Multi-modal Deep Learning Platform for IoT Data

The multi-modal deep learning platform for IoT data performs AI learning on a combination of images, text, and numeric data and uses it in applications such as highly accurate similar image retrieval, image classification, and object detection. This article describes two of the techniques used, namely triplet network learning and the application of deep learning to image data by utilizing a distributed representation of words. The article also describes what Hitachi sees as the likely applications for the platform and the value it offers customers as it prepares for deployment in solutions.

Yasumichi Ikeura
Koichi Okamoto
Ryohei Kashima
Yusuke Hijikata
Atsushi Hiroike, Ph.D.

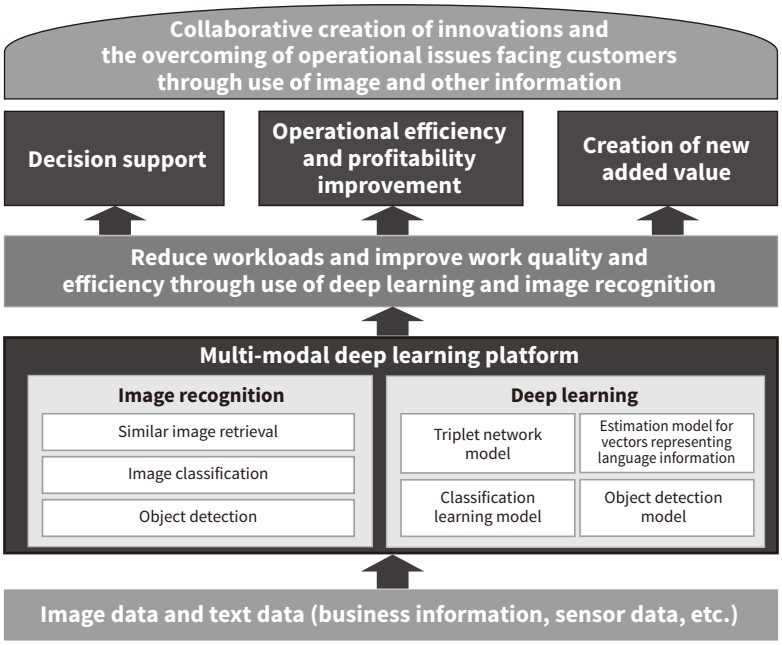
1. Introduction

Since the development of higher storage capacity, improved computing performance, and faster data transfer speeds over networks, it has become possible for various businesses to store up large electronic archives. It is not only conventional bibliographic data, but also various electronic information that can be acquired, including media data such as video or images, and Internet of Things (IoT) data from sensors, etc. This in turn creates a need for information processing technology that can utilize these electronic archives. In the examination of patents, trademarks, and designs, for example, a wide variety of information is contained in archives of examinations such as application documentation, patent drawings, and bibliographic information compiled and associated with the examination. If these archives can be leveraged to improve the efficiency of examination and of the work involved in submitting an application, it would make a major contribution to progress in the industrial sector.

Meanwhile, deep learning neural networks have attracted attention in recent years as an information processing technology means of processing media data. In particular, in the field of image and video recognition, this has included the proposal of numerous models able to overcome problems considered intractable in the past. However, while models have been shown to work in an academic setting, numerous difficulties arise when an attempt is made to apply them as is in business applications. For example, unlike the data for benchmarking in academic fields, business data is diverse and suffers from issues such as the presence of exceptions and incomplete data, so it is often difficult to apply models as-is to the sort of archives used in business. In response, Hitachi has been undertaking research and development of learning models and the framework (basic technologies) used to implement deep learning neural networks.

Figure 1 — Diagram of Multi-modal Deep Learning Platform

The application of deep learning to business information, sensor data, and so on which customers hold can enhance the precision of similar image retrieval, image classification, and object detection.



2. Overview of Multi-modal Learning Platform for IoT Data

This section gives an overview of the multi-modal deep learning platform (see diagram in **Figure 1**). The platform uses deep learning to provide image recognition services including similar image retrieval, image classification, and object detection. Use of these functions improves efficiency by automating the task of searching or classifying large numbers of images, a task that in the past would be done manually. It also includes a graphical user interface (GUI) for initiating deep learning and monitoring its progress. The interface is easy to use and check even by those who lack experience with this technology, providing an environment that facilitates proof-of-concept (PoC) trials of machine-learning-based image recognition.

The task of machine learning involves more than just setting it running, it also requires interaction with the application to prepare the training data (annotating it with metadata) and coordination of the various applications including those that put the learning results to use in actual solutions. The learning platform uses the application framework of Hitachi's similarity-based image retrieval system[(1), (2)] to perform the sequence of (a) Annotation, (b) Learning, and (c) Deployment in an efficient manner and deliver unified applications. The main features are as follows.

(1) Similar image retrieval using Hitachi's similarity-based image retrieval system

The similar image retrieval engine developed by Hitachi combines technologies from three different fields: image recognition, databases, and platforms. In the case of image recognition, this includes image feature extraction. Moreover, the database and platform technologies used have a strong affinity with deep learning neural networks.

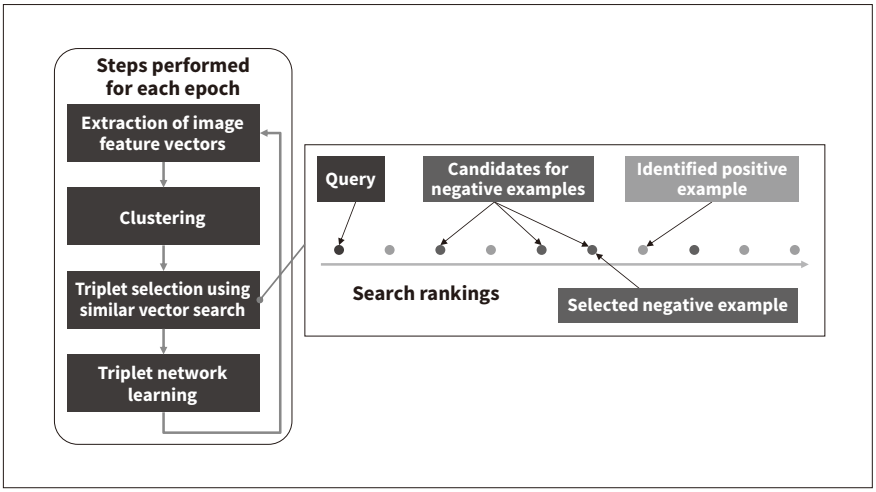
Deep learning neural networks can be characterized as a means of generating one vector from another. A key feature of the similar image retrieval engine is its high-speed function for searching for similar vectors. This approximate similarity vector search, uses Hitachi's proprietary clustering algorithm, and the search engine works by incorporating it into a database management system (DBMS) as a standard function. It can rapidly search for similar data in a large dataset made up of high-dimensional vectors with tens to thousands of dimensions. With deep learning neural network output vectors like this, a wide variety of functions become possible when deep learning neural networks and similar image retrieval are used together.

(2) Support for various learning models

The multi-modal deep learning platform is equipped with a number of learning models. Along with the model mentioned earlier for extracting image feature values used in similar image retrieval, these also include learning models that provide the ability to perform image classification and object detection. Other models include an original triplet network model and an estimation model for a variable number of vectors representing language information, the latter being used for multi-modal learning. The following section provides more information about these models.

Figure 2 — Triplet Network Learning

The triplets best suited to learning are identified by applying feature vector extraction and similar vector search to the network obtained at each learning epoch.



3. Learning Model Implementation Examples

This section describes the distinctive learning models included in the multi-modal learning platform.

3. 1 Triplet Network Model

Triplet network models are based on "triplets" made up of a query and positive and negative examples for that query (data that should or should not be included in the query's search results). Model learning is performed on feature vectors and seeks to minimize a loss function defined in terms of the distance relationships of triplets in the feature vector space (triplets being a combination of a query with positive and negative examples of the desired data).

The number of triplets in a given training data set is very large. One of the key challenges for triplet network model learning is the problem of triplet selection. The learning platform implements a learning method that automatically selects effective triplets for learning by using similar vector search. **Figure 2** shows an overview of one learning "epoch" (one cycle of learning using a particular training data set).

The first step is to obtain the feature vectors for all images using the network in its current state (extraction of the image feature vectors). Clustering is performed on these feature vectors and then a similar vector search performed on all images with the sequential selection of queries from the full set of images. Ideally, positive examples will tend to predominate near the top of the search results for each image, with negative examples appearing either near the bottom or not at all. Next, triplets are formed by selecting those negative examples that rank higher than positive examples. When these triplets are then used for learning, the network should evolve to a new state in which positive examples appear higher in the results and negative examples appear lower down.

When this technique was tested by searching for images of people in surveillance video, it generated feature vectors with greater accuracy than was achieved by classification learning.

3. 2 Estimation Model for Variable Number of Vectors Representing Language Information

A common technique used in natural language representation is to use word2vec* to produce a distributed representation of words in the form of vectors. This section describes a model that performs inference on word vectors generated from image keywords using word2vec (see **Figure 3**).

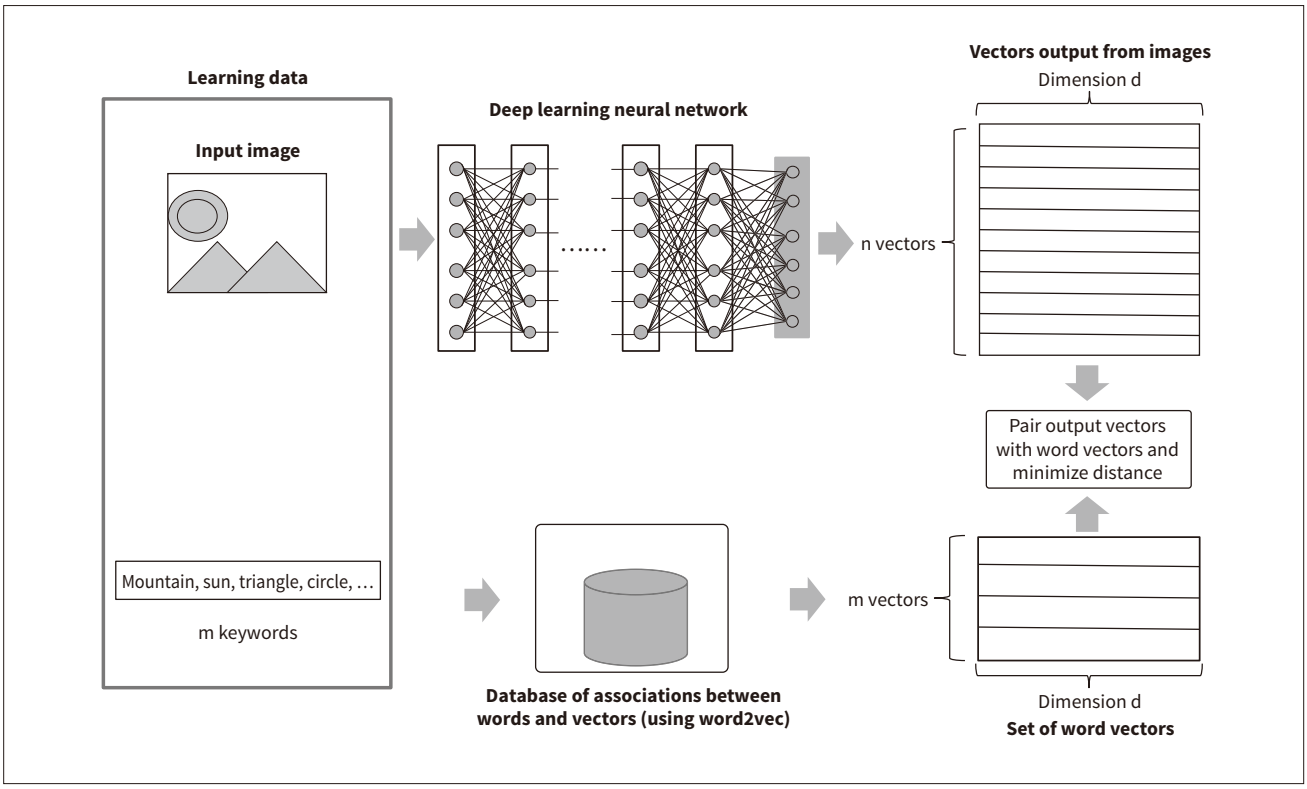
The model is intended for applications where the relevant images have been tagged with an arbitrary number of keywords. For example, given a word vector of dimensionality d, the neural network is designed to output n × d-dimensional vectors from an input image. The keywords, meanwhile, are converted into a set of d-dimensional word vectors, with the associations between words and vectors having been stored in a database beforehand. Next, the distance in vector space between the vectors in each output vector/word vector pair is calculated for each learning iteration, with m being the number of keywords. The pairs are then ranked in order of this calculated distance (shortest first) and learning performed so as to minimize the sum of the distances between pairs. As m < n in most cases, there will be some output vectors left over after pairing with word vectors. Learning is done so as these vectors will tend toward the zero vector.

When this model is used in practice, a similar vector search is performed in which model output vectors are submitted as queries to the word vector database. Those vectors that appear near the top of the results indicate words that are relevant to the input image.

* A method that analyzes large amounts of text data and converts the words into a numeric vector with low dimensionality (several hundred dimensions) to enable meaning and grammar to be determined. The ability to use this vector representation as a basis for performing addition or subtraction operations on words to infer or calculate their degree of similarity means that the method can be used for tasks such as label estimation or classification.

Figure 3 — Estimation Model for Variable Number of Vectors Representing Language Information

Image keywords converted to vectors by word2vec are input to a deep learning neural network. A fixed number of output vectors from the deep learning neural network and their associations with a variable number of keywords are then selected dynamically based on the state of the neural network at that iteration.



4. Example Applications

This section describes example applications of the multimodal deep learning platform that are expected to improve operational efficiency.

4. 1 Application to Prior Art Survey of Advanced Technologies

The surveys of "prior art" (previous instances of a technology) conducted for patents involve searches using the text information in documents or the classification information assigned to documents to narrow down the literature to be reviewed. As the literature identified as relevant by the survey is screened manually, reducing the amount of material to review will improve efficiency.

Since 2016, Hitachi has been involved in a research project at Japan's Patent Office, which aims to make prior art surveys more efficient through use of artificial intelligence (AI), as well as research into the classification and searching of patent drawings.

The idea is that this work can be made more efficient by using the similar image retrieval capability of the deep learning platform to perform searches directly using drawing information, using this to exclude literature from manual screening.

4. 2 Application to Assist Inspection and Assessment of Bridges and Other Infrastructure

A large amount of image data showing patterns of damage such as cracking is available from past inspection and assessment surveys of bridges and other infrastructure that were conducted manually. Using this pre-existing archive of assessed images and text-format assessment results as training data, application of the deep learning platform to this work can assist with recording the type and severity of damage, reliably and without omissions, by performing image classification to automatically identify cracks in need of remedial action. By rationalizing the work in a way that avoids the misallocation of human resources to the inspection and assessment of bridges in good condition, this allows these resources to instead focus on those bridges with high importance or urgency.

4. 3 Linking with High-speed Person Detection and Tracking Solutions

Hitachi's high-speed person detection and tracking solution uses AI image analysis to help detect and track persons of interest based on their description or a full-body image. Hitachi is looking into deploying the technology in applications such as police investigations[3]. Linking the learning

platform with this solution is planned for FY2021 and it is hoped that doing so and applying the learning results will enable the detection and tracking of people to be accomplished in ways that better suit the customer's environment.

5. Conclusions

This article has described a multi-modal deep learning platform currently under development by Hitachi, the technologies it uses, and where it is to be deployed.

Based on the results of PoC trials and other studies, Hitachi plans to continue adding new functions that will deliver efficiency gains for users and deploying them in services with high added value for customers.

References

1) A. Hiroike et al., "Information Retrieval for Large-scale Image and Audio Archive," Hitachi Hyoron, 95, pp. 200–205 (Feb. 2013) in Japanese.
2) A. Hiroike, "Similarity-based Image Retrieval System 'EnraEnra'," Journal of the Japanese Society for Artificial Intelligence, Vol. 29, No. 5, pp. 430–438 (Sep. 2014) in Japanese.
3) H. Okita et al., "AI-based Video Analysis Solution for Creating Safe and Secure Society," Hitachi Review, 69, pp. 687–693 (Sep. 2020).

Authors

Yasumichi Ikeura
Government & Public Systems Department 3, Government & Public Corporation Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Development of multi-modal deep learning platform. *Society memberships*: The Society of Project Management (SPM).

Koichi Okamoto
Government & Public Systems Department 3, Government & Public Corporation Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Development of multi-modal deep learning platform.

Ryohei Kashima
Government & Public Systems Department 3, Government & Public Corporation Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Development of multi-modal deep learning platform.

Yusuke Hijikata
Government & Public Systems Department 3, Government & Public Corporation Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Development of multi-modal deep learning platform.

Atsushi Hiroike, Ph.D.
Lumada Data Science Laboratory, Research & Development Group, Hitachi, Ltd. *Current work and research:* Research and development of image retrieval and deep learning. *Society memberships*: The Japanese Psychological Association (JPA).