

Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities

Daisuke Hamuro, Koji Iida, Kiyotada Usami, Shunsuke Yura, Yoshinori Matsuo, Takeharu Eda, Akira Sakamoto, Masashi Toyama, Keita Mikami, Noriaki Inoue, Ryuji Nakayama, Shohei Enomoto, Taku Sasaki, Xu Shi, Yutaka Hirokawa, and Katsuo Inaya

Abstract

This article introduces efficient inference technology as an important element in applying deep learning to business and an inference cloud service that is combined with NTT Group assets such as telephone exchange buildings and base stations.

Keywords: cloud-based inference, regional edge, deep learning optimization

1. Solving social problems with deep learning

It has been almost eight years since the overwhelming win by Geoffrey Hinton and his group at the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). As a result, various deep learning technologies are now being investigated worldwide.

Past research on deep learning revealed that trials and proof of concept demonstrations were early topics. Today, however, there is now much discussion about solving social problems through deep learning. As a disruptive technology, deep learning is no longer a topic limited to researchers—it has become a technology for solving real social problems [1].

This trend began with use cases applying image recognition as a substitute for the human eye. Many use cases now include speech recognition and lan-

guage processing. As a result, deep learning is on the road to becoming a commonly accepted technology.

2. Necessary element for accelerating the solving of social problems

At NTT Software Innovation Center (SIC), we have developed critical technology for accelerating the solving of social problems ahead of the competition. Specifically, we have redefined the meaning of *surveillance camera service* by giving Takumi Eyes [2], a commercial surveillance service launched by NTT Communications in 2017, the capability of conducting real-time analysis of surveillance-camera video. In the past, such video simply served as material to be examined after the occurrence of an event to determine what happened before being handed over to a

law enforcement agency.

Needless to say, this ability to conduct real-time analysis of video has significantly broadened the range of social problems that can be solved (effective use cases). The managing of surveillance cameras in commercial facilities and office buildings is typical of the work conducted today by security businesses, but trials have been held on searching for elderly individuals suffering from dementia [3], a phenomenon expected to become a major problem as society ages in Japan.

3. Specific technologies enabling real-time video analysis

Real-time surveillance camera service was achieved through the development of the following key technologies for achieving efficient inference processing in deep learning.

(1) Technology for densification of inference tasks

This technology includes the batch transfer of multiple inference tasks to a graphics processing unit (GPU), a close-packing processing method for parallelizing post-processing after inference,^{*1} and a stream-merge method for reducing GPU memory through batch processing of multiple data streams. The idea is to decrease the cost per task by multiplexing inference tasks in a high-density manner through a variety of patent-pending efficiency-enhancement technologies.

(2) Lightweight filter technology for inference

When analyzing stream data, of which video is one example, there are many cases in which the entire stream does not need to be analyzed due, for example, to the absence of individuals in certain segments of that video. However, processing without taking this into consideration means that computing resources will be monopolized even for video not requiring any analysis. This problem is solved by applying a lightweight filter that determines whether analysis is necessary according to the inference model being used so that only those locations that require analysis are targeted for inference processing. This technology reduces processing cost.

(3) Server/edge distributed processing technology

This technology makes it possible to use the same query language to describe server/edge linking without having to worry about the individual roles of edge devices, servers, or other components. For example, combining this technology with lightweight filter technology means that simple analysis tasks can be processed at the edge while more detailed analyses

can be offloaded to servers, which reduces network and facility costs. At the same time, preprocessing conducted at the edge in this manner makes it possible to protect highly confidential information that should not be uploaded to an external server (**Fig. 1**).

(4) Deep-learning-model-optimization technology supporting heterogeneous devices

This technology makes it possible to deploy a model that makes maximum use of the performances of individual devices by preparing an execution environment for multiple inference accelerators (CPUs (central processing units), GPUs, etc.) and making calls to these devices from a stream-processing engine.

Combining this technology with a training-model compiler (such as NVIDIA TensorRTTM or Intel OpenVINOTM) can also improve capacity by conducting different types of optimization such as model compression and low-precision processing.

(5) Technology for building inference microservices




This technology enables dedicated processes that perform only inference processing to be built as inference microservices on separate servers. Combining this technology with server/edge distributed processing technology also makes it possible to uncouple computationally intensive inference processing from relatively powerless edge devices and to apply inference-task densification technology on the inference-microservice side where many inference tasks are concentrated.



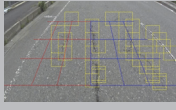


Combining all technologies described above improves capacity by more than ten times and enables real-time video analysis.

4. Service creation toward a golden era of deep learning and business scaling

In creating a commercial service, it is essential to compare the value that can be obtained from such a service with what must be paid for it (cost effectiveness) from the customer's perspective. Regardless of whatever benefits a customer can receive from a service using deep learning technology, if the cost of just an inference infrastructure reaches about 100 million yen, it is very difficult for a mid-sized company to decide on whether to introduce such a service. In other words, the ability to inexpensively construct

*1 Inference: Data analysis processing using deep learning technology. The way in which inference is used determines how the means of inference, inference environment, inference cloud, and other elements are used.

Area	Crime prevention	Cashier-free store	Marketing
Use case	Detect suspicious activities in stores/facilities (e.g. shop lifting) from surveillance cameras. Challenges: AI cameras alone cannot support varied analysis requirements. 	Capture customers' purchase activities from multiple high-resolution cameras in stores. Challenges: Networking and analysis cost is very high. 	Aggregate customer records, activity records, and POS data and analyze purchase trends and sales achievements. Challenges: Privacy must be protected when connecting to external systems. 
Edge process	Human/object/anomaly/intrusion detection	Human/object detection	Human/object detection
Server process	Face/full-body recognition, posture/activity/attribute estimation	Face/full-body recognition, posture/activity/attribute estimation	Face/full-body recognition, activity/attribute estimation, time-series analysis

Area	Hospital/Elderly care	Others			
Use case	Detect person falling or lying facedown in hospitals or nursing homes and notify staff. Challenges: Requires many cameras in one facility and accurate analysis in life-or-death situations. 	 AR-supported work	 Monitoring	 Drones	 Agriculture
Edge process	Human detection	Partial detection such as object detection			
Server process	Face/full-body recognition, posture/activity/attribute estimation	Detailed analysis based on use cases			

AR: augmented reality
 POS: point of sale

Fig. 1. Use cases of regional artificial intelligence edge services.

and use an infrastructure for an execution environment (inference environment) is key. In response to this need, the technologies introduced above for achieving real-time processing have come to the forefront. For a mid-sized company, the application of these technologies to enable efficient, real-time use of an inference environment can bring the cost effectiveness of using a deep-learning service up to a level commensurate with its benefits. For this reason, providing a service that enables efficient use of an inference environment on an inference cloud*2 to all types of customers at an appropriate price should enable NTT to gain a competitive edge over its competitors.

*2 Inference cloud: A general name given to a FaaS (function as a service) that enables efficient use of an inference environment in deep learning and machine learning.

5. Carrier Cloud for Deep Learning—expanding from surveillance cameras to deep learning

Application of technologies for achieving an efficient inference environment described in section 3 is not limited to surveillance-video-analysis services. It can also be applied to nearly all services that use deep learning. The means of generalizing these technologies is called Carrier Cloud for Deep Learning (Fig. 2). Given expectations that models using deep learning and machine learning will increase in number and continue to be used in the years to come, Carrier Cloud for Deep Learning is an execution environment for running such models when they are put to commercial use [4].

At the same time, a framework conducive to these technologies is taking shape, as summarized below.

- (1) Acceleration of service development using deep

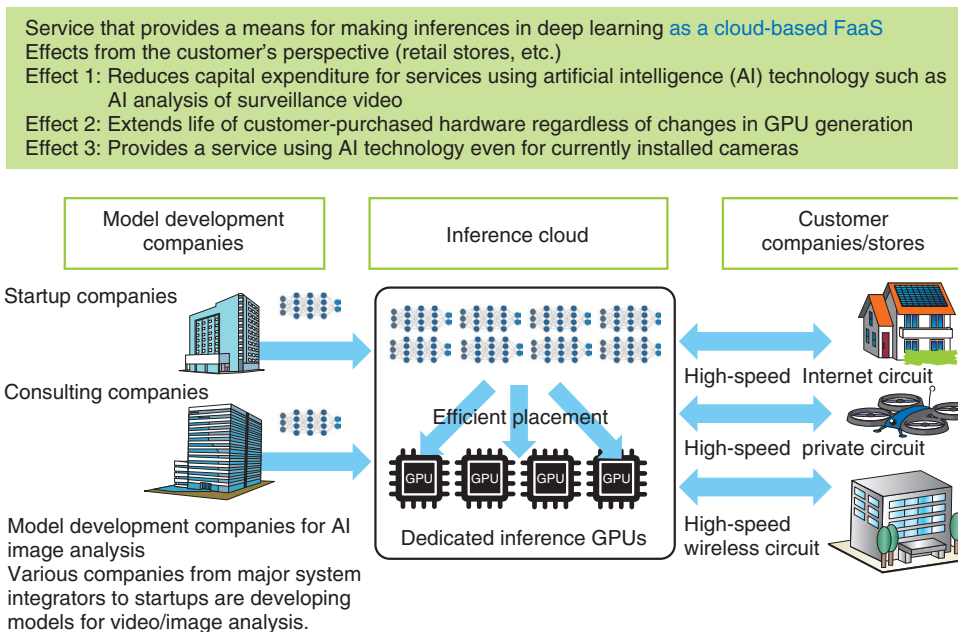


Fig. 2. Overview of Carrier Cloud for Deep Learning.

learning

We can expect even services that have so far been provided in the form of ordinary rule-based programs to be developed on a deep learning basis together as huge volumes of data are accumulated. For example, translation services based on deep learning are more powerful than rule-based forms of these services. This trend is expected to accelerate.

(2) Unbundling of training (learning) and runtime (inference)

It has been necessary to consistently use the same deep learning framework (TensorFlow, Caffe, etc.) from training to inference, but technical standards (such as ONNX: Open Neural Network Exchange) for exporting/importing previously trained models are progressing, making it easy to select the means of training and inference separately.

(3) Development of many accelerators (semiconductors) for inference processing

Only GPUs from NVIDIA were previously used for deep-learning purposes, but a variety of companies are developing and selling accelerators [5]. Regarding major companies, Intel provides the Nervana Neural Network Processor for Inference (NNP-I) and Myriad X, while Google provides Edge TPU. However, more than 100 companies including startups are now providing accelerators.

6. Regional carrier edge providing enhanced security and low latency for a more competitive inference cloud

Placing the Carrier Cloud for Deep Learning in regional NTT telephone exchange buildings and other NTT assets enables the provision of a low-cost, high-security, and low-latency service called *regional carrier edge*.

What can be provided to customers through low-latency services? We introduce some use cases.

The first use case relates to *xR*, which is the general term for the combination of virtual reality (VR), augmented reality (AR), and mixed reality (MR). The term *VR sickness* is well known. This is a phenomenon in which a user using a VR headset experiences nausea, drowsiness, or other disorientating effects when the processing speed lags, generating a delay. Regional carrier edge may be able to eliminate such VR sickness, so this may be one use case of a low-latency service.

The second use case is cloud gaming. This is a service that runs a game at a datacenter and forwards screens and operations to a terminal. In cloud gaming, a large delay limits the extent to which a game can be played. While a game such as a puzzle can be played and enjoyed regardless of delay, a game with real-time characteristics cannot be played with a large

delay. For example, if the user sees that a bullet is coming his/her way and operates the controller to avoid the bullet, a large delay would cause the bullet to hit the user before that operation information arrives at the datacenter.

Technologies for constructing inference clouds with regional carrier edge in NTT telephone exchange buildings and between 5G (fifth-generation mobile communication) antenna and the Internet are being developed at SIC. Therefore, many services that can be provided thanks to low latency, such as quality inspection on factory production lines, will be provided in the future.

7. Wanted! Partners wanting a game changer

The inference cloud introduced in this article includes technologies that can be developed by other companies or using open source software, but on the whole, it is a world that presents a new challenge that no company has ever achieved.

At SIC, we strongly believe that technology can change the world and are investigating game-changing technology regarding inference clouds. To this end, we are seeking partners to achieve such game-changing breakthroughs together.

Some courage is probably needed to propose technologies with no established track record to custom-

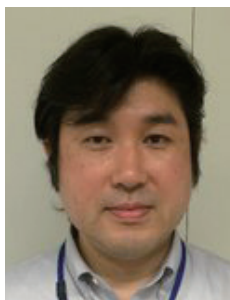
ers. Proposing technology that a customer is not familiar with, managing a project, and delivering it on time are very difficult. With this in mind, our plan is to first provide such technologies to the NTT Group then conduct tests so that we can later include them in services we provide to customers. We would like to create such an environment together with our partners.

References

- [1] T. Yukawa, "Stanford University Professor Says, 'AI is entering a phase of solving social issues'," AI Shinbun, Mar. 2019 (in Japanese). <https://aishinbun.com/clm/20190330/2018/>
- [2] Press release issued by NTT Communications on July 12, 2017 (in Japanese). <https://www.ntt.com/about-us/press-releases/news/article/2017/0712.html>
- [3] J. Hirono, "Issues Surrounding the Introduction of Deep Learning and Use Cases," The 2nd Deep Learning Lab, July 2017 (in Japanese). <https://www.slideshare.net/hironojumpei/ss-78291832>
- [4] AWS re:Invent 2018 - Keynote with Andy Jassy, <https://www.youtube.com/watch?v=ZOIkOnW640A>
- [5] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and Benchmarking of Machine Learning Accelerators," Aug. 2018. <http://arxiv.org/abs/1908.11348>

Trademark notes

All brand, product, and company/organization names that appear in this article are trademarks or registered trademarks of their respective owners.



Daisuke Hamuro

Executive Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in physics from Tokyo Institute of Technology in 1992 and 1994 and joined NTT Network Service Systems Laboratories in 1994. His current research interests include network security and privacy control technologies. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).



Takeharu Eda

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. in mathematics from Kyoto University in 2001 and an M.S. in engineering from Nara Institute of Science and Technology in 2003. He joined NTT in 2003. His research interests include a wide range of topics in SysML (Systems and Machine Learning). He is a member of the Information Processing Society of Japan (IPSI) and the Association for Computing Machinery.



Koji Iida

Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.E. and M.Sc. from Keio University, Kanagawa, in 1993 and 1995. He joined NTT Information Platform Laboratories in 1995 and studied enterprise communication middleware and distributed object technologies. He moved to NTT Information Sharing Platform Laboratories in 2007 and investigated identity management technology and cloud computing technology. As a result of organizational changes in July 2012, he is now with NTT Software Innovation Center.



Akira Sakamoto

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in information science from Hokkaido University in 1991 and 1993 and joined NTT in 1993. His current research interests include deep learning and data science.



Kiyotada Usami

Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

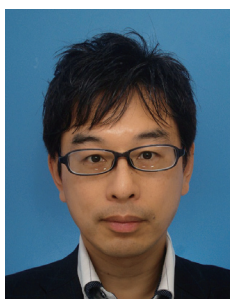
He received a B.S. and M.S. in electrical engineering from Keio University, Kanagawa, in 1993 and 1995. He joined NTT Human Interface Laboratories in 1995 and moved to NTT Software Innovation Center in 2019. His current research interests include deep learning and computer vision.



Masashi Toyama

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in information and computer science from Keio University, Kanagawa, in 2003 and 2005 and joined NTT in 2005. His current research interests include data science and software engineering.



Shunsuke Yura

Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S., M.S., and Ph.D. in information science from the University of Tokyo in 1994, 1996, and 1999 and joined NTT in 1999. He has mainly been researching and developing service collaboration platforms and cloud service platforms. His research interests include service platforms and software engineering.



Keita Mikami

Senior Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in information and computer science from Waseda University, Tokyo, in 2005 and 2007 and joined NTT in 2007. His current research interests include data science and software engineering. He is a member of IPSJ.



Yoshinori Matsuo

Senior Research Engineer, Supervisor, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in physics from Tokyo Institute of Technology in 1999 and 2001 and joined NTT Cyber Space Laboratories in 2001. His research interests include a wide range of topics in network security and system engineering.



Noriaki Inoue

Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

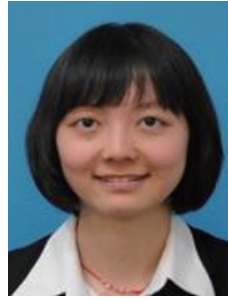
He received a B.S. and M.S. from Osaka University in 1995 and 1997 and joined NTT in 1997. His current research interests include network engineering and deep learning.



Ryuji Nakayama

Research Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. in computer science from Yamanashi University in 1988 and 1990 and joined NTT in 1990. His current research interests include cloud computing and software engineering. He is a member of IPSJ.



Xu Shi

Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

She joined NTT Software Innovation Center in 2014. Her current research interests include deep learning and computer vision.



Shohei Enomoto

Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. from Tokyo Institute of Technology in 2014 and 2016 and joined NTT Software Innovation Center in 2016. His current research interests include deep learning.



Yutaka Hirokawa

Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.E. and M.E. in computer science from Tohoku University, Miyagi, in 2003 and 2005 and joined NTT in 2005. His research interests include anomaly network traffic detection.



Taku Sasaki

Engineer, Distributed Data Processing Platform SE Project, NTT Software Innovation Center.

He received a B.S. and M.S. from Tokyo Institute of Technology in 2014 and 2016 and joined NTT Software Innovation Center in 2016. His current research interests include attention-based deep learning and computer vision.



Katsuo Inaya

Senior Research Engineer, Supervisor, Planning Section, NTT Software Innovation Center.

He joined NTT in 1995. He is an experienced engineer with a long history of working in the information technology and services industry. He has experience in the areas of enterprise software, business development, strategy, strategic partnerships, and mobile devices. His current research interests include deep learning.