ColBERT-Att: Late-Interaction Meets Attention for Enhanced Retrieval

Anonymous ACL submission

Abstract

Vector embeddings from pre-trained language models form a core component in Neural Information Retrieval systems across a multitude of knowledge extraction tasks. The paradigm of *late interaction*, introduced in ColBERT, demonstrates high accuracy along with runtime efficiency. However, the current formulation fails to take into account the attention weights of query and document terms, which intuitively capture the "importance" of similarities between them, that might lead to a better understanding of relevance between the queries and documents. This work proposes ColBERT-Att, to explicitly integrate *attention mechanism* into the late interaction framework for enhanced retrieval performance. Empirical evaluation of ColBERT-Att depicts improvements in recall accuracy on MS-MARCO as well as on a wide range of BEIR and LoTTE benchmark datasets.

1 Introduction & Background

Semantic similarity or relevance between queries and documents using high-dimensional dense vector representations of texts, from large language models, has become ubiquitous in Information Retrieval (IR) and also forms a core component in Retrieval Augmented Generation (RAG) (Lewis et al., 2020). Retrieval of relevant documents or information has transitioned from lexical and text matching (e.g., BM25 (Robertson et al., 1995)) to semantic retrieval based on neural models (e.g., ColBERT (Khattab and Zaharia, 2020)) capturing the semantics and context of information contents.

Traditional systems like BM25 using simplistic text matching (e.g., TF-IDF (Spärck Jones, 1972; Salton and Buckley, 1988) measure) based on *sparse encoding* and corpora statistics, failed to capture similarities beyond surface form equivalence. SPLADE (Formal et al., 2021) provides an explicit sparsity regularization and a log-saturation effect for obtaining term weights, eliminating the need of statistics and hyper-parameters. However, it suffers from language dependency along with expensive memory and compute requirements. BM42 (Vasnetsov, 2024) aims to combine the strengths of lexical matching and attention mechanism from language models. The use of document token attention weights as a proxy to documents term importance was shown to be effective.

Orthogonally, neural IR methods (Karpukhin et al., 2020; Qu et al., 2021; Zhan et al., 2020; Gao et al., 2021; Luan et al., 2021) allow semantic and contextual matching by encoding queries and documents into *single-vector dense representations* from language models to gauge the relevance between queries and documents. *Late interaction* strategy used in ColBERT, produces *multi-vector representations* at token or subtoken level granularity, and computes fine-grained relevance scores between query and document tokens using efficient and scalable token-level computations.

Specifically, the original ColBERT architecture computes the maximum cosine similarity (i.e., *MaxSim* operation) between a query token and all the document tokens, using the token-level dense representations. Finally, a summation of the maximum similarities over all the query tokens is computed to obtain the final relevance scores between query and documents. Consider, $\mathcal{E}_{q_i}$ and $\mathcal{E}_{d_j}$ to denote the embeddings of the i^{th} token of query $\mathcal{Q}$ and the j^{th} token in document $\mathcal{D}$ respectively. Mathematically, the score is thus computed as,

$$\mathcal{S}_{\mathcal{Q},\,\mathcal{D}} = \sum_{q_i \in \mathcal{Q}} \max_{d_j \in \mathcal{D}} \mathcal{E}_{q_i} \cdot \mathcal{E}_{d_j} \qquad (1)$$

ColBERT has been shown to perform extremely well on retrieval tasks by leveraging the semantic expressiveness of language model embeddings, along with the ability to pre-compute document representations for speeding up query processing latency. To improve upon the accuracy and memory footprint of the late interaction architecture,

vector compression techniques and denoising training strategy were proposed in ColBERTv2 (Santhanam et al., 2022b). The use of centroid interaction and pruning approach were introduced in ColBERTv2$_{\text{PLAID}}$ (Santhanam et al., 2022a), to improve the search latency with comparable performance. Rotary positional embeddings (Su et al., 2024) and advanced activation functions for enhanced retrieval performance were recently incorporated in ModernColBERT (Chaffin, 2025).

Motivation. Observe that the *MaxSim* operation in ColBERT does not explicitly factor in the importance of either the query or document terms. As such, all term matches between query and documents are considered to be of equal importance, which is not necessarily true and might degrade the overall performance. Although, vector embeddings obtained from language models implicitly leverage attention weights, we posit that explicit incorporation of terms importance via attention mechanism within the relevance score computation could be beneficial. As an intuition of how attention weights can enable better understanding of *query-document relevance*, consider the following toy example.

Assume a query and candidate documents as:
$\mathcal{Q}$: `Who is going to study?`
$\mathcal{D}_1$: `Alice is walking to school.`
$\mathcal{D}_2$: `Bob is going to buy apples.`
$\mathcal{D}_3$: `Only studying makes Jack a dull boy.`
Here, we highlight probable key terms (underlined) within the query and documents, which would ideally have higher attention weights from a contextual language model.

Observe, the phrase *"is going to"* in $\mathcal{Q}$ has a high (embedding based cosine) similarity with $\mathcal{D}_2$, since it is present in both the texts. However the attention weights of these terms are relatively low in $\mathcal{Q}$, as they are not important to the overall context and intent of the query. Thus, incorporation of query term attention would effectively and accurately reduce the relevance between $\mathcal{Q}$ and $\mathcal{D}_2$.

Similarly, the term *"studying"* having low attention weight in document $\mathcal{D}_3$ would diminish its relevance to the query, in spite of a high similarity match with the term "study" in $\mathcal{Q}$.

On the other hand, although the terms *"study"* and *"school"* in $\mathcal{Q}$ and $\mathcal{D}_1$ resp. are related (having moderate cosine similarity), the high query and document term attention weights would *boost* the relevance score between the query and document – thus accurately retrieving $\mathcal{D}_1$ for query $\mathcal{Q}$.

Thus, incorporation of query and document attention weights within the late-interaction framework could potentially improve the overall retrieval performance, as proposed here in `ColBERT-Att`.

2 `ColBERT-Att` Training

Consider a query $\mathcal{Q}$ and a document $\mathcal{D}$ to be composed of n and m tokens respectively. Thus, $\mathcal{Q} = \{q_1, q_2, \cdots, q_n\}$ and $\mathcal{D} = \{d_1, d_2, \cdots, d_m\}$. The relevance score of the document to the query ($\mathcal{S}_{\mathcal{Q},\,\mathcal{D}}$) is then computed as,

$$\mathcal{S}_{\mathcal{Q},\,\mathcal{D}} = \sum_{i=1}^{n} e^{\,\mathcal{A}_{q_i}} \cdot \max_{j=1}^{m} (\mathcal{E}_{q_i} \odot \mathcal{E}_{d_j}) \cdot (e^{\,\mathcal{A}_{d_w}})^{\delta} \tag{2}$$

where, $\mathcal{E}$ and $\mathcal{A}$ represents the corresponding vector embeddings and attention weights respectively. Further, $\odot$ denotes the cosine similarity operator, and the document token that depicts the highest similarity to the query token q_i is represented as $d_w = \text{argmax}_{j \in [1,\cdots,m]} (\mathcal{E}_{q_i} \odot \mathcal{E}_{d_j})$. Finally, δ signifies a document length based *attention weight regularizer*, which we discuss later in this section.

In other words, we augment the *MaxSim* operation of Eq. (1) with the corresponding query token attention weight along with the associated document token attention weight that depicts the highest (cosine) similarity to the query token. Since the attention weights are typically small values, we accentuate their values (and relative differences) by taking the exponent. Following the original framework of ColBERT, we consider the queries to comprise 32 tokens and documents to be represented by 300 tokens (including special and mask tokens) for most datasets unless specified otherwise.

To obtain our `ColBERT-Att` model, we trained ColBERTv2$_{\text{PLAID}}$ (obtained from `https://github.com/stanford-futuredata/ColBERT`) with the modified objective of Eq. (2) (with $\delta = 1$) using the official train queries and corresponding positive/negative triples of MS-MARCO dataset. Training was performed for 1M steps with default parameter settings, and the best checkpoint was considered. The document token embeddings and attention weights are computed offline and stored as a pre-processing step, while the query token embeddings and attentions are obtained on-the-fly during inference. Observe, that the attention weights are technically free (in terms of compute), as they are an inherent artifact of the encoding process – thus has *no impact on inference latency*.

	R@50	R@100	R@1K
ColBERTv2$_{\text{PLAID}}$	86.76	91.36	97.58
`ColBERT-Att`	**86.78**	**91.54**	**97.64**

Table 1: Results on **MS-MARCO** Passage Ranking dev set.

Attention Weight Regularizer. It is important to note that document length plays a significant effect on the values of the attention weights, with longer documents demonstrating lower token attention weights compared to shorter ones. Thus, substantial difference between the attention weights encountered during training of `ColBERT-Att` and those during inference would introduce discrepancies and might degrade the overall retrieval performance. In fact, the average document length for MS-MARCO (used for training) is around 55, while the average document lengths range from 10 (for Quora) to 230 (in NFCorpus) across other datasets in the BEIR evaluation benchmark.

To alleviate the above, in the formulation of Eq. (2) we introduce the *attention weight regularizer* (δ), defined as $\delta = \min(1, doc\_len/l)$. We empirically set the document length clipping hyper-parameter $l = 150$, discussed later in Section 3. Effectively, this regularizer aims to scale-down high attention weights (for shorter documents), while keeping the original values for others (using $\min$).

For model training we used 2 NVIDIA A100 GPUs (with 80 GB each), while inference was conducted on an NVIDIA Quadro RTX GPU (16 GB).

3 Empirical Results

We evaluate our proposed approach on a wide variety of retrieval tasks from open-source benchmark datasets. Specifically, we compare the performance across several existing methodologies using the **MS-MARCO** (Nguyen et al., 2016) (dev split), **BEIR** (Thakur et al., 2021) (search and semantic relatedness tasks), and **LoTTE** (Santhanam et al., 2022b) (search and forum) datasets. In terms of evaluation measures, we report *Recall@k*, *nDCG@10*, and *Success@5* respectively for the different benchmarks, as shown in literature.

From Table 1, we observe that `ColBERT-Att` achieves a performance improvement of 0.2% on Recall@100 even for the challenging **MS-MARCO** dataset, wherein baseline methods perform quite high. This constitutes *in-domain* evaluation, as the model has been trained on MS-MARCO, and here we set $\delta = 1$ during inference.

To showcase the efficacy of our framework on *out-of-domain* datasets, we evaluate on **LoTTE**, that focuses on natural search queries on documents with long-tailed topics, unlike open-ended QA of the BEIR dataset. From Table 2, we observe `ColBERT-Att` to consistently outperform the existing approaches on all the datasets with an avg. improvement of $\sim 1\%$ on the Success@5 metric.

For completeness, we also evaluate the different methods on a range of **BEIR** datasets spanning search and semantic retrieval tasks as presented in Table 3. The baseline results reported are from (Santhanam et al., 2022b), while ColBERTv2$_{\text{PLAID}}$ results are obtained by executing the code repository available at `github.com/stanford-futuredata/ColBERT`. Here, we observe `ColBERT-Att` to perform better (on most datasets) than the original ColBERTv2$_{\text{PLAID}}$ model (which uses the MaxSim without any attention weights). In fact, on *ArguAna*, we obtain a significant gain of around 2%. Overall, `ColBERT-Att` is seen here to be comparable to the other baselines.

Observe that ColBERTv2 has been shown to perform better than all existing baselines (including SPLADE) (Santhanam et al., 2022b). Due to the unavailability of the original code of ColBERTv2, our current implementation is based on ColBERTv2$_{\text{PLAID}}$ (which performs slightly worse compared to ColBERTv2) (Santhanam et al., 2022a). Overall, we present that within the current framework, inclusion of attention weights tends to improve the accuracy for different retrieval tasks on multiple benchmark datasets. Incorporation of our objective in *ModernColBERT* framework, an interesting direction of future study, provides hope of achieving state-of-the-art retrieval results.

Ablation Study. Table 4(a) depicts how the inclusion of both the query and document attention weights in the *MaxSim* operation of Eq. (2) provides the best performance for `ColBERT-Att`.

To evaluate the effect of *attention regularizer* (δ) in `ColBERT-Att`, we vary the document length clipping hyper-parameter l and report the observed results in Table 4(b). We observe that this strategy can efficiently handle document length (i.e., attention weight value) mismatches between training and inference – leading to an impressive 5% nDCG@10 improvements on Quora (having $5\times$ lower avg. document length compared to MS-MARCO training data). Overall, `ColBERT-Att` is seen to be quite robust across a wide range of values, and we set $l = 150$ for our experimental setup.

	ColBERT	BM25	ANCE	RocketQAv2	ColBERTv2$_{\text{PLAID}}$	ColBERT-Att
LoTTE Search Test Queries (*Success@5*)						
Lifestyle	80.2	63.8	82.3	82.1	84.3	**84.9**
Science	53.6	32.7	53.6	55.3	56.6	**56.9**
Writing	74.7	60.3	74.4	78.0	79.5	**80.2**
Recreation	68.5	56.5	64.7	72.1	71.6	**72.3**
Technology	61.9	41.8	59.6	63.4	66.1	**67.8**
Weighted Av.	68.82	52.73	72.88	71.42	72.7	**73.5**
LoTTE Forum Test Queries (*Success@5*)						
Lifestyle	73.0	60.6	73.1	73.7	76.7	**77.2**
Science	41.8	37.1	36.5	38.0	46.1	**46.5**
Writing	71.0	64.0	68.8	71.5	75.7	**77.1**
Recreation	65.6	55.4	63.8	65.7	70.6	**70.7**
Technology	48.5	39.4	46.8	47.3	53.2	**54.3**
Weighted Av.	59.94	51.27	57.76	59.20	64.4	**65.1**

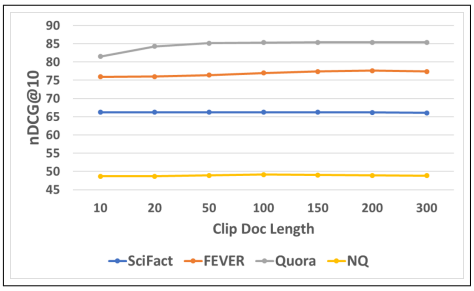
Table 2: Evaluation results on **LoTTE** *Search and Forum* datasets. (Best results are presented in **bold**.)

	ColBERT	DPR-M	ANCE	MoDIR	TAS-B	RocketQAv2	ColBERTv2$_{\text{PLAID}}$	ColBERT-Att
BEIR Search Tasks (*nDCG@10*)								
FiQA	31.7	27.5	29.5	29.6	30.0	30.2	**35.1**	34.8
NFCorpus	30.5	20.8	23.7	24.4	31.9	29.3	33.0	**33.1**
NQ	**52.4**	39.8	44.6	44.2	46.3	50.5	48.8	49.0
HotpotQA	59.3	37.1	45.6	46.2	58.4	53.3	**66.1**	65.9
BEIR Semantic Relatedness Tasks (*nDCG@10*)								
ArguAna	23.3	41.4	41.5	41.8	42.7	**45.1**	42.06	44.3
SciFact	**67.1**	47.8	50.7	50.2	64.3	56.8	67.1	66.2
SCIDOCS	14.5	10.8	12.2	12.4	**14.9**	13.1	14.3	14.5
Quora	85.4	84.2	85.2	**85.6**	83.5	74.9	84.9	85.4
FEVER	77.1	58.9	66.9	68.0	70.0	67.6	76.53	**77.4**
C-FEVER	18.4	17.6	19.8	20.6	**22.8**	18.0	16.7	17.6

Table 3: Evaluation results on **BEIR** *Search and Semantic Relatedness* datasets. (Best results are presented in **bold**, while second-best results are underlined. For *ArguAna*, the query was represented with 300 tokens, as used in the literature.)

	No Attn.	Only $\mathcal{A}_q$	Only $\mathcal{A}_D$	ColBERT-Att
Lifestyle	84.72	84.42	84.87	**84.87**
Science	56.89	57.05	**58.02**	56.89
Writing	79.08	79.65	79.74	**80.21**
Recreation	71.86	72.29	72.19	**72.29**
Technology	66.78	66.94	67.62	**67.79**

(a)



(b)

Table 4: Ablation study for ColBERT-Att: **(a)** Results of Success@5 with different *attention inclusion* on **LoTTE**, and **(b)** Impact of *attention regularizer* (δ) in $\mathcal{A}_D$ with varying document length clipping on nDCG@10 for **BEIR**.

4 Conclusion

This work presented a novel framework, ColBERT-Att, that explicitly integrates the *late interaction* mechanism with attention weights. We show that this incorporation of query and document term importance through attention weights within the *MaxSim* operation along with document length based attention regularizer, provides improved accuracy on diverse retrieval tasks from multiple benchmark datasets.

Limitations

In this work, we showcase that `ColBERT-Att` provides efficient retrieval performance on a variety of benchmark datasets. However, the incorporation of both query and document token attention weights (along with the regularizer) tends to significantly increase the training time of our proposed framework. For instance, the original ColBERTv2 was trained for 400K steps, while we trained for 1M steps. Further, storage of document token attention weights (computed offline) also increases the memory footprint of `ColBERT-Att`. Although model training and attention weight storage is an offline and one-time process, it imposes higher resource requirements.

Additionally, the performance of our model depends on the attention weight regularizer parameter, which might need to be appropriately set for different training and inference datasets and use-cases.

References

Antoine Chaffin. 2025. GTE-ModernColBERT. https://huggingface.co/lightonai/GTE-ModernColBERT-v1.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. *arXiv preprint arXiv:2104.07186*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human-Generated Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. PLAID: An Efficient Engine for Late Interaction Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Karen Spärck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced Transformer with Rotary Position Embedding. *NeuroComputing*, 568:127063.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663*.

Andrey Vasnetsov. 2024. BM42: New Baseline for Hybrid Search. https://qdrant.tech/articles/bm42/.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv preprint arXiv:2006.15498*.