Contrastive Ranking-Aware Learning

CoRAL – Decoupled Representations for Retrieval

Lukas Gienapp Niklas Deckers Martin Potthast

Leipzig University & ScaDS.AI

April 8, 2023

Background

Overview on Approaches

Generic Retrieval Model

"Given a query, *(Representations)* induce a *(Relevance Estimation)*, which orders *(Identifiers)* that map to *(Results)*. "

Background

Overview on Approaches

BM25

"Given a query, *sparse representations* induce a *lexical matching*, which orders *doc IDs* that map to *documents*. "

	Traditional
Example	BM25
Representations	Sparse Repr.
Relevance Estim.	Lexical Match
Identifiers	Doc-IDs
Results	Documents

Background

Overview on Approaches

Bi-Encoder

"Given a query, *dense representations* induce an *inner product space*, which orders *doc IDs* that map to *documents*. "

	Traditional	Neural
		Representation L.
Example	BM25	Bi-Encoder [5]
Representations	Sparse Repr.	Dense Repr.
Relevance Estim.	Lexical Match	Inner Product Sp.
Identifiers	Doc-IDs	Doc-IDs
Results	Documents	Documents

Background

Overview on Approaches

Cross-Encoder

"Given a query, *directly* order *doc IDs* that map to *documents*. "

	Traditional	Neural	
		Representation L.	Metric Learning
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]
Representations	Sparse Repr.	Dense Repr.	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs
Results	Documents	Documents	Documents

Background

Overview on Approaches

Differentiable Index

"Given a query, *generate doc IDs* that map to *documents*. "

	Traditional	Neural		
		Representation L.	Metric Learning	"Index Learning"
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]	Diff. Index [6]
Representations	Sparse Repr.	Dense Repr.	–	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct	–
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs	Gen. Doc-IDs
Results	Documents	Documents	Documents	Documents

Background

Overview on Approaches

Infinite Index

“Given a query, *generate documents*. ”

	Traditional	Neural			
		Representation L.	Metric Learning	“Index Learning”	“Generative L.”
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]	Diff. Index [6]	Infinite Index [1]
Representations	Sparse Repr.	Dense Repr.	–	–	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct	–	–
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs	Gen. Doc-IDs	–
Results	Documents	Documents	Documents	Documents	Gen. Docs.

Background

Overview on Approaches

Generic Retrieval Model

“Given a query, *(Representations)* induce a *(Relevance Estimation)*, which orders *(Identifiers)* that map to *(Results)*. ”

	Traditional	Neural			
		Representation L.	Metric Learning	“Index Learning”	“Generative L.”
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]	Diff. Index [6]	Infinite Index [1]
Representations	Sparse Repr.	Dense Repr.	–	–	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct	–	–
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs	Gen. Doc-IDs	–
Results	Documents	Documents	Documents	Documents	Gen. Docs.

Efficiency

Background

Overview on Approaches

Generic Retrieval Model

“Given a query, *(Representations)* induce a *(Relevance Estimation)*, which orders *(Identifiers)* that map to *(Results)*. ”

	Traditional	**Neural**			
		Representation L.	Metric Learning	“Index Learning”	“Generative L.”
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]	Diff. Index [6]	Infinite Index [1]
Representations	Sparse Repr.	Dense Repr.	–	–	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct	–	–
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs	Gen. Doc-IDs	–
Results	Documents	Documents	Documents	Documents	Gen. Docs.

Efficiency

Effectiveness

Background

Overview on Approaches

Generic Retrieval Model

"Given a query, *(Representations)* induce a *(Relevance Estimation)*, which orders *(Identifiers)* that map to *(Results)*. "

	Traditional	**Neural**			
		Representation L.	Metric Learning	"Index Learning"	"Generative L."
Example	BM25	Bi-Encoder [5]	Cross-Encoder [5]	Diff. Index [6]	Infinite Index [1]
Representations	Sparse Repr.	Dense Repr.	–	–	–
Relevance Estim.	Lexical Match	Inner Product Sp.	Direct	–	–
Identifiers	Doc-IDs	Doc-IDs	Doc-IDs	Gen. Doc-IDs	–
Results	Documents	Documents	Documents	Documents	Gen. Docs.

Efficiency

Effectiveness

Bi-Encoders offer a good tradeoff between efficiency and effectiveness.

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

2. Domain discrepancy
   - Queries: short, simple; representation computed live
   - Documents: long, complex; representations can be cached

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

2. Domain discrepancy
   - Queries: short, simple; representation computed live
   - Documents: long, complex; representations can be cached

3. Scale discrepancy
   - Usually multiple (q, d) form a batch, with docs from other queries being implicit negatives; but: one query may has multiple relevant docs in reality
   - This training setup is mostly due to sparsity of ground truth labels

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

   Contribution: contrastive ranking-aware loss

2. Domain discrepancy
   - Queries: short, simple; representation computed live
   - Documents: long, complex; representations can be cached

3. Scale discrepancy
   - Usually multiple (q, d) form a batch, with docs from other queries being implicit negatives; but: one query may has multiple relevant docs in reality
   - This training setup is mostly due to sparsity of ground truth labels

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

   Contribution: contrastive ranking-aware loss

2. Domain discrepancy
   - Queries: short, simple; representation computed live
   - Documents: long, complex; representations can be cached

   Contribution: decoupled encoders with compatible latent spaces

3. Scale discrepancy
   - Usually multiple (q, d) form a batch, with docs from other queries being implicit negatives; but: one query may has multiple relevant docs in reality
   - This training setup is mostly due to sparsity of ground truth labels

Motivation

Problems of Bi-Encoders

Current Bi-Encoders are subject to three problems:

1. Task discrepancy
   - Training: either one or multiple, equally relevant positives
   - Inference: multiple positives with graded relevance

   Contribution: contrastive ranking-aware loss

2. Domain discrepancy
   - Queries: short, simple; representation computed live
   - Documents: long, complex; representations can be cached

   Contribution: decoupled encoders with compatible latent spaces

3. Scale discrepancy
   - Usually multiple (q, d) form a batch, with docs from other queries being implicit negatives; but: one query may has multiple relevant docs in reality
   - This training setup is mostly due to sparsity of ground truth labels

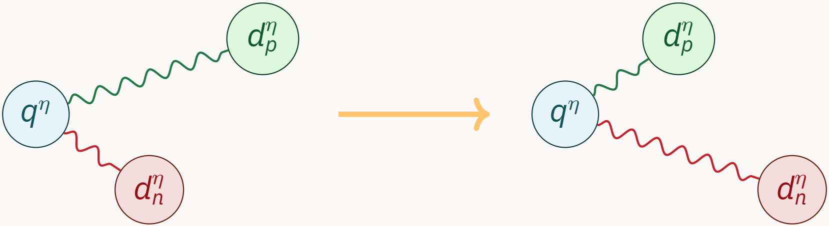
   Contribution: knowledge distillation with graded, single-query batches

(I) Fixing Task Discrepancy

Model Architecture

Contrastive Learning

Objective: given an anchor (query q) and a positive (document d_p) and negative (document d_n) example, minimize the distance between anchor and positive and maximize the distance between anchor and negative.



Contrastive learning can be extended to multiple positives and negatives.
But: does not discriminate in-class (i.e., trains set retrieval only).

Model Architecture

Contrastive Ranking-Aware Learning

Ranking information can be directly integrated into the loss [2, 7]:

$$l_{\tau,k}(q, D) = \log \frac{\exp(q^\eta \cdot d_i^\eta / \tau)}{\sum_{j=1}^{b} \exp(q^\eta \cdot d_j^\eta / \tau)}$$

For each query q...

... using a standard contrastive loss

Model Architecture

Contrastive Ranking-Aware Learning

Ranking information can be directly integrated into the loss [2, 7]:

$$l_{\tau,k}(q, D) = \log \frac{\mathbb{1}_{[r_q(d_i) \leq k]} \exp(q^\eta \cdot d_i^\eta / \tau)}{\sum_{j=1}^{b} \mathbb{1}_{[r_q(d_j) \geq r_q(d_i)]} \exp(q^\eta \cdot d_j^\eta / \tau)}$$

For each query q...

... using a standard contrastive loss

... we inject distant ranking supervision $r_q(\cdot)$ (oracle),

Model Architecture

Contrastive Ranking-Aware Learning

Ranking information can be directly integrated into the loss [2, 7]:

$$l_{\tau,k}(q, D) = \frac{-1}{k} \sum_{i=1}^{|D|} \log \frac{\mathbb{1}_{[r_q(d_i) \leq k]} \exp(q^\eta \cdot d_i^\eta / \tau)}{\sum_{j=1}^{b} \mathbb{1}_{[r_q(d_j) \geq r_q(d_i)]} \exp(q^\eta \cdot d_j^\eta / \tau)}$$

For each query q...

... using a standard contrastive loss

... we inject distant ranking supervision $r_q(\cdot)$ (oracle),

... such that each of the top-k documents (positives)

Model Architecture

Contrastive Ranking-Aware Learning

Ranking information can be directly integrated into the loss [2, 7]:

$$l_{\tau,k}(q, D) = \frac{-1}{k} \sum_{i=1}^{|D|} \log \frac{\mathbb{1}_{[r_q(d_i) \leq k]} \exp(q^\eta \cdot d_i^\eta / \tau)}{\sum_{j=1}^{b} \mathbb{1}_{[r_q(d_j) \geq r_q(d_i)]} \exp(q^\eta \cdot d_j^\eta / \tau)}$$

For each query q...

... using a standard contrastive loss

... we inject distant ranking supervision $r_q(\cdot)$ (oracle),

... such that each of the top-k documents (positives)

... is contrasted by each document following it (negatives).

Model Architecture

Contrastive Ranking-Aware Learning

Ranking information can be directly integrated into the loss [2, 7]:

$$l_{\tau,k}(q, D) = \frac{-1}{k} \sum_{i=1}^{|D|} \log \frac{\mathbb{1}_{[r_q(d_i) \leq k]} \exp(q^\eta \cdot d_i^\eta / \tau)}{\sum_{j=1}^{b} \mathbb{1}_{[r_q(d_j) \geq r_q(d_i)]} \exp(q^\eta \cdot d_j^\eta / \tau)}$$

For each query q...

... using a standard contrastive loss

... we inject distant ranking supervision $r_q(\cdot)$ (oracle),

... such that each of the top-k documents (positives)

... is contrasted by each document following it (negatives).

Standard BERT-based text encoders are used for η_q and η_d.

Model Architecture

Contrastive Ranking-Aware Learning

$$\sum_{i=3}^{k} \quad d_{14} \quad d_1 \quad d_{27} \quad d_5 \quad d_9 \quad d_{12} \quad d_{127} \quad d_{62} \quad d_{49} \quad d_{45}$$

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from D,

Model Architecture

Contrastive Ranking-Aware Learning

$$\sum_{i=3}^{k} \quad \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ d_{14} & d_{1} & d_{27} & d_{5} & d_{9} & d_{12} & d_{127} & d_{62} & d_{49} & d_{45} \end{array}$$

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from D,

... ranked by $r_q(\cdot)$ given as above,

Model Architecture

Contrastive Ranking-Aware Learning

$$\sum_{i=3}^{k} \quad \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ d_{14} & d_{1} & d_{27} & d_{5} & d_{9} & d_{12} & d_{127} & d_{62} & d_{49} & d_{45} \end{array}$$

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from D,

... ranked by $r_q(\cdot)$ given as above,

... documents at ranks $1, 2$ are ignored (treated in previous iterations),

Model Architecture

Contrastive Ranking-Aware Learning

$\sum_{i=3}^{k}$	1	2	3	4	5	6	7	8	9	10
	d_{14}	d_{1}	d_{27}	d_{5}	d_{9}	d_{12}	d_{127}	d_{62}	d_{49}	d_{45}

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from *D*,

... ranked by $r_q(\cdot)$ given as above,

... documents at ranks $1, 2$ are ignored (treated in previous iterations),

... the document at rank 3 is treated as positive,

Model Architecture

Contrastive Ranking-Aware Learning

$$\sum_{i=3}^{k} \quad \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ d_{14} & d_{1} & d_{27} & d_{5} & d_{9} & d_{12} & d_{127} & d_{62} & d_{49} & d_{45} \end{array}$$

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from D,

... ranked by $r_q(\cdot)$ given as above,

... documents at ranks $1, 2$ are ignored (treated in previous iterations),

... the document at rank 3 is treated as positive,

... and is contrasted by documents at ranks $4 \ldots b$ as negatives.

Model Architecture

Contrastive Ranking-Aware Learning

$$\sum_{i=3}^{k}$$

1	2	3	4	5	6	7	8	9	10
d_{14}	d_1	d_{27}	d_5	d_9	d_{12}	d_{127}	d_{62}	d_{49}	d_{45}

For example, at iteration 3 of the loss computation ...

... with a batch of 10 documents from D,
... ranked by $r_q(\cdot)$ given as above,
... documents at ranks $1, 2$ are ignored (treated in previous iterations),
... the document at rank 3 is treated as positive,
... and is contrasted by documents at ranks $4 \ldots b$ as negatives.

Analogously, at iteration 4, ranks $1, 2, 3$ are ignored,
rank 4 is treated positive, and ranks $5 \ldots b$ as negatives.

Model Architecture

Loss Properties

The loss requires the model to learn a latent space such that:

(1) maximize similarity of query to positive documents $\left(q^{\eta} \cdot d_p^{\eta} \gg 0\right)$

(2) minimize similarity sum of query to negative documents $\left(\sum q^{\eta} \cdot d_n^{\eta} \to 0\right)$

(3) (1) and (2) are competing because of ranking supervision

- each negative sample up to *k* becomes positive in a later iteration
- thus (1) and (2) need to balance out between iterations dependent on rank
- the earlier in the ranking, the more important (1) is over (2) for loss minimum

The global loss is the average over all top-*k* ranks over all queries.

Model Architecture

Summary

In summary, the proposed **CoRAL** loss resembles the target task of retrieval task more closely than previous contrastive pretraining approaches.

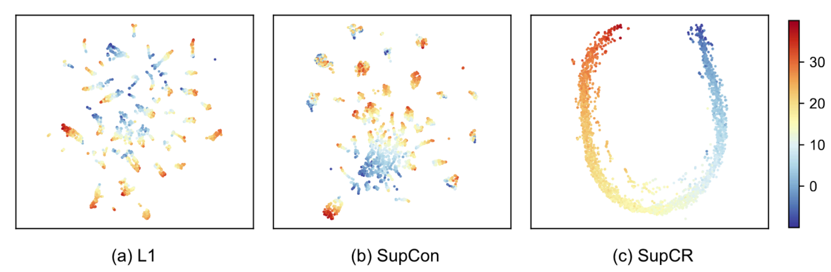


Figure 1: UMAP representation of latent space from models trained with L1, InfoNCE, and Ranked Contrastive Loss for temperature classification from webcam images [7].

Model Architecture

Summary

In summary, the proposed **CoRAL** loss resembles the target task of retrieval task more closely than previous contrastive pretraining approaches.

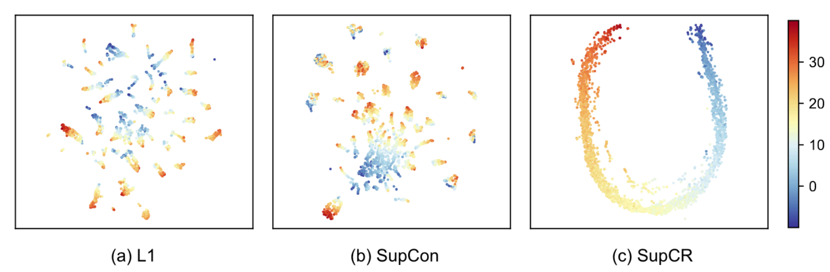


Figure 1: UMAP representation of latent space from models trained with L1, InfoNCE, and Ranked Contrastive Loss for temperature classification from webcam images [7].

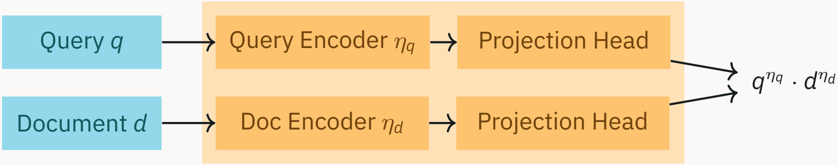
Contribution

Ranked contrastive loss has only been applied for single target rank concepts; application to multi-faceted rank objectives (retrieval) is novel.

(II) Fixing Domain Discrepancy

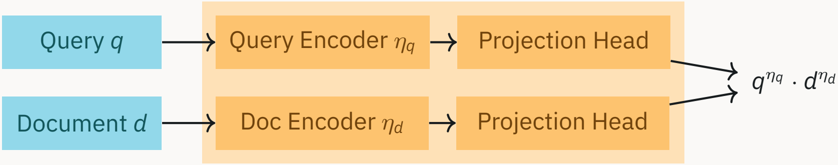
Decoupled Representations

Multimodal Training



Decoupled Representations

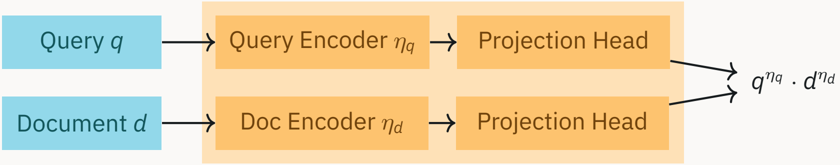
Multimodal Training



- **Query encoder** can be small & fast for efficiency
- **Document encoder** can be large & complex for effectiveness
- **Multimodal training** with projection heads allows for joint latent space

Decoupled Representations

Multimodal Training

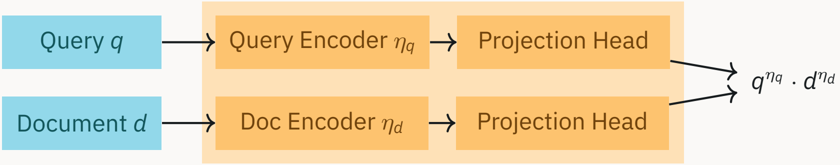


- **Query encoder** can be small & fast for efficiency
- **Document encoder** can be large & complex for effectiveness
- **Multimodal training** with projection heads allows for joint latent space

We can even omit the projection head of the query encoder and utilize a freezed pre-trained model (e.x. `distilBERT`).

Decoupled Representations

Multimodal Training



- **Query encoder** can be small & fast for efficiency
- **Document encoder** can be large & complex for effectiveness
- **Multimodal training** with projection heads allows for joint latent space

We can even omit the projection head of the query encoder and utilize a freezed pre-trained model (e.x. `distilBERT`).

Contribution

Utilize multimodal training to derive both efficient and effective bi-encoder models, taking inspiration from recent multimodal text/image models.

(III) Fixing Scale Discrepancy

Training Setup
Batch Construction

Traditional Setup

- Construct batch from (q, d) positive pairs; documents from other queries are treated as implicit negatives
- Problem: we can not ensure ‘correct’ negatives; we only learn top-1 retrieval

Improved Setup

- Single-query batches based on rank supervision
- Rank supervision induced by an oracle Ω (teacher model, ground truth, ...)
- Top-k are used as positives, rest of ranking as negatives

Training Setup

Sources of Rank Supervision

- Synthetic rankings from teacher models (e.x. `monoT5/duoT5` [3])
  - infinitely available since it can be synthesized at training time
  - but: trained model cannot exceed the effectiveness of the teacher model
- Direct rankings from human annotations (e.x. `TREC`)
  - sparse, and not suitable for training; evaluation only
- Pseudo rankings from large-scale query-logs (e.x. `AQL` [4])
  - allows for generalization beyond teacher model
  - vast amount of queries, but: limited depth per query

Idea

Can we generate training data by combining “real” results from query logs and augment with “synthetic” results from teacher models?

Current Status

Done

- Literature review, theoretical foundation
- Model implementation
- Convergence tested on small toy data

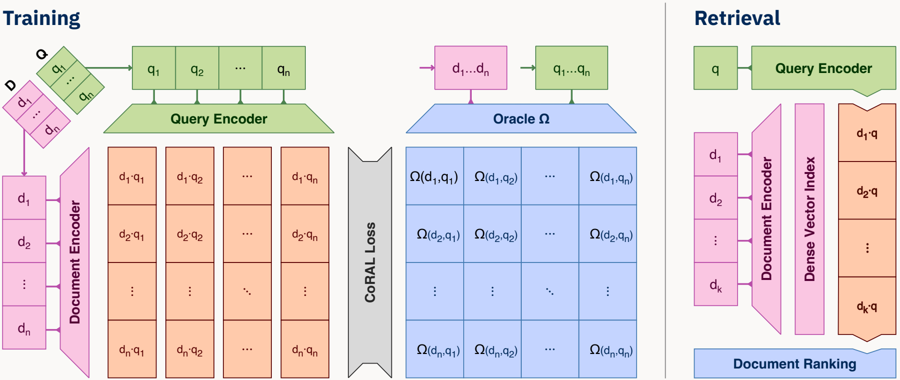
In progress

- Data curation & pretrained model selection
- Code optimization for large-scale training

Todo

- Batch sampling & training
- Ablation studies & evaluation

Conclusion



Summary

We adress the three main challenges of representation learning for retrieval using a ranked contrastive loss in conjunction with decoupled encoders and knowledge distillation for data augmentation.

References I

[1] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023)*. ACM, March 2023. doi: 10.1145/3576840.3578327.

[2] David T. Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 897–905. AAAI Press, 2022.

[3] Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667, 2021. URL `https://arxiv.org/abs/2101.05667`.

References II

[4] Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. The Archive Query Log: Mining millions of search result pages of hundreds of search engines from 25 years of web archives. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. ACM, 2023.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL `https://doi.org/10.18653/v1/D19-1410`.

References III

[6] Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. *CoRR*, abs/2202.06991, 2022. URL `https://arxiv.org/abs/2202.06991`.

[7] Kaiwen Zha, Peng Cao, Yuzhe Yang, and Dina Katabi. Supervised contrastive regression. *CoRR*, abs/2210.01189, 2022. doi: 10.48550/arXiv.2210.01189. URL `https://doi.org/10.48550/arXiv.2210.01189`.