Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks

Link: https://arxiv.org/pdf/2010.08240.pdf

Cross Encoders: perform full attention, too slow

- both sentences passed to the network and full attention applied,
- quadratic complexity while computing similarity scores
- easier training

Bi Encoder: mapping input to a vector space

- map input to a dense vector space for efficient indexing
- challenging training and requires more training data to map to a vector space

Use cross encoder training scores to sbert bi encoder for training.

Previous Work

- Bert's cross encoder performs full attention of the input seperated by SEP, there are **no independent representations**. SBERT solved it by creating input embeddings.
- Poly encoder does embedding generation of context and candidate and then does full attention of these two, cannot be applied for symmetric similarity, doesnt help in info retrieval.
- DiPair same as AugSBERT but focuses on speed based on biencoder archtiecture, AugSBERT focusses on sampling (**sampling appropriate pairs has a decisive impact on performance**)

Sampling Techniques

- Pairwise Sampling, reusing individual sentences and recombine the sentences,

- Random sampling would lead to dissimilar pairs mostly negative,
- KDE, for classification randomly sampling positive and negative pairs in same proportion, for regression they use KDE for gold and silver set and minimize KL Divergence
- BM25, uses lexical overlap, select unique sentences and retireve top k sentences,
- Semantic Search Sampling, BM25 only selects sentences with lexical overlap, here they use cosine similarity to rank top k sentences,
- BM25 + SS sampling, aggregating the strategies skews data to negative pairs

Random seeds in BERT converge to different minima and generalize differently for unseen data. Optimize seeds with 5 random seeds, select the best model on the deployment set, and stop other models at 20% while the best model is trained completely.

Datasets

- Single Domain, STS, Arugment Similarity, Quora QP duplicates, Microsoft's paraphrase identification dataset.
- Multi Domain, question duplicates detection

Model

- Cross Encoder is bert uncased add a linear layer with sigmoid activation on top of the [CLS] token to output scores 0 to 1.
- Bi Encoder - Finetuned SBERT
- BM25 and Semantic Search Sampling
- Evaluation: Regression tasks, spearman's rank correlation is used. Classification, f1 score for positively labelled pairs.
    - Baselines, Jaccard similarity to measure the word overlap of the two input sentences for regression
- In sampling, KDE and BM25 individually give high performances but BM25 + Semantic Search sampling gives lower similarity.