A Quantum Annealing-Based Instance Selection Approach for Transformer Fine-Tuning

Andrea Pasin[1],*, Washington Cunha[2], Marcos André Gonçalves[2] and Nicola Ferro[1]

[1]University of Padua, Padua, Italy
[2]Federal University of Minas Gerais, Belo Horizonte, Brazil

Abstract

Currently, *Deep Learning (DL)* is widely used to solve very complex tasks. However, the training of DL models requires huge datasets and long training times. We introduce a novel quantum *Instance Selection (IS)* approach that reduces training dataset sizes by up to 28% while maintaining effectiveness, enhancing training efficiency and scalability. Our method leverages *Quantum Annealing (QA)*, a specific Quantum Computing paradigm, that can address optimization problems. This is the first attempt to tackle the IS problem using QA, and we propose a new *Quadratic Unconstrained Binary Optimization (QUBO)* formulation for it. Extensive experiments with several *Automatic Text Classification (ATC)* datasets show our solution's feasibility and competitiveness with current state-of-the-art IS solutions.

Keywords

Instance Selection, Quantum Annealing, Deep Learning, Transformer

1. Introduction

Deep Learning (DL) models, especially Transformers and *Large Language Models (LLMs)*, excel in tasks like document ranking, semantic labeling, and sentiment analysis. Models like BERT [2], RoBERTa [3], and GPT [4] are able to capture word meanings and contexts. However, they require huge datasets and significant computational resources for both training and fine-tuning. IS, which involves selecting a subset of relevant data points from a dataset to improve training efficiency, can be used to mitigate this issue [5]. This work presents *Balanced Cosine (BCos)* [1], a novel IS approach leveraging QA, a specific *Quantum Computing (QC)* paradigm that is based on special-purpose devices (quantum annealers) able to tackle specific optimization problems. The basic idea of a quantum annealer is to represent a problem as the energy of a physical system and then leverage quantum-mechanical phenomena to let the system find a state of minimal energy, which corresponds to the solution of the original problem.

Our approach formulates the IS problem with a novel QUBO formulation, demonstrating significant training dataset reduction (up to 28%) and speedups (up to 1.35x) while maintaining

IIR2024: 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy

*Corresponding author.

This paper is an extended abstract of [1].

andrea.pasin.1@phd.unipd.it (A. Pasin); washingtoncunha@dcc.ufmg.br (W. Cunha); mgoncalv@dcc.ufmg.br (M. A. Gonçalves); nicola.ferro@unipd.it (N. Ferro)
https://www.dei.unipd.it/~pasinandre/ (A. Pasin); https://www.dei.unipd.it/~ferro/ (N. Ferro)
0009-0007-5193-0741 (A. Pasin); 0000-0002-1988-8412 (W. Cunha); 0000-0002-2075-3363 (M. A. Gonçalves); 0000-0001-9219-6239 (N. Ferro)

model effectiveness. This is the first attempt at using QA for IS. Extensive experiments on various datasets show that BCos is competitive with current *State-Of-The-Art (SOTA)* IS methods such as E2SC, CNN, LSSm, and LSBo. E2SC [6] uses weak classifiers to remove redundant data efficiently, achieving a good balance between effectiveness, training size reduction, and efficiency, with $O(log(n))$ complexity. CNN [7] focuses on hard-to-classify instances with $O(n^3)$ complexity. LSSm and LSBo [8] leverage local sets for IS, both with $O(n^2)$ complexity. Furthermore, we compare our QA approach with *Simulated Annealing (SA)*, a traditional optimization approach that solves the same QUBO problem but using classical hardware. Despite current QC hardware limitations, our experiments reveal the potential of QA in practical optimization problems, showing a glimpse of the future capabilities of quantum technologies.

This work is organized as follows. Section 2 details our approach. Section 3 describes the experimental setup and the results. Section 4 presents conclusions and potential future work.

2. Methodology

Our BCos approach aims to reduce a training set T by a factor $p \in]0, 1[$, resulting in a subset t where $|t| = \lfloor p \cdot |T| \rfloor$. We try to achieve this by removing redundant samples and identifying difficult instances, such as outliers or samples near class boundaries, using a heuristic method. Our QUBO formulation follows the general QUBO framework

$$\min \quad y = x^T Q x \tag{1}$$

where $x \in \{0, 1\}$ represents whether a document should be removed (0) or kept (1), and the matrix Q is defined as follows:

$$Q_{i,j} = \begin{cases} \cos(doc_i, doc_j) & \text{if } lbl_i = lbl_j \text{ and } i \neq j & \text{(2a)} \\ -\cos(doc_i, doc_j) & \text{if } lbl_i \neq lbl_j \text{ and } i \neq j & \text{(2b)} \\ \dfrac{|T[lbl_i]|}{|T|} & \text{if } i = j & \text{(2c)} \end{cases}$$

where doc_i is the vectorized version of the i-th document in the training set T, lbl_i is the label associated with the i-th document in T, $|T[lbl_i]|$ is the number of documents in T having the same label as the i-th document. Furthermore, $\cos(doc_i, doc_j)$ is the cosine similarity between two documents. Each element in equations 2a, 2b, and 2c has its own intuitive interpretation, remembering that we are solving a minimization problem:

- 2a: Highly similar documents within the same class are more likely to be removed from T. They are assigned a **positive** value in the QUBO matrix based on their cosine similarity;
- 2b: Highly similar documents from different classes are more likely to be retained in T. Therefore, we assign them a **negative** value in the QUBO matrix corresponding to their cosine similarity as they likely represent difficult instances near classification boundaries;
- 2c: To avoid worsening class imbalance in T, we penalize the removal of documents from minority classes, which could bias the classification toward majority classes[9].

To complete the QUBO formulation (Equation 1) we added the constraint of selecting a subset of instances of a predefined size. In this case, we need to set opportune constraints that allow us to keep only $\lfloor p \cdot |T| \rfloor$ documents out of $|T|$ documents. This can be achieved as follows:

$$\gamma \cdot (\sum_i a_i \cdot x_i - \lfloor p \cdot |T| \rfloor)^2 = 0 \,, \tag{3}$$

with γ representing a *penalty* sufficiently high. These constraints are summed to the original QUBO formulation. In our case, the percentage of documents that are kept is 75% of the total collection since a reduction of 20-25% has been shown to be a good trade-off between efficiency and effectiveness [6].

QC faces hardware limitations, particularly in the size of problems it can handle. Quantum annealers are restricted by the number of qubits and interconnections available in their *Quantum Processing Unit (QPU)*, limiting them to problems with 100-150 variables. Since the size of a training set is typically in the order of thousands of documents, we need to process them in batches. Thus, we opted for a batch size of $B = 80$ document instances to use most of the capacity of the QPU. We then split the overall problem into $n = \lceil |T|/B \rceil$ batches; note that the last batch might contain less than B document instances, equal to $|T| \mod B$. For each batch, we extract a subset sub_i of the most relevant documents. Their union forms the final subset S of documents that should be fed to the Transformer during the training phase:

$$S = \bigcup_{i=1}^{n} sub_i \,. \tag{4}$$

To minimize the minor embedding (i.e., mapping the problem to the QPU topology) time, we compute the minor embedding twice, as the batch graph structure remains unchanged for each B apart from the last batch that may differ if it has fewer documents.

3. Results

The experiments were executed on an AMD Ryzen 5 5600X Processor with 6-Core and 12 Threads, running at 3.70GHz, 64Gb RAM, and a NVIDIA RTX 3090 24 GB. The quantum annealer employed for the experiments is the D-Wave Advantage ($\approx 5,000$ qubits). The considered transformer is the BERT (base) model. We adopted a 5-fold validation technique for each dataset. E2SC and BCos allow to set fixed reduction rates, thus it was set to 25% as explained in Section 2. However, BCos(QA) often exceeded this, achieving around 28% due to noise and randomness, slightly lowering its effectiveness. The other baselines use automatic reduction criteria. In Table 1, we see that E2SC is the only IS baseline that maintained the effectiveness of the trained BERT model across all datasets. LSBo, LSSm, and CNN experienced losses on at least one dataset and faced scalability issues with large datasets like AG News. Both BCos(SA) and BCos(QA) maintain effectiveness comparable to E2SC and to the full datasets. However, in Table 2 we can notice that BCos(QA) offers better overall efficiency for large datasets than BCos(SA). Table 3 shows the BERT training time according to the considered IS approaches. In particular, it can be seen that by reducing the original datasets' sizes, it is possible to train BERT much more efficiently. Overall, it is possible to see that by keeping a reduction rate around 20-25%, ES2C and BCos represented the best tradeoff between effectiveness and training efficiency.

Table 1
Macro-F1 scores achieved by BERT on different datasets processed using SOTA IS approaches. Elements marked with ● indicate statistical equivalence to the model trained on the original dataset.

Dataset	Original Dataset	E2SC	LSBo	LSSm	CNN	BCos (SA)	BCos (QA)
Vader NYT	80.9(1.1)	81.0(2.0) ●	79.9(2.4) ●	80.5(3.3) ●	79.4(3.1) ●	80.9(0.8) ●	81.7(1.9) ●
Yelp Reviews	95.8(0.3)	96.0(0.7) ●	94.5(0.6) ●	94.7(0.9) ●	94.2(1.4) ●	95.7(0.7) ●	95.5(0.6) ●
WebKB	83.3(2.8)	82.5(2.5) ●	77.1(1.4) ▼	80.1(3.8) ●	80.2(1.8) ●	82.6(1.8) ●	82.5(3.0) ●
20 Newsgroups	82.3(1.1)	80.4(0.7) ●	80.9(0.5) ●	82.2(0.7) ●	77.4(1.6) ▼	80.6(0.5) ●	80.6(0.6) ●
OHSUMED	75.5(0.9)	74.4(0.7) ●	67.3(2.2) ▼	72.2(1.5) ▼	69.8(2.1) ▼	74.2(0.6) ●	73.5(1.6) ●
AG News	91.7(0.2)	91.5(0.2) ●	-	-	-	91.4(0.3) ●	91.4(0.2) ●

Table 2
Time required (seconds) to execute the IS baselines and our BCos approach, considering SA and QA.

Dataset	E2SC	LSBo	LSSm	CNN	BCos (SA)	BCos (QA)
Vader NYT	0.20	18.67	9.89	5.94	13.29	65.78
Yelp Revi.	0.61	16.98	10.36	10.40	13.68	39.38
WebKB	1.07	46.53	29.81	35.75	21.96	94.31
20 NewsG.	3.71	337.07	192.19	159.76	51.40	86.25
OHSUMED	3.50	256.09	161.59	146.27	43.58	92.91
AG News	12.11	-	-	-	295.88	287.33

Table 3
BERT training time (seconds) on the datasets and extracted subsets according to the IS approaches.

Dataset	Original Dataset	E2SC	LSBo	LSSm	CNN	BCos (SA)	BCos (QA)
Vader NYT 2L	204.69	156.56	119.69	202.22	136.14	152.18	143.34
Yelp Reviews 2L	234.11	161.88	76.93	173.07	123.16	166.18	165.96
WebKB	391.20	309.34	138.04	307.58	239.04	306.17	295.97
20 Newsgroups	986.52	745.92	753.28	1002.16	711.30	750.51	723.94
OHSUMED	942.66	774.13	328.42	723.13	517.67	674.02	692.60
AG News	4675.70	3801.80	-	-	-	3773.78	3800.97
Average Speedup	-	1.31x	2.35x	1.18x	1.65x	1.33x	1.35x

4. Conclusions

We proposed a novel IS approach using QA, reducing training datasets while maintaining transformer effectiveness and speeding up fine-tuning. Our method performs comparably to SOTA IS approaches, demonstrating QC's potential. Despite the current limitations of quantum annealers, advancements in technology can improve results and enable larger problem-handling. Future work will explore new QUBO formulations and other models including LLMs.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work is also partially funded by CNPq, CAPES, FAPEMIG, FAPESP, and AWS.

References

[1] A. Pasin, W. Cunha, M. A. Gonçalves, N. Ferro, A quantum annealing instance selection approach for efficient and effective transformer fine-tuning, in: The 10th ACM SIGIR / The 14th International Conference on the Theory of Information Retrieval, 2024.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint 1907.11692 (2019).

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[5] W. Cunha, F. Viegas, C. França, T. Rosa, L. Rocha, M. A. Gonçalves, A comparative survey of instance selection methods applied to nonneural and transformer-based text classification, ACM Computing Surveys (2023).

[6] W. Cunha, C. França, G. Fonseca, L. Rocha, M. A. Gonçalves, An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 665–674.

[7] P. Hart, The condensed nearest neighbor rule (corresp.), IEEE transactions on information theory 14 (1968) 515–516.

[8] E. Leyva, A. González, R. Pérez, Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective, Pattern Recognition 48 (2015) 1523–1537.

[9] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: 2008 Fourth international conference on natural computation, volume 4, IEEE, 2008, pp. 192–201.