Data-Efficient French Language Modeling with CamemBERTa

Wissam Antoun Benoît Sagot Djamé Seddah
Inria, Paris
{firstname,lastname}@inria.fr

Abstract

Recent advances in NLP have significantly improved the performance of language models on a variety of tasks. While these advances are largely driven by the availability of large amounts of data and computational power, they also benefit from the development of better training methods and architectures. In this paper, we introduce CamemBERTa, a French DeBERTa model that builds upon the DeBERTaV3 architecture and training objective. We evaluate our model's performance on a variety of French downstream tasks and datasets, including question answering, part-of-speech tagging, dependency parsing, named entity recognition, and the FLUE benchmark, and compare against CamemBERT, the state-of-the-art monolingual model for French. Our results show that, given the same amount of training tokens, our model outperforms BERT-based models trained with MLM on most tasks. Furthermore, our new model reaches similar or superior performance on downstream tasks compared to CamemBERT, despite being trained on only 30% of its total number of input tokens. In addition to our experimental results, we also publicly release the weights and code implementation of CamemBERTa, making it the first publicly available DeBERTaV3 model outside of the original paper and the first openly available implementation of a DeBERTaV3 training objective.[1]

1 Introduction

Advances in natural language processing (NLP) have been driven mainly by scaling up the size of pre-trained language models, along with the amount of data and compute required for training (Raffel et al., 2020; Radford et al., 2019; Rae et al., 2021; Fedus et al., 2021; Hoffmann et al., 2022). However, these are not the only factors to determine a model's downstream performance, as the model's architecture and training objective are also important. He et al. (2021b) showed that we can improve a model's performance by using disentangled attention, which uses two vectors to represent a token, one for position and one for content. He et al. (2021a) later showed that performance could be further improved by using ELECTRA's (Clark et al., 2020) self-supervised and sample-efficient replaced token detection objective. Another crucial aspect lies in the ability to train models faster, which allows for quick iteration and thus accelerates the research process and allows for more efficient exploration of new ideas (Izsak et al., 2021; Pan et al., 2022; Geiping and Goldstein, 2022).

This research aims to develop data-efficient and optimized training techniques that can improve performance in downstream tasks, while reducing the required training corpus size and compute. To achieve this goal, we propose a new data-efficient French language model based on DeBERTaV3 (He et al., 2021a). Our proposed model aims to optimize the training process by using a sample-efficient training objective, a state-of-the-art model architecture, and an efficient implementation. We evaluate downstream performance with a variety of NLP tasks, including dependency parsing, part-of-speech tagging, named entity recognition, text classification, and question answering. We compare our model to a BERT model trained with the masked language modeling (MLM) objective using the same tokenizer and training corpus, and to the state-of-the-art French language model, CamemBERT (Martin et al., 2020), which required three times as many training iterations. Our results show that our proposed model reaches or establishes a new state-of-the-art using one third of the computational budget of its main predecessors.

Our contributions can be summarized as follows:

- We propose a new data-efficient French language model, which we train based on our DeBERTaV3 re-implementation with our optimized training recipe.

[1] https://gitlab.inria.fr/almanach/CamemBERTa

- We empirically show that under the same conditions, our model outperforms Transformer models trained with MLM on most tasks, and that it reaches or establishes a new state-of-the-art even when compared with models trained for three times as long.
- Our release is the only publicly available implementation of DeBERTaV3's training objective, and the first for a monolingual DeBERTaV3 model other than the original paper.

Our code and models are available under an open-source license[2], making it easy for researchers to reproduce our results and build upon our work.

2 Related Works

Transformers. This architecture has been widely adopted in NLP tasks such as language modeling, mainly due to the use of the self-attention mechanisms (Vaswani et al., 2017), which allow the model to weigh the importance of different parts of the input when making predictions. A downside of the Transformer block is that it is permutation-invariant, which inhibits the model from encoding word order information. Originally, the authors proposed to add either a fixed sinusoidal pattern or a learned positional embedding as positional bias the input token embedding. Later studies have shown that using relative positional embeddings is more effective (Shaw et al., 2018; Dai et al., 2019; Qu et al., 2021). Recently, He et al. (2021b) proposed a new disentangled attention mechanism, which considers both the relative position and the content of the input tokens as separate vectors.

Pre-trained French Language Models. Current language models available for French are either trained using Masked Language Modeling (MLM) or Causal Language Modeling (CLM). CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) are two of the most popular contemporary French models, both trained with masked language modeling. Other models include FrALBERT (Cattan et al., 2021), a French version of ALBERT (Lan et al., 2020), LePetit (Micheli et al., 2020) which is a small version of CamemBERT, and D'AlemBERT (Gabay et al., 2022), a RoBERTa (Liu et al., 2020) based language model targeted towards Early Modern French. BARThez (Kamal Eddine et al., 2021) is a sequence-to-sequence model trained with BART's objective (Lewis et al., 2020), and PAGnol (Launay et al., 2022) and Cedille (Müller and Laurent, 2022) are models trained with the CLM objective.

To the best of our knowledge, there is no prior effort in developing language models with this improved disentangled attention mechanism and objectives other than MLM/CLM beyond English.

3 CamemBERTa: Methodology

The following section details our proposed architecture and pre-training objective, along with descriptions for the downstream tasks.

Architecture CamemBERTa is based on the DeBERTaV3 (He et al., 2021b) architecture which uses two vectors to encode the word and its position, with the premise being that the relative position of a word pair should also directly affect the computed attention weights. The V3 version optimizes the initial DeBERTa architecture by sharing the relative position embedding projection layers across all the encoder layers, and by adding a convolution layer aside the first encoder layer.[3] We use a base model configuration with 12 layers and 12 attention heads, 768 hidden dimensions with 32k for vocabulary size.

Training Objective We follow the DeBERTaV3 (He et al., 2021a) pretraining strategy by using the replaced token detection (RTD) pre-training loss first introduced in ELECTRA (Clark et al., 2020), with a generator and discriminator based on the DeBERTa architecture. During pre-training we project the generator embeddings to 256 dimensions and keep the generator model at 12 layers.

During pre-training the generator model is trained using the MLM objective where we dynamically mask 15% of the input tokens. We then sample from the generator the masked tokens, and feed the output along with the unmasked tokens to the discriminator which is tasked to identify tokens that were replaced by the generator. The RTD objective increases sample efficiency since the model is predicting over all input tokens instead of the 15% masked tokens.

In DeBERTaV3, the authors hypothesized and showed that sharing token embeddings between the generator and the discriminator results in a tug-of-war situation, where the MLM and RTD tasks

[2]https://gitlab.inria.fr/almanach/CamemBERTa

[3]See Section 5.3 of the DeBERTa paper (He et al., 2021b)

pull the embedding vectors into opposing directions. To alleviate this problem, the authors implemented Gradient-Disentangled Embedding Sharing (GDES), a method that re-parameterize the discriminator's token embeddings as $E_D = sg(E_G) + E_\Delta$, where sg stops the gradient flow from the RTD loss to the generator token embeddings E_G, and hence the loss gradient only updates a Difference Embedding matrix E_Δ that is added to E_G to form the discriminator token embeddings E_D. After pre-training, E_Δ and E_G are summed to get the final E_D and E_Δ is then discarded.

Pre-Training We pre-train on the French subset of CCNet[4] (Wenzek et al., 2020), the same corpus used to pre-train CamemBERT$_{CCNet}$ (Martin et al., 2020).[5] Moreover we reuse CamemBERT$_{CCNet}$'s tokenizer (Kudo and Richardson, 2018). By reusing the pre-training corpus and tokenizer, we isolate the performance differences to the model architecture and training objective variables.

Optimization To speed up the pre-training experiments, we split the pre-training into two phases; in phase 1, the model is trained with a maximum sequence length of 128 tokens for 10,000 steps with 2,000 warm-up steps and a very large batch size of 67,584. In phase 2, maximum sequence length is increased to the full model capacity of 512 tokens for 3,300 steps with 200 warm-up steps and a batch size of 27,648. Because we use very large batch sizes, we optimize the model using the LAMB optimizer (You et al., 2020) with a learning rate of $6e^{-3}$, $\beta_1 = 0.878$, and $\beta_2 = 0.974$.

4 Experiments and Results

Pre-Training Setup We re-implement the DeBERTaV3 RTD pre-training objective with GDES, since no public implementation was available at the time of writing. Our training implementation is based on Nvidia's ELECTRA and BERT TensorFlow2 implementations.[6] We train our models for 8 days on 6 Nvidia A40 with Horovod (Sergeev and Balso, 2018), and make use of XLA compilation, mixed-precision and gradient accumulation to speed-up training and to fit large batch sizes with our limited compute.

During pre-training, our model would have seen 133B tokens compared to 419B tokens for CamemBERT$_{CCNet}$ which was trained for 100K steps. This represents roughly 30% of CamemBERT's full training. Hence for a fair comparison, we train a RoBERTa model, which we dub CamemBERT$_{30\%}$, using our same exact pre-training setup but with the MLM objective.

Downstream Evaluation We compare our models, CamemBERT$_{CCNet}$, and CamemBERT$_{30\%}$, on a diverse set of French downstream tasks and datasets, namely: Question Answering (QA) on FQuAD 1.0 (d'Hoffschmidt et al., 2020), Part-Of-Speech (POS) tagging and Dependency Parsing on GSD (McDonald et al., 2013), Rhapsodie (Lacheret et al., 2014), Sequoia (Candito and Seddah, 2012; Candito et al., 2014) in their UD v2.2 versions and the French Social Media Bank[7] (Seddah et al., 2012), Named Entity Recognition (NER) on the 2008 version of FTB (Abeillé et al., 2000; Candito and Crabbé, 2009) with NER annotation by Sagot et al. (2012), and the FLUE benchmark (Le et al., 2020).

We use the dataset splits as provided by their respective authors, and we finetune using well-tested scripts from the Hugging Face *Transformers* library and the HOPS parser (Grobol and Crabbé, 2021). We only perform hyper-parameter tuning for the NER and QA tasks. See Appendix C for task-specific details. **Bold** text shows the best statistically significant score over 5 seeds.

Question Answering. We evaluate our model on the FQuAD 1.0 dataset (d'Hoffschmidt et al., 2020), which is a SQuAD (Rajpurkar et al., 2016) style French question-answering dataset with 20731 examples for training, and 3188 for evaluation.

The results shown in Table 2 show that our model outperforms CamemBERT$_{30\%}$ by 6.01 F1 points, but shows no statistically significant improvement over CamemBERT$_{CCNet}$ F1 score, and exact match (EM) score.

Part-of-Speech and Dependency Parsing. We report our results on 4 diverse French treebanks. For the parser training, we make use of the HOPS parser (Grobol and Crabbé, 2021) implementation, which is a graph-based dependency parser inspired by Dozat and Manning (2017). Our configuration uses the Transformer model's last layer in addi-

[4] See Appendix 4 for more information on dataset choice.

[5] We go over the pertaining dataset choice in the experiments section.

[6] https://github.com/NVIDIA/DeepLearningExamples/

[7] We follow Riabi et al. (2021) and use their shuffled version of the treebank, which they split into around 2000 sentences for training, and 1000 for each the dev and test sets

Model	GSD		Rhapsodie		Sequoia		FSMB		NER
	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1
CamemBERT$_{30\%}$	98.55±0.05	**94.26**±0.03	**97.61**±0.12	83.19±0.62	**99.32**±0.08	94.09±0.06	94.63±0.11	80.13±0.41	**91.04**±0.76
CamemBERT$_{CCNet}$	**98.57**±0.07	**94.35**±0.15	**97.62**±0.08	**84.29**±0.56	**99.35**±0.09	**94.78**±0.12	**94.80**±0.16	**81.34**±0.63	89.97±0.50
CamemBERTa	**98.55**±0.05	**94.38**±0.15	**97.52**±0.14	**84.23**±0.08	**99.44**±0.07	**94.85**±0.14	**94.80**±0.09	80.74±0.25	**90.33**±0.54

Table 1: **POS tagging**, **dependency parsing** and **NER** results on the test sets of our French datasets. *UPOS (Universal Part-of-Speech) refers here to POS tagging accuracy, and LAS measures the overall accuracy of labeled dependencies in a parsed sentence.*

Model	F1	EM
FrALBERT	72.6*	55.1*
CamemBERT$_{30\%}$	75.14±0.17	56.19±0.27
CamemBERT$_{CCNet}$	**80.98**±0.48	**62.51**±0.54
CamemBERTa	**81.15**±0.38	**62.01**±0.45

Table 2: Question Answering results on FQuAD 1.0.

Model	CLS	PAWS-X	XNLI
FrALBERT	72.17±3.32	76.29±1.28	66.87±0.42
FlauBERT	93.22*	89.49*	80.6*
CamemBERT$_{30\%}$	93.28±0.19	88.94±0.14	79.89±0.64
CamemBERT$_{CCNet}$	94.62±0.04	**91.36**±0.38	**81.95**±0.51
CamemBERTa	**94.92**±0.13	**91.67**±0.17	**82.00**±0.17

Table 3: Text classification results (Accuracy) on the FLUE benchmark. **Results taken from Le et al. (2020).*

tion to FastText embeddings (Bojanowski et al., 2017), character-level bi-directional RNN embeddings, and word embeddings trained during the fine-tuning phase.

Table 1 shows that our proposed model consistently outperforms CamemBERT$_{30\%}$, and competes with CamemBERT$_{CCNet}$ on all 4 treebanks.

Named Entity Recognition is performed on the French Treebank (FTB) which contains 350k tokens in 27k sentences extracted from news articles. Our results in Table 1, surprisingly show that CamemBERT$_{30\%}$ outperforms CamemBERT$_{CCNet}$, while not being statistically better than our model.

FLUE Benchmark We use datasets from the French Language Understanding Evaluation (FLUE) benchmark (Le et al., 2020), namely the French part of the paraphrase identification dataset PAWS-X (Yang et al., 2019), and of XNLI (Conneau et al., 2018), in addition to CLS, a binary classification dataset with Amazon reviews taken from Amazon.

Our results (Table 3) show that our model outperforms all models on the CLS movie classification task, and matches the performance of CamemBERT$_{CCNet}$ on the other FLUE tasks.

Pre-training Dataset Choice We choose CCNet as our pre-training dataset instead of the more common OSCAR dataset (Ortiz Suárez et al., 2019), as (i) it was shown to produce less offensive output (Launay et al., 2022) and (ii) it allowed us to be fully comparable with many of the CamemBERT models (Martin et al., 2020), enabling thus meaningful comparisons. Nevertheless, we also ran experiments with CamemBERT$_{OSCAR}$, and found that it performed slightly worse than CamemBERT$_{CCNet}$, as shown in Table 5 Appendix A.

Pre-training Compute and CO2 Impact Our model was trained for 8 days on 6 A40 GPUs, compared to CamemBERT which was trained on 256 V100 GPUs for one day, which is roughly equivalent to 28 days of training on 6 A40 GPUs, since an NVIDIA A40 GPU is about 1.5x faster than a V100 GPU on language modeling tasks according to recent benchmarks.[8]

Following the reports by Luccioni et al. (2022) and Cattan et al. (2022) on the environmental impact of language model training, we use Lannelongue et al.'s (2021) online carbon footprint calculator to provide the following estimates: CamemBERTa's pre-training used 700kWh and emitted 36kg CO_2 compared to 3.32MWh and 170kg for CamemBERT.[9]

5 Discussion

Our experiments clearly show that given the same training corpus, tokenizer, and total number of examples seen during training, CamemBERTa outperforms the MLM trained CamemBERT model

[8]See https://lambdalabs.com/blog/nvidia-rtx-a40-benchmarks.

[9]These estimates are specific to our training infrastructure situated in France. These estimates highlight the remarkable efficiency achieved by CamemBERTa's pretraining process.

on all tasks except NER on FTB and POS tagging on Rhapsodie. Moreover, our model implementation is able to match or outperform a fully trained CamemBERT model, trained on around 3 times more samples and more compute. The strong performance of our model on higher level FLUE tasks suggest that lower level tasks such as POS tagging and dependency parsing are less challenging for current generation models, since they mostly require surface level information which the model can capture early in the training process, as suggested by Martin et al. (2020), compared to tasks such as question answering and text classification which require more complex processing.

Taking a step back and looking at the only DeBERTa model that includes French, mDeBERTa (He et al., 2021a) we can see (cf. Table 4) that our model only requires 6.6% of its multilingual counterpart training samples to achieve competitive performance while additionally also outperforming the XLM-R model (Conneau et al., 2020) trained on a much larger training sample size.

	XNLI	Steps	# tokens†	Size‡
mDeBERTa*	84.4	500k	2T	0.295T
CAMEMBERTA	82.0	33k††	0.139T	0.032T
XLM-R**	81.4	1.5M	6T	0.295T
C.BERT$_{CCNet}$	81.95	100k	0.419T	0.032T

Table 4: Comparison of XNLI results for different pretraining settings. ††step count was converted assuming 8k batch size. †the total number of tokens seen during training. ‡Total dataset size in tokens. *He et al. (2021a), **Conneau et al. (2020).

This confirms the interest in using such training paradigms in compute limited scenarios for semantically demanding tasks such as question-answering or natural-language inference.

Last but not least, other competitive language models for French are available and although not the primary focus of this paper, we conducted a comparative analysis involving FlauBERT (Le et al., 2020) and FrALBERT (Cattan et al., 2021). The results, presented in Table 5 in Appendix A, demonstrate the better performance of our model across all evaluated tasks in comparison to these French models. Additionally, it is worth noting that FlauBERT was trained for 17 days with 32 V100 GPUs, which is equivalent to 60 days of training on 6 A40 GPUs. This represents a 7.5-fold increase in computational resources employed compared to CAMEMBERTA.

6 Conclusion

We presented CAMEMBERTA, a data-efficient French language model trained on a large corpus of French text and the first publicly available DeBERTaV3-style pretrained model and implementation. For a fair evaluation we reused the same corpus and tokenizer as CamemBERT$_{CCNet}$, but using only 30% of the total number of input training tokens. We compared the performance of both models in addition to an MLM model trained from scratch under the same setup as CAMEMBERTA, CamemBERT$_{30\%}$, on a variety of downstream tasks. Our experiments showed that our model outperforms CamemBERT$_{30\%}$ on all tasks except NER on FTB, and that it is able to match and even surpass CamemBERT$_{CCNet}$. Furthermore, we have also made our optimized code implementation and pretrained model weights publicly available for others to use.

Limitations

Although our model is more efficient than previous models trained using the MLM objective and the standard transformer architecture, we notice that the models runs around 30% slower. This is due to the disentangled attention mechanism, which is more computationally expensive than the standard attention mechanism. We also note that at the time of writing, the DeBERTaV3 TensorFLow 2 implementation available on HuggingFace's Transformers library (Wolf et al., 2020) experiences heavy slowdowns with TPU backends. Our attempts to solve this issue were unsuccessful, and we were unable to train our model on TPUs.

Ethics Statement

We propose a model trained using DeBERTaV3 style pre-training along with an optimized training implementation, which reduces training computation cost when compared to previous models, and hence greatly reduces the energy cost and environmental impact of language model training. We trained our model using the CCNet dataset, for which we direct the reader to for further discussion on bias and ethical considerations. Our experiments do not include any additional data collection or human annotators. Like other language models trained on massive corpora, there may be potential biases present in the training data, which could affect the output of our models. Therefore, we advise against using these models in production without thorough testing. All our experiments were carried out on clusters with energy sources consisting of nuclear (65–75%), 20% renewable, and the remaining from gas.

Acknowledgements

This work was partly funded by Benoît Sagot's chair in the PRAIRIE institute funded by the French national reseach agency (ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. This work also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021607. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France. Association for Computational Linguistics.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Eric De La Clergerie. 2014. Deep syntax annotation of the sequoia french treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.

Oralie Cattan, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2022. Benchmarking transformers-based models on french spoken language understanding tasks. *arXiv preprint arXiv:2207.09152*.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. On the Usability of Transformers-based models for a French Question-Answering task. In *Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French Question Answering Dataset. *arXiv e-prints*, page arXiv:2002.06071.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Jonas Geiping and Tom Goldstein. 2022. Cramming: Training a language model on a single gpu in one day.

Loïc Grobol and Benoît Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés. In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, pages 2675–2689, Berlin, Germany.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Julien Launay, E.l. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. PAGnol: An extra-large French generative model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé 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.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive french language 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.

Rui Pan, Shizhe Diao, Jianlin Chen, and Tong Zhang. 2022. Extremebert: A toolkit for accelerating pretraining of customized bert.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Anlin Qu, Jianwei Niu, and Shasha Mo. 2021. Explore better relative position embeddings from encoding perspective for transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2989–2997, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

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.

Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021. Can character-based language models improve downstream task performances in low-resource and noisy language scenarios? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 423–436, Online. Association for Computational Linguistics.

Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the Paris 7 French TreeBank) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 535–542, Grenoble, France. ATALA/AFCP.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Alexander Sergeev and Mike Del Balso. 2018. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*.

Appendix

A Experiments Results on OSCAR and Dropout

Model	UPOS	LAS	NER	CLS	PAWS-X	XNLI	F1_{FQuAD}	EM_{FQuAD}
FrALBERT	93.53	78.89	69.83	72.17	76.29	66.87	72.6*	55.1*
FlauBERT	97.51	87.92	-	93.22*	89.49*	80.6*	-	-
CamemBERT_{OSCAR}	97.50	88.24	88.19	94.61	90.87	81.38	79.92	61.15
CamemBERT_{CCNet}	<u>97.59</u>	<u>88.69</u>	<u>89.97</u>	<u>94.62</u>	<u>91.36</u>	<u>81.95</u>	<u>80.98</u>	<u>62.51</u>
CAMEMBERTA	<u>97.57</u>	88.55	<u>90.33</u>	<u>94.92</u>	<u>91.67</u>	<u>82.00</u>	<u>81.15</u>	<u>62.01</u>
$\text{CAMEMBERTA}_{dropout}$	97.56	<u>88.57</u>	90.03	94.46	91.42	81.91	79.37	60.29

Table 5: Comparison results of CamemBERT_{OSCAR} and CamemBERT_{CCNet}, and our model CAMEMBERTA, with and without dropout. Due to compatibility issues with FlauBERT's tokenizer, we were unable to conduct FlauBERT testing on FQuAD and NER using standard finetuning scripts. **Results from the models' respective papers Cattan et al. (2021) and (Le et al., 2020).*

B Negative Results

In addition to our main results, we attempted to improve the performance of our model by adding BPE-Dropout (Provilkov et al., 2020) to the tokenization process, as it was shown that this method of subword regularization improves performance on translation tasks. We retrain our model with BPE-Dropout, dubbed $\text{CamemBERTa}_{dropout}$, and compare the results to our original model in Table 5. We observe that by adding BPE-Dropout, we obtain a decrease in performance on most tasks, except for POS tagging and dependency parsing, where the performance does not change.

C Hyper-parameters

Hyper-parameter	Value
Max sequence length	512
Batch size	16
FP16	Enabled
Learning rate	{1.5e-5,2e-5,3e-5}
Epochs	8
Scheduler	linear
Warmup steps	{0,0.1%}
Seed	{1,25,42,666,1337}

Table 6: Hyper-parameters used for the Question Answering and Named Entity Recognition experiments.

For experiments on the FLUE benchmark we use the same hyper-parameters as the authors of CamemBERT on the NLI task. As for POS tagging and dependency parsing, we use the same configurations as the one used in Riabi et al. (2021).