Wang et al., 2017), and product-user embedding (Barkan and Koenigstein, 2016).

The key idea behind negative sampling is as follows: Given a training data point (x_i, y_i), the standard training algorithm updates the weights of the output neurons for *all* y's $\in Y$, not just for the training label y_i, making the training cost proportional to the output-class size $|Y|$. Negative sampling avoids this high cost by adjusting the weights for (1) the given training label, $y = y_i$ ("the positive sample") and (2) just a few y's that are randomly sampled from $Y - \{y_i\}$ ("negative samples"). Clearly, taking a few negative samples reduces the training cost by several orders of magnitudes when the output class size is large.

In general, it is reported that using a larger negative sample size leads to better downstream performance. For example, when Mikolov used negative sampling to embed words into high-dimensional vectors in (Mikolov et al., 2013c), he reported between 2-15% increase in downstream task performance when he used the 15 negative samples ($k = 15$) compared to 5 negative samples ($k = 5$). Unfortunately, the training cost of $k = 15$ is three times as large as that of $k = 5$, making its use significantly less appealing in practice. For instance, since training on a larger corpus generally improves the downstream performance as well, it may be the case that using a smaller k on a larger corpus may be just as good as or even better than using a larger k on a smaller corpus. This is the primary topic of this chapter. Is it possible to get the best of both worlds? Can we achieve a higher-quality model trained on a larger k without paying its training cost?

In this chapter, we give our answer to this question by finding a surprisingly simple yet effective modification to the negative sampling technique, named *amplified negative sampling*. Compared to the standard negative-sampling technique, our proposed technique can be used to either (1) *improve the prediction accuracy* of the trained model *for the same training cost* or (2) *lower the training cost for the same prediction accuracy*. We demonstrate the effectiveness of our proposed technique through extensive experiments on real-world data sets.

In summary, we make the following contributions in this chapter: (i) We propose *amplified negative sampling*, a simple yet effective modification to the widely-used negative-sampling technique that can improve its accuracy and lower its training cost based on our rigorous mathematical analysis. (ii) We compare the effectiveness of our proposed technique with standard negative sampling by conducting an extensive set of experiments on real-world data sets. Our results show that the effectiveness of our technique is in line with our theoretical prediction and can often be *twice as sample efficient* as the standard technique.

The rest of this chapter is organized as follows. In Section 4.3, we formally describe the multi-class classification problem and review the standard negative sampling technique. Then in Section 4.4, we propose amplified negative sampling and give the rigorous mathematical analysis of the method. In Section 4.5, we present the results of our experiments. We review related work in Section 4.2 and wrap up the chapter in Section 4.6.

4.2 Related Work

Softmax has been widely used in various models. In large output class classification problem, the intractable normalizing constant of softmax function will slow down the computation efficiency greatly. There are three major types of strategies have been investigated by the research community, including sampled softmax(Bengio, 2008), hierarchical softmax(Morin and Bengio, 2005) and spherical softmax(Vincent et al., 2015). We are not aiming to approximate the softmax function. This work makes one related yet distinct contribution: an efficient training method based on negative sampling strategy for a large classifier. The (Ruiz et al., 2018) share a similar scope with us which is not targeting on softmax approximation. Since the amplifying factor is distribution agnostic, this method actually can be applied to all the sampling-based softmax approximation method(Blanc and Rendle, 2017; Rawat et al., 2019).

Sampled Softmax This category contains the method which generates a subset of negative samples to avoid the high overhead of full negative sampling. Among this category, one class of methods try to generate samples from the softmax distribution. An adaptive sampling method was proposed by Bengio (Bengio, 2008) inspired by the importance sampling. Another prominent

example is the negative sampling (Mikolov et al., 2013a) which uses a simple noisy distribution to generate negative samples. TAPAS (Bai et al., 2017) uses a two pass scheme to generate samples from two different size candidate pools to reduce the sample overhead. Hashing method (Bakhtiary et al., 2015; Vijayanarasimhan et al., 2014) is also applied to either find the closest class or partial computation. (Rawat et al., 2019). Another class of methods instead focus on the sampled loss, including Noisy Contrastive Estimation(NCE) (Gutmann and Hyvärinen, 2010) by assuming the partition as an extra parameter to be computed during the computation and Adversarial Contrastive Estimation(ACE) (Bose et al., 2018) and (Schroff et al., 2015; Mussmann et al., 2017) selects the hardest negative examples.

Hierachical Softmax Hierarchical softmax was introduced in (Goodman) by utilizing the cluster structure to reduce the computation cost of softmax function. Bengio (Morin and Bengio, 2005) extend it into the tree structure. Due to the different inference procedure, the hierarchical softmax need extra steps to update the tree structure and maintain its property. Various method are proposed to stabilize this process such as class similarity, frequency binning and other optimization techniques. Zweig did some experiments to compare various tree structures.

Spherical softmax and kernel method The spherical softmax was proposed in (Vincent et al., 2015; De Brébisson and Vincent, 2015) which use quadratic function to replace the exponential function and enable faster computation of the gradients. However, these method seems not quite stable when the output label size is large. Kernel-based methods are also explored in (Blanc and Rendle, 2017; Rawat et al., 2019). These works introduces quadratic kernel and random Fourier features which show very promising results.

4.3 Framework

In this section, we briefly go over the general problem formulation of multi-class classifier learning and negative sampling to introduce key notation used in this chapter.

4.3.1 Preliminaries

We are given a dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in X$ is an *input feature* and $y_i \in Y$ is an *output class label*. We assume a discrete space of feature values x_i and output labels y_i, such that x_i and y_i take an integer value between $1 \leq x_i \leq m$ and $1 \leq y_i \leq m$. The multi-class classifier learning problem is to find a classifier $C : X \rightarrow Y$ that returns the correct label y_i given the input feature x_i: $y_i = C(x_i)$. Due to noise in the dataset and the uncertainty in predicting the correct label, this problem is often formulated as finding a conditional probability distribution $P(y|x)$ from the dataset D, which is interpreted as the probability that the correct output label is y given the input feature x.

Note that this formulation encompasses not just the multi-class classifier learning problem, but also most of the "data embedding" problems, such as word embedding (Mikolov et al., 2013a,c), graph embedding (Grover and Leskovec, 2016b), and item embedding (Barkan and Koenigstein, 2016). For example, the well-known skip-gram model for word2vec (Mikolov et al., 2013c) falls under this formulation by defining a *context word* as an input feature x_i and any *target word* that appears near the context word as an output label y_i.

Learning the conditional probability function $f(x,y) = P(y|x)$ from the dataset D is done by assuming a parameterized hypothesis space $f_\theta(x,y) = P_\theta(y|x)$, where the hypothesis space $f_\theta : (x,y) \rightarrow [0,1]$ is a space of differentiable functions parameterized by $\theta \in R^d$. Among all possible parameters $\theta \in R^d$, an *optimal parameter* θ^* is chosen to minimize the loss function $L(f_\theta, D)$, where $L(f_\theta, D)$ captures the "loss of f_θ" or the difference between f_θ and D. Multiple definitions of the loss function $L(f_\theta, D)$ are used in practice, including $L1$, $L2$, and *cross entropy*:

$$L_1 : \sum_{(x_i,y_i)\in D} \sum_{y\in Y} |\mathbb{1}(y = y_i) - f_\theta(x_i, y)| \tag{4.1}$$

$$L_2 : \sum_{(x_i,y_i)\in D} \sum_{y\in Y} (\mathbb{1}(y = y_i) - f_\theta(x_i, y))^2 \tag{4.2}$$

$$L_{CE} : - \sum_{(x_i,y_i)\in D} \sum_{y\in Y} [\mathbb{1}(y = y_i) \log f_\theta(x_i, y) + (1 - \mathbb{1}(y = y_i)) \log (1 - f_\theta(x_i, y))] \tag{4.3}$$

Here, $\mathbb{1}(y = y_i)$ is an indicator function that takes the value 1 if $y = y_i$ and 0 otherwise. In order to make our discussion concrete, we primarily assume $L2$ as our loss function in the rest of this chapter and simply state the result of our analysis for the other loss functions.

The gradient-descent method is often utilized to identify the parameter θ^* that minimizes the loss function $L(f_\theta, D)$. Given the definition of the $L2$ loss function, its gradient is:

$$\nabla_\theta L_2(f_\theta, D) = -2 \sum_{(x_i,y_i)\in D} \left[\sum_{y\in Y} (\mathbb{1}(y = y_i) - f_\theta(x_i, y)) \nabla_\theta f_\theta(x_i, y) \right] \tag{4.4}$$

Note that the inner summation of the above equation makes its computation prohibitively expensive: For each training data $(x_i, y_i) \in D$, we take the inner sum over *every* output label $y \in Y$, not just the training label y_i.[1] This makes the computational cost *proportional to the output class size* $|Y|$. We refer to the training method that computes the full gradient of Equation 4.4 as *full-gradient training*.

4.3.2 Negative Sampling

Negative sampling is a technique that tries to reduce the high computational cost of full-gradient training. The idea of negative sampling was originally proposed in 2010 as Noisy Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010), which was generalized for natural language processing by Mihn in (Mnih and Teh, 2012). It was used as part of the word2vec computation (Mikolov et al., 2013a), which led to a wide adoption for general vector-embedding problems (Grover and Leskovec, 2016a; Barkan and Koenigstein, 2016; Grover and Leskovec, 2016b).

The basic idea of negative sampling can be summarized as follows: Given a training data (x_i, y_i), we refer to y_i as the "*positive sample*" and all other label $y \in Y - \{y_i\}$ as "*negative samples*." Full-gradient training sums up the gradients from (1) the positive sample y_i and (2) *all* negative samples $y(\neq y_i)$. Negative sampling, instead, sums up the gradients from (1) the positive sample y_i and (2) *just a few randomly-selected negative samples* $y \in Y - \{y_i\}$.

[1] In certain cases, this sum over every $y \in Y$ is implicitly added to the hypothesis space $f_\theta(x, y)$. For example, when implemented as a neural network, the final layer is typically implemented as a softmax layer with one neuron per output label, which has the same effect as summing over ever $y \in Y$.

More precisely, we use NEG-$k(i)$ to represent the k randomly-chosen negative samples for the ith training data (x_i, y_i). That is, NEG-$k(i)$ is a size-k random subset of $Y - \{y_i\}$. Negative sampling then *approximates* the $L2$ loss function $L_2(f_\theta, D)$ as follows:

$$
\begin{aligned}
& L_2(f_\theta, D) \\
&= \sum_{(x_i,y_i)\in D} \left[\sum_{y\in Y} (\mathbb{1}(y = y_i) - f_\theta(x_i, y))^2 \right] && (4.5) \\
&= \sum_{(x_i,y_i)\in D} \left[(\mathbb{1}(y = y_i) - f_\theta(x_i, y))^2 \Big|_{y=y_i} \right. \\
&\qquad \left. + \sum_{y\in Y-\{y_i\}} (\mathbb{1}(y = y_i) - f_\theta(x_i, y))^2 \right] && (4.6) \\
&\approx \sum_{(x_i,y_i)\in D} \left[(1 - f_\theta(x_i, y_i))^2 + \sum_{y\in \text{NEG-}k(i)} f_\theta(x_i, y)^2 \right] && (4.7)
\end{aligned}
$$

From Equation 4.5 to 4.6, the sum over $y \in Y$ is expanded into the sum of $y = y_i$ and $y \in Y - \{y_i\}$. From Equation 4.6 to 4.7, the sum over all negative samples $y \in Y - \{y_i\}$ is approximated by the sum over NEG-$k(i)$. Clearly, this approximation can decrease the cost of computing the loss function significantly when $|Y|$ is large, which is the key reason for its efficiency. But what is the accuracy of this approximation? Will we still be able to get the same accurate model despite this approximation? If not, what is its exact impact? The next section will show the answers for those issues.

4.3.3 Optimal Model of Negative Sampling and Full-gradient Model

In this section, we investigate the achievable optimal model of the NEG-$k(i)$ and compare it with the model training with all $Y - \{y_i\}$ negative samples (Full-gradient) to see how accurate the NEG-$k(i)$ can approximate. For a better explanation, we use $\#(x)$ to represent the number of times that the input feature value x appears in the training data D and $\#(x, y)$ to represents the number of

times that the feature-and-label-value pair (x, y) appears in D. More formally,

$$\#(a) = |\{(x,y) \in D \mid x = a\}|$$
$$\#(a,b) = |\{(x,y) \in D \mid x = a \text{ and } y = b\}|$$

When we select the k negative samples for the ith training data (x_i, y_i), NEG-$k(i)$, we assume that a negative sample $y \in Y - \{y_i\}$ is sampled with replacement with probability p_y. Two popular choices of the sampling probability p_y are (1) the uniform distribution $p_y = c$ for some constant c and (2) according to the frequency of y in D. Our results are stated with the generic symbol p_y without making any explicit assumption on the sampling distribution. With this notation, we now could show the analysis to the negative sampling.

Theorem 1 (Optimal Model of Negative Sampling)**.** *When the hypothesis space f_θ has sufficient capacity,*[2] *the optimal model f_{θ^*} trained with k negative samples is the following with high probability:*

- *For L1 loss:* $f_{\theta^*}(x,y) = \mathbb{1}\left(\frac{\#(x,y)}{\#(x)} > \frac{k \cdot p_y}{k \cdot p_y + 1}\right)$
- *For L2 or cross-entropy loss:* $f_{\theta^*}(x,y) = \frac{\#(x,y)}{k \cdot p_y [\#(x) - \#(x,y)] + \#(x,y)}$

That is, for example, let θ_t^* be the parameter that minimizes the L1 loss function after t training epochs. Then for any $\varepsilon > 0$,

$$\lim_{t \to \infty} P\left(\left| f_{\theta_t^*}(x,y) - \mathbb{1}\left(\frac{\#(x,y)}{\#(x)} > \frac{k \cdot p_y}{k \cdot p_y + 1}\right)\right| < \varepsilon\right) = 1,$$

where $\mathbb{1}(a > b)$ is an indicator function whose value is 1 if $a > b$ and 0 otherwise.[3] *Similar statements can be made for L2 and cross entropy loss functions.*

While for the Full-gradient model training with all the negative samples instead of only k samples, we can get a corollary as follows.

[2] By having sufficient capacity, we mean that the hypothesis space has an independently adjustable parameter per every discrete (x, y) value pair following the assumption of (Goldberg and Levy, 2014).

[3] More precisely, $\mathbb{1}(a > b)$ may take any value between 0 to 1 when $a = b$.

Corollary 1.1 (Full-Gradient Model)**.** *When the hypothesis space f_θ has sufficient capacity, the respective optimal models trained with the full-gradient method are the following:*

- *For L1 loss:* $f_{\theta^*}(x,y) = \mathbb{1}\left(\frac{\#(x,y)}{\#(x)} > \frac{1}{2}\right)$
- *For L2 or cross-entropy loss:* $f_{\theta^*}(x,y) = \frac{\#(x,y)}{\#(x)}$

Note that the theorem and corollary mentioned above are achievable global optimal model but it does not guarantee that the global optimal can be always achieved in practice. In our another work, we provide a rigorous mathematical analysis and proof to the Theorem 1, the Corollary 1.1 and more extended corollaries. For this chapter, we do not show the analysis as we focus proposing a more efficient sample-based training method based on the Theorem 1 and the Corollary 1.1. In next section, we will show the design of the new sample-efficient training method and the analysis for it.

4.4 Amplified Negative Sampling

Our analysis and some related works like PMI (Levy and Goldberg, 2014) have shown that the optimal model trained with the full-gradient method is equivalent to the maximum likelihood estimator $f_{\theta^*}(x,y) = \frac{\#(x,y)}{\#(x)}$ while the model from negative sampling is not. Assuming that the dataset D is a representative sample from the true underlying distribution $P(y|x)$ (that is, $P(y|x) \approx \frac{\#(x,y)}{\#(x)}$) we can expect that full-gradient training are likely to result in a better model than negative sampling for a classification task.

Indeed, this is the general trend reported in the literature – not just for classification tasks but also for embedding tasks where negative sampling is frequently used. For example, when Mikolov used negative sampling to embed words into high-dimensional vectors in (Mikolov et al., 2013c), he reported noticeable accuracy increase in the word-analogy-task performance when he used $k = 15$ compared to $k = 5$.[4] Unfortunately, the training cost of using k negative samples is proportional to k, making the use of a higher k value unappealing in practice; since training on a

[4]For embedding tasks, the ultimate goal is to obtain vector representations of words that lead to high accuracy for downstream tasks. For this reason, obtaining the MLE model may not necessarily be "better." However, in a relatively small range of k, it is generally observed that higher k leads to better downstream task performance as well.

larger corpus generally increases the embedding quality as well, one may prefer using a smaller k on a larger corpus than using a larger k on a smaller corpus, assuming that it is bound by the same computational cost. This is where our study of *amplified negative sampling* started. Can we obtain the high quality model of higher k for the low computational cost of lower k? Can we get the best of the both worlds? We now explain how this can be achieved using amplified negative sampling.

Amplifying Factor. The key idea behind our amplified negative sampling comes from Theorem 1. From its analytic form, we observe that the optimal model $f_{\theta^*}(x,y)$ depends *only on the negative-sample size k*, not on how the samples are obtained. Given this, can we use *one negative sample multiple times during training*, pretending that it is the result from multiple random sampling? More formally, what will happen if we change the loss function $L(f_\theta, D)$ of Equation 4.7 (standard negative sampling) to the following?

$$L_2(f_\theta, D) = \sum_{(x_i,y_i)\in D} \left[(1 - f_\theta(x_i,y_i))^2 + \beta \sum_{y\in \text{NEG-}k(i)} f_\theta(x_i,y)^2 \right] \tag{4.8}$$

Note that Equation 4.8 is different from Equation 4.7 just by a constant factor β of the second term. We refer to β as a *amplifying factor*, since its intended role is to artificially "amplify" the effect of the negative samples NEG-k(i) by making its size look larger than they really are. Surprisingly, the following corollary shows that adding a amplifying factor produces this exact outcome.

Corollary 1.2 (Amplified Negative Sampling)**.** *When the hypothesis space f_θ has sufficient capacity, the respective optimal models trained with k negative samples with amplifying factor β under the three loss functions are the following with high probability:*

- *For L1 loss:* $f_{\theta^*}(x,y) = \mathbb{1}\left(\frac{\#(x,y)}{\#(x)} > \frac{\beta\cdot k\cdot p_y}{\beta\cdot k\cdot p_y + 1}\right)$
- *For L2 loss:* $f_{\theta^*}(x,y) = \frac{\#(x,y)}{\beta\cdot k\cdot p_y[\#(x) - \#(x,y)] + \#(x,y)}$
- *For cross-entropy loss:*
  $f_{\theta^*}(x,y) = \frac{\#(x,y)}{\beta\cdot k\cdot p_y[\#(x) - \#(x,y)] + \#(x,y)}$

Note that the factor k in Theorem 1 is replaced with βk in Corollary 1.2. That is, the amplifying factor β makes the optimal model trained with NEG-k effectively identical to the one from NEG-βk! Simply by multiplying a constant β to the loss function, we get the optimal model from a much larger sample size.

Proof for Corollary 1.2: We assume a discrete domain of input data D, where each data point $(x,y) \in D$ is an integer value pair of $1 \leq x \leq m$ and $1 \leq y \leq l$. Within the discrete domain, the most general parameterization of the function $f_\theta(x,y)$ is to assign an independent parameter per every pair of values (x,y) within $1 \leq x \leq m$ and $1 \leq y \leq l$. We use the symbol θ_{ab} to represent these parameters, i.e., $f_\theta(a,b) = \theta_{ab}$ for $1 \leq a \leq m$ and $1 \leq b \leq l$, where θ_{ab} can take any value in $[0,1]$.

Given the notation, the L_2 loss function of negative sampling is:

$$L_2(f_\theta, D) = \sum_{(x,y)\in D} \left[(1 - f_\theta(x,y))^2 + \sum_{y' \in \text{NEG-}k(x,y)} f_\theta(x,y')^2 \right] \tag{4.9}$$

$$= \sum_{(x,y)\in D} \left[(1 - \theta_{xy})^2 + \sum_{y' \in \text{NEG-}k(x,y)} \theta_{xy'}^2 \right] \tag{4.10}$$

Now, the $L2$ loss function for amplified negative sampling is

$$L_2(f_\theta, D) = \sum_{(x,y)\in D} \left[(1 - f_\theta(x,y))^2 + \beta \sum_{y' \in \text{NEG-}k(x,y)} f_\theta(x,y')^2 \right] \tag{4.11}$$

Since the only difference from standard negative sampling (4.9) is the coefficient β of the second summation, we can show that

$$\frac{\partial L_2(f_\theta, D)}{\partial \theta_{ab}} \xrightarrow{P} \#(a,b)(-2)(1-\theta_{ab}) + [\#(a) - \#(a,b)]\beta\, k\, p_b\, 2\theta_{ab} \tag{4.12}$$

as $t \to \infty$. By setting $\frac{\partial L_2}{\partial \theta_{ab}} = 0$, we can show that

$$\theta^*_{t,ab} \xrightarrow{P} \frac{\#(a,b)}{\beta\, k\, p_b [\#(a) - \#(a,b)] + \#(a,b)} \qquad \text{as } t \to \infty. \tag{4.13}$$

The proof for other loss functions can be done similarly.

Computational Cost of Amplified Negative Sampling. Note that both the standard negative sampling and our amplified negative sampling are *sampling methods* that are independent of the particular choice of the training method. They simply provide a straightforward recipe for selecting a few negative samples and incorporating them in the computation of the loss function. Therefore,

both the standard and the amplified versions use the same training algorithm,[5] making their algorithmic and computational complexity identical. That is, as long as they use the same negative sample set NEG-k, their computational costs are (almost) identical.[6] At the same time, our mathematical analysis indicates that even though the same negative samples are used, the amplified version is likely to produce a significantly more accurate model than the standard version by the factor β. In the later experiment section, we evaluate the validity of this analytical result both in terms of the computational cost and the model accuracy through an extensive set of experiments on real-world datasets.

Amplifying Factor vs Learning Rate. Conceptually, our amplifying factor may look similar to the *learning rate* used for the gradient-descent algorithm; at epoch t, the gradient-descent algorithm updates the parameter θ from the current value θ_t to the new value θ_{t+1} via the the following equation:

$$\theta_{t+1} = \theta_t + \alpha \cdot \nabla_\theta L(f_\theta, D), \tag{4.14}$$

where α, the learning rate, controls how quickly and reliably the updates converge. As we can see from Equation 4.8, our amplifying factor β is also multiplied to (a part of) the loss function, so it indeed plays a role very similar to the learning rate. The only difference is that α is multiplied to the *entire* loss function $L(f_\theta, D)$ while β is multiplied to its *negative-sample terms only* $\sum_{y \in \text{NEG-}k(i)} f_\theta(x_i, y)^2$. Interestingly, our analysis and later experiments show that this seemingly minor change leads to an enormous difference in terms of the final trained model.

Optimal Amplifying Factor and Model Quality. The similarity of the learning rate and the amplifying factor raises another interesting question. How big a amplifying factor can we safely use? It is well known that a higher learning rate generally leads to a faster convergence rate initially, but it makes the training process less stable at a later stage. Will using a large amplifying factor lead to similar behaviour? Or will it always be better to use a larger amplifying factor? Our analysis indicates that the optimal value of β might be $\beta = \frac{1}{kp_y}$ since this value leads to the MLE model.

The results from our experiments do not provide a single answer to this question. In the ex-

[5] Perhaps, the most popular choice is the stochastic-gradient descent algorithm

[6] The amplified version has the overhead of multiplying β compared to the standard version, but this cost negligible in practice.

periments conducted with our own code, where we measure the model accuracy in terms of the difference of the trained model from MLE, we observe that using a large amplifying factor always produces a model closer to MLE. It also does not introduce much instability to the training process all the way from 1 through $\beta = \frac{1}{kp_y}$. In the experiments conducted with existing codes for *other downstream tasks*, we observe that using a amplifying factor up to $\beta = 3$ reduces the training time and improves downstream-task performance, but starting from $\beta > 3$, we sometimes observe reduced downstream-task performance. This may be due to the fact that downstream-tasks performance does not necessarily correlate with how well our model estimates the conditional probability $P(y|x)$. Given these two results, we find using a reasonably small amplifying factor, say $\beta = 3$, may be a safe choice in general; it reduces training time significantly and improves downstream-task performance.

In our experiments, we also explored a few other directions, including (1) decaying the amplifying factor over epochs similarly to decaying the learning rate, (2) dynamically setting the amplifying factor based on the "loss value" of the negative sample similarly to the idea of importance sampling and (3) early stopping, where we stop using the amplifying factor after a few epochs. We find that these changes do not introduce a meaningful difference to the results.

4.5 Experiments

The primary goal of this section is to experimentally investigate the following two issues: (1) Does the result of our analysis hold in practice? We want to examine how well our theoretical results match with experiments. We also want to experimentally explore a few questions raised in this chapter, including the choice of the optimal amplifying factor and the difference between the learning rate and the amplifying factor. (2) Does amplified negative sampling help other downstream tasks as well? The performance of other downstream tasks may not necessarily depend on how accurately the trained model captures the conditional probability $P(y|x)$, so we want to experimentally check whether amplified negative sampling has positive effects on other downstream tasks or not.

In the Section (Model Accuracy and Training Efficiency) 4.5.1, we explore the first issue by

measuring the difference between the maximum likelihood estimator (MLE) $\tilde{P}(y|x) = \frac{\#(x,y)}{\#(x)}$ and the models trained with (a) full-gradient training (b) negative sampling and (c) amplified negative sampling. The results of our experiments show that the conclusions of our analysis hold in practice to a surprising degree of accuracy. They also show that the learning rate and the amplifying factor have vastly different effects on the trained model. In the subsection (Experiments on Other Downstream Tasks) 4.5.2, we investigate the second issue by running experiments on downstream tasks: word-analogy tasks (Mikolov et al., 2013a). Here, we observe that amplified negative sampling leads to improved performance on downstream tasks as well.

In summary, our experimental results strongly indicate that there really is not much downside to using amplified negative sampling; as long as we use a reasonably small amplifying factor, say $\beta = 3$, amplified negative sampling leads to lower training time and higher model accuracy.

4.5.1 Model Accuracy and Training Efficiency

Experimental Settings. In this subsection, we experimentally compare four training algorithms, full-gradient training (*FullGrad*), 5 negative samples (*Neg5*), 5 negative samples with the amplifying factor 3 (*Neg5-Amplify3*), and 15 negative samples (*Neg15*), under three different loss functions, $L1$, $L2$, and cross entropy. For the choice of the hypothesis space $f_\theta(x,y)$ and the training set, we use a setting similar to (Mikolov et al., 2013a). That is, as our hypothesis space we use the skip-gram model of (Mikolov et al., 2013a) with a 100-dimensional hidden layer. As our dataset, we use a subset of Text8 corpus from (Mikolov et al., 2013a) by extracting the first 75,000 words and applying the same min_count filter of 5 in (Mikolov et al., 2013a).[7] We use the stochastic-gradient descent (SGD) with the batch size of 500 as the training algorithm. All our experiments use the window size 3 and the learning rate 0.025 unless noted otherwise. All other parameter settings are the same as in (Mikolov et al., 2013a). All results reported are the average of three independent runs with identical settings. All codes were implemented using PyTorch v1.0.1.

Model Accuracy and Convergence Rate. In Figure 4.1, we compare the model accuracy

[7]Using a subset of Text8 here is due to the high training cost of *FullGrad* and our desire to keep the training time at a manageable level. In our next experiments on other downstream tasks, we run report our results from experiments on much larger datasets.

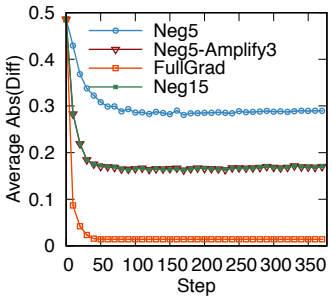


Figure 4.1: Model accuracy

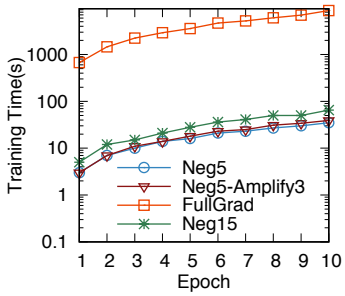


Figure 4.2: Training time.

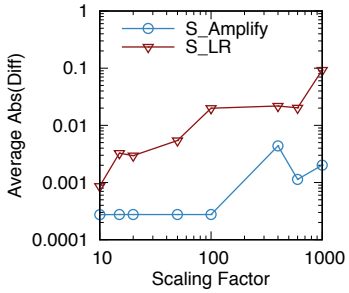


Figure 4.3: Amplifying factor vs learning rate

and convergence rate when the model is trained with the four algorithms (FullGrad, Neg5, Neg5-Amplify3, Neg15) under the $L2$ loss function.[8] In the graph, the horizontal axis corresponds to the stochastic-gradient-descent training batch steps (with roughly 70 steps corresponding to one training epoch) and the vertical axis corresponds to the average absolute difference between the trained model $f_{\theta^*}(x,y)$ and MLE $\tilde{P}(y|x) = \frac{\#(x,y)}{\#(x)}$, i.e., $\sum_{(x,y)\in D} \frac{1}{|D|} \left| f_{\theta^*}(x,y) - \frac{\#(x,y)}{\#(x)} \right|$.

From the graph, a few things are clear: (1) Full-gradient training converges to MLE. Even at epoch 1 (step 70), the mean absolute difference of *FullGrad* is close to zero, indicating that it converged to MLE. (2) The amplifying factor β effectively "increases" the negative sample size by the factor β in terms of model accuracy. The mean absolute difference of Neg5-Amplify3 and Neg15 are the same at every training step — they overlap so closely and it is difficult to tell them apart in the graph — indicating that they both converge to the same model at the same rate. This result is what our theoretical analysis predicts: $(k = 15, \beta = 1)$ leads to the same optimal model as $(k = 5, \beta = 3)$. (3) A model trained on larger k approximates MLE better. The mean absolute difference of Neg15 is significantly smaller than that of Neg5.

Training Time and Computational Cost. In Figure 4.2, we compare the training time of the four algorithms. The horizontal axis corresponds to training epochs and the vertical axis corre-

[8] While we performed experiments under all three loss functions, we report the results only from $L2$ here. The conclusions from other loss functions are essentially the same.

sponds to training time, which roughly captures the computational cost of each algorithm. The vertical axis is logarithmic; since the training time of *FullGrad* is two orders of magnitude larger than others, its result is not visible in the same graph otherwise. From the graph, we again observe what is predicted by our analysis. The training time of Neg5-Amplify3 is practically the same as that of Neg5. That is, Neg5-Amplify3 works almost like Neg5 in terms of its training time and computational cost, but it works almost like Neg15 in terms of its model accuracy and convergence rate! Amplified negative sampling indeed gives the best of both worlds.

Learning Rate vs Amplifying Factor. In Figure 4.3, we compare the effect of using different learning rates and amplifying factors. The graph is from training the model using Neg15 under $L1$ loss. The curve labeled as *S_LR* is obtained by multiplying the default learning rate of 0.025 by a factor between 10 and 1,000. The curve labeled as *S_Amplify* is obtained by including the amplifying factor β between 10 and 1,000. The vertical axis is again in the logarithmic scale due to the high difference between the two curves and represents the model accuracy (the mean absolute difference from MLE) at the given learning rate/amplifying factor. From the graph, we see that changing the learning rate and changing the amplifying factor lead to vastly different results. As we increase the learning rate, the trained model diverges further away from MLE. When we increase the amplifying factor, however, the trained model stays close to MLE all the way through $\beta = 100$. Only after $\beta > 100$, the model starts to diverge and becomes unstable. This result is consistent with our analysis; according to Corollary 1.2, amplifying converges to MLE at $\beta \approx 140$ under the current setting,[9] so its divergence beyond $\beta > 140$ is expected.

4.5.2 Experiments on Other Downstream Tasks

In the previous set of experiments, we investigated the effect of amplified negative sampling on the trained model in terms of its difference from MLE. In the next set of experiments, we investigate its effect on downstream tasks: word-analogy tasks. In all our experiments, we compare the results from Neg5, Neg5-Amplify3, and Neg15 at training epoch 3.

[9] Amplified negative sampling converges to MLE when $\beta k p_y = 1$. Given $k = 15$ and the uniform probability $p_y \approx 1/2000$ for this experiment, $\beta k p_y = 1$ at $\beta \approx 140$.

Word2vec: Word analogy. This test is conducted with the CBOW model trained on the Text8 corpus (Mikolov et al., 2013c) using the code downloaded from (Word2vec, 2019) after we add amplifying code. The performance is evaluated by the 14 word-analogy tasks in (Mikolov et al., 2013c). We use the default parameter settings of the downloaded code.

In Table 4.1 we show the accuracy of Neg5, Neg5-Amplify3, and Neg15 for the first 5 word-analogy semantic tasks of (Mikolov et al., 2013c). From the results, the trend is clear: the performance of Neg5-Amplify3 is higher than Neg5 and is close to Neg15. In Table 4.2, we show the training times of Word2vec. We can find the training time of Neg5-Amplify is close to Neg5 and is significantly smaller than that of Neg15. In the cases of Word2vec, the difference is by a factor 2. From our experiments, we observe the general trend: The downstream-task performance of Neg5-Amplify3 is close to Neg15 while its training cost is close to Neg5.

Table 4.1: Word-analogy semantic-task top-1 accuracy.

Task Name	Neg5	Neg5-Amplify3	Neg15
capital-common-countries	38.14	43.02	47.43
capital-world	24.66	29.34	32.14
city-in-state	15.17	17.29	14.93
currency	15.72	22.26	22.21
family	56.86	60.30	64.27

Table 4.2: Training time (seconds).

Model	Neg5	Neg5-Amplify3	Neg15
Word2vec	17.35	17.37	36.93

4.6 Conclusion

In this chapter, we proposed *amplified negative sampling*, a new sample-efficient method for training multi-class classifiers with a large output-class size. Our proposed method was based on our rigorous mathematical analysis. Our extensive set of experiments demonstrated that amplified negative sampling gives us the best of both worlds: It leads to the higher-accuracy model of a larger sample size without paying its high computational cost. Given its simplicity, and experimental effectiveness, we believe our proposed method will be an important extension to the widely-popular technique that results in meaningful improvements in practice.

CHAPTER 5

Expressive Library for Recursive Queries: LLib and LFrame

In this chapter, we focus on designing a succinct expressing interface to facilitate complex data analysis, especially the recursive algorithms, which are commonly utilized to develop applications in various domains including graph data analytics, language models, etc.

Due to the ever-increasing volume of data, there is an urgent need to provide expressive and efficient tools to support Big-Data analytics. The declarative logical language BigDatalog has proven very effective at expressing concisely graph, machine learning, and knowledge discovery applications via recursive queries that execute with superior performance and scalability on Apache Spark. To help data scientists benefit from these advances in full synergy with Spark's rich libraries and programming environs, we develop the Spark DataFrame extension LFrame and the Logical Algorithm Library LLib.

LLib provides a wide range of Datalog algorithms written using BigDatalog on Apache Spark. LLib encapsulates complex logic-based algorithms into high-level APIs, which simplify the development and provide a unified interface akin to the one of Spark MLlib. As a fully compatible DataFrame-based API, LLib enables the integrated utilization of LLib applications and new Datalog queries with existing Spark functions, such as those provided by MLlib and Spark SQL. To facilitate the development of new applications not contained in LLib, our system provides an LFrame-based programming interface to Datalog. LFrame objects convert directly to and from DataFrame objects to support both relational operations and logic-based operations. In addition, both LLib and LFrame support interoperability with multiple programming languages, whereby users can now express succinctly powerful recursive queries in Scala, Java, or Python. As a result, BigDatalog becomes a very attractive software development tool in the Apache Spak ecosystem. This chapter utilizes several running examples to demonstrate the power and versatility of LLib

and LFrame.

5.1 Introduction

In the era of big data, the demand for flexible analytics on large-scale data has driven researchers to build various general-purpose user-friendly platforms like Apache Spark (Zaharia et al., 2012), AsterixDB (Alsubaiee et al., 2014), Pig (Olston et al., 2008), Hive (Thusoo et al., 2009), etc. Among these systems, Spark is getting more and more attractive due to its efficient in-memory computation and abundant APIs (i.e. Spark SQL, GraphX, MLlib and SparkR) for sophisticated analytics to extract rich information encapsulated in the data.

However, for iterative applications like identifying transitive closures or connected components of millions of vertexes, there are no dedicated designs for optimization among recursions in Spark. For more sophisticated recursive analytics, the programming needs deep understanding and extensive knowledge of the platform. To solve these issues, researchers have attempted to implement Datalog (Consens and Mendelzon, 1990) systems.

Datalog, a well known logical programming language with superior expressive power on recursive algorithms, consists of a set of rules and facts. The DeALS project of UCLA (Yang et al.) implements a unified Datalog programming language and provides a parallel evaluation on multi-core machines. For the distributed logical computing on clusters, the BigDatalog (Shkapsky et al., 2016) system is further developed on Spark. While considering the SQL programming customs, the Recursive-aggregate-SQL (RaSQL) (Gu et al., 2019) is proposed as a simple extension of Spark SQL for Datalog.

This torrent of Datalog platforms, however, underscores the needs to make Datalog as an integral part of data processing pipeline and provide high-level APIs to simplify the development. In those systems, one Datalog application run independently as one job with input rules and datasets. It requires much programming for users to convert the output of one Datalog program to the required input of another Datalog program. Similarly, the collaboration with other libraries like machine learning (MLlib), graph computation (GraphX) in Spark is inconvenient. Also, it is necessary for users to learn about Datalog or get used to a SQL extension (RaSQL) when they actually want

a simple function wrapping all the logic of a common Datalog algorithm.

In this chapter, we focus on making the Datalog module as the first-class citizen in Spark, that can easily collaborate with Spark libraries. We proposes LLib, wrapping all key Datalog algorithms with a unified interface for adding new ones, and LFrame, the DataFrame extension supporting the logical operations. LLib and LFrame are implemented on BigDatalog and Spark. They can support multiple programming languages, such as Python, Scala and Java. The wide audience of Spark community should be familiar with the interface of LLib (like Spark MLlib) and LFrame (like Spark DataFrame). With LFrame and LLib, data scientists could have access to the data manipulated in Datalog applications and continuously do subsequent processing like machine learning algorithm within one job. This ecosystem makes the end-to-end developing Datalog algorithms with a high-level API (LLib) possible. For a user-defined recursive application outside LLib, with LFrame, it can be as friendly as Pandas (McKinney et al.)/Spark DataFrame.

Our contributions can be concluded as following:

- Usability. Both LLib and LFrame are tailored to data scientists and support multiple programming languages including Python, Scala and Java. In addition, LLib provides functions for a wide range of typical Datalog algorithms and makes it possible for the end-to-end recursive development with high-level APIs.

- Interoperability. LLib is the DataFrame-based API, which takes the DataFrames as the input and generates also DataFrames. This facilitates the collaborations between LLib applications and Spark MLlib, Spark SQL or GraphX. As for LFrame, we provide the functions for flexible conversions between the DataFrame and LFrame, which also smooth the collaborations between LFrame-basesd algorithms and Spark libraries. Both LLib and LFrame make the Datalog module as an integral part of data processing pipeline.

- Extendability and flexibility. In LLib, there is a template and several utility functions to help contributing extra Datalog algorithms. We also allow user-defined Datalog function on LLib to wrap any possible Datalog algorithm. LFrame data structure is associated with various general Datalog operations to develop any recursive algorithm.

The chapter is organized as follows. Section 5.2 reviews the basics about the Datalog language and related platforms including Apache Spark, BigDatalog and RaSQL. Section 5.3 describes the working paradigm of LLib, user-defined Datalog functions and the collaborations with other Spark libraries. Section 5.4 presents the conversion between LFrame and DataFrame, the basic unary and N-ary operations supported by LFrame, and some examples by LFrame. Section 5.5 demonstrates our design to support multi-language programming. Section 5.6 discusses the performance overhead. Section 5.7 draws conclusion and maps out plans for future work.

5.2 Preliminaries

5.2.1 Datalog

A Datalog application is comprised of a finite set of rules. Each rule r is formed as H $\leftarrow$ $l_1, l_2, ...l_n$, where H is the head of r, $l_{1..n}$ (the *body*) are *literals* and the $\leftarrow$ means implication. The literals ($l_{1..n}$) are positive or negated atoms. One atom (H or l_i) can be formed as $p(t_1, .., t_k)$, where p is a *predicate* and $(t_1, .., t_k)$ terms can be *variables*, *constants* or *functions*. So, the r is a rule to infer H. However, if r does not have the body $l_1, l_2, ...l_n$, it becomes the *fact*, which corresponds to a tuple in a relation . The comma separating literals means the logical conjunction (AND). To evaluate a Datalog application, we need a *query* indicating which predicate to evaluate.

Next, we will illustrate a classic example in Datalog, single source shortest path (SSSP), with more terms covered below. The SSSP is to calculate the length of shortest paths from one source vertex to all other vertices in a graph with weighted edges.

Query 1 - Single source shortest path (SSSP).

1 : $database(\{\ warc(A : integer,\ B : integer,\ Cost : integer)\ \}).$

2 : $sp(B,\ mmin < C >) \leftarrow B = \{startvertex\},\ C = 0.$

3 : $sp(B2,\ mmin < C >) \leftarrow sp(B1,C1),\ warc(B1,B2,C2),\ C = C1 + C2.$

4 : $result(B,\ min < C >) \leftarrow sp(B,C).$

5 : $query\ result(T,C).$

As shown in Query 1 line 1, the input relation (*base relation*) is warc with the schema (A:integer, B:integer, Cost:integer). One fact of this relation can be warc(1, 2, 5), which shows the cost from the vertex 1 to 2 is 5. The database is a keyword specifying the base relation. In the first rule (line 2), it initializes the shortest distance from the source vertex to itself as 0, where the "{startvetex}" can allow user to input the source vertex ID. The second rule (line 3) recursively produces all the minimum distances for all possible paths from source node to another node. The monotonic aggregate (Zaniolo et al., 2019; Das et al., 2019), mmin is utilized, which allows the aggregation inside the recursion when set containment is satisfied. The mmin will get a new lower value with a larger set of possible paths. And a normal aggregate min is finally exploited (line 4) to obtain the minimum cost path. The fifth line denotes the predicate (*relation*), result(T,C) will be evaluated and become the output of the application.

5.2.2 Related Platforms: Apache Spark, BigDatalog and RaSQL

Apache Spark. Apache Spark is a DISC system with various modules like Spark SQL (Armbrust et al., 2015b), MLlib (Meng et al., 2016) and GraphX (Gonzalez et al., 2014) to support analytics on structured data, machine learning algorithms and graph computation algorithms respectively. All the Spark applications are eventually represented by a series of transformations and actions on Resilient Distributed Datasets (RDDs) (Dean and Ghemawat, 2004), the main abstraction provided by Spark. The operators like groupBy and filter are lazily evaluated until an output action like count trigger evaluations on RDDs. In this way, before execution, the Spark Optimizer can design a better physical plan by avoiding some duplicate computation or pipelining opera-

tions. RDDs in Spark are actually Python or Java objects stored in memory and can be processed in parallel. The RDD is fault-tolerant due to its lineage. The lineage graph of RDDs records the transformations applied to them, which facilitates the tracking and debugging through data transformations (Gulzar et al., 2017; Interlandi et al., 2018). The Spark has attracted wide audiences from database community due to its high usability (Armbrust et al., 2015a) and continuously optimized query planner (Li et al., 2018c).

BigDatalog. With the popularity of Spark and the benefits for recursive query evaluation and optimization brought by Datalog, the new requirements have been re-emerged to support the Datalog programming on Spark. As far as we can find, BigDatalog (Shkapsky et al., 2016) is the first platform implementing DeALS, a Datalog platform, on Spark. It supports the execution of recursive operations on multi-core machines and clusters. BigDatalog also proposes optimizations on physical planning for recursive queries to obtain performance improvement.

RaSQL. BigDatalog enables development of Datalog on Spark, but users should get used to the Datalog syntax. With the continuing popularity of SQL, it is beneficial to design a language similar to SQL for Datalog queries. RaSQL (Gu et al., 2019) proposes a new language following and extending SQL standards, and utilizes some novel optimizations on fixpoint operators for the Datalog platform built on Apache Spark.

5.2.3 Spark MLlib and DataFrame

MLlib is the machine learning library of Spark. It consists of classification and regression algorithms. Users can flexibly build a pipeline of a sequence of algorithms to process data with the abundant libraries in MLlib. The pipeline could span across data cleaning, model initialization, training, prediction, and evaluation.

DataFrame is a popular facility for data scientists and has been supported by various trendy data analytics platforms and languages such as Spark, Pandas (McKinney et al., 2010), R, etc. It is utilized as a data structure and the data stored there is organized in rows and columns like a table in Excel. This increases the visibility of development for non-expert users. The supported operations for the DataFrame are similar to the relational algebra, but are exposed as pre-defined functions

like libraries of Java, Python, etc.

5.3 LLib

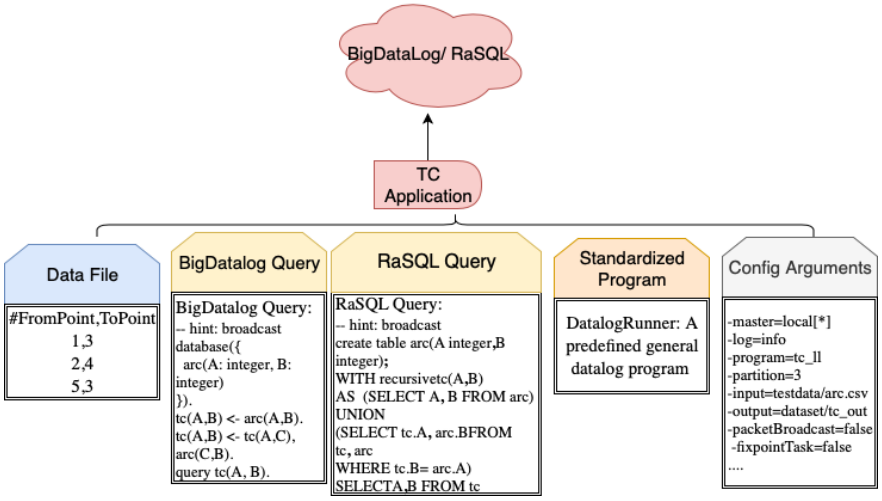
In this section, we discuss LLib, which works in a similar way as MLlib to develop Datalog applications. It does not require users to be familiar with logical programming. LLib contains a wide spectrum of recursive applications including graph algorithms (Transitive Closure, i.e. *TC*), temporal database queries (Interval Coalesce), financial applications (MLM), machine learning algorithms (Logistic Regression), etc. The data analysts could easily take one Datalog algorithm as one step within their complex data processing process, which is more flexible than the previous Datalog programming interfaces.

5.3.1 Working Paradigms Comparison: LLib, BigDatalog and RaSQL

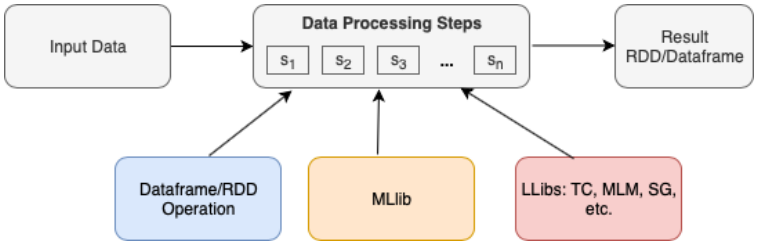
LLib grants the access to the processed data to users through the Spark DataFrame, which greatly reduces the required boilerplate code and improves the readability. In this part, we would like to firstly provide a high level picture of the differences while working on two popular Spark-based distributed Datalog platforms (BigDatalog and RaSQL) and our LLib.

As shown in Figure 5.1 (a) about Transitive Closure application, the previous Datalog programs are made up of (1) source data file(s), which should follow a determined format, (2) a query file following RaSQL or BigDatalog syntax, (3) a standardized program (script) to execute the query on the source data, and (4) a set of arguments to guide the execution. This is a general-purpose design for all potential applications that can be expressed by a combination of rules running on several data files, but it causes a troublesome programming considering the following aspects.

- *Learning new language.* Users are required to learn the syntax of Datalog or an extension of SQL while they actually want to use a well-wrapped Datalog function.
- *Data preparation.* For each application on the previous platforms, the required data need to follow a determined format. It costs extra time for the data preparation.



(a) The composition of TC application on custom Datalog platforms.



(b) LLib paradigm

Figure 5.1: Working paradigm comparison between existing Spark-Based Datalog platforms and LLib.

- *Isolation from other applications.* The development mode of the original logical programming works in a way like a black box. Users provide data and rules to standardize program and get results. They only have access to the generated file and can hardly involve preprocessing or subsequent processing.

To tackle those issues, we develop the DataFrame-core LLib as shown in Figure 5.1 (b), where Datalog applications are seamlessly integrated into the data processing pipeline. Within an LLib-based application, there can be many steps made up of the permutations of regular DataFrame operations, MLlib's ML algorithms and LLib's Datalog algorithms. Users do not need to manually program to connect $s_1 ... s_n$ by preparing Datalog input data or persisting and loading Datalog execution results for subsequent processing. Later in next subsection, we will describe the underlying architecture and working pipeline of LLib through a running example, Transitive Closure.

5.3.2 LLib Processing Pipeline and Underlying Architecture

5.3.2.1 Working Session and Acquiring Data

In LLib, the first step is to construct a working environment for the Datalog queries and libraries. We respect the customs of building Spark Session and exploit the similar way as following (with Transitive Closure, i.e. TC as a running example):

Example 1.1 - Transitive closure with LLib: Working session.

$$session = LLibSession.builder().appName(\text{“}TC\text{”}).master(\text{“}local[*]\text{”}).getOrCreate()$$

$$schema = StructType(List(StructField(\text{“}Point1\text{”}, IntegerType, true),$$

$$StructField(\text{“}Point2\text{”}, IntegerType, true)))$$

$$df = session.read.format(\text{“}csv\text{”}).option(\text{“}header\text{”}, \text{“}false\text{”}).schema(schema).load(\text{“}arc.csv\text{”})$$

LLibSession synthesizes the Spark environment and the special designs for logical programming. Within the same session, users are free to utilize existing Spark libraries like data loading functions to create a DataFrame *df* (line 4) or our Datalog libraries i.e. LLib or LFrame.

5.3.2.2 Initializing an Executable Object of LLib and Mapping the Schema

Initialization: In LLib, the pipeline of the data processing for a typical application is wrapped to an executable object, which users can initialize and set parameters of. In Example 1.2, we initialize a transitive closure object tc with the TC library, which is pre-defined and included in the LLib. Then, we can set the property by built-in functions.

Example 1.2 - Transitive closure with LLib: Initializing TC.

$$import\ edu.ucla.cs.wis.bigdatalog.spark.LLib.TC$$

$$val\ tc = new\ TC()$$

$$tc.setDirection(FromCol = "Node1", ToCol = "Node2")$$

Query 2 - Transitive closure.

$$database(arc(From : integer, To : integer)).$$

$$tc(From, To) \leftarrow arc(From, To).$$

$$tc(From, To) \leftarrow tc(From, Tmp), arc(Tmp, To).$$

$$query\ tc(From, To).$$

Schema Mapping: Among the built-in functions, all the libraries are required to contain one function *setDirection* for schema mapping. This is necessary because we want a more general design to accept DataFrames with various schemas as the input. One attribute may have different names in different DataFrames. With schema mapping, we could know the corresponding relationship between input DataFrame's attributes and the attributes in our library's computing logic. For example, Query 2 is a set of Datalog rules used by the transitive closure of our library. We have two attributes, From and To, in the arc table. The two attributes can be called differently like (Node1, Node2) as shown in the df initialized in Example 1.1. The mapping between (From, To) and (Node1, Node2) can be provided by the setDirection function as shown in the third line of Example 1.2. If there is a long mapping list for attributes, we can store the mapping information within a hash table.

Schema Recovering: While processing the data with our library, we need the mapping information. But after processing, the output data's format should be consistent. There is one mechanism

to rollback the schema. At the beginning of data processing, we store the schema of the input data. Then, in the end, we could use the pre-stored schema to recover.

5.3.2.3 Execution and Persistence

With the executable object and imported data, the execution and persistence can be merely a one-line execution with a pre-defined function (run, genDF or genRDD) in LLib. These three functions can support basic requirements to operate data and store the result to a variable or a file. As shown in Example 1.3, the function *run* is to run the logical programming and persist the result directly to the target address. If users would like to do subsequent processing, it can output the data into a DataFrame or RDD as shown in second and third line of the Example 1.3.

Example 1.3 - Transitive closure with LLib: Execution and persistence.

$$tc.run(df, output = "File", session)$$

$$val\ dfNew = tc.genDF(df, session)$$

$$val\ rddNew = tc.genRDD(df, session)$$

These three functions are implemented for each library of LLib. They expect the input data (df) and the environment (session) as inputs. The session information is needed for during execution, we want the program running in an environment with ability to support logical programming.

5.3.2.4 Multiple Data Sources

The previous TC example operates on only one relation, however there are many applications requiring more than one relation, which brings changes to the pipeline. We illustrate the Multi Level Market (MLM) Bonus as a typical Datalog application (MLM, 2008) acquiring more than one dataset. The application is to calculate the bonus for members of a hierarchical structural Multi Level Marketing organization. In the organization, new members are recruited by and get products from old members (sponsors). One member's bonus is based on his own sales and a proportion

of the sales from the people directly or indirectly recruited by him. The scale of the proportion is user-defined.

There are two relations in MLM Bonus, including the *sponsor* and *sales*. The sponsor relation describes the recruiting relationship among members, while the sales relation records the profits for each member. In Datalog syntax, the base case should be calculating the member's bonus by the sales table. And the recursive rule is to calculate the bonus based on the basic profits and the profits derived from the downstream members.

With the help of LLib, users could implement MLM Bonus by a pre-defined class in LLib ignoring the complex logic. The program can be as easy as follows.

Example 2 - LLib with more than one input relation: Multi Level Market Bonus.

$$val\ MLM = new\ MLM()$$

$$MLM.setDirection(MCol = "Member", ProfitCol = "Bonus")$$

$$MLM.setSecDirection(MCol = "Member1", M2Col = "Member2")$$

$$MLM.run(Array(dfSales, dfSponsor), output = "resMLM", session)$$

Suppose we already have the two relations stored in dfSales and dfSponsor. In the first line, we build an executable object of MLM. Then, we set the schema mappings for two relations in the next two lines. To operate the data and persist to resMLM file, we use the forth line with the run function. The run function of previous TC application only expects one relation as the input. While dealing with multiple relations, we aggregate the relations in an array as the input. The schema mapping is implemented by adding a new function for the second relation. However, it is possible to maintain the schema mapping for each relation in a hash function (h_r) and use another hash function with the relation's name as the key and the schema mapping information (the hash function, h_r) as the value.

5.3.3 LLib Categories and an Example of Machine Learning with LLib

The supported common Datalog algorithms and utilities in LLib can be categorized into five groups including the graph algorithms (Connected Components, Count Paths, etc.), machine learning algorithms (Linear Regression, Logistic Regression, etc.), financial applications (MLM Bonus, Bill of Materials, etc.), temporal database queries (like Interval Coalesce) and other applications. Although the machine learning algorithms already exist in Spark MLlib, we also provide the Datalog version in LLib because in our another work, it has been proven that the Datalog implementation can bring performance boost.

The working pipelines of all those libraries are unified and similar to the previous introduced TC and MLM example in Section 5.3.2. Here, we describe one machine learning example (Example 3) to train a linear regression model with LLib as following. In this example, the training data is stored in the verticalized format (a format often used for sparse data processing), whose schema is (item, column, value, prediction). The *item* represents the training instance. The *column* and *value* record each item's feature ID and feature value, while the *prediction* is the estimated value. To train on the loaded DataFrame (*dfTrain*), we could simply establish a training object *lr*, which wraps all the computation rules for a Datalog version of linear regression. We can also set the properties like learning rate or initial values of parameters before training. And the trained model (parameters values) could be either stored to a file (lr.run) or to a DataFrame (lr.genDF) for further usage. It can be observed that the LLib's machine learning library usage is as succinct as the usage of machine learning libraries in Spark MLlib.

Example 3 - Training of Linear Regression model with LLib.

$$
\begin{aligned}
&// \ Import\ data.\\
&var\ vs = StructType(List(StructField(\text{"}item\text{"}, IntegerType, true),\\
&StructField(\text{"}column\text{"}, IntegerType, true), StructField(\text{"}value\text{"}, DoubleType, true),\\
&StructField(\text{"}prediction\text{"}, IntegerType, true)))\\
&var\ dfTrain = spark.read.format(\text{"}csv\text{"}).schema(vs).load(\text{"}dataV\text{"})\\
&\\
&// \ Training\ on\ the\ input\ relation\ df.\\
&import\ edu.ucla.cs.wis.bigdatalog.spark.LLib.LL\_LinearRegression\\
&val\ lr = new\ LL\_LinearRegression().setMaxIter(10).setLearningRate(lr = 0.01).setInitial(0)\\
&lr.run(dfTrain, output = \text{"}model.out\text{"}, session)
\end{aligned}
$$

5.3.4 Extension of LLib

Besides a wide range of custom recursive algorithms, LLib also provides a unified template, *TempLib*, to follow when followers contribute new algorithms. TempLib contains abundant utility functions to facilitate development and some requirements to follow. With the template, the implementation becomes quite uncomplicated. As long as users have the Datalog rules for the algorithm, they can add their own algorithm by doing minor changes to the existing codes for algorithms like TC.

5.3.5 Collaboration with Other Applications

The input and output relations in LLib can be both DataFrames, which makes it possible to add a preprocessing algorithm which generates the input DataFrames or a subsequent processing algo-

rithm which takes the DataFrame generated from LLib. LLib allows users to exploit the Datalog application as a simple step at any place of their processing pipelines. In this section, we illustrate with a concrete example to show the collaborations between LLib and other applications like Spark MLlib or Spark DataFrame Operations. The showed LLib application is the MLM Bonus application mentioned previously, and we want to use the linear regression library from MLlib to get the relation between the bonus and the working days for a member. To save space, we do not show the process to build the session (LLibSession), and load input relations (dfSales, dfSponsor, dfWorkTime). The dfWorkTime stores the working days for each member. With MLM class from LLib, we could get the *dfRes* storing the bonus for each member. Joining the result relation with the dfWorkTime on the member column will generate a new relation with columns of the day and bonus. The Linear Regression function from MLlib could train on the result DataFrame. In this example, we can find the LLib is able to collaborate fluently with other operations in Spark like MLlib (LinearRegression), DataFrame Operations (join). Similarly, the collaboration among multiple Datalog applications is also possible.

Example 4 - Collaboration between LLib and other Spark libraries.

$$val\ MLM = new\ MLM()$$

$$MLM.setDirection(MCol = "Member", ProfitCol = "Bonus")$$

$$MLM.setSecDirection(MCol = "Member1", M2Col = "Member2")$$

$$val\ dfRes = Delivery.genDF(Array(dfSales, dfSponsor), session)$$

$$val\ dfRes = dfRes.join(dfWorkTime)$$

$$val\ parsedData = dfRes.rdd.map(row =>$$

$$LabeledPoint(row.getAs[Int]("Days"), Vectors.dense((row.getAs[Int]("Bonus")))))$$

$$val\ numIterations = 10$$

$$val\ stepSize = 0.00000001$$

$$val\ model = LinearRegressionWithSGD.train(parsedData, numIterations, stepSize)$$

$$model.save(session.sparkContext, "scalaLinearRegressionWithSGDModel")$$

5.3.6 User Defined Datalog Function

While developing, users may want to define one function (user-defined function) for once and call it many times for modular programming. To serve the user-defined datalog function (UDDF), we provide two general classes, SingleTableUDDF and MultipleTablesUDDF, which wrap the necessary components to execute the Datalog queries on single relation or multiple relations. While utilizing the two classes to specify new Datalog functions, the only required information are Datalog rules and the schema of the basic table (or input table).

5.4 LFrame

DataFrame, a widely used data structure, is always supported in different data analytic facilities like Spark SQL, Python, but not in Datalog platforms. In this section, we propose a DataFrame-like object, LFrame, wrapping the Datalog transactions and the common DataFrame transactions.

5.4.1 Conversion from DataFrame to LFrame

To endow a DataFrame with the ability of logical operations, we build a bridge for converting a DataFrame to LFrame. As shown in Example 4.1 below, the entry point of the application using LFrame is the LLibSession, same as the one in LLib, which makes it possible to use both LLib and LFrame within one execution environment. The session can load the data to a DataFrame variable df, but the df cannot execute any logical transaction. To construct an LFrame variable with the df, a built-in function wrapperDF in LLibSession can be utilized with the session and df as input arguments.

Example 5.1 - Development with LFrame: Conversion from DataFrame to LFrame.

$$val\ session = LLibSession.builder().appName(\text{"}LFrame\text{"}).master(\text{"}local[*]\text{"}).getOrCreate()$$

$$var\ schema = StructType(List(StructField(\text{"}Parent\text{"}, IntegerType, true),$$

$$StructField(\text{"}Child\text{"}, IntegerType, true)))$$

$$var\ df = session.read.format(\text{"}csv\text{"}).option(\text{"}header\text{"}, \text{"}false\text{"}).schema(schema).load(\text{"}sg\text{"})$$

$$var\ lframe = LLibSession.wrapperDF(df, session)$$

5.4.2 LFrame: Unary Operation

The constructed LFrame in the Example 4.1 could support various Datalog operations that can be categorized to unary operations and N-ary operations. The unary operation only involves one LFrame object each time. With the lframe, we implement the SG application as following to show

the typical unary operations.

Example 5.2 - Development with LFrame: Unary operations with SG application.

1 : $lframe = lframe.registerCurLFrame(\text{"}lframe1(Parent : integer, Chile : integer)\text{"})$

2 : $lframe = lframe.rules(\text{"}sg(X,Y) \leftarrow lframe1(Parent,X), lframe1(Parent,Y), X = Y\text{"})$

3 : $lframe = lframe.rules(\text{"}sg(X,Y) \leftarrow rel(A,X), sg(A,B), rel(B,Y)\text{"})$

4 : $lframe = lframe.rules(\text{"}sg(X,Y) \leftarrow sg(X,Y)\text{"})$

5 : $lframe = lframe.delRule(3)$

6 : $lframe = lframe.query(\text{"}sg(X,Y)\text{"})$

7 : $// \ Execution \ and \ persistence.$

8 : $lframe.run(output = \text{"}SG\text{"})$

9 : $val \ dfRes = lframe.genDF()$

10 : $val \ rddRes = lframe.genRDD()$

11 : $// \ Normal \ DataFrame \ operations.$

12 : $lframe.nonRecursive().where(\text{"}X < 10\text{"}).select(\text{"}X\text{"}).collect().foreach(println)$

There are five categories of unary operations included in the above example.

- Registering one LFrame as a base relation. In line 1, the registerCurLFrame is to register the lframe as a base relation so that the Datalog rules can utilize it as a given dataset. While registering, users are free to rename the dataset's columns and LFrame will map the columns according to the order of appearance.
- Appending (or removing) rules. The main body (line 2 to 4) of a Datalog program is a finite set of rules. To state the rules, the function called "rules" can be exploited. Whenever

the function is utilized, a new rule will be appended to the existing rule sets owned by the lframe. If one rule is wrongly appended, it can be removed by the delRule (line 5) function using its index. The provided index variable can be an array to remove more than one rule.

- Specifying the output relation. To evaluate the application, the query function (line 6) assists to point out the output relation, which is the sg relation in the SG. The generated result LFrame will store the output dataset in the schema specified in query, (X, Y).

- Lazy execution and persistence. In DataFrame, the evaluation is lazy. Similarly, the LFrame's evaluation is lazy until the run action is triggered. As shown in the second part of the code (line 8 to 10), the run function will store the result relation to a file and if a user wants to restore to a DataFrame or RDD, genDF or genRDD function can be considered.

- Non-recursive transactions. The design of LFrame is to wrap both the Datalog transactions and the normal DataFrame transactions. When a normal DataFrame transaction is required, like the line 12 shows, the nonRecusive function will convert LFrame to the normal DataFrame for all the DataFrame operators afterwards.

Compare the LFrame and DataFrame, we can find they work in analogous ways but LFrame can support more operations. A more compact way to express the transactions from line 2 to 7 is to list them back to the lframe one by one like in line 15.

5.4.3 LFrame: N-ary Operation

In the previous SG example, only one relation is contained in the Datalog application, while in this section, we illustrate the transaction (registerMoreDFs) embroiling more relations like the join operator of the relational database. We adopt the MLM Bonus application as a running example (introduced in Example 2), for it contains both the sales and the sponsor relations.

Example 5.3 - Development with LFrame: *N*-ary operations with MLM Bonus application.

1 : $var\ lfSales = DatalogLibSession.wrapperDF(dfSales, session)$

2 : $lfSales = lfSales.registerCurDF(\text{"}sales(Member : integer, Bonus : integer)\text{"})$

3 : $lfSales = lfSales.registerMoreDFs(otherDF = Array(dfSponsor),$

4 : $registers = Array(\text{"}sponsor(Member1 : integer, Member2 : integer)\text{"}))$

5 : $lfSales = lfSales.rules(\text{"}bonus(M, msum < (M, B) >)$

6 : $\leftarrow sales(M, P), B = (P * 0.1)\text{"})$

7 : $lfSales = lfSales.rules(\text{"}bonus(M1, msum < (M2, B) >)$

8 : $\leftarrow bonus(M2, B2), sponsor(M1, M2), B = (B2 * 0.5)\text{"})$

9 : $lfSales = lfSales.query(\text{"}bonus(M, B)\text{"})$

10 : $lfSales.run(output = \text{"}bonus\text{"})$

In the example, we do not show the process of establishing the execution environment (session) or loading data to DataFrame to save space. We have two DataFrame objects dfSales and dfSponsor. The dfSales is converted to an LFrame object with Datalog functionalities encapsulated. To exploit the relation as the base relation, the registerCurDF function is exploited in line 2. To involve more datasets (dfSponsor), the registerMoreDFs function (line 3) is utilized. The input parameters include an array of DataFrames to be registered and another array of schemas exploited to register them. The two arrays are ordered and have one-to-one correspondence. Thereupon, the rules can take these DataFrames as the base relations to use. To construct the rule sets, the rules function is adopted multiple times. Eventually, the output relation specified by query function will be stored to the address contained in the run function (line 9 to 10).

5.5 Multi-language Programming

To make our interface more appealing, we consider supporting multiple programming languages including Python, Scala and Java. The LLib and LFrame is written in Scala and so Scala is the default interface. Since Java is interoperable with Scala, it is relatively simple to support Java LLib (*JLLib*) and Java LFrame (*JLFrame*). The remaining gaps between Scala and Java version are mainly the conversions of data collections. We bridge these gaps via using collections known by both languages in Scala implementation and using the implicit converting mechanism in Scala.

For the Python version of LLib (*PLLib*) and LFrame (*PLFrame*), we utilize Py4J (Dagenais, 2009), a bridge between Java and Python language, and design a Removal and Recovery mechanism to transfer complex objects with schema information. With the gateway module, Py4J enables the Python interpreter to transfer objects to or access objects from the JVM. Follow the previous design, both the PLLib and PLFrame should support operations on PySpark DataFrame layer. But for the DataFrame object, which is complex and contains schema information, the Py4J can hardly transfer it. We design a Removal and Recovery mechanism, where the transferred DataFrame in PySpark will be converted to a common data collection type acceptable by both Java and Python languages. Then, during execution in Java, we could infer the schema or directly get the schema from users. For example, the PLFrame supports the operation to register a relation, which allows users to provide the schema information of the relation. Since the interfaces of Python and Java are designed similarly, we do not show more examples.

5.6 Performance Overhead

With LLib or LFrame, users can conveniently exploit existing Datalog algorithm or develop new recursive algorithm during the data processing pipeline. We believe the convenience is not based on huge performance sacrifice. We perform experiments on several custom Datalog algorithms on a machine with Ubuntu 14.04 LTS, an Intel i7-4770 CPU (3.40GHz, 4 core/8 thread), 32GB memory and a 1 TB 7200 RPM hard drive. The conducted experiments include SSSP, Reach, Connected Components (CC) and Same Generation (SG). SSSP is introduced previously in Query

1. Reach is to find all nodes which are reachable from a given source node. Connected Components is to identify the connected components of a graph. And Same Generation looks for the pairs of nodes in a tree-like structure with the same distance to a common ancestor.

Table 5.1: Execution time comparison for typical Datalog algorithms.

Algorithm	Dataset	BD (s)	LLib (s)	RelDiff(%)	Dataset	BD (s)	LLib (s)	RelDiff(%)
SSSP	RMAT-1M	19.64	21.2	7.94	RMAT-4M	61.5	64.3	4.55
Reach		6.9	7.9	14.49		15.07	16.7	10.82
CC		12.7	13.42	5.67		36.63	38.6	5.38
SG	Grid-150	18.07	18.95	4.87	Grid-250	36.72	38.28	4.25

The exploited datasets include RMAT-1M, RMAT-4M, Grid-150 and Grid-250. RMAT-1M and RMAT-4M are synthetic graphs generated with GTgraph generator (Bader and Madduri, 2006) using parameters (a, b, c) = (0.45, 0.25, 0.15). RMAT-n (n =1, 4 Million) contains n nodes and 10n directed edges with uniform integer weights range from [0,100). Grid-150 is a 151 by 151 grid with 45,300 edges and Grid-250 is a 251 by 251 grid with 125,500 edges. Comparing the BigDatalog (BD) (Shkapsky et al., 2016) and LLib execution time in Table 5.1, we can find the performance overhead brought by LLib is always minor. The overhead could come from steps like DataFrame schema mapping and recovering for a general design. With a larger dataset (e.g. RMAT-4M or Grid-250), there is a trend of relatively smaller overhead considering the RelDiff.

5.7 Conclusion

In this chapter, we have shown the LLib to encapsulate typical Datalog algorithms, which follows the data scientists' customs and requires less logical programming expertise. For the possible extensions of LLib, we provide not only a unified template for normalizing the contribution, but also some utilities to simplify the extending process. With some running examples, we demonstrate the benefit of designing the LLib as a DataFrame-based API is that the Datalog algorithm can flexibly collaborate with other Spark operations. For the audience who has more logical programming experience and would like to design their own applications, we also provide an LFrame API containing both logical and DataFrame-based operations. A simple conversion function between the DataFrame and LFrame is also provided. Both LFrame and LLib make the Datalog as an integral

part of data processing pipeline and are implemented to support general programming languages like Python, Scala and Java. Moving forward, we have plenty of data processing algorithms and logical oeprations to add for the LLib and LFrame separately. For example, the feature extraction (Ramírez-Gallego et al., 2017) or data cleaning (Wang et al., 2016c) or entity extraction (Wang et al., 2019, 2020) algorithms for preprocessing. Further plans call for the implementation of both our interfaces on other distributed platforms and across various platforms in ways akin to those of the polystore system (Duggan et al., 2015).

CHAPTER 6

Conclusions and Future Work

In this dissertation, we have presented models to extract information with involvement of unused data for the optimal downstream task performance. We also discussed the approaches to improve the efficiency of developing analytics algorithms with complex recursions and training large-class classifiers for the big data. In this section, we will wrap up the dissertation and suggest some avenues for future research.

In Chapter 2, we discussed the way to include the IP address sequences in the internet embedding by deep learning, which we found could tackle the issue of missing information for locating unknown IP addresses and provide sufficient coverage of protected networks in denial-of-service attacks. We discovered the hidden structural information encoded in a node's IPv4 address. To extract the encoded information, we designed a deep learning based framework, DIP, which is a ten-layer neural network and a variant of RNN. During the training of DIP, we jointly utilized the IP address, hop count and routing information and used the distance estimation as the objective. With experiments on test data, we found the learned embedding vectors could preserve the real-world clustering of the associated nodes and predict distance between them accurately. Moreover, for unknown hosts, DIP could accurately impute hop count distance to them merely by their IP addresses and routable prefixes. These findings inspired us to apply DIP on hop count filtering based spoofing detection, a classical framework in network security. We reviewed the previous design of spoofing detection with the explicitly computed IP maps, the collection of immutable structural network properties among IPs like hop counts. Since DIP could learn the embedding of internet and predict the structural properties of arbitrary IPs, the DIP-based detection mechanism could save the time of measuring the structural properties to build up IP maps and increase the coverage of protected hosts.

In Chapter 3, we discussed the way to include the 2^{nd}-n^{th} best hypotheses generated by ASR module in SLU pipeline, which we found could improve the SLU system robustness to ASR errors. In a conventional SLU system, the ASR module transcribes the input speech to sentences (hypotheses) and the hypothesis with highest confidence score will be transferred to natural language understanding module. We argued that solely relying on the best hypothesis could be erroneous, which could be revealed from the spoken recognition quality distribution. We reviewed the existing approaches to utilize n hypotheses, among which the reranking model is the most popular one. With motivating examples, we demonstrated the condition that reranking model cannot figure out. To tackle the issue, we tried integrating the hypothesized texts and embedding vectors in numerous models. Among the developed models, the PoolingAvg, which concatenated the embedding vectors and used a average pooling layer to generate a unified vector, outperformed all the others. The PoolingAvg achieved significant classification accuracy improvement for downstream tasks including domain classification and intent classification. We also observed that with more hypotheses combined, the performance could be further improved.

In Chapter 4, we addressed the expensive computational cost to train a large-class classifier for a big data set. When a classifier is modeled as a neural network, it is always represented by a softmax layer. The softmax layer training is the main reason for the high cost due to its intractable normalization constant. We gave a thorough overview of the negative sampling, hierarchical softmax, adaptive softmax to approximate the softmax. We proposed the amplified negative sampling by introducing the amplifying factor to the known negative sampling. With experiments on real-world datasets and tasks, we showed the efficiency of the amplified negative sampling on both the sampling cost savings and performance boosting.

In Chapter 5, we demonstrated an expressive interface for succinct development of complex analytics with recursions. We introduced the superiority of Datalog systems on expressing recursive queries and overviewed the latest Datalog systems, including BigDatalog, RaSQL. Although much effort has been taken to optimize the scalable logical operations within distributed Datalog environment and to design a better interface to extend the expressive power, there is still space to improve the usability considering ”normal” data analysts. We provide the cross-language libraries, i.e. LLib and LFrame. The LLib, similar to Spark MLlib, encapsulates common Datalog applica-

tions for an end-to-end development. The LFrame, an extension to Spark DataFrame, supported both relational and logical operations. With running examples, we showed the interface helped data scientists efficiently develop succinct recursive analyzing algorithms with a familiar environment.

All together, we are really excited about all the progress made in the big data analytics systems and algorithms and glad to be able to contribute to this. We do think there is still a long way to go and would like to point out the avenues for future research in our mind.

The first direction is to use more available but unused information. In the dissertation, we introduce the unrecognized information in conventional algorithms of two domains, however we believe this is common in other domains. Here, we would like to keep sharing our other observations of the two domains and hope these could encourage more upcoming research works in those areas or more.

As for the SLU, the first type of information can be helpful is the acoustic-model information like confidence score, which is ignored in our current hypotheses integration models but shown to be informative in other speech applications (Kumar et al., 2014; Fiscus, 1997). The recognized hypotheses from ASR module are associated with confidence scores, which tell the quality of each hypothesis. The confidence scores exist in different layers, for example, the confidence score of i^{th} best hypothesis or the confidence score of the j^{th} word in the i^{th} best hypothesis. The PoolingAvg approach treats each hypothesis equally although the quality of the hypotheses actually varies. We thus need to consider a new design to hierarchically involve the multi-layer acoustic-model information for a more efficient integration. The second direction is to use deep learning framework on word lattice (Liu et al., 2014) or confusion network (Hakkani-Tür et al., 2006; Tur et al., 2002). The hypotheses are derived from the word lattice or confusion network, so they may contain more information like times.

As for the internet embedding, since we have proposed the DIP and the neural network can be easily extended for other data sources, we could use the other latency measurements (Dabek et al., 2004) more than hop counts by simply changing the cost function of the DIP. In addition, the AS membership information could provide a coarse indication of locality of IPs (Eriksson et al., 2009). We could adapt the AS membership as another estimating error within the internet

embedding algorithm.

The second direction is to consider multi-task learning. Besides involving the extra information, there is also opportunity to consider more relevant tasks by multi-task learning (MTL), which generalizes the learned model and broadcasts the knowledge among multiple fields. MTL (Zhang and Yang, 2017; Liu et al., 2019; Caruana, 1997) is a widely used machine learning paradigm for training multiple related tasks in the same time. The superiority of MTL is to avoid overfitting and transfer knowledge. This could inspire us to add more tasks for the model training. For example, in SLU, we can consider a new task to reconstruct the ground-truth transcription using the hypotheses. The common tasks for hypothesis embedding model are natural language understanding tasks (domain or intent classification or slot filling), while the transcription reconstruction can help recover the error contained in hypotheses for a better understanding.

The third direction is to consider multi-platform data analytics. Our LLib and LFrame could support data scientists' familiar development with multiple programming languages, while another need to support multi-platform programs keeps increasing. For example, while detecting abnormal exchanges of stock market with real-time platforms such as Spark Streaming (Zaharia et al., 2013), users may be interested in retrieving historical stock data stored in a NoSQL databases like MongoDB (Chodorow, 2013). To support the cross-storage-system queries, the ploystore (Duggan et al., 2015) architecture was proposed. Following this trend, for more succinct recursive query analysis, we could try to build a Datalog-based polystore system, which should be extensible to various platforms like Spark, MongoDB, etc.

I hope this dissertation could inspire the research in deep learning and scalable data analytics. Furthermore, I hope the embedding approaches, training algorithms and libraries constructed could contribute to both academic and industrial applications.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283. 1

Alsubaiee, S., Altowim, Y., Altwaijry, H., Behm, A., Borkar, V., Bu, Y., Carey, M., Cetindil, I., Cheelangi, M., Faraaz, K., et al. (2014). Asterixdb: A scalable, open source bdms. *Proceedings of the VLDB Endowment*, 7(14):1905–1916. 71

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. 1

Ark IPv4. Ark ipv4 routed topology dataset. http://www.caida.org/data/active/ipv4_routed_24_top-ology_dataset.xml. 16, 24, 30, 38

Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Or, A., Rosen, J., Stoica, I., Wendell, P., Xin, R., and Zaharia, M. (2015a). Scaling spark in the real world: performance and usability. *Proceedings of the VLDB Endowment*, 8(12):1840–1843. 75

Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., Meng, X., Kaftan, T., Franklin, M. J., Ghodsi, A., et al. (2015b). Spark sql: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1383–1394. 74

Bader, D. A. and Madduri, K. (2006). Gtgraph : A synthetic graph generator suite. 91

Bai, Y., Goldman, S., and Zhang, L. (2017). Tapas: Two-pass approximate adaptive sampling for softmax. *arXiv preprint arXiv:1707.03073*. 53, 55

Baker, F. and Savola, P. (2004). Ingress filtering for multihomed networks. BCP 84, RFC Editor. http://www.rfc-editor.org/rfc/rfc3704.txt. 21, 23

Bakhtiary, A. H., Lapedriza, A., and Masip, D. (2015). Speeding up neural networks for large scale classification using wta hashing. *arXiv preprint arXiv:1504.07488*. 55

Barkan, O. and Koenigstein, N. (2016). Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE. 53, 56, 57

Bengio, Y. e. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722. 2, 52, 54

Blanc, G. and Rendle, S. (2017). Adaptive sampled softmax with kernel based sampling. *arXiv preprint arXiv:1712.00527*. 52, 54, 55

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bose, A. J., Ling, H., and Cao, Y. (2018). Adversarial contrastive estimation. *arXiv preprint arXiv:1805.03642*. 55

Bremler-Barr, A. and Levy, H. (2005). Spoofing Prevention Method. In *IEEE Infocom*. 21, 23

Bruzzone, L. and Prieto, D. F. (1999). An incremental-learning neural network for the classification of remote-sensing images. *Pattern Recognition Letters*. 20

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75. 42, 96

Chan, H. Y. and Woodland, P. (2004). Improving broadcast news transcription by lightly supervised discriminative training. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–737. IEEE. 41

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics. 41, 44

Chen, W., Grangier, D., and Auli, M. (2015). Strategies for training large vocabulary neural language models. *arXiv preprint arXiv:1512.04906*. 52

Chen, Y., Das, S., Dhar, P., Saddik, A. E., and Nayak, A. (2008). Detecting and preventing ip-spoofed distributed dos attacks. *International Journal of Network Security*, 7(1):70–81. 21

Chodorow, K. (2013). *MongoDB: the definitive guide: powerful and scalable data storage*. " O'Reilly Media, Inc.". 96

Cloudflare. The root cause of large DDoS IP Spoofing. https://blog.cloudflare.com/the-root-cause-of-large-ddos-ip-spoofing/. 21, 24

Collins, M., Roark, B., and Saraclar, M. (2005). Discriminative syntactic language modeling for speech recognition. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 507–514. Association for Computational Linguistics. 41

Consens, M. P. and Mendelzon, A. O. (1990). Low complexity aggregation in graphlog and datalog. In *International Conference on Database Theory*, pages 379–394. Springer. 71

Costa, M., Castro, M., Rowstron, A., and Key, P. (2004). PIC: Practical Internet coordinates for distance estimation. In *ICDCS*. 10, 24, 29

Cunha, I., Teixeira, R., Veitch, D., and Diot, C. (2011). Predicting and tracking internet path changes. In *ACM Sigcomm*. 37

Dabek, F., Cox, R., Kaashoek, F., and Morris, R. (2004). Vivaldi: a decentralized network coordinate system. In *SIGCOMM*. 7, 8, 10, 17, 19, 20, 21, 24, 29, 95

Dagenais, B. (2009). Py4j - a bridge between python and java. https://www.py4j.org/index.html. 90

Das, A., Li, Y., Wang, J., Li, M., and Zaniolo, C. (2019). Bigdata applications from graph analytics to machine learning by aggregates in recursion. *arXiv preprint arXiv:1909.08249*. 74

DDoS incident. February 28th DDoS incident. https://githubengineering.com/ddos-incident-report/. 21

De Brébisson, A. and Vincent, P. (2015). An exploration of softmax alternatives belonging to the spherical loss family. *arXiv preprint arXiv:1511.05042*. 55

Dean, J. and Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. 74

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 1

Dikici, E., Semerci, M., Saraçlar, M., and Alpaydin, E. (2012). Classification and ranking approaches to discriminative language modeling for asr. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):291–300. 41

Duan, Z., Yuan, X., and Chandrashekar, J. (2008). Controlling IP Spoofing through Interdomain Packet Filters. *IEEE Transactions on Dependable and Secure Computing*, 5(1). 21, 23

Duggan, J., Elmore, A. J., Stonebraker, M., Balazinska, M., Howe, B., Kepner, J., Madden, S., Maier, D., Mattson, T., and Zdonik, S. (2015). The bigdawg polystore system. *ACM Sigmod Record*, 44(2):11–16. 92, 96

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*. 8

Eriksson, B., Barford, P., and Nowak, R. (2008). Network Discovery from Passive Measurements. In *ACM Sigcomm*. 2, 7, 10

Eriksson, B., Barford, P., and Nowak, R. (2009). Estimating Hop Distance Between Arbitrary Host Pairs. In *IEEE Infocom*. 2, 8, 10, 18, 19, 20, 22, 24, 29, 38, 95

Eriksson, B., Barford, P., Nowak, R., and Crovella, M. (2007). Learning Network Structure from Passive Measurements. In *IMC*. 7

FastText (2019). Facebook fasttext project source code.

Feamster, N. and Rexford, J. (2017). Why (and how) networks should run themselves. *arXiv preprint arXiv:1710.11583*. 7

Feng, W., Kaiser, E., Feng, W., and Luu, A. (2005). The Design and Implementation of Network Puzzles. In *Infocom*. 21, 23

Ferguson, P. and Senie, D. (2000). Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing. BCP 38, RFC Editor. http://www.rfc-editor.org/rfc/rfc2827.txt. 21

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE. 95

Frankel, S. and Krishnan, S. (2011). IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap. RFC 6071, RFC Editor. http://www.rfc-editor.org/rfc/rfc6071.txt. 21, 23

Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*. 59

Gong, P., Ye, J., and Zhang, C. (2013). Multi-stage multi-task feature learning. *The Journal of Machine Learning Research*, 14(1):2979–3010. 42

Gonzalez, J. E., Xin, R. S., Dave, A., Crankshaw, D., Franklin, M. J., and Stoica, I. (2014). Graphx: Graph processing in a distributed dataflow framework. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 599–613. 74

Gonzalez, R., Manco, F., Garcia-Duran, A., Mendes, J., Huici, F., Niccolini, S., and Niepert, M. (2017). Net2vec: Deep learning for the network. In *ACM Big-DAMA'17*. 36

Goodman, J. Classes for fast maximum entropy training. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. 55

Grave, E., Joulin, A., Cissé, M., Jégou, H., et al. (2017). Efficient softmax approximation for gpus. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1302–1310. JMLR. org. 52

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE. 1

Grover, A. and Leskovec, J. (2016a). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 53, 57

Grover, A. and Leskovec, J. (2016b). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM. 56, 57

Gu, J., Watanabe, Y. H., Mazza, W. A., Shkapsky, A., Yang, M., Ding, L., and Zaniolo, C. (2019). Rasql: Greater power and performance for big data analytics with recursive-aggregate-sql on spark. In *Proceedings of the 2019 International Conference on Management of Data*, pages 467–484. ACM. 3, 71, 75

Gulzar, M. A., Interlandi, M., Han, X., Li, M., Condie, T., and Kim, M. (2017). Automated debugging in data-intensive scalable computing. In *Proceedings of the 2017 Symposium on Cloud Computing*, pages 520–534. 75

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. 55, 57

Hakkani-Tür, D., Béchet, F., Riccardi, G., and Tur, G. (2006). Beyond asr 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514. 51, 95

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 1

Interlandi, M., Ekmekji, A., Shah, K., Gulzar, M. A., Tetali, S. D., Kim, M., Millstein, T., and Condie, T. (2018). Adding data provenance support to apache spark. *The VLDB Journal*, 27(5):595–615. 75

Jin, C., Wang, H., and Shin, K. G. Hop-count filtering: An effective defense against spoofed ddos traffic. 21, 22, 24, 25, 30, 31, 32

Jin, C., Wang, H., and Shin, K. G. (2003). Hop-count filtering: An effective defense against spoofed ddos traffic. In *CCS*. 7, 11, 20

John, W., Dusi, M., and claffy, k. (2010). Estimating Routing Symmetry on Single Links by Passive Flow Measurements. In *ACM International Workshop on TRaffic Analysis and Classification (TRAC)*. 38

Jyothi, P., Johnson, L., Chelba, C., and Strope, B. (2012). Large-scale discriminative language model reranking for voice-search. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 41–49. Association for Computational Linguistics. 41

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 8

Killalea, T. (2000). Recommended internet service provider security services and procedures. BCP 46, RFC Editor. 21

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 10, 16

Kumar, K., Liu, C., and Gong, Y. (2014). Normalization of asr confidence classifier scores via confidence mapping. In *Fifteenth Annual Conference of the International Speech Communication Association*. 95

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*. 10

Levchenko, K., Dhamdhere, A., Huffaker, B., claffy, k., Allman, M., and Paxson, V. (2017). PacketLab: A Universal Measurement Endpoint Interface. In *Internet Measurement Conference (IMC)*. 7

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185. 60

Li, M., Lumezanu, C., Zong, B., and Chen, H. (2018a). Deep Learning IP Network Representations. In *ACM Sigcomm Big-DAMA*. xiii, 22, 25, 26, 27, 28, 31, 32, 38

Li, M., Lumezanu, C., Zong, B., and Chen, H. (2018b). Learning IP Network Representations. *SIGCOMM CCR*. 22, 26

Li, M., Ruan, W., Liu, X., Soldaini, L., Hamza, W., and Su, C. (2020). Improving spoken language understanding by exploiting asr n-best hypotheses. *arXiv preprint arXiv:2001.05284*. xiii, 40

Li, Y., Li, M., Ding, L., and Interlandi, M. (2018c). Rios: Runtime integrated optimizer for spark. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 275–287. 75

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*. 96

Liu, X., Wang, Y., Chen, X., Gales, M. J., and Woodland, P. C. (2014). Efficient lattice rescoring using recurrent neural network language models. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4908–4912. IEEE. 51, 95

Lone, Q., Luckie, M., Korczynski, M., and van Eeten, M. (2017). Using loops observed in traceroute to infer the ability to spoof. In *PAM*. 23, 30

Lumezanu, C., Levin, D., and Spring, N. (2007). PeerWise discovery and negotiation of faster paths. In *HotNets*. 21, 29, 37

Luong, T., Socher, R., and Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

McKinney, W. et al. Data structures for statistical computing in python. 72

McKinney, W. et al. (2010). Data structures for statistical computing in python. 75

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al. (2016). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 17(1):1235–1241. 74

Miao, R., Zeng, H., Kim, C., Lee, J., and Yu, M. (2017). SilkRoad: Making Stateful Layer-4 Load Balancing Fast and Cheap Using Switching ASICs. In *ACM Sigcomm*. 7

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 2, 52, 55, 56, 57, 65

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech*. 13

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*. 8

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 52, 53, 56, 60, 68

MLM (2008). Multi-level_marketing. 80

Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*. 57

Morbini, F., Audhkhasi, K., Artstein, R., Van Segbroeck, M., Sagae, K., Georgiou, P., Traum, D. R., and Narayanan, S. (2012). A reranking approach for recognition and classification of speech input in conversational dialogue systems. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 49–54. IEEE. 2, 41, 42, 44

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer. 2, 52, 54, 55

Mussmann, S., Levy, D., and Ermon, S. (2017). Fast amortized inference and learning in log-linear models with randomly perturbed nearest neighbor search. *arXiv preprint arXiv:1707.03372*. 55

Ng, T. S. E. and Zhang, H. (2002). Predicting Internet network distance with coordinates-based approaches. In *INFOCOM*. 10, 17, 19, 24, 29

Node2vec, G. (2019). Node2vec project souce code. https://github.com/snap-stanford/snap.git.

Oba, T., Hori, T., and Nakamura, A. (2007). An approach to efficient generation of high-accuracy and compact error-corrective models for speech recognition. In *Eighth Annual Conference of the International Speech Communication Association*. 42

Olston, C., Reed, B., Srivastava, U., Kumar, R., and Tomkins, A. (2008). Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1099–1110. ACM. 71

Padmanabhan, R., Dhamdhere, A., Aben, E., kc claffy, and Spring, N. (2016). Reasons dynamic addresses change. In *IMC*. 37

Park, K. and Lee, H. (2001). On the Effectiveness of Route-Based Packet Filtering for Distributed DoS Attack Prevention in Power-Law Internets. In *ACM Sigcomm*. 21

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. 1

Peng, F., Roy, S., Shahshahani, B., and Beaufays, F. (2013). Search results based n-best hypothesis rescoring with maximum entropy classification. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 422–427. IEEE. 2, 41, 42, 44

Pyxida. Pyxida. http://pyxida.sourceforge.net/. 10, 24, 29

Ramírez-Gallego, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Benítez, J. M., Alonso-Betanzos, A., and Herrera, F. (2017). An information theory-based feature selection framework for big data under apache spark. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9):1441–1453. 92

Rawat, A. S., Chen, J., Yu, F. X. X., Suresh, A. T., and Kumar, S. (2019). Sampled softmax with random fourier features. In *Advances in Neural Information Processing Systems*, pages 13834–13844. 2, 52, 54, 55

RouteViews. http://www.routeviews.org. 16

Ruiz, F. J. R., Titsias, M. K., Dieng, A. B., and Blei, D. M. (2018). Augment and reduce: Stochastic inference for large categorical distributions. 54

Sak, H., Saraclar, M., and Güngör, T. (2010). On-the-fly lattice rescoring for real-time automatic speech recognition. In *Eleventh annual conference of the international speech communication association*. 41

Sak, H., Saraclar, M., and Gungor, T. (2011a). Discriminative reranking of ASR hypotheses with morpholexical and n-best-list features. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011*, pages 202–207. 41

Sak, H., Saraçlar, M., and Güngör, T. (2011b). Discriminative reranking of asr hypotheses with morpholexical and n-best-list features. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 202–207. 41

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823. 55

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681. 43

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*. 43

Shkapsky, A., Yang, M., Interlandi, M., Chiu, H., Condie, T., and Zaniolo, C. (2016). Big data analytics with datalog queries on spark. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1135–1149. ACM. 3, 71, 75, 91

Shue, C., Shin, Y., Gupta, M., and Choi, J. Y. (2005). Analysis of ipsec overheads for vpn servers. In *1st IEEE ICNP Workshop on Secure Network Protocols, 2005. (NPSec)*. 21

Sommer, R. and Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*. 23

Spearman, C. (1904). The proof and measurement of association between two things. *American journal of Psychology*, 15(1):72–101.

Spoofing-ps. Port scanning techniques and the defense against them. https://www.sans.org/reading-room/whitepapers/auditing/port-scanning-techniques-defense-70. 21

Spoofing-state. State of IP Spoofing. https://spoofer.caida.org/summary.php. 23, 30

Spring, N., Mahajan, R., and Anderson, T. (2003). Quantifying the Causes of Path Inflation. In *ACM Sigcomm*. 7

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., and Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment*, 2(2):1626–1629. 71

Tur, G. and De Mori, R. (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons. 2, 41

Tur, G., Wright, J., Gorin, A., Riccardi, G., and Hakkani-Tür, D. (2002). Improving spoken language understanding using word confusion networks. In *Seventh International Conference on Spoken Language Processing*. 51, 95

Vijayanarasimhan, S., Shlens, J., Monga, R., and Yagnik, J. (2014). Deep networks with large output spaces. *arXiv preprint arXiv:1412.7479*. 55

Vincent, P., De Brébisson, A., and Bouthillier, X. (2015). Efficient exact gradient update for training deep networks with very large sparse targets. In *Advances in Neural Information Processing Systems*, pages 1108–1116. 54, 55

Wang, D., Cui, P., and Zhu, W. (2016a). Structural Deep Network Embedding. In *KDD*. 20

Wang, D., Cui, P., and Zhu, W. (2016b). Structural deep network embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 25, 37

Wang, H., Jin, C., and Shin, K. G. (2007). Defense against spoofed ip traffic using hop-count filtering. *IEEE/ACM Transactions on Networking*, 15:40–53. 21, 24, 29

Wang, H., Li, M., Bu, Y., Li, J., Gao, H., and Zhang, J. (2016c). Cleanix: A parallel big data cleaning system. *ACM SIGMOD Record*, 44(4):35–40. 92

Wang, J., Lin, C., Li, M., and Zaniolo, C. (2019). An efficient sliding window approach for approximate entity extraction with synonyms. 92

Wang, J., Lin, C., Li, M., and Zaniolo, C. (2020). Boosting approximate dictionary-based entity extraction with synonyms. *Information Sciences*. 92

Wang, L., Li, Y., Huang, J., and Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 8

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743. 53

Word2vec, G. (2019). Word2vec project source code. https://github.com/tmikolov/word2vec.git. 68

Yaar, A., Perrig, A., and Song, D. (2006). Stackpi: New packet marking and filtering mechanisms for ddos and ip spoofing defense. *IEEE Journal on Selected Areas in Communications*, 24(10):1853–1863. 21

Yang, M., Shkapsky, A., and Zaniolo, C. Parallel bottom-up evaluation of logic programs: Deals on shared-memory multicore machines. 3, 71

Zafarani, R. and Liu, H. (2009). Social computing data repository at ASU.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association. 71

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., Stoica, I., et al. Spark: Cluster computing with working sets. 1

Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., and Stoica, I. (2013). Discretized streams: Fault-tolerant streaming computation at scale. In *Proceedings of the twenty-fourth ACM symposium on operating systems principles*, pages 423–438. 96

Zaniolo, C., Das, A., Gu, J., Li, Y., Wang, J., et al. (2019). Monotonic properties of completed aggregates in recursive queries. *arXiv preprint arXiv:1910.08888*. 74

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. 96

Zhao, X., Sala, A., Zheng, H., and Zhao, B. Y. (2011). Efficient shortest paths on massive social graphs. In *CollaborateCom*. 10