On the Usability of Transformers-based models for a French Question-Answering task

Oralie Cattan, Christophe Servan, Sophie Rosset

On the Usability of Transformers-based models for a French Question-Answering task - extended abstract

Oralie Cattan[1,2,*,†], Christophe Servan[1,2,†] and Sophie Rosset[2,†]

[1]Paris-Saclay University, CNRS, LISN
[2]QWANT

Abstract

Transformers have sparked a paradigmatic shift in question-answering training practices by simplifying its architectures. As models became larger and better important usability shortcomings appeared. This includes their computational costs and degraded performance with limited training data (e.g., domain-specific or low-resourced language tasks). Considering this resource trade-off, we (i) explore training strategies such as data augmentation, hyperparameter optimization, cross-lingual transfers and cross-dataset mixing, (ii) perform an in-depth analysis to understand the contribution of each on model performance maintenance and (iii) provide a question-answering corpus and a compressed pre-trained model for French[1]. Our experimental results attest to the merit of a flexible paradigm for a low-resource scenario.

Keywords

question answering, transformer architectures, pre-trained models and scalability, language resources

1. Introduction

Question-Answering (QA) consists in extracting an answer given a question and a context document such as Wikipedia articles. Until recently, most of the proposed approaches have relied on an architectural complexification of recurrent neural network integrating increasingly complex attention mechanisms to model the semantic interdependencies between the embedded inputs. The use of pre-trained language models based on the Transformer architecture [1] such as BERT [2] have allowed both to obtain significantly higher performance, but also to remove the recurrence of previous architectures in order to achieve parallelization efficiency.

However, while these methods achieved state-of-the-art results across multiple natural language processing (NLP) tasks, the ever-increasing number of model parameters has raised usability doubts such as their excessively high resource costs (temporal, financial and environmental) [3, 4] and the need for on-device models [5]. For example, speech-related applications have some known problems related to communication and privacy. Providing compact models using compression has therefore become an active research area.

Compact models have been evaluated so far, only on large-scale language understanding tasks [6], on high-resource languages such as English.

Recently, [7, 8] shown that adopted fine-tuning practices are inappropriate under resource-constrained conditions and adversely affect model performance stability. For this purpose, we first establish in sections 2 and 3 a broad state-of-the-art of current research on the usability of Transformers and low-resourced QA. Then, we propose in section 4 to overcome the lack of data by investigating training strategies such as data augmentation, hyperparameters optimization, cross-lingual transfers and cross-dataset mixing on *FQuAD* [9] and *PIAF* [10], two small-scale QA datasets in French. Finally, we release a 12 million (M) parameter compact model for French, which proves to be as competitive as a large French model of 110M parameters pre-trained on the same amount of text. We also provide to the community a high-quality French translated version of the SQuAD corpus [11], the QA reference corpus in English.

2. The question of Transformer usability

Various works have been oriented towards the pre-training of lightweight, responsive and energy-efficient models. To this end, methods based on compression (quantization, pruning or distillation) or architecture optimization have been introduced in order to build compact models with comparable performances to large models.

The idea of quantization [12] is to take advantage of the use of lower precision bit-width floats to reduce memory usage and increase computational density. Following the same objective, pruning [13] consists in removing parts

CIRCLE (Joint Conference of the Information Retrieval Communities in Europe) 2022, July 04–07, 2022, Samatan, France
*Corresponding author.
†These authors contributed equally.
oralie.cattan@lisn.fr (O. Cattan); christophe.servan@lisn.fr (C. Servan); sophie.rosset@lisn.fr (S. Rosset)
0000-0003-2805-5620 (O. Cattan); 0000-0003-2306-7075 (C. Servan); 0000-0002-6865-4989 (S. Rosset)

of a model with minimal precision losses. Finally, knowledge distillation [14] enables the generation of models that mimic the performance of a large model while having fewer parameters. Self-attention is the model core where the most time and memory consuming operations are concentrated. It grows quadratically in respect to the input length. Common approach to this issue consists of approximating the dot-product attention matrix [15] or its factorization [16]. However these solutions have demonstrated they suffer from important computational overheads for tasks with smaller sequence lengths, such as QA.

3. Low-Resource Question Answering

In recent years, low-resource NLP has drawn an increasing amount of attention with solutions ranging from developing new data collection methodologies from crowdsourcing or machine translation (MT) to cross-lingual and transfer learning approaches for which information is shared across languages or tasks. Neural MT as made considerable progress in recent years such as translating large-scale QA datasets from a high-resourced to under-resourced languages [17, 18, 19] or conversely [20] has become an intuitive way of generating annotated datasets in a cost-effective manner. However, since the performance of these approaches depends strongly on the quality of the MT models and due to the lack of reliable models for some language pairs, approaches that foster the transfer of knowledge from other languages or tasks while requiring fewer data have been developed [21, 22, 23]. More recently, efforts focused on pre-training Transformer models multilingually such as mBERT [2] or XLM-R [24] to learn cross-lingual representations. If they are sufficient for well-sourced languages, they remain less efficient for poorly endowed languages [25, 24] which, moreover, often do not have a sufficient amount of annotated data, which also limits the transfer [7, 8].

Concerning French, there are two small QA corpora, and the existing pre-trained language models: CamemBERT [26] and FlauBERT [27] are large. Consequently, in this paper we propose a new compact model FrALBERT and a new QA dataset for French we present in the next section.

4. Methodology

We pre-train a new version of ALBERT [16] from scratch on the French Wikipedia (4 GB of text, 17M of sentences) we called FrALBERT. It is based on parameter sharing/reduction techniques that allow to reduce the computational complexity and speed up the fine-tuning and inference phases.

We use alongside the large monolingual French model CamemBERT [26], the large multilingual models: XLM-R and mBERT pre-trained on 100+ languages, the distilled version of mBERT: distil-mBERT [14] and small-mBERT [28], a mBERT model whose the original vocabulary has been reduced to French. Table 1 gives a comparison of the models.

model	data	vocab.	params.	size
CamemBERT$_{\text{base}}$	OSCAR (138 GB)	32K	110 M	445 MB
CamemBERT$_{\text{large}}$	CCNet (135 GB)	32K	335 M	1.3 GB
CamemBERT$_{\text{base}}$	Wikipedia (4 GB)	32K	110 M	445 MB
FrALBERT$_{\text{base}}$	Wikipedia (4 GB)	32K	12 M	50 MB
XLM-R$_{\text{base}}$	CC-100 (2.5 TB)	250K	278 M	1.1 GB
XLM-R$_{\text{large}}$	CC-100 (2.5 TB)	250K	559 M	1.2 GB
mBERT$_{\text{base}}$	Wiki-100	119K	177 M	714 MB
small-mBERT$_{\text{base}}$	Wiki-100	33K	111 M	447 MB
distil-mBERT$_{\text{base}}$	Wiki-100	119K	134 M	542 MB

Table 1
Characteristics of the pre-trained models.

Our experiments are conducted on *SQuAD* (v1.1) [11] (*SQuAD-en*) comprising of 100K+ English QA pairs, *FQuAD*[1] (v1.0) [9] consisting of 25K+ French pairs, *PIAF* (v1.0) [10] with only 3K+ French pairs; and *SQuAD-fr*$_{\text{train}}$ our translated-to-French version of *SQuAD-en*. To alleviate the data-gathering burden we explore four training strategies with: population-based hyperparameter optimization [29], data augmentation through the use of NMT, cross-lingual transfer ability of multilingual models [2, 24] and cross-dataset mixing using *SQuAD-en* + *FQuAD*. Performances are evaluated using the Exact Match (EM) and F1 scores, 10% of the training data served as validation set.

5. Results

Tables 2 and 3 present the performance results of the monolingual and cross-lingual models respectively. The cross-lingual transfer-based approaches using multilingual models outperform the monolingual approaches. Large models achieve better results than their base or compact version.

Tuning the hyperparameter tends to make models more accurate with gains in terms of EM scores. Highest F1 / EM scores are 90.2 / 75.5 on *FQuAD*$_{\text{dev}}$ and 71.0 / 44.8 on *PIAF*. Data augmentation got nearly the best results in both F1 and EM scores except for the CamemBERT$_{\text{large}}$. The performance gains are up to 11 and 20 of F1 and EM points, respectively, on *FQuAD*$_{\text{dev}}$ and up to 4 points on both metrics on *PIAF*. When no French data is used for training (*SQuAD-en*$_{\text{train}}$ and *SQuAD-en*$_{\text{train}}$ w/ optim), the models confirm the outstanding crosslingual ability with

[1] *FQuAD* test set are not made public, so we use the development set instead

testing data	$FQuAD_{dev}$			$PIAF$		
model \ training strategy	$FQuAD_{train}$	$FQuAD_{train}$ w/ optim.	$FQuAD_{train}$ + $SQuAD\text{-}fr_{train}$	$FQuAD_{train}$	$FQuAD_{train}$ w/ optim.	$FQuAD_{train}$ + $SQuAD\text{-}fr_{train}$
CamemBERT$_{base}$	77.6 / 52.5	85.5 / 70.3	**86.7 / 71.7**	62.0 / 37.5	63.8 / 38.9	**64.3 / 39.2**
CamemBERT$_{large}$	81.2 / 55.9	**90.2 / 75.5**	89.9 / 75.2	68.1 / 42.2	**71.0 / 44.8**	68.9 / 42.5
CamemBERT$_{base}$ (wiki 4 GB)	74.2 / 49.5	80.7 / 61.8	**85.1 / 69.5**	61.7 / 37.3	62.9 / 37.9	**65.9 / 41.0**
FrALBERT$_{base}$ (wiki 4 GB)	72.6 / 55.1	75.6 / 64.8	**84.3 / 70.5**	61.0 / 38.9	62.1 / 39.5	**66.9 / 43.7**
XLM-R$_{base}$	82.1 / 66.8	83.1 / 67.9	**84.2 / 68.8**	65.0 / 39.6	66.9 / 41.2	**68.6 / 42.7**
XLM-R$_{large}$	86.8 / 71.5	**89.5 / 75.8**	87.3 / 72.5	70.4 / 43.8	**73.2 / 45.8**	72.6 / 45.2
mBERT$_{base}$	78.6 / 61.8	82.5 / 65.7	**84.1 / 68.6**	62.5 / 37.8	64.1 / 38.0	**64.8 / 40.0**
small-mBERT$_{base}$	75.1 / 55.7	78.0 / 62.2	**81.6 / 64.6**	60.8 / 35.6	62.2 / 37.7	**63.7 / 39.8**
distil-mBERT$_{base}$	72.8 / 56.0	73.0 / 55.1	**78.1 / 61.5**	52.3 / 30.1	53.6 / 31.4	**58.3 / 34.9**

Table 2
F1/EM performance on the baseline training ($FQuAD_{train}$), using hyperparameter optimization ($FQuAD_{train}$ w/ optim) and with data augmentation ($FQuAD_{train}$ + $SQuAD\text{-}fr_{train}$), on two French QA tasks ($FQuAD_{dev}$ and *PIAF*).

testing data	$FQuAD_{dev}$			$PIAF$		
model \ training strategy	$SQuAD\text{-}en_{train}$	$SQuAD\text{-}en_{train}$ w/ optim.	$SQuAD\text{-}en_{train}$ + $FQuAD_{train}$	$SQuAD\text{-}en_{train}$	$SQuAD\text{-}en_{train}$ w/ optim.	$SQuAD\text{-}en_{train}$ + $FQuAD_{train}$
XLM-R$_{base}$	81.3 / 65.0	82.5 / 66.5	**83.6 / 67.5**	61.4 / 37.2	62.7 / 38.5	**64.9 / 39.9**
XLM-R$_{large}$	82.8 / 64.8	84.4 / 67.8	**87.1 / 72.0**	65.1 / 39.1	66.3 / 40.5	**69.0 / 43.2**
mBERT$_{base}$	76.0 / 59.3	79.5 / 62.3	**83.5 / 67.6**	61.6 / 37.2	62.1 / 36.9	**64.5 / 39.6**
small-mBERT$_{base}$	73.1 / 49.0	76.0 / 59.1	**81.4 / 62.1**	59.6 / 36.5	61.0 / 37.8	**63.0 / 38.9**
distil-mBERT$_{base}$	65.4 / 47.4	68.6 / 48.5	**75.9 / 56.3**	48.8 / 28.1	52.0 / 29.2	**56.5 / 33.1**

Table 3
F1/EM cross-language transfer performances on the baseline ($SQuAD\text{-}en_{train}$), using hyperparameter optimization ($SQuAD\text{-}en_{train}$ w/ optim) and with data augmentation ($SQuAD\text{-}en_{train}$ + $FQuAD_{train}$), on two French QA tasks ($FQuAD_{dev}$ and *PIAF*).

performances that can exceed the performances of the monolingual models in French.

Finally, compact models can reach comparable results to larger ones at a lower cost. Results in Table 4 allow us to gauge the cost/performance trade-off. The F1 performances of the FrALBERT$_{base}$ model are close to those of the CamemBERT$_{base}$ model, both pre-trained on Wikipedia (4 GB). In terms of watt usage, carbon emissions and training time, FrALBERT is two times less the distilled version of BERT. More generally, large models although better by an average of 4 F1 points have a footprint up to 3 times more than their base versions, with longer training times of over 5 hours.

model	params.	size	time	energy	CO_2
CamemBERT$_{base}$	110 M	445 MB	7,2	1.0	317.8
CamemBERT$_{large}$	335 M	1.35 GB	19,4	3.1	914.2
FrALBERT$_{base}$	**12 M**	**50 MB**	**3,8**	**0.5**	**167.8**
XLM-R$_{base}$	278 M	1.1 GB	7,6	1.1	337.7
XLM-R$_{large}$	559 M	1.2 GB	21,1	3.3	973.2
mBERT$_{base}$	177 M	714 MB	7,3	1.0	317.0
small-mBERT$_{base}$	111 M	447 MB	7,1	1.0	321.4
distil-mBERT$_{base}$	134 M	542 MB	6,4	1.0	314.1

Table 4
Comparison of models by computational costs on $FQuAD_{train}$

6. Conclusion and future work

We propose an assessment of the comparative advantage gains in performance when using different training strategies (data augmentation, hyperparameter search and cross-lingual transfer) over monolingual and multilingual pre-trained models, large and compact for a QA task in French under resource constraints.

In this study, we have shown that a number of significant shortcomings of usability have recently been pointed out and that some solutions have been drawn up with compact models. Comparing performances on a French QA task using large and compact models provides insight into the usability of these models when data scarcity problems arise as is the case for under-resourced languages.

Our experimental results suggest that hyperparameter tuning or data augmentation can help to alleviate the data-gathering burden, with performances close to those of a high-resourced language such as English. Finally, compact models provide alternatives to high-energy consumption models by showing comparable performance while reducing size and computational complexity.

In a future work, we aim to employ meta-learning to enhance transferability [30].

Acknowledgments

This work has been partially funded by the French National Research Agency (ANR) through the project Text-ToKids (AAPG2019).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, At-

tention is all you need, in: Advances in Neural Information Processing Systems, 2017.
[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NACL, 2019.
[3] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: ACL, 2019.
[4] N. S. Moosavi, A. Fan, V. Shwartz, G. Glavaš, S. Joty, A. Wang, T. Wolf (Eds.), SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, 2020.
[5] D. Gündüz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, M. van der Schaar, Machine learning in the air, IEEE (2019).
[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, in: EMNLP, 2018.
[7] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, Revisiting few-sample bert fine-tuning, in: ICLR, 2021.
[8] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines, in: ICLR, 2021.
[9] M. d'Hoffschmidt, W. Belblidia, Q. Heinrich, T. Brendlé, M. Vidal, Fquad: French question answering dataset, in: EMNLP, 2020.
[10] R. Keraron, G. Lancrenon, M. Bras, F. Allary, G. Moyse, T. Scialom, E.-P. Soriano-Morales, J. Staiano, Project piaf: Building a native french question-answering dataset, in: LREC, 2020.
[11] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: EMNLP, 2016.
[12] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, K. Keutzer, Q-bert: Hessian based ultra low precision quantization of bert, in: AAAI, 2020.
[13] P. Michel, O. Levy, G. Neubig, Are sixteen heads really better than one?, in: Advances in Neural Information Processing Systems, 2019.
[14] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: Advances in Neural Information Processing Systems, 2019.
[15] N. Kitaev, L. Kaiser, A. Levskaya, Reformer: The efficient transformer, in: ICLR, 2020.
[16] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: ICLR, 2020.
[17] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), AI*IA, Springer International Publishing, 2018.
[18] H. Mozannar, E. Maamary, K. El Hajal, H. Hajj, Neural Arabic question answering, in: WANLP, 2019.
[19] C. P. Carrino, M. R. Costa-jussà, J. A. R. Fonollosa, Automatic spanish translation of squad dataset for multi-lingual question answering, in: LREC, 2020.
[20] A. Asai, A. Eriguchi, K. Hashimoto, Y. Tsuruoka, Multilingual extractive reading comprehension by runtime machine translation (2018). arXiv:1809.03275.
[21] A. M. Dai, Q. V. Le, Semi-supervised sequence learning, in: Advances in Neural Information Processing Systems, 2015.
[22] S. Min, M. Seo, H. Hajishirzi, Question answering through transfer learning from large fine-grained supervision data, in: ACL, 2017.
[23] G. Wiese, D. Weissenborn, M. Neves, Neural domain adaptation for biomedical question answering, in: CoNLL, 2017.
[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: ACL, 2020.
[25] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, in: ACL, 2019.
[26] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, in: ACL, 2020.
[27] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: LREC, 2020.
[28] A. Abdaoui, C. Pradel, G. Sigel, Load what you need: Smaller versions of mutililingual bert, in: SustaiNLP, 2020.
[29] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al., Population based training of neural networks (2017).
[30] O. Cattan, S. Rosset, C. Servan, On the cross-lingual transferability of multilingual prototypical models across nlu tasks, in: MetaNLP, 2021.