ditlab at the NTCIR-18 Transfer-2 Task

Yuuki Tachioka and Yasunori Terao

DENSO IT LAB

Abstract

- Participation in RAG (Retrieval Augmented Generation) and DMR (Dense Multimodal Retrieval) subtasks
- Challenges of RAG subtask:
  - efficient context integration for LLM
- Challenges of DMR subtask:
  - cross-modal retrieval
  - geolocation encoding

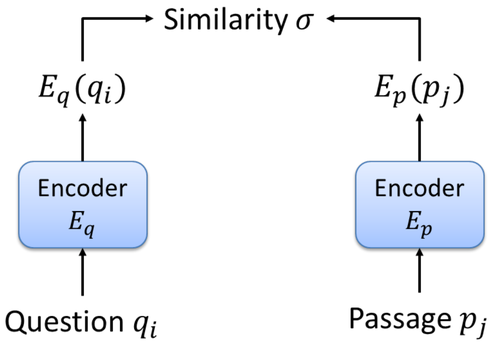
Methods for RAG Subtask (first stage)

Dense Passage Retrieval

- DPR for efficient context retrieval using inner product similarity
  - Encodes queries and passages into dense vectors
  - Loss function for training:

$$L(q_i, p_i^+, P_i^-) = -\log \frac{e^{\sigma(q_i, p_i^+)}}{e^{\sigma(q_i, p_i^+)} + \sum_{p^- \in P_i^-} e^{\sigma(q_i, p^-)}}$$

  - where p_i^+ is a relevant passage and P_i^- is a set of n irrelevant passages



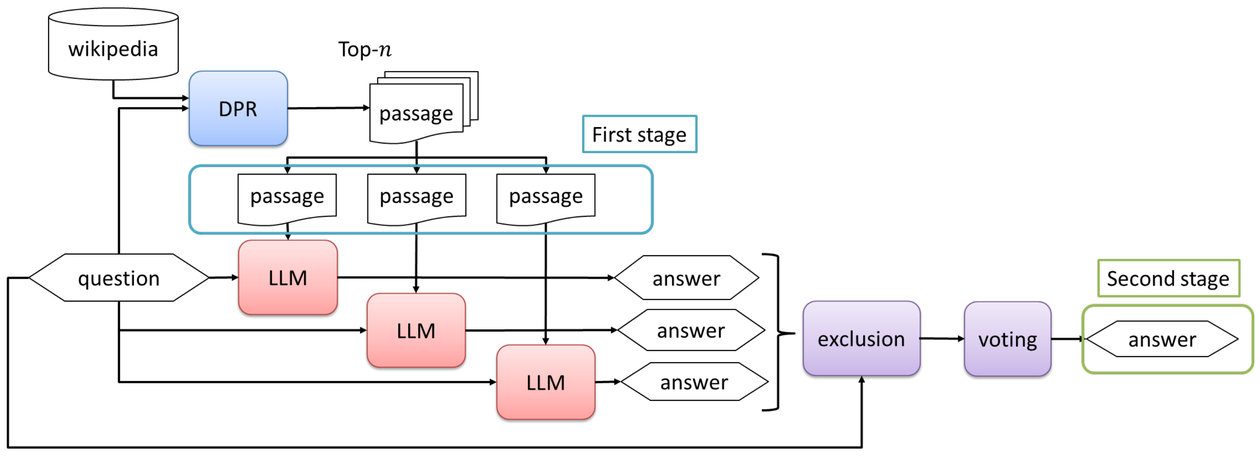
Methods for RAG Subtask (second stage)

Finetuning of LLM for Quiz Answering

- LoRA-based fine-tuning to improve answer consistency and accuracy
- Trains LLMs to handle short, concise answers for quiz-style tasks on the data where the DPR retrieved relevant passage but LLM generated wrong answer

LLM fusion to handle multiple contexts

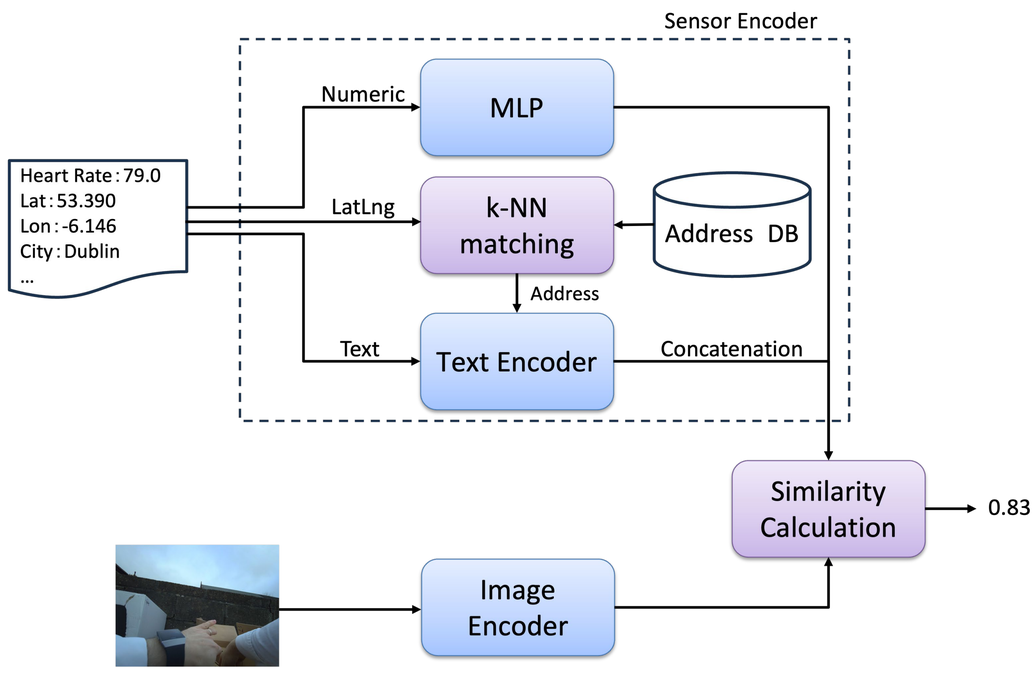
- Generation of Answer Candidates by Multiple LLMs
  - Each passage independently processed by separate LLMs
- Majority Voting to Select Final Answer
  - Aggregates multiple LLM outputs to reduce noise and improve accuracy



Methods for DMR Subtask

Modality-aware sensor encoder

- Numerical and textual features are encoded separately using MLP and Sentence-BERT to preserve modality-specific information
- Concatenated embeddings are used to compute similarity with image embeddings for multi-modal retrieval



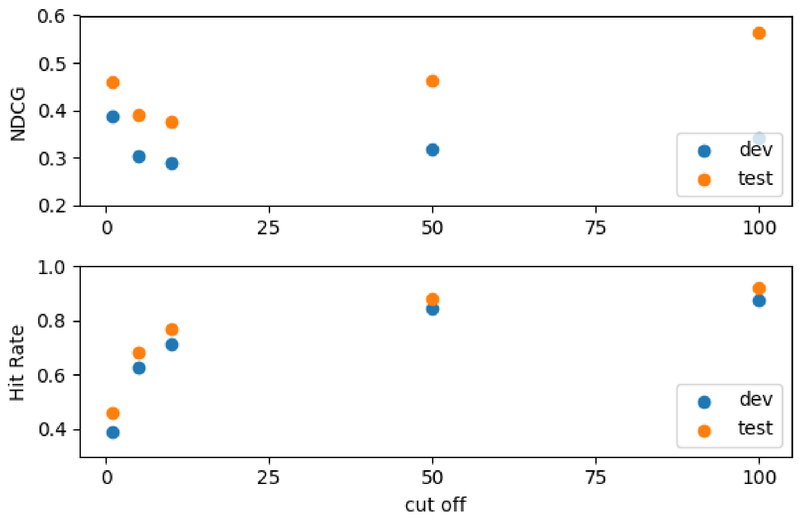
Geolocation encoding via address mapping

- Latitude and longitude are converted into address strings to provide semantically rich input for retrieval
- During inference, addresses are retrieved from a reference database using k-NN matching and encoded as text

Experiments for RAG Subtask

First Stage (passage retrieval performance)

- NDCG initially decreased to a cut-off of 10, after which it began to improve
  - Hit Rate reached near saturation around @ 50
  - Hit Rate @ 1 = 0.388 for dev and 0.461 for the test



Second Stage (accuracy of the final answer)

- Accuracy improvements with LoRA (run3) and majority voting (run4)

Run ID	Description	Dev Accuracy	Test Accuracy
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

Experiments for DMR Subtask

Performance degradation compared to baseline

- Our best model achieved MRR 0.0947 (img2sen) and 0.0414 (sen2img), compared to baseline scores of 0.2829 and 0.2788
- Using a broader 18-month training set may have led to poorer specialization for evaluation data limited to Mar-Jun 2020

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

Post-submission experimental results

- Code and data refinements led to reproduced performance comparable to the baseline (using the same training data)
- Slight increase in img2sen and decrease in sen2img suggest that the proposed encoder (hybrid) may amplify distributional gaps between sensor and image embeddings
- To improve both directions of retrieval, better alignment between sensor and image embeddings may be essential

Conclusion

- We participated in the RAG and DMR subtasks
- RAG subtask
  - our proposed method (LoRA-based fine-tuning and late fusion) significantly improved the answer accuracy compared to baseline in more than 10 points
- DMR subtask
  - our approach using ViT-based image encoders and a modality-aware sensor encoder underperformed but can improve the official baseline performance when the same training data are used