UC SANTA BARBARA

Code Link: https://github.com/heshanxiu/sbert.git **Department:** Computer Science **Affiliation:** CMPSC 291K Project

Zero-Shot Dense Retrieval with Contrastive Dual Learning

Shanxiu He

shanxiuhe@ucsb.edu

Motivation: Improve Dense Retrieval

Learn P(d|q) `Query` ⇔ `Document⁺, Document⁻, …, Document⁻`

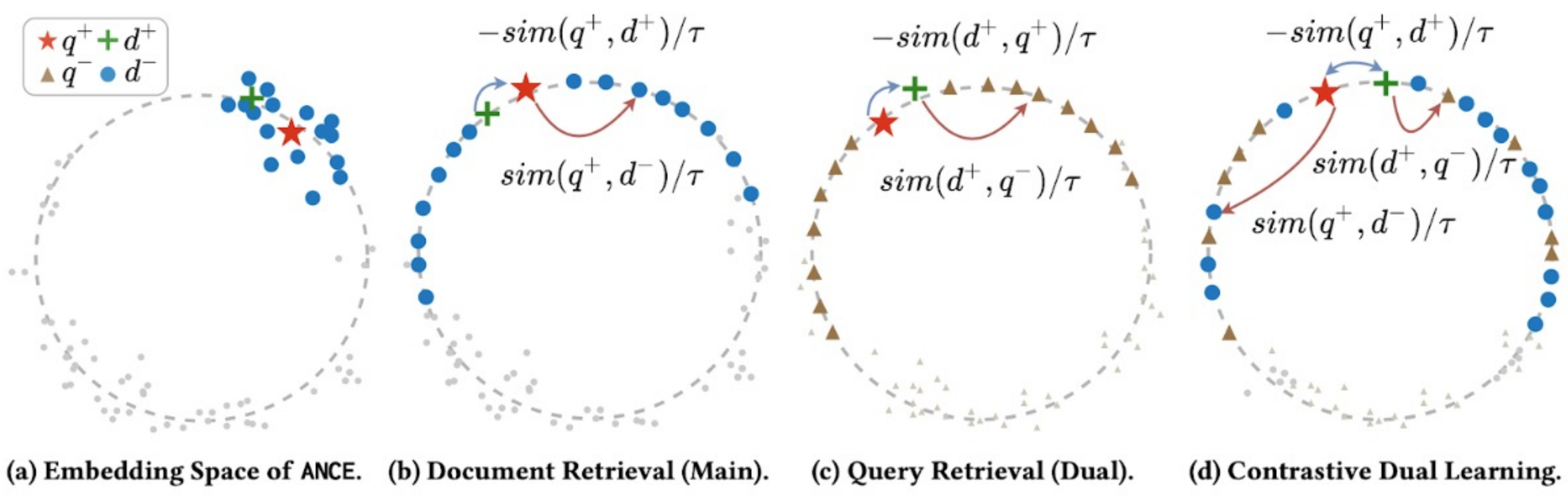
Only aligning query to relevant documents,

What about query embedding space?

Learn P(q|d) `Document` ⇔ `Query⁺, Query⁻, …, Query⁻`

- Both **query retrieval** and **document retrieval**
- Learn better query embedding space

Goal: Better Embedding Space

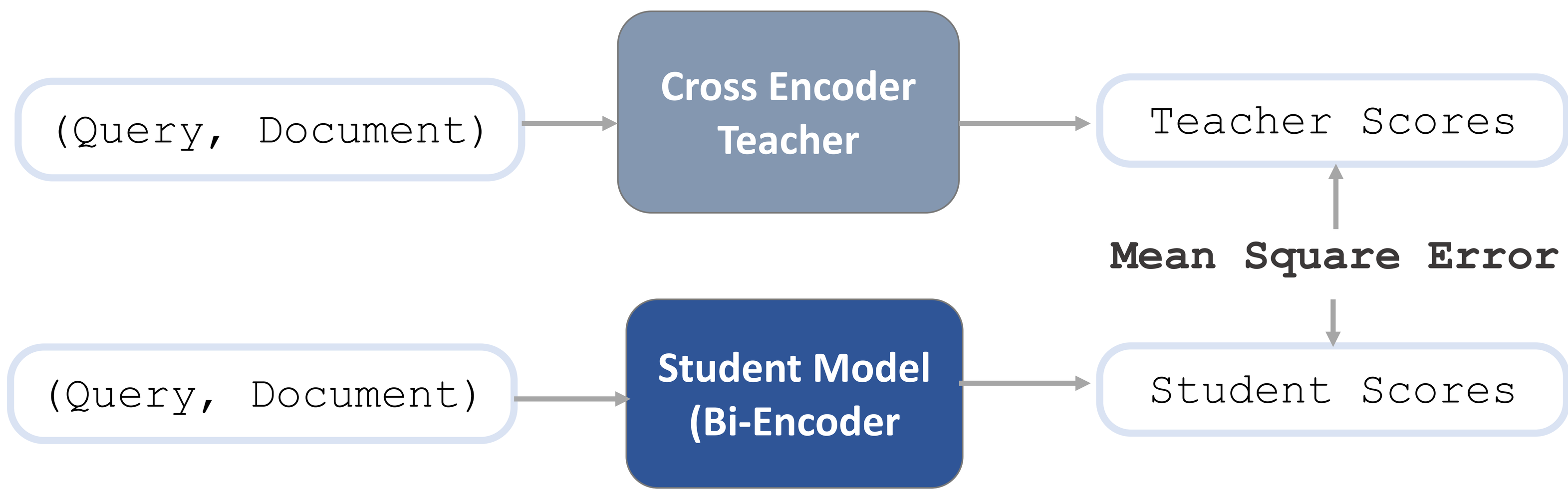


[1*] Diagram from Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE)

Representation that **pushes document away from negative queries**

Methods & Directions

Part 1. Cross-Encoder Knowledge Distillation



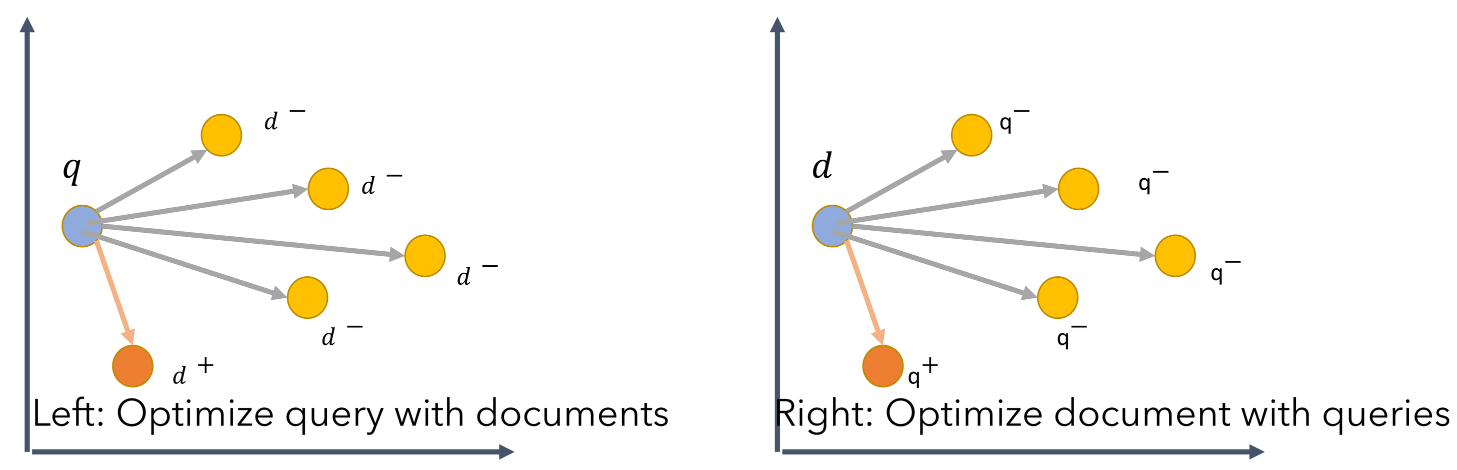
Directly learn how "teacher" positioned the embeddings

Margin-MSE-Dual, A Weighted Sum of

$$l(q, d^+, d^-) = \mathrm{MSE}(M_s(q, d^+) - M_s(q, d^-), M_t(q, d^+) - M_t(q, d^-))$$
$$l_d(d, q^+, q^-) = \mathrm{MSE}(M_s(d, q^+) - M_s(d, q^-), M_t(d, q^+) - M_t(d, q^-))$$

Note: Teacher never learns how to position document with negative queries; might hurt performance compared to pure MMSE.

Part 2. Multiple Negative Ranking Loss



MNRL-Dual, Weighted Sum of

$$l_d(d, q^+, Q^-) = -\log \frac{e^{f(d^+,q^+)}}{e^{f(d^+,q^+)} + \sum_{q^- \in Q^-} e^{f(d^+,q^-)}}$$
$$l(q, d^+, D^-) = -\log \frac{e^{f(q^+,d^+)}}{e^{f(q^+,d^+)} + \sum_{d^- \in D^-} e^{f(q^+,d^-)}}$$

Anticipate: Improve Zero-Shot performance

Experiments and Results

Training Dataset: MS MARCO Teacher: cross-encoder/ms-marco-MiniLM-L-6-v2 Model: Part 1. MiniLM-L12 Part 2. DistillBERT

In-domain Evaluation on MS MARCO

Method	MRR@10	NDCG@10	MAP@100	Recall@10
MMSE	0.3620	0.4268	0.3676	0.6437
$MMSE_D$	0.3581	0.4244	0.3636	0.6457

Table 1: MS MARCO Performance on MiniLM-L12 with Margin-MSE loss.

Method	MRR@10	NDCG@10	MAP@100	Recall@10
MNRL	0.3318	0.3894	0.3369	0.5846
$MNRL_D$	0.3309	0.3884	0.3355	0.5844

Table 2: MS MARCO Performance on DistillBERT with Multiple Negatives Ranking loss.

Zero-Shot Evaluation Observation

Part 1. Directly imitate teacher on query retrieval is not preferrable.

Part 2. Learning query retrieval (from scratch) contrastively improves on zero-shot setting

Zero-Shot Evaluation on BEIR

Corpus	Baselines		MiniLM12		DistillBERT	
	DPR	BM25	MMSE	$MMSE_D$	MNRL	$MNRL_D$
DBPedia	0.236	0.313	**0.367**	0.365	0.304	**0.309**
FiQA-2018	0.275	0.236	**0.307**	0.304	0.238	**0.245**
NQ	0.398	0.329	0.452	**0.471**	**0.448**	0.446
NFCorpus	0.208	0.325	0.308	**0.308**	**0.269**	0.267
TREC-COVID	0.561	0.656	**0.476**	0.454	0.443	**0.479**
Torche-2020	0.243	0.367	0.174	**0.185**	0.194	**0.194**
ArguAna	0.414	0.315	**0.453**	0.451	**0.404**	0.402
Climate-FEVER	0.176	0.213	**0.239**	0.211	**0.185**	0.184
Quora	0.842	0.789	**0.869**	0.867	0.839	**0.839**
SCIDOCS	0.108	0.158	**0.180**	0.176	0.124	**0.126**
SciFact	0.478	0.665	**0.627**	0.598	0.525	**0.527**
Avg	0.359	0.397	**0.405**	0.399	0.361	**0.365**

Table 3: NDCG@10 results on BEIR (HotpotQA and FEVER excluded)

Differences from Previous Dual Learning Training

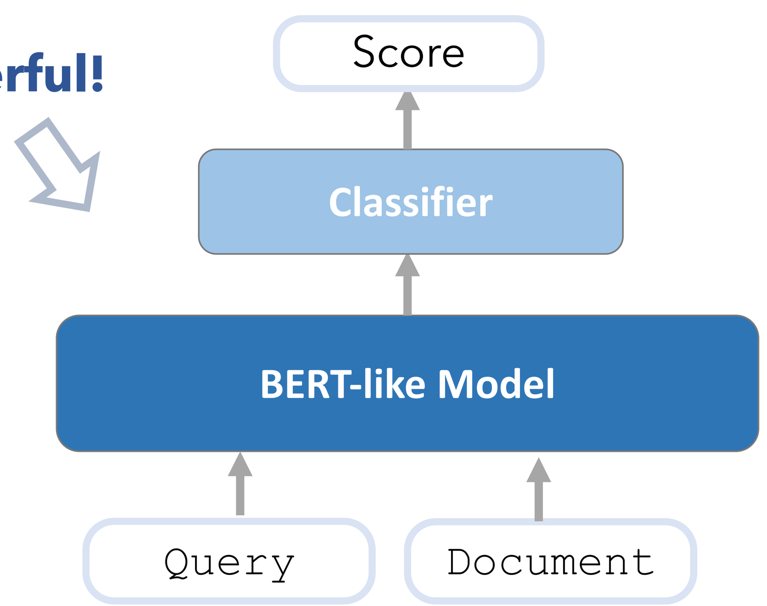
- Exact Match Instead of Approximate Nearest Search
- Better Hard Negatives
- Efficient (No Need to Rebuild Indexes) and Synchronously Update

Observation

1) Distillation provides better baselines
2) For in-domain, dual training produces similar results

Next Step: Distillation + Multiple Negative Loss

Powerful!



Cross Encoder Model Architecture

1) MMSE overperforms MNRL on baselines
2) $MNRL_D$ improves zero-shot performance

Leverage released teacher:

Objective: MMSE + $MNRL_D$

Results Pending

Contribution

- First to attempt dual training on zero-shot domain and demonstrates its efficacy
- Lighter computation needs than previous attempt
- Deploy better hard negatives

Conclusion & Future Work

- Dual training might benefit zero-shot performance of dense retrieval model
- Employ techniques on better baselines and even improve training of cross encoder teachers