NTT's LLM "tsuzumi": Capable of Comprehending Graphical Documents

Ryota Tanaka, Taichi Iki, Taku Hasegawa, and Kyosuke Nishida

Abstract

Large language models (LLMs) are being applied to fields such as healthcare, customer support, and office digital transformation. Information handled in such fields includes not only text but also a variety of visual content such as figures and diagrams. To develop LLMs as the core of artificial intelligence, their capabilities must be expanded so they can comprehend visual information. NTT Human Informatics Laboratories has been researching and developing NTT's LLM called "tsuzumi." In this article, we discuss our efforts related to tsuzumi's visual machine reading comprehension technology for comprehending the content of a document from visual information.

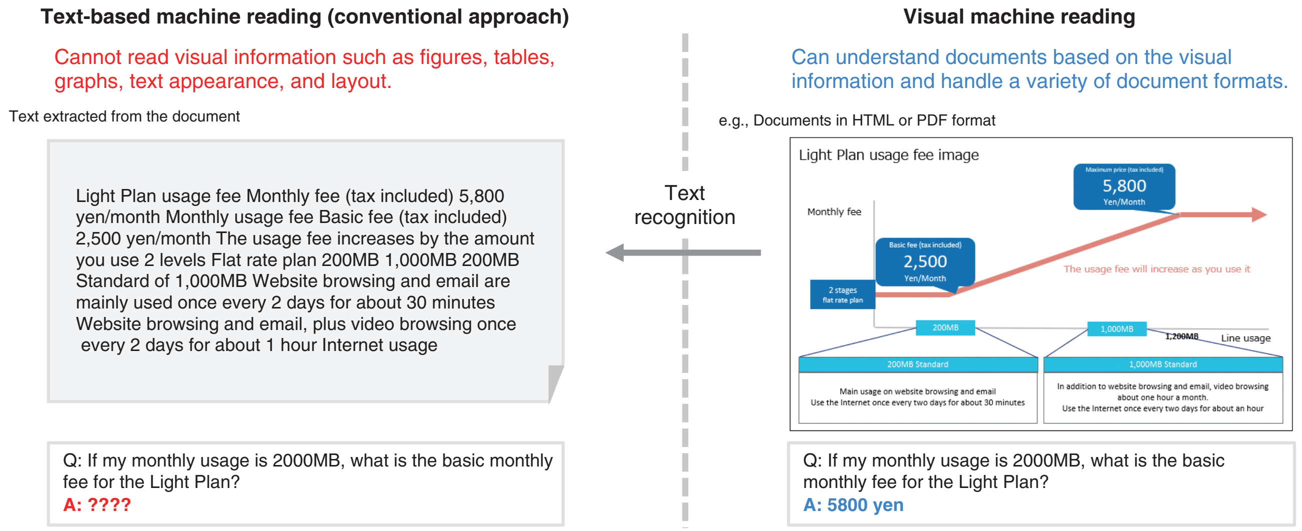
Keywords: tsuzumi, LLM, visual machine reading comprehension

1. Visual machine reading comprehension technology for visual document understanding

Besides text, the documents we handle every day include visual elements (e.g., icons, figures, and diagrams). Such information exists in a variety of forms and layouts. Developing technology to read and comprehend real-world documents is one of the most urgent challenges in the field of artificial intelligence (AI). Many AI models, including large language models (LLMs), which have the ability to achieve general-purpose language understanding and generation, have been developed. Although AI capabilities have expanded vastly, surpassing, for example, humans' reading ability, AI models still face the limitation of being able to understand only textual information in documents. To address this issue, NTT has proposed visual machine reading comprehension technology (**Fig. 1**). This technology enables comprehension of documents from visual information in a manner similar to how humans understand information.

We constructed datasets, such as VisualMRC [1] and SlideVQA [2], to make this technology possible. These datasets contain question-answering pairs on single and multiple document images, such as screenshots of webpages and presentation materials. Comprehending document images requires comprehension of not only linguistic information but also visual information such as the size and color of characters, figures and diagrams, graphs, and layout. We proposed LayoutT5 [1], a visual machine reading comprehension model that integrates two sets of inputs. It first applies object-recognition technology to extract regions in a document (titles, paragraphs, images, captions, lists, etc.) and applies text-recognition technology to extract the position and appearance information of text as additional input. We also proposed M3D [2], which understands the relationships between multiple document images. These models, which take into account visual information, perform better than models that only handle text, confirming the effectiveness of this technology inspired by human information processing.

Using the knowledge we obtained from constructing the datasets and models, we participated in the

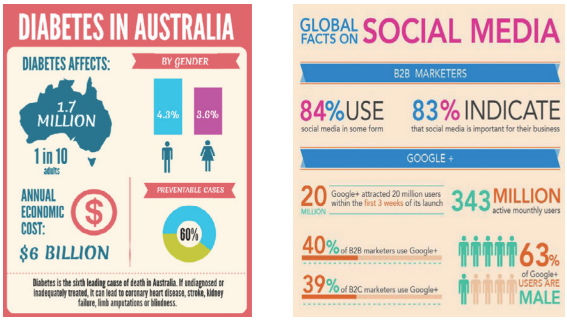

Text-based machine reading (conventional approach)
Cannot read visual information such as figures, tables, graphs, text appearance, and layout.
Text extracted from the document
Visual machine reading
Can understand documents based on the visual information and handle a variety of document formats.
e.g., Documents in HTML or PDF format
Text recognition
Light Plan usage fee image
Monthly fee
Maximum price (tax included)
5,800
Yen/Month
Basic fee (tax included)
2,500
Yen/Month
The usage fee will increase as you use it
2 stages flat rate plan
200MB
1,000MB
1,200MB
Line usage
200MB Standard
1,000MB Standard
Main usage on website browsing and email
Use the Internet once every two days for about 30 minutes
In addition to website browsing and email, video browsing about one hour a month.
Use the Internet once every two days for about an hour

Light Plan usage fee Monthly fee (tax included) 5,800 yen/month Monthly usage fee Basic fee (tax included) 2,500 yen/month The usage fee increases by the amount you use 2 levels Flat rate plan 200MB 1,000MB 200MB Standard of 1,000MB Website browsing and email are mainly used once every 2 days for about 30 minutes Website browsing and email, plus video browsing once every 2 days for about 1 hour Internet usage

Q: If my monthly usage is 2000MB, what is the basic monthly fee for the Light Plan?
A: ????

Q: If my monthly usage is 2000MB, what is the basic monthly fee for the Light Plan?
A: 5800 yen

Fig. 1. Overview of visual machine reading comprehension technology.



Q1: How many females are affected by diabetes?
A: 3.6%

Q2: What total percent of B2B and B2C markets use Google+?
A: 79% (40% + 39%)

Fig. 2. Example of questions and answers in InfographicVQA.

Document Visual Question Answering (DocVQA) competition at the International Conference on Document Analysis and Recognition in 2021 (ICDAR 2021). This challenge tests a model's ability to answer questions presented in infographics containing visual representations of information, data, and knowledge. Examples of questions given at this competition and their answers are shown in **Fig. 2**. To answer Q1 in the figure, the model must understand that the icon shown in the center right of the document represents women. To answer Q2, the model must be able to extract the numerical values from the document and calculate "40% + 39% = 79%".

Addressing these challenging questions requires a wide range of capabilities: understanding both textual content and visual information (e.g., icons, figures, and diagrams), comprehending the spatial relationships between text and visual elements, and executing arithmetic operations. For this competition, we thus proposed a model for answering infographic questions called IG-BERT [3]. We introduced a new method for learning layout relationships between text and visual objects in document images and a data-augmentation method for generating reasoning processes. IG-BERT achieved the highest performance among models of similar size while curbing the amount of pre-training data needed to 1/22 that of conventional models. It won second place out of 337 models submitted by 18 teams.

2. Issues with conventional visual machine reading comprehension technology

Conventional techniques in visual machine reading comprehension struggled with handling diverse tasks, such as extracting information from invoices. Typically, achieving high performance on specific tasks required extensive training on task-specific datasets, resulting in high data creation and training costs. This approach created barriers to developing models that could effectively meet user needs across different applications. We thus sought to develop a visual machine reading comprehension model that is effective in following instructions by using an LLM,

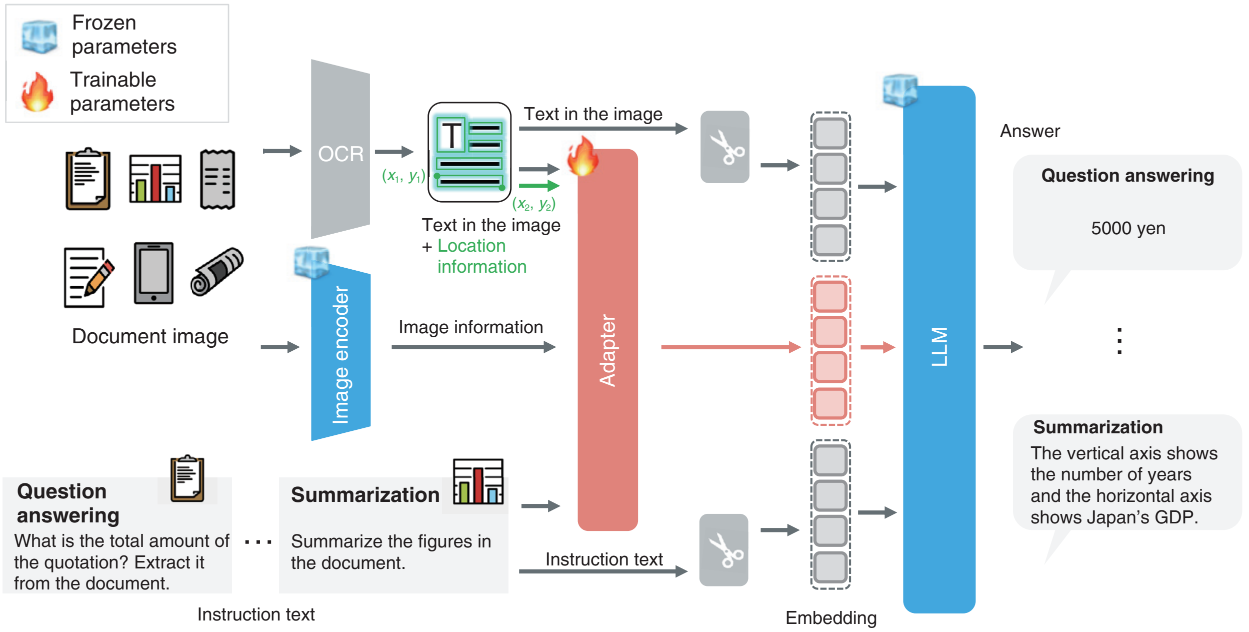


Fig. 3. Overview of LLM-based visual machine reading comprehension technology.

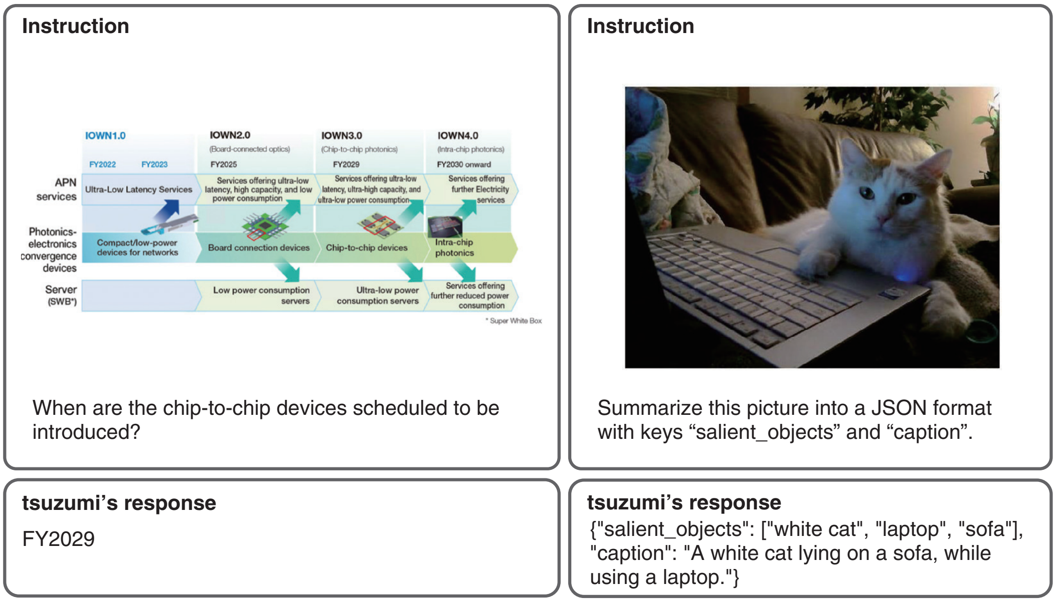
which is endowed with the general-purpose ability to understand and generate language. Our model would be able to respond to questions without training on each comprehension task. Specifically, our developmental challenge was to determine how to integrate visual information such as figures and diagrams contained in document images with text so the data can be understood by the LLM, which can only understand text information, without degrading the LLM's reasoning capability.

3. LLM-based visual machine reading comprehension technology

To enable an LLM to comprehend visual information such as figures and diagrams included in document images, visual information represented as a collection of pixels must be converted into a form that the LLM can easily process. Regarding visual document understanding of NTT's LLM "tsuzumi," we achieved the addition of visual understanding while maintaining text comprehension skills by combining an image encoder and lightweight adapter, as shown in **Fig. 3** [4]. The image encoder maps a set of pixels to textual meaning. The adapter transforms the meaning so it can be processed by tsuzumi.

3.1 Image encoder capable of understanding Japanese images characteristics

An image encoder processes the visual information of what appears in an image. We prepared an image encoder that converts images to vectors as well as a text encoder that converts text sequences to vectors. We train the image encoder so that the distance between an image vector and vector of the text representing the content of the image is close, and the distance is far when the image and text have no relationship. Therefore, visual information obtained by the image encoder can be connected to textual information. To train the image encoder, we constructed a dataset of several hundred million pairs of text and images. We collected not only general images and English captions but also images particular to Japan and their Japanese captions. The Japan-specific images contain, for example, Japanese writing and landscapes found only in Japan. By purposefully including Japanese captions during training, we enable the model to learn expressions particular to the Japanese language, such as "*aoi shingo* (*blue* traffic light)" and "*makkana taiyo* (bright *red* sun)." We are also engaged in constructing models that robustly support both English and Japanese by developing techniques that allow encoders trained on English text and images to adapt to the Japanese language [5].


IOWN1.0
FY2022
FY2023
IOWN2.0
(Board-connected optics)
FY2025
IOWN3.0
(Chip-to-chip photonics)
FY2029
IOWN4.0
(Intra-chip photonics)
FY2030 onward
APN services
Ultra-Low Latency Services
Services offering ultra-low latency, high capacity, and low power consumption
Services offering ultra-low latency, ultra-high capacity, and ultra-low power consumption
Services offering further Electricity services
Photonics-electronics convergence devices
Compact/low-power devices for networks
Board connection devices
Chip-to-chip devices
Intra-chip photonics
Server (SWB*)
Low power consumption servers
Ultra-low power consumption servers
Services offering further reduced power consumption
* Super White Box

Instruction

When are the chip-to-chip devices scheduled to be introduced?

tsuzumi's response

FY2029

Instruction

Summarize this picture into a JSON format with keys "salient_objects" and "caption".

tsuzumi's response

{"salient_objects": ["white cat", "laptop", "sofa"], "caption": "A white cat lying on a sofa, while using a laptop."}

Fig. 4. Example of visual machine reading comprehension by tsuzumi.

3.2 Training adapter for understanding visual documents

An adapter has the role of translating a document representation into the "words" understood by the LLM. An LLM splits a text sequence into several strings called tokens. To input tokens into the neural network, they are converted to vectors. The vector corresponding to the token, called an embedding, is the "word" received by the LLM. The adapter communicates the image to the LLM by converting the output of the image encoder to an embedding.

Because the adapter is a neural network with a small number of parameters, training is necessary. In tsuzumi's visual machine reading comprehension, we maintain tsuzumi's reasoning capability by fixing its parameters and those of the image encoder and training only the parameters of the adapter. An adapter suited for images is also achieved by carrying out multi-stage training. Initially, large-scale image-caption pairs are used to train tsuzumi to recognize general visual concepts such as objects, scenery, and locations. Subsequently, tsuzumi is further trained using optical character recognition (OCR) from image and textual instructions as inputs, with answers as outputs. To this end, we created the dataset InstructDoc, which is a large-scale collection of 30 publicly available visual document understanding datasets, each with diverse instructions in a unified format [4]. For example, it includes tasks such as information extraction, wherein given a question like "Tell the total amount" and a invoice image, the system provides answers such as "5000 yen." We experimentally verified that an LLM trained on InstructDoc can achieve a high rate of success for unseen tasks (tasks not included in the training data).

3.3 Output examples

An example of tsuzumi's visual machine reading comprehension is shown in **Fig. 4**. The left side shows a question-answering task on the figure. The figure depicts the Innovative Optical and Wireless Network (IOWN) roadmap and the question "When are the chip-to-chip devices scheduled to be introduced?" is posed. Our model responds with "FY2029," the correct answer. This demonstrates that it is capable of reading the visual structure of the diagram, in which the roadmap is divided into columns showing fiscal years. Because the training dataset contains images of various figures and diagrams, our model understands standard visual layouts, thus could answer the question. The right side of the figure shows a photo-recognition task. A photo of a cat is shown, and the instruction "Summarize this picture into a JSON format with the keys 'salient_objects' and 'caption'" is given. JSON is a standard text-based format for representing structured data. Our model responded with "{"salient_objects": ["white cat", "laptop", "sofa"], "caption": "A white cat lying on a sofa, while using a laptop."}.” The model could not only output text in JSON format with the given keys, it could also extract content from the image that matches the meaning of each key. Controlling output format based on text-image understanding can be used for diverse applications such as image tagging. Therefore, the flexibility of tsuzumi’s visual machine reading comprehension connects text and image understanding to accomplish tasks that meet the needs of users.

4. Future goals

We seek to further expand the capabilities of the current document comprehension model. We will also expand the application range of tsuzumi by connecting it with other modalities besides the visual modality, with the goal of advancing research and development and commercialization to ultimately achieve a society where humans and AI coexist.

References

[1] R. Tanaka, K. Nishida, and S. Yoshida, “VisualMRC: Machine Reading Comprehension on Document Images,” Proc. of the 35th Annual AAAI Conference on Artificial Intelligence (AAAI 2021), pp. 13878–13888, Feb. 2021.

[2] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, “SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images,” Proc. of the 37th Annual AAAI Conference on Artificial Intelligence (AAAI 2023), pp. 13636–13645, Washington, D.C., USA, Feb. 2023.

[3] R. Tanaka, K. Nishida, J. Xu, and S. Nishioka, “Infographic Question Answering Based on Integrated Understanding of Text and Visual Information,” Proc. of the 28th Annual Meeting of the Association for Natural Language Processing (NLP 2022), pp. 52–57, Mar. 2022 (in Japanese).

[4] R. Tanaka, T. Iki, K. Nishida, K. Saito, and J. Suzuki, “InstructDoc: A Dataset for Zero-shot Generalization of Visual Document Understanding with Instructions,” Proc. of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024), pp. 19071–19079, Vancouver, Canada, Feb. 2024.

[5] T. Hasegawa, K. Nishida, K. Maeda, and K. Saito, “DueT: Image-Text Contrastive Transfer Learning with Dual-adapter Tuning,” Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), pp. 13607–13624, Singapore, Dec. 2023.

Ryota Tanaka

Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E. and M.E. in computer science and engineering from Nagoya Institute of Technology, Aichi, in 2018 and 2020. He has also been pursuing a Ph.D. at Tohoku University, Miyagi, since 2023. He joined NTT Media Intelligence Laboratories in 2020. His research topics lie in multimodal machine learning, in particular, visual document understanding. He received the Best Paper Award at the 27th Annual Meeting of the Association of Natural Language Processing (NLP2021) and the Outstanding Paper Award at NLP2023 and 2024. He was a runner-up at the ICDAR 2021 DocumentVQA competition. He is a member of NLP.

Taichi Iki

Research Scientist, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.S. and M.S. in physics from the University of Tokyo in 2013 and 2015. After working as an AI engineer in research and development of natural language processing and dialogue systems, he entered a doctoral program. He received a Ph.D. in informatics from the Graduate University for Advanced Studies, SOKENDAI, Kanagawa, in 2022. He joined NTT in 2022. His research interests include vision-language foundation models and AI. He is a member of NLP and the Japanese Society for Artificial Intelligence (JSAI).

Taku Hasegawa

Research Scientist, Human Insight Laboratory, NTT Human Informatics Laboratories.

He skipped undergraduate studies and received an M.S. and Ph.D. in computer science and intelligent systems from Osaka Prefecture University in 2016 and 2018. He joined NTT in 2019. His research interests include vision-language foundation models and AI. He received the Paper Award from the Japanese EC Society in 2017, the Paper Award from JSAI in 2020, and an award from NLP in 2023. He is a member of NLP and JSAI.

Kyosuke Nishida

Senior Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E., M.I.S., and Ph.D. in information science and technology from Hokkaido University in 2004, 2006, and 2008. He joined NTT in 2009. His current interests include natural language processing, vision-and-language, and AI. He is currently serving as the technical lead for NTT’s LLMs. He received the Paper Award from NLP from 2018 to 2024. He is a member of the Association for Computing Machinery (ACM), NLP, Information Processing Society of Japan (IPSJ), and the Database Society of Japan (DBSJ).