The Significance of Computational Performance for Enterprise AI in Question-Answering

Jan Meiswinkel[a, b], Patrick Levi[b] and Rainer Hoch[a]

[a] *Baden-Württemberg Cooperative State University Mannheim, Coblitzallee 1-9, Mannheim, 68163, Germany*
[b] *Schaeffler Group, Industriestraße 1-3, Herzogenaurach, 91074, Germany*

Abstract

Recent publications in Machine Learning follow the trend of trading small quality improvements for huge increases in computational requirements. Most often, the optimization of the computational performance of models is neglected, even though it is a major component in actual value creation – which is the primary factor for investment decisions and adoption of new technologies by enterprises. The small quality improvements of scaling pure ML-based models often do not justify the exponential rise in computing requirements. We argue that this trend hinders the adoption of Enterprise AI. With the example of a highly optimized question-answering service based on BERT, this paper demonstrates the impact of performance and quality on value creation in comparison with each other. The service has been created in early 2020 to lower the effort of company-internal information retrieval from long documents like press statements, reports, publications, handbooks or manuals. As the service's potential value creation lies in its superiority over manual information retrieval, an experiment with 45 participants was carried out in which service and participants had to solve the same question-answering tasks. The service performs 49% faster than the average participant and achieves a consistent F1-score of 100%. The results of the experiment show that in order to increase the service's potential benefits, its computational performance must be increased. This contradicts the widespread search for improved quality. Furthermore, it is shown how model quality and computational efforts affect the added business value of Enterprise AI. A reduced model quality can often bring major performance improvements, exceeding the price of lower quality. The combination of different AI disciplines, like Knowledge Engineering, with Machine Learning bear high potential to become more attractive to companies, reducing the trend of pure scaling in Machine Learning.

Keywords

Enterprise AI, Information Retrieval, Performance, Cost-Benefit, BERT, NLP, Question-Answering

1. Introduction

Over the last few years, researchers of the NLP ("Natural Language Processing") community are competing on pushing the qualitative capabilities of models – precision and recall. With BERT, short for "Bidirectional Encoder Representations from Transformers", Devlin et al. achieved state-of-the-art results on several NLP challenges [1], succeeding ELMo ("Embeddings from Language Models") [2]. While BERT allows remarkably better text understanding, it follows the trend of increasingly expensive models [3]. For SQuAD 1.1 ("Stanford Question Answering Dataset"), the F1-value improvement from 90% to 91.5% increased the hardware burden by factor 10. The past years show a clear trend of exponentially increasing computation requirements: From 2012 to 2019, the processing

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

✉ jan.meiswinkel@schaeffler.com (J. Meiswinkel); patrick.levi@schaeffler.com (P. Levi); hoch@dhbw-mannheim.de (R. Hoch)
0000-0002-1030-4011 (J. Meiswinkel); 0000-0002-5216-4555 (P. Levi); 0000-0002-3839-1583 (R. Hoch)

power used to train and run models rose by factor 10 each year, greatly exceeding the performance gains of hardware [4].

When it comes to developing business applications based on state-of-the-art techniques, the added monetary value is the most decision-relevant indicator. While researchers excel at explaining and demonstrating their technique's advantages on qualitative indicators, it is often unclear at what price in performance they come. With information retrieval, for example, the *correct* information should be retrieved with *minimum* effort. If a user can find an answer faster than the application, a potentially perfect result would yet be worth as much as a wrong answer.

The Schaeffler Group, a global automotive and industrial supplier, developed a BERT-based question-answering service (= Q&A service) to support their employees at retrieving information from documents like press statements, reports, publications, handbooks, or manuals. A dedicated experiment, comparing the service's performance with that of users on question-answering tasks, allows an approximate potential analysis of such services. It is also shown that the benefits of the Q&A service are dependent on underlying text characteristics.

We argue that the trend of rising complexity and often sole scaling of models in Machine Learning greatly hinders practical adoption in businesses. Rather than accepting the increasing cost for small quality improvements, it may be necessary to close the final quality gap in Machine Learning by adopting strengths from other, less compute-intense disciplines of Artificial Intelligence.

2. Related Work

Continuous effort is already being invested into optimizing neural network architectures in order to increase model performance while keeping resource consumption reasonable. For example, the ALBERT model achieves better qualitative results than the original BERT model and yet requires fewer resources by applying several parameter reduction techniques, which leads to reduced training time [9]. In terms of a business software solution that includes a question-answering functionality, usually pre-trained models are used. Training time is therefore of lower importance compared to inference time, despite their usually positive correlation.

Regarding the trend of vastly increasing computation requirements mentioned before, it is evident that most models are primarily optimized for improved quality rather than overall performance. The negligence of reporting details on computational performance, and thus also valuing its significance, has also been observed by Thompson et al. [4], who have examined 1,058 research papers on deep learning for information on computation and hardware burden. Only very few papers included any information on these aspects.

In this paper, we advocate for thinking more value-oriented, rather than forcing quality improvements while ignoring their effects on business value. Furthermore, we argue that in order to allow widespread practical adoption, the whole architecture of a Q&A service must be kept in mind. While the neural network architecture performance remains an important aspect, there are many other crucial factors from software architecture down to hardware resources and elasticity of the infrastructure. They all influence the time required to get an answer without affecting answer quality. Those aspects carry great impact on the final business value. We found that there is hardly any literature focusing particularly on the aspect of business value of Q&A services, which we consider our contribution.

3. Aspects of Developing a BERT-based Question-Answering Service Regarding Performance and Quality

The main purpose of a Q&A service is information retrieval from unstructured documents. The major competitor of a Q&A service is the "CRTL+F"-search implemented in most text editors and browsers. If this remains the quickest way of finding answers while yielding an adequate quality, users are expected to stick to this proven technique. Consequently, any Q&A service must be compared to it. From a user's perspective, there are two important factors: The time it takes to receive an answer and the perceived quality of the answer. Neither of the two may be optimized at the greater expense of the other. A perfect answer that takes a long time is as inacceptable as a fast answer, that is trivial or even wrong, which is also shown in the following equation:

Duration d describes the time required by the Q&A service to find an answer, regardless of its correctness. Whenever an answer is incorrect, the user must again wait for the duration d. This factor is described by Q, depending on the rate at which an answer is incorrect, being described as the probability P of the answer's F1-score being smaller than 100%. The average required time to find the correct answer t is the result of the equation:

$$t = Q * d \tag{1}$$

with

$$Q = 1 + P(F1 < 100\%)$$

No business value is added if t is greater than the average time a user needs to find the answer without the Q&A service. Furthermore, since Q cannot be smaller than 1, t can never be smaller than d. Thus, business value can also not be added if said average time by the user is smaller than d, regardless of the rate at which answers are correct. The equation shows that the high-end quality improvements, coming at exponential increases of duration, cannot justify the marginal gains in quality. Only in cases, where the relative increase of either factor is greater than the relative decrease of the other, the change can be considered an optimization thus worth implementing.

Other important aspects in developing a Q&A service are the infrastructure and resource consumption. While the training of a BERT model requires vast computational power (GPU or TPU clusters), the resulting model can be used for a wide range of different tasks through fine-tuning. The model is trained only once on a dedicated powerful resource [1]. For the deployed service, expensive computing resources can cause immense costs. The basic service infrastructure must run permanently, causing fixed costs, with computational nodes scaling based on demand. To limit the resource consumption of the Q&A service, several measures were taken. First, the service was built upon the smaller BERT model [1] which was fine-tuned for Q&A with SQuAD 1.1 [5]. With the prototypical source code of a BERT-based question-answering service released by Google [6] forming the base of the application, several optimizations were introduced to increase the application's computational efficiency. For example, the loading and compilation of libraries, especially complex ones like TensorFlow, was improved. Using the extended instruction set AVX2 brought major improvements of up to 50%. Furthermore, it has been found that the combination of AVX2 and ML on CPUs eliminates all benefits of hyperthreading – the underlying shared processing unit is almost fully utilized already due to AVX2 and does not leave gaps for hyperthreading. The overhead of the hypervisor thus only reduces the efficiency. A gain of about 20% could be observed by disabling it. The observation that vectorized operations lead to lower hyperthreading efficiency was first made by Saini et al. at NASA [7] but we did not observe a connection to NLP in any papers yet.

4. Experiment: Comparing the Q&A Service with User Performance

4.1.Setup and Execution

Devlin et al. themselves have already proven that BERT-based question-answering services provide answers with excellent quality, for example on SQuAD [1]. For more insights into optimization potentials and comparability of the developed service, an experiment was conducted. The aim was to find out how the Q&A service and users compare in their performance on texts with special linguistic characteristics: Extreme misspelling, high context relevance, academic vocabulary, short texts, and long texts. Inspired by SQuAD 1.1 [5], the task was to select the part of the given article that contains the answer. In total, five questions on five articles had to be answered, each with one of the properties listed above. The duration, the answer itself, and general demographics of the users were recorded.

The service ran in its target environment in a container with 4 cores and 8 GB RAM. For the Q&A service, the average of three runs was taken as the result of each task. The human participants solved the tasks on a website created for the experiment. They were encouraged to use the native search function built into their browser (CTRL+F) and got immediate feedback on their answer quality and speed compared to the service after each task.

Exemplary excerpt from a presented text-question pair [8]:

> *Question:* What is the frequency of the radio station WBT in North Carolina?
> *Text:* In Denver, KOA (850 AM) and KRFX (103.5 FM) will carry the game, with Dave Logan on play-by-play and Ed McCaffrey on color commentary. In North Carolina, WBT (**1110 AM**) will carry the game, with Mick Mixon on play-by-play and Eugene Robinson and Jim Szoke on color commentary. WBT will also simulcast the game on its sister station WBT-FM (99.3 FM), which is based in Chester, South Carolina [...]

4.2. Results

In total, 45 people participated with the average age being 25,7. The BERT-based question-answering service solved the tasks 49% faster than the average participant. In 10% of all records, a user was able to find the correct answer faster than BERT. Figure 1 shows that the service's answer speed varies only about 20% at maximum. The user's answer speed is greatly impacted by the text features with users taking longest on the long and scientific texts.

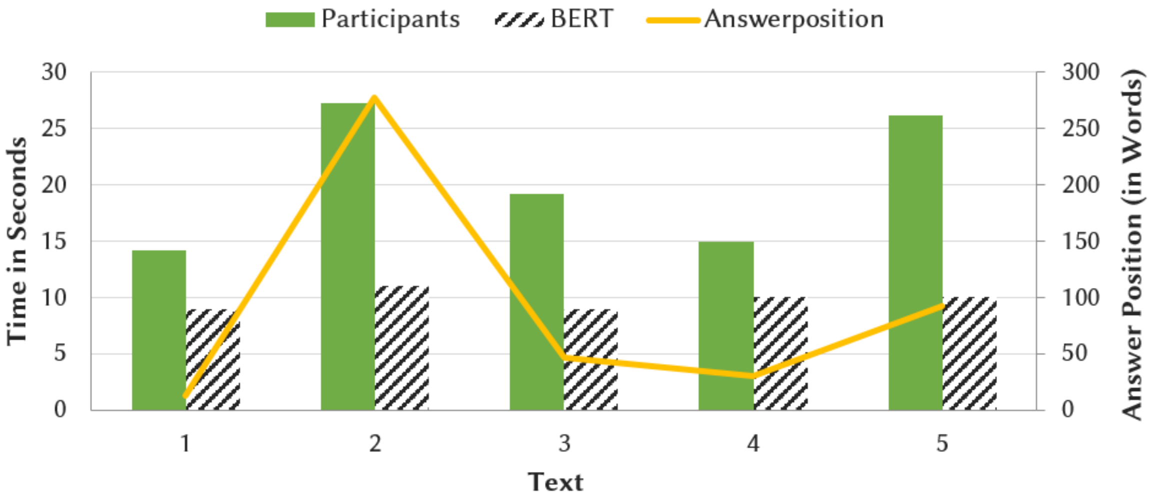

Figure 1: Average duration per text with answer position.

The Q&A service also excels in answer quality. Even in a text full of misspelled words, it scores a F1-value of 100%, as it does on every other task of the experiment. The users' answer quality is significantly worse: With an average F1-value of 64.6%, users struggled to be precise but complete in their answer selection. Most remarkably, 25 of the 45 participants scored an F1-value of 0% on question three – a text that required careful reading, as a radio frequency annotation was searched and the relevant one was third in the list. This shows another advantage of the question-answering service: The complete text is processed, allowing certainty that there are no other potentially relevant passages after the answer, that have not been read yet, which is often the case in manual information retrieval.

It is remarkable that some few users were able to select the correct answer faster than the Q&A service with the perfect F1-value of 100% in every text except for the last. All results are summarized in Table 1.

It is unknown, how much lower the model quality of the question-answering service could be to still achieve adequate quality or even maintain the constant F1-score of 100%. While it was proven, that BERT is able to outperform humans even on a sole CPU, the margins could presumably be much higher if the complexity of the model was reduced. The current margin of a few seconds could already be eliminated by network issues or the phrasing of the question.

Table 1

Result Summary

n = 45	BERT		Human (Average)		Human (Best)	
Text	Time	F1	Time	F1	Time	F1
#1	8,96	100%	14,2	69%	6,4	100%
#2	10,4	100%	27,24	82%	7,31	100%
#3	8,7	100%	19,17	41%	4,67	100%
#4	9,07	100%	14,96	94%	4,13	100%
#5	9,92	100%	26,14	37%	15,33	100%

5. Conclusion

In the experiment, it was proven that the BERT-based question-answering service outperforms the average human participant in quality and speed. Regarding its achieved F1-scores of 100%, it is unknown how much lower the model's quality could be to still yield an F1-value of 100%. Obviously, the model is not perfect – it scores an F1-value of 93.2% on the far more complex SQuAD 1.1 set [1]. Coming back to the perspective of added business value, perfection is not required though. Regarding the increasingly high price of improved answer quality, there are great optimization- and cost-saving potentials which would result in further adoption of AI in businesses.

Assuming the trend of increasing computational requirements in Machine Learning is further followed to close the final quality gap towards 100%, especially Enterprise AI will become scarcer. Methods combining strengths from multiple fields of Artificial Intelligence to increase computational efficiency and reducing costs will, as a result, become more attractive to companies.

6. References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), NAACL, Association for Computational Linguistics, Minneapolis, MN, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[2] M. E. Peters, Neumann, Mark, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), NAACL, Association for Computational Linguistics, New Orleans, LA, 2018, pp. 2227-2237. doi:10.18653/v1/N18-1202.

[3] O. Kovaleva, A. Romanov, A. Rogers and A. Rumshisky, Revealing the Dark Secrets of BERT, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4365-4374. doi:10.18653/v1/D19-1445.

[4] N. C. Thompson, K. Greenewald, K. Lee and G. F. Manso, The Computational Limits of Deep Learning, MIT Initiative on the Digital Economy Research Brief 2020 Vol. 4, MIT Initiative on the Digital Economy, Cambridge, MA, 2020.

[5] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Austin, TX, 2016, pp. 2383-2392. doi:10.18653/v1/D16-1264.

[6] J. Devlin, A. Rao, D. Markowitz, M.-W. Chang and K. Lee, BERT, 2020. URL: https://github.com/google-research/bert.

[7] S. Saini, H. Jin, R. Hood, D. Barker, P. Mehrotra and R. Biswas, The impact of hyper-threading on processor resource utilization in production applications, in: 18th International Conference on High Performance Computing, IEEE, Bengaluru, India, 2011, pp. 1-10, doi:10.1109/HiPC.2011.6152743.

[8] P. Rajpurkar, “SQuAD2.0”, 2020. URL: https://rajpurkar.github.io/SQuAD-explorer/

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 2020.