

ditlab at the NTCIR-18 Transfer-2 Task

Yuuki Tachioka
Denso IT Laboratory
tachioka.yuki@core.d-itlab.co.jp

Yasunori Terao
Denso IT Laboratory
terao.yasunori@core.d-itlab.co.jp

ABSTRACT

The ditlab team participated in the Retrieval Augmented Generation (RAG) and Dense Multimodal Retrieval (DMR) subtasks of the NTCIR-18 Transfer-2 task. For the RAG subtask, we proposed a late fusion method for answer generation that uses multiple contexts retrieved by the dense passage retriever. Unlike sequential approaches that input contexts sequentially into large language models (LLM), our method processes contexts in parallel and employs majority voting to determine the final answer. We also finetuned the LLM using a LoRA-based method to better handle quiz-style questions, achieving over 10 points gains against the baseline in terms of accuracy. For the DMR subtask, we introduce a modality-aware sensor encoder that processes numerical and textual sensor features separately, and enhance geolocation features by converting latitude/longitude data into address strings via k-nearest neighbor matching. Although our baseline performance is degraded from the official baseline due to the mismatch of data between the training and evaluation data, our approach improved the image-to-sensor retrieval performance from our baseline.

KEYWORDS

large language model, late fusion, modality-aware sensor encoder

TEAM NAME

ditlab

SUBTASKS

Retrieval Augmented Generation (Japanese)
Dense Multimodal Retrieval (English)

1 INTRODUCTION

The ditlab team participated in the Retrieval-Augmented Generation (RAG) and Dense Multimodal Retrieval (DMR) subtasks of the NTCIR-18 Transfer-2 task [9, 10]. The RAG subtask aims to develop efficient RAG techniques for open-domain question answering (QA). RAG improves large language models (LLMs) by incorporating external knowledge and addressing limitations in their training data [3, 4, 13]. Passage retrieval plays a vital role in open-domain QA [5, 8, 12], but managing multiple retrieved contexts is challenging. Quiz-style QA is particularly difficult due to intentionally obscured answers and misleading or irrelevant information in the retrieved passages. Keywords in such questions often do not directly indicate the correct answer, and inappropriate contexts in retrieval lead to incorrect predictions [6, 7, 17].

Sequential RAG methods, such as those in LlamaIndex, process top-ranked contexts one at a time, allowing LLMs to revise initial answer. However, this approach degrades accuracy when using multiple passages (e.g. top-2 or top-3), because irrelevant information causes the model to incorrectly revise correct answers. Moreover, this sequential process increases computational time.

To overcome these limitations, we propose a late fusion of LLM results, inspired by fusion-in-decoder (FiD) [1, 11, 16]. This method runs LLMs in parallel, generating answers independently for each context. The final answer is selected by majority voting, improving accuracy, and reducing computation time.

DMR subtask addresses the challenge of retrieving relevant information across different data modalities. In the Transfer-2 task, this involves working with the dataset of the NTCIR18-Lifelog6 task [19] that combines camera images and sensor data collected from wearable devices. Since image and sensor data differ significantly in terms of distribution, dimensionality, and sampling frequency, it is essential to construct retrieval methods that can interpret and align both modalities, in order to accurately capture the contextual correspondence.

Vector representations are often used to align different modalities as they allow heterogeneous data types to be compared with each other in a shared embedding space. However, sensor data include a mixture of numerical values and textual information, which makes it non-trivial to generate embeddings that are both meaningful and discriminative. To address this, we propose a sensor encoder that combines a multi-layer perceptron (MLP) with a text embedding model for numerical and string-based data, respectively. Furthermore, to improve the quality of embeddings for accurate retrieval, we transform latitude and longitude values into textual address representations using the k-nearest neighbor (k-NN) matching before encoding. In the retrieval process, either an image or a sensor record is used as a query, and similarity scores are computed against the data in the other modality to identify the most relevant matches.

2 METHODS FOR RAG SUBTASK

2.1 Dense passage retrieval

The Dense Passage Retrieval (DPR) [11] is used to retrieve the relevant passages. A question q and a set of passages p (e.g. from Wikipedia) are encoded into fixed-dimensional embeddings $E_q(q)$ and $E_p(p)$ using separate encoders. The similarity is calculated through the inner product $\sigma(q, p) = E_q(q)^\top E_p(p)$, and the upper n passages are retrieved. These are then provided to the LLM as a context along with the question. The encoders are trained using the following loss function (1), which maximizes the score of the positive passage relative to negatives:

$$\mathcal{L}(q_i, p_i^+, \mathcal{P}_i^-) = -\log \frac{e^{\sigma(q_i, p_i^+)}}{e^{\sigma(q_i, p_i^+)} + \sum_{p^- \in \mathcal{P}_i^-} e^{\sigma(q_i, p^-)}}, \quad (1)$$

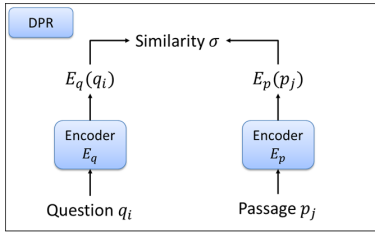


Figure 1: Deep passage retriever.

You are a quiz show contestant. Using the provided context, answer the query concisely in under 20 characters.

Context: ...

Query: ...

###Answer:

Figure 2: Prompt for conducting RAG.

where p_i^+ is a relevant passage and \mathcal{P}_i^- is a set of n irrelevant passages p^- . Since retrieval is a well-established process, the main challenge lies in how to efficiently incorporate diverse contexts into the LLM.

2.2 Finetuning for quiz answering

To improve answer accuracy, the LLM is fine-tuned using a low-rank approximation (LoRA). Training data consist of context-answer pairs in which the retrieved context was relevant but the generated answer was incorrect. The correct answers are added after the prompt ### Answer in Fig. 2, and the LLM is trained to reproduce these answers.

Since QA tasks often require short and concise answers, finetuning helps ensure consistency in the answer style, especially when specific formats are expected.

2.3 Generation of answer candidates by multiple LLMs

In the FiD method [11], the question and top- n passages are concatenated and processed jointly by the decoder. Inspired by this, we propose a late fusion approach utilizing LLM as shown in Fig. 3: each context is fed into separate LLM in parallel, and after multiple answers are prepared for answer candidates, a final answer is selected from the candidates by majority voting. Parallel processing reduces latency compared to the sequential approach such as LlamaIndex.

2.4 Majority voting to select final answer

In addition, another key benefit of our method is the ability to filter out invalid answers, such as those that simply repeat part of the query. For example, for the query “In the movie ‘West Side Story,’ what is the name of the other gang besides the Sharks?”, the answer “Sharks” is clearly incorrect and invalid.

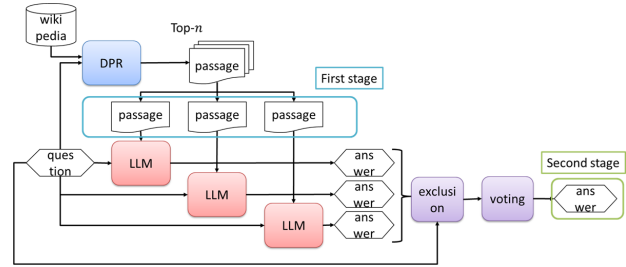


Figure 3: Late fusion of LLM results.

Up to n answers are generated from the top- n contexts. However, missing or invalid responses can reduce this number. The final answer is selected by majority voting. In case of a tie, the answer corresponding to the higher-ranked context is chosen. For example, if the top 5 answers are [a, a, b, b, c], the final answer would be a. Detailed descriptions are found in [15].

3 METHODS FOR DMR SUBTASK

The DMR subtask is evaluated using Mean Reciprocal Rank (MRR) in both image-to-sensor and sensor-to-image retrieval settings. We built our system on top of the baseline implementation [18], introducing several modifications to create effective embeddings for multimodal retrieval. In addition to the changes described below, we also replaced the ResNet-based image encoder with a vision transformer (ViT) model [2].

3.1 Modality-aware sensor encoder

In the baseline system, all sensor features including categorical text values are converted into numerical vectors and encoded using an MLP [10]. In contrast, we proposed a modality-aware sensor encoder that separately processes numerical and textual information to better preserve informative textual content for accurate retrieval.

Specifically, numerical features such as heart rate and sleep metrics are processed using a two-layer MLP, which outputs a 128-dimensional embedding. Categorical features including activity labels and address information are first converted into prompt-style text strings and then encoded using a lightweight Sentence-BERT variant to obtain sentence embeddings [14], which produces a 384-dimensional embedding. The numerical and textual embeddings are concatenated and passed through another MLP to project the sensor record into the same space as the image embeddings, allowing for cross-modal similarity computation.

3.2 Geolocation encoding via address mapping

While latitude and longitude provide accurate spatial coordinates, they do not inherently capture semantic cues that are useful for multimodal retrieval. To enrich the representation of location features, we converted raw coordinates into textual address strings.

Since online access to address information is unavailable during inference, we construct a reference database by associating each

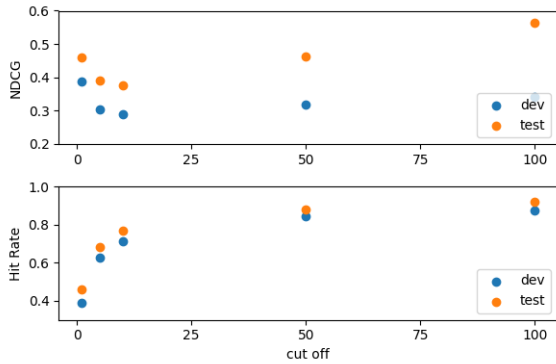


Figure 4: NDCG and Hit Rate of first stage of RAG subtask.

Table 1: Accuracy of second stage for each run on the RAG subtask.

Run ID	Description	Accuracy	
		dev	test
run1	baseline (LlamaIndex)	30.0	39.1
run2	top1 (no LoRA)	29.7	39.5
run3	top1 (LoRA)	35.7	43.7
run4	top1-7 (LoRA, majority voting)	40.4	50.3

coordinate with its corresponding address during training. At inference time, we retrieve the nearest address using k-NN matching and encode it as part of the textual input.

4 EXPERIMENTS FOR RAG SUBTASK

The RAG subtask is divided into two stages, and performance was evaluated for each stage: first stage (retrieval stage) and second stage (answer generation stage).

4.1 First stage

Figure 4 shows the Normalized Discounted Cumulative Gain (NDCG) and the Hit Rate of the first stage. In this analysis, we observed that the test accuracy tends to be higher than that of the development set across different evaluation metrics. The evaluation was carried out using various cut-off values: 1, 5, 10, 50, and 100.

In terms of NDCG, the score initially decreased to a cut-off of 10, after which it began to improve. On the other hand, the Hit Rate reached near saturation around a cut-off of 50, showing only a marginal difference of approximately 3 to 4 points between Hit Rate @ 50 and Hit Rate @ 100. The value of Hit Rate @ 1 was approximately 0.388 for the development set and 0.461 for the test set, suggesting that the use of multiple contexts is necessary to achieve greater recall and better performance.

For reference, LlamaIndex yielded NDCG @ 1 (= Hit Rate @ 1) of 0.241 (dev) and 0.407 (test), although, since the passage division method differs between DPR and LlamaIndex, the performance cannot be compared.

Table 2: MRR scores for each run on the DMR subtask.

Run ID	Description	img2sen	sen2img
baseline	Official baseline (100 epochs)	0.2830	0.2788
run1	Full data + ViT + sensor encoder	0.0523	0.0410
run2	run1 + address k-NN matching	0.0710	0.0392
run3	run2 + larger batch size	0.0746	0.0414
run4	run3 + larger network	0.0947	0.0326

4.2 Second stage

Table 1 shows the results of the second stage, showing that the baseline using the LlamaIndex (run1) and the RAG model with the passage of the highest similarity score (run2) achieved similar levels of accuracy¹. From run2 to run4, the same documents retrieved in the first stage were used. When the LLM was fine-tuned using LoRA (run3), the accuracy improved by 6.0 points on the development set and 4.2 points on the test set. This result shows that fine-tuning the model is an effective way to teach how LLM answers the question. Furthermore, aggregation of the top-7 answers with the final answer selection by majority voting (run4) achieved an additional improvement of 4.7 points in the development set and 6.6 points in the test set compared to run3.

5 EXPERIMENTS FOR DMR SUBTASK

Table 2 shows the retrieval performance in terms of MRR for both image-to-sensor and sensor-to-image tasks. The best-performing variant of our system achieved an MRR of 0.0947 for image-to-sensor and 0.0414 for sensor-to-image retrieval, while the baseline scores were 0.2829 and 0.2788, respectively.

One likely reason for the observed performance degradation is the difference in the training data used. Although the baseline was trained on sensor and image data from March 2020, our system was trained on the full dataset covering an 18-month period from 2019 to 2020. If the evaluation is restricted to data from a specific time window, training on the full dataset may have caused the model to generalize across a broader range of distributions. As a result, the learned representations may not have been sufficiently specialized for the evaluation data, leading to a reduced retrieval performance.

6 CONCLUSIONS

We participated in the RAG and DMR subtasks of the NTCIR-18 Transfer-2 task. In the RAG subtask, our proposed method that combined LoRA-based fine-tuning and late fusion by majority voting significantly improved the answer accuracy compared to baseline in more than 10 points. The late fusion of different contexts allowed us to mitigate the risk of irrelevant contexts, demonstrating the effectiveness of model fusion strategies in open-domain QA. In the DMR subtask, although our approach using ViT-based image encoders and a modality-aware sensor encoder underperformed relative to the baseline, we observed that training on full data was ineffective for this subtask. This result highlights the importance of the selection of useful data from the entire dataset. However,

¹LLMs were common in all runs.

our approach improved our baseline (run1) in image-to-sensor retrieval, which can improve the official baseline performance when the same training data are used.

REFERENCES

- [1] Michiel de Jong, Yury Zemlyanskiy, Joshua Ainslie, Nicholas FitzGerald, Sumit Sanghai, Fei Sha, and William Cohen. 2023. FIDO: Fusion-in-Decoder optimized for stronger performance and faster inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11534–11547. <https://doi.org/10.18653/v1/2023.findings-acl.732>
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).
- [3] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) (KDD '24). Association for Computing Machinery, New York, NY, USA, 6491–6501. <https://doi.org/10.1145/3637528.3671470>
- [4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint* (2023). <https://arxiv.org/abs/2312.10997>
- [5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 368, 10 pages.
- [6] Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with Retrieval: Faithful Large Language Model Inference. *arXiv:2301.00303 [cs.CL]* <https://arxiv.org/abs/2301.00303>
- [7] Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasaratopoulou, and Jeff Pan. 2023. Retrieval Augmented Generation with Rich Answer Encoding. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 1012–1025. <https://doi.org/10.18653/v1/2023.ijcnlp-main.65>
- [8] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfay (Eds.). Association for Computational Linguistics, Online, 874–880. <https://doi.org/10.18653/v1/2021.eacl-main.74>
- [9] Hideo Joho, Atsushi Keyaki, Yuuki Tachioka, and Shuhei Yamamoto. 2024. Building test collections for Japanese dense information retrieval technologies and beyond. In *Proceedings of The First Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research (EMTCIR '24)*. NII.
- [10] H Joho, A Keyaki, Y Tachioka, and S Yamamoto. 2025. Overview of the NTCIR-18 Transfer-2 Task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*. Tokyo.
- [11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-Tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [12] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020). <https://arxiv.org/abs/2004.12832>
- [13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, Canada) (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- [15] Yuuki Tachioka. 2025. Fusion-in-LLM: A Parallel Approach to Improve Retrieval-Augmented Generation for Quiz-Based Question Answering. In *The 30th International Conference on Natural Language & Information Systems*. Kanazawa, Japan, accepted.
- [16] Cunxiang Wang, Haofei Yu, and Yue Zhang. 2023. RFID: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2473–2481. <https://doi.org/10.18653/v1/2023.findings-acl.155>
- [17] Zizhong Wei, Dengrong Huang, Jichen Zhang, Chen Song, Sijia Zhang, Jianing Zhang, Zhaochuan Li, Kai Jiang, Rui Li, and Qiang Duan. 2024. GARAG: A General adaptive question-answering system based on RAG. In *Proceedings of the 2024 International Conference on Cloud Computing and Big Data* (Dali, China) (IC-CBD '24). Association for Computing Machinery, New York, NY, USA, 442–447. <https://doi.org/10.1145/3695080.3695156>
- [18] Shuhei Yamamoto and Noriko Kando. 2025. Temporal Closeness for Enhanced Cross-Modal Retrieval of Sensor and Image Data. In *MultiMedia Modeling*, Ichiro Ide, Ioannis Kompatsiaris, Changsheng Xu, Keiji Yanai, Wei-Ta Chu, Naoko Nitta, Michael Riegler, and Toshihiko Yamasaki (Eds.). Springer Nature Singapore, Singapore, 170–183.
- [19] Liting Zhou, Quang-Linh Tran, Hoang Bao Le, Duc-Tien Dang-Nguyen, Healy Graham, ChenYang Lyu, Tianbo Ji, Hsin-Hung Chen, Longyue Wang, Hideo Joho, Ly-Duyen Tran, Dajung Jung, Alice Schoeffmannand Klaus, Thanh Binh Nguyen, and Cathal Gurrin. 2025. Overview of the NTCIR-18 Lifelog-6 Task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-18)*. Tokyo, Japan.