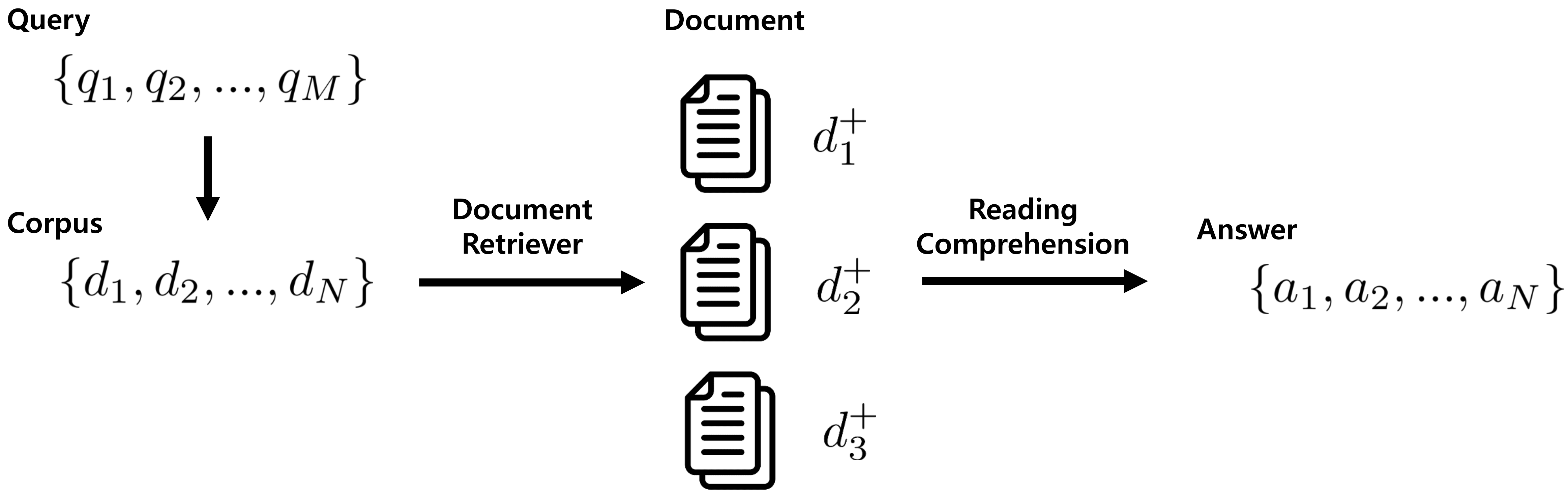
CAPSTONE: Curriculum Sampling for Dense Retrieval with Document Expansion

EMNLP 2023

The University of Hong Kong
Microsoft Research Asia
Microsoft

Background

- **Text Retrieval**
  - Retriever-Reader Model
  - Open-Domain Question Answering



Document by Lano Yado from Noun Project (CC BY 3.0)
Danqi Chen et al. Reading Wikipedia to Answer Open-Domain Questions. ACL, 2017.

Background

- **Text Retrieval**
  - Sparse Retrieval
    - Lexical term matching

$$\{0, 1, ..., 1, 0, ..., 1\} \in d_1$$

  - Dense Retrieval
    - Neural network-based model

$$\{-1.03, 1.72, ..., 3.42, -2.32, ..., 2.34\} \in d_1$$

Part 1.

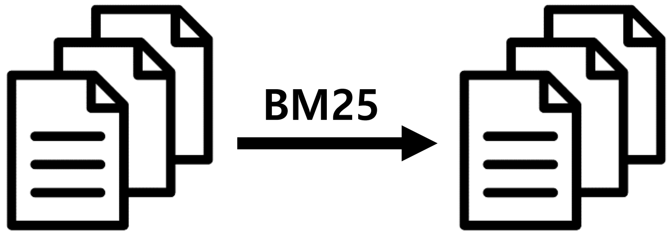
Background

- **Sparse Retrieval**
  - Lexical term matching
  - Bag-of-Words methods

TF-IDF: Importance of a word

$$W_{t,d} = tf_{t,d} \cdot \log(\frac{N}{df_t})$$

Reranking: Top-K Document



$\{d_1, d_2, ..., d_{20}\}$ $\{d_{11}, d_2, ..., d_6\}$

BM25: Ranking the Relevance of Document to Query

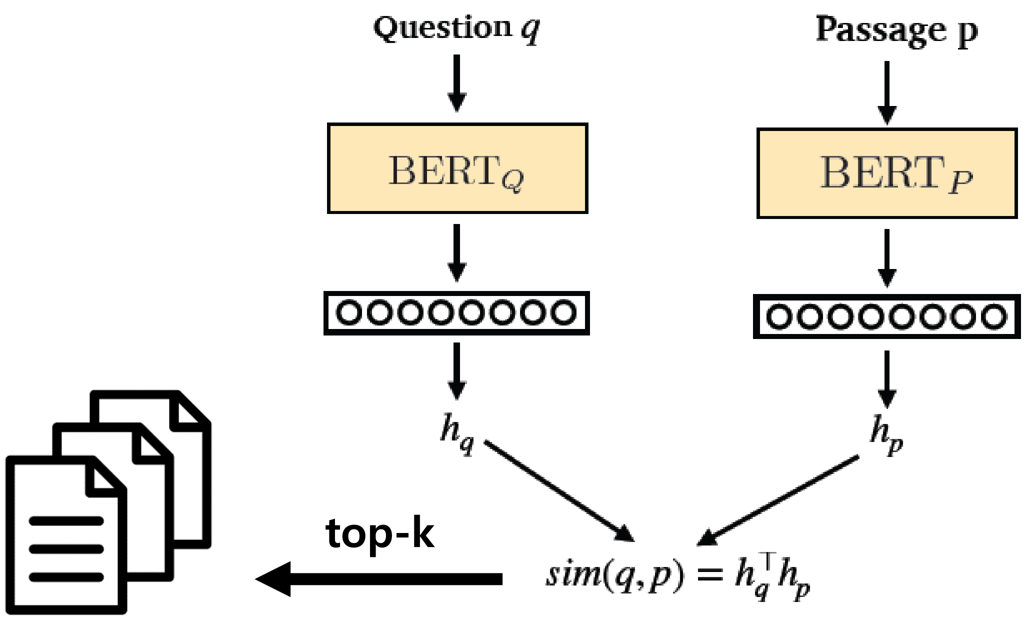
$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science. 1996.
Stephen Robertson et al. Okapi at TREC-3. In Overview of the Third Text REtrieval Conference(TREC-3). pages 109–126. 1995.

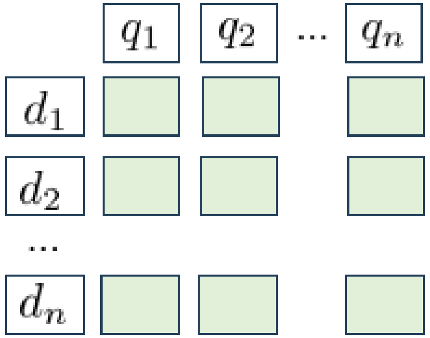
Background

- **Dense Retrieval**
  - Coarse-grained Representation
  - Neural network-based model

Dense Passage Retrieval (DPR)



Similarity Score



$$sim(q,p) = h_q^\top h_p$$

Training Objective

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$$

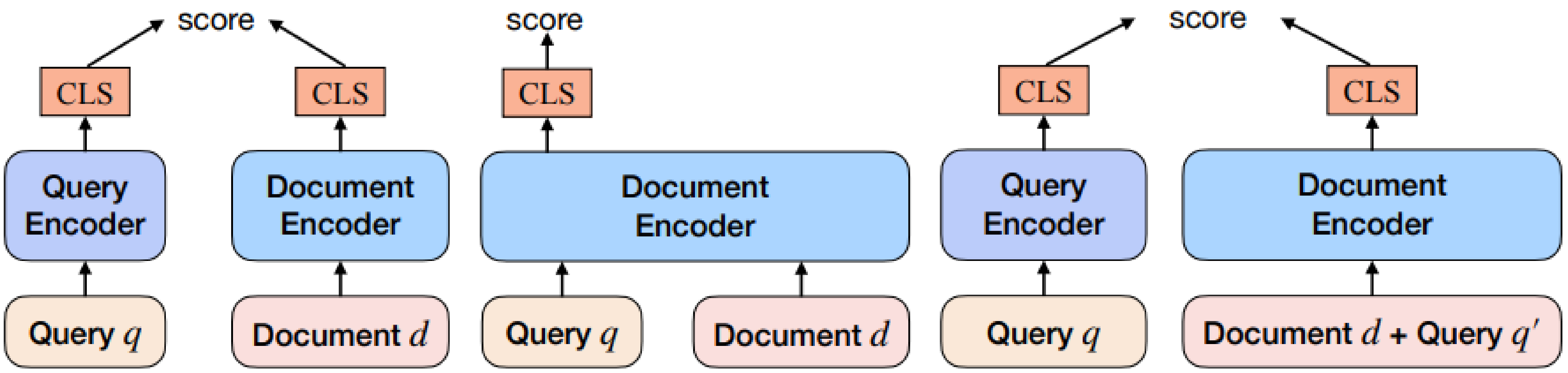
$$\boxed{\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) \\ = & -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\mathrm{sim}(q_i, p_{i,j}^-)}} \end{aligned}}$$

Document by DinosoftLabs from Noun Project (CC BY 3.0)
Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020.
Danqi Chen et al. Open-Domain Question Answering. ACL, 2020.

Background

- **Dense Retrieval**
  - Encoder Architecture

Dense Passage Retrieval (DPR) **Li Et Al. (2022)**



(a) Dual-encoder (b) Cross-encoder (c) Dual-cross-encoder

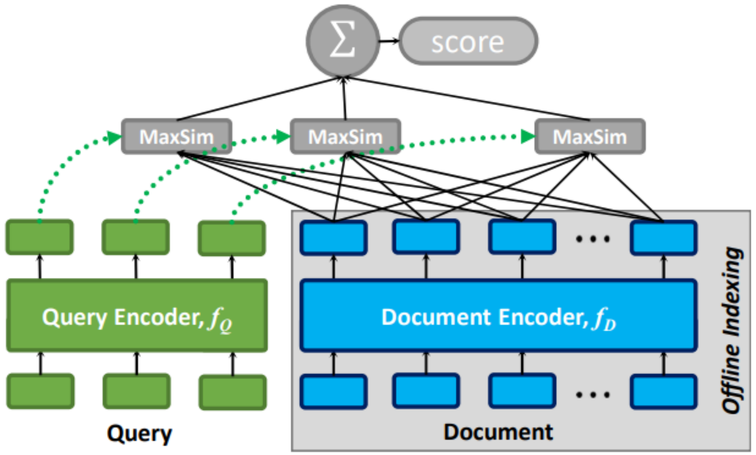
$$sim(q,d) = E_q(q)^T E_d(d). \quad sim(q,d) = FCL(E(d+q)), \quad sim(q,d) = E_q(q)^T E_d(d+q\prime).$$

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020
Rodrigo Nogueira et al. Document Expansion by Query Prediction. arXiv preprint arXiv:1904.08375, 2019.
Zehan Li et al. Learning Diverse Document Representations with Deep Query Interactions for Dense Retrieval. arXiv preprint arXiv:2208.04232, 2022.

Introduction

- **Dual-cross-encoder**
  - Query-related document representation

ColBERT



Late Interaction

$$S_{q,d} := \sum_{i \in [|E_q|]} \max_{j \in [|E_d|]} E_{q_i} \cdot E_{d_j}^T$$

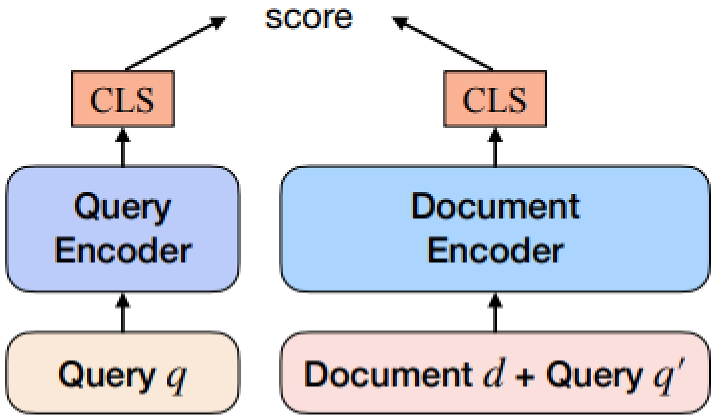
Maximum Inner Product Search (MIPS)

Trade-Off

Retrieval Efficiency ↓

Performance ↑

Dual-cross-encoder(Li Et al, 2022)



Pre Interaction

$$sim(q,d) = E_q(q)^T E_d(d + q\prime).$$

Interaction-Aware Document Representation

Document Expansion

Training

$$d + q'$$

Inference

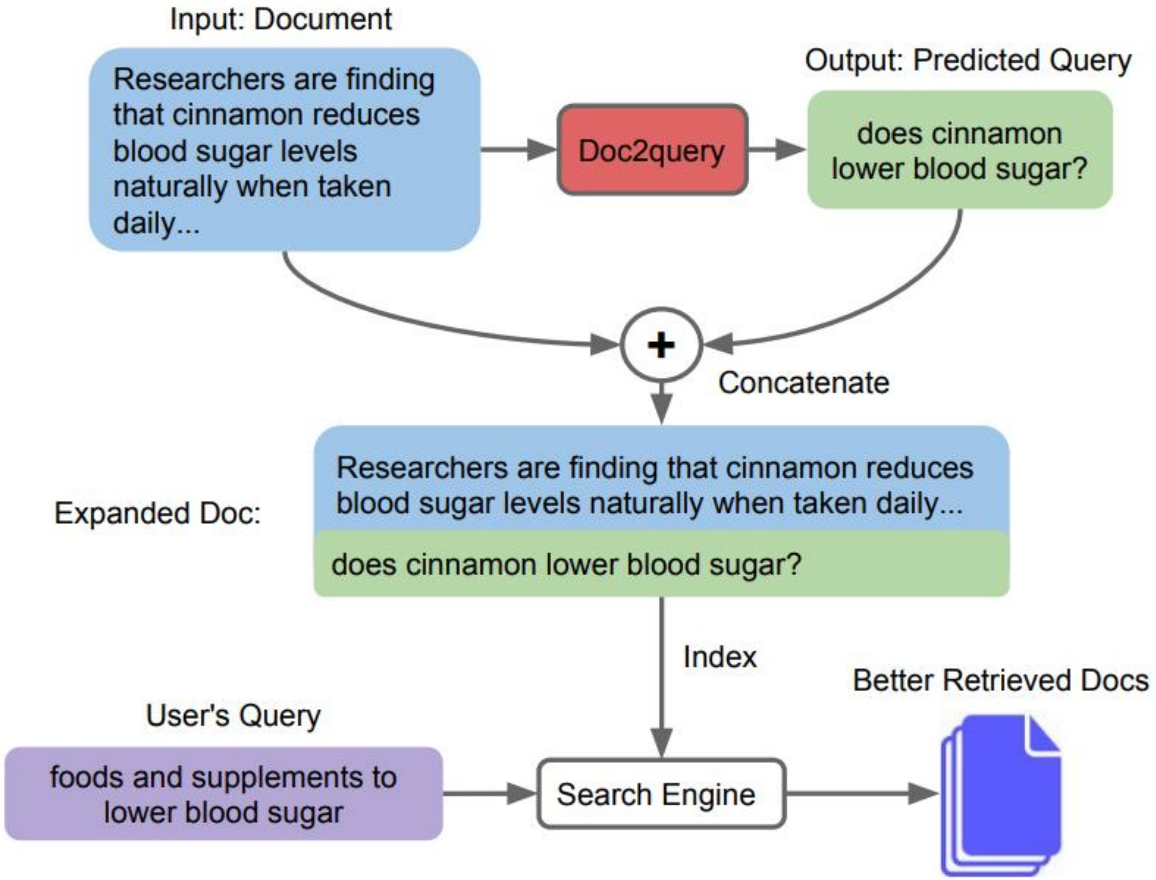
only d

Retrieval Efficiency ↑

Rodrigo Nogueira et al. Document Expansion by Query Prediction. arXiv preprint arXiv:1904.08375, 2019.
Omar Khattab et al. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR, 2020.
Zehan Li et al. Learning Diverse Document Representations with Deep Query Interactions for Dense Retrieval. arXiv preprint arXiv:2208.04232, 2022.

Introduction

- **Dual-cross-encoder**
  - Document Expansion



Doc2Query

- After training, predict a set of queries
- 10 queries from top-k random sampling

Effect

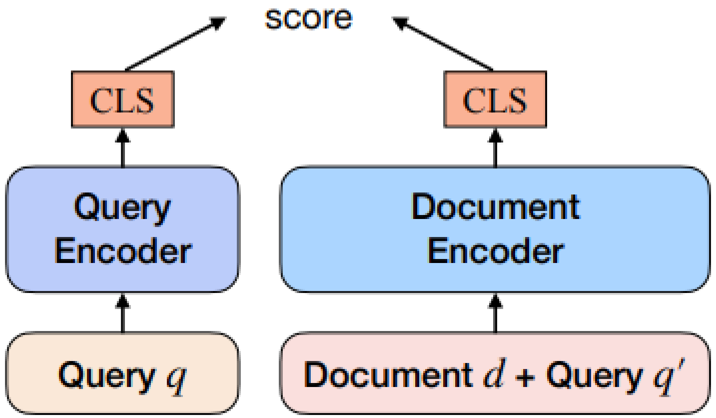
- Domain adaptation in data scarcity scenarios
- Query-informed document representations
- Multiview document representations

Rodrigo Nogueira et al. Document Expansion by Query Prediction. arXiv preprint arXiv:1904.08375, 2019.
Zehan Li et al. Learning Diverse Document Representations with Deep Query Interactions for Dense Retrieval. arXiv preprint arXiv:2208.04232, 2022.

Introduction

- **Dual-cross-encoder**
  - Document Expansion

Li Et al. (2022)



Pre Interaction

$$sim(q,d) = E_q(q)^T E_d(d+q\prime).$$

Interaction-Aware Document Representation

Discrepancy

Training

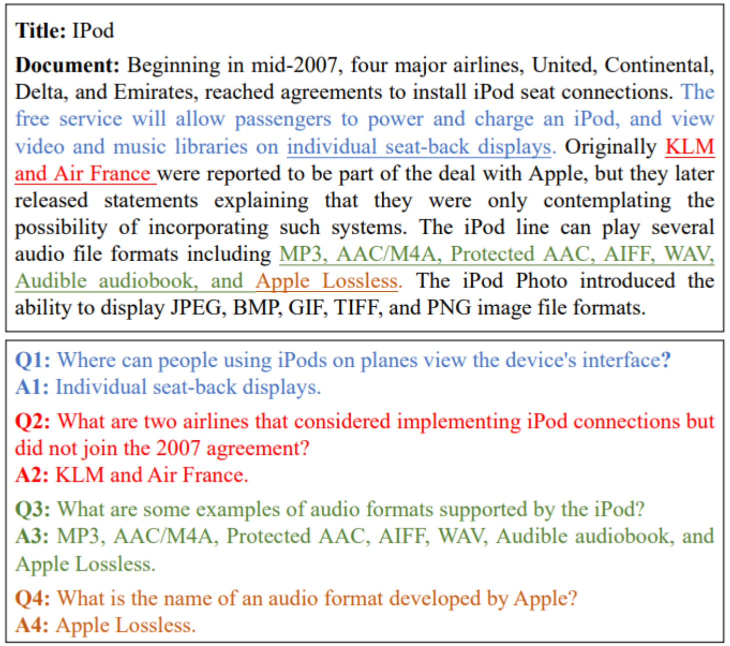
$q, \ q'$

Inference

q

Specific Task Performance ↓

Multiview document representations


Title: IPod

Document: Beginning in mid-2007, four major airlines, United, Continental, Delta, and Emirates, reached agreements to install iPod seat connections. The free service will allow passengers to power and charge an iPod, and view video and music libraries on individual seat-back displays. Originally KLM and Air France were reported to be part of the deal with Apple, but they later released statements explaining that they were only contemplating the possibility of incorporating such systems. The iPod line can play several audio file formats including MP3, AAC/M4A, Protected AAC, AIFF, WAV, Audible audiobook, and Apple Lossless. The iPod Photo introduced the ability to display JPEG, BMP, GIF, TIFF, and PNG image file formats.

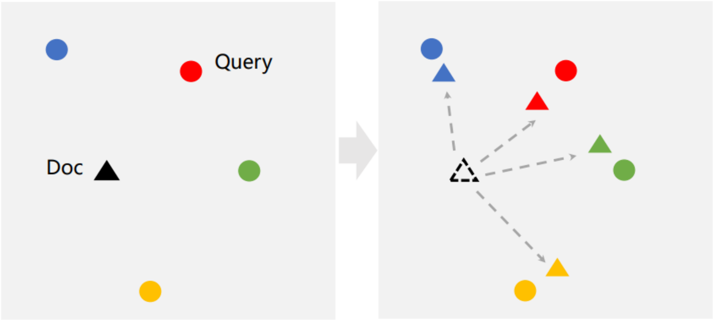
Q1: Where can people using iPods on planes view the device's interface?
A1: Individual seat-back displays.

Q2: What are two airlines that considered implementing iPod connections but did not join the 2007 agreement?
A2: KLM and Air France.

Q3: What are some examples of audio formats supported by the iPod?
A3: MP3, AAC/M4A, Protected AAC, AIFF, WAV, Audible audiobook, and Apple Lossless.

Q4: What is the name of an audio format developed by Apple?
A4: Apple Lossless.

(a) An example from SQuAD Open Dataset.



(b) Our proposed MVR method.

Similar effect to MIPS

Performance ↑

Rodrigo Nogueira et al. Document Expansion by Query Prediction. arXiv preprint arXiv:1904.08375, 2019.
Shunyu Zhang et al. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. ACL, 2022.
Zehan Li et al. Learning Diverse Document Representations with Deep Query Interactions for Dense Retrieval. arXiv:2208.04232, 2022.

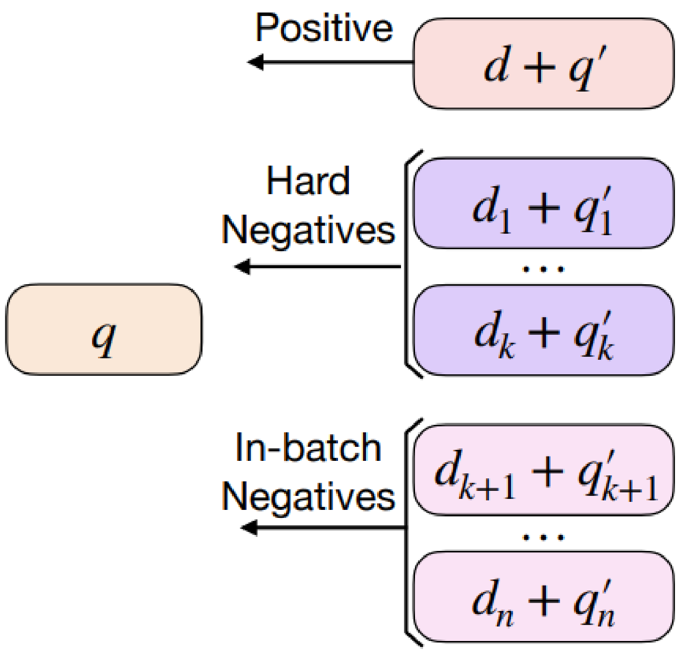
Introduction

- **CAPSTONE**
  - Training stage: Curriculum sampling
    - Bridge the gap between training and inference from document expansion
    - Query-informed document representation
  - Inference stage: Take the average pooling of different document views
    - Corpus expansion by generated query
    - Compute the typical document representation
  - Model Performance
    - Experiments on in-domain retrieval datasets and zero-shot BEIR benchmark
    - Improve the performance without sacrificing retrieval efficiency

Method

- **Discrepancy in Training and Inference**
  - Document expansion in the training phase

Dual-cross-encoder (Li Et al, 2022)



Similarity Score in minibatch

	q_1	q_2	q_3	q_4	…	q_n
d_1	P	HN	BN	BN		BN
d_2	HN	P	HN	BN		BN
d_3	HN	HN	P	HN		BN
d_4	PN	HN	BN	P		BN
…						
d_n	HN	PN	HN	BN		P

Dense Passage Retrieval (DPR)

Hard Negatives

- Don't contain the answer
- Match most question tokens

In-batch Negatives

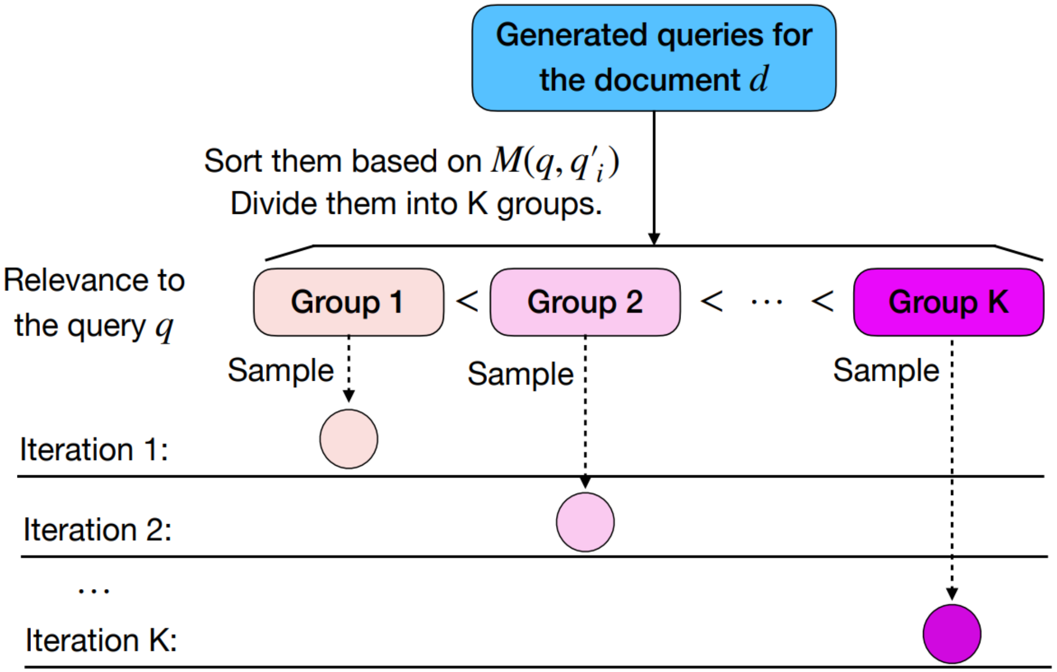
- Sort negative queries by random sampling
- Produce more pairs of training examples
- Contribute to the model performance
- Efficient training without creating new ones

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020.
Rodrigo Nogueira et al. Document Expansion by Query Prediction. arXiv preprint arXiv:1904.08375, 2019.
Zehan Li et al. Learning Diverse Document Representations with Deep Query Interactions for Dense Retrieval. arXiv preprint arXiv:2208.04232, 2022.

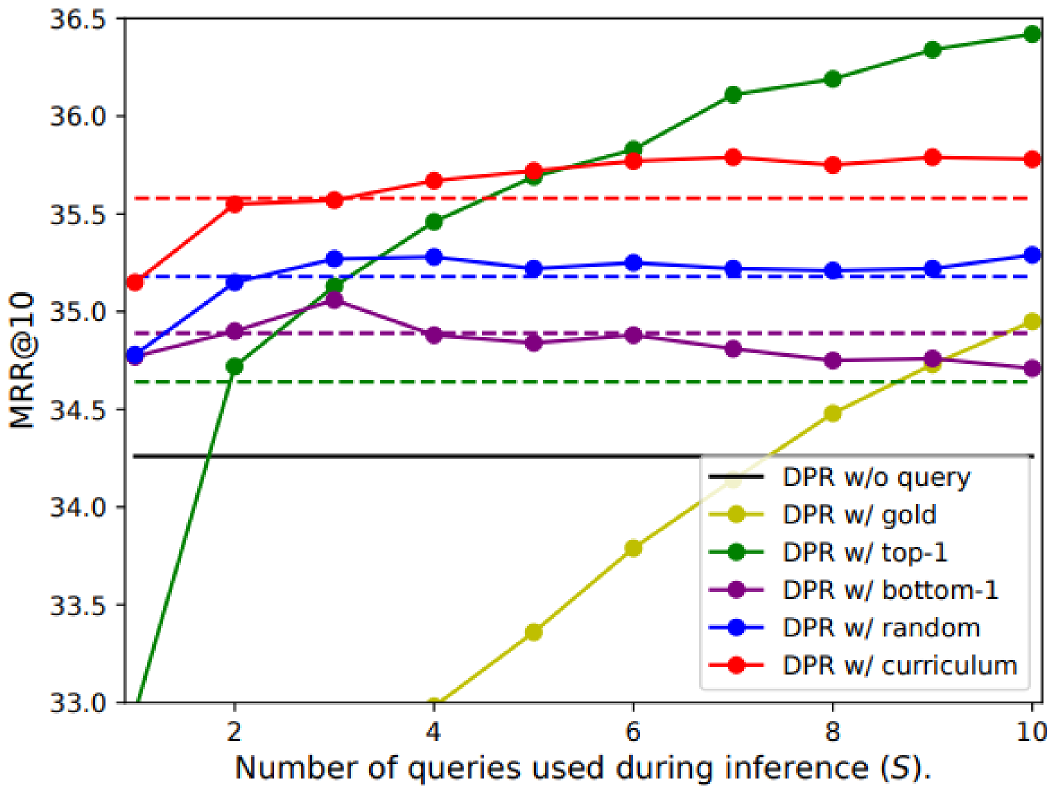
Method

- **Bridging the Gap with Curriculum Sampling**
  - Document expansion in the training phase

Query selection process



Ablation Study

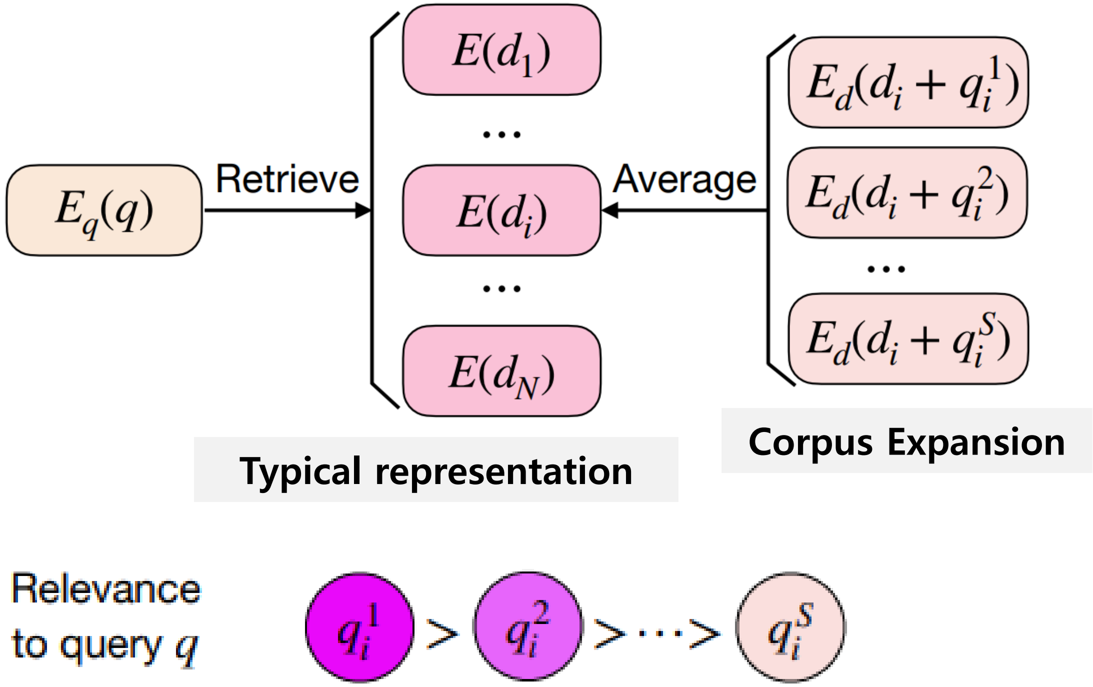


Dataset: MS-MACRO

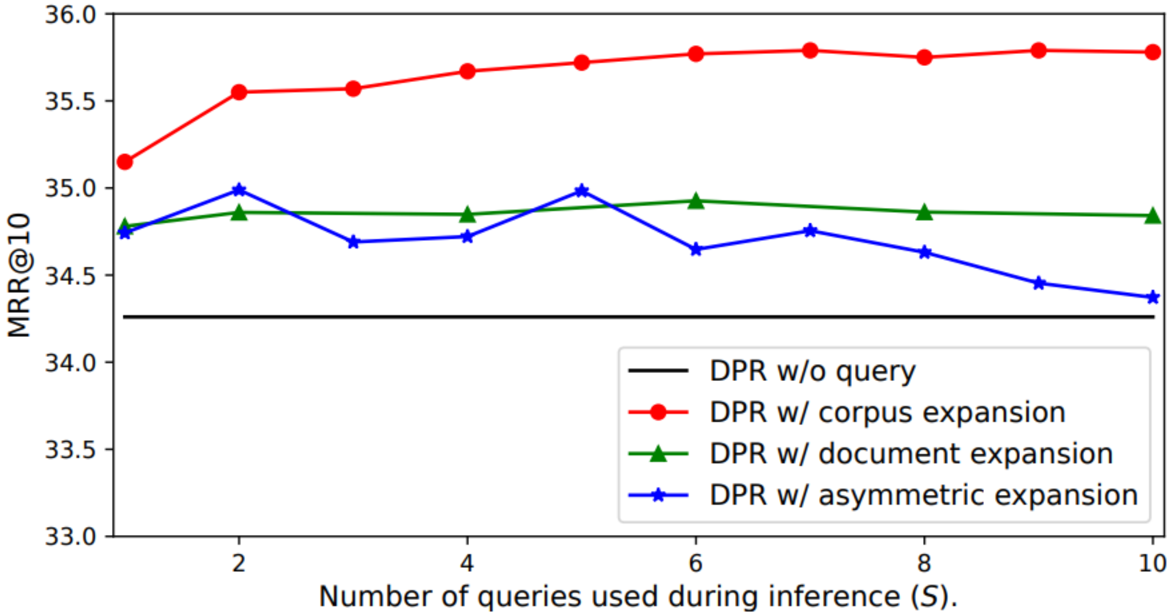
Method

- **Computing the Typical Representation of Different Views**
  - Take the average pooling of views in the inference stage

Retrieval stage



Ablation Study



Dataset: MS-MACRO

Method

- **Computing the Typical Representation of Different Views**
  - Take the average pooling of different document views

Ablation Study

Variants	MRR@10	R@1000
DPR	34.26	97.02
CAPSTONE+DPR w/ $S = 1$	35.15	97.19
CAPSTONE+DPR w/ average	**35.66**	**97.28**
CAPSTONE+DPR w/ max	35.45	97.25
CAPSTONE+DPR w/ median	35.64	97.20

Experiment

- **Comparison with baseline model**
  - In-domain Performance
    - Dataset: MS-MARCO, TREC-2019, TREC-2020
  - Zero-shot Performance
    - Dataset: BEIR benchmark

- **Ablation Study**
  - Corpus Expansion vs. Document Expansion
  - Effect of Query Selection Strategies
  - Comparison of Methods for Computing the Typical Representation
  - Multi-stage Retrieval Performance

Experiment

- **In-domain Performance**
  - Passage retrieval results on MS-Marco Dev, and TREC datasets

Models	MS-MARCO MRR@10	R@50	R@1000	TREC DL 19 nDCG@10	TREC DL 20 nDCG@10
Sparse retrieval					
BM25 (Yang et al., 2017)	18.5	58.5	85.7	51.2	47.7
DeepCT (Dai and Callan, 2019)	24.3	69.0	91.0	57.2	-
DocT5Query (Nogueira and Lin, 2019)	27.7	75.6	94.7	64.2	-
Dense retrieval					
DPR (Karpukhin et al., 2020)	31.4	-	95.3	59.0	62.1
ANCE (Xiong et al., 2021)	33.0	-	95.9	64.5	64.6
SEED (Lu et al., 2021)	33.9	-	96.1	-	-
STAR (Zhan et al., 2021)	34.7	-	-	68.3	-
TAS-B (Hofstätter et al., 2021)	34.0	-	97.5	**71.2**	69.3
RocketQA (Qu et al., 2021)	37.0	85.5	97.9	-	-
COIL (Gao et al., 2021)	35.5	-	96.3	70.4	-
ColBERT (Khattab and Zaharia, 2020)	36.0	82.9	96.8	-	-
DCE (Li et al., 2022)	36.0	-	96.4	68.3	68.9
RetroMAE (Xiao et al., 2022)	35.0	-	97.6	-	-
Condenser (Gao and Callan, 2021)	36.6	-	97.4	69.8	-
coCondenser (Gao and Callan, 2022)*	37.9	86.3	98.4	70.7	69.8
CAPSTONE	**38.6**	**86.6**	**98.6**	71.1	**70.3**

Experiment

• Zero-shot Performance

◦ Performances on BEIR benchmark (measured with nDCG@10)

Task
Bio-Medical IR
Question Answering
Tweet Retrieval
News Retrieval
Augment Retrieval
Duplicate Question Retrieval
Entity Retrieval
Citation Prediction
Fact Checking

Dataset	BERT	LaPraDoR	SimCSE	DiffCSE	SEED	Condenser	coCondenser*	CAPSTONE
TREC-COVID	64.9	49.5	52.4	49.2	61.2	75.4	74.0	**77.9**
BioASQ	26.2	23.9	26.4	25.8	29.7	31.7	34.1	**34.3**
NFCorpus	25.7	28.3	25.0	25.9	25.6	27.8	32.4	**33.0**
NQ	43.8	41.5	41.2	41.2	42.5	45.9	50.5	**50.5**
HotpotQA	47.8	48.8	50.2	49.9	52.8	53.7	56.4	**56.7**
FiQA-2018	23.7	26.6	24.0	22.9	24.4	26.1	30.0	**30.4**
Signal-1M (RT)	21.6	24.5	**26.4**	26.0	24.6	25.8	24.7	23.1
TREC-NEWS	36.2	20.6	36.8	36.3	33.5	35.3	39.1	**40.3**
Robust04	36.4	31.0	35.3	34.3	34.8	35.2	40.3	**40.7**
ArguAna	35.7	**50.3**	43.6	46.8	34.7	37.5	40.9	39.2
Touche-2020	27.0	17.8	17.8	16.8	18.0	22.3	27.0	**31.0**
CQADupStack	28.4	**32.6**	29.5	30.5	28.5	31.6	30.0	30.0
Quora	78.2	84.3	84.8	85.0	84.9	**85.5**	84.3	83.8
DBPedia	29.8	32.8	30.4	30.3	32.4	33.1	37.2	**38.0**
SCIDOCS	11.5	**14.5**	12.5	12.5	11.7	13.6	14.3	14.3
FEVER	68.4	51.8	65.1	64.1	65.3	68.2	72.4	**72.7**
Climate-FEVER	20.5	17.2	**22.2**	20.0	17.6	19.9	19.4	19.3
SciFact	50.4	48.3	54.5	52.3	55.6	57.0	58.3	**60.5**
Avg. Performance	37.6	35.8	37.7	37.2	37.7	40.3	42.7	**43.3**

Experiment

- **Ablation Study**
  - Multi-stage Retrieval Performance (at two training stages)
    - Initialize the retriever at each stage
    - Evaluate the last model training checkpoint on the retrieval datasets

	First Stage		Second Stage	
Models	**BM25 Negatives**		**Mined Negatives**	
	MRR@10	**R@1000**	**MRR@10**	**R@1000**
DPR	34.26	97.02	36.44	97.65
CAPSTONE+DPR	35.66	97.28	37.28	97.82
coCondenser	35.91	98.21	37.94	98.41
CAPSTONE+coCondenser	**36.75**	98.21	**38.65**	98.60

Luyu Gao et al. Unsupervised corpus aware language model pre-training for dense passage retrieval. ACL, 2022.

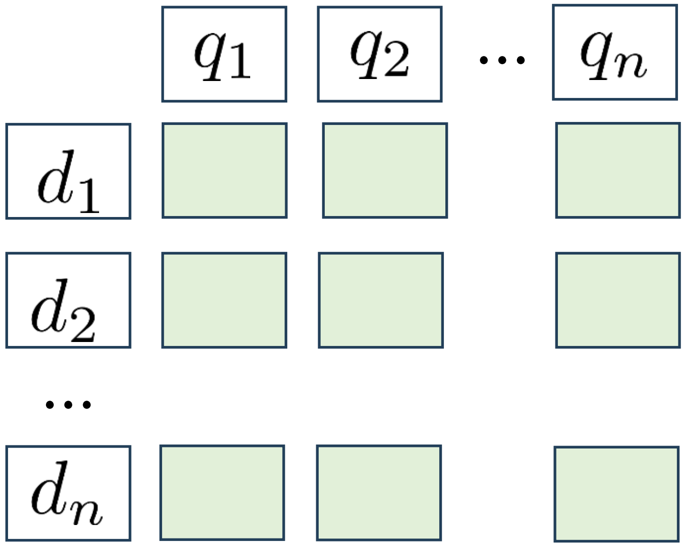
Conclusion

- **CAPSTONE**
  - Curriculum sampling for dense retrieval with document expansion
    - Learn much better query-related document representations
  - Typical representation of different document views
    - Balance between inference efficiency and effectiveness

- **Limitation**
  - Generating synthetic queries for each document is time-consuming work
  - Only verify our method on vanilla DPR and coCondenser
  - We plan to verify our method on other dense retrieval models

Appendix

- **Similarity Score for Dense Retrieval**



Appendix

- **Dense Passage Retrieval (DPR)**
  - Similarity Score in minibatch

	q_1	q_2	q_3	q_4	…	q_n
d_1	P	HN	BN	BN		BN
d_2	HN	P	HN	BN		BN
d_3	HN	HN	P	HN		BN
d_4	PN	HN	BN	P		BN
…						
d_n	HN	PN	HN	BN		P

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020