Article Presentation

IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas

Student: Marcus Vinícius Borela de Castro

ColBERTv2 Effective and Efficient Retrieval via Lightweight Late Interaction

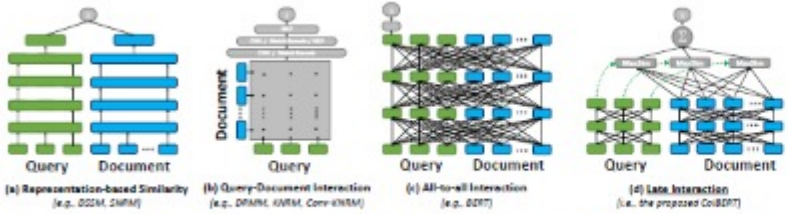


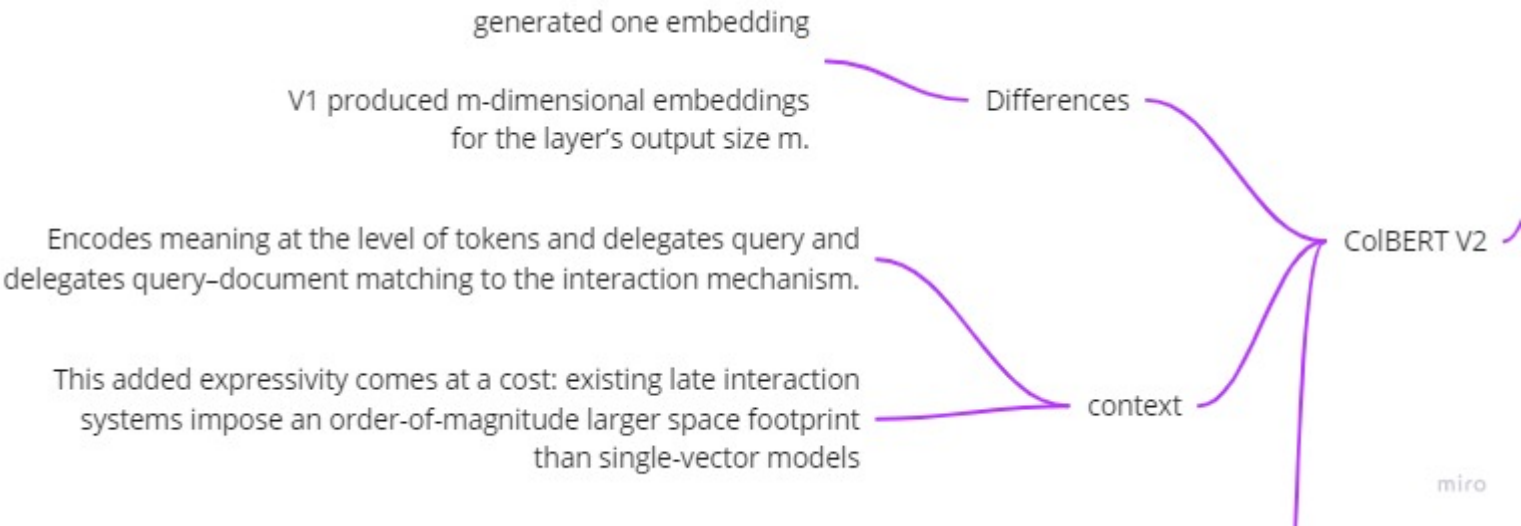
Figure 2: Schematic diagrams illustrating query–document matching paradigms in neural IR. The figure contrasts existing approaches (sub-figures (a), (b), and (c)) with the proposed late interaction paradigm (sub-figure (d)).

ColBERT
a ranking model based on
COntextualized Late interaction
over BERT (figure 2.d)
remembering

Main concepts

ColBERT introduces a late interaction architecture (as a paradigm for e□cient and e□ective neural ranking) that independently encodes the query and the document using BERT and then employs a cheap yet powerful interaction step that models their one-grained similarity.

Results show that ColBERT's efectiveness is competitive with existing BERT-based models (and outperforms every non-BERT baseline), while executing two orders-of-magnitude faster and requiring four orders-of-magnitude fewer FLOPs per query


generated one embedding
V1 produced m-dimensional embeddings for the layer's output size m.
Differences
Encodes meaning at the level of tokens and delegates query and delegates query–document matching to the interaction mechanism.
This added expressivity comes at a cost: existing late interaction systems impose an order-of-magnitude larger space footprint than single-vector models
context
ColBERT V2
miro

proposes

residual compression mechanism
to reduce the space footprint of late interaction
by 6–10 while preserving quality.

- Codebook:
  - i=1 : $y_1 = (0,0)$ $(0-0)^2+(0-1)^2=1$
  - i=2 : $y_2 = (2,1)$
  - i=3 : $y_3 = (1,3)$
  - i=4 : $y_4 = (1,4)$ $(1-0)^2+(4-1)^2=10$

- Signal : 0 1 2 3 2 0
- Transmit to decoder : 1 3 2
- Decoded signal : 0 0 1 3 2 1
- Quantization error : 0 -1 -1 0 0 1

each token in a clusters is represented by its residual r, such that v = Ct + r (an approximation of the real vector)

Uses vector quantization technique (data/image compression technique) (figure in link)

To support fast nearest neighbor search, they group the embedding IDs that correspond to each centroid together, and save this inverted list to disk

27 000 dif tokens

different meaning indicate different clusters

MSMarco

into C=260 000 clusters (2**18 centroids)
90% of clusters have <= 16 distinct tokens

Article contribution

- Introduces LoTTE, a new resource for out-of-domain evaluation of retrievers. LoTTE focuses on natural information-seeking queries over **long-tail topics**,
- Establishes SOTA retrieval quality both within and outside its training domain with a competitive space footprint with typical single vector models.

Interesting/unexpected results

On 22 of 28 out-of-domain tests, achieves the highest quality, outperforming the next best retriever by up to 8% relative gain

The residual compression approach (§3.3) preserves approximately the same quality as the uncompressed embeddings even in test with other pipelines (Colbert, Colbert QA and HoVer)

Basic doubts that may arise

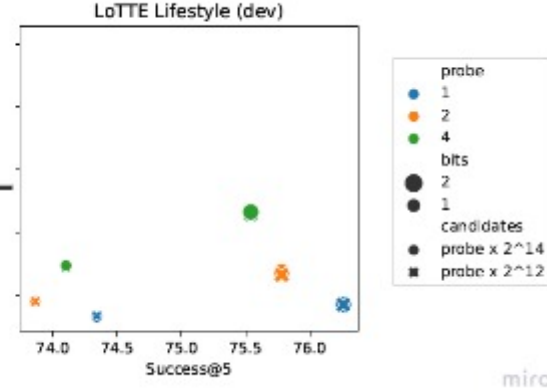
How b of just 1 or 2 bits may represent the residual?

R.: It is a known technique "Vector Quantization" used form image compression

b is the quantization error. If b = 2, I imagine:
bit1: +/-
bit2: 0/1 (values: k and m)
What if b=1?
Does it assume a constant K and keeps just the signal?

What means the probe factor in Figure 3?

"searching by probing the nearest 1, 2, or 4 centroids to each query vector"
"performance for larger probe values tends to require scoring a larger number of candidates."



Advanced topic to discuss

According to these sentences:
We use a 22M-parameter MiniLM cross encoder trained with distillation
(...)
We use a KL-Divergence loss to distill the cross-encoder's scores into the ColBERT architecture.

What means distillation?

Is the meanings of distillation and "denoised" correlated?