Meeting Note

Author: Henry Lin
Date: 2024/08/05 – 2024/08/09

Weekly Summary
Retrieval-Augmented Generation (RAG) [1].
LLM Hallucination Mitigation.

Plan for Next Week
IMAGDressing-v1: Customizable Virtual Dressing (arXiv: 2407.12705)
Shape of Motion: 4D Reconstruction from a Single Video (arXiv: 2407.13764)

Details
To address the problem of hallucination, one of the methodologies is Retrieval-Augmented Generation (RAG) [1]. The Large Language Model (LLM) can store knowledge in the trained parameters, as they can answer questions without accessing the external memory. But they cannot easily expand or revise their parameters, the authors introduced a hybrid parametric and non-parametric memory called RAG.

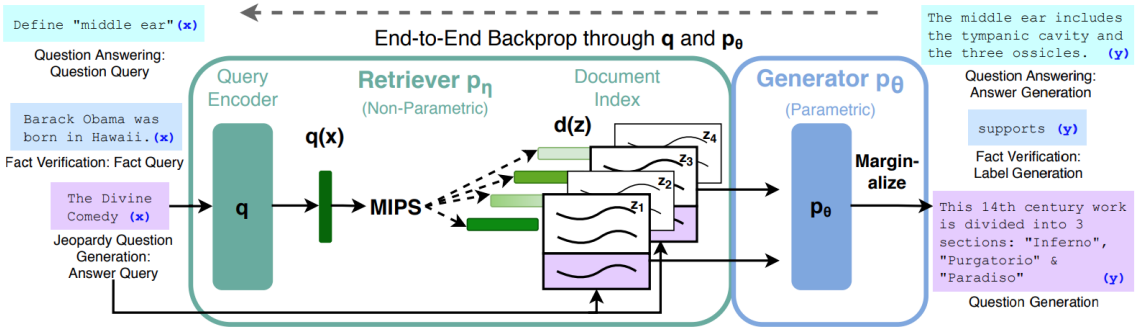


Figure 1. The overall proposed approach. They combine pre-trained Retriever p_η and seq2seq Generator p_θ, and fine-tune end-to-end. The Maximum Inner Product Search (MIPS) is used to find the top-K documents z_i, and z is inputted into the Generator as latent variable.

The non-parametric memory is a dense vector index of any data (e.g. the Wikipedia) which is accessed by a pre-trained neural retriever that will find the top-K latent documents related to the query. And the parametric memory is pre-

trained generator (e.g. seq2seq Transformer) which take the query and the top-K latent documents as input and output the answers.

Since RAG is trained end-to-end, they treat the retrieved document as a latent variable, and two models to marginalize over the latent documents are proposed, that is, RAG-Sequence Model and RAG-Token Model.
Both RAG-Sequence Model and RAG-Token Model retrieve the top-K documents at first, but while generating the output, the former using the retrieved document to generate the entire output y; and the latter can choose content from different documents for each token y_i. And the marginalization process is also different for each method, for RAG-Sequence Model, it treats the retrieved document as a single latent variable z and marginalize over it to calculate the probability $p(y|z)$; for RAG-Token Model, it marginalizes over different latent documents for each token to calculate the probability $p(y|z)$. The following are the formulas for these two methods, the p_η representing the Retriever and p_θ representing the Generator.

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$

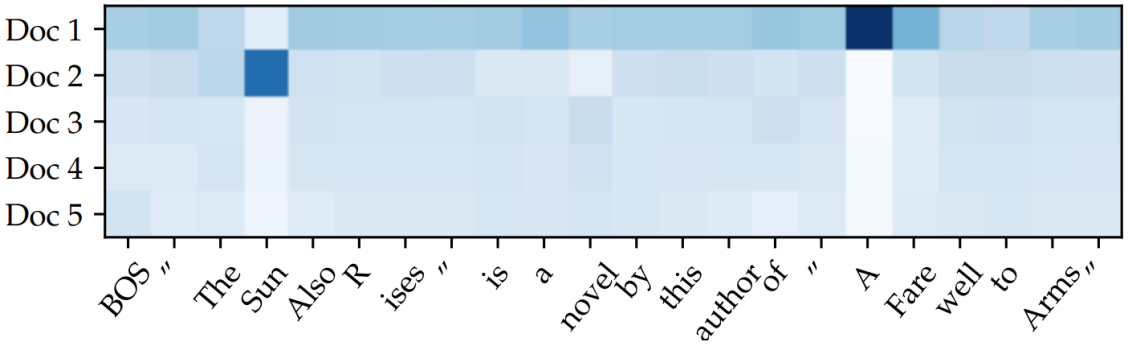
The authors experiment four types of tasks, that is, open-domain question answering (QA), abstractive QA, Jeopardy question generation and fact verification. Table 1 shows the results of open-domain QA, and Table 2 shows the results of the remaining tasks. For each benchmark, please refer to the Remarks.

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set.

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		



Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel **"A Farewell to Arms"** (1929) ...

Document 2: ... artists of the 1920s "Lost Generation" expatriate community. His debut novel, **"The Sun Also Rises"**, was published in 1926.

Figure 2. RAG-Token document posterior for each generated token for input "Hemingway" for Jeopardy generation with 5 retrieved documents, showing why RAG-Token may perform better than RAG-Sequence since it can generate question that combine content from different documents.

In this study, the retriever p_η is based on Dense Passage Retrieval (DPR) [7]. Two BERT-based encoders are used to obtain the query and document representations, and then Maximum Inner Production Search (MIPS) is used to find the top-K documents; and the generator p_θ could be any generative model.

Remarks

"Marginalize" means to eliminate hidden variables by summing or integrating over them, so that the focus is only on the observed variables. Specifically, the latent documents here refer to hidden data that is not directly observed. By marginalizing, we can obtain a distribution of generated text based only on the observed data. In short, "marginalize" means to "remove the data we don't care about".

Natural Questions (NQ) [2]: A dataset containing real user questions issued to Google search, along with answers found from Wikipedia. It's designed for training and evaluating automatic question-answering systems.

TriviaQA (TQA) [3]: A realistic text-based question-answering dataset with 950K question-answer pairs from 662K documents collected from Wikipedia and the web. It's more challenging than standard QA benchmarks like SQuAD due to longer context and indirect answers.

WebQuestions (WQ) [4]: A dataset using Freebase as the knowledge base, containing 6,642 question-answer pairs. Questions are centered around named entities and are answerable by Freebase.
CuratedTrec (CT) [5]: Introduced in TREC 2019, it provides large reusable datasets for training and evaluating deep learning and traditional ranking methods in document and passage retrieval tasks.
FEVER [6]: The Fact Extraction and Verification (FEVER) dataset focuses on fact-checking and verification. It contains claims, evidence, and labels for factuality1.
QB: Unfortunately, the paper does not explicitly define "QB." It might refer to a specific benchmark or task not widely known.

References

[1] LEWIS, Patrick, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 2020, 33: 9459-9474.
[2] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*, 2019. URL https://tomkwiat.users.x20web.corp.google.com/papers/natural-questions/main-1455-kwiatkowski.pdf.
[3] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://www.aclweb.org/anthology/P17-1147.
[4] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/ D13-1160.
[5] Petr Baudiš and Jan Šedivy. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 222–228. Springer, 2015. URL

https://link.springer.com/chapter/10.1007% 2F978-3-319-24027-5_20.
[6] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://www.aclweb.org/anthology/N18-1074.
[7] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. URL https://arxiv.org/abs/2004.04906.