Distance Dilemma: Siamese vs. Triplet SBERT Networks for Information Retrieval

Andreas S. H. Rygaard (s220886)[2], Anna Bramsløw (s194656)[1], Nikolai Beck Jensen (s194639)[1], Rasmus Bryld Bagger (s194668)[1]

1 DTU Management, Technical University of Denmark; 2 DTU Bioengineering, Technical University of Denmark

Introduction

Since its introduction in 2019, BERT models have shown top tier results in various NLP tasks [2]. However, BERT is unsuited for similarity search in a large corpus due to combinatorial explosion of pairwise comparisons. In order to use the excellent NLP capabilities of BERT, the model can be fine-tuned to produce semantically meaningful embeddings by implementing it in a Siamese or Triplet network structure. This approach has yielded models with excellent sentence embedding capabilities, enabling similarity search using cheap distance metrics [3]. In this work, we investigate which of these network architectures are most suitable for information retrieval tasks across different similarity measures, respectively cosine similarity and euclidean distance.

Key Points

- We use a subset of the public **MS MARCO dataset** [1]. Our subset is comprised of 125k anonymized queries, sampled from Bing's search query logs, as well as a corpus, comprising 8.8 mio short text passages with information retrieved from Bing webdocuments. For each query, a cross encoder (CE) has identified the similarity to a set of passages in the corpus. From this list, a positive passage containing the answer is identified as well as a set of 'hard negatives' - passages that are hard to classify, given that they do not contain the answer but have a high similarity to the query.
- We develop a **Siamese BERT-Network** based on a distilled BERT model [4], which takes a query and a passage as input. The network is trained to produce sentence embeddings with a similarity corresponding to the CE-scores.
- We develop a **Triplet BERT-Network** based on a distilled BERT model [4], which takes a query (the anchor) and a related (positive) and less-related (negative) passage as an input. The network is trained to embed the sentences, minimizing the distance between the anchor and the positive passage, while maximizing the distance between the anchor and the negative passage.
- We evaluate models on **information retrieval** from the full passage corpus based on approx. 200 MS MARCO test queries and corresponding human annotated relevant passages.
- We make a qualitative inspection of the embedding space, using a **Principal Component Analysis** (PCA).

SBert Model Architechture

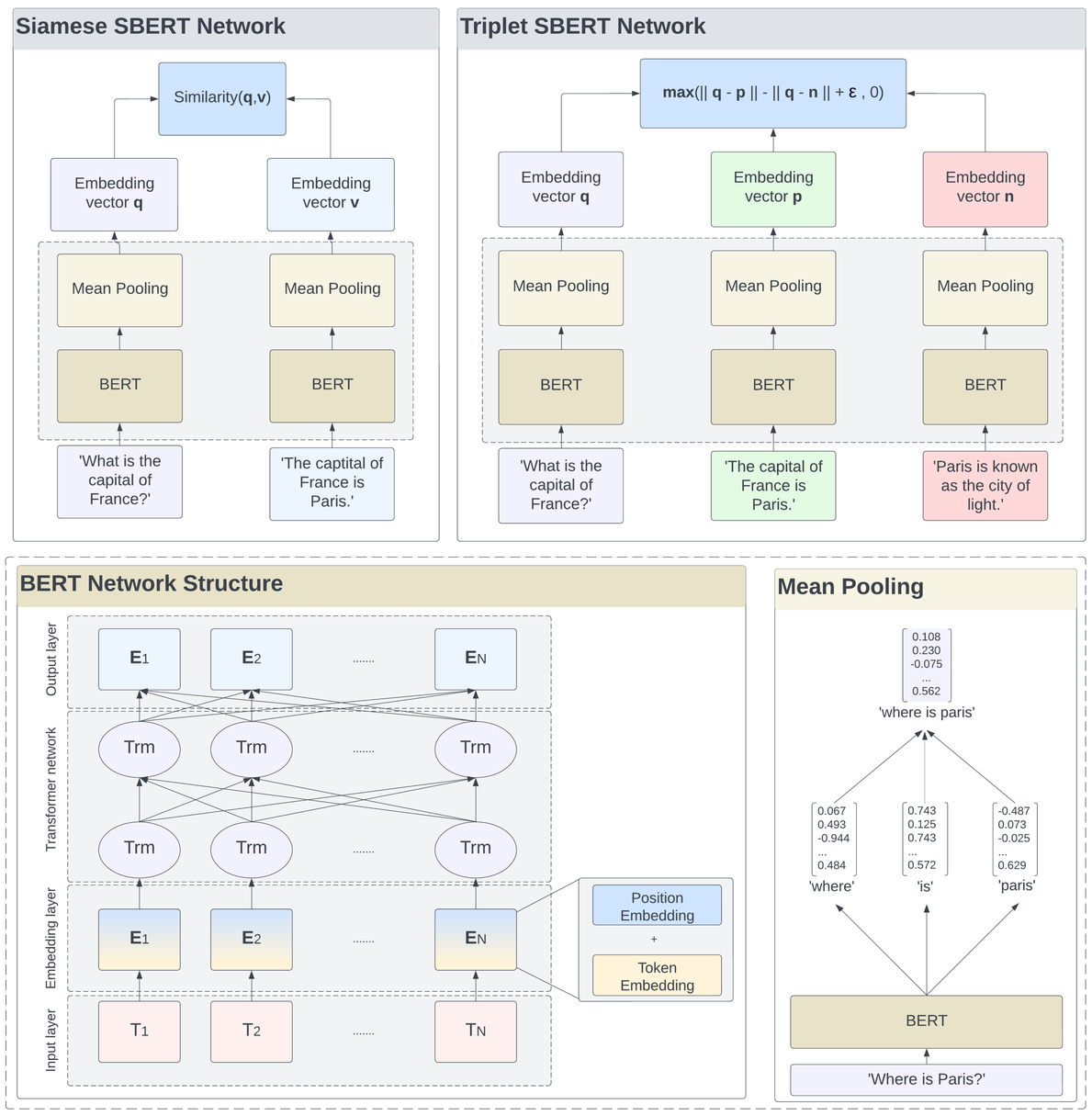

Figure 1: Model architecture representation.

Loss Functions

For the **Siamese network**, given anchor embedding q and passage embedding v, which can be either positive or negative, and cross-encoder similarity score y, we compute the loss as:

- **Cosine-Siamese loss**:
  $MSE(cosine\_similarity(q, v), y)$
- **Euclidian-Siamese loss**:
  $MSE(euclidean\_distance(q, v), 1 - y)$

For the **Triplet network**, given anchor embedding a and positive passage embedding p and negative n, we compute the loss as:

- **Cosine-Triplet loss**:
  $MAX((1 - cosine\_similiarity(q, p)) - (1 - cosine\_similiarity(q, n)) + 5, 0)$
- **Euclidian-Triplet loss**:
  $MAX(euclidean\_distance(q, p) - euclidean\_distance(q, n) + 5, 0)$

Evaluating Information Retrieval

For each test query, the models each retrieve the top 1, 5 and 10 relevant passages. These retrievals are evaluated across 200 test queries by the following metrics:

- **Precision**: Average share of relevant passages in top k across test queries.
- **Accuracy**: Share of retrievals with at least 1 relevant passage across test queries.
- **Mean Reciprocal Rank** (MRR): Average reciprocal rank of first relevant passage.
- **Normalized Discounted Cumulative Gain** (NDCG): The cumulative gain (sum of binary relevance across top k) is discounted with a logarithmic factor depending on position in retrieval. The discounted cumulative gain (DCG) is normalized by the ideal DCG to give NDCG.

Information Retrieval Performance

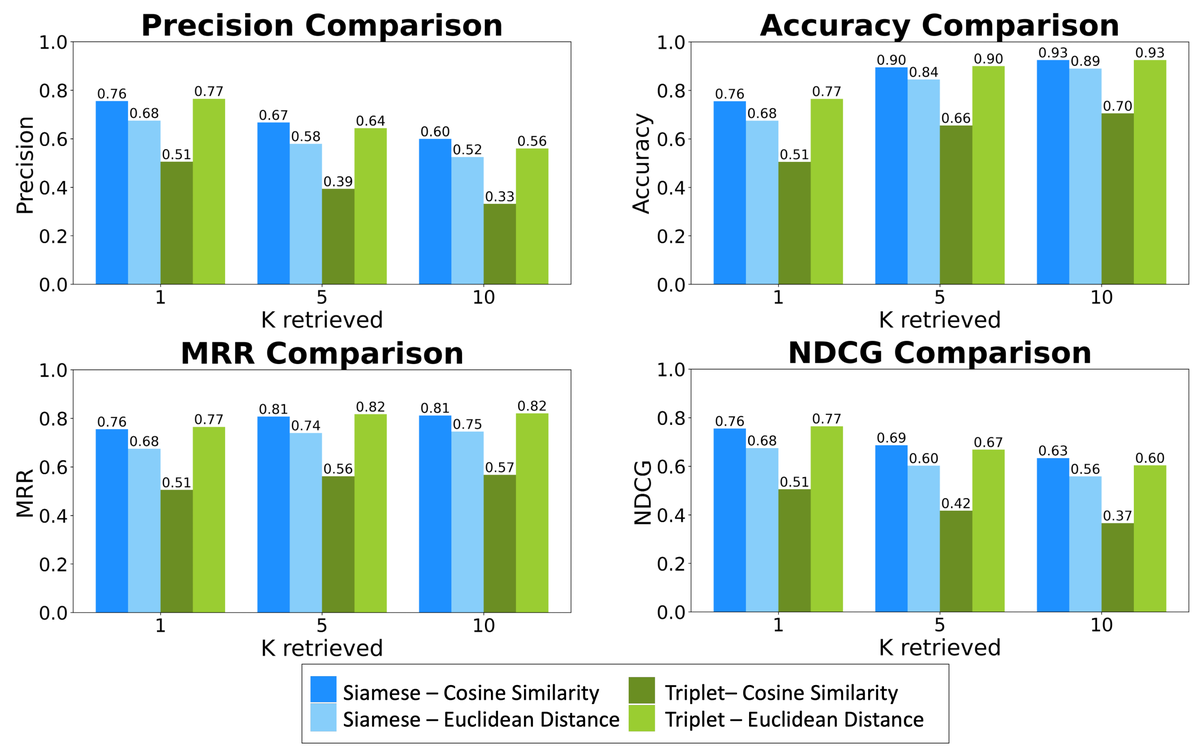

Figure 2: Performance of Siamese and Triplet models on four different information retrieval metrics. The models were trained and evaluated on the same similarity measure.

Results

- **Siamese network** performs best when trained with **cosine similarity** while **Triplet network** performs best, when trained with **euclidean distance**.
- Best performing models are Siamese trained with cosine similarity and Triplet trained with euclidean distance - **performing on par**.
- Training time was approx. **11 hours for the Siamese network** and **15 hours for the Triplet network**.

Embedding Visualizations

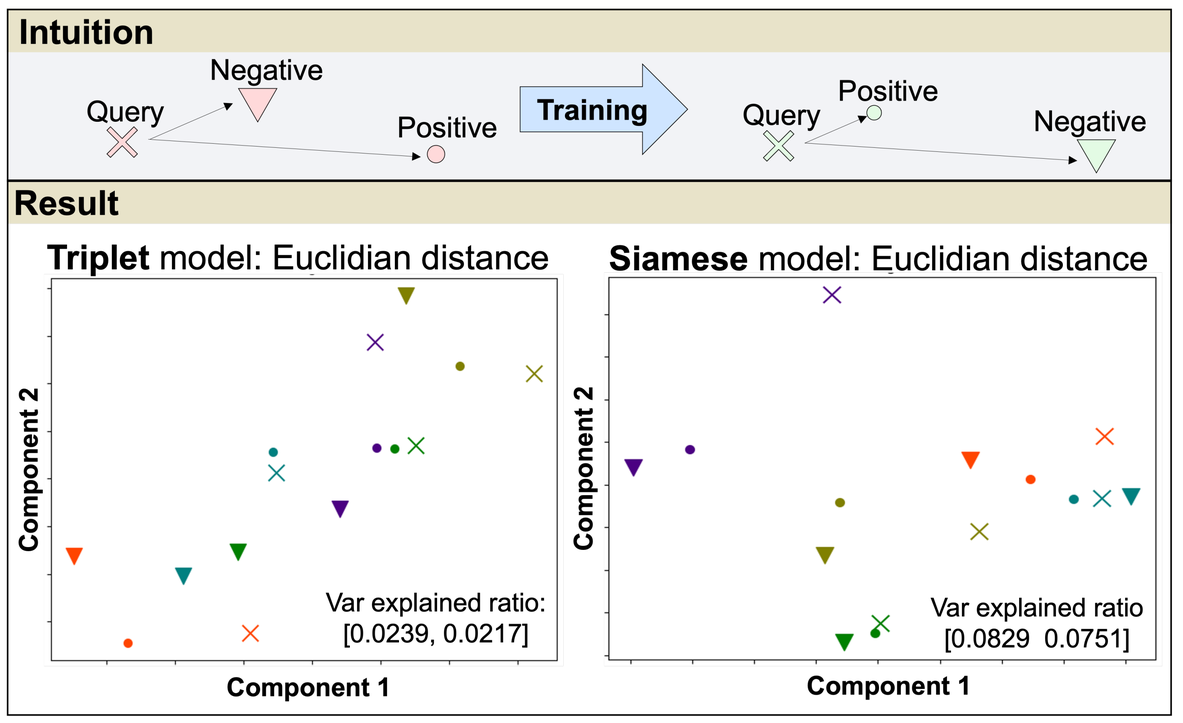

Figure 3: PCA visualization of Triplet embeddings using encodings from Siamese and Triplet networks trained with euclidian distance.

Intuitively, training **moves the query closer to the positive passages than the negative passages**. Seen in Figure 3 is a PCA of the embedding-spaces from Triplet and Siamese models trained with euclidean distance. The Siamese model, **places the triplets in distinct groups**. The Triplet model has **less variance explained in the PCA**. The points are less grouped, but **the intuitive pattern is seen**

Acknowledgments

We would like to thank our supervisor Beatrix Miranda Ginn Nielsen for excellent supervision.

References

[1] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[3] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.