ColBERTv2 in Portuguese

Júlia Tessler and Manoel Veríssimo

July 2023

Abstract

With the exponential growth of digital information, effective retrieval of relevant data has become crucial. ColBERTv2, an efficient single-stage solution, utilizes deep language models and employs late interaction retrievers to enhance retrieval quality while maintaining efficiency. Through distillation and hard-negative mining techniques, ColBERTv2 achieves superior performance, surpassing traditional single vector models. This study explores the effectiveness of ColBERTv2 trained on Portuguese-language data in improving information retrieval quality for Portuguese-language queries and documents, contributing to advancements in information retrieval for this language. The findings provide valuable insights into leveraging ColBERTv2 for efficient and accurate retrieval in Portuguese.

1 Introduction

In the era of information overload, effective retrieval of relevant information from vast document collections has become a critical challenge. To address this challenge, recent advancements in information retrieval have relied on natural language processing (NLP) techniques, albeit with increased latency during inference. Such cost can be reduced using sparse retrievers as a first step of the information retrieval pipeline, then applying a more powerful reranker on the top-k documents.

ColBERT (Contextualized Late Interaction over BERT) [4] has emerged as an efficient single-stage solution for information retrieval, utilizing deep language models such as BERT [2] for ranking. The model architecture independently encodes queries and documents into multi-vector representations and late interaction retrievers improve the quality of multi-vector retrieval models while naturally creating compact token representations that can be easily stored and are highly efficient. Figure 1 shows the overall architecture of ColBERT.

Building upon ColBERT, ColBERTv2 [5] incorporates distillation from a cross-encoder and incorporates hard-negative mining during the creation of the training dataset to enhance quality. Furthermore, it employs a residual compression mechanism to reduce the space footprint of late interaction while preserving vector quality when representing text. These improvements have yielded

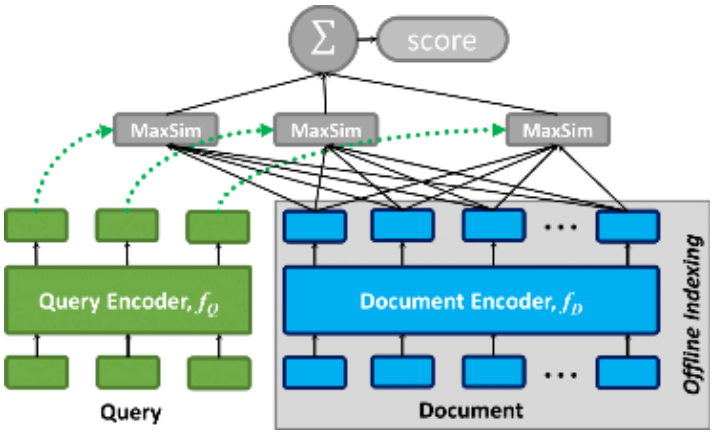

Figure 1: The general architecture of ColBERT given a query q and a document d. Source: [4][5]

state-of-the-art retrieval results within and outside the training domain while maintaining a competitive space footprint compared to typical single vector models.

This report focuses specifically on ColBERTv2 trained and applied to the Portuguese language domain. Portuguese, spoken by millions worldwide and an official language in multiple countries, poses unique challenges for information retrieval. By training ColBERTv2 on a Portuguese dataset, the objective is to explore its effectiveness in improving retrieval accuracy for Portuguese-language queries and documents. The findings of this study shed light on the potential of ColBERTv2 in enhancing retrieval systems for the Portuguese language, thereby contributing to the advancement of information retrieval in this domain.

2 Methodology

The implementation of ColBERTv2 for training on Portuguese data was facilitated by the availability of the complete code provided by Santhanam et al. [5]. The code, which can be accessed on GitHub [1], offered the necessary tools for data preparation, training, and indexing. In the subsequent sections, we present an overview of the key steps involved in training ColBERTv2 specifically for the Portuguese language.

Despite our efforts, we encountered difficulties when attempting to run the code on Google Colab[2]. Although we had sufficient GPU credits and attempted various optimizations, including adjusting runtimes and resource allocation, executing non-interactive code proved challenging. It is worth noting that these

[1] `https://github.com/stanford-futuredata/ColBERT`
[2] `https://colab.research.google.com/`

issues might be resolved by utilizing a Google Colab Pro+ subscription, which we did not explore due to our existing surplus of credits.

Throughout the majority of this project, except for training the model, we utilized an NVIDIA GTX-1070 Ti GPU. For the model training phase, we were fortunate to have access to an NVIDIA V100 GPU provided by CEIA - UFG (*Centro de Excelência em Inteligência Artificial* - Universidade Federal de Goiás)[3].

All of our code and accompanying documentation have been made publicly available on GitHub (`https://github.com/juliatessler/P_IA368DD_2023S1-colbertv2-ptbr`). The code repository includes the necessary dependencies from ColBERT's original codebase and incorporates Docker files specifically created for utilizing CEIA's computational resources.

2.1 Data set & Preparation

In our study, we utilized the mMARCO Portuguese data set (mMARCO-PT) [1] as is common in many information retrieval works. Training ColBERTv2 requires distillation as part of the training data set. While the original author's repository provides code for obtaining distillation scores from an MSMARCO-like triples file and a BERT model, there is no documentation or guidance on its usage, and it was not referenced in other parts of the code. Therefore, we were unable to complete this step using the provided code. However, we would like to acknowledge the helpfulness of the code owners in addressing our queries on GitHub.

To generate the distillation scores, we employed Pyserini's[4] index and BM25 implementations. We utilized Pyserini's functionalities to retrieve the top 200 documents for each query. Subsequently, we reranked these documents using unicamp-dl/mMiniLM-L6-v2-en-pt-msmarco-v2 [1][5][6]

The process of generating distillation scores for the mMARCO-PT training set required approximately 47 hours on the V100 GPU. The resulting file amounted to 4.1GB in size and contained 80,868,810 rows.

2.2 Training the model & Indexing data

Once we obtained the distillation scores, training ColBERTv2 using the provided code and the same set of hyperparameters was a straightforward process. We employed the neuralmind/bert-base-portuguese-cased (BERTimbau Base) [6][7] as the encoder for both documents and queries.

The training of ColBERTv2 required approximately 5 days on the V100 GPU. The code implemented 500,000 training steps and saved checkpoints ev-

[3] `https://ceia.ufg.br/`

[4] `https://github.com/castorini/pyserini`

[5] Available from `https://huggingface.co/unicamp-dl/mMiniLM-L6-v2-en-pt-msmarco-v2`

[6] In retrospect, considering that we exclusively used Portuguese data, it would have been more appropriate to use mMiniLM trained solely on Portuguese rather than the multi-lingual version as we did.

[7] Available from `https://huggingface.co/neuralmind/bert-base-portuguese-cased`

Table 1: Comparision of MRR@10 results on mMARCO-PT dev set for BM25 and several checkpoints of ColBERTv2's trained model

	MRR@10
BM25	0.152
Checkpoint 7,000	0.146
Checkpoint 200,000	0.233
Checkpoint 300,000	**0.238**
Checkpoint 500,000	0.227

Table 2: Comparision of metrics for the best checkpoint and BM25 on mMARCO-PT dev set. Missing values for BM25 weren't reported by [1]

	MRR@10	nDCG@10	nDCG@10	Recall @1000
BM25	0.152	-	-	0.744
Checkpoint 300,000	**0.238**	0.283	0.305	**0.815**

ery 10,000 steps. We performed indexing and evaluation using three different checkpoints: the final checkpoint, as well as those at 300,000 steps, 200,000 steps, and 7,000 steps. The indexing process for mMARCO-PT was completed in approximately 33 hours on the V100 GPU, with memory usage peaking at 40GB. During the retrieval stage, we observed memory peaks of 48GB.

3 Experiments

We conducted an evaluation of multiple checkpoints of the trained model using the mMARCO-PT dev set. For comparison, we used the BM25 baseline, as reported by Bonifácio et al. [1], on the same data set. The results, specifically for the MRR@10 metric, are presented in Table 1. Surprisingly, the latest checkpoint, obtained after training for 500,000 steps, did not yield the best results. The checkpoint that demonstrated the highest performance on the mMARCO-PT dev set was the one obtained after 300,000 steps, achieving an MRR@10 of 0.238.

For the checkpoint 300,000 (the best one amongst the ones we evaluated), some other results can be found in Table 2. For both MRR@10 and Recall@1000, the model beat BM25 on the mMARCO-PT dev set.

As previously mentioned, ColBERTv2 offers the flexibility to be employed as either a single-stage retrieval model or a reranker. In our experimentation, we conducted a test utilizing ColBERTv2 as a reranker, but the obtained results were disappointing. The MRR@10 score was recorded at 0.007. Regrettably, we encountered difficulties in identifying the underlying cause of this poor per-

formance. Further investigation is required to pinpoint the specific issues that contributed to these suboptimal results.

4 Conclusion

In this report, we explored the effectiveness of ColBERTv2, a powerful information retrieval model, trained and applied to the Portuguese language domain. Leveraging the advancements in deep language models, ColBERTv2 demonstrated its potential for improving retrieval accuracy in Portuguese-language queries and documents.

We trained ColBERTv2 on mMARCO-PT with distillation scores generated for this task with Pyserini and the unicamp-dl/mMiniLM-L6-v2-en-pt-msmarco-v2. Our evaluation involved comparing ColBERTv2 against the BM25 baseline on the mMARCO-PT dev set. The results highlighted the effectiveness of ColBERTv2, with the checkpoint obtained after 300,000 training steps demonstrating the best performance, outperforming the latest checkpoint trained for 500,000 steps and the BM25 baseline. However, it is worth noting that when utilizing ColBERTv2 as a reranker, our initial test yielded unsatisfactory results, indicating the presence of certain challenges that require further investigation.

Despite encountering some limitations and challenges along the way, our project can contribute to the field of information retrieval for the Portuguese language domain. It sheds light on the potential of ColBERTv2 as an efficient retrieval model and provides insights into its performance characteristics.

By understanding the strengths, limitations, and areas for improvement of ColBERTv2, we can advance the field of information retrieval and ultimately enhance the retrieval accuracy and user experience for Portuguese-speaking users.

5 Future Work

In terms of future work, addressing the challenges encountered during the application of ColBERTv2 as a reranker is of paramount importance. Further research and analysis are warranted to identify the specific factors that influence its performance and to devise potential solutions.

One of the primary limitations we encountered in this project was the computational bottleneck. Due to limited access to large and expensive GPUs, the generation of data, training, and indexing processes proved to be time-consuming. Consequently, we were unable to evaluate the model on the mRobust data set [3], as originally planned, or explore the performance of different checkpoints of the trained model.

We have observed interesting developments from the other group that also trained ColBERTv2 on Portuguese as part of their final project. Collaborating with them presents an opportunity to enhance our results and, ideally, publish the model and data set together.

Furthermore, it is worth noting that while the original code for ColBERT functions adequately when applied to MSMARCO-like data, it encountered difficulties when handling mRobust data. This suggests that there is room for improvement in the code's robustness to accommodate various information retrieval datasets more effectively.

References

[1] Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of ms marco passage ranking dataset, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. mrobust04: A multilingual version of the trec robust 2004 benchmark, 2022.

[4] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020.

[5] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction, 2022.

[6] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.