Efficient Multi-Vector Document Retrieval with Adaptive Representation Size

Author: Jakub Štětina

Supervisor: Ing. Martin Fajčík, Ph.D.

Motivation

Problem: Multi-vector retrieval models like ColBERTv2 produce large indexes with per-token embeddings → high storage and retrieval cost.

Goal: Introduce learned sparsity during indexing to retain only important token representations.

Idea: Add a lightweight scoring layer to decide which token embeddings to keep.

Training

- Fine-tune on MS MARCO
- Enforce sparsity during training.
- Grid search λ parameter - sparsity intensiry

$$\mathcal{L} = \mathcal{L}_{\text{ColBERT}} + \lambda \cdot \frac{\sum_i |s_i|}{|S|}$$

Formula 1: Modified loss function with added L1 regularization term for sparsity control.

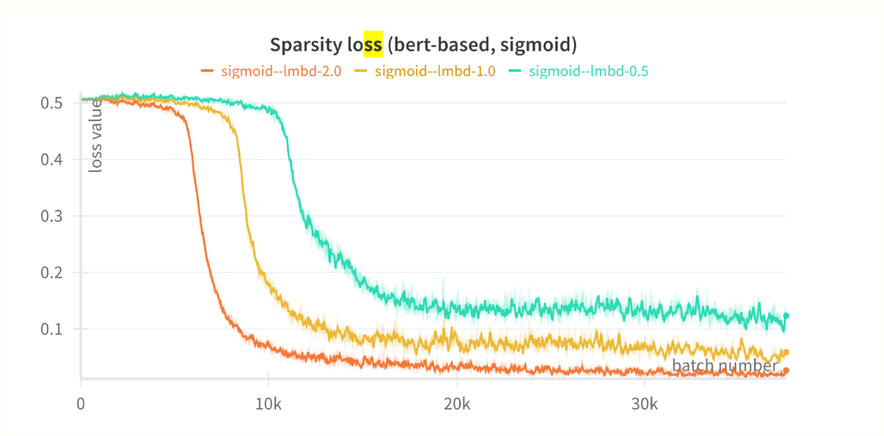

Figure 1: Effect of varying sparsity loss coefficient λ on the auxiliary loss curve during training.

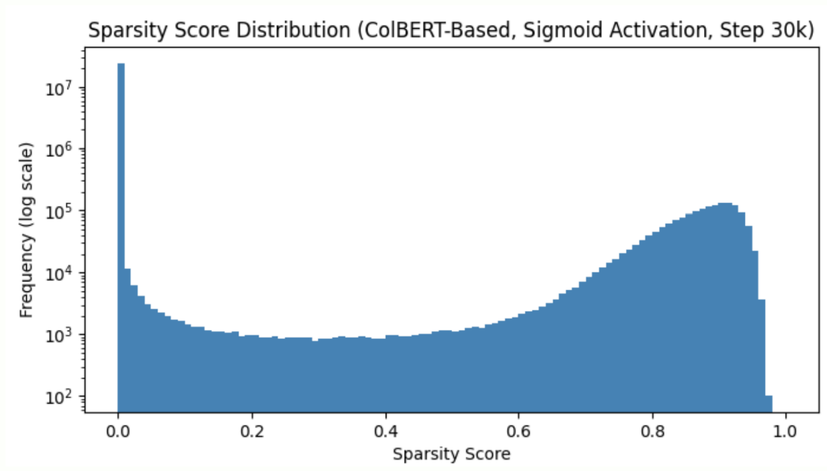

Figure 2: Distribution of learned sparsity scores after 30,000 training steps.

Sparsity Mechanism

- Sparisty scoring Layer on top of token embeddings.

- Each token gets a sparsity score $s \in [0,1]$ via sigmoid/softmax activation.
- Decision: Threshold-based or Quantile-based

$$\text{keep decision} = \mathbb{I}\left(\phi(W \cdot h + b) > \tau\right)$$

Formula 2: Decision rule for token retention based on activation score and threshold.

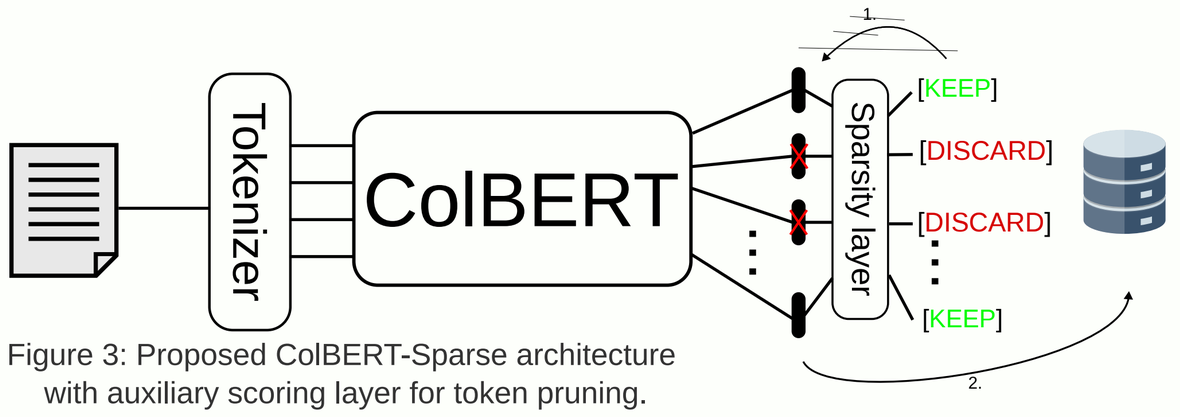

Figure 3: Proposed ColBERT-Sparse architecture with auxiliary scoring layer for token pruning.

Results

Model	Index Reduction (%)	R@10	Best Step
ColBERT-based	87.7	0.933	140k
BERT-based	63.2	0.944	35k
Baseline: ColBERTv2	0.0	0.957	-
Baseline: ColBERT-single-vect	98.5	0.517	70k

Table 1: Retrieval performance and index compression comparison against original ColBERT and single-vector baselines.

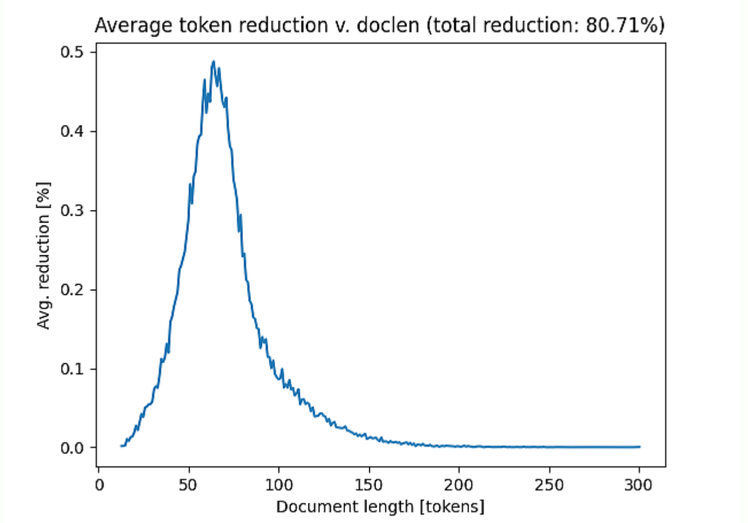

Figure 4: Average token reduction across documents of varying length

POS Category	Original (%)	Kept (%)	Retention Ratio (Kept / Original)
NOUN	29.66	52.83	1.78x
VERB	12.41	17.94	1.45x
ADJ	10.84	21.41	1.98x
ADV	2.74	2.15	0.78x
PRONOUN	2.38	0.38	0.16x
DETERMINER	8.51	0.36	0.04x
PREPOSITION	11.34	1.53	0.13x
CONJUNCTION	3.27	0.09	0.03x
MODAL	0.84	0.36	0.43x
NUMERAL	4.13	2.28	0.55x
PARTICLE	0.15	0.11	0.76x
INTERJECTION	0.01	0.00	0.24x
PUNCTUATION*	13.44	0.09	0.01x

Table 2: Token retention ratios for different POS categories