YMX2L at the NTCIR-18 Transfer-2 Task

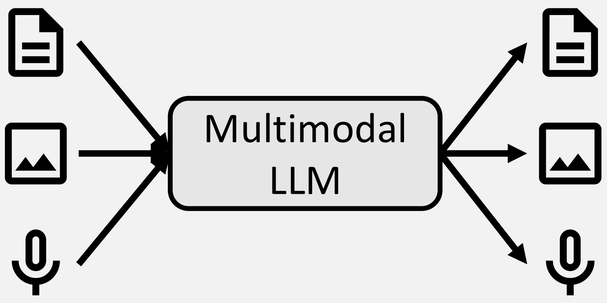
Riku Mizuguchi [†1], Takeshi Yamazaki [†1], and Shuhei Yamamoto [†2]

[†1] College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba, Japan
[†2] Institute of Library, Information and Media Science, University of Tsukuba, Japan

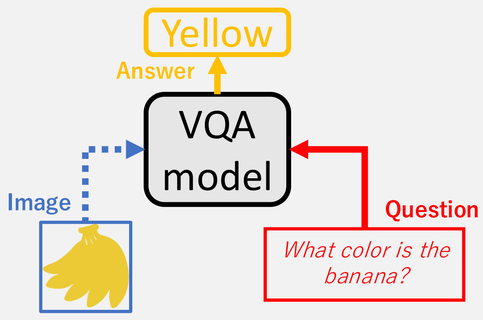
Background

- The advancements in LLMs have paved the way for cross-modal information processing, enabling interactions across various modalities, including text, images, and audio.

Multi modal LLM can handle multiple modalities.

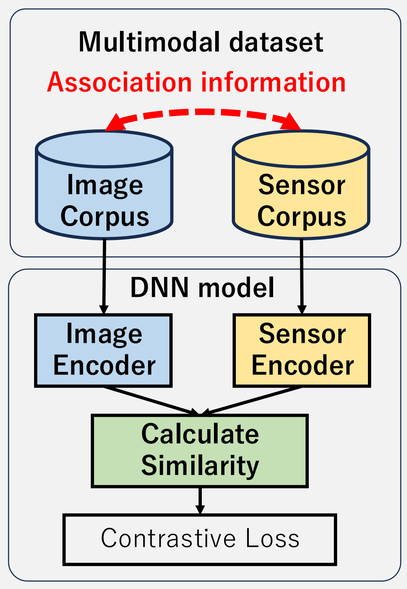


Visual Question Answering (VQA)

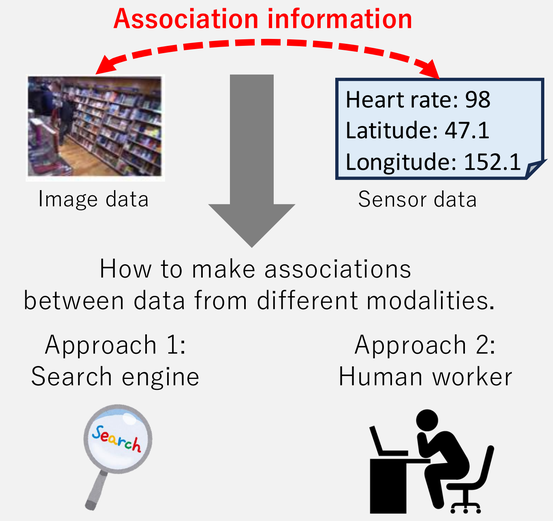


- While large text and image corpora exist (e.g., Microsoft COCO and Flickr30k), they often lack corresponding data from other modalities, limiting their utility in developing robust cross-modal information access technologies.

Cross-modal info. access construction flow

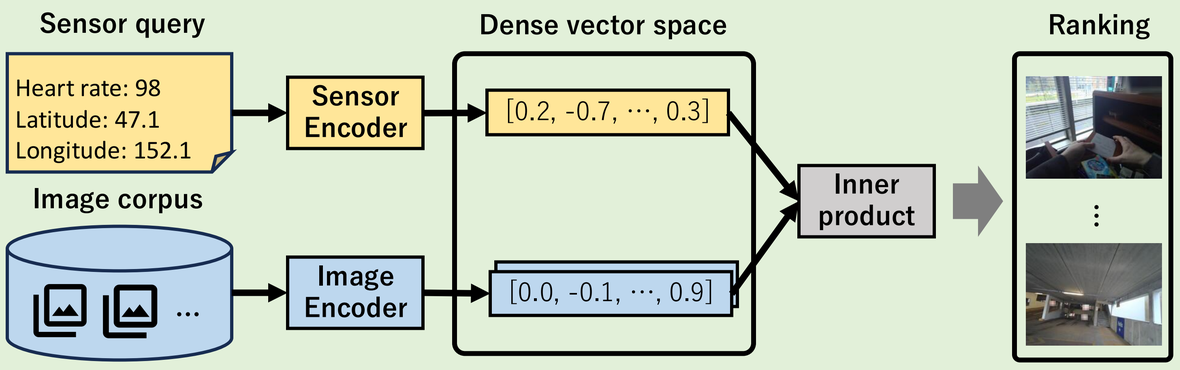


Challenges in multimodal dataset construction



Our Task

- We introduce a dense retrieval method for retrieving relevant data across multiple modalities.
- In our approach, image and sensor data are projected into this shared vector space using modality-specific encoders.



Contrastive learning for dense retrieval

Positive sample: A relevant image to sensor query
Negative samples: Set of irrelevant image to the query

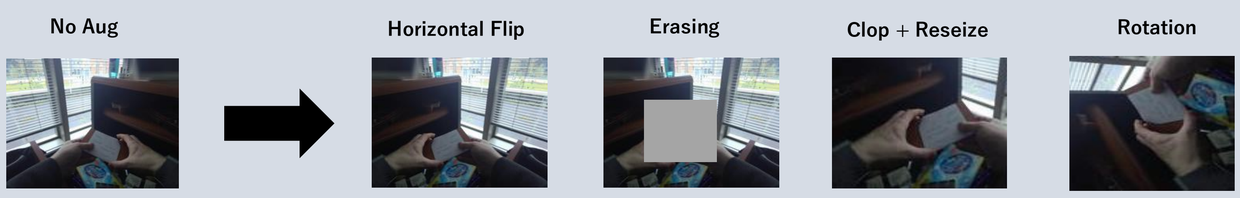
$$\mathcal{L}\left(\mathbf{q}^{\text{sen}}, \mathbf{d}_{+}^{\text{img}}, D_{-}^{\text{img}}\right) = -\log \frac{\exp\left(\text{sim}(\mathbf{q}^{\text{sen}}, \mathbf{d}_{+}^{\text{img}})\right)}{\exp\left(\text{sim}(\mathbf{q}^{\text{sen}}, \mathbf{d}_{+}^{\text{img}})\right) + \sum_{\mathbf{d}_{-}^{\text{img}} \in D_{-}^{\text{img}}} \exp\left(\text{sim}(\mathbf{q}^{\text{sen}}, \mathbf{d}_{-}^{\text{img}})\right)}$$

- Dense retrieval training tends to benefit from a larger set of hard negatives and in-batch negatives.
- In our task, positive (relevant) and hard negative (irrelevant) samples are not directly available.
- Therefore, it is necessary to extract an appropriate set of hard negative samples and positive sample.

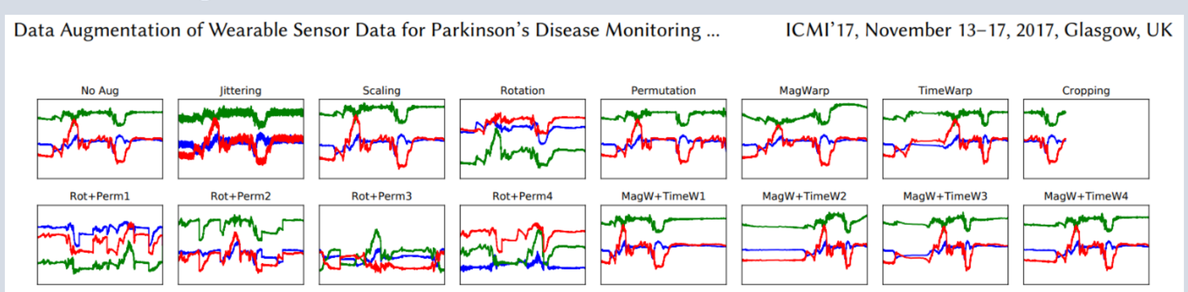
Method

Data augmentation-based approach

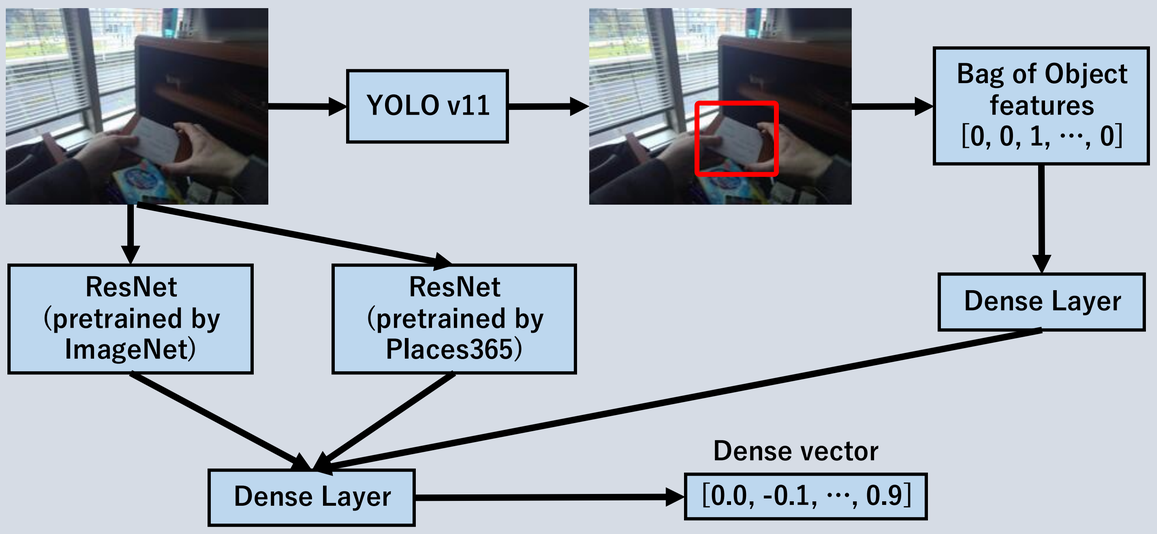
Image data augmentation



Sensor data augmentation



Object detection-based approach



Results

- The experimental evaluation utilizes LSC'24 dataset, which is a multi-modal dataset capturing users' daily activities.
- The dataset consists of images captured by egocentric camera and sensor data by a smart tracker.

Details of the experimental dataset

Split	Period	# Data Points	# Days
Training	Jan. 01, 2020 ~ Feb. 29, 2020	27,907	60
Validation	Mar. 01, 2020 ~ Mar. 16, 2020	21,198	16

Examples of created topics

Topic ID	Timestamp	Query	Relevant Data
sen2img_0316	2020-03-16 16:50:00	hr: 0.015 lat: -0.121 lng: 0.055	20200316_165048.jpg
img2sen_0319	2020-03-19 09:03:00	20200319_090305.jpg	hr: 1.891 lat: -0.121 lng: 0.055

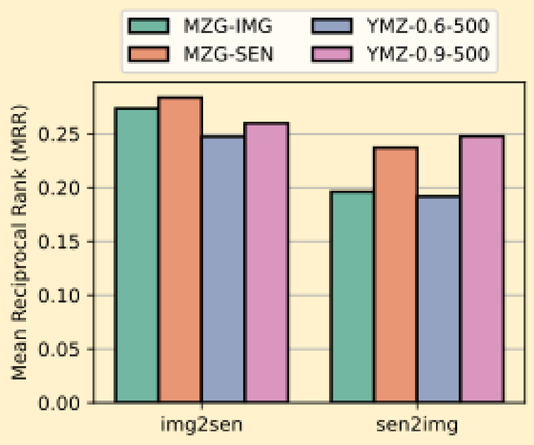

20200316_165048.jpg


20200319_090305.jpg

Experimental setup

Image Encoder	■ Linearly transformed 3 x 224 x 224. ■ Two Resnet50 pretrained by ImageNet/Places365
Sensor Encoder	■ Normalized to avg. 0 and s.d. 1. ■ Two MLP layers.
# dense dimentions	512
Mini-batch	256
Optimizer	Adam
# Iterations	100, 200, 500
Evaluation metrics	$MRR = \frac{1}{\lvert Q \rvert}\sum_{i=1}^{Q}\frac{1}{rank_i}$

Official Results (evaluated by Mean Reciprocal Rank)



Our methods

MZG-IMG	Image augmentation-based
MZG-SEN	Sensor augmentation-based
YNZ-0.6-500	Object detection with threshold α =0.6, trained for 500 epochs
YMZ-0.9-500	Object detection with threshold α =0.9, trained for 500 epochs

Monthly DMR evaluation results