ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction

Thiago Coelho Vieira

1.1 Main Concepts

- late interaction
- MaxSim largest cosine similarity between each query token embedding and all passage token embeddings.
- multi-vector representation

1.2 Interactions

2.1 Contribution

- improvements on ColBERTv1
 - dense vectors compressions + better negative selection
 - o ColBERTv1
 - 128dim vectors with 2 bytes = 256 bytes/vector
 - ColBERTv2
 - dimensionality reduction by arranging vectors in clusters indexed by 4 bytes (2^{32}) clusters)
 - improvement that enable 20-36bytes/vector
 - memory improvement ~6-10x (residual compression)
- multi-vectors are stored in cluster based on MaxSim
- new dataset LoTTE (Long-Tail Topic-stratified Evaluation)

2.2 How it works

Training

o add

Dimensionality Reduction - Product Quantization

- high dim vectors splitted in same size smaller vectors
- each sub-vector is associated with the nearest centroid on vector space
- replace the values of the centroids by the unique ids
- outputs a vector of unique ids for each centroid

Inverted Index

o centroids ids

Search

 \circ At search time, the query q is encoded into a multi-vector representation and its similarity to a passage d is computed as the summation of query-side MaxSim operations.

3. interesting/unexpected results

- in-domain
 - \circ beats DPR and SPLADEv2
- gigantic index
 - ColBERTv1 154GiB 🔻
 - \circ ColBERTv2 16GiB(1bit) and 25GiB(2bit)
- *MMR*@10
 - 1bit 36.2
 - o 2bit 35.5
- success@5 metric
- LoTTE dataset

4. Basic Doubts

long-tail topics (ask gpt)