Scaling Up Deep Learning on Clusters

Quanlai Li

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-74
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-74.html

May 12, 2017

Scaling Up Deep Learning on Clusters

by

Quanlai "Qualia" Li

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Master of Engineering

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jitendra Malik, Chair
Professor John F. Canny
Assistant Professor Joseph E. Gonzalez

Spring 2017

Scaling Up Deep Learning on Clusters

Quanlai “Qualia” Li

Contents

List of Figures

Acknowledgments

I would first like to thank my thesis advisor Prof. John Canny at UC Berkeley. The door to Prof. John Canny office was always open whenever I ran into a trouble spot or had a question about my research or writing.

I would also like to express my very profound gratitude to my project teammates Aleksandre Kamko, Max "Jiaqi" Xie and Abdelrahman Elbashandy. Without their passionate participation and input, the research could not be properly conducted.

I would also like to acknowledge Dr. Alex Beliaev for passionately guiding me towards the thesis, and Prof. Joseph Gonzalez at UC Berkeley as the second reader of this thesis.

Finally, I must express my very profound gratitude to my family and friends who are always supporting me.

Chapter 1

Technical Contribution

> The challenge of the 21st century is to find out what works and scale it up.[1]
>
> *Bill Clinton*

1.1 Introduction

Capstone Members

Our capstone project is Scaling Up Deep Learning on Clusters, advised by Professor John F. Canny[1], read by Assistant Professor Joseph E. Gonzalez[2]. The project is focused on a noval machine learning framework called BIDMach[3]. BIDMach is a component of BID Data

[1]https://people.eecs.berkeley.edu/ jfc/
[2]https://people.eecs.berkeley.edu/ jegonzal/
[3]https://github.com/BIDData/BIDMach/

Project[4], by Berkeley Institute of Design.

Our capstone team has three MEng students, Quanlai "Qualia" Li[5], Jiaqi "Max" Xie and Aleks Kamko. An undergraduate Kevin Peng and a software engineer Abdelrahman Elbashandy from Laurence Lab are also working for the project. Different backgrounds in Computer Science, Mathematics and Software Engineering help us better understand and solve the problems in this project.

BID Data Project

BID Data is an open-source[6] project initiated by Prof. John Canny. It offers resources for fast big data tools. BID Data is equipped with interactive environment, thus easy to build and use machine learning. BID Data suite has three major elements[2][3]:

1. Underlying hardware (e.g. CPU, GPU[7]).

2. Software:

    a) BIDMat, a matrix library[8] that takes care of data management, calculation and hardware acceleration.

    b) BIDMach, a machine learning framework that includes efficient algorithms[4].

[4]http://bid2.berkeley.edu/bid-data-project/

[5]https://www.liquanlai.com

[6]Open-source projects are projects that share their code online, and are sometimes open for modification and redistribution

[7]Central Processing Unit and Graphics Processing Unit, GPUs are good at matrix computation and thus widely used in machine learning

[8]A software library is a reusable programming component

3. Scaling Up:

    a) Butterfly Mixing: an efficient inter-machine communication strategy.

    b) Sparse AllReduce: a MapReduce[9] like primitive for scalable communication[5].

It has several features. One feature of BIDMach is its low cost of energy. OpenChai wants to provide a machine learning solution that runs on local machines using mobile processors, rather than on cloud. This solution will be more affordable, since mobile processors are cheaper. It will also be more secure for the enterprises and individuals who do not want to upload their data to cloud services controlled by big companies.

BIDMach before the Capstone

Before our capstone project started, BIDMach could run on one single machine. It supported several machine learning algorithms, like K-Means, Random Forest and General Linear Model[2]. We want to scale up these algorithms on a cluster of machines. BIDMach did not support Neural Network models completely at that time. For example, it did not support the Convolutional Layer, thus unable to process images. We need to complete different types of neural network by ourselves first, and then scale it up[6][7].

[9]MapReduce is a programming model that deals with big data operation

Objectives of the Capstone Project

We have two main objectives. The first is to scale up BIDMach on clusters. Spark[10], a communication layer, is used for this objective[4]. And the second is to cooperate with our industry partner, OpenChai[11] to provide machine learning solutions to enterprises and individuals (Li, 2016). The objectives can be broken down to several parts, as listed below:

1. Familiarize with system and tools (e.g. Spark framework, TX1 hardware, Scala[12] language).
2. Extend model-parallel algorithms (e.g. KMeans, Random Forest) to Spark.
3. Extend data-parallel algorithms (e.g. General Linear Model) to Spark.
4. Bring Neural Network algorithms to BIDMach (e.g. Convolutional Neural Network, Sequence to Sequence Model, etc.)
5. Implement the Mechanism to Send Worker Progress to Master
6. Benchmark algorithms on OpenChai hardware, TX1.
7. Optimize parallel algorithms on TX1.
8. Help marketing OpenChai's machine learning solution.

[10]http://spark.apache.org/
[11]http://openchai.com/
[12]https://www.scala-lang.org/

The purpose of this chapter is to demonstrate my technical contributions to this project, and the importance of the contributions. Specifically, I focused on objectives 1, part of 3, part of 4, and 5.

1.2 Elastic Average Stochastic Gradient Descent

Elastic Average Stochastic Gradient Descent is a mechanism that enables General Linear Model to run on a cluster of machines. Predictions and classifications are made much faster in this way.

Three members of our team worked together to accomplish the first three tasks. We got familiar with the system and tools, and then started extending machine learning algorithms on BIDMach. Aleks Kamko first focused on data-parallel algorithms[8]. Max worked on a function called ParCall, which is communication method for distributed machines, useful for model-parallel workers[9].

A distributed system (a.k.a. cluster) usually has one master machine and several worker machines. In our project, the master machine can assign tasks to worker machines. In model parallel models, there are worker-machine-wise communication.

In machine learning context, a matrix of weights is updated when the algorithm is taking more data. Most machine learning models want to find a representation of weights and find a direction to faster update the matrix.

Under General Linear Model, the master machine distributes a partition of data to every

worker machine and asks them to update the matrix using data. The challenge is to construct a communication method for the worker machines. GLM is indispensable for BIDMach because they are widely applied and effective. BIDMach will not be a legitimate machine learning framework if GLM is not supported.

Elastic Stochastic Gradient Descent (ESGD) is an algorithm to update the matrix for worker machines[10]. There is a master machine who periodically collects matrixes from worker machines and calculates the average. After each pass the master machine broadcasts the matrix and worker machines update their matrix elastically based on average matrix. With elasticity, a worker machines update its matrixes using a weighted average of previous matrix from itself and received matrix from the master.

We scrutinized the research by Zhang and implemented the ESGD function on our models. ESGD deserves our attention since it finds a balance between keeping each worker machine's independence and enabling them to communicate. According to our experience, models with ESGD have better prediction accuracy and lower running time.

Results

The major innovations of our capstone project are EASGD and robust communication framework[13]. These techniques are used for coordinating communication between machines. After applying these methods, we got a significant performance boost on multiple algorithms, as shown in figure 1.1.

[13]Discussed in Jiaqi's report

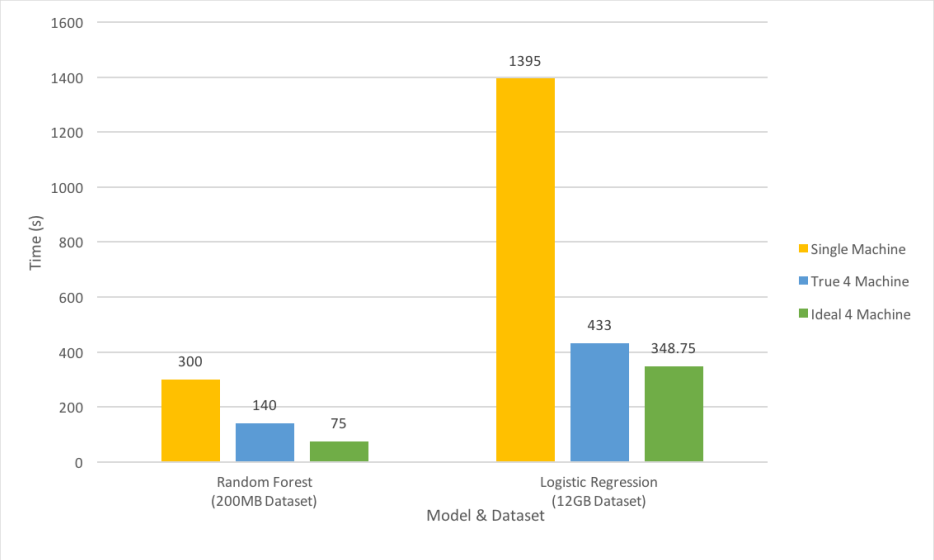

Figure 1.1: Scaling Up: 4 Distributed Machines vs. 1 Machine

While we increased the training speed much, we did not decrease accuracy.

1.3 Convolutional Neural Network

Description

Apart from General Linear Model, there are other effective machine learning models that we also want to run on BIDMach. Deep Neural Network, or the Deep Learning model is the most notable one.

We focused more on one type of Neural Network, Convolutional Neural Network (CNN). CNN is able to process images, with Convolutional Layer taking pixels of an image as input. CNN also consists other layers like Pooling Layer, ReLU Layer and Fully Connected

Layer. These layers process information in different ways. By combining these layers in different orders, we can orchestrate Convolutional Neural Network models for different image processing tasks[11][12].

Implementation of Convolutional Neural Network models is necessary, in that it can do high-performance image processing.

Implementation

Most of our work on CNN is to implement the Convolutional layer. Implementing Convolutional Layer requires a lot of computation. We are using an underlying library, BIDMat, which is also designed by Prof. Canny. BIDMat provides us with efficient matrix operation APIs[14].

I worked together with Jiaqi on this part. While doing this, we first read through the code of BIDMat and got a better comprehension of its underlying design, and found out useful APIs that could help us build BIDMach. Meanwhile, we referred to some code of other Neural Network layers. For example, Linear Layer was already implemented, and shared some similarity with Convolutional Layer. Later we utilized these APIs to write code for Convolutional Layers.

[14]Application Program Interface, used as standardized protocols and tools in programming

Scripts

In order to test the code written for Convolutional Layer. I wrote a script[15] to test it. With references to other scripts, I transplanted an image classification problem from TensorFlow[16] to BIDMach. An image prediction dataset, rcv1[17] is used for this task. Below is the design of my neural network:

(design of neural network)

With this script, we can do more than just testing. On the one hand, this script serves as an instruction for new users to design their CNN for another specific task. It is easier to deploy another CNN by just changing some lines in this script.

On the other hand, since we have the same network design, configuration (a.k.a. hyper-parameters) and dataset, we can compare our performance with that of TensorFlow.

Benchmarking of Convolutional Neural Network

Once we have finished this Convolutional Neural Network, we can benchmark BIDMach on some datasets like Cifar10 or Rcv1. Cifar10 is a standardized dataset with 60,000 labeled images. We can train BIDMach and other systems (e.g. MXNet, TensorFlow) on this dataset and compare our prediction accuracy, time-consumption and energy-consumption.

[15] A small piece of easy-modifiable code
[16] https://www.tensorflow.org/
[17] www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf

Other Neural Networks to Implement on BIDMach

Besides Convolutional Neural Network, we implemented other types of Neural Networks, like Sequence-to-Sequence (Seq2Seq) Model. This model has same type of input and output. A typical application is human language translation, where both input and output are human languages. Our team member Aleks puts more effort in this part.

1.4 Worker Progress Sending

Users of BIDMach may also want to have a better control and understanding of how BIDMach is working. Sending worker progress to master can give user who controls master more information about the whole system.

We want to provide basic information about worker progress (e.g. number of iterations, training and validation accuracy) and performance criteria (e.g. current calculation speed and data throughput). Calculation speed is measured in GFLOPS[18], data throughput (a.k.a. bandwidth) is measured in MB/S.

There are several components in BIDMach related to data throughput. Technically, java classes like Machine, Learner, Worker, Master are all related to throughput. I first needed to decide with class to put my measurement in. Machine is a class at lower level of this systems, and has more interaction with the underlying communication layers (e.g. Spark, or Direct Memory Access, DMI). Thus, I decided to put my measurement in this part.

[18]Number of billion floating point operations per second

Class Machine in BIDMach is running asynchronously, which means different Machine instances are running at the same time, and could influence each other. This design is faster for the whole system, but made it harder for me to evaluate the throughput, since the running time cannot be precisely calculated. By gauging the total number of nanoseconds spent during a socket data transmission, I got the overall time spent on one transmission. I could also get the number of bytes transmitted one time. Therefore, calculating the average bandwidth (a.k.a. throughput speed) is possible by doing the division.

1.5 Conclusion

In this chapter, I introduced the tasks of our capstone project first. Then I discussed my accomplishments and solved problems. To be specific, I tried to implement elastic average gradient descent, convolutional neural network script, and built a mechanism to send the worker progress to master machine.

Meanwhile, I demonstrated the validity of my contributions. Convolutional Neural Network is an essential part of deep learning able to processing images. Elastic gradient descent enables General Linear Model to run on a cluster of machines, which speeds up classification and prediction. Sending worker progress to master can help users better analyze the status of the whole system. Users can further improve the design of the machine learning model according to the status.

To move one step forward, I would like to collaborate with Prof. Canny and other

teammates, and construct an algorithm running on a master machine that sends timeout value to worker machines with regard to its progress. This will make the whole system more efficient.

1.6 Appendix

Machine Learning and Deep Learning

Machine learning is a way to solve artificial intelligence problems. It usually has a mathematical or statistical model and takes a large amount of data as input. A machine learning problem generally has two phases, training phase and predicting phase. Training is to fit the model to data, in other words, to change the numbers (a.k.a. weights) in the model with regard to data, in order to better represent the real world. Predicting is to calculate the value given a weighted model and data. While predicting is trivial, training takes a longer time and needs more human work to optimize[12][13][14].

Distributed System and Scalability

A distributed system, or a cluster, is a group of computers working on one task. With large number of small machines, a distributed system could have a synergized computational power. Scalability is how the computational power increases with regard to the number of machines. Ideally the computational power could be proportional to the number of machines.

However, oftentimes it is not the case. For some tasks (e.g. machine learning algorithms), good scalability is a research topic.

Chapter 2

Engineering Leadership

> A breakthrough in machine learning would be worth ten Microsofts.[15]
>
> *Bill Gates*

2.1 Introduction

"Big Data" is a growing trend in the world of technology and business. Enormous amounts of data is collected, stored, and ultimately analyzed. Companies are in eager need of large - scale Machine Learning. The Data Processing industry comes as a necessity, providing data storage and the fast - growing cloud computing services.

However, several problems emerge: computing power efficiency, energy consumption, and data privacy are some of the most important issues.

BID Data is a toolkit for machine learning developed by our advisor - Prof. John Canny. It has the potential to alleviate these problems. BID Data is currently the fastest Big Data tool running on single machine, with a high energy efficiency[16]. Our team will be scaling the toolkit to work on clusters, while maintaining its advantages in the field.

This chapter addresses how this capstone project will make a significant impact to the industry of Machine Learning and Big Data, in the sense of improving speed and efficiency of computation, conciliating customers' concerns on privacy, and helping companies solve big data analytics problems.

2.2 Machine Learning Trends

Booming Machine Learning Industry

After six decades of research since its conception, artificial intelligence (i.e. machine learning) is receiving unprecedented attention. Leading technology giants, like Google, Microsoft, and Uber[17] are in intense competition with each other to build the most intelligent systems. Search engines, autonomous vehicles, language translation services, and even more are becoming more advanced every day[18].

Other industries besides software are also catching up by integrating machine learning algorithms into their products and services. These newcomers, ranging from the financial industry to the manufacturing industry, are increasingly investing in AI[19].

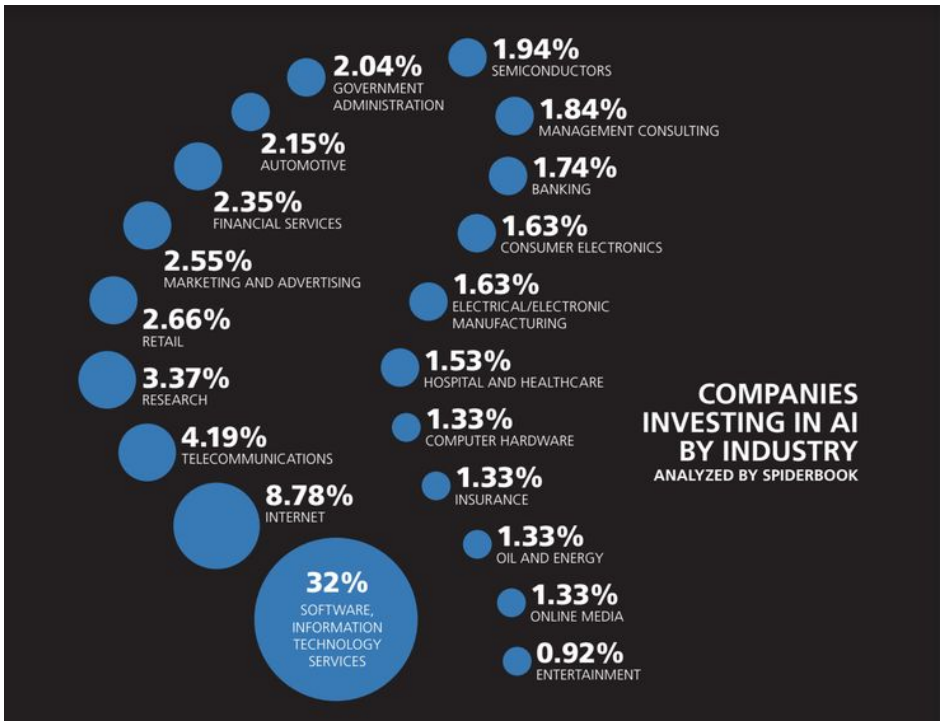

Figure 2.1: Companies Investing In AI[19]

These investments seem to be worthwhile, driven by lucrative projected revenues. A market forecast by Tractica shows the momentum of artificial intelligence revenue in the following decade[20].

Naturally, the booming of machine learning and artificial intelligence necessitates more research into better methods, models, and algorithms.

New Machine Learning Research Topics

To meet industry needs, machine learning models are becoming more sophisticated. This fact is especially apparent with the recent popularity in neural networks. A popular computer

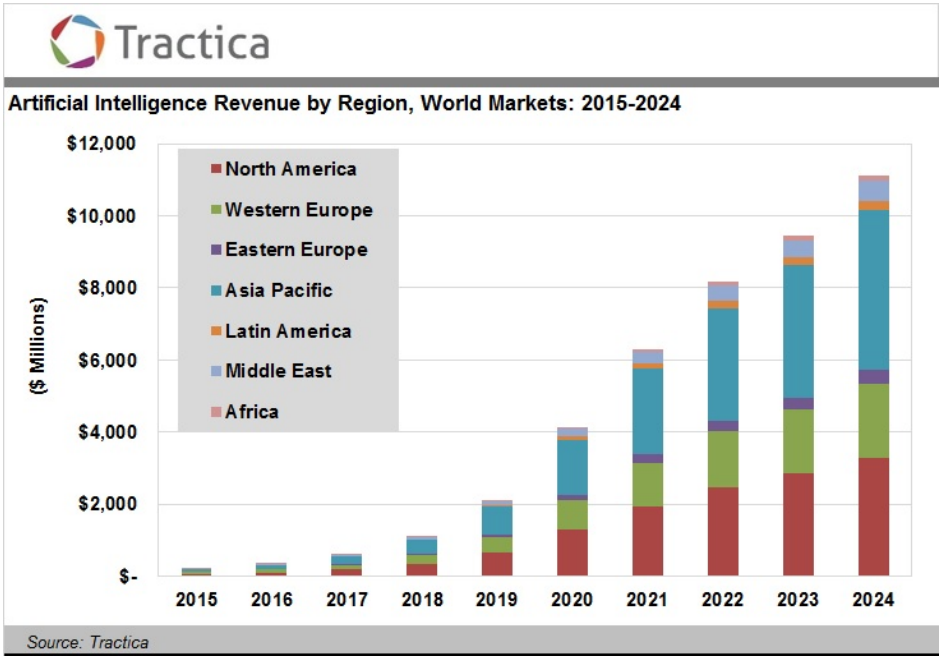

Figure 2.2: AI Revenue by Region[20]

vision competition, the ImageNet challenge, shows an increase in the depth (i.e. complexity) of neural networks, correlated with a significant decrease in classification error[21].

This increasing complexity calls for better utilizations and management of computational resources.

Traditional machine learning algorithms are made faster by hand - tuning algorithms and by updating hardware resources (e.g. leveraging a new GPU's computation power). However, these methods are reaching their limits. Efficient scaling, an optimization technique at the crossroads of algorithm and hardware improvement, is gaining traction.

An algorithm's scalability is a measure of how well it is able to run on a distributed system, like cluster of machines. A machine learning model with perfect scalability can

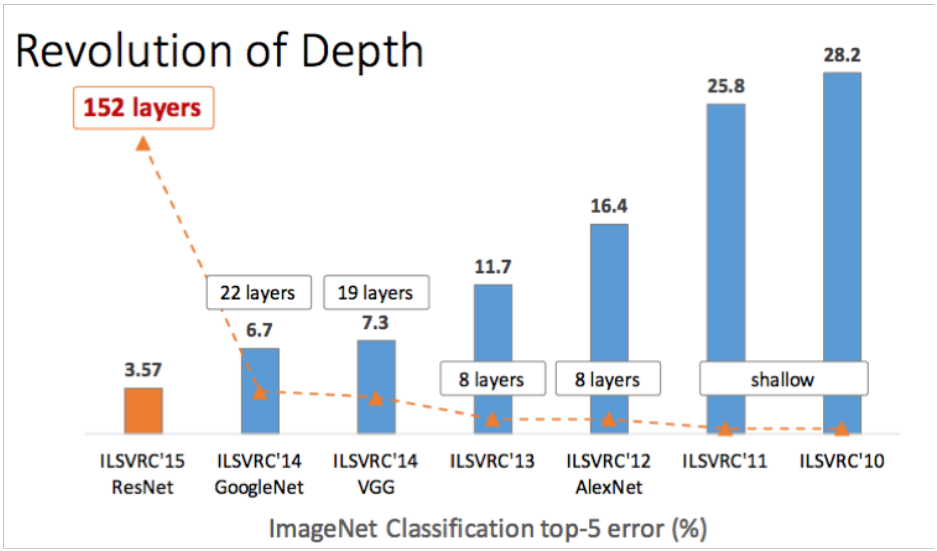

Figure 2.3: Revolution of Depth[22]

run at a speed proportional to the size of cluster. Scalable machine learning would allow us to learn from massive datasets at unprecedented speeds, enabling us to solve seemingly unsolvable problems[23].

2.3 Big Data Trends

Big Data

The Big Data industry lies at an intersection of Business Analytics and Technology. Analytics teams of large companies have been using data mining and other predictive analytic techniques for a long time[24]. With the rapid development of the Internet of Things industry in the 21st century, massive volumes of data - 50,000GB per second - are being created every day[25]. Companies leverage this data by using powerful Machine Learning algorithms

to extract meaningful information, creating real business value. As predicted by McKinsey Global Institute 5 years ago, "Big data [is becoming] a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus"[26].

We anticipate a significant growth potential in the data analytics market. U.S. Industry Report predicts that in 5 years, increasingly powerful computing technology will drive revenues for the data analytics industry to $53.9 billion, with an annual increasing rate of 5.5%[24]. Consequently, today is an opportune time to make an impact in the industry.

Data Center and Cloud Service

The Data Center industry is long established, helping companies store and process data since the 1950s. However, in the current era of Big Data, data centers have been evolving to fit a booming demand, resulting in modern day Cloud Service Providers. These providers modularly rent out their networked data center machines for expensive computing tasks.

The Cloud Service Provider industry is in rapid development, and its customers have a variety of interesting requirements[24]. Mainstream cloud solutions are far from perfect. Our capstone team aims to improve these services.

Problems

The first challenge our project attempts to address is computational throughput maximization. One advantage of a cloud compute cluster is its higher computational capability

compared to a single machine. In theory, a cluster of 1,000 computers could achieve a peak performance equivalent to 1,000 times that of a single machine. However, in practice, this is not the case. Communication and synchronization bottlenecks between the machines in a cluster cause latency and reduce the overall computing speed. This problem is exacerbated as data sizes grow and the system is scaled to more machines, causing diminishing improvement. By maximizing network throughput and lowering communication overhead, our capstone is able to improve upon the status quo.

Power consumption and energy waste in data centers is a second challenge. An environmental action organization - NRDC[1] - pointed out the problem in a recent report stating that, in 2013 alone, U.S. data centers used an amount of energy equivalent to the annual output of 34 large [2] power plants. This amount of energy could be used to provide two years'worth of power for all of New York City's households[27]. Pierre Delforge, an expert on energy efficiency from NRDC, claims that the Data Center industry is "one of the few large industrial electricity uses which are growing"[28]. This is growth in energy consumption is likely caused by the increasing growth in the size of datasets, so the problem will continue to compound unless preventative measures are taken. Our capstone project aims to alleviate energy waste by utilizing data centers more efficiently, requiring fewer machines to do the same data analytics and therefore using less energy.

Finally, data security and privacy is also becoming an important issue in this emerging

[1]Natural Resources Defense Council
[2]500 - megawatt

industry. As stated previously, companies collect massive amounts of data in order to extract useful insights for their business. The drawback here is that a malicious organization could extract private information about customers if it were to get access to the such data. Since cloud services require network connectivity, many company's data is not protected by physical boundaries and there may always be a possibility of private information being exposed via a leak or a hack. Furthermore, as the market expands and more organizations begin to collect data about their clients, the attack surface will only broaden. For industries like health - care, banking, and consulting, where data contains highly confidential information about clients, this issue becomes a top priority. Our capstone also targets these industries, and this is where our industry partner enters the picture.

2.4 Tackling the Data Privacy Issue with OpenChai

Our capstone team is partnering with OpenChai to tackle the privacy and security concerns which surface from sending data into the cloud. Together, we aim to avoid this issue by enabling enterprise customers to run their machine learning models entirely offline and in - house. OpenChai is using mobile GPUs to craft a energy - efficient yet computationally powerful desktop product that is optimized for machine learning; essentially, OpenChai is building a cloud - in - a - box[29]. This means that OpenChai customers get total visibility and control of their information assets. Our team is working to adapt the BIDData suite to run efficiently on OpenChai hardware, which will maximize the hardware's computational

throughput while minimizing its energy footprint.

Smartphone Market Analysis

OpenChai's product is only feasible because to their novel use of mobile (e.g. smartphone) processors. Consequently, OpenChai's market strategy rides on the crest of the global proliferation of smartphones. As shown in figure 2.4 below, in China alone, the number of smartphones has increased from 189 million in 2012 to over 600 million in 2015, and is projected to grow to 1.6 billion by 2021.

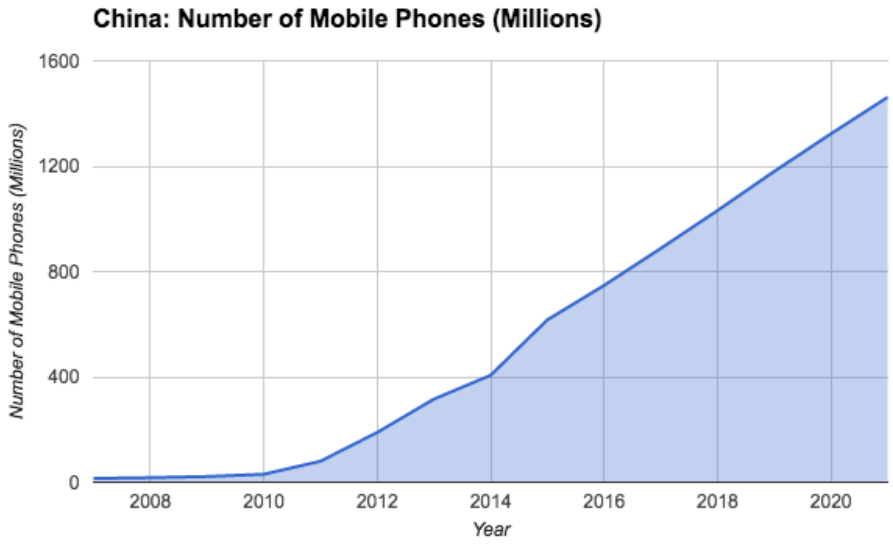

Figure 2.4: The number of mobile phones in China is growing quickly[30]

India, too, is likely to follow a similar trajectory according to Morgan Stanley Research[31]. To sustain a competitive advantage under this rising demand, manufacturers are pushed to innovate and develop improved products[30]. One crucial avenue for innova-

tion lies in developing more powerful mobile processors. ARM and Nvidia are two of the most prolific producers of mobile CPUs and GPUs, respectively; they are also the main suppliers of the mobile processors OpenChai is putting into their product. We extrapolate that the rapid growth of the global smartphone market trickles down to pave the way for OpenChai. As the smartphone market expands, ARM, Nvidia, and by extension OpenChai, will continue to innovate with better, faster products.

Nvidia and the TX1

Nvidia in particular, a company specializing in GPU design, is a key enabler for OpenChai's strategy. Nvidia hit the machine learning world in a blaze in 2012 when Krizhevsky et [3] with a neural - network - based model using Nvidia GPUs. The research group used these GPUs to engineer a novel computer vision method, producing the most outstanding result in ILSVRC to date[32]. Following this event, the use of Nvidia GPUs in machine learning exploded, correlating with continuing improvements in machine vision accuracy, as shown in figure 2.5.

Fast forward to 2015: Nvidia unveils the TX1, one of the first processors that brings the same machine learning capabilities of high - end desktop GPUs to a mobile chip[34]. The TX1 boasts impressive performance while staying up to 4x more efficient than its desktop counterpart on heavy machine learning workloads[34].

The TX1 forms the backbone of OpenChai's product. Using multiple enhanced deriva-

[3] ilsvrc

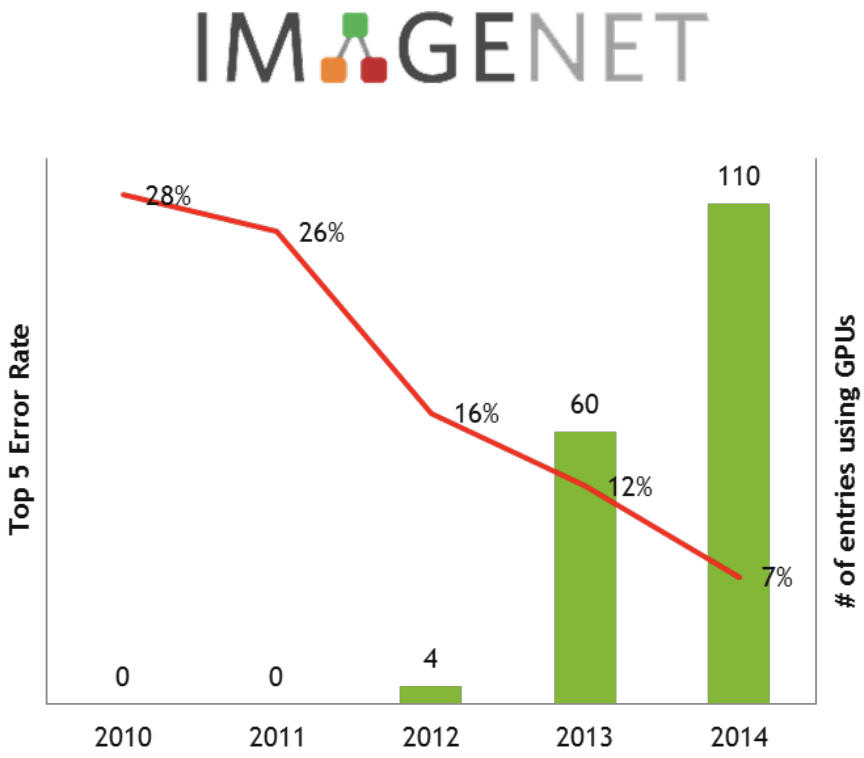

Figure 2.5: Following the "Krizhevsky result" of 2012, the use of GPUs in computer vision has exploded[33]

tives of the mobile TX1 chips, OpenChai can perform swift machine learning computations on large datasets offline and at a fraction of the power and cost of the GPUs provided by cloud computing platforms.

Conclusion

Through our analysis of the expanding smartphone landscape and the machine learning space, we believe that OpenChai is poised for growth and success. Nvidia, the main GPU hardware supplier for OpenChai, is fueled by the these two markets. Any innovation in mobile

GPU technology for these factors will be realized in better performance and efficiency of machine learning algorithms on mobile GPUs, transparently improving OpenChai's product.

Bibliography

[1] C. Volkmann, K. Tokarski, and K. Ernst, "Social entrepreneurship and social business," *An Introduction and Discussion with Case Studies. Gabler. Wiesbaden*, 2012.

[2] J. Canny, "Interactive machine learning," *University of California, Berkeley*, 2014.

[3] J. Canny, H. Zhao, B. Jaros, Y. Chen, and J. Mao, "Machine learning at the limit," in *Big Data (Big Data), 2015 IEEE International Conference on*, IEEE, 2015, pp. 233–242.

[4] J. Canny and H. Zhao, "Bidmach: Large-scale learning with zero memory allocation," in *BigLearn Workshop, NIPS*, 2013.

[5] H. Zhao and J. Canny, "Kylix: A sparse allreduce for commodity clusters," in *Parallel Processing (ICPP), 2014 43rd International Conference on*, IEEE, 2014, pp. 273–282.

[6] J. Jia, P. Kalipatnapu, R. Chiou, Y. Yang, and J. F. Canny, "Implementing a gpu-based machine learning library on apache spark," 2016.

[7] R. Chiou, J. Jia, P. Kalipatnapu, Y. Yang, and J. F. Canny, "Building a distributed, gpu-based machine learning library," 2016.

[8] A. Kamko, "Aleksandr kamko's final report," 2017.

[9] J. Xie, "Jiaqi xie's final report," 2017.

[10] S. Zhang, A. E. Choromanska, and Y. LeCun, "Deep learning with elastic averaging sgd," in *Advances in Neural Information Processing Systems*, 2015, pp. 685–693.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[14] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[15] J. Barnes, "Microsoft azure essentials: Azure machine learning," *E-kirja. Mi*, 2015.

[16] BIDData. (2015). Bidmach benchmarks, [Online]. Available: `https://github.com/BIDData/BIDMach/wiki/Benchmarks` (visited on 02/05/2017).

[17] C. Mercer. (2016). Nine tech giants investing in artificial intelligence: Microsoft, google, uber and more are investing in ai: What is their plan and who are other key players? techworld, [Online]. Available: `http://www.techworld.com/picture-gallery/big-`

data/tech-giants-investing-in-artificial-intelligence-3629737 (visited on 02/05/2017).

[18] R. Merrett. (2015). Where is machine learning heading in 2016? [Online]. Available: http://www.cio.com.au/article/590834/where-machine-learning-headed-2016 (visited on 02/05/2017).

[19] A. Naimat. (2016). The new artificial intelligence market., [Online]. Available: https://www.oreilly.com/ideas/the-new-artificial-intelligence-marke (visited on 02/05/2017).

[20] Tractica. (2015). Artificial intelligence for enterprise applications to reach $11.1 billion in market value by 2024, [Online]. Available: https://www.tractica.com/newsroom/press-releases/artificial-intelligence-for-enterprise-applications-to-reach-11-1-billion-in-market-value-by-2024 (visited on 02/05/2017).

[21] A. Vieira. (2016). The revolution of depth, [Online]. Available: https://medium.com/@Lidinwise/the-revolution-of-depth-facf174924f5#.mansn7ey (visited on 02/05/2017).

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] M. Braun. (2014). What is scalable machine learning? [Online]. Available: http://blog.mikiobraun.de/2014/07/what-is-scalable-machine-learning.html (visited on 02/05/2017).

[24] G. Blau. (2016). Ibisworld industry report 51121c: Business analytics & enterprise software, [Online]. Available: https://www.ibisworld.com (visited on 02/05/2017).

[25] VCloudNews. (2015). Every day big data statistics - 2.5 quintillion bytes of data created daily, [Online]. Available: http://www.computerworld.com/article/2598562/data-center/data-centers-are-the-new-polluters.html (visited on 04/26/2017).

[26] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. (2011). Big data: The next frontier for innovation, competition, and productivity, [Online]. Available: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (visited on 04/26/2017).

[27] P. Delforge and J. Whitney. (2014). America's data centers consuming and wasting growing amounts of energy, [Online]. Available: https://www.nrdc.org/sites/default/files/data-center-efficiency-assessment-IP.pdf (visited on 02/05/2017).

[28] P. Thibodeau. (2014). Data centers are the new polluters, [Online]. Available: http://www.computerworld.com/article/2598562/data-center/data-centers-are-the-new-polluters.html (visited on 04/26/2017).

[29] OpenChai. (2016). Openchai overview, [Online]. Available: http://openchai.org.%20Customers%20and%20Features%20sections (visited on 10/11/2016).

[30] IBISWorld. (2016). Smart phone manufacturing in china, [Online]. Available: https://www.ibisworld.com/industry/china/smart-phone-manufacturing.html (visited on 04/26/2017).

[31] A. Truong. (2016, April 26). Why india could be the new china for smartphone companies., [Online]. Available: http://www.huffingtonpost.com/entry/india-smartphone-market-new-china_us_571f82c2e4b0b49df6a8fe4 (visited on 10/11/2016).

[32] O. Russakovsky. (2015). Imagenet large scale visual recognition challenge, [Online]. Available: https://arxiv.org/pdf/1409.0575v3.pdf (visited on 04/26/2017).

[33] A. Gray. (2015, August 13). Nvidia and ibm cloud support imagenet large scale visual recognition challenge [web log post]., [Online]. Available: https://devblogs.nvidia.com/parallelforall/nvidia-ibm-cloud-support-imagenet-large-scale-visual-recognition-challenge/.%20Figure (visited on 10/11/2016).

[34] Nvidia. (2015, January 5). Nvidia ces 2015 press conference: Tegra x1, [Online]. Available: https://www.youtube.com/watch?v=ao47RQvCZw (visited on 10/11/2016).