NLU & IR: NEURAL IR (III)

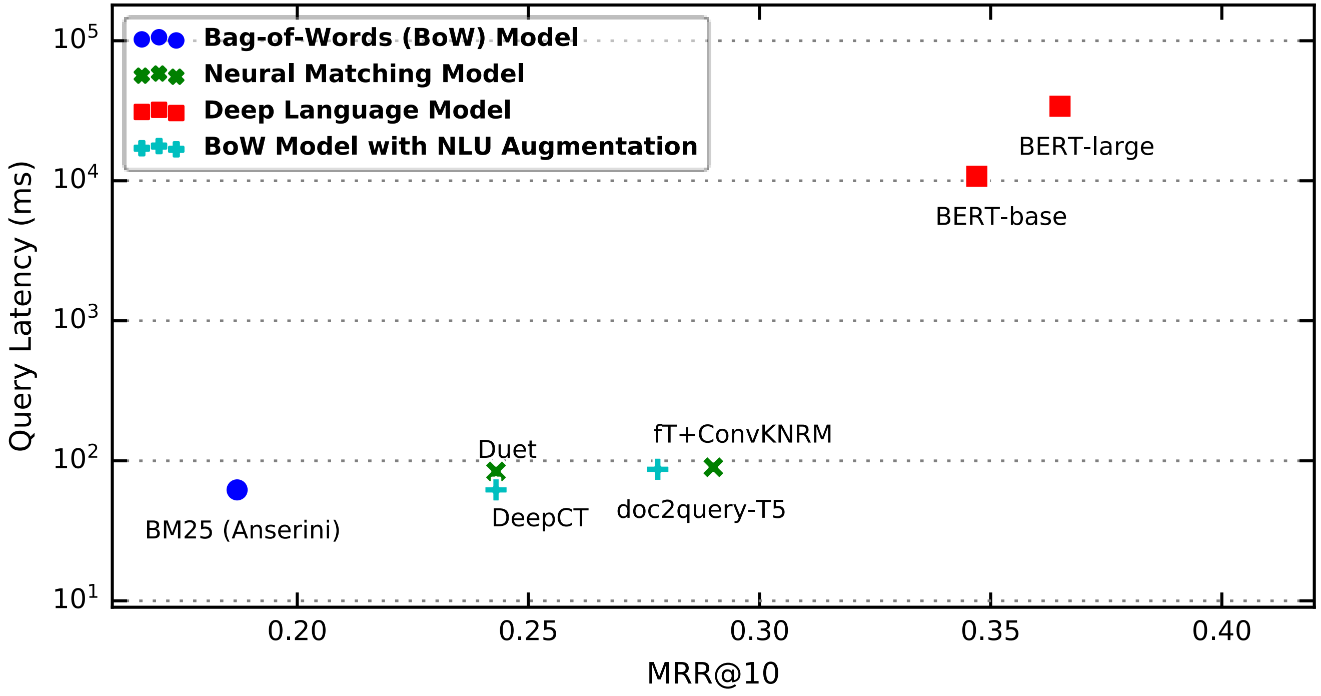
Omar Khattab

CS224U: Natural Language Understanding

Spring 2021

Learning term weights: DeepCT and doc2query

- We get to learn the term weights with BERT and to **re-use** them!
- But our query is back to being a “bag of words”.



Can we do better?

Neural IR Paradigms: **Representation Similarity**

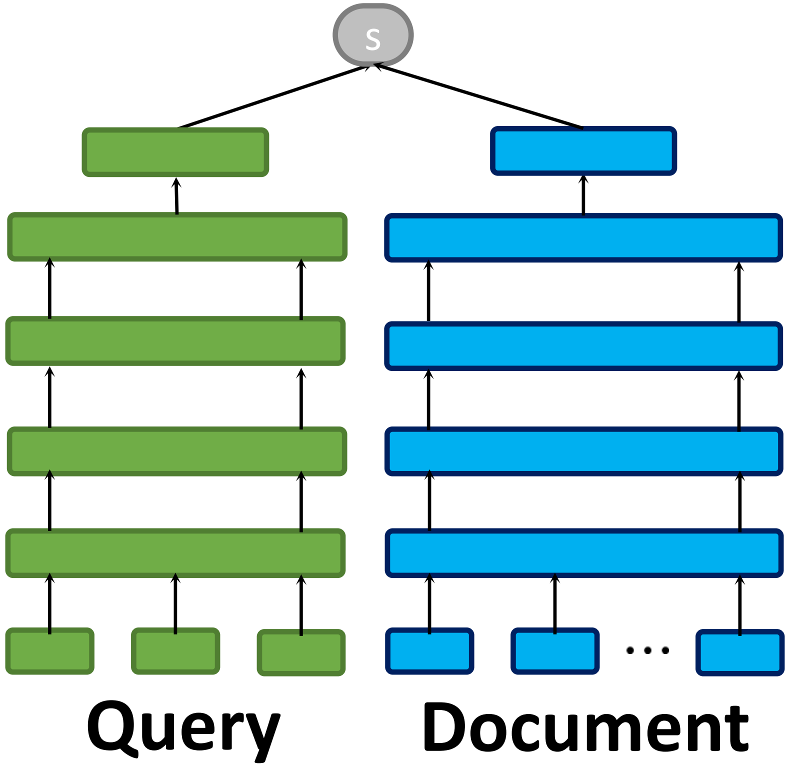
- Tokenize the query and the document
- **Independently** encode the query and the document
- … into a **single-vector** representation each
- Estimate relevance a dot product
  - Or a cosine similarity

Like learning term weights, this paradigm offers strong **efficiency** advantages:

✓ Document representations can be pre-computed!

✓ Query computations can be amortized.

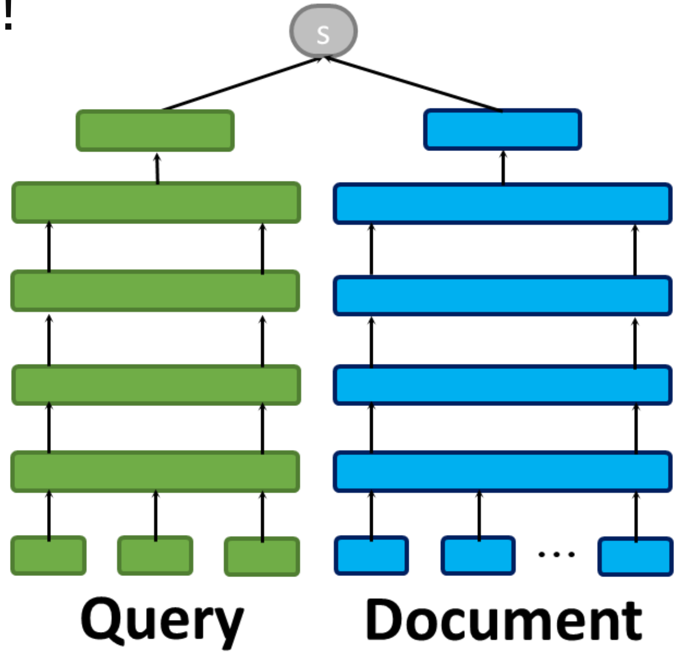
✓ Similarity computations are very cheap.



Representation Similarity: Models

- Many pre-BERT IR models fall under this paradigm!
  - DSSM and SNRM

- Numerous BERT-based models exist
  - SBERT, ORQA, **DPR**, DE-BERT, RepBERT, ANCE



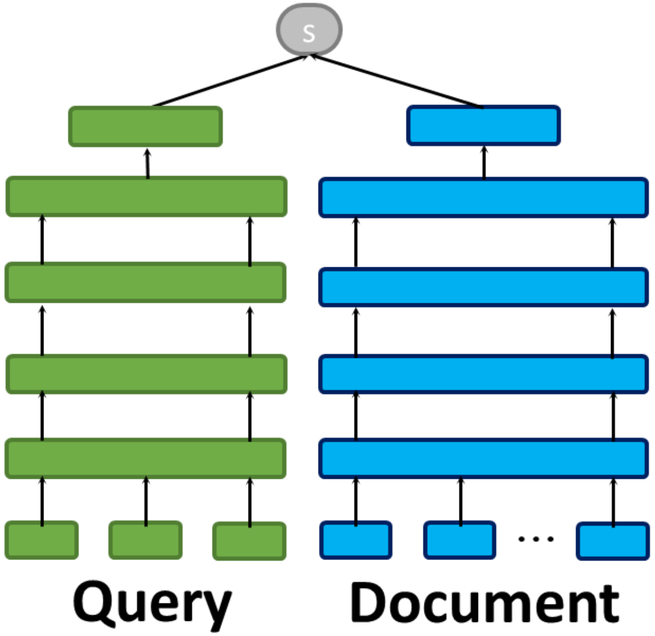
Many of these BERT-based representation similarity models are *concurrent* to one another (late 2019 / early 2020).

The largest differences are in the **specific tasks** each targets and the **supervision approach**.

Representation Similarity: DPR

Dense Passage Retriever (DPR) by *Karpukhin et al.*

- Encodes each passage into a 768-dimensional vector
- Encodes each query into a 768-dimensional vector
- Trained with N-way cross-entropy loss, over the similarity scores between the query and:
  - A positive passage
  - A negative passage, sampled from BM25 top-100
  - Many in-batch negative passages
    - the positive passages for the *other* queries in the same training batch



Vladimir Karpukhin, et al. "Dense passage retrieval for open-domain question answering." EMNLP'20

Representation Similarity: DPR

Dense Passage Retriever (DPR) by *Karpukhin et al.*

- Encodes each passage into a 768-dimensional vector
- Encodes each query into a 768-dimensional vector

Xiong et al. (2020) test a DPR-style retriever on MS MARCO: **31% MRR**. They show that a sophisticated supervision scheme can achieve **33%**.

Both constitute progress over “learned term weights” like DeepCT, but they are still considerably lower than standard BERT’s **>36% MRR**.

  - Many in-batch negative passages
    - the positive passages for the *other* queries in the same training batch

Vladimir Karpukhin, et al. "Dense passage retrieval for open-domain question answering." EMNLP’20
Lee Xiong, et al. "Approximate nearest neighbor negative contrastive learning for dense text retrieval." ICLR’21

Representation Similarity: Downsides

✕ **Single-Vector Representations**

- They “cram” queries and documents into a **coarse-grained** representation!

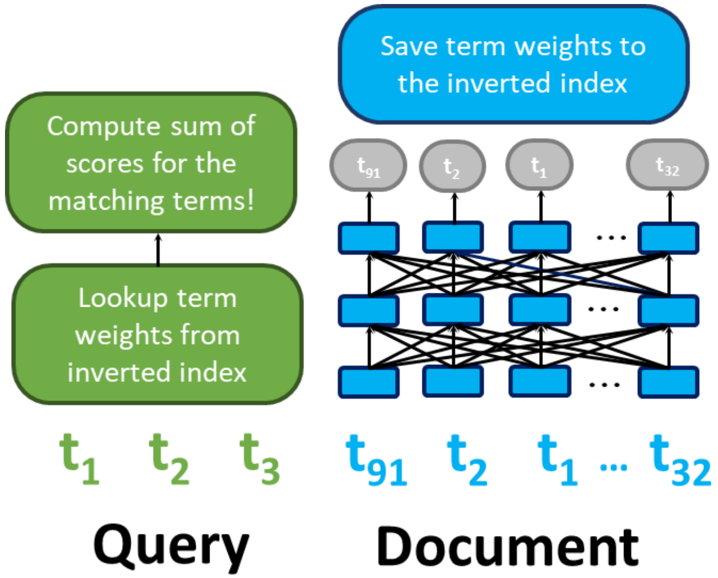
✕ **No Fine-Grained Interactions**

- They estimate relevance as **single dot product**!
- We lose **term-level interactions**, which we had in:
  - Query–Document interaction models (e.g., BERT or Duet)
  - And even term-weighting models (e.g., DeepCT and BM25)

Can we keep precomputation and still have fine-grained interactions?

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR’20.

Summary: Neural Ranking Paradigms



Compute sum of scores for the matching terms!

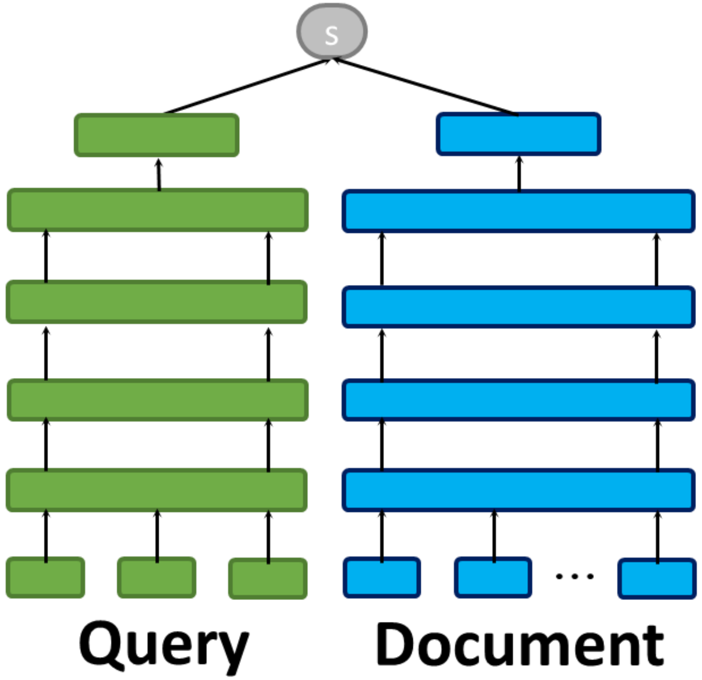
Lookup term weights from inverted index

Save term weights to the inverted index

(a) Learned Term Weights

✓ Independent Encoding

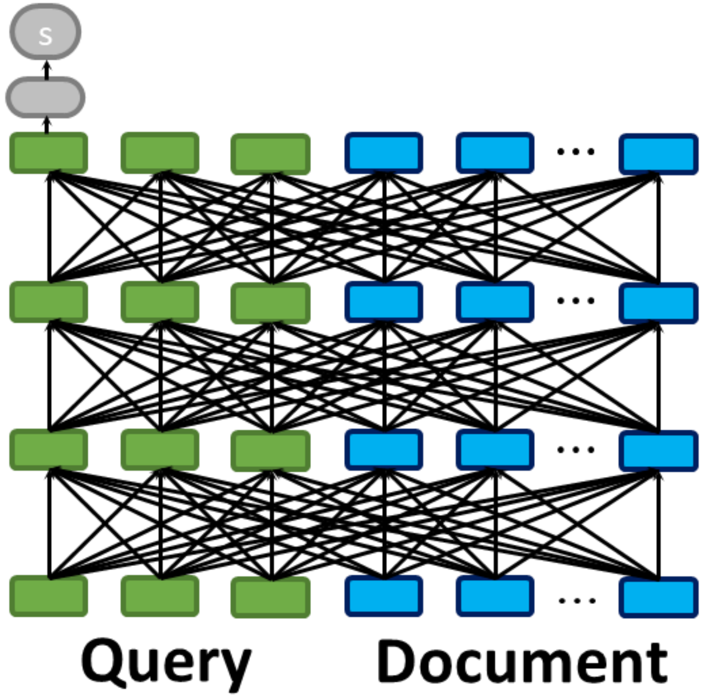
✗ Bag-of-Words Matching



(b) Representation Similarity

✓ Independent, Dense Encoding

✗ Coarse-Grained Representation



(c) Query–Document Interaction

✓ Fine-Grained Interactions

✗ Expensive <u>Joint</u> Conditioning

Beyond Re-ranking: End-to-end Retrieval

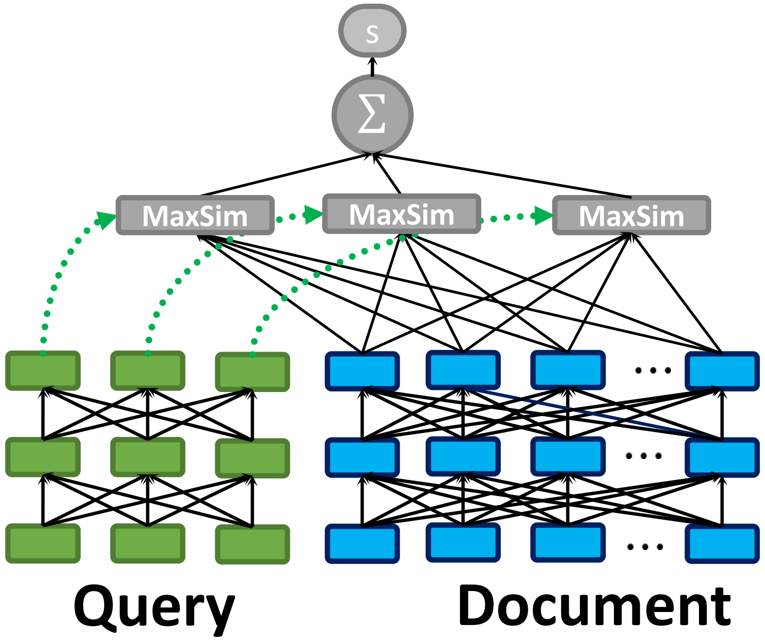
- **Query–Document Interaction** models forced us to use a re-ranking pipeline, where we just re-scored the top-1000 documents retrieved by BM25.

End-to-end retrieval is essential toward improving RECALL.

- **Learning Term Weights** and **Representation Similarity** models alleviate this!
  - They allow us to do end-to-end retrieval: quickly searching over all documents directly.
  - We can save **term weights** in the **inverted index**. This means that we do NOT need a re-ranking pipeline.
  - We can also index **vector representations** for **fast vector-similarity search**, which allows **PRUNING** to find the top-K matches without exhaustive enumeration.
    - Libraries like **FAISS** abstract away the details.

https://github.com/facebookresearch/faiss/

Neural IR Paradigms: **Late Interaction**



(d) Late Interaction

(i.e., ColBERT)

Can we keep precomputation and still have fine-grained interactions?

Desired Properties:

- ✓ Independent Encoding
- ✓ Fine-Grained Representations
- ✓ End-to-End Retrieval (pruning!)

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR’20.

Late Interaction: Real Example of Matching

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] **on** August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The **Transformers** […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

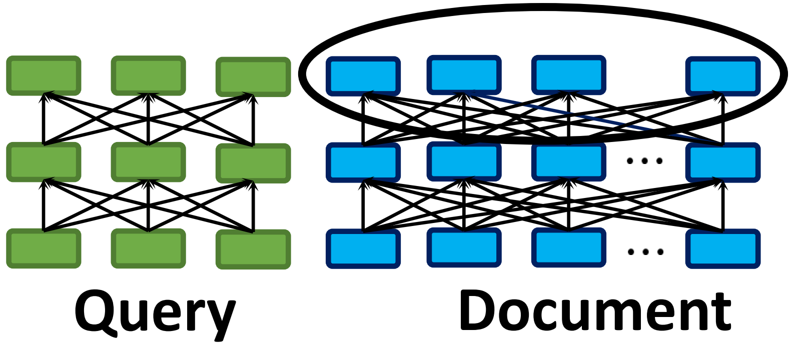
[…] the **animated** […] The Transformers […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was **released** […] on August 8, 1986

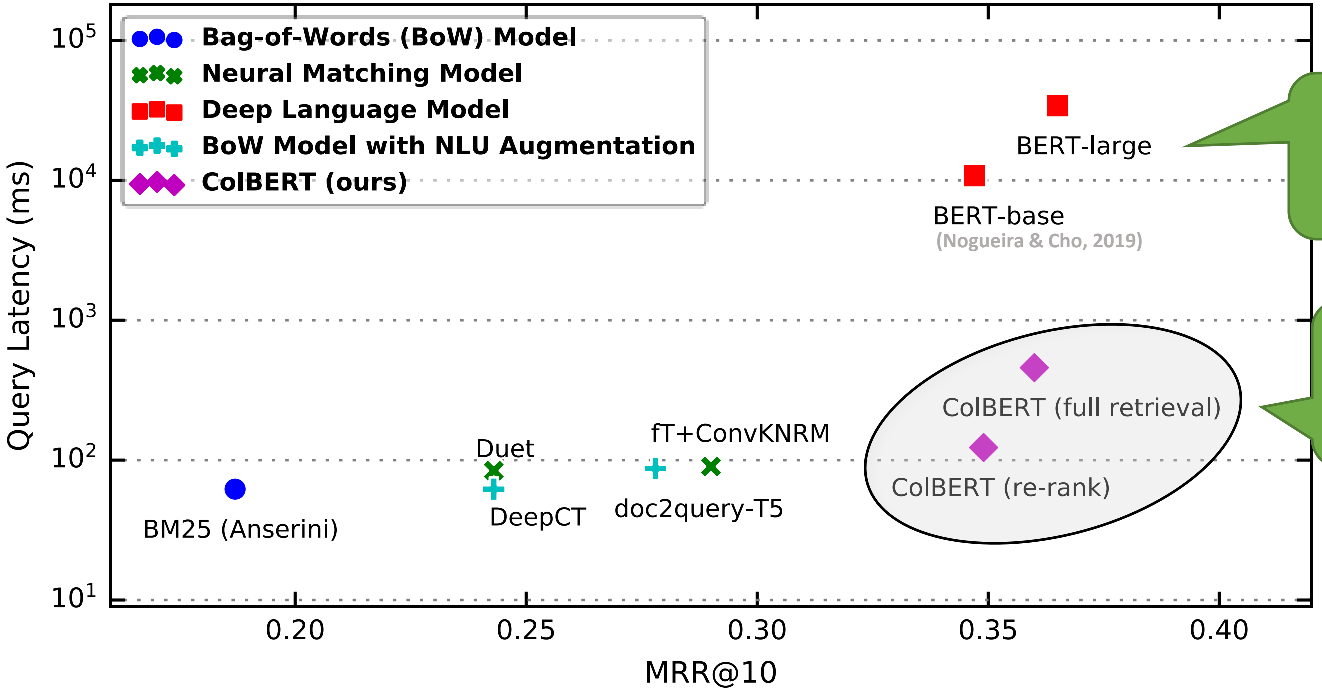
Late Interaction: ColBERT

Notice that ColBERT represents the document as a MATRIX, not a vector.



(d) Late Interaction
(i.e., ColBERT)

Late Interaction: ColBERT



Recall@1000 = 81%

Recall@1000 = 97%
Recall@50 = 83%

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR’20.

Robustness: Out-of-Domain Quality

- So far, we’ve looked at <u>in-domain</u> effectiveness evaluations.
  - We had training and evaluation data for MS MARCO.

- We often want to use retrieval in new, out-of-domain settings.
  - … with NO training data and NO validation data.
  - This is sometimes called a “zero-shot” setting; it emphasizes transfer.

- BEIR is a recent benchmark for IR models in “zero-shot” scenarios

Thakur, Nandan, et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.“ *arXiv:2104.08663* (2021)

Robustness: Out-of-Domain **NDCG@10**

- **Fine-grained interaction** is key to robustly high precision

	Classical IR	**Interaction Models**	**Representation Similarity**	**Representation Similarity**	**Late Interaction**
IR Task	**BM25**	**ELECTRA re-ranker**	**DPR**	**SBERT**	**ColBERT**
BioMed	48	**49**	22	34	**49**
QA	38	**51**	33	41	48
Tweet	**39**	31	16	26	27
News	37	**43**	16	37	39
Arguments	**52**	35	15	34	25
Duplicates	53	56	20	58	**60**
Entity	29	38	26	34	**39**
Citation	**16**	15	8	13	15
Fact-Check	48	52	34	47	**54**
Overall Avg	42	**45**	23	39	44

Table aggregated from the BEIR results (Table 2) by Thakur, Nandan, et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.“ *arXiv:2104.08663* (2021)

Robustness: Out-of-Domain **Recall@100**

- **Scalable** fine-grained interaction is key to robustly high recall

	Classical IR	**Interaction Models**	**Representation Similarity**	**Representation Similarity**	**Late Interaction**
IR Task	**BM25**	**ELECTRA re-ranker**	**DPR**	**SBERT**	**ColBERT**
BioMed	**45**	**45**	23	35	**45**
QA	67	67	60	68	**75**
Tweet	**38**	**38**	16	26	28
News	**40**	**40**	22	37	37
Arguments	**70**	**70**	46	62	61
Duplicates	77	77	44	79	**81**
Entity	38	38	35	40	**46**
Citation	**35**	**35**	22	30	34
Fact-Check	71	71	65	74	**75**
Overall Avg	59	59	43	57	**61**

Table aggregated from the BEIR results (Table 4) by Thakur, Nandan, et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.“ *arXiv:2104.08663* (2021)

Final Thoughts on Neural IR

- Speed vs. **Scalability**: not always the same!
  - Inductive biases are crucial to **effective** models that **scale**.

- Next…
  - **Can scalability drive new gains in quality?**
    - YES! We will see examples of this in the Open-QA screencast.

  - **How can we tune a neural IR model for open-domain NLU tasks?**

References

Vladimir Karpukhin, et al. "Dense passage retrieval for open-domain question answering." EMNLP’20

Lee Xiong, et al. "Approximate nearest neighbor negative contrastive learning for dense text retrieval." ICLR’21

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT.“ SIGIR’20

Nandan, et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.“ arXiv:2104.08663 (2021)