In defense of dual-encoders for neural ranking

Aditya Krishna Menon

Sadeep Jayasumana

Ankit Singh Rawat

Seungyeon Kim

Sashank Reddi

Sanjiv Kumar

Google Research

Information retrieval

- Given a query, and a document corpus, find *k* most relevant documents

Information retrieval

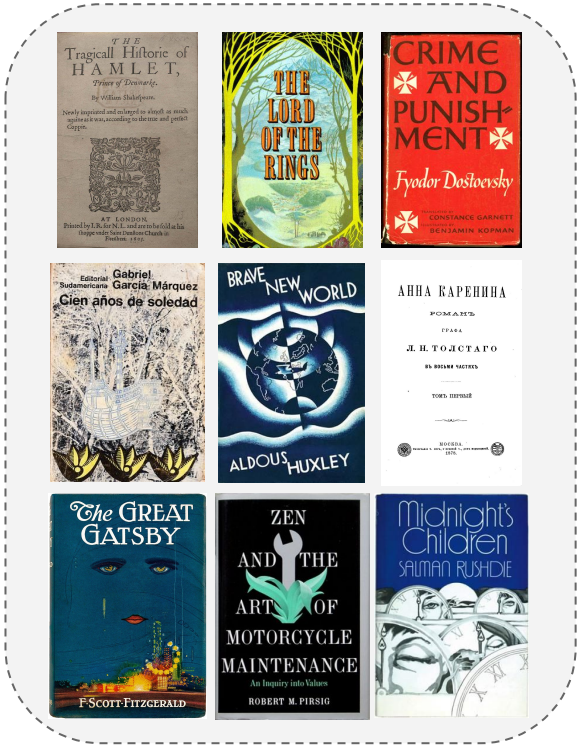
- Given a query, and a document corpus, find *k* most relevant documents

“books with
sad endings”

Information retrieval

- Given a query, and a document corpus, find *k* most relevant documents

"books with sad endings"



Information retrieval

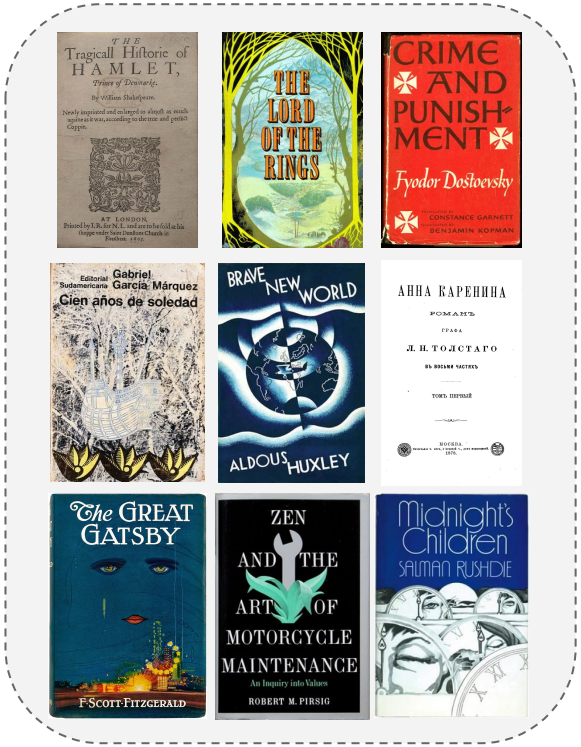
- Given a query, and a document corpus, find *k* most relevant documents

"books with sad endings"

Retrieval and scoring phases

- Typically, we first **retrieve** a set of **candidate** documents

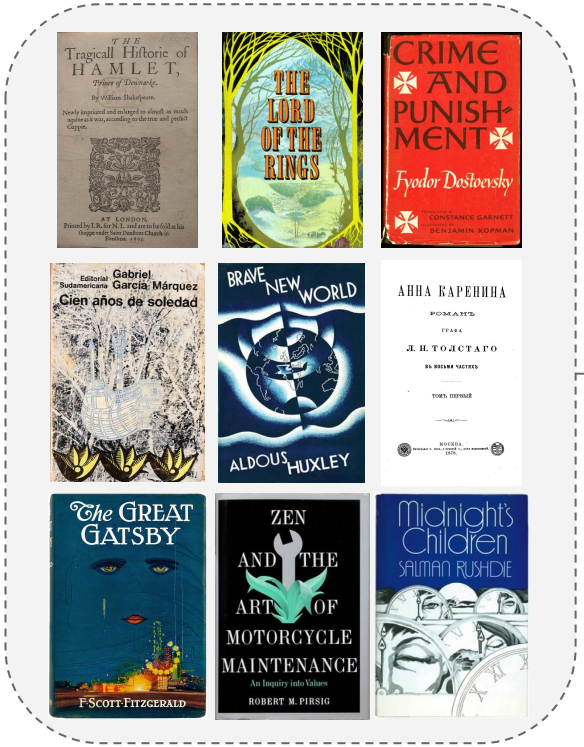
"books with sad endings"



Retrieval and scoring phases

- Typically, we first **retrieve** a set of **candidate** documents

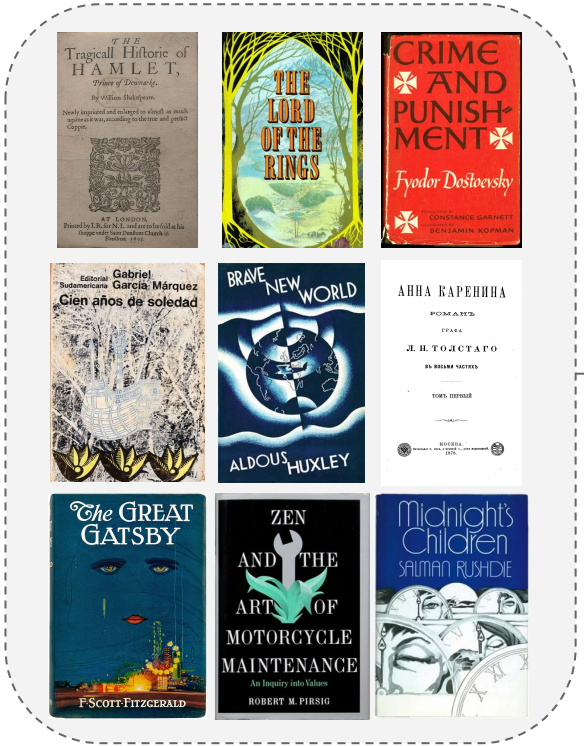
"books with sad endings"



Retrieval and scoring phases

- We then **score** and **re-rank** these documents to obtain the final results

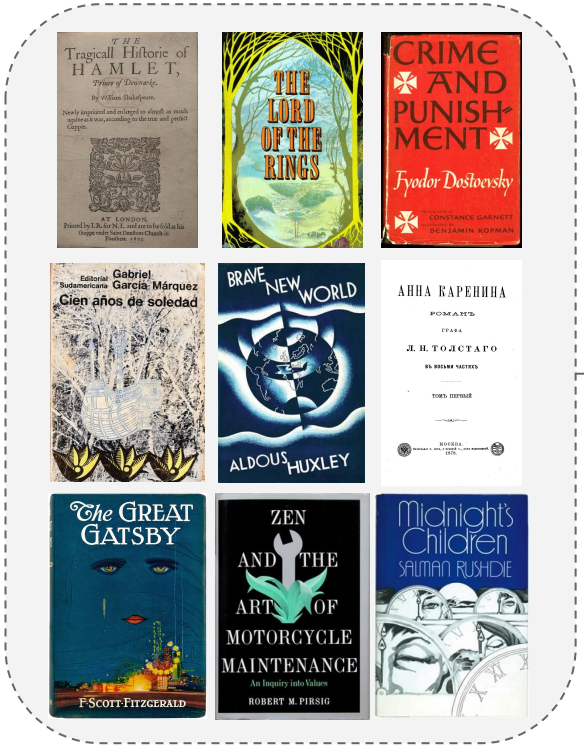
"books with sad endings"



Retrieval and scoring phases

- We then **score** and **re-rank** these documents to obtain the final results

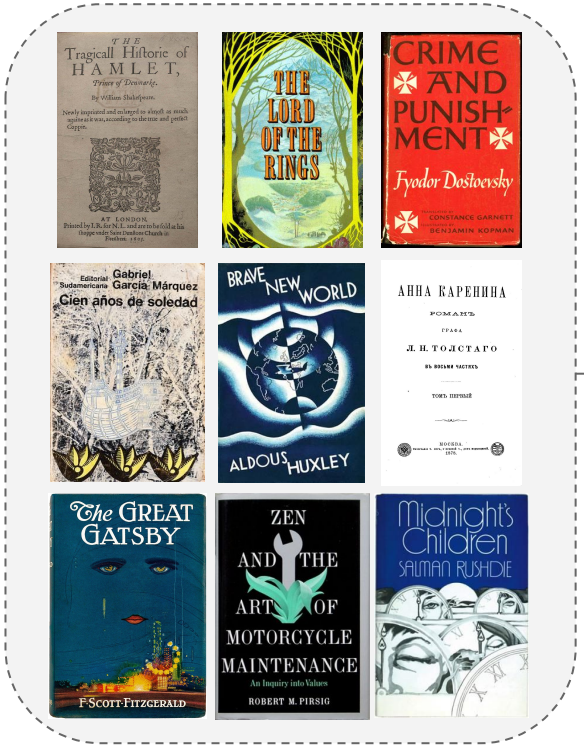
"books with sad endings"



Retrieval and scoring phases

- We then **score** and **re-rank** these documents to obtain the final results

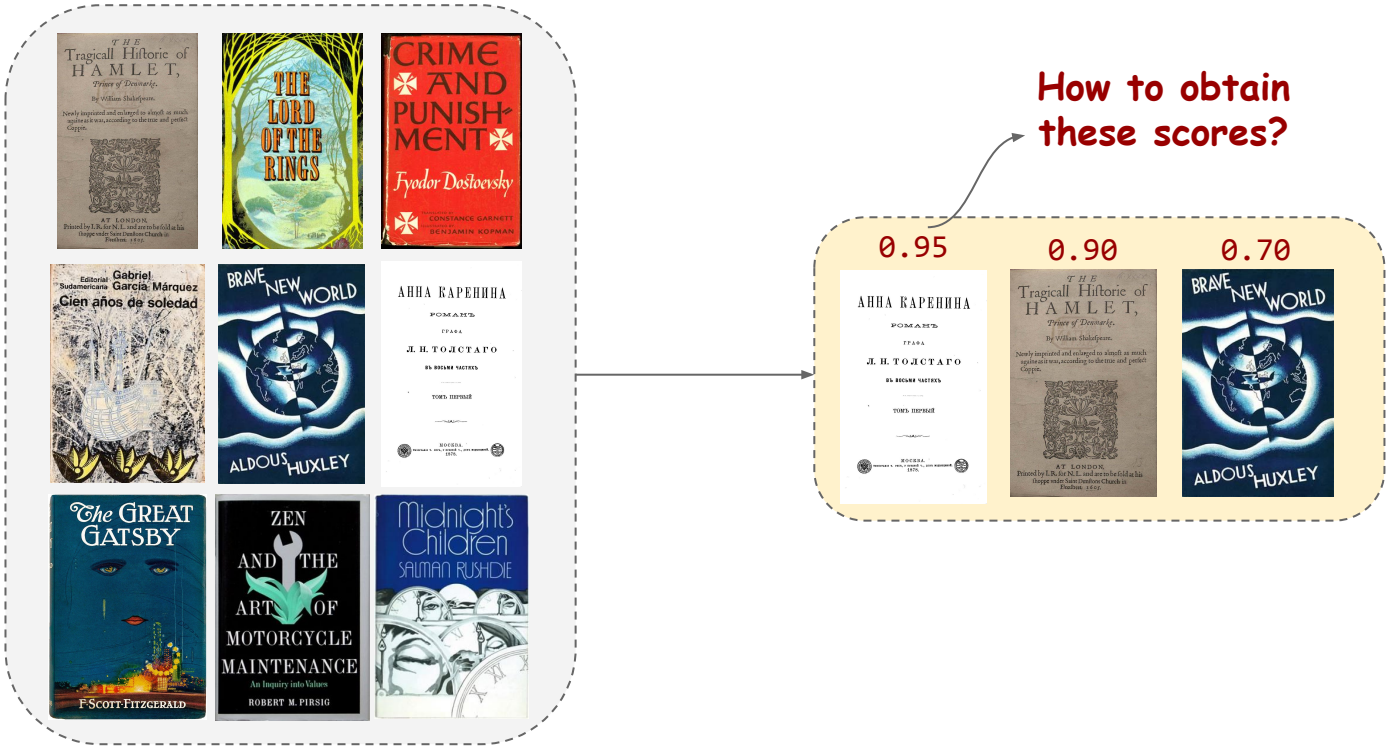
"books with sad endings"



Retrieval and scoring phases

- We then **score** and **re-rank** these documents to obtain the final results

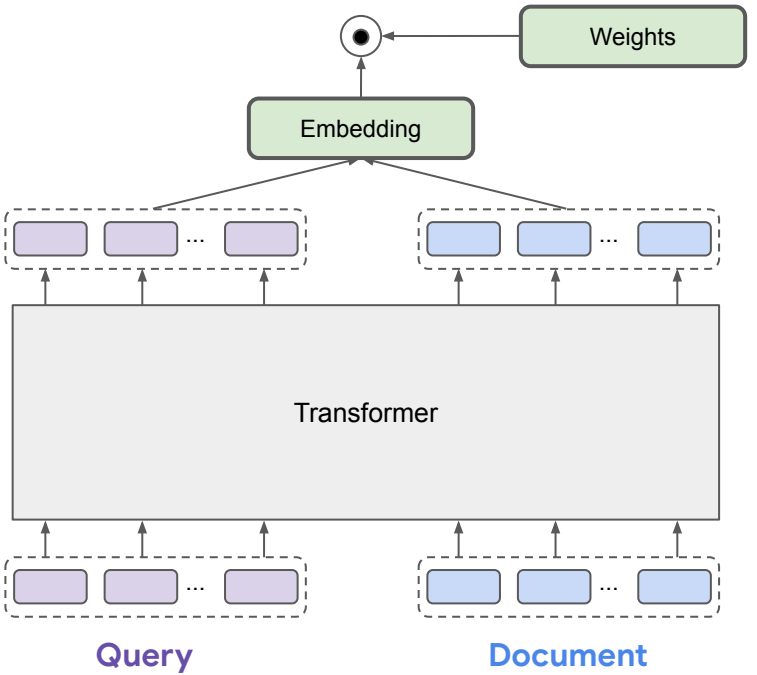
"books with sad endings"



Neural ranking via transformer models

- **Transformer** (e.g., BERT) based neural models are a popular choice

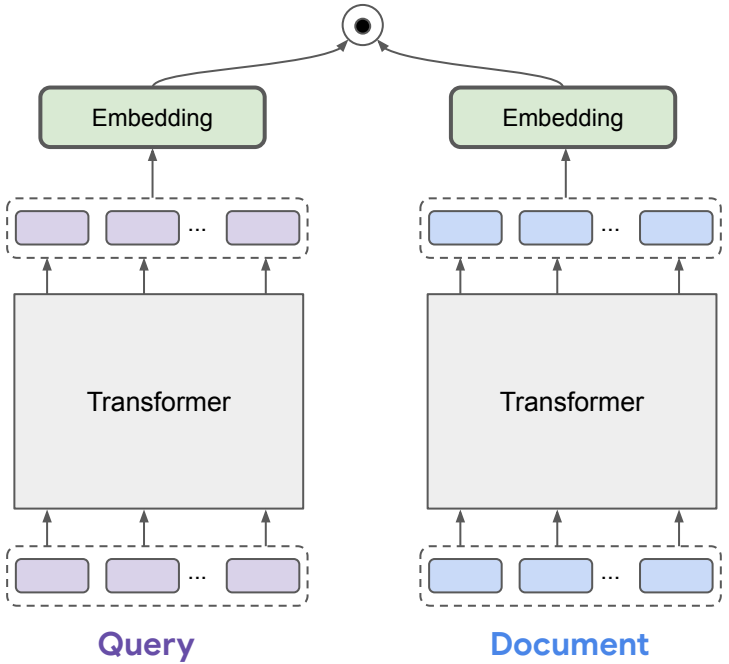
- **Cross-attention (CA)** models apply a transformer to the concatenation of query and document
  - Score = Embed(Query, Doc)^T Weight
  - **Joint** query-document interaction



Neural ranking via transformer models

- **Transformer** (e.g., BERT) based neural models are a popular choice

- **Dual-encoder (DE)** models apply a transformer to the query and document separately
  - Score = Embed(Query)^T Embed(Doc)
  - **Factorised** query-document interaction



CA versus DE models

- Empirically, CA models outperform DE models for re-ranking

	MSMARCO re-rank		TREC DL19 re-rank		NQ re-rank	
Model	**MRR**	**nDCG**	**MRR**	**nDCG**	**MRR**	**nDCG**
Cross-attention BERT (12-layer)	0.370	0.430	0.829	0.749	0.746	0.673
Dual-encoder BERT (6-layer)	0.310	0.360	0.834	0.677	0.676	0.601

CA versus DE models

- Empirically, CA models outperform DE models for re-ranking

	MSMARCO re-rank		TREC DL19 re-rank		NQ re-rank	
Model	**MRR**	**nDCG**	**MRR**	**nDCG**	**MRR**	**nDCG**
Cross-attention BERT (12-layer)	0.370	0.430	0.829	0.749	0.746	0.673
Dual-encoder BERT (6-layer)	0.310	0.360	0.834	0.677	0.676	0.601

What causes this performance gap?

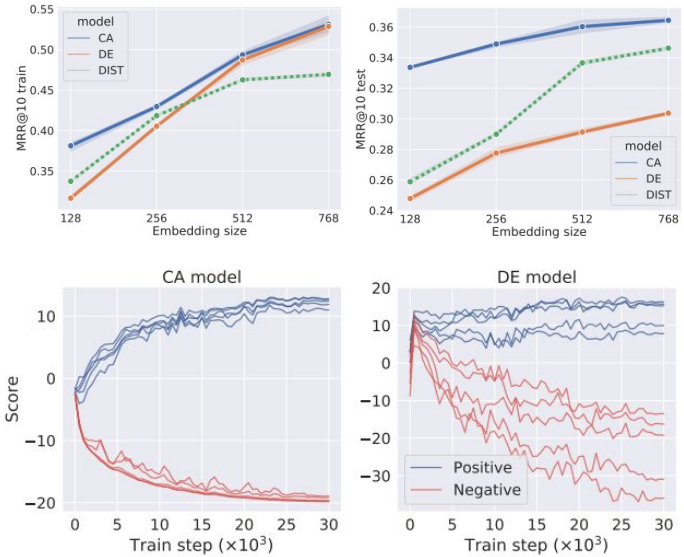
Summary of our work

Q: Why do CA models outperform DE models?

Poorer model capacity, or poorer model training?

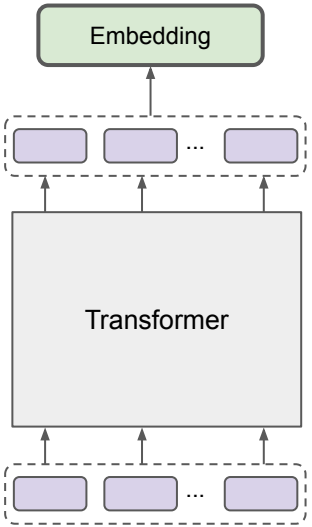
A: Model capacity may not be the cause; DE models exhibit a strong generalisation gap!

This can be alleviated by careful use of distillation



How good are DE models in theory?

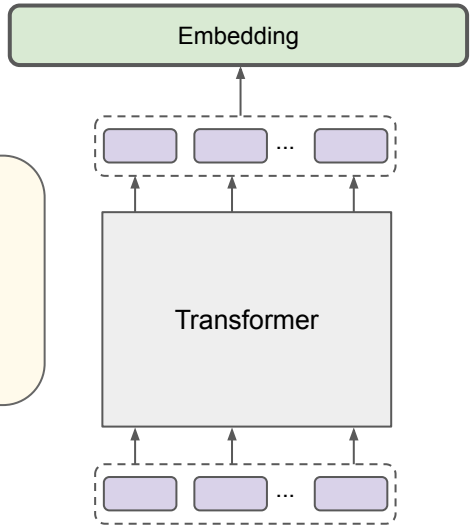
- Can DE models fit any reasonable relevance function (in principle)?



How good are DE models in theory?

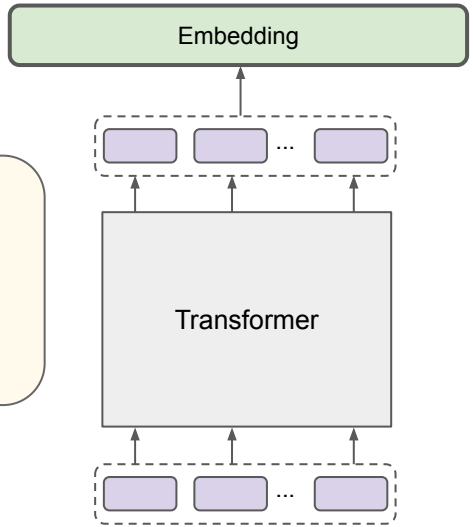
- Can DE models fit any reasonable relevance function (in principle)?
- **Yes**, with sufficiently high embedding dimension!

Proposition. Under mild technical conditions, any continuous query-document score function $s(q, d)$ can be approximated by some $Z(q)^{\mathsf{T}} W(d)$, where $Z(q)$, $W(d)$ have at most countably infinite dimension.



How good are DE models in theory?

- Can DE models fit any reasonable relevance function (in principle)?
- **Yes**, with sufficiently high embedding dimension!

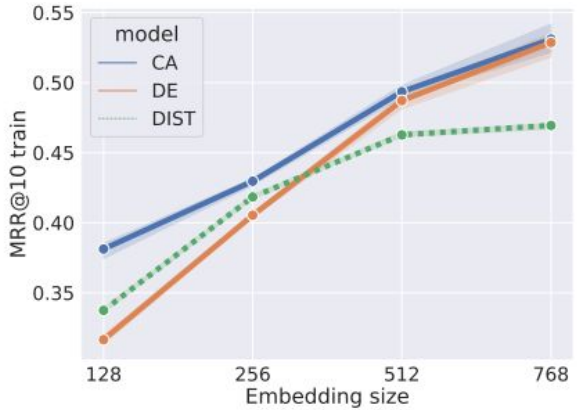


Proposition. Under mild technical conditions, any continuous query-document score function $s(q, d)$ can be approximated by some $Z(q)^{\mathsf{T}} W(d)$, where $Z(q)$, $W(d)$ have at most countably infinite dimension.

Do we see this in practice?

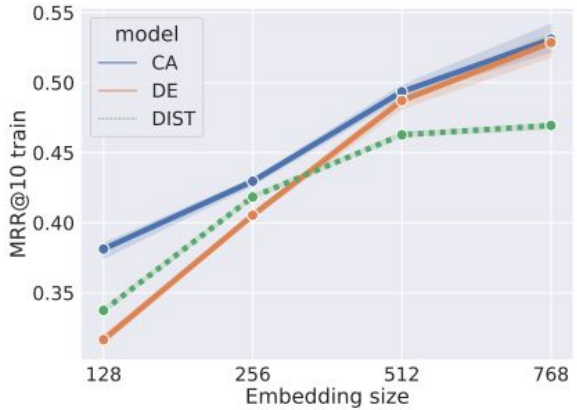
How good are DE models in practice?

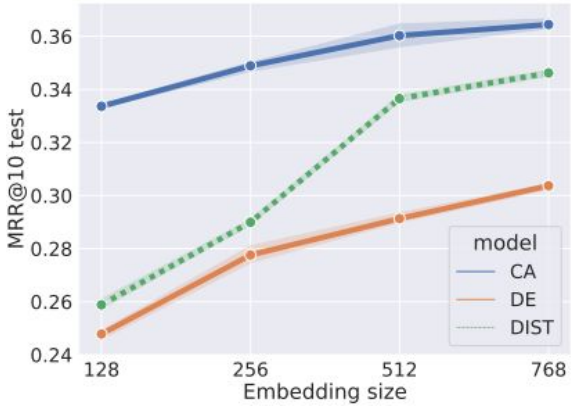
- With large embedding size, DE models work well on **training** set!



How good are DE models in practice?

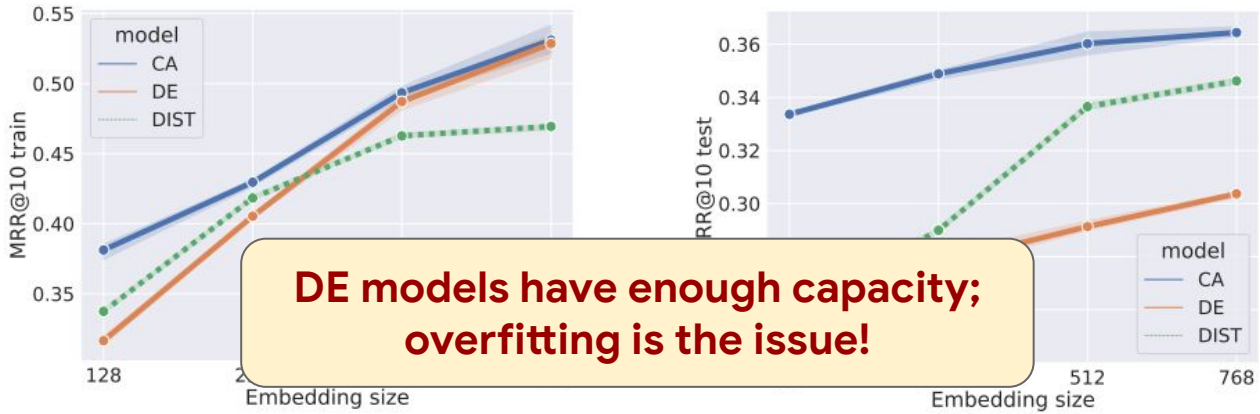
- With large embedding size, DE models work well on **training** set!
- However, there is a significant generalisation gap on the **test** set!





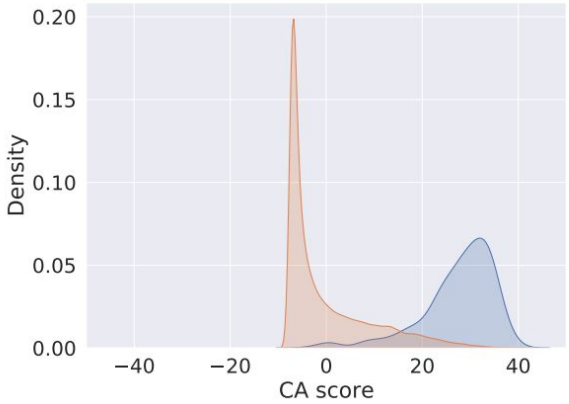
How good are DE models in practice?

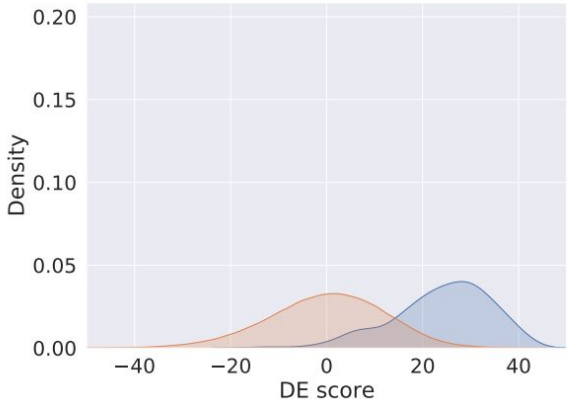
- With large embedding size, DE models work well on **training** set!
- However, there is a significant generalisation gap on the **test** set!

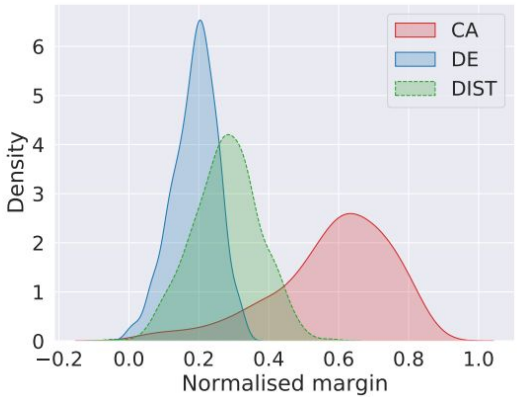


Why is there a generalisation gap?

- DE models yield poorer **margins**

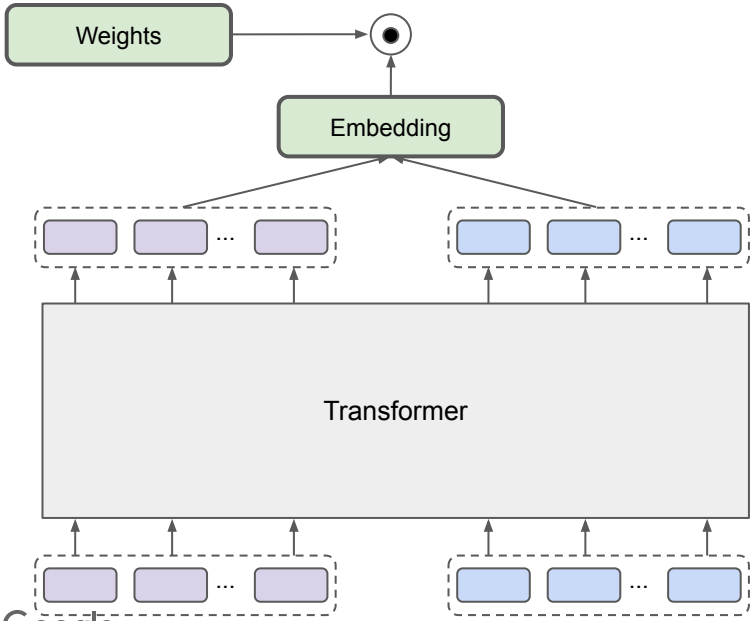


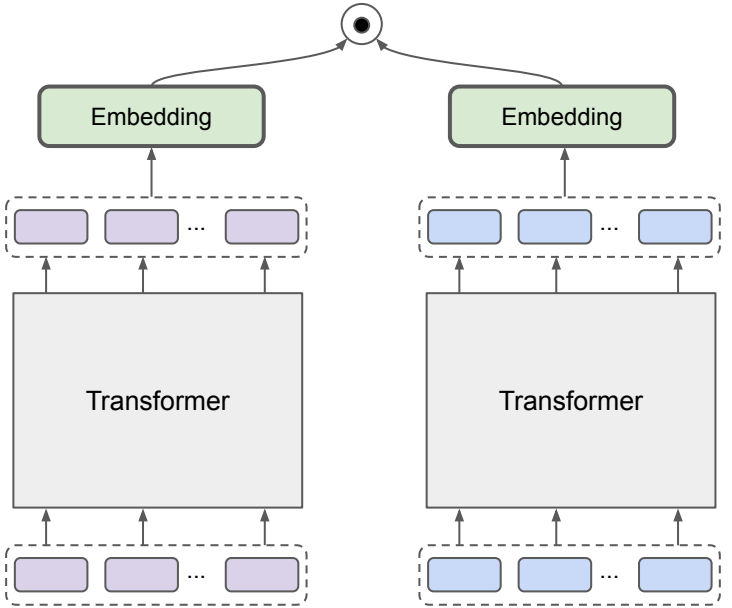




How do we mitigate the generalisation gap?

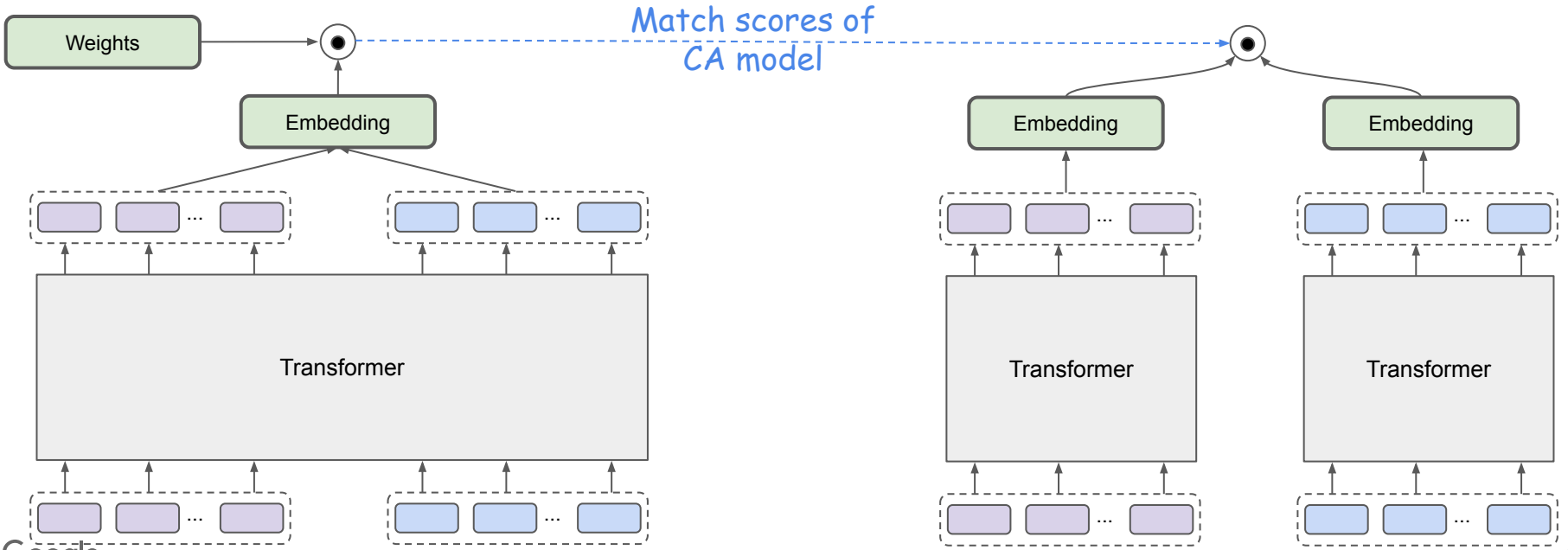
- We **distill** predictions from a CA to DE model





How do we mitigate the generalisation gap?

- We **distill** predictions from a CA to DE model



Distillation via multi-margin MSE (M3SE)

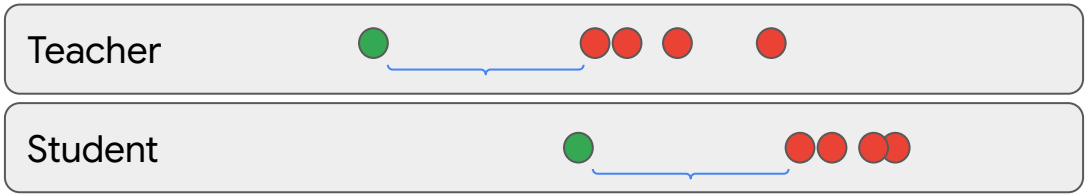
- Generalises margin MSE loss of (Hofstatter et al., ‘20)
- Encourages matching teacher margin

$$\ell_{\mathrm{m3se}}(\mathbf{t}, \mathbf{s}) = \sum_{i \in P}((t_i - t_{j^*}) - (s_i - s_{j^*}))^2 + \sum_{j \in N}[s_j - s_{j^*}]_+^2$$

Teacher score

Student score

Highest scoring negative



Empirical results

- Distillation can help mitigate the generalisation gap!

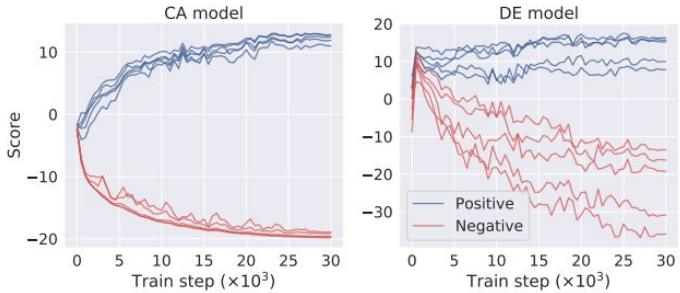
Model	MSMARCO re-rank MRR	MSMARCO re-rank nDCG	TREC DL19 re-rank MRR	TREC DL19 re-rank nDCG	NQ re-rank MRR	NQ re-rank nDCG
One-hot models						
BM25 (Robertson & Zaragoza, 2009)	0.194†	0.241†	0.689†	0.501†	—	—
ANCE (Xiong et al., 2021)	—	—		—	0.677†	—
Cross-attention BERT (12-layer)	0.370	0.430	0.829	0.749	0.746	0.673
Dual-encoder BERT (6-layer)	0.310	0.360	0.834	0.677	0.676	0.601
Distilled dual-encoders						
MSE (Hofstätter et al., 2020a)	0.289	0.343	0.781	0.693	0.659	0.591
Margin MSE (Hofstätter et al., 2020a)	0.334	0.392	0.867◇	0.718	0.673	0.594
RankDistil-B (Reddi et al., 2021)	0.249	0.301	0.852	0.708	0.649	0.561
Softmax CE (Equation 1)	0.346	0.405	0.846	0.726◇	0.682	0.607
M^3SE (Equation 4)	0.349	0.406	0.852	0.714	0.699	0.625

Empirical results

- More results in paper, including:
  - Use of ColBERT model as teacher
  - Insufficiency of alternate regularisation strategies
  - Noisy score updates of DE versus CA models

	Scoring function	
Teacher	**Dot**	**ColBERT**
One-hot	0.310	0.356
Dot	0.316	0.351
ColBERT	0.334	0.368
CA	0.334	0.376

Strategy	**Train MRR@10**	**Test MRR@10**
Baseline DE	0.619	0.310
Increased embedding dropout	0.588	0.299
Token dropout	0.572	0.291
Masked language loss	0.548	0.299
Focal loss	0.546	0.307



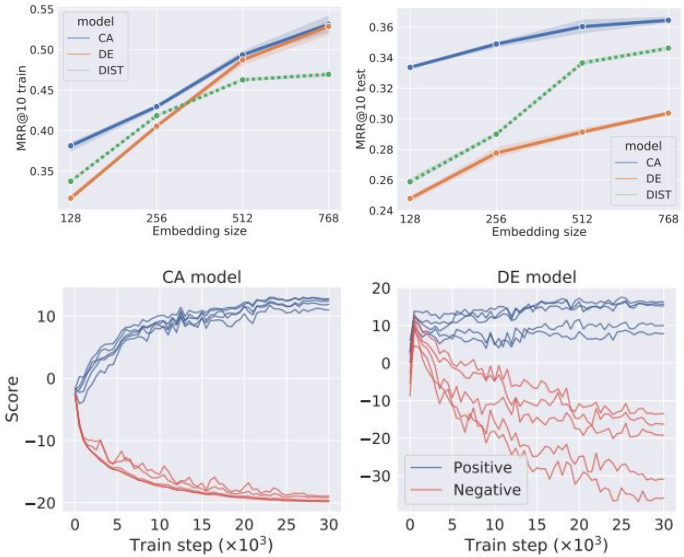
Summary of our work

Q: Why do CA models outperform DE models?

Poorer model capacity, or poorer model training?

A: Model capacity may not be the cause; DE models exhibit a strong generalisation gap!

This can be alleviated by careful use of distillation



See paper for more!