

On the importance of pre-training data volume for compact language models

Vincent Micheli

Illuin Technology / EPFL

vincent.micheli@illuin.tech

Martin d’Hoffschildt

Illuin Technology

martin.dhoffschmidt@illuin.tech

François Fleuret

UNIGE

francois.fleuret@unige.ch

Abstract

Recent advances in language modeling have led to computationally intensive and resource-demanding state-of-the-art models. In an effort towards sustainable practices, we study the impact of pre-training data volume on compact language models. Multiple BERT-based models are trained on gradually increasing amounts of French text. Through fine-tuning on the French Question Answering Dataset (FQuAD), we observe that well-performing models are obtained with as little as 100 MB of text. In addition, we show that past critically low amounts of pre-training data, an intermediate pre-training step on the task-specific corpus does not yield substantial improvements.

1 Introduction

Over the past year, pre-trained language models have become the norm in Natural Language Processing. These large-scale Transformer-based (Vaswani et al., 2017) networks considerably advanced the state-of-the-art in language understanding (Devlin et al., 2019) via a two-step process: self-supervised learning on a vast text corpus followed by fine-tuning on a specific downstream task.

Following these advances, the ongoing trend has been to build bigger models with an ever-increasing amount of data (Liu et al., 2019; Raffel et al., 2020; Radford et al., 2019; Brown et al., 2020). However, pre-training models with billions of parameters over hundreds of gigabytes of text requires tremendous computational resources that only a few companies and institutions can afford. Besides, many languages and specific corpora (e.g. legal, scientific) are currently under-resourced. Hence, our goal is to explore model architectures and data volumes lowering the entry barrier to new research and practical applications.

We conduct experiments on French corpora in order to release the first French compact language

models and to illustrate the training process in another language than English. Furthermore, we consider the question answering task since compact models may find their purpose in low-latency/fault-tolerant information retrieval systems.

2 Problem statement

We intend to study the impact of pre-training data volume when training compact bidirectional Transformers (Devlin et al., 2019). We assume a scarce resources setting, both in terms of data and computing power. Two key aspects are explored:

- The amount of pre-training data required to train high-performing compact language models.
- The importance of corpus-specific MLM before fine-tuning.

We use the French part of the OSCAR corpora (Ortiz Suárez et al., 2019) for pre-training and the FQuAD dataset¹ (d’Hoffschildt et al., 2020) for machine reading comprehension fine-tuning. Moreover, the models under consideration are based on the CamemBERT (Martin et al., 2020) language model.

3 Related work

A wealth of work has recently been released (Ganesh et al., 2020) on compressing Transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) through the pre-training of compact models (Turc et al., 2019), distillation (Hinton et al., 2015; Jiao et al., 2019; Sun et al., 2020), pruning (Li et al., 2020; McCarley et al., 2019; Sanh et al., 2020; Fan et al., 2020a) and quantization (Shen et al., 2019;

¹[https://illuin-tech.github.io/
FQuAD-explorer/](https://illuin-tech.github.io/FQuAD-explorer/)

Fan et al., 2020b). Nevertheless, absolute performance is not the end goal of this study. Rather, we investigate the training process of compact models in the absence of larger ones to distillate or prune. Furthermore, Sanh et al. (2020) acknowledge the difficulty of speeding up sparse models due to the absence of specialized hardware. Therefore, from an inference speed standpoint, it is currently preferable to train compact models.

Language models have been successfully pre-trained on domain-specific corpora (Beltagy et al., 2019; Lee et al., 2019) and outperform their general-purpose counterparts on targeted downstream tasks. Still, training these models involved large datasets and computational resources out of reach for most.

Multilingual models (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020) have been released to alleviate the need for language-specific pre-training. While they offer competitive results, they usually lag behind monolingual models and require larger architectures.

Martin et al. (2020) observed that large models did not improve on evaluation tasks when increasing the amount of pre-training data from 4 GB to 138 GB. They left as future work to question the need for large scale pre-training corpora with other model architectures and tasks.

4 Datasets

4.1 OSCAR

OSCAR² (Ortiz Suárez et al., 2019) is a large-scale multilingual open source collection of corpora obtained by language classification and filtering of the Common Crawl corpus³. The whole French part amounts to 138 GB of text and it has already been used to train French language models (Martin et al., 2020). In this work, we only extract a sample of 4 GB of shuffled lines.

4.2 FQuAD

FQuAD (d’Hoffschild et al., 2020) is a recently introduced open source French native reading comprehension dataset. It consists of 60,000 questions and answers gathered on a set of 1,769 high-quality Wikipedia articles. In many aspects, it is the French equivalent of SQuAD 1.1 (Rajpurkar et al., 2016). Given a question and a paragraph, the task consists

²<https://oscar-corpus.com/>

³<https://commoncrawl.org/about/>

Model	Size	Time
CamemBERT _{SMALL}	72 MB	157 ms
CamemBERT _{BASE}	440 MB	705 ms
CamemBERT _{LARGE}	1340 MB	2376 ms

Table 1: Model size and inference time on an Intel Xeon 2.30GHz Quad core CPU with batch size 1 and max sequence length 384 tokens (average over 1000 samples).

in extracting from the paragraph the span of text answering the question.

We chose FQuAD as the fine-tuning dataset because it allows one to draw a direct parallel with its English counterpart (d’Hoffschild et al., 2020) and is one of the largest annotated French datasets. However, question answering is a notoriously difficult task for compact models (McCarley et al., 2019). While distillation has shown to improve their results on the GLUE benchmark (Wang et al., 2018) substantially, machine reading comprehension remains difficult to speed-up without incurring a significant drop in accuracy.

5 CamemBERT_{SMALL}

CamemBERT (Martin et al., 2020) is a multi-layer bidirectional Transformer (Vaswani et al., 2017) with two architectures: base (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters) and large (24 layers, 1024 hidden dimensions, 16 attention heads, 355M parameters). It is very similar to RoBERTa (Liu et al., 2019). The main differences are the use of whole-word masking and SentencePiece tokenization (Kudo and Richardson, 2018) instead of subword-masking and byte-level Byte-Pair encoding (Sennrich et al., 2016; Radford et al., 2019). RoBERTa itself improves upon BERT by aggregating several modifications on top of the original architecture such as removing the next sentence prediction task, dynamic masking, and training with larger batches on more data.

We introduce CamemBERT_{SMALL}⁴, a CamemBERT-based language model with a small architecture (12 layers, 256 hidden dimensions, 4 attention heads, 17M parameters). The main difference with the original CamemBERT lies in the use of subword-masking. Indeed, the authors later found out that whole-word masking had

⁴The pre-trained models are made available in the Hugging Face collection: <https://huggingface.co/illuin/lepetit>.

Hyperparameter	Pre-train	Fine-tune
Train steps	200k	30k
Warmup steps	10k	3k
Batch size	128	32
Learning rate	1e-4	1e-4
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Weight decay	0.01	0.0
Max gradient norm	1.0	1.0
Dropout	0.1	0.1
Mask percent	15	n/a
Max sequence length	512	384

Table 2: Pre-training and fine-tuning hyperparameters. In the corpus-specific MLM step, we take the same hyperparameters as in pre-training except that we decrease the number of steps to 2.5k and drop the warmup.

at best a marginal impact on downstream task performance.

Apart from inference speed and size considerations, two main factors explain this architectural choice:

- This is the same architecture as ELECTRA_{SMALL++} (Clark et al., 2020), a recently released compact language model. Even though ELECTRA and CamemBERT differ in many regards (ELECTRA being trained as a discriminator rather than a generator), prior experiments conducted by Clark et al. (2020) give us an acceptable set of hyperparameters when pre-training and fine-tuning the model.
- Turc et al. (2019)’s empirical results suggest that depth should be prioritized over width when pre-training compact models.

Table 1 shows that CamemBERT_{SMALL} is much smaller and faster than its larger siblings. In a plausible setup for question answering systems, it provides, respectively, a 4.5-fold and 15-fold inference speed-up compared to CamemBERT_{BASE} and CamemBERT_{LARGE} while being 6.2 and 18.8 times smaller.

6 Experiments

Six overlapping subsets are built from the 4 GB OSCAR sample. They are denoted as OSC₁₀, OSC₁₀₀, OSC₅₀₀, OSC₁₀₀₀, OSC₂₀₀₀ and OSC₄₀₀₀ (the numbers indicating the number of MB). We extract an

additional 10 MB sample from the corpus, which serves as a validation set for the self-supervised pre-training task. On the other hand, FQuAD consists of a train/dev split of 50,741 and 5,668 question/context pairs.

For each OSCAR subset, we pre-train a CamemBERT_{SMALL} model with the standard masked language modeling (MLM) objective. Then we fine-tune the pre-trained models on the question answering task with the same span prediction method as BERT (Devlin et al., 2019). Between those two steps, an optional MLM step over the FQuAD train set is included.

Table 2 shows the pre-training, intermediate MLM (if any) and fine-tuning hyperparameters. Fine-tuning being a brittle process (Dodge et al., 2020), fine-tuning results are averaged over 3 seeds.

The experiments described were implemented using Hugging Face’s Transformers library (Wolf et al., 2019) and were conducted on an NVidia V100 16 GB.

7 Analysis

Martin et al. (2020) observed that complex downstream tasks may require more pre-training steps. Since for each OSCAR subset the validation loss is still slowly decreasing after 200k steps, we assume that training longer might increase performance on the difficult question answering task. On the other hand, corpus-specific MLM fine-tuning quickly converged for all models. Table 3 reports the entirety of the results.

7.1 How much data does one need to pre-train a compact language model?

As we increase the amount of pre-training data, perplexity on the OSCAR dev set decreases in every instance but one (OSC₄₀₀₀). Nevertheless, aside from OSC₁₀, discrepancies are small and the models show almost identical learning curves. OSC₁₀ is underperforming in terms of MLM perplexity and question answering F1 score when compared to larger subsets. However, past this smallest dataset, pre-training data volume does not exhibit any strong monotonic relationship with downstream performance. The only OSCAR subset displaying a noticeable performance gap is OSC₂₀₀₀, with a +2.46 average F1 score increase over OSCAR₁₀₀. For anchoring, a randomly initialized CamemBERT_{SMALL} model ”fine-tuned” directly on the FQuAD train set achieves an F1 score

Subset	Perplexities	F1 score	Subset	Perplexities	F1 score
OSC ₁₀	45.20 / 43.34	58.18 (0.60)	OSC ₁₀	40.31 / 18.95	62.33 (0.58)
OSC ₁₀₀	14.22 / 11.91	68.50 (0.25)	OSC ₁₀₀	16.35 / 9.41	69.04 (0.16)
OSC ₅₀₀	12.75 / 10.58	69.50 (0.41)	OSC ₅₀₀	15.09 / 8.77	70.25 (0.43)
OSC ₁₀₀₀	12.56 / 10.57	69.35 (0.64)	OSC ₁₀₀₀	14.74 / 8.83	69.84 (0.27)
OSC ₂₀₀₀	12.45 / 10.41	70.96 (0.66)	OSC ₂₀₀₀	14.72 / 8.75	70.71 (0.08)
OSC ₄₀₀₀	12.49 / 10.35	69.76 (0.61)	OSC ₄₀₀₀	14.90 / 8.68	69.84 (0.79)

(a) Without MLM fine-tuning.

(b) With MLM fine-tuning.

Table 3: Dev OSCAR / FQuAD perplexities and FQuAD F1 score (average token overlap between predicted and ground truth answers) for each pre-training subset.

of only 17.76, i.e. 40 F1 points less than OSC₁₀. This result indicates that even if a small amount of pre-training data is available, one should not neglect that step. Regarding larger architectures, CamemBERT_{BASE} and CamemBERT_{LARGE} models from Martin et al. (2020) obtain an F1 score of 88 and 92, respectively, after fine-tuning.

Due to computational constraints, we could not investigate smaller or larger datasets as well as a prolonged pre-training phase. It could be the case that for a 200k pre-training steps budget, data volume is not the bottleneck. In fact, additional training steps may be even more beneficial for larger datasets. Nonetheless, a preliminary experiment pushing the pre-training phase of CamemBERT_{SMALL} on OSC₂₀₀₀ to 300k steps revealed that while the MLM loss decreased, the F1 score on the downstream task did not improve.

7.2 Is corpus-specific MLM beneficial?

Again, we observe a contrast between OSC₁₀ and larger subsets. OSC₁₀ is the only pre-training dataset significantly improving on the downstream task (+4.15 F1) and experiencing a decrease in perplexity on both pre-training and fine-tuning data when complemented with an intermediate MLM step. However, this corpus-specific MLM step is not truly intermediate since FQuAD contexts contain 10MB of raw text. This implies a 2-fold increase in pre-training data rather than a specific domain adaptation step. Therefore, we turn our focus to larger subsets for the rest of this analysis.

In these cases, MLM fine-tuning results in a net FQuAD perplexity decrease at the cost of an OSCAR perplexity increase. Domain shift may be the root cause of this trade-off. Indeed, as there exists a mismatch between pre-training and fine-tuning sets, the language model has to adapt to the specificity of descriptive paragraphs. In addition,

perplexity is higher on the OSCAR dev set than on the FQuAD one. This is most likely due to the difficulty of predicting masked words in an heterogeneous web-crawled dataset compared to a set of high quality Wikipedia articles.

For every pre-training subset but one (OSC₂₀₀₀), MLM fine-tuning induced a slight F1 score increase on the downstream task. However, these gains are marginal with at most a +0.75 average F1 score increase in the case of OSC₅₀₀. Additional experiments are required to consolidate these findings, especially on larger task-specific datasets such as scientific or legal corpora. In those instances, a greater domain shift would probably legitimate an intermediate MLM fine-tuning step.

8 Conclusion

We investigated the importance of pre-training data volume when training compact Transformer-based models. We made the observation that 100 MB of raw text are sufficient to reach similar performance as with larger datasets on a question answering task, and that corpus-specific self-supervised learning does not bring significant improvements on that particular problem. These preliminary results pave the way for further experiments with other language models, various architectures and new downstream tasks.

Acknowledgments

We gratefully thank Quentin Heinrich for his reviewing and helpful discussions. We also thank Illuin Technology for its technical support and funding.

References

- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [Electra: Pre-training text encoders as discriminators rather than generators](#). In *International Conference on Learning Representations*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Martin d’Hoffschildt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. [Fquad: French question answering dataset](#).
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. [Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping](#).
- Angela Fan, Edouard Grave, and Armand Joulin. 2020a. [Reducing transformer depth on demand with structured dropout](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Angela Fan, Pierre Stock, Benjamin Graham, Edouard Grave, Remi Gribonval, Herve Jegou, and Armand Joulin. 2020b. [Training with quantization noise for extreme model compression](#).
- Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Deming Chen, Marianne Winslett, Hassan Sajjad, and Preslav Nakov. 2020. [Compressing large-scale transformer-based models: A case study on bert](#).
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. [Distilling the knowledge in a neural network](#). In *NIPS Deep Learning and Representation Learning Workshop*.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. [Tinybert: Distilling bert for natural language understanding](#).
- Taku Kudo and John Richardson. 2018. [Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *EMNLP (Demonstration)*, pages 66–71. Association for Computational Linguistics.
- Guillaume Lample and Alexis Conneau. 2019. [Cross-lingual language model pretraining](#).
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. 2020. [Train large, then compress: Rethinking model size for efficient training and inference of transformers](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamel Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- J. S. McCarley, Rishav Chakravarti, and Avirup Sil. 2019. [Structured pruning of a bert-based question answering model](#).
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). Proceedings of the Workshop on

- Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. [Movement pruning: Adaptive sparsity by fine-tuning](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2019. [Q-bert: Hessian based ultra low precision quantization of bert](#). *CoRR*, abs/1909.05840.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. [MobileBERT: a compact task-agnostic BERT for resource-limited devices](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Well-read students learn better: The impact of student initialization on knowledge distillation](#). *CoRR*, abs/1908.08962.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#).