Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[⋆],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;

HUMANE Lab 박현빈

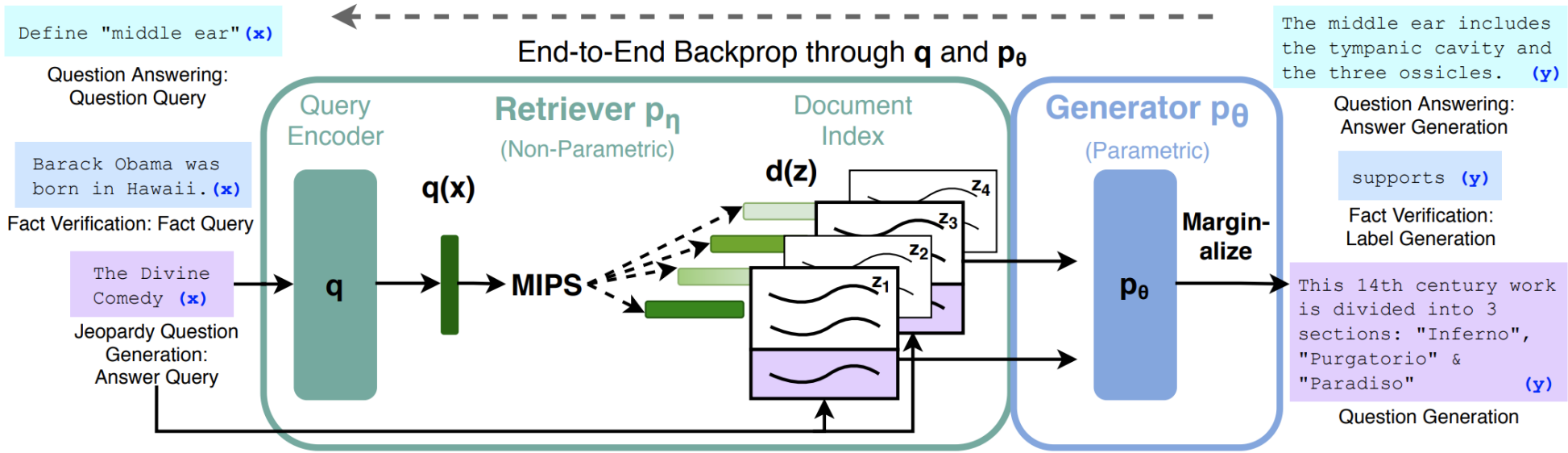
NeurIPS 2020

25.02.07

Background

- Previous research has used non-parametric memory to provide clear interpretability of the prediction process and reduce hallucinations.
- However, non-parametric memory has only been used for extractive tasks and has not been applied to generation tasks.

Model


Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
Retriever pη
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z1
z2
z3
z4
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Model

- RAG-Sequence Model
  - Referencing a single document z when generating a sequence

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

- RAG-Token Model
  - Referencing multiple documents z when generating each token

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

Model

- Retrieval : DPR
  - Retrieving documents with high MIPS scores for a given query
  - Using two encoders: a document-specific BERT encoder and a query-specific BERT encoder
  - Trained to retrieve documents containing answers to TriviaQA and Natural Questions

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right) \qquad \mathbf{d}(z) = \mathrm{BERT}_d(z), \;\; \mathbf{q}(x) = \mathrm{BERT}_q(x)$$

- Generator : BART-Large

Model

- Training
  - train without any direct supervision on what document should be retrieved
  - minimize the negative (marginal) log-likelihood
  - Adam optimizer
  - keep the document encoder fixed, only fine-tuning the query encoder BERT and BART generator

Model

- Decoding
  - RAG-Token : searching for the optimal answer using standard beam search
  - RAG-Sequence : performing beam search for each document z

Experiments

- Open-domain QA
  - Train RAG by directly minimizing the negative log-likelihood of answers

	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**

EM scores

  - RAG can generate correct answers even when the correct answer is not in any retrieved document

Experiments

- Abstractive QA

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		

- RAG outperforms BART and achieves results close to state-of-the-art models that use gold passages
- RAG has fewer hallucinations than BART and tends to generate factually correct text more often

Experiments

- Jeopardy Question Generation
  - A task that generates the question **"In 1986 Mexico scored as the first country to host this international sports competition twice."** given the answer entity **"The World**

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		

  - Q-BLEU is a variant of BLEU with a higher weight for matching entities and has higher correlation with human judgement
  - Human evaluation also determines that RAG generates more factual and specific questions than BART.

Experiments

- Fact Verification
  - FEVER requires classifying whether a claim is supported or refuted by Wikipedia, or whether there is not enough information to decide
  - Unlike most other approaches to FEVER, supervision on retrieved evidence is not used.
  - In real-world applications, retrieval supervision signals aren't available, and models that do not require such supervision will be applicable to a wider range of tasks.

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		

  - RAG achieved a score similar to Sota models, which are trained using retrieval supervision, which RAG does not require

Additional Results

- Effect of Retrieving more documents