Master Internship offer - Spring 2022
Does fine-tuning make compressed language models racist and sexist?

Antoine Gourru, Christophe Gravier, Charlotte Laclau

Information

General Overview: Does fine-tuning make compressed language models racist and sexist? Pre-trained language models (e.g., BERT) are usually compressed to reduce inference time. These compressed models are then fine-tuned to solve specific tasks. The internship aims at studying the effect of fine tuning on the fairness and bias of compressed model. See details below.

Supervisors: antoine.gourru@univ-st-etienne.fr, christophe.gravier@univ-st-etienne.fr, charlotte.laclau@univ-st-etienne.fr
Localization: Laboratoire Hubert Curien, UMR CRNS 5516, Saint-Étienne, France
Duration: 6 months, between February and August 2022.
Stipend: 573,30 euros / month[1]

Expected profile: Master or engineering degree in Computer Science or Applied Mathematics related to machine learning/natural language processing. The candidate should have a strong scientific background with good technical skills in programming, and be fluent in reading and writing English.

How to apply? Send a CV, a motivation letter and Master records to:
`antoine.gourru@univ-st-etienne.fr` and `christophe.gravier@univ-st-etienne.fr`.
Recommendation letters would be appreciated. Interviews will be conducted as they arise and the position will be filled as soon as possible – the latest application date is set to 15th January.

Important remark This internship is within the framework of a new ANR project between Laboratoire Hubert Curien, Laboratoire ERIC from U. Lyon 2, and NaverLabs Europe. Two PhD positions are funded by the project, and we therefore hire an internship student with the strong possibility to continue as a PhD student in the framework of the project as the PhD funding is already secured from 2022 up to 2025. That possibility obviously depends on how the internship goes – from the perspectives of both sides!

[1]Standard internship stipend in France – Computed on Government Website: `https://www.service-public.fr/simulateur/calcul/gratification-stagiaire`, new law to be published that should make it higher in 2022 but actual figures are yet unknown.

Context

The transformer architecture [25] has led to very large language models such as BERT [6] or RoBERTa [16], which are able to solve wide range of text analysis applications (classification [23], sentiment analysis, grammar checker, spam detection, etc.) but also generation tasks such as: Information Retrieval (for instance using the French ColBERT model [12]), Machine Translation [4], Question Answering [26], Chatbots [15] or Machine Comprehension [20] – to name a few. It is the building block of many NLP contributions nowadays (the "transformer" paper is cited $28,403$ as of September 2021!). Indeed, while training BERT is very expensive (dozens thousands of euros), it shall only be made once, and then the model can be fine-tuned cheaply for the NLP downstream task at hand. BERT [27] takes its name as a follow-up of the previous contextualized embedding model named ELMO [21] (another character from the Sesame street show). The BERT architecture has been refined several times, each time a follow-up improved architecture was also mischievously coined as a character from the Muppet show: ROSITA [18], KERMIT [2], ERNIE [28], ALBERT [14], RoBERTa [16] (henceforth referred as "Muppet models"). All these models are transformer-based, with specificities, and we will conveniently refer to them as the "Muppet models". The beauty of the Muppet models is that they are not reserved to GAFAMs, but one can **easily download them and fine-tune them to its own task with as few as 10 lines of open source code**[2].

BERT represents 110 million parameters, BERT Large represents 340 million parameters. Yet, there are ways to improve the applicability of these large neural networks. The weights can be either compressed using standard float quantization techniques [5] (or even recently integer quantization [13]), removed if they have very limited impact in the system using weight pruning techniques [7], or a small model can be trained from scratch as long as we have access to its large counterpart that it will be trained to mimic (this is referred as Knowledge Distillation [8]). Those compressed models will mostly be the ones going into products in our society. A question is therefore how compression affect the fairness of these large models – which is already challenged for the large model versions [1].

In [9, 10], the authors have shown (in the context of image classification) that compressed models impact much more underrepresented features, which often coincide with notion of fairness and to the consideration of an imbalance in the representation of the different categories. [10] demonstrates that compressed models can indeed amplify bias of image classification algorithms. [24] studies the gender bias in pre-trained language models (including compressed models such as DistilBERT and ALBERT). It introduces two level of bias evaluation (skewness vs stereotype bias). One of the findings of this work is that DistilBERT has significantly higher "skew" compared to original BERT. This is inline with [10] findings that compressed model amplifies certain features present in the original model.

Objectives

When using a large or compressed model, it is a common practice to fine-tune that model on the corpus at hand using the masked language modeling task, for instance text classification [23]. This fine-tuning step has called for a large number of studies on how to best perform especially to preserve and hopefully increase the downstream accuracy for the given NLP task, e.g. text classification [23]. In this internship, we are first interested in how the standard step of model fine-tuning affect compressed models fairness. Fine-tuning can be done at two different step of the learning pipeline. First, it can be performed on the large language model, prior its compression. Second, it can also be performed after the model has been compressed (aka fine-tuning the compressed model itself, regardless of the compression technique used either distillation or model pruning. Ultimately, it is worth saying that investigating a double fine-tuning

[2] https://huggingface.co/transformers/quickstart.html – Kudos to HuggingFace to make this psosible.

step combining the two steps where fine-tuning can be applied, has discussed above, can be investigated, as the second fine-tuning could help to handle catastrophic forgetting [3] due to model compression. While this leads to as many fine-tuning opportunities, that is the number of compressed models times four fine-tuning strategies (the three mentioned above and the no fine-tuning case). The question is then to come up with fairness evaluation of these models. A promising approach to profile how a language model encode linguistic information is to use probing tasks [17]. We want to devise a set of fairness related probing tasks, that would allow us to investigate the effect of the different fine-tuning strategies.

The work plan proposed to the student is as follows :

1. Literature review on language model compression, including distillation and prunning.
2. Identify a set of fairness related datasets and metrics that can be used to evaluate compressed models. A popular dataset has already been used in related studies in the frame of our project using the StereoSet dataset [19].
3. Devise and implement fine-tuning strategies for these datasets using common compressed language models such as DistilBert [22] and TinyBERT [11]. Lab. The intern will have acces to the Hubert Curien computing cluster.
4. If the internship leads to publish work, support to go present your work in a conference.

Recommendation for applicants

If you want to know more about the direction of this research and this internship, you may consider first reading the following articles[3]:

- On the impact of model compression: Sara Hooker et al. *What Do Compressed Deep Neural Networks Forget?* 2019. arXiv: 1911.05248 [cs.LG]
- On profiling language models using probing tasks: Alessio Miaschi et al. "Linguistic Profiling of a Neural Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 745–756. DOI: 10.18653/v1/2020.coling-main.65
- On quantifying gender bias in pre-trained and fine-tuned language models: Daniel de Vassimon Manela et al. *Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models*. 2021. arXiv: 2101.09688 [cs.CL]

References

[1] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proc. of the 2021 ACM FAccT*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

[2] William Chan et al. *KERMIT: Generative Insertion-Based Modeling for Sequences*. 2019. arXiv: 1906.01604 [cs.CL].

[3] Chen et al. *Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting*. 2020. arXiv: 2004.12651 [cs.CL].

[3] This is for the sake of providing a focus on major related papers to what we want to achieve, and you are not expected to master these papers end-to-end prior making contact with us!

[4] Stephane Clinchant et al. "On the use of BERT for Neural Machine Translation". In: *Proc. of the 3rd Work. on Neural Generation and Translation*. Hong Kong: ACL, Nov. 2019, pp. 108–117. DOI: 10.18653/v1/D19-5611.

[5] William Dally. "High-performance hardware for machine learning". In: *NIPS Tutorial* 2 (2015).

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Jonathan Frankle et al. "The lottery ticket hypothesis: Finding sparse, trainable networks". In: *arXiv:1803.03635* (2018).

[8] Geoffrey Hinton et al. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[9] Sara Hooker et al. *Characterising Bias in Compressed Models*. 2020. arXiv: 2010.03058 [cs.LG].

[10] Sara Hooker et al. *What Do Compressed Deep Neural Networks Forget?* 2019. arXiv: 1911.05248 [cs.LG].

[11] Xiaoqi Jiao et al. "TinyBERT: Distilling BERT for Natural Language Understanding". In: *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*. Ed. by Trevor Cohn et al. Vol. EMNLP 2020. Findings of ACL. Association for Computational Linguistics, 2020, pp. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.372.

[12] Omar Khattab et al. "Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval". In: *CoRR* abs/2101.00436 (2021). arXiv: 2101.00436. URL: https://arxiv.org/abs/2101.00436.

[13] Sehoon Kim et al. *I-BERT: Integer-only BERT Quantization*. 2021. arXiv: 2101.01321 [cs.CL].

[14] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *Proc. of ICLR 2020*. 2020, TBA.

[15] Xiujun Li et al. "End-to-End Task-Completion Neural Dialogue Systems". In: *Proc. of IJCNLP 2017, vol. 1*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 733–743.

[16] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[17] Alessio Miaschi et al. "Linguistic Profiling of a Neural Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 745–756. DOI: 10.18653/v1/2020.coling-main.65.

[18] Phoebe Mulcaire et al. "Polyglot Contextual Representations Improve Crosslingual Transfer". In: *Proc. of NAACL 2019, vol. 1*. Minneapolis, Minnesota: ACL, June 2019, pp. 3912–3918. DOI: 10.18653/v1/N19-1392.

[19] Moin Nadeem et al. "StereoSet: Measuring stereotypical bias in pretrained language models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong et al. Association for Computational Linguistics, 2021, pp. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. URL: https://doi.org/10.18653/v1/2021.acl-long.416.

[20] Yasuhito Ohsugi et al. "A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension". In: *Proc. of 1st Workshop on NLP for Conversational AI*. Florence, Italy: ACL, Aug. 2019, pp. 11–17. DOI: 10.18653/v1/W19-4102.

[21] Matthew E Peters et al. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[22] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. eprint: arXiv:1910.01108.

[23] Chi Sun et al. *How to Fine-Tune BERT for Text Classification?* 2020. arXiv: 1905.05583 [cs.CL].

[24] Daniel de Vassimon Manela et al. *Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models*. 2021. arXiv: 2101.09688 [cs.CL].

[25] Ashish Vaswani et al. "Attention is all you need". In: *Proc. of NIPS 2017*. 2017, pp. 5998–6008.

[26] Zhiguo Wang et al. "Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering". In: *Proc. of the EMNLP/IJCNLP*. Hong Kong, China: ACL, Nov. 2019, pp. 5878–5882. DOI: 10.18653/v1/D19-1599.

[27] Patrick Xia et al. "Which *BERT? A Survey Organizing Contextualized Encoders". In: *Proc. of EMNLP 2020*. ACL, Nov. 2020, pp. 7516–7533.

[28] Zhengyan Zhang et al. "ERNIE: Enhanced Language Representation with Informative Entities". In: *ACL 2019*. Florence, Italy: ACL, July 2019, pp. 1441–1451. DOI: 10.18653/v1/P19-1139.