Research on Asynchronous Distributed Deep Learning Technology—Optimizing Machine Learning Models in the Age of Distributed Data Storage

Kenta Niwa
Distinguished Researcher, NTT Communication Science Laboratories

Overview

While modern deep learning often requires aggregating data into a single datacenter to train models, in the near future data will be distributed due to increased data volume and privacy protection concern. In this article, we spoke to Kenta Niwa, a distinguished researcher working on asynchronous distributed deep learning technology. This technology allows us to optimize machine learning models as if the data was aggregated in a single datacenter, even in the modern era of distributed data.

Keywords: asynchronous distributed deep learning, machine learning, decentralized system

What is asynchronous distributed deep learning technology?

—Please tell us about your research.

Deep learning is often used for e.g., speech/image recognition. Modern deep learning involves overconcentration of data, namely, all the data is aggregated in a single huge datacenter and then used to train the model. However, considering several industrial applications, such as self-driving vehicles, factory automation, distributed power-grid, and highly personalized models, the volume of data will continue to increase and it will become much more difficult to collect, process, and deploy all the data in a single datacenter. Data aggregation is also becoming more difficult from a privacy perspective due to the effects of legislations, e.g., GDPR (the European Union's General Data Protection Regulation). Because of these factors, we believe that data storage and inference processing will be carried out in a distributed manner in the near future.

We are conducting research to develop an advanced algorithm to virtually manage data sets stored on multiple servers. For example, we recently achieved results with technology that allows us to train

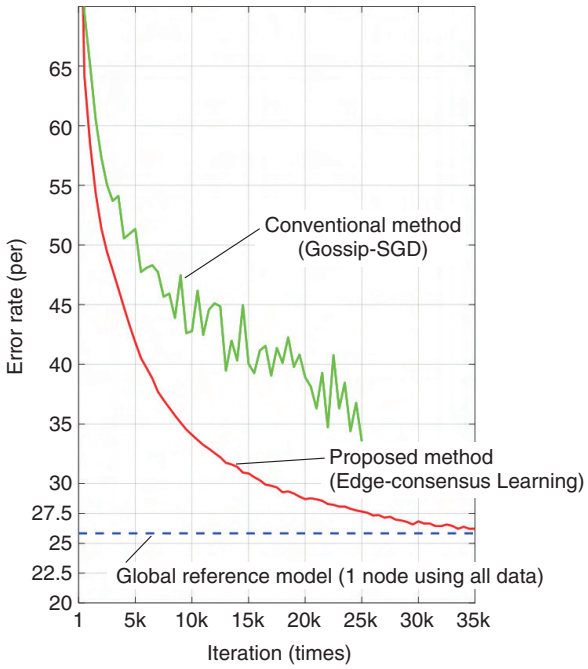


Fig. 1. Comparison among conventional and proposed methods and the global reference model.

machine learning models as if the data was concentrated in one place, even data sets are not aggregated.

—What kinds of methods are specifically being used?

Today, distributed deep learning often uses a method of exchanging and averaging models between servers (forming an average consensus). The formulation of an average consensus is a very simple and effective operation. However, while this method works well if each server has statistically homogeneous data, training often fails to progress if the data on each server is statistically biased (heterogeneous). In addition, as the number of servers increases, synchronous communication becomes more and more difficult.

We built a deep learning algorithm where all the servers connected to the network communicate asynchronously and collaboratively to train the model. I'll spare you the details of the formulas and algorithms, but to put it simply, the research expresses the idea of "multi-node collaboration" in mathematical formula.

For example, it's easy for people who get along well with each other to come together, talk it over, and draw a conclusion. However, if a group of people with strong personalities who don't get along together, everyone will go in different directions. There's not really any point in performing averaging in a situation like this. Think of this algorithm as a way to skillfully express how well the individual elements work together.

Figure 1 shows a comparison between the proposed method (Edge-consensus Learning) and the conventional method (gossip-based stochastic gradient descent (Gossip-SGD)) using a data set commonly used for testing an image classifier called CIFAR-10, with the data distributed in eight servers with statistical bias. The vertical axis represents the classification error rate, with smaller values meaning better performance. The blue dotted line shows the performance of the global reference model that trained using all data collected in one server, the solid green line shows the performance of the conventional method (Gossip-SGD), and the solid red line shows the performance of the new method being proposed (Edge-consensus Learning).

While performance does not reach to the global reference model when using the conventional method, the proposed method gets closer to the performance of the global reference model as learning progresses. I believe we can take this to mean that we have obtained a model more suited to the entirety of the data, even with asynchronous communication.

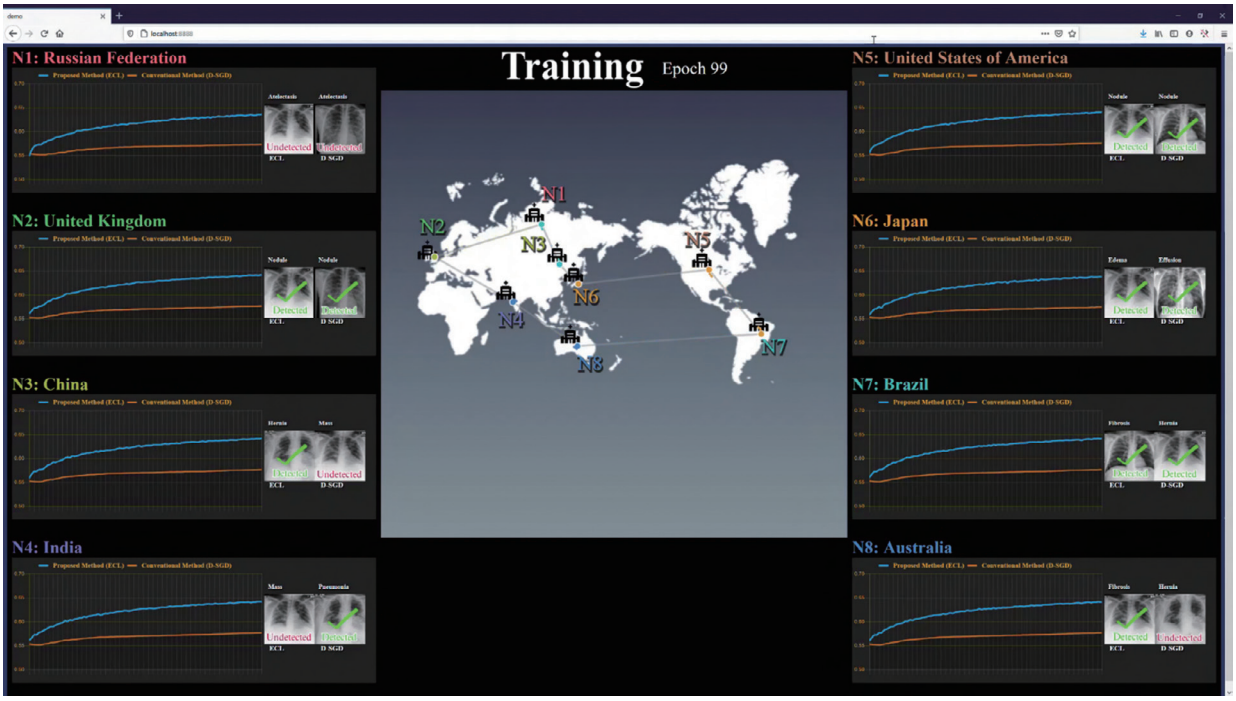


Fig. 2. Combined training of medical image analysis model across multiple hospitals.

—What kinds of possibilities does this research unlock?

Figure 2 shows a demonstration of medical image analysis model using data from multiple hospitals. Medical data is the most important example from the perspective of privacy concerns, and in practice it generally should not leave the hospital. And taking it out of the country is next to impossible. So we considered a network connecting eight hospitals (N1 to N8), where we train the model without taking the imaging data out of the hospital. Specifically, this model is a medical imaging diagnostic aid that uses chest x-ray images to detect the presence of a disease and identify the disease from among 14 different diseases. In light of the differences in diseases handled by hospitals in the real world and regional differences in conditions, we encounter a situation where the number of data items at N1–N8 and the diseases being recorded are statistically biased.

The blue lines show the recognition performance using new model at each of N1 through N8, and the orange lines show the progress of the conventional training method. The medical image data spread across eight hospitals works together well to create a highly advanced level of knowledge, like some kind of "super doctor."

Asynchronous distributed deep learning technology allows us to reap the benefits of machine learning without sacrificing privacy. This technology is not particularly application-dependent, so the algorithm is highly flexible and can be used in various applications such as text translation using smartphone data, training autonomous driving models, voice recognition in call centers, and anomaly detection in industrial factories.

Providing services that leverage NTT's locational advantages

—What do you consider NTT's strengths to be?

NTT has communication stations throughout Japan, and each station is connected by a network. Why not use this locational/physical advantage to install servers in each station to store and process data? Of course, each station will have very different types of data stored, and the system as a whole will be huge. Leveraging this data to bring services such as data processing closer to users is likely to be a viable option for NTT's business in the future. It's what's known as "edge computing." When we started thinking

about this, we thought it might be useful to provide asynchronous distributed machine learning services.

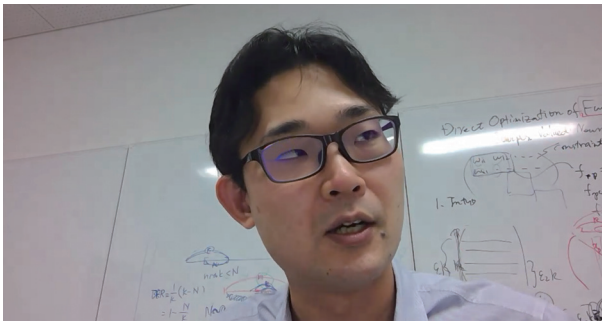
—What do you think about the future direction of your field?

I was originally conducting research related to acoustics, especially communication using microphones and loudspeakers. About 10 years after joining the company, I started studying machine learning after studying abroad in New Zealand, particularly the theme of distributed optimization field.

Many modern systems are constructed in a centralized manner. This is because of the amount of data that can be handled in one place, and because what is needed is considered to be universal, so the same service is distributed to all. However, I believe that services will continue to become more personalized as time goes on. I think we will also offer training and inference for models tailored to individual customers. With this reality, it begs the question "Is the current centralized system still the right one?" It's only natural to consider decentralized manner.

When you think about it, our brains are also "decentralized" in a manner of speaking. We can talk about one thing while thinking about something completely different. I don't think my own mind is doing heavy computing like deep learning and inference in the areas that aren't already burned out! It feels like the distributed computing groups in our heads are connected asynchronously, and they can perform high-level tasks by combining a lot of very light calculations. In this regard, I believe the systems that support the next generation of communications and society as a whole can also be processed at an advanced level with low power consumption and high flexibility, and be durable enough that they won't break even if they get a little damaged.

For example, the IOWN (Innovative Optical and Wireless Network) concept discusses concepts such as traffic coordination of self-driving vehicles in a large-scale smart city in the context of optimizing society as a whole. Of course, every car has a different destination, so I think they should also be driven differently. That said, working only for one's own benefit is no good, and I don't feel like averaging is particularly desirable either. I don't think there are systems that can coordinate and control every car yet, but I would like to create decentralized systems and core software that contribute to these things, keeping "distribution" in mind as a theme.



—What would you say to anyone hoping to become involved in basic research in the future?

The field of machine learning is incredibly competitive and changes every few months. New papers are coming out nearly every day. To be more specific about the intense competition in distributed systems, I think that there is too much competition in distributed learning, but I don't think that's the case when talking about asynchronous distributed systems like the one I'm proposing, which can flexibly obtain a high level of knowledge. Centralized and end-to-end are mainstream right now, and asynchronous distributed systems are still relatively unexplored, and I want to see the possibilities there.

Opening up new fields requires energy. It's easy for an expert on the topic to say "This is how we should do it," but 99.9% of people aren't experts, so it's incredibly important to find the challenges we have in common and discuss, refer to and compare them as we compete with each other. On the other hand, it's also important to create your own unique niches in your policy. I think it's important to work on the things you want to express and build gradually over time.

Nowadays, anyone who can gather data (even high school students) can create models for whatever application they want. In this reality, many of the researchers around me are also struggling to decide whether to focus themselves on basic research or practical development. I have had the opportunity to study abroad, and I learned a great amount about distribution there, so I ended up focusing on basic research. I'm sure I could just as easily have chosen the opposite. I can't say which is better as some people are particularly suited or unsuited to a specific position, but I think in times like these it's best to focus on one end of the spectrum.

■ **Interviewee profile**
Kenta Niwa
Distinguished Researcher, NTT Communication Science Laboratories.

He joined NTT in 2008, engaged in research and development on sound recording processing at NTT Media Intelligence Laboratories. Achieved results in contributing to commercialization of microphone array based speech enhancement technology and the "zoom-in microphone" that can pick up sound clearly from far away. After studying abroad at Victoria University of Wellington in New Zealand from 2017 to 2018, he began research into machine learning, including distributed optimization, at NTT Communication Science Laboratories. He is currently focusing on asynchronous distributed deep learning technology. He also works at NTT Computer and Data Science Laboratories.