Scalable Attentive Sentence-Pair Modeling via Distilled Sentence Embedding

Oren Barkan*, **Noam Razin***, **Itzik Malkiel**, **Ori Katz**, **Avi Caciularu**, **Noam Koenigstein**

Microsoft R&D

*Equal Contribution

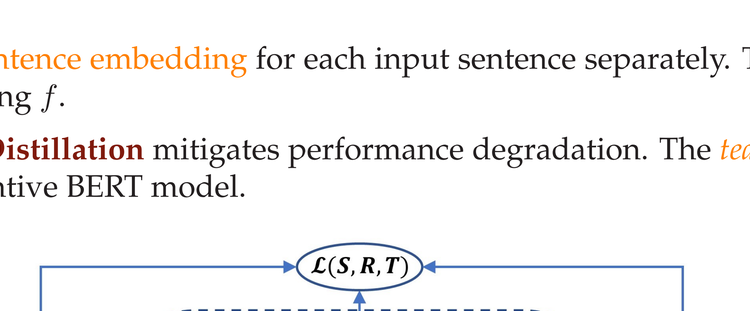
1) Motivation: Speeding Up Transformers

Large Transformer Models are Impractical for Production Systems:

- Current NLP Transformer models, e.g. BERT (Devlin et al. 2019), score a pair of sentences using multiple ***cross-attention*** operations.
- **Cross-attention**: each word in sentence Y attends all words in sentence Z and vice versa.
- Scoring a set of query-candidate sentences requires propagating all pairs throughout the whole network.

Enabling Fast Online and Offline Sentence-Pair Scoring:

- Replacing **cross-attention** operations with **self-attention** decouples the computation for each sentence. Decoupling allows fast online and offline sentence-pair scoring.



2) Setting: Sentence-Pair Tasks

Online Query-Candidates Scoring:

- Computing all scores for a new query sentence and an existing large candidates set.
- Highly relevant for search and text retrieval applications.

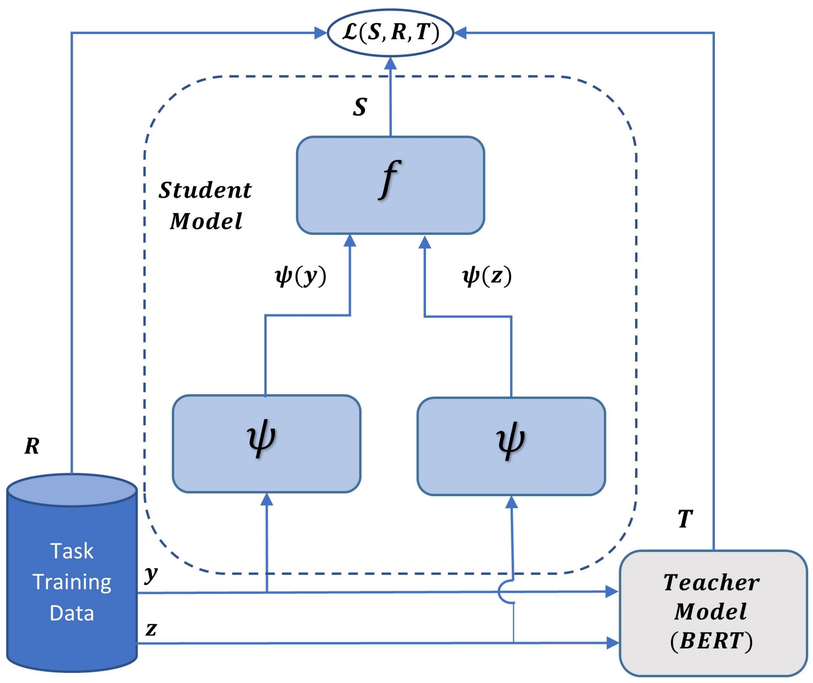
Offline All-Pairs Scoring:

- Computing similiarty scores for all pairs of sentences in a large database.
- Usages in text-based recommender and retrieval systems.

3) Distilled Sentence Embedding (DSE)

Model Description:

- DSE consists of a *student* self-attentive only BERT model ψ, and a low-cost similarity function f.
- ψ creates a sentence embedding for each input sentence separately. The final score is computed using f.
- **Knowledge Distillation** mitigates performance degradation. The *teacher* is the original cross-attentive BERT model.



y, z - Pair of sentences R - True label T - Teacher score S - DSE score

4) Computational Speedup

Online Query-Candidates Scoring:

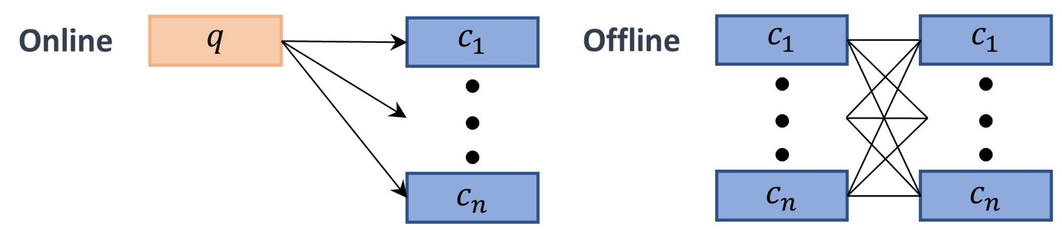
DSE allows precomputing the candidates' sentence embeddings. Online scoring requires only a single BERT forward pass.

> 1 hour → 0.2 seconds for 100K query-candidate pairs. ∼**13500x** speedup!

Offline All-Pairs Scoring:

Computing all sentence-pair scores requires only $O(n)$ BERT forward passes, instead of $O(n^2)$.

> 9.6 hours → 37 seconds for 1M sentence-pairs. ∼**900x** speedup!



5) Performance Evaluation

Sentence-Pair Tasks:

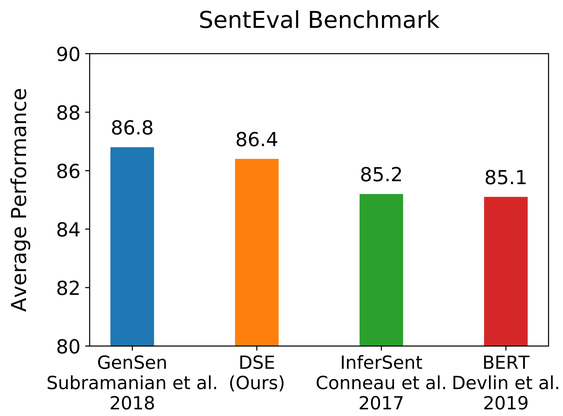
- We evaluate DSE on sentence-pair tasks from the GLUE benchmark (Wang et al. 2018).

Model	Avg Performance	DSE Improvement
BERT	86.82	-4.6%
DSE	82.83	-
ELMo + Attn (MT)[1]	76.45	8.3%

- Average relative degradation of only 4.6% compared to BERT. Significantly outperforms existing sentence embedding methods.

Universal Sentence Embeddings:

- DSE can also be used to create general purpose sentence embeddings.
- Extracted embeddings are competitive with state-of-the-art methods on the SentEval benchmark (Conneau and Kiela 2018).



[1]Peters et al. 2018

6) Contributions Summary

- A **sentence embedding** model that is **supervised by the outputs of large cross-attentive models**.
- Significant **speedup in online and offline query-candidate scoring**, with a relatively small degradation in performance.
- Can be used to create **high quality general purpose sentence embeddings**.